Universidad Rey Juan Carlos

MUSA

DETECCIÓN DEL ROL DE LOS JUGADORES MEDIANTE ALGORITMOS DE CLUSTERING

PROYECTO FINAL

ASIGNATURA: INTRODUCCIÓN AL SPORT ANALYTICS

PROFESOR: JAVIER MARTÍN BULDÚ

FECHA: 18/12/2025

ALUMNOS (@FutbolMUSA):

- DANIEL GRANIZO
- SARA MAS
- ALEJANDRO PASCUAL

ÍNDICE

ÍNDICE DE FIGURAS Y TABLAS

1. RESUMEN

Hoy en día, la presencia e influencia cada vez más grande del sports analytics en el deporte profesional ha transformado la forma de entender el rendimiento de los jugadores, en este caso de jugadores de fútbol. Sin embargo, la asignación de posiciones y roles de los jugadores sigue estando muy ligada a la observación subjetiva del cuerpo técnico. Con este proyecto, se propone una metodología objetiva para detectar el rol real de los futbolistas a partir de sus acciones en el campo, utilizando datos de eventos de StatsBomb de LaLiga EA Sports de la temporada 24/25

En una primera fase del proyecto se construye una base de datos unificada que integra los ficheros JSON de datos filtrando aquellos jugadores que hayan disputado menos de 1000 minutos en la competición, obteniendo de estos ficheros todos los datos de eventos de los partidos de la temporada excluyendo la información 360º. A continuación, se divide el terreno de juego en 20 sectores (4x5) y se asigna a cada evento el sector correspondiente según sus coordenadas, pudiendo así caracterizar especialmente el comportamiento de cada jugador.

A partir de las posiciones más genéricas de jugadores (portero, defensa, centrocampista y delantero) se hacen dos subdivisiones más para poder realizar el algoritmo de clustering correctamente, detectar el rol exacto para cada jugador y así conseguir un ranking e incluso poder hacer comparaciones entre jugadores.

En cuanto a los resultados esperados podemos incluir una nueva clasificación funcional de los jugadores basada en su comportamiento real en el campo y no solo en su posición teórica y por otra parte un ranking por posición construido a través de métricas ponderadas de rendimiento. Estos resultados pueden ayudar a apoyar procesos de scouting, toma de decisiones tácticas y evaluación objetiva del rendimiento individual.

2. PLANTEAMIENTO DEL PROBLEMA

2.1 ¿Qué se quiere analizar?

En este proyecto se pretende identificar el rol funcional que desempeña cada futbolista en este caso de la liga española en base a los eventos que se generan durante un partido. Aunque los sistemas de juego (4-3-3, 4-2-3-1…) suelen hablar de jugadores en posiciones fijas, en la práctica muchos futbolistas se comportan con perfiles híbridos (lateral ofensivo, pivote destructor, interior con llegada, extremo pasador…). El objetivo es cuantificar ese comportamiento mediante variables de juego (pases, tiros, recuperaciones, duelos, conducciones, presiones…y utilizar técnicas de reducción de dimensión y clustering para agrupar jugadores con patrones de acción similares, definiendo así roles específicos dentro de cada posición.

2.2 Relevancia deportiva

Desde el punto de vista del rendimiento deportivo, conocer el rol real de cada jugador permite al cuerpo técnico evaluar si sus acciones se alinean con lo propuesto en el modelo de juego del equipo. A este nivel táctico, los clusters de roles pueden ayudar a construir plantillas equilibradas, como por ejemplo lograr combinaciones de centrales o de delanteros más adecuados a la idea del entrenador. Por otra parte, para el scouting, disponer de un ranking según los diferentes roles facilita la búsqueda de jugadores con esas características en el mercado reduciendo la dependencia de la observación subjetiva.

2.3 Hipótesis de trabajo

- H1: Los jugadores que comparten posición nominal exhiben patrones similares en tipos de eventos y zonas del campo, lo cual permitirá identificar roles mediante clustering.
- H2: La combinación de sector principal más sector vecino mejora la representación espacial del rendimiento.
- H3: PC1 y PC2 reducirán adecuadamente la dimensionalidad manteniendo la estructura táctica de los datos.
- H4: Para comparar jugadores, es indispensable que todos tengan exactamente los mismos 10 tipos de eventos por posición; los eventos no realizados deben introducirse manualmente como 0.
- H5: No es necesario dividir las acciones por 90 minutos, dado que el filtro de más de 1000 minutos proporciona estabilidad estadística suficiente.

3. DATOS UTILIZADOS

3.1 Fuente de los datos

Los datos proceden de la base de datos de StatsBomb para LaLiga EA Sports de la temporada 24/25, utilizando para ello los distintos ficheros JSON que contienen la información global de jugadores y los eventos de los diferentes partidos de toda la temporada.

3.2 Tipo de datos

- Datos de evento: cada acción registrada durante el partido (pases, tiros, recuperaciones, duelos, conducciones…) con información sobre el tipo del evento, jugador, equipo, minuto de juego, resultado de la acción y coordenadas (x, y) del punto del terreno de juego donde se produce
- Datos de contexto de jugadores: posición, equipo, minutos disputados a lo largo de la temporada y demás metadatos necesarios para poder filtrar y agruparlos.

Cabe destacar que no se utilizan datos de tracking ni datos biométricos, el análisis se basa exclusivamente en datos de eventos (sin tener en cuenta el componente 360º)

3.3 Período de observación y cobertura

El periodo de estudio abarca toda la temporada 2024/2025 de LaLiga EA Sports. Se incluyen todos los partidos de liga con datos disponibles en StatsBomb. Para garantizar estabilidad estadística, solo se consideran los jugadores que acumulan al menos 1.000 minutos de juego durante la temporada.

3.4 Limpieza y preparación de los datos.

Clasificación y organización de las posiciones. A partir de las posiciones más genéricas de jugadores (portero, defensa, centrocampista y delantero) se hacen dos subdivisiones más para poder realizar el algoritmo de clustering correctamente.

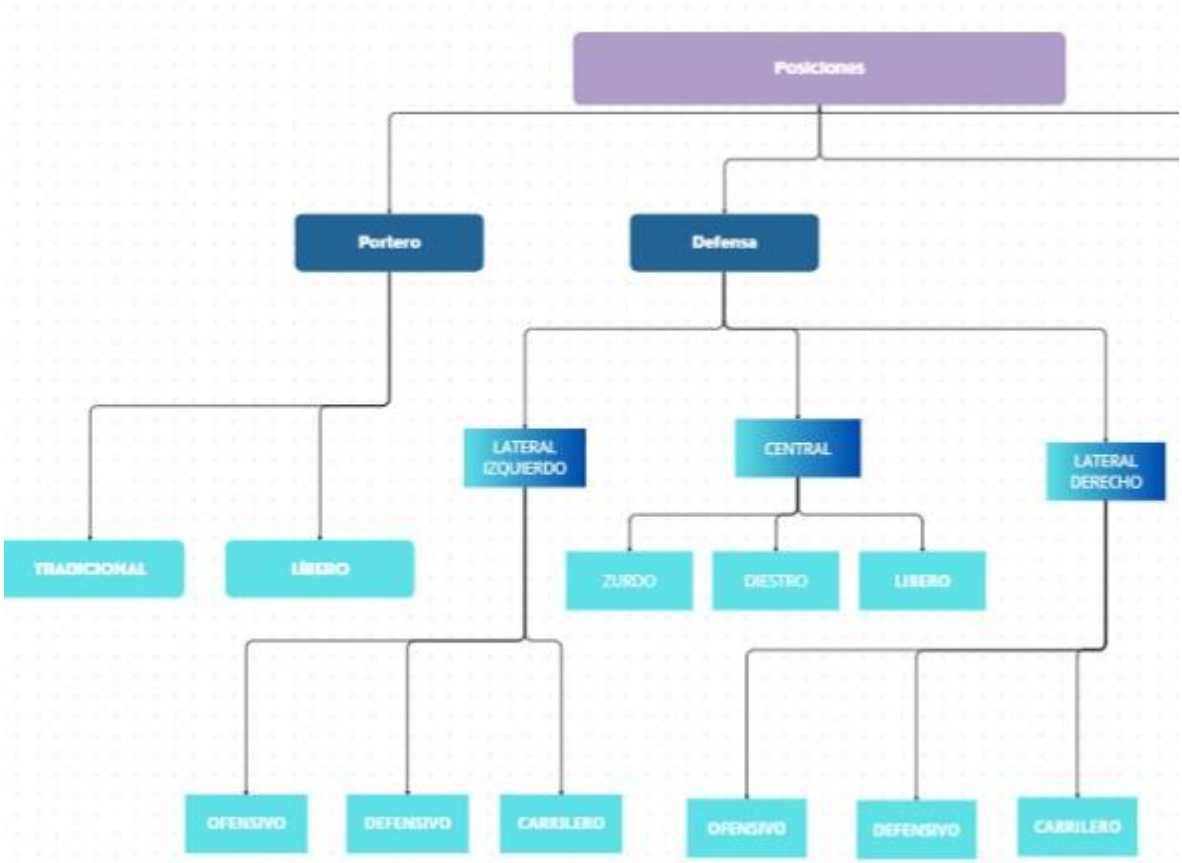


Figura 1: clasificación y organización de las posiciones. Porteros y defensas.

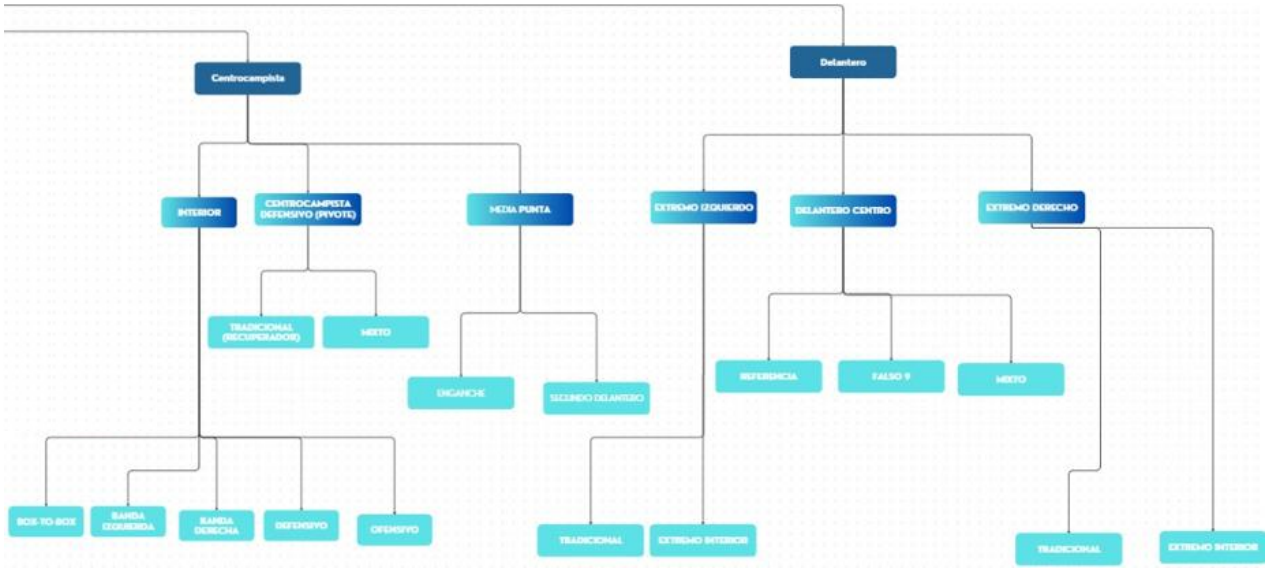


Figura 2: clasificación y organización de las posiciones. Mediocentros y atacantes.

Filtrado de jugadores por minutos disputados. A partir del fichero global de jugadores se calcula el total de minutos jugados y se eliminan aquellos con menos de 1.000 minutos, reduciendo ruido en el análisis.

Unificación de posiciones. Muchas etiquetas de posición se agrupan en categorías más generales (p. ej., “Left Center Back” y “Right Center Back” → “Center Back”), lo que simplifica el análisis y evita clases demasiado específicas con pocos casos.

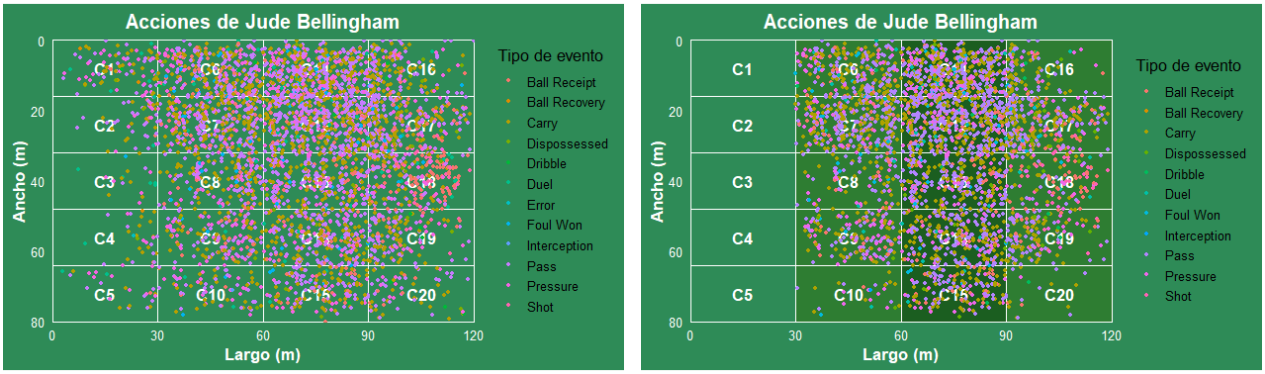
Sectorización del campo. Se definen dimensiones estándar del terreno de juego (120×80 m) y se divide en 20 sectores (4 filas × 5 columnas). Mediante una función se asigna a cada evento el sector correspondiente según sus coordenadas, generando una base de datos donde cada acción está geolocalizada en un sector concreto.

Figura 3: sectorización del campo

Asignación de eventos relevantes por posición: Para cada posición se han definido 10 eventos considerados representativos del comportamiento táctico característico del rol. Estos eventos provienen tanto de la literatura táctica como de la observación de patrones empíricos en los datos.

Para evitar que jugadores con muy poca participación distorsionen el análisis, se estableció un filtro mínimo de 70 acciones relevantes por jugador en su posición. Este umbral asegura que los datos reflejen un comportamiento estable y representativo del estilo del jugador. Sin este filtro, jugadores con pocos eventos podrían sesgar el PCA, los percentiles y el clustering.

Asignación de sectores principales y vecinos: Cada posición ocupa unos sectores principales que reflejan su espacio natural de actuación. Por otro lado, los sectores vecinos son aquellos adyacentes (arriba, abajo, izquierda o derecha) al sector principal de un jugador. Introducir sectores vecinos permite capturar acciones que, aunque no se producen en la zona habitual del jugador, siguen siendo parte de su comportamiento real.



Figuras 4 y 5: Asignación sectores principales y vecinos

Asignación de pesos a los eventos: A cada tipo de evento se le asigna un peso táctico denominado Valor Ajustado, que puede ser positivo (evento beneficioso) o negativo (evento perjudicial). La puntuación final se calcula como:

- Evento en sector principal = Valor Ajustado × número de acciones.
- Evento en sector vecino = (Valor Ajustado × número de acciones) / 2.

Esta estructura permite que acciones negativas como pérdidas o errores penalicen adecuadamente al jugador. La puntuación de estos eventos es menor que la del sector principal, ya que se considera que la acción está menos alineada con el rol natural.

Las tres asignaciones se reflejan en estas tablas:

Portero → Sector C3		
Eventos	**Definición**	**Valor ajustado**
Goal Keeper	Paradas y acciones clave	30

Clearance	Despejes de peligro	10
Block	Bloqueos de tiros	8
Pass	Distribución con balón	5
Ball Recovery	Recuperaciones defensivas	7
Pressure	Presión recibida	4
Carry	Conducción con balón	3
Own Goal Against	Goles en propia	-15
Error	Fallos graves	-10
Dispossessed	Pérdidas de balón	-8

Central → Sectores C2, C3, C4		
Eventos	**Definición**	**Valor ajustado**
Clearance	Despejes defensivos	20
Block	Bloqueos de disparos	15
Interception	Cortes de pase	15
Duel	Duelos defensivos	12
Error	Fallos defensivos	8
Pass	Pérdidas en salida	6
Ball Recovery	Recuperaciones	-10
Dribbled Past	Superado en 1vs1	-8
Dispossessed	Presión recibida	-4
Foul Committed	Salida de balón	-2

Lateral → Sectores C1, C6 (izquierdo) y C5, C10 (derecho)		
Eventos	**Definición**	**Valor ajustado**
Pass	Centros y pases	18
Dribble	Regates por banda	15
Carry	Conducciones ofensivas	12
Interception	Cortes defensivos	12
Duel	Duelos en banda	10
Clearance	Despejes	10
Ball Recovery	Recuperaciones	8
Foul Won	Faltas recibidas	7
Foul Committed	Faltas cometidas	-3
Dispossessed	Pérdidas	-5

Pivote → Sectores C7, C8, C9		
Eventos	**Definición**	**Valor ajustado**
Interception	Cortes clave	20
Ball Recovery	Recuperaciones en medio	18
Pass	Distribución del juego	15

Pressure	Presión tras pérdida	12
Duel	Duelos defensivos	10
Carry	Progresión con balón	7
Shield	Protección del balón	5
Error	Fallos en base	-8
Dispossessed	Pérdidas peligrosas	-3
Foul Committed	Faltas tácticas	-2

Interior → Sectores C11, C12, C14, C15		
Eventos	**Definición**	**Valor ajustado**
Pass	Circulación ofensiva	20
Carry	Conducción hacia adelante	15
Pressure	Presión alta	12
Dribble	Desborde	12
Ball Recovery	Recuperación ofensiva	10
Foul Won	Faltas recibidas	8
Interception	Anticipación	8
Duel	Duelos en medio	7
Ball Receipt	Recepción en espacio	5
Dispossessed	Pérdidas	-3

Media punta → Sectores C12, C13, C14		
Eventos	**Definición**	**Valor ajustado**
Pass	Último pase	22
Shot	Finalización	18
Dribble	Desborde ofensivo	15
Ball Receipt	Recepción peligrosa	12
Carry	Conducción ofensiva	10
Foul Won	Faltas cerca del área	8
Pressure	Presión tras pérdida	6
Duel	Duelos ofensivos	3
Error	Fallos ofensivos	-4
Dispossessed	Pérdidas clave	-2

Extremo → Sectores C11, C16 (izquierdo) y C15, C20 (derecho)		
Eventos	**Definición**	**Valor ajustado**
Dribble	Desborde por banda	22
Carry	Progresión vertical	18
Pass	Centros y asistencias	15
Shot	Finalización desde banda	12
Ball Receipt	Recepción ofensiva	10

Foul Won	Faltas provocadas	8
Pressure	Presión alta	6
Duel	Duelos ofensivos	4
Interception	Recuperación tras pérdida	3
Dispossessed	Pérdidas ofensivas	-2

Delantero centro → Sectores C17, C18, C19		
Eventos	**Definición**	**Valor ajustado**
Shot	Finalización	22
Dribble	Regates completados	17
Ball Receipt	Recepción en área	15
Pass	Participación ofensiva	12
Carry	Conducción hacia portería	10
Foul Won	Faltas en zona gol	6
Error	Fallos en definición	4
Offside	Posicionamiento	-3
Duel	Duelos ofensivos	-3
Dispossessed	Pérdidas clave	-2

Tablas 1-8: Asignaciones de eventos, sectores y pesos por posiciones

Construcción de la base analítica. Finalmente, se agregan los eventos por jugador, posición y sector. Esta tabla agregada será la entrada para el PCA y los algoritmos de clustering en fases posteriores del proyecto.

Para que PCA, clustering y distancias euclídeas funcionen correctamente, todos los jugadores deben tener las mismas variables. Por ello, incluso si un jugador no ha realizado un evento, este se incorpora manualmente como valor **0**. Esto es especialmente importante para eventos negativos: que un jugador no haya cometido errores no debe interpretarse como ausencia de dato (NA), sino como valor real igual a cero.

4. METODOLOGÍA

4.1 Componentes Principales

El Análisis de Componentes Principales (PCA) es una técnica de reducción de la dimensionalidad que permite resumir un conjunto grande de variables en un número menor de componentes que capturan la mayor parte de la variabilidad original. Para ello, transforma las variables iniciales en nuevas variables no correlacionadas entre sí, llamadas componentes principales (PC), que se ordenan según la cantidad de variabilidad que explican: el primer componente recoge la mayor parte de la información, el segundo la mayor parte de la restante, y así sucesivamente.

Aplicar PCA por posición permite identificar qué combinaciones de variables son más relevantes para describir a los jugadores de cada rol, eliminando redundancias y facilitando la visualización de las diferencias entre ellos. Además, el PCA sirve como paso previo ideal para técnicas como el clustering, ya que ofrece un espacio reducido, menos ruidoso y más interpretable para agrupar jugadores según sus características.

Para aplicar esta técnica estadística resulta necesario pivotar la base de datos original. Con ello se obtiene un nuevo dataset en el que las variables son, además del nombre del jugador y su posición, todos los eventos disponibles. Las observaciones corresponden a las puntuaciones asociadas a cada evento, calculadas como el número de acciones multiplicado por su peso, el cual varía según si el sector es principal o vecino.

Posteriormente, se hace un PCA de cada una de las posiciones, utilizando únicamente los eventos relevantes asignados a cada posición. Se escogerán únicamente dos componentes principales (PC1 y PC2), pues al disponer únicamente de 10 eventos por posición se asume que dos componentes son suficientes para explicar un mínimo del 75% de la varianza de la posición y facilita la interpretación en gráficos en 2D.

Por último, a partir de los scores obtenidos, se calcularán las nuevas coordenadas de cada jugador según su posición, lo que permitirá representar de manera resumida y precisa sus características dentro del espacio definido por los componentes principales.

4.2 Clustering

Como se mencionó en el apartado 2, se utilizará la técnica del clustering para identificar patrones de juego dentro de una misma posición. Esta técnica de aprendizaje no supervisado permite agrupar observaciones similares entre sí sin necesidad de una variable objetivo. El algoritmo analiza las características de cada jugador y forma grupos (clusters) en los que los elementos del mismo grupo comparten rasgos parecidos, mientras que los de grupos distintos presentan diferencias más marcadas. Entre los métodos disponibles, se empleará k-means, que divide los datos en k clusters previamente establecidos. Este método asigna cada observación al centroide más cercano y recalcula estos centroides de forma iterativa hasta estabilizar la clasificación, generando así grupos compactos y fáciles de interpretar, lo que permite detectar distintos estilos o patrones de juego dentro de una misma posición.

Esta técnica se aplicará a cada posición utilizando los scores derivados de PC1 y PC2 de los componentes principales. De este modo, es posible generar un gráfico en 2D que facilite la visualización y permita observar con mayor claridad la separación de los jugadores en distintos grupos y poder asignar los roles de cada posición de forma más gráfica y sencilla.

Una vez calculados los clusters para cada posición, se incorporarán a la base de datos original indicando el cluster al que pertenece cada jugador en su respectiva posición en el campo. Esto permitirá elaborar un ranking de jugadores basado en sus puntuaciones, teniendo en cuenta tanto la posición como el rol (cluster) que desempeñan dentro de esa posición.

4.3 Percentiles

Los percentiles son una medida estadística que permite ubicar un valor dentro de una distribución en relación con el resto de los datos, indicando el porcentaje de observaciones que se encuentran por debajo de dicho valor. Son especialmente útiles para comparar posiciones relativas dentro de un grupo y para transformar distintos valores a una escala uniforme de 0 a 100.

Es otra forma de realizar un ranking de los jugadores, ya que permite identificar en qué percentil se encuentra cada jugador para cada uno de sus eventos en cada posición. Esto facilita comparar su rendimiento relativo dentro del grupo, destacando en qué aspectos sobresale y en cuáles podría mejorar, independientemente de la escala original de los indicadores.

Se puede aplicar el mismo enfoque, pero en lugar de comparar a un jugador con todos los que comparten su posición, se le compara únicamente con los jugadores que comparten posición y rol (cluster). De esta manera, el análisis es más preciso y contextualizado, ya que permite evaluar el rendimiento relativo dentro de grupos homogéneos de jugadores que desempeñan funciones similares, facilitando la identificación de fortalezas y áreas de mejora específicas para cada rol dentro de la posición.

4.4 Distancia euclídea

La distancia euclídea es una medida de disimilitud que permite cuantificar qué tan diferentes son dos jugadores en función de sus características. Se calcula como la raíz cuadrada de la suma de los cuadrados de las diferencias entre cada una de sus variables. Cuanto más similares sean los valores de sus características, menor será la distancia; cuanto más distintos sean, mayor será. Esta medida facilita comparar jugadores de manera objetiva, identificando cuáles se parecen más entre sí y cuáles presentan diferencias significativas en su rendimiento o estilo de juego.

De esta manera, se puede seleccionar un jugador y compararlo con un número n de jugadores de tal manera que se calcule la distancia entre ellos y, posteriormente, un porcentaje de similaridad (p = (1 - dist / max(dist)) * 100)

Para llevar a cabo este procedimiento, primero se seleccionan los eventos relevantes de cada jugador según su posición, así como los sectores del campo en los que dichos eventos ocurren. A continuación, se calculan las puntuaciones correspondientes en cada sector. Después, se crea una nueva columna que combina cada evento con su sector, de modo que el jugador solo se compare con aquellos que presentan exactamente las mismas combinaciones evento–sector. Posteriormente, se pivota el dataset. Este proceso se repite con la base de datos original (excluyendo al jugador analizado), obteniéndose finalmente un único dataset que contiene las puntuaciones de los jugadores que comparten, como mínimo, la misma combinación evento–sector que el jugador inicial. Con este dataset se pueden calcular las distancias euclídeas.

5. RESULTADOS

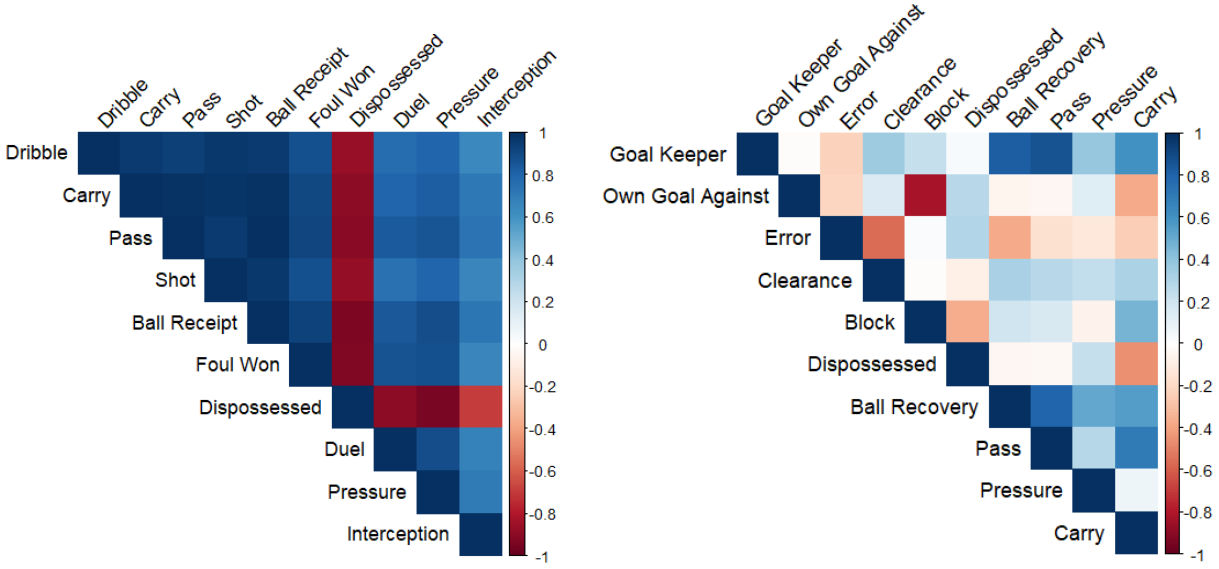
5.1 Hallazgos principales

Como se preveía en la metodología, los componentes principales permiten agrupar los 10 eventos principales de cada posición en dos únicos componentes (PC1 y PC2). La única posición que no alcanza el 75% de varianza explicada es la de portero. Esto puede deberse a que el único evento propiamente específico de esta posición es *"Goal Keeper"*. Dicho evento reúne en una sola categoría todas las estadísticas relacionadas con el portero —paradas, balones blocados, intervenciones, despejes, etc.—, lo que implica que concentra una gran cantidad de información heterogénea. Esta agregación puede dificultar que los componentes principales capten adecuadamente toda la variabilidad asociada al rendimiento del guardameta, reduciendo así el porcentaje de varianza explicada.

Posición	Porcentaje varianza acumulada explicada	
	PC1	PC2
Portero	38.3%	61%
Central	65.7%	75.5%
Lateral izquierdo	70.3%	84%
Lateral derecho	79.6%	88.6%
Pivote	65.2%	75.2%
Interior	75.8%	84.9%
Mediapunta	75.3%	85.4%
Extremo izquierdo	81.9%	88%
Extremo derecho	87.1%	92.3%
Delantero	75.5%	86.5%

Tabla 9: Porcentaje varianza acumulada de PC1 y PC2

Este porcentaje de varianza explicada se puede prever en la matriz de correlaciones, pues permite identificar el grado de relación entre las variables.



Figuras 6 y 7: matriz de correlaciones de portero vs de extremo derecho

En esta representación gráfica los cuadrantes azules representan correlación positiva entre variables, los rojos correlación negativa y los blancos incorrelación. A partir de estas matrices se observar que los eventos de los porteros no están muy correlacionados, mientras que los eventos de los extremos derechos sí están altamente correlacionados. Esto permite predecir si la información de los eventos de los jugadores puede estar bien representados con dos componentes.

La interpretación de los autovectores es fundamental, ya que indican el peso y la dirección con la que cada evento contribuye a cada componente principal. Esto permite comprender qué combinación de variables define mejor las diferencias entre jugadores dentro de una misma posición y cuáles son los aspectos del juego que explican la mayor parte de la variabilidad observada.

Eventos	Autovectores	
	PC1	PC2
Dribble	0.630	0.705
Pass	0.921	
Interception	0.859	-0.298
Clearence	0.793	-0.511
Duel	0.876	-0.363
Carry	0.869	0.291
Ball Recovery	0.954	0.115
Dispossessed	-0.783	-0.330
Foul Committed	-0.818	0.376
Foul Won	0.938	0.195

Tabla 10: autovectores eventos lateral izquierdo

Tomando como ejemplo la posición de lateral izquierdo, los loadings muestran cómo cada evento contribuye a PC1 y PC2. El primer componente (PC1) muestra cargas altas y positivas en la mayoría de las variables analizadas, por lo que representa un factor general de actividad y rendimiento en acciones con balón. Además, *Dispossessed* y *Foul Committed* tienen cargas negativas, los jugadores con valores altos en PC1 tienden a perder menos balones y cometer menos faltas, lo que refuerza la interpretación de PC1 como un componente de efectividad y control en el juego.

Por otro lado, el segundo componente (PC2) presenta un patrón diferente: variables como *Dribble* y *Foul Committed* cargan positivamente, mientras que *Clearence*, *Duel* y *Interception* lo hacen negativamente. Este contraste sugiere que PC2 está más relacionado con la conducción y la progresión ofensiva.

En conjunto, estos dos componentes proporcionan una visión clara de las dimensiones latentes del rendimiento: un primer eje que mide la eficiencia global y participación con balón, y un segundo eje que distingue tendencias ofensivas frente a defensivas.

Una vez calculados los nuevos valores según los score del PCA, se puede realizar con estos nuevos valores el análisis cluster con los k-clusters definidos anteriormente para cada posición. Como cada jugador y cada posición tiene PC1 y PC2, se puede graficar en un gráfico 2D señalando los jugadores y qué cluster pertenece cada uno, permitiendo visualizar cómo se distribuyen los jugadores según los patrones identificados por el PCA agrupando los jugadores según su proximidad en el espacio de los dos componentes. Volveremos a poner como ejemplo la posición de lateral izquierdo:

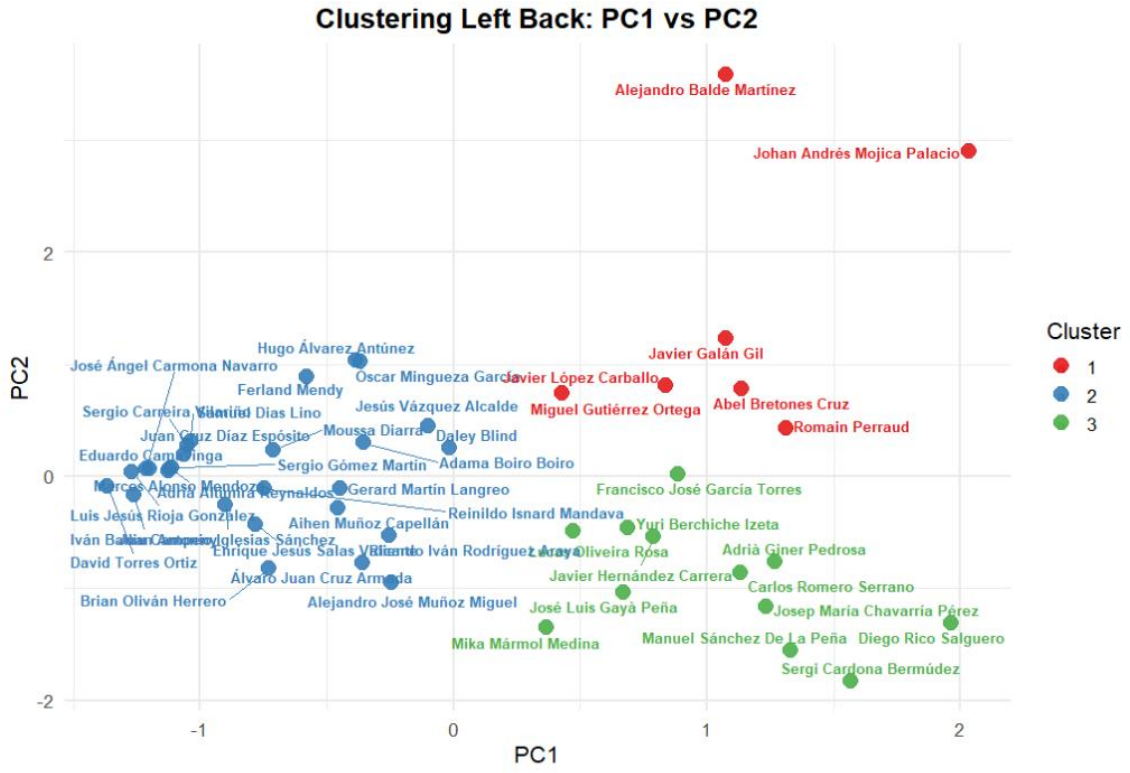


Figura 8: gráfico de clustering lateral izquierdo

En el gráfico resultante se observan tres clusters diferenciados:

- **Cluster 1 (rojo)**: agrupa a laterales con valores elevados en PC1 y PC2, lo que indica perfiles con mucha participación en el juego y muy ofensivos. Se asigna este cluster al rol de Carrilero.
- **Cluster 2 (azul)**: corresponde al grupo más numeroso y compacto. Sus jugadores tienden a situarse en valores moderados o bajos de PC1 y PC2, lo que sugiere poca participación de balón y una participación ofensiva moderada. Se asigna este cluster al rol de Defensivo.
- **Cluster 3 (verde)**: reúne jugadores que se desplazan hacia valores positivos de PC1 pero negativos de PC2. Este comportamiento indica un perfil con mayor influencia en el juego y moderados en ataque. Se asigna este cluster al rol de Ofensivo.

Una vez que todos los jugadores han sido asignados a un clúster y, por tanto, se ha definido su rol dentro de cada posición, es posible elaborar un ranking basándose en la puntuación acumulada de cada rol. Sin embargo, esta aproximación resulta limitada, ya que únicamente identifica quién es el mejor dentro de cada perfil sin mostrar el nivel relativo del resto.

Una alternativa más precisa es construir el ranking mediante percentiles, lo que permite comparar a los jugadores según su rendimiento en cada uno de los eventos analizados. De esta manera, se obtiene una evaluación más detallada y contextualizada del rendimiento. A continuación, se muestra un ejemplo aplicado a un jugador:

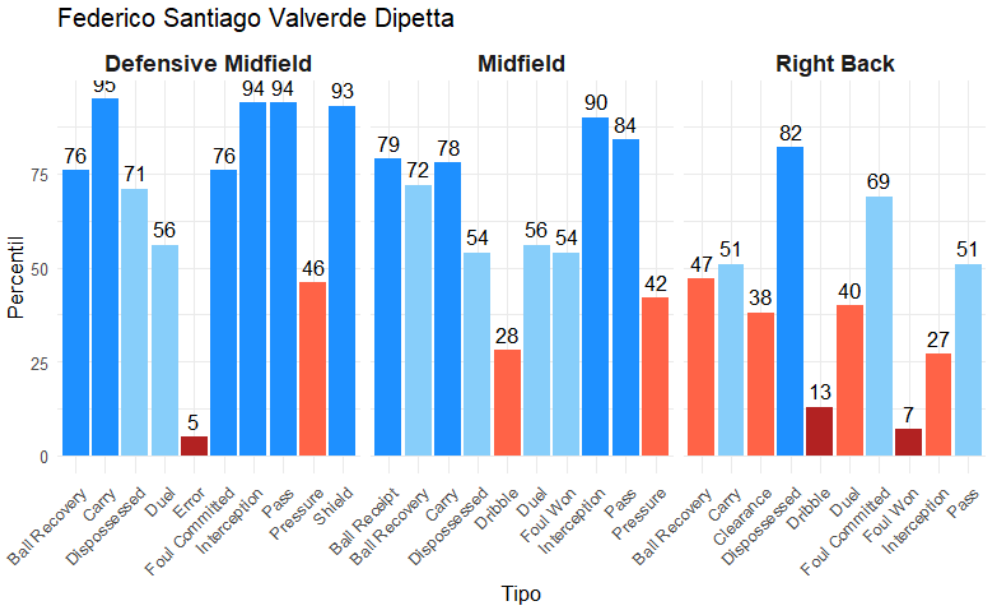


Figura 9: percentiles posiciones Federico Valverde

Este jugador tiene tres posiciones: pivote, interior y lateral derecho

- En la posición de pivote, se observa que está en percentiles muy altos en acciones con balón, pero en percentiles medios o bajos en acciones sin balón. Esto sugiere que actúa como un pivote mixto
- En la posición de mediocentro, mantiene valores altos en múltiples acciones clave. Esto refleja que en esta posición rinde como un box-to-box, capaz de aportar tanto en la creación como en la presión y el sostén del equipo.
- En la posición de lateral derecho, sus percentiles son claramente más bajos en comparación con las otras posiciones. Su rendimiento es más discreto en acciones defensivas y ofensivas propias del rol, lo que sugiere que se trata de una posición no natural para él, probablemente desempeñada por necesidades tácticas o bajas dentro del equipo. Desempeña un rol defensivo.

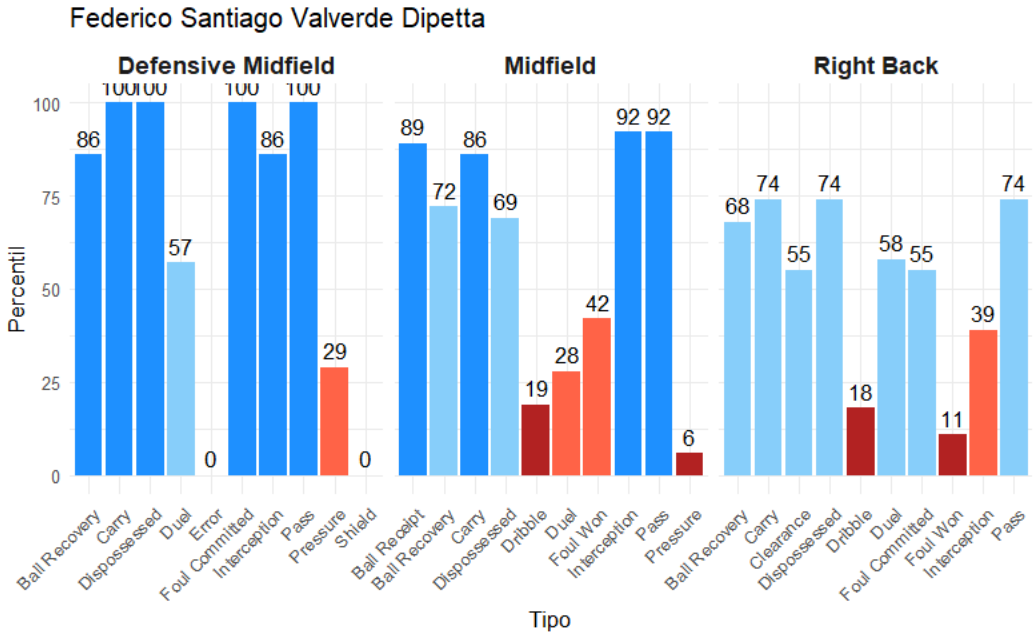


Figura 10: percentiles roles Federico Valverde

Se observan diferencias claras respecto a sus percentiles en las posiciones independientemente de su rol:

- En el rol de pivote mixto, Valverde alcanza percentiles máximos en casi la mitad de las acciones, pero más bajos en acciones defensivas puras. Este contraste indica que, dentro de su cluster, se consolida como un pivote de gran impacto con balón, muy dominante en la recuperación y progresión, pero algo menos destacado en intervenciones defensivas directas.
- En el rol de box-to-box, mantiene valores altos en recuperación, pase, presión e intercepción. Es un jugador enfocado en la presión y la progresión, aunque menos fuerte que su *cluster* en disputas y pérdidas.
- En el rol de lateral defensivo, se ve una clara mejora respecto a sus valores generales en esta posición, pero siguen siendo bastante menores respecto el nivel del jugador en el resto de las posiciones. Dentro de su rol, es un lateral más participativo que la media.

Finalmente, se lleva a cabo la comparación entre jugadores utilizando la distancia euclídea y calculando el porcentaje de similitud. Como se indicó en la metodología, cada jugador debe incluir todas las combinaciones de sector y evento del jugador de referencia. Como ejemplo práctico, se determinarán los jugadores más parecidos a Vinícius Jr., quien actualmente se encuentra en proceso de negociación con el Real Madrid para su renovación. Se buscarán cinco jugadores que puedan remplazarle si las negociaciones fracasan.

Jugador de remplazo	Distancia euclídea	Porcentaje similaridad
Nico Williams	19.07	61.79%
Jesús Rodríguez	20.8	58.32%
Bryan Zaragoza	20.85	58.22%
Abdessamad Ezzalzouli	21.84	56.23%
Bryan Gil	22.73	54.45%

Tabla 11: distancia euclídea y similaridad Vinícius Jr.

Al analizar los jugadores más similares a Vinícius Jr., se identifican cinco candidatos principales: Nico Williams, Jesús Rodríguez, Bryan Zaragoza, Abdessamad Ezzalzouli y Bryan Gil. A pesar de que todos estos jugadores comparten ciertas cualidades con Vinícius Jr., como la habilidad para encarar rivales y su capacidad de desborde por banda, sus porcentajes de similitud no son especialmente altos. Esto sugiere que, aunque pueden desempeñar roles ofensivos similares, ninguno posee exactamente el mismo perfil completo que caracteriza a Vinícius Jr., destacando así la singularidad de sus habilidades y su influencia en el juego.

6. CONCLUSIONES Y RECOMENDACIONES

Los resultados obtenidos confirman lo planteado en la metodología: los eventos registrados para cada jugador presentan relaciones claras entre sí, lo que permite reducir su complejidad mediante un Análisis de Componentes Principales (PCA). En la mayoría de posiciones, los diez eventos principales pueden agruparse en solo dos componentes (PC1 y PC2), acumulando entre un 75% y un 92% de la varianza total. Esta elevada capacidad de síntesis indica que el comportamiento estadístico de los jugadores de campo está bien estructurado y que sus acciones se organizan de forma coherente dentro del modelo.

La única excepción significativa es la posición de portero, que no alcanza el umbral del 75% de varianza explicada. Este resultado se debe, principalmente, a la naturaleza del evento “Goal Keeper”, que agrupa bajo una única etiqueta una gran variedad de acciones específicas. Al concentrar información heterogénea en una sola categoría, los datos del portero presentan una estructura mucho más compleja y menos reducible. Esto limita la capacidad del PCA para capturar adecuadamente la variabilidad del rendimiento del guardameta y, en consecuencia, dificulta la clusterización posterior.

En este sentido, los resultados ponen de manifiesto una carencia en la forma en que se registran los datos de porteros. Las empresas encargadas de la recolección, como en este caso StatsBomb, deberían considerar una clasificación más detallada y específica de las acciones del portero.

Por otro lado, el análisis cluster demuestra que los jugadores desempeñan distintos roles dentro de una misma posición y que dichos roles son reconocibles e interpretables. Esta clusterización permite identificar patrones de comportamiento específicos, diferenciar perfiles tácticos y comprender cómo se manifiestan las distintas funciones dentro del juego. Además, facilita la comparación objetiva entre jugadores, ayuda a detectar sustitutos potenciales y aporta una visión más detallada del rendimiento individual dentro del colectivo.

En conjunto, este análisis no solo enriquece la evaluación del desempeño individual y colectivo de los jugadores, sino que también proporciona una herramienta de gran utilidad para la toma de decisiones deportivas. Su aplicación resulta especialmente relevante en ámbitos como la búsqueda de talento, la comparación objetiva de perfiles a partir de un porcentaje de similaridad (calculado mediante la distancia euclídea) o la adaptación táctica de la plantilla en función de las necesidades del equipo. Además, la incorporación de comparaciones basadas en percentiles permite contextualizar el rendimiento de cada jugador dentro de la distribución global de su posición, facilitando una interpretación más precisa y estandarizada de sus fortalezas y debilidades. Al ofrecer una visión estructurada y basada en datos, el enfoque empleado contribuye a optimizar los procesos de planificación deportiva y a mejorar la interpretación del rendimiento en diferentes contextos de juego.

6.1 Líneas futuras de análisis

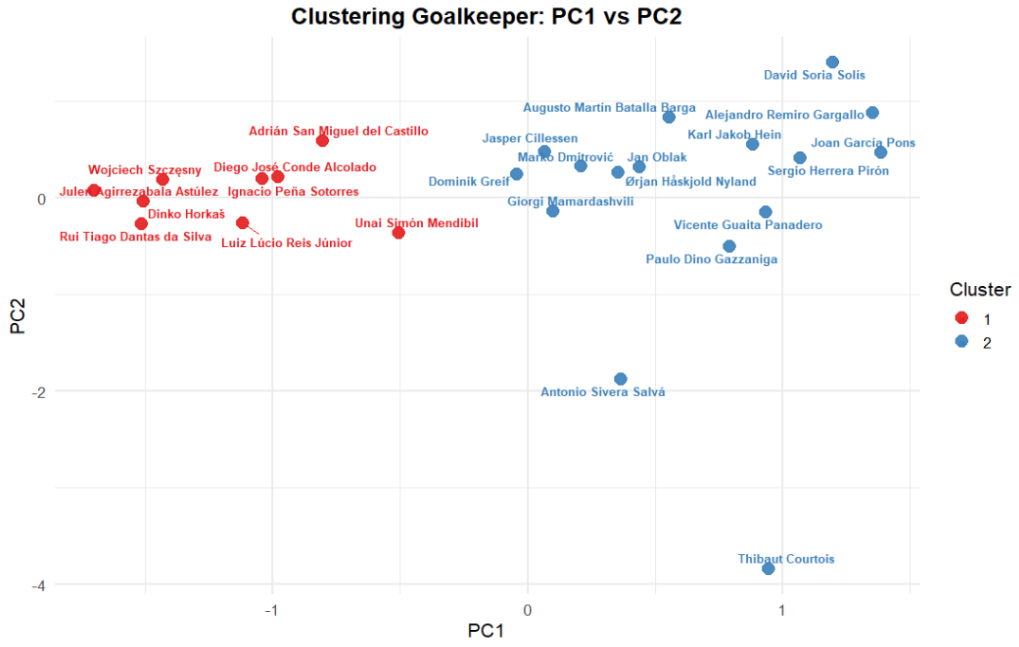
De cara a investigaciones posteriores, este enfoque podría ampliarse a través de diversas vías:

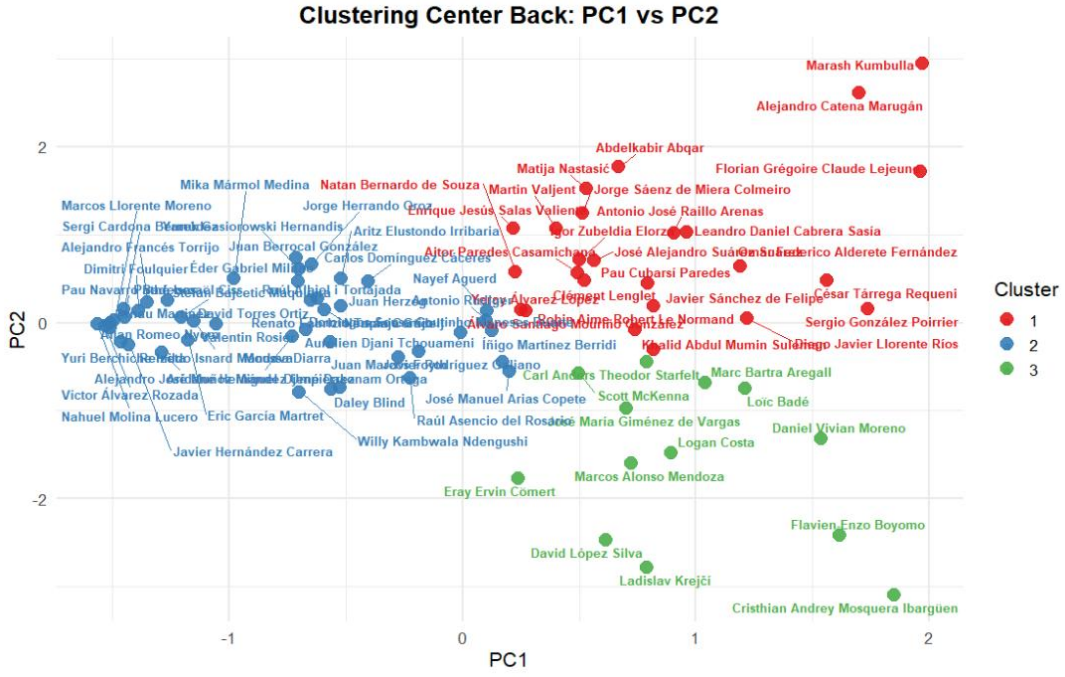
- Desagregar con mayor detalle los eventos asociados a los porteros, con el fin de clasificar de manera más precisa sus acciones y mejorar la calidad del análisis en esta posición.
- Desarrollar análisis longitudinales que permitan estudiar la evolución de los roles, los perfiles y las similitudes entre jugadores a lo largo de varias temporadas.
- Evaluar el impacto de los roles identificados sobre métricas colectivas relevantes, como los goles esperados (xG), la presión exitosa o el control territorial.

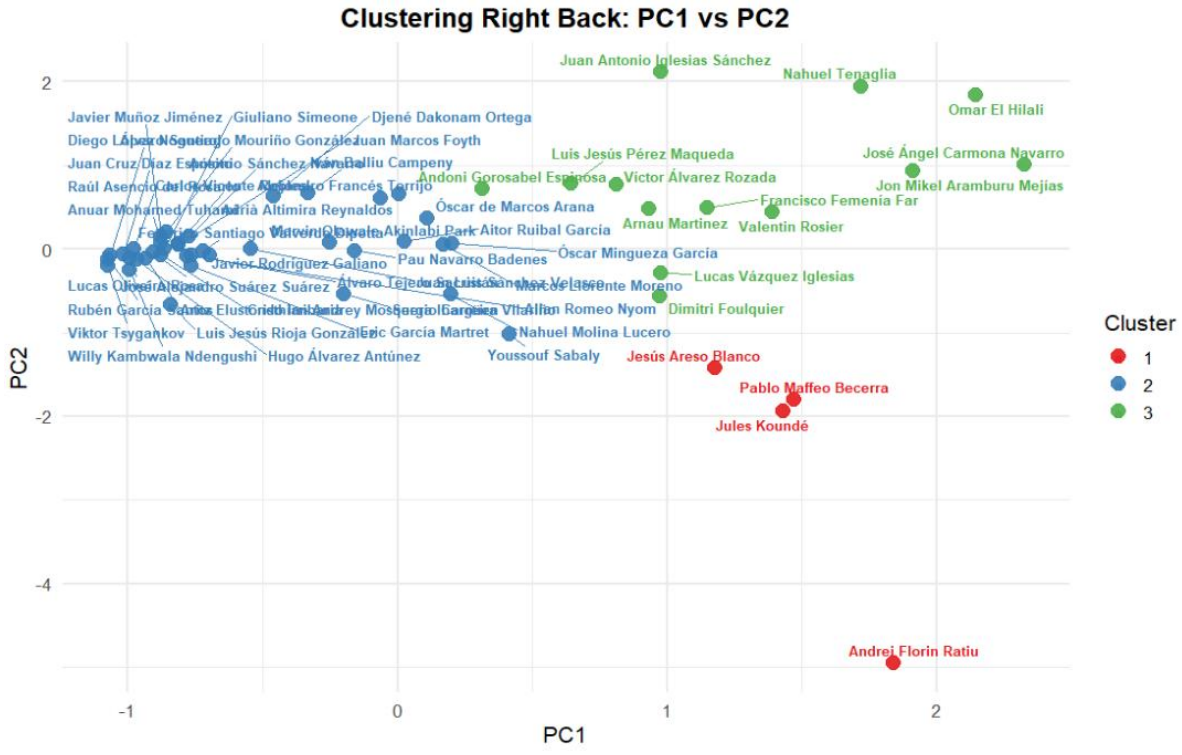
7. ANEXOS

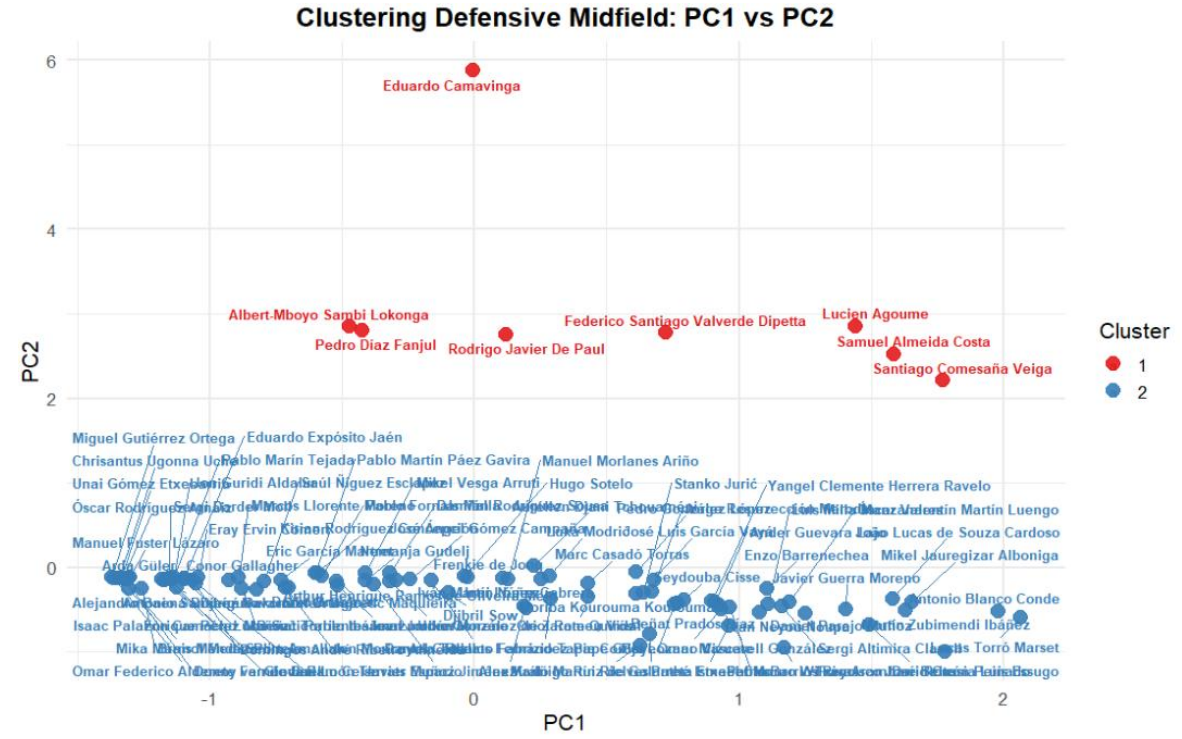
7.1 Código.

7.2 Gráficos extendidos

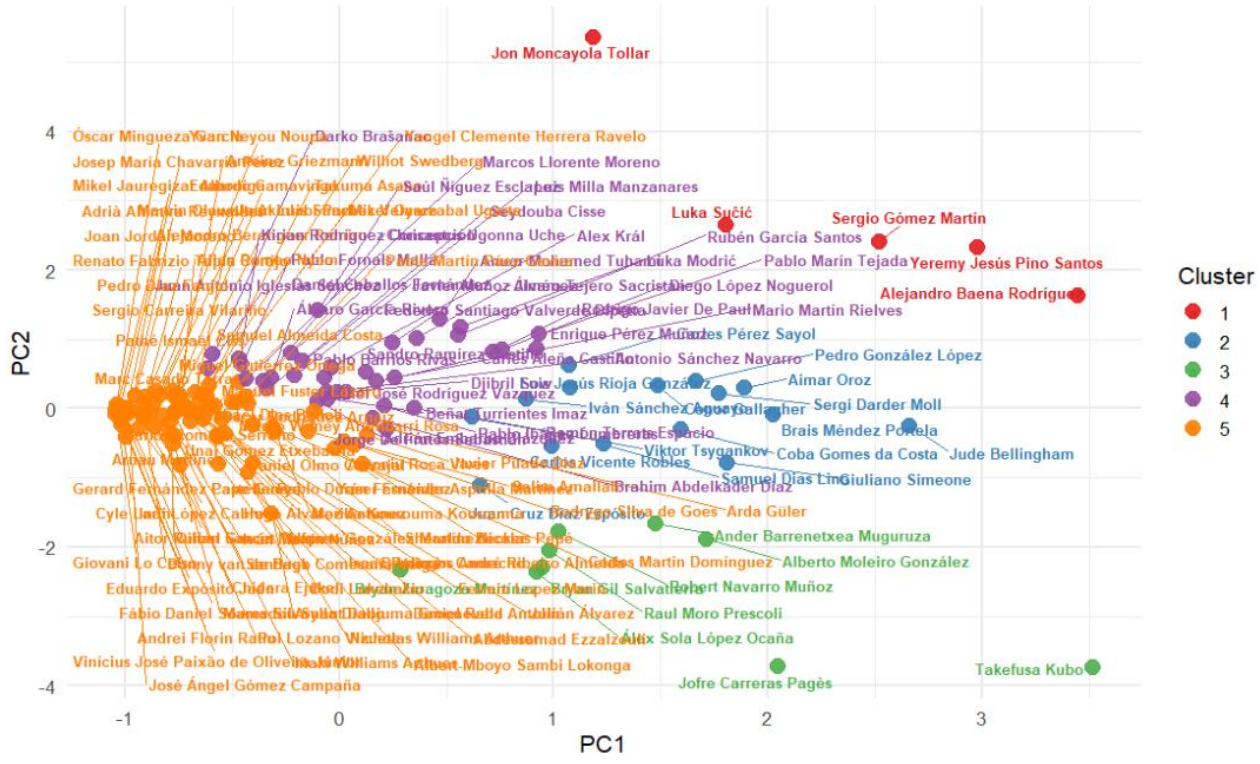




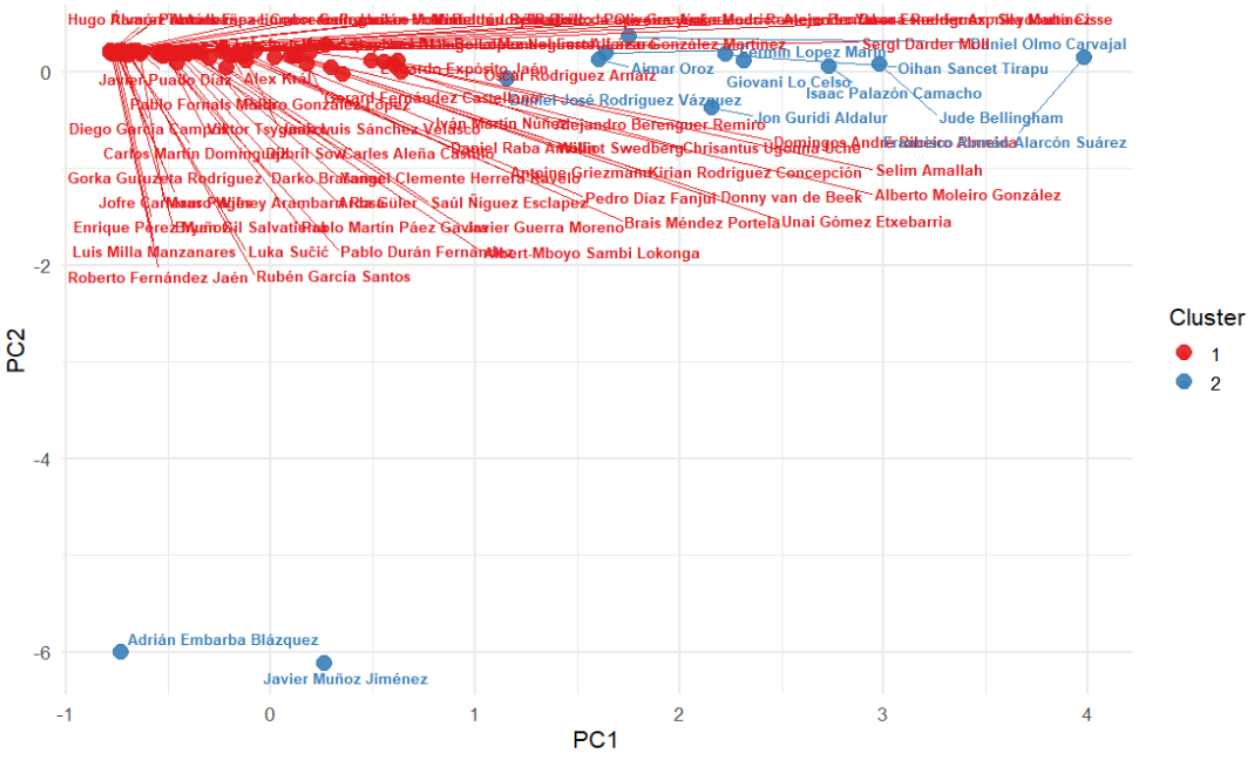

Clustering Right Back: PC1 vs PC2
PC1
PC2
Cluster
1
2
3
Juan Antonio Iglesias Sánchez
Nahuel Tenaglia
Omar El Hilali
José Ángel Carmona Navarro
Jon Mikel Aramburu Mejías
Luis Jesús Pérez Maqueda
Andoni Gorosabel Espinosa
Víctor Álvarez Rozada
Francisco Femenia Far
Arnau Martínez
Valentín Rosier
Aitor Ruibal García
Óscar Mingueza García
Lucas Vázquez Iglesias
Dimitri Foulquier
Nahuel Molina Lucero
Jesús Areso Blanco
Pablo Maffeo Becerra
Jules Koundé
Andrei Florin Ratiu
Javier Muñoz Jiménez
Giuliano Simeone
Djené Dakonam Ortega
Óscar de Marcos Arana
Pau Navarro Badenes
Javier Rodríguez Galiano
Marcos Llorente Moreno
Eric García Martret
Youssouf Sabaly
Hugo Álvarez Antúnez
Viktor Tsygankov
Willy Kambwala Ndengushi

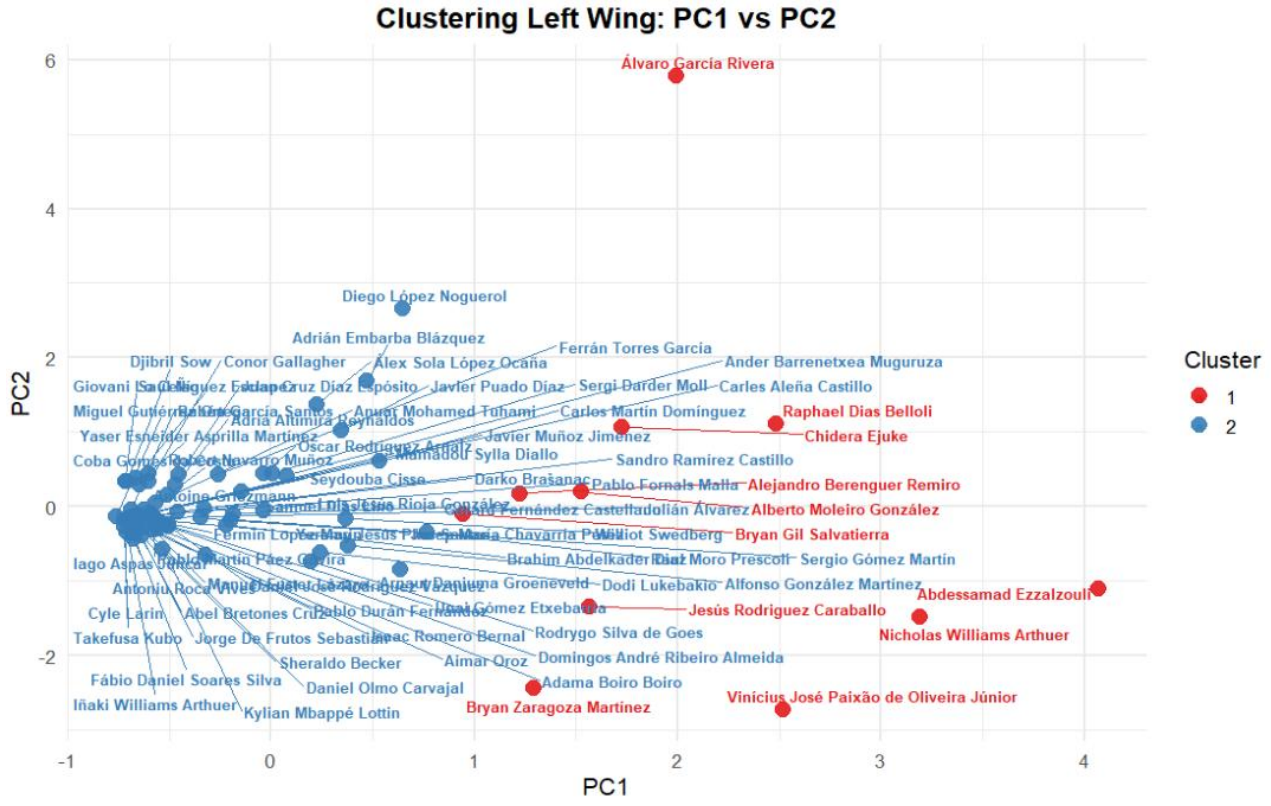

Clustering Defensive Midfield: PC1 vs PC2
PC1
PC2
Cluster
1
2
Eduardo Camavinga
Albert-Mboyo Sambi Lokonga
Pedro Díaz Fanjul
Federico Santiago Valverde Dipetta
Rodrigo Javier De Paul
Lucien Agoume
Samuel Almeida Costa
Santiago Comesaña Veiga
Miguel Gutiérrez Ortega
Eduardo Expósito Jaén
Pablo Marín Tejada
Pablo Martín Páez Gavira
Manuel Morlanes Ariño
Hugo Sotelo
Stanko Jurić
Yangel Clemente Herrera Ravelo
Manuel Fuster Lázaro
Eric García Martret
Marc Casadó Torras
Enzo Barrenechea
Mikel Jauregizar Alboniga
Javier Guerra Moreno
Antonio Blanco Conde
Sergi Altimira Clarís
Torró Marset

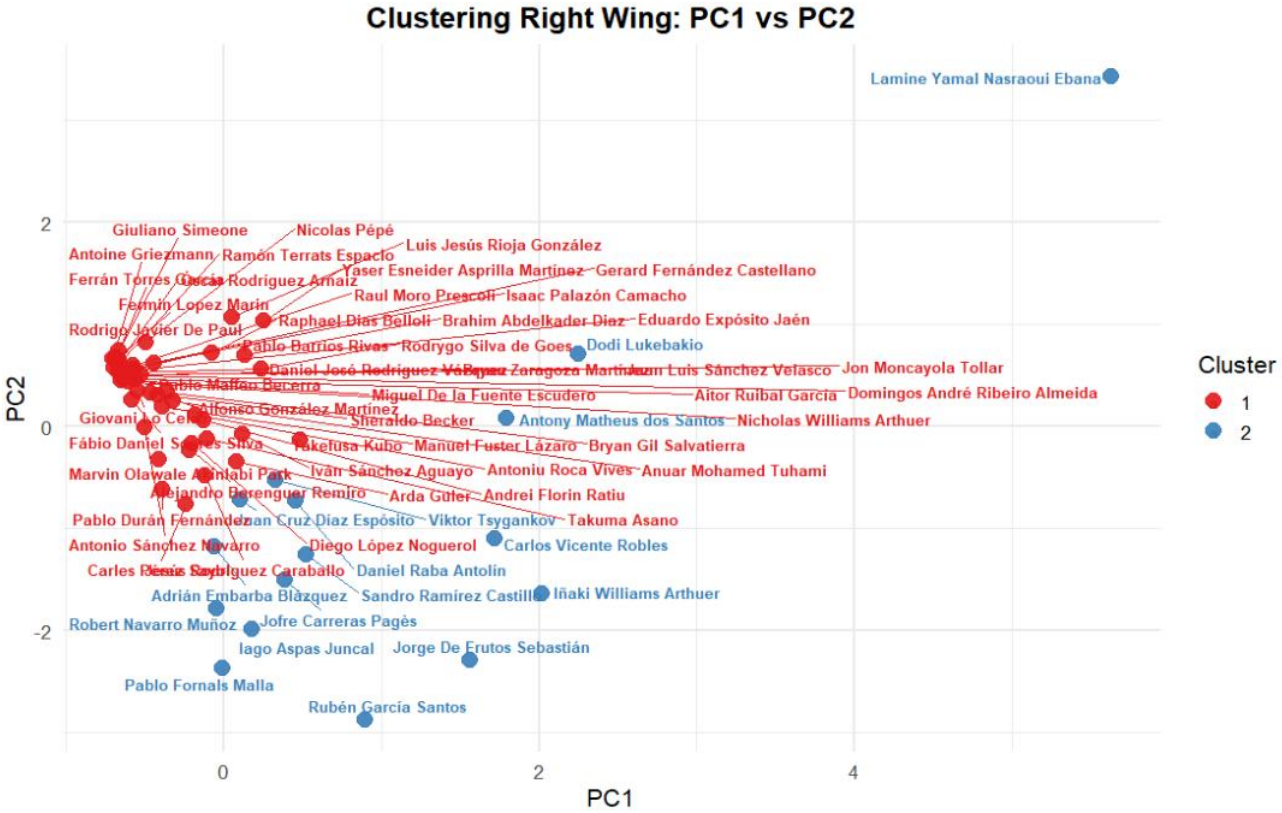
Clustering Midfield: PC1 vs PC2

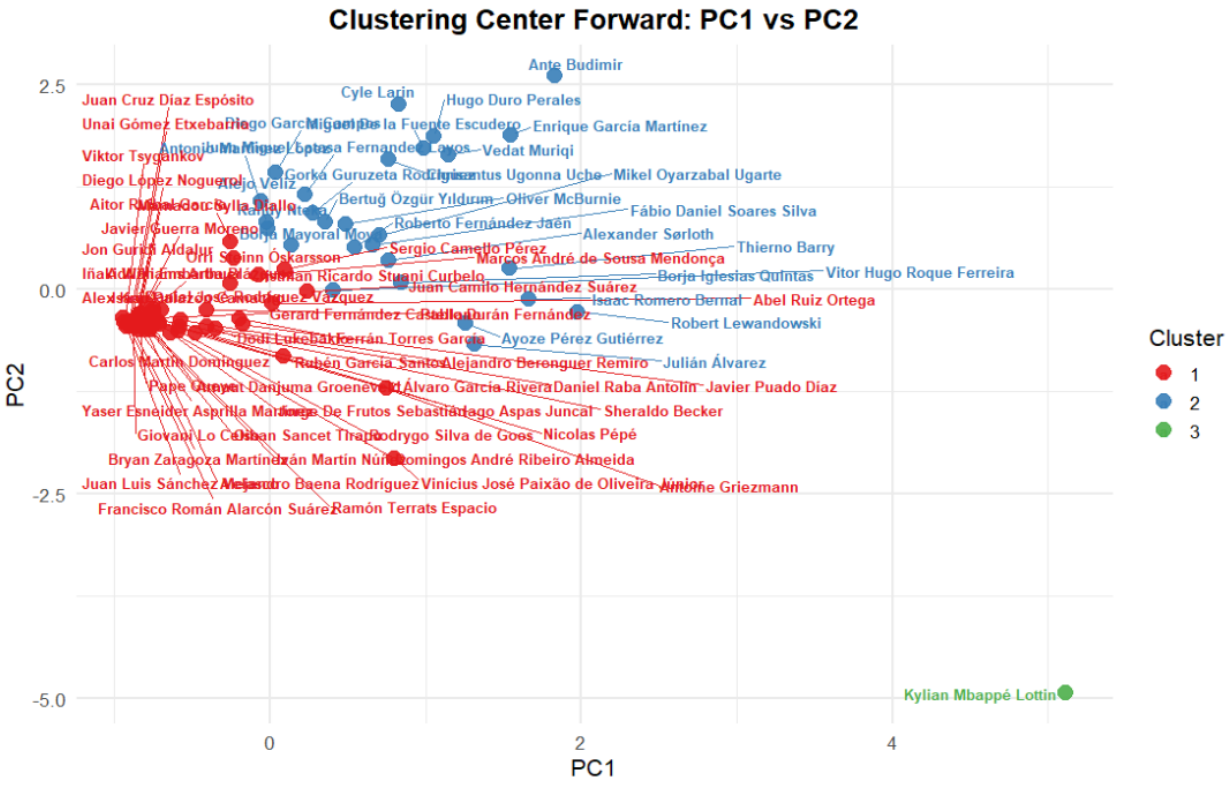


Clustering Attacking Midfield: PC1 vs PC2




Clustering Left Wing: PC1 vs PC2
PC1
PC2
Cluster
1
2
Álvaro García Rivera
Diego López Noguerol
Adrián Embarba Blázquez
Djibril Sow
Conor Gallagher
Álex Sola López Ocaña
Ferrán Torres García
Ander Barrenetxea Muguruza
Javier Puado Díaz
Sergi Darder Moll
Carles Aleña Castillo
Raphael Dias Belloli
Chidera Ejuke
Javier Muñoz Jiménez
Sandro Ramírez Castillo
Pablo Fornals Malla
Alejandro Berenguer Remiro
Alberto Moleiro González
Bryan Gil Salvatierra
Sergio Gómez Martín
Dodi Lukebakio
Alfonso González Martínez
Abdessamad Ezzalzouli
Jesús Rodríguez Caraballo
Nicholas Williams Arthuer
Rodrygo Silva de Goes
Domingos André Ribeiro Almeida
Adama Boiro Boiro
Vinícius José Paixão de Oliveira Júnior
Bryan Zaragoza Martínez
Kylian Mbappé Lottin
Iñaki Williams Arthuer
Fábio Daniel Soares Silva
Daniel Olmo Carvajal
Sheraldo Becker
Aimar Oroz
Takefusa Kubo
Cyle Larin
Antoniu Roca Vives
Yaser Esneider Asprilla Martínez
Oscar Rodríguez Arnaiz
Coba Gomes
Seydouba Cissa
Darko Brašanac
Antoine Griezmann
Fermín López Marín
Brahim Abdelkader Díaz
Dani Gómez Etxebarria
Isaac Romero Bernal
Jorge De Frutos Sebastián
Iago Aspas Juncal
Pablo Martín Páez Gálvez
Abel Bretones Cruz
Pablo Durán Fernández
Arnaut Danjuma Groeneveld
Miguel Gutiérrez Rodríguez
Adria Altimira Reynaldos
Anuar Mohamed Tuhami
Carlos Martín Domínguez
Mamadou Sylla Diallo
Robert Navarro Muñoz
Luis Rioja González
Gerard Fernández Castelladolián Álvarez
Raul Moro Prescoli
Williot Swedberg
Ramón Terrats Espacio
Rodrigo Javier De Paul
Eduardo Expósito Jaén
Pablo Barrios Rivas
Daniel José Rodríguez Váquez
Iván Luis Sánchez Velasco
Jon Moncayola Tollar
Pablo Maffeo Becerra
Miguel De la Fuente Escudero
Aitor Ruibal García
Giovani Lo Celso
Alfonso González Martínez
Antony Matheus dos Santos
Marvin Olawale Akinlabi Park
Iván Sánchez Aguayo
Antoniu Roca Vives
Anuar Mohamed Tuhami
Alejandro Berenguer Remiro
Arda Güler
Andrei Florin Ratiu
Pablo Durán Fernández
Iván Cruz Díaz Espósito
Viktor Tsygankov
Takuma Asano
Antonio Sánchez Navarro
Diego López Noguerol
Carlos Vicente Robles
Carles Reus Rodríguez
Daniel Raba Antolín
Adrián Embarba Blázquez
Sandro Ramírez Castillo
Iñaki Williams Arthuer
Robert Navarro Muñoz
Jofre Carreras Pagès
Iago Aspas Juncal
Jorge De Frutos Sebastián
Pablo Fornals Malla
Rubén García Santos


Clustering Right Wing: PC1 vs PC2
PC1
PC2
Cluster
1
2
Lamine Yamal Nasraoui Ebana
Giuliano Simeone
Nicolas Pépé
Antoine Griezmann
Ramón Terrats Espacio
Luis Jesús Rioja González
Ferrán Torres García
Yaser Esneider Asprilla Martínez
Gerard Fernández Castellano
Fermín López Marín
Raul Moro Prescoli
Isaac Palazón Camacho
Rodrigo Javier De Paul
Raphael Dias Belloli
Brahim Abdelkader Díaz
Eduardo Expósito Jaén
Pablo Barrios Rivas
Rodrygo Silva de Goes
Dodi Lukebakio
Pablo Maffeo Becerra
Miguel De la Fuente Escudero
Aitor Ruibal García
Domingos André Ribeiro Almeida
Giovani Lo Celso
Alfonso González Martínez
Sheraldo Becker
Antony Matheus dos Santos
Nicholas Williams Arthuer
Fábio Daniel Soares Silva
Takefusa Kubo
Manuel Fuster Lázaro
Bryan Gil Salvatierra
Marvin Olawale Akinlabi Park
Iván Sánchez Aguayo
Antoniu Roca Vives
Anuar Mohamed Tuhami
Alejandro Berenguer Remiro
Arda Güler
Andrei Florin Ratiu
Pablo Durán Fernández
Iván Cruz Díaz Espósito
Viktor Tsygankov
Takuma Asano
Antonio Sánchez Navarro
Diego López Noguerol
Carlos Vicente Robles
Carles Reus Rodríguez
Daniel Raba Antolín
Adrián Embarba Blázquez
Sandro Ramírez Castillo
Iñaki Williams Arthuer
Robert Navarro Muñoz
Jofre Carreras Pagès
Iago Aspas Juncal
Jorge De Frutos Sebastián
Pablo Fornals Malla
Rubén García Santos


Clustering Center Forward: PC1 vs PC2
PC1
PC2
Cluster
1
2
3
Ante Budimir
Cyle Larin
Hugo Duro Perales
Enrique García Martínez
Vedat Muriqi
Mikel Oyarzabal Ugarte
Oliver McBurnie
Fábio Daniel Soares Silva
Alexander Sørloth
Thierno Barry
Vitor Hugo Roque Ferreira
Abel Ruiz Ortega
Robert Lewandowski
Julián Álvarez
Javier Puado Díaz
Sheraldo Becker
Nicolas Pépé
Vinicius José Paixão de Oliveira Júnior
Antoine Griezmann
Kylian Mbappé Lottin
Juan Cruz Díaz Espósito
Viktor Tsygankov
Diego López Noguerol
Ramón Terrats Espacio
2.5
0.0
-2.5
-5.0
0
2
4