# ANÁLISIS DE JUGADORES EN LIGAS EUROPEAS POR MEDIO DE PCA, CLUSTERING JERÁRQUICO Y DBSCAN

Karina Durán Bautista código: **200824840**
David Castrillón Montaño código: **202120603**
Edna Julieth Mora código: **202124435**
Misael García Díaz código: **202124621**

**RESUMEN**

En este trabajo se aplican herramientas del análisis no supervisado para el apoyo en la toma de decisiones por parte de las empresas de fútbol que tienen gran cantidad de datos y desean extraer de estos información para mejorar sus diferentes políticas. Esto, se realiza mediante agrupamiento de jugadores de ligas europeas, implementando PCA, clustering jerárquico aglomerativo y DBSCAN.

La base de datos analizada es de las temporadas 2019-2020 de ligas europeas, después de realizar una limpieza y transformación de algunas variables, se estandariza la base de datos y se usa PCA de forma iterativa dos veces y se decide dejar 50 variables totales y 13 componentes principales que explican el 81.4% de la varianza. Por último, se aplican los métodos de clustering jerárquico aglomerativo y DBSCAN y se obtienen las diferentes subdivisiones que fueron analizados y clasificados. A partir de los resultados se realizan estudio detallado sobre la base de datos principal, se encuentran subdivisiones de jugadores que pueden ser clasificados en tipo defensivo, ofensivo y de armador de juego. Esta información puede ser usada por el director técnico y determinar si hay potencial adicional en determinado jugador o posibles fichajes de jugadores. Y así se da respuesta al: "análisis de jugadores en ligas europeas por medio de PCA, Clustering Jerárquico y DBSCAN".

## INTRODUCCIÓN

La industria del fútbol es una de las más proliferas en el mundo, alrededor del fichaje de jugadores, transmisiones de televisión y streaming, merchandising, entradas a estadios, apuestas, entre otras. De acuerdo con [7] en la temporada 2020/21 los 20 clubes que producen mayor ingreso produjeron 8.2M€ entre partidos jugados, retransmisiones e ingresos comerciales. Esta industria no es ajena a la necesidad de contar con herramientas para el soporte en la toma de decisiones como son el fichaje de jugadores y la proposición de estrategias de juego según el oponente. Con este trabajo se responde parcialmente a esta necesidad, utilizando como medio metodologías basadas en clustering, encontrando relaciones entre la calidad de los jugadores, su posición y la influencia de su valor comercial.

Un acercamiento al uso de este tipo de herramientas en la industria del fútbol se presenta en [6], donde a partir de datos de los partidos se desarrolla un modelo que clasifica las cualidades individuales de los futbolistas, detectando primero las posiciones de juego: delanteros, centrocampistas, extremos, laterales, centrales y porteros, mediante la agrupación de K-Means, con una precisión de 0.89 (para la Premier League 2021/2022 y 0.84 para la Allsvenskan 2021. Se utilizó un enfoque de modelado no supervisado en el que se aplica el análisis PCA de manera iterativa. Para la posición de juego 'Delantero' se encontró 4 estilos de juego diferentes de juego. Como resultado de este trabajo, se encontró que es más fácil encontrar estilos de juego que tengan acciones claras y evidentes con el balón que los distinguen de otros jugadores dentro de su respectiva posición.

Bonilla y otros en [1]. realizan una serie de mediciones antropométricas de atletas olímpicos jóvenes en el Urabá antioqueño de Colombia bajo condiciones controladas, logrando generar agrupamientos con el método de K-medoides para identificar formas eficientes de entrenar a los atletas, descubrir talentos, focalizar la habilidad o especialización de los atletas y de entender las características de los agrupamientos. Las variables que miden y que les permiten caracterizar a los atletas están asociados a los tamaños de ciertos miembros del cuerpo, masa corporal, masa ósea, tejido adiposo, grosor de las diferentes capas de la piel, entre otras. Bonilla y los demás identifican dos fenotipos en el agrupamiento del algoritmo para los atletas, uno con contexturas ligeras, pequeñas y biomecánicamente más eficientes, pero con poca madurez ósea; el segundo grupo con contexturas gruesas, más altos y biomecánicamente menos eficientes y con una madurez ósea más avanzada.

Una estrategia, más robusta y completa la realizan Tosato y Wu en [2], donde se establece un sistema de recomendación con aprendizaje no supervisado de agrupamiento conocido como Projective Adaptive Resonance Theory (PART), que aplica redes neuronales. Tosato y Wu aplican el sistema de recomendación al mercado de jugadores del equipo AS Roma de Italia para determinar atributos críticos y características al quedar agrupados de cierta manera. El algoritmo recomienda aspectos importantes de la formación del

equipo, y de forma especial la compra o venta de jugadores entre los diferentes agrupamientos, teniendo en cuenta restricciones económicas. El algoritmo forma agrupamientos de baja dimensionalidad en la base de datos completos, es decir, agrupa en un subconjunto de atributos de la base de datos completa.

En el presente trabajo, usando reducción de la dimensionalidad con PCA, se da interpretación a los diferentes clústeres que aparecen debido a que se consigue trabajar con una cantidad reducida de variables obtenidas de PCA identificando si el clúster comparte en su mayoría propiedades defensivas, ofensivas o de armador de juego. También, se identifican ciertos jugadores atípicos o que inicialmente parecen no pertenecer o estar alineados con la mayoría de las propiedades de los demás integrantes de su clúster y que en algunos casos pueden ser interpretados como jugadores versátiles que podrían desarrollar mejor su potencial, apoyando esta información la decisión del técnico de ubicarlo en otra posición dentro de la cancha o cambiar su foco de entrenamiento. Una de las principales limitaciones del estudio es que presenta un agrupamiento heurístico en el clustering jerárquico aglomerativo ya que la cantidad de clústeres se encuentra de forma visual a partir del dendograma.

**MATERIALES Y MÉTODOS**

Los datos seleccionados corresponden a la evaluación de los jugadores de las ligas europeas de fútbol durante la temporada 2019 – 2020, realizada por la compañía Transfermarket [3], firma especializada en seguimiento estadístico y valoración de jugadores según su rendimiento. Obtenidos del conjunto de dataset públicos de Kaggle [4] página especializada en competencias de ciencias de datos. El dataset cuenta con 2644 observaciones y 400 variables 9 de ellas categóricas (Información general del jugador), y el resto numéricas (Estadísticas del jugador).

El proceso de limpieza se centró en realizar en primer lugar un análisis descriptivo para identificar si dentro de las 400 variables existen algunas que no estén relacionadas directamente con el desempeño del jugador o si existen correlaciones entre las mismas para reducir la cantidad de variables a incluir en el modelo. Se encuentran solo 9 variables tipo objeto entre las cuales se encuentran 124 jugadores repetidos, 102 nacionalidades distintas, 10 posiciones de juego diferentes y 14 de posición secundaria, 124 jugadores que no referencia pie preferido, 98 equipos diferentes y existe una columna con el ranking del equipo en la liga en la que participa liga en la que juega (5 ligas), asistencia (valores únicos coinciden con cantidad de equipos) y temporada.

Se realizó una exploración de datos con valores de variables categóricas faltantes o duplicados. Algunos jugadores se encuentran repetidos en el conjunto de dato. Para tratar estos casos se escoge el registro con mayor tiempo del jugador para conservar y asegurar así los valores únicos. Se identifica la variable 'CLBestScorer' con datos faltantes, dado que relaciona si el jugador es goleador en una competición especifica donde no todos los jugadores participan, se excluye del modelo. Algunos jugadores con información faltante en el campo de pie dominante son muy jóvenes y carecen de información en varios campos. Se imputaron estos datos con el valor más común en la posición del jugador, puesto que esta tiene una fuerte relación con la pierna dominante del mismo.

Para las variables numéricas, se encuentra que hay cierta concentración de jugadores con pocos minutos de juego. Esto puede ser ocasionado al número de jugadores sustitutos que no siempre son convocados o jugadores nuevos que llegan terminando la temporada. Su falta de apariciones podría afectar el correcto análisis por lo que se establece un mínimo de minutos jugados esperados para ser incluidos en el modelo.

Se presentan valores generales de juegos en los que participo, cantidad de juegos en los que fue titular, minutos jugados y total de minutos jugados en 90 minutos, lo que da un valor del total de 90 minutos jugados, una forma de presentar el total de juego realizado independiente de no jugar los 90 minutos completos de los partidos en que participo. Estas variables presentan información distinta, al tener un jugador titular frente uno que suplente puede afectar su segmentación, de igual forma la cantidad de minutos es importante para comparar entre jugadores el rendimiento, no es lo mismo un jugador con 100 pases en 1 partido que uno con 100 pases en 100 partidos. Luego del procedimiento descrito, se tienen 127 variables de interés y 2169 jugadores en la base de datos.

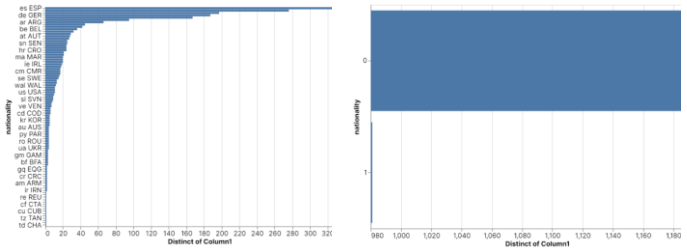

*Figura 1. Jugadores por nacionalidad.*

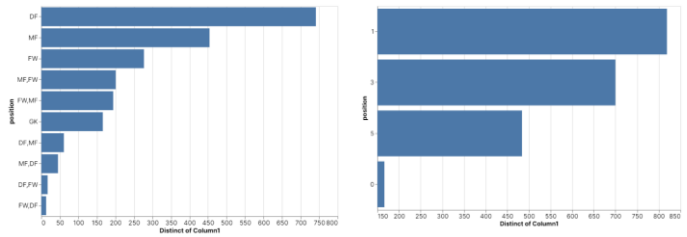

*Figura 2. Posiciones principales.*

1: porteros
2: defensas
3. mediocampistas
4: delanteros

```
# Cambio de valores de posicion segun la ubicación en el campo de
data_new["position"] = data_new["position"].replace({"GK": 0,
                                                     "DF": 1, "DF,MF": 1, "DF,FW": 1,
                                                     "MF":3, "MF,DF":3, "MF,FW":3,
                                                     "FW":5, "FW,DF":5, "FW,MF":5})
```

*Figura 3. Descripción de las nuevas posiciones.*

Con los datos filtrados, se exploran los campos de interés, para encontrar nuevos filtros o tratamientos a las variables. Las principales nacionalidades de los jugadores son España, Francia, Italia, Inglaterra y Alemania, que coincide con las ligas incluidas en el set de datos, y son casi el 50 % de los jugadores presentes, por lo que la variable nacionalidad para el resto de los jugadores filtraría demasiado el set de datos y podría afectar el desempeño del algoritmo. La nacionalidad se transforma en una variable binaria si juega en la liga de su país o en una extranjera, Figura 1. Para el caso de las posiciones se identifican jugadores que no juegan en una única posición y que pueden jugar en una segunda, o q su posición se combina entre estas dos. Sin embargo, esta distribución de valores genera ambigüedad entre un defensa medio campista y campo defensa, por lo que se agrupan estas posiciones en un nuevo grupo para reducir los valores presentes en las variables y poder cambiarla a una numérica, Figura 2.

Para la variable de liga, se encuentran más de 400 jugadores en cada una de ellas, para poder convertir esta variable en una numérica, se busca clasificar cada liga según su nivel de dificultad, como una forma de diferenciar el rendimiento similar de dos jugadores que jueguen en ligas distintas. No sería igual un jugador con porcentaje alto de goles en una liga con alta competición, a un jugador con el mismo porcentaje de gol en una liga menos exigente. Se decide unificar la información de las ligas entonces en tres grupos. Con esto finaliza el proceso de limpieza de los datos.

Debido a la alta cantidad de variables se propone realizar PCA con el fin de reducir la dimensionalidad de la base de datos y facilitar el análisis de los resultados [5]. Se realiza el cálculo de PCA mediante eigenvalores y se estandarizan los datos antes de realizar la descomposición, para elegir el número de componentes se toma una varianza mínima del 80%, para lo cual se requiere tomar los 15 primeros componentes, para una varianza explicada del 80.2%, Figura 4.

Con los componentes obtenidos se realiza una inspección del aporte de cada una de las 127 variables a cada componente, se observa que el aporte de una gran cantidad de las variables es muy pequeño, se procede a realizar un filtrado de las variables que aportan un porcentaje mayor o igual al 2% a cada componente y se eliminan las demás. Se encuentran 50 columnas de mayor aporte entre estas: asistencias, faltas, tiros de esquina errores, tiros, goles, pases, penaltis, tiros de creación y abordajes. Con estas columnas identificadas como ponderadas se procede a filtrar de nuevo el data set para recalcular los componentes a utilizar. Se obtienen 13 componentes para una varianza explicada acumulada de 81.4%, Figura 5.

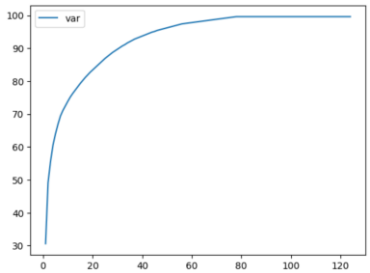


*Figura 4. Varianza acumulada PCA.*

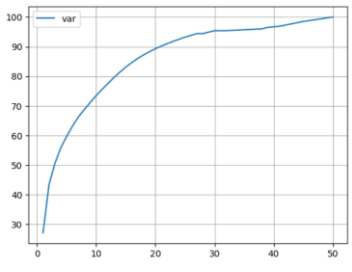


*Figura 5. Varianza acumulada PCA filtrado por aporte.*

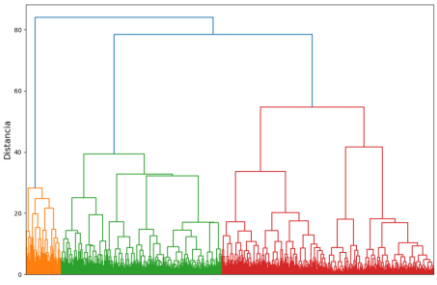


*Figura 6. Dendograma*

El proceso de agrupamiento se inicia implementando el algoritmo de clustering jerárquico aglomerativo, en donde se identifican para una distancia de 45, 4 clústeres, (como se observa en el dendograma de la Figura 4) utilizando la distancia “euclidiana” y el enlace “ward” que fueron los que brindaron una separación con más sentido. También se implementa el algoritmo basado en densidad DBSCAN, donde al utilizar radios (*eps*) muy cortas, o muchas muestras (*min_samples*) para formar un clúster, la cantidad de clústeres eran muy pocos debido a que al aumentar la dimensionalidad del problema se tienen los datos mucho más separados entre sí (maldición de la dimensionalidad). Finalmente, una buena separación se consigue de tal forma que se tiene también 4 clústeres. El resultado obtenido con el algoritmo DBSCAN es similar en rasgos generales y cualitativos a la distribución que brinda el algoritmo jerárquico aglomerativo, por lo tanto, se van a omitir gráficas y análisis en este documento (pero se pueden ver en el cuaderno del código). Luego, se procede a analizar los resultados.

**RESULTADOS Y DISCUSIÓN**

Al implementar el algoritmo jerárquico aglomerativo con la métrica “euclidiana”, el enlace “ward” y 4 clústeres de acuerdo con el corte a una altura aproximada de 45 en el dendograma de la Figura 6 tenemos una separación adecuada y que permite identificar claramente 3 tipos de jugadores: los ofensivos, defensivos y los armadores de juego. Los *ofensivos* mostraban pesos mayores en variables como goles por 90 minutos y tiros al arco; los *jugadores defensivos* tenían pesos mayores en las variables del tipo de interceptaciones y faltas; por último, el otro grupo con mayoría fueron los *armadores de juego* que poseían mayor peso en las variables asociadas a los pases cortos y pases completos. En la Figura 4 del anexo podemos ver como se genera una agrupación parcial en dos dimensiones al graficar las componentes principales 0 y 2 (imagen de la izquierda) y las componentes 0 y 8 (imagen de la derecha). Debido a que estamos trabajando datos en 13 dimensiones (las 13 componentes principales seleccionadas) una proyección en dos dimensiones no parece tener una agrupación clara y distintiva, los puntos se solapan de forma exagerada, pero en gran medida se debe al hecho de que se están proyectando en un espacio de menos dimensiones que el original. En la Figura 4 del anexo se puede ver una imagen en 3D de los 4 clústeres, donde claramente se ven 3 de los clústeres y el cuarto clúster parece oculto o detrás de los otros tres.

A partir de los resultados mencionados para el algoritmo jerárquico aglomerativo se hace segmentación de los datos de acuerdo con los 4 clústeres definidos, se realizan estadísticas teniendo en cuenta el total de jugadores y las 50 variables seleccionadas, para cada uno de los clústeres se obtienen los promedios desviaciones estándar y percentiles con el fin de analizar y detectar patrones en los clústeres arrojados por el algoritmo. La variación de los datos es alta, luego la media y la desviación no son concluyentes, sin embargo, se observa una diferenciación entre clúster para el percentil 50 y 75. Con base en el percentil 70 se identifican 5 categorías asociadas al clúster 4 (corner_kicks_straight, pens_made, pens_att, corner_kicks_straight, pens_won), 5 asociadas a los clústers 4 y 2 (gca_shots, gca_fouled, corner_kicks_in, corner_kicks_out errors) y una exclusiva al cluster 1 (pens_conceded) y 27 variables que generan categorías internas dentro de todos los clusters:

- passes_switches
- passes_offsides
- cards_yellow
- goals_per90
- assists_per90
- assists_per90
- npxg_net
- crosses_into_penalty_area
- crosses
- sca
- sca_dribbles
- sca_shots
- fouled
- interceptions
- clearances
- carry_progressive_distance
- throw_ins
- fouls
- blocked_passes
- dribbles_completed
- xg_xa_per90
- sca_passes_dead
- passes_right_foot
- asses_pressure
- tackles_won
- tackles_mid_3rd
- tackles_att_3rd

Con estas variables identificadas resulta más sencillo analizar la base de datos de jugadores respondiendo al planteamiento inicial, encontrando por ejemplo jugadores de desempeño similar para cierta posición dada, pero con grandes diferencias en su valor comercial. Por ejemplo, queremos identificar delanteros que tengan un desempeño similar y seleccionar los de más bajo valor comercial para un posible fichaje. A partir de los datos obtenidos para el cluster 4, se grafican los valores de los jugadores vs 'goals_per90' (goles en a 90 minutos) que sería una medida significativa para la posición delantero y se encuentra dentro del set de variables categorizadas en los clústeres, la posición 5 corresponde a delantero en la Figura 4.

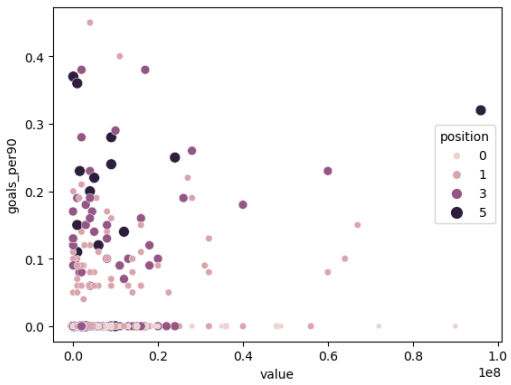


*Figura 7. Efectividad en goles por 90 minutos vs valor.*

El resultado en la gráfica nos permite ver que la posición 5 (delantero) contiene jugadores de alto valor con 'goals_per90'por encima de 0.2 pero también de esta misma medida los hay de un valor muy bajo en la escala, lo que permite intuir que existen jugadores de calidad similar con valor comercial bajo que resultarían interesantes para un fichaje o en el caso de jugadores de valor intermedio, pero desempeño por debajo de 0.2 identificar que técnicas o estrategias deberían trabajarse en el entrenamiento para potenciar su efectividad. Basándose en las variables identificadas, asociadas únicamente al clúster 4 se filtra la base de datos total para buscar los delanteros con un potencial de 'goals_per90' mayor a 0.27:

| *índex* | *Jugador* | *Valor* | *edad* | *goles_per90* | *pen_gan* | *Tiros de esquina _straight* | *penalties _att* | *penalties_ made* |
|---|---|---|---|---|---|---|---|---|
| ***82*** | Antoine Griezmann | 96000000 | 28 | 0.32 | 0 | 0 | 0 | 0 |
| ***21*** | Harry Wilson | 17000000 | 22 | 0.38 | 0 | 1 | 0 | 0 |
| ***540*** | Nianzou Kouassi | 11000000 | 17 | 0.40 | 0 | 0 | 0 | 0 |
| ***341*** | MickaÃ«l Cuisance | 10000000 | 19 | 0.29 | 1 | 5 | 0 | 0 |
| ***277*** | GrÃ©goire Defrel | 9000000 | 28 | 0.28 | 0 | 0 | 1 | 1 |
| ***417*** | Giorgio Chiellini | 4000000 | 34 | 0.45 | 0 | 0 | 0 | 0 |
| ***101*** | Antonin Bobichon | 2000000 | 23 | 0.28 | 0 | 4 | 0 | 0 |
| ***109*** | Junior Stanislas | 2000000 | 29 | 0.38 | 0 | 1 | 1 | 1 |
| ***315*** | Jonathan Burkardt | 1000000 | 19 | 0.36 | 0 | 0 | 0 | 0 |
| ***307*** | Juan Villar | 65000 | 31 | 0.37 | 0 | 0 | 1 | 1 |

*Tabla 1. Desempeño de delanteros clúster 4, algoritmo jerárquico.*

De la Tabla 1 se puede concluir que Antoine Griezmann, delantero del Real Madrid tiene una alta efectividad en goles por partido, pero en las demás métricas de desempeño no registra datos, el caso de Juan Villar es bastante interesante, tiene un desempeño mayor a Griezmann, pero su valor no es ni el 1% del de Griezmann y su edad para la época era de 31 años. Para un fichaje se ve interesante el jugador Jonathan Burkardt, con tan solo 19 años y un desempeño cercano al de Griezmann.

Los resultados obtenidos son satisfactorios, permiten responder la pregunta planteada y propone una futura implementación de, una herramienta para el soporte en la toma de decisiones como son el fichaje de jugadores y de estrategias de mejora técnica para los jugadores, además, posiblemente también se pudiera lograr proponer estrategias de juego contra distintos oponentes. Es bastante interesante el resultado del algoritmo, ya que agrupa los jugadores por características que apuntan a la calidad del jugador y su posición en el campo, con las 27 variables identificadas transversales a los clústeres, importantes en la ponderación de una media de calidad del jugador, así como algunas propias de cada clúster. Es posible navegar en la base de datos para encontrar fácilmente jugadores con ciertas características de interés. Como limitación del trabajo desarrollado no se cuenta con algoritmos adicionales de comparación y una medida de desempeño entre ellos, aunque se logró implementar el clustering jerárquico, aglomerativo y DBSCAN sin encontrar diferencias significativas al analizar agrupaciones de 4 clústeres.

Como trabajo futuro es posible automatizar la elección de las características brindándole al usuario la posibilidad de elegirlas a partir de su objetivo o pregunta de negocio, así mismo implementar diferentes algoritmos, para que el usuario pueda seleccionar y comparar las respuestas a su pregunta a través de cada abordaje a través del resultado y una medida de desempeño entre los mismos.

**CONCLUSIONES**

El estudio aquí presentado muestra un posible camino para implementar una herramienta de recomendación para la industria del fútbol basada en estadísticas por jugador y algoritmos para clusterización. La agrupación de jugadores en 4 clústeres que se implementaron tanto con el algoritmo jerárquico aglomerativo y el basado en DBSCAN permitieron señalar jugadores similares y de alto nivel junto con otros nuevos que pueden ser fichados. Los agrupamientos permitieron distinguir características que estaban asociados con los delanteros (como número de goles o disparos al arco). Un segundo grupo implementaba más las variables de pases de todo tipo, siendo identificados con armadores y creadores de juego. Por último, otra clasificación que aparece en un grupo es la de defensores, que contaban en su mayoría con variables del tipo interceptaciones en área de peligro y faltas.

## BIBLIOGRAFÍA


[1] Diego A. Bonilla, Javier O. Peralta-Alzate, Jhonny A. Bonilla-Henao, Wilson Urrutia-Mosquera, Roberto Cannataro, Jana Koci, Jorge L. Petro. Unsupervised machine learning analysis of the anthropometric characteristics and maturity status of young Colombian athletes. Journal of Physical Education and Sport (JPES), Vol. 22 (issue 1), Art 33, pp. 256 - 265, January 2022. doi:10.7752/jpes.2022.01033

[2] Marco Tosato, Jianhong Wu. An application of PART to the Football Manager data for players clusters analyses to inform club team formation. Big Data & Information Analytics, 2018, doi: 10.3934/bdia.2018002

[3] https://www.transfermarkt.co

[4]https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics?select=transfermarkt_fbref_201920.csv

[5] Ignacio Sarmiento-Barbieri. Cuaderno de aprendizaje no supervisado, Análisis de Componentes Principales. Fundamentos Teóricos.

[6] Jakob E. Persson, Emil Danielsson. Avatar Playing Style, From analysis of football data to recognizable playing styles, UPPSALA UNIVERSITET, Master Thesis, 2022.

[7] https://www2.deloitte.com/co/es/pages/consumer-business/articles/football-money-league-2022.html


## ANEXO

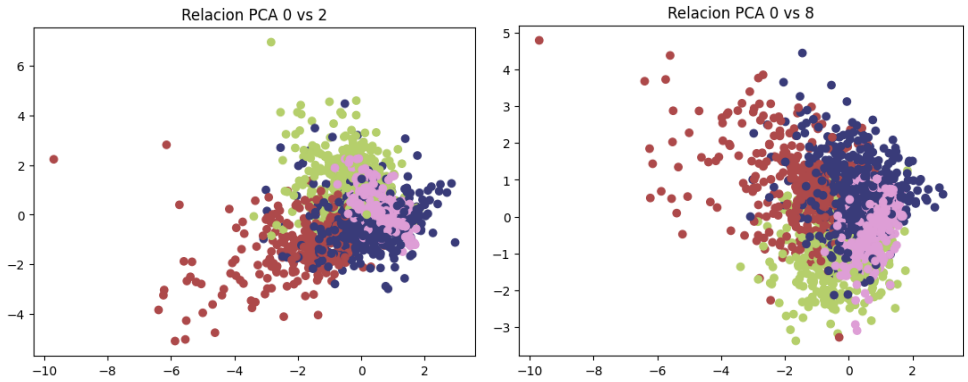


*Figura 8. Visualización en dos dimensiones de las componentes principales 0 y 2 (izquierda) y componentes principales 0 y 8 (derecha).*

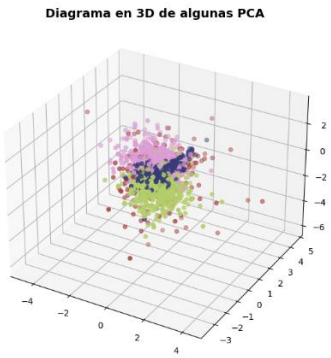


*Figura 9. Proyección en 3D de los 4 clústeres del algoritmo jerárquico aglomerativo.*