UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Estudio de algoritmos y modelos para la detección de patrones similares de juego en futbolistas profesionales

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Simó Vidal, Andreu

Tutor/a: Sánchez Anguix, Víctor

Cotutor/a: Alberola Oltra, Juan Miguel

CURSO ACADÉMICO: 2022/2023

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Estudio de algoritmos y modelos para la detección de patrones similares de juego en futbolistas profesionales

Trabajo Fin de Grado

Grado en Ciencia de datos

Autor: Andreu Simó Vidal

Tutor: Víctor Sánchez Anguix, Juan Miguel Alberola Oltra

2022-2023

Resumen

Este trabajo aborda la tarea de la búsqueda de jugadores similares, en concreto, de los jugadores de campo de las cinco grandes ligas durante la temporada 2017-18 a partir de técnicas de aprendizaje automático, como el análisis de componentes principales, UMAP, t-SNE y técnicas de clustering. Para que estos resultados se comprendan por el usuario estos se mostrarán junto a distintos gráficos y razonamientos que muestren el porqué de la elección de los perfiles. Esta propuesta trata de aumentar la información que se tiene en cuenta para buscar perfiles similares incluyendo información más disgregada y añadiendo la componente espacial a algunas de las características. Resumiendo, este trabajo trata de ofrecer una herramienta para ayudar a los clubes de fútbol profesionales en forma de soporte en la búsqueda de reemplazos a la hora de elaborar la planificación de la plantilla.

Palabras clave: analítica deportiva, clustering, reducción de dimensionalidad, optimización.

Abstract

This work addresses the task of searching for similar players, specifically, out- field players in the five major leagues during the 2017-18 season using big data and machine learning techniques, such as principal component analysis, UMAP, t-SNE and clustering techniques. For these results to be displayed to the user, they must also be shown together with different graphs and a justification of why the profiles were chosen. This proposal tries to increase the information considered to search for similar profiles by including more disaggregated information and adding the spatial component to some of the features. To sum up, this work tries to offer a tool to help professional clubs in the search for replacements when planning their squad.

Keywords: big data, machine learning, dimensionality reduction, sports analytics, clustering, optimization.

Table de contenidos

Índice de tablas

Índice de ecuaciones

Índice de figuras

1 Introducción

En este capítulo introductorio se contemplan algunos de los aspectos más relevantes en el inicio de una investigación. Las secciones que lo componen son, en primer lugar, el aliciente detrás de este trabajo, los objetivos que se persiguen, el impacto que se espera que tenga y por último la metodología de trabajo que se ha seguido para lograr los objetivos propuestos.

1.1 Motivación

En el último medio siglo, el acelerado avance tecnológico ha abocado a las organizaciones a una transición en el proceso de toma de decisiones desde un paradigma basado en la intuición hacia otro dirigido por los datos relevantes y el conocimiento que se puede extraer de estos (Marr, 2016). Tal ha sido el cambio que, gracias al conocimiento extraído de los datos, muchas empresas han conseguido una ventaja estratégica mientras que otras que no se adaptaron a este cambio de paradigma acabaron por desaparecer.

En las entidades deportivas en general y en los equipos de fútbol en concreto, este cambio también se ha producido, aunque con una adopción más tardía. Desde hace una década, el crecimiento de departamentos de análisis de datos en los equipos de fútbol profesionales ha sido exponencial y hoy en día, estos departamentos constituyen el núcleo de la planificación deportiva de los clubes (Rejec, 2016).

Es en este contexto donde toma relevancia la tarea de búsqueda de grupos y similitudes, habitual en los proyectos de Ciencia de datos, en su aplicación en diversos campos. Estas técnicas de aprendizaje no supervisado[1] como las técnicas de agrupación o Clustering devuelven información valiosa para el usuario la cual le permite conocer por ejemplo los distintos tipos de compradores que acuden a un establecimiento (Kansal, Bahuguna, Singh, & Choudhury, 2018)o para agrupar imágenes similares basadas en características visuales (Chang, Wang, Meng, Xiang, & Pan, 2017).

En la disciplina de la analítica deportiva puede haber varias motivaciones posibles detrás de un proyecto que aborde la tarea de búsqueda de perfiles similares de jugadores de fútbol como por ejemplo mejorar el reclutamiento de jugadores o apoyar a la toma de decisiones tácticas ayudando al equipo técnico a la hora de realizar sustituciones durante un partido (de la Torre, Lopez-Lopez,

[1] El aprendizaje no supervisado se refiere al uso de técnicas de aprendizaje automático para analizar y agrupar conjuntos de datos no etiquetados.

Juan, & Clavet, 2022). El auge de las técnicas de aprendizaje automático[2] en el fútbol profesional ha permitido a clubes de la élite obtener una ventaja competitiva, especialmente en el área de confección de plantillas con los clubes realizando traspasos constantemente.

Por otra parte, afrontar este tipo de tarea supone un reto en sus distintas fases. Empezando por la obtención de los datos, es complicado poder encontrar un conjunto de datos abiertos con información temporalmente completa y con una cantidad relevante de clubes y jugadores, además del trabajo de adaptar la información para esta tarea concreta. En segundo lugar, dada la gran cantidad de atributos que se obtienen de los jugadores a lo largo de un partido, esto significa que se cuenta con una alta dimensionalidad lo cual implica la necesidad del uso de técnicas para trabajar en estos entornos. Por último, en lo referente al modelado y la visualización, hay varios desafíos como el de encontrar las técnicas que mejor se adaptan a la naturaleza de los datos y el de buscar la forma más completa de mostrar la información para los clubes de tal forma que se entienda el porqué de las similitudes entre futbolistas.

Más allá de lo mencionado anteriormente, el fútbol es un deporte que da mucho juego para desarrollar proyectos que involucren el uso de datos, ya que al ser un deporte en equipo ofrece muchas más posibilidades de tratar los datos desde diferentes puntos de vista, mientras que, por otro lado, los deportes individuales tienen más limitaciones puesto que no permiten el estudio de dinámicas de plantilla ni de planificación deportiva.

Resumidamente, la motivación detrás de este trabajo es la de ofrecer una herramienta de soporte a los clubes profesionales en el abordaje de una de las áreas más importantes de la planificación deportiva.

En este documento se cuenta cómo se ha desarrollado esta herramienta en sus varias fases.

1.2 Objetivos

El objetivo global del presente trabajo es el de comparar diferentes técnicas para modelar el comportamiento de los jugadores en el terreno de juego y agruparlos por similitud con la finalidad de ofrecer recomendaciones de perfiles con los que reemplazar jugadores salientes o para completar planteles con jugadores que puedan adaptarse un determinado estilo de juego.

Para la consecución de este objetivo se ha planteado partir de un datos abiertos y adaptarlos con la finalidad de conseguir el objetivo general.

[2] El aprendizaje automático (Machine learning) es un campo de la inteligencia artificial que permite a los sistemas aprender y mejorar a partir de la experiencia sin ser programados explícitamente.

1.3 Impacto esperado

El principal público objetivo del producto son los clubes profesionales, los cuales disponen de los medios y departamentos necesarios para efectuar planificaciones de plantillas. El producto resultante de la investigación servirá como herramienta de apoyo para estos.

La ventaja de este servicio respecto a otras alternativas es que actualiza las metodologías, añadiendo modelos más recientes y comparándolos. Además de que se dispone también de una metodología aplicable a distintos conjuntos de eventos.

Resumidamente, para las instituciones deportivas la herramienta sirve como apoyo para la tarea de la confección de plantillas, lo cual es especialmente interesante para aportar un respaldo técnico en la toma de decisiones del área de planificación, siendo esta una de las áreas de más peso en un equipo de fútbol, tanto desde el punto de vista deportivo como económico.

1.4 Metodología de trabajo

En esta sección se discute el criterio de metodología trabajo que se ha seguido con sus distintas fases.

La metodología llamada Crisp DM (Cross Industry Standard Process for Data Mining) es la más utilizada para proyectos dedicados a extraer valor de los datos. Esta metodología se divide en seis fases que son: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. Se puede ver la metodología esquematizada visualmente en la figura 1.

Figura 1 Pasos de la metodología Crisp DM. Fuente: healthdataminer.

1.4.1 Entendimiento del negocio

La primera fase consiste en realizar una análisis acerca del marco teórico en el que se va a desarrollar la tarea. Esto implica la búsqueda de foros y artículos del sector, con el fin de determinar cuáles son las formas habituales de trabajar de los profesionales, qué áreas ya han sido exploradas y cuáles son las limitaciones u oportunidades que quedan vacantes para iniciar líneas de investigación. De esta fase es especialmente importante salir con un conocimiento extenso del área en que se trabaja y empezar a formarse una idea de qué es aquello que se va a proponer.

En este caso, la fase de entendimiento del negocio empezó por establecer una definición de la analítica deportiva y estudiar trabajos de analítica deportiva en general y de analítica en el mundo del fútbol con la enumeración de las posibilidades de mejora o de ampliación de propuestas anteriores.

1.4.2 Entendimiento de los datos

De la anterior fase se dijo que era relevante salir con una idea clara de la tarea que se quiere plantear. Esto es imprescindible para el correcto entendimiento de los datos, ya que esto implica que se seleccionen conjuntos de datos adecuados para posteriormente explorarlos.

Con el objetivo en mente, se buscaron conjuntos de datos de tipo evento, los cuales indican características referentes a la forma de jugar de los futbolistas. El análisis exploratorio para evaluar la información disponible en esta fase fue necesario para decidir la preparación del conjunto de datos en el siguiente paso.

1.4.3 Preparación de los datos

Tomar la decisión de trabajar con datos abiertos significa que se cuenta con una mayor libertad para poder trabajar sobre dichos conjuntos, no obstante, también significa que uno se debe adaptar a la información que se encuentra disponible.

En esta fase se decidió alterar la forma en la que se almacenan los eventos para que en lugar de tenerlos a estos disgregados se tengan los eventos agrupados por jugadores para obtener una imagen de la forma de jugar de estos a lo largo de la temporada. Pero además con la ayuda de la técnica PCA se pudo extraer conocimiento sobre un conjunto de datos que no era abordable dada su elevado número de columnas, pudiendo así saber que algunos individuos se encontraban duplicados puesto que fueron traspasados a principio de temporada y uno de los duplicados contenía información sobre pocos partidos la cual no representaba su forma de jugar.

No es extraño que en proyectos de esta índole se combinen distintas fuentes. El hecho de añadir los datos de valor de mercado a la fuente de eventos supuso el uso de técnicas de similitud al no contar con una variable coincidente en los dos conjuntos provenientes de distintas fuentes, pero

fue de especial utilidad para agregar valor añadido destinado al público objetivo de esta investigación que son los clubes de fútbol.

1.4.4 Modelado

La fase de modelado implica la selección de técnicas y modelos que se van a seleccionar para poder extraer el conocimiento útil de los datos. Esta fase conforma un proceso iterativo con respecto a la fase previa. La aplicación de los modelos a el conjunto de datos definitivo puede motivar nuevos procesados de los datos.

Por ejemplo, con el conjunto de datos finalizado al principio, los eventos de pase se transformaron según la frecuencia con la que un jugador efectuaba uno desde una zona a todas las zonas discretizadas del campo (pares de tipo [Zona_origen, Zona_destino]), teniendo en principio ochenta y una variables sólo para los eventos de tipo pase. Se comprobó al aplicar técnicas de reducción de dimensionalidad que incluso con el escalado de los datos, las únicas variables que se consideraban relevantes para la creación del espacio de baja dimensionalidad eran los eventos de tipo pase. Se decidió finalmente que para paliar el exceso de relevancia de este tipo de eventos se optaría por otra transformación donde, en lugar de contar con los pares de zona, se contabilizará la frecuencia de pases que se efectúan desde cada zona de origen y a cada zona de destino, consiguiendo no perder la componente de localización de los pases, pero con sólo diez y ocho variables.

Por supuesto, la selección de modelos se debe hacer teniendo en cuenta el objetivo planteado en la fase uno y con la ayuda de los conocimientos adquiridos en la misma fase. En esta investigación la metodología que se sigue para encontrar similitudes entre grupos consiste en la aplicación de técnicas de agrupación, pero a causa de la alta dimensionalidad de los datos es pertinente aplicar una reducción de dimensionalidad previa.

1.4.5 Evaluación

A la hora de plantear los modelos, se debe decidir también que técnicas y algoritmos concretos se van a comparar con la finalidad de ver cuáles son las técnicas que mejor resultado dan para los datos.

Para esto se tiene que plantear una arquitectura que explore el rendimiento de las distintas combinaciones de modelos, en este caso combinaciones de técnicas de reducción de dimensionalidad y algoritmos de clustering y establecer una métrica de evaluación. Dado que se busca alguna métrica de la calidad de los grupos, se optó por definir una métrica y buscar su valor óptimo en una arquitectura de optimización en la cual se van probando para TSNE y UMAP (técnicas de reducción de dimensionalidad), las combinaciones con algoritmos de agrupamiento

y finalmente estudiar cuál es el modelo óptimo de entre el flujo de UMAP con la técnica de agrupación óptima para su espacio y el flujo análogo con TSNE.

1.4.5.1 Conclusiones de la evaluación

Es importante en este contexto extraer las conclusiones acerca de los resultados que ha devuelto la fase de evaluación. Se concluyó que los resultados con UMAP son significativamente mejores que los de TSNE. Esto se debe a que la proyección que realiza UMAP se adapta mejor a la naturaleza de estos datos. Sin embargo, esto no significa que una técnica vaya a ser siempre superior a la otra y esto siempre está supeditado a la naturaleza del conjunto que se estudie.

1.4.6 Despliegue

Es necesario estructurar el conocimiento obtenido a través del proceso de minería de datos y exponerlo de manera que pueda ser aplicado en el entorno empresarial. Esta es esencialmente la fase de despliegue. Una fase en la cual se debe determinar una estrategia para la puesta en operación del modelo y resumir y organizar los pasos seguidos, así como los resultados que se obtienen.

Para que la investigación sea relevante en el uso empresarial, una forma habitual de transmitir resultados es mediante una visualización. Esta visualización debe contener toda la información relevante teniendo en cuenta que esta debe ser comprendida por un usuario que no tiene por qué compartir los mismos conocimientos del campo de la ciencia de datos. Esta información consiste en: los grupos que se han formado, los nombres, rol, valor de los jugadores de cada grupo y la caracterización de cada clúster.

1.5 Estructura

Vista gran parte de la introducción, lo que resta del documento está organizado de la siguiente forma:

- La sección segunda, Estado del arte, contiene una revisión de trabajos del ámbito de la analítica deportiva, así como una crítica razonable al trabajo en este campo y por último la aproximación que se ha tomado en esta investigación en el apartado de propuesta.
- En el tercer capítulo se analiza la problemática sobre la que se ha trabajado desde un punto de vista del marco legal y ético que lo envuelve.
- El capítulo cuarto es el de propuesta, donde se detalla cada paso que se ha seguido en el apartado técnico de esta investigación, siendo estas la obtención de los datos, procesado de los datos, la metodología, la arquitectura de experimentación propuesta y la visualización.

- En el quinto capítulo se detallan los resultados obtenidos de la fase de experimentación introducida en el capítulo previo junto con las conclusiones que se extraen de los experimentos.
- El sexto capítulo de conclusiones repasa el estado de cumplimiento de los objetivos, así como las secciones de legado que revisa la herencia de este trabajo y la relación del trabajo desarrollado con los conocimientos adquiridos en el grado.
- Seguidamente el capítulo séptimo enumera mejoras y nuevas líneas de investigación a partir del presente estudio.
- Finalmente, en los capítulos ocho y nueve se incluyen las referencias que se han incluido en el documento y los anexos con código e información adicional respectivamente.

2 Estado del arte

En esta sección se comentará el marco teórico en que se desarrolla este trabajo, así como diversos artículos académicos relacionados con este trabajo, exponiendo, de cada artículo, su objetivo, el conjunto de datos que utiliza, las técnicas utilizadas, los resultados, la validación y una breve crítica. Este capítulo finaliza con la exposición de la propuesta de este trabajo.

Seguidamente se contará con más detalle qué es la disciplina de la analítica deportiva y cuáles son algunos de los foros más relevantes de ésta. Después se muestran trabajos de este campo, empezando por uno externo al fútbol y seguidamente, más centrados en la temática de este trabajo.

2.1 Marco teórico

Aquí se contemplan las distintas técnicas que se han requerido en las distintas fases de este trabajo, así como el papel concreto que han tomado.

2.1.1 Aprendizaje no supervisado

El aprendizaje no supervisado es una técnica de aprendizaje automático que busca patrones en un conjunto de datos sin utilizar etiquetas preexistentes. A diferencia del aprendizaje supervisado, este enfoque no requiere la intervención humana y el trabajo de anotación previa. La figura 2 muestra la diferencia entre ambos tipos de aprendizaje, el aprendizaje supervisado muestra que los datos ya están etiquetados y sus clases aparecen en la leyenda en la parte inferior derecha.

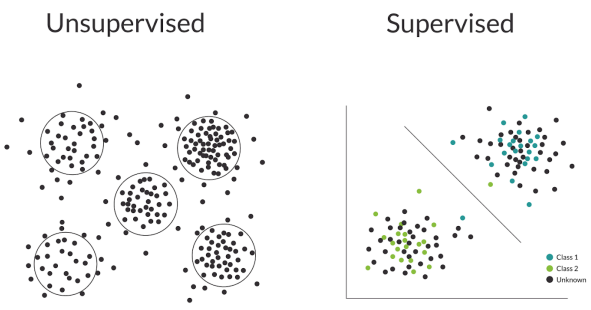


Figura 2 Diferencia entre aprendizaje no supervisado y supervisado. Fuente: ExtraHop.

Los modelos de aprendizaje se dividen en tres tipos:

- Agrupación: proceso de encontrar similitudes entre datos y agruparlos.
- Asociación: encontrar relaciones entre los datos de un conjunto de datos determinado.

- Reducción de la dimensionalidad: reducción del número de entradas a un tamaño más manejable preservando la integridad de los datos.

Entrando en más detalle, es la tarea encargada de encontrar un número de grupos con los datos originales llamados clústeres. Estos clústeres se forman en base a la similitud de las características de los datos, de manera que los objetos dentro de un mismo clúster son más similares entre sí que con aquellos en otros clústeres.

El análisis de clúster permite descubrir patrones ocultos y estructuras inherentes en los datos, identificar grupos homogéneos y heterogéneos, y obtener una comprensión más profunda de la distribución y la relación entre los datos.

Existen diferentes algoritmos y enfoques para realizar el análisis de clúster, y cada uno tiene sus propias suposiciones y métodos de cálculo para determinar los clústeres óptimos. A continuación, en la figura 3 se puede visualizar la diferencia entre la aplicación de K-Means y DBSCAN con varios conjuntos de datos, en concreto se muestra que las diferentes suposiciones y formas de calcular las técnicas llevan a la obtención de distintos resultados para el mismo conjunto de datos.

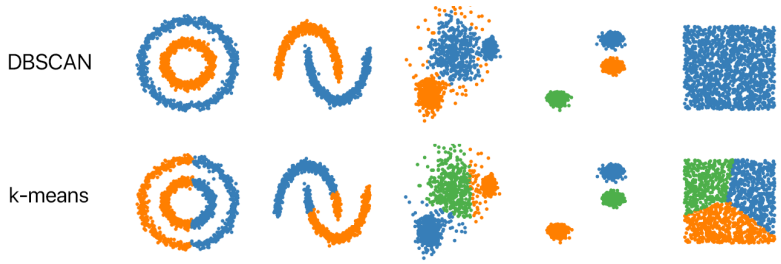


Figura 3 Resultados de kMeans y DBSCAN para distintos datos. Fuente: Towards Data Science.

Es necesario para entender que los algoritmos son distintos, que se exponga el funcionamiento de cada uno de los que han aparecido en la metodología.

KMeans:

- Inicialización: El algoritmo comienza eligiendo habitualmente al azar k centroides iniciales, siendo k es el número de clústeres que se desea encontrar.
- Asignación de puntos: Luego, cada punto de datos en el conjunto se asigna al clúster cuyo centroide está más cerca en función de una métrica de distancia.
- Actualización de centroides: Después de asignar todos los puntos a clústeres, los centroides se recalculan como el promedio de todos los puntos asignados a cada clúster.

- Repetición y resultado: Los pasos de asignación y actualización se repiten hasta que los centroides convergen o hasta que se alcance un número máximo de iteraciones. Cuando el algoritmo converge, obtienes k clústeres donde cada punto está asignado a uno de ellos.

Jerárquico:

- Construye una jerarquía de clústeres en lugar de asignar directamente puntos a clústeres. Comienza considerando que cada punto es un clúster individual y luego fusiona gradualmente los clústeres más cercanos en uno solo.
- Se puede realizar un clustering jerárquico de dos maneras: aglomerativa (comienza con clústeres individuales y los fusiona) o divisiva (comienza con un único clúster que contiene todos los puntos y lo divide gradualmente).

DBSCAN:

- En lugar de requerir un número de clústeres k como entrada, DBSCAN encuentra automáticamente el número de clústeres. Funciona identificando regiones densas de puntos en el espacio de datos y agrupando puntos que están cerca en función de una medida de densidad y una distancia máxima definida por el usuario llamada épsilon.
- Los puntos se clasifican como nucleares, de borde o como ruido en función de su densidad y su proximidad a otros puntos.
- Este algoritmo es eficaz para identificar clústeres de formas y tamaños irregulares y es resistente al ruido en los datos.

Ya hemos mencionado que el aprendizaje no supervisado requiere de muy poca intervención humana, no obstante, en el análisis clustering el usuario ha de decidir el número de grupos u otros parámetros.

La agrupación y la reducción de la dimensionalidad son los dos enfoques de principal interés puesto que ambas se combinan en la fase de metodología para encontrar grupos en cada una de las posibles posiciones de futbolistas.

2.1.2 Reducción de dimensionalidad

Las técnicas de reducción de dimensionalidad son métodos utilizados en el campo del análisis de datos y la estadística para disminuir la cantidad de variables o características en un conjunto de datos, mientras se intenta retener la mayor cantidad de información relevante posible. En otras palabras, estas técnicas buscan simplificar la representación de los datos al reducir el número de dimensiones en las que se encuentran.

Hay dos tipos de técnicas de reducción de dimensionalidad: las lineales, que asumen que las relaciones entre las variables se pueden expresar de manera aproximadamente lineal y las no lineales las cuales abordan situaciones donde las relaciones entre las variables no pueden ser aproximadas de manera efectiva mediante transformaciones lineales.

2.1.2.1 Técnicas lineales

Empezando con las técnicas lineales, el análisis de componentes principales (PCA) es la técnica de aprendizaje no supervisada principalmente utilizada para tareas de reducción de dimensionalidad, aunque su uso no se limita a esta tarea y también es una técnica muy útil para análisis descriptivo y tratamiento de datos atípicos.

Para poder conseguir la reducción de dimensionalidad, esta técnica busca relaciones lineales entre variables para poder así obtener nuevas variables independientes, llamadas componentes principales de tal forma que con un número pequeño de componentes principales (respecto al número de columnas originales) se pueda explicar una gran parte de la varianza de los datos originales. El orden de estas componentes principales es relevante y las sucesivas componentes irán recogiendo cada vez un menor porcentaje de varianza de los datos, ya que cada componente irá recogiendo un porcentaje de la varianza no recogida por componentes previas (a excepción de la primera).

Desde un punto de vista más formal, el PCA basa sus fundamentos en conceptos del álgebra lineal, en concreto de los valores y vectores propios.

Las matrices son muy útiles por la facilidad de operar con ellas, al multiplicar una matriz (A) por un vector (v) se obtiene un nuevo vector (u), se puede decir en este caso que la matriz ha llevado a cabo una transformación lineal al vector de entrada (ecuación 1).

$$Av = u$$

Ecuación 1 Transformación lineal al vector de entrada

De todos los vectores afectados por una matriz que atraviesa un espacio, el vector propio es aquél que cambia de longitud, pero no de dirección; es decir, el vector propio ya apunta en la misma dirección hacia la que la matriz empuja a todos los vectores.

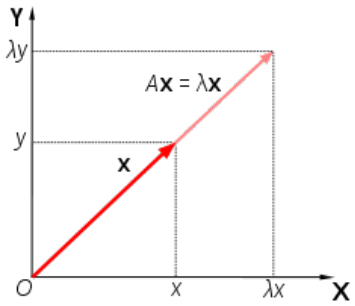


Figura 4 Vector propio de una matriz. Fuente: Wikipedia.

En la figura 4 el vector v es vector propio de la matriz A si el multiplicar dicho vector v por la matriz A es equivalente a multiplicar un valor escalar lambda (valor propio asociado a dicho vector propio) por el vector v.

La definición de un vector propio, por tanto, es un vector que responde a una matriz como si esa matriz fuera un coeficiente escalar, esta definición se corresponde con la ecuación 2.

$$Av = \lambda v$$

Ecuación 2 Definición de vector propio

Las matrices cuadradas pueden tener tantos vectores propios como dimensiones tengan; una matriz de n x n podría tener n vectores propios, cada uno de los cuales representaría su línea de acción en una dimensión.

Pues bien, los vectores propios asociados a la matriz de covarianzas de los datos originales es lo que permite el cómputo de las componentes principales. Para comprender esto, debemos definir el concepto de "score" en el contexto del PCA.

Los scores de las componentes principales son las coordinadas de cada punto, con respecto a los ejes de las componentes principales, estos son rectas y ortogonales respecto al resto de componentes. En la ecuación 3 podemos ver la representación correspondiente a los puntos en el espacio original p-dimensional:

$$x_i = (x_{1i}, x_{2i}, \dots, x_{pi})$$

Ecuación 3 Representación de los individuos en los datos originales

En estas coordinadas, esto significa que por ejemplo a través del eje x_1, el punto se encuentra a una distancia x_{1i} del origen. Pues bien, PCA es en esencia otra manera de describir este punto con respecto a un eje acorde a las componentes principales en lugar de describir dicho punto de

acuerdo con los ejes de las variables originales. Esto da lugar a la siguiente expresión en la ecuación 4:

$$z_i = (z_{1i}, z_{2i}, \dots, z_{pi}) = A(x_i - \bar{x})$$

Ecuación 4 Representación de los individuos en el nuevo espacio

Donde A es la matriz p x p de pesos de las componentes principales (con los vectores propios en cada una de sus filas) y $\bar{x}$ es el centroide de los datos (o vector de medias).

Así que puede pensar que los vectores propios describen dónde se encuentran las "líneas rectas" que describen los componentes principales. Luego, los scores de las componentes principales describen dónde se encuentra cada punto de los datos en cada línea recta, en relación con el "centroide" de los datos.

La selección del número de componentes principales es crítica, ya que un número elevado de estas puede llevar a sobreajuste del modelo, dañando su capacidad predictiva.

En este trabajo, la implicación de esta técnica se ve claramente en la fase de procesado y limpieza de los datos, ayudando a comprobar que los individuos estén correctamente etiquetados y que no haya individuos duplicados.

2.1.2.2 Técnicas no lineales

Uniform Manifold Approximation and Projection o UMAP (McInnes, Healy, & Melville, 2020) y t-Distributed Stochastic Neighbor Embedding o t-SNE (van der Maaten & Hinton, 2018) son ambas técnicas de reducción de dimensionalidad no lineales, ambos algoritmos se basan en la construcción de distribuciones de probabilidad que capturan relaciones y similitudes en los datos. La diferencia principal entre ellos radica en cómo modelan y minimizan estas distribuciones para lograr la representación en el espacio de menor dimensión.

UMAP permite crear nuevos parámetros (UMAP X y UMAP Y) a partir de los datos originales con alta dimensionalidad. Es una técnica reciente que ofrece ventajas frente a otras técnicas de reducción de dimensionalidad como t-SNE como un aumento considerable de la velocidad de cómputo, así como la capacidad de una mejor conservación de la estructura original de los datos. Visualmente, las diferencias entre ambas técnicas aparecen en la figura 5.

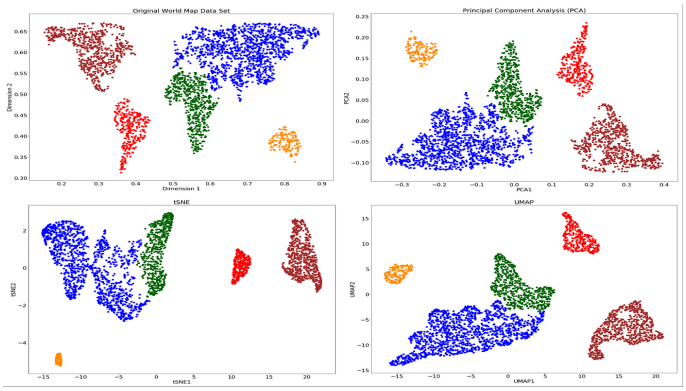


Figura 5 Preservación de estructura global en un dataset del mundo con PCA vs tSNE vs UMAP. Fuente: Towards data science.

UMAP es una técnica que utiliza algoritmos de representación de grafos para organizar los datos en un espacio de baja dimensión a partir de otro de alta dimensión. Esta técnica empieza por computar, a partir de los datos originales, las distancias por pares entre cada uno de los individuos en base a una métrica (habitualmente la distancia euclídea). A partir de estas distancias, se construye un conjunto llamado "fuzzy simplicial set", el cual representa la estructura del vecindario de los puntos, capturando un concepto de similitud entre puntos tomando en cuenta la conectividad y densidad de los datos. El objetivo de UMAP es el de mantener las relaciones de vecindad capturadas por el conjunto previamente mencionado, esto se hace en un proceso de optimización en el cual se busca disminuir la discrepancia entre las similitudes por pares en los espacios de alta y baja dimensionalidad. UMAP entonces construye una representación en grafo del "fuzzy simplicial set" el cual captura la estructura global de los datos. Se lleva entonces a cabo un refinamiento de para mejorar el conjunto, ajustando las posiciones de los puntos basándose en la estructura del grafo y la conectividad de los datos. Los procesos de optimización y refinado se repiten hasta que se consigue una convergencia por número de iteraciones o un criterio de parada.

A pesar de tener ventajas frente a t-SNE, UMAP no es una técnica a prueba de balas y su funcionamiento depende en gran medida de la selección de parámetros por parte del usuario.

Por otra parte, t-SNE es el otro algoritmo de reducción de dimensionalidad previamente mencionado. Su objetivo principal es tomar datos con muchas características y representarlos en un espacio de menor dimensión, manteniendo las relaciones y estructuras entre los datos lo mejor posible.

Al igual que en el caso de UMAP, TSNE empieza por calcular la distancia entre pares de puntos, a partir de las cuales se calculan distribuciones de probabilidad para cada punto que representan la similitud entre dicho punto y el resto; estas se transforman en probabilidades utilizando una

función softmax[3]. Se genera una representación de los datos originales en un nuevo espacio de baja dimensionalidad aleatoriamente.

De forma similar al espacio de alta dimensión, las similitudes entre pares de puntos de datos se calculan en el espacio de baja dimensión. Sin embargo, en el espacio de baja dimensión se utiliza una distribución diferente, denominada distribución t de Student, para modelar las similitudes entre pares.

El objetivo del t-SNE es minimizar la divergencia entre las similitudes por pares en el espacio de alta dimensión y las similitudes por pares en el espacio de baja dimensión. Esto se consigue mediante un proceso de optimización iterativo que suele realizarse mediante descenso por gradiente.

Como nota adicional, cada algoritmo tiene sus propias fortalezas y debilidades, por lo que es recomendable probar ambos para ver cuál se adapta mejor a los datos y al problema en cuestión.

En esta investigación la reducción de dimensionalidad se combinará con los algoritmos de agrupación en la fase de experimentación.

2.1.3 Optimización bayesiana

Es habitual que cuando se plantee un problema, para este exista más de una posible decisión o solución. Ante este tipo de problemas, es necesario decidir qué proceso de toma de decisiones vamos a seguir. Este proceso puede ser cualitativo (en base a juicios personales y sin una formalización matemática clara) o cuantitativo (donde el problema es formalizado matemáticamente, y el problema matemático es resuelto mediante un algoritmo o técnica determinado).

Para aquellos problemas cuantificables, la optimización matemática puede ser de gran ayuda. Esta está relacionada con la toma de decisiones y ayuda a escoger la mejor o mejores decisiones para una determinada situación en la que debemos decidir.

La técnica de optimización presente en la metodología ha sido la optimización Bayesiana. La optimización bayesiana es un enfoque para encontrar la configuración óptima de parámetros de un modelo o algoritmo que se basa en la combinación de métodos estadísticos y de aprendizaje automático. Se utiliza cuando la función objetivo (aquello que deseamos optimizar) es costosa de evaluar y no se dispone de su forma analítica.

[3] La finalidad de la función softmax es convertir un vector de números reales en una distribución de probabilidad, lo que significa que asigna probabilidades a cada elemento del vector de manera que la suma de todas las probabilidades sea igual a 1.

En la optimización bayesiana, se construye un modelo probabilístico denominado "modelo de respuesta" que captura la relación entre los parámetros de entrada y la función objetivo. Este modelo se actualiza a medida que se recopilan más evaluaciones de la función objetivo, utilizando un proceso llamado inferencia bayesiana.

Mediante la exploración inteligente de la función objetivo, la optimización bayesiana puede encontrar la configuración de parámetros que maximiza o minimiza la función objetivo, incluso cuando hay ruido o incertidumbre en las evaluaciones. En lugar de probar sistemáticamente cada posible combinación de parámetros, como se hace en la búsqueda en cuadrícula o en la búsqueda aleatoria, la optimización bayesiana se enfoca en regiones prometedoras, lo que la hace más eficiente en términos computacionales.

La optimización bayesiana en esta investigación se ha visto implicada en la selección del modelo óptimo acorde a una función objetivo definida para encontrar lo métodos que maximicen parámetros de evaluación de calidad de clustering.

2.1.4 Clustering supervisado

El clustering es una técnica de las ya mencionadas en el apartado de aprendizaje no supervisado, esto es así porque en el clustering habitualmente el algoritmo de agrupación organiza los datos en grupos basándose en patrones o similitudes de los propios datos.

El clustering supervisado, de forma opuesta al clustering clásico (no supervisado) es otra rama del aprendizaje automático donde el algoritmo aprende de datos etiquetados para realizar predicciones.

Hay distintas formas de atajar una tarea de clustering supervisado, siendo una de las opciones la que implica el uso de valores SHAP, con esta metodología habiendo probado ser capaz de obtener buenos resultados aplicada a conjuntos de datos de sintomatología del COVID-19 (Doyle, Doyle, & Bourke, 2021). A diferencia del enfoque tradicional, donde se aplica el clustering directamente sobre los datos originales, el clustering supervisado primero transforma estos valores en los mencionados valores SHAP.

Para computar estos nuevos valores, se requiere de una variable objetivo adecuada. Con las variables explicada y explicativa, se puede entrenar un modelo supervisado del cual posteriormente se extrae una nueva base de datos con las mismas dimensiones que la original, pero con unos nuevos valores que vienen determinados por la cantidad de información que aportan los individuos a la predicción del modelo supervisado.

Entrando en más detalle en cómo funciona SHAP, este es un método que explica cómo un modelo de aprendizaje automático realiza predicciones individuales. SHAP descompone una predicción en una suma de contribuciones de cada una de las variables de entrada del modelo. Para cada instancia de los datos, la contribución de cada variable de entrada a la predicción del modelo variará en función de los valores de las variables para esa instancia concreta.

La combinación de estas contribuciones se combina para explicar una predicción, se puede ver esto en el caso más simple, el de un modelo de aprendizaje automático que realiza una clasificación binaria (0 o 1), la salida del modelo es un valor entre cero y uno. Se sitúa un límite a partir del cual, valores superiores resultarían en la clase positiva y menores en la negativa. Esta salida puede ser interpretada como una probabilidad, donde un valor mayor coincide con predicciones positivas más seguras y uno menor con más negativas.

Lo que SHAP proporciona en última instancia es un conjunto de datos con las mismas dimensiones que el original, pero en lugar de con los valores originales, con los ya mencionados valores SHAP.

Lo más importante es que las predicciones del modelo de aprendizaje automático para cada instancia pueden reproducirse como la suma de estos valores SHAP, más un valor base fijo, tomando la siguiente forma:

$$f(x) = valor\_base + sum(valores\_SHAP)$$

Ecuación 5 Cálculo de valores shap en clasificación binaria.

La forma en la que se ha utilizado el clustering supervisado es como mejora al enfoque tradicional de reducción de dimensionalidad más clustering, que en algunos casos puede proporcionar grupos que no resultan demasiado informativos.

2.1.5 Reglas de decisión

Una regla de decisión es una sentencia del tipo si-entonces, consistente de una condición o antecedente y una predicción. Estas reglas siguen una estructura del tipo, si se cumplen las condiciones, entonces se hace una cierta predicción.

Este tipo de reglas son modelos de predicción muy interpretables, ya que siguen una estructura semántica muy similar al lenguaje natural y por tanto a la forma en que pensamos los seres humanos. Esto se cumple siempre y cuando las reglas se construyan con variables comprensibles y que la longitud de la condición sea breve con pocos pares característica = valor unidos por una condición lógica de tipo AND.

La utilidad de una regla de decisión se mide con dos valores: precisión y recall (precisión y apoyo). Siendo el recall el porcentaje de instancias para las cuales se cumple la condición de una regla y la precisión una medida de lo acertada que es la regla de decisión a la hora de predecir la clase correcta para los casos en los que aplica la condición de la regla.

Suele haber un equilibrio entre precisión y compatibilidad: Si añadimos más características a la condición, podemos lograr una mayor precisión, pero perdemos apoyo.

No se deben confundir las reglas de asociación con las de decisión, las reglas de asociación se centran en descubrir relaciones entre elementos o atributos de un conjunto de datos, mientras que las reglas de decisión se utilizan para tomar decisiones o realizar predicciones en tareas de clasificación.

La implicación de las reglas de decisión en el trabajo se sitúa en los esfuerzos de explicabilidad de los grupos, en concreto, se aprenden reglas de decisión para cada uno de los grupos resultantes diferenciando a cada grupo del resto. Estas reglas se aplican sobre los datos originales y no sobre los datos obtenidos con la mencionada metodología SHAP, ya que lo que se persigue es caracterizar a los grupos y esto no sería posible con valores SHAP, ya que lo que nos interesa es describir a los grupos con respecto a sus características de juego en el campo.

2.2 Sports analytics

Como ya hemos comentado, la disciplina de Sports analytics se encarga de aplicar distintas técnicas de analítica de datos al campo de los deportes. La finalidad más habitual del Sports analytics es el de proporcionar información relevante a las instituciones deportivas para obtener una ventaja competitiva frente al resto.

Por ejemplo, el equipo de fútbol alemán Bayern Múnich ha utilizado análisis deportivo avanzado para evaluar el desempeño de sus jugadores y oponentes, lo que les ha permitido tomar decisiones más informadas sobre tácticas y estrategias (McKenna, 2014). Esto no se limita únicamente a los equipos que participan en las competiciones, la marca de ropa deportiva Nike utiliza análisis de datos para evaluar las preferencias de los consumidores y diseñar productos que se ajusten mejor a sus necesidades y gustos (Barseghian, 2019).

Independientemente del deporte concreto del que estemos hablando, hay una serie de temas que son comunes a distintos deportes. Uno de los ejemplos recurrentes a través de las distintas disciplinas del deporte en equipo es el de la planificación de plantillas, en la cual, se buscan perfiles concretos para sustituir jugadores propios o incrementar la plantilla, es decir, partiendo de un perfil, la entidad se lanza a la búsqueda de otros similares.

Otro de los más habituales es el de la aplicación de técnicas estadísticas para la mejora de estrategias deportivas. Esto puede ser o bien el análisis de las debilidades del rival o el perfeccionamiento de la estrategia del propio equipo. Un caso real de esto es el equipo de baloncesto de la NBA, los Houston Rockets, el cual ha utilizado análisis de datos para optimizar su estrategia de tiro y maximizar la eficacia de sus jugadores (Sarnoff, 2018).

Alejándonos del contexto previo del Sports analytics, hay otras aplicaciones de esta disciplina relativas a las predicciones, en especial las de resultados. Habitualmente para esta tarea se cuenta con datos históricos o indicadores de rendimiento y se utilizan modelos de Machine learning para la predicción de resultados.

Como es habitual en las disciplinas técnicas y académicas, en la analítica deportiva hay multitud de foros dedicados a la difusión de trabajos e innovaciones en el campo. De entre las más relevantes, cabe destacar:

- **MIT Sloan Sports Analytics Conference**: una conferencia anual organizada por el Instituto de Tecnología de Massachusetts que reúne a líderes en el campo de análisis deportivo para discutir las últimas tendencias y desarrollos.
- **Sports Analytics World Series**: una serie de eventos globales que se enfocan en el análisis deportivo y el uso de tecnologías emergentes en el deporte.
- **Journal of Sports Analytics**: una revista académica que publica investigaciones originales en el campo del análisis deportivo.
- **Sports Analytics & Fan Engagement Summit**: una conferencia anual en Nueva York que se enfoca en cómo los equipos y las marcas pueden mejorar la experiencia del fanático utilizando análisis de datos y tecnología.
- **OptaPro Analytics Forum**: un evento anual organizado por OptaPro, que reúne a expertos en análisis deportivo para compartir ideas y discutir los últimos avances en el campo.

2.2.1 Casos de uso

Con el fin de ilustrar lo expuesto anteriormente, se presentan a continuación una serie de artículos con casos concretos de uso práctico de Sports analytics con aplicaciones más allá del fútbol.

El trabajo de (Tzai Lampisa, 2023) tiene la finalidad de construir y comparar distintos métodos para la predicción de resultados de partidos de básquet con datos proporcionados por OPTA. Los datos abarcan cinco torneos en Europa y el periodo 2013-2018 con más de 5000 partidos, y contienen, relativo a los equipos con tres tipos de información: general, relativa al estadio donde se juega cada partido, la fecha, el resultado y el ganador; de carácter ofensivo, tiros ejecutados y

con éxito, asistencias y rebotes en ataque y por último de carácter defensivo, rebotes bloqueos, faltas y robos. Para la selección de las características, los autores se han preguntado cuáles son los aspectos importantes que determinan el resultado final de un partido, la selección final cuenta con información histórica, sistemas de valuación (como "Elo rating"), la forma actual del equipo y las características concretas de la competición. Se cuenta finalmente con 110 características para las cuales se ha realizado el escalado que aparece en la figura 6.

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)}$$

Figura 6 Escalado estándar de características (Tzai Lampisa, 2023).

Los modelos que se han utilizado son del tipo explicable dado el interés que se tiene en interpretar el funcionamiento de las técnicas. Se han utilizado tres: regresión logística, random forest y extreme gradient boosting trees. Por último, la valoración de los resultados se realiza comparando los ofrecidos por los modelos con un modelo "vanilla" utilizando sólo la información del nombre del equipo para cada juego, asumiendo el efecto positivo de jugar como local a través de todas las competiciones. Los resultados muestran que cada uno de modelos se adecua mejor a una de las competiciones respecto al resto, que no hay ningún modelo que destaque positivamente sobre el resto, aunque sí que es verdad que en conjunto la regresión logística funciona peor que los otros dos métodos. Como posible mejora a futuro se podría tal vez comparar la efectividad de los modelos que se han utilizado con otros de tipo "caja negra", teniendo en cuenta que los autores posiblemente los hayan excluido dadas las limitaciones que estos últimos tienen a la hora de interpretar resultados.

Pasamos ahora a artículos de Sports Analytics concretos del futbol, empezando por con el estudio de (Bransen & Van Haaren, 2018), el objetivo de este documento es presentar un nuevo método para medir la contribución de los jugadores de fútbol a través de los pases. El conjunto de datos incluye datos de 9.061 partidos de siete ligas, abarcando las temporadas 2014/2015 a 2017/2018 y fue proporcionado por Wyscout. Para cada partido, el conjunto de datos contiene información sobre los jugadores (nombre, fecha de nacimiento y posición) y los equipos (es decir, alineación inicial y sustituciones), así como datos de eventos jugada a jugada que describen los acontecimientos más notables ocurridos en el terreno de juego. Para cada evento, el conjunto de datos proporciona la siguiente información: marca de tiempo, equipo y jugador que realiza el evento, tipo (por ejemplo, pase o disparo) y subtipo (por ejemplo, pase cruzado o alto), y lugar de inicio y fin. El documento describe varias técnicas utilizadas en el enfoque propuesto para medir

las contribuciones de los jugadores de fútbol en los pases durante los partidos. Estas técnicas incluyen:

1. Agrupación de las secuencias de posesión utilizando técnicas sofisticadas para identificar patrones en datos de series temporales.
2. Utilización del algoritmo XGBoost para entrenar el modelo de goles esperados.
3. Cálculo de las distancias entre las secuencias de posesión utilizando la deformación temporal dinámica (DTW) como función de distancia.
4. Búsqueda de vecinos más cercanos con DTW para determinar la recompensa esperada de una secuencia de posesión.

En general, el enfoque propuesto combina varias técnicas de aprendizaje automático, minería de datos y análisis deportivo para valorar cada pase calculando la diferencia entre la recompensa esperada de la secuencia de posesión que constituye el pase antes y después del pase. La evaluación empírica sobre el conjunto de datos del mundo real demostró que el enfoque propuesto es capaz de identificar diferentes tipos de jugadores con impacto. Una posible crítica al enfoque adoptado por el trabajo es que se basa en gran medida en el modelo de goles esperados para valorar los tiros, que puede no reflejar siempre el verdadero valor de un tiro. El modelo de goles esperados se basa en datos históricos y puede no tener en cuenta el contexto específico de un partido, como la calidad de la defensa del equipo contrario o el marcador actual.

Alejándonos un poco del pase como única métrica y abordando un problema distinto dentro de la analítica en el fútbol tenemos la investigación de (García-Aliaga, Marquina, Coreton, Rodriguez-Gonzalez, & Luengo-Sanchez, 2021) cuyo propósito es analizar los datos de rendimiento de los jugadores de fútbol utilizando técnicas de aprendizaje automático para identificar las principales variables que caracterizan las distintas posiciones en el juego y profundizar en la dinámica real del juego. Se cuenta con una base de datos proporcionada por Opta con tres grupos de variables, las de acciones ofensivas, defensivas y acciones relativas a la construcción del juego. Los jugadores además han sido agrupados en una de las 9 posiciones que se proporcionan en el conjunto de datos original. El artículo utiliza técnicas de aprendizaje automático, en concreto aprendizaje no supervisado y reducción de dimensionalidad, para analizar datos de rendimiento de jugadores de fútbol. Los autores utilizan estas técnicas para identificar los comportamientos técnico-tácticos de los jugadores en función de sus estadísticas y determinar las variables más influyentes en cada posición. También utilizan estas técnicas para identificar jugadores anómalos y valores atípicos en los datos. El artículo menciona el t-SNE y el UMAP (cuyas proyecciones aparecen posteriormente en la figura 7) como técnicas habituales de reducción de la dimensionalidad, cuyo objetivo es proyectar datos de alta dimensión en menos dimensiones

preservando la estructura de vecindad de los datos. La puntuación de Kullback-Leibler se utiliza para reubicar correctamente a los jugadores en el espacio de baja dimensión.

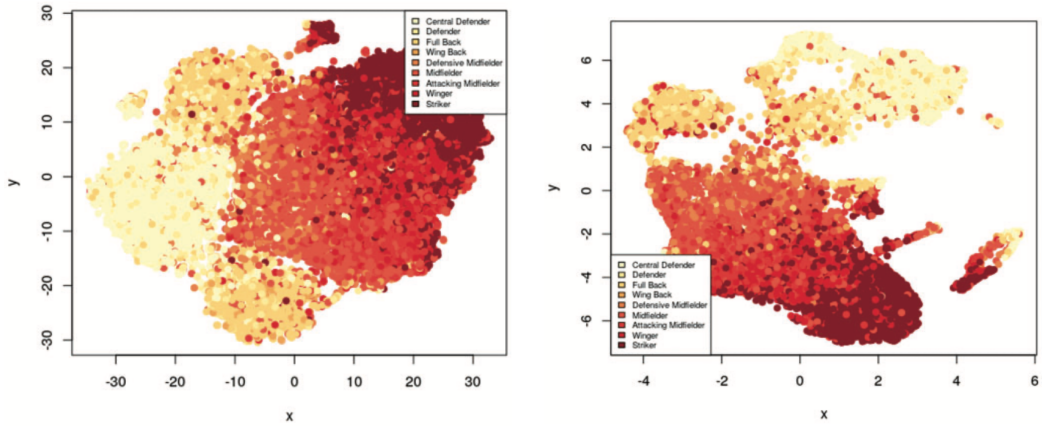


Figura 7 Comparación de proyecciones con UMAP y t-SNE (García-Aliaga, Marquina, Coreton, Rodriguez-Gonzalez, & Luengo-Sanchez, 2021).

El artículo menciona que se utilizó el algoritmo RIPPER para generar un conjunto de reglas de clasificación con el fin de identificar las variables que mejor caracterizan cada posición de los jugadores en el campo. Los autores validaron las reglas de clasificación mediante un proceso de validación cruzada de 10 veces. Los resultados mostraron que el algoritmo RIPPER podía clasificar con exactitud a los jugadores en sus respectivas posiciones con altos valores de precisión y recuerdo. También se utilizaron técnicas de aprendizaje automático para reducir la dimensionalidad e identificar jugadores anómalos y valores atípicos en los datos. Los resultados mostraron cuatro grupos de valores atípicos, y los autores pudieron identificar el origen de los errores y excluir estos datos del análisis. En general, los resultados del estudio proporcionan una visión más profunda de los comportamientos técnico-tácticos de los jugadores de fútbol y de las variables que mejor caracterizan cada posición en el campo. El estudio también demuestra la utilidad de las técnicas de aprendizaje automático para analizar datos de rendimiento deportivo e identificar valores atípicos y errores en los datos. Como posible crítica, hay que señalar que el estudio sólo se centra en los comportamientos técnico-tácticos y no tiene en cuenta los descriptores espaciotemporales, que podrían aportar información adicional sobre el rendimiento de los jugadores. Por último, el estudio no ofrece aplicaciones prácticas ni recomendaciones para entrenadores o equipos basadas en los resultados. Aunque el estudio proporciona datos interesantes sobre el rendimiento de los jugadores, sería útil ver cómo se podría utilizar esta información para mejorar el rendimiento del equipo o el desarrollo de los jugadores.

El artículo de (Stanojevic & Gyarmati, 2016) se encarga de predecir el valor monetario de los jugadores a partir de los atributos técnicos y tácticos de los jugadores. Aquí se obtiene información de distintas fuentes; el valor estimado de los jugadores se obtiene de la página

TransferMarkt (estos valores sirven para medir la bondad de las predicciones) mientras que los datos relativos al rendimiento de los jugadores provienen de InStat y contienen seis temporadas además de una cantidad enorme de partidos (+100000), se tienen en cuenta variables como pases, goles, pases precisos, duelos, entradas… La metodología consiste en extraer características de los datos de rendimiento, incluidos los números absolutos de los distintos eventos en los que participaron los jugadores, como asistencias, entradas, pases, desafíos, pases clave, disparos, regates, desafíos aéreos y centros. Los autores también extraen características relacionadas con la edad, la posición, la altura y el equipo del jugador. A partir de estas características, los autores construyen un modelo de regresión para predecir el valor de mercado del jugador. El modelo se entrena utilizando las estimaciones del valor de mercado de transfermarkt.com como señales supervisoras, consideradas valores de mercado reales perturbados por una variable de ruido. Los autores evalúan la solidez de su estimación del valor de mercado en función del rendimiento (PDMVE) para predecir los resultados de los equipos y la comparan con las estimaciones del valor de mercado de transfermarkt.com, esto aparece ejemplificado en la figura 8.

Player name	TMVE (M £)	PDMVE (M £)
neymar	60.0	69.55
eden hazard	52.5	57.72
cesc fabregas	37.5	49.54
sergio aguero	45	48.81
lionel messi	90	48.53
nolito	7.5	46.14
luis suarez	60.0	40.65
thomas muller	41.25	38.08
marco verratti	30.0	36.07
diego costa	37.5	35.79

Player name	TMVE (M £)	PDMVE (M £)
nolito	7.5	46.14
dani alves	7.5	25.39
karim bellarabi	9.0	23.39
dries mertens	13.5	26.73
cesc fabregas	37.5	49.54
willian	22.5	34.32
graziano pelle	8.25	19.90
arjen robben	21	32.00
mikel san jose	3.75	14.56
mario gaspar	4.5	14.49

Figura 8 Valor de mercado real contra predicho (Stanojevic & Gyarmati, 2016).

Como posible punto de mejora, se podrían añadir al modelo factores externos como las lesiones o los conflictos con los entrenadores, que pueden afectar significativamente al valor de un jugador.

Ahondando en algunas de las ideas anteriormente mencionadas, tenemos el TFM de (Malagón Selma, 2019), en este caso, la base de datos con 2662 jugadores con variables cualitativas (nombre, posición, equipo y liga) así como cuantitativas (número de pases realizados con éxito, número de robos, número de goles…) contabilizadas por temporada, estas últimas se han transformado a acciones por partido, siguiendo la siguiente fórmula (ecuación 5):

$$(Variable\ /Minutes\ played)\ *\ 90minutos$$

Ecuación 6 Fórmula de acciones por partido (Malagón Selma, 2019).

además, se eliminaron las variables con alta correlación entre ellas. Este trabajo tiene dos objetivos, el de encontrar jugadores similares a uno dado y el de analizar las variables más

relativas asociadas a cada posición. La primera técnica que se aplica es la de PCA, previo rescaldado y centrado de los datos, se comprueban los residuos para eliminar los atípicos. Además, se han seleccionado sólo los jugadores con más del 20% de los minutos disputados para reducir la distorsión. Los scores del PCA se transforman en una matriz de correlación a partir de la cual se computarán las distancias entre jugadores. Para más detalle en la selección de perfiles similares se combina el "Loading plot" con el "Score plot" (figura 9) que proporciona el PCA, con lo cual se obtienen de entre los perfiles más cercanos, los más similares de acuerdo con una serie de características relevantes para el conjunto anterior.

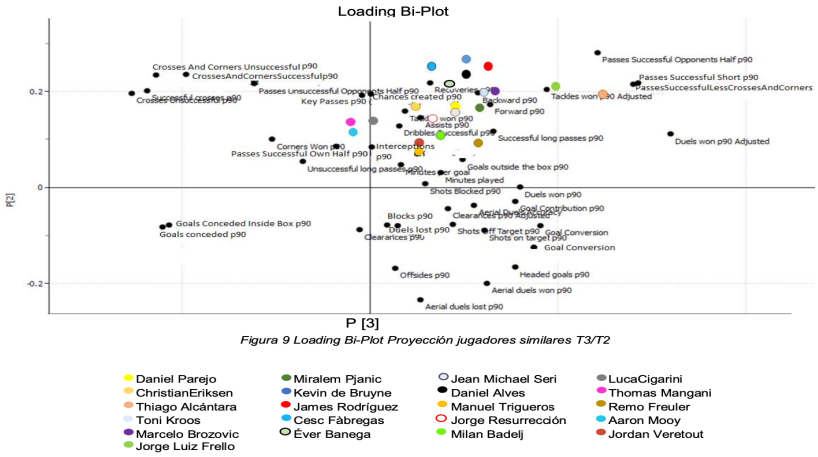


Figura 9 Loading Bi-Plot Proyección jugadores similares T3/T2

Figura 9 Proyección PCA de jugadores similares (Malagón Selma, 2019).

Además, en caso de querer comparar al detalle, se pueden seleccionar dos jugadores para compararlos mediante un gráfico de radar, donde las variables seleccionadas son las más relevantes para el jugador referencia, tal y como se muestra en la figura 10.

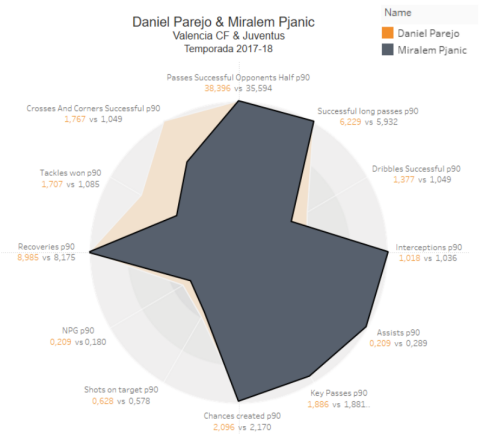


Figura 10 Comparación en un gráfico de radar de los centrocampistas Daniel Parejo y Miralem Pjanic (Malagón Selma, 2019).

Por otro lado, en relación con el segundo objetivo, se elabora un gráfico de contribuciones para cada una de las tres posiciones de la base de datos original (defensa, mediocentro o delantero). A través del análisis de este gráfico será posible identificar las variables que destacan en posiciones disgregadas al comparar los gráficos de los jugadores que compiten en estas demarcaciones. De este modo, se comprobará si existen diferencias en el estilo de juego y, en su caso, se determinarán qué variables resultan características para cada tipo de jugador en cada posición original, en concreto en la memoria del TFM se menciona la diferencia entre las variables que caracterizan a los defensas centrales y a los laterales (figura 11).

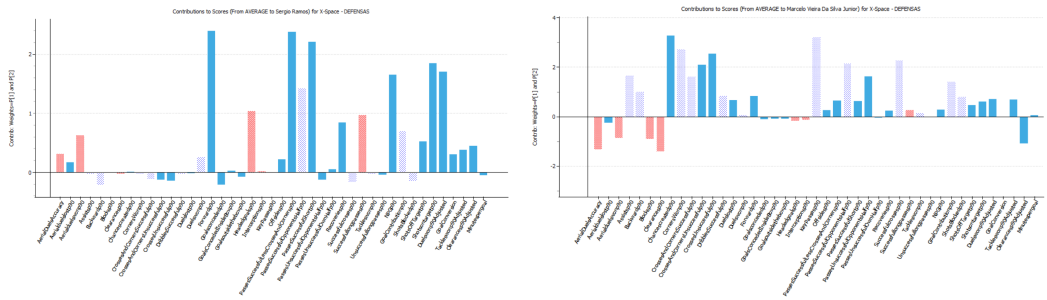


Figura 11 Comparación de defensas central y lateral (Malagón Selma, 2019).

Para la validación de los resultados relativos al segundo objetivo se utiliza un modelo Random Forest con los datos ya etiquetados.

En el contexto de la búsqueda de jugadores similares contamos con el artículo de (Lopez Peña & Sanchez Navarro, 2015) cuyo objetivo es el de buscar el reemplazo idóneo del centrocampista Xavi Hernández basándose en secuencias de pases de longitud tres como los que se pueden ver en la figura 12.

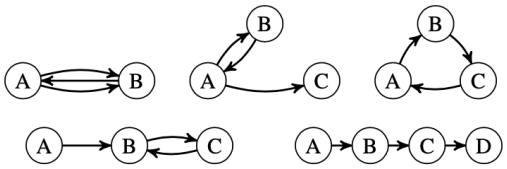


Figura 12 Conjunto de pases de longitud tres (Lopez Peña & Sanchez Navarro, 2015).

El dataset con el que se trabaja ha sido proporcionado por la empresa Opta y en su forma original contiene información de eventos. A partir del dataset original, se ha obtenido, para cada uno de los jugadores del dataset, se calcula el número medio de apariciones de cada uno de los motivos de pase y se utilizan los resultados como vector de características para describir el estilo de cada jugador. La recopilación de datos, el ajuste de modelos, el análisis y el trazado de gráficos se realizaron con IPython y con librerías como matplotlib.

Tras computar estos valores, se aplican las técnicas de Clustering y PCA. Con el PCA se reduce la dimensionalidad y además se aprovecha para mostrar las dos primeras componentes que son aquellas que interesan para encontrar perfiles similares al sujeto de estudio, Xavi Hernández. Y con el Clustering se consiguen mostrar más de 30 grupos con su correspondiente representante (un jugador) y también aquí se comprende visualmente la similitud entre jugadores (dada la distancia que representa la similitud) y los distintos tipos de perfiles en el terreno de juego. La distancia entre jugadores es la distancia euclídea entre los valores del vector de características, la cual es el tipo de distancia más habitual.

Figura 13 Gráfico comparativo de las dos primeras componentes del PCA (Lopez Peña & Sanchez Navarro, 2015).

Los resultados visibles en la figura 13 muestran en primer lugar que Xavi, el jugador de referencia se encuentra muy separado del resto (haciendo de su perfil uno verdaderamente único), los perfiles más similares ya se encuentran en el mismo equipo (en la temporada del estudio) y que de entre el resto, el mejor candidato por restricciones de edad u otras es Danny Blind. Como posible mejora se podría señalar la falta de validación de los resultados, que bien podría haber sido realizada separando el datasest en entrenamiento o validación o bien mediante opiniones de expertos en el campo.

Otro de los artículos centrados en la búsqueda de perfiles similares es el de (Mazurek, 2018). El objetivo de este trabajo es utilizar el análisis estadístico para determinar qué jugador de fútbol actual se parece más a Lionel Messi. Para ello utiliza la base de datos que se utiliza en este caso es la de WhoScored, la cual cuenta con 24 parámetros relativos a cada jugador entre los cuales se encuentran goles, disparos por partido, asistencias, precisión de los pases, regates, faltas… relativos a las 5 ligas más importantes de Europa durante la temporada 2017/18. Los autores

recurren a la normalización (escalado de características) de los datos para tener en cuenta las distintas unidades y escalas de los criterios seleccionados. El artículo también incluye análisis de correlación y visualizaciones (una de las cuales se muestra en la figura 14) para ilustrar las relaciones y patrones de los datos. En general, el artículo utiliza una combinación de técnicas estadísticas para analizar y comparar el rendimiento de los futbolistas. Se calcula la distancia entre los vectores de características asociados a los jugadores mediante la distancia de Manhattan dada su sencillez de uso y evaluación.

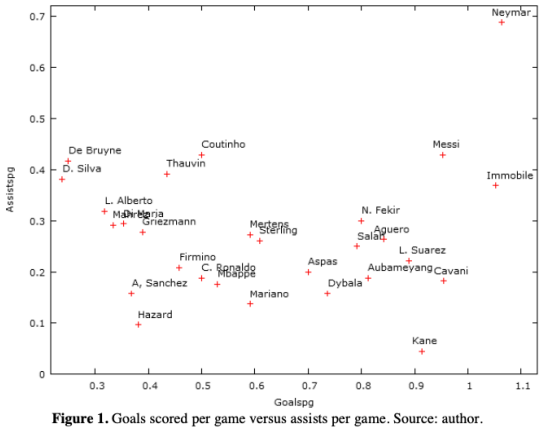


Figure 1. Goals scored per game versus assists per game. Source: author.

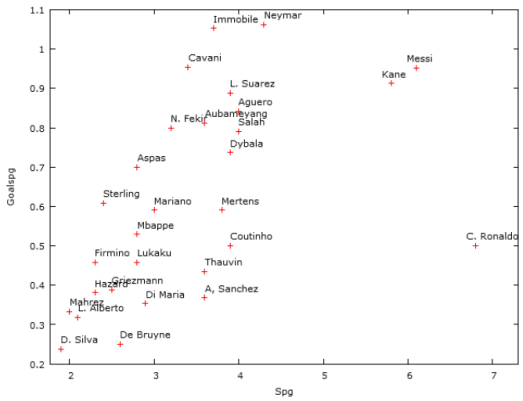


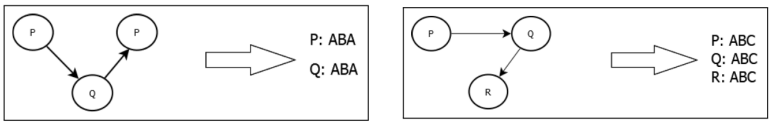
Figure 2. Shots per game versus goals per game. Source: author.

Figura 14 Proyecciones por pares de variables de futbolistas (Mazurek, 2018).

El documento concluye que Neymar es el jugador que más se parece a Messi según los criterios seleccionados. Como posibles críticas al documento cabe señalar la falta de validación, ya que, aunque los autores utilizan el análisis estadístico para comparar el rendimiento de los futbolistas en función de los criterios seleccionados, no está claro hasta qué punto son fiables los resultados sin un proceso de validación; y la subjetividad de la selección de criterios, pues el proceso de selección de criterios puede ser subjetivo y estar abierto a la interpretación. Otros investigadores pueden seleccionar criterios diferentes o ponderarlos de forma distinta, lo que podría dar lugar a resultados diferentes.

En el mismo sentido que el anterior artículo encontramos el trabajo de (Barbosa, Ribeiro, & Dutra, 2022), en este, el objetivo es el de proponer un método para medir la similitud entre jugadores de fútbol basándose en sus secuencias de pases, utilizando motivos como bloques de construcción de redes complejas. Los autores pretenden proporcionar una forma más objetiva de puntuar y comparar diferentes enfoques, así como de identificar jugadores similares basándose en sus estilos de pase y de juego. También exploran el papel de la posición del jugador en los motivos en los que participa, y prueban su enfoque con datos de diferentes ligas. Los datos abarcan la temporada 2017/18 y cuentan con partidos desglosados por eventos. Se preprocesaron los datos brutos para

transformarlos en secuencias de pases entre los jugadores implicados y para eliminar a los jugadores que no participaron en al menos el 80% de los partidos de la temporada.



(a) ABA motif (b) ABC motif

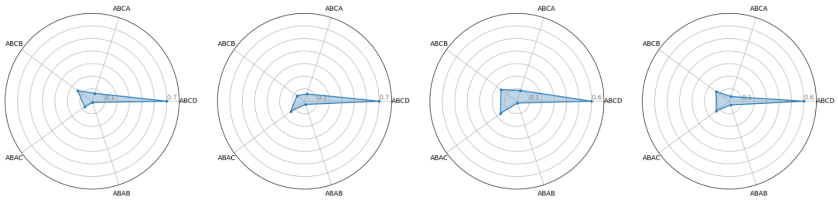
Figura 15 Motivos de pase considerados en este estudio (Barbosa, Ribeiro, & Dutra, 2022).

Los autores entonces extraen motivos de tamaño 3 y 4 de las secuencias de pases de los jugadores algunos de los cuales aparecen en la figura 15 anterior y cuentan su frecuencia para cada jugador. A continuación, calculan la distancia entre los jugadores en función de los motivos en los que participan, utilizando la distancia cuadrática con la fórmula que se muestra en la figura 16.

$$D(A,B) = \sqrt{\sum_{m \in M} (A_m - B_m)^2}$$

Figura 16 Distancia entre 2 jugadores (A y B) donde M es el conjunto de todos los motivos y A_m es la frecuencia del motivo m para el jugador A (Barbosa, Ribeiro, & Dutra, 2022).

También se tiene en cuenta la posición específica que cada jugador representa en el motivo, lo que aporta más variabilidad y características para ayudar a separar a los jugadores entre sí. Se evalúa el modelo dividiendo el conjunto de datos en dos mitades y calculando el número de jugadores en el top 10 de similitud en las diferentes mitades. Los autores también realizan un análisis visual de los gráficos de radar de algunos jugadores para identificar su estilo de juego tal y como se puede visualizar en la figura 17.



(a) Ederson (b) Cédric Soares (c) Jorginho (d) Messi

Figura 17 Gráfico de radar para 4 jugadores (Barbosa, Ribeiro, & Dutra, 2022).

Los resultados muestran que el método propuesto identifica eficazmente a jugadores similares basándose en sus estilos de pase y juego. Para finalizar, hay que decir que el enfoque adoptado en el artículo se basa en gran medida en el uso de motivos. Aunque los motivos pueden ayudar a identificar patrones en las secuencias de pases, es posible que no capten toda la complejidad del

estilo de juego de un jugador. Por ejemplo, un jugador puede tener una forma única de colocarse en el campo o de regatear que los motivos no captan.

Además de ayudar a comprender la estructura y el tipo de experimentación que se realizan, las limitaciones que se exponen en los distintos artículos nos ayuda a poder seleccionar tareas que no han podido ser realizadas en estos; así, tras consultar los distintos trabajos, se ha decidido incluir más características además de los pases e incluir los datos relativos al valor de mercado de los jugadores, lo cual añade una funcionalidad para el usuario final.

2.3 Crítica al estado del arte

Revisando los diversos artículos de analítica deportiva y especialmente los de fútbol, queda claro que mucha de la información de este tipo no se encuentra disponible para el público, dado el elevado valor de dicha información. No es por tanto de extrañar que se encuentren pocos trabajos de este tipo. Otra consecuencia de esto es que los trabajos no pueden contar con la información actualizada, ya que los pocos datos abiertos son de hace algunas temporadas.

Por otro lado, en cuanto a la evolución de los trabajos sí que se puede observar que se va aumentando la complejidad de los análisis y la metodología, así como de la inclusión de nuevas técnicas como es el caso de las técnicas de reducción de la dimensionalidad TSNE y UMAP. También en las técnicas de validación se ve una mejoría que empieza con la validación clásica dividiendo los datos en conjuntos de entrenamiento y validación para acabar con el uso de herramientas de aprendizaje supervisado (random forest) para validación.

Para concluir, los diferentes artículos han servido de guía para conocer la forma de atajar esta clase de problemas, así como para conocer posibles mejoras en el campo de interés particular de búsqueda de perfiles similares.

2.4 Solución Propuesta

Esta investigación gira entorno a la tarea de automatizar la búsqueda de similitudes en conjuntos de datos con un gran número de dimensiones, al proporcionar un enfoque sistemático y basado en los datos para identificar jugadores con atributos técnicos similares, lo que puede ser valioso para los clubes de fútbol en la toma de decisiones relacionadas con la contratación y el desarrollo de jugadores.

Esto se enmarca en el uso de técnicas de preprocesado y de aprendizaje automático, para expandir el número de sujetos que se pueden comparar (ya que, de otra manera, con las técnicas clásicas de reclutamiento, se debían invertir muchos más recursos económicos en desplazamientos y

seguimientos, así como en ojeadores para poder abarcar una mayor área geográfica) así como para comprender la justificación de las similitudes.

Este trabajo, se encarga de una tarea recurrente en la analítica deportiva, y, aunque la bibliografía disponible al respecto no es todo lo extensa que cabría esperar por el enfoque en la privacidad de los clubes, sí que se pueden establecer tanto nexos como divergencias con otras investigaciones similares. Como similitud principal, hay que destacar que este trabajo utiliza algunas técnicas que ya se han propuesto para abordar esta tarea, como es el caso de las técnicas de reducción de dimensionalidad (PCA, t-SNE y UMAP) o el uso de atributos similares (pases, disparos, duelos…). No obstante, el planteamiento que se ha tomado innova en el uso de una búsqueda de hiperparámetros óptimos, así como el uso del clustering supervisado para la mejora en la obtención de grupos.

Resumidamente, a pesar de que la tarea en cuestión no es una propuesta desconocida, esta aproximación pretende ahondar e incorporar avances en el campo de la ciencia de datos para resolver una tarea especialmente relevante para los clubes en la planificación deportiva.

2.4.1 Software

Para llevar a cabo la propuesta anteriormente mencionada se ha utilizado el lenguaje de programación Python y varias de sus librerías.

Python es un lenguaje de programación de alto nivel que se caracteriza por su sintaxis clara y legible. Este lenguaje cuenta con una amplia biblioteca estándar y una gran cantidad de paquetes y módulos de terceros, lo que permite a los desarrolladores acceder a una amplia gama de herramientas y funcionalidades.

Para poder adaptar la base de datos original a nuestra investigación se ha hecho uso de la librería Pandas. Pandas proporciona estructuras de datos flexibles y eficientes, los DataFrames, que permiten almacenar y manipular datos de manera tabular. Además, esta librería facilita tareas como la limpieza y preparación de conjuntos de datos, la manipulación de columnas y filas… así como funciones avanzadas como la capacidad de fusionar y combinar conjuntos de datos y realizar operaciones de agregación y agrupamiento.

Para las tareas de clustering y reducción de dimensionalidad ha sido utilizada la librería sklearn, también conocida como scikit-learn (Buitinck, y otros, 2011). Esta aglutina diversos algoritmos y herramientas para realizar tareas de aprendizaje supervisado y no supervisado.

Para poder visualizar los resultados de las tareas previamente mencionadas han sido de gran utilidad las librerías Matplotlib (Hunter, 2007), Seaborn (Waskom, 2021) y Plotly.

- Matplotlib es una biblioteca de trazado 2D que permite crear diversos gráficos. Proporciona una gran flexibilidad y control sobre la apariencia de los gráficos, lo que permite personalizar casi todos los aspectos. Matplotlib sin embargo puede requerir un poco más de código para crear visualizaciones más complejas.
- Por otro lado, Seaborn se basa en Matplotlib y proporciona una interfaz de alto nivel para crear visualizaciones atractivas y estilizadas. Seaborn simplifica la creación de gráficos estadísticos complejos. Además, Seaborn tiene una mayor capacidad de personalización en comparación con Matplotlib, ya que ofrece temas visuales predefinidos que mejoran la estética de los gráficos.

En resumen, Matplotlib es una biblioteca más generalizada que ofrece un control detallado sobre la apariencia de los gráficos, mientras que Seaborn es una biblioteca especializada en visualizaciones estadísticas y proporciona una interfaz más sencilla y estilizada.

Optuna (Akiba, Sano, Yanase, Ohta, & Koyama, 2019) se ha aplicado en la tarea de optimización de los modelos, su objetivo principal es encontrar la configuración óptima de hiperparámetros[4] para modelos de aprendizaje automático mediante la búsqueda automática y sistemática en el espacio de búsqueda definido. La ventaja principal de Optuna es su capacidad para automatizar y simplificar el proceso de ajuste de hiperparámetros. En lugar de realizar ajustes manuales y tediosos, Optuna realiza una búsqueda inteligente en el espacio de hiperparámetros definido, explorando y evaluando diferentes combinaciones de valores de forma eficiente. Esto ahorra tiempo y recursos al encontrar rápidamente los hiperparámetros que optimizan el rendimiento del modelo.

En cuanto a la metodología de clustering supervisado, esta tarea se ha podido abordar en Python gracias a la ayuda de dos librerías, en primer lugar, la librería SHAP, con la cual se obtienen valores Shapley y se muestran varias visualizaciones de la propia librería, y las reglas de decisión, con la librería Skope-Rules.

[4] Los hiperparámetros son parámetros cuyos valores controlan el proceso de aprendizaje y determinan los valores de los parámetros del modelo que un algoritmo de aprendizaje acaba aprendiendo.

3 Análisis del problema

En cualquier trabajo que implique el uso de datos, es necesario como una buena práctica profesional y de confianza llevar a cabo una revisión del marco legal, así como la traza de los datos en conjunto.

3.1 Análisis del marco legal y ético

Dado el alto valor económico que reportan los datos recogidos por las principales empresas dedicadas a la analítica del fútbol no es de extrañar que la gran mayoría de estos se mantengan privados y sólo se comercialicen a los clubes y, por tanto, dichos datos no están al alcance de usuarios individuales.

Los archivos de datos disponibles en abierto son conjuntos de información que están disponibles para que el público en general los utilice, reutilice y redistribuya de manera libre. Estos datos se proporcionan en formatos legibles por máquina y se publican en portales o plataformas en línea.

Los datos de Wyscout utilizados, a pesar de que vengan generados por una empresa privada, se han publicado bajo la licencia CC BY 4.0 y están a disposición del público en la página web figshare. Cada conjunto de datos se proporciona en formato JSON (JavaScript Object Notation), un formato estándar abierto que utiliza un lenguaje legible por humanos y procesable por máquinas.

Para lo obtención de estos datos se llevó a cabo un proceso de etiquetado de los eventos futbolísticos a partir de los vídeos de los partidos por parte de los autores. Esto se hace en los siguientes pasos:

(a) Captura de pantalla del software de etiquetado. Una acción es etiquetada por un operador a través de un teclado especial personalizado, creando así un nuevo evento en la línea de tiempo del partido (figura 18).

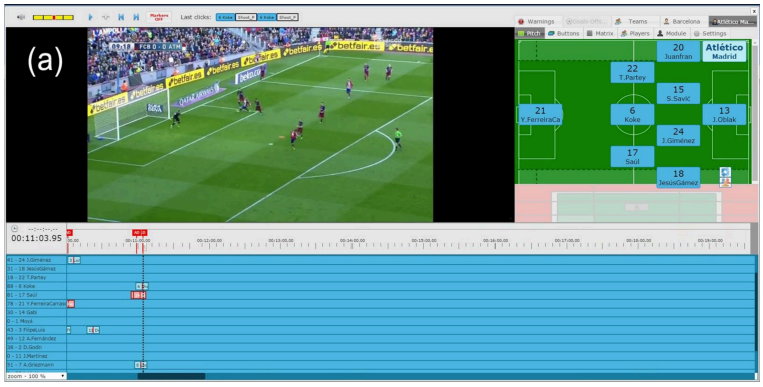


Figura 18 Software de etiquetado de los datos originales. Fuente: Wyscout.

(b) Cuando se establece la posición del evento en el terreno de juego, aparece el módulo de introducción de datos específicos (figura 19 arriba). También aparecen módulos de entrada relacionados con el evento para establecer atributos adicionales del evento que se está produciendo (figura 19 abajo).

Figura 19 Registro de eventos. Fuente: Wyscout.

Toda esta información y más detalles respectivos a los datos originales se encuentra disponible en el trabajo original (Pappalardo, y otros, 2019).

4 Propuesta

Este trabajo consta de un conjunto de tareas que parten del conjunto de datos originales hasta llegar a la visualización de los resultados obtenidos de la experimentación. Este proceso se puede ver ilustrado en la siguiente imagen (figura 20), donde aparecen las tareas representadas gráficamente, así como el orden y el paso de una a otra tarea representado por las flechas, uno de los pasos el de la transformación en valores SHAP no es estrictamente necesario, de hecho, se han llevado a cabo los experimentos sobre los datos en bruto y posteriormente con estos valores.

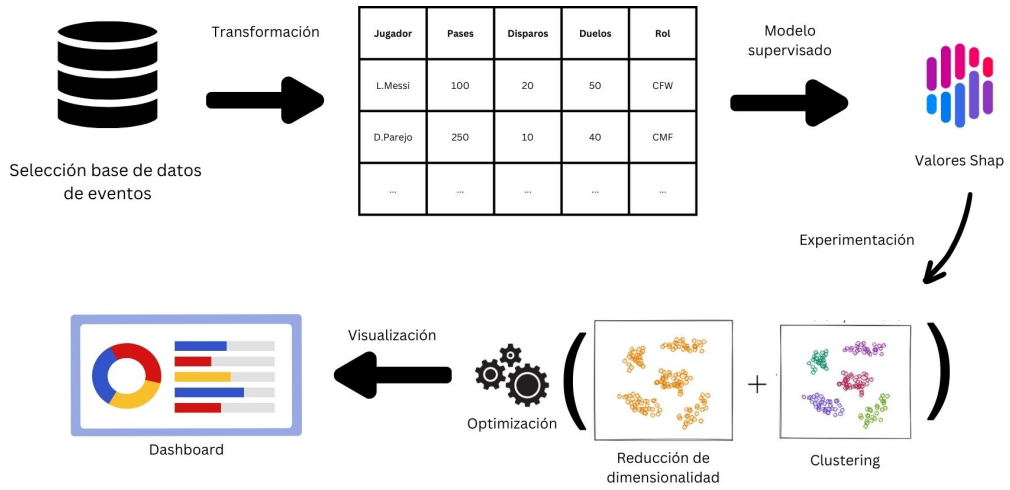


Figura 20 Propuesta de trabajo con sus distintas fases. Elaboración propia.

Se entrará a continuación al detalle en cada una de las tareas mostradas para comprender qué se lleva a cabo en cada una de ellas.

4.1 Preparación y comprensión de datos

La materia prima alrededor de la cual gira esta investigación son los datos. Este primer caso consiste en, primero buscar una fuente de datos disponible y luego adaptarla para el cumplimiento de los objetivos.

4.1.1 Selección de datos

Como ya se ha establecido previamente, las empresas que ofrecen datos del mundo del fútbol los mantienen lejos del público por norma general, los datos actualizados se mantienen en privado y

mucha de la información que se publica no persigue la finalidad de ser útil para un eventual análisis, sino que más bien, se muestran conjuntos de datos muy limitados que sirven como reclamo publicitario. Esto añade una capa de dificultad a la hora de encontrar información útil para tareas de análisis.

Los pocos conjuntos de datos abiertos abordan distintas temáticas que no se limitan a información de eventos de futbolistas, sino que también se incluyen datos de equipos, estadios, resultados de partidos e incluso de videojuegos con temática futbolística, lo cual no es de utilidad en este trabajo. Las bases de datos que sí se centran en características de juego de futbolistas tienen diversos defectos que han hecho que fueran descartadas. Algunas de ellas no contaban con datos información completa (ausencia de partidos, individuos, características…) mientras que otras estaban muy simplificadas en cuánto a las características seleccionadas.

Teniendo en el foco el hecho de que se buscan grupos de jugadores partiendo de información de eventos que permitan analizar el comportamiento de un futbolista durante un partido, se ha seleccionado la base de datos que más se ajusta a estos parámetros. El conjunto de datos seleccionado cuenta con registros completos de todos partidos de una temporada, con una diversidad de eventos y además se puede seguir todo el proceso que se realizó en la toma de dichos datos lo cual añade garantías acerca de la calidad de la información.

4.1.2 Fuente de datos

La base de datos de (Wyscout, s.f.) cuenta con eventos de tipo espaciotemporal de partidos de fútbol (pases, entradas, disparos…) de las ligas de primera división española, inglesa, alemana francesa e italiana, respectivos a la temporada 2017-2018. Cada partido contiene para todo evento, información acerca de su posición, el tiempo en el contexto del partido, el resultado el jugador implicado además de otras características.

De todos los conjuntos de datos disponibles del repositorio público, los más relevantes en este caso son los que contienen información acerca de los jugadores y los eventos. Se verá a continuación al detalle cada uno de estos conjuntos de datos.

En primer lugar, el conjunto de eventos cuenta con eventos generados a lo largo de los partidos y almacena las características que se enumeran a continuación en la tabla 1.

Tabla 1 Fuente de datos: conjunto de eventos

eventId	Identificador del tipo de evento. Cada identificador tiene asociado un nombre de evento.

eventName	Nombre del evento, hay siete posibles: pase, falta, disparo, duelo, tiro libre, fuera de juego y toque.
subEventId	Identificador del sub-evento, el cual añade más detalle al evento, al igual que con el identificador del evento, el del sub-evento tiene asociado un nombre de sub-evento.
subEventName	Nombre del sub-evento. Cada evento tiene asociado un conjunto de sub-eventos distinto.
tags	Una lista de etiquetas de evento, cada una da información adicional sobre el evento y cada evento tiene asociado un conjunto de etiquetas distinto asociado. Un ejemplo de etiqueta podría ser la etiqueta de pase completado o en el caso de los disparos, que acabe entre los tres palos.
eventSec	El tiempo en el que el evento ocurre, se cuenta en segundos desde el inicio de la correspondiente parte en la que se encuentre el encuentro.
playerId	El identificador del jugador que genera el evento. Este valor se corresponde con el valor del campo “wyid” en el conjunto de datos de jugadores.
positions	En esta columna se almacena la información acerca del origen y el destino de cada uno de los eventos en coordinadas. Estas coordinadas son un par de valores (x, y) ambas en un rango que va de cero a cien (porcentaje), e indican el porcentaje del campo desde la perspectiva del equipo atacante. En particular, el valor de la coordenada x indica la proximidad a la portería contraria, mientras que el valor de la coordenada e indica la proximidad al lado derecho del campo.

En cuanto al conjunto de datos de jugadores, este describe los futbolistas desde un punto de vista tanto biográfico, como futbolístico y cuenta diversos atributos, esta información se muestra en la taba 2.

Tabla 2 Fuente de datos: Información de jugadores.

birthArea	Información geográfica acerca del lugar de nacimiento del jugador.
birthDate	Fecha de nacimiento del jugador en formato YYYY-MM-DD.
firstName	Nombre del jugador.

lastName	Apellido del jugador.
foot	El pie preferido o dominante del jugador.
height	Altura del jugador en centímetros.
role	Posición habitual del jugador.
wyId	Identificador del jugador asignado por Wyscout.

A los conjuntos de datos se les ha tenido que aplicar una serie de modificaciones para así poder adaptarlos a la tarea de la búsqueda de jugadores similares, este proceso se verá a continuación detalladamente, pero, en resumen, lo que se ha hecho es, concatenar tablas, pivotarlas y realizar modificaciones a las columnas desde valores absolutos a valores relativos o frecuencias.

4.1.3 Adaptación de los datos

El primer paso, una vez obtenidos los datos es realizar un análisis exploratorio sobre estos, este consiste en examinar y visualizar los datos de manera sistemática para obtener una comprensión inicial de su distribución y características. En este análisis previo, lo primero que se ha buscado obtener es la frecuencia de los eventos.

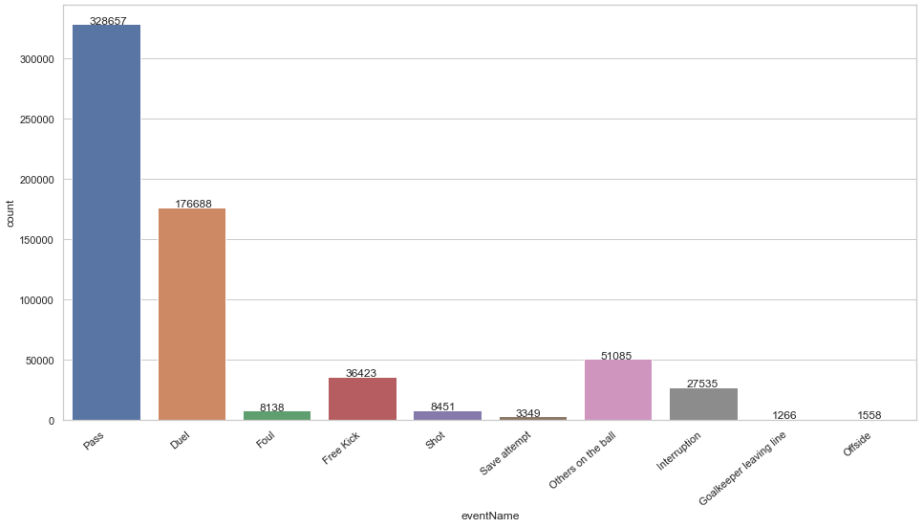


Figura 21 Gráfico de barras con el número total de apariciones de los eventos.

De la imagen previa (figura 21) se puede observar que hay un desbalanceo en la cantidad de apariciones de los distintos eventos, dada la naturaleza del deporte, no es de extrañar que la mayoría del tiempo, los partidos se desarrollan en pases o duelos entre los jugadores para el control de la pelota.

Del total de los eventos, se ha decidido eliminar las variables “Goalkeeper leaving de line” y “save attempt” para sólo considerar en nuestra investigación a los jugadores de campo puesto que los porteros difieren en gran medida en la evaluación de sus características técnicas del resto. Además, otros atributos como “Interruption” u “Offside” también han sido descartados ya que las interrupciones del partido y los fueras de juego no son relevantes con respecto a los patrones técnicos de un jugador.

Originalmente, cada una de las filas representaba un evento concreto, para cada evento, dado que se cuenta con un valor de “playerId” (esta forma original se muestra con algunos ejemplos de pases en la figura 22), se ha podido concatenar la tabla de eventos con la de jugadores, de tal manera que además del identificador también se cuente con su nombre y posición.

	eventId	subEventName	tags	playerId	eventName	subEventId
0	8	Simple pass	[{'id': 1801}]	25413	Pass	85
1	8	High pass	[{'id': 1801}]	370224	Pass	83
2	8	Head pass	[{'id': 1801}]	3319	Pass	82
3	8	Head pass	[{'id': 1801}]	120339	Pass	82
4	8	Simple pass	[{'id': 1801}]	167145	Pass	85

Figura 22 Estado original de los eventos.

La información del jugador ya ha sido asociada a cada uno de los eventos de tal manera que ahora podemos agrupar por jugador y obtener el número total de los distintos eventos asociados al futbolista.

Los eventos han sido posteriormente tratados individualmente de acuerdo con la información relevante de cada uno de ellos. Así, para los pases y los disparos se ha obtenido la frecuencia de estos de acuerdo con la zona (discretizada) en la que ocurren, mientras que, para las jugadas a balón parado y los duelos se ha hecho una disgregación por sub-eventos.

En el caso de los pases, la discretización de las localizaciones de estos se hizo dividiendo el terreno de juego en nueve zonas de igual área, se dividió el espacio vertical y horizontal en tres ejes dando lugar a la siguiente partición representada en la figura 23:

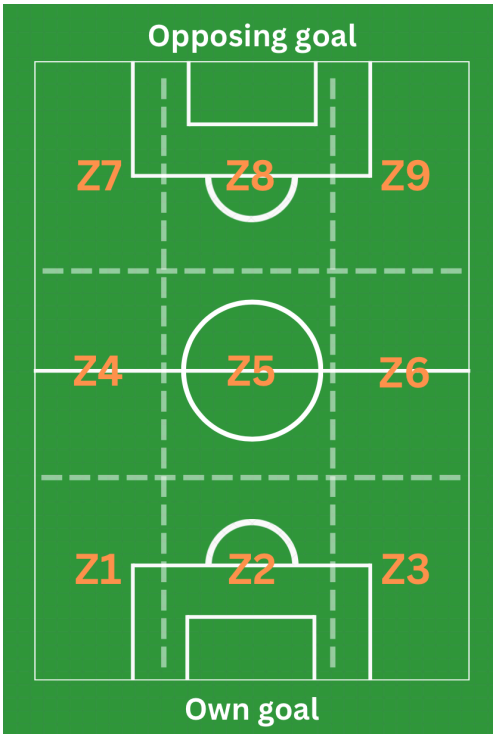


Figura 23 División del campo en nueve zonas (Zn) para los pases. Fuente: recurso propio.

En un primer momento se decidió contabilizar la frecuencia de pase por pares origen destino, aunque esto finalmente se descartó ya que se generaban ochenta y una variables con valores muy reducidos, con lo cual finalmente se optó por considerar nueve variables contabilizando la frecuencia de pases (respecto al total de pases del jugador) que tienen como origen cada una de las zonas y otras nueve con las frecuencias de destinos de los pases.

Además de la localización de los pases, también es interesante conocer la calidad de estos, para lo cual se ha añadido a nuestra base de datos final otro atributo llamado “PassAccuracy” con el porcentaje de pases totales de cada jugador que cuentan con la etiqueta “accurate”.

Para los disparos se ha seguido una metodología semejante, se ha dividido el último tercio del campo en seis zonas de igual área y se han clasificado el resto de los disparos como lejanos y se ha contado la frecuencia de disparo de cada una de las zonas. Finalmente, sólo nos hemos quedado con los disparos efectuados en el último tercio del campo ya que la cantidad de disparos efectuados desde más lejos es residual.

De forma semejante a los eventos de tipo pase, se ha añadido una variable llamada “shotAccuracy” resultante de dividir el número de goles que anota un jugador entre el número de disparos que efectúa.

Para el resto de los eventos, en lugar de contar con localizaciones se ha decidido darle más importancia a los sub-eventos, esto se debe a que, en el caso de las jugadas a balón parado, es más relevante que un jugador sea lanzador de faltas, de córneres o de penaltis que el hecho de que los lance desde una localización determinada.

Este razonamiento es extensible a los duelos en los que se involucra el jugador, donde nos puede interesar si el jugador es más propenso a duelos aéreos, duelos en ataque, en defensa… Este tipos de evento se almacenan en forma de número de eventos por partido (x90) y la forma de calcularlos consiste en dividir el total de un sub-evento entre el número de partidos jugados, este valor total de partidos no se incluye en la base de datos, pero se puede calcular sencillamente dividiendo el total de minutos jugados entre 90 (la duración de un partido).

Se ha observado que algunas de las variables de jugadas a balón parado siguen una distribución asimétrica, para estas variables se ha hecho una transformación con raíz cuadrada, para que la distribución de estas características se asemeje más a la distribución normal.

Dadas las diferentes escalas de las variables (frecuencia, valores x90, porcentajes), también se ha aplicado una función de escalado (StandardScaler de la librería Sklearn). Esta función, para cada muestra x calcula un nuevo valor z (como se ve en la siguiente ecuación) con:

$$z = (x - u)/s$$

Ecuación 7 Fórmula del escalado estándar

donde u es la media de la variable y s es su desviación estándar.

Para que las comparativas entre los jugadores sean justas, se ha decidido establecer un requisito de tiempo mínimo de juego para el cual seleccionar a los jugadores, ya que, pocos partidos no pueden ser representativos del estilo de juego de un futbolista y sesgaría mucho el estudio por estados de forma. En otros trabajos, este límite se establece en la cantidad de minutos equivalente a la mitad de los partidos de la temporada, pero en nuestro caso se ha establecido un límite inferior de novecientos minutos (diez partidos) para poder contar con más individuos.

Este conjunto ya es el definitivo a partir del cual se llevará a cabo la experimentación, no obstante, se comentaba al principio de la sección, que uno de los pasos que se mostraba (Clustering supervisado) implica otra transformación de los datos. Esto se verá más al detalle cuando se hable de la metodología final que se ha utilizado conocida como clustering supervisado.

4.2 Metodología

Entrando ahora a los métodos que se utilizan para la obtención de información de los datos, se utilizarán técnicas de reducción de dimensionalidad y algoritmos de agrupamiento.

4.2.1 Clustering

La búsqueda de similitudes en una base de datos es el propósito de ser de los algoritmos clustering. Estos se encargan de crear grupos a partir de una medida de distancia entre los puntos. Distintos algoritmos ofrecen distintos resultados en cuánto a las separaciones que realizan sobre un mismo conjunto de datos.

Como se mencionó previamente en el marco teórico, KMeans realiza un proceso iterativo de asignación de puntos a grupos según la distancia al centroide más cercano y actualización de centroides a partir de los nuevos puntos hasta que se llega a la convergencia. Por otro lado, el agrupamiento jerárquico, como su nombre indica construye una jerarquía en la que se fusionan o dividen grupos gradualmente.

La otra técnica que se ha incluido es la de DBSCAN, la cual toma una aproximación diferente a las dos anteriores en el sentido de que a diferencia de las anteriores donde es el usuario el que decide el número de grupos, DBSCAN encuentra automáticamente estos grupos identificando regiones densas en espacios de datos y posteriormente agrupando puntos en función de parámetros que sí que indica el usuario.

Existen más alternativas de algoritmos de clustering, sin embargo la variedad propuesta engloba la mayoría de familias de estos algoritmos, ya que KMeans pertenece a la familia de agrupación basada en centroides (rápida y eficiente), clustering jerárquico forma parte de los algortimos basados en jerarquías (destinada a datos de tipo jerárquico) y por último DBSCAN se basa en densidades, donde los puntos se consideran parte de un grupo basándose en la probabilidad de que pertenezcan a dicho grupo.

En el contexto del aprendizaje automático, a los parámetros que selecciona el usuario se les llama hiperparámetros, y en el caso de las técnicas de clustering, los hiperparámetros de cada una de estas técnicas, así como su descripción se muestran en la tabla 3.

Tabla 3 Hiperparámetros de los algoritmos de clustering.

Algoritmo	**Hiper parámetro**	**Tipo**	**Descripción**
KMeans	Clusters	Entero	El número de clústeres a formar, así como el número de centroides a generar.

Jerárquico	Clusters	Entero	El número de clusters a encontrar.
	Enlace	Texto	El criterio de vinculación determina qué distancia utilizar entre los conjuntos de observación.
DBSCAN	Épsilon	Decimal	Distancia máxima entre dos muestras para que una se considere vecina de la otra.
	Mínimo muestras	Entero	El número de muestras en un vecindario para que un punto se considere un punto central.

Habitualmente, cuando se trata con conjuntos de alta dimensionalidad como es el caso, se realiza un paso previo a la aplicación de las técnicas clustering, la transformación mediante técnicas de reducción de dimensionalidad. Evitar que se apliquen directamente las técnicas de agrupación se debe a que los algoritmos de agrupación al dependen de algún tipo de medida de distancia y es conveniente llevar a cabo una reducción sobre los espacios de alta dimensionalidad para que la métrica de distancia tenga sentido.

4.2.2 Reducción de dimensionalidad

Las técnicas de reducción de dimensionalidad son aquellas que consiguen la transformación de espacios de alta latencia a otros de menor latencia, estos pueden ser lineales como PCA donde se consigue esta reducción mediante la búsqueda de relaciones lineales u otra opción son las no lineales las cuales se basan en la construcción de distribuciones de probabilidad para capturar relaciones y similitudes entre los datos, como es el caso de UMAP.

La elección inicial de estas técnicas se debe a que, por un lado, PCA es la técnica más extensa por su uso en cuanto a técnicas de reducción de dimensionalidad, además, a pesar de que en las técnicas de reducción de dimensionalidad la interpretación de los nuevos valores no es sencillo, pero es cierto que hay más facilidad para explicar las técnicas lineales que las no lineales.

Sin embargo, a diferencia de PCA, la cual es una técnica antigua que apareció en el año 1901 por el estadístico Karl Pearson, la técnica UMAP fue desarrollada en el año 2020 y es actualmente estado del arte ofreciendo resultados excelentes con conjuntos de datos muy variados.

De nuevo, se cuentan con hiperparámetros en los casos de PCA y UMAP los cuales tienen una elevada influencia en los resultados que devuelven los modelos. Estos parámetros aparecen resumidos en la tabla 4.

Tabla 4 Hiper parámetros de las técnicas de reducción de dimensionalidad.

Modelo	Nombre	Tipo	Descripción
UMAP	Vecinos	Entero	Tamaño de la vecindad local (número de puntos) para la aproximación de la matriz. Los valores más grandes dan como resultado vistas más globales, los más pequeños hacen que se conserven más datos locales.
	Métrica	Texto	Métrica para calcular distancias en un espacio de alta dimensión.
PCA	Dimensiones	Entero	Número de componentes que se mantienen.

4.3 Arquitectura de experimentación

Recapitulando, la fase de modelado implica que en primer lugar se lleva a cabo una reducción de dimensionalidad y en segundo, en este nuevo espacio se buscan los grupos mediante algoritmos de clustering. Hay distintos hiperparámetros que pueden tomar diferentes valores numéricos o de texto, lo cual implica la creación de una estructura que permita explorar todas las combinaciones posibles.

Para la selección de hiperparámetros, una de las opciones más populares es la búsqueda en rejilla, pero esta plantea algunos problemas, en primer lugar, esta búsqueda no sigue una estrategia de búsqueda que llegue a un valor óptimo, sino que prueba todas las combinaciones que se indiquen. En su lugar, se ha decidido realizar una optimización con la librería de Python Optuna que trata de maximizar el valor de la función objetivo previamente mencionada.

Esta estrategia cuenta con algunas ventajas como pueda ser una búsqueda orientada donde se deciden los candidatos a próximos hiperparámetros a partir de los resultados de la última iteración.

Es sencillo comprender las diferencias entre la búsqueda en rejilla y la optimización en la representación gráfica de la figura 24 a la izquierda, en la búsqueda en rejilla, no hay ninguna ordenación de las pruebas, mientras que, en la derecha, optimización bayesiana, el pase de una prueba a la siguiente viene dado de un proceso de mejora iterativa del valor de la función objetivo.

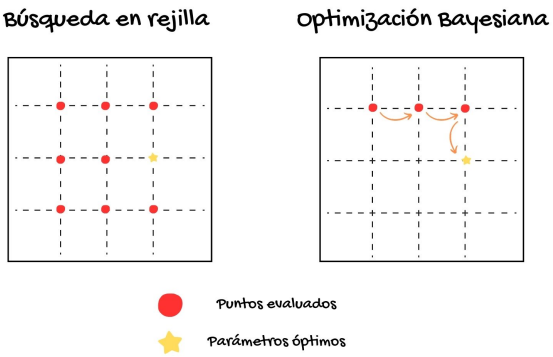


Figura 24 Diferencias entre la búsqueda en rejilla y la optimización bayesiana.

Se han creado dos experimentos, uno por cada técnica de reducción de dimensionalidad, y para cada uno de estos experimentos hay un proceso de dos partes, primero, a partir de los datos se aplica la reducción de dimensionalidad y en este espacio se buscan los grupos con tres algoritmos de agrupamiento: K-Means, clustering jerárquico y DBSCAN. Para cada experimento se ejecutan quinientas pruebas y finalmente se devuelve la prueba que mejor rendimiento ha dado de acuerdo con la función objetivo, así como los hiperparámetros de dicha prueba.

La función objetivo anteriormente mencionada es lo que en concreto deseamos optimizar. Para examinar la calidad de las metodologías, se ha definido dicha función objetivo a partir de otras dos de evaluación interna de clustering. La métrica definida ha sido:

$$(S + 1)/2 * 1/(1 + D)$$

Ecuación 8 Definición de la función objetivo

Donde S es el coeficiente de Silhouette y D es lel coeficiente de Davies-Bouldin, ambas utilizadas para evaluar la calidad del clustering.

El coeficiente de Silhouette se calcula utilizando la distancia media intra-clúster (a) y la distancia media al clúster más cercano (b) de cada muestra. El coeficiente de Silhouette de una muestra es:

$$(b - a)/max(a, b)$$

Ecuación 9 Coeficiente de Silhouette

Visto de otra forma, b es la distancia entre una muestra y el clúster más cercano del que no forma parte la muestra. Nos interesa maximizar este valor, y, por lo tanto, el coeficiente de Silhouette está situado en el numerador de la primera parte de la ecuación propuesta.

Por otra parte, el coeficiente de Davies-Bouldin se define como la medida de similitud media de cada clúster con su clúster más similar, donde la similitud es la relación entre las distancias dentro de un clúster y las distancias entre clústeres. Así, los clústeres más alejados y menos dispersos obtendrán una mejor puntuación. Este coeficiente puede tomar valores entre cero e infinito, con valores más cercanos a cero indicando mayor bondad del clustering, esta es la razón por la que este valor se sitúa en el denominador de la segunda parte de la ecuación.

4.3.1 Clustering supervisado

Volviendo sobre la primera figura que se mostró en el capítulo de propuesta, la modificación sobre la estructura de experimentos original se llevó a cabo debido a que como veremos en los resultados de experimentación, aplicar directamente la reducción de dimensionalidad sobre los datos originales no ha proporcionado separaciones por parte de los algoritmos de clustering muy informativas en los espacios de baja dimensión.

Es por eso por lo que se decidió optar por el enfoque del clustering supervisado. Esto implica que se debe añadir un paso previamente a la aplicación de las técnicas de reducción de dimensionalidad consistente en obtener unos nuevos valores, los valores shapley a partir de la aplicación de un modelo de aprendizaje supervisado. Estos nuevos valores tienen las mismas dimensiones que la base de datos original.

El cálculo de los valores de Shapley implica promediar las contribuciones marginales[5] de cada jugador (o característica) en todas las permutaciones potenciales de jugadores. Esto significa que se han de evaluar todas las combinaciones posibles de características y determinar el impacto que tiene cada característica en la predicción del modelo cuando se incluye en estas combinaciones.

Al promediar estas contribuciones en todas las posibles combinaciones de características, podemos lograr una evaluación equilibrada e interpretable de la importancia de cada característica en la predicción del modelo.

Dado que, para conseguir estos valores, se necesita entrenar un modelo supervisado, hay que definir una variable objetivo (“y”) adecuada, esta variable en el conjunto de datos sería el rol que toman los futbolistas de una determinada demarcación. Por ejemplo, en el caso de los delanteros, los roles son delantero centro, extremo derecho y extremo izquierdo.

La variable objetivo se ha transformado en una variable de tipo numérica y se ha llevado a cabo la tarea de clasificación utilizando un modelo Gradient Boosted Tree. Este modelo es de tipo

[5] Las contribuciones marginales en el contexto del aprendizaje automático se refieren a una medida que cuantifica el impacto o la influencia que cada característica o atributo tiene en la capacidad de un modelo de machine learning para hacer predicciones.

ensemble formado por un conjunto de árboles de decisión individuales. La razón de escoger este modelo es que estos en primer lugar son aplicables a tareas de clasificación, al ser no paramétricos no dependen del cumplimiento de distribuciones específicas y, por último, porque tienen una buena escalabilidad.

Con los nuevos valores computados, se decide utilizar TSNE en lugar de PCA como la otra técnica de reducción junto a UMAP, esto es debido a que tras la obtención de los primeros resultados PCA no arrojaba una transformación con la misma separabilidad que las técnicas de reducción de dimensionalidad no lineales.

Se puede ver este nuevo proceso en la siguiente figura 25 que además de la nueva estructura de experimentación contiene los hiperparámetros que se van a probar. La única adición respecto a los hiperparámetros mencionados anteriormente es el valor de "Perplexity" de TSNE, este valor está relacionado con el número de vecinos más cercanos que se utiliza en otros algoritmos de aprendizaje múltiple, diferentes valores de Perplexity pueden dar lugar a resultados significativamente diferentes.

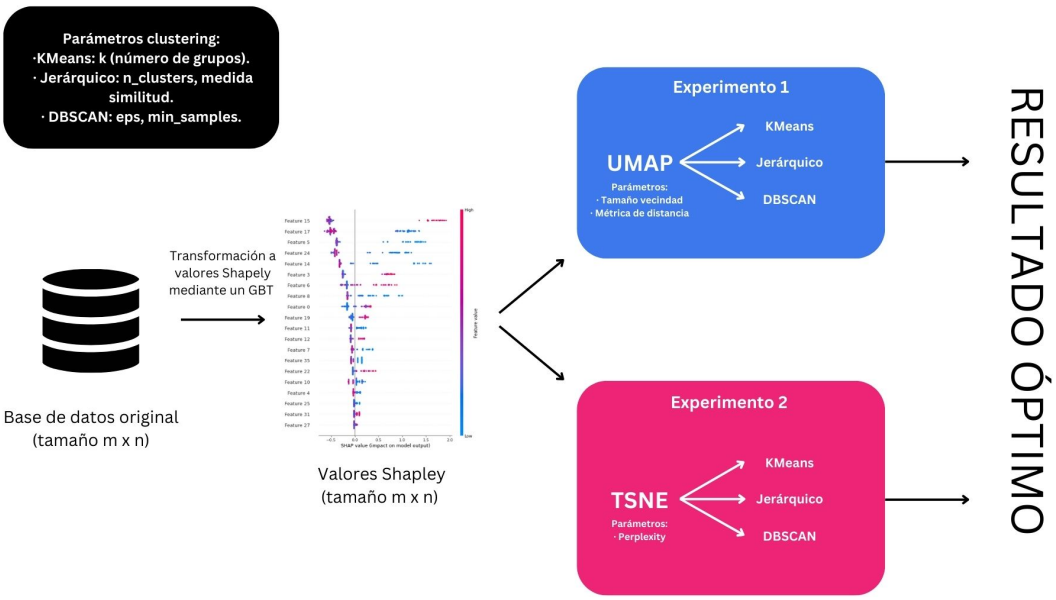


Figura 25 Estructura de la experimentación con Clustering supervisado.

Como paso adicional, se decidió además caracterizar los grupos resultantes de la tarea de clustering, esto se consiguió mediante el uso de reglas de decisión aplicadas sobre los datos originales. Estas reglas se calculan con una metodología de uno contra todos, convirtiendo este enfoque en una tarea de clasificación binaria supervisada.

Para esta tarea, la librería de Python Skope-rules ha sido muy útil porque proporciona una buena interpretación de un grupo la cual se basa en una expresión sencilla de la frontera que aísla el grupo.

4.4 Visualización

La salida de los experimentos es una proyección en un espacio bidimensional con distintos grupos caracterizados por reglas de decisión. Para añadir más detalle acerca de estas reglas, se debe añadir una información de referencia para poder saber si una condición de las que devuelve la regla de decisión está por encima del valor promedio del rol de los individuos que forman parte del grupo en concreto. Otra información con la que se debe contar es con las métricas de calidad de clustering, las cuales son las que se ha buscado optimizar en el proceso de experimentación.

Para poder visualizar en un único lugar toda la información relevante se ha diseñado un panel con todos los gráficos y texto que ayude a clarificar dichos gráficos, esto ha sido posible gracias a la librería Dash de Python.

El panel incluye un gráfico interactivo con los distintos grupos coloreados, en el cual el usuario puede mover el cursor sobre los puntos y aparecerá un cuadro con información relevante sobre el jugador como el nombre y el valor de mercado y otro gráfico similar en menor tamaño en el que se puede ver la proyección coloreada por roles.

En la parte superior derecha hay un cuadro de texto deslizable en el cual se muestran las reglas de asociación y en la parte inferior una tabla con los valores medios según el rol lo cual ayuda a comprender los valores de las reglas.

Poor último, también se incluyen los valores de las métricas internas de evaluación de clustering (coeficientes Silhouette y Davies Bouldin).

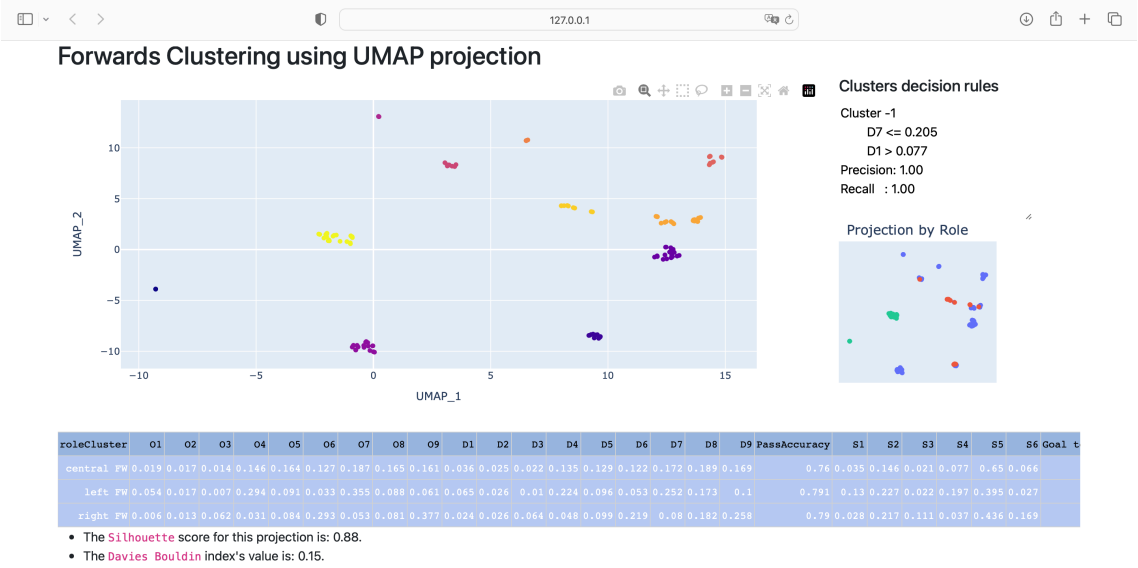


Figura 26 Dahboard de los delanteros.

La vista mostrada anteriormente (figura 26) se corresponde a los delanteros y la proyección con UMAP, la cual es la que proporciona mejores resultados en esta demarcación.

4.5 Limitaciones

A pesar de que la fuente de datos utilizada cuenta con una elevada cantidad y variedad de eventos y jugadores y de ser uno de los pocos conjuntos con información completa a nivel de temporada; los datos dejan fuera eventos relevantes como puedan ser los relativos a regates, pases clave o contribuciones de gol.

Además, la base de datos que se empleó es en cierta medida mejorable en cuanto a la caracterización de posiciones. Comúnmente se categoriza la distribución táctica de los jugadores en el campo en siete posiciones distintas: portero, defensa central, defensa lateral, mediocentro defensivo, mediocentro, extremo y delantero. Sin embargo, la base de datos utilizada solo clasifica a los jugadores en tres posiciones (defensa, mediocentro y delantero). Esta circunstancia ha generado una limitación significativa en el análisis de los jugadores, ya que en la fase de experimentación la división se ha hecho teniendo en cuenta esta división de posiciones.

Como remedio a futuro de esta limitación, la metodología es fácilmente adaptable para la replicabilidad con distintos conjuntos de datos, con lo cual de disponer con conjuntos más extensos y con información que complete el apartado de regates se puede aplicar el proceso seguido con el conjunto de datos y comparar los resultados, con especial énfasis en el estudio de que relevancia tienen los nuevos atributos.

4.6 Identificación y análisis de soluciones posibles

Dado que la búsqueda de jugadores similares es una tarea popular en la analítica deportiva, hay distintas soluciones que se pueden considerar. Para finalizar esta sección es relevante llevar a cabo un estudio comparativo de otras soluciones que se podrían haber llevado a cabo y porqué se eligió esta.

Con el uso de técnicas estadísticas se puede comparar el rendimiento de los jugadores en diferentes métricas y se pueden utilizar medidas de similitud como la distancia euclidiana o el coeficiente de correlación para encontrar jugadores con perfiles similares.

Otra alternativa es la de utilizar el análisis clustering, para agrupar a los jugadores en función de características similares. Esto permite identificar grupos de jugadores con perfiles similares y encontrar nuevos talentos con características deseadas.

Si se aborda esta tarea desde el punto de vista del aprendizaje supervisado se pueden entrenar modelos que predigan características o rendimiento de los jugadores en función de variables de entrada. Por ejemplo, se pueden utilizar algoritmos como árboles de decisión, bosques aleatorios o redes neuronales para predecir el rendimiento de un jugador en base a sus estadísticas pasadas.

Las técnicas de análisis de video también son útiles para identificar patrones de juego y características específicas en los jugadores. Se pueden utilizar algoritmos de visión por computadora y reconocimiento de patrones para extraer información relevante de los videos de los partidos y comparar jugadores en base a estas características.

Además de los datos estadísticos, también se pueden considerar otros tipos de datos no estructurados, como comentarios de expertos, opiniones de entrenadores u ojeadores. Se pueden utilizar técnicas de procesamiento de lenguaje natural y minería de texto para extraer información útil de estos datos y encontrar jugadores con perfiles similares.

Cada una de las posibles soluciones tiene sus puntos a favor y en contra, la tabla 5 muestra los pros y los contras de cada aproximación.

Tabla 5 Posibles soluciones al problema planteado

Solución	**Ventajas**	**Inconvenientes**
Análisis estadístico	Utiliza datos numéricos fácilmente disponibles.	No tiene en cuenta el contexto y las circunstancias del juego.

	Permite comparar jugadores en función de métricas objetivas. Puede proporcionar una visión general del rendimiento de los jugadores.	No considera aspectos cualitativos o intangibles del rendimiento. Puede ser limitado en su capacidad para identificar jugadores con habilidades específicas no capturadas por las estadísticas.
Clustering	Permite agrupar jugadores con perfiles similares de manera automática. Ayuda a identificar patrones y segmentos de jugadores en función de características específicas. Puede descubrir talentos ocultos o jugadores infravalorados.	La elección del número óptimo de clústeres puede ser subjetiva y afectar los resultados. Puede no ser efectivo si los datos disponibles no capturan suficiente variabilidad o no son representativos. No proporciona una clasificación individualizada de los jugadores, sino más bien grupos de jugadores similares.
Aprendizaje automático supervisado	Puede proporcionar predicciones precisas y personalizadas para jugadores individuales. Permite considerar una amplia gama de variables de entrada, incluidas las estadísticas y otros datos relevantes. Capaz de capturar relaciones complejas y no lineales entre variables.	Requiere un conjunto de datos grande y de alta calidad para entrenar modelos precisos. Puede ser susceptible al sesgo inherente en los datos de entrenamiento. La interpretación de los modelos de aprendizaje automático puede ser difícil debido a su naturaleza "caja negra".
Análisis de video	Permite capturar aspectos técnicos y tácticos del juego que no están presentes en los datos estadísticos. Proporciona información visual y contextual valiosa sobre los jugadores.	Requiere un procesamiento y análisis intensivo de grandes cantidades de videos. La extracción de características y el análisis de video pueden ser subjetivos y depender de la interpretación humana. La disponibilidad y calidad de los videos pueden variar, lo que puede afectar la

	Puede revelar detalles sutiles y habilidades específicas de los jugadores.	consistencia y la precisión de los resultados.
Análisis de datos no estructurados	Permite aprovechar información adicional y no cuantitativa sobre los jugadores. Puede capturar la intuición y el conocimiento experto que no se reflejan en los datos numéricos. Ayuda a obtener una imagen más completa y multidimensional de los jugadores.	Requiere técnicas de procesamiento de lenguaje natural y minería de texto, que pueden ser complejas y requerir recursos computacionales adicionales. La calidad y la subjetividad de los datos no estructurados pueden influir en los resultados. La interpretación y el análisis de datos no estructurados pueden ser más difíciles y menos objetivos que los datos estructurados.

Las opciones presentadas anteriormente, se abren como posibilidades de líneas abiertas de investigación que merecen una toma en consideración. Sin embargo, es importante destacar que, hasta el momento, existe una limitada cantidad de investigaciones y trabajos previos relacionados con la aplicación de técnicas de clustering para agrupar futbolistas basándose en sus similitudes. Esta brecha en la literatura científica y deportiva señala una interesante oportunidad de investigación que aún no ha sido explotada de manera exhaustiva.

Para elegir la solución que se desarrollará, se deben tener en cuenta los datos disponibles, con lo cual las dos soluciones posibles pasan por realzar un análisis estadístico o clustering. Se opta por esta segunda opción dado que es menos habitual en la literatura y para paliar algunos de los inconvenientes como el sesgo del número de clústeres se consideran varias técnicas de clustering se hace una optimización del hiperparámetro número de clústeres.

5 Experimentos y resultados

Con la propuesta planteada queda ahora comentar la fase de experimentación, evaluar la calidad de los resultados y explicar estos resultados al detalle. Los experimentos de los dos siguientes apartados se muestran con la primera de las distribuciones con las que se hicieron experimentos, con los jugadores de la posición delantero.

Posteriormente se habla de los experimentos en las otras posiciones de la base de datos, centrocampista y central.

5.1 Reducción de dimensionalidad y clustering (no supervisado)

Ya se ha mencionado previamente que, inicialmente, se optó por el enfoque clásico de clustering no supervisado, en el cual, aplicando las técnicas de reducción de dimensionalidad sobre los datos originales se obtienen en este nuevo espacio los grupos aplicando algún algoritmo de clustering.

Los primeros resultados de experimentación no arrojaron separaciones de datos muy informativos. Las técnicas de reducción de dimensionalidad seleccionadas en un primer momento fueron PCA y UMAP.

Para PCA, el resultado óptimo se obtuvo con cuatro dimensiones en conjunto con el algoritmo de clustering KMeans con el valor tres como número de grupos. Además de que el valor de la métrica de Silhouette tenía un valor poco de prometedor de 0.5, los grupos devueltos eran, por un lado, el formado por todos los extremos izquierdos junto con algunos delanteros centro, otro formado por únicamente delanteros centro y el otro por extremos derechos y de nuevo algunos delanteros centro.

En la figura 27 se muestra a la izquierda, la división por roles y a la derecha por grupos, no habiendo entre ambas una diferencia significativa, más allá del hecho de que algunos delanteros centro se han clasificado en grupos de los extremos.

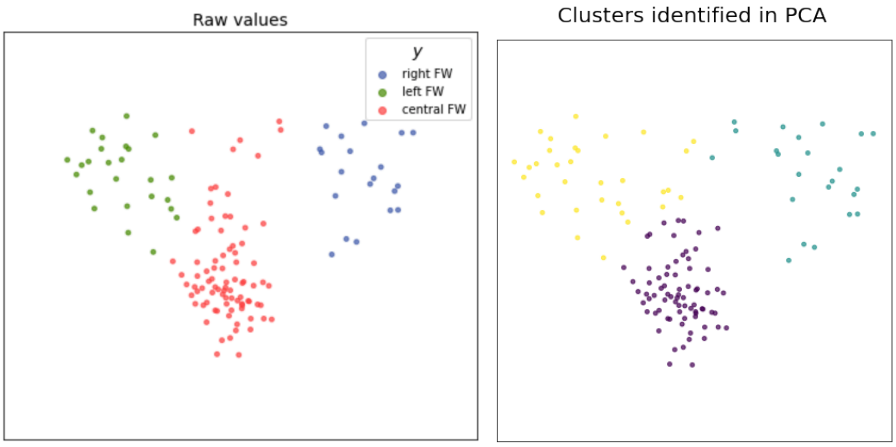


Figura 27 Proyecciones PCA coloreadas por rol y grupos para clustering no supervisado.

Con UMAP los resultados eran incluso menos prometedores. En primer lugar, el valor del coeficiente de Silhouette era menor, por otra parte, visualmente los resultados carecían de sentido, con un agrupamiento de tipo jerárquico que dividía el espacio en dos grupos, uno que sólo cuenta con algunos delanteros centro y otro que engloba al resto, tal y como se puede ver en la figura 28.

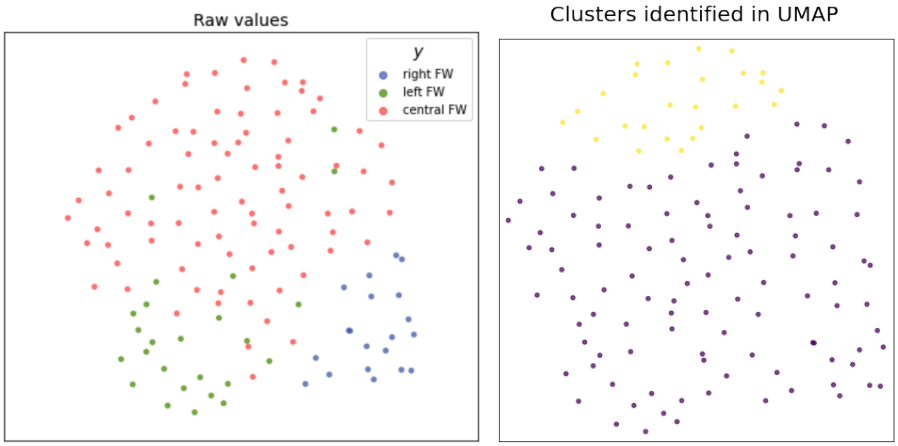


Figura 28 Proyecciones UMAP coloreadas por rol y grupos para clustering no supervisado.

Estos pobres resultados eran similares con todas las distribuciones de datos (delanteros, mediocentros y centrales).

De esta primera experimentación no todo fueron aspectos negativos, ya que los pobres resultados llevaron a la investigación del método de clustering supervisado, diseñado para casos como este en el que no se puede extraer conocimiento útil de las proyecciones.

Este cambio implicó cambiar la técnica de PCA por TSNE puesto que las técnicas de reducción de dimensionalidad lineales no se adaptan muy bien con el uso de los valores Shapley y sin embargo las técnicas no lineales sí que ofrecen mejores resultados.

5.2 Clustering supervisado

Pasando ahora a las técnicas de clustering supervisado, se utiliza en primer lugar un modelo Gradient Boosted Trees para la obtención de los valores shapley. Con estos valores se ha llevado a cabo el mismo procedimiento que en la fase anterior, sólo que aplicando la reducción de dimensionalidad a estos nuevos valores y con el añadido de la caracterización de los grupos mediante reglas de decisión como se verá posteriormente.

Mediante las siguientes gráficas, se pueden ver cuáles han sido las variables más relevantes para cada uno de los tres roles (delantero centro, extremo izquierdo y extremo derecho).

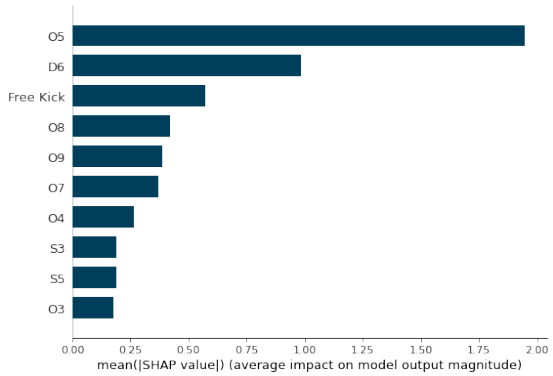


Figura 29 Valores medios SHAP para delanteros centro.

Empezando por el gráfico de delanteros centro (figura 29), podemos ver que las variables que más importancia tienen en la decisión que lleva al modelo a clasificar a un individuo como delantero centro son O5, D6, Free Kick, O8, O9 y O7, siendo las variables O_i las que representan la frecuencia de pases desde el origen i y Free Kick las faltas lanzadas por el jugador.

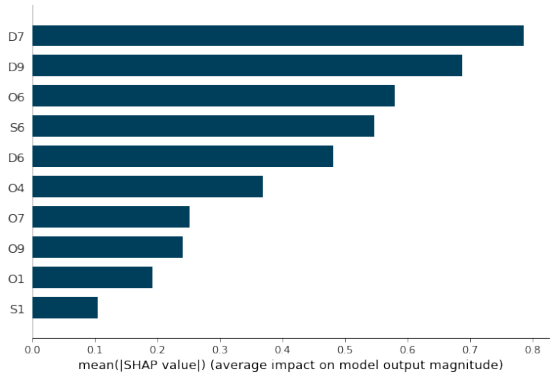


Figura 30 Valores medios SHAP para extremos izquierdos.

En la clase de extremos izquierdos, se puede ver en la figura 30 que los atributos que se consideran más relevantes para que el modelo prediga que un jugador es delantero centro son D7, D9, O6, S6, D6 y O4. No obstante, la magnitud máxima del eje X es menor que en el caso de los delanteros

centro, donde la variable O5 tenía una magnitud del doble respecto a la siguiente variable más relevante.

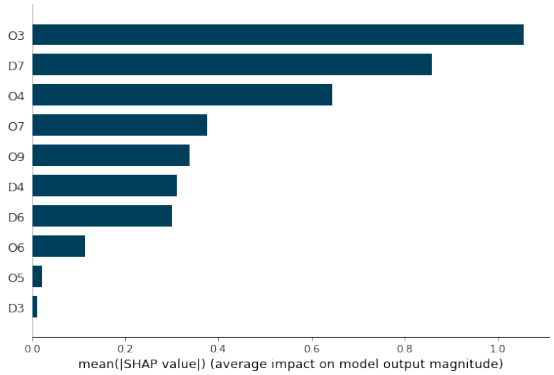


Figura 31 Valores medios SHAP para extremos derechos.

Finalizando con los extremos derechos en la figura 31, las variables con más impacto en que se prediga esta clase son O3, D7, O4 y O7.

Esto en definitiva muestra tres conjuntos de valores SHAP que representan a cada uno de los roles, obteniendo el promedio de estos, se puede obtener una imagen de la demarcación que enmarca a todos los delanteros, y sobre estos valores, se puede realizar una proyección de todos los delanteros.

Se empieza de nuevo por seleccionar los hiperparámetros para explorar el conjunto de estos que resulta óptimo a nivel de calidad de clustering.

5.2.1 UMAP

Para comprender como se ha llevado a cabo el proceso de selección de hiperparámetros son de utilidad algunas visualizaciones, algunas de estas no son comprensibles con un elevado número de pruebas, por eso, estas visualizaciones se corresponden con un estudio realizado con cincuenta pruebas con jugadores de ataque y con la técnica UMAP.

El primer gráfico (figura 32) muestra el histórico de las pruebas del experimento, en este ejemplo, el valor óptimo se obtiene en el noveno intento, esto se puede ver en la línea roja, mientras que los intentos completados (aquellos que no han sido podados) se muestran en los puntos en azul.

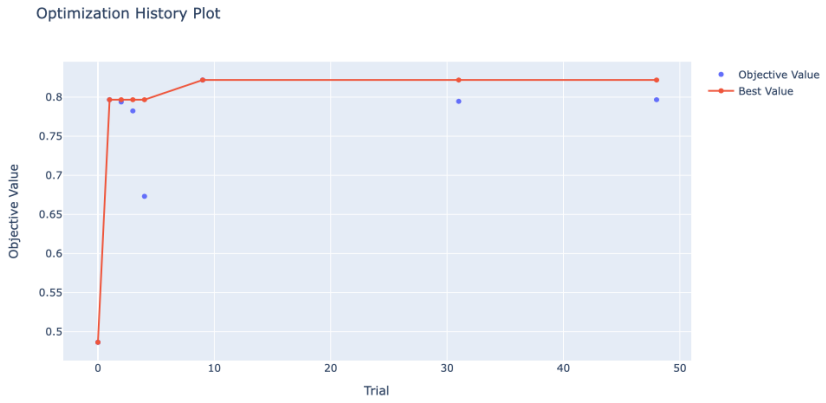


Figura 32 Gráfico histórico de optimización UMAP con delanteros.

Otra cuestión interesante es conocer cuáles son los parámetros más relevantes con respecto a la función objetivo, siguiendo en el mismo ejemplo, los hiperparámetros relativos a la reducción de dimensionalidad tienen similar importancia al algoritmo de clustering y de entre los hiperparámetros de reducción de dimensionalidad, el número de vecinos tiene el triple de impacto que la métrica a considerar (figura 33).

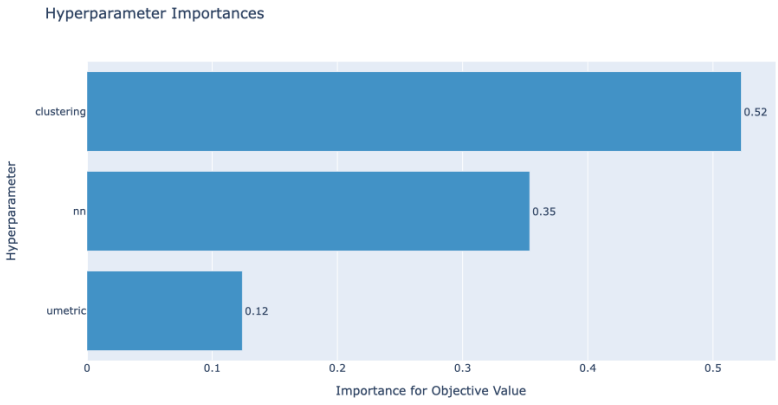


Figura 33 Importancia de hiperparámetros para la función objetivo UMAP de delanteros.

Por último, el gráfico temporal (figura 34) muestra las pruebas que han sido completadas con éxito en azul y las que se han podado en naranja, así como el tiempo que ha tomado cada prueba representada por la longitud de la barra. Este gráfico sirve para resaltar que, añadiendo una poda se disminuye el tiempo de ejecución sin perder la calidad de los resultados.

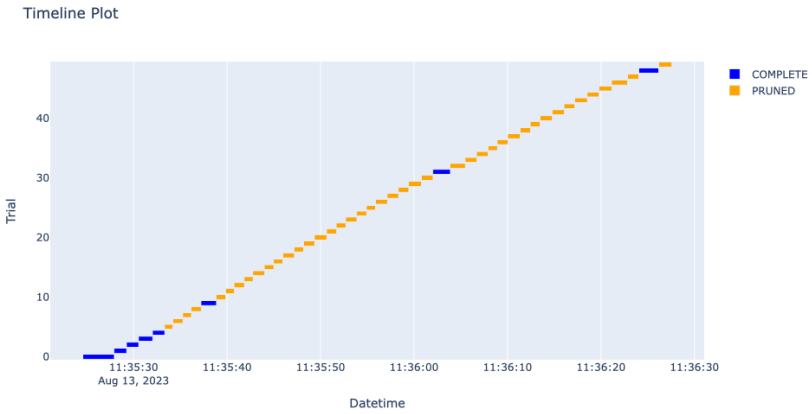


Figura 34 Pruebas completadas y podadas en optimización.

Para los experimentos con un elevado número de pruebas, se devuelve el conjunto de hiperparámetros que maximiza la función objetivo y el histórico de pruebas, así como la relevancia de dichos parámetros. Siguiendo con los jugadores de ataque, estos son los resultados.

Empezando por la combinación de UMAP y clustering, el modelo óptimo con un valor de la función objetivo de 0.8 con un valor de tres como tamaño de vecindad local y la métrica Manhattan para la técnica de reducción de dimensionalidad y con la técnica de clustering DBSCAN con valor de épsilon de 1.1 y un valor de mínimo de muestras de cuatro. Este valor óptimo se alcanza muy pronto, en la novena prueba.

La proyección de los puntos en el nuevo espacio muestra algunas cuestiones relevantes. Los extremos izquierdos se engloban en su mayoría en un grupo, mientras los tres extremos restantes (Çalhanoğlu, Sisto y Perišić) forman un grupo por ellos mismos, en concreto el grupo en azul más a la izquierda en el eje X en la gráfica en la parte derecha de la figura 35. Los extremos derechos y delanteros centro aparecen mezclados en varios grupos.

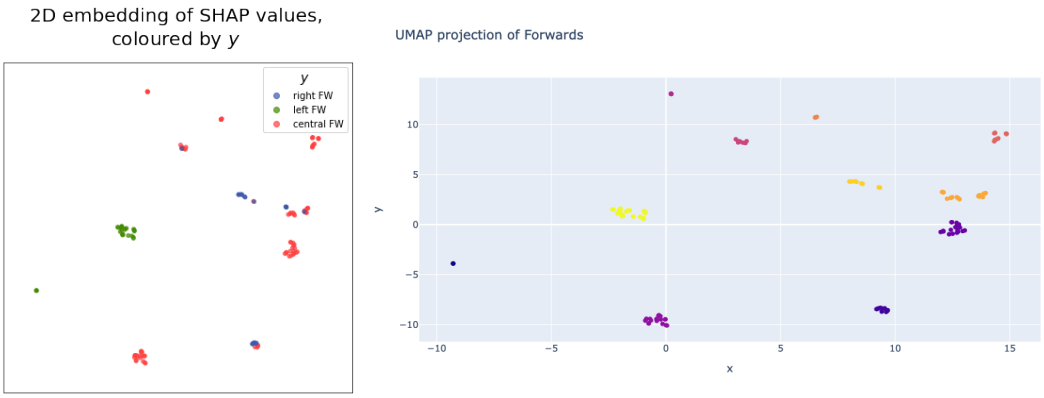


Figura 35 Proyecciones UMAP por roles y grupos de delanteros.

5.2.2 TSNE

De nuevo se han seguido los mismos pasos que en el apartado anterior con el único cambio de que en lugar de los hiperparámetros de UMAP, el único parámetro en el caso de t-SNE es el parámetro de perplejidad ("perplexity").

El resultado óptimo se obtiene con un valor de perplejidad de cinco y un clustering jerárquico con doce grupos y el criterio de vinculación de la media[6]. Este obtiene un valor de la función objetivo de 0.65, significativamente menor que con el experimento anterior. En la función objetivo el impacto que ha tenido la elección de la técnica de clustering ha sido determinante para encontrar el valor óptimo (figura 36).

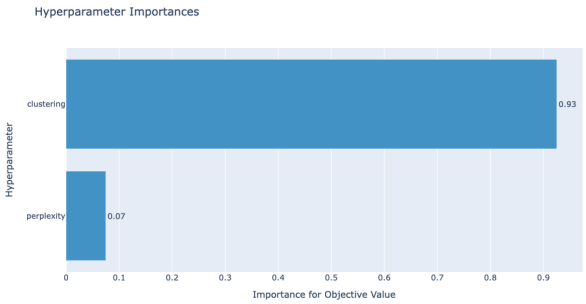


Figura 36 Importancia de hiperparámetros para la función objetivo TSNE de delanteros.

En la proyección en el nuevo espacio (figura 37) se forman los mismos grupos con extremos izquierdos, aunque esta vez, la separación con el grupo de delanteros más cercanos es menor. Además; Christian Eriksen, Kevin De Bruyne y Mesut Özil (los tres jugadores que suelen jugar un poco más atrasados que el delantero centro típico, en la posición de punta), caen en grupos distintos cuando por similitudes en el estilo de juego lo habitual sería que se encontraran en el mismo grupo.

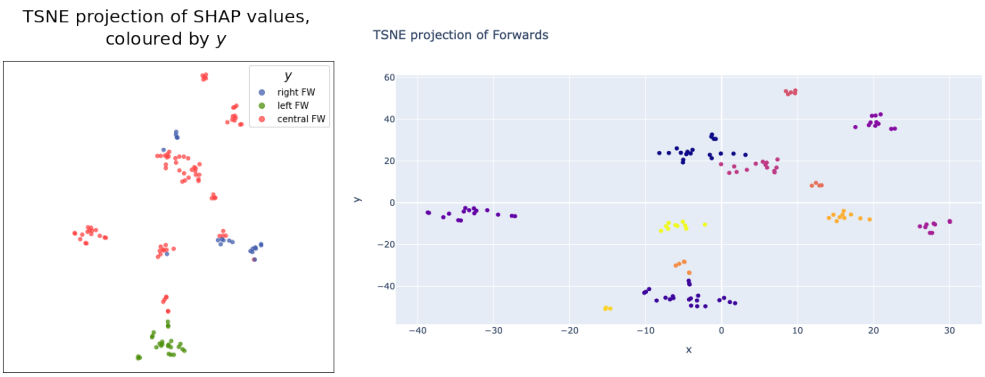


Figura 37 Proyecciones TSNE por roles y grupos de delanteros.

[6] La vinculación media minimiza la media de las distancias entre todas las observaciones de pares de grupos.

Las reglas de decisión en este caso muestran que la característica que comparten los grupos uno y ocho en morado oscuro y naranja respectivamente (en la parte inferior del eje vertical y cerca del cero en el eje horizontal) es la alta frecuencia de pases que tienen como destino la zona del extremo izquierdo, pero cuentan con la diferencia que los extremos izquierdos del grupo uno también se mueven más en la zona izquierda de la defensa realizando pases desde dicha parte del campo mientras que en el grupo ocho con delanteros centro, estos no ejecutan pases desde allí. Otra aportación interesante es que el grupo dos (en morado a la izquierda en el eje horizontal) se diferencia del resto de grupos de delanteros centro al no dirigir tantos pases al extremo derecho.

5.2.3 Análisis de los grupos

Para un análisis más detallado de los grupos, se ha optado por explorar aquellos grupos resultantes de la técnica de reducción de dimensionalidad que mejores resultados ofrece en cuanto a reducción de dimensionalidad, siendo esta UMAP.

La cantidad de muestras de cada grupo, así como las coordinadas de sus centroides se muestran a continuación en la tabla 6. Se puede ver que el primer, cuarto y séptimo cuentan con pocos individuos mientras que los grupos tercero cuarto octavo y décimo cuentan con alrededor de la veintena de individuos.

Tabla 6 Información de grupos de delanteros

Número	Cantidad de miembros	Coordinadas del centroide
-1	3	[-9.28198 -3.885299]
0	10	[9.441344 -8.487299]
1	18	[12.563259 -0.42232004]
2	18	[-0.41835102 -9.551821]
3	4	[0.22512075 13.072672]
4	9	[3.2530608 8.298501]
5	11	[14.501172 8.8237]
6	5	[6.5294504 10.7221]
7	20	[13.009008 2.8978384]
8	9	[8.494054 4.126626]
9	22	[-1.6188222 1.192023]

5.2.3.1 Reglas de decisón

Podemos caracterizar cada uno de los grupos con reglas de decisión. Esto se puede hacer con una metodología de uno contra todos, con los datos originales, se aprenden reglas que diferencian a un grupo de el resto, pudiendo así caracterizar los grupos con independencia de los valores SHAP.

Las reglas (tabla 7) muestran que el grupo que forman los tres extremos derechos tiene una menor frecuencia de pase a la de ataque izquierda con respecto al resto de extremos derechos (D7 <= 0.205) y, además, estos tres jugadores son los que envían sus pases a la zona de defensa izquierda (D1 > 0.07) con una tendencia mayor a la habitual para este rol.

Otras reglas relevantes son las correspondientes a los grupos 1, 2, 3, 5, 8 y 9.

- El grupo uno contiene delanteros centro, con las peculiaridades de que la frecuencia con la que pasan hacia la zona derecha del capo es superior a la media de su posición y además participan en menos duelos de defensa de balones rasos que la media.
- El grupo dos también es de delanteros centro, pero estos tienen otras particularidades como una frecuencia menor de pases hacia las zonas derechas del centro del campo y del último tercio de ataque.
- De nuevo, el grupo tres contiene delanteros centro, con la diferencia de que estos cuentan con una frecuencia de pases hacia la zona del extremo izquierdo mayor de lo que se cabría esperar de la posición.
- El grupo cinco cuenta con la particularidad de que sus jugadores efectúan más pases de lo habitual hacia la zona del lateral izquierdo.
- El grupo ocho lo componen extremos derechos y los antecedentes de las reglas de decisión muestran que su frecuencia de pases desde el centro del ataque es menor que lo que cabría esperar en dicha posición.
- Por último, el grupo nueve cuya regla señala que la frecuencia con la que sus jugadores realizan los pases hacia el extremo izquierda es mayor que la del otro grupo que forman los tres exrtemos izquierdos restantes.

Hay que tener en cuenta, no obstante, que para que estas comparaciones entre grupos sean correctas se debe tomar en cuenta los valores de precisión y exhaustividad de las reglas, tomando por ejemplo el segundo clúster, a pesar de que esta regla tiene en cuenta al noventa por cien de los individuos, sólo la mitad de ellos la cumplen, con lo cual, esta regla no será representativa del grupo al completo. Se pueden ver las reglas de decisión para los delanteros en la tabla 7.

Tabla 7 Reglas de decisión de los grupos de delanteros

Cluster -1	Cluster 5
D7 <= 0.205	O1 > 0.0228

```
        D1 > 0.0769
Precision: 1.00
Recall   : 1.00


Cluster 0
        O3 > 0.026
        Penalty <= 0.02
Precision: 0.53
Recall   : 0.90


Cluster 1
        D9 > 0.155
        Ground defending duel <= 3.2
8
Precision: 0.71
Recall   : 0.83


Cluster 2
        D6 > 0.0729
        D9 <= 0.141
Precision: 0.90
Recall   : 1.00


Cluster 3
        D7 > 0.226
        D9 > 0.146
Precision: 1.00
Recall   : 1.00


Cluster 4
        S5 > 0.874
Precision: 1.00
Recall   : 0.33
```

```
        O4 <= 0.17
Precision: 0.77
Recall   : 0.91


Cluster 6
        O2 > 0.002
        Ground defending duel > 9.43
Precision: 0.50
Recall   : 0.40


Cluster 7
        S5 > 0.666
        S6 > 0.076
Precision: 0.90
Recall   : 0.45


Cluster 8
        D6 > 0.204
        O8 <= 0.076
Precision: 1.00
Recall   : 0.89


Cluster 9
        O1 > 0.0178
        D7 > 0.2198
Precision: 1.00
Recall   : 1.00
```

5.2.3.2 Análisis Cualitativo

De los grupos que presentan reglas de decisión con unos valores de precisión y recall elevados, podemos ver cuáles son sus integrantes y clasificar cada uno de ellos más allá de los datos numéricos, evaluando sus estilos de juego, esto es lo que se muestra en la tabla 8.

Tabla 8 Análisis cualitativo de delanteros.

Grupo	Miembros	Descripción
-1	Perisic, Çalanoglu y Sisto	Extremos veloces y con buen regate.
1	Aduriz, Ben Yedder, Zapata, Morata, Aubameyang, Ronaldo, Messi, Lewandowski, Agüero…	Delanteros de élite goleadores con muy buenas capacidades de regate.

2	Suárez, Immobile, Belotti, Stuani, Giroud, Cavani…	Delanteros de élite poco móviles y con una excelente capacidad goleadora.
3	Mertens, Vardy, Dzeko, Munir…	Delanteros centro goleadores y veloces.
5	Griezmann, Maxi Gómez, G.Simeone.	Delanteros con tendencia a jugar por el centro.
8	Candreva, Lucas Vázquez, Suso, El Zhar.	Extremos derechos veloces.
9	Mané, Neymar, A.Sánchez, Guedes, Coutinho, Richarlison…	Extremos izquierdos de élite.

5.3 Experimentación con centrocampistas

En la base de datos hay más de trescientos centrocampistas divididos entre mediocentros y centrocampistas izquierdos y derechos. Esta división queda clara al proyectar los datos con la técnica PCA (figura 38), donde los colores representan los tres roles.

La modificación que se ha llevado a cabo en el preprocesado de estos datos ha sido la eliminación del valor duplicado relativo a Sergi Darder, el cual al igual que Gonçalo Guedes (delantero) fue traspasado de un club a otro al principio de la temporada.

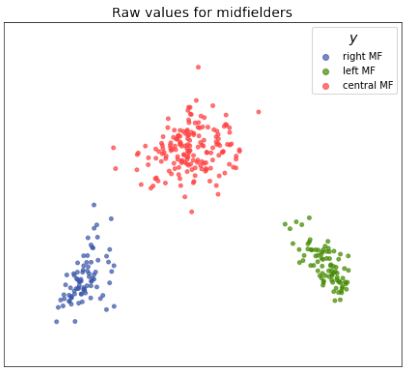


Figura 38 Proyección PCA de mediocentros.

Tras entrenar el modelo supervisado, el gráfico de impacto de los valores SHAP para cada uno de los roles se pudo ver las variables que más relevancia toman para predecir cada uno de los roles.

En primer lugar, en la clase de mediocentros mostrada en la figura 39, sólo hay una variable que tiene una enorme importancia de predecir a un centrocampista como mediocentro y esto es la frecuencia con la que se realizan pases desde el centro del campo.

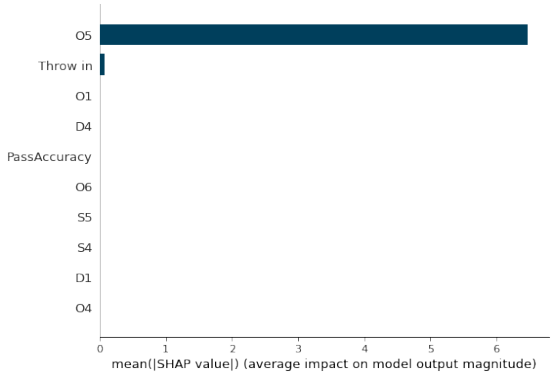


Figura 39 Valores medio SHAP de mediocentros.

De forma similar, con lo que en la base de datos original se considera como centrocampistas derechos, hay de nuevo una única variable que toma una enorme importancia a la hora de realizar las predicciones del modelo. Esta es la variable O3, o, dicho de otra forma, la variable que mide la frecuencia de pases desde la zona del lateral derecho. Esto se puede ver en la figura 40.

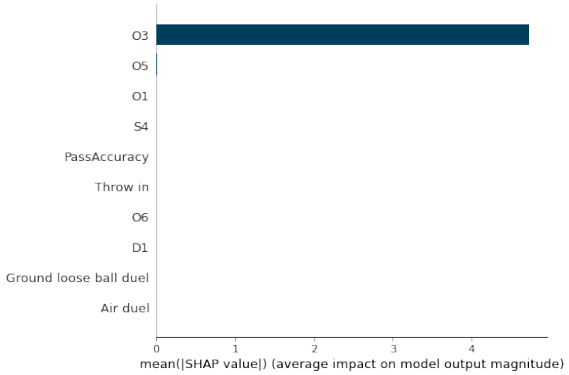


Figura 40 Valores medio SHAP de centrocampistas derechos.

Por último, para los centrocampistas izquierdos (figura 41) sí que hay un número más elevado de características relevantes, siendo estas la variable O1 (frecuencia de pases desde la zona del lateral izquierdo) y las variables D3, D4 Y D6. Respecto a D6 y D3, esta importancia a que los centrocampistas tienen frecuencias bajas de pases con estos destinos mientras que respecto a D4 la importancia reside en que los valores sean elevados.

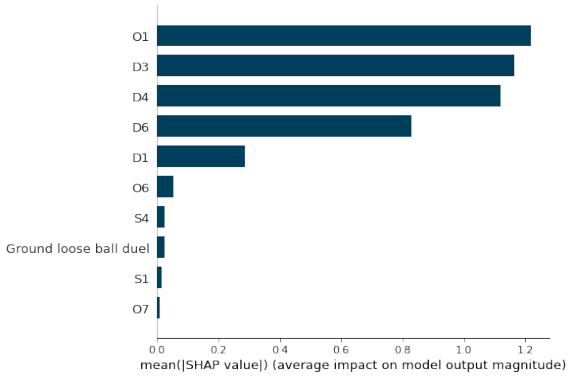


Figura 41 Valores medio SHAP de centrocampistas izquierdos.

La fase de optimización ha devuelto los mejores resultados con UMAP con tres como tamaño de vecindad y la métrica euclídea y con t-SNE con perplejidad de catorce. Con las técnicas de agrupación DBSCAN con un valor épsilon de 1.0 y tres como mínimo de muestras en el caso de UMAP y agrupamiento jerárquico con treinta y un grupos y vinculación sencilla para TSNE.

Como aspectos relevantes de los resultados de optimización, los gráficos de importancia para la función objetivo (figura 42) muestran que con UMAP los parámetros de esta técnica tienen un peso muy elevado (más del ochenta por cien) para la función objetivo con la técnica de clustering teniendo un peso en la función objetivo de alrededor de un veinte por cien.

En el caso de TSNE al igual que con los delanteros, el hiper parámetro de TSNE es irrelevante y la los hiper parámetros de las técnicas de agrupación son las que tienen toda la importancia a la hora de calcular la función objetivo.

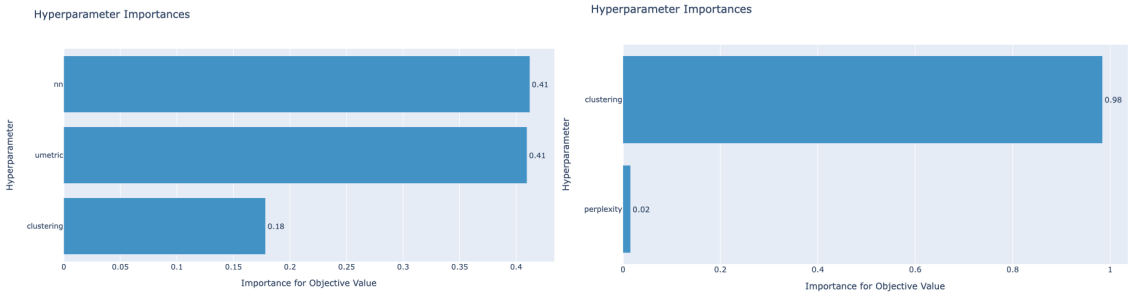


Figura 42 Gráficos de importancia de hiper parámetros para la función objetivo de centrocampistas.

El resultado con UMAP mostrado en la figura 43 en términos de calidad del clustering es significativamente mejor que el resultado con t-SNE mostrado en la figura 44.

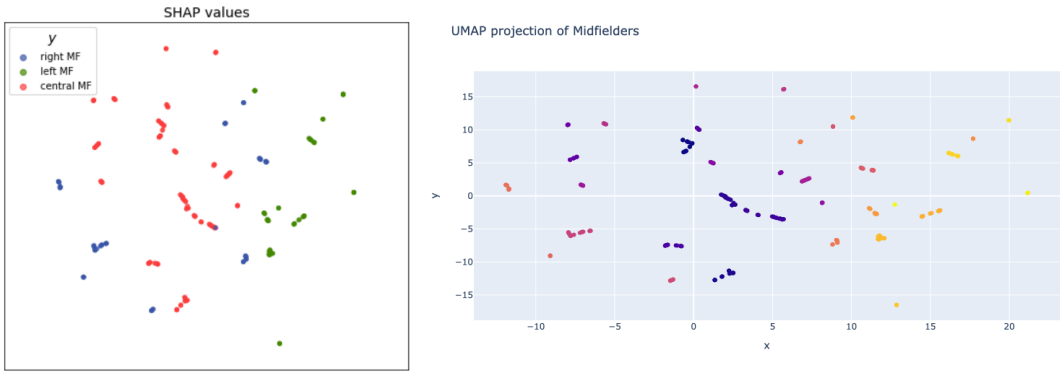


Figura 43 Proyecciones UMAP por roles y grupos de centrocampistas.

En líneas generales, se destaca que hay pocos grupos que incorporen a jugadores en más de un rol. Los roles distintos que comparten grupos son mediocentros y centrocampistas derechos, mientras que, en el caso de los centrocampistas derechos, estos aparecen claramente separados del resto tanto en la proyección con UMAP como en la proyección con PCA.

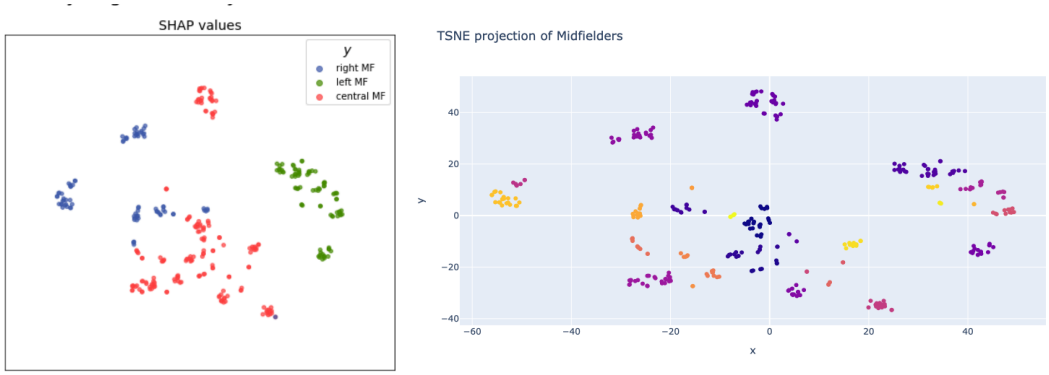


Figura 44 Proyecciones TSNE por roles y grupos de centrocampistas.

5.3.1 Análisis de los grupos

Dado que UMAP ha vuelto a obtener un mejor rendimiento que TSNE, se pasará a examinar al detalle los grupos relevantes que ha devuelto el algoritmo de clustering en el espacio bidimensional de UMAP.

El elevado número de jugadores que se clasifican como centrocampistas viene acompañado de un igualmente elevado número de grupos. Esto perjudica la explicabilidad de dichos grupos mediante el uso de reglas de decisión.

Muchos de los grupos no tienen reglas definidas, sin embargo hay algunos grupos que sí que cuentan con reglas con valores suficientemente altos de precisión y recall en sus reglas.

5.3.1.1 Reglas de decisión

Las reglas relevantes en el caso de los centrocampistas son aquellas que corresponden con los grupos diez, doce, veinte y treinta.

- El grupo veinte lo forman centrocampistas derechos. La particularidad de este grupo con respecto al resto se encuentra en que su frecuencia de pases desde la zona del lateral derecho es más alta que lo habitual (mayor a treinta por cien, cuando la media se sitúa en veinticuatro) y que además participan en duelos por la disputa de un balón por el suelo en los que ningún equipo tiene la posesión clara.
- Los grupos diez y doce son grupos que cuentan con mediocentros. La particularidad del grupo diez es que sus jugadores efectúan menos saques de banda de lo que suelen llevar a cabo los mediocentros y por otro lado el grupo doce que tiene una frecuencia casi tres veces superior a la media de disparo en la parte exterior izquierda del área.
- Finalizando con el grupo de centrocampistas izquierdos, el grupo treinta, su particularidad es que los jugadores efectúan un número de pases desde la zona del lateral izquierdo respecto a la media de centrocampistas izquierdos y que su frecuencia de disparo en la zona S4, el borde derecho central del área, es significativamente menor (menos de la mitad) que el respectivo promedio. Esto indica que los jugadores son de un corte más defensivo que el resto, puesto que se mueven más por la zona del lateral y llegan menos al área.

A pesar de que se siguen captando las particularidades de algunos grupos, la mayoría de ellos han quedado sin caracterizar. Esto es por la ya mencionada limitación que es para las reglas de decisión el contar con un número elevado de grupos y aplicar esta metodología del tipo uno contra todos.

5.3.1.2 Análisis cualitativo

De nuevo, tras el análisis cuantitativo llevado a cabo mediante reglas de decisión, se pueden analizar los grupos obtenidos desde un punto de vista cualitativo. Estos son algunos de los grupos más relevantes con sus particularidades.

Grupo	Miembros	Descripción
27	Marcos Alonso, Jaume Costa, Kurzawa, Ben Davies	Laterales izquierdos con resistencia y buena capacidad de centros.
26	Álex Sandro, Pedraza, Lucas Hernández, F.Delph	Laterales izquierdos veloces al esprint.

19	Barragán, Vrsalijko, Trippier, Bruno Saltor, Juanfran	Laterales derechos con buena capacidad para centrar.
6	Krychowiak, Sergi Darder, D. De Rossi, Fabinho, Guardado, Jorginho	Centrocampistas de corte defensivo.
	Khedira, Modric, Paulinho, Tolisso	Centrocampistas con buen control de balón y visión de juego.
22	Carvajal, Kimmich, Bellerín	Laterales derechos defensivos.
17	Dani Alves, Kyle Walker	Laterales derechos con excelentes atributos físicos y fuerza de disparo.
5	Kanté, Emre Can, M.Vecino	Centrocampistas de corte defensivo, resistentes y hábiles para las entradas de pie.
0	Ndidi, L.Cook, Wijnaldum, Lemina, Hojbjerg, Schneiderlin, Lemina, I.Gueye	Centrocampistas de corte defensivo.

5.4 Experimentación con centrales

Finalmente, en los defensas se cuenta con dos roles, central derecho y central izquierdo. De nuevo, la proyección de las dos primeras componentes PCA se muestra con los centrales con el color verde representando a los centrales derechos y el rojo a los izquierdos en la figura 45.

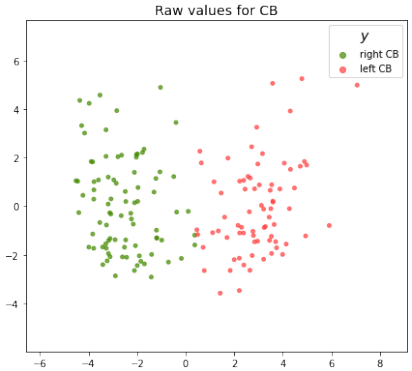


Figura 45 Proyecciones PCA de centrales.

Como único cambio adicional, se ha cambiado de rol a Rodrigo Ely, según Transfermarkt, el jugador juega tanto en el central izquierdo como en el derecho, originalmente se le asignó el rol de central izquierdo, pero por su localización en la primera componente PCA, parece indicar que en esta temporada jugó de central derecho.

Esta separación de roles de defensa no es la única que se podría hacer, siendo otra posibilidad entre la separación por laterales (izquierdos y derechos) y centrales.

No obstante, la división de roles de la base de datos original realizó la separación ya mencionada de centrales y los laterales se clasificaron como centrocampistas derechos e izquierdos. La justificación detrás de esta decisión no es explicada en la base de datos original, aunque puede ser debida a que los laterales ocupan en algunas alineaciones el rol de carrilero, el cual se asemeja al rol del centrocampista de banda ya que se implica en tanto tareas de defensa como de ataque.

Tras entrenar el modelo supervisado se pudo ver que las variables más relevantes en la predicción de roles para centrales derechos e izquierdos son las mismas (figura 46), indicando que esta separación de los defensas no es una separación tan adecuada como la de los delanteros.

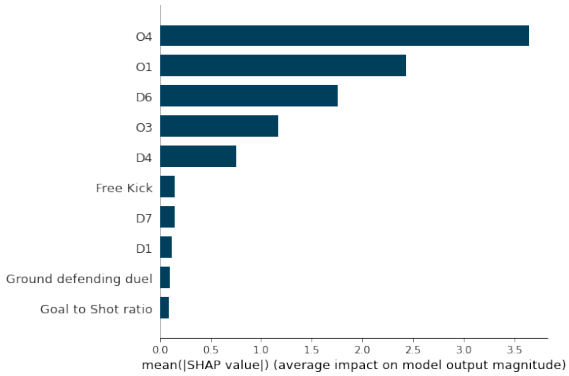


Figura 46 Gráfico de valores SHAP de centrales.

No obstante, a pesar de que las variables más importantes sean las mismas para ambos roles, un análisis más detallado de los valores SHAP con los "Beeswarm plots" indica que la importancia de las variables para los valores SHAP es inversa.

El gráfico siguiente (figura 47) está ordenado descendentemente según la su valor shap medio absoluto para todos los individuos, esto es lo mismo que el gráfico de barras mostrado previamente. En este tipo de gráficos, el valor SHAP de cada instancia en cada una de las variables está representada por un punto. Los puntos se distribuyen en el eje horizontal y en aquellas zonas donde hay una elevada cantidad de valores los puntos se apilan verticalmente. Por último, la barra de colores corresponde a con los valores originales de los datos, los puntos en azul significan

valores relativamente pequeños para la variable en concreto y los puntos en morado valores elevados para la variable.

Examinando el color de la distribución de las variables en el eje horizontal permite conocer cuál es la relación entre los valores originales de una variable y los valores SHAP.

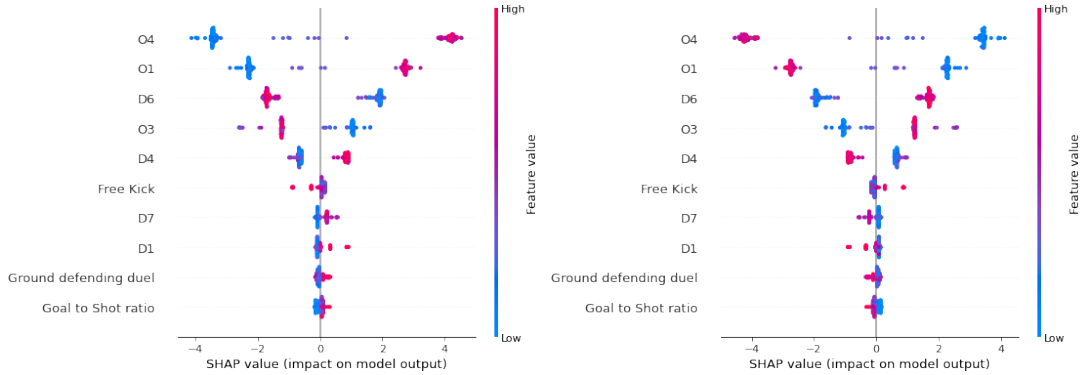


Figura 47 "Beeswarm plots" para centrales.

Así, mientras los centrales zurdos se caracterizan por un elevado número de pases desde la parte izquierda de las zonas de defensa y centro del campo y poca frecuencia de pases hacia la zona derecha del centro del campo y desde la zona derecha de la defensa, los centrales diestros se caracterizan por los opuesto.

Del proceso de selección de parámetros óptimos hay que destacar que, para UMAP tiene más relevancia los parámetros de la técnica UMAP que la selección de técnica de agrupación (0.8 frente a 0.2, tal y como muestra la figura 48) y con t-SNE los papeles se invierten, la técnica de agrupamiento seleccionada es mucho más importante para la función objetivo que el parámetro de t-SNE. El mejor resultado que devuelve UMAP es de nuevo mejor en términos de calidad de clustering que el de t-SNE con un valor de la función objetivo de 0.78 y 0.64 respectivamente.

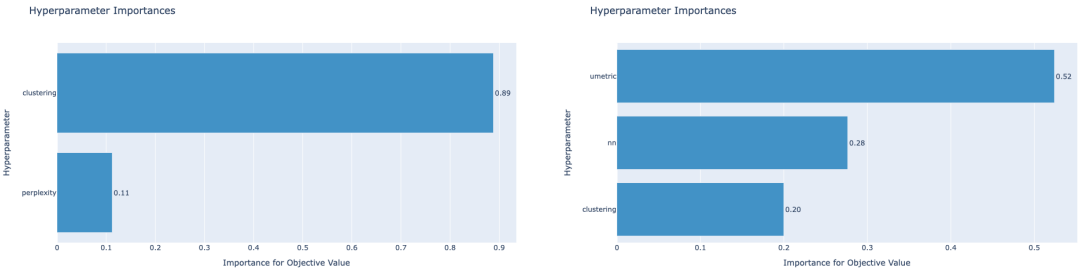


Figura 48 Importancia de hiperparámetros con TSNE y UMAP para centrales.

El resultado óptimo con UMAP, se obtienen 20 grupos con la técnica de agrupamiento DBSCAN. Los resultados se presentan en la figura 49 con la gráfica de la izquierda coloreada por posiciones

y la de la derecha por grupos. Este modelo cuenta con siete reglas de decisión relevantes de entre estos veinte grupos.

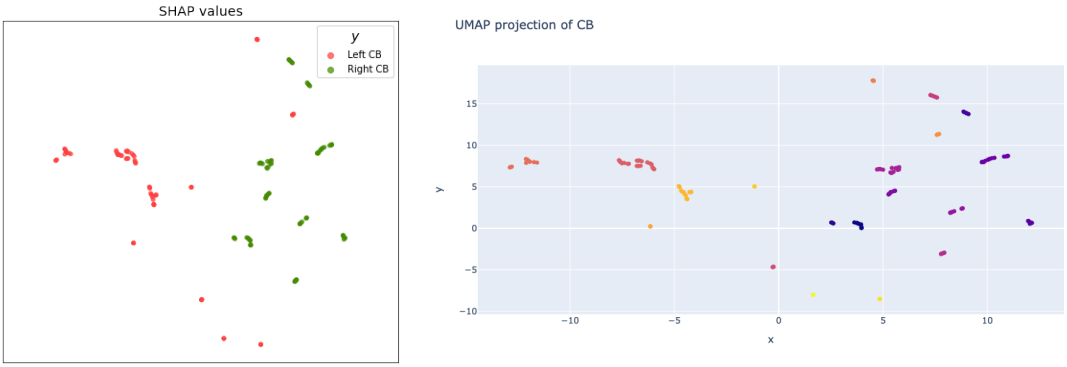


Figura 49 Proyecciones UMAP por roles y grupos de centrales.

Con t-SNE, el modelo óptimo se ha obtenido con agrupamiento jerárquico y diez como el número de grupos. Al contar con menos clústeres se hace más fácil la búsqueda de reglas que los caractericen, puesto que estas reglas se obtienen comparando un grupo con el resto. Algunas peculiaridades se pueden ver (en la figura 50 a continuación) en el grupo dos (morado oscuro), en el cual aparecen jugadores con un elevado número de duelos aéreos (más de ocho por partido) o el grupo siete, en el cual los jugadores son de los que con más frecuencia lanzan los pases desde la zona central de la defensa.

En el caso de TSNE, de los diez grupos que se forman, siete son caracterizados correctamente por reglas de decisión. Esto es una mejora respecto a la capacidad de explicabilidad del resultado con UMAP, aunque a cambio, el valor de calidad del clustering es significativamente peor.

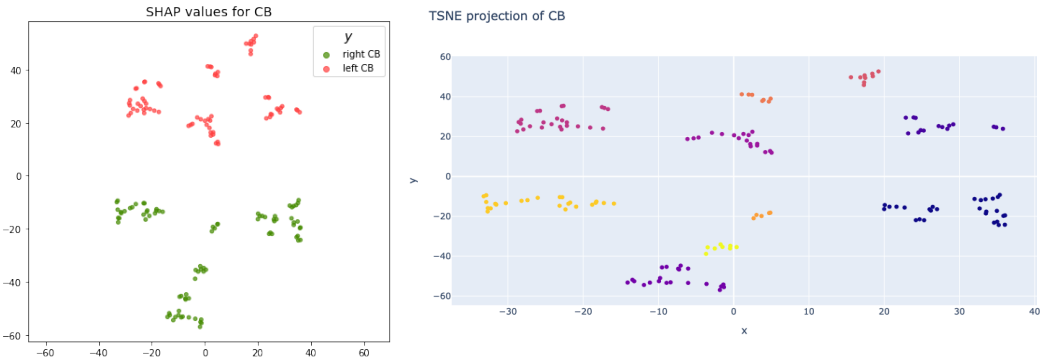


Figura 50 Proyecciones TSNE por roles y grupos de centrales.

5.4.1 Análisis de grupos

Se procede a continuación al análisis de los grupos resultantes de centrales, tanto desde el punto de vista cuantitativo cómo el cualitativo.

5.4.1.1 Reglas de decisión

Entrando de nuevo al detalle de las reglas de decisión con el resultado óptimo (UMAP + DBSCAN), de los veinte grupos obtenidos contamos con reglas de decisión válidas para los grupos dos, cinco, ocho, nueve, catorce, quince y dieciséis.

- Empezando por los grupos de centrales derechos, el grupo dos cuenta en la parte del antecedente con la condición O2>0.35, esto se traduce en que la frecuencia de pases que efectúan partiendo de la zona central de la defensa es superior a la media de centrales derechos (0.22).
- Siguiendo con los centrales derechos, el grupo cinco parte de las condiciones del antecedente de D3 >= 0.1 y O7<=0.001, la primera por encima de la media de D3 para centrales derechos y la segunda por debajo de la media (0.04).
- En cuanto al grupo ocho, la condición del antecedente Free Kick > 2.45 significa que los jugadores se encuentran muy por encima de la media de número de faltas lanzadas por partido, la cual es de aproximadamente 1.4 para todos los defensas.
- El grupo nueve cuenta con jugadores los cuáles reportan los valores más altos de precisión de pase, con un valor mayor al noventa y dos por cien respecto a la media de ochenta y seis.
- Por parte del grupo quince, en este grupo se catalogan tres centrales izquierdos, los cuales destacan respecto al resto por una frecuencia de pases mayor hacia la zona de defensa izquierda y también por que participan en menos intentos de quitar la posesión al contrario cuando este está en un proceso de ataque.
- En cuanto a las reglas de los grupos catorce, y dieciséis, a pesar de ser válidas en cuanto a precisión y recuperación, no son muy informativas, ya que las magnitudes en las que difieren la media y las condiciones del antecedente no son muy relevantes. Estos grupos tienen en sus reglas antecedentes muy similares, valores de D6 por debajo de la media para centrales izquierdos y, por otra parte, valores de ratio goles entre disparos más bajo que la media de centrales.

5.4.1.2 Análisis Cualitativo

Entre los grupos que se han formado, en un análisis más allá de los datos numéricos que recogen los valores numéricos, se puede observar que algunos de los grupos cuentan con jugadores de un corte por cualidades físicas o formas de juego.

Estos grupos se muestran a continuación en la tabla 9 junto a sus integrantes y su perfil.

Tabla 9 Análisis cualitativo de defensas centrales.

Grupo	Miembros	Descripción
1	Alderweireld, Boateng, Issa Diop	Centrales derechos con fuerza física y con muy buena capacidad para las entradas en parado
4	Lovren, Dante, Pezzela, Mustafi, Smalling	Centrales con buenos atributos de entradas, fuerza e intercepciones.
5	Varane, Garay, Tah	Centrales de corte clásico, buena capacidad de intercepciones y reacciones.
6	Matip, Albiol, Skriniar, Manolas, Marquinhos, Morgan, Savic	Mejores atributos defensivos en robo, anticipación y cabezazos.
7	Glik, Kjaer, Rami, Koundé, Cook	Jugadores agresivos con buena capacidad de robo.
9	Benatia, Rugani, Piqué, N'Koulou, Cabral	Buen marcaje y capacidad de luchar balones por alto.
11	Upamecano, Sidnei, Sergio Ramos, Lewis Dunk, Mario Hermoso, Rüdiger	Centrales de corte físico y con buen juego aéreo.
15	Rolando, Sakho, Gamberini	Jugadores fuertes y con buenas entradas en parado.
18	Astori, Bender, Hummels	Los robos son su mejor cualidad defensiva y como mejor cualidad mental las intercepciones.

5.5 Conclusiones de la experimentación

El conocimiento que se ha podido extraer de los diversos experimentos en cuanto a las pruebas realizadas con los demarcaciones de delanteros, centrocampistas y centrales han arrojado información relevante tanto en cuanto a los procesos de optimización y la formación de los grupos.

En primer lugar, se debe destacar que, en el caso de las técnicas de reducción de dimensionalidad, UMAP y TSNE se comportan de forma distinta en el sentido de que en la primera técnica se le da mucha más importancia a la forma de obtener la proyección en la baja dimensionalidad que a la técnica de agrupación que se utiliza, mientras que en con la segunda, los hiperparámetros relativos a la reducción de la dimensión son casi irrelevantes y toma más importancia el algoritmo de clustering que se use.

Esto se puede explicar con el hecho de que, en términos de calidad de los grupos, con UMAP, la separabilidad de estos viene dada por los parámetros de vecindad local y la métrica que se utiliza para computar las distancias, mientras que en el caso de TSNE, el parámetro "Perplexity" no entra en juego en la separabilidad de los grupos, sino que esto en su lugar viene dado por la técnica de agrupamiento que busca los grupos en este nuevo espacio.

UMAP se adapta mejor a las tres distribuciones de datos en cuanto a calidad interna de clustering, resultando en una mejor separabilidad (grupos mejor distinguidos del resto) y compacidad (como de cerca se sitúan los elementos dentro de un mismo grupo).

Por otra parte, en referencia a los grupos resultantes, se pude ver que con un número menor de grupos las reglas de decisión son más efectivas para ayudar a comprender las diferencias entre grupos. Esto se debe a que, en el caso de los centrocampistas (el cual es el que cuenta con más grupos resultantes), la formación del elevado número de grupos se traduce en una formación excesiva de fronteras, lo que causa que las reglas de decisión no sean capaces de encontrar en muchos de los casos fronteras que diferencien a un grupo del resto.

Para paliar este problema, una posible solución es la de aumentar el número mínimo de vecindad local para disminuir el número de grupos, o bien reestructurar la base de datos, reclasificando a los laterales que se etiquetaron como centrocampistas derechos o izquierdos y realizando el estudio de defensas con centrales y laterales.

6 Conclusiones

Esta sección se encarga de resumir las ideas clave y la revisión de los objetivos marcados al principio en base al trabajo realizado en las distintas fases de este trabajo.

Recapacitando, el objetivo planteado era, tras adaptar un conjunto de datos abiertos para obtener las características técnicas relevantes de los jugadores, comparar diferentes metodologías para encontrar grupos de jugadores en cada demarcación, así como las características más relevantes a la hora de encontrar los grupos.

El primer paso se consiguió al adaptar los datos abiertos de eventos de la empresa Wyscout, esto se consiguió llevando a cabo un análisis exploratorio de los datos originales y la posterior transformación y limpieza de la base de datos final, donde han entrado en juego librerías para tratamiento de datos, PCA para una primera visualización de los individuos y el estudio del dominio de la analítica deportiva. Esto supuso un trabajo adicional respecto a otras bases de datos que hubieran necesitado de menos modificaciones, no obstante, es importante adaptarse a la información que se encuentra disponible ya que no es siempre posible obtener conjuntos de datos privados y menos aún en el fútbol donde unas pocas empresas son las que mantienen el almacenamiento de los datos y son muy herméticas acerca de a quién los proporcionan dado el elevado nivel de ingresos que les reportan.

Por otra parte, lograr el objetivo plateado implicó utilizar un enfoque de clustering supervisado, la búsqueda óptima de hiperparámetros mediante optimización y en último lugar, la aplicación de reglas de decisión a los grupos, así como la presentación de los resultados en una única vista interactiva para poder extraer conocimiento relevante, la cual no sólo incluye las características técnicas de los jugadores sino que también cuenta con la información del valor económico de los jugadores con lo cual se puede reducir la lista de individuos que pueden resultar interesantes a un club añadiendo una restricción de precio.

El clustering supervisado en conjunto con el uso de optimización en la experimentación es un enfoque innovador en este campo el cual ha permitido no sólo limitarse a la representación de los grupos, sino que, además, con el añadido de las reglas de decisión, se pueden caracterizar los grupos acorde a cuáles son las características que los distinguen del resto y en el caso de la transformación aplicada con la librería SHAP, se puede saber cuáles han sido las variables más relevantes para distinguir los distintos roles en una determinada posición.

En cuanto a los resultados de experimentación, UMAP obtiene un mejor rendimiento con los conjuntos de datos y que en cuánto a las técnicas de agrupación que t-SNE, y en cuanto a algoritmos de agrupación, el jerárquico y DBSCAN han sido las más recurrentes. Esto no tiene por qué ser así con otros conjuntos de datos y es relevante a futuro considerar todas las técnicas de reducción de dimensionalidad, así como los algoritmos de agrupación y actualizar el estudio con nuevas metodologías.

No hay que olvidar que, en cualquier deporte, los datos no siempre cuentan toda la verdad, no se deben basar las decisiones únicamente en estos. Hay una serie de factores que van más allá de las características de juego y que tienen un impacto determinante en el rendimiento de los jugadores. Estados de ánimo, sinergias entre individuos, estado físico son sólo algunos de los factores que pueden tener un enorme impacto en el impacto que tiene un jugador al incorporarse a un nuevo grupo y esto es algo que no se puede contemplar en los datos. Este tipo de herramientas debe siempre servir como apoyo y en ningún caso sustituir a grupos de trabajo de profesionales en los distintos sectores que se ven implicados en los clubes deportivos.

6.1 Legado

Al finalizar cualquier proyecto es necesario conocer cuál es la herencia que este deja tanto al autor como a los destinatarios.

En el caso del autor, realizar esta investigación ha supuesto un aporte relevante al trabajo en la analítica deportiva, sirviendo tanto de punto de partida como para profundizar en el campo de conocimiento. Trabajos de esta naturaleza son de utilidad para ayudar en el desarrollo académico ya que además del apartado tecnológico, también lo son para poder llegar a una audiencia más genérica, con el beneficio que esto reporta tanto a nivel personal como en conjunto a la disciplina de la ciencia de datos.

Desde un punto de vista académico, este TFG ofrece un aporte al campo de la ciencia de datos y su aplicación en el deporte. Explorando nuevas metodologías y técnicas para analizar los aspectos técnicos del juego y encontrar agrupaciones de jugadores similares. Se desarrollan nuevas perspectivas y enfoques que podrían ser útiles para futuros investigadores y profesionales en el campo.

También esta tarea tiene su aplicación práctica en la industria deportiva, al lograr desarrollar un modelo para identificar perfiles de futbolistas similares basándose en los aspectos técnicos del juego, los clubes podrían utilizar esta herramienta para identificar jugadores con características similares a los que ya tienen en sus filas, lo que les permitiría tomar decisiones más fundamentadas en el proceso de selección de jugadores.

Asimismo, supone una mejora en la toma de decisiones, que podría ayudar a los entrenadores y a los responsables de la toma de decisiones en los clubes a tomar decisiones más informadas en cuanto a la formación de equipos.

Por último, también se pretende que este TFG pueda servir de inspiración para futuros estudios en áreas similares. Conduciendo al aumento de las investigaciones en el campo, para permitir que la analítica deportiva se pueda mantener constantemente actualizada con respecto al contexto tecnológico.

6.2 Relación del trabajo desarrollado con los estudios cursados

Al respecto del nexo entre este trabajo y los estudios del grado de Ciencia de Datos, este trabajo hace uso de los conocimientos adquiridos tanto para el abordaje de problemas, así como del marco teórico que permite abordar cada fase de este trabajo ordenadamente y con la justificación de cada una de las decisiones tomadas junto con la comprensión de los resultados obtenidos en cada fase.

Empezando por la fase de obtención y tratamiento de datos, los conocimientos de bases de datos han permitido, en primer lugar, discernir la calidad de la fuente primaria de datos para poder así elegir una fuente con la suficiente adaptabilidad por la tarea que nos concierne, así como con la cantidad adecuada de sujetos (jugadores) y también con la suficiente extensión temporal (temporadas). En lo referente al tratamiento de los datos, el procesado necesario para adecuar los datos para poder obtener la información disgregada por jugadores se ha hecho con la ayuda de los conocimientos tantos de bases de datos como de programación, puesto que este tratamiento se ha hecho en el lenguaje Python y la librería Pandas. Aquí se establece una conexión con las asignaturas de bases y gestión de datos.

Una vez obtenida la base de datos, se ha realizado un análisis descriptivo de esta, para este análisis se ha hecho uso tanto de las técnicas de visualización junto con el análisis de componentes principales para la selección de características y el estudio de datos anómalos. Este análisis ha sido relevante puesto que ha permitido que los resultados que ofrecen los modelos han sido más relevantes y que se aplique el preprocesado dadas las diferentes escalas de las características.

Además de con el análisis descriptivo, las técnicas de reducción de dimensionalidad han sido necesarias para trabajar con la alta dimensionalidad de nuestros datos, haciendo comprensibles los resultados en dos dimensiones y para el estudio de la colinealidad de las características. Estas técnicas entran dentro del área de modelos descriptivos y predictivos.

En el proyecto de búsqueda de jugadores de fútbol similares, se evalúa la calidad de los grupos, así como la similitud de los jugadores recomendados. Esto implica el uso de métricas de evaluación y de nuevo de técnicas de visualización de datos.

El uso de los diferentes modelos de aprendizaje no supervisado y supervisado se enmarca en las asignaturas de aprendizaje automático, siendo estas técnicas las que se han utilizado en la fase de modelado.

En cuanto a la infraestructura de los experimentos, aquí entra en juego el conocimiento adquirido en la asignatura de optimización.

La presentación de los resultados se enmarca en el uso de técnicas de visualización, así como la explicación de estos resultados estrechamente relacionada con la comprensión y explicación de conocimientos del dominio específico.

Por último, en lo relativo a la legalidad que se ve implicada en la obtención y el uso de datos, se establece un nexo con las áreas de profesionalidad y marco laboral tratadas en el grado.

En conclusión, las habilidades obtenidas en el grado han permitido la especialización del dominio, así como de la capacidad de abordaje de un proyecto de ciencia de datos con todas sus fases.

7 Trabajos futuros

Durante el desarrollo de este TFG, se ha llevado a cabo un análisis de los aspectos fundamentales relacionados con el uso de técnicas de aprendizaje automático para la detección de patrones similares de juego en futbolistas profesionales. Sin embargo, como es característico de cualquier estudio científico, siempre hay espacio para nuevas investigaciones y mejoras. Esta sección tiene como objetivo identificar y plantear diversas propuestas para investigaciones futuras que podrían expandir los horizontes y ampliar los conocimientos en este campo.

En primer lugar, sería interesante, buscar una forma de considerar la compatibilidad de jugadores, considerando las áreas en las que puede haber sinergias o aquellas en las que las debilidades de un jugador pueden ser suplidas por las fortalezas de otro. Esto añade una capa de complejidad debido al elevado número de combinaciones que se pueden dar entre un par de sujetos. De la misma forma, también se podría buscar la forma de medir la compatibilidad de un perfil concreto con un equipo.

Otra posible línea de investigación puede ser la búsqueda de posiciones con las cuales el jugador puede ser compatible. Esto implicaría contrastar los atributos técnicos del jugador con los atributos relevantes para un rol concreto. Se debería añadir una variable con los roles que se pueden desarrollar en cada posición y a partir de ahí obtener valores de referencia para los roles y llevar a cabo un estudio estadístico para establecer un criterio de similitud. Los resultados se podrían mostrar en gráficos de radar u otras gráficas que permitan saber las características en las que el jugador tendría que trabajar.

En último lugar, con la información disponible de Transfermarkt y añadiendo datos sobre los jugadores como la edad o aspectos relacionados con la comerciabilidad de un jugador, para predecir la evolución en el valor de los futbolistas. Esta información podría ser de interés para aquellos clubes que busquen jugadores que se vayan a revalorizar.

Las limitaciones a la hora de obtener datos de jugadores han resultado en que sólo se consideren los datos relativos a una temporada, lo cual significa que no se obtiene una imagen muy realista de toda la carrera futbolística, no es extraño que los jugadores se lesionen o ya no cuenten con las mismas cualidades físicas y cambien su estilo de juego o que por decisiones del técnico cambien de posición. Para estas futuras investigaciones, puede que sea interesante actualizar la base de datos de la que se dispone con más temporadas y más ligas, para conseguir resultados más

completos y poder añadir una capa más de complejidad pudiendo llegar a incluir análisis con evoluciones temporales.

8 Bibliografía

(s.f.). Obtenido de Wyscout: https://www.hudl.com/products/wyscout

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining*.

Barbosa, A., Ribeiro, P., & Dutra, I. (2022). Similarity of football players using passing sequences. *Machine Learning and Data Mining for Sports Analytics*.

Barseghian, A. (7 de Octubre de 2019). How Nike Is Using Analytics To Personalize Their Customer Experience. *Forbes*.

Bransen, L., & Van Haaren, J. (2018). Measuring football players' on-the-ball contributions from passes during games. *SciSports*, 1-13.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., . . . Holt, B. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 2825-2830.

Chang, J., Wang, L., Meng, G., Xiang, S., & Pan, C. (2017). Deep Adaptive Image Clustering. *International Conference on Computer Vision*, 5879-5887.

de la Torre, R., Lopez-Lopez, D., Juan, A. A., & Clavet, L. O. (2022). Business Analytics in Sport Talent Acquisition: Methods, Experiences, and Open Research Opportunities. *International Journal of Business Analytics*, 1-20.

Doyle, A., Doyle, O., & Bourke, A. (2021). Supervised Clustering for Subgroup Discovery: An Application to COVID-19 Symptomatology. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 408-422.

García-Aliaga, A., Marquina, M., Coreton, J., Rodriguez-Gonzalez, A., & Luengo-Sanchez, S. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching*, 148-157.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 90-95.

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. *International Conference on Computational Techniques, Electronics and Mechanical Systems*, 129-135.

Lopez Peña, J., & Sanchez Navarro, R. (2015). Who can replace Xavi? A passing motif analysis of football players. *Physycs and Society*, 1-9.

Lopez Peña, J., & Sanchez Navarro, R. (2015). Who can replace Xavi? A passing motif analysis of football players. *Physycs and Society*, 1-9.

Malagón Selma, M. d. (Septiembre de 2019). *Machine Learning en el mundo del fútbol*. Valencia, España: Universitat Politècnica de València.

Marr, B. (2016). Rolls-Royce: How Big Data is Used to Drive Success in Manufacturing. En *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results* (págs. 25-30). Wiley.

Mazurek, J. (2018). Which Football Player Bears Most Resemblance to Messi? A Statistical Analysis.

McInnes, L., Healy, J., & Melville, J. (2020). Uniform Manifold Approximation and Projection for Dimension Reduction. 1-63.

McKenna, B. (19 de Agosto de 2014). Bayern Munich teams up with SAP to hit sporting and commercial goals. *ComputerWeekly.com*.

Pappalardo, L., Cintia, P., Rossi, A., Emanuele, M., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatiotemporal match events in soccer competitions. *Scientific data*.

Rejec, J. (2016). How Big Data is Changing the World of Football. *Datafloq*.

Sarnoff, Z. (5 de Abril de 2018). Obtenido de Harvard: https://d3.harvard.edu/platform-digit/submission/moreyball-the-houston-rockets-and-analytics/

Stanojevic, R., & Gyarmati, L. (2016). Towards data-driven football player assessment. *IEEE 16th International Conference on Data Mining Workshops*, 167-172.

Tzai Lampisa, N. I. (2023). Predictions of european basketball match results with machine learning algorithms. *Journal of Sports Analytics*, 1-20.

van der Maaten, L., & Hinton, G. (2018). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 2579-2605.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, 1-3021.

9 Anexos

9.1 Anexo I - Base de datos

Tabla 10 Base de datos final

wyIndex	Índice asignado por wyScout
shortName	Abreviatura del nombre completo del jugador
playerId	
totalMins	Minutos totales jugados a lo largo de la temporada.
roleCluster	Posición del jugador.
Oi (O1, O2, O3, O4, O5, O6, O7, O8, O9)	Frecuencia (entre cero y uno) de pases que se efectúan con origen en cada una de las zonas del campo. Para un jugador, la suma de los valores desde O1 hasta O9 es 1.
Di (D1, D2, D3, D4, D5, D6, D7, D8, D9)	Frecuencia (entre cero y uno) de pases que se efectúan con destino a cada una de las zonas del campo. Para un jugador, la suma de los valores desde O1 hasta O9 es 1.
PassAccuracy	Porcentaje de acierto de pases, resultado de dividir el total de pases completados con éxito entre el total de pases intentados.
Si (S1, S2, S3, S4, S5, S6)	Frecuencia (entre cero y uno) de disparos que se efectúan desde cada una de las zonas en las que se divide el último tercio del campo contrario. Para un jugador, la suma de los valores desde S1 hasta S6 es 1.
Goal to Shot Ratio	División entre el número de goles y el número de disparos efectuados.
AirDuel	Cantidad de disputas de balones aéreos con algún jugador del equipo contrario. Este valor se representa en la cantidad de acciones por 90 minutos.
Ground attacking duel	Cantidad de disputas en acciones de ataque de balones rasos con algún jugador del equipo

	contrario. Del tipo x90 (cantidad de acciones por cada 90 minutos).
Ground defending duel	Cantidad de disputas en acciones de defensa de balones rasos con algún jugador del equipo contrario. Del tipo x90.
Ground loose ball duel	Cantidad de pérdidas de balones rasos. Del tipo x90.
Free Kick	Tiros libres efectuados. Del tipo x90.
Throw in	
Corner	Córneres efectuados. Del tipo x90.
Free kick cross	
Free Kick shot	
Penalty	
Goal Kick	
Initial Value	Valor de mercado de Transfermarkt al inicio de la temporada.
League	Liga en la que participa el equipo del jugador.

Obtención de la base de datos

Pases

```
def Passpr(passes, db):

    discrPass = []
    for i in passes["positions"]:
        discrPass.append(discretisation(i))

    passes['Discr'] = discrPass
    Transform = passes[["playerId","Discr"]]

    g1 = pd.DataFrame({'count':Transform.groupby(["playerId", "Discr"])['Discr'].count()}).reset_index()

    players = db["playerId"].unique()
    zones = [("Z1","Z1"),("Z1","Z2"),("Z1","Z3"),('Z1','Z4'),('Z1','Z5'),('Z1','Z6'),('Z1','Z7'),('Z1','Z8'),('Z1','Z9'),
            ("Z2","Z1"),("Z2","Z2"),("Z2","Z3"),('Z2','Z4'),('Z2','Z5'),('Z2','Z6'),('Z2','Z7'),('Z2','Z8'),('Z2','Z9'),
            ("Z3","Z1"),("Z3","Z2"),("Z3","Z3"),('Z3','Z4'),('Z3','Z5'),('Z3','Z6'),('Z3','Z7'),('Z3','Z8'),('Z3','Z9'),
            ("Z4","Z1"),("Z4","Z2"),("Z4","Z3"),('Z4','Z4'),('Z4','Z5'),('Z4','Z6'),('Z4','Z7'),('Z4','Z8'),('Z4','Z9'),
            ("Z5","Z1"),("Z5","Z2"),("Z5","Z3"),('Z5','Z4'),('Z5','Z5'),('Z5','Z6'),('Z5','Z7'),('Z5','Z8'),('Z5','Z9'),
            ("Z6","Z1"),("Z6","Z2"),("Z6","Z3"),('Z6','Z4'),('Z6','Z5'),('Z6','Z6'),('Z6','Z7'),('Z6','Z8'),('Z6','Z9'),
            ("Z7","Z1"),("Z7","Z2"),("Z7","Z3"),('Z7','Z4'),('Z7','Z5'),('Z7','Z6'),('Z7','Z7'),('Z7','Z8'),('Z7','Z9'),
            ("Z8","Z1"),("Z8","Z2"),("Z8","Z3"),('Z8','Z4'),('Z8','Z5'),('Z8','Z6'),('Z8','Z7'),('Z8','Z8'),('Z8','Z9'),
            ("Z9","Z1"),("Z9","Z2"),("Z9","Z3"),('Z9','Z4'),('Z9','Z5'),('Z9','Z6'),('Z9','Z7'),('Z9','Z8'),('Z9','Z9')]

    main = []
    for player in players:
        totalPass = 0
        second = []
        for zone in zones:
            try:
                second.append(g1.loc[(g1['playerId'] == player) & (g1['Discr'] == zone),'count'].item())
                totalPass += int(g1.loc[(g1['playerId'] == player) & (g1['Discr'] == zone),'count'].item())
            except:
                second.append(0)

        second = [x / (totalPass+0.01) for x in second]
        main.append(second)

    df = pd.DataFrame(main, columns = zones)
    df.index = players
    df.reset_index(inplace=True)
    df = df.rename(columns = {'index':'playerId'})
    df = df.sort_values(by=['playerId'])
    df.sort_index(ascending=True)
    return df
```

Figura 51 Transformación original de los eventos de tipo pase.

```python
def uniPass(db):
    starts = ["('Z1'","('Z2'","('Z3'","('Z4'","('Z5'","('Z6'","('Z7'","('Z8'","('Z9'"]
    finishes = ["Z1')","Z2')","Z3')","Z4')","Z5')","Z6')","Z7')","Z8')","Z9')"]
    startdb = pd.DataFrame()
    finishdb = pd.DataFrame()
    i,j=0,0
    for start in starts:
        i += 1
        zn = db[db.columns[[x.startswith(start) for x in db.columns]]]
        zn = zn.sum(axis=1)
        zn.name = start
        startdb[f"O{i}"] = zn
    for finish in finishes:
        j+=1
        zfn = db[db.columns[[x.endswith(finish) for x in db.columns]]]
        zfn = zfn.sum(axis=1)
        zfn.name = finish
        finishdb[f"D{j}"] = zfn

    part1 = db.iloc[:,0:4]
    part2 = db.iloc[:,85:]

    frames = [part1,startdb,finishdb,part2]
    result = pd.concat(frames, axis=1)

    return result
```

Figura 52 Simplificación de los eventos de pase

Disparos

```python
def Shotpr(df, playerdb):

    shots = df[df['eventName'].str.contains('Shot')]
    dsh = []

    for i in shots["positions"]:
        dsh.append(ShotDiscretisation(i))
    shots['Discr'] = dsh
    ShotTransform = shots[["playerId","Discr"]]

    g3 = pd.DataFrame({'count':ShotTransform.groupby(["playerId", "Discr"])['Discr'].count()}).reset_index()

    zones = [("Z1"),("Z2"),("Z3"), ("Z4"),("Z5"), ("Z6"),("LD")]
    players = playerdb["playerId"].unique()

    main = []
    for player in players:
        totalShot = 0
        second = []
        for zone in zones:
            try:
                second.append(g3.loc[(g3['playerId'] == player) & (g3['Discr'] == zone),'count'].item())
                totalShot += int(g3.loc[(g3['playerId'] == player) & (g3['Discr'] == zone),'count'].item())
            except:
                second.append(0)
        second = [x / (totalShot+0.01) for x in second]
        main.append(second)

    shotsdf = pd.DataFrame(main, columns = zones)
    shotsdf.index = players
    shotsdf.reset_index(inplace=True)
    shotsdf = shotsdf.rename(columns = {'index':'playerId'})
    shotsdf = shotsdf.sort_values(by=['playerId'])
    shotsdf.sort_index(ascending=True)
    return shotsdf
```

Figura 53 Transformación de los eventos de disparo

```python
def shotAcc(tags):
    res = []
    value = "ND"
    for tag in tags:
        res.append(list(tag.values()))

    try:
        for i in res:
            if i[0] == 101:
                value = "Goal"
    except:
        pass

    return value
```

```python
def ShotsGoal(db):
    strList = []
    for i in db["tags"]:
        strList.append(shotAcc(i))
    db["ShotAccuracy"] = strList
    return db
```

```python
def accShot(db):
    db = ShotsGoal(db)
    ongoal = db[db['ShotAccuracy'].str.contains('Goal')]


    gdb = pd.DataFrame({'count':db.groupby(["playerId"])['playerId'].count()})
    gac = pd.DataFrame({'count':ongoal.groupby(["playerId"])["playerId"].count()})


    gdb["count"] = gac["count"]/gdb["count"]
    gdb['count'] = gdb['count'].replace(np.nan, 0)
    gdb = gdb.rename(columns={"count": "Goal to Shot ratio"})

    return(gdb)
```

Figura 54 Obtención de la columna ShotAccuracy

Duelos y jugadas a balón parado

```python
def countSub(db, evType):
    tdb = db[db['eventName'].str.contains(evType)]
    tdb = tdb[["playerId","subEventName"]]
    g = pd.DataFrame({'Total':tdb.groupby(["playerId", "subEventName"])['subEventName'].count()}).reset_index()

    players = db["playerId"].unique()
    subev = g["subEventName"].unique()
    counts = []

    for player in players:
        sub = []
        for sev in subev:
            try:
                sub.append(g.loc[(g['playerId'] == player) & (g['subEventName'] == sev),'Total'].item())
            except:
                sub.append(0)
        sub = [x / (db.loc[(db['playerId'] == player),'x90'].values[0]) for x in sub]
        counts.append(sub)

    res = pd.DataFrame(counts, columns = subev)
    res.index = players
    res.reset_index(inplace=True)
    res = res.rename(columns = {'index':'playerId'})
    res = res.sort_values(by=['playerId'])
    res.sort_index(ascending=True)

    return res
```

Figura 55 Transformación de los eventos duelo y jugada a balón parado

Valores de TransferMarkt

```
def similarity(df1, df2, col1, col2, colname, threshold = 70):
    # Empty lists for storing the matches later
    mat1, mat2, p = [],[],[]

    # Converting df column to list to do fuzzy matching
    list1 = df1[col1].tolist()
    list2 = df2[col2].tolist()

    # Iterating through list1 to extract it's closest match from list2
    for i in list1:
        mat1.append(process.extractOne(i, list2, scorer=fuzz.token_set_ratio))

    # Iterating through closest matches to filter out the closest
    if threshold:
        for j in mat1:
            if j[1]>= threshold:
                p.append(j[0])
            mat2.append(",".join(p))
            p = []
    else:
        for j in mat1:
            p.append(j[0])
            mat2.append(",".join(p))
            p = []

    # Storing the resultant matches back to df1
    df1[colname] = mat2

    return df1
```

Figura 56 Incorporación de los datos de valor de traspaso

9.2 Anexo II – ODS

Grado de relación con los objetivos de desarrollo sostenible

Tabla 11 Relación con los ODS

Objetivos de Desarrollo Sostenibles	**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**		

Los Objetivos de Desarrollo Sostenible (ODS) son una serie de 17 metas globales establecidas por las Naciones Unidas en 2015 para abordar desafíos mundiales, como la pobreza, el hambre, la igualdad de género, la educación y el medio ambiente, con el objetivo de mejorar la calidad de vida y proteger el planeta para las generaciones futuras.

Este trabajo en el campo de la ciencia de datos se relaciona indirectamente con varios Objetivos de Desarrollo Sostenible establecidos por las Naciones Unidas. A través de la aplicación de técnicas de aprendizaje automático y clustering al análisis de datos de fútbol.

Este trabajo se enmarca en el contexto de la analítica deportiva y por tanto no encuentra nexos con los objetivos de preservación del planeta o la mayoría de los proyectos contra la pobreza, no obstante, sí se pueden ver algunas relaciones con objetivos relacionados con salud, educación e innovación, a pesar de que estos no resulten directamente de los objetivos planteados a lo largo de esta investigación.

A continuación, se enumeran los objetivos con los que se han encontrado vínculos respecto al presente trabajo:

- En primer lugar, en relación con el objetivo **cuarto** de las ODS de educación de calidad, el análisis de datos deportivos puede ser utilizado como una herramienta educativa para entrenadores, jugadores y analistas. Al entender mejor los patrones de juego, los equipos técnicos pueden proporcionar una educación de calidad a los jugadores, mejorando sus habilidades y conocimientos en el campo.
- También se establece un nexo con el **décimo** objetivo de reducción de desigualdades, promoviendo la identificación de talentos en regiones subrepresentadas, lo que contribuye a reducir las desigualdades en el acceso a oportunidades deportivas y a fomentar la diversidad en equipos y competiciones.
- Por último, en el marco del objetivo **decimoséptimo** de Alianzas para lograr los objetivos, la colaboración entre académicos, analistas de datos, entrenadores y jugadores es esencial para lograr avances significativos en el análisis deportivo. Este trabajo fomenta la colaboración entre diferentes actores interesados en el fútbol, con el objetivo de mejorar el rendimiento y la toma de decisiones.
- Otro objetivo que también guarda relación con esta investigación es el objetivo **tercero**, de salud y bienestar. Una herramienta de apoyo en la planificación deportiva se traduce en una mejor capacidad de organizar plantillas y que los clubes se aseguren de disponer con un número suficiente de jugadores de calidad para encontrar un balance de la carga de partidos de cada jugador y así poder prevenir lesiones.

A pesar de que, con el desarrollo actual no se puede hablar de una búsqueda de la igualdad de género, el trabajo de experimentación es fácilmente adaptable a los ensayos con nuevos conjuntos de datos. De poder obtener datos referentes tanto al fútbol masculino cómo al femenino, se contribuiría a la promoción de la igualdad de género a través del deporte. Con la disponibilidad

de datos actuales esto no ha sido posible, sin embargo, esta es una línea de investigación a futuro que merece la pena tener en cuenta.

En resumen, este TFG se alinea indirectamente con varios Objetivos de Desarrollo Sostenible a través de la mejora del rendimiento, la toma de decisiones informadas y la promoción de la igualdad, contribuyendo al avance de la sociedad desde el ámbito del deporte