GRADO EN INTELIGENCIA Y ANALÍTICA DE NEGOCIOS

WEBSCRAPING Y CLUSTERING APLICADO A LA LIGA ESPAÑOLA

ALUMNO/A: Javier Carapeto, Alberto España, Carles Gomar y Mario Guillem
TUTOR/A: Francisco Gabriel Morillas Jurado

CURSO ACADÉMICO: Datos No Estructurados
FECHA DE DEPÓSITO: 4º BIA (08/01/2024)

VNIVERSITAT DE VALÈNCIA Facultat d'Economia

Índice

Resumen

1. **Introducción**
   - **1.1 Contexto general**
   - **1.2 Objetivos**
   - **1.3 Hipótesis**
2. **Marco teórico: analítica en el fútbol**
3. **Metodología**
   - **3.1 Recogida y limpieza de datos**
   - **3.2 Análisis Exploratorio de Datos**
     - **3.2.1 Descripción de variables**
     - **3.2.2 Análisis descriptivo por posición en campo**
   - **3.3 Técnicas empleadas**
     - **3.3.1 Webscraping**
     - **3.3.2 Modelos de regresión**
     - **3.3.3 Clustering**
     - **3.3.4 Análisis de varianza (ANOVA)**
     - **3.3.5 Método de Scheffé**
4. **Resultados**
   - **4.1 Análisis exploratorio**
   - **4.2 Regresiones**
   - **4.3 Desempeño de los jugadores por posición de campo**
     - **4.3.1 Porteros**
     - **4.3.2 Defensas**
     - **4.3.3 Mediocentros**
     - **4.3.4 Delanteros**
5. **Conclusiones**
   - **5.1 Futuras líneas de investigación**
6. **Bibliografía**
7. **Anexo**

Índice de tablas y figuras

Tabla 1. Posiciones en campo de juego
Tabla 2. Top asistencias
Tabla 3. Top goleadores
Tabla 4. Top 3 porteros
Tabla 5. Defensas. Bloqueos
Tabla 6. Defensas. Tackles
Tabla 7. Top 5 pasadores
Tabla 8. Tiros por 90 minutos

Figura 1. Número óptimo de clusters Porteros
Figura 2. Resultados clustering K-means Porteros
Figura 3. Número óptimo de clusters Defensas
Figura 4. Resultados clustering K-means Defensas
Figura 5. Número óptimo de clusters Mediocentro
Figura 6. Resultados clustering K-means Mediocentro
Figura 7. Número óptimo de clusters Delanteros
Figura 8. Resultados clustering K-means Delanteros

Resumen

En el vertiginoso escenario del fútbol contemporáneo, la recopilación y análisis de datos se ha convertido en una herramienta esencial para comprender a fondo el rendimiento de los equipos y jugadores.

En este contexto, el presente trabajo estudiará el rendimiento de los diferentes equipos y jugadores que conforman LaLiga EA Sports (Primera División Española de Fútbol), más concretamente centrándose en la última temporada completada, es decir, la temporada 2022- 2023.

En este contexto, la aplicación de técnicas avanzadas de análisis, como el clustering, se presenta como una herramienta para descubrir patrones, perfiles y relaciones ocultas entre los jugadores que conforman la competición. Además, se tendrán en cuenta diferentes técnicas como ANOVA o modelos de regresión que nos permitan comprender las estadísticas de nuestra base de datos.

El principal objetivo de este estudio será el uso de técnicas de aprendizaje supervisado como no supervisado, para encontrar perfiles de jugadores que puedan destacar sobre el resto. Dentro de este enfoque, se tendrán en cuenta también las estadísticas de los equipos a los que pertenecen los jugadores. Asimismo, previamente se realizará sobre el conjunto de datos un análisis exploratorio, mediante el cual poder comprender mejor dichos datos.

Abstract

In the fast-paced scenario of contemporary football, data collection and analysis has become an essential tool to fully understand the performance of teams and players.

In this paper we will study the performance of the different teams and players that make up LaLiga EA Sports (Spanish First Division Football), more specifically we will focus on the last completed season, i.e. the 2022- 2023 season.

In this context, the application of advanced analysis techniques, such as clustering, is presented as a tool to discover patterns, profiles and hidden relationships among the players that make up the competition. In addition, different techniques such as ANOVA or regression models that allow us to understand the statistics of our database will also be taken into account.

The main objective of this study will be the use of supervised and unsupervised learning techniques to find player profiles that can stand out from the rest. Within this approach, the statistics of the teams to which the players belong will also be taken into account. Moreover, an exploratory analysis will be previously carried out on our dataset, in order to better understand the data.

Palabras clave */ Keywords*

Webscraping, cluster, fútbol, regresión, ANOVA

1. Introducción

1.1 Contexto general

El fútbol, como deporte universal, ha evolucionado más allá de ser simplemente un juego; se ha convertido en un fenómeno social y cultural que capta la atención de millones de aficionados en todo el mundo. La liga española, conocida por su competitividad y calidad técnica, alberga algunos de los equipos y jugadores más destacados a nivel global.

La riqueza de información contenida en los datos estadísticos de fútbol puede ofrecer perspectivas valiosas sobre estrategias de juego, rendimiento individual y colectivo, así como tendencias tácticas que delinean el curso de una temporada. Para obtener estos datos de manera eficiente y completa, se emplea la técnica de web scraping en la página web fbref.com, una plataforma reconocida por su exhaustividad y precisión en la recopilación de estadísticas de fútbol.

A lo largo de este trabajo, se explorará el proceso de web scraping como herramienta para la obtención sistemática de datos, centrándonos específicamente en la Liga española de fútbol durante la temporada 2022-2023. La metodología utilizada permitirá la extracción de información detallada sobre equipos, jugadores, partidos y otras métricas relevantes, proporcionando así una base sólida para el análisis y la interpretación de los fenómenos que moldean la competición.

Así pues, este estudio no solo aspira a ofrecer una radiografía detallada de los datos recopilados, sino también a contribuir al campo de la ciencia de datos aplicada al fútbol, brindando nuevas perspectivas para entender y apreciar el juego desde una perspectiva analítica. Y es que, a medida que avanzamos en la era digital del deporte, el análisis de datos se erige como un pilar fundamental para el desarrollo estratégico y la toma de decisiones informada en el ámbito futbolístico. Precisamente, este trabajo se sitúa en la vanguardia de esta evolución.

1.2 Objetivos

Tras contextualizar y poner en relieve el problema de estudio, se pasará a definir los objetivos que se persiguen y, posteriormente, las distintas hipótesis iniciales que se pretenden contrastar.

El **objetivo principal** de este proyecto es analizar las diferencias en el desempeño de los jugadores de la liga española por posición en el campo (portería, defensa, mediocentro y delantera) durante la temporada 2022-2023 proponiendo una clasificación óptima de los mismos atendiendo a una serie de variables. Se pretende distinguir diferentes perfiles de jugadores en cada posición del campo, de manera que puedan servir de ayuda a hipotéticos directivos u ojeadores de los distintos clubes de fútbol a la hora de evaluar la

eficacia de los jugadores actuales o, incluso, enfocar un traspaso para un tipo de jugador determinado.

Por otro lado, se incluye como **objetivo secundario** señalar indicios del desempeño de los equipos de la liga española durante la temporada 2022-2023.

1.3 Hipótesis

Como hipótesis de partida para el desarrollo del estudio se han tenido en cuenta los siguientes supuestos:

Habrá jugadores de países considerados como "menos convencionales" en el mundo del fútbol cuyo desempeño en la consecución de goles sea significativo. Países asiáticos, del este de Europa o de Norteamérica.

En segundo lugar, se espera que los delanteros y los porteros sean significativos en relación a los goles, con coeficientes positivos y negativos respectivamente.

Los equipos que mejor porcentaje de pases completado tienen son el Real Madrid y el FC Barcelona.

Asimismo, se espera encontrar un grupo de porteros con menos goles en contra por 90 minutos y un mayor porcentaje de partidos con portería a 0 y de salvadas de forma significativa respecto al resto de grupos.

Se espera encontrar un grupo de defensas con un número de goles significativamente superior al resto correspondiente a los defensas que más suben y mejor van de cabeza, con un perfil más atacante. Por otro lado, se espera encontrar un grupo de defensas con un número de bloqueos y porcentaje de regateadores tacleados significativamente superior al resto correspondiente a los defensas más físicos.

Se espera encontrar un grupo de mediocentros con un número de goles significativamente superior al del resto correspondiente a los mediocentros más ofensivos; y por otra parte un grupo con un porcentaje de pases completados y asistencias significativamente superior al del resto, correspondiente a los mediocentros con mejor visión de juego.

Se espera encontrar un grupo de delanteros con un número de goles y un número de tiros a porteria por cada 90 minutos significativamente superior al del resto de grupos, correspondiendo a los delanteros que más disparan y aciertan; además esperamos otro grupo con un número de asistencias significativamente superior, correspondiendo a los delanteros que destacan más por hacer asistencias de gol.

2. Marco teórico

La obtención de datos de fútbol se encuentra en constante evolución, pudiendo conseguirse tanto manualmente mediante empleados dedicados a anotar todas las acciones que tienen lugar en un partido, como mediante tecnología de inteligencia artificial.

La evolución tecnológica también ha llegado al mundo del fútbol y de los datos. Existen cámaras como las SportsVU que mediante inteligencia artificial reconocen a los jugadores, sus dorsales y permiten registrar toda la información que se requiere. Otras técnicas que permiten la obtención de datos es la de "*trackers*" con las que medir diferentes métricas.

Empresas como Stats Perform o FBref, que se dedican a la explotación de datos en el mundo del deporte, ofrecen datos públicos con el objetivo de atraer clientes. Sin embargo, la otra vertiente de la empresa se dedica a la venta de servicios y de datos mucho más precisos.

En el contexto de un análisis de estadísticas de jugadores de fútbol utilizando R, este trabajo se propone en primera instancia recopilar datos. Existen diferentes técnicas gratuitas para ello como la descarga directa, las APIs o el webscraping. En este estudio se utilizará esta última alternativa. Una vez obtenidos, se pasará a realizar un análisis descriptivo de los mismos, para posteriormente seleccionar características relevantes, normalizar y transformar datos, y aplicar técnicas de clustering que permitan categorizar jugadores en grupos de acuerdo con un desempeño similar.

El uso de herramientas como R permite una comprensión más profunda del rendimiento de los jugadores. Trabajos previos, como "*A Review of Data Mining Techniques for Result Prediction in Sports*" (Haghighat y Rastegari, 2013) ofrecen una visión general de metodologías en análisis de datos en fútbol, mientras que "*Clustering of football players based on performance data and aggregated clustering validity indexes*" (Akhanii y Henning, 2022) se centra específicamente en técnicas de clustering para la categorización de jugadores. Estos trabajos se centran en varias ligas en general pero ninguna realiza un análisis más detallado de la liga española. Por lo que se decide adentrarse en esta en exclusivo.

3. Metodología

3.1 Recogida y limpieza de datos

Para la elaboración de este informe, se llevaron a cabo extracciones de datos a través de la técnica de webscraping. Esta técnica se aplicó a la plataforma de estadísticas de fútbol, fbref.com. El proceso de recolección de datos se centró específicamente en la Liga de Fútbol Profesional de España, correspondiente a la temporada 2022-2023. La elección de fbref.com como fuente principal se debe a su reputación por proporcionar información detallada y confiable sobre diversos aspectos del rendimiento de los jugadores y equipos en el ámbito futbolístico.

Los datos recopilados abarcan una variedad de variables cruciales que permitirán un análisis exhaustivo del desempeño individual de los jugadores en la mencionada temporada. Entre estas variables se incluyen aspectos fundamentales como la nacionalidad de los jugadores, los clubes en los que militan, los minutos jugados en la temporada, el número de goles anotados, la cantidad de pases completados con éxito, los tiros realizados a lo largo de los encuentros, entre otras variables. La diversidad de estas estadísticas ofrece una visión integral de las habilidades y contribuciones de los jugadores en diferentes aspectos del juego, proporcionando así una base sólida para el análisis posterior.

Cabe destacar que la elección de estas variables específicas se fundamenta en su importancia para comprender no solo el rendimiento global de los jugadores, sino también para desentrañar patrones y tendencias que podrían influir en la clasificación y perfilado de los mismos. La minuciosidad en la extracción de datos provenientes de fbref.com garantiza la calidad y la actualidad de la información recopilada, siendo esencial para la fiabilidad y robustez de los análisis posteriores en este proyecto.

3.2 Análisis Exploratorio de Datos

3.2.1 Descripción de variables

El presente estudio se ha centrado sobre cuatro temas principales, los cuales son:

- Disparos de jugadores (tiros)
- Pases en equipo
- Acciones defensivas
- Portería del equipo

Cada tema ha generado una base de datos, para finalmente unificar las 4 bases de datos en una única. Así pues, las variables extraídas sobre las cuales se centrará el presente estudio pueden agruparse en dos grupos principalmente:

- Características básicas del jugador (variables en común en todas las bases de datos):
    - *Jugador:* Nombre y apellido del futbolista.
    - *País*: Iniciales de la nacionalidad del futbolista (Por ejemplo ESP equivale a España).
    - *Posc*: Posición que ocupa en el campo de fútbol (portero -PO-, delantero -DL-, mediocentro/centrocampista -CC-, defensa -DF- o una combinación de éstas).
    - *Equipo*: Club deportivo al que pertenece.
    - *Edad* y *nacimiento*: edad y año de nacimiento del jugador.
    - *90 s.x:* minutos jugados dividido entre la duración de un partido (90 min).

- Estadísticas del desempeño del jugador (que agrupadas por las secciones temáticas anteriormente presentadas quedan de la siguiente manera):
    - Disparos de jugadores (tiros)
        - *Gls*.: goles marcados
        - *T/90*: tiros al arco cada 90 minutos
    - Pases en equipo
        - *por_pases_completados*: porcentaje de cumplimiento de pases (completados entre intentados).
        - *Ass*: número de asistencias. Éstas se contabilizan cuando un jugador asiste a un compañero dejándole en ocasión manifiesta de gol.
    - Acciones defensivas
        - *Bloqueos*: número de veces que se bloquea el balón poniéndose en su camino.
        - *TKL%:* porcentaje de regateadores "*tacleados*" (hacerle una entrada a un jugador).
    - Portería de equipo
        - *GC90*: goles en contra por 90 minutos.
        - *%Salvadas*: porcentaje de salvadas (tiros mayoritariamente tapados por el portero -sin incluir penales-).
        - *PaC%*: porcentaje de partidos que dan como resultado una portería a 0.

Alternativamente, las estadísticas del desempeño de los jugadores agrupadas por la posición en el campo de juego quedarían de la siguiente manera.

- Portería
    - *GC90*
    - *% Salvadas*
    - *PaC%*
- Defensa
    - *Gls*.
    - *Bloqueos*
    - *TKL%*

- Mediocentro
    - *Gls*.
    - *Ass*
    - *por_pases_completados*
- Delantera
    - *Gls*.
    - *Ass*
    - *T/90*

Esta división de variables resulta fundamental para realizar los futuros agrupamientos y es la que se tendrá en cuenta a la hora de realizar los análisis posteriores y extraer las conclusiones pertinentes.

3.2.2 Análisis descriptivo por posición en campo

Se ha llevado a cabo un estudio descriptivo de los datos disponibles en función de las posiciones que ocuparan los jugadores. Se ha tomado como base el *database* único para después segmentar en 4 diferentes atendiendo a los valores ofrecidos por la variable "*Posc".* Precisamente, en la Tabla 1 puede observarse la equiparación entre valores/siglas y posición en campo de juego.

- Porteros
    - Valor: "PO"
- Defensas
    - Valores: "DF", "DF-CC", "DF-DL".
- Mediocentros
    - Valores: "CC", "CC-DF" , "CC-DL"
- Delanteros
    - Valores: "DL", "DL-CC", "DL-DF"

Tabla 1. Posiciones en el campo de juego

VALOR	*SIGNIFICADO*
PO	Porteros
DF	Defensas
CC	Centrocampistas
DL	Delanteros

Fuente: Elaboración Propia

Una vez segmentado el "data-frame" inicial en 4, se han realizado medidas de estadística descriptiva como la media o varianza de determinadas variables, cuantiles, máximos y mínimos, representación gráfica mediante histogramas y tablas para poder mostrar de una forma más visual la información extraída.

3.3 Técnicas empleadas

3.3.1 Webscraping

El webscraping con R es una técnica utilizada para extraer datos de páginas web de manera automatizada.
R es un lenguaje de programación y un entorno de desarrollo estadístico que cuenta con herramientas y bibliotecas especializadas para realizar estas tareas de extracción de información web de manera eficiente.

El proceso implica enviar solicitudes HTTP a la página web de interés, analizar el código HTML resultante y extraer los datos específicos de interés.

En R, las bibliotecas más comunes para llevar a cabo web scraping incluyen rvest, httr, xml2, y tidyverse. La combinación de estas bibliotecas permite cargar y analizar el contenido HTML de una página web, seleccionar y extraer los elementos deseados, y luego organizar esos datos para su posterior análisis.

En nuestro estudio, definimos una función de web scraping llamada *extraer_datos_automatico* que se utiliza para extraer datos de estadísticas de fútbol de la web fbref.com. La función toma una palabra clave como argumento y utiliza esa palabra clave para construir la URL y el XPath necesario para encontrar y extraer la tabla de datos correspondiente. Se extraen datos de tiros, pases, defensa y porteros, utilizando las palabras clave "shooting", "passing", "defense" y "keepers" respectivamente.

En primer lugar, la función verifica si la palabra clave es "keepers". Si es así, ajusta la URL y el XPath específicamente para la sección de porteros. Esto se debe a que en la URL utiliza la palabra “keepers” mientras que en el xpath utiliza “keeper”, esto no sucede en el resto de palabras clave.

Si la palabra clave no es "keepers", construye la URL y el XPath utilizando la palabra clave en minúsculas.

Seguidamente, se utiliza la función read_html del paquete rvest (Wickham H (2023)) para leer la página web correspondiente a la URL construida. A continuación, para poder acceder a las tablas de los datos las cuales se encuentran comentadas, se tiene que extraer el contenido de los comentarios en la página HTML y eliminar los caracteres de comentario <!-- y --> del texto completo. Después, se convierte el texto sin comentarios a un objeto HTML utilizando la función read_html. Entonces, se utiliza el XPath para encontrar el nodo HTML que contiene la tabla de datos y se convierte la tabla a un dataframe utilizando la función html_table. Finalmente, se realizan algunos ajustes en los datos.

3.3.2 Modelos de regresión

Los modelos de regresión lineal son muy populares en diversos campos de investigación gracias a su rapidez y facilidad de interpretación.

Debido a su capacidad para transformar datos, pueden utilizarse para simular una amplia gama de relaciones, y debido a su forma, que es más simple que la de las redes neuronales, sus parámetros estadísticos se analizan y comparan con facilidad, lo que permite que se les extraiga información valiosa (José Ángel Saavedra, 2023). En este sentido, los modelos de regresión que implementaremos en el análisis posterior serán:

- *Modelo Lineal (LM):* enfoque matemático que describe la relación entre una variable dependiente y una o más variables independientes mediante una función lineal. La forma más básica de un modelo lineal con una variable independiente se expresa de la siguiente manera:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Donde Y es la variable independiente, β_0 y β_1 son dos constantes desconocidas que representan el punto de intersección y la pendiente respectivamente y ε es el residuo o error. La función de ε es explicar la posible variabilidad de los datos que no pueden explicarse a través de la relación lineal de la fórmula.

- *Modelos Lineal Generalizado (GLM):* asume que la influencia de las variables explicativas sobre la variable respuesta se produce de forma lineal, siguiendo la siguiente expresión (Cayuela, 2009):

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

3.3.3 Clustering

El término Clustering hace referencia a las técnicas de aprendizaje no supervisado en las cuales su objetivo principal es agrupar a las diferentes observaciones en grupos o clusters. Se busca la homogeneidad entre las observaciones de cada grupo y heterogeneidad entre grupos. Debido al desarrollo de esta técnica durante los años, se han diferenciado 3 grandes familias:

- *Partitional Clustering*: Indicando previamente el número de clusters que el usuario desea. Técnicas: K-means, K-medoids…
- *Hierarchical Clustering:* No es necesario indicar el número de clusters al inicio del proceso. Dentro de esta familia se engloban:
    - Agglomerative clustering: Cada observación comienza como un cluster y se van agrupando.
    - Divise clustering: Todas las observaciones comienzan siendo parte del mismo cluster y se van dividiendo fase a fase.
- Métodos que combinan los dos métodos anteriores: Fuzzy Clustering.

El presente estudio ha optado por el método K-means (MacQueen, 1967), el cual agrupa todas las observaciones de una base de datos en un número determinado *K* que el usuario establece antes de desarrollar el algoritmo. El objetivo de este algoritmo es encontrar los K mejores clusters en los cuales la varianza interna a cada cluster sea la mínima posible.

La aplicación de la técnica K-means se lleva a cabo en el software Rstudio utilizando el paquete stats (R Core Team (2022)) el cual incluye la función kmeans que realiza el algoritmo de agrupamiento K-means en un conjunto de datos.

En el contexto del estudio, se pretende realizar agrupamientos de jugadores para cada una de las posiciones que puede ocupar un jugador en el campo (portería, defensa, mediocampo, delantera) según el número óptimo que se considere. Para ello evalúa las diferentes variables que se han definido para cada posición.

El número de agrupamientos que realice se decidirá utilizando el método del codo (Elbow method), el cual se utiliza para identificar el punto en el cual la disminución de la suma de cuadrados dentro de los clusters (WSS) comienza a aplanarse, formando un "codo" en el gráfico. Claude, J. (2008). En R aplicaremos la función *fviz_nbclust* del paquete factorextra (Kassambara, A. and Mundt, F. (2020)).

3.3.4 Análisis de varianza (ANOVA)

Analizar la variación en una variable de respuesta (variable continua aleatoria) medida en circunstancias definidas por factores discretos (variables de clasificación) (Dagnino, 2014).

La técnica de análisis de varianza (ANOVA) también conocida como análisis factorial y desarrollada por Fisher en 1930, constituye la herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la media de una variable continua. Es por lo tanto el test estadístico a emplear cuando se desea comparar las medias de dos o más grupos.

La hipótesis nula de la que parten los diferentes tipos de ANOVA es que la media de la variable estudiada es la misma en los diferentes grupos, en contraposición a la hipótesis alternativa de que al menos dos medias difieren de forma significativa.

En este estudio, se aplica la técnica análisis de varianza (ANOVA) en el software Rstudio. La función *aov* en R del paquete stats (R Core Team (2022)) se utiliza para realizar análisis de varianza (ANOVA). El output de esta función proporciona información sobre la variabilidad entre grupos y dentro de grupos. Se busca el valor p asociado con la estadística F, el cual indica la probabilidad de observar el resultado si la hipótesis nula (no hay diferencias significativas) es verdadera. Si el valor p es menor que un umbral (por ejemplo, 0.05), se considera evidencia de diferencias significativas.

En este estudio, se aplica ANOVA a cada una de las variables utilizadas en el algoritmo K-means, para evaluar si las medias de esas variables son diferentes entre los clusters

generados por el algoritmo K-means. Dado que la variable que se está analizando con ANOVA es la misma que se utilizó para formar los clusters, es probable que se encuentren diferencias significativas entre los grupos simplemente debido al proceso de clusterización. No obstante, en el caso del estudio encuentra el sentido porque se está interesado en analizar diferencias más allá de las estructuras de clusters originales a través de pruebas post hoc que se definirán a continuación.

3.3.5 Método de Scheffé

La prueba de Scheffé (1959) es un procedimiento post hoc utilizado en el análisis de varianza (ANOVA) para comparar todas las combinaciones posibles de medias entre grupos después de haber encontrado diferencias significativas en el análisis global. Se incluye dentro de las llamadas pruebas de comparaciones múltiples. La prueba de Scheffé es conocida por ser conservadora, lo que significa que tiende a ser menos propensa a cometer errores tipo I (rechazar falsamente la hipótesis nula). (Hinton, P. R. (1995))

En este estudio, se aplica utilizando la función *scheffe.test* del paquete agricolae (Felipe de Mendiburu and Muhammad Yaseen (2020)).

En el contexto del estudio, la utilidad del test de Scheffé radica en la necesidad de explorar estadísticamente las diferencias entre los grupos identificados mediante K-means en relación con las variables específicas que estás analizando.

4. Resultados

4.1 Análisis exploratorio

En primer lugar, se estudió el número de minutos que disputaron los jugadores. En LaLiga EA Sports se disputan un total de 38 encuentros por cada equipo, siendo sólo 3 jugadores aquellos que los jugaron todos completos, es decir, disputaron un total de 3420 minutos de juego. Los futbolistas fueron:

- Giorgi Mamardashvili, portero del Valencia CF.
- Álex Remiro, portero de la Real Sociedad de Fútbol.
- David Soria, portero del Getafe CF.

El jugador de campo que más minutos realizó fue Fran García, lateral izquierdo del Rayo Vallecano con tan solo 10 minutos de descanso en toda la competición.

En segundo lugar, se estudiaron los jugadores con más asistencias, que podemos ver en la Tabla 2.

Tabla 2: TOP ASISTENCIAS

Jugador	Equipo	Ass
Antoine Griezmann	Atlético Madrid	16
Vinicius Júnior	Real Madrid	9
Mikel Merino	Real Sociedad	9
Rodrygo	Real Madrid	8
Rodrigo De Paul	Atlético Madrid	7
Ousmane Dembélé	Barcelona	7
Robert Lewandowski	Barcelona	7

Fuente: Elaboración Propia

El jugador con más asistencias es Antoine Griezmann, siendo significativamente superior al resto de jugadores con un total de 16 asistencias. La media de asistencias por jugador durante toda una temporada se encuentra en 1,09 asistencias.

Tabla 3: TOP GOLEADORES

Jugador	Equipo	Gls.
Robert Lewandowski	Barcelona	23
Karim Benzema	Real Madrid	19
Joselu	Espanyol	16
Antoine Griezmann	Atlético Madrid	15
Borja Iglesias	Betis	15

Fuente: Elaboración Propia

Por lo que respecta al número de goles marcados en la temporada, en la Tabla 3, encontramos nombres conocidos como Lewandowski, Benzema o Griezmann.

Cabe destacar la aparición del jugador del Atlético de Madrid como el máximo asistente además de ser uno de los máximos goleadores.

Seguidamente se estudió la variable de la Edad de los jugadores. El jugador más joven en disputar la competición fue el jugador del FC Barcelona Lamine Yamal con 15 años de edad (13 de julio de 2007). En el otro extremo encontramos a Joaquín, jugador del Betis, con 41 años de edad (21 de julio de 1981).

Un dato a destacar es que Joaquín cuenta con más años como jugador profesional (23 años), que la propia edad de Lamine Yamal.

Una vez analizadas estás características genéricas de los jugadores, se procede a realizar el análisis exploratorio de cada posición. Se han tenido en cuenta solo aquellos jugadores que hayan disputado más de 10 partidos con el objetivo de obtener resultados consistentes.

Porteros

Tan solo 27 porteros de 44 inscritos disputaron más de 10 partidos, es decir, el 61%.

Tabla 4: TOP 3 PORTEROS

Jugador	Equipo	PaC%	GC90
Marc-André ter Stegen	Barcelona	68.4	0.48
Gerónimo Rulli	Villarreal	50.0	0.71
Álex Remiro	Real Sociedad	39.5	0.92

Fuente: Elaboración Propia

En la tabla 4, podemos constatar los porteros con mejores estadísticas, tanto de % de porterías a 0 (Pac%), como de goles en contra por 90 minutos. Podemos ver como el guardameta del Barcelona ocupa una holgada primera posición, seguido por Rulli y Álex Remiro.

Cabe destacar que Álex Remiro cuenta con unas destacables estadísticas siendo el único de estos 3 porteros que ha disputado todos los partidos.

Defensas

Por lo que respecta a los defensas, destaca notablemente la actuación de Javi Galán con un total de 104 bloqueos o cortes de jugada durante toda la competición. Muy de lejos le siguen Alfonso Espino del Cádiz con 87 jugadas cortadas. En la Tabla 5 podemos ver estas estadísticas

Tabla 5: DEFENSAS

Bloqueos totales

Jugador	Bloqueos
Javi Galán	104
Alfonso Espino	87
Fran Garcia	84
José Luis Gayà	80
Carlos Clerc	75

Fuente: Elaboración Propia

En las jugadas de uno para uno los defensas que mejores resultados tienen son los mencionados en la Tabla 6.

Tabla 6: DEFENSAS

Tackles

Jugador	Equipo	Tkl%
David García	Osasuna	86.4
Sergi Gómez	Espanyol	81.8
Karim Rekik	Sevilla	80.0
Daniel Vivian	Athletic Club	80.0
Éder Militão	Real Madrid	78.8

Fuente: Elaboración Propia

Mediocentros

En referencia a los mediocentros que tienen un mejor porcentaje de pases completados durante toda la temporada se encuentran jugadores de renombre como los que encontramos en la Tabla 7.

Tabla 7: TOP 5 PASADORES

% pases compeltados

Jugador	Equipo	por_pases_completados
Aurélien Tchouaméni	Real Madrid	92.8
Toni Kroos	Real Madrid	90.9
Eduardo Camavinga	Real Madrid	90.7
Dani Ceballos	Real Madrid	90.6
Frenkie de Jong	Barcelona	90.0

Fuente: Elaboración Propia

En efecto, 4 de los 5 futbolistas con mejores porcentajes son del Real Madrid, siendo el único jugador del equipo contrario Frenkie de Jong del FC Barcelona.

Delanteros

Finalmente, tras analizar el resto de posiciones se procede a comentar los resultados de los delanteros. Los 5 jugadores con más goles como se indica en la Tabla 3, son en su totalidad delanteros.

Tabla 8: Tiros por 90 minutos

Jugador	Equipo	T/90
Raúl García	Athletic Club	4.84
Karim Benzema	Real Madrid	4.37
Robert Lewandowski	Barcelona	4.24
Ansu Fati	Barcelona	3.94
Rodrygo	Real Madrid	3.67

Fuente: Elaboración Propia

En la Tabla 8, se aprecia como el jugador que más dispara a puerta es el español Raúl García con una media de casi 5 disparos por partido, el otro español del top es Ansu Fati con casi 4 disparos por partido, ambos no aparecen entre los 10 máximos goleadores de la liga.

4.2 Regresiones

Se realizaron 3 modelos de regresión lineal generalizada que nos permitieran comprender las estadísticas.

En primer lugar y en relación a la primera hipótesis se explicaron los goles en función de las nacionalidades de los jugadores. Los países que son significativamente diferentes al resto son Canadá, Japón, Corea del Sur, Kosovo, Noruega, Polonia y Turquía.

En segundo lugar, se realizó una regresión con el objetivo de entender los goles en función de las posiciones de los jugadores y las posiciones a destacar fueron los delanteros y los defensas; los atacantes con un coeficiente de goles positivo, mientras que los defensores obtuvieron un coeficiente negativo. Se esperaba que los porteros fueran significativos al realizar todos 0 goles, pero no fue así. Intuimos que al tener una muestra tan grande de jugadores el modelo no toma como outliers a los porteros

Finalmente se estudió el % de pases completado en función del equipo de los jugadores. Se obtuvieron significativos 2 equipos, Real Madrid y FC Barcelona. Los resultados concuerdan con los establecidos en la Tabla 7.

4.3 Desempeño de los jugadores por posición de campo

Tal y como se ha definido anteriormente, se busca analizar la heterogeneidad de los jugadores de la liga española por posición en el campo proponiendo una clasificación óptima de los mismos atendiendo a una serie de variables.

En primer lugar, se agrupan los jugadores por posición en el campo (portería, defensa, mediocentro y delantera) y a partir de ahí se repite el procedimiento que sigue para cada posición.

El presente estudio, tal y como se ha comentado anteriormente, ha optado por el método K-means (MacQueen, 1967). K-means agrupa todas las observaciones disponibles en la base de datos en diferentes grupos en función de su desempeño en las variables definidas como relevantes para cada posición. La aplicación de la técnica K-means se lleva a cabo en el software Rstudio utilizando el paquete stats (R Core Team, 2022) el cual incluye la función kmeans.

No obstante, antes de aplicar la técnica k-means se debe determinar el número de agrupamientos óptimo. Para ello se utiliza el método del codo (Claude, J., 2008) explicado anteriormente.

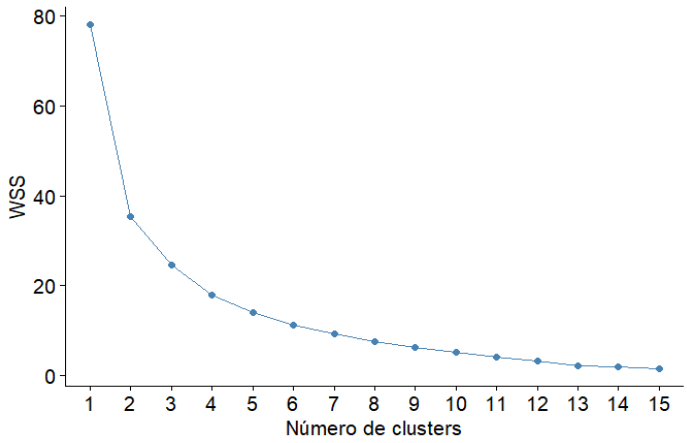
A continuación, se aplica la técnica K-means (MacQueen, 1967) para obtener el número de agrupamientos definido.

Seguidamente se aplica ANOVA a cada una de las variables utilizadas en el algoritmo K-means mediante la función aov (R Core Team, 2022) , para evaluar si las medias de esas variables son diferentes entre los clusters generados por el algoritmo K-means.

Finalmente, se aplica el test de Scheffé (Hinton, P. R., 1995) mediante la función scheffe.test (Felipe de Mendiburu and Muhammad Yaseen, 2020). con tal de explorar estadísticamente las diferencias entre los grupos identificados mediante K-means en relación con las variables específicas que estás analizando.

4.3.1 Porteros

Figura 1: Número óptimo de clústeres Porteros

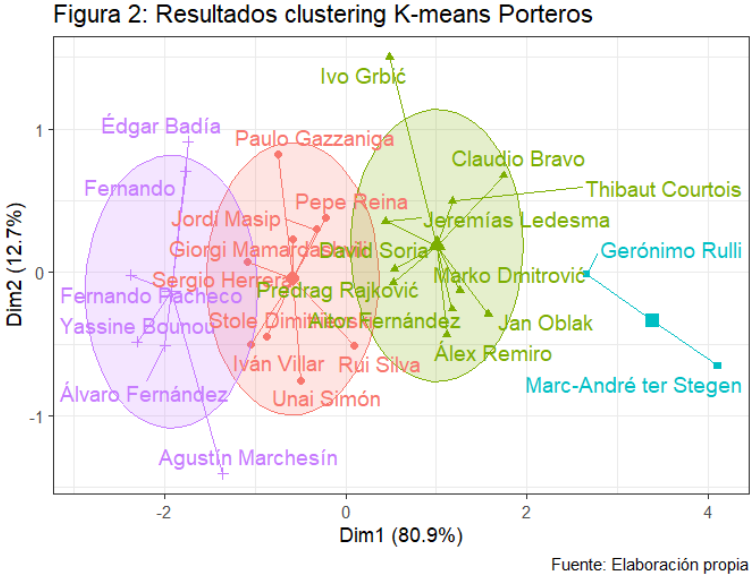


Fuente: Elaboración propia

El punto en el cual la disminución de la suma de cuadrados dentro de los clusters (WSS) comienza a aplanarse, formando un "codo" en el gráfico representado en la Figura 1, es a partir de cuatro agrupamientos.

Seguidamente se aplica la técnica K-means para el número de agrupamientos definido y se obtiene los siguientes resultados:

Figura 2: Resultados clustering K-means Porteros



Fuente: Elaboración propia

Según la Figura 2, los cuatro agrupamientos construidos por el método K-means son los siguientes:

- **Cluster 1:** Giorgi Mamardashvili, Iván Villar, Jordi Masip, Paulo Gazzaniga, Pepe Reina, Rui Silva, Sergio Herrera, Stole Dimitrievski y Unai Simón.
- **Cluster 2:** Aitor Fernández, Álex Remiro, Claudio Bravo, David Soria, Ivo Grbić, Jan Oblak, Jeremías Ledesma, Marko Dmitrović, Predrag Rajković y Thibaut Courtois.
- **Cluster 3:** Gerónimo Rulli y Marc-André ter Stegen.
- **Cluster 4:** Agustín Marchesín, Álvaro Fernández, Édgar Badía, Fernando Pacheco y Yassine Bounou.

Seguidamente se aplica ANOVA mediante la función aov (R Core Team (2022)) para cada una de las variables utilizadas para definir los agrupamientos.

`% de salvadas` (porcentaje de salvadas)

La hipótesis nula (H0) es que no hay efecto significativo del factor cluster en la variable % Salvadas. El valor p extremadamente bajo (0.000000259) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se demuestra que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable porcentaje de salvadas.

`PaC%` (porcentaje de partidos con portería a 0)

La hipótesis nula (H0) es que no hay efecto significativo del factor cluster en la variable % Salvadas. El valor p extremadamente bajo (0.000000541) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se demuestra que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable porcentaje de porterías a cero.

`GC90` (goles en contra por 90 minutos)

La hipótesis nula (H0) es que no hay efecto significativo del factor cluster en la GC90. El valor p extremadamente bajo (0.0000000206) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable goles en contra por 90 min.

Dado que se han detectado diferencias significativas en todas las variables tal y como era de esperar, se aplica el test de Scheffé (Hinton, P. R. (1995)) mediante la función scheffe.test (Felipe de Mendiburu and Muhammad Yaseen(2020)) con tal de explorar estadísticamente estas diferencias detectadas entre los grupos identificados.

`% de salvadas` (porcentaje de salvadas)

El análisis de Scheffé ha agrupado las medias en dos grupos significativamente diferentes. Por un lado tenemos los cluster 2 y 3 que conforman el grupo "a" y los clusters 1 y 4 que conforman el grupo "b".

El método de las comparaciones múltiples scheffe.test muestra que los porteros del cluster 2 y el cluster 3 tienen un porcentaje de paradas significativamente superior a los porteros que conforman los clusters 1 y 4.

`PaC%` (porcentaje de partidos con portería a 0)

El análisis de Scheffé ha agrupado las medias en tres grupos significativamente diferentes. Por un lado tenemos el cluster 3 que conforma el grupo "a", el cluster 2 que conforma el grupo "b" y los clusters 1 y 4 que conforman el grupo "c".

El método de las comparaciones múltiples scheffe.test muestra que los porteros del cluster 3 tienen en media un porcentaje de porterías a cero significativamente superior a los porteros que conforman los cluster 1 y 4 y también sobre el cluster 2 el cual quedaría en un punto intermedio.

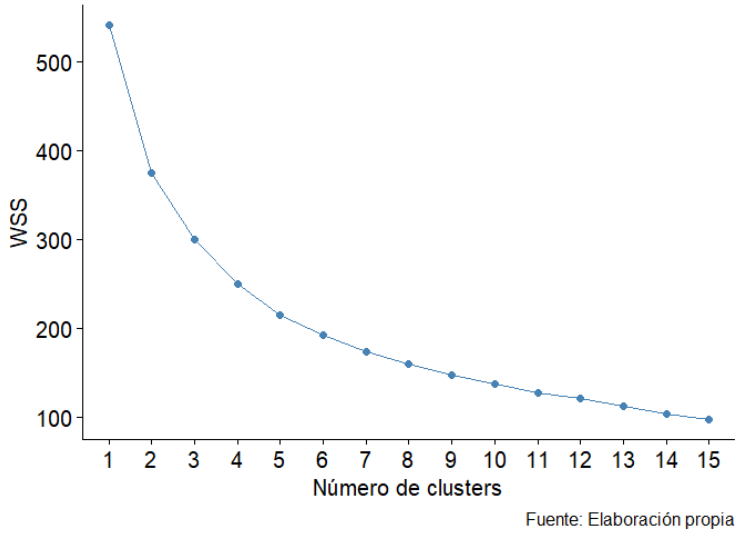
`GC90` (goles en contra por 90 minutos)

El análisis de Scheffé ha agrupado las medias en cuatro grupos significativamente diferentes. Por un lado tenemos el cluster 4 que conforma el grupo "a", luego el cluster 1 que conforma el grupo "b", el cluster 2 que conforma el grupo "c" y finalmente el cluster 3 que conforma el grupo "d"

El método de las comparaciones múltiples scheffe.test muestra que los porteros del cluster 4 tienen en media el mayor número de goles en contra por cada 90 minutos jugados. La diferencia en la media es significativamente superior a los demás grupos, siendo los porteros del grupo 3 los que en media menos goles en contra tienen por cada 90 minutos jugados.

4.3.2 Defensas

Figura 3: Número óptimo de clústeres Defensas

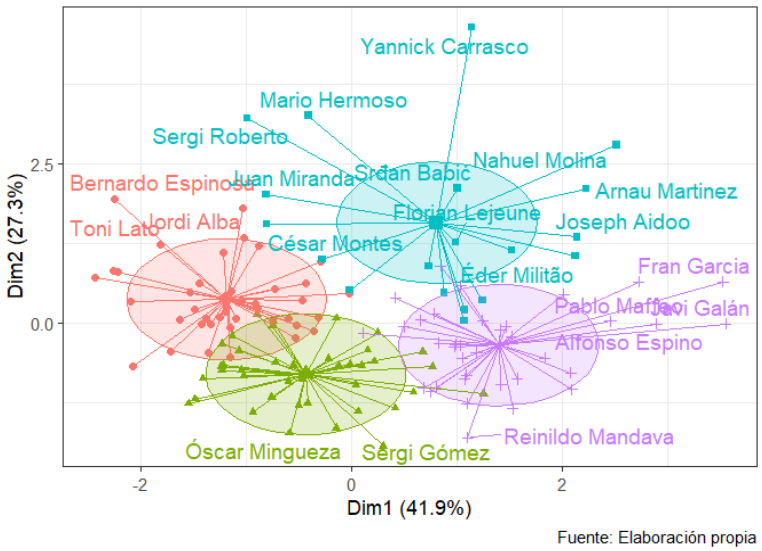


Fuente: Elaboración propia

La Figura 3 proporciona una guía visual para determinar el número óptimo de clusters en el análisis de K-means. En este caso, el "codo" en 4 agrupaciones sugiere que ese número puede ser una elección apropiada para equilibrar la cohesión dentro de los clusters y la simplicidad del modelo.

Seguidamente se aplica la técnica K-means para el número de agrupamientos definido y se obtiene los siguientes resultados:

Figura 4: Resultados clustering K-means Defensas



Fuente: Elaboración propia

De acuerdo con la Figura 4, los cuatro agrupamientos construidos por el método K-means son los siguientes:

- **Cluster 1:** Aïssa Mandi, Aritz Elustondo, Dani Carvajal, Éray Cömert, Eric García, Ferland Mendy, Jesús Navas, Jordi Alba, Nacho, Nacho Vidal, Omar Alderete, Rubén Peña, Toni Lato, Víctor Chust y Yan Couto entre otros.
- **Cluster 2**: Alfonso Pedraza, Andreas Christensen, Damián Suárez, David Alaba, Fali, Gabriel Paulista, Iñigo Lekue, Iñigo Martínez, Jules Koundé, Raúl Albiol, Robin Le Normand, Stefan Savić, Unai García, Yeray Álvarez, Ronald Araújo entre otros.
- **Cluster 3:** Éder Militão, Hugo Mallo, Marcos Acuña, Marcos Alonso, Mario Hermoso, Miguel Gutiérrez, Mouctar Diakhaby, Nemanja Gudelj, Nacho, Nahuel Molina, Pedro Bigas, Sergi Roberto y Yannick Carrasco entre otros.
- **Cluster 4:** Alejandro Balde, Alfonso Espino, Axel Witsel, Jaume Costa, Juan Foyth, Lucas Vázquez, Óscar de Marcos, Pablo Maffeo, Pau Torres,Ronald Araújo, Santiago Bueno, Unai Núñez, Thierry Correia y Yuri Berchiche.

Seguidamente se aplica ANOVA mediante la función aov (R Core Team, 2022) para cada una de las variables utilizadas para definir los agrupamientos.

`Gls.` (goles)

La hipótesis nula (H0) es que no hay efecto significativo del factor cluster en la variable goles. El valor p extremadamente bajo (<0.0000000000000002) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se concluye que hay al menos

una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable goles.

`Bloqueos` (El número de veces que se bloquea el balón poniéndose en su camino)

La hipótesis nula (H0) es que no hay efecto significativo del factor cluster en la variable bloqueos. El valor p extremadamente bajo (<0.0000000000000002) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable bloqueos.

`Tkl %` (Porcentaje de regateadores tacleados)

La hipótesis nula (H_0) es que no hay efecto significativo del factor cluster en la variable `Tkl%`. El valor p extremadamente bajo (0.00000000000000143) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable tacleados.

Dado que se han detectado diferencias significativas en todas las variables tal y como era de esperar, se aplica el test de Scheffé (Hinton, P. R., 1995) mediante la función scheffe.test (Felipe de Mendiburu and Muhammad Yaseen, 2020) con tal de explorar estadísticamente estas diferencias detectadas entre los grupos identificados.

`Gls.` (goles)

El análisis de Scheffé ha agrupado las medias en dos grupos significativamente diferentes. Por un lado tenemos el cluster 3 que conforma el grupo "a", y por otro lado los clusters 1,2 y 4 que conforman el grupo "b".

El método de las comparaciones múltiples scheffe.test muestra que los defensas del cluster 3 tienen en media un números de goles significativamente superior respecto a los defensas que conforman los cluster 1, 2 y 4.

`Bloqueos` (El número de veces que se bloquea el balón poniéndose en su camino)

El análisis de Scheffé ha agrupado las medias en cuatro grupos significativamente diferentes. Por un lado tenemos el cluster 4 que conforma el grupo "a",luego el cluster 3 que conforma el grupo "b", el cluster 2 que conforma el grupo "bc" y finalmente el cluster 1 que conforma el cluster "c".

El método de las comparaciones múltiples scheffe.test muestra que los defensas del cluster 4 tienen en media un número de bloqueos significativamente superior respecto a los defensas que conforman los grupos 1, 2 y 3, respectivamente, en ese orden.

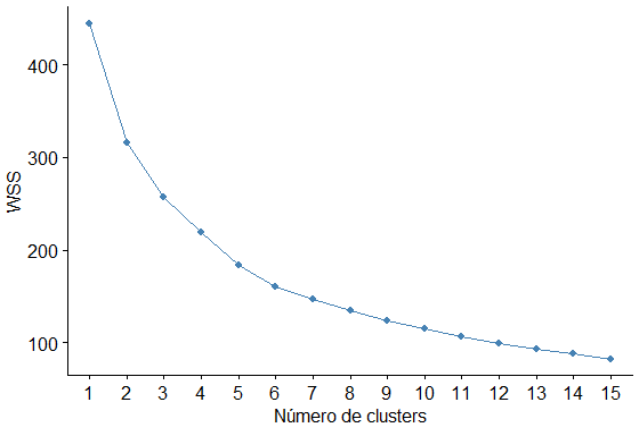
`Tkl %` (Porcentaje de regateadores tacleados)

El análisis de Scheffé ha agrupado las medias en cuatro grupos significativamente diferentes. Por un lado tenemos el cluster 2 que conforma el grupo "a", luego el cluster 4 que conforma el grupo "b", el cluster 3 que conforma el grupo "bc" y finalmente el cluster 1 que conforma el cluster "c".

El método de las comparaciones múltiples scheffe.test muestra que los defensas del cluster 2 tienen en media un número de "tacleos" a regateadores significativamente superior respecto a los defensas que conforman los clusters 1, 3 y 4, respectivamente, en ese orden.

4.3.3 Mediocentro

Figura 5: Número óptimo de clústeres Mediocentros

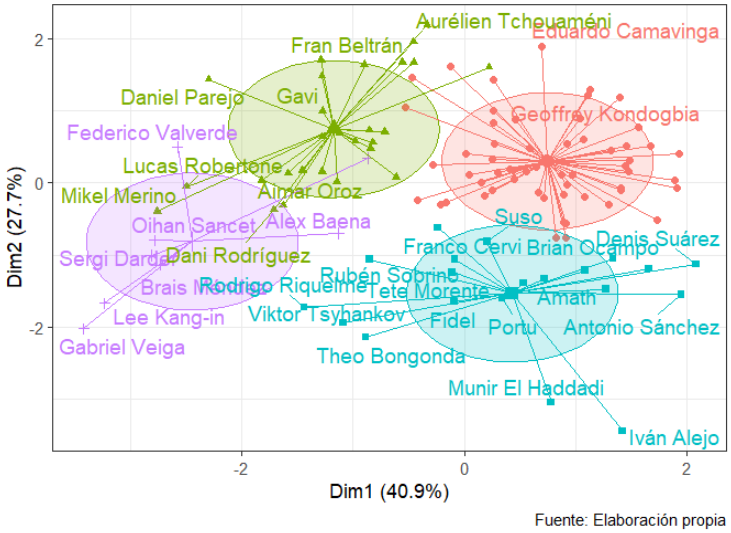


Fuente: Elaboración propia

La Figura 5 proporciona una guía visual para determinar el número óptimo de clusters en el análisis de K-means. En este caso, el "codo" en 4 agrupaciones sugiere que ese número puede ser una elección apropiada para equilibrar la cohesión dentro de los clusters y la simplicidad del modelo.

Seguidamente se aplica la técnica K-means para el número de agrupamientos definido y se obtiene los siguientes resultados:

Figura 6: Resultados clustering K-means Mediocentros



Fuente: Elaboración propia

Según la Figura 6, los cuatro agrupamientos construidos por el método K-means son los siguientes:

- **Cluster 1:** Asier Illarramendi, Eduardo Camavinga, Franck Kessié, Geoffrey Kondogbia, Giovani Lo Celso, Guido Rodríguez, Iker Muniain, Ivan Rakitić, Joan Jordán, Nabil Fekir, Nemanja Maksimović, Renato Tapia, Roque Mesa, Rubén Alcaraz, Samu Costa, Saúl Ñíguez, Thomas Lemar, Unai López, William Carvalho, Yangel Herrera y Yunus Musah entre otros.
- **Cluster 2:** Aleix García, Aurélien Tchouaméni, Dani Ceballos, Dani Rodríguez, Daniel Parejo, David Silva, Fran Beltrán, Frenkie de Jong, Gavi, Koke, Luka Modrić, Martín Zubimendi, Mikel Merino, Mikel Vesga, Rodrigo De Paul, Santi Comesaña, Sergio Busquets y Toni Kroos entre otros.
- **Cluster 3:** Aleix Vidal, Carles Aleñá, Denis Suárez, Edu Expósito, Gonzalo Escalante, Iván Alejo, Munir El Haddadi, Pablo Ibáñez, Portu, Rodrigo Riquelme, Rubén Sobrino y Suso entre otros.
- **Cluster 4:** Alex Baena, Brais Méndez, Federico Valverde, Gabriel Veiga, Lee Kang-in, Oihan Sancet, Pedri y Sergi Darder entre otros.

Seguidamente se aplica ANOVA mediante la función aov (R Core Team, 2022) para cada una de las variables utilizadas para definir los agrupamientos.

`Gls.` (goles)

La hipótesis nula (H_0) es que no hay efecto significativo del factor cluster en la variable `Goles`. El valor p extremadamente bajo (<0.0000000000000002) indica que hay evidencia significativa para rechazar la hipótesis nula.

Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable goles.

`por_pases_completados`

La hipótesis nula (H_0) es que no hay efecto significativo del factor cluster en la variable porcentaje de pases completados. El valor p extremadamente bajo (<0.0000000000000002) indica que hay evidencia significativa para rechazar la hipótesis nula.

Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable porcentaje de pases completados.

`Ass` (Asistencias)

La hipótesis nula (H_0) es que no hay efecto significativo del factor cluster en la variable asistencias. El valor p extremadamente bajo (0.00000000000000322) indica que hay evidencia significativa para rechazar la hipótesis nula.

Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable asistencias.

Dado que se han detectado diferencias significativas en todas las variables tal y como era de esperar, se aplica el test de Scheffé (Hinton, P. R., 1995) mediante la función scheffe.test (Felipe de Mendiburu and Muhammad Yaseen, 2020) con tal de explorar estadísticamente estas diferencias detectadas entre los grupos identificados.

`Gls.` (goles)

El análisis de Scheffé ha agrupado las medias en cuatro grupos significativamente diferentes. Por un lado tenemos el cluster 4 que conforma el grupo "a",luego el cluster 2 que conforma el grupo "b", el cluster 3 que conforma el grupo "bc" y finalmente el cluster 1 que conforma el cluster "c".

El método de las comparaciones múltiples scheffe.test muestra que los mediocentros del cluster 4 marcan en media un número de goles significativamente superior respecto a los mediocentros que conforman los clusters 1, 3 y 2, respectivamente, en ese orden.

`por_pases_completados`

El análisis de Scheffé ha agrupado las medias en cuatro grupos significativamente diferentes. Por un lado tenemos el cluster 2 que conforma el grupo "a",luego el cluster 1 que conforma el grupo "ab", el cluster 4 que conforma el grupo "b" y finalmente el cluster 3 que conforma el cluster "c".

El método de las comparaciones múltiples scheffe.test muestra que los mediocentros del cluster 2 completan en media un porcentaje de pases significativamente superior respecto a los mediocentros que conforman los clusters 3, 4 y 1 respectivamente, en ese orden.

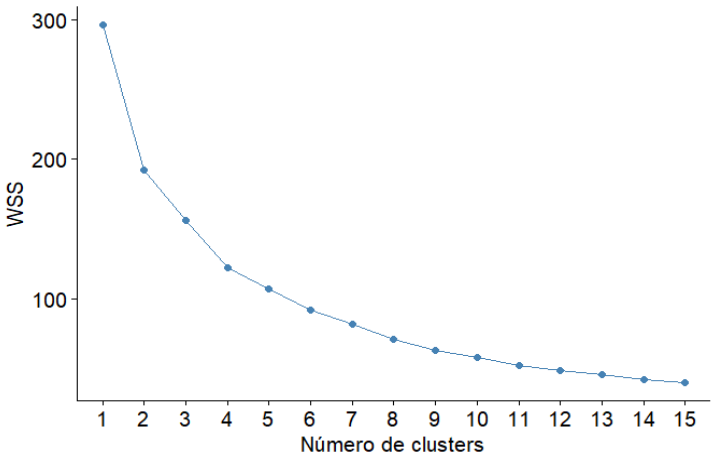
`Ass` (Asistencias)

El análisis de Scheffé ha agrupado las medias en cuatro grupos significativamente diferentes. Por un lado tenemos el cluster 2 que conforma el grupo "a", luego el cluster 4 que conforma el grupo "ab", el cluster 3 que conforma el grupo "bc" y finalmente el cluster 1 que conforma el cluster "c".

El método de las comparaciones múltiples scheffe.test muestra que los mediocentros del cluster 2 asisten en media un número de goles significativamente superior respecto a los mediocentros que conforman los clusters 1,3 y 4 respectivamente, en ese orden.

4.3.4 Delantera

Figura 7: Número óptimo de clústeres Delantera

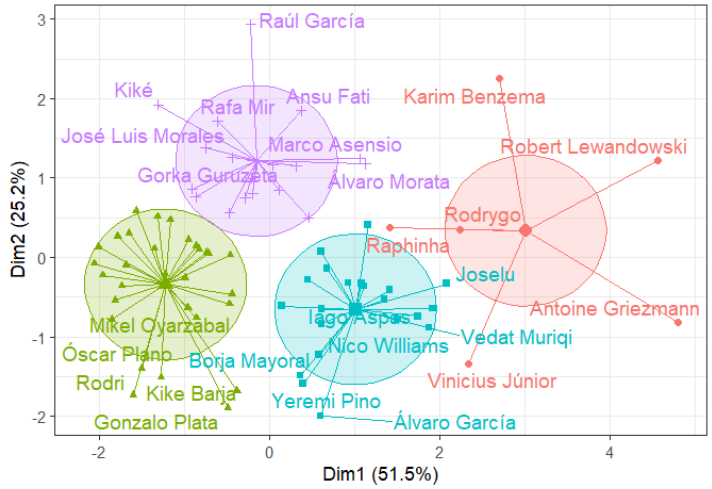


Fuente: Elaboración propia

La Figura 7 proporciona una guía visual para determinar el número óptimo de clusters en el análisis de K-means. En este caso, el "codo" en 4 agrupaciones sugiere que ese número puede ser una elección apropiada para equilibrar la cohesión dentro de los clusters y la simplicidad del modelo.

Seguidamente se aplica la técnica K-means para el número de agrupamientos definido y se obtiene los siguientes resultados:

Figura 8: Resultados clustering K-means Delantera



Fuente: Elaboración propia

De acuerdo con la Figura 8, los cuatro agrupamientos construidos por el método K-means son los siguientes:

- **Cluster 1 :** Alexander Sørloth, Borja Iglesias, Borja Mayoral, Enes Ünal, Iago Aspas, Iñaki Williams, Isaac Palazón, Joselu, Lucas Boyé, Martin Braithwaite, Nico Williams, Samuel Chukwueze, Samuel Lino, Vedat Muriqi y Yeremi Pino entre otros.
- **Cluster 2 :** Antoine Griezmann, Karim Benzema, Raphinha, Robert Lewandowski, Rodrygo y Vinicius Júnior.

- **Cluster 3 :** Álvaro Morata, Ángel Correa, Ansu Fati, Ante Budimir, Cristhian Stuani, Gerard Moreno, José Luis Morales, Justin Kluivert, Marco Asensio, Ousmane Dembélé, Rafa Mir, Raúl García y Youssef En-Nesyri entre otros.
- **Cluster 4 :** Bryan Gil, Edinson Cavani, Ferrán Torres, Hugo Duro, Juanmi, Lucas Ocampos, Mikel Oyarzabal, Papu Gómez, Pere Milla, Samu Castillejo, Sergio Canales y Sergio León entre otros.

Seguidamente se aplica ANOVA mediante la función aov (R Core Team, 2022) para cada una de las variables utilizadas para definir los agrupamientos.

`Gls.` (goles)

El valor p extremadamente bajo (0.00000000000277) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable goles.

`T/90` (tiros al arco cada 90 minutos)

La hipótesis nula (H_0) es que no hay efecto significativo del factor cluster en la variable `T/90` . El valor p extremadamente bajo (0.00000000000000268) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable disparos por cada 90 min.

`Ass` (Asistencias)

La hipótesis nula (H_0) es que no hay efecto significativo del factor cluster en la variable `Ass`. El valor p extremadamente bajo (0.00000000158) indica que hay evidencia significativa para rechazar la hipótesis nula. Por lo tanto, se concluye que hay al menos una diferencia significativa entre los grupos definidos por la clusterización en términos de la variable asistencias.

Dado que se han detectado diferencias significativas en todas las variables tal y como era de esperar, se aplica el test de Scheffé (Hinton, P. R., 1995) mediante la función scheffe.test (Felipe de Mendiburu and Muhammad Yaseen, 2020) con tal de explorar estadísticamente estas diferencias detectadas entre los grupos identificados.

`Gls.` (goles)

El análisis de Scheffé ha agrupado las medias en tres grupos significativamente diferentes. Por un lado, tenemos el cluster 2 que conforma el grupo "a", luego el cluster 1 y 3 que conforma el grupo "b", y finalmente el cluster 4 que conforma el grupo "c".

El método de las comparaciones múltiples scheffe.test muestra que los delanteros del cluster 2 marcan en media un número de goles significativamente superior respecto a los delanteros que conforman los clusters 4, 3 y 1 respectivamente, en ese orden.

`T/90` (tiros al arco cada 90 minutos)

El análisis de Scheffé ha agrupado las medias en tres grupos significativamente diferentes. Por un lado tenemos el cluster 2 y 3 que conforman el grupo "a",luego el cluster 1 que conforma el grupo "b", y finalmente el cluster 4 que conforma el grupo "c".

El método de las comparaciones múltiples scheffe.test muestra que los delanteros del cluster 2 disparan en media un número de veces significativamente superior respecto a los delanteros que conforman los clusters 4, 1 y 3 respectivamente, en ese orden.

`Ass` (Asistencias)

El análisis de Scheffé ha agrupado las medias en dos grupos significativamente diferentes. Por un lado, tenemos el cluster 2 conforma el grupo "a", y por otro lado los cluster 1, 3 y 4 que conforman el grupo "b".

El método de las comparaciones múltiples scheffe.test muestra que los delanteros del cluster 2 asisten en media un número de goles significativamente superior respecto a los delanteros que conforman los clusters 4, 3 y 1 respectivamente, en ese orden.

5. Conclusiones

En relación a la primera hipótesis de partida, a la hora de la consecución de goles existen nacionalidades "menos convencionales" significativas como es el caso de Canadá, Japón, Kosovo, Noruega, Polonia y Turquía.

Como se planteaba al inicio del estudio la posición de delantero es significativa para explicar la cantidad de goles que marcan los jugadores. Sin embargo, se esperaba que la posición de defensa fuera significativa, pero con un coeficiente negativo, pero no lo ha sido.

Como se ha comentado en el apartado de Regresiones y como se ha reflejado en la Tabla número 7 los equipos que mejor % de pases exitosos tienen son el FC Barcelona y el Real Madrid.

Según los resultados anteriormente expuestos se puede concluir que existen diferencias estadísticamente significativas entre agrupaciones de jugadores según su posición en el campo y atendiendo a variables básicas como número de goles, asistencias, pases completados… Atendiendo a las hipótesis inicialmente planteadas, podemos afirmar que:

Respecto a la posición de portería, es precisamente el clúster número 3 formado por los jugadores Gerónimo Rulli y Marc-André ter Stegen aquel que presenta el mejor desempeño general de toda la liga de fútbol español; es decir, presentan un % de salvadas mayor y han conseguido finalizar los partidos con un menor número de goles en la portería que defienden. Asimismo, presentan una proporción considerable de partidos con portería a 0.

Por lo que respecta a los defensas, no se observa la predominancia de un mismo cluster sobre las tres estadísticas analizadas (goles, bloqueos y porcentaje de regateadores tacleados) tal y como se planteaba en la hipótesis. En efecto, para la variable goles los defensas del tercer cluster el cual está formado por jugadores como Éder Militão, Mario Hermoso, Sergi Roberto y Yannick Carrasco entre otros presenta un número mayor de goles; mientras que para las variables defensivas de el número de bloqueos y el porcentaje de regateadores tacleados no se encuentra el grupo dominador que en principio se suponía.

En relación a la posición de centrocampista, se ha encontrado un grupo de mediocentros (cluster 4) con un número de goles significativamente superior al del resto correspondiente a los mediocentros más ofensivos entre los que destacamos a Federico Valverde, Lee Kang-in, Pedri y Sergi Darder entre otros. Por otra parte, también se ha encontrado el grupo (cluster 2) con un porcentaje de pases completados y asistencias significativamente superior al del resto que se planteaba, correspondiente a los mediocentros con mejor visión de juego entre los que destacamos a Aurélien Tchouaméni, Dani Ceballos, Dani Parejo, David Silva, Toni Kroos, Sergi Busquets, Gavi y Mikel Merino entre otros.

Por su parte, tal y como se esperaba, se ha encontrado un grupo de delanteros (cluster 2) con un número de goles y un número de tiros a portería por cada 90 minutos significativamente superior al del resto de grupos, correspondiendo a los delanteros que más disparan y aciertan entre los que figuran las grandes estrellas de la Liga 2022-2023 Antoine Griezmann, Karim Benzema, Raphinha, Robert Lewandowski, Rodrygo y Vinicius Júnior. Sin embargo, se esperaba encontrar un grupo de delanteros diferente con un número de asistencias significativamente superior, correspondiendo a los delanteros que destacan más por hacer asistencias de gol. No obstante, se ha encontrado que los jugadores anteriormente mencionados también han sido los más destacados en este aspecto.

5.1 Futuras líneas de investigación

Como futuras líneas de investigación, se proponen diversas vertientes.

En primer lugar, comparar la variación de las estadísticas estudiadas (2022-23) con las de temporadas pasadas o temporadas futuras, con el objetivo de ver cómo varían los datos de determinados jugadores, equipos y la liga en general.

En segundo lugar, se podría realizar el mismo estudio sobre las denominadas "5 grandes ligas de Europa", Bundesliga (Alemania), Premier League (Inglaterra), Serie A (Italia), Ligue 1 (Francia) y la LaLiga (España), para poder compararlas entre sí y poder determinar qué liga es la que mejores jugadores reúne.

Por otro lado, también se podrían obtener los valores de mercado de los jugadores mediante plataformas como *transfermarkt.com* y realizar modelos que expliquen y nos permitan predecir valores de los jugadores gracias a las estadísticas dadas.

6. Referencias bibliográficas

Akhanli, Serhat Emre and Hennig, Christian. "Clustering of football players based on performance data and aggregated clustering validity indexes" Journal of Quantitative Analysis in Sports, vol. 19, no. 2, 2023, pp. 103-123. https://doi.org/10.1515/jqas-2022-0037 Consultado el 13 de diciembre de 2023

Cayuela, L. (2009). Modelos lineales generalizados (GLM). Materiales de un curso del R del IREC. Consultado el 2 de diciembre de 2023

Claude, J. (2008). Morphometrics with R. Países Bajos: Springer New York. https://www.researchgate.net/publication/258885152_Morphometrics_With_R Consultado el 13 de diciembre de 2023

Felipe de Mendiburu and Muhammad Yaseen (2020). agricolae: Statistical Procedures for Agricultural Research. R package version 1.4.0, Consultado el 15 de diciembre de 2023 https://myaseen208.github.io/agricolae/https://cran.r-project.org/package=agricolae.

Haghighat, Maral & Rastegari, Hamid. (2013). A Review of Data Mining Techniques for Result Prediction in Sports. Advances in Computer Science: an International Journal. 2. Consultado el 13 de diciembre de 2023

Hinton, P. R. (1995). Statistics explained: a guide for social science students. Reino Unido: Routledge. Consultado el 15 de diciembre de 2023 http://196.188.170.250:8080/jspui/bitstream/123456789/4155/1/Statistics_Explained_A_Guide...2004.pdf

Kassambara, A. and Mundt, F. (2020) Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_nbclust

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

S., J. D. (2014). ANÁLISIS DE VARIANZA. Revista Chilena de Anestesia. Consultado el 6 de diciembre de 2023, en https://revistachilenadeanestesia.cl/PII/revchilanestv43n04.07.pdf

Saavedra, José Ángel (2023). Regresión Lineal: teoría y ejemplos. Consultado el 6 de diciembre de 2023. Disponible en: https://ebac.mx/blog/regreson-lineal

Wickham, H. (2019). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.0. Disponible en: https://CRAN.R-project.org/package=rvest

7. Anexo

Código: https://github.com/marioguillem/DatosNoEstructurados