Caracterización y partición de jugadores de ligas europeas buscando mejoras de entrenamiento, preparación y fichajes

Karina Durán Bautista código: **200824840**

Edna Julieth Mora código: **202124435**

David Castrillón Montaño código: **202120603**

Misael García Díaz código: **202124621**

RESUMEN

El presente trabajo es una muestra del conocimiento adquirido en las últimas semanas en la materia aprendizaje no supervisado, a través de las cuales inicialmente el equipo de trabajo ha hecho lluvia de ideas para definir 3 ideas posibles de abordar en este proyecto y luego de analizar las tres propuestas, seleccionamos:

"**Caracterización y partición de jugadores de ligas europeas buscando mejoras de entrenamiento, preparación y fichajes"**

Viabilidad de los datos, disponibilidad, y tiempo de ejecución son factores que se tienen cuenta para la definición. En ese sentido, se procede a revisión preliminar en la literatura, exploración y revisión de diferentes fuentes que poseen información valiosa para este documento, como GitHub y páginas web, conservando las más relevantes. Primer paso análisis exhaustivo exploratorio de datos, gestión de datos faltantes, variables redundantes. Paso dos, en esta sección se aplica aprendizaje no supervisado utilizando algoritmos de reducción de dimensión y clustering, una combinación de los dos. En razón al alto volumen de variables (400), por un lado, reducir las dimensiones de los datos asociados a rendimiento para identificar los componentes principales o valores singulares en los que el cuerpo técnico podría priorizar esfuerzos. Así como clusterizar jugadores por desempeño en partidos, desempeño de posición, desempeño de la liga, lesiones, etc. Clusterizar equipos rivales según datos históricos de encuentros de manera de poder identificar grupos de rivales que puedan ser tratados con estrategias distintas entre ellos.

Este algoritmo es preliminar y puede ser modificado para la entrega final, es posible que a lo largo del curso incorporemos nuevas herramientas que pueden resultar más apropiadas, para este caso.

Finalmente se hacen conclusiones de la metodología utilizada.

INTRODUCCIÓN

La analítica de datos es un campo que no es ajeno a la práctica deportiva y específicamente al Fútbol. El desempeño de los jugadores es seguido con variedad de métricas como la cantidad de minutos jugados, errores cometidos, jugadas exitosas y detalle de desempeño por partido, temporada incluso por minuto jugado [1,2,3,4]. Resultado del análisis de esta información podría dar perspectivas del desempeño individual de cada jugador, diferentes equipos o clubes [6]. Identificar grupos de jugadores dentro de la nómina, liga o incluso a nivel mundial, puede servir a los equipos para ajustar estrategias de entrenamiento, recuperación o alineación. De acuerdo con esto, se busca agrupar y/o clasificar jugadores europeos basados en sus atributos físicos y desempeño histórico, mediante algoritmos de aprendizaje no supervisado, tal que permita identificar agrupaciones que resulten aplicables en la toma de decisiones dentro e inter-equipos de fútbol.

REVISIÓN PRELIMINAR DE LA LITERATURA (ESTADO DEL ARTE)

En [6] se toman los datos de los partidos de fútbol y se desarrolla un modelo que clasifica las cualidades individuales de los futbolistas, detectando primero las posiciones de juego: delanteros, centrocampistas, extremos, laterales, centrales y porteros, mediante la agrupación de K-Means, con una precisión de 0.89 (para la Premier League 2021/2022 y 0.84 para la Allsvenskan 2021. En segundo lugar, se creó un modelo binario simplificado que solo clasifica el estilo de juego del jugador como ofensivo/defensivo, a partir de los malos resultados de este modelo se demostró que existen más de dos estilos de juego. Finalmente, se utilizó un enfoque de modelado no supervisado en el que se aplica el análisis PCA de manera iterativa. Para la posición de juego 'Delantero' se encontró 4 estilos de juego diferentes de juego. Como resultado de este trabajo, se encontró que es más fácil encontrar estilos de juego que tengan acciones claras y evidentes con el balón que los distinguen de otros jugadores dentro de su respectiva posición.

Otra propuesta similar e interesante a la que se busca implementar, esta vez a nivel nacional, se describe por Bonilla y otros en [1]. Allí se realizan una serie de mediciones antropométricas de atletas olímpicos jóvenes en el Urabá antioqueño de Colombia bajo condiciones controladas, logrando a partir de ellas generar agrupamientos con el método de K-medoides para identificar posibles formas eficientes de entrenar a los atletas, descubrir talentos, focalizar la habilidad o especialización de los atletas y de entender las características que los llevaron a agruparse en esos conjuntos específicos. Las variables que ellos miden y que les permiten caracterizar a los atletas están asociados a los tamaños de ciertos miembros del cuerpo, la cantidad de masa corporal, masa ósea, tejido adiposo, grosor de las diferentes capas de la piel, entre otras. Bonilla y los demás identifican dos fenotipos en el agrupamiento del algoritmo para los atletas, uno con contexturas ligeras, pequeñas y biomecánicamente más eficientes, pero con poca madurez ósea; el segundo grupo con contexturas gruesas, más altos y biomecánicamente menos eficientes y con una madurez ósea más avanzada. Recordemos que los atletas son jovencitos por lo tanto se considera como variable la madurez de sus cuerpos. Finalmente, al comparar los dos agrupamientos ellos encuentran que la madurez explica la mayoría de la varianza. Nuestra propuesta sigue líneas similares en la forma de analizar los datos y de realizar los agrupamientos, pero difieren en que ellos realizaron y prepararon las mediciones, nosotros tenemos los datos ya listos y sin que hayamos preparado o controlado su medición. por último, hay otra diferencia que radica en que nosotros realizaremos el agrupamiento para un rango de edades mucho mayor y no exactamente solo jóvenes, y aunque son deportistas no son atletas, sino futbolistas.

Una estrategia mucho más robusta y completa la realizan Tosato y Wu en [2], donde se establece un sistema de recomendación con aprendizaje no supervisado de agrupamiento conocido como Projective Adaptive Resonance Theory (PART), que aplica redes neuronales. Tosato y Wu aplican el sistema de recomendación al mercado de jugadores del equipo AS Roma de Italia para determinar atributos críticos y características al quedar agrupados de cierta manera. El algoritmo completo finalmente recomienda aspectos importantes de la formación del equipo, y de forma especial la compra o venta de jugadores entre los diferentes agrupamientos, teniendo en cuenta restricciones económicas porque hay variables de ese tipo dentro de la dimensionalidad del conjunto de datos. El algoritmo forma agrupamientos de baja dimensionalidad en la base de datos completos, es decir, agrupa en un subconjunto de atributos de la base de datos completa. Esta solución tiene implicaciones monetarias y de programación complejas y en nuestra aproximación no esperamos tener un nivel de recomendación tan sofisticado, ni usar el mismo algoritmo, pero si poder realizar un agrupamiento que permita caracterizar y analizar donde estaría mejor un jugador o si un jugador puede tener diferentes posiciones en cual tendría un mejor desempeño.

DESCRIPCIÓN DETALLADA DE LOS DATOS

Los datos seleccionados corresponden a la evaluación de los jugadores de las ligas europeas de futbol durante la temporada 2019 – 2020, realizada por la compañía Transfermarket [3], firma especializada en seguimiento estadístico y valoración de jugadores según su rendimiento. Obtenidos del conjunto de dataset públicos de Kaggle [4] página especializada en competencias de ciencias de datos.

El dataset cuenta con 2644 observaciones y 400 variables 9 de ellas categóricas (Información general del jugador), y el resto numéricas (Estadísticas del jugador).

En la tabla 1, se presenta la descripción de las primeras 6 columnas del dataset como ejemplo del análisis preliminar realizado, en general solo una columna presenta datos vacíos (CLBestScorer) que indica si el jugador ha sido goleador en el torneo de champions league.

	games float64	games_starts f…	minutes float64	goals float64	assists float64	pens_made floa…
count	2644.0	2644.0	2644.0	2644.0	2644.0	2644.0
mean	18.476172465960666	14.329803328290469	1285.9145234493192	1.7692889561270801	1.2310892586989 41	0.1785173978819 9697
std	10.939718892104723	10.927726860555 53	947.34688343241 56	3.3383579003406 94	2.0113003177530007	0.8122547450096267
min	1.0	0.0	1.0	0.0	0.0	0.0
25%	9.0	4.0	424.5	0.0	0.0	0.0
50%	19.0	13.0	1181.5	0.0	0.0	0.0
75%	28.0	23.0	2050.25	2.0	2.0	0.0
max	38.0	38.0	3420.0	36.0	21.0	14.0

Tabla 1. Ejemplo estadísticas descriptivas del set de datos

Dentro de las variables numéricas se encuentran las estadísticas de cantidad de pases, efectividad de pases, partidos jugados, etc. Se encuentran algunas variables que son resultado de operaciones entre otras variables como porcentaje de efectividad del pase que se construye de las variables: cantidad de pases realizados y cantidad de pases completados. Esto nos indica la probabilidad de encontrar correlaciones entre los datos y la necesidad de ejecutar modelos de reducción de dimensiones para el manejo eficiente de los datos. Evidencia de esto es la Figura 1, donde se presentan las correlaciones más fuertes entre las variables.

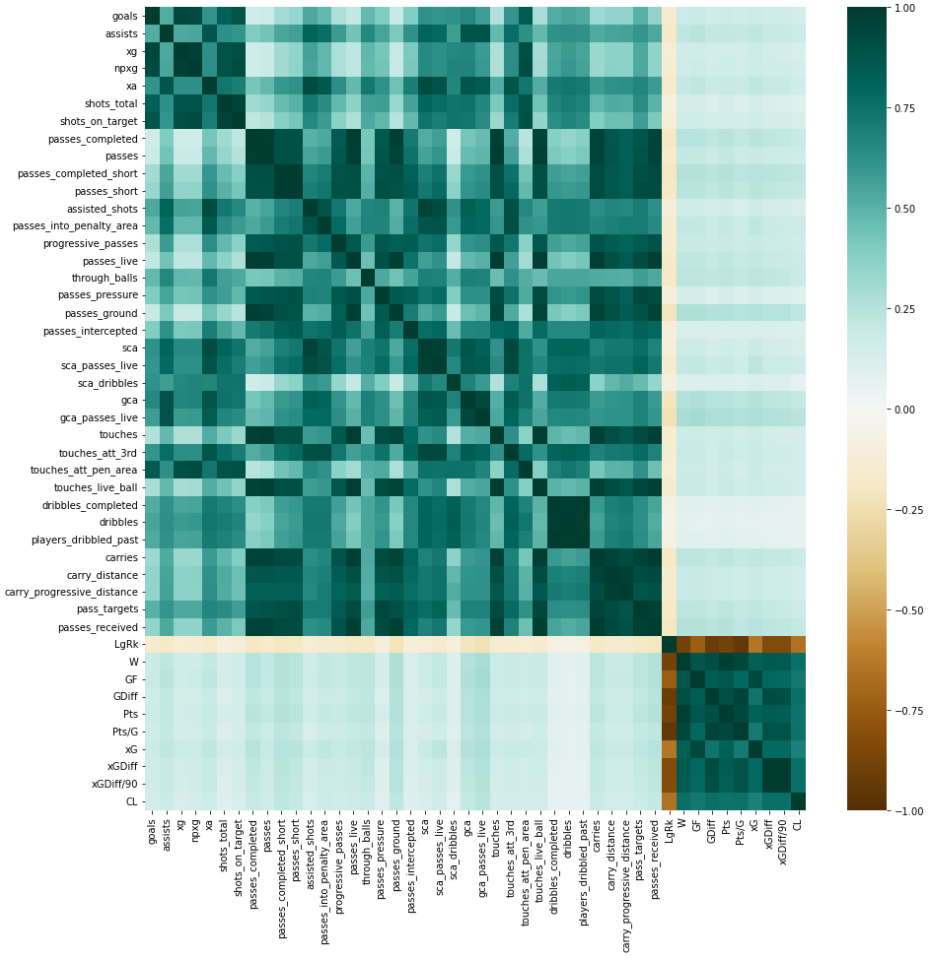

Figura 1. Mapa de calor entre variables con correlación mayor a 0.4

Los grandes conjuntos de datos son cada vez más comunes y, a menudo, difíciles de interpretar. Para hacerlo, se requieren métodos que permitan reducir drásticamente su dimensionalidad de una manera interpretable, de modo que se conserve la mayor parte de la información de los datos.

El análisis de componentes principales, o PCA por sus siglas en inglés, es una técnica de aprendizaje no supervisado que permite reducir la dimensionalidad de tales conjuntos de datos, aumentando la interpretabilidad, pero al mismo tiempo minimizando la pérdida de información. Se han desarrollado muchas técnicas para este propósito, pero el análisis de componentes principales es uno de los más antiguos y más utilizados. Su idea es simple: reducir la dimensionalidad de un conjunto de datos, mientras se preserva la mayor "variabilidad" posible.[5]

En un acercamiento a esta situación se realiza un PCA inicial con todo el conjunto de datos estandarizado identificando que con 70 componentes podría explicarse hasta un 90% de la variabilidad.

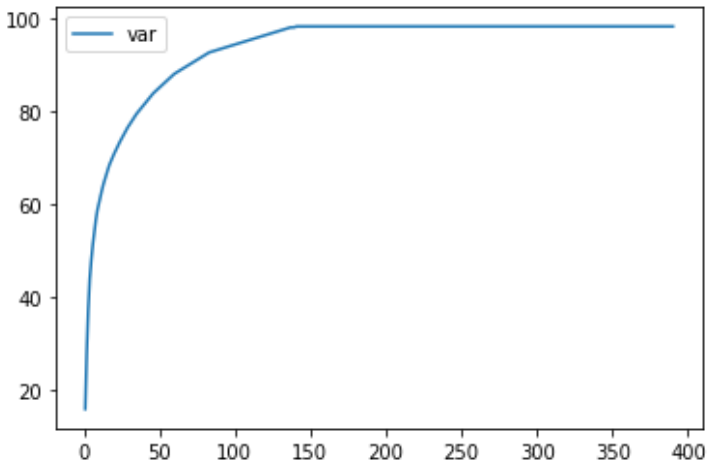

Figura 2. Explicación de la varianza por componentes calculados

Este acercamiento nos confirma la posibilidad del uso de algoritmos de aprendizaje no supervisado para reducir el dimensionamiento de los datos y teniendo como primer hallazgo una posible agrupación de variables importantes para el análisis de jugadores.

PROPUESTA METODOLÓGICA

En esta sección se aplica aprendizaje no supervisado utilizando algoritmos de reducción de dimensión y clustering, una combinación de los dos. En razón al alto volumen de variables (400), por un lado, reducir las dimensiones de los datos asociados a rendimiento para identificar los componentes principales o valores singulares en los que el cuerpo técnico podría priorizar esfuerzos. Así como clusterizar jugadores por desempeño en partidos, desempeño de posición, desempeño de la liga, lesiones, etc. Clusterizar equipos rivales según datos históricos de encuentros de manera de poder identificar grupos de rivales que puedan ser tratados con estrategias distintas entre ellos.

Para iniciar vamos a utilizar un clúster jerárquico, son excelentes estructuras para organizar los datos, y nos muestran la relación anidada entre puntos. Así mismo nos muestra la relación de elementos. También la caracterización de los elementos. Otra ventaja es que no es necesario a priori elegir el número de clústeres, sin embargo, será necesario especificar enlaces y distancia.

Como sabemos que el clúster jerarquizado no es tan fuerte en datos atípicos u outliers como alternativa vamos a utilizar un método más robusto, que los tenga en cuenta y no altere los resultados.

Como un segundo método para trabajar y comparar con el rendimiento del clúster jerárquico es el algoritmo DBSCAN con el ánimo de encontrar datos atípicos dentro del conjunto de jugadores y sus diversas habilidades. El algoritmo DBSCAN incorpora densidad, tiene ventajas frente a clúster jerarquizado que tiene otro enfoque, frente a al manejo de ruido u outliers. Con la densidad que incorpora DBSCAN dejamos de lado los datos atípicos, quedan fuera de los clústeres, quedando como datos atípicos no agrupados. Para elaborarlo vamos a utilizar los componentes necesarios del algoritmo como n_sample, eps

Los datos y técnicas deberían extraer la información de las estadísticas de los jugadores y permitir finalmente una herramienta adicional para técnicos, entrenadores y negociadores del fútbol.

BIBLIOGRAFÍA

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

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

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

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

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

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