UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Un estudio de modelos y algoritmos para la detección de patrones similares de juego en equipos de fútbol profesional

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Carmona Salido, Mario

Tutor/a: Sánchez Anguix, Víctor

Cotutor/a: Alberola Oltra, Juan Miguel

CURSO ACADÉMICO: 2022/2023

Resumen

En esta investigación se presenta una metodología innovadora destinada a analizar exhaustivamente el estilo de juego adoptado por los equipos en las cinco principales ligas europeas. Para llevar a cabo este estudio, se emplearon minuciosamente los datos de eventos proporcionados por Wyscout, recopilados durante la temporada 2017/2018, permitiendo la identificación de los 25 tipos de posesiones más prevalentes en dichas competiciones. Posteriormente, se ejecutó un proceso de agrupación que englobó a los 98 equipos en categorías definidas por las características que componen su estilo de juego. Los resultados obtenidos revelan diferencias estadísticamente significativas entre los grupos generados mediante el empleo de tres métodos de agrupación distintos, a saber, HDBSCAN, aglomerativo y K-Means. Con el uso de HDBSCAN, se logró una división en dos grupos: uno de élite y otro de equipos menos destacados, basándose en la calidad general de los equipos. En cambio, el método aglomerativo permitió la identificación de un grupo por cada liga analizada, agrupando de esta forma a los equipos que pertenecían a la misma liga. Por último, gracias a K-Means, se pudieron identificar hasta 13 estilos de juego más específicos, brindando una visión sumamente detallada del panorama futbolístico europeo. Esta investigación representa un importante avance en la comprensión y análisis de las dinámicas de juego en el fútbol europeo, aportando valiosos insights que pueden ser de utilidad para entrenadores, analistas y aficionados al deporte en todo el mundo.

Palabras clave: Fútbol, análisis futbolístico, estilo de juego, posesiones, clustering, Big Data, ciencia de datos, HDBSCAN, aglomerativo, K-Means, ANOVA.

Abstract

This research presents an innovative methodology aimed at thoroughly analyzing the playing style adopted by teams in the top five European leagues. To conduct this study, data from events provided by Wyscout, collected during the 2017/2018 season, was meticulously employed, allowing for the identification of the 25 most prevalent possession types in these competitions. Subsequently, a clustering process was executed that encompassed the 98 teams into categories defined by the characteristics that define their playing style. The results obtained reveal statistically significant differences among the groups generated using three different clustering methods, namely, HDBSCAN, agglomerative, and K-Means. With the use of HDBSCAN, a division into two groups was achieved: one for elite teams and another for less prominent teams, based on the overall quality of the teams. In contrast, the agglomerative method allowed for the identification of one group for each league studied, thus grouping teams belonging to the same league. Lastly, thanks to K-Means, up to 13 more specific playing styles could be identified, providing an extremely detailed view of the European football landscape. This research represents a significant advancement in the understanding and analysis of football dynamics in Europe, offering valuable insights that can be useful for coaches, analysts, and sports enthusiasts worldwide.

Key words: Soccer, soccer analytics, style of play, possessions, clustering, Big Data, data science, HDBSCAN, Agglomerative, K-Means, ANOVA.

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

El mundo del deporte viene siendo caracterizado en las últimas décadas por un aumento en la competitividad debido, entre otros motivos, al compromiso para con el espectáculo y los espectadores. La ambición competitiva de los deportistas y el interés de las televisiones en acaparar el mayor número de televidentes crean un caldo de cultivo perfecto para el desarrollo de todos aquellos factores que permiten aumentar la intensidad competitiva cada vez más.

En este contexto, el análisis deportivo viene adquiriendo cada vez un papel más fundamental en el día a día tanto de preparadores como de deportistas. El análisis deportivo es un meticuloso proceso de investigación que abarca la recopilación, estudio e interpretación de datos, métricas y estadísticas relacionados con el deporte, desde aspectos tácticos y rendimiento físico hasta estrategia y toma de decisiones, contribuyendo a la mejora continua en el mundo deportivo. Este tipo de análisis permite, por un lado, la toma de decisiones informadas, basadas en la evidencia; y por otro, optimizar el rendimiento de los deportistas, por ejemplo, identificando áreas de mejora y riesgo de lesiones. La combinación de estas dos características permite a los deportistas conseguir una ventaja competitiva respecto al resto. A este respecto, «Moneyball: The Art of Winning an Unfair Game» (Lewis, 2004) es uno de los libros más conocidos en la literatura sobre la trayectoria del análisis deportivo, pues cuenta la historia de un evento que marcó un antes y un después en el deporte en general y en el béisbol en particular. Moneyball relata cómo el equipo de béisbol de los Oakland Athletics revolucionó la industria implementando un enfoque sistemático basado en el análisis estadístico para reclutar jugadores contando con un presupuesto muy inferior al de sus competidores. Este enfoque permitió al equipo conseguir resultados muy exitosos y romper con las convenciones clásicas establecidas en el deporte hasta la fecha.

En la NBA encontramos otro claro ejemplo, Daryl Morey implantó la filosofía de la toma de decisiones basada en datos en los Houston Rockets. El resultado fue notorio, los Houston Rockets, equipo del cual el Sr. Morey era director ejecutivo, logró romper el récord de porcentaje de tiros de tres anotados en una temporada.

En el fútbol, por su parte, la implementación de este tipo de técnicas y metodologías ha recibido un gran rechazo desde los sectores más conservadores durante los últimos años, lo que ha complicado enormemente la irrupción de estas. No obstante, entidades de categoría como La Liga están cada vez más comprometidas con el análisis y procesado de estadísticas en tiempo real. En la temporada 2021/22, La Liga y Microsoft llegaron a un acuerdo para lanzar Beyond Stats, un proyecto de estadísticas de fútbol avanzadas. Según la propia compañía, «Beyond Stats ofrece un análisis de grandes conjuntos de datos dinámicos tratados por la plataforma Mediacoach, parte del portfolio de LaLiga Tech, la compañía de soluciones tecnológicas especializadas para el sector del deporte

y entretenimiento lanzada este 2021 al abrigo de LaLiga. Estas métricas llegan ahora al usuario gracias a la Inteligencia Artificial y el Machine Learning que proporciona Microsoft Azure» (Prensa, 2021). Este es el primer paso en firme de cara a ofrecer este tipo de estadísticas al público de forma abierta. No obstante, algunos equipos ya contaban con departamentos de análisis de datos de forma, pero llevando a cabo este tipo de empresas siempre desde la privacidad. Este tipo de datos son rara vez accesibles de forma abierta y generalmente cuentan con un precio muy privativo, que sólo está al alcance de las entidades más grandes.

1.1 Motivación

La irrupción del aprendizaje automático y la inteligencia artificial en este deporte ha abierto un abanico de posibilidades que antes eran inimaginables. La confluencia de estas tecnologías con el fútbol, una disciplina que se caracteriza por su riqueza táctica y su constante evolución, plantea desafíos que deberán ser superados para exprimir al máximo el potencial de este deporte. En este sentido, la identificación de estilos o patrones de juegos es un campo que aún se encuentra en desarrollo, y la exploración del uso de diversas herramientas y tecnologías, como la inteligencia artificial y el aprendizaje automático, es esencial para aprovechar al máximo el potencial de este deporte.

El fútbol se distingue por su carácter global y altamente competitivo, en el que incluso la más mínima ventaja puede ser determinante. En este contexto, la automatización de los procesos de análisis táctico representa un objetivo de gran envergadura. La idea de crear herramientas que ofrezcan a entrenadores y equipos acceso a análisis detallados y personalizados sobre el juego de sus rivales y su propio equipo en cuestión de minutos se antoja fascinante. Esto no solo ahorraría tiempo valioso, sino que también permitiría una toma de decisiones más precisa y estratégica.

Los aspectos tácticos del fútbol son cruciales en la búsqueda de la victoria. Anticipar las estrategias del adversario y adaptarse en consecuencia es un desafío constante. En este contexto, surge la pregunta: ¿Es posible utilizar herramientas avanzadas de análisis de datos para identificar patrones y tendencias en el juego que escapan a la percepción humana? Este trabajo pretende hacer uso de este tipo de técnicas para el desarrollo de una metodología que permita explorar a fondo las características comunes y diferenciales del estilo de juego de cada equipo. Esta información puede resultar particularmente valiosa para lograr, por un lado, una mayor comprensión acerca de este complejo deporte, y por otro, identificar rápidamente las características de un equipo cuando no se dispone de conocimiento previo sobre el mismo.

En este sentido, se ha considerado que aprovechar la disponibilidad de los datos de eventos de Wyscout de la temporada 2017/2018 supone el punto de anclaje ideal para iniciar este proyecto. Estos datos ofrecen una enorme cantidad de información acerca de los equipos de las cinco grandes ligas europeas, de hecho, el alcance de este proyecto no contemplará el procesado de toda la información disponible en los ficheros de datos debido a las limitaciones temporales. En la siguiente sección se detallan los objetivos de este estudio.

1.2 Objetivos

El objetivo principal de este estudio será el estudio de los estilos de juego de los equipos de las cinco grandes ligas. Para ello, se definirán algunos objetivos específicos a cumplir para valorar el éxito del estudio.

- Crear una metodología para la recopilación y procesamiento de los datos de eventos que permita identificar el estilo de juego de un club. Esta metodología debe combinar técnicas propias de la estadística clásica aplicada a la ciencia del deporte y técnicas informáticas de aprendizaje más modernas relacionadas con el *Big Data*.
- Identificar los patrones de posesiones observables más comunes entre las cinco grandes ligas europeas: La Liga, Bundesliga, Ligue 1, Premier League y Serie A.
- Estudiar si existen diferencias entre los estilos de juego de cada una de estas ligas.
- Proponer una clasificación de estilos de juego en los que englobar a cada uno de los clubes.

1.3 Marco legal y ético

Desde una perspectiva legal, este trabajo hace uso de datos completamente abiertos para uso académico. El autor del artículo donde se publican únicamente explicita que será necesario referenciar debidamente el documento en el trabajo. En la sección 3.1 se detallará más a fondo el contenido y origen de los datos; sin embargo, es oportuno recalcar que la fuente de los datos ha sido debidamente referenciada en este trabajo.

A nivel ético no se presentan grandes implicaciones más allá de la garantía de la integridad y la precisión de los resultados extraídos en el análisis. Esto incluye la veracidad en la representación de los eventos deportivos y características de los equipos estudiados. Los resultados han sido considerados siempre desde un punto de vista objetivo y procurando el máximo respeto a todas las entidades involucradas en el estudio.

1.4 Estructura de la memoria

A continuación, se describirán cada uno de los capítulos que conforman la estructura de la memoria junto con una pequeña descripción de los mismos.

- **Marco teórico:** En este apartado se presenta el contexto teórico en el que se enmarca el trabajo o proyecto. Se introducirán conceptos clave, teorías relevantes y antecedentes que sirvan como base para comprender el trabajo.
- **Herramientas utilizadas y análisis exploratorio previo:** En este apartado se describen las herramientas y técnicas utilizadas para llevar a cabo la investigación o proyecto. Además, se incluirá el análisis exploratorio inicial de los datos que se utilizaron.
- **Mi propuesta:** Aquí se presenta la metodología propuesta para llevar a cabo el estudio. Se trata del capítulo central del trabajo, donde se detallan todos los aspectos de interés a contemplar para abordar el problema.
- **Experimentación y validación de resultados:** En este apartado se describe cómo se llevó a cabo la experimentación para la optimización de los parámetros de los clusterings aplicados. Se incluyen detalles sobre la metodología utilizada y cómo se validaron los resultados obtenidos.
- **Resultados:** Se presentan los resultados de la experimentación o investigación de manera clara y concisa. Se incluyen gráficos y justificaciones desde una visión deportiva para respaldar los hallazgos.

- **Conclusiones:** Aquí se resumen las conclusiones clave derivadas del trabajo realizado. Se pueden destacar las implicaciones de los resultados y posibles direcciones futuras de investigación.

CAPÍTULO 2
Marco teórico

En el marco teórico, se introducirá el análisis deportivo como concepto y se explorarán diferentes casos de uso en el mundo de fútbol para la toma de decisiones basadas en datos. Finalmente, se expondrá el estado del arte en cuestión de búsqueda de similitudes entre patrones de juego junto con las herramientas más utilizadas en este ámbito.

2.1 Análisis deportivo

Uno de los libros que aborda el análisis deportivo en profundidad es «Sports Analytics: A Guide for Coaches, Managers and Other Decision Makers» (Alamar, 2013) donde se define este término como «la gestión de datos históricos estructurados, la aplicación de modelos analíticos predictivos que utilizan esos datos y el uso de sistemas de información para informar a los tomadores de decisiones y permitirles ayudar a sus organizaciones a obtener una ventaja competitiva en el campo de juego». Se trata de una definición extensa que persigue abarcar todos los aspectos clave del proceso.

La base fundamental de cualquier tipo de análisis basado en la evidencia es la recopilación de datos, que, en este caso, puede incluir registros de eventos durante un partido, mediciones biométricas de los atletas, datos de rendimiento físico y muchas otras fuentes de información relevante. La calidad y la cantidad de los datos recopilados son aspectos fundamentales para obtener resultados confiables y significativos.

Los recientes avances tecnológicos han permitido la miniaturización de sensores y dispositivos de monitoreo, lo que ha hecho de la captura de datos una tarea mucho más accesible para atletas de diferentes niveles. En la actualidad existen en el mercado diferentes tipos de sistemas biométricos altamente sofisticados, basados en visión, en ondas de radio o sistemas de posicionamiento global (GPS) entre otros (Leser et al., 2011). No obstante, este campo sigue siendo un área de estudio de interés en la actualidad. Boluarte Pretell (2022) propone dos sistemas inalámbricos, «Sistema de análisis de gases y Sistema de aceleraciones», el primero encargado de medir la concentración de CO2 en la respiración y el volumen respirado y, el segundo, encargado de medir la aceleración a la que se somete el usuario.

Asimismo, el desarrollo de aplicaciones que permitan procesar y visualizar los datos recogidos por los sistemas de captura también se ha visto acelerado. El auge de la tecnología portátil y el desarrollo de la capacidad computacional de los teléfonos móviles han permitido el diseño de aplicaciones capaces de recoger y procesar los datos biométricos para ser visualizados por pantalla en tiempo real. En este contexto se presenta, por ejemplo, Dron-fit (los Ríos et al., 2017) que es una aplicación que permite al usuario hacer un seguimiento de su actividad física en vídeo mediante grabación aérea con un dron. Esta

aplicación permite obtener en tiempo real información biométrica mediante una aplicación móvil. Además, está diseñada para enviar alertas al detectar anomalías en alguno de los parámetros recogidos.

Por otro lado, encontramos que el análisis deportivo no engloba únicamente el estudio de las características físicas de los deportistas. El estudio de la estrategia intrínseca de cada deporte es otro de los grandes objetivos que se persigue desarrollar cuando se realizan este tipo de análisis. En este contexto, el desarrollo de aplicaciones capaces de detectar patrones de juego que den soporte a las decisiones tácticas de los técnicos es una de las áreas de estudio que más viene desarrollándose en los últimos años. Wu et al. (2021) propone un sistema de visualización basado en minería de patrones capaz de «obtener de manera efectiva conocimientos sobre la progresión de tácticas en la mayoría de los deportes de raqueta». Este tipo de herramientas con capacidad para facilitar la identificación y análisis exploratorio de tácticas diferentes son muy demandadas por, entre otros, equipos técnicos de clubes y entrenadores de deportistas.

Finalmente, es conveniente hablar de un campo que, a pesar de no guardar una relación tan estrecha con el análisis deportivo como lo anteriormente mencionado, viene cobrando una importancia cada vez más capital en el deporte: la predicción de resultados para apuestas deportivas. Se trata de un área en la que el exponencial crecimiento del aprendizaje automático Mitchell et al. (2007) ha permitido aumentar la precisión y, por ende, la rentabilidad de las apuestas. Se trata de otro claro ejemplo de toma de decisiones basadas en evidencia y que, en cierta medida, trata de extraer toda la información posible de los deportistas para pronosticar el resultado de cualquier enfrentamiento. Horvat and Job (2020) realizan una revisión de más de 100 papers analizando cuáles son los modelos más utilizados en la predicción de resultados en diferentes deportes y su rendimiento. Cualquier tipo de información de esta índole que pueda ser extraída toma cada vez más valor tanto para deportistas como para las entidades deportivas.

2.2 Analítica en fútbol

El fútbol es considerado el deporte más practicado a nivel mundial, lo que lo convierte, por tanto, en uno de los deportes con mayor exigencia competitiva, especialmente cuando hablamos del fútbol europeo. Es por este motivo que los clubes buscan constantemente herramientas que sean capaces de otorgarles esa ventaja estratégica que les permita permanecer en la élite y competir por ser el mejor. Dentro del contexto de este deporte podemos identificar diversas ramas de estudio:

En primer lugar, la prevención de lesiones, que cuenta con un papel fundamental en el fútbol de élite donde los calendarios son altamente exigentes. En este sentido, el estudio sobre la rotura del ligamento cruzado anterior es sin duda uno de los temas tendencia. Gupta et al. (2020) analiza las diferencias en la incidencia de este problema entre hombres y mujeres que practican fútbol, siendo estas últimas tres veces más propensas a sufrir una rotura de ligamento cruzado anterior. Además, se proponen mecanismos de prevención para evitar esta lesión, que cuenta con la capacidad de acabar con la carrera deportiva de cualquier deportista que la sufra. En la misma dirección apunta Villa et al. (2020) que realiza un exhaustivo estudio sobre diferentes características de las lesiones de ligamento cruzado anterior. Mediante vídeo análisis se estudian mecanismos, patrones y la biomecánica de este tipo de lesiones y las situaciones en las que se producen. También podemos encontrar otros estudios como Peel et al. (2020) que indagan en la posición y rotación del pie. No es de extrañar, por tanto, que el trabajo específico para la prevención de lesiones o para la posterior recuperación sea parte del día a día de los jugadores y suponga un esfuerzo capital para los equipos médicos.

Siguiendo en esta línea, cuando hablamos de fútbol de élite, otra de las grandes preocupaciones de los cuerpos técnicos es la recuperación física de los jugadores. Los cada vez más saturados calendarios obligan a los clubes a contar con plantillas amplias en las que cada uno de ellos debe estar en perfecta disposición para jugar cualquier encuentro en cualquier minuto. Por ello, en la actualidad los futbolistas se ven sometidos a fuertes controles sobre su preparación. Desde la dieta hasta las horas de sueño, los equipos médicos vigilan las condiciones a las que se exponen tras cada partido con el fin de optimizar la recuperación de cada uno de ellos.

No obstante, no es sencillo encontrar en la literatura mucha información al respecto, Altarriba-Bartes et al. (2021) compara los diferentes métodos de recuperación empleados por los clubes de la primera división española en la temporada 2018/19 y los tres que ascendieron para la 2019/20. Se trata de una investigación que pretende arrojar algo de luz en este ámbito y en la que destaca una variabilidad significativa entre los diferentes equipos. La diferencia de recursos entre los distintos equipos conlleva que no todos sean capaces de usar métodos cuya eficacia haya sido demostrada por evidencia científica.

En segundo lugar, el desarrollo de análisis tácticos es fundamental para brindar a los cuerpos técnicos de los clubes herramientas que les permitan desempeñar su labor de manera efectiva. Estos análisis tienen como objetivo comprender y evaluar las diferentes estrategias, sistemas de juego y movimientos tácticos empleados por los equipos. Dentro del análisis táctico podemos identificar dos aspectos clave, el juego individual y el juego en equipo.

El juego individual hace referencia al desarrollo de tácticas que busquen maximizar o minimizar el impacto de algún jugador concreto en el desarrollo del juego. Para acometer esta tarea es necesario realizar informes precisos sobre cada uno de los jugadores de interés, desde sus capacidades físicas hasta sus movimientos naturales dentro del terreno de juego. Las aplicaciones encargadas de realizar este tipo de tareas suelen ser denominadas como «aplicaciones de *scouting*» ya que los *scouts* (ojeadores en español) son las personas encargadas de realizar este trabajo basando la toma de decisiones en la intuición y experiencia. Esta labor viene siendo complementada en los últimos años por aplicaciones que buscan dar un enfoque más objetivo a la toma de decisiones en esta área. Desde organizar y planificar la plantilla para la temporada buscando aquellos jugadores que mejor encajen en el juego del equipo, hasta rastrear jóvenes jugadores con capacidades prometedoras. Son diversos los casos de uso que podemos encontrar para aplicaciones como «The Scouting App» avalada por el equipo de *scouting* de diversos clubes de élite. No obstante, hoy en día este campo sigue en vías de desarrollo y podemos encontrar en la literatura trabajos basados en el desarrollo de aplicaciones para la detección de jóvenes talentos como «Sport Search» Zulyaden et al. (2022) o sistemas de apoyo a la búsqueda de jugadores que ocupen las vacantes libres de una plantilla de la forma más idónea mediante un exhaustivo análisis multifactorial como el que presenta Ghar et al. (2021).

Por su parte, el análisis del juego colectivo se centra en el estudio de movimientos y estrategias utilizadas por un equipo como unidad, lo que comúnmente se denomina «estilo de juego». La caracterización del estilo de juego implica un análisis detallado de cómo el equipo juega, sus patrones con la pelota y sin ella, estrategias utilizadas y fortalezas y debilidades dentro del césped. Esta información es crucial para la toma de decisiones informadas por parte de los entrenadores en cuanto a la alineación y la estrategia de juego a la hora de enfrentar un rival. En este sentido, es fundamental resaltar que el fútbol, debido a su gran versatilidad, ofrece una amplia variedad de estilos. Cada equipo cuenta con una combinación única de jugadores y cuerpo técnico. Como resultado, es posible observar miles de configuraciones tácticas distintas e incluso diferencias en el rendimiento de los propios jugadores según el enfoque táctico adoptado.

Antes de finalizar esta sección, es importante destacar el artículo de Rein and Memmert (2016) que indaga sobre el impacto del *Big Data* y el aprendizaje automático en el análisis táctico. En este artículo se aborda cómo, gracias al desarrollo de estas disciplinas, se han superado las limitaciones previas (principalmente relacionadas con la calidad de los datos), permitiendo avances significativos en la investigación. Además, se contextualiza el análisis táctico y se presentan diversos casos de uso, consolidando gran parte de la información existente sobre este tema y la influencia de las nuevas técnicas de aprendizaje automático en él.

2.3 Similitud entre patrones de juego

La caracterización del estilo de juego de un club es un elemento clave en el análisis táctico del fútbol de élite moderno. Este proceso implica el estudio detallado de la forma en que un equipo juega y se desenvuelve en el campo. Con esta información, se pretenden identificar aquellos atributos propios del estilo de juego de cada club que contribuyen a la generación de situaciones ventajosas para el mismo, tanto ofensiva como defensivamente.

La revisión de la caracterización en el fútbol es un tema relevante en la actualidad, ya que la tecnología y el análisis de datos están transformando la forma en que se estudia el juego. Con el fin de seguir desarrollando este campo es necesario evaluar la eficacia de los métodos actuales de caracterización y explorar nuevas técnicas y herramientas para mejorar la precisión y la utilidad de los análisis.

Las metodologías para abordar el problema de la caracterización del estilo de juego de un club que podemos encontrar en la literatura son bastante diversas. En este sentido, Goes et al. (2021) realiza una revisión de las diferentes metodologías aplicadas en artículos que proponen herramientas para el análisis táctico distinguiendo dos tipos de enfoques, la «ciencia deportiva» (eminentemente estadístico) y el enfoque «informático» (aplicando herramientas propias del *Big Data*). Se discuten las diferencias metodológicas entre ambas aproximaciones, desde el objetivo buscado hasta las herramientas más utilizadas. Finalmente, se propone la idea de realizar análisis que combinen ambos enfoques con la idea de combinar las ventajas que ofrece cada uno. A continuación, se procede a la descripción de algunas de las metodologías más comunes en la actualidad.

En primer lugar, Clemente et al. (2015a) apunta a la posibilidad de modelizar el estilo de juego de un equipo como una red donde cada nodo representa a cada uno de los jugadores que intervienen en una acción ofensiva y cada arista representa un pase entre dos jugadores. A partir de esta modelización, se calcula la matriz de adyacencia total mediante la suma de todas las redes creadas a lo largo del partido. Finalmente, se proponen tres métricas (densidad, centralidad y heterogeneidad) que pueden ser calculadas a partir de la matriz de adyacencia y se proporciona el significado de los valores que puede tomar cada una de estas aplicadas al mundo del fútbol. Posteriormente, Clemente et al. (2015b) amplía este estudio incluyendo otras variables notacionales, como el ganador del partido o los goles marcados para buscar relación entre las métricas computadas y las variables relacionadas con el rendimiento del equipo en el partido.

En segundo lugar, Gyarmati et al. (2014) apunta a la posibilidad de hacer un estudio del estilo de juego de un club a partir del análisis de los patrones de pases realizados durante las posesiones de este. De este modo, se especifican cinco posibles patrones y se cuantifica el número de veces que se observa cada uno en los diferentes equipos de La Liga, de forma que se identifique cada club con la frecuencia con la que desarrolla cada uno de estos patrones. Tras un clustering posterior, se pueden observar similitudes entre equipos que tienden a usar patrones de pases similares y cuáles son los que más usan.

Cabe destacar que, en el ámbito del scouting, Peña and Navarro (2015) y Barbosa et al. (2022) proponen una metodología muy similar, basada en los patrones de pases en la que el centro es el jugador y no el club. Es decir, se compara la frecuencia con la que diferentes jugadores aparecen en diferentes patrones de pases para buscar similitudes entre ellos.

Las dos aproximaciones anteriormente mencionadas comparten un punto de vista común, ambas centran su análisis en los pases ejecutados. Es decir, el análisis del estilo de juego se centra en la forma de mover el balón de los diferentes equipos, obviando en su gran mayoría el resto de las acciones del juego.

Cuando hablamos de estilos de juego, existen otras variables como el número de disparos a puerta o faltas cometidas que otorgan una visión más detallada del juego de un equipo y que quedan relegadas a un papel secundario o incluso nulo en este tipo de modelizaciones. Además, la faceta defensiva queda totalmente diluida en este tipo de estudios, ya que únicamente tiene en consideración aquellos momentos del partido en los que el equipo en cuestión tiene en posesión la pelota.

Sin embargo, hoy en día empresas como Wyscout juegan un papel vital en el proceso de análisis técnicos para los clubes ya que se encargan de la recolección y puesta a punto de los diferentes datos recogidos en cada partido disputado. La información contenida en este tipo de datos abarca todas las acciones que se llevan a cabo durante un partido (llamadas «eventos»), incluyendo las coordenadas x e y de donde ocurre cada acción. El formato de datos de eventos permite a los clubes hacer análisis mucho más exhaustivos, tanto de los clubes rivales como del propio.

En este sentido, Decroos et al. (2018) realiza un análisis sumamente exhaustivo, estudiando diversos eventos de partido (pases, tiros, recuperaciones, etc.) para desarrollar una aplicación que pretende dar soporte a las decisiones de los cuerpos técnicos de los clubes. Tras dividir cada encuentro en fases (algo similar a lo que comúnmente se denominan «posesiones» en la jerga futbolística), se procede a la agrupación de estas fases por similaridad mediante *clustering*, para finalmente asignar una importancia a cada uno de estos clústeres. Una vez completado todo ese proceso se procede a la aplicación de *pattern mining* y *pattern ranking* sobre cada una de las fases de los clústeres más importantes para identificar qué sucesiones de eventos son las más comunes en un equipo durante un partido. Cabe destacar que, a partir de las coordenadas de cada uno de los eventos, resulta particularmente sencillo graficar las jugadas más características sobre un terreno de juego del club sobre el que se realice el análisis. No obstante, este tipo de metodologías requieren del apoyo de expertos en la materia que den soporte a la toma de decisiones arbitrarias, como puede ser la asignación de importancia a los diferentes clústeres. Además, este enfoque mucho más típico del campo de la informática Goes et al. (2021) conlleva el uso de herramientas computacionalmente más costosas.

Finalmente, también cabe la posibilidad de contar con las coordenadas de cada jugador en el terreno de juego recogidas con cierta frecuencia (ej. 15 Hz), lo que se denomina como «datos de seguimiento». En este caso, el rastreo constante de los jugadores permite estudiar la formación del equipo y el comportamiento posicional de los jugadores dentro de ella. Merlin et al. (2020) identifica tres tipos de variables «notacional» (goles, tiros, etc.), «ocupación de espacio» y «sincronización de desplazamiento». El objetivo del estudio es agrupar aquellas posesiones que sigan patrones similares e identificar aquellas variables que permiten discriminarlas en mayor medida.

Por otro lado, Shaw and Glickman (2019) propone un sistema de identificación de formaciones dinámico mediante la medición de la posición relativa de los jugadores en el terreno de juego. La elección de la formación no es algo trivial, los entrenadores deben tener en cuenta factores muy diversos, desde las características de sus propios jugadores hasta el estilo de juego del rival. Este tipo de trabajos cobran aún más importancia hoy en

día donde las formaciones son mucho más fluidas, especialmente en los clubes de élite, donde la disciplina juega un papel más importante. A lo largo del tiempo de juego de un encuentro es común encontrar diferentes formaciones en un mismo equipo, de hecho es común que los equipos adopten diferentes formaciones para la fase ofensiva y para la defensiva. Asimismo, una formación «4-3-3» puede tener diferentes características (anchura de líneas de medio, posición de laterales, bloque bajo, bloque alto, etc) en función del equipo que la adopte.

Estos dos últimos formatos de datos presentados (eventos y seguimiento) son más propicios a ser analizados haciendo uso de herramientas propias del área del *Big Data*. Esto tiene dos principales implicaciones:

En primer lugar, la combinación de diferentes tipos de datos podría dar lugar a enfoques combinados desde la perspectiva clásica de la ciencia del deporte y la informática Goes et al. (2021) para proponer estudios que realmente tengan interés práctico en el mundo del fútbol.
En segundo lugar, y por contra, que este tipo de datos requieren lidiar con problemas clásicos del manejo de grandes cantidades de datos Sagiroglu and Sinanc (2013), el almacenamiento y procesado de los mismos.

Resulta complicado hallar en la literatura investigaciones que complementen la información obtenida del estudio de diferentes formatos de datos para el análisis táctico. Tanto los datos de eventos, como los de seguimiento, a pesar de ser cada vez más accesibles, tienden a ser muy costosos, lo cual dificulta la realización de estudios de esta índole mediante el empleo de datos de acceso público. La combinación de múltiples formatos y disciplinas, por ende, se convierte en una labor en muchos casos inviable debido a la gran cantidad de inconvenientes que se presentan. Esto se traduce en una gran cantidad de artículos que no logran cumplir con el objetivo de ofrecer información realmente útil en la práctica para los cuerpos técnicos de los clubes.

2.4 Análisis clustering

A lo largo de la revisión bibliográfica se han expuesto algunos artículos como el de Decroos et al. (2018) que hacen uso de clustering para agrupar, en este caso, fases del juego por similaridad. Cuando se habla de buscar agrupaciones o patrones en los datos, el análisis de clustering es la herramienta más fundamental. Esta técnica permite identificar relaciones y similitudes entre elementos en un conjunto de datos, agrupándolos en clústeres que comparten características comunes. Al utilizar algoritmos matemáticos y estadísticos, el análisis de clustering busca revelar estructuras ocultas, lo que puede conducir a una comprensión más profunda de la naturaleza de los datos y brindar información valiosa para la toma de decisiones.

La historia del clustering se remonta a los albores del siglo XX, cuando pioneros en campos como la psicología y la biología comenzaron a organizar datos en grupos según similitudes. Sin embargo, fue en la década de 1950 cuando el término *clustering* se consolidó y se buscaron enfoques más sistemáticos. Durante los años 1960, figuras como J. A. Hartigan y E. W. Forgy desarrollaron algoritmos esenciales, como K-Means (Hartigan and Wong, 1979), que sentaron las bases para técnicas de agrupamiento.

A medida que la informática y las capacidades de procesamiento mejoraron, la década de 1980 marcó una explosión en la diversidad de algoritmos de clustering. El campo continuó evolucionando con la integración de enfoques de aprendizaje automático y análisis de redes, permitiendo una exploración más profunda de datos complejos y la detección de patrones sutiles. Hoy en día, esta técnica es ampliamente utilizada en diversas

disciplinas, desde marketing y biología hasta análisis de imágenes y exploración de datos masivos.

En la actualidad existen diversos tipos de técnicas de clustering:

Clustering Jerárquico: En este enfoque, se construye una jerarquía de clústeres, donde los clústeres se agrupan o dividen en función de la similitud entre los datos. Pueden ser aglomerativos (comenzando con datos individuales y fusionándolos en clústeres más grandes) o divisivos (comenzando con todos los datos en un clúster y dividiéndolos en clústeres más pequeños).

Clustering de Partición: En este enfoque, los datos se dividen directamente en un número predefinido de clústeres sin construir una estructura jerárquica. Uno de los ejemplos más comunes de clustering de partición es el algoritmo K-Means, que divide los datos en k clústeres.

Métodos híbridos: En este enfoque, se combinan diferentes técnicas.

Clustering difuso *(fuzzy)*: Es un enfoque en el análisis de datos donde los puntos se asignan a clústeres con grados de membresía en lugar de asignaciones binarias «pertenece/no pertenece» en el clustering rígido. En el *fuzzy* clustering, cada punto tiene una distribución de pertenencia a todos los clústeres, lo que refleja la incertidumbre o ambigüedad en su asignación.

Clustering basado en mixturas: Este enfoque asume que los datos provienen de una combinación (mezcla) de múltiples distribuciones de probabilidad. Cada una de estas distribuciones representa un posible clúster en los datos. El proceso de clustering basado en mixturas generalmente se realiza utilizando el algoritmo *Expectation-Maximization* (EM) (Dempster et al., 1977), que busca encontrar los parámetros de las distribuciones que mejor describen los datos observados y las asignaciones probabilísticas de puntos a clústeres.

Clustering basado en densidad: Este enfoque busca agrupar datos en función de su densidad de distribución en el espacio de características. Los puntos que están cerca entre sí y tienen una densidad suficientemente alta se agrupan en un clúster, mientras que las áreas de baja densidad actúan como separadores naturales entre los clústeres.

2.4.1. K-MEANS

K-Means posiblemente se trate del algoritmo de clustering más popular. MacQueen et al. (1967) fue el primer autor en hacer mención al término K-Means, a pesar de que el algoritmo para el cómputo de los clusters fue propuesto por Stuart LLoyd en 1957. Más tarde Hartigan and Wong (1979) propondría una versión más eficiente del algoritmo.

Algoritmo de LLoyd

1. Inicializar K centroides $C_1, C_2, \ldots, C_K$ de forma aleatoria o utilizando algún método de inicialización.

2. Asignar cada punto del conjunto de datos al centroide más cercano.

3. Repetir los siguientes pasos hasta que los centroides converjan o se alcance un número máximo de iteraciones:

a) Para cada clúster $i = 1$ hasta K, calcular el nuevo centroide C_i como el promedio de los puntos asignados al clúster i.

b) Para cada punto j en el conjunto de datos, asignar j al centroide más cercano.

El objetivo fundamental de K-Means es la minimización de la suma de cuadrados intracluster (SSQ) (Ward Jr, 1963). El objetivo es agrupar los datos en K clústeres de manera que la variabilidad interna dentro de cada clúster sea mínima. Esta medida se logra al minimizar la suma de los cuadrados de las distancias Euclidianas entre cada punto x y el centroide μ_i del clúster C_i al que pertenece. La función objetivo se expresa como:

$$\min_{C_1, C_2, \ldots, C_K} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2.1}$$

Donde $C_1, C_2, \ldots, C_K$ representan los K clústeres y μ_i es el centroide del clúster C_i. La función objetivo busca asignar cada punto x al clúster cuyo centroide μ_i minimiza la distancia Euclidiana al cuadrado entre x y μ_i. El objetivo es encontrar la disposición de clústeres que minimice la variabilidad interna y maximice la coherencia dentro de cada clúster.

En la ecuación, el término $\|x - \mu_i\|^2$ representa la distancia Euclídea al cuadrado entre el punto x y el centroide μ_i. El algoritmo K-Means itera a través de la asignación de puntos a clústeres y la actualización de los centroides para alcanzar una configuración que minimice la suma de cuadrados. Esto conduce a la formación de clústeres donde los puntos dentro de cada clúster son cercanos entre sí y distantes de los puntos de otros clústeres.

2.4.2. HDBSCAN

Un ejemplo popular de clustering basado en densidad es el algoritmo DBSCAN (Ester et al., 1996) (Density-Based Spatial Clustering of Applications with Noise). DBSCAN no requiere especificar el número de clústeres de antemano. En cambio, se basa en la densidad de puntos: un clúster se forma cuando un número mínimo de puntos está cerca unos de otros, y los puntos aislados se consideran ruido.

Sin embargo, en 2013 Campello et al. (2013) propuso una modificación de DBSCAN, HDBSCAN. HDBSCAN se diseñó para abordar una limitación de DBSCAN, que es la dificultad para manejar clústeres de diferentes densidades. HDBSCAN construye una estructura jerárquica de clústeres utilizando el concepto de *condensation tree*, lo que permite identificar clústeres de diferentes escalas y densidades. Además, el algoritmo ofrece una forma de determinar automáticamente el número de clústeres significativos y reduce la sensibilidad a la elección de parámetros.

2.4.3. Comparación de técnicas

Como se puede apreciar, las técnicas de clustering se encuentran en constante evolución y adaptación para enfrentar los desafíos cambiantes en el análisis de datos. A medida que surgen nuevos tipos de datos, se generan problemas más complejos y se desarrollan tecnologías avanzadas, las técnicas de clustering evolucionan para brindar soluciones más precisas y versátiles.

En este contexto, podemos encontrar varias comparaciones en la literatura entre diversas técnicas de clustering. A fin de mostrar el rendimiento que ofrece cada una en un *dataset* académico como el archiconocido *iris*, Kanagala and Krishnaiah (2016) propone

una comparativa entre los clusters creados por K-Means, DBSCAN y OPTICS (Ordering Points To Identify the Clustering Structure), otro tipo de clustering basado en densidad. Se trata de una comparación muy sencilla; no obstante, suficiente para mostrar los puntos flacos de cada una de las técnicas. En el caso de K-Means, la necesidad de presuponer un número de clústeres en los datos implica que la elección del parámetro K puede ser un desafío crítico. Una elección incorrecta de K puede llevar a una segmentación inadecuada de los datos y a la interpretación errónea de los resultados. Además, K-Means no es una técnica capaz de detectar valores anómalos, al contrario que DBSCAN. DBSCAN ofrece una mayor versatilidad, pues presenta la capacidad de crear clusters de diversas formas y tamaños, además permite clasificar valores atípicos y ruido. Sin embargo, cuando los clusters se encuentran demasiado cercanos entre sí, DBSCAN muestra no ser capaz de realizar una diferenciación clara. OPTICS, por su parte, sí es capaz de garantizar agrupaciones de calidad manteniendo el orden en que se procesan los objetos de datos. Aun así es necesario destacar que ambos métodos deben ser debidamente parametrizados para extraer resultados óptimos.

Por otra parte, cuando se trata con *datasets* reales y, especialmente, de gran tamaño, es común encontrar que las divisiones entre los grupos son muy difusas e incluso nulas. En problemas donde se encuentran una gran variedad de individuos la tarea de agruparlos se complica especialmente. Es por este motivo que en la actualidad se vienen desarrollando métodos de incrustación *embeddings*, capaces de preservar la estructura local y, al mismo tiempo, revelar la estructura global de los datos. En este sentido podemos encontrar artículos como el de Allaoui et al. (2020) donde no sólo se comparan diferentes técnicas de clustering (K-Means, HDBSCAN, GMM y Clustering aglomerativo), sino que se compara el propio rendimiento de las técnicas con el rendimiento que ofrecen cuando previamente se aplica UMAP. UMAP (Uniform Manifold Approximation and Projection) se define como «una novedosa técnica de aprendizaje de múltiples para la reducción de dimensiones. UMAP se construye a partir de un marco teórico basado en la geometría de Riemann y la topología algebraica» según se explica en el artículo original (McInnes et al., 2018).

Volviendo al artículo anterior, la previa aplicación de UMAP sobre varios conjuntos de datos de imágenes permite a cada uno de los algoritmos de clustering estudiados mejorar su rendimiento, llegando a alcanzar una mejora de hasta el 60 % en el *Accuracy* computado. Por tanto, UMAP es capaz de mejorar notablemente la eficacia de las diferentes técnicas de clustering, especialmente en espacios de muy alta dimensionalidad. A su vez, esta reducción de la dimensionalidad permite una visualización efectiva de los datos y sus posibles agrupaciones.

2.4.4. Métricas de calidad

Hasta ahora, únicamente se ha hecho referencia a la necesidad de extraer «agrupaciones de calidad», pero para medir la calidad de una agrupación será necesario, por un lado especificar qué se entiende por calidad en las agrupaciones creadas y qué métricas son las más comunes en el contexto del clustering.

Cuando se habla de calidad de las agrupaciones, usualmente se tienen en cuenta dos características a maximizar:

- Compacidad intraclúster: es deseable que los clusters sean homogéneos, es decir, los individuos dentro de un clúster deben ser lo más similares posible en términos de características.

- Separación interclúster: a su vez, es deseable que el conjunto de clusters sea heterogéneo, es decir, cada clúster debe ser lo más diferente del resto posible.

En adición, cuando hablamos de las técnicas utilizadas para el clustering, idealmente los resultados deben ser consistentes en diferentes ejecuciones o en diferentes subconjuntos de los datos, lo que se denomina estabilidad.

Existen una vasta cantidad de métricas en la literatura para medir la calidad de las agrupaciones de muy diversos modos.

- **Coeficiente de silhouette**
  - Para cada punto i en el conjunto de datos:
    - $a(i)$ es la distancia promedio entre el punto i y los otros puntos en el mismo clúster.
    - $b(i)$ es la distancia promedio más pequeña entre el punto i y los puntos en otros clústeres.
    - El coeficiente de silhouette para el punto i se calcula como:

$$S(i) = \frac{b(i) - a(i)}{\text{máx}\{a(i), b(i)\}}$$

    - El denominador $\text{máx}\{a(i), b(i)\}$ asegura que el valor del coeficiente esté en el rango de -1 a 1.
    - Un valor cercano a 1 indica que el punto está bien asignado a su clúster y lejos de otros clústeres, mientras que cerca de -1 indica una mala asignación.
    - El coeficiente de silhouette se puede calcular para cada punto y luego se puede promediar para obtener una medida global de la calidad del clustering.
- **Índice de Davies-Bouldin**
  - Se calcula para cada clúster C_i en función de su distancia promedio a otros clústeres y la suma de los radios intraclúster.
  - Para cada clúster C_i:
    - Se calcula la distancia promedio R_i entre los puntos en C_i y el punto central del clúster.
    - Se calcula la suma de los radios intraclúster $R(C_i)$, que es la distancia máxima entre cualquier par de puntos en C_i.
    - Para cada otro clúster C_j:
      - Se calcula la distancia promedio R_j entre los puntos en C_j y el punto central del clúster.
    - El índice de Davies-Bouldin para el clúster C_i se calcula como:

$$DB(C_i) = \frac{R_i + R(C_i)}{R_j}$$

    - Se elige el clúster C_j con el valor más alto de $R_i + R(C_i)$, reflejando la similitud mínima entre C_i y otros clústeres.
    - El índice de Davies-Bouldin global es el promedio de los índices de todos los clústeres, evaluando la calidad general del clustering.

  - Un índice de Davies-Bouldin más bajo indica mejor calidad, con clústeres más separados y coherentes.

- **Índice de Calinski-Harabasz**
  - Se calcula considerando la dispersión entre clústeres (B) y la dispersión dentro de los clústeres (W).
  - La fórmula del índice CH es:
$$CH = \frac{B}{W} \times \frac{N-K}{K-1}$$
  - Donde N es el número total de puntos y K es el número de clústeres.
  - Un valor más alto de CH indica clústeres más separados y compactos, considerando una mejor calidad de clustering.

- **Índice Dunn**
  - Evalúa la relación entre la distancia mínima entre clústeres y la máxima distancia intraclúster.
  - Para calcular el índice Dunn (D), se considera la distancia mínima entre clústeres C_i y C_j dividida por la máxima distancia dentro de cualquier clúster C_k.
  - La fórmula del índice Dunn es:
$$D = \min_{1 \leq i \leq K} \left(\min_{j \neq i} \left(\frac{d_{\text{mín}}(C_i, C_j)}{\text{máx}_{1 \leq k \leq K} d_{\text{máx}}(C_k)} \right) \right)$$
  - Donde C_i es el clúster i, $d_{\text{mín}}(C_i, C_j)$ es la distancia mínima entre clústeres y $d_{\text{máx}}(C_k)$ es la máxima distancia dentro del clúster C_k.
  - Un valor más alto de Dunn indica mejor calidad de clustering, con clústeres más separados y compactos.

Todo lo anteriormente mencionado se tratan de métricas de información interna, basadas en las características internas del clustering, como la cohesión y la separación dentro y entre clústeres. No requieren una partición de referencia y se centran en la estructura de los propios datos. Sin embargo, si se posee información previa sobre las particiones que se desean realizar, en la literatura existen otros índices como el **Índice de Jaccard**, que calcula la relación entre los elementos que están en el mismo clúster en ambas particiones y la cantidad total de elementos:

- **Índice de Jaccard**
  - El índice de Jaccard se utiliza para medir la similitud entre dos clústeres C_i y C_j (en el contexto del clustering).
  - Se calcula como la proporción de puntos que están en ambos clústeres sobre el total de puntos que están en al menos uno de los clústeres.
  - La fórmula del índice de Jaccard es:
$$J(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$
  - Donde $|C_i \cap C_j|$ es el número de puntos en común entre C_i y C_j, y $|C_i \cup C_j|$ es el número de puntos en total en los dos clústeres.
  - El índice de Jaccard proporciona una medida de similitud entre clústeres, donde 1 indica que los clústeres son idénticos y 0 indica que no comparten ningún punto.

CAPÍTULO 3
Herramientas utilizadas y análisis exploratorio previo

En este capítulo se introducirán las características del *dataset* utilizado para el desarrollo del proyecto, la colección de datos de Wyscout para la temporada 2017-18. Posteriormente se procederá con el detallado de las herramientas utilizadas para la manipulación de estos datos. Finalmente, se indagará más a fondo en los datos mediante una primera exploración de estos a fin de explicar con detalle su naturaleza y forma.

3.1 Conjunto de datos

Como ya se mencionó anteriormente, actualmente, los datos futbolísticos más completos son propiedad de empresas especializadas en el ámbito, y por tanto, es difícil encontrar *datasets* de dominio público para la investigación científica. Sin embargo, en los artículos Pappalardo et al. (2019b) y Pappalardo et al. (2019a) se publica una completísima colección de eventos de partidos disputados durante la temporada 2017-2018. En el propio artículo se detalla más a fondo cualquier tipo de detalle sobre la colección; sin embargo, podemos encontrar un pequeño resumen del autor: «proporcionamos al público la colección de logs de fútbol más grande jamás liberada, recopilada por Wyscout (https://wyscout.com/), que contiene todos los eventos espaciotemporales (pases, tiros, faltas, etc.) que ocurren durante todos los partidos de una temporada completa de siete competiciones (La Liga, Serie A, Bundesliga, Premier League, Ligue 1, Copa Mundial de la FIFA 2018, Eurocopa de la UEFA 2016). Un evento de partido contiene información sobre su posición, tiempo, resultado, jugador y características. Este conjunto de datos ha sido utilizado recientemente durante el Soccer Data Challenge y, hasta donde sabemos, es la mayor colección pública de logs de fútbol»(Pappalardo and Massucco, 2019).

Esta colección contempla 1941 partidos, 3251294 eventos y 4299 jugadores distribuidos en los siguientes ficheros de datos:

Competiciones: Siete ficheros correspondientes a cada una de las cinco grandes ligas europeas (mencionadas anteriormente), el Mundial 2018 y la Eurocopa 2016.

Partidos: El *dataset* de los partidos incluye información de todos los partidos de cada una de las competiciones anteriores.

Equipos: El *dataset* de los equipos incluye información de todos los equipos (incluyendo selecciones) que participan en cada una de las competiciones anteriores.

Fichero *JSON*	Campos utilizados
events_Spain	eventID, eventName, subEventName, positions, playerID, matchID, matchPeriod, teamID, eventSec, subEventID, id
events_England	
events_Italy	
events_France	
events_Germany	
teams	wyId, name, officialName, area[name]
players	wyId, firstName, middleName, lastName, currentTeamId
matches_Spain	teamsData
matches_England	
matches_Italy	
matches_France	
matches_Germany	

Tabla 3.1: Campos utilizados en los ficheros *JSON*

Jugadores: El *dataset* de jugadores proporciona información sobre cada uno de los jugadores de cada una de las plantillas de todos los equipos contemplados.

Eventos: Este es el principal *dataset*, que contiene toda la información de todos los eventos registrados en cada uno de los partidos de las competiciones contempladas.

Entrenadores: Este *dataset* proporciona información sobre los entrenadores de cada uno de los equipos.

Árbitros: Este *dataset* incluye información sobre cada uno de los árbitros encargados de dirigir cada uno de los partidos.

Esta colección de datos será la utilizada para llevar a cabo el desarrollo de este proyecto.

Cabe recordar que en el artículo original Pappalardo et al. (2019b) proporciona información más precisa acerca de los campos encontrados dentro de cada fichero. Sin embargo, y puesto que no todos los *datasets* son interesantes para este trabajo, en la tabla 3.1 se exponen de forma clara los ficheros de los que se ha decidido finalmente hacer uso junto con los campos de interés correspondientes. Estos campos serán explicados de forma detallada más adelante en la sección 3.3.

3.2 Herramientas utilizadas

Tras la elección del *dataset*, se describirán detalladamente las herramientas empleadas, se explicará su papel en el desarrollo del proyecto y se justificará su elección en función de los requisitos específicos del trabajo. Se proporcionará una visión general de las capacidades y el valor que aportan estas herramientas al abordar las problemáticas planteadas y alcanzar los objetivos establecidos.

En este sentido, ha sido Python la principal herramienta en este estudio, debido a su versatilidad y potencia en el ámbito del análisis de datos y la programación científica. Con su amplia variedad de bibliotecas, este lenguaje de programación proporciona las capacidades necesarias para la manipulación eficiente de datos, cálculos numéricos, visualización de resultados y desarrollo de algoritmos complejos. Además, Python se ha convertido en los últimos años en un lenguaje ampliamente utilizado tanto en el ámbito

académico como en el laboral, lo que lo convierte en un lenguaje ideal para la comunicación de resultados y la replicabilidad de este estudio. La elección de Python como la herramienta principal asegura una base sólida y flexible para abordar las tareas críticas en el estudio, permitiendo una exploración profunda de los datos, la implementación de modelos, y la presentación efectiva de resultados. Además, Python es un proyecto *open source*, por lo que la replicabilidad de este estudio no estará sujeta a ningún tipo de costo o licencia. En resumen, el uso de un único lenguaje de programación facilita las tareas de integración y replicabilidad, en base a una estructura sólida pero versátil como la que ofrece el uso de este lenguaje de programación para proyectos de Ciencia de datos.

3.2.1. Librerías utilizadas

A continuación, se detallarán más a fondo todas y cada una de las librerías de las que se ha hecho uso para la realización del estudio.

- **JSON:** es pertinente recordar que los datos de la colección están codificados en formato *JSON*. La librería *JSON* en Python es un módulo incorporado que permite trabajar con datos en este formato de una forma sencilla.
- **pandas:** una biblioteca de código abierto en Python que proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento y fáciles de usar. Esta librería permite transformar los campos de los ficheros *JSON* a formato tabular donde su manipulación será mucho más conveniente.
- **NumPy:** una biblioteca para el cálculo numérico y la manipulación de matrices y arreglos multidimensionales.
- **Matplotlib:** una biblioteca para la visualización de datos.
- **mplsoccer:** una biblioteca Python que se especializa en la visualización de datos relacionados con el fútbol utilizando Matplotlib, una popular librería de trazado y visualización en Python. La finalidad principal de mplsoccer es facilitar la creación de gráficos y visualizaciones específicas para analizar y representar aspectos tácticos y estadísticas del fútbol.
- **scikit-learn:** una biblioteca de aprendizaje automático en Python, en este caso, se ha utilizado para el clustering y el cálculo de métricas.
- **HDBSCAN:** una biblioteca de Python que implementa el algoritmo de clustering Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN).
- **umap-learn:** una biblioteca de Python que implementa el algoritmo UMAP (Uniform Manifold Approximation and Projection) para la reducción de dimensionalidad no lineal y la visualización de datos en un espacio de menor dimensión.
- **optuna:** una biblioteca de optimización de hiperparámetros para *machine learning* y ciencias de datos. Automatiza la búsqueda de los mejores hiperparámetros mediante técnicas como optimización bayesiana y estudios de hiperparámetros.
- **seaborn:** una biblioteca de visualización de datos basada en Matplotlib que proporciona una interfaz de alto nivel para crear gráficos atractivos y informativos.
- **networkx:** una biblioteca de Python utilizada para la creación, manipulación y estudio de la estructura, dinámica y funciones de redes complejas.

- **statsmodel:** una biblioteca de Python que proporciona clases y funciones para la estimación de muchos modelos estadísticos diferentes, así como para realizar pruebas estadísticas y explorar datos.

- **random:** un módulo incorporado en Python que proporciona funciones para la generación de números aleatorios.

- **scipy:** una biblioteca de Python utilizada para matemáticas, ciencia e ingeniería. Proporciona funcionalidades para optimización, estadísticas, procesamiento de señales y más.

3.2.2. Funcionamiento de Optuna

Vale la pena mencionar que esta librería es fundamental para todo el proceso experimental llevado a cabo en este trabajo. Optuna automatiza el proceso de optimización de modelos al buscar la combinación óptima de hiperparámetros en función de los resultados de las métricas de rendimiento. El funcionamiento interno de Optuna se basa en la optimización bayesiana, que utiliza probabilidades y distribuciones para explorar de manera eficiente el espacio de hiperparámetros. He aquí un pequeño esquema donde se detalla cómo funciona:

1. Definición del Espacio de Búsqueda: Antes de comenzar la optimización, se define el espacio de búsqueda de hiperparámetros. Esto incluye especificar las distribuciones para cada hiperparámetro, como uniforme, loguniforme o discreta.

2. Creación de un estudio: Un «estudio» en Optuna es una colección de experimentos de optimización. Cada experimento representa una combinación de hiperparámetros y su resultado asociado. Optuna busca encontrar la mejor combinación de hiperparámetros al minimizar o maximizar la métrica de rendimiento específica.

3. Selección de pruebas: Optuna selecciona una combinación de hiperparámetros para probar en función de un algoritmo de selección. Estos hiperparámetros se utilizan para entrenar y validar el modelo en un conjunto de datos.

4. Evaluación del rendimiento: Después de entrenar y validar el modelo con los hiperparámetros seleccionados, se calcula la métrica de rendimiento (por ejemplo, precisión, AUC, pérdida) en los datos de validación. Este resultado se registra en el estudio.

5. Actualización del modelo probabilístico: Optuna utiliza un modelo probabilístico (generalmente *Gaussian Process*) para estimar la relación entre los hiperparámetros y las métricas de rendimiento. A medida que más experimentos se realizan, el modelo se actualiza para reflejar mejor las relaciones.

6. Selección de próximos hiperparámetros: Basándose en el modelo probabilístico, Optuna selecciona la próxima combinación de hiperparámetros para probar de manera inteligente. La elección se basa en el equilibrio entre explorar nuevas áreas y explotar áreas prometedoras.

7. Iteración y mejora: Optuna repite el proceso de selección, evaluación y actualización del modelo probabilístico varias veces. A medida que avanza, se enfoca cada vez más en las combinaciones de hiperparámetros que tienen el potencial de ser óptimas.

8. Resultados y mejor modelo: Una vez que se completa el estudio, Optuna proporciona los resultados finales, incluida la mejor combinación de hiperparámetros y su métrica de rendimiento asociada. Esto permite que los usuarios implementen el modelo con los hiperparámetros optimizados.

En caso de requerir información más precisa sobre algún aspecto no detallado en esta memoria, es posible encontrar información más específica en la web oficial de Optuna o, en su defecto en la documentación oficial.

3.2.3. Desarrollo de código

Por otro lado, para el desarrollo de los códigos y el análisis, se utilizará el entorno de *Jupyter Notebooks* dentro de *Visual Studio Code* (VSCode). Esta combinación proporciona una potente plataforma para la creación y ejecución de código en Python, junto con la capacidad de documentar y visualizar los resultados de manera interactiva.

Visual Studio Code es un popular editor de código que ofrece soporte para múltiples lenguajes de programación, complementos personalizables y una interfaz amigable para el desarrollo. La integración de Jupyter Notebooks en VSCode amplía esta funcionalidad, permitiendo combinar la creación de código con la capacidad de documentar y explicar cada paso del análisis de datos.

Mediante los Jupyter Notebooks en VSCode, es posible escribir y ejecutar fragmentos de código de manera interactiva, visualizar gráficos, incluir comentarios detallados utilizando *Markdown* y, al mismo tiempo, mantener un registro ordenado de los pasos realizados en el análisis. Esta combinación de funcionalidades facilita la colaboración, el seguimiento de cambios y la presentación de resultados, lo que es especialmente valioso para proyectos de análisis de datos y programación científica. La elección de Jupyter Notebooks dentro de Visual Studio Code como el entorno de desarrollo garantiza una experiencia integrada y eficiente para la creación, análisis y documentación de código. El objetivo final de todo esto será la facilitación al máximo posible de la comprensión y el seguimiento de todo el proceso de desarrollo en el proyecto.

Todo el código desarrollado junto puede ser encontrado en el siguiente enlace. También se adjuntan otros archivos de interés como los ficheros de datos utilizados.

https://drive.google.com/file/d/1BjGbaLYixxDvkUlGvGOL7SGgwUjwwlP4/view?usp=drive_link

3.3 Análisis exploratorio

Finalmente, tras la exposición de los datos y herramientas utilizadas para el estudio se procede con el análisis exploratorio. Este análisis permitirá ahondar en los datos y mejorar la comprensión de estos, se comenzará por la presentación de la estructura de los ficheros (previamente expuestos) que han sido utilizados. El fichero de jugadores únicamente ha sido usado a con fines identificativos y, por tanto, no se tendrá en cuenta.

- **Fichero de eventos:** Este fichero es el eje central de todo el trabajo puesto que contiene toda la información de los partidos objeto de estudio. Debido a su importancia, se extraerá toda la información posible (incluyendo todos los identificadores).
- **Fichero de partidos:** Este fichero permite extraer el número total de partidos disputados de cada equipo (a fin de realizar una comprobación de completitud de los datos).

- **Fichero de equipos:** Este fichero permite asociar el id de cada equipo con su nombre oficial a fin de poder ser identificado, además se ha extraído la región (liga) a la que pertenece cada uno.

El primer paso para comenzar con la exploración de los datos es buscar información sobre los mismos y comprenderlos. En este sentido, Pappalardo et al. (2019b) expone en el propio artículo donde publica la colección de *datasets* una pequeña muestra informativa de los eventos dividida en tres dimensiones. En primer lugar, realiza una pequeña exploración a nivel técnico, donde observamos que los partidos están compuestos de media por 1682 eventos, separados por una media de 3.59 segundos. Además, en la figura 3.1 se observa cómo los pases y los duelos son los eventos que se observan con mayor frecuencia.

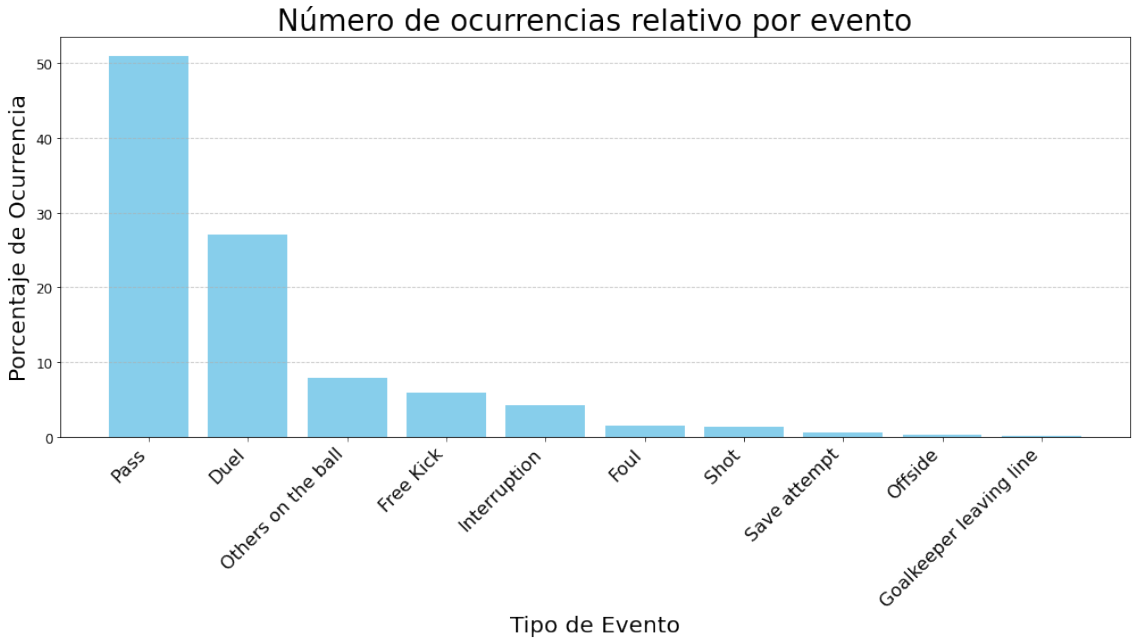

Figura 3.1: Frecuencia relativa de registro de cada tipo de evento.

Finalmente, también se incluye un gráfico (figura 3.2) donde se visualizan todos los eventos registrados según su geolocalización para un partido de ejemplo, en este caso "Lazio - Internazionale" de la primera división italiana (20 de mayo de 2018).

En segundo lugar, realiza un pequeño estudio descriptivo a nivel espacial, donde muestra para diferentes eventos, dónde se producen con mayor frecuencia dentro del terreno de juego mediante un mapa de calor. Y, en tercer lugar, a nivel temporal, expone la distribución de frecuencias de los registros de diferentes tipos de eventos en función del tiempo del partido, destacando la división entre primer y segundo tiempo. Además, el autor también realiza dos pequeños análisis a nivel de equipo y jugador, que no se detallarán aquí.

Este trabajo del autor original nos permite extraer información muy genérica acerca de los eventos y comprenderlos mejor. Sin embargo, y, a pesar de ser la base para este estudio, este tipo de análisis resulta incompleto para este trabajo. Por tanto, se ha decidido realizar una ampliación de este análisis poniendo enfoque en aquellos detalles que pudiesen suponer un obstáculo para el desarrollo del trabajo.

En esta línea, se comenzará por el estudio de los datos faltantes en el *dataset* de eventos. Como cabe esperar, los datos faltantes no serán un problema, ya que, como se men-

[1]Fuente: Pappalardo et al. (2019b) - Fig. 2 c)

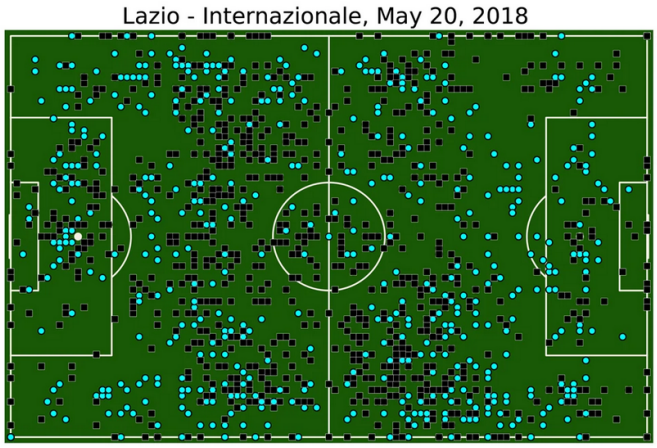

Figura 3.2: Eventos generados por los dos equipos en el partido Lazio (puntos cian) vs. Internazionale (puntos negros). Los eventos se representan en la posición del campo donde ocurrieron.
[1]

cionó anteriormente, el origen de los datos es una fuente altamente fiable, Wyscout. Sin embargo, encontramos que todos los eventos de tipo *fuera de juego* tienen a nulo los campos *subEventName* y *subEventId* ya que no existen subeventos para el evento fuera de juego y será necesario considerar esta condición para el estudio.

En segundo lugar, a pesar de contar con posición inicial y final para cada evento, no todos los eventos son susceptibles de contar con ambas posiciones. Por ejemplo, en el caso de los fuera de juego anteriores, únicamente se considera la posición del jugador que comete la infracción, la posición final en este caso simplemente se rellena con valores incoherentes como (0, 0) o (100, 100). Este es otro elemento a tener en cuenta, especialmente a la hora de visualizar las posesiones puesto que podría dar lugar a confusiones.

A continuación, es conveniente comprobar que en los ficheros de partidos todos los equipos han jugado el número correspondiente de partidos en función de la liga a la que pertenezcan (34 para los equipos de la Bundesliga, y 38 para el resto) a fin de asegurar que están presentes los datos de todos y cada uno de los partidos.

Para finalizar esta sección se procede con el estudio de los eventos, pero en esta ocasión enfocados en la agrupación por equipos. El objetivo del estudio es encontrar posibles diferencias entre equipos, para ello estudiaremos si, a simple vista, es posible encontrar algunas diferencias entre la distribución de diferentes eventos en cada equipo. Esto podría reforzar la hipótesis de la existencia de diferentes estilos de juego identificables entre los diferentes equipos objeto de estudio. No obstante, llevar a cabo un análisis exploratorio que abarque a todos los equipos sería poco factible. En su lugar, se optará por seleccionar a dos representantes de cada liga, lo que conformará una muestra total de 10 equipos para llevar a cabo este análisis. Los equipos escogidos de cada liga serán el primero y el segundo clasificado en sus respectivas ligas.

Antes de comenzar es necesario tener en cuenta que para realizar una comparación justa antes se ha de atender a uno de los detalles mencionados con anterioridad, a pesar de que en el resto de las ligas se disputen 38 jornadas, en la Bundesliga se disputan 34, por lo que es lógico pensar que en menos jornadas, los equipos alemanes contarán con menos eventos totales. Por tanto, el primer paso será normalizar el número de eventos por jornada disputada, de este modo se consigue una visión justa de la realidad a la hora de comparar el desempeño de cada uno de los 10 equipos.

LIGA	REPRESENTANTES
Bundesliga	Bayern de Múnich y Schalke 04
Serie A	Juventus y Napoli
La Liga	Barcelona y Atlético de Madrid
Premier League	Manchester City y Manchester United
Ligue 1	PSG y Mónaco

Tabla 3.2: Equipos representantes por liga.

3.3.1. Tiros

Cuando se atiende al número de tiros medio registrado por jornada de cada uno de los equipos se observan grandes diferencias entre los distintos equipos. Se puede observar como la diferencia entre el equipo que más veces patea al arco (Manchester City con 15.86 tiros por partido) y el que menos (Schalke 04 con 9.2 tiros por partido) asciende a 6.6 tiros. Es decir, el Manchester City dispara un 172 % más que el Schalke 04. Obviamente, existe un gran salto de calidad entre ambos equipos; sin embargo, este salto de calidad puede estar relacionado no sólo con la calidad de los jugadores sino con el estilo de juego que permite desplegar la plantilla. En este sentido, un equipo como el Atlético de Madrid logró quedar segundo en La Liga gracias a su estilo particularmente defensivo, llegando a ser el equipo menos goleado de la temporada. Hecho que también resalta en la figura 3.3 donde se observa cómo el Atlético de Madrid se coloca en penúltimo lugar en tiros realizados. Por otro lado, el mismo Manchester City arrasó en la Premier League, llegando a colocarse con 100 puntos y 106 goles a favor durante la competición liguera gracias a un estilo de juego totalmente diferente, mucho más ofensivo. Con este pequeño contexto se pretende dar a entender que diferentes estilos de juego pueden dar lugar a resultados muy similares en lo que a posición liguera se refiere.

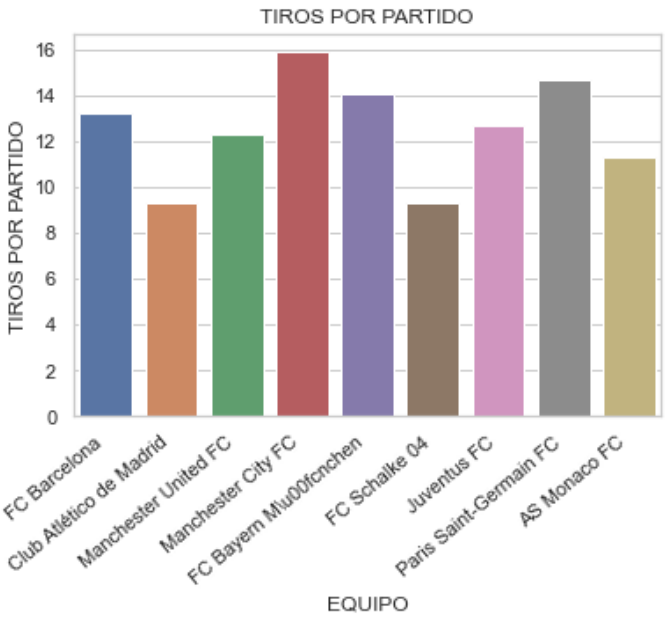

Figura 3.3: Tiros por partido de cada uno de los equipos representantes.

3.3.2. Pases

Por otro lado, cuando se pone el foco en los pases realizados (figura 3.4), llama claramente la atención que la distribución final resulta particularmente similar a la distribución observada en la figura 3.3 de los tiros registrados por partido. Aquellos equipos que fueron más agresivos también se vieron en la necesidad de gozar por más tiempo de la posesión del balón para poder elaborar su juego y crear las ocasiones necesarias para obtener un remate fiable. Por otro lado, los equipos más defensivos no necesitan de la posesión de la pelota para crear ocasiones, el juego directo les permite alcanzar la meta rival sin necesidad de preparar la jugada.

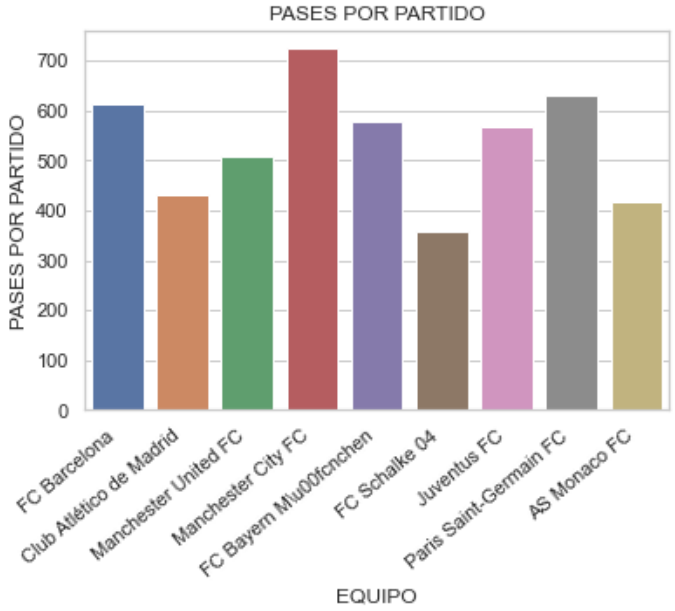

Figura 3.4: Pases por partido de cada uno de los equipos representantes

3.3.3. Duelos

Finalmente, en la figura 3.5 se muestra la distribución de duelos disputados de cada uno de los equipos. En esta ocasión sí que se presenta una distribución diferente, mucho más uniforme, pero en la que se observa un cambio radical. Los equipos que anteriormente se disponían últimos tanto en pases como en tiros por partido, en esta ocasión se encuentran a la cabeza. Cabe recordar que, con anterioridad, se mencionó que estos equipos despliegan un estilo de juego más defensivo, por lo que se ven envueltos en muchas más disputas por el balón, este tipo de juego requiere de más capacidad física y corpulencia. Sin embargo, los equipos más ofensivos intentarán evitar estos encuentros haciendo gala de su calidad con la pelota. Aun así, es preciso destacar que no encontramos una diferencia tan grande como la que se apreciaba en el estudio de los pases y los tiros, pues los partidos de élite son siempre muy intensos y disputados, por lo que las disputas por el balón son altamente frecuentes. Este hecho fue comprobado con anterioridad, cuando se realizó el estudio de la frecuencia de registro de cada uno de los diferentes eventos, donde los duelos eran los eventos más frecuentes, sólo por detrás de los pases.

Este primer acercamiento ha permitido entrar en contexto y extraer una primera conclusión: no todos los clubes presentan el mismo patrón de juego. Sin embargo, y a pesar de que esto parece obvio para cualquier entendido de la materia, existe una inmensa cantidad de estilos de juegos diferentes. Catalogar un equipo simplemente como ofensivo o defensivo es demasiado fundamental y poco informativo para el fútbol actual. El fin

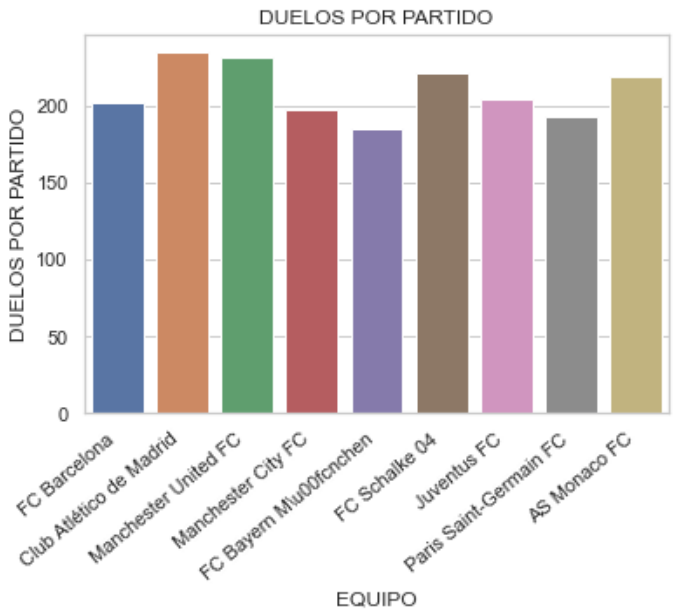

Figura 3.5: Duelos por partido de cada uno de los equipos representantes

de este trabajo será ahondar en estas diferencias y en cómo se desarrolla el estilo de cada club. Identificar qué variables marcan la diferencia, y cuáles deberían ser tenidas en cuenta a la hora de analizar un rival.

CAPÍTULO 4
Mi propuesta

Este capítulo se erige como la piedra angular de este proyecto, donde se desplegará una descripción detallada del enfoque, diseño y desarrollo de la solución. Se ha llevado a cabo un análisis exhaustivo de las necesidades y desafíos identificados, y se ha fundamentado en una combinación de metodologías comprobadas y tecnologías innovadoras con una visión global.

A lo largo de este capítulo, se desglosará cada componente de la solución, explicando su funcionamiento, sus interconexiones y cómo se relacionan con los objetivos establecidos. Se prestará especial atención a las innovaciones y ventajas competitivas que esta propuesta aporta al campo, destacando cómo aborda las limitaciones de soluciones anteriores y abre nuevas perspectivas.

4.1 Metodología

Dado que se requiere que el proceso sea claro y esté bien definido, se ha considerado conveniente presentar la figura 4.1, que detalla de forma esquemática cada uno de los procesos a seguir.

Este diagrama pretende brindar apoyo al lector en el seguimiento de la metodología propuesta en este trabajo. En él se detallan los procesos que se han seguido de forma secuencial hasta alcanzar los resultados. A continuación, se presenta un breve resumen:

1. Procesado del conjunto de datos original. Dado que el *dataset* original viene presentado en formato *JSON*, se ha decidido procesarlos y transformarlos en archivos *CSV*. Este formato tabular ofrece una mayor facilidad para el manejo de los datos. Además, en esta etapa se realiza una primera exploración de los datos.

2. Procesado de posesiones. En esta etapa se procesarán los eventos, agrupándolos por posesiones. Esto permitirá obtener el *dataset* de posesiones con el que identificar las posesiones más comunes y extraer algunas estadísticas de interés para cada equipo, como podría ser, por ejemplo, el número de faltas.

3. Clustering de posesiones. Una vez obtenido el *dataset* de posesiones, se preprocesará y se procederá con la experimentación y análisis de resultados.

4. Clustering de equipos. Para crear el vector de características de cada equipo se deberán unir los resultados del clustering de posesiones junto con las estadísticas deseadas para cada equipo y las métricas derivadas de la modelización de los propios equipos como grafos. Una vez todas las características han sido recogidas para cada

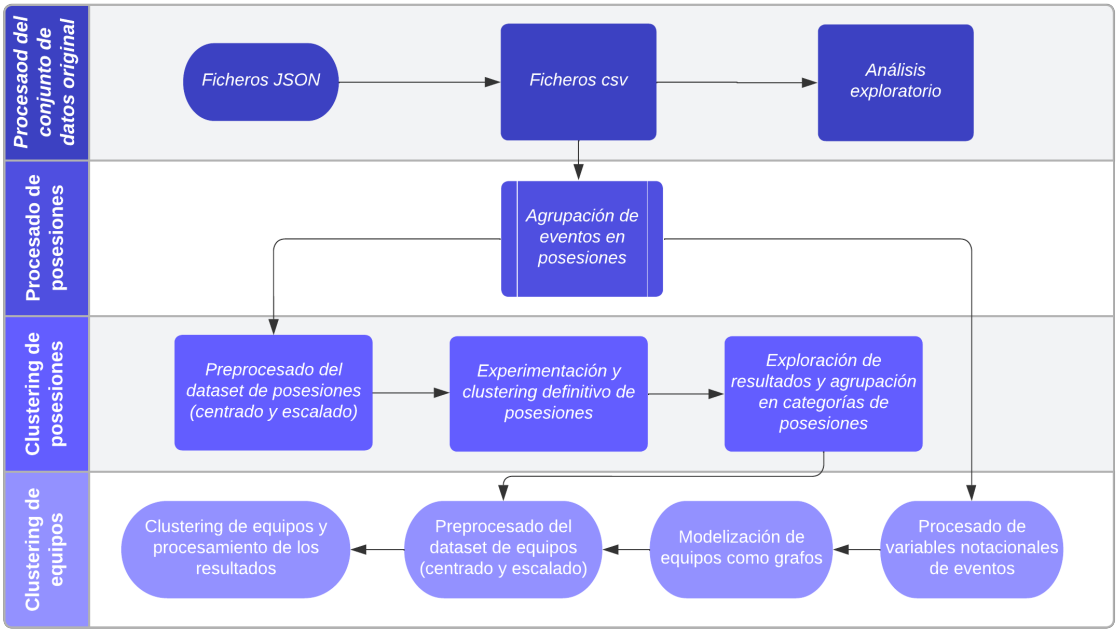

Figura 4.1: Diagrama de flujo de la metodología propuesta

equipo, se conforma del *dataset* de equipos. Este dataset deberá ser preprocesado al igual que el de las posesiones y posteriormente se aplicarán el clustering necesario para la obtención de resultados.

En las secciones subsiguientes, se desglosarán minuciosamente todos estos procesos, proporcionando una información amplia y detallada que posibilitará la replicación del procedimiento de manera precisa.

4.2 Preprocesado

Previo al inicio del estudio será necesario llevar a cabo un pequeño preprocesado de los datos que supondrá la puesta a punto de estos para su posterior manipulación. Este proceso ha resultado bastante simple y simplemente se detallarán los aspectos más importantes.

Comenzando con el preprocesado de los datos de eventos y, como ya se ha mencionado en las anteriores secciones, el primer paso será transformar el formato de original de los datos (*JSON*) a un formato tabular para facilitar el análisis. Para ello, se han cargado los datos de eventos de cada una de las 5 grandes ligas en formato *JSON* como objetos de Python, y, posteriormente, se han introducido los campos citados en la tabla 3.1 en un *dataframe* de la librería pandas para ser posteriormente exportados a *CSV*. Se ha tomado la decisión de no crear un fichero *CSV* con la concatenación de todos los eventos de todos los partidos; sin embargo, haber realizado este paso no supondría ningún tipo de inconveniente a fin de replicar el proyecto.

Por otro lado, el preprocesado de los ficheros de equipos y partidos ha sido realizado de manera muy similar, haciendo uso de la librería *JSON* para cargar los ficheros y posteriormente extrayendo a un *dataframe* los campos de interés de cada uno de los ficheros. Una vez realizado el primer paso, ya no se trabajará con ningún archivo en formato *JSON*.

4.3 División en posesiones

El primer elemento clave de este trabajo será la agrupación de eventos en lo que se denominarán «posesiones». Estas posesiones actúan como unidades fundamentales que permiten una segmentación lógica de las secuencias de acciones de un partido, es decir, se puede entender un encuentro entre dos equipos como una sucesión de posesiones. En función de los eventos registrados de cada posesión, será posible agruparlas por su similaridad, y, de este modo, identificar qué clubes son más propensos a usar cada uno de los diferentes tipos de posesión.

No obstante, previo al procesado de las procesiones será necesario implementar algunos cambios en el *dataframe* de eventos. En primer lugar, se añadirán tres variables derivadas de las coordenadas de los eventos. Estas variables permitirán categorizar las coordenadas de un evento en una zona dentro del campo, de este modo las las localizaciones se agruparán en una sola variable. Esta transformación permitirá la simplificación y la agregación de esta variable mediante la división imaginaria del campo en lo que se ha denominado «posiciones». Existen nueve posiciones diferentes, fruto de dividir en tres zonas el campo tanto a lo largo, como a lo ancho. Las divisiones imaginarias del terreno de juego en su eje más largo serán denominadas «zonas», identificando de esta manera tres zonas: defensa, medio y ataque. Por otro lado, las divisiones del terreno en su eje más corto serán denominados «carriles», comprendiendo izquierda, centro y derecha. Finalmente, las posiciones son creadas mediante las combinaciones de las variables zona y carril con los valores «defensa_izquierda, defensa_centro, defensa_derecha, medio_izquierda, medio_centro, medio_derecha, ataque_izquierda, ataque_centro, ataque_derecha», figura 4.2. La división se realiza por tercios de campo, es decir, el carril izquierdo ocupa desde la coordenada 0 hasta la 33 del eje y, el carril centro ocupa desde el 34 hasta el 66 y desde el 67 hasta el 100 será el carril derecho. De forma similar, la zona de defensa alcanzará hasta la coordenada 33 del eje x, la zona del medio desde la 34 hasta la 66 y la zona de ataque desde la 67 hasta la 100.

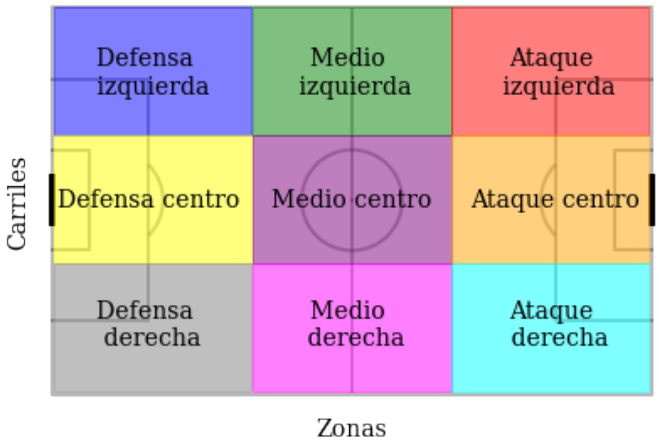

Figura 4.2: División en posiciones

Además, se introducirá una nueva variable *longXY* que recogerá la distancia euclídea entre emisor y receptor del evento en cuestión. Esta nueva característica permitirá estimar la longitud de los pases, el tipo de evento objetivo de esta variable.

$$longXY = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4.1}$$

Donde (x_1, y_1) representan las coordenadas de la posición inicial y (x_2, y_2) representan las coordenadas de la posición final. Por ejemplo, en un evento de pase serían las coordenadas de emisor y receptor. Cabe resaltar que no todos los eventos son susceptibles de presentar una posición inicial y otra final.

Volviendo al procesado de las posesiones, una vez creadas todas las variables auxiliares necesarias, se procederá con la división de los datos de eventos de Wyscout en posesiones. Sin embargo, es necesario definir previamente lo que se considerará una posesión para este trabajo. Una posesión será aquella sucesión de eventos de duración indefinida que tocará su fin cuando se cumpla alguna de las siguientes condiciones:

- El equipo dueño del balón pierde la pelota.
- Se comete algún tipo de interrupción en el juego (faltas, córneres, etc.)
- El equipo dueño del balón ejecuta un tiro.

La duración mínima de una posesión debe ser de cuatro o más eventos para ser tenida en cuenta para este estudio. Por otro lado, no se tendrá en cuenta ningún tipo de evento del tipo "Others on the ball" o "Duel"; en el caso del primero, son eventos que aportan información de jugadores que no tienen control sobre la pelota, y en el caso del segundo, los eventos únicamente mencionan la existencia de una lucha por la posesión del balón. Se considera que la inclusión de estos eventos para la identificación de las posesiones únicamente obstaculiza y ralentiza el proceso. Además, no aporta ningún tipo de información ya que, es conveniente recordar que el objetivo de este procesado es extraer patrones de movimiento del balón en posesión de un equipo. Se despreciará cualquier tipo de información que tenga que ver con el jugador o jugadores que participen en cualquiera de los eventos.

Además, una última consideración que debe tenerse en cuenta es que, a fin de aliviar el coste computacional del clustering posterior y evitar considerar posesiones genéricas con poca importancia, sólo se tendrán en cuenta aquellas cuyo último evento se haya producido en, al menos, el último tercio del campo (exactamente, cuando la posición de inicio del último evento sea mayor que 70 para el eje x). Se considera que estas posiciones no tienen relevancia para el estudio ya que las posesiones acabadas en zonas más defensivas podrían ser fruto del simple movimiento del balón sin objetivo claro o no ser representativas del juego. El objetivo en el fútbol es anotar más goles que el rival, por tanto, las posesiones que no sean propicias de crear una ocasión de gol serán descartadas.

Una vez definida de forma clara qué se considerará una posesión se procede a definir qué características se recogerán de cada una de estas posesiones.

- **Número de eventos de la posesión** Esta característica se incluye con la finalidad de discernir entre posesiones largas y cortas.
- **Coordenadas inicial y final de la posesión** Con estas variables podremos identificar el patrón de la posesión. Se ha decidido tener en cuenta la altura inicial y final de la posesión (coordenada x del primer y último evento) y el carril del último evento de la posesión (eje y).
- **Porcentaje de eventos registrados en cada una de las nueve posiciones del campo** De nuevo, conocer, las zonas más frecuentadas en cada posesión para una mejor representación de esta.

- **Binaria que indique si la posesión acaba en tiro o no** Para distinguir posesiones en base a su importancia, aquellas con mayor número de remates serán las que idóneamente persiga realizar cualquier equipo.

Finalmente, con objetivos meramente identificativos se incluirán los identificadores de la posesión, del equipo dueño de la misma y el instante en el que esta se inicia.

4.4 Clustering de posesiones

Una vez almacenados los datos de posesiones ya procesados será necesario someterlos a un tratamiento previo a la aplicación del clustering. En este caso, el centrado y escalado de los datos es completamente necesario para realizar un clustering de calidad. Esto es debido a la diversidad de variabilidad existente entre las métricas utilizadas para medir las características, que introduce una diferencia artificial entre las diferentes variables que sesgará la importancia de cada una de las variables dentro del propio clustering. El centrado permite plasmar todas las características en un mismo origen de coordenadas (media 0) y el escalado permitirá igualar la varianza de todas las variables (varianza 1) para que no tenga influencia en el algoritmo de agrupación. Este proceso también es conocido como estandarización o conversión a *z-score*. El proceso de estandarización se realiza mediante la siguiente fórmula:

$$z = \frac{x - \mu}{\sigma} \tag{4.2}$$

Donde:

z es el valor escalado a varianza unitaria de la característica.
x es el valor original de la característica.
μ es la media de la característica en el conjunto de datos.
σ es la desviación estándar de la característica en el conjunto de datos.

Una vez realizada la puesta a punto de los datos se diseñará un pequeño experimento cuya finalidad será el de encontrar aquella combinación de técnicas e hiperparámetros que maximicen la calidad de los clústeres obtenidos. De este modo, tras obtener los resultados, se procederá con la aplicación del clustering sobre los datos usando la mejor combinación encontrada en el experimento. Todo este proceso se detallará más en profundidad en la sección 5.2.

Los resultados del clustering de posesiones permitirán sacar las primeras conclusiones sobre los tipos de posesiones que podemos encontrar durante los encuentros disputados en las 5 competencias ligueras más importantes de Europa. Esta información permitirá arrojar luz sobre las estrategias más utilizadas en el fútbol europeo, y observar si realmente existen diferencias claras entre las diferentes jugadas. No cabe duda de que encontrar dos posesiones iguales es muy complicado, y que, agrupar posesiones por similaridad es todo un reto. Sin embargo, a cada clúster de posesiones se le asignará una «posesión referencia» que representará a todo el clúster y permitirá describir de una forma gráfica qué tipo de jugadas podemos encontrar dentro del mismo con el objetivo de facilitar la comprensión de las características de cada clúster. La caracterización de cada uno de los clústeres permitirá agrupar estos clústeres en categorías creadas de forma arbitraria. Es decir se reagruparán los n clústeres resultado en m categorías ($m < n$). La creación de estas categorías será fundamentada en la sección 6.1.

Una vez identificados los clústeres que agrupan y dividen todas las posesiones, se asignará a cada equipo una puntuación relacionada con cada una de las categorías. Esta puntuación se calculará como el porcentaje de posesiones del equipo correspondiente que se engloban dentro de cada categoría. De este modo, suponiendo que sólo existen dos categorías (categoría1 y categoría2) y que el FC Barcelona engloba el 60 % de sus posesiones en la primera categoría y el 40 % restante en la segunda; el FC Barcelona tendrá asignado un *score* de 0.6 para la categoría1 y de 0.4 para la categoría2. Esta asignación permite obtener una medida identificativa del tipo de juego que despliega cada equipo en función de la división creada con el clustering de posesiones.

4.5 Clustering de equipos

Para el segundo clustering será necesario definir las características que se considerarán para identificar el juego de un equipo. Esta elección ha sido propuesta atendiendo a diferentes criterios, el movimiento del balón del equipo, el comportamiento colectivo y las acciones defensivas. Sin embargo, el estilo de juego es algo realmente complejo de definir y caracterizar y las variables escogidas pueden estar sometidas a debate ya que no se trata de un tema objetivo. A continuación, se muestra la selección escogida para este estudio:

- **Número de faltas:** Esta característica permite identificar los equipos con un estilo de juego más físico e intenso.
- **Número de paradas:** Esta característica supone una representación de la capacidad defensiva del equipo, un mayor número de paradas puede ser indicio de fragilidad defensiva.
- **Longitud media de los pases** Esta característica permitirá discriminar aquellas posesiones que hacen uso de pases más cortos (juego de toque) de aquellas con pases más largos (transiciones).
- **Número de pases clave:** Esta característica discrimina entre equipos con posesiones más elaboradas y peligrosas de aquellos con posesiones más planas.
- **Número de intercepciones:** Esta característica supone una representación de la intensidad en la presión, las intercepciones no implican necesariamente robos de pelota.
- **Número de duelos disputados:** De nuevo, esta característica permite identificar equipos más físicos e intensos que no rehuyen las disputas del balón.
- **Número de tiros divididos por distancia:** Para diferenciar entre equipos con preferencia por la larga o la corta distancia.
- **PageRank:** El valor del PageRank para el nodo con el PageRank más alto, esta característica permite identificar la importancia relativa del nodo más importante de la red de jugadores del equipo.
- **Densidad:** Esta característica ofrece una visión sobre la compacidad del estilo de juego del equipo.
- **Longitud de la mayor comunidad:** Esta característica permite identificar la cantidad de jugadores implicados en el juego principal del equipo. Podría ser un modo de definir el número de jugadores que conforman el «esqueleto» del equipo.

- **Puntuación para cada categoría de posesión:** Que ofrecerá la visión del movimiento del balón dentro del terreno de juego del equipo.

Para la extracción de características de los equipos se han realizado diversos procesos en paralelo. Por un lado, se han recorrido todos los eventos y, contador mediante, se han registrado cada uno de los eventos de interés para cada equipo. Cabe recordar que los eventos de interés son el número de faltas, paradas, pases clave, intercepciones y el número de duelos disputados. La longitud media de los pases, por su parte, se ha computado de forma similar, calculando la media de la variable long_XY (mencionada previamente) para todos los eventos de tipo pase de cada equipo.

Por otro lado, se han registrado el número de tiros en función del carril en función de la distancia. Para la clasificación según la distancia se ha seguido un criterio simple, cualquier tiro desde detrás de la línea frontal del área (coordenada 85 del eje x en Wyscout) se considerará tiro lejano y cercano en cualquier otro caso. Las características notacionales (número de faltas, paradas, pases clave, intercepciones y número de duelos disputados) han sido normalizadas por partido ya que la Bundesliga (liga alemana) disputa 34 jornadas de competición a diferencia de las 38 jornadas que disputan el resto de las ligas europeas. Esta normalización permitirá eliminar el sesgo introducido debido a la diferencia de partidos disputados por los equipos alemanes del mismo modo al descrito en el análisis exploratorio.

Finalmente, se recoge la puntuación asociada a cada categoría de cada equipo a modo de representación del comportamiento dentro del terreno de juego.

El preprocesado en este caso se realizará de forma similar al preprocesado aplicado para el primer clustering, se centrarán y escalarán los datos a media cero y varianza unitaria para igualar la importancia de todas las características. Una vez finalizado el proceso de clusterización, se procederá con la validación y análisis de los resultados, que serán detallados en los siguientes capítulos.

4.5.1. Equipos como red

Tres de las variables utilizadas (PageRank, Densidad y Longitud de la máxima comunidad) provienen de la modelización del comportamiento de un equipo de fútbol como una red. Se ha considerado que este tipo de enfoque permite extraer características de la estructura colectiva del equipo de la forma más óptima posible. A continuación, se detallará el proceso de modelización de cada uno de los clubes como un grafo dirigido y las métricas de las que se ha decidido hacer uso junto su correspondiente justificación.

Para cada equipo, se ha calculado una matriz de adyacencia a partir de los jugadores que han intervenido en cada una de las posesiones registradas a lo largo de la temporada liguera. Es importante destacar que en este caso cambiarán las características y el procesado de las posesiones: sólo se considerará un cambio de posesión cuando el equipo dueño del balón lo pierda y no se contemplarán posesiones menores de tres eventos; además, sólo se contemplarán los eventos de tipo pase para este proceso. Esta relajación de las restricciones se debe a que lo que se pretende modelizar es únicamente el movimiento del balón entre los jugadores del equipo, es decir, identificar el patrón de pases del equipo. Se ha decidido no tener en cuenta posesiones de menos de tres eventos ya que se considera que no son lo suficientemente trascendentales como para ser tenidas en cuenta dentro de este patrón de pases que se busca modelizar.

De este modo, se construye la matriz de adyacencia como una matriz cuadrada con tantas filas y columnas como jugadores haya en la plantilla del club en cuestión. Cada vez que se registre un pase de un jugador x a un jugador y, se sumará 1 en la posición

x,y de la matriz de adyacencia. Por tanto, la matriz final contendrá el número de pases totales registrados a lo largo de la temporada entre cada par de jugadores. En este caso los valores de la diagonal principal tienen poco interés y se situarán a 0 ya que no se contempla la posibilidad de que un jugador se realice un pase a sí mismo.

De este modo, se puede transformar la matriz de adyacencia en una red dirigida tal que:

El grafo dirigido ponderado $G(V, E, W)$ se define como una tupla ordenada de conjuntos, donde W es el conjunto de pesos asociados a cada arco:

$$G(V, E, W) = (V, E, \{w_{ij} \mid (v_i, v_j) \in E\})$$

La notación V representa el conjunto de nodos (jugadores):

$$V = \{v_1, v_2, \ldots, v_n\}$$

La notación E representa el conjunto de arcos (pases de un jugador x a un jugador y):

$$E = \{(v_i, v_j) \mid v_i, v_j \in V \text{ y } v_i \text{ realiza un pase a } v_j\}$$

La notación W representa el conjunto de pesos asociados a cada arco:

$$W = \{w_{ij} \mid (v_i, v_j) \in E\} = \{w_{12}, w_{23}, w_{34}, \ldots\}$$

Donde V es el conjunto de nodos (jugadores), E es el conjunto de aristas (pases entre jugadores) y W es el conjunto de pesos que asigna un peso w_{ij} a cada arco (v_i, v_j).

Cada peso w_{ij} en W representa la frecuencia de pases del jugador v_i al jugador v_j.

Para la manipulación de este tipo de grafos se ha hecho uso del paquete NetworkX (Hagberg et al., 2008) que ofrece todas las herramientas necesarias para calcular las métricas propuestas en este trabajo.

Densidad

Una vez la red del equipo ha sido modelizada como un grafo, es posible extraer la información deseada mediante teoría de grafos. Comenzando por la densidad, que permitirá extraer una medida de la conexión de los jugadores de un equipo, una densidad mayor indicará que ese equipo tiende a realizar muchos pases entre todos sus miembros, estará densamente conectado. Se trata de una medida simple pero eficaz para evaluar la calidad de las interconexiones de un equipo.

La densidad del grafo ponderado se calcula como:

$$\text{Densidad} = \frac{1}{m} \sum_{w \in W} w \tag{4.3}$$

Donde:

n es el número de nodos (jugadores) en el equipo,
$m = n \cdot (n-1)$ es el número total de arcos posibles en el equipo,
W es el conjunto de pesos asociados a los arcos,
w es un peso individual perteneciente al conjunto W.

PageRank

El algoritmo PageRank (Brin and Page, 1998) se utiliza para medir la importancia relativa de los nodos en una red. El cálculo del PageRank en un grafo ponderado se realiza mediante iteraciones que redistribuyen la importancia de los nodos basada en los enlaces y sus pesos. La fórmula general para el cálculo del PageRank en una iteración es:

$$PR(v) = \sum_{w \in C_{\text{in}(v)}} \frac{PR(w) \cdot w_{vw}}{\sum_{u \in C_{\text{out}(w)}} w_{wu}}$$

Donde:

$PR(v)$ es el PageRank del nodo v,
$C_{\text{in}(v)}$ es el conjunto de nodos que apuntan a v,
$C_{\text{out}(w)}$ es el conjunto de nodos a los que w apunta,
w_{vw} es el peso de la arista entre v y w.

En NetworkX, este cálculo puede ser fácilmente realizado por la función pagerank(). Sin embargo, para realizar una comparación justa entre redes, se ha tomado la decisión de normalizar el valor del PageRank dividiéndolo entre el número de nodos de la red. Para el caso que ocupa en este trabajo, únicamente se almacenará el valor más alto del PageRank de cada red (equipo) y este valor servirá como referencia de la importancia relativa en la red del jugador más importante (valga la redundancia) del equipo. Entendiéndose el jugador más importante como aquel que participa en mayor número de posesiones, aquel jugador por cuyas botas pasan la mayoría de las jugadas. En algunos equipos este jugador puede ser fácilmente identificable y en otros esta figura puede estar más diluida. Este hecho también supone una característica del juego del equipo, de cómo de centralizado está. Valores más altos supondrán más importancia del jugador clave y valores más bajos supondrán mayor repartición en el protagonismo de la pelota.

Comunidad más grande

Para explicar esta métrica es necesario explicar, en primer lugar, qué es una partición. Una partición en un grafo se refiere a la división de los nodos del grafo en grupos o comunidades, de manera que los nodos dentro de un mismo grupo estén más densamente conectados entre sí que con nodos en otros grupos. Es decir, una partición busca agrupar todos los nodos que comparten características o conexiones similares.

En este contexto se encuentra el algoritmo de Detección de Comunidades Louvain (Blondel et al., 2008), que es un método sencillo para extraer la estructura de comunidades en una red. Se trata de un método heurístico basado en la optimización de la modularidad. La modularidad es una medida que cuantifica la calidad de una partición de un grafo en comunidades. Se basa en la comparación entre las conexiones reales dentro de las comunidades y las conexiones esperadas en una red aleatoria equivalente.

Según la propia explicación encontrada en la documentación de NetworkX: «El algoritmo funciona en dos pasos. En el primer paso, asigna cada nodo para que esté en su propia comunidad, y luego, para cada nodo, intenta encontrar la ganancia máxima de modularidad positiva moviendo cada nodo a todas sus comunidades vecinas. Si no se logra ninguna ganancia positiva, el nodo permanece en su comunidad original».

La fórmula de la modularidad se expresa como (Dugué and Perez, 2015):

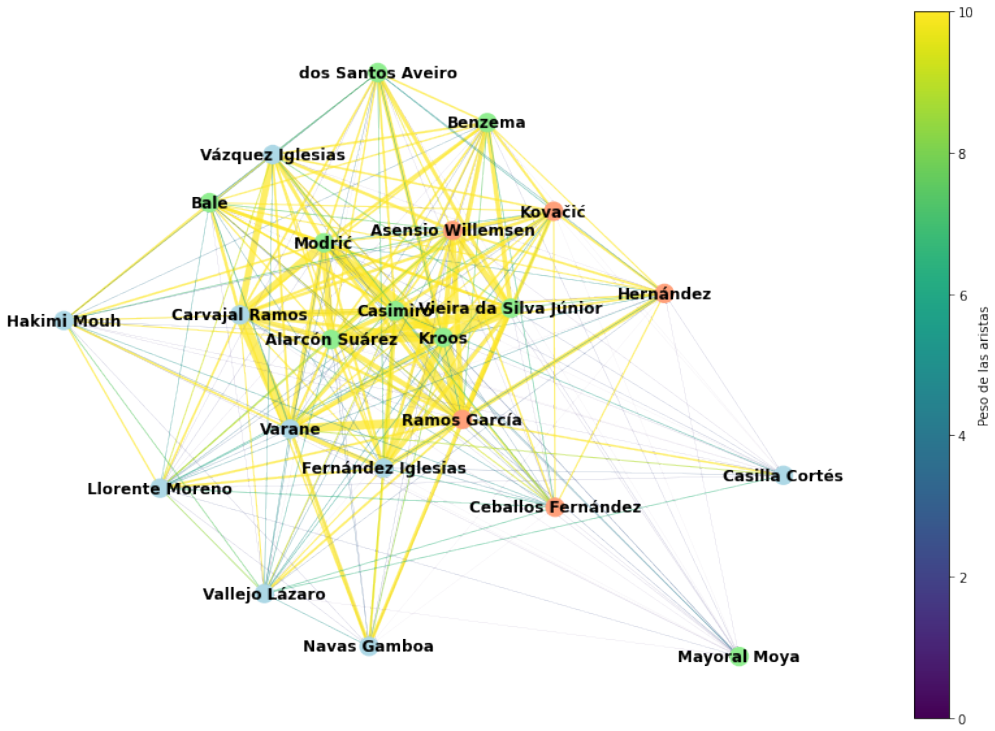

Figura 4.3: Grafo creado a partir de la modelización del Real Madrid CF

$$\Delta Q = \frac{k_{i,\text{in}}}{m} - \gamma \frac{k_i^{\text{out}} \cdot \Sigma_{\text{tot}}^{\text{in}} + k_i^{\text{in}} \cdot \Sigma_{\text{tot}}^{\text{out}}}{m^2} \tag{4.4}$$

Donde:

$k_{i,\text{in}}$ es la suma de los pesos de las aristas que entran en el nodo i (grado ponderado interno),
γ es un factor de ajuste,
k_i^{out} es la suma de los pesos de las aristas que salen del nodo i (grado ponderado externo),
$\Sigma_{\text{tot}}^{\text{in}}$ es la suma total de los pesos de todas las aristas que entran en todos los nodos en la comunidad,
$\Sigma_{\text{tot}}^{\text{out}}$ es la suma total de los pesos de todas las aristas que salen de todos los nodos en la comunidad,
m es el t.

La fórmula compara la densidad de conexiones reales dentro de la comunidad (primer término) con la densidad esperada de conexiones en una red aleatoria (segundo término). Un valor positivo de ΔQ indica que la comunidad está más densamente conectada de lo esperado por azar, sugiriendo la presencia de estructuras de comunidad significativas. Cabe mencionar que se han extraído tanto la modularidad como el número de comunidades de la partición a modo informativo, aunque finalmente se ha decidido no incluirlo en el estudio.

Esta métrica proporciona una visión de la unidad del equipo, la existencia de una gran comunidad implicará un equipo altamente unido, con un estilo de juego donde todos sus miembros participan, de hecho, si el tamaño de la comunidad más grande es mayor que 11 implica que, incluso los suplentes, presentan una alta integración en el grupo principal. Sin embargo, se encontrarán equipos cuya comunidad principal esté formada por no más de ocho jugadores, lo que implicará la existencia de jugadores que no están claramente integrados en el grupo, incluso a pesar de formar parte del «11 titular».

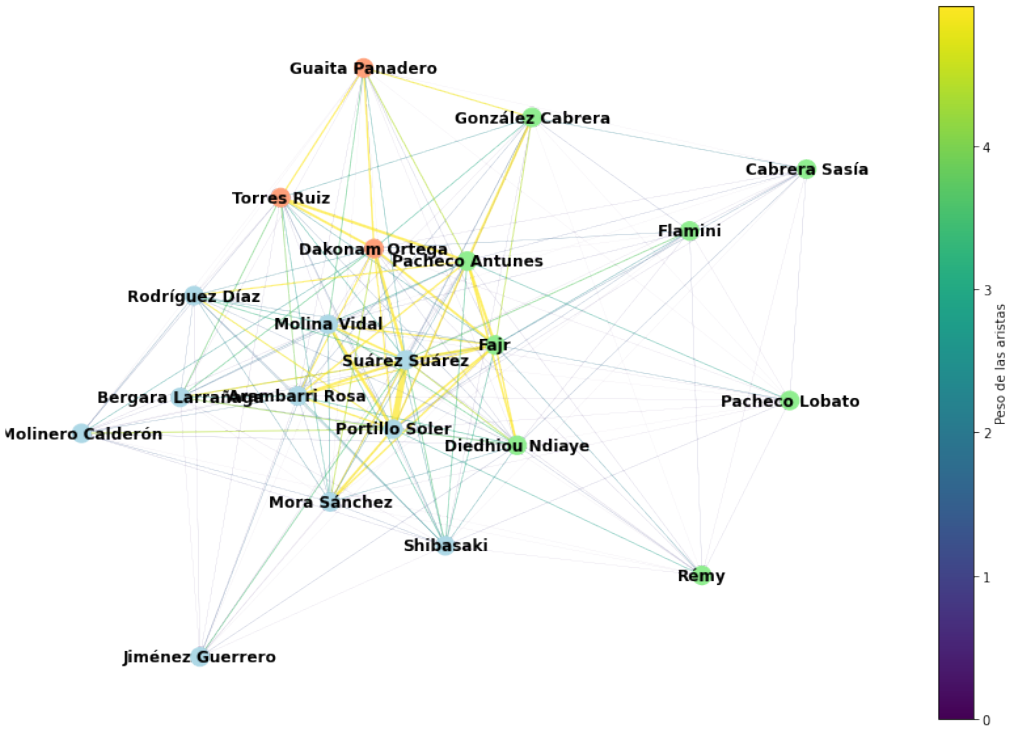

Figura 4.4: Grafo creado a partir de la modelización del Getafe CF

Puesto que todo lo relatado hasta ahora puede resultar bastante abstracto, se ha decidido incluir las figuras 4.3 y 4.4. Estas figuras resultan de la modelización del Real Madrid CF y Getafe CF como un grafo, los arcos han sido coloreadas según el peso asociado. Para una mejor visualización se ha decidido normalizar este peso respecto del máximo, los valores normalizados oscilan entre 0 y 10. En adición, para poder comparar ambos equipos, la normalización de los pesos se ha realizado en función del peso máximo entre los dos grafos. En este caso, el arco de mayor peso ha sido el que conecta a los jugadores $Kroos- > Ramos$ con un peso de 215 (que equivaldría al valor máximo una vez relativizado). Además a cada nodo se le ha asignado un color en función de la comunidad correspondiente tras la partición realizada. Comparando ambas ilustraciones, se observa rápidamente la diferencia de densidad que existe entre ambos equipos, siendo el Real Madrid un equipo mucho más conexo que el Getafe. Además, en el Real Madrid se observa que existen tres comunidades y que, en general, están asociadas a la posición, es decir una para defensas, otra para mediocampistas y otra para delanteros. Sin embargo, en el Getafe las tres comunidades vienen son claramente distinguidas por el número de minutos disputados entre sus jugadores.

CAPÍTULO 5
Experimentación y validación de resultados

En este capítulo se detallarán exhaustivamente, por un lado, los experimentos llevados a cabo para la optimización de parámetros del clustering y, por otro, el proceso de validación de los resultados del mismo.

5.1 Consideraciones previas

La experimentación ha supuesto una de las tareas con mayor consumo de tiempo y recursos informáticos. En este trabajo se han diseñado dos experimentos, cada uno de ellos asociado al clustering de posesiones y de equipos, respectivamente. La finalidad de ambos experimentos ha sido la de optimizar los parámetros para extraer agrupaciones de la mayor calidad posible. Así pues, se comenzará por recordar los tres tipos de técnicas de clustering con las que se ha experimentado:

K-Means: K-Means es un algoritmo de agrupamiento que organiza datos en grupos (clústeres) basados en similitudes. Selecciona centroides iniciales, asigna puntos al centroide más cercano y actualiza los centroides iterativamente para minimizar la varianza dentro de los clústeres. Para su implementación en Python se ha hecho uso del paquete Scikit-Learn.

HDBSCAN: HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de agrupamiento que encuentra clústeres en datos con diferentes densidades y formas. Utiliza la densidad local para formar clústeres y puede identificar puntos ruidosos o atípicos. También construye una estructura de árbol de clústeres para diferentes escalas de densidad. Para su implementación en Python se ha hecho uso del paquete hdbscan.

Aglomerativo: En este método, cada elemento se considera inicialmente como un clúster independiente y luego se combinan gradualmente en clústeres más grandes basados en su similitud. El proceso de clustering aglomerativo se lleva a cabo de manera jerárquica, es decir, se crea una jerarquía de clústeres anidados. En cada etapa, los clústeres más cercanos entre sí se fusionan para formar un nuevo clúster más grande. Esto continúa hasta que todos los elementos estén agrupados en un solo clúster o en un conjunto predeterminado de clústeres. Para su implementación en Python se ha hecho uso, de nuevo, del paquete Scikit-Learn.

Además, previo a la aplicación del clustering, también se ha optado por ensayar con técnicas de reducción de la dimensionalidad, en este caso UMAP. De nuevo, cabe recordar que UMAP (Uniform Manifold Approximation and Projection) es un algoritmo de reducción de dimensionalidad no lineal que preserva las relaciones de proximidad entre datos. A continuación, se realiza una breve justificación de la elección de cada una de las técnicas propuestas:

Por una parte, K-Means ofrece la posibilidad de establecer un número de clústeres arbitrario; lo que supone una ventaja cuando queremos extraer una amplia cantidad de agrupamientos, pero los datos son bastante similares entre sí. Sin embargo, esta técnica también es bastante influenciable por el ruido en los datos, en este sentido, UMAP podría ayudar disminuir el sesgo introducido y facilitar la creación de clústeres de mayor calidad.

Por otra parte, HDBSCAN ofrece la ventaja de la robustez frente a datos ruidosos; sin embargo, en datos que no pueden ser claramente agrupados puede presentar el inconveniente de no poder elegir el número de clústeres de forma arbitraria, y por tanto, crear pocos clústeres de grandes dimensiones. Cuando se enfrenta un problema de las características del agrupamiento de las posesiones, es necesario tener siempre en cuenta que la diversidad de tipos de posesiones es infinita y la línea divisoria entre dos clústeres puede ser demasiado delgada.

Finalmente, el clustering aglomerativo destaca por su simplicidad y capacidad para formar jerarquías de clústeres, lo que permite una interpretación intuitiva de la estructura de los datos. Este hecho presenta un especial interés para el caso que aquí se presenta tanto para la agrupación de posesiones como para la agrupación de equipos. Ambas agrupaciones podrían presentar una estructura jerárquica, por ejemplo, una división en dos grandes clústeres, uno con jugadas que acaban en remate y otro que no, y, a su vez, encontraríamos un nivel inferior en el que encontramos un clúster por cada uno de los carriles predominantes en las posesiones. Sin embargo, el enfoque aglomerativo puede volverse computacionalmente costoso en conjuntos de datos grandes debido a su complejidad $O(n^3)$ en algunas implementaciones. Además, puede ser sensible a ruido y valores atípicos, ya que las fusiones iniciales pueden influir en las etapas posteriores.

Todos los métodos testeados tienen, a priori, ventajas y desventajas al ser aplicados para abordar la tarea en cuestión. No obstante, existe una barrera clara e incuestionable a tener en cuenta: los recursos computacionales. El *dataset* de posesiones cuenta con 86235 filas y resulta el más restrictivo claramente, las consecuencias de esto son, en primer lugar, el uso obligatorio de UMAP para la reducción de la dimensionalidad, y, por ende, el coste computacional y; en segundo lugar, la imposibilidad de aplicar clustering aglomerativo por falta de memoria. Por otro lado, para el clustering de los equipos sí se realizarán pruebas con las tres técnicas, de nuevo se hará uso de UMAP previo para mantener la consistencia en la experimentación.

5.2 Diseño de los experimentos

Como ya se ha mencionado, encontrar agrupaciones de calidad que satisfagan las expectativas resulta un reto particularmente difícil, la experimentación será la solución planteada para atacar este problema. Mediante la combinación de las posibles configuraciones y la optimización de sus hiperparámetros se pretende encontrar la mejor agrupación posible. No obstante, como se acaba de mencionar, no sólo se trata de encontrar la mejor configuración, sino de optimizar cada una de las configuraciones y elegir cuál es la más conveniente para este caso. Esto implica explorar una inmensa cantidad de combinaciones de hiperparámetros para cada configuración que lastrará el proceso de

experimentación y donde el límite de recursos a invertir viene definido por los propios límites del proyecto. Por tanto, a continuación, se citarán los hiperparámetros a optimizar para cada una de las técnicas mencionadas anteriormente a fin de dilucidar las posibles combinaciones a explorar para cada configuración y exponer cuáles han sido las que finalmente se han decidido experimentar.

5.2.1. UMAP

Los hiperparámetros que permite modificar el modelo UMAP de la librería umap-learn y que se incluirán dentro de las combinaciones a explorar son:

- **n_neighbors:** Define la cantidad de vecinos cercanos considerados durante la construcción del grafo de vecinos. Más vecinos tienden a capturar estructuras globales, mientras que menos vecinos pueden enfocarse en estructuras locales.
  Clustering de posesiones: Se han considerado todos los valores entre 10 y 40.
  Clustering de equipos: Se han considerado todos los valores entre 2 y 20.

- **min_dist:** Controla la compresión de los datos en el espacio de representación. Valores bajos conservan estructuras locales, pero pueden causar superposición. Valores altos refuerzan la separación de clústeres, pero pueden sacrificar detalles locales.
  Clustering de posesiones: Se han considerado todos los valores entre 0 y 0.25.
  Clustering de equipos: Se han considerado todos los valores entre 0 y 0.25.

- **n_components:** Indica la dimensión del espacio reducido deseado. Determina cuántas dimensiones conservar después de la reducción de dimensionalidad.
  Clustering de posesiones: Se han considerado todos los valores entre 3 y 6.
  Clustering de equipos: Se han considerado todos los valores entre 3 y 6.

- **metric:** Especifica la métrica de distancia para medir la similitud entre puntos en el espacio original. Afecta a cómo se construye el grafo de vecinos.
  Clustering de posesiones: Se han considerado las distancias euclídea y de Manhattan.
  Clustering de equipos: Se han considerado las distancias euclídea y de Manhattan.

5.2.2. K-Means

Los hiperparámetros que permite modificar el K-Means de la librería Scikit-learn y que se incluirán dentro de las combinaciones a explorar son:

- **n_clusters:** Define la cantidad de clústeres que se intentan encontrar en los datos. Especifica cuántos grupos distintos se esperan en el conjunto de datos.
  Clustering de posesiones: Se han considerado todos los valores entre 10 y 25.
  Clustering de equipos: Se han considerado todos los valores entre 5 y 15.

- **distance:** Se utiliza para definir la función de distancia para medir la similitud entre puntos. Afecta a cómo se asignan los puntos a los clústeres..
  Clustering de posesiones: Se han considerado las distancias euclídea y de Manhattan.

Clustering de equipos: Se han considerado las distancias euclídea, de Manhattan.

5.2.3. HDBSCAN

Los hiperparámetros que permite modificar el HDBSCAN de la librería hdbscan y que se incluirán dentro de las combinaciones a explorar son:

- **min_cluster_size:** Define el tamaño mínimo de un clúster válido. Los clústeres con menos puntos que este valor se consideran ruido o atípicos.
  Clustering de posesiones: Se han considerado todos los valores entre 100 y 500.
  Clustering de equipos: No se ha considerado tamaño mínimo de clústeres.
- **min_samples:** Establece la cantidad mínima de puntos requeridos en una vecindad para que se considere un punto como núcleo. Influencia la detección de densidad y la formación de clústeres.
  Clustering de posesiones: Se han considerado todos los valores entre 50 y 250.
  Clustering de equipos: No se ha considerado mínimo de muestras.
- **clúster_selection_epsilon:** Este parámetro controla la selección de clústeres cuando se utiliza el método «eom» (exceso de masa) para elegir clústeres finales.
  Clustering de posesiones: Se han considerado todos los valores entre 0 y 0.5.
  Clustering de equipos: Se han considerado todos los valores entre 0 y 0.5.

5.2.4. Clustering aglomerativo

Los hiperparámetros que se pueden modificar en el algoritmo de Clustering aglomerativo de la librería scikit-learn, y que serán considerados en las combinaciones exploradas, son los siguientes:

- **n_clusters:** El número de clústeres deseado al final del proceso de clustering.
  Clustering de equipos: Se han considerado todos los valores entre 5 y 15.

- **linkage:** El método de enlace utilizado para calcular las distancias entre clústeres.
  Clustering de equipos: Se han considerado los valores *ward, complete, average y single*.

Para la creación de los experimentos se ha hecho uso de la librería optuna como ya se ha mencionado. Asimismo, también se ha detallado el espacio de exploración, en este sentido poco más queda que añadir al respecto sobre el diseño de los estudios de Optuna. Se ha tomado la decisión de intentar un máximo de 100 combinaciones para las posesiones y 300 para los equipos (puesto que las pruebas requieren un coste computacional muy inferior) y escoger la mejor de entre todas ellas.

Sin embargo, hasta ahora simplemente se ha hablado de la búsqueda de clústeres o agrupaciones de calidad. Previamente en este trabajo, en concreto en la sección 2.4.4, ya se mencionaron diferentes medidas de la calidad de un clúster y se debatió sobre las fortalezas y debilidades de cada una. En este trabajo se ha buscado una forma de aplacar las deventajas de hacer uso de un sólo índice mediante la combinación de dos: el coeficiente de Silhouette y el índice de Davies-Bouldin.

$$\text{Índice combinado} = \frac{\text{silhouette} + 1}{2} \times \frac{1}{1 + \text{davies\_bouldin}} \tag{5.1}$$

Esta métrica está normalizada entre 0 y 1 y permite comparar diferentes configuraciones combinando las ventajas de ambos índices. El coeficiente de Silhouette mide la cohesión y separación de los clústeres considerando tanto la distancia promedio dentro de un clúster como la distancia promedio al clúster más cercano. El Índice Davies-Bouldin, por otro lado, se centra en la separación entre clústeres en función de la distancia y la dispersión. Combinar ambas medidas permite evaluar tanto la cohesión interna como la separación interclúster, lo que proporciona una imagen más completa de la calidad de los clústeres.

Esta métrica ha sido implementada en el diseño de los experimentos de Optuna como métrica a maximizar en el proceso de optimización de hiperparámetros. Antes de concluir esta sección, es necesario mencionar que se ha tomado una precaución significativa para garantizar la replicabilidad de los resultados obtenidos durante la experimentación mediante el uso de semillas.

Las semillas aseguran que las operaciones aleatorias y las inicializaciones sean consistentes entre diferentes ejecuciones, especialmente en algoritmos con alta variabilidad en los resultados como son los de clustering. De esta manera, se busca minimizar la variabilidad introducida por factores aleatorios y permitir la reproducción coherente de los resultados.

5.3 Validación de los resultados

El proceso de validación de resultados se centra en la verificación de la robustez y significancia de los resultados obtenidos en una investigación o análisis. Su objetivo fundamental es garantizar que los datos recopilados y las conclusiones extraídas sean precisos, coherentes y fielmente representativos de la realidad bajo estudio.

5.3.1. Clustering posesiones

ANOVA

En este caso, los resultados a validar provienen de la aplicación de clustering, mediante la experimentación se aseguró que los grupos obtenidos tengan la mayor calidad posible en base a la métrica propuesta. No obstante, una vez han sido etiquetados los datos es necesario comprobar que los clústeres resultantes son lo suficientemente diferentes. Sin duda alguna, el ANOVA es la técnica apropiada cuando se presenta este problema.

El ANOVA (Análisis de Varianza) es una técnica estadística que se utiliza para comparar las medias varios grupos en un conjunto de datos. El objetivo es determinar si las diferencias observadas entre los grupos son estadísticamente significativas o si podrían ser el resultado del azar. Esta técnica analiza la variabilidad entre los grupos y la variabilidad dentro de los grupos para evaluar si hay suficiente evidencia para concluir que al menos un grupo difiere significativamente de los demás en términos de la variable que se está estudiando. Además, el ANOVA presenta también su versión multivariante, MANOVA, que permite estudiar diferencias entre grupos en función de múltiples variables.

Sin embargo, ANOVA es una prueba paramétrica, está basada en supuestos específicos sobre la distribución de los datos. En concreto, se asume que los datos siguen una distribución normal y que se cumple la homocedasticidad (la varianza de cada grupo

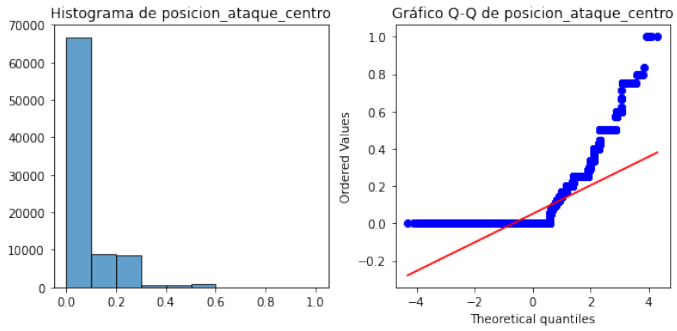

Figura 5.1: Estudio de la normalidad para la variable posicion_ataque_centro

para cada variable es aproximadamente igual). Estos supuestos son fundamentales para garantizar la validez y la interpretación adecuada de los resultados. Por tanto, el primer paso antes de aplicar un ANOVA es verificar que los datos cumplen ambos supuestos.

Comenzando por la normalidad, generalmente, y siempre y cuando la cantidad de variables objeto de estudio no vuelva esta tarea inviable. El estudio de la normalidad se realiza mediante la combinación de visualización y pruebas estadísticas. Las visualizaciones consideradas normalmente para la verificación de la normalidad en una variable son el histograma, la función de densidad y el papel probabilístico normal. Por otro lado, la prueba para la verificación de la normalidad es el test de Shapiro-Wilk.

El test de Shapiro-Wilk se basa en la hipótesis nula (H0) de que los datos provienen de una distribución normal. La hipótesis alternativa (H1) es que los datos no siguen una distribución normal. La prueba calcula una estadística de prueba y un valor p asociado que indica la probabilidad de obtener los resultados observados si los datos se distribuyeran normalmente. Si el valor p es menor que un umbral predefinido, que se denominará nivel de significancia (α) (por ejemplo, $\alpha = 0{,}05$), se rechaza la hipótesis nula y se concluye que los datos no siguen una distribución normal.

Por otro lado, para verificar la homocedasticidad es posible atender a gráficos de caja y bigotes o visualizar la varianza de cada una de las variables. En adición, para verificar si existe o no homocedasticidad es común el uso de la prueba de Levene.

La prueba de Levene es una prueba estadística utilizada para evaluar si las varianzas de dos o más grupos son estadísticamente iguales. La hipótesis nula (H0) en la prueba de Levene establece que las varianzas entre los grupos son iguales. La hipótesis alternativa (H1) sugiere que al menos una varianza es diferente de las demás. Si el p-valor es menor que el nivel de significancia predefinido (α), entonces se rechaza la hipótesis nula. Esto significa que hay suficiente evidencia para concluir que al menos una de las varianzas entre los grupos es diferente. En caso contrario, se asume que las varianzas entre los grupos son estadísticamente iguales.

Continuando en esta dirección, resulta particularmente sencillo comprobar si las variables del clustering de posesiones cumplen las suposiciones del ANOVA. Todas las pruebas realizadas han lanzado una conclusión firme, no se cumple ninguna de las suposiciones del ANOVA y, por tanto, no será posible aplicarlo. El test de Shapiro-Wilk realizado para las variables ha lanzado un p-valor de 0 para todas ellas. Por otro lado, la prueba de Levene, del mismo modo, ha lanzado un p-valor de 0 para todas las variables. Es decir, las variables no siguen una distribución normal y no podemos suponer que tienen una varianza estadísticamente igual. Los gráficos por su parte, así lo corroboran (figuras 5.1 y 5.2).

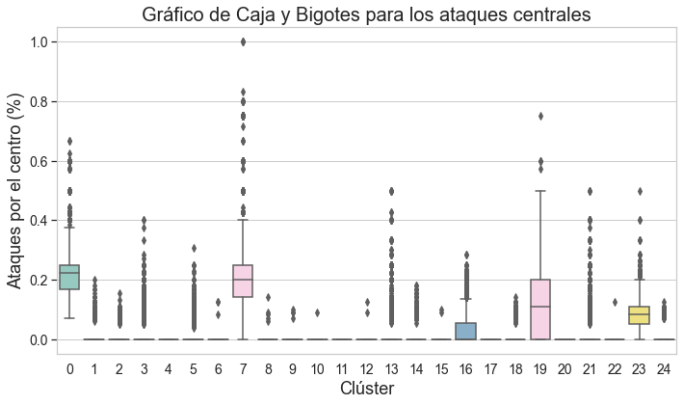

Figura 5.2: Estudio de la homocedasticidad para la variable posicion_ataque_centro

Pruebas no paramétricas

Una vez descartamos la posibilidad de aplicar ANOVA se proponen dos soluciones diferentes, transformar los datos o aplicar pruebas no paramétricas que persigan el mismo fin que el ANOVA. Para tomar esta decisión es necesario atender al contexto y conocer las ventajas y desventajas de cada una de las dos soluciones. Si se opta por aplicar transformaciones, es necesario destacar que los datos se encuentran muy lejos de cumplir las suposiciones de normalidad y homocedasticidad. Por tanto, este proceso será potencialmente ineficiente en términos de recursos y no asegura resultados exitosos. Por otro lado, si se opta por aplicar técnicas no paramétricas será necesario tener en cuenta que este tipo de técnicas resultan menos informativas y suelen lanzar resultados sesgados cuando se dispone de un conjunto escaso de datos. Sin embargo, esto no supone un problema ya que, a pesar de no contar con una distribución uniforme en el número de individuos entre los clústeres, todos tienen una cantidad de individuos suficiente como para que esto no sea un problema.

Finalmente, se opta por aplicar el test de Kruskal-Wallis, que es una prueba estadística no paramétrica utilizada para comparar las medianas de varios grupos independientes. Esta prueba se utiliza para determinar si al menos uno de los grupos difiere significativamente en términos de sus distribuciones poblacionales. La prueba de Kruskal-Wallis se basa en los rangos de los datos y opera bajo la hipótesis nula de que las medianas de todos los grupos son iguales. La hipótesis alternativa sugiere que al menos una de las medianas es diferente.

Cuando se aplica la prueba a todos los clústeres en cada una de las variables se aprecia que sí existen diferencias estadísticamente significativas para todas ellas entre los diferentes grupos, con un p-valor de 0.

En adición, para verificar entre qué grupos se encuentran estas diferencias se realiza un procedimiento *post-hoc* como es la prueba de Dunn para cada variable y para cada par de clústeres. La prueba de Dunn es una extensión de la prueba de rangos con signos de Wilcoxon para comparaciones múltiples. La hipótesis nula es que no hay diferencias significativas entre ninguno de los pares de grupos comparados y la hipótesis alternativa es que al menos un par de grupos comparados presenta diferencias significativas. La prueba de Dunn busca determinar si hay al menos un par de grupos que tienen diferencias significativas en sus medianas. No obstante, esto supone realizar 625 pruebas por cada una de las variables estudiadas, es decir, el estudio individual de cada caso es una operación totalmente inviable. Es por este motivo que se decide construir una matriz en forma de visualización donde se colorea el p-valor en función de la significancia, siendo verde un p-valor de 0 (diferencias estadísticamente significativas entre las medianas de los dos

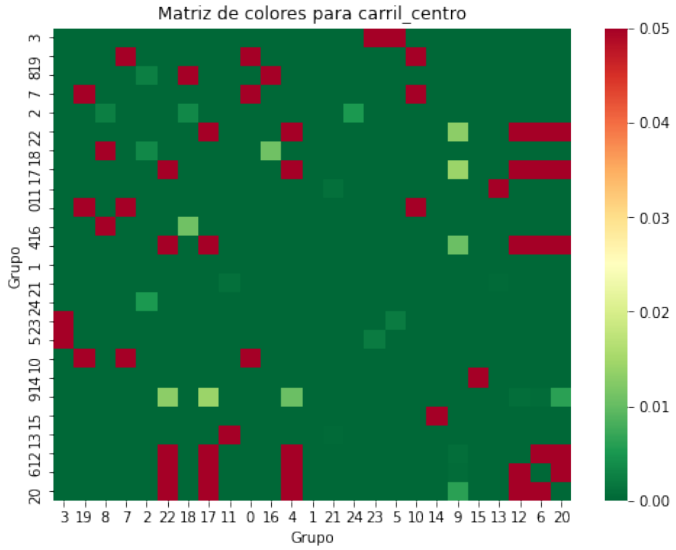

Figura 5.3: Estudio de las diferencias entre clústeres para la variable carril_centro mediante la prueba de Dunn

clústeres comparados en la variable estudiada) y un p-valor de 0.05 rojo (no existen tales diferencias ya que supera el $\alpha = 0{,}05$). Esta visualización proporciona de forma rápida una visión de lo diferentes que son cada uno de los clústeres para cada una de las variables. A este respecto, cabe destacar que variables como los carriles (figura 5.3), las zonas (excepto la zona de defensa) y el número de eventos, son aquellas en las que podemos encontrar mayor disparidad entre los clústeres. Sin embargo, en otras variables como el remate (figura 5.4) se aprecia que gran parte de los clústeres no presentan diferencias significativas. Esto, por ejemplo, es debido a que los clústeres se pueden englobar en dos grandes, los que contienen mayoritariamente posesiones con remates y los que contienen pocas o ninguna.

5.3.2. Clustering equipos

ANOVA

Por otro lado, será necesario validar los resultados provenientes de la aplicación de clustering. Para ello, una vez han sido etiquetados los datos es necesario comprobar que los clústeres resultantes son lo suficientemente diferentes. Volveremos a recurrir al ANOVA en este caso como primera opción ya que se considera una opción más sólida que el uso de pruebas no paramétricas. Sin embargo, será necesario comprobar que se cumplen las suposiciones de normalidad y homocedasticidad en el *dataset* de las características de los equipos.

Comenzando por el estudio de la normalidad, de nuevo se recurrirán al test de Saphiro-Wilk y el estudio de los histogramas y gráficos de cuantiles. En este caso se presenta una situación diferente, ya que, como se puede observar en la tabla 5.1 de las 19 variables, el test Saphiro-Wilk considera que siete de ellas no lo son. Sin embargo, atendiendo al gráfico de cuantiles y el histograma (figura 5.5), se observa que la falta de normalidad viene producida por la existencia de algunos valores anómalos (equipos con valores extremadamente altos para estas variables). Por tanto, se ha decidido asumir que estas variables pueden ser consideradas normales o no violan el supuesto de normalidad significativamente. Aun así, será necesario tener en cuenta los posibles efectos que estas asunciones puedan tener sobre la eficacia del ANOVA durante el estudio. En este sentido, son varios

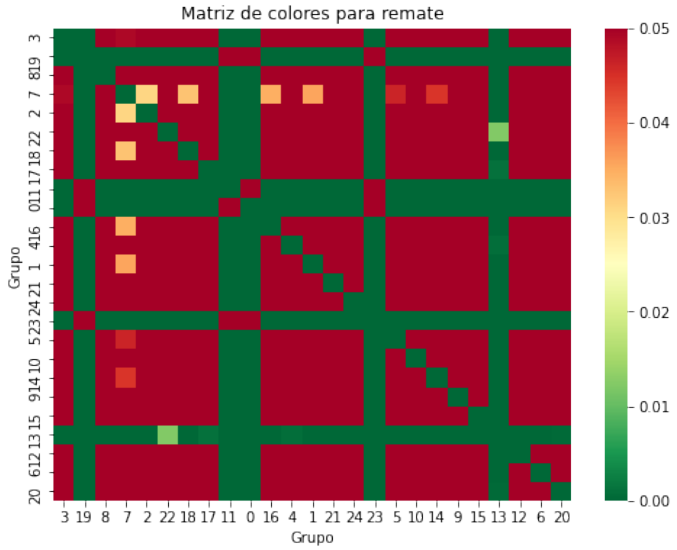

Figura 5.4: Estudio de las diferencias entre clústeres para la variable remate mediante la prueba de Dunn

los artículos que discuten la influencia de la violación de los supuestos de esta técnica, Feir-Walsh and Toothaker (1974) aporta evidencias a favor del uso de ANOVA cuando se violan estos supuestos. En este caso, de nuevo, se ha decidido aceptar este riesgo y se ha comprobado la hipótesis de homocedasticidad, con la prueba de Levene y visualmente.

Lo primero a destacar en esta prueba es que se estudiarán los resultados de tres tipos diferentes de clustering K-Means, HDBSCAN y Clustering aglomerativo. Lo segundo a destacar es que, como se observa en la tabla 5.2, entre los grupos del clustering de K-Means se puede suponer homocedasticidad sin ningún tipo de problema. Por otro lado, se puede suponer homocedasticidad entre los grupos resultado del clustering aglomerativo a excepción de las cuatro variables resaltadas (*Remates, num_faltas, pageRank y tiros_largos*). Los gráficos de caja y bigotes, por su parte, verifican esta hipótesis (figura 5.6). Además, en este caso, la prueba de Levene indica que las posibles violaciones del supuesto de la homocedasticidad son leves, con p-valores cercanos al $\alpha = 0{,}05$ en las cuatro variables a destacar, por lo que estas posibles violaciones no resultan un problema particularmente importante. Finalmente, sí que destacan otras cuatro variables (Elaboradas, densidad, duelos, pases_clave) entre los grupos resultado de HDBSCAN que presentan grandes indicios de tener diferencias estadísticamente significativas entre sus varianzas. Esto se debe a que los grupos presentan un gran desbalance en el número de muestras y los resultados de la prueba t de Student deberán ser examinados con extrema precaución.

Por tanto, tras las debidas justificaciones aportadas, se ha decidido aplicar MANOVA para la comparación de los grupos resultados de los clusterings aglomerativo y K-Means. El resultado de la aplicación de HDBSCAN se validará mediante la prueba T-Student (Student, 1908) que compara dos medias bajo los mismos supuestos del ANOVA.

Los dos MANOVA aplicados han lanzado un p-valor de 0, menor que el $\alpha = 0{,}05$ considerado. Por lo que se concluye que al menos uno de los clústeres obtenidos tiene una diferencia estadísticamente significativa respecto del resto para los clusterings aglomerativo y K-Means. De nuevo, para estudiar de qué variables y entre qué clústeres se observan estas diferencias será necesario aplicar un estadístico *post-hoc*. Este estadístico será la prueba de Dunn, que será aplicada de forma análoga a la utilización de esta para el clustering de equipos. Es decir, se creará una matriz para cada variable donde se podrán comparar cada uno de los grupos y se coloreará en función del nivel de significancia de

Test Saphiro-Wilk	
Variable	**P-valor ($\alpha = 0{,}05$)**
Centro	0.8133
Contragolpes	0.2120
Derecha	0.1023
Elaboradas	0.0000
Izquierda	0.1327
Mixtas	0.5680
Picos	0.2949
Remates	0.0002
densidad	0.0000
duelos	0.1179
longitud_max	0.0015
longitud_pase	0.7699
num_faltas	0.4707
num_intercepciones	0.0540
num_paradas	0.1102
pageRank	0.2172
pases_clave	0.0000
tiros_cortos	0.0000
tiros_largos	0.0082

Tabla 5.1: Tabla de p-valores para las variables. Los p-valores $\leq 0{,}05$ están resaltados en rojo.

Test de Homocedasticidad (Levene)			
Variable	**aglomerativo**	**K-Means**	**HDBSCAN**
Centro	0.4341	0.4547	0.0746
Contragolpes	0.0775	0.4609	0.0715
Derecha	0.7552	0.6321	0.6235
Elaboradas	0.4521	0.1465	0.0115
Izquierda	0.3507	0.4093	0.1471
Mixtas	0.7477	0.0970	0.5772
Picos	0.3999	0.1362	0.1191
Remates	0.0457	0.1922	0.2394
densidad	0.5427	0.0698	0.0034
duelos	0.7132	0.8340	0.0201
longitud_max	0.7604	0.4810	0.5946
longitud_pase	0.3629	0.3474	0.9139
num_faltas	0.0375	0.9002	0.1701
num_intercepciones	0.4714	0.8732	0.8839
num_paradas	0.1656	0.1095	0.4282
pageRank	0.0361	0.9859	0.8770
pases_clave	0.0834	0.0785	0.0214
tiros_cortos	0.4260	0.2393	0.1018
tiros_largos	0.0453	0.2801	0.0967

Tabla 5.2: P-Valores para las variables con $\alpha = 0{,}05$, p-valores asociados a los grupos derivados de la aplicación de los tres tipos de clustering. Los p-valores $\leq 0{,}05$ están resaltados en rojo.

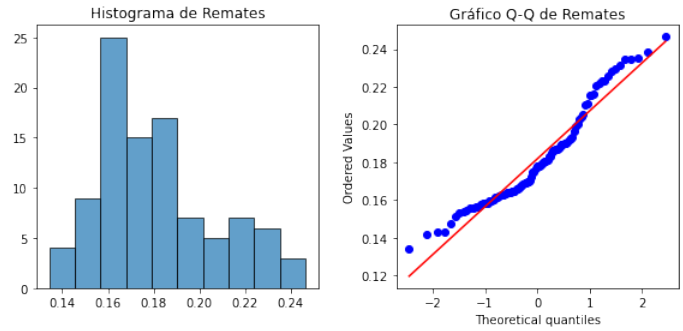

Figura 5.5: Estudio de la normalidad para la variable Remates

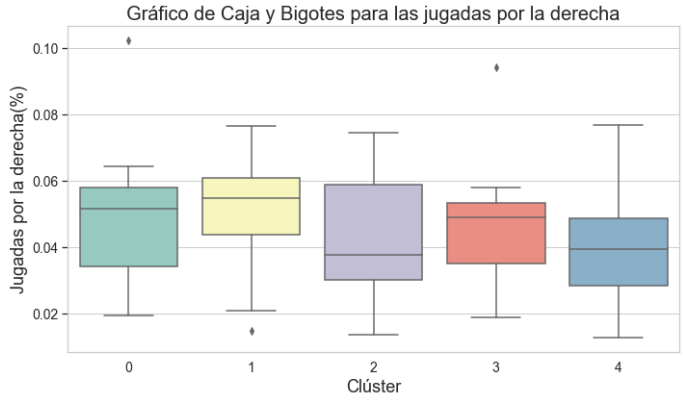

Figura 5.6: Estudio de la homocedasticidad para la variable Derecha

la propia prueba. Cabe destacar, que los propios MANOVA han sido, a su vez, validados mediante el estudio de la normalidad y homocedasticidad de los residuos. Los resultados fruto del estudio de las diferencias entre las agrupaciones se debatirá más a fondo en el siguiente capítulo.

Para concluir, la validación de los resultados del clustering del HDBSCAN será realizada mediante la prueba T-Student. Como se mencionó anteriormente, la prueba t compara la diferencia entre las medias de dos grupos con respecto a la variabilidad dentro de cada grupo. Evalúa si la diferencia observada entre las medias es lo suficientemente grande como para ser estadísticamente significativa o si podría deberse al azar. En esta ocasión, puesto que sólo se distinguen dos clústeres la propia validación de los resultados es información clave para los resultados del trabajo por lo que será descrita más a fondo en el siguiente capítulo.

CAPÍTULO 6
Resultados

En este capítulo se presentarán los resultados derivados de todo el proceso descrito hasta ahora en este trabajo. Se aportarán visualizaciones y hallazgos clave y se añadirán las pertinentes interpretaciones derivadas del estudio.

6.1 Clustering de posesiones

El principal objetivo de este primer clustering es claro: identificar los patrones de posesiones más comunes en las cinco grandes ligas. Para la extracción de estos patrones se han considerado cuatro tipos de características de una posesión: su longitud, las coordenadas inicial y final del primer y último evento, las zonas del campo en las que se ha desarrollado y si acabó en remate o no. Cualquier diferencia que pueda existir entre alguna de estas características es susceptible de dar lugar a un tipo de posesión diferente, por lo que cabe esperar que exista un número de grupos relativamente grande. Sin embargo, y como se mencionó en el apartado 5.1, resulta casi imposible encontrar dos posesiones iguales y agruparlas requiere de cierta arbitrariedad ya que la escala de grises es prácticamente infinita y las líneas divisorias no están nada claras. Este aspecto ha sido considerado en la experimentación mediante la asignación de hiperparámetros cuyos valores sean más propicios a lanzar grupos de las características más o menos esperadas, o en su defecto, evitar que los grupos no cumplan con los estándares deseables.

En este sentido, el clustering cuyos resultados se aproximan más a las expectativas ha sido el K-Means. Las tablas 6.1 y 6.2 muestran los mejores parámetros de las dos configuraciones testeadas, UMAP-K-MEANS y UMAP-HDBSCAN. Con sendas selecciones de parámetros, los resultados son bastante dispares, por un lado, K-Means contempla 25 clústeres, que varían en tamaño desde las 7000 posesiones hasta las 500, por otro HDBSCAN contempla 5 clústeres, con aproximadamente 18000 posesiones en el mayor y 14 en el más pequeño. Sin embargo, al explorar individualmente los cada uno de los grupos resultantes, se observa que el K-Means consigue dividir mucho mejor las posesiones en base a las características contempladas. Un dato interesante para resaltar es que el clúster realizado con HDBSCAN detecta 1288 puntos como datos anómalos.

Con las condiciones planteadas, se optará por trabajar con los resultados del K-Means ya que tienen mayor potencial para ajustarse a la división de las posesiones que se persigue obtener en este trabajo. A continuación, se procederá con el análisis y caracterización de cada uno de los clústeres resultado. Además, se acompañará la explicación de una visualización de la posesión centroide del clúster correspondiente. Esta visualización está disponible para todos los clústeres en el anexo A

UMAP-HDBSCAN	
Puntuación combinada	0.258323
Duración	00:06:02.685698
cluster_selection_epsilon	0.224258
min_cluster_size	315
min_samples	118
n_neighbors	35
n_components	4
min_dist	0.221282
metric	manhattan

Tabla 6.1: Parámetros y resultados de UMAP-HDBSCAN.

UMAP-K-Means	
Puntuación combinada	0.178609
Duración	00:05:41.777524
distance	manhattan
n_clusters	25
n_neighbors	30
n_components	3
min_dist	0.243687
metric	manhattan

Tabla 6.2: Parámetros resultado de UMAP-K-Means.

- **Clúster 0:** Este clúster engloba jugadas muy directas que concentran gran parte de sus eventos en la zona delantera del ataque, especialmente en el centro.
- **Clúster 1:** Este clúster engloba jugadas que se originan en el medio del campo en zona centro-izquierda y avanzan hacia la zona derecha del ataque.
- **Clúster 2:** Este clúster engloba jugadas que se originan en el medio y avanzan hacia la zona izquierda del ataque.
- **Clúster 3:** Este clúster engloba jugadas más elaboradas que inician en el centro de la zaga (incluso desde el portero) y avanzan hasta acabar por la zona delantera derecha.
- **Clúster 4:** Este clúster engloba jugadas muy cerradas y enfocadas en la esquina izquierda del campo, generalmente para llegar a línea de fondo, figura 6.1.
- **Clúster 5:** Este clúster engloba jugadas más elaboradas que inician en el centro de la zaga (incluso desde el portero) y avanzan hasta acabar por la zona delantera izquierda, figura 6.2.
- **Clúster 6:** Este clúster engloba jugadas muy enfocadas en el carril derecho que se originan en la zona de la defensa o el medio del campo, figura 6.3.
- **Clúster 7:** Este clúster engloba jugadas más directas originadas en medio del campo y que cuentan con un elevado número de acciones en la delantera y por el carril del centro.
- **Clúster 8:** Este clúster engloba jugadas por los carriles centro e izquierda del campo.

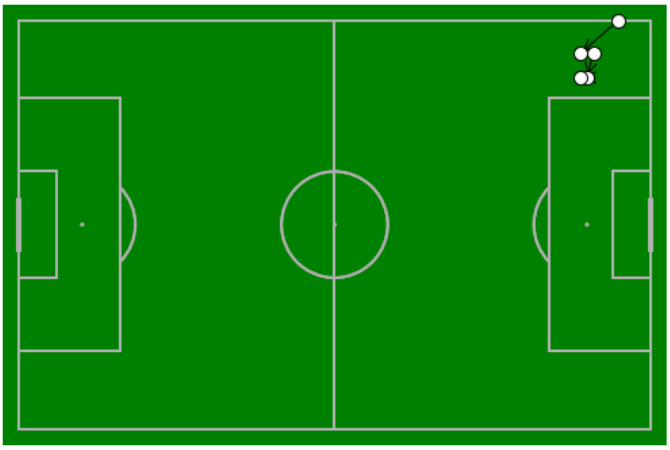

Figura 6.1: Posesión centroide del Clúster 4

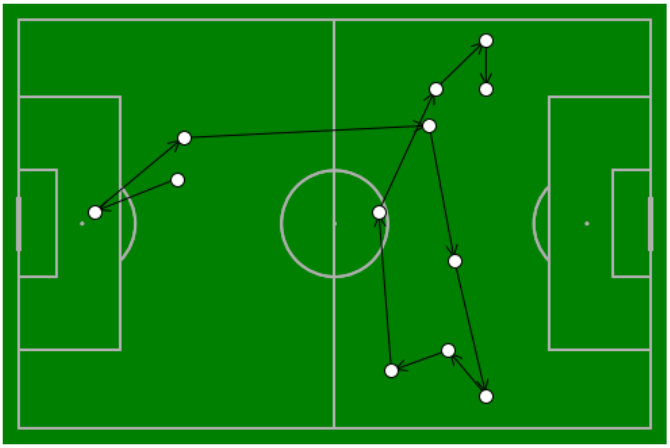

Figura 6.2: Posesión centroide del Clúster 5

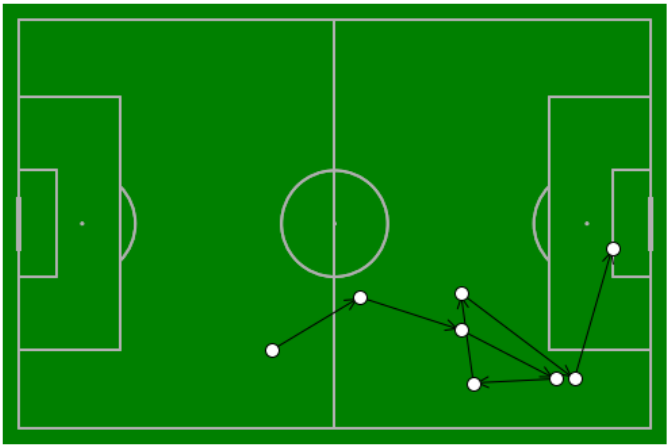

Figura 6.3: Posesión centroide del Clúster 6

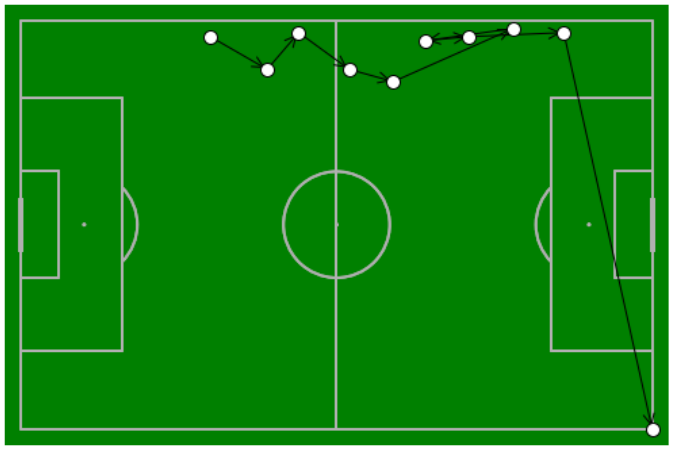

Figura 6.4: Posesión centroide del Clúster 12

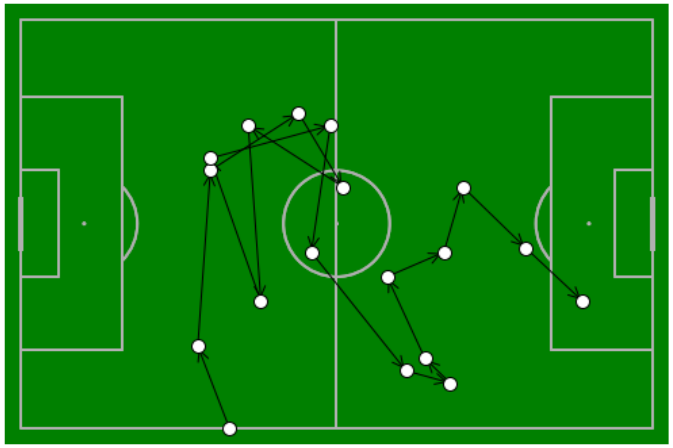

Figura 6.5: Posesión centroide del Clúster 16

- **Clúster 9:** Este clúster engloba jugadas nacidas en el medio del campo, pero que tienen un gran influencia en la esquina derecha.
- **Clúster 10:** Este clúster engloba jugadas por los carriles centro y derecha del campo.
- **Clúster 11:** Este clúster engloba jugadas originadas en el medio del campo con rápida transición al ataque pero por ambas bandas.
- **Clúster 12:** Este clúster engloba jugadas muy enfocadas en el carril izquierdo que se originan en la zona de la defensa o el medio del campo, figura 6.4.
- **Clúster 13:** Este clúster engloba jugadas de transición muy rápida desde la defensa hasta el ataque.
- **Clúster 14:** Este clúster engloba jugadas de elaboración en zona de tres cuartos de campo sin apenas profundidad.
- **Clúster 15:** Este clúster engloba jugadas enfocadas en zona delantera izquierda pero originadas en el carril derecho.
- **Clúster 16:** Este clúster engloba jugadas muy elaboradas, moviendo la pelota de izquierda a derecha del campo, esencialmente en el medio del campo, figura 6.5.
- **Clúster 17:** Este clúster engloba jugadas cortas por el carril izquierdo muy cortas y con poca importancia en general.

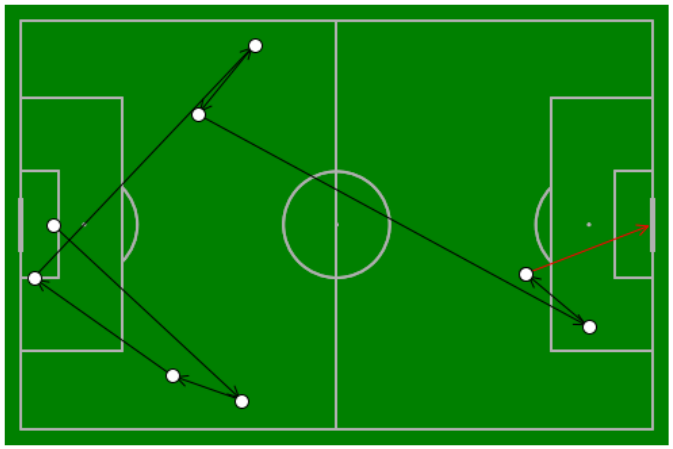

Figura 6.6: Posesión centroide del Clúster 19

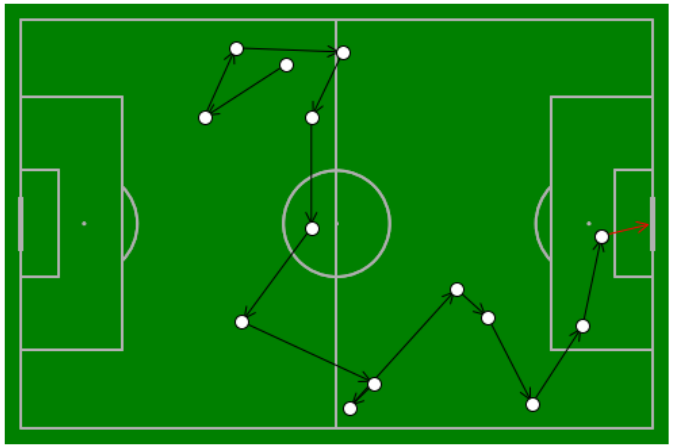

Figura 6.7: Posesión centroide del Clúster 23

- **Clúster 18:** Este clúster engloba jugadas algo elaboradas que parten desde el medio centro del campo hasta la zona delantera derecha del ataque.
- **Clúster 19:** Este clúster engloba jugadas de transición muy rápida desde la defensa hasta el ataque, figura 6.6.
- **Clúster 20:** Este clúster engloba jugadas muy cerradas y enfocadas en la esquina derecha del campo, generalmente para llegar a línea de fondo.
- **Clúster 21:** Este clúster engloba jugadas de transición muy rápidas que parten desde la zona de la defensa izquierda y avanzan por el carril central.
- **Clúster 22:** Este clúster engloba jugadas cortas por el carril derecho muy cortas y con poca importancia en general.
- **Clúster 23:** Este clúster engloba jugadas bastante elaboradas que acaba en remate y que suelen tener tendencia a la banda derecha, figura 6.7.
- **Clúster 24:** Jugada de movimiento del balón por el medio en los tres carriles pero poco elaborada, figura 6.8.

También es importante resaltar que la visualización proporcionada es simplemente un apoyo para la identificación de un clúster, el centroide no es más que una posible representación de las posesiones contenidas en el mismo. Sin embargo, no siempre debe

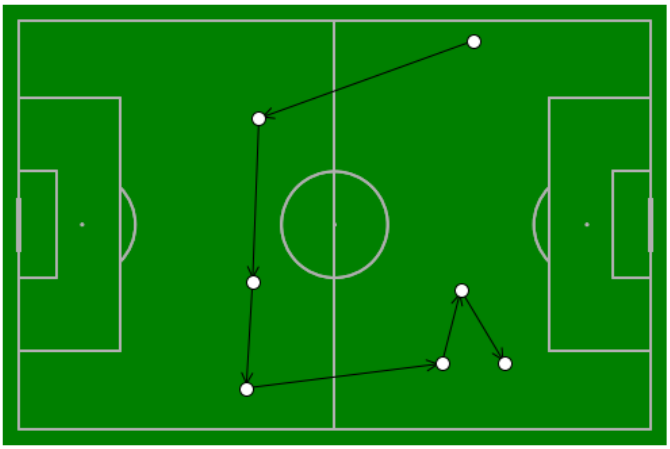

Figura 6.8: Posesión centroide del Clúster 24

que coincidir con la descripción propuesta ya que la descripción está realizada en base a un análisis individual de algunas de las posesiones del clúster y la comparación de las medias de cada una de las variables con respecto al resto. Esta comparación ha sido realizada mediante un escalado Min-Max. El escalado Min-Max, también conocido como normalización Min-Max, es una técnica utilizada a fin de transformar características numéricas, generalmente, entre 0 y 1, para obtener un intervalo uniforme en el que comparar sus valores. El proceso de escalado Min-Max se realiza utilizando la siguiente fórmula:

$$x_{\text{escalado}} = \frac{x - \min(X)}{\max(X) - \min(X)} \tag{6.1}$$

Donde:

x_{escalado} es el valor escalado de la característica.
x es el valor original de la característica.
$\min(X)$ es el valor mínimo de la característica en el conjunto de datos.
$\max(X)$ es el valor máximo de la característica en el conjunto de datos.

La normalización Min-Max mapea los valores originales al intervalo [0, 1], donde el valor mínimo se asigna a 0 y el valor máximo se asigna a 1. Los valores intermedios se escalan proporcionalmente dentro de ese rango. Esto permite comparar los valores de cada una de las medias de todas las variables entre los diferentes grupos de una forma simple. Si se aplica un proceso similar al aplicado con la prueba de Dunn, se obtiene una matriz por colores donde el color azul es asignado al 0 y el color rojo al 1. De este modo, se consigue una comparación más amigable entre los diferentes clústeres, figura 6.9.

Una vez identificados los patrones más comunes en las posesiones de las 5 grandes ligas europeas, el siguiente paso es utilizar esta información para aplicar un clustering a los equipos. No obstante, no parece tan interesante el hecho de contar con la información de cada uno de los clústeres, sino que basta con contemplar si el equipo vuelca su ataque por un carril específico, tiende a generar posesiones que acaben en remates o cuenta con un estilo de juego que combine diferentes tipos de posesiones. Por lo que se ha decidido reagrupar los clústeres en ocho categorías que resumen de forma certera las características requeridas. Aun así, la información obtenida en los 25 clústeres es muy valiosa y resultaría especialmente interesante plantear otro clustering que siga en la dirección de aprovechar la información desagregada de los mismos. El objetivo de este estudio sería

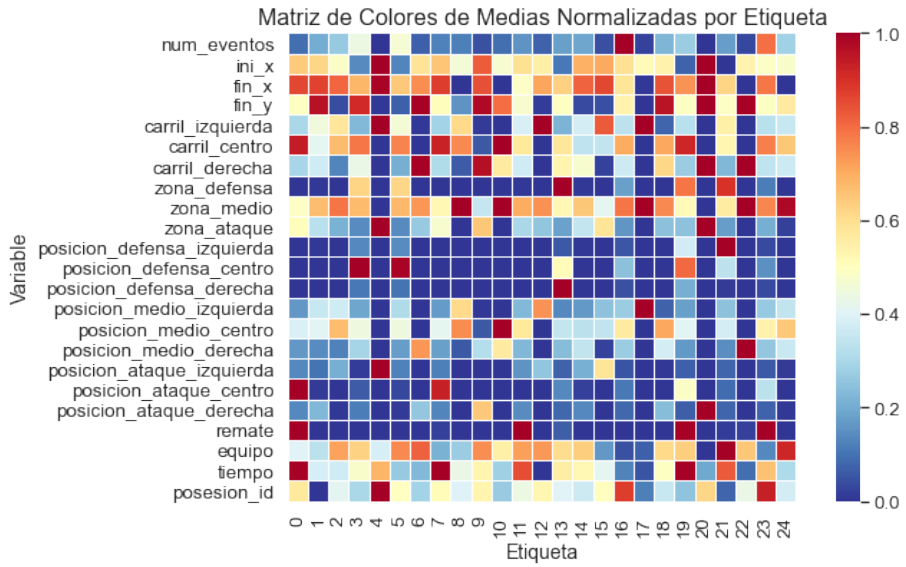

Figura 6.9: Matriz comparativa con escalado Min-Max para los clústeres de posesiones

realizar un análisis específico sobre las jugadas más comunes en cada equipo, pero lejos queda esto de los límites del presente trabajo. Prosiguiendo con la categorización, sólo resta mencionar que ha sido creada en función del conocimiento teórico en la materia y en concordancia con las características de los clústeres resultantes.

A continuación, se muestra la distribución de cada clúster entre las diferentes categorías:

- **Centro:** Posesiones que eminentemente se desarrollan por el carril central. Se han incluido el clúster 3, clúster 5, clúster 7, clúster 8 y clúster 10.
- **Izquierda:** Posesiones que eminentemente se desarrollan por el carril izquierdo. Se han incluido el clúster 12, clúster 15 y clúster 17.
- **Derecha:** Posesiones que eminentemente se desarrollan por el carril derecha. Se han incluido el clúster 6, clúster 9 y clúster 22.
- **Contragolpes:** Posesiones caracterizadas por una rápida transición defensa-ataque. Se han incluido el clúster 13, clúster 19 y clúster 21.
- **Picos o esquinas:** Posesiones caracterizadas por presentar una alta densidad de eventos cercanos a línea de fondo y por las bandas. Se han incluido el clúster 4 y clúster 20.
- **Remates:** Posesiones que son más propicias a acabar en un remate. Se han incluido el clúster 0, clúster 11 y clúster 23.
- **Elaboradas:** Posesiones muy largas que se desarrollan generalmente en medio del campo y por todos los carriles. Se ha incluido el clúster 16.
- **Mixtas:** Posesiones con combinación de carriles (p.ej. izquierda-centro) y que no pueden ser claramente englobadas en ninguna de las categorías previamente mencionadas. Se han incluido el clúster 1, clúster 2, clúster 14, clúster 18 y clúster 24.

Esta categorización facilitará la comprensión y el manejo de los resultados obtenidos en el clustering de las posesiones.

6.2 Clustering de equipos

El objetivo del segundo clustering será identificar los diferentes estilos de juego presentes en las cinco ligas y qué equipos se incluyen dentro de cada uno, además se estudiarán las posibles diferencias que existan entre las propias ligas. Para ello, se han recogido algunas características relacionadas con el rendimiento y la creatividad, estructura colectiva y categorías de posesión de los clubes. La situación que se presenta tras la experimentación resulta particularmente diferente de la que se observaba en el clustering de posesiones ya que las tres configuraciones testeadas han originado resultados interesantes. A continuación, se describen cada uno junto con las diferencias encontradas entre los grupos y las implicaciones de los propios resultados.

6.2.1. Clustering HDBSCAN

En primer lugar, se expondrán los resultados del clustering realizado con HDBSCAN para los equipos. A continuación se muestran los parámetros de la mejor configuración encontrada para el clustering con HDBSCAN para los equipos (tabla 6.3). Es relevante destacar que el valor del parámetro *n_neighbors*, pues únicamente se están considerando tres vecinos. Esto significa que UMAP está prestando menos atención a los puntos cercanos en el espacio de alta dimensionalidad al mapearlos en un espacio de dimensionalidad menor. Es decir, resulta en una representación más simplificada y menos detallada de los datos originales, lo que equivale a una proyección más general de los mismos.

UMAP-HDBSCAN	
Puntuación combinada	0.269434
Duración	00:00:01.834798
n_neighbors	3
min_dist	0.155062
n_components	5
cluster_selection_epsilon	0.299053
metric	manhattan

Tabla 6.3: Parámetros resultado de UMAP-HDBSCAN para los equipos.

Se ha decidido comenzar por este ya que ha resultado ser la agrupación más simple (figura 6.10, con únicamente dos clústeres bien diferenciados, el clúster 0 con los «equipos élite» y el clúster 1 de «equipos regulares». Esto es debido a la proyección general de los datos que se discutía con anterioridad. Una rápida exploración entre los clubes englobados en cada uno de los grupos basta para percatarse de este hecho, en el que si se explora más a fondo se pueden encontrar algunas sorpresas. Por ejemplo, el Atlético de Madrid (2º clasificado de La Liga 2017/18) se encuentra englobado dentro del clúster de los equipos regulares. Para encontrar una explicación a este hecho será necesario atender a las diferencias entre ambos grupos, para lo cual se ha hecho uso de la prueba t de Student como se mencionó en el capítulo anterior.

Los resultados de esta prueba muestran que las variables para las cuáles se encuentran diferencias estadísticamente significativas entre ambos grupos son las siguientes:

- Contragolpes ($p - valor = 0$)
- Picos ($p - valor = 0{,}0474$)
- Remates ($p - valor = 0{,}0288$)

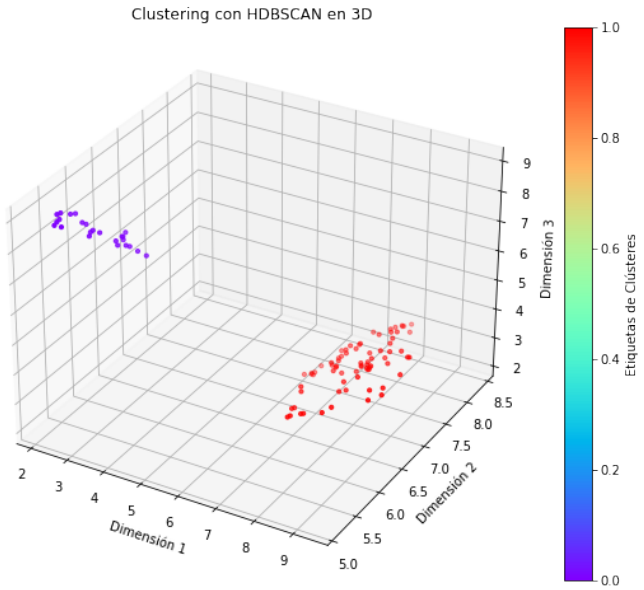

Figura 6.10: Visualización 3D de los equipos coloreados por etiqueta en el clustering HDBSCAN.

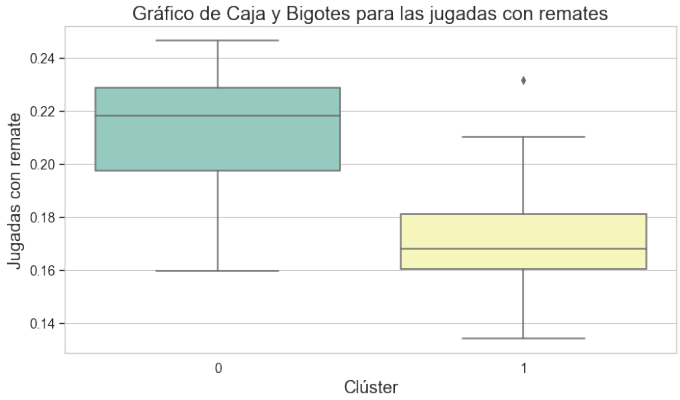

Figura 6.11: Boxplot comparación para los clústeres resultado de HDBSCAN para la variable Remates

- duelos ($p - valor = 0{,}0484$)
- num_faltas ($p - valor = 0$)

Es relativamente sencillo encontrar coherencia entre los resultados estadísticos y la explicación deportiva de este hecho ya que, generalmente, dentro de las ligas europeas encontramos un grupo pequeño de clubes que, por lo general consigue desarrollar un estilo de juego más ofensivo, como se puede apreciar en la cantidad de posesiones acabadas en remates, figura 6.11. Este tipo de juego, a pesar de encontrar diferencias entre los diferentes clubes, se caracteriza por ahogar al rival en su propio campo, sometiéndolo a sucesivos ataques. Es por este motivo que los clubes regulares recurren en mayor medida al uso de contragolpes o transiciones rápidas para aprovechar adelantamientos o desajustes defensivos de los equipos técnicamente superiores. Además, también recurren a jugadas más cortas, cercanas a línea de fondo, para colocar centros al área, que son el tipo de posesiones englobadas en los *Picos*. Además, otra diferencia clara que se encuentra son los remates, ya que el equipo superior se vuelca constantemente en el ataque y, por continuidad, consigue realizar un número de posesiones que acaben en remate significativamente superior. Finalmente, el número de duelos y faltas es superior en los equipos

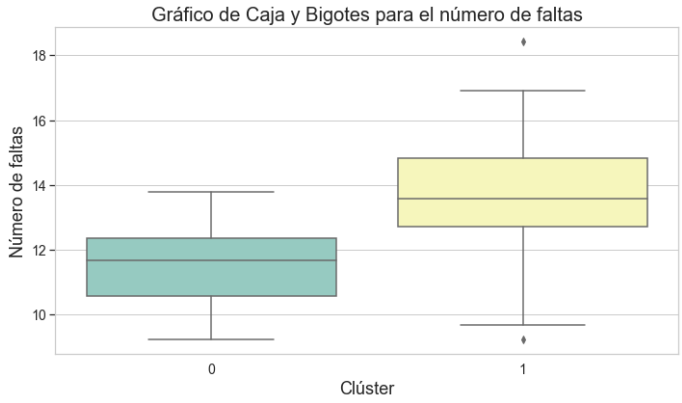

Figura 6.12: Boxplot comparación para los clústeres resultado de HDBSCAN para la variable num_faltas

más humildes como se aprecia en la figura 6.12, que intentan compensar la falta de calidad llevando el partido a un plano más físico. Es común apreciar como estos equipos hacen uso de las faltas para cortar situaciones potencialmente peligrosas o comprometidas para la defensa.

6.2.2. Clustering aglomerativo

Este segundo clustering obtenido mediante la técnica jerárquica de clustering aglomerativo, es posiblemente el más peculiar de las tres agrupaciones diferentes obtenidas para los equipos de las cinco grandes ligas europeas. En este caso, la mejor configuración obtenida en la experimentación se puede observar en la tabla 6.4. El valor que destaca es el *min_dist* pues resulta ser el mayor de las tres configuraciones del clustering de equipos, es decir, en este caso los clústeres resultantes estarán claramente separados y definidos, esto es rápidamente apreciable en la figura 6.13.

UMAP-AGLOMERATIVO	
Puntuación combinada	0.259102
Duración	00:00:02.906992
n_neighbors	9
min_dist	0.214693
n_components	3
n_clusters	5
linkage	single
metric	euclidean

Tabla 6.4: Parámetros resultado de UMAP-aglomerativo para los equipos.

Los resultados de este clustering han sido cinco grupos correspondientes, precisamente, a cada una de las ligas estudiadas. Sin más dilación, se presentan los clústeres resultado:

- Clúster 0: Premier League
- Clúster 1: Ligue 1
- Clúster 2: Bundesliga
- Clúster 3: La Liga

- Clúster 4: Serie A

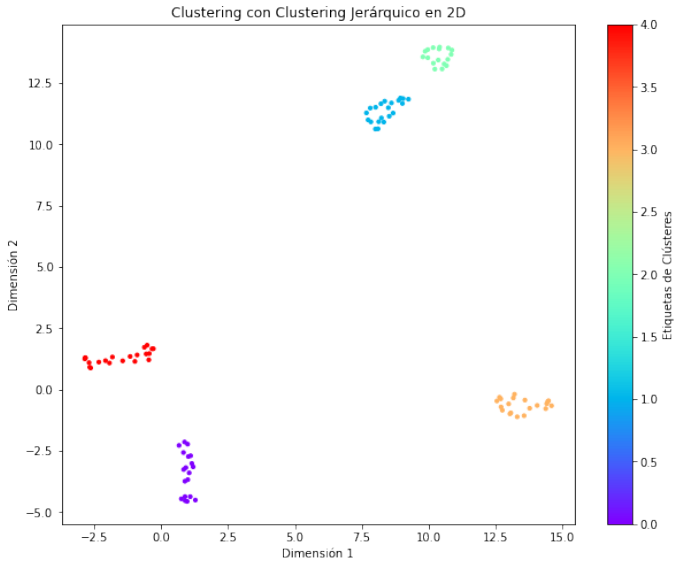

Figura 6.13: Visualización 2D de los equipos coloreados por etiqueta en el clustering Jerárquico aglomerativo.

Cabe destacar que la clasificación no es correcta al cien por cien, pues se encuentran algunos equipos incluidos en el clúster erróneo. Ejemplo de esto es el Swansea City, integrado dentro de la Ligue 1 o el Rasen Ballsport Leipzig, dentro de la Serie A. Sin embargo, el resto de los equipos presentan una clasificación totalmente correcta respecto a su liga, es por este motivo que resultará particularmente interesante el estudio de las variables que más ayudan a discriminar los equipos por sus ligas.

A continuación, se detallarán cada una de ellas y se proporcionará un posible razonamiento que dé sentido a los resultados obtenidos:

- Contragolpes: Los equipos de la Ligue 1 destacan sin duda alguna por ser aquellos que más posesiones categorizadas como contragolpes presentan. En el lado opuesto se encuentra la Premier League, liga cuyos equipos optan en menor medida por este tipo de jugadas. En un tercer grupo intermedio se encuentran el resto de las ligas sin diferencias significativas.

  Los equipos franceses son conocidos por su estilo de juego particularmente físico, donde las transiciones ofensivas son el punto fuerte de gran parte de los equipos, que encuentran mayor comodidad en la fase defensiva. Por otro lado, la Premier League, también caracterizada por su juego físico y directo presenta el menor porcentaje de jugadas de contraataque, esto puede ser debido al creciente aumento de la calidad técnica de los jugadores y a la influencia del juego de toque cuya bandera ostenta el Manchester City de Pep Guardiola.

- Mixtas: Aquí se encuentra La Liga claramente diferenciada del resto de ligas entre las cuáles no se aprecia ninguna diferencia importante. Los equipos de la liga tienden a elaborar jugadas haciendo uso de diversos carriles para preparar su ofensiva, la circulación del balón a lo ancho del terreno de juego permite a los equipos españoles permear en las rocosas defensas de los equipos más humildes de la categoría. En este sentido, la Ligue 1 presenta un patrón similar, pero sin una diferencia significativa respecto al resto.

- Picos: Si antes La Liga se presentaba protagonista, en esta ocasión es la Premier League la que toma el papel diferencial respecto al resto. El juego directo y las acciones individuales permiten a los equipos ingleses llegar a línea de fondo por las bandas con mayor frecuencia que los equipos del resto de ligas, generalmente, para poner el balón en el área para un remate con ventaja del delantero. La Liga y la Serie A son las que presentan este patrón con menor frecuencia.

- Remates: Las jugadas que acaban en remates permiten diferenciar entre la mayoría de las ligas estudiadas. En primer lugar se encuentra la Bundesliga, una liga con una gran flexibilidad táctica pero que viene destacando en los últimos años por el desarrollo de un juego particularmente ofensivo que hace que no sorprenda para nada este dato. En segundo y tercer lugar se encuentran la Serie A (con una variabilidad enorme) y la Premier League, respectivamente. La Ligue 1 y La Liga se colocan en cuarto y quinto puesto, pero por diferentes motivos. La Ligue 1, en general, ofrece un estilo de juego más defensivo y, por contraparte, La Liga presenta un estilo de juego algo más ofensivo, pero sin llegar a encontrar un remate claro en gran parte de las ocasiones. Esta comparación se aprecia de forma más clara en la figura 6.14.

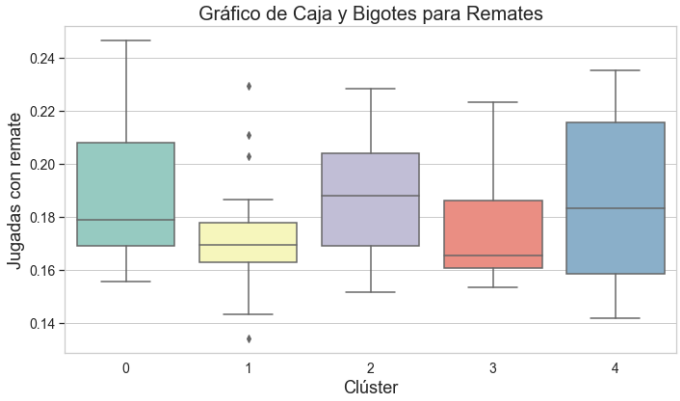

Figura 6.14: Boxplot comparación para los clústeres resultado de HDBSCAN para la variable Remates

- duelos: De nuevo, se encuentran los equipos ingleses a la cabeza, junto con los alemanes, son las dos ligas que buscan disputar mayor número de duelos, donde encontrar la superioridad ofensiva mediante acciones individuales, ya sea por superioridad técnica o haciendo gala del físico de los jugadores. En último lugar encontramos la Serie A, que, generalmente rechaza este tipo de duelos y suele optar por ataques y defensas mucho más posicionales.

- longitud_pase: Esta variable no ayuda claramente a distinguir entre grupos, sin embargo, destaca la comparación entre la Bundesliga y la Serie A. Pues, a pesar de encontrar diferencias estadísticamente significativas, estas diferencias son realmente pequeñas, siendo la liga alemana la que mayor tendencia presenta a realizar pases largos y la italiana encontrándose en mitad del ranking. Sin embargo, un aspecto a destacar interesante es que, la variabilidad que se encuentra en la Serie A es particularmente pequeña, es decir, podríamos afirmar que apenas existe diferencia entre la longitud de los pases de los equipos italianos. Esto no ocurre en el caso de la liga española, que presenta la mayor variabilidad de entre todas las ligas estudiadas.

- num_faltas: Esta variable es especial y debe ser estudiada con cautela. El número de faltas es, sin duda, una de las mayores diferenciadoras entre ligas. No obstante, las faltas dependen de la subjetividad de los árbitros de la liga correspondiente;

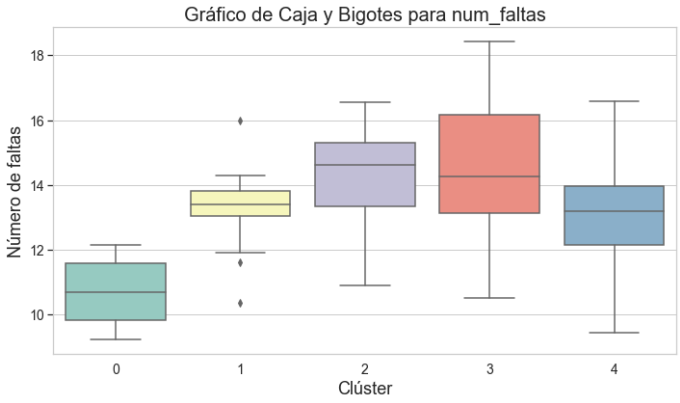

Figura 6.15: Boxplot comparación para los clústeres resultado del clustering aglomerativo para la variable num_faltas.

es decir, el número de faltas es una variable que claramente muestra las diferencias culturales que existen a nivel futbolístico entre las grandes ligas. La Premier League es la liga con diferencia que menos falta registra (figura 6.15), pero también tiene fama de ser una liga muy permisiva, donde no son sancionadas los contactos que en otras ligas sí serían. Por tanto, para tener una referencia real de qué equipos tienden a provocar mayor número de faltas se deberían comparar en el territorio más neutral posible, es decir, en competiciones europeas.

- num_intercepciones: Poco que destacar de esta variable salvo la comparación entre la Serie A y La Liga, las ligas que presentan mayor y menor número de intercepciones respectivamente. La defensa posicional italiana favorece este tipo de estadísticas, mientras que en la liga no es tan común encontrar equipos que prioricen este tipo de movimientos.

- tiros_largos: La última variable a destacar en las comparativas. La Serie A y la Ligue 1, conocidas por ser más defensivas son las que presentan un mayor número de disparos desde fuera del área. Este tipo de disparos es otra opción para intentar anotar goles cuando el equipo rival presenta una defensa contundente y disciplinada. La Liga, de nuevo, vuelve a destacar por ser aquella cuyos equipos presentan menor número de disparos desde fuera del área. Como se mencionó anteriormente, los equipos españoles suelen optar por una circulación del balón rápida entre diferentes carriles para crear huecos en las defensas.

El clustering aglomerativo ha aportado unos resultados muy interesantes, a la par que coherentes con la realidad. Cada una de las ligas ha podido ser descrita en base a las variables que presentaban diferencias estadísticamente significativas para alguno de los grupos según la prueba de Dunn. Se ha puesto el foco en las diferencias entre las diferentes ligas y se ha propuesto una posible explicación deportiva a los resultados estadísticos, que, sin lugar a duda, han apreciado diferencias significativas entre cada una de las cinco grandes ligas.

6.2.3. Clustering K-Means

Para finalizar el estudio, se explorarán los resultados del clustering K-Means. Los parámetros escogidos para este último clustering pueden ser apreciados en la tabla 6.5. De esta configuración destaca el número de componentes de UMAP, con un valor de seis, se trata del mayor valor obtenido a lo largo del estudio. Esto implica que este es el

clustering es en el que se ha aplicado una menor reducción de la dimensionalidad y, por ende, mayor cantidad de información será considerada para el clustering.

UMAP-K-MEANS	
Puntuación combinada	0.18961
Duración	00:00:02.599903
n_neighbors	12
min_dist	0.047268
n_components	6
n_clusters	14
metric	euclidean
distance	manhattan

Tabla 6.5: Parámetros resultado de UMAP-K-Means para los equipos.

Las agrupaciones obtenidas en este caso no corresponden a una clasificación tan clara como en los dos descritos anteriormente. Sin embargo, no por ello se presentan con menor interés, pues podría decirse que se trata de la clasificación más compleja encontrada en este trabajo ya que se trata de la que más dimensiones contempla.

Es por este motivo que los grupos resultado tienen unas características mucho más específicas y definidas que en los casos anteriores. Por tanto, se volverá a recurrir a la visualización de la matriz de las medias con escalado Min-Max (figura 6.16). En este caso, la división ha sido producida en base a lo que se podría entender como estilo de juego en el sentido en el que fue planteado originalmente este estudio. Con esto se pretende dar a entender que para la clasificación intervendrán la mayor parte de las variables propuestas y sus diferentes combinaciones. A continuación, se detallan las características de cada uno de los clústeres junto un equipo de ejemplo:

- **Clúster 0**: Equipos muy defensivos, con alto número de duelos disputados, con el recurso de tiros de larga distancia y que también atacan en contraataque por las bandas. Rechazan el uso de jugadas elaboradas. Un equipo ejemplo de este clúster es el Hannover 96.

- **Clúster 1**: Equipos muy ofensivos, con alto porcentaje de jugadas elaboradas con gran número de pases clave. Destacan por disparar al arco con gran frecuencia, recuperar el balón rápidamente y presentar una estructura de equipo muy densa, con una gran cantidad de pases entre sus jugadores. Se trata de los mejores equipos de las ligas. Un equipo ejemplo de este clúster es el Real Madrid CF.

- **Clúster 2**: Equipos muy defensivos que hacen uso de posesiones categorizadas como mixtas, que tienden a acabar por alguna de las bandas en línea de fondo. Un equipo ejemplo de este clúster es el Levante UD.

- **Clúster 3**: Equipos muy defensivos, que apuestan por contragolpes y ataques por el carril central como arma ofensiva. Un equipo ejemplo de este clúster es el Lille OSC.

- **Clúster 4**: Equipos que apuestan por el ataque por las bandas con acciones individuales. Un equipo ejemplo de este clúster es el Leicester City FC.

- **Clúster 5**: Muy similar al Clúster 4. Un equipo ejemplo de este clúster es el Club Atlético de Madrid.

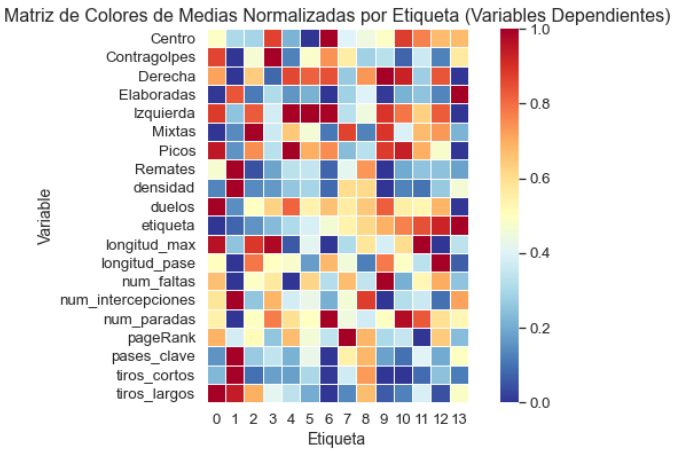

Figura 6.16: Matriz comparativa con escalado Min-Max para los clústeres de equipos con K-Means

- **Clúster 6**: Equipos muy defensivos, con un gran número de paradas, pero que apuestan por ataques directos por cualquiera de los tres carriles, especialmente el central. Un equipo ejemplo de este clúster es el Stoke City FC.
- **Clúster 7**: Equipos que apuestan por jugadas mixtas y que destacan por tener un jugador más importante, con alta participación en el juego colectivo. Un equipo ejemplo de este clúster es el Valencia CF.
- **Clúster 8**: Clúster similar al 1, pero con equipos de menor calidad. Un equipo ejemplo de este clúster es la Real Sociedad de Fútbol.
- **Clúster 9**: Equipos que vuelcan sus ataques por las bandas, especialmente la derecha y que, a pesar de apenas realizar remates, presentan una gran cantidad de duelos y una longitud de pases ligeramente mayor al resto. Destacan por realizar un gran número de faltas. Un equipo ejemplo de este clúster es el Getafe CF.
- **Clúster 10**: Similar al Clúster 6, pero con la excepción de no optar por ataques por el carril central. Un equipo ejemplo de este clúster es el Everton FC.
- **Clúster 11**: Equipos débiles, muy poco conectados y que reciben una gran cantidad de remates. Un equipo ejemplo de este clúster es la UD Las Palmas.
- **Clúster 12**: Equipos defensivos que apuestan por pases largos para la salida de balón con transiciones rápidas. Prefieren volcar su ofensiva por las bandas. Un equipo ejemplo de este clúster es el Athletic Club Bilbao.
- **Clúster 13**: Equipos que hacen un gran uso de posesiones elaboradas y tiros cortos, no realizan ataques directos por los carriles. Un equipo ejemplo de este clúster es el Real Betis Balompié.

Esta clasificación hace uso de la gran mayoría de las variables planteadas para el clustering, atendiendo a los p-valores de la prueba de Dunn, podemos encontrar diferencias estadísticamente significativas entre al menos dos de los grupos para cada una de las variables. En este sentido encontramos variables como Picos, que presentan diferencias entre casi todos los grupos y otras como num_paradas que no permiten realizar una diferenciación tan severa. Un dato peculiar para destacar es que resulta complicado encontrar un clúster que destaque en el uso de posesiones por el carril izquierdo o por el carril diestro, es decir que presente una asimetría en el ataque. Una desventaja que se encuentra en este clustering es que no logra encontrar una clara división entre el estilo

de juego de los clubes más importantes, pues pone el foco en las diferencias encontradas entre el grupo mayoritario de equipos más humildes.

Cuando se estudian los equipos en competiciones ligueras, el desnivel de los equipos más importantes respecto de los más humildes es tal que el clustering no es capaz de diferenciar entre estilos de juego que podríamos considerar diferentes como son el del Real Madrid y el del Manchester City. A la hora de realizar comparaciones, la diferencia que puede haber entre los dos equipos anteriormente citados es mínima en comparación a la que se encuentra respecto de los equipos de mitad baja de tabla. Sería interesante realizar este tipo de estudios exclusivamente entre equipos de élite y exclusivamente entre equipos humildes para llegar al fondo de la cuestión y encontrar las diferencias esperables entre equipos de juego que presentan estilos de juego dispares, pero que el clustering realizado no ha sido capaz de plasmar.

Además, otro detalle a tener en cuenta es que, en las competiciones ligueras, la mayor parte de las jornadas que disputan los equipos de élite son contra equipos que, por necesidad, deben realizar un esfuerzo extra en fase defensiva a costa de sacrificar la ofensiva. Es decir, los equipos grandes disputan la gran mayoría de los partidos sometiendo a su rival en su propio campo y apenas tienen necesidad de defender. Esto implica que equipos que son especialmente potentes en transiciones como el Real Madrid, apenas puede explotar esta característica porque el equipo rival pocas veces tendrá la capacidad de encerrarlo en su propio campo. Con esta reflexión se pretende dar a entender que el estudio del estilo de juego de cualquier equipo realizado en este trabajo debe ser contextualizado dentro de la competición liguera correspondiente. Para comparar el estilo de juego entre los equipos de élite sería más idóneo hacer uso de enfrentamientos directos que puedan darse en competiciones como puede ser la Champions League.

En conclusión, a lo largo de este capítulo se han presentado tres posibles agrupaciones en función de las variables propuestas. Estas agrupaciones han resultado en dividir los equipos por calidad, por liga y por estilo de juego. Entre todas ellas se han encontrado diferencias estadísticamente significativas, es decir, se puede afirmar que hay dos grupos abstractos de equipos, uno de equipos de élite y otro de equipos más modestos, el primero suele tender a presentar un comportamiento mucho más ofensivo en el campo y el segundo algo más defensivo. Por otro lado, también se han encontrado diferencias significativas entre los estilos de juegos propios de cada una de las cinco grandes ligas, cada una caracterizada por una o varias de las variables propuestas. Finalmente, se ha presentado una posible división en estilos de juego en la que englobar cada uno de los equipos en base a las características estudiadas.

CAPÍTULO 7
Conclusiones

En este capítulo se presentan las conclusiones alcanzadas tras el estudio en relación con los objetivos especificados. Además, se incluirán tanto una pequeña valoración personal, como las limitaciones y el trabajo futuro propuesto para continuar con el estudio.

7.1 Conclusiones y objetivos

Todo el desarrollo de este proyecto, desde el planteamiento inicial hasta la extracción de resultados deben ser evaluados y discutidos en función a los objetivos iniciales planteados. Comenzando por la identificación de los patrones de posesión, se han identificado satisfactoriamente 25 patrones de movimiento de la pelota dentro del terreno de juego en los que englobar cada una de las posesiones estudiadas. Estos patrones han sido descritos y se ha propuesto una posesión de ejemplo a modo de representación visual del clúster en cuestión. La división ha resultado satisfactoria, y se ha decidido englobar, a su vez, cada uno de estos patrones en ocho categorías a fin de simplificar el estudio: *Centro, Izquierda, Derecha, Contragolpes, Picos, Remates, Elaboradas y Mixtas*. Esta nueva agrupación ha sido realizada de forma arbitraria en base a las características de cada clúster y el conocimiento en la materia.

Una vez recogidas todas las características especificadas para definir el estilo de juego de un equipo, se ha procedido a agrupar cada uno de los equipos en función de estas características. En este sentido, se pueden extraer tres conclusiones derivadas de los tres clusterings aplicados. En primer lugar, existen diferencias estadísticamente significativas entre los estilos de juego que presentan los equipos de cada una de las cinco grandes ligas. Es decir, cada liga posee un estilo de juego general que puede caracterizarla. En segundo lugar, se encuentran también diferencias estadísticamente significativas entre el estilo de juego que proponen los equipos de élite y los equipos más humildes. Y finalmente, también se pueden diferenciar 13 estilos de juego más concretos, en los que pueden ser englobados los 98 equipos de las cinco grandes ligas. Estos 13 estilos de juego pueden ser perfectamente caracterizados y se ha propuesto un equipo de ejemplo de ese estilo de juego. Sin embargo, este clustering no fue capaz de encontrar diferencias significativas entre los estilos de juego de los equipos de élite.

Finalmente, en cuanto a la discusión sobre la metodología presentada, se considera que cumple satisfactoriamente con los requisitos deseados, a lo largo de esta memoria se han detallado uno por uno los pasos realizados hasta llegar al estudio de los resultados a fin de que pueda ser replicada por cualquier tipo de persona interesada. Se ha considerado exitoso este objetivo ya que el sistema propuesto ha sido capaz de dar respuesta de forma satisfactoria al resto de objetivos, además se ha hecho uso de la combinación de diferentes técnicas, desde técnicas propias del ámbito del *Big Data* y la informática

como es el clustering hasta técnicas de la estadística más clásica como la prueba t de Student, pasando por modelización de grafos incluso. El propósito consistía en integrar datos provenientes de diversas fuentes para generar un análisis exhaustivo, lo que permite clasificar el enfoque propuesto como innovador.

7.2 Aportación personal

A nivel personal, estoy muy satisfecho con el trabajo realizado, pero soy consciente de que el estudio presenta una metodología muy abierta que tiene, sin lugar a dudas, espacio a mejoras. Como ya he mencionado, el objetivo era incorporar diferentes metodologías encontradas en el estado del arte para crear un enfoque más completo y novedoso que permita incluir las ventajas de cada uno de los ámbitos en el estudio del estilo de juego. Sin embargo, las limitaciones encontradas a lo largo del trabajo han impedido que pueda desarrollar este enfoque con tanto rigor estadístico como me hubiese gustado. En adición a esto, este tipo de estudios requieren de la toma de decisiones arbitrarias en base al conocimiento de la materia ya que la definición de una posesión o del estilo de juego no es objetiva.

Aun así, este proyecto me ha aportado una gran cantidad de conocimientos sobre herramientas estadísticas, teoría de grafos y otros ámbitos que considero han sido muy enriquecedores para mi carrera como científico de datos. He tenido la oportunidad de aplicar todos los conocimientos adquiridos en la carrera sobre estadística (en las asignaturas de Análisis exploratorio de datos, Modelos estadísticos para la toma de decisiones I y Modelos estadísticos para la toma de decisiones II), sobre Python (Fundamentos de programación y Programación), sobre clustering (Modelos descriptivos y predictivos I) y sobre teoría de grafos (Matemática discreta y Modelado discreto).

7.3 Limitaciones

A pesar de los resultados positivos de este trabajo, es importante destacar algunas limitaciones que afectaron su desarrollo y resultados:

- **Limitación de Tiempo:** El tiempo disponible para llevar a cabo este proyecto fue limitado. Una mayor cantidad de tiempo podría haber permitido una investigación más exhaustiva y la realización de experimentos adicionales y la implementación de un mayor rigor estadístico. Este trabajo abre varias puertas para continuar en diferentes direcciones desde diferentes puntos a fin de cumplir otro tipo de objetivos.
- **Limitación de Recursos:** Los recursos disponibles, principalmente computacionales han supuesto un claro cuello de botella de cara, esencialmente, a la exploración de configuraciones en la experimentación. Una mayor cantidad de recursos computacionales hubiese aligerado este proceso y hubiese evitado estancamientos temporales debido a los largos periodos de espera para la obtención de resultados.
- **Limitación de Datos:** A pesar de contar con un vector de características que se ha considerado satisfactorio, cabe explorar otro tipo de variables como la posición del resto de jugadores de campo en el momento del evento. Carecer de datos relacionados con otro tipo de variables implica que estas no pueden ser tenidas en cuenta para el estudio.
- **Limitación de Herramientas:** Las herramientas y tecnologías utilizadas pueden contar con limitaciones específicas, a pesar de contar con una amplísima variedad

de funcionalidades, Python aún cuenta con algunos puntos flacos para la implementación de ciertas técnicas estadísticas.

- **Limitación de Conocimientos Previos:** A pesar de haber realizado una revisión del marco teórico, la falta de experiencia en el ámbito y el desconocimiento de algunas técnicas utilizadas han provocado ralentizaciones y posibles fallos metodológicos en el desarrollo del proyecto.

Estas limitaciones son comunes en muchos proyectos de investigación y deben ser consideradas al evaluar la validez y la aplicabilidad de los resultados.

7.4 Legado

La relevancia académica de este trabajo ha radicado en la inclusión de metodologías propias de diferentes ámbitos, incluyendo características que no suelen estar presentes en el estado del arte a la hora de analizar el estilo de juego. Mediante el presente estudio se ha pretendido abrir una puerta a la ampliación de horizontes en el análisis deportivo en general y en el análisis futbolístico en particular.

Como se ha mencionado a lo largo de esta memoria, se ha intentado, en la medida de lo posible, justificar cada uno de los elementos que conforman el estudio. Además, se ha procurado la replicabilidad del proyecto tomando las debidas precauciones. El código Python desarrollado para llevar a cabo el estudio estará a disposición de cualquier usuario que así lo desee, el enlace al fichero con todo el código y otros archivos puede ser encontrado en la sección 3.2.3.

7.5 Trabajo futuro

Como ya se ha mencionado, este proyecto tiene un gran potencial para ser desarrollado en diferentes direcciones. Se espera que la metodología aquí presentada, a su vez, también pueda servir de inspiración para futuros proyectos similares. A continuación, se presenta trabajo futuro relacionado con este trabajo que puede ser de particular interés.

- Revisión estadística del proyecto, para verificar que las técnicas estadísticas aplicadas lo han sido de la forma correcta y considerar incluir algunas más de cara a fortalecer el rigor estadístico del proyecto. Un ejemplo de esto podría ser ajustar el nivel de significación para evitar errores multitest.

- Enfoque en posesiones, podría ser interesante realizar una revisión de la forma de procesar las posesiones, a fin de contemplar otras características o cambiar la forma en que se procesa alguna de las ya presentes. También podría ser interesante probar diferentes configuraciones de clustering de cara a explorar nuevos tipos de posesiones.

- Enfoque en equipos, como ya se ha mencionado, podría ser interesante considerar realizar estudios sobre el estilo de juego sobre equipos que comparten un nivel de competición similar, intentando en la medida de lo posible comparar los equipos más humildes con los más potentes. Este tipo de comparaciones se ve opacado por la gran diferencia de nivel que evita encontrar diferencias entre los equipos más fuertes.

- Contemplar únicamente el patrón del movimiento del balón, para realizar un estudio sobre cómo son los ataques de cada uno de los equipos de las cinco grandes ligas. Poniendo el foco al máximo en los patrones de movimiento del balón y obviando el resto de las características contempladas en el estudio.
- Creación de una aplicación, con capacidad para reproducir la metodología propuesta y brindar las estadísticas computadas para cada uno de los equipos en este estudio. Esta aplicación podría ser de interés para entrenadores o usuarios con conocimientos futbolísticos.

Bibliografía

Alamar, B. (2013). Sports analytics. In *Sports Analytics*. Columbia University Press.

Allaoui, M., Kherfi, M. L., and Cheriet, A. (2020). Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In *International conference on image and signal processing*, pages 317–325. Springer.

Altarriba-Bartes, A., Peña, J., Vicens-Bordas, J., Casals, M., Peirau, X., and Calleja-González, J. (2021). The use of recovery strategies by spanish first division soccer teams: a cross-sectional survey. *The Physician and Sportsmedicine*, 49(3):297–307.

Barbosa, A., Ribeiro, P., and Dutra, I. (2022). Similarity of football players using passing sequences. In *Machine Learning and Data Mining for Sports Analytics: 8th International Workshop, MLSA 2021, Virtual Event, September 13, 2021, Revised Selected Papers*, pages 51–61. Springer.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Boluarte Pretell, V. (2022). Desarrollo de un sensor biométrico inalámbrico orientado a aplicaciones deportivas.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Clemente, F. M., Couceiro, M. S., Martins, F. M. L., and Mendes, R. S. (2015a). Using network metrics in soccer: a macro-analysis. *Journal of human kinetics*, 45(1):123–134.

Clemente, F. M., Martins, F. M. L., Kalamaras, D., Wong, P. d., and Mendes, R. S. (2015b). General network analysis of national soccer teams in fifa world cup 2014. *International Journal of Performance Analysis in Sport*, 15(1):80–96.

Decroos, T., Van Haaren, J., and Davis, J. (2018). Automatic discovery of tactics in spatio-temporal soccer match data. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 223–232.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Dugué, N. and Perez, A. (2015). *Directed Louvain: maximizing modularity in directed networks*. PhD thesis, Université d'Orléans.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Feir-Walsh, B. J. and Toothaker, L. E. (1974). An empirical comparison of the anova f-test, normal scores test and kruskal-wallis test under violation of assumptions. *Educational and Psychological Measurement*, 34(4):789–799.

Ghar, S., Patil, S., and Arunachalam, V. (2021). Data driven football scouting assistance with simulated player performance extrapolation. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1160–1167. IEEE.

Goes, F., Meerhoff, L., Bueno, M., Rodrigues, D., Moura, F., Brink, M., Elferink-Gemser, M., Knobbe, A., Cunha, S., Torres, R., et al. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, 21(4):481–496.

Gupta, A. S., Pierpoint, L. A., Comstock, R. D., and Saper, M. G. (2020). Sex-based differences in anterior cruciate ligament injuries among united states high school soccer players: an epidemiological study. *Orthopaedic Journal of Sports Medicine*, 8(5):2325967120919178.

Gyarmati, L., Kwak, H., and Rodriguez, P. (2014). Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308*.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

Horvat, T. and Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1380.

Kanagala, H. K. and Krishnaiah, V. J. R. (2016). A comparative study of k-means, dbscan and optics. In *2016 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE.

Leser, R., Baca, A., and Ogris, G. (2011). Local positioning systems in (game) sports. *Sensors*, 11(10):9778–9797.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

los Ríos, D., Andrés, J., Gamba Cárdenas, W. E., Junco Hurtado, D. S., Siza Correa, S., et al. (2017). Dron-fit: sistema de monitoreo, seguimiento y generación de alertas enfocado en la práctica deportiva, basado en información capturada con sensores presentes en el dispositivo microsoft band y registro de video a través de un drone a través de la plataforma openmtc.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Merlin, M., Cunha, S. A., Moura, F. A., Torres, R. d. S., Gonçalves, B., and Sampaio, J. (2020). Exploring the determinants of success in different clusters of ball possession sequences in soccer. *Research in Sports Medicine*, 28(3):339–350.

Mitchell, T. M. et al. (2007). *Machine learning*, volume 1. McGraw-hill New York.

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., and Giannotti, F. (2019a). Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–27.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., and Giannotti, F. (2019b). A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):236.

Pappalardo, L. and Massucco, E. (2019). Soccer match event dataset.

Peel, S. A., Thorsen, T. A., Schneider, L. G., and Weinhandl, J. T. (2020). Effects of foot rotation on acl injury risk variables during drop landing. *Journal of Science in Sport and Exercise*, 2:59–68.

Peña, J. L. and Navarro, R. S. (2015). Who can replace xavi? a passing motif analysis of football players. *arXiv preprint arXiv:1506.07768*.

Prensa, M. (2021).

Rein, R. and Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, 5(1):1–13.

Sagiroglu, S. and Sinanc, D. (2013). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)*, pages 42–47. IEEE.

Shaw, L. and Glickman, M. (2019). Dynamic analysis of team strategy in professional football. *Barça sports analytics summit*, 13.

Student (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

Villa, F. D., Buckthorpe, M., Grassi, A., Nabiuzzi, A., Tosarelli, F., Zaffagnini, S., and Villa, S. D. (2020). Systematic video analysis of acl injuries in professional male football (soccer): injury mechanisms, situational patterns and biomechanics study on 134 consecutive cases. *British Journal of Sports Medicine*, 54(23):1423–1432.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Wu, J., Liu, D., Guo, Z., Xu, Q., and Wu, Y. (2021). Tacticflow: Visual analytics of ever-changing tactics in racket sports. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):835–845.

Zulyaden, A., Dewi, R., and Tantri, A. (2022). Football talent scouting application development "sport search" method based on android. In *Proceedings of the 7th Annual International Seminar on Transformative Education and Educational Leadership, AISTEEL 2022, 20 September 2022, Medan, North Sumatera Province, Indonesia*.

APÉNDICE A
Centroides posesiones

En este apéndice se mostrarán las posesiones centroide de cada uno de los clústeres obtenidos en el clustering de posesiones:

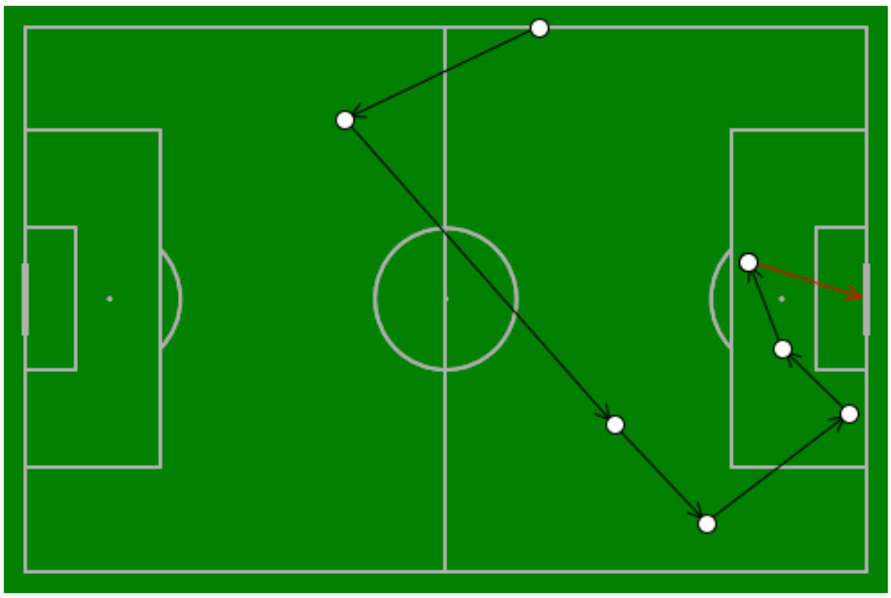

Figura A.1: Posesión centroide del Clúster 0

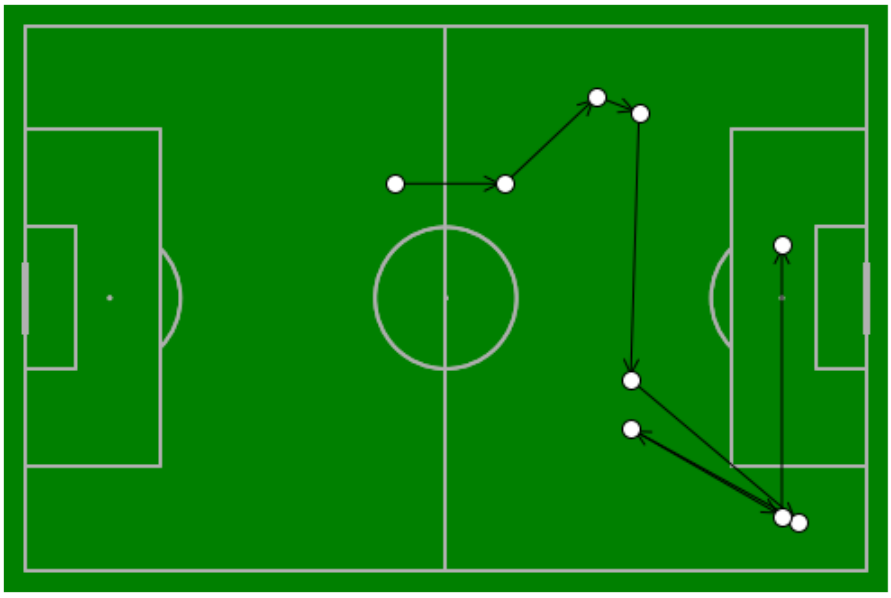

Figura A.2: Posesión centroide del Clúster 1

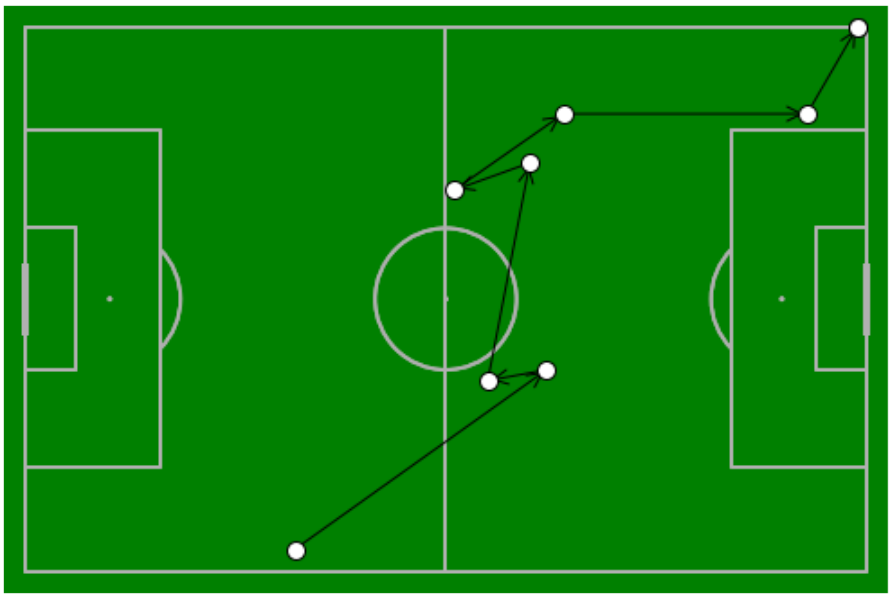

Figura A.3: Posesión centroide del Clúster 2

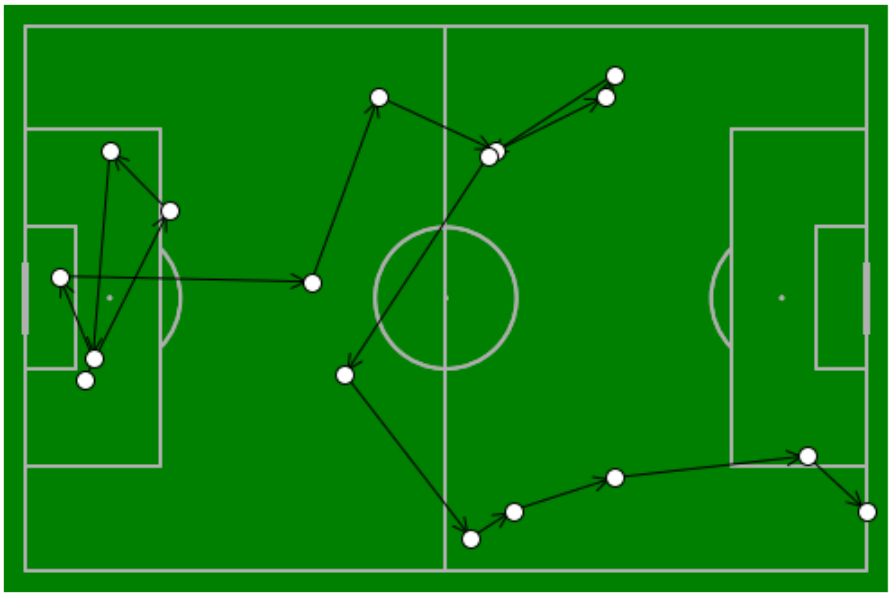

Figura A.4: Posesión centroide del Clúster 3

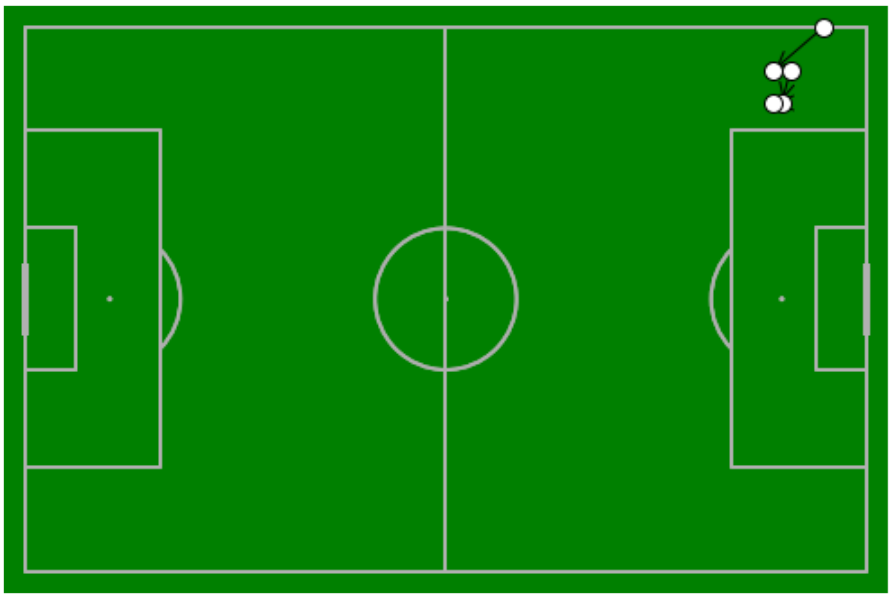

Figura A.5: Posesión centroide del Clúster 4

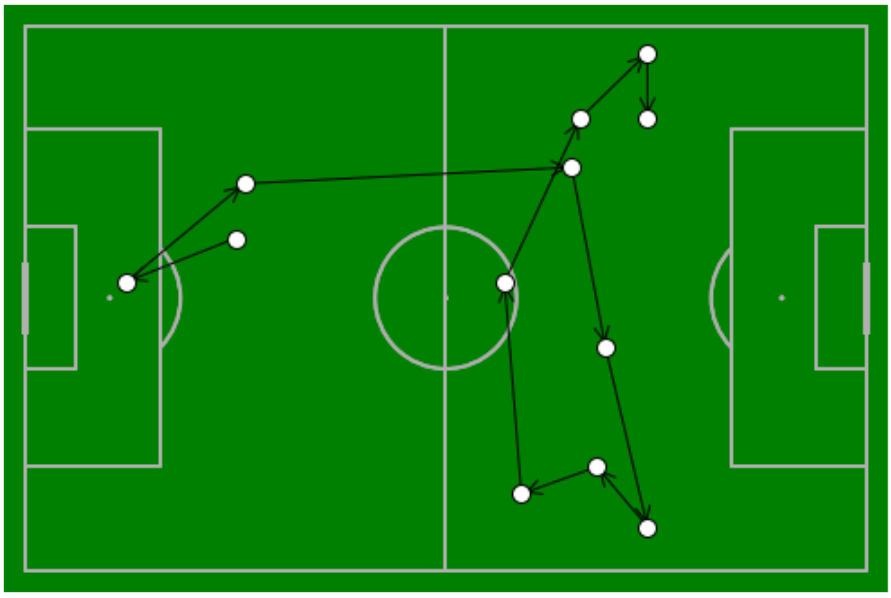

Figura A.6: Posesión centroide del Clúster 5

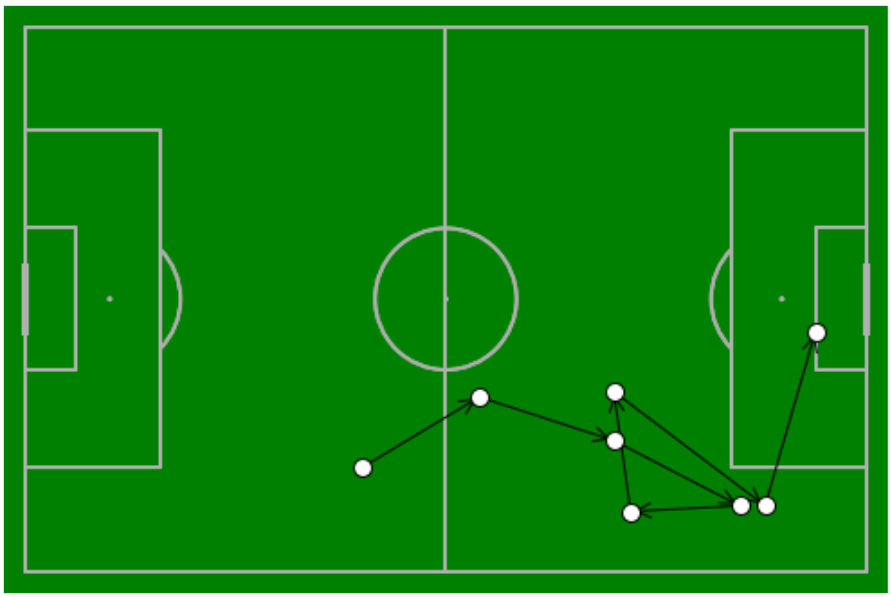

Figura A.7: Posesión centroide del Clúster 6

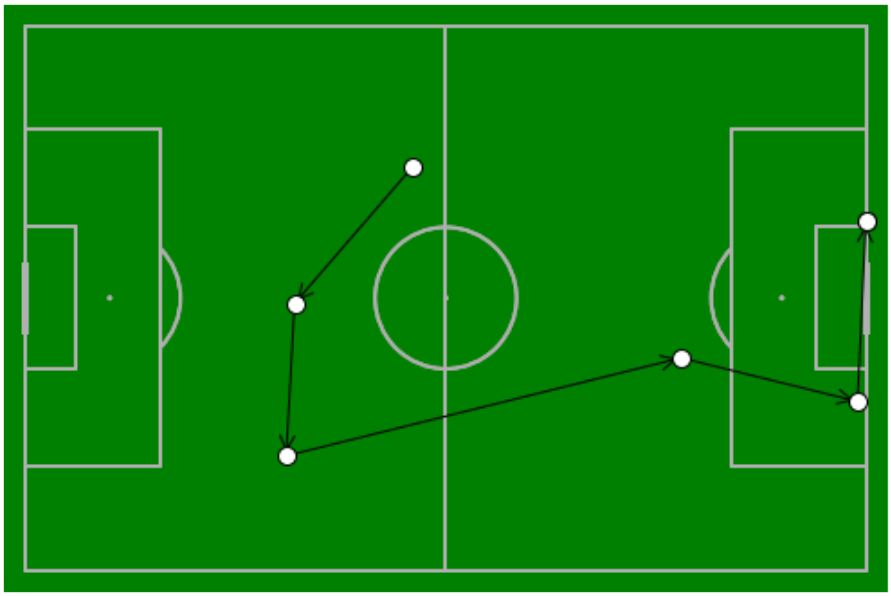

Figura A.8: Posesión centroide del Clúster 7

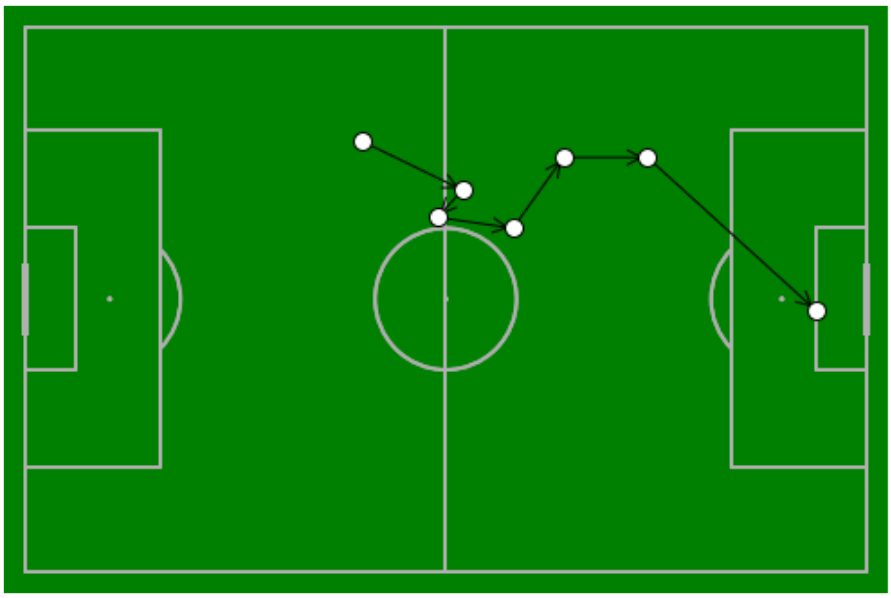

Figura A.9: Posesión centroide del Clúster 8

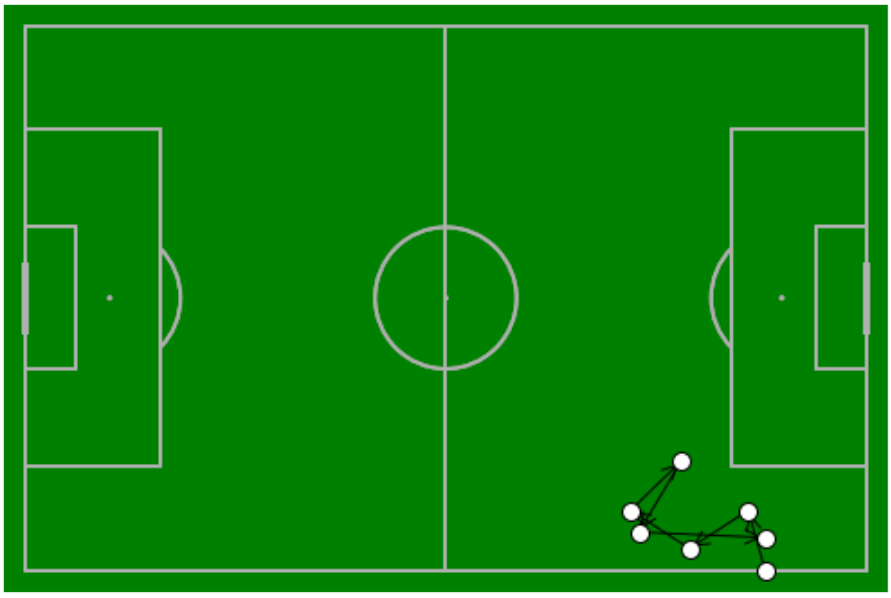

Figura A.10: Posesión centroide del Clúster 9

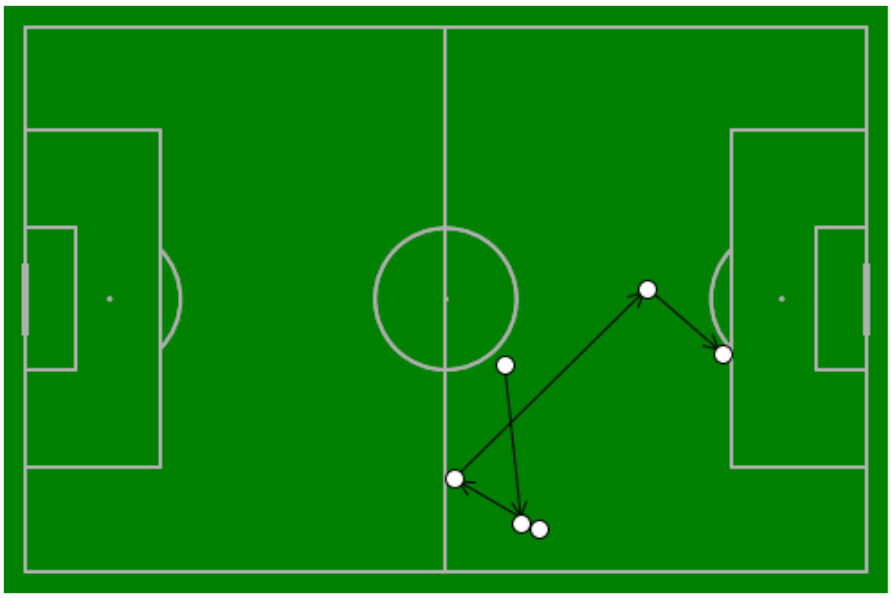

Figura A.11: Posesión centroide del Clúster 10

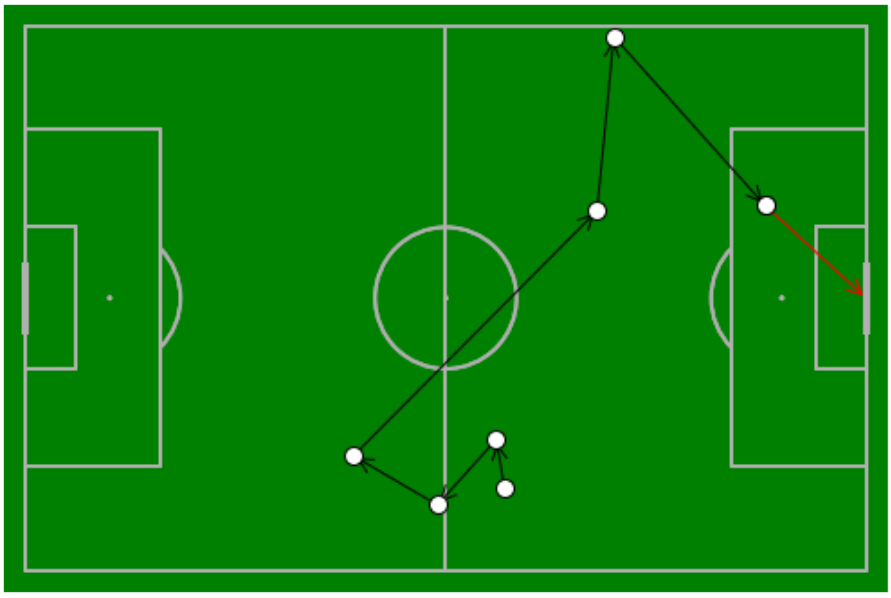

Figura A.12: Posesión centroide del Clúster 11

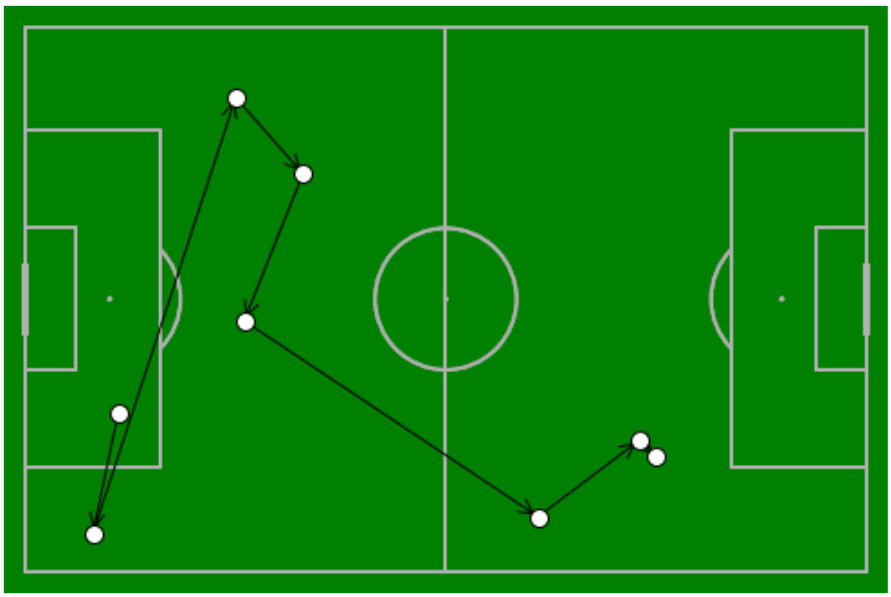

Figura A.13: Posesión centroide del Clúster 13

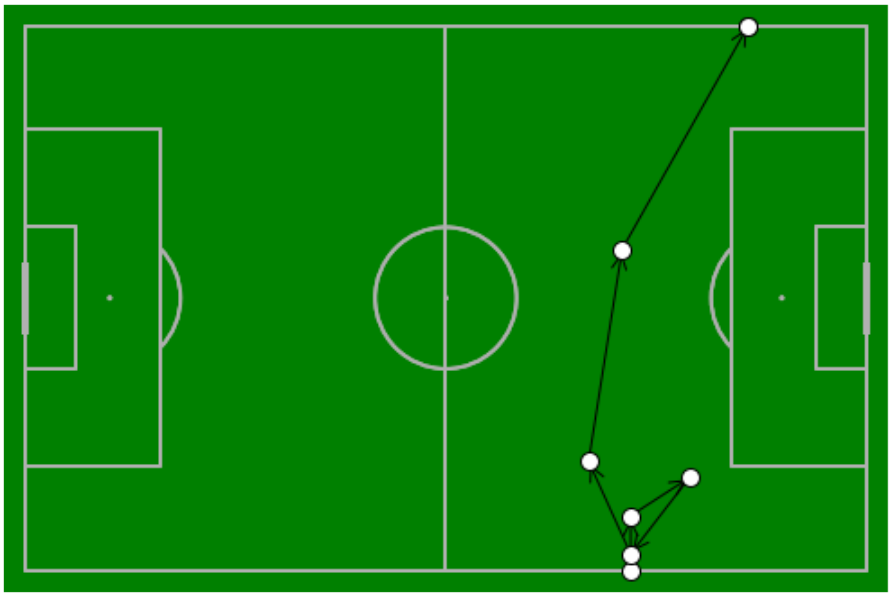

Figura A.14: Posesión centroide del Clúster 14

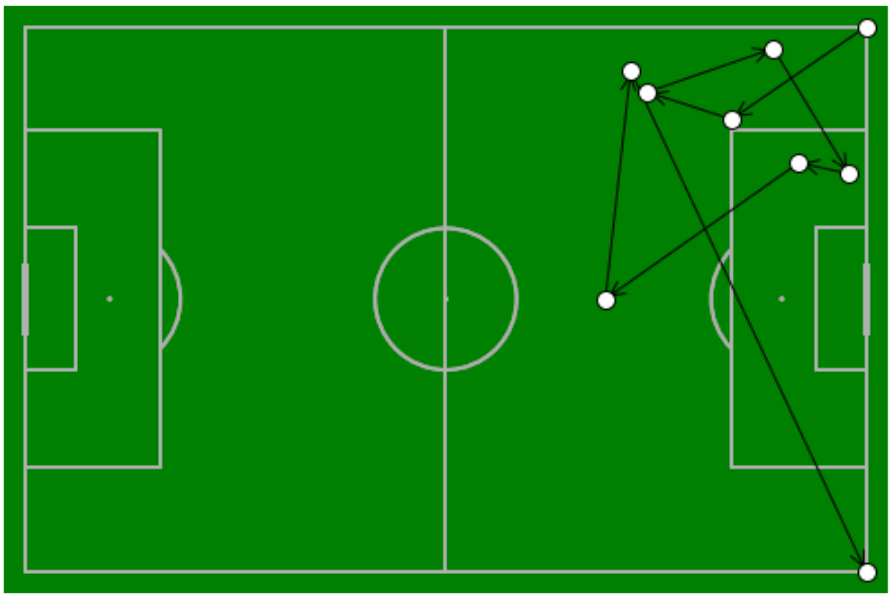

Figura A.15: Posesión centroide del Clúster 15

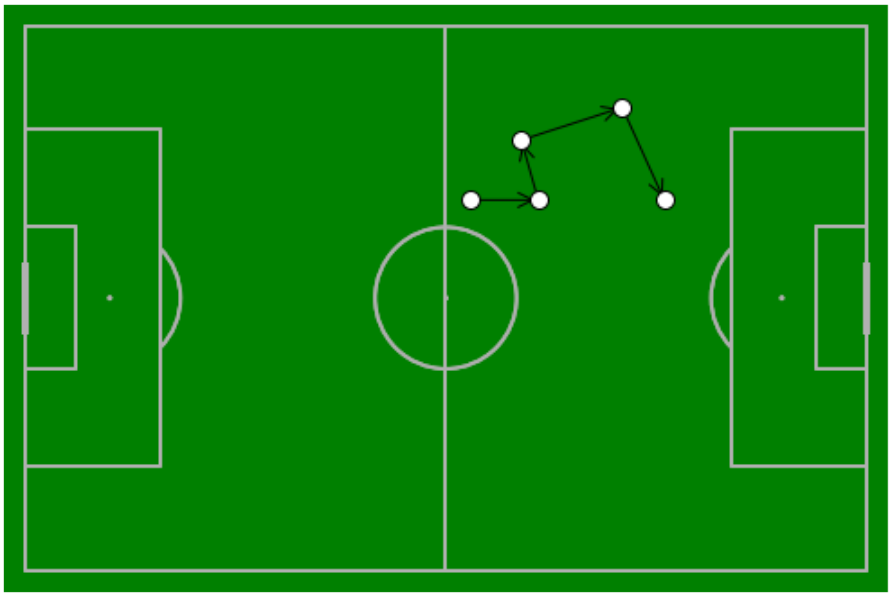

Figura A.16: Posesión centroide del Clúster 17

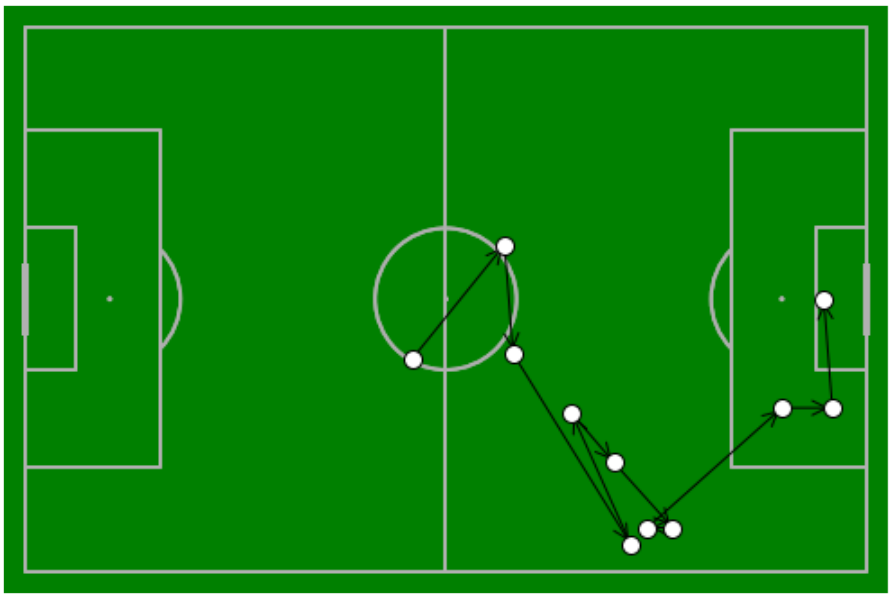

Figura A.17: Posesión centroide del Clúster 18

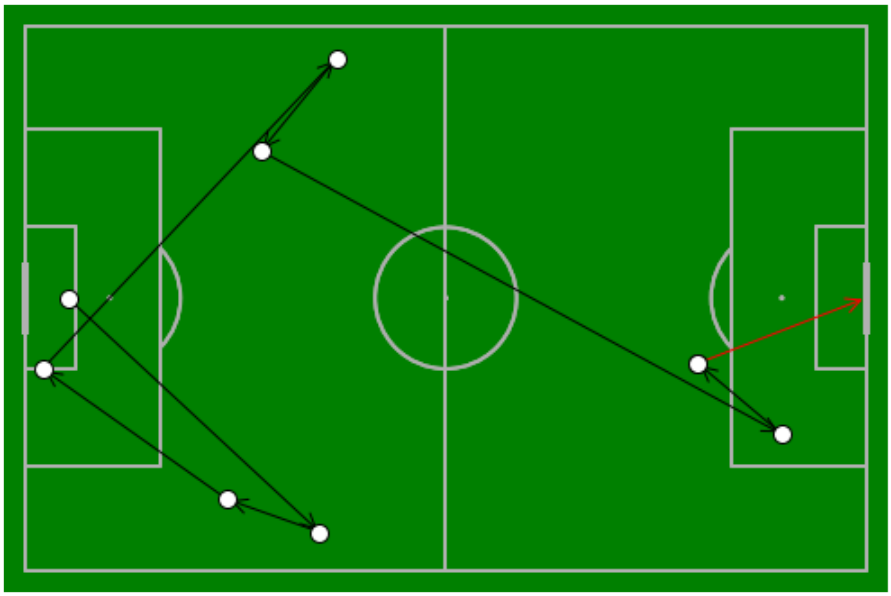

Figura A.18: Posesión centroide del Clúster 19

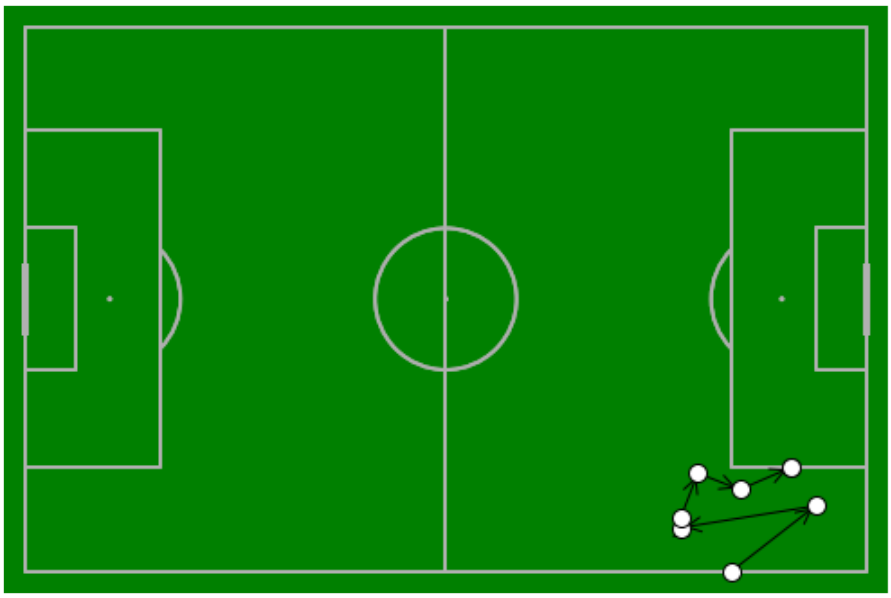

Figura A.19: Posesión centroide del Clúster 20

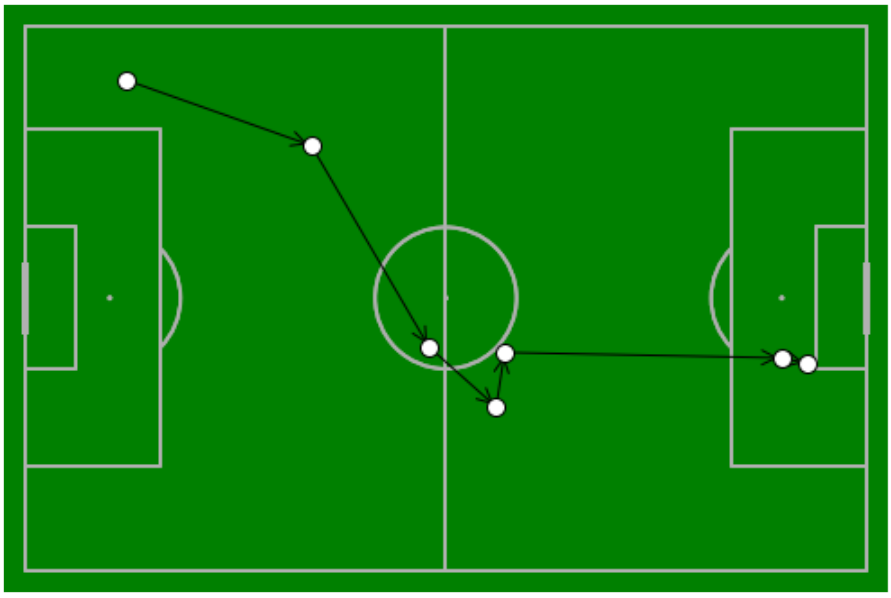

Figura A.20: Posesión centroide del Clúster 21

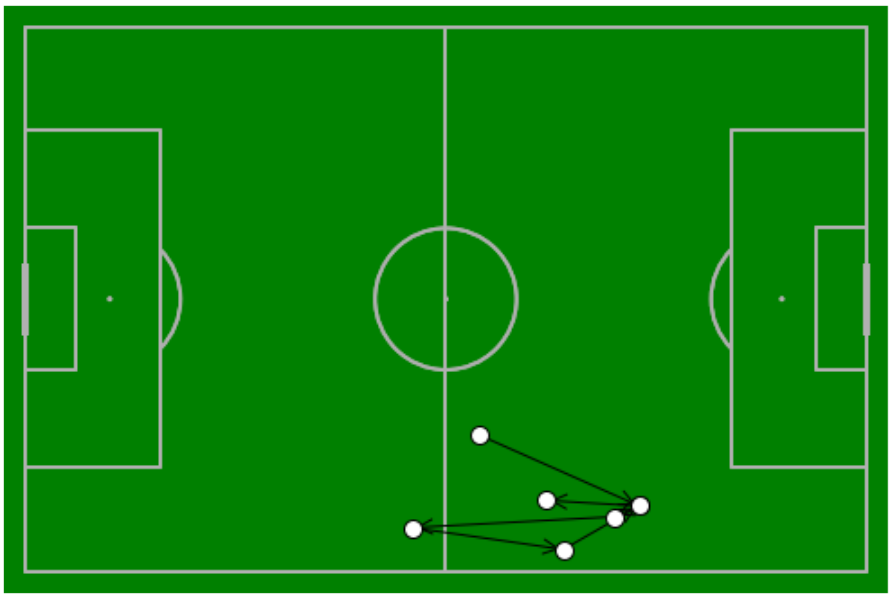

Figura A.21: Posesión centroide del Clúster 22

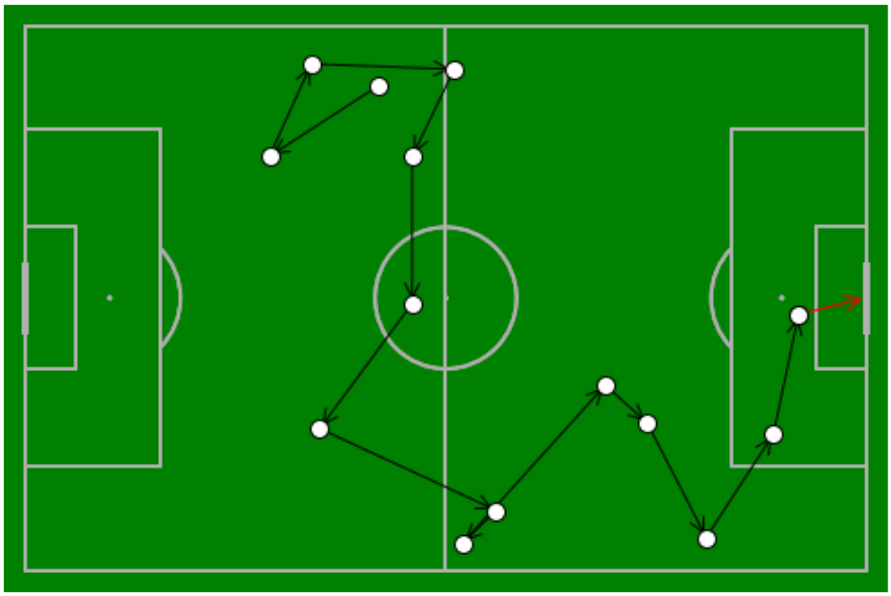

Figura A.22: Posesión centroide del Clúster 23

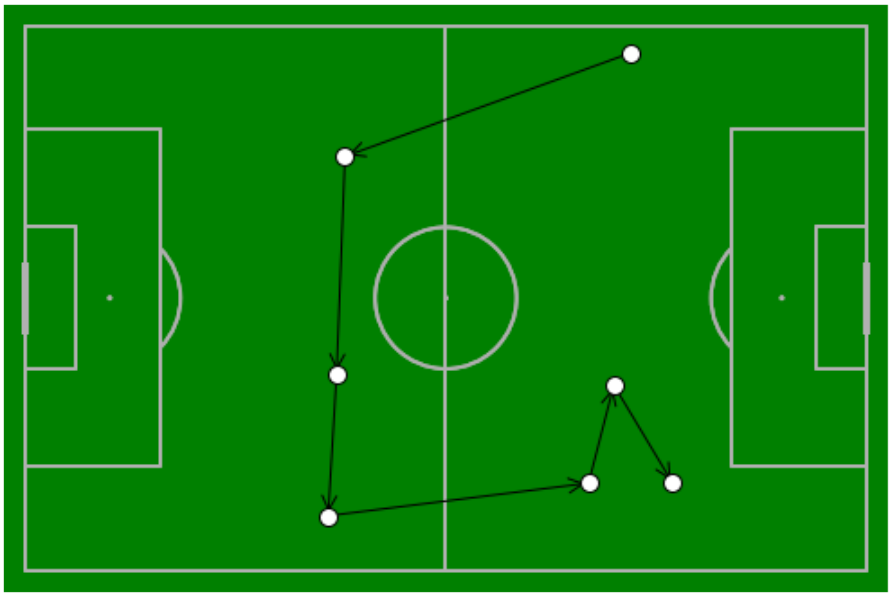

Figura A.23: Posesión centroide del Clúster 24

APÉNDICE B
ODS

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenible	**Alto**	**Medio**	**Bajo**	**No procede**
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**	X			
ODS 4. **Educación de calidad.**				X
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**		X		
ODS 9. **Industria, innovación e infraestructuras.**				X
ODS 10. **Reducción de las desigualdades.**				X
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**				X

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

En un mundo donde el fútbol se ha convertido en mucho más que un simple deporte, mi Trabajo de Fin de Grado (TFG) se sumerge en la esencia de este apasionante juego. Mi objetivo principal es analizar los estilos de juego de los equipos que participan en las cinco principales ligas de fútbol en Europa y cómo estos estilos pueden estar relacionados con los Objetivos de Desarrollo Sostenible (ODS) propuestos por las Naciones Unidas. Si bien a primera vista puede parecer que el fútbol profesional y los ODS no tienen una conexión evidente, este análisis busca explorar las posibles intersecciones y contribuciones que este deporte puede hacer hacia un mundo más sostenible.

El fútbol ha evolucionado más allá de ser simplemente un juego. Es un fenómeno cultural, económico y social que une a personas de todas las edades, géneros y orígenes étnicos en todo el mundo. Las cinco grandes ligas de fútbol europeas (La Liga española, la Premier League inglesa, la Serie A italiana, la Bundesliga alemana y la Ligue 1 francesa) atraen a una audiencia global masiva y generan enormes ingresos. En este sentido, mi trabajo está relacionado con el "Trabajo Decente y Crecimiento Económico" (ODS 8), pues el fútbol profesional crea empleo y oportunidades económicas. Desde jugadores hasta personal de apoyo, el fútbol genera trabajos y estimula el crecimiento económico en las ciudades y regiones que albergan equipos de fútbol.

Por otro lado, el simple hecho de fomentar la práctica de algún deporte ya guarda relación con la "Salud y Bienestar" (ODS 3). Aunque el fútbol profesional a menudo está rodeado de lesiones y tensiones físicas, promueve la salud y el bienestar en la sociedad en general. Fomenta la actividad física y el ejercicio, lo que contribuye a la prevención de enfermedades y promueve un estilo de vida saludable. Además, el fútbol no solo beneficia la salud física, sino también la salud mental. El hecho de seguir a un equipo favorito, asistir a partidos o simplemente disfrutar de un partido en la televisión puede tener un impacto positivo en el bienestar emocional de las personas. La emoción que el fútbol genera entre los aficionados pueden aliviar el estrés y mejorar el estado de ánimo, lo que refuerza aún más su vínculo con el ODS 3, promoviendo una vida saludable y bienestar.

En cuanto a la relación con el ODS 8, el impacto económico del fútbol no se limita solo a la creación de empleo. Los clubes de fútbol generan ingresos significativos a través de la venta de entradas, patrocinios, derechos de televisión y merchandising, entre otros. Estos ingresos no solo benefician a los clubes, sino que también tienen un efecto multiplicador en las economías locales. Los negocios locales, como restaurantes, bares y hoteles, experimentan un aumento en la clientela los días de partido, lo que impulsa sus ingresos. Además, el turismo relacionado con el fútbol atrae a visitantes de todo el mundo, lo que contribuye al crecimiento económico de las ciudades anfitrionas. Así, el fútbol se convierte en un motor económico que respalda directamente el ODS 8, fomentando el trabajo decente y el crecimiento económico.

En resumen, mi TFG sobre el análisis de estilos de juego en las cinco grandes ligas de fútbol no solo revela la complejidad y diversidad del deporte, sino que también destaca su relevancia en la promoción de la salud, el bienestar y el crecimiento económico. El fútbol va más allá de ser un simple entretenimiento; es un fenómeno global que tiene el potencial de impactar positivamente en la sociedad y en la consecución de los Objetivos de Desarrollo Sostenible propuestos por las

Naciones Unidas. A través de este análisis, se destaca la importancia de reconocer y aprovechar el poder del fútbol como una fuerza para el cambio y la sostenibilidad en nuestro mundo actual. El deporte puede ser mucho más que un juego; puede ser un agente de transformación social y un aliado en la búsqueda de un futuro más sostenible para todos.

Escola Tècnica
Superior d'Enginyeria
Informàtica

ETS Enginyeria Informàtica
Camí de Vera, s/n, 46022, Vaència
T +34 963 877 210
F +34 963 877 219
etsinf@upvnet.upv.es - www.inf.upv.es