APRENDIZAJE NO SUPERVISADO PARA LA DETECCIÓN DE SIMILITUDES ENTRE FUTBOLISTAS PROFESIONALES

Luis Rodríguez Rico
30 de junio de 2023

BIG DATA International Campus
ENIIT INNOVA IT BUSINESS SCHOOL
Sports Data Campus
FDIF Football Data International Forum

Índice

1. INTRODUCCIÓN, OBJETIVOS Y METODOLOGÍA

El deporte, y en concreto el fútbol, está sufriendo un enorme cambio en la forma de administrar los recursos en cualquiera de sus áreas, bien sea en el apartado físico, técnico-táctico o de scouting. El factor fundamental que produce esto es el gran volumen de diferentes datos o tipos de información que se recopilan a causa del potencial tecnológico actual.

La circunstancia previamente descrita hace referencia al término *Big Data* que, según la compañía informática Oracle, puede ser definido como aquellos conjuntos de datos de mayor tamaño y complejidad, procedentes de nuevas fuentes de conocimiento, que no pueden ser procesados a través de software convencional pero que, en cambio, facilitan abordar problemas empresariales que antes hubieran sido imposibles de solucionar de forma eficaz y eficiente.

Este hecho obliga a los clubes o entidades deportivas a poseer profesionales capacitados que conozcan la realidad que rodea al equipo o deportista en particular, para extraer valor de estas grandes cantidades de datos y facilitar la toma de decisiones. En consecuencia, la puesta en marcha de una cultura *data-driven* dentro de la organización y la modernización de los materiales utilizados en el departamento en cuestión, es clave dentro de múltiples escenarios, como puede ser la preparación de los partidos, la gestión económica de la entidad o la detección de un jugador desconocido de alto rendimiento.

La dirección deportiva es uno de los departamentos que más influyen en el desarrollo y crecimiento de una entidad, dado que, no solo permite confeccionar la platilla con la que el equipo competirá para tratar de alcanzar las aspiraciones definidas a principio de temporada, sino que añade valor a esta, a partir de la contratación de nuevos jugadores o entrenadores. Por ello, cabe la posibilidad de que, tras una excelente gestión, en algún momento se localice a la persona que genere un salto de calidad en ella, tanto a nivel de resultados en las diferentes competiciones como económico.

De modo que, el objetivo fundamental de este proyecto final de master es construir una herramienta que facilite la detección de jugadores similares a uno dado. Así pues, esta no solo ofrecerá a las entidades deportivas la capacidad de sustituir a un jugador que se ha marchado por otro similar, sino que también dispondrá de otros usos para la propia organización. Entre ellos, por un lado, está la posibilidad de encontrar a un jugador semejante a uno objetivo en caso de que el fichaje de este ha fallado o por capacidades económicas no se pueda aspirar a él y, por tanto, se deban rebajar las pretensiones. Por otro lado, existe el riesgo de que cuando se realice un proceso de búsqueda de jugadores de forma intuitiva se deje de incluir a un jugador que al final resulte ser más propicio que los del conjunto analizado, por lo que con ella se evitarían esos "olvidos". Por último, su empleo sería de gran utilidad para la detección de jóvenes promesas que sea similares a un jugador de alto nivel pero que, sin embargo, compitan en lugares o entidades menos conocidas por el gran público.

Desde otro punto de vista, relacionado con el apartado técnico táctico, este instrumento también da viabilidad para detectar potenciales cualidades o atributos desconocidos en un deportista de la propia plantilla. Este escenario podría ser crucial para corregir un posible bajo rendimiento en el jugador. Dado que, si se compara con otro futbolista, de quien se piensa que posee un estilo parecido y que se encuentra en un buen estado de forma, por medio del análisis de los resultados en la aplicación, se podría detectar que el primer jugador no posee una categorización similar al segundo y, de acuerdo con ello, concluir que no se está haciendo un

uso correcto de él en los esquemas. Por lo que, serviría de gran ayuda para el cuerpo técnico, quienes podrían cambiar su posición o estilo de juego del equipo para adaptarlo mejor.

Para la obtención de resultados a través de este mecanismo de detección de jugadores similares, se necesita aplicar un proceso ingesta, transformación y procesamiento de los datos. De esta manera, en primer lugar, se tienen que extraer los diferentes conjuntos de información por medio de fuentes abiertas. En segundo lugar, se requiere aplicar sobre ellos una selección de los campos más importantes para, posteriormente, realizar una transformación y limpieza de los datos. Lo que, a su vez, posibilite la generación de nuevas tablas por puesto y tipo de registro, ya definitivas para el procesamiento. En tercer lugar, se aplican los diferentes métodos estadísticos y matemáticos para el análisis no supervisado, que garanticen la reducción del gran número de características contenidas en los datasets por puesto, la agrupación de los jugadores, para extraer información acerca de sus características de juego; y las puntuaciones de similitud, que conlleven la creación de un porcentaje definitivo de semejanza. Finalmente, estos nuevos valores, contenidos en unos archivos finales para cada posición, junto con los datos generales de los deportistas, son los utilizados sobre la herramienta final que da sentido a este proyecto.

Para terminar con esta introducción a la memoria del proyecto, se debe recalcar que existen factores intangibles que el análisis avanzado de datos no puede considerar ni cuantificar de manera objetiva (Soria Polo, 2021, p. 2)., como por ejemplo la adaptabilidad del futbolista a una nueva cultura o al equipo, las responsabilidades familiares del jugador o simplemente su manera de ser. Por tanto, hay que puntualizar que la utilidad que posee este instrumento y, por tanto, el proyecto, es la ofrecer a la dirección deportiva mayores facilidades en la toma de decisiones en base a resultados científicos. Así pues, siempre va a ser tarea de esta y del cuerpo técnico decidir quién es el futbolista que mayor potencial de rendimiento aportará al desarrollo deportivo del equipo.

2. EXTRACCIÓN Y PREPROCESAMIENTO DE LOS DATOS

2.1. EXTRACCIÓN

La extracción de datos consiste en la recuperación de diferentes tipos de información desde varias fuentes, para su posterior procesamiento, almacenamiento y análisis. Para este proyecto se ha utilizado fundamentalmente la librería *worldfootballR* que como su propio nombre indica se encuentra diseñada para el lenguaje de programación R.

Ilustración 2.1: Logos de R y worldfootballR

Fuentes: www.r-project.org y github.com/JaseZiv/worldfootballR

Esta librería permite extraer datos relacionados con resultados de fútbol y estadísticas de los jugadores procedentes de diversos portales web (FBref, Transfermarkt, Understat y Fotmob). Dicho paquete construido por Jason Zivkovich posee su propio repositorio de GitHub donde se pueden encontrar las últimas actualizaciones y correcciones efectuadas en este. Además, dispone de su propia página web con las instrucciones sobre el uso de cada una de las funciones para la obtención de los datos desde las fuentes mencionadas anteriormente.

En este estudio se ha hecho uso, exclusivamente, de un par de funciones que permiten extraer datos de FBref y Transfermarkt. La primera página ha permitido la consecución de un conjunto muy completo de diferentes estadísticas de los jugadores. Mientras que la segunda, ha posibilitado la obtención de información relacionada con el propio jugador como pudiera ser su Valor de Mercado (€), la fecha de fin de contrato, su posición específica, su actual club, su altura, su pie bueno, etc.

Ilustración 2.2:Logos de FBref y Transfermarkt

Fuentes: fbref.com y www.transfermarkt.com

Las competiciones seleccionadas para la obtención de los jugadores que han formado parte del análisis y por ende sus estadística de rendimiento han sido aquellas que dentro de FBref poseen todos los tipos de métricas posibles *(standard, shooting, passing, passing types, keeper, keeper advanced, goal creation actions, defense, possession y miscellaneous)*. De este modo, para el análisis se han considerado las siguientes ligas: La primera y segunda división inglesa *(Premier League y Championship)*, además de, las primeras divisiones española *(La Liga)*, italiana *(Serie A)*, alemana *(Bundesliga)*, francesa *(Ligue 1)*, portuguesa *(Primeira Liga)*, holandesa *(Eredivisie)*, brasileña *(Série A)*, mexicana *(Liga MX)* y norteamericana *(MLS)*.

Para el propio proceso de extracción se ha decidido crear un *script* de R llamado *Extract.R*. El propio documento se encuentra formado por las siguientes fases:

La primera parte consiste en fijar el directorio del proyecto, para así posibilitar al código, sin importar la máquina desde donde se esté ejecutando, guardar los diferentes conjuntos de datos obtenidos en la carpeta *Data*.

En segunda lugar, se procede a extraer los datos del portal FBref, de tal manera que se opta por crear una función denominada *getPlayerStats()* que permita crear los diferentes *datasets* en función de los tipos de métricas de rendimiento comentadas anteriormente, utilizando el parámetro *type* de esta función para tal fin. Para ello, se crea un vector con todas las *URLs* de las competiciones ligueras, del cual se hará uso en un bucle "*for*", aplicando el link que dirige a cada liga sobre la función *fb_teams_urls()* para así obtener los equipos que la conforman.

Una vez conseguido este vector de entidades para una liga específica, se pasa a aplicarlo en la función *fb_team_player_stats(),* junto con el tipo de estadística. Con el objetivo de obtener un dataset que guarde las métricas para ese tipo y esa liga en específico. Posteriormente se procede a concatenar los resultados obtenidos por cada competición para que, de este modo, se consiga el conjunto completo de datos con todas los torneos ligueros para este tipo de métrica de rendimiento en particular. Por tanto, y según lo comentado en párrafos anteriores, se obtendrán 10 archivos de datos por cada tipo de métrica con todas las ligas incluidas.

En tercer lugar, se generan los 10 conjuntos de datos, aplicando la función *getPlayerStats()* ese números de veces e indicándole el tipo de estadística necesaria *(standard,*

passing, shooting, etc.). Seguidamente, una vez creado el dataset se guarda en formato CSV dentro de la carpeta *Results* del proyecto para su posterior tratamiento.

En cuarto lugar, se extraen los datos de *Transfermarkt*, en este caso algo más sencillo, pues no es necesario crear una función propia y tan solo se genera un archivo de datos. Aun así, si es necesario crear un vector con las URLs de las competiciones para este portal. Dicho conjunto de links es igualmente aplicado sobre un bucle *"for"* para así, indicarle a la función *tm_player_market_values()* el link y el año de comienzo de la temporada, siendo 2022 a tal efecto para este proyecto. Posteriormente, sobre este bucle se van concatenando los diferentes sets de datos para cada liga y así obtener el *dataframe* final, que también será guardado en formato CSV dentro de la misma carpeta *Data.*

Por último, pero no menos importante, se extrae un archivo de mapeo que permite unir los datos de ambos portales para los distintos jugadores, donde se utiliza como clave los links de sus perfiles para las dos webs. Estos datos son obtenidos, gracias a la función disponible en el propio paquete de R llamada *player_dictionary_mapping()* que no necesita recibir ningún parámetro ya que directamente descarga el archivo en el propio repositorio de GitHub. Finalmente, y de igual forma, este es guardado en un CSV dentro de la carpeta *Data.*

Así pues, en este directorio se encuentra toda la información extraída de la cual beberá el proyecto y que será necesaria procesar para obtener unos datos limpios de cara a aplicar en el análisis no supervisado.

2.2. TRANSFORMACIÓN Y LIMPIEZA DE LOS DATOS

Para la transformación y limpieza de los datos obtenidos en FBref y Transfermarkt, se crea en primer lugar un archivo Excel que permita la selección de los campos que se han considerado requeridos para el estudio, junto con la explicación y definición de cada uno ellos. Este archivo se denomina *"PFM Columns"* y se encuentra guardado en la carpeta *"Documents"* del proyecto. Dentro de este archivo, existen cuatro hojas que serán requeridas para la transformación y creación de los datasets por posiciones y datos generales de los jugadores.

La primera de ellas es *"To Select"*, que como su propio nombre indica, posibilita seleccionar todas las columnas requeridas para este proceso, además de, renombrar el nombre original de los campos por otros elegidos para el análisis y así poder tener un mejor entendimiento de cada uno de ellos.

La segunda, *"To Calculate",* está relacionada con las columnas que tendrán que ser recalculadas, principalmente, aquellas cuyo tipo de valor es numérico y está representado o bien en porcentaje o por 90 minutos. Se debe realizar esto dado que, existen jugadores que han disputado minutos a los largo de esta temporada en equipos diferentes, sea el caso de João Félix (Atlético de Madrid - Chelsea) o João Cancelo (Manchester City – Bayern München) y, por tanto, posee estadísticas diferentes para un mismo campo que serán necesarias volver a calcular. Posteriormente, en la explicación del código se describirá este proceso.

La tercera se trata de "*Definitive Ones"* cuya utilidad es bastante simple, dentro de la limpieza del dataset principal sobre el que se van realizando todas la transformaciones y cálculos. Así pues, se hará uso de esta hoja para seleccionar todas las columnas que son

definitivas para el estudio y eliminar o tirar todas aquellas que han servido para los cálculos pero que no son de interés.

Por último, la cuarta hoja, nombrada *"Position Columns"* cumple la función de clasificar todos los campos que existen en la tabla principal de este proceso en cado uno de los datasets creados para este estudio *(Keepers, Defenders, Midfilders, Attackers y General)*. De este modo, se ha establecido una primera columna llamada *"Column Dataset"* que permite identificar el tipo de campo según posición o dato general del jugador, para que así, el código posteriormente sea capaz de introducirlo en cada una de las cinco tablas creadas según su categoría.

Además, dentro de esta columna de la hoja, explicada anteriormente, se incluyen dos categorías especiales "ID" y "FILTER", la primera de ellas aparece en los cinco conjuntos de datos y hace referencia al link del jugador en FBref, lo que posibilita su diferenciación y la unión de las métricas de rendimiento con los datos generales en el esquema de la herramienta de visualización de Power BI. Por otra parte, la segunda se utiliza como filtro y alude a la posición/es del jugador en FBref (GK, DF, MF, FW), de este modo, una vez seleccionados los campos que van dentro de cada tabla de posiciones se utilizará esto para solo incluir a aquellos jugador que poseen al menos la posición del dataset al que hace referencia. Es decir, por ejemplo, cuando se crea la tabla de defensores se seleccionan todos los jugadores que dentro de la posición de FBref cuenta con el término "DF".

Por otra parte, se considera importante destacar que una misma métrica puede ir a dos o más conjuntos de datos por puesto, dado que, dicho campo puede ser relevante en el estudio para diversas posiciones. Por ejemplo, se han contemplado las asistencias tanto para los centrocampistas como para los atacantes. Relacionado con lo anterior, también es significativo reflejar que un jugador tiene la capacidad de estar en dos tablas diferentes en base a la información obtenida en FBref y según lo explicado anteriormente en la parte del filtro, un ejemplo de esto es el mismo João Cancelo quien según este portal puede ser defensor (DF) o atacante (FW), dada su posición de lateral/carrilero derecho.

Una vez creado este archivo, se procede a la transformación y limpieza de los datos, haciendo uso del lenguaje de programación de Python y sus librerías Pandas y Numpy principalmente, con el objetivo de obtener y guardar los datasets definitivos que se utilizarán en los análisis no supervisados aplicados en este estudio.

Ilustración 2.1: Logo de Python

Fuente: python.org

Dentro de este proceso, en primer lugar, se procede a leer cada una de las hojas explicadas anteriormente y se crea una lista con cada uno de los archivos CSV obtenidos en la extracción de los datos con R, para su posterior lectura a través de un bucle *for*, excepto para las métricas de tipo "*Standard"*.

En segundo lugar, se leen los archivos CSV con las métricas *"Standard"* de FBref, los datos de Transfermarkt y aquel que contiene los links de ambos portales para la unión de los datos. Posteriormente para todos ellos, se pasa a renombrar sus campos según lo establecido en el documento de Excel para la selección de columnas y, se une cada uno de ellos en un mismo conjunto de datos denominado *"actual_df"*.

En tercer lugar, se aplica un bucle *"for"* sobre los archivos restantes, los cuales reflejan cada uno de los tipos de estadísticas de rendimiento de FBref, exceptuando la comentada en el paso anterior. En dicho proceso, se efectúan cada una de las acciones explicadas en el párrafo anterior, es decir, renombramiento y unión de datos, hasta conseguir un primer dataset configurado por 8157 filas y 150 campos, que deberá ser tratado, en etapas posteriores.

En cuarto lugar, ya hechas todas las uniones se procede a eliminar todos los jugadores que no tuvieron minutos (valores nulos en *"Minutes"*), principalmente son jugadores que aparecen en FBref pero que realmente forman parte del equipo filial o sub21 de las entidades a las que pertenecen. De tal manera, que al aplicar este filtro el dataset pasa a tener 6930. Además, por la forma en la que aparecen los datos, para algunas competiciones el número de partidos jugados aparece en el campo *"MP"* y para otras en *"MP2"*, por tanto, es necesario harmonizar estos valores en un mismo campo, eligiendo para ello el primero (MP).

En quinto lugar, se pasa a la fase de los cálculos para obtener un único valor en los jugadores que aparecen dos veces por haber estado en dos equipos en la misma temporada. Esto se aplicará sobre las variables relacionadas con porcentajes, medias o valores por 90 minutos. Para ello, se seleccionan todas la métricas de rendimiento (valores numéricos) y se agrupan mediante una suma, obteniendo el agregado de cada jugador individual. Posteriormente, se calcula sobre el código cada una de esas estadísticas mediante su fórmula. Por tanto, el jugador que tan solo aparece una vez obtendrá el mismo valor, pero para el jugador que ha estado en más de un equipo su métrica será aquella que proceda de los valores agregados de la temporada. Un ejemplo de todo lo comentado anteriormente son los goles por 90 minutos. Así pues, para este tipo de jugador, primeramente, se deberá sumar los minutos disputados en ambos equipos y luego dividirlos entre 90. En segundo término, se tendrán que sumar los goles anotados en las dos clubes, para finalmente dividir la suma de los goles entre los partidos disputados exactos (minutos/90). Después de haber aplicado todos los cálculos se deberá unir este dataset agrupado con los del dataset anterior para tener el resto de los campos que no son numérico y que no formaban parte de esto (Equipo, Nacionalidad, Valor de Mercado, etc.), obteniendo un dataset sin jugadores duplicados.

En sexto lugar, se seleccionan las columnas definitivas requeridas para el estudio y se aplican una serie de correcciones sobre los datos como eliminar los días de la edad y seleccionar solo los años, transformar esta misma a numérica, eliminar del nombre del equipo el año de la temporada para la liga brasileña y norteamericana, introducir el nombre de la liga en el campo competición para los equipos de estás ligas, además de para algunos clubes de Países Bajos, Portugal y México.

En séptimo lugar, se procede a establecer el equipo más actual del jugador, para la fecha en la que se han extraído los datos, en base al equipo que aparece en Transfermarkt. Asimismo, se seleccionan los jugadores que han disputado al menos un 30% del total de minutos de su competición.

Finalmente, en función de la posición del jugador y el tipo de variable, se procede a crear cada uno de los cinco archivos finales requeridos para la elaboración de este estudio, en base a lo descrito anteriormente en la explicación de la hoja *"Position Columns"*. Estos archivos, con los datos limpios y preparados, están guardados en formatos CSV dentro de la carpeta *"Results"*. El esquema o estilo que sigue sus nombres es el siguiente; *<<Type>>_Dataset.csv*. De igual modo, en función de los campos seleccionados y jugadores para cada posición, las dimensiones de cada uno ellos son diferentes:

- *General_Dataset.csv* → 3110 filas y 17 columnas (incluye todos los jugadores)
- *Keepers_Dataset.csv* → 247 filas y 64 columnas (incluye todos los porteros)
- *Defenders_Dataset.csv* → 1299 filas y 67 columnas (los que pueden ser defensas)
- *Midfilders_Dataset.csv* → 1357 filas y 91 columnas (incluye los centrocampistas)
- *Attackers_Dataset.csv* → 996 filas y 115 columnas (los que pueden jugar en el ataque)

Así pues, tras haber aplicado esta fase de selección, transformación y limpieza de los datos, se han conseguido una serie de datasets que serán utilizados en las próximas etapas del estudio. Dentro de la fase de los algoritmos de análisis no supervisado, se hará uso de los archivos relacionados con las métricas de rendimientos por posiciones con el objetivo de obtener unos documentos de datos con los resultados finales para alimentar a los cuadros de mando de la herramienta en Power BI, mientras que los datos de carácter general se aplicarán directamente en esta.

3. REDUCCIÓN DE LA DIMENSIONALIDAD: ACP

El análisis de componentes principales (ACP) es una técnica de reducción de la dimensionalidad que posibilita pasar de una gran cantidad de variables con una alta correlación entre ellas a un nuevo conjunto de características denominado componentes y, que proceden de las originales, pero sin ser correlativas entre sí. Estas también se encuentran limitadas, es decir, la primer componente albergará la máxima variabilidad posible de los datos, mientras que la segunda la siguiente y así sucesivamente.

Así pues, este método trata de obtener combinaciones lineales de las variables originales que expliquen los mejor posible la información o variabilidad contenida en los datos. De tal manera que, con menos variables, componentes principales, es suficiente para poder comprender aquello que estaba contenido en los datos con las dimensiones originales.

Por tanto, este algoritmo de aprendizaje no supervisado es de gran aplicabilidad dentro de los cuatro datasets por puestos, dado que, están formados por una alta dimensionalidad como se ha visto en el apartado anterior. Asimismo, el procedimiento a través del cual se construyen estos nuevos campos y su relación con las variables originales permite tener un mejor entendimiento de la correlación inherente en los datos originales, algo que sería muy difícil de comprender dada sus altas dimensionalidades.

Finalmente, las nuevas componentes principales, obtenidas a partir de la aplicación de esta metodología, formarán un vector aleatorio de dimensión menor que será empleado para obtener las puntuaciones de los jugadores en cada componente. De modo que, posteriormente se hará uso de esto para calcular la correlación entre jugadores y así obtener las primeras puntuaciones de similitud de este estudio en base al propio ACP.

A continuación, se describirán los pasos para su aplicación en cada uno de los cuatro conjuntos de datos. Se ha decidido utilizar R como lenguaje de programación para su ejecución, dado el alto nivel de representación gráfica que posee la librería empleada para obtener los resultados *(factoextra)*.

3.1. RELACIONES LINEALES ENTRE VARIABLES

Una vez leídos los cuatro datasets por posiciones sobre los que va a ser aplicado este método, se procede a calcular el determinante de la matriz de correlación (R) para que de forma científica se demuestre el grado de similitud entre los campos incluidos en cada tabla. Por consiguiente, se necesita obtener la propia matriz de correlación para cada par de variables numéricas incluidas en cada conjunto de datos, haciendo uso de la librería *corrplot*.

De tal forma que, un valor bajo de este determinante indicará una alta correlación entre los campos, lo que demuestra la idoneidad de aplicar el ACP sobre estos. Dado que existe un alto grado de información común entre ellos, pudiendo reducir su alto espacio de dimensiones.

Ilustración 3.1: Determinantes de la matriz de correlación para las cuatro posiciones

```
[1] "Determinante de (R) para Porteros: 1.03933518059957e-136"
[1] "Determinante de (R) para Defensores: -7.08770053982771e-120"
[1] "Determinante de (R) para Centrocampistas: 0"
[1] "Determinante de (R) para Atacantes: 0"
```

Fuente: Elaboración propia con R

En este sentido, se puede observar que el resultado para todas las posiciones es muy próximo a cero, incluso siendo nulo para los centrocampistas y atacantes, lo que pone de manifiesto que los campos de las diferentes tablas estás altamente correlacionados y que por tanto es viable aplicar el ACP.

3.2. APLICACIÓN Y SELECCIÓN DEL NÚMERO ÓPTIMO DE COMPONENTES

Como se ha explicado, el objetivo de este algoritmo es el de crear un nuevo conjunto de variables (componente principales), no correlacionadas entre sí y de máxima varianza posible, para que el dataset resultante, con una menor dimensionalidad, mantenga la mayor parte de la información original.

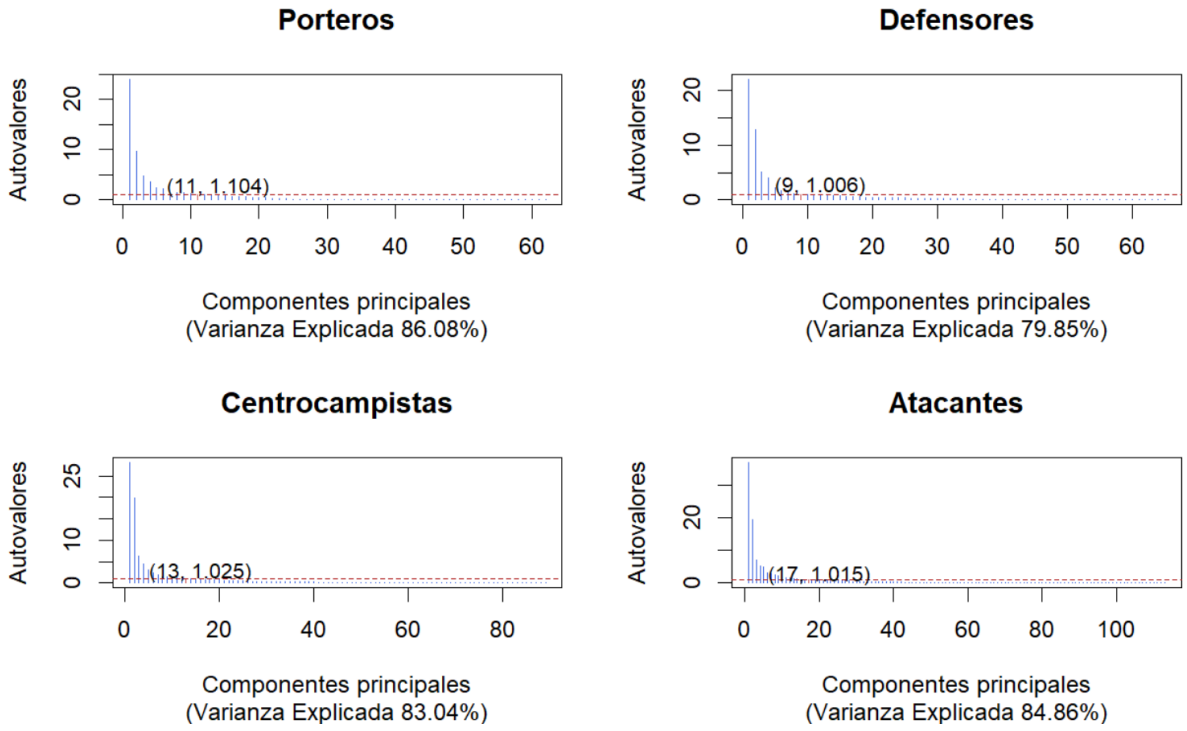
Para su aplicación se utiliza la función *prcomp()* del paquete *factoextra*, estableciendo como verdadero el parámetro que permite el escalamiento de los datos directamente en la propia ejecución. Dicho proceso es necesario pues los campos del dataset se encuentran en diferentes escalas o medidas, además pueden existir valores extremos entre ellos; los cuales no se han decidido tratar pues no existen errores de medición.

Por tanto, si antes de calcular las componentes principales no se estandarizan todos los campos a media cero y desviación típica uno, aquellas variables cuya medición/escala sea mayor va a dominar sobre el resto en la aplicación del algoritmo, dado que la varianza se calcula como potencia al cuadrado en su misma escala.

Una vez obtenida la proporción de varianza explicada para cada componente principal en los datasets para los cuatro puestos, se procede a la selección del número óptimo de componentes. Para ello se ha decidido seleccionar aquel procedimiento que considera las componentes cuyos autovalores son superiores a la media, pues es una técnica de selección que puede ser automatizada, no dejando de ser analizada y debiendo considerar si son suficientes.

Para ello, se representarán una serie de gráficos de sedimentación para las posiciones, siendo un gráfico de barra que representa en el eje X cada una de las componentes y en el eje Y los autovalores, es decir varianza explicada por cada componente. Además, se muestra conjuntamente la media de todos los autovalores, para mostrar el número óptimo de componentes a utilizar.

Ilustración 3.2: Gráficos de las componentes con autovalores superiores a la media para las cuatro posiciones



Fuente: Elaboración propia con R

Así pues, se observa que para los porteros el número óptimo de componentes es 11 con un porcentaje de varianza explicada del 86%. Con relación a los defensores, se observa que reduciendo la dimensionalidad del dataset a nueve nuevas variables se consigue contener el 80% de la información procedente de los primeros datos. Para los centrocampista, se alcanza el 83% de la varianza explicada con 13 componentes, mientras que para los atacantes se necesitan 17 nuevas componentes para obtener el 85% de la información contenida en los datos iniciales.

De esta forma, para los próximos pasos dentro de este análisis y para las siguientes aplicaciones de algoritmos en el estudio, se considerarán los nuevos datasets para cada posición con estas nuevas variables o dimensiones.

3.3. COEFICIENTES Y CORRELACIÓN CON LAS ORIGINALES

Justo después de haber determinado el número óptimo de componentes a considerar para reducir el espacio de variables de los cuatro datasets por posiciones, la próxima fase es interpretar estos nuevos campos construidos.

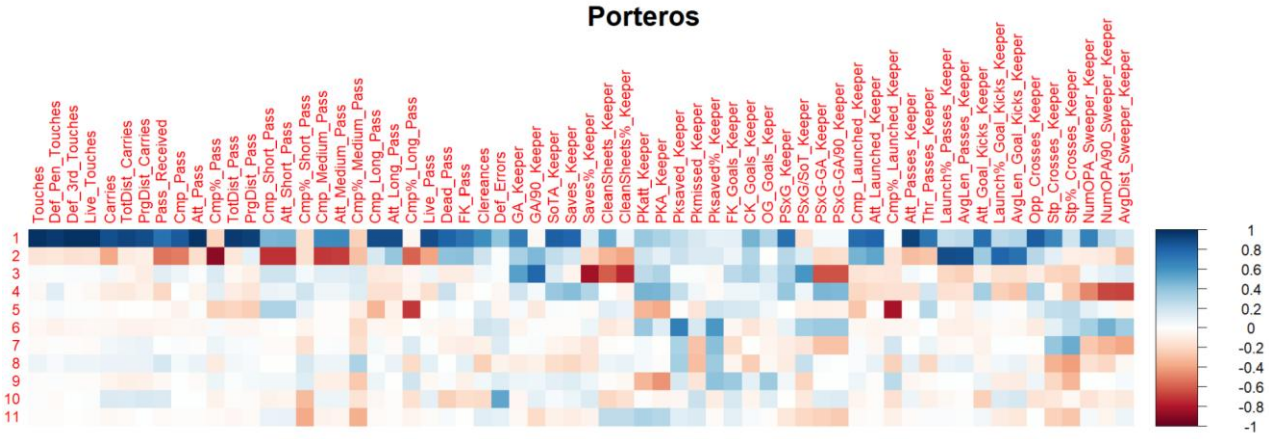
Las nuevas componentes son combinaciones lineales de las variables originales, como se ha venido explicando. Por tanto, los coeficientes sobre estas coinciden con los autovectores de los autovalores de la matriz de correlación, puesto que, los vectores propios de una matriz

son todos aquellos que resultan en el mismo o en un múltiplo entero de este, al multiplicarlos por la propia matriz.

Una vez obtenido el coeficiente de cada componente, a través del método *$rotation* del PCA, que multiplica a cada campo original para calcular las nuevas componentes principales; se procede a conocer cómo se correlacionan con estas variables reales. El objetivo de ello es ver que peso tienen las primeras métricas con las nuevas componentes extraídas, especialmente con las 2 primeras ya que posee la mayor parte de la variabilidad que se llega a explicar.

Así pues, el cálculo de las correlaciones entre las variables originales y la nuevas componentes, consiste en multiplicar el coeficiente de cada componente (*PCA$rotation*) por la desviación típica de la componente correspondiente. Seguidamente, se procede a mostrar un gráfico de correlación para los porteros que muestra estas correlaciones entre cada par de variables originales con las 11 nuevas componentes construidas.

Ilustración 3.3: Gráfico de correlación entre variables originales y componentes principales para los porteros

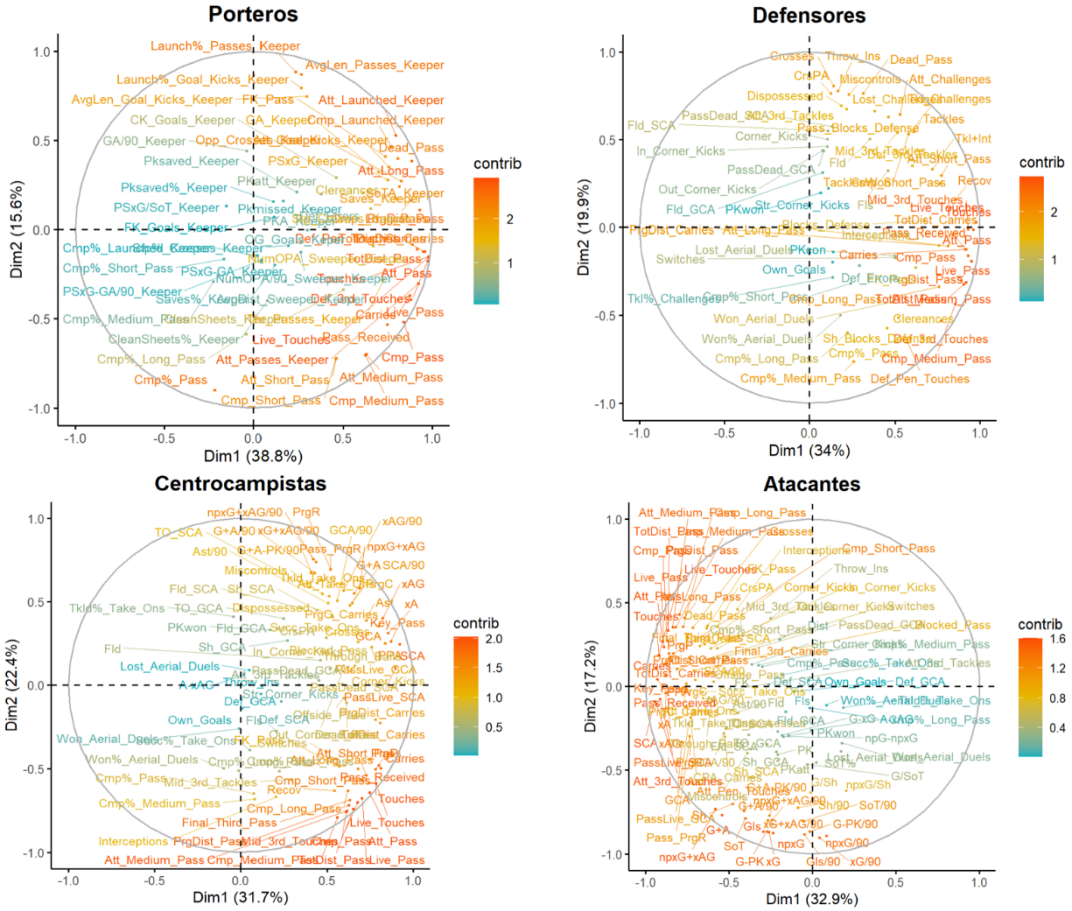


Fuente: Elaboración propia con R

Por consiguiente, y como se viene explicando a lo largo de este apartado, la primera de las componentes es la más correlacionada con la variables iniciales, lo que indica que es aquella que posee mayor capacidad para explicar la variabilidad original. Además, gracias a este gráfico se puede empezar a determinar la relación que tienen los campos iniciales con las nuevas componentes y así ver la importancia que tendrán respecto a ellas posteriormente.

Con ayuda de la propia librería *factoextra* y haciendo uso de la función *fviz_pca_var()*, se representan las contribuciones que las variables originales poseen respecto a las dos primeras componentes, además permite conocer, en la parte inferior de los ejes, el porcentaje de variabilidad explicada que posee cada una.

Ilustración 3.4: Representación de las variables originales sobre las dos componentes principales por puestos



Fuente: Elaboración propia con R

De tal manera que, para los porteros la primera componente, con una capacidad de explicar el 39% de la variabilidad contenida en los datos originales, se encuentra altamente influenciada por todas aquellas variables relacionadas con el contacto del balón con los pies para dar pases, independientemente de la longitud de desplazamiento. Mientras que para segunda (15,6%) el acierto y el total de distancia recorrida tiene un mayor peso.

Para el caso de los defensores, se observa que los duelos ganados, como recuperaciones, y el juego de balón, tanto en pases como en salida de balón, tienen una gran contribución en la primera de ellas, la cual explica el 34% de la variabilidad contenida en los datos originales. En cambio, para la segunda se analiza una mayor influencia de los errores, tales que duelos perdidos, controles de balón malos o pases que no van a ninguna parte.

En lo relacionado con los centrocampistas, se contempla para la primera componente (31,7%) que todos aquellos campos correspondientes a métricas de posesión en ataque como son conducciones que progresan en distancia, asistencias o toques de balón, influyen altamente sobre esta. Así pues, para la segunda (22,4%) las variables relativas a la finalización en ataque contribuyen más, dichos campos son goles más asistencias esperadas por 90 minutos, acciones creadoras de gol (GCA) o, también pases muy progresivos recibidos o que son alcanzados dentro del área rival (PrgR), etc. Lo que de que los centrocampista ofensivos estarán altamente representados en dicha dimensión.

Finalmente, en cuanto a los atacantes, se puede señalar que la primera componente (33%) se encuentra opuestamente contribuida por aquellas variables originales relacionadas con el contacto con el balón, como son los pases, conducciones, toques, etc. Por tanto, todos aquellos delanteros que participen altamente en la creación del juego estarán correlacionados negativamente con dicha dimensión. De igual modo, la segunda (17,2%) se encuentra influida de manera contraria por las principales métricas ofensivas de los atacantes, es decir, goles, disparos a puerta o acciones creadoras de gol.

Conocida la contribución de las variables originales sobre las componentes principales para cada puesto, el último paso es determinar los valores que tomarán los jugadores sobre estas nuevas variables para así determinar la correlación entre ellos y, de forma última, su similitud

3.4. PUNTUACIONES INDIVIDUALES SOBRE LAS COMPONENTES

Para calcular el valor que toma cada uno de los futbolistas en las nuevas componentes, que están contenidas cada uno de los nuevos datasets por puesto, se debe utilizar el elemento *$x* del PCA, el cual permite acceder a esas puntuaciones o *scores.* Seguidamente, se hace ver, a modo de ejemplo, una muestra del dataframe de los defensores. Este presenta el valor que toma cada jugador en las nueve componentes obtenidas para ellos.

Ilustración 3.5: Muestra de un dataframe de defensores sobre las puntuaciones de los jugadores en las componentes

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Ohis.Felix.Uduokhai	-4.3480496	-1.0661139	1.4468817	-1.1122231	-0.3176581	0.1664406	-1.1230322	-0.5301185	-0.2868781
Iago	-5.8495520	2.8101465	0.4977725	-0.8250020	-1.4860901	2.3133807	-0.9964326	0.9482879	0.4977971
Mads.Pedersen	-4.5566421	3.5973654	1.4011567	-0.5940390	-1.2580579	-0.2601329	0.7999137	0.2119127	-0.6670320
Maximilian.Bauer	-3.6128739	-0.7353300	1.4097437	-2.0454100	-1.0755182	-0.5115691	-1.0410057	-0.3566294	0.0245092
Robert.Gumny	-4.5508017	2.3884920	1.7646843	-1.1293490	-2.1027998	1.7633809	-0.9545543	-0.1214459	-0.3420963
Jeffrey.Gouweleeuw	4.1131315	-1.6618981	2.1024885	-3.9513159	-1.4689937	0.6147682	-2.0688745	0.2120320	0.3145840
Odilon.Kossonou	-3.3557775	-1.8572089	-0.6238234	1.1190921	-0.1643460	0.4192864	-0.5364553	-1.1702886	0.0355667
Mitchel.Bakker	-4.5658079	0.4558227	-1.5408014	1.8767927	-1.2527896	1.2099412	-0.9653969	-0.2809086	-0.2161807
Piero.Hincapie	4.9627523	-0.4111139	0.8909409	1.2646915	-0.1732569	0.8247576	0.4953631	-2.1529184	0.8346446
Jonathan.Tah	2.9042488	-4.8799852	-1.2787574	0.7333852	-0.5369788	0.5411283	0.2655455	-0.9787815	0.1975944
Jeremie.Frimpong	0.3347505	7.0231460	0.8124728	5.6169046	-5.2597309	-3.2458056	0.7588544	-0.5627832	-0.5710769
Edmond.Tapsoba	9.5135009	-2.7983856	-0.4255598	1.4139337	-1.5615791	-1.5082663	1.6251513	-0.2021419	1.2256708
Daley.Blind	-2.1057444	-2.6055458	-2.8821288	2.8255383	-0.1450455	-0.5810093	0.9637523	1.1853663	-0.4714399
Joao.Cancelo	6.3563969	2.0730124	-1.6741951	4.5525135	0.0513314	1.2385233	-0.2273378	0.2817977	-0.6171740
Noussair.Mazraoui	-5.0843186	0.0988226	-0.7836046	2.6773923	1.2166267	0.3290335	0.5391382	-0.8759216	0.3451929
Alphonso.Davies	2.5725716	1.4285073	-1.2688082	4.9356421	0.3002713	0.4589816	-1.7494722	-1.1721073	-1.7188665
Dayot.Upamecano	10.3616364	-3.6875675	-1.2341805	2.4029942	1.1477107	0.0786076	-1.0441191	0.0867869	-1.8944054
Matthijs.de.Ligt	6.6690874	-5.3702346	-1.9991184	1.8026617	-0.7150067	0.6321136	-0.3113966	-0.7905351	0.8537488
Benjamin.Pavard	9.8986904	-1.6002500	-0.3809300	2.3371240	2.1743573	1.2575696	1.6987636	-0.3338268	1.1472101
Kostas.Stafylidis	-7.4988397	1.0996342	0.1245013	-0.8555091	-0.0420175	0.0751170	-1.3302815	0.4288003	-0.0468880

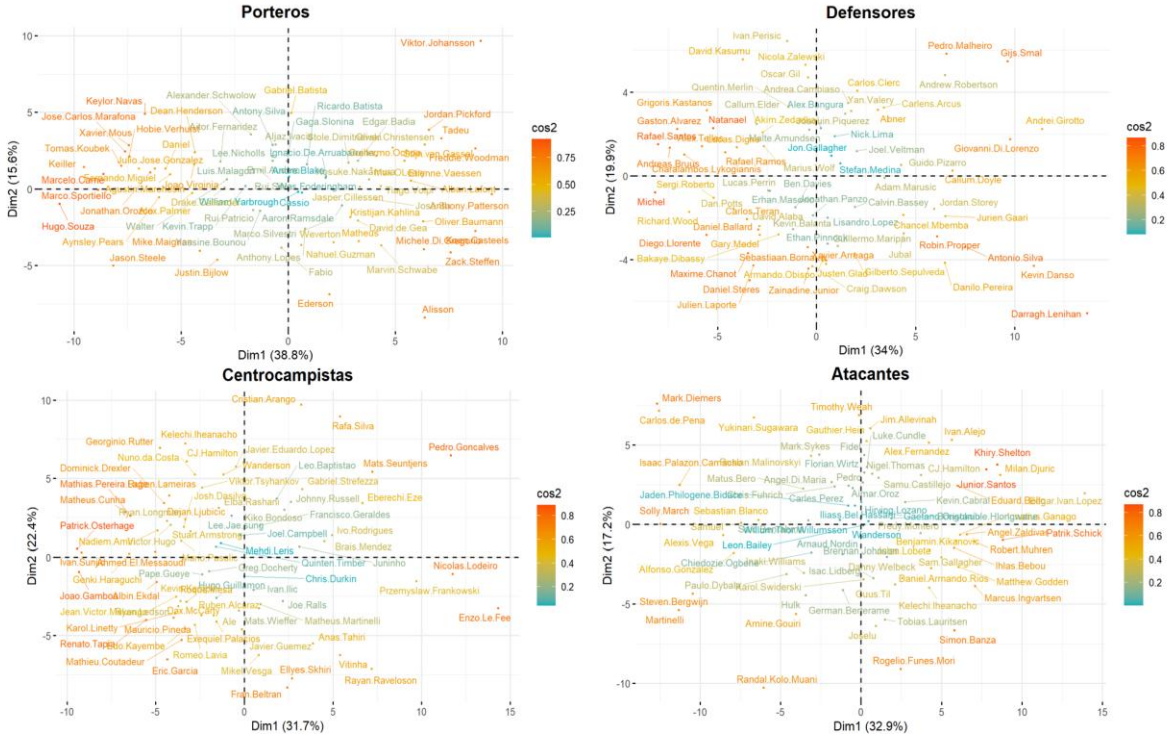
Fuente: Elaboración propia con R

Después de haber revisado el valor que toman cada uno de los futbolistas en sus respectivos datasets reducidos por posición, se procede a representarlos en gráficos de dispersión gracias a la función *fviz_pca_ind()* del mismo paquete *factoextra.* Donde, al igual que ocurría en la contribución de los campos originales sobre las componentes, el eje X hace referencia a la primera componente mientras que el eje Y a la segunda.

En dichas figuras el color de los puntos y los nombres de los jugadores está determinado por el valor que toma su coseno al cuadrado *(cos2)*, que indica la calidad de la representación de los datos. Así pues, para un individuo poseer un valor alto en este parámetro hace que se

encuentre muy bien representado en tal dimensión. Además, gracias a haber entendido la influencia que tienen las variables iniciales sobre la nueva, se puede interpretar fácilmente que campos originales están influyendo en la representación del jugador y, por tanto, obtener una primera percepción sobre las características de juego del propio futbolista.

Ilustración 3.6: Gráficos de contribuciones individuales de los jugadores a las componentes por puesto



Fuente: Elaboración propia con R

De este modo, se toma como ejemplo para explicar lo descrito previamente, la representación de los atacantes. Por lo tanto y en base a lo concluido en la contribución de las originales sobre estas componentes para la posición tomada, se puede observar que existen jugadores muy bien representados sobre la primera dimensión, tanto de forma directa como inversa. Para aquellos que guardan una relación positiva, se puede destacar a Patrick Schick. Por tanto, sobre el delantero checo cabe la posibilidad de afirmar, en base a los resultados obtenidos, que es un atacante poco participativo en el juego. En cambio, si se selecciona a Martinelli, dada la posición habitual del jugador brasileño (extremo izquierdo) se describe como un futbolista con alta participación en el juego y que suele generar bastante peligro al rival, esto último se asocia con su relación indirecta sobre la segunda dimensión.

3.5. SCORES DE SIMILITUD ENTRE JUGADORES

Una vez obtenidas los valores de cada jugador sobre las componentes para su posición y analizada la calidad de su representación en las dos primeras dimensiones, se pasa a calcular las puntuaciones de similitud respecto al resto de jugadores dentro de su puesto específico.

Para ello, se transpone cada uno de los datasets reducidos, con el fin de poder aplicar el Coeficiente de Correlación de Pearson sobre los valores de los jugadores y obtener sus *scores* de similitud en base a los resultados del ACP (*PCA_Score*).

Una vez conseguido esto, se transforma todo el dataset a una estructura vertical para tener todos los nombres de los jugadores por filas (observaciones), es decir, tanto aquel a comparar (*Player*) como con quien es comparado (*Player_Comp*), además de sus respectivos valores de similitud. A modo de ejemplo, se expone una tabla con los resultados para Sergio Busquets, haciendo uso del conjunto de datos reducido de los centrocampistas.

Ilustración 3.7: Tabla Scores de Similitud por PCA para Sergio Busquets

Player	Player_Comp	PCA_Score
Sergio.Busquets	Sergio.Busquets	1.0000
Sergio.Busquets	Jordy.Clasie	0.9552
Sergio.Busquets	Koke	0.9490
Sergio.Busquets	Marcel.Ruiz	0.9483
Sergio.Busquets	Rodri	0.9435
Sergio.Busquets	Pierre.Hojbjerg	0.9419
Sergio.Busquets	Victor.Wanyama	0.9417
Sergio.Busquets	Michael.Bradley	0.9383
Sergio.Busquets	Josh.Cullen	0.9360
Sergio.Busquets	William.Carvalho	0.9340

Fuente: Elaboración propia con R y Power Point. Imágen obtenida de fbref.com

Por consiguiente, para el que fue pivote del club blaugrana y la selección española, se observa que la mayor parte de los futbolistas con los que guarda alta similitud, para la muestra analizada, ocupan su misma posición o similar. Entre los jugadores a destacar, se encuentran Rodri o Koke con cerca de un 0.95 de correlación respecto a los valores de Busquets.

De esta manera, la gran utilidad que posee este análisis de reducción de la dimensionalidad es el hecho de que a partir de estos nuevos conjuntos de datos se ha podido extraer un score de similitud entre jugadores. Asimismo, se podrán aplicar nuevos algoritmos y técnicas que permitan obtener conclusiones mucho más específicas acerca de las cualidades de estos futbolistas, su estilo de juego y nuevas métricas de similitud.

Así pues, posteriormente se hará uso del análisis de conglomerados, en concreto del algoritmo K-medias, que posibilitará clasificar a los diferentes futbolistas en diferentes grupos dentro de su posición en función de las características contenidas en las nuevas componentes.

4. ANÁLISIS DE CONGLOMERADOS: K-MEDIAS

El análisis de conglomerados es una técnica estadística multivariante que tiene como objeto agrupar los individuos de un conjunto de datos intentando lograr la máxima homogeneidad en cada grupo y la mayor diferenciación entre estos.

Al igual que el ACP se tratan de una técnica estadística de aprendizaje no supervisado, siendo en este caso de clasificación. Es decir, a partir de un conjunto de observaciones no etiquetadas, trata de situar/clasificar estos registros en grupos homogéneos, denominados *clusters* o conglomerados. De manera que, los jugadores que son considerados con cualidades similares son asignados a un mismo grupo, mientras que, individuos disimilares se localicen en otros distintos.

Para este estudio se ha optado por seleccionar una técnica de reasignación dentro de los métodos no jerárquicos de agrupación, como es el algoritmos K-medias o *K-means*, en inglés. Este enfoque se diferencia de aquellos que, si son jerárquicos, principalmente, en que se encuentran diseñados para la agrupación en un número determinado de grupos, necesitando este ser especificado a priori (criterio de selección). Además, otra diferencia importante es que aquí se trabaja con el conjunto de datos original y no se necesita convertir a una matriz de distancias.

Como se ha descrito anteriormente, este algoritmo basa su funcionamiento en el método de la reasignación, lo que permite que un individuo que es clasificado en un grupo, en un determinado paso del proceso, pueda ser reasignado a otro *cluster* en otro momento si se cumple la optimización del criterio de selección. Así pues, el algoritmo finaliza una vez que ya no existen individuos (jugadores) cuya reasignación pueda mejorar (optimizar) el resultado conseguido.

El fin con el que se aplica K-medias es el de conseguir que los datos se clasifiquen en un número de conglomerados determinado y fijado a priori en "*K"*. El algoritmo opera en bucle, es decir de manera iterativa, para asignar y reasignar cada registro en uno de los *K* grupos, en función de las características proporcionadas y la similitud que poseen los individuos sobre estas.

Un elemento clave de estos conglomerados son sus centroides, que pueden ser definidos como el punto representativo de cada *cluster* y cuya distancia con los individuos que pertenecen a cada uno de los grupos es la menor posible (criterio de optimización).

La etapas que se deben aplicar para proceder con su ejecución son las siguientes:

1. Seleccionar el número de grupos/centroides óptimo a alcanzar (*"K")*
2. Indicar *K* e inicializar el algoritmo para que asigne cada punto al centroide cuya distancia sea la menor
3. Actualizar la posición de los *K* centroides, calculando la posición promedio de todos los individuos que pertenecen a esa clase
4. Comenzar el proceso de reasignación, esto es repetir pasos dos y tres hasta que los centroides no cambien de posición y, por tanto, la clasificación de los puntos sea la óptima, o se alcance el número máximo de iteraciones (criterio de parada)

4.1. SELECCIÓN DEL NÚMERO DE *CLUSTERS*: MÉTODO DEL CODO

La selección del número de *clusters* consiste en encontrar el número de particiones óptimas (conglomerados) en los datos cuando se desconoce de antemano la cantidad necesaria. Es importante destacar que no existe una técnica verdadera o exacta que demuestre el número específico de centroides a seleccionar. Sin embargo, se puede obtener una estimación fiable y precisa utilizando algunas de las siguiente técnicas; como son el método del codo, la validación cruzada, los criterios de información o el de la silueta, siendo esta última medida de gran utilidad para medir la calidad de la clasificación.

En este análisis se ha optado por seleccionar el Método del Codo (*Elbow Method*) dado que es el más habitual entre la comunidad científica y basa la selección en aquel punto donde la suma de la distorsión de cada cluster no genera variaciones significativas. Esto es lo que se conoce como la inercia, que es el valor obtenido del sumatorio de las distancias al cuadrado de

cada individuo al centroide de su cluster correspondiente. Por lo tanto, tras aplicar varias veces K-medias para diferentes valores de *K*, la elección óptima está en aquel número de conglomerados donde ya no existe una mejora considerable en la reducción de la inercia (codo).

Ilustración 4.1: Fórmula de la Inercia

$$Inertia = \sum_{j=1}^{K} \sum_{i=1}^{N} \left\| x_i^{(j)} - c_j \right\|^2$$

donde:

K : *número máximo de clusters,*

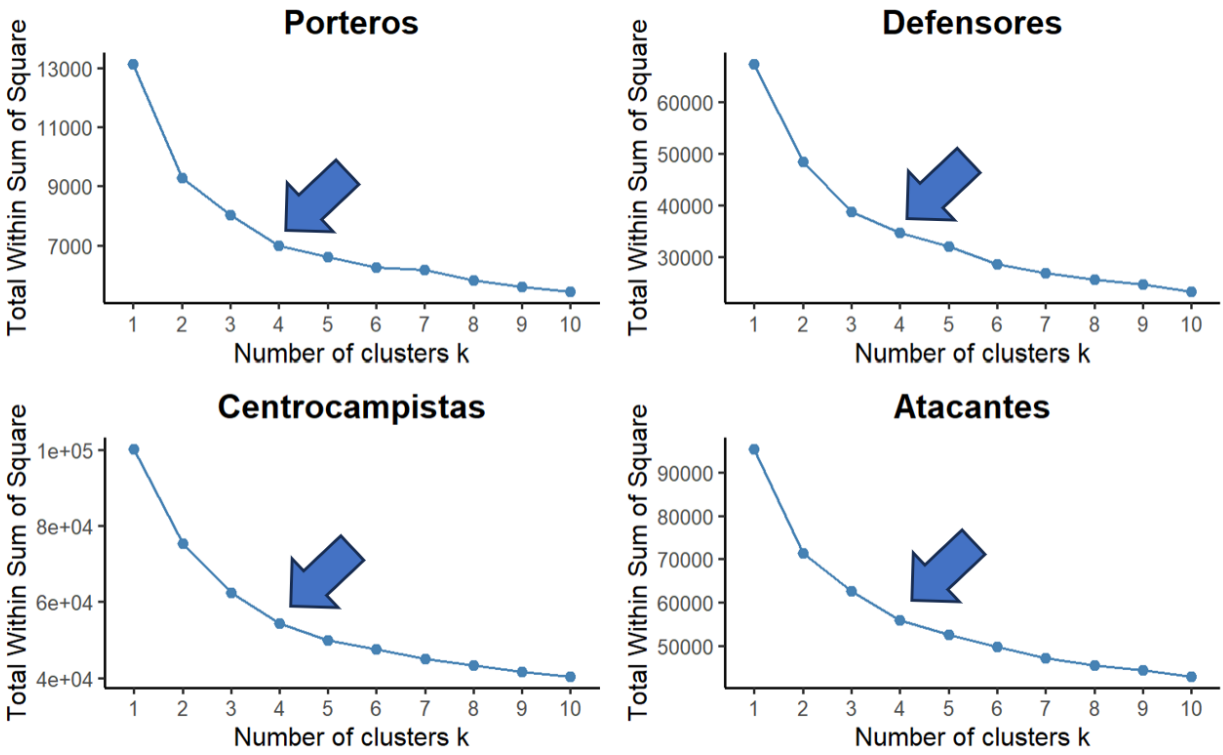
N : *número de registros*

$\left\| x_i^{(j)} - c_j \right\|$: *distancia del individuo$_i$, perteneciente al cluster j, al centroide*

Fuente: Estadística y Matemáticas aplicado al deporte con R. M3. Modelización y análisis final (p. 29), por J. Fernández, 2021, Sport Data Campus.

A continuación, se procede a mostrar cada uno de los gráficos por puesto para los que se aplica está técnica de selección. Como se podrá observar en las siguientes representaciones, su línea tiene una tendencia decreciente, ya que, a mayor número de grupos, existirá una menor distancia entre los registros y el centroide del grupo, siendo la inercia menor.

Ilustración 4.2: Representación del Método del Codo por posición



Fuente: Elaboración propia con R

Analizados los cuatro gráficos por puestos, se concluye que para todos ellos el número óptimo de *clusters* a seleccionar son cuatro, puesto que es el punto donde no hay variaciones importante en la reducción de la inercia.

Una vez decidido cuantos *clusters* se van a utilizar en la aplicación del algoritmo para cada posición, se procede a evaluar la calidad del agrupamiento para así determinar si la cantidad elegida es la más idónea, cabiendo la posibilidad de optar por otro número.

4.2. EVALUACIÓN Y CREACIÓN DE LAS PARTICIONES

Antes de aplicar el propio algoritmo K-medias, se considera esencial evaluar la calidad del agrupamiento para los cuatro conglomerados elegidos en todas las posiciones y así poder decidir definitivamente si la elección de K es la óptima. Para ello, se usa el método de la silueta.

Esta métrica se encuentra comprendida entre -1 y 1, de tal manera que asigna un valor en dicho rango a cada individuo en función del *cluster* en el que se encuentre. Valores próximos a 1 significan que el jugador se encuentra bien clasificado en dicha clase y que, por tanto, guarda poca relación con los grupos próximos (vecinos). En cambio, un registro cercano a -1 señala que ese individuo debería de haber formado parte de otro grupo vecino, dado que la distancia del individuo al cluster más próximo es menor que a puntos de su propio conglomerado.

Se puede definir el valor de la silueta de la siguiente forma:

Ilustración 4.3: Fórmula de la Silueta

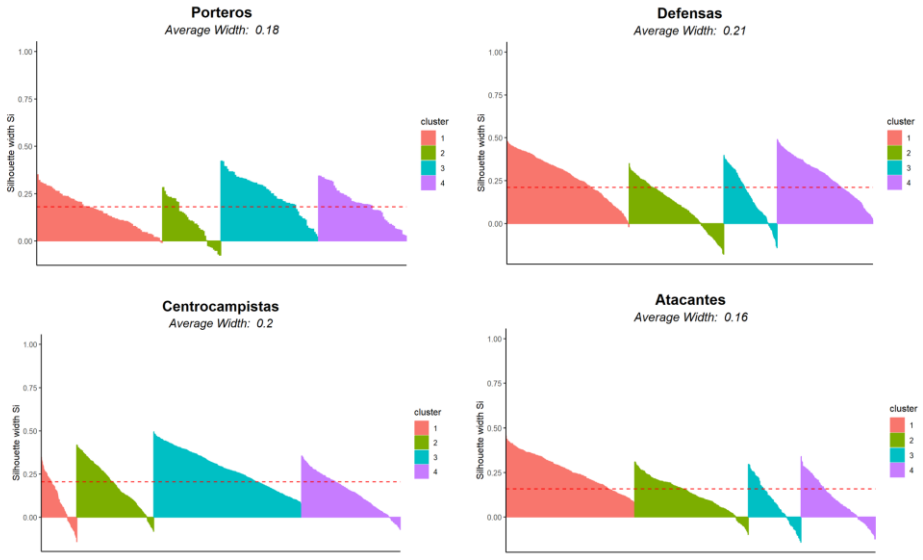
$$s(x) = \frac{b(x) - a(x)}{max|a(x), b(x)|}$$

donde:
b(x) = distancia promedio del elemento x a todos los demás puntos del mismo cluster.
a(x) = distancia promedio del elemento x a todos los demás puntos en el cluster más cercano.

Fuente: Estadística y Matemáticas aplicado al deporte con R. M3. Modelización y análisis final (p. 65), por J. Fernández, 2021, Sport Data Campus.

Así pues, una vez explicada la técnica para la evaluación de la calidad del agrupamiento, mediante el valor de la silueta, se procede a representar para cada posición cuatro gráficos que permiten analizar esto. En el eje X de los gráficos se representan a cada uno de los individuos (jugadores), mientras que el Y se expone su valor de silueta para el grupo en el que se encuentra. La línea roja horizontal y discontinua representa el promedio resultante del valor de la silueta para el total para todas las agrupaciones.

Ilustración 4.4: Gráficos por puesto para la evaluación de las agrupaciones a través del método de la silueta



Fuente: Elaboración propia con R

Representados los cuatro diagramas, se observa que ninguna posición obtiene un valor medio de silueta próximo a cero, lo que es indicativo de que los agrupamientos no poseen una calidad elevada. Sin embargo, se debe destacar el hecho de que se está tratando de agrupar aproximadamente más de mil de jugadores para los diferentes puesto, por lo que la probabilidad de que existan jugadores mal clasificados (valores negativos) es más alta, lo que hace que se reduzca dicho promedio. Además, se tiene que señalar que se ha intentado utilizar otras cantidades de clusters distintas, siendo esta la que producido los valores óptimos.

Para obtener estas representaciones, se ha utilizado la función *eclust()* del paquete *factoextra* la cual permite indicar el tipo de clusterización a utilizar ("k*means"*) y el número de conglomerados (cuatro). Además, dicha función no solo permite la evaluación de las particiones si no que posibilita la creación directa de los datasets con los resultados obtenidos por medio de haber aplicado el algoritmo. En los conjuntos de datos resultantes se encuentra incluido el cluster al que pertenece cada jugador y su vecino más cercano, pero no el valor de silueta. Para acceder a este, se debe indicar sobre los resultados alcanzados el método *$silinfo$widths*, y de esta manera pasar a concatenarlo al dataset anterior.

A continuación, se procede a mostrar un ejemplo de la estructura del dataset resultante para los defensores:

Ilustración 4.5: Dataset de defensores sobre los resultados del algoritmo K-medias

Player	Cluster	Neighbor	Sil_Score
Luis.Manuel.Orejuela	1	4	0.4797380
Hugo	1	4	0.4781148
Nacho.Vidal	1	4	0.4754093
Jere.Uronen	1	4	0.4733013
Robin.Gosens	1	4	0.4685521
Emerson.Palmieri	1	4	0.4670157
Paulo.Vitor	1	4	0.4650087
Franco.Escobar	1	4	0.4635695
Pedrinho	1	4	0.4625165
Banzouzi.Locko	1	4	0.4614895
Samuele.Birindelli	1	4	0.4603684
Ray.Gaddis	1	4	0.4603481
Zaidu.Sanusi	1	4	0.4557766
Junior.Dina.Ebimbe	1	4	0.4550980
Salvador.Reyes.Chavez	1	4	0.4530792

Fuente: Elaboración propia con R

Asimismo, se considera importante destacar y recordar, que la aplicación de este algoritmo se realiza sobre los nuevos datos obtenidos a partir de la reducción de la dimensionalidad, aplicada en el apartado anterior (ACP). Por ello, se debe remarcar que los datos ya se encuentran escalados (media cero y desviación estándar uno), algo esencial, ya que, si hay campos en escalas muy diferentes, los atributos con métricas altas dominarán sobre el cálculo de las distancias.

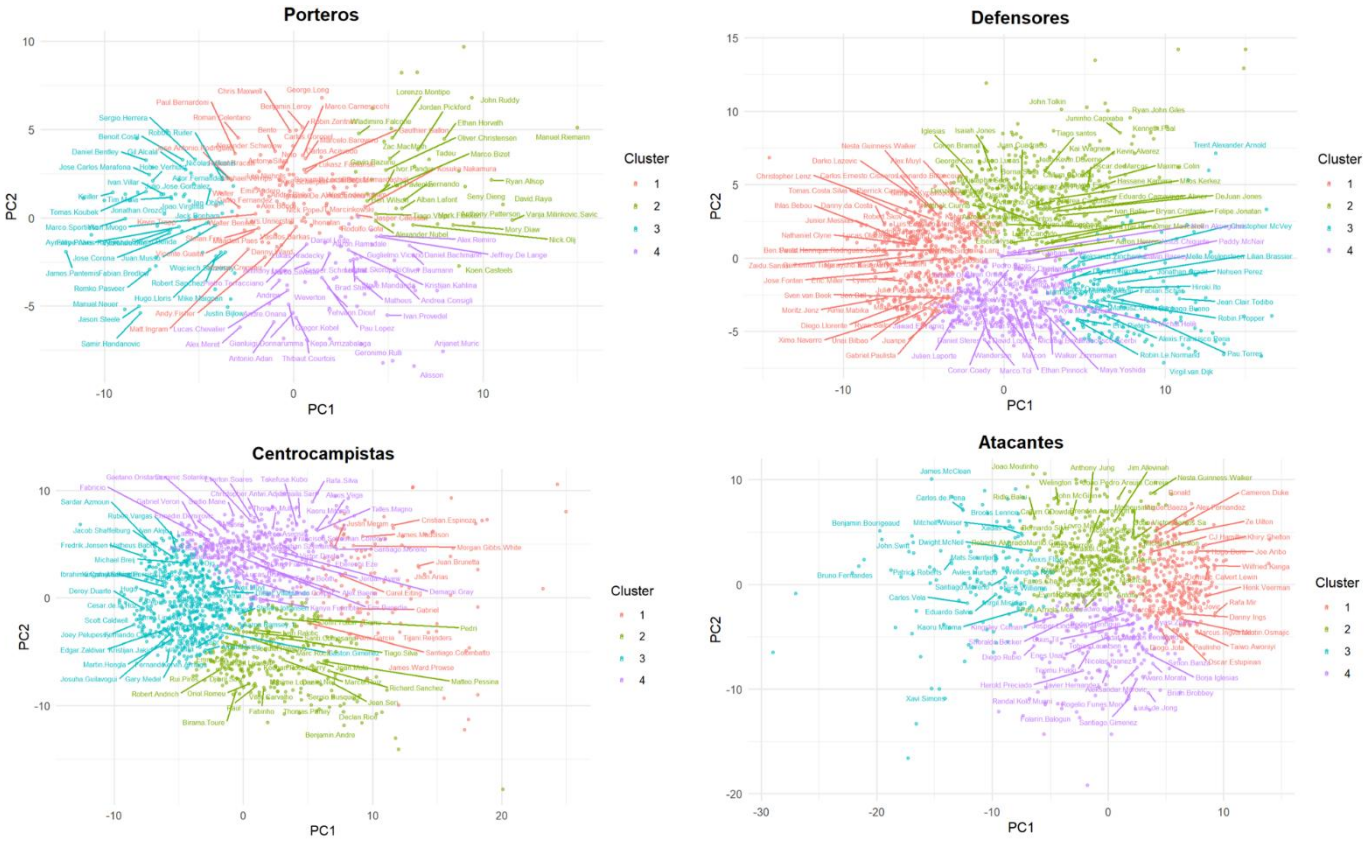
Por otra parte, también se utiliza el menor número de campos posibles, haciendo frente al problema de dimensionalidad. Esto significa que a medida que aumenta el número de características las distancias tienen menor capacidad para discriminar entre clases y por tanto clasificar mejor el punto en los diferentes grupos, dado que la variabilidad de las distancias entre los datos disminuye significativamente.

Después de haber obtenido los resultados, se procede a visualizar estos con el objetivo de extraer una serie de conclusiones generales sobre las características de los jugadores en función del conglomerado en el que haya sido clasificado.

4.3. VISUALIZACIÓN DEL ALGORITMO K-MEDIAS

Se opta por mostrar a los jugadores para cada posición sobre las primera dos dimensiones obtenidas en la aplicación del ACP.

Ilustración 4.6: Visualización de los clusters para las cuatro posiciones



Fuente: Elaboración propia con R

Así pues, se observa que los cuatro conglomerados para las diferentes posiciones se encuentran bastante bien diferenciados, aunque con algunas excepciones dada la gran cantidad de jugadores a representar. Dado que, se están utilizando las dimensiones del ACP en la representación, esto posibilita comprender mejor la información contenida en cada *cluster*, gracias a la interpretación realizada sobre la contribución de los campos originales en las nuevas componentes.

De tal manera que, para los guardametas, teniendo en cuenta las dos primeras dimensiones, se extrae que en el primer grupo (rojo) se encuentran aquellos porteros que posee una alta precisión en el juego largo con los pies y están habituados a este tipo de estilo. En el segundo de los conglomerados (verde) están aquellos que no quieren tomar riesgos cuando

poseen el balón y que, por tanto, buscan el juego en largo sin importar el resultado del pase. En el tercero de los grupos (azul) hace referencia a los cancerberos con características opuesta a los del segundo, es decir, jugadores que tratan de jugar el balón en corto, intentando ser eficaces en la salida. Por último, la cuarta de las clases (violeta) hace referencia a aquellos porteros altamente participativos en el juego, es decir, que poseen muchos toques de balón durante los partidos y que además suelen mantener una posición avanzada durante los partidos. Así pues, como se ha explicado anteriormente la dos primeras dimensiones para el caso de los porteros están altamente influenciadas por campos relacionados con el juego de pies.

Considerando a los defensores, se define el primer grupo (rojo) como aquel donde se encuentran los jugadores que no son tan participativos en el juego en cuanto pases, toques de balón o conducciones, cuyo estilo de juego es encontrarse replegados y ser contundentes en fases sin balón. En el segundo (verde), se localizan aquellos jugadores de la zaga que poseen un juego más arriesgado y ofensivo, cometiendo mayores errores con balón, como son controles malos o perdidas y, que buscan continuamente centros en zonas de ataque. Por consiguiente, en este conglomerado se hace sitúan principalmente los laterales. El tercero (azul) es la clase opuesta respecto de la primera, es decir son futbolistas que son muy participativos en el juego con balón, es decir, suelen dar muchos pases en los partidos, se involucran en la salida en corto de la pelota y realizan conducciones. De igual modo, acostumbran a jugar en posiciones muy adelantadas necesitando ser muy precisos en los duelos, intercepciones o robos. En el cuarto *cluster* (violeta) se sitúan aquellos jugadores que son también muy participativos pero que no arriesgan tanto, es decir, tratan de jugar más en largo. Además, suelen estar más implicados en tareas defensivas del tipo, duelos aéreos, bloqueos o despejes. Probablemente, sea esto debido a que sus equipos no buscan llevar tanto el peso del partido como en los del tercer grupo.

En el primero de los conglomerados para los centrocampistas (rojo), se encuentran aquellos futbolistas que tratan de realizar conducciones largas con el balón y que juegan cerca del área rival tratando de dar pases claves o crear acciones del gol. En el segundo de los grupos (verde), se encuentran los mediocentros posicionales, aquellos que dan el mayor número de pases y, por tanto, con los que el balón recorre más kilómetros tratando de conectar la salida del juego con las zonas de finalización. El tercer *cluster* (azul) es para los centrocampistas que no son tan participativos en el juego con posesión, en otras palabras, sus puntos fuertes están en el juego sin balón, seguramente dada la táctica de sus equipo, los cuales no le dan tanta importancia a la posesión. Finalmente, el cuarto (violeta) alude a los centrocampistas ofensivos, esto es tanto mediocentros de banda como volantes, aquellos que podrían jugar como delanteros o extremos. Son los jugadores que conducen las jugadas de ataque de sus equipos, en definitiva, aquellos que dan los pases progresivos que introducen el balón dentro del área y que, además, posee altos registros en métricas ofensivas, como asistencias, goles, etc.

Por último, en cuanto a los atacantes, se observa que en el primer grupo (rojo) están aquellos delanteros que son poco interventores del juego, es decir, quienes pueden ser calificados como atacantes puramente rematadores, que fijan a los centrales rivales. En el segundo (verde) se encuentran los atacantes habilidosos, que son altamente dribladores y que se caracterizan por superar rivales o líneas defensivas ellos mismo, ya sea por potencia y velocidad, o puro talento. En el tercer conglomerado (azul) se encuentran los atacantes, que pueden ocupar la posición de delantero centro, pero que son mucho más móviles y que, por tanto, también suelen jugar como centrocampistas ofensivos o enganches. Por consiguiente,

son aquellos futbolistas que permiten el último pase, que generan asistencias, anotan goles, y que habitualmente tratan de crear peligro con el disparo a media distancia, etc. Finalmente, en la cuarta categoría (violeta), al igual que en el primero, se asientan los puntas más comunes o que buscan más ese último toque para introducir el balón a la red. Sin embargo, estos jugadores son más participativos del juego de sus equipos, viniendo a recibir del centro del campo y tratando de dar muchos más pases.

5. APLICACIÓN DE ALGORITMOS DE SIMILITUD

Tras haber agrupado a cada uno de los jugadores en base a sus posiciones y descrito estos grupos, se pasa a aplicar nuevos algoritmos de similitud en base a las distancias euclídea y coseno. Estos métodos seguirán siendo ejecutados sobre los datasets con dimensiones reducidas, con el objetivo de obtener nuevos resultados que permitan aproximar y encontrar puntuaciones similares de un futbolista con otro, en función de las dimensiones obtenidas en el ACP a partir de las métricas originales de rendimiento.

5.1. MÉTRICAS DE SIMILARIDAD

Como se ha mencionado previamente, las métricas de similitud que van a ser empleadas se basan en las distancias euclídea y coseno.

Puesto que, existen diferencias entre ambas con relación a la forma de ser calculadas y a su interpretabilidad, a continuación, se procede a realizar una explicación detallada de lo que conlleva el uso de cada una de ellas.

5.1.1. Similitud a través de distancia euclídea

La distancia euclídea es la medida de distancia más habitual, y mide la diferencia absoluta entre dos individuos en un espacio multidimensional, en el caso de que existan más de dos características en el estudio. Así pues, a mayor distancia en el plano para cada par de observaciones, mayor diferencia habrá entre ellos y, por tanto, menor será el valor de la similitud para ellos.

Ilustración 5.1: Fórmula distancia euclídea

$$d(X,Y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \ \epsilon \ [0, Infinito)$$

Fuente: Estadística y Matemáticas aplicado al deporte con R. M3. Modelización y análisis final (p. 38), por J. Fernández, 2021, Sport Data Campus.

Como se observa en la fórmula, los valores que resultan de esta se encuentran comprendidos entre cero e infinito. De tal manera que, se deben transformar estos resultados en una medida de similitud, sin disponer de un rango definido. Para ello, se utilizará una transformación basada en el percentil 95 con el objetivo de retirar posibles valores extremos de la muestra que estén afectado en los resultados de las distancias (Fernández, 2021, p.42).

Ilustración 5.2: Fórmula similitud en base a distancia euclídea

$$Similitud_{euclidea} = \left(1 - \frac{d}{d_{P95}}\right) * 100$$

Fuente: Estadística y Matemáticas aplicado al deporte con R. M3. Modelización y análisis final (p. 41), por J. Fernández, 2021, Sport Data Campus.

Por ende, esta fórmula busca obtener la similitud a partir del valor del cociente entre la diferencia de puntos del jugador objetivo con el futbolista comparar y el valor de distancia que mantiene con el deportista en el percentil 95. De tal manera, que se consigue su transformación porcentual y, así acotar o definir el rango de similitudes.

5.1.2. Similitud a través de distancia coseno

Para el cálculo de la otra distancia considerada en el análisis, se utiliza el valor del coseno procedente del ángulo entre dos vectores en el espacio dado. Así pues, esta medida se encuentra más condicionada por la dirección que toma la línea entre los dos puntos, que simplemente por la longitud de la recta.

Ilustración 5.3: Fórmula distancia euclídea

$$sim(X,Y) = \cos\theta = \frac{\vec{x} * \vec{y}}{\|x\| \, \|y\|} \epsilon [-1,1]$$

Fuente: Estadística y Matemáticas aplicado al deporte con R. M3. Modelización y análisis final (p. 38), por J. Fernández, 2021, Sport Data Campus.

Como se observa en su fórmula, la distancia coseno si tiene un rango de valores definido, que se encuentra entre menos uno y uno. De manera que, un valor positivo de la unidad significaría una semejanza exacta. Al igual que para el caso anterior, también será necesaria su transformación a valores porcentuales de similitud (0 a 100%). De todos modos, su cambio es más sencillo, ya que se utiliza la normalización *MinMax*, para si transformar los valores a una escala comprendida entre cero y uno, para posteriormente multiplicar por 100 y así pasar definitivamente los valores a porcentajes.

Ilustración 5.4: Fórmula similitud en base a distancia coseno (normalización MinMax)

$$Similitud_{coseno} = \left(\frac{d - d_{min}}{d_{max} - d_{min}}\right) * 100$$

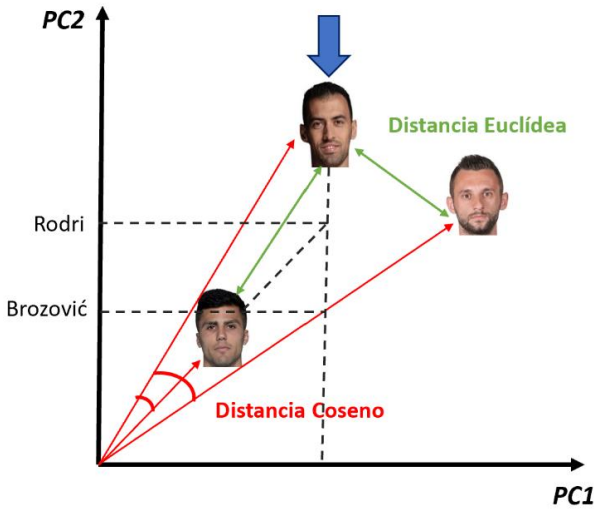
Fuente: Estadística y Matemáticas aplicado al deporte con R. M3. Modelización y análisis final (p. 43), por J. Fernández, 2021, Sport Data Campus.

5.1.3. Interpretación de las distancias

Una vez explicado el modo de calcular ambas similitudes, a partir de esas dos distancias. Se considera de gran importancia no solo aclarar las discrepancias existentes entre ambas si no también la forma de interpretarlas.

A modo de ejemplo, se hará uso de una ilustración muy simple donde se muestran y comparan ambas distancias, representando sobre el plano de las dos componentes principales y de forma esquemática las propias distancias entre Busquets, Rodri y Brozović.

Ilustración 5.5: Gráfico ilustrativo para la interpretación de distancia euclídea y coseno



Fuente: Elaboración propia con Power Point, a partir de información en Estadística y Matemáticas aplicado al deporte con R. M3. Modelización y análisis final (p. 43), por J. Fernández, 2021, Sport Data Campus.

En base a esta representación y tomando como jugador objetivo al "ex-cinco" blaugrana, se observa que si se selecciona la distancia euclídea como métrica de estudio el jugador más similar a este (menor diferencia) es Marcelo Brozović, dado que por definición de esta el jugador croata se encuentra más próximo en el plano que el centrocampista "*cityzen*". En cambio, si se opta por la distancia coseno, quien tendría mayor parecido con el jugador analizado es el propio Rodri Hernández, dado que el ángulo formado por sus vectores es más pequeño que para el jugador balcánico.

Así pues, se ha demostrado la importancia de interpretar ambas métricas de manera correcta. En este escenario, se demuestra que Brozović es superior a Rodri en ambas dimensiones y, por tanto, en base a la distancia en el plano, más parecido. De todos modos, si en este caso ilustrativo Rodri goza de unas oportunidades de juego similares a las de Sergio este sería más similar, en base a la definición de la distancia coseno.

Finalmente, es importante remarcar que ambas medidas pertenecen a categorías distintas, siendo la distancia coseno una métrica para calcular la similitud en el plano, mientras que, la euclídea, como se ha visto, la distancia entre dos puntos. Es por ello, que para este tipo de análisis se utiliza de manera más habitual la similitud coseno, además de que posibilita la obtención de unos porcentajes más altos y genera más fiabilidad a la hora de interpretar los resultados.

De cualquier forma, esto no significa que exista una métrica más robusta que otra, por tanto, la práctica más eficaz es utilizar los métodos expuestos y analizar sus resultados. Por

tanto, una vez hecho esto se puede concluir que para aquellos futbolistas que posean un porcentaje alto en ambas técnicas son similares al analizado.

5.2. ELABORACIÓN DEL ALGORITMO

En esta sección se explica la forma en la que se ha construido la función en R que permite aplicar el algoritmo de similitud empleado para este proyecto. Las métricas de similitud adoptadas sobre este son las vistas previamente, euclídea y coseno.

La función que controla el algoritmo se denomina similarity_algorithm() y tan solo contiene tres parámetros, con el fin de simplificar su uso para la ejecución de ambas similitudes en todos los puestos. Asimismo, dado todos los procedimientos expuestos a lo largo de este documento, de selección de características, limpieza de datos o de reducción de la dimensionalidad; no es necesario aplicar ningún criterio de filtrado sobre esta.

En consecuencia, los valores a recibir por la función son los siguientes:

- *df_pca* → Conjunto de datos con las nuevas componentes y dimensionalidad reducida.
- *player* → Jugador objetivo a encontrar otros similares a este.
- *distance* → Distancia a seleccionar entre euclídea y coseno

Establecidos los parámetros de la función. Se procede a explicar su puesta en marcha. Debido a que es necesario aplicar dos métricas distintas sobre cuatro datasets por puestos, se ha optado por crear una función denominada *dataset_similarity()* que posibilite, de la manera lo más eficiente posible, su propia ejecución. Así pues, los parámetros que deben ser ingresados en esta función son los mismos que en la anterior, exceptuando que en este caso se le aplica un vector con los jugadores en lugar de un único jugador.

La inclusión de este es necesaria, ya que, la función contiene un bucle *"for"* a través del cual se inicializa el algoritmo anterior para cada jugador que forma parte de este conjunto. Posteriormente, se aplica cada uno de los métodos para el cálculo de similitudes, previamente fijados en la propia función. Para finalmente, concatenar los resultados de cada futbolista, contenidos en ese vector, en un dataset final por posición y métrica de similitud, que será lo que devuelve la propia función.

La ejecución de la función en la que reside el propio algoritmo (*similarity_algorithm*) y, contenida en la iteración explicada previamente, es bastante simple. Por un lado, se aplica la métrica de similitud seleccionada. Para posteriormente aplicar la transformaciones porcentuales de los valores originales. De tal manera que, por último, se prepara un conjunto de datos donde aparece el nombre del jugador a comparar, aquel con quien es comparado y su métrica de similitud transformada.

Por consiguiente, definidas estas dos funciones, los conjuntos de datos resultantes, por puesto y por medida de similitud, se obtienen en base a aplicar dataset_similarity(). Dicha función, va a ser aplicada un total de ocho veces sobre el código, es decir, para cada puesto y tipo de medida; de manera que se le indique, el conjunto de datos reducido para la posición, el puesto y la distancia seleccionada.

En último término, se decide exponer dos ejemplos de datasets obtenidos a partir de la aplicación de esta función y el algoritmo de similitud incluido en ella, para ambos métodos. Así pues, para el primero de ellos, se opta por presentar los resultados del Ilkay Gündogan con

relación a los valores obtenidos en la aplicación de la similitud euclídea, en base a los datos para los centrocampistas. Asimismo, en dichos resultados se puede observar que los valores obtenidos son claramente inferiores en comparación con la similitud por coseno. Para esta última, se decide mostrar los resultados de Karim Benzema, alcanzados de acuerdo con los datos contenidos en el dataset reducido para los atacantes.

Ilustración 5.6: Ejemplos de resultados similitudes euclídea y coseno para Gündogan y Benzema, respectivamente

Player	Player_Comp	Euc_Similarity
Ilkay.Gundogan	Ilkay.Gundogan	100.00000
Ilkay.Gundogan	Daichi.Kamada	70.86642
Ilkay.Gundogan	Granit.Xhaka	70.24054
Ilkay.Gundogan	Angel.Gomes	65.04518
Ilkay.Gundogan	Fredrik.Aursnes	64.86141
Ilkay.Gundogan	Luka.Modric	61.78915
Ilkay.Gundogan	Yimmi.Chara	61.15595
Ilkay.Gundogan	Bernardo.Silva	60.96059
Ilkay.Gundogan	Rodrigo.De.Paul	60.71223
Ilkay.Gundogan	David.Silva	59.67171
Ilkay.Gundogan	Dani.Ceballos	59.62490
Ilkay.Gundogan	Alexander.Ring	59.36610
Ilkay.Gundogan	Ganso	59.26207
Ilkay.Gundogan	Cristian.Roldan	58.93650
Ilkay.Gundogan	Adrien.Thomasson	58.74676

Player	Player_Comp	Cos_Similarity
Karim.Benzema	Karim.Benzema	100.000
Karim.Benzema	Gabriel.Barbosa	90.970
Karim.Benzema	Jonathan.David	89.978
Karim.Benzema	Mehdi.Taremi	89.337
Karim.Benzema	Christopher.Nkunku	87.535
Karim.Benzema	Robert.Lewandowski	87.286
Karim.Benzema	Folarin.Balogun	86.517
Karim.Benzema	Alexandre.Lacazette	85.984
Karim.Benzema	Andrej.Kramaric	85.840
Karim.Benzema	Javier.Hernandez	84.891
Karim.Benzema	Kylian.Mbappe	84.439
Karim.Benzema	Wissam.Ben.Yedder	84.025
Karim.Benzema	Erling.Haaland	84.004
Karim.Benzema	Jeremy.Ebobisse	83.774
Karim.Benzema	Romell.Quioto	83.168

Fuente: Elaboración propia con R y Power Point. Imágenes obtenidas de fbref.com

6. CASO DE USO: HERRAMIENTA EN POWER BI

Una vez obtenidos todos los resultados finales tras la aplicación de las diferentes técnicas de análisis no supervisado y algoritmos de similitud, se procede a crear una herramienta en Power BI que sirva de cuadro de mandos y, sobre la que se aplican los distintos filtros automatizados para la búsqueda de jugadores similares a uno objetivo.

6.1. PREPARACIÓN DE LOS DATOS FINALES

Antes de seguir adelante con la construcción de la aplicación, se deben realizar un pequeño procesamiento en R sobre los conjuntos de datos definitivos, que han sido conseguidos en los apartados previos, con el objetivo de preparar estos de cara a su correcto formato y funcionamiento en Power BI.

Para esta limpieza y transformación de los datos finales se construye y se hace uso de la función *join_sim_general_data()*. Así pues, el funcionamiento de esta consiste, primeramente, en unir todos los datasets obtenidos para cada análisis (ACP, K-medias, Similitud euclídea y Similitud coseno) por puesto, junto con la URL de FBref para el jugador comparado con el objetivo. De tal manera que, se puedan unir sus datos generales en el esquema de la aplicación, dentro de Power BI.

Posteriormente se realiza una limpieza sobre los nombres de los jugadores, tanto del jugador a comparar como de aquellos que son evaluados sobre este para conocer su similitud. En este tratamiento se elimina el punto que existe entre el nombre y el apellido de los

futbolistas, además de algunos números que aparecen en sus nombres. Esto fue debido a que, en el proceso de reducción de la dimensionalidad, se decidió establecer como índice de los datasets el nombre que aparece al final de los links de FBref, para poder representar a estos en las visualizaciones de las contribuciones individuales. Además, a causa de que el nombre de una fila en R debe ser único, se tuvo que aplicar la función *make.names()* del paquete base, con el fin de transformar dichos registros a valores exclusivos, generando estos símbolos de puntuación y números para diferenciarlos.

El siguiente paso que se aplica, es el de crear un *score* de similitud general para todos los jugadores y, sobre el que se basa el análisis en la herramienta de visualización. En otras palabras, se considera que es el valor más importante para determinar el grado de similitud de un jugador respecto a otro. Para su cálculo, se realiza una normalización *MinMax* los resultados del ACP, puesto que son correlaciones comprendidas entre menos uno y uno; además de la similitud euclídea. Una vez hecho esto, para obtener el valor definitivo de esta métrica se aplica la media aritmética sobre las tres medidas de similitud.

El último paso, aplicado sobre dicha función para finalizar el procesamiento, es el de añadir el nombre del equipo, entre paréntesis, al propio nombre del jugador objetivo, para que así, se posibilite dentro del cuadro de mandos buscar al jugador por el equipo o ver el conjunto de futbolistas por puesto que existen para ese equipo.

Una vez, aplicada la función para cada conjunto de datos final por puesto, se obtienen unas tablas con estructura exacta a la muestra expuesta a continuación para Sergio Ramos como jugador objetivo, en base a los datos para los defensores:

Ilustración 6.1: Ejemplo de dataset final para uso en herramienta con Sergio Ramos como jugador objetivo

UrlFBref	Player	Player_Comp	Sim_Score	PCA_Score	Euc_Similarity	Cos_Similarity	Cluster	Neighbor	Sil_Score
https://fbref.com/en/players/08511d65/Sergio-Ramos	Sergio Ramos (Paris Saint-Germain)	Sergio Ramos	100.00	100.00	100.00	100.00	3	4	0.327
https://fbref.com/en/players/df8d6029/Arthur-Theate	Sergio Ramos (Paris Saint-Germain)	Arthur Theate	96.59	98.12	93.67	97.99	3	4	0.328
https://fbref.com/en/players/33651873/Goncalo-Inacio	Sergio Ramos (Paris Saint-Germain)	Goncalo Inacio	96.43	98.34	92.61	98.33	3	4	0.339
https://fbref.com/en/players/d248cd8f/Dayot-Upamecano	Sergio Ramos (Paris Saint-Germain)	Dayot Upamecano	95.74	97.58	92.25	97.39	3	4	0.374
https://fbref.com/en/players/6f5ec8bb/Facundo-Medina	Sergio Ramos (Paris Saint-Germain)	Facundo Medina	95.08	96.66	91.95	96.64	3	4	0.291
https://fbref.com/en/players/4b3e1a38/Amir-Rrahmani	Sergio Ramos (Paris Saint-Germain)	Amir Rrahmani	95.08	97.52	90.54	97.17	3	4	0.240
https://fbref.com/en/players/20030c06/Willi-Orban	Sergio Ramos (Paris Saint-Germain)	Willi Orban	94.99	96.61	91.76	96.59	3	4	0.378
https://fbref.com/en/players/47327321/Taylor-Harwood-Bellis	Sergio Ramos (Paris Saint-Germain)	Taylor Harwood Bellis	94.92	96.55	91.69	96.52	3	4	0.343
https://fbref.com/en/players/06c48679/Jonathan-Gradit	Sergio Ramos (Paris Saint-Germain)	Jonathan Gradit	94.71	97.18	90.33	96.61	3	4	0.343
https://fbref.com/en/players/e0f8151c/Kim-Min-jae	Sergio Ramos (Paris Saint-Germain)	Kim Min jae	94.60	96.45	91.10	96.24	3	4	0.357

Fuente: Elaboración propia con R

Dichos datos serán guardados dentro de la carpeta *Results* del proyecto en formato CSV. El tamaño de estos conjuntos de datos, en registros, es el número de filas de los originales al cuadrado, puesto que, para cada jugador se calculan sus similitudes y se concatenan dentro de un bucle. El esquema que seguirán sus nombres, dentro de dicha carpeta, será *<<Posición>>_Final_Dataset.csv:*

- *Keeperes_Final_Dataset.csv → 61.009 observaciones*
- *Defenders_Final_Dataset.csv → 1.687.401 observaciones*
- *Midfilders_Final_Dataset.csv → 1.841.449 observaciones*
- *Attackers_Final_Dataset.csv → 992.016 observaciones*

6.2. PRESENTACIÓN DE LA HERRAMIENTA

Una vez creados y guardados los conjuntos de datos definitivos por puestos. En la pestaña habilitada para la modelización de los datos en Power BI, se deben conectar estos a los datos genéricos obtenidos en el segundo apartado de esta documentación. Dicha conexión se

hace a partir de la URL del jugador en FBref, contenida en todos los datasets de los que bebe la propia herramienta.

Esta herramienta consta de cuatro hoja de visualización, una para cada demarcación analizada. Adicionalmente, se establece una primera presentación que hace la función de inicio, a través de la cual, el usuario puede seleccionar la posición del jugador a analizar de su interés. Además de esto, la persona a cargo de utilizarla también debe considerar el hecho de que el jugador se encuentre en dos puestos diferentes, como se ha comentado al inicio y, por lo tanto, los jugadores con los que va a ser comparado o poseer mayores similitudes serán distintos en función de esta. Véase el caso de João Cancelo quien en el estudio es defensor y atacante. Es decir, para los defensores este será similar a los laterales, sin embargo, para los atacantes lo será a jugadores que potencialmente ocupen las posiciones de extremo.

Ilustración 6.2: Pestaña de inicio herramienta de visualización

Fuente: Elaboración propia con Power BI Desktop

Las cuatro hojas de visualización siguen exactamente la misma estructura. En la parte superior izquierda se encuentra un botón para acceder a la pestaña de inicio. A su lado, un cuadro de texto indicando el tipo de posición genérica de los jugadores. Debajo de ella, aparece el nombre del jugador objetivo junto con un cuadro de búsqueda para escribir su nombre o el del equipo al que pertenece, en base a la columna "Player" de la tabla mostrada con anterioridad a modo de ejemplo. Una vez hecho esto, el futbolista objetivo aparecerá en el filtro, expuesto debajo del propio cuadro de búsqueda. En la parte central, se encuentran el resto de los filtros, los cuales se establecen por competición/liga, edad, valor de mercado y la fecha fin de su contrato.

En otro orden de ideas, en la parte superior central y derecha se puede localizar una pequeña tabla que hace referencia a los datos más significativos del jugador a analizar, como son el equipo, la liga, su valor de mercado, etc. Además, dentro de ella, se ha decido mostrar el *cluster* al que pertenece, cuyos significados pueden ser encontrados en la parte de visualización de los conglomerados dentro de esta documentación (apartado 4.3); junto con su vecino y el valor de silueta. Esto se ha hecho con el fin de poder comparar estos valores con los de sus jugadores más similares.

Finalmente, en la zona inferior de la visualización, se localiza la tabla más importante, donde se muestran los jugadores con mayores valores de similitud al estudiado, en base a ese score de calculado anteriormente con un rango de entre cero y 100. En dicho cuadro de datos, ordenado de mayor a menor valor de semejanza, se pueden ver campos como el nombre de los

jugadores similares al analizado, la edad, la nacionalidad, el equipo, la posición específica (extremo derecho, defensa central, centrocampista ofensivo, etc.), el valor de mercado, la fecha fin de contrato y otras más; junto con las métricas obtenidas de la ejecución del algoritmo K-Media (*cluster*, vecino y silueta), además del *score* de similitud, siendo el valor clave dentro de cada representación.

Como punto adicional se aprovecha para reflejar que los nombres de los individuos, tanto en la tabla del jugador a analizar (parte superior) como la de los futbolistas similares a este (parte inferior) poseen insertado un vínculo al link de FBref, por lo que el usuario dispone de la capacidad para hacer click sobre él y visitar el perfil del jugador en este portal.

Presentada y explicada brevemente la estructura de la herramienta final. Hay que destacar que, en apartados posteriores, se van a mostrar cada una de estas presentaciones para las cuatros posiciones, tratando de usar un caso de un jugador real para cada una de ellas, con el objetivo de ser más ilustrativo con relación al funcionamiento de esta visualización presentada en el estudio.

6.3. PORTEROS: GIORGI MAMARDASHVILI

Con relación a los guardametas, se opta por seleccionar como ejemplo real de uso a Giorgi Mamardashvili. Entre las razones principales de su elección está que es uno de los cancerberos jóvenes con mayor potencial a nivel mundial y que, debido a esto, es el jugador de la plantilla valencianista con más probabilidades de abandonar el club en este mercado de fichajes veraniego para la temporada 2023-2024. Además, si esto sucede, se ingresaría una cantidad de traspaso muy necesaria, dada la situación actual del club.

Por consiguiente, el club "che", en caso de que no quiera dar la oportunidad de la titularidad a uno de los porteros que tiene actualmente en su club, va a verse obligado a tener que contratar un nuevo jugador que ocupe ese puesto. Debido a ello, se utilizará esta visualización, a modo de ayuda, con el objetivo de encontrar un jugador similar a este en función de sus características y filtros fijados en el propio cuadro de mandos.

Los filtros que se han establecido, pensando especialmente en el plan ahorrador que están imponiendo sus dirigentes, tienen por objetivo encontrar a un portero de rendimiento inmediato y que además pueda ofrecer varios años de alto desempeño en la elite, es decir, alguien que haya disputado partidos en las mejores ligas del mundo (*big fives*) y que no supere los 30 años de edad. Respecto al precio, este no puede ser muy elevado ya que el club no tiene una alta capacidad de gasto, de este modo, se consideran como máximo 10 millones de euros en su valor de mercado. Finalmente, se considera también de gran relevancia la duración de su contrato, pues cuanto menos tiempo quede para su finalización mayor poder negociador tendrá la entidad. Por tanto, se aspira a aquellos perfiles que además de cumplir las condiciones anteriores, también posean un año o menos de duración de contrato.

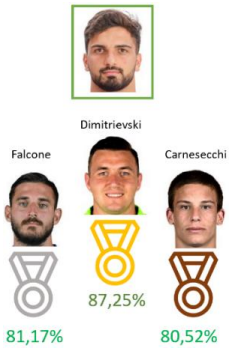
Ilustración 6.3: Caso de uso de la herramienta para porteros: Giorgi Mamardashvili

Fuente: Elaboración propia con Power BI Desktop

Analizados los resultados, se observa que el portero más similar a Mamardashvili es Stole Dimitrievski, jugador de la liga española en el Rayo Vallecano, con un 87.25% según el score de similitud calculado para este estudio y los filtros impuestos en la herramienta. En segundo y tercer lugar se encuentran los porteros de la liga italiana Falcone y Carnesecchi, respectivamente. Ambos poseen un score bastante similar, que se encuentra entorno al 81%. Una vez se pasa de los tres primeros, el resto de los guardametas ya no consigue alcanzar ni el 70% de similitud respecto al portero analizado.

Con relación a los porteros de la Serie A, cabría destacar que el tercero de ellos es mucho más joven, tan solo 22 años, pero posee un valor de mercado más elevado que los otros dos porteros (8 millones de euros), por lo que se puede extraer que va a alcanzar un alto potencial de rendimiento en su carrera y, por tanto, una alta probabilidad de revalorización. Por último, si se contempla el portero del Lecce, este se encuentra en un *cluster* diferente al resto de porteros más similares, es decir se localiza en el conglomerado número dos y el resto está en el uno. Aunque, por el valor de la silueta (negativo) se concluye que se encuentra mal clasificado, posiblemente debiendo estar en ese uno, dada su similitud con el resto de los cancerberos contemplados.

Ilustración 6.4: Pódium de jugadores más similares a Giorgi Mamardashvili



Fuente: Elaboración propia con Power Point. Imágenes obtenidas de fbref.com

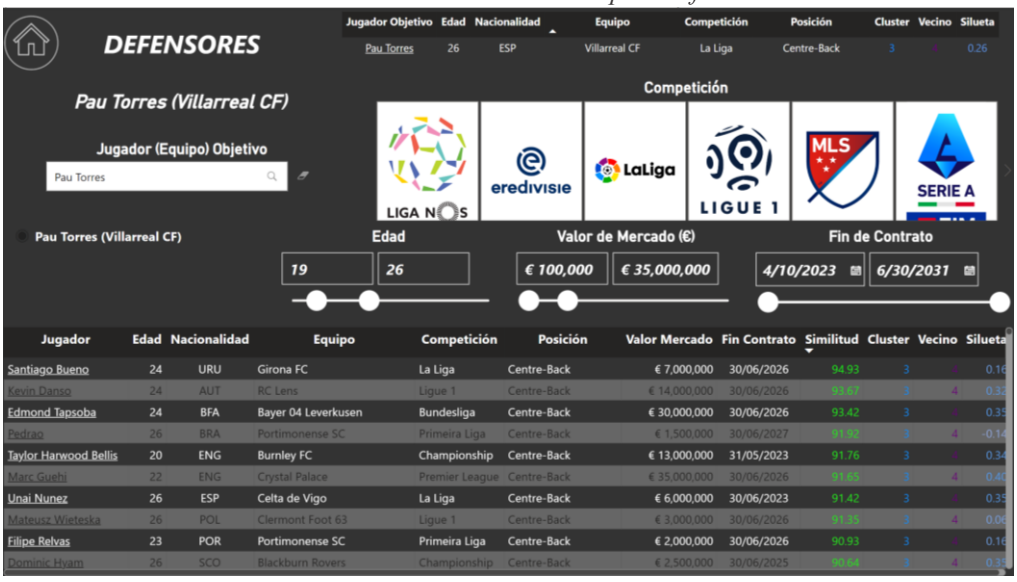
6.4. DEFENSORES: PAU TORRES

Para los defensores se ha elegido a Pau Torres como jugador de ejemplo para un caso real de uso, dado que, el central del Villarreal, actualmente, acaba de de abandonar la entidad *groguet* para recalar en el Aston Villa de la Premier League. En consecuencia, los dirigentes tendrán que localizar a un sustituto similar a Pau Torres, pues durante las ultimas temporadas ha sido una pieza clave dentro de los esquemas del equipo en los partidos, con los diferentes entrenadores que ha tenido. Uno de ellos es Unai Emery quien hoy en día entrena al equipo de Birmingham.

Con el objetivo de encontrar a ese sustituto se fijan una serie de filtros dentro de la propia herramienta que posibiliten acotar las características generales y que no son propias del rendimiento deportivo del propio jugador a hallar. De acuerdo con ello, se amplía el abanico con relación al valor de mercado. Dado que, se tiene que sustituir a un jugador fundamental en la plantilla y, por el que se espera obtener una cifra cuantiosa por su traspaso. Así pues, se opta por fijar este valor en un máximo de 35 millones de euros, siempre con el propósito de obtener una diferencia positiva entre ingresos y gastos. Respecto a la competición y a la fecha de final de contrato no se considera ningún filtro, pues se decide que para la primera se puede localizar a un jugador desconocido con alto potencial y, para la segunda, sea cual sea la duración se está dispuesto a realizar un desembolso importante para la entidad si es necesario. Por el contrario, para la edad se establece que el futbolista sustituto no puede superar los años que tiene Pau, en este caso 26.

Tras aplicar estos filtros en el cuadro de mandos se muestran los resultados obtenidos:

Ilustración 6.5: Caso de uso de la herramienta para defensores: Pau Torres



Fuente: Elaboración propia con Power BI Desktop

Una vez evaluados los datos que aparecen en la tabla de jugadores similares, se contempla que Santiago Bueno, central uruguayo que disputa sus partidos en la primera división española con el Girona, es el jugador que más similitudes guarda en conformidad con los filtros aplicados. Seguido de este y a muy poca distancia, de acuerdo con el valor de similaridad, se encuentran el resto de los zagueros, visualizados en el propio cuadro. Por lo que, la elección del sustituto puede resultar complicada dada la poca diferencia entre ellos.

Así pues, se debe recalcar que los dos siguientes defensores, Kevin Danso y Edmond Tapsoba, poseen un valor bastante más elevado que el primero de ellos, llegando el futbolista del Bayern Leverkusen a los 30 millones de euros, según *Transfermarkt*. Además, este jugador se encuentra actualmente muy seguido por otros clubes de alto nivel dado su potencial de crecimiento en su rendimiento. Entre los equipos que le siguen cabe hacer mención del Tottenham Hotspur, ya que el propio Pau también estuvo muy cerca de vestir la camiseta de los *Spurs*. Lo que puede ser indicativo de que el conjunto inglés está buscando indirectamente un sustituto al central del submarino amarillo, a través del central de la Bundesliga, dado su fichaje fallido.

De cualquier forma, se recomendaría la contratación de Santi Bueno, ya que, no solo posee un valor de mercado bastante inferior a los otros dos centrales, sino que también tiene experiencia dentro de la competición y conoce el país. Además, esta pasada temporada demostró un muy buen rendimiento con el conjunto catalán, recién ascendido. Por otra parte, también se podría tratar de fichar a uno de los otros dos, en caso de que se deseé dar un salto de calidad a la plantilla. En cambio, habría que entrar a la puja por ellos junto a entidades europeas con un elevado potencial económico y no se debería olvidar el riesgo que supone el desconocimiento de la competición liguera o el modo de vida del país.

Ilustración 6.6: Pódium de jugadores más similares a Pau Torres

Fuente: Elaboración propia con Power Point. Imágenes obtenidas de fbref.com

6.5. CENTROCAMPISTAS: SERGIO BUSQUETS

Tras haber abandonado Sergio Busquets el FC Barcelona, el equipo de la ciudad condal necesita encontrar un sustituto de rendimiento inmediato para el que seguramente ha sido el mejor pivote de su historia.

Así pues, se considera que el club debe hacer el mayor de los esfuerzos económicos dentro de su situación actual, dado que se está buscando un reemplazo para una de las posiciones más significativas dentro del esquema de juego culé. Por tanto, se considera que el máximo valor de mercado al que puede optar el club se encuentra en alrededor de 50 millones de €, además se impone un mínimo de 15 millones de €, como límite, para segmentar a jugadores de un cierto nivel mundial. Relacionado con esto, solo se van a tener en cuenta a futbolistas procedente de las cinco grandes ligas del mundo. Considerando, la duración del contrato se piensa que esta variable no es importante dado que se acotaría demasiado el rango de posibles sustitutos para un jugador tan importante y donde se necesita rendir desde el primer minuto. Por último, para la edad se decide seleccionar aquellos que tenga menos de la treintena de edad, con el fin de asegurar futbolistas con una cierta experiencia y que ofrezca años de un buen rendimiento.

Se procede a mostrar los resultados obtenidos, aplicados los filtros sobre la herramienta:

Ilustración 6.7: Caso de uso de la herramienta para centrocampistas: Sergio Busquets

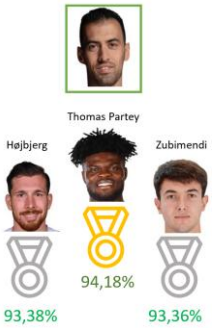
Fuente: Elaboración propia con Power BI Desktop

De esta manera, se atiende a que el mediocentro ghanés del Arsenal, Thomas Partey, es el jugador más similar a Busquets, según los filtros aplicados en el propio cuadro de mandos. Seguido de este, se encuentran el centrocampista de los *Spurs*, Højbjerg y, el pivote *txuri-urdin*, Martín Zubimendi. Todos ellos con valores cercanos al 94% de similitud. Asimismo, si se tienen en cuentan los resultados obtenidos por medio del algoritmo K-medias, se observa que la totalidad de jugadores representados en la visualización pertenecen al *cluster* 2, es decir, aquel destinado a los centrocampistas posicionales según la explicación dada en el apartado de visualización de conglomerados. De todos modos, es importante destacar que el centrocampista danés es el único que posee un vecino diferente al resto (el uno frente al tres para los demás), además su valor de silueta no se aproxima al "0,4" como si lo hacen el resto de los jugadores más similares. Esto es debido a que su estilo no es tan "posicional" en comparación con los otros futbolistas.

Por consiguiente, si se debe recomendar la contratación de un jugador para sustituir a Sergio, en base al análisis efectuado, este es Thomas Partey. Las razón principal es su alta proximidad tanto en los resultados ofrecidos por el *score* de similitud como por el análisis de conglomerados. Dentro de esto, también cabe mencionar a Zubimendi, en cambio se debe anteponer que su fichaje es altamente improbable dadas las altas pretensiones del club que posee sus derechos (Real Sociedad).

Con relación al centrocampista del conjunto londinense, se debe reflejar la probable contratación de otro jugador de su competencia como es Declan Rice, por lo que es probable que desee abandonar el propio club. Por último, también mencionar, que el jugador africano ha realizado una gran temporada en el conjunto *gunner*, club que comparte un modo de juego similar a la entidad blaugrana. Además, este jugador no solo posee un excelente trato de balón, sino que también puede aportar físico al centro del campo, cualidad muy demanda en la actualidad, independientemente del estilo.

Ilustración 6.8: Pódium de jugadores más similares a Sergio Busquets



Fuente: Elaboración propia con Power Point. Imágenes obtenidas de fbref.com

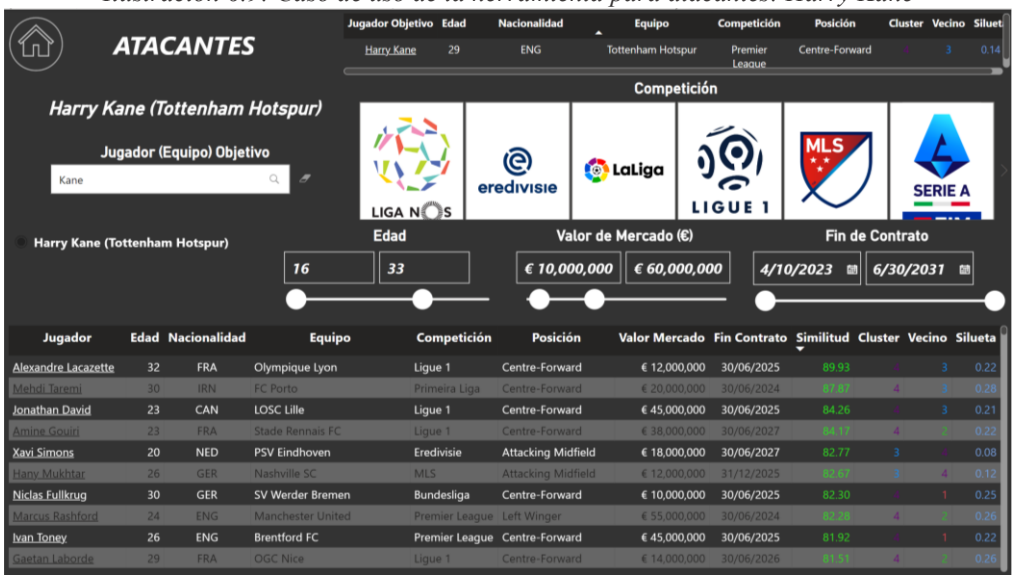
6.6. ATACANTES: HARRY KANE

Con relación a la situación de los atacantes, se ha elegido a Harry Kane como ejemplo de uso de la herramienta. El principal motivo de la elección es la alta probabilidad que existe para que, en el actual mercado de fichajes de verano, abandone el Tottenham Hotspur, dado que el conjunto del norte de Londres no disputará competición europea la próxima temporada. Así pues, se opina que un jugador de tan alto nivel no puede permitirse dicha situación.

Para localizar a su sustituto, no se ha estimado altamente primordial el imponer filtros claves dentro del cuadro de mandos. Únicamente, se ha acotado la edad a los 33 años, ya que se considera que a partir de dicha franja el jugador suele decrecer en su rendimiento y, se ha fijado un valoración máxima de 60 millones de euros, puesto que esta entidad no suele realizar inversiones muy cuantiosa, teniendo en cuenta el nivel de gasto de la Premier League.

Implementados estos filtros se procede a enseñar los resultados obtenidos:

Ilustración 6.9: Caso de uso de la herramienta para atacantes: Harry Kane



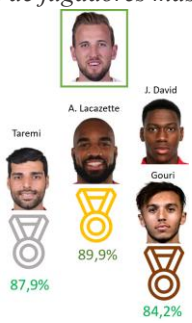
Fuente: Elaboración propia con Power BI Desktop

De acuerdo con la visualización previa, se contempla que el jugador que posee una mayor similitud (90%) es el jugador del O. Lyon, Alexandre Lacazette con 32 años, quien ya dispone de experiencia en la competición tras su paso por el equipo rival de los *spurs*, el Arsenal. Seguidamente, con aproximadamente un 88% de similitud se encuentra Taremi,

delantero iraní de 30 años perteneciente al Oporto. A continuación de ellos, aparecen otros dos atacantes procedentes de la liga francesa y que también poseen cualidades similares de juego, pues pertenecen al *cluster* 4, dichos delanteros son Jonathan David del Lille y Gouiri del Stade de Rennais. A esto se debe añadir, que son mucho más jóvenes, ambos con 23 años.

Por consiguiente, se cree que en esta ocasión la elección del jugador no debe ser basada tanto en la puntuación de similitud, sino más en el rendimiento futuro que puedan ofrecer estos jugadores. Por ello, se toma como candidatos a los jugadores de la *Ligue 1*, aunque su valor de mercado sea bastante superior a los dos primeros. Por último, hay que señalar que Taremi puede ser una buena opción y de más fácil contratación, ya que, además de ser más barato, se le estarían dando la oportunidad de jugar en un "grande" de Inglaterra, tras haber disputado competición europea con el club portugués. Para el caso de Lacazette, aunque sea el más similar, se cree que es el supondría el fichaje más improbable, pues tendría que abandonar su país e ir a un club rival de su antiguo equipo.

Ilustración 6.10: Pódium de jugadores más similares a Harry Kane



Fuente: Elaboración propia con Power Point. Imágenes obtenidas de fbref.com

Por ende, se ha podido demostrar con cuatro ejemplos para cada posición como podría ejecutarse un estudio de *scouting* dentro de la herramienta, contemplando en todo momento las limitaciones que posee y extrayendo posibles ideas que sirvan de mejora.

7. CONCLUSIONES Y TRABAJO FUTURO

A lo largo de este proyecto se ha demostrado como técnicas avanzadas de análisis estadístico no supervisadas pueden ser también aplicadas dentro del ámbito deportivo y, de manera específica, en el campo del *scouting* y el análisis para el fútbol. De igual forma, se ha puesto de manifiesto la capacidad de obtener resultados óptimos en el contexto profesional, a través de grandes cantidades de datos abiertos, cuya extracción es totalmente gratuita.

De este modo, se evidencia que estos métodos llevados a cabo sobre elevados volúmenes de información son de gran ayuda, no solo para la contratación de jugadores similares a uno dado, sino que también para la detección de nuevas características en un jugador de la propia plantilla.

Así pues, con la intención de guardar la mayor parte de la información posible contenida en los datos, se ha optado por reducir las altas dimensiones de los conjuntos de datos originales, una vez preprocesados, mediante la aplicación del ACP. Lo que ha posibilitado tener una mejor interpretación de los datos y conseguir las primeras puntuaciones de similitud sobre cada par de jugadores.

Posteriormente con esos nuevos registros, a raíz de la ejecución del algoritmo K-medias, se pudo agrupar a los individuos en una serie de conglomerados, especificados con anterioridad, para detectar los atributos y cualidades de su juego, en función de las dos primeras componentes y el *cluster* al que perteneciese por puesto.

En último término, con el fin de añadir mayor fiabilidad al estudio se optó por alcanzar otros valores de similitud en base a las distancias euclídea y coseno, para que, junto con los resultados de semejanza a través de las nuevas componentes, pudiera ser creado un porcentaje de similitud promedio. Dicho valor, unido a las métricas obtenidas en el análisis de conglomerados, son los registros principales sobre los que se han basado las decisiones tomadas en la herramienta, toda vez los filtros se encuentren aplicados.

En otro orden de ideas, se ha decidido proponer una serie de puntos de mejora con el objetivo de optimizar el proceso por el que este proyecto es construido y aumentar la calidad de visualización de la herramienta.

En primer lugar, se plantea construir tres bases de datos estructuradas (SQL) donde almacenar las tablas que contienen los siguientes conjuntos de información. Por un lado, los datos extraídos directamente desde FBref o *Transfermarkt (raw data)*. Por otro lado, aquellos que son preprocesados por posición y que son necesarios para el análisis. Finalmente, aquellos que guardan los resultados definitivos, después de haber sido aplicado el análisis no supervisado, y la información general necesaria para los cuadros de mando.

En segundo lugar, se sugiere crear una aplicación o un portal web que sustituya a la herramienta creada como ejemplo de caso de uso en Power BI. Dicho sitio web tendría una estructura similar a la propia herramienta, sin embargo, se le añadirían nuevas visualizaciones como gráficos de radar o de barras que comparen algunos campos originales del jugador analizado con uno similar. Asimismo, se ofrecería al usuario la posibilidad de seleccionar entre los campos originales que más le interesen.

En tercer lugar, el consumidor de este sistema web también tendría la posibilidad de ejecutar cada uno de los análisis desde la propia página, decidiendo el modo de seleccionar las componentes, ya sea manual o con valores superiores a la media, o la forma de indicar el número de clusters, en caso de que los resultados no sean los óptimos.

En cuarto lugar, también se proporcionaría un filtro para seleccionar las características originales, procedentes de la extracción y transformación de los datos, en función de la posición a analizar. De este modo, se da la posibilidad de controlar la información o variación explicada en las componentes y mejorar los resultados de los algoritmos aplicados a posteriori.

Por último, se repara en que puede ser de gran utilidad, para dotar de mayor solidez al análisis, añadir información o puntuaciones de similitud entre clubes o entrenadores. Así pues, las decisiones se basarían en resultados más robustos pues se estarían considerando elementos claves, con miras a la adaptabilidad del jugador en el estilo de juego del equipo.

Estas ha sido una serie de ideas que con el tiempo suficiente y en base a la estructura del proyecto puede ser llevadas a cabo sin alterar demasiado la idea fundamental de este. De igual forma, y como conclusión definitiva, hay que destacar que a partir de un proceso bastante sencillo de almacenamiento de datos y aplicación de algoritmos no supervisados se ha podido desarrollar un producto que puede ser viable en un contexto real de dirección deportiva.

8. BIBLIOGRAFÍA

Carlos Soria, P. (2021). *Diseño y aplicación de técnicas de machine Learning para optimizar el Scouting en clubes de fútbol* (pp. 1-2). Sevilla. Escuela Técnica Superior de Ingeniería Universidad de Sevilla.

Claudio Curia, L. (24 de diciembre de 2010). *Hugo Lloris y misterio del algoritmo de similitud*. Marca. https://www.marca.com/blogs/master-big-data-deportivo/2021/12/24/hugo-loris-y-el-misterio-del-algoritmo.html

Delgado, R. (23 de junio de 2018). *Introducción a los Modelos de Agrupamiento (Clustering) en R*. Rpubs. https://rpubs.com/rdelgado/399475

Fernández, J. (2021). *Diploma Universitario en Estadística y matemáticas aplicado al deporte con R. Módulo 2. Tratamiento de los datos* (pp. 33-40, 71-81). Sport Data Campus (ENIT – Big Data Internacional Campus)

Fernández, J. (2021). *Diploma Universitario en Estadística y matemáticas aplicado al deporte con R. Módulo 3. Modelización y análisis final* (pp. 21-44, 54-81). Sport Data Campus (ENIT – Big Data Internacional Campus)

FBREF. https://fbref.com/en/

Gil Martínez, C. (junio de 2018). *Análisis de componentes principales (PCA)*. Rpubs. https://rpubs.com/Cristina_Gil/PCA

Oracle. *¿Qué es big data?*. https://www.oracle.com/es/big-data/what-is-big-data/

Sánchez Pantigoso, C.F. (18 de noviembre de 2019). *Análisis de componentes*. Rpubs. https://rpubs.com/Csanchez15/551258

Zivkovic, J. (23 de junio de 2023). *Extracting data from FBreft*. worldfootballR. https://jaseziv.github.io/worldfootballR/articles/extract-fbref-data.html

Zivkovic, J. (23 de junio de 2023). *Extracting data from Transfermarkt*. worldfootballR. https://jaseziv.github.io/worldfootballR/articles/extract-transfermarkt-data.html

Zivkovic, J. (23 de junio de 2023). *worldfootballR*. worldfootballR. https://jaseziv.github.io/worldfootballR/