Clasificación de futbolistas mediante clustering

Alumno : Nicolas Bousquet

Presentación y objetivos

En el fútbol estamos acostumbrados a clasificar a los jugadores en tres categorías: delanteros, defensas y centrocampistas. Sin embargo, estos términos no significan mucho, ya que dos defensas, centrocampistas o delanteros pueden ser muy diferentes. Entonces, ¿sería posible clasificar a los jugadores de una manera más precisa y objetiva? Este es el objetivo de este estudio, que se basa en el método de clustering.

Los datos

He utilizado una base de datos (ver *"original_data.csv")* que hace referencia a un montón de datos sobre cada jugador que ha jugado en la Liga de Campeones entre las temporadas 2013/2014 y 2018/2019. Hay más de 50 parámetros registrados para cada jugador. Entre ellos se encuentran los básicos (goals, assists ...), pero también datos mucho más avanzados (successfulDribbles, groundDuelsWon, aerialDuelsWon ...).

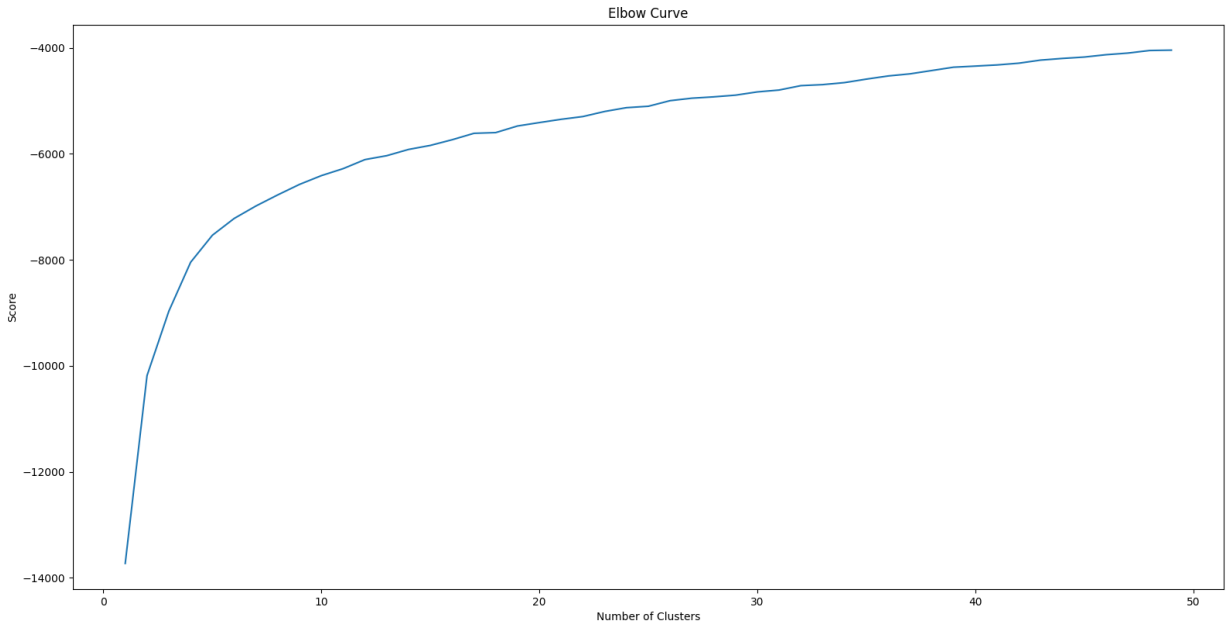
		name	season year	team	rating	goals	bigChancesCreated	bigChancesMissed	assists	
1	0	Robin van Persie	13/14	manchester-united		4				
2	1	Ashley Cole	13/14	chelsea		0			0.0	0
3	2	José Antonio Reyes	15/16	sevilla	7.25	0	0.0	1.0	0.0	0
4	3	Roman Weidenfeller	13/14	borussia-dortmund		0			0.0	0
5	4	Roman Weidenfeller	14/15	borussia-dortmund		0			0.0	0
6	5	Roman Weidenfeller	16/17	borussia-dortmund	6.6	0	0.0	0.0	0.0	0
7	6	Roman Weidenfeller	17/18	borussia-dortmund	0.0	0	0.0	0.0	0.0	0
8	7	Sebastian Kehl	13/14	borussia-dortmund		1			0.0	1
9	8	Sebastian Kehl	14/15	borussia-dortmund		0			1.0	1
10	9	Daniel Braaten	13/14	fc-kobenhavn		1			0.0	1

Como los datos estaban incompletos antes de la temporada 2015/2016, solo he mantenido las temporadas entre 2015/2016 y 2018/2019 (4 temporadas). A continuación, procesé los datos para eliminar todos los posibles sesgos (seleccionando a los jugadores que habían jugado un mínimo de 1000 minutos, dividiendo sus estadísticas por su número de minutos jugados * 90) y para eliminar todos los datos irrelevantes ("penaltyGoals", "substitutionIn/Out" ...). He aquí un extracto de los datos procesados.

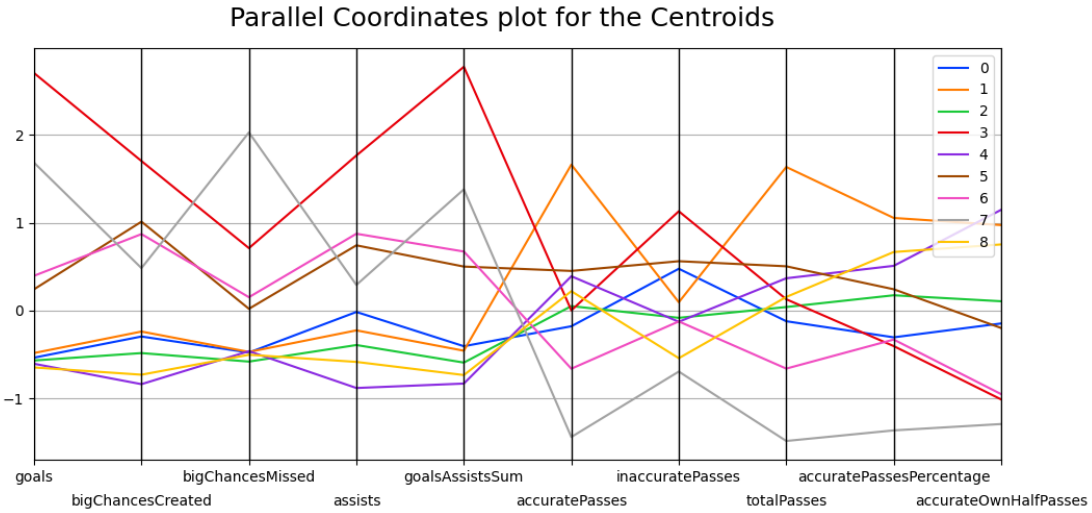
	name	goals	bigChancesCreated	bigChancesMissed	assists	goalsAssistsSum	accuratePasses	inaccuratePasses	
1	Adriano	0.154772141	0.154772141	0.077386071	0	0.154772141	37.60963027	8.744625967	46.
2	Adrien Rabiot	0.196078431	0.049019608	0.049019608	0.147058824	0.343137255	64.85294118	6.323529412	71.
3	Alan Dzagoev	0.202398801	0.067466267	0.067466267	0	0.202398801	34.67766117	9.445277361	44.
4	Alejandro Grimaldo	0.147420147	0.073710074	0	0	0.147420147	42.6044226	6.412776413	49.
5	Aleksandar Kolarov	0.033682635	0.134730539	0	0.202095808	0.235778443	41.53068862	10.10479042	51.
6	Aleksandr Golovin	0	0.070532915	0.141065831	0.141065831	0.141065831	36.04231975	7.970219436	44.
7	Alessandro Florenzi	0.0456621	0.091324201	0.091324201	0.0456621	0.091324201	28.35616438	10.3196347	38.
8	Alex Sandro	0	0.06232687	0	0.093490305	0.093490305	39.42174515	7.167590028	46.
9	Alex Telles	0.084626234	0.084626234	0.042313117	0.253878702	0.338504937	29.28067701	9.055007052	38.
10	Alexis Sánchez	0.325105358	0.270921132	0.325105358	0.433473811	0.758579169	30.34316677	8.886213125	39.
11	Allan	0	0	0	0	0	47.57142857	6.928571429	54.
12	Álvaro Morata	0.44813278	0.298755187	0.970954357	0.22406639	0.67219917	19.12033195	6.124481328	25.
13	André Almeida	0	0.052113492	0	0.104226983	0.104226983	40.59640996	9.015634047	49.
14	Andrea Barzagli	0	0.043562439	0	0.043562439	0.043562439	56.84898354	5.532429816	62.

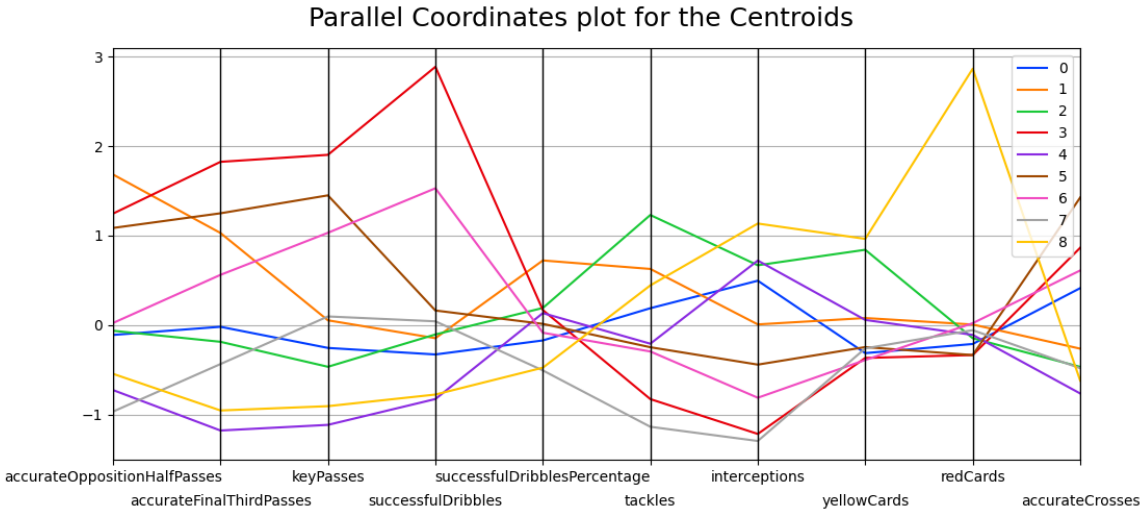
Algoritmo de k-means

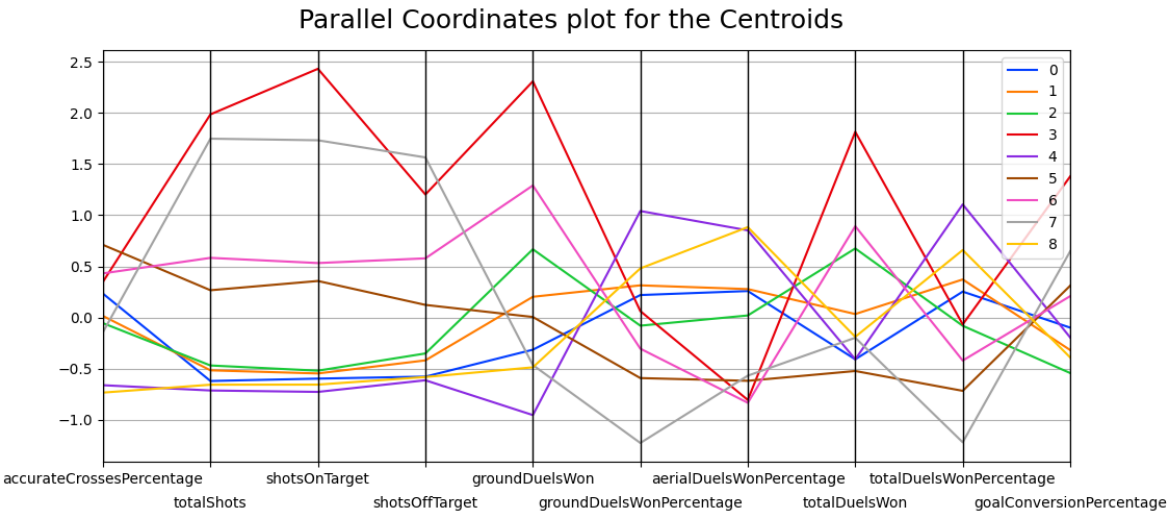
Para elegir el número correcto de clusters, he utilizado el método del "elbow". Aquí podemos ver que el número óptimo de clusters está a 9.

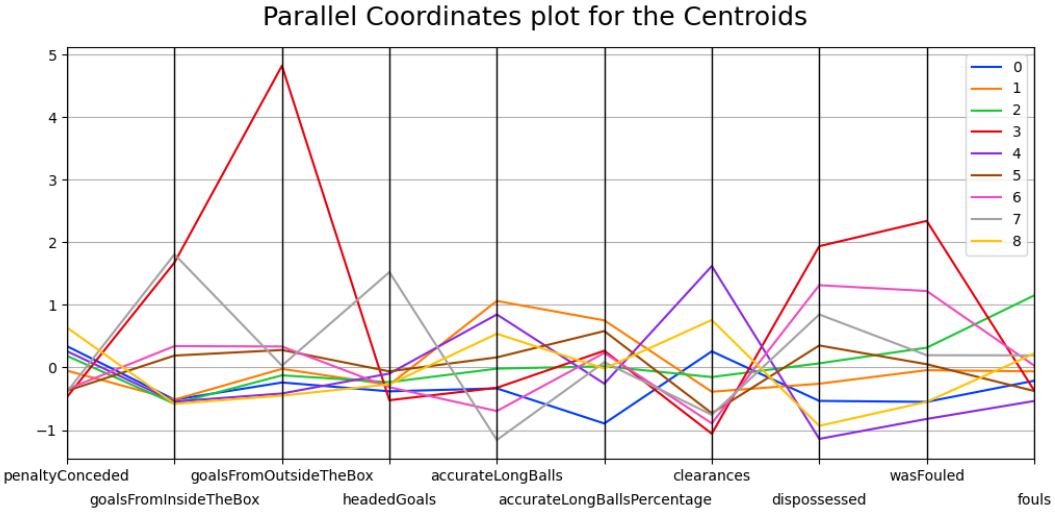


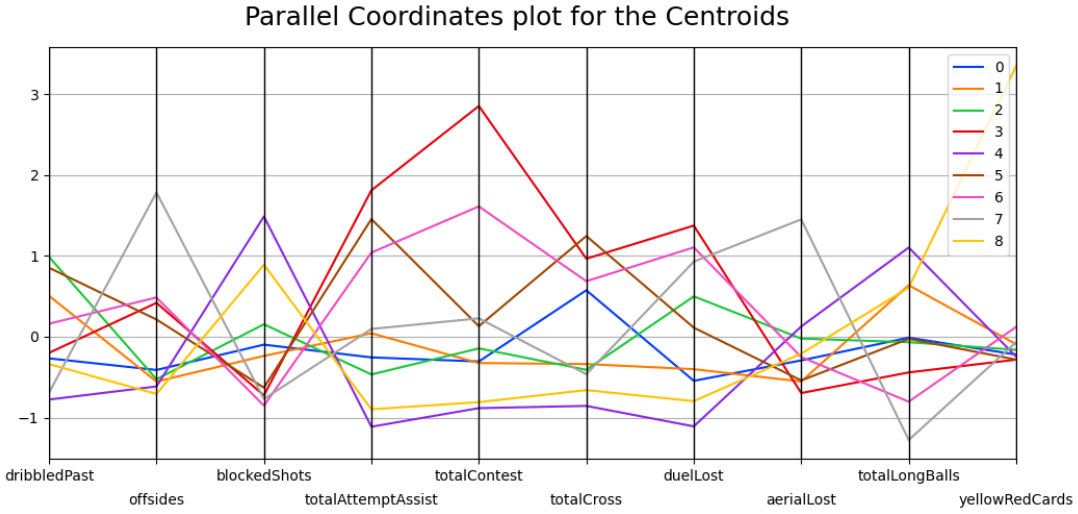
Cuando ejecutamos el algoritmo, a simple vista, cualquiera que sepa algo de fútbol puede ver que la consistencia de los resultados es bastante impresionante (ver grupos de jugadores a continuación).






Parallel Coordinates plot for the Centroids
accurateCrossesPercentage
totalShots
shotsOnTarget
shotsOffTarget
groundDuelsWon
groundDuelsWonPercentage
aerialDuelsWonPercentage
totalDuelsWon
totalDuelsWonPercentage
goalConversionPercentage
0
1
2
3
4
5
6
7
8


Parallel Coordinates plot for the Centroids
penaltyConceded
goalsFromInsideTheBox
goalsFromOutsideTheBox
headedGoals
accurateLongBalls
accurateLongBallsPercentage
clearances
dispossessed
wasFouled
fouls
0
1
2
3
4
5
6
7
8


Parallel Coordinates plot for the Centroids
dribbledPast
offsides
blockedShots
totalAttemptAssist
totalContest
totalCross
duelLost
aerialLost
totalLongBalls
yellowRedCards
0
1
2
3
4
5
6
7
8

Jugadores conocidos en cada grupo (ver archivo *"k_means_clusters.csv"*):

Grupo 0 (41 jugadores): Los laterales. Son medios en todas partes. Interceptan muchos balones y también hacen muchos centros.

Marcelo, César Azpilicueta, Kyle Walker, Juanfran, Juan Bernat, Daniel Carvajal, João Moutinho, David Alaba, Jesús Navas, Filipe Luis, Sami Khedira, Trent Alexander-Arnold, Maxwell ...

Grupo 1 (25 jugadores): Los metrónomos, los que hacen muchos pases, con un porcentaje de éxito muy alto, incluidos los pases largos.

Toni Kroos, Thiago Alcántara, Thiago Motta, Ivan Rakitic, Xabi Alonso, Javier Mascherano, Gabi, Jordi Alba, Philipp Lahm, Luka Modric, Marco Verratti, Sergio Busquets, Yaya Touré, Adrien Rabiot, Aymeric Laporte, Axel Witsel ...

Grupo 2 (33 jugadores): Los numéro 6, los que hacen muchas entradas e intercepciones, que ganan muchos duelos también.

Moussa Sissoko, Blaise Matuidi, Casemiro, Arturo Vidal, Alex Sandro, Thomas Partey, Fernandinho, Fabinho, Saúl Ñíguez ...

Grupo 3 (4 jugadores): Los más decisivos. Marcan muchos goles y dan muchos pases, pero sobre todo intentan y consiguen muchos regates. Juegan muchos duelos y reciben faltas.

Philippe Coutinho, Willian, Lionel Messi, Neymar.

Grupo 4 (48 jugadores): Los centrales, los que ganan muchos duelos, incluidos los aéreos, que hacen muchos pases en su propio campo y despejes.

Lucas Hernández, Thiago Silva, Leonardo Bonucci, Andrea Barzagli, Raphaël Varane, Pepe, Sergio Ramos, Mats Hummels, Marquinhos, Jérôme Boateng, Virgil van Dijk, Giorgio Chiellini, Diego Godín, Gerard Piqué ...

Grupo 5 (29 jugadores): Los creativos, los que hacen pases clave, a menudo en la mitad del campo del adversario, también hacen centros.

Mesut Özil, Arjen Robben, Ángel Di María, James Rodríguez, Mario Götze, Miralem Pjanic, Ilkay Gündogan, Julian Draxler, Kevin De Bruyne, Koke, Joshua Kimmich, Christian Eriksen, Dani Alves, Isco, Marco Asensio, Cesc Fàbregas, David Silva …

Grupo 6 (34 jugadores): Son similares al grupo 3 pero no tan fuertes. Provocan mucho a las defensas, regatean y ganan duelos, marcan y dan pases decisivos, pero menos que los del grupo 3.

Yannick Carrasco, Juan Cuadrado, Dušan Tadic, Radja Nainggolan, Raheem Sterling, Alexis Sánchez, Eden Hazard, Ricardo Quaresma, Riyad Mahrez, Paulo Dybala, Bernardo Silva, Paul Pogba, Christian Pulišic, Lucas Moura, Leroy Sané, Lucas Vázquez, Gareth Bale, Douglas Costa, Franck Ribéry, Ousmane Dembélé …

Grupo 7 (35 jugadores): Las puntas, Son los que marcan mucho, pero pasan poco y regatean poco. Disparan mucho, pero también fallan muchas ocasiones.

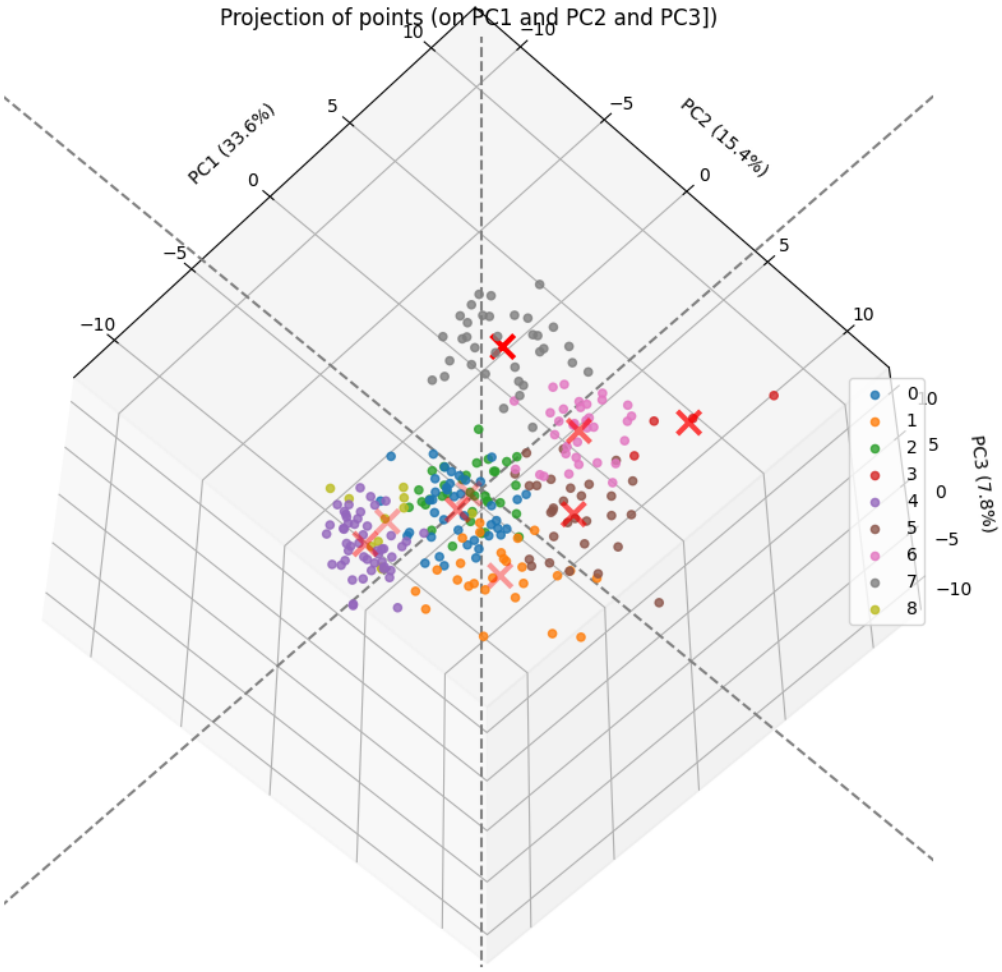
Álvaro Morata, Sergio Agüero, Thomas Müller, Edinson Cavani, Romelu Lukaku, Luis Suárez, Kylian Mbappé, Karim Benzema, Sadio Mané, Cristiano Ronaldo, Harry Kane, Gonzalo Higuaín, Fernando Torres, Roberto Firmino, Robert Lewandowski, Radamel Falcao, Pierre-Emerick Aubameyang, Mohamed Salah, Antoine Griezmann …

Grupo 8 (11 jugadores): Los expulsados, estos son los que recibieron tarjetas rojas.

Nacho Fernández, Javi Martínez, Stefan Savic, Samuel Umtiti …

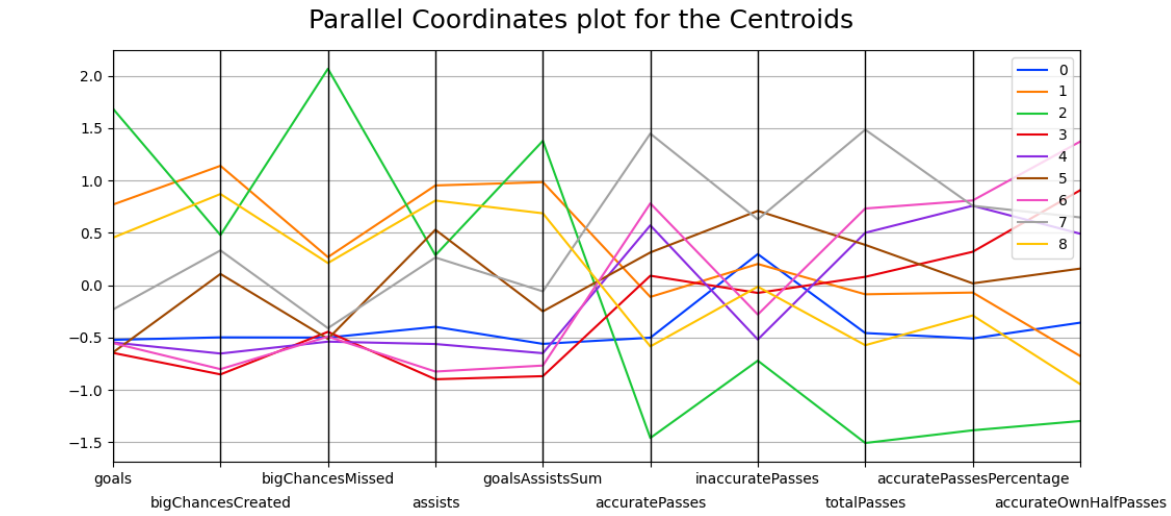
Análisis de componentes principales (opcional)

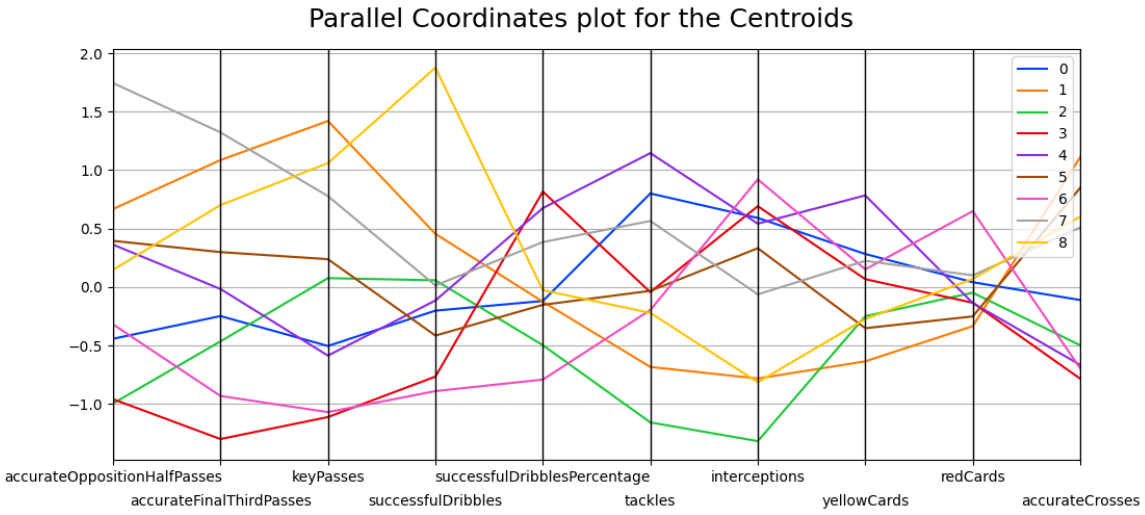
En nuestro caso, utilizar un análisis de componentes principales para representar la distribución en un espacio 2D o 3D no es muy útil. En efecto, al tener 52 dimensiones en la base, reducir el espacio a 2 o 3 dimensiones nos hace perder demasiada información. Sin embargo, aquí está a título informativo (ver archivo *"k_means_PCA.csv"*).



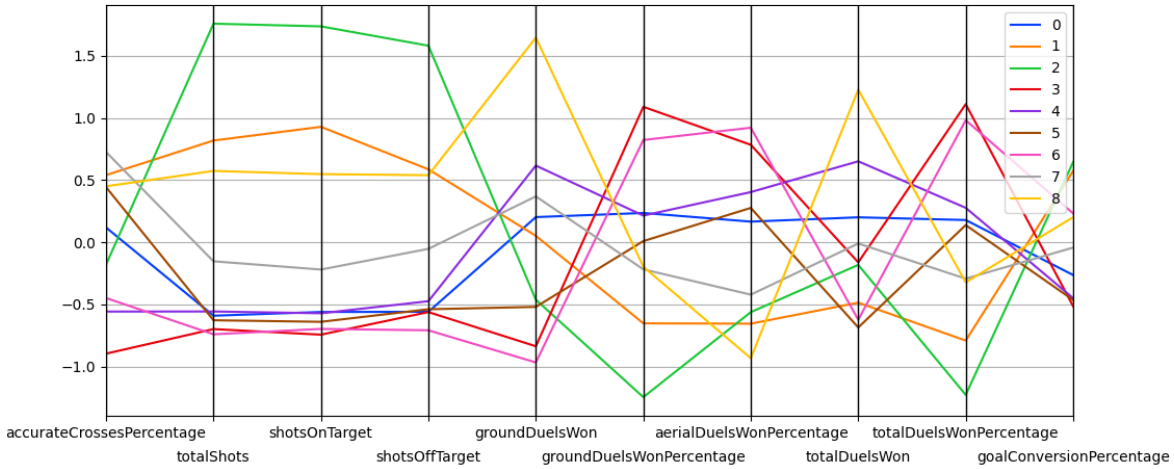
El modelo de mixtura de Gaussianas

Hacemos lo mismo con el modelo de mixtura de Gaussianas.

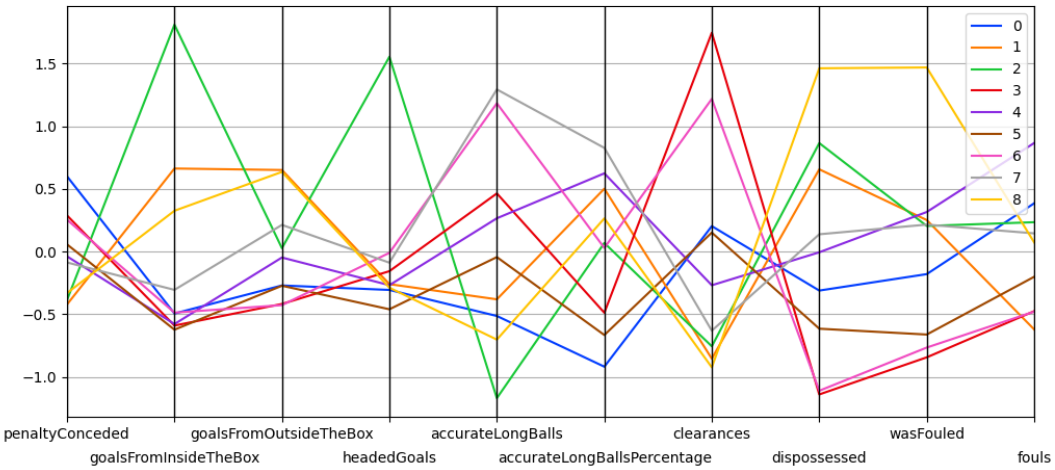




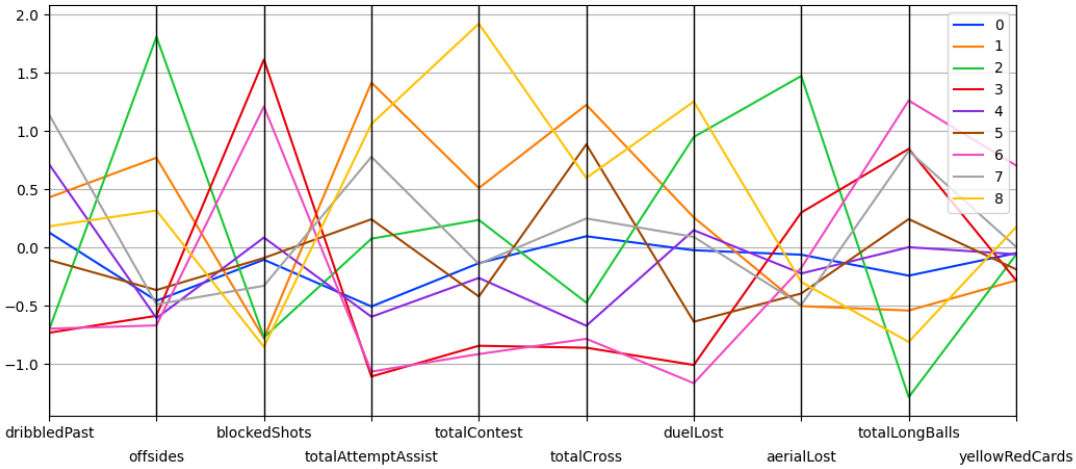
Parallel Coordinates plot for the Centroids



Parallel Coordinates plot for the Centroids



Parallel Coordinates plot for the Centroids



Grupo 0 (33 jugadores): Los laterales. Son medios en todas partes. Atacan e interceptan un poco más que los demás.

Juanfran, Bacary Sagna, Georginio Wijnaldum, Nélson Semedo, Djibril Sidibé, Lucas Digne, Filipe Luis, Elseid Hysaj, Juan Bernat, Alex Sandro, Alessandro Florenzi, Serge Aurier, Saúl Ñíguez ...

Grupo 1 (27 jugadores): Los que son muy decisivos. Marcan, crean ocasiones, hacen muchos pases clave. Sin embargo, no defienden mucho y pierden muchos duelos.

Mesut Özil, James Rodríguez, Lionel Messi, Gareth Bale, David Silva, Mario Götze, Ángel Di María, Antoine Griezmann, Marco Asensio, Arjen Robben, Kevin De Bruyne, Philippe Coutinho, Christian Eriksen, Isco, Radja Nainggolan, Riyad Mahrez ...

Grupo 2 (33 jugadores): Las puntas. Marcan mucho, disparan mucho, pero también fallan muchas ocasiones. No pasan mucho y pierden muchos balones.

Luis Suárez, Harry Kane, Karim Benzema, Gonzalo Higuaín, Kylian Mbappé, Fernando Torres,, Cristiano Ronaldo, Roberto Firmino, Romelu Lukaku, Radamel Falcao, Pierre-Emerick Aubameyang, Marcus Rashford, Sadio Mané, Sergio Agüero, Álvaro Morata, Thomas Müller, Mohamed Salah, Edinson Cavani, Robert Lewandowski ...

Grupo 3 (30 jugadores): Los centrales. Ganan muchos duelos, incluso aéreos, interceptan y hacen muchos despejes.

Virgil van Dijk, Raphaël Varane, Pepe, Mats Hummels, Lucas Hernández, Giorgio Chiellini, Clément Lenglet, Diego Godín ...

Grupo 4 (28 jugadores): Los numéro 6. Atacan, interceptan, sacan muchas tarjetas amarillas pero también pasan mucho con poco desperdicio.

Casemiro, Axel Witsel, Blaise Matuidi, Fernandinho, Fabinho, Sergio Busquets, Thomas Partey, Adrien Rabiot, Danilo ...

Grupo 5 (26 jugadores): Son muy similares al grupo 0 pero son un poco más ofensivos.

João Moutinho, Jesús Navas, Maxwell, Trent Alexander-Arnold, Jordi Alba, Marcelo, James Milner, Sami Khedira, Philipp Lahm, Andrew Robertson, Daniel Carvajal, David Alaba, Kyle Walker

Grupo 6 (30 jugadores): Son muy similares al grupo 3. La única diferencia es que el grupo 3 hace más regates.

Jérôme Boateng, Aymeric Laporte, Samuel Umtiti, Andrea Barzagli, Sergio Ramos, Thiago Silva, César Azpilicueta, Nacho Fernández, Marquinhos, Gerard Piqué, Leonardo Bonucci, Javi Martínez, Javier Mascherano

Grupo 7 (25 jugadores): Los metrónomos, los que hacen muchos pases con mucha precisión, incluidos los pases largos.

Gabi, Koke, Thiago Alcántara, Thiago Motta, Ivan Rakitic, Joshua Kimmich, Toni Kroos, Xabi Alonso, Luka Modric, Andrés Iniesta, Marco Verratti, Dani Alves, Miralem Pjanic, Arturo Vidal, Cesc Fàbregas …

Grupo 8 (17 jugadores): Los que regatean mucho, provocan y reciben faltas. Ganan muchos duelos en el suelo, crean ocasiones pero pierden muchos balones.

Yannick Carrasco, Willian, Ousmane Dembélé, Paul Pogba, Paulo Dybala, Neymar, Alexis Sánchez, Leroy Sané, Juan Cuadrado, Franck Ribéry, Douglas Costa, Raheem Sterling, Eden Hazard, Marlos, Bernardo Silva, Lucas Vázquez …

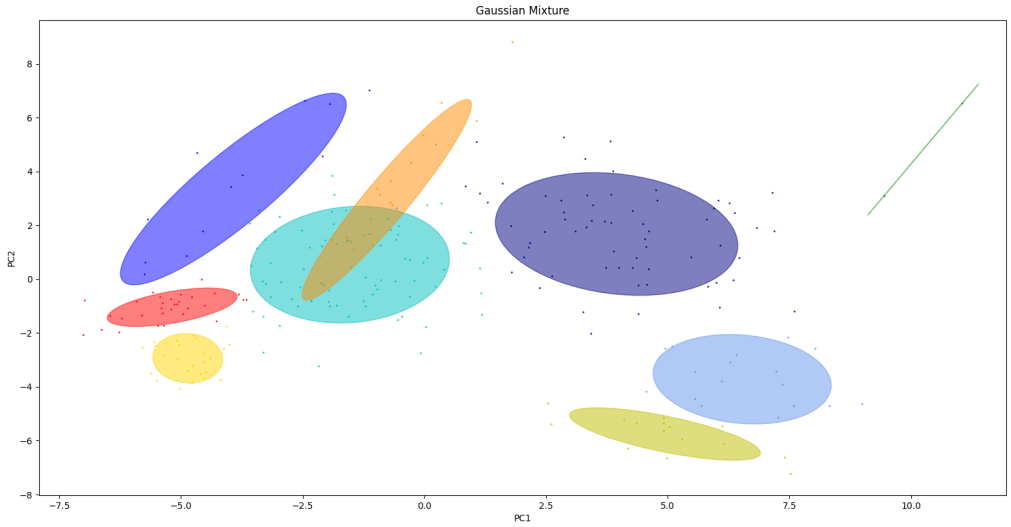
Análisis probabilístico

Lo interesante son las probabilidades de pertenecer a cada grupo. Esperaba ver jugadores situados en la frontera entre dos grupos, pero no es el caso en absoluto. Efectivamente, todos pertenecen a su clúster con una probabilidad muy cercana o igual a 1 (ver archivo “clusters_gmm.csv”). Al principio pensé que se trataba de un error, pero observando la matriz de covarianza (ver archivo “covariances_gmm.csv”), vemos que los valores que la componen son muy

pequeños. Por lo tanto, la probabilidad de pertenecer a una gaussiana disminuye muy rápidamente al alejarse de su centro. Por tanto, no se trata de un error. Esto puede explicarse por el gran número de dimensiones que permiten la formación de clusters muy distintos.

Análisis de componentes principales (opcional)

De la misma manera que para los k-means, utilizar un análisis de componentes principales para representar la distribución en un espacio 2D o 3D no es muy útil. La pongo aquí a título informativo (ver archivo *"gmm_PCA.csv"*).



Conclusión

Es posible clasificar a los jugadores de fútbol de forma muy precisa utilizando métodos de clustering. Aquí, los métodos k-means y el modelo de mixtura de gaussianas muestran resultados muy satisfactorios. Este tipo de algoritmo podría utilizarse, y quizás ya se utiliza, para reclutar jugadores, y encontrar jugadores poco conocidos que tengan características de grandes jugadores.