



**VNiVERSiDAD
D SALAMANCA**
CAMPUS DE EXCELENCIA INTERNACIONAL

TRABAJO FIN DE GRADO

GRADO EN ESTADÍSTICA

FACULTAD DE CIENCIAS

CURSO 2023/24

EXPLOTACIÓN ESTADÍSTICA DE ALMACENES DE DATOS EN LAS CINCO GRANDES LIGAS EUROPEAS

AUTOR

LUCIO ALEJANDRO MOUZO MONTEAGUDO

TUTORA

MARÍA TERESA CABERO MORÁN

Salamanca, 2024



**VNiVERSiDAD
D SALAMANCA**
CAMPUS DE EXCELENCIA INTERNACIONAL

TRABAJO FIN DE GRADO

GRADO EN ESTADÍSTICA

FACULTAD DE CIENCIAS

CURSO 2023/24

EXPLOTACIÓN ESTADÍSTICA DE ALMACENES DE DATOS EN LAS CINCO GRANDES LIGAS EUROPEAS

AUTOR

LUCIO ALEJANDRO MOUZO MONTEAGUDO

TUTORA

MARÍA TERESA CABERO MORÁN

Firmado por MOUZO
MONTEAGUDO, LUCIO
ALEJANDRO (FIRMA) el
día 11/01/2024 con un
certificado emitido
por AC DNIE 005

Firmado por CABERO MORAN MARIA
TERESA - ***6717** el día
11/01/2024 con un certificado
emitido por AC FNMT Usuarios

Salamanca, 2024



Certificado de la tutora de TFG

Dña. **María Teresa Cabero Morán**, profesora del Departamento de Estadística de la Universidad de Salamanca,

HACEN CONSTAR:

Que el trabajo titulado “EXPLOTACIÓN ESTADÍSTICA DE ALMACENES DE DATOS EN LAS CINCO GRANDES LIGAS EUROPEAS”, que se presenta, ha sido realizado por **Lucio Alejandro Mouzo Monteagudo**, con DNI ****7946 y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, 10 de enero de 2024

Firmado por CABERO MORAN MARIA TERESA -
***6717** el día 11/01/2024 con un certificado
emitido por AC FNMT Usuarios

Fdo.: María Teresa Cabero Morán

CONTENIDOS

1. INTRODUCCIÓN.....	1
1.1. OBJETIVOS	2
1.2. ESTRUCTURA DE TRABAJO.....	2
2. EL FÚTBOL Y LA EVOLUCIÓN DESDE SUS INICIOS EN LOS GRANDES PAÍSES EUROPEOS ...	4
2.1. EL FÚTBOL Y SU EVOLUCIÓN	4
2.2. LA EVOLUCIÓN DEL FÚTBOL EN ESPAÑA.....	6
3. EXPLOTACIÓN ESTADÍSTICA Y ALMACENES DE DATOS	8
3.1. BASES DE DATOS.....	8
3.2. ALMACÉN DE DATOS	8
3.3. MINERÍA (TÉCNICAS Y ALGORITMOS).....	9
4. TÉCNICAS UTILIZADAS.....	11
4.1. ÁRBOL DE CLASIFICACIÓN	11
4.2. ANÁLISIS DE COMPONENTES PRINCIPALES.....	14
4.3. ANÁLISIS FACTORIAL.....	17
5. METODOLOGÍA	21
5.1. BASE DE DATOS	21
5.2. <i>SOFTWARE R</i>	24
5.3. <i>SOFTWARE SPSS</i>	24
5.4. <i>SOFTWARE EXCEL</i>	25
6. ANÁLISIS ESTADÍSTICO	26
6.1. ESTADÍSTICOS DESCRIPTIVOS	26
6.2. TÉCNICAS ESTADÍSTICAS.....	28
6.2.1. ÁRBOL DE CLASIFICACIÓN	29
6.2.2. ANÁLISIS DE COMPONENTES PRINCIPALES.....	37
6.2.3. ANÁLISIS FACTORIAL	40
7. CONCLUSIONES	44
8. BIBLIOGRAFÍA.....	46
8.1. LIBROS Y ARTÍCULOS.....	46
8.2. PÁGINAS WEB Y PDF.....	46

1. INTRODUCCIÓN

El mundo del fútbol ha sufrido una evolución muy grande en los últimos años gracias a la introducción de la estadística. Actualmente todos o prácticamente todos los partidos que se juegan a nivel profesional o semiprofesional cuentan con un análisis en directo de cada uno de los jugadores y de lo que ocurre en el terreno de juego a cada instante.

Esta introducción del mundo de la estadística lo que ofrece es un amplio enfoque para la mejora en muchos ámbitos del deporte, pudiendo ser en la forma física de los jugadores o en el trabajo que se debe aplicar para poder llegar a unas condiciones óptimas a todos los encuentros a través de la medición de las cargas de trabajo. También se utiliza para la prevención de lesiones, al poder medir mediante los chalecos de trabajo y los distintos tipos de medidores, tanto las cargas como las intensidades de los jugadores a lo largo de la semana.

Otra de las aplicaciones que tiene la estadística en el fútbol es en el *scouting*, que utiliza esta ciencia para realizar fichajes y en las comparaciones entre jugadores para poder encontrar y fichar a futbolistas que cumplan unas características necesarias dentro del estilo de juego del equipo o que sean similares a jugadores que rinden actualmente en el club pero que finalizan contrato, que tienen una edad alta o que van a ser vendidos, dando paso a la compra de otro que supla su puesto.

Este deporte no ha sido de los primeros en incorporar el mundo de la estadística, si no que ha seguido los pasos de otros deportes como el béisbol o el baloncesto que han sido de los primeros que la han incorporado ya en el siglo pasado.

El fútbol es uno de los deportes mayoritarios en gran parte del mundo llegando a ser influyente en la vida de muchas personas. Es un deporte que contiene una gran cantidad de jugadores y aficionados, ya que muchos niños lo realizan desde una temprana edad y son aficionados de equipos profesionales o de alto rendimiento. Mientras que también hay que tener en cuenta gente adulta, es decir, mayor de los 18 años, que a pesar de terminar su etapa en el fútbol base sigue jugando por afición y sin contraprestación económica de ninguna índole por parte del deporte.

Por otra parte, se abordarán las cinco grandes ligas europeas ya que son las que cuentan con jugadores de un mayor valor de mercado. Y se seguirá con el análisis estadístico de los jugadores y de los datos que hay en la base creada para realizar el trabajo; y al analizarlo se obtienen los principales estadísticos que explican que contiene dicha base y como se distribuyen los deportistas de estas ligas por sus distintas cualidades o características, que se ven reflejadas en los números que han realizado mediante sus actuaciones o acciones que han tenido lugar en el terreno de juego.

De seguido se llevarán a cabo las técnicas multidimensionales, que en este caso son tres, tratándose del árbol de clasificación, el análisis factorial y el análisis de componentes principales. Cada una de estas técnicas tienen su propio objetivo, siendo el objetivo de dos de ellas similar, el del análisis factorial y el del análisis de componentes principales, ya que ambas consisten en la reducción de la dimensionalidad creando o formando unas nuevas variables latentes denominadas factores o componentes principales, respectivamente. La otra técnica empleada es el árbol de clasificación cuyo objetivo es distinto al de las anteriores, el cual consiste en la clasificación de las variables mediante la significación de éstas, pasando a llamarse variables predictoras a las que dividen el árbol de un nodo a otros.

Con la aplicación de estas técnicas se quiere enseñar que variables son más significativas entre todos los jugadores, siendo observados respecto sus ligas; además de las relaciones que tienen las variables entre ellas, las relaciones respecto de los componentes o de los factores que se obtengan y la explicación de las variables latentes que se han obtenido.

1.1. OBJETIVOS

El objetivo principal del trabajo es realizar un análisis de las estadísticas de los jugadores de fútbol de las cinco grandes ligas europeas, que son la Premier League, LaLiga, la Serie A, la Ligue 1 y la Bundesliga, en la temporada 2021-2022, intentando realizar una comparación entre las ligas y las posiciones de los jugadores.

Este objetivo principal se puede sintetizar en varios objetivos específicos, que son:

- Revisar bibliográficamente publicaciones relacionadas con el fútbol y con la evolución de la estadística en este deporte.
- Crear una base de datos a través del *software* Excel.
- Desarrollar una base de datos en la que se puedan realizar estadísticos y gráficos mediante dos *softwares*, R y SPSS.
- Observar cuáles son las variables significativas en el árbol de clasificación respecto las ligas que se tienen.
- Observar los grupos complejos que se forman del árbol de clasificación.
- Realizar un análisis de componentes principales para unificar las estadísticas de los jugadores en un mínimo de componentes y explicar a que se pueden referir.
- Realizar un análisis factorial para unir las estadísticas de los jugadores en un número de factores mínimo y explicar estos factores.

1.2. ESTRUCTURA DE TRABAJO

El trabajo está estructurado en 8 capítulos en los que se habla tanto del fútbol, que es el tema a tratar en la base de datos, como de las técnicas y *softwares* que se utilizan para poder trabajar sobre la base y comprobar los objetivos iniciales.

El primero es la introducción, en la cual se abordan los objetivos que se buscan obtener de la explotación de la base de datos y la estructura con la que cuenta el trabajo de fin de grado.

El segundo es el fútbol y la evolución que ha sufrido desde sus inicios en los grandes países europeos. Hablando de en qué consiste el deporte y en cómo fueron sus inicios hace unos siglos, y como ha ido evolucionado a lo largo de los años en los distintos países del mundo, y más concretamente en España.

El tercer capítulo es la explotación estadística y los almacenes de datos, en el cual se explica lo que se va a realizar en la base, y se define la explotación estadística de datos, la base de datos, los almacenes y la minería; así como una explicación de las técnicas y algoritmos más conocidos o importantes.

El cuarto capítulo trata sobre las técnicas que se utilizan para intentar lograr los objetivos. Estas tres técnicas se pueden dividir en dos grupos, una técnica de clasificación como es el

árbol de clasificación y dos técnicas de reducción de la dimensionalidad como son el análisis de componentes principales (ACP) y el análisis factorial (AF).

El quinto trata sobre la base que se ha creado y en la que se van a realizar las técnicas estadísticas mencionadas en el punto anterior. En este punto se describe detalladamente la base de datos, con que variables cuenta y como ha sido recogida. También se abordan los tres *softwares* utilizados, de los que se explica su procedencia y su evolución.

El sexto punto es la interpretación de los datos, tanto de manera descriptiva como de manera estadística. En éste se realiza una matriz de correlaciones con la que se lleva a cabo la decisión de los descriptivos que se han de realizar con variables con una relación alta o significativa. Y de seguido se realizan los tres métodos para poder obtener respuestas a los objetivos marcados al inicio del trabajo.

El séptimo capítulo que tiene el trabajo cuenta con las conclusiones que se han obtenido después de la realización de los descriptivos y de las técnicas multivariantes con las que se ha analizado la base de datos, comprobando de este modo si se han cumplido los objetivos marcados.

El octavo punto, y último, es la bibliografía, en éste se encuentran los libros y páginas web que se han utilizado para poder llevar a cabo el trabajo.

2. EL FÚTBOL Y LA EVOLUCIÓN DESDE SUS INICIOS EN LOS GRANDES PAÍSES EUROPEOS

En este punto se trata el deporte en cuestión que es el fútbol y la evolución que ha tenido este deporte a lo largo de la historia, y en España a lo largo de los años. El primer punto trata de en qué consiste el deporte y como se juega, seguido de su evolución a lo largo de los años empezando hace varios siglos hasta la actualidad. En el segundo punto se habla de cómo ha evolucionado el fútbol en España desde finales del siglo XIX hasta el momento actual, tratando momentos importantes o que han podido influir en algún momento de la historia del deporte.

2.1. EL FÚTBOL Y SU EVOLUCIÓN

El fútbol es un deporte colectivo que se juega entre dos equipos, que están conformados por once jugadores cada uno; siendo el objetivo de estos marcar goles en la portería del equipo rival. Los partidos se llevan a cabo en un campo rectangular, que consta con dos porterías, una de cada equipo, en los extremos opuestos del campo. Al fútbol se juega principalmente con los pies, pero también se utilizan otras partes del cuerpo, excluyendo las manos y brazos, estas partes solo podrán ser usadas por los porteros dentro de sus respectivas áreas.

El objetivo del deporte es anotar goles, que se llevan a cabo cuando el balón cruza la línea de meta rival. El equipo que gana el partido es el que al finalizar éste lleve más goles marcados. Este deporte es uno de los más practicados a nivel mundial y uno de los más populares en audiencia, contando con una gran cantidad de seguidores y una rica historia.

El fútbol es uno de los deportes más jugados, llegando a jugarse desde los niveles más bajos que podrían ser en las pistas de los parques, hasta los más altos, pudiendo mencionar como el partido más importante la final de la Copa del Mundo. El fútbol tiene establecidas unas reglas que se siguen a nivel mundial, creadas por la federación más importante del deporte que es la FIFA (*Fédération Internationale de Football Association*), donde se determinan tanto el tamaño del campo, el tiempo del partido, las faltas y sanciones, y más acciones que están establecidas en las reglas.

El fútbol ha sido y es muy importante en la sociedad, promoviendo la competitividad, la estrategia, el trabajo en equipo y el acondicionamiento físico, y también tiene un papel muy importante en los comportamientos de la sociedad, en el entretenimiento y en la cultura en muchas partes del mundo, donde llega a formar parte de la vida de las personas o una dependencia tanto moral como de afición por un club o por el deporte mismamente.

Lo siguiente que se abordará es la historia de este deporte que es muy extensa y que comienza hace muchos siglos y que ha ido evolucionando desde ese inicio hasta el día de hoy, y que seguramente siga evolucionando y variando en los próximos años.

Los inicios del fútbol se cree que se encuentran en las civilizaciones antiguas, como en la civilización china, la griega y la romana, donde realizaban juegos similares en los que incluían pelotas infladas que se pateaban con los pies. En la Edad Media, en toda Europa, se jugaban diversas actividades de balón con reglas variables dependiendo del lugar. Estos juegos eran comunes en las calles y se jugaban a ellos en equipos, pero con una estandarización muy pobre de las reglas, siendo éstas muy pocas.

A pesar de todo, el fútbol moderno no comenzó a tomar forma hasta el siglo XIX en Inglaterra. Durante esta época, aparecieron distintas reglas locales, como las Reglas de Cambridge en

1848 y las Reglas de Sheffield en 1857, que establecieron los fundamentos del juego. Unos años más adelante, en 1863 tuvo lugar la fundación de la *Football Association* (FA), es decir, la Asociación de Fútbol, en Inglaterra, siendo un momento importante en la historia del deporte ya que establece un conjunto de reglas comunes y un sistema de gobierno, haciendo evolucionar el deporte a un fútbol moderno. La FA se convirtió en la federación que mandaba en el fútbol inglés y un tiempo más adelante se convirtió en la mandataria en todo el mundo del fútbol.

A lo largo del siglo XIX, el fútbol llegó a otros países, principalmente a los más cercanos a Inglaterra, que son los que fundaron las siguientes asociaciones nacionales, como ocurrió en Escocia en 1873, Gales en 1876 e Irlanda en 1880. Pero a mayores que en estos países vecinos de Inglaterra, donde comenzó a ganar una gran popularidad llegando a convertirse en un elemento importante en la sociedad y en la cultura fue en América del Sur, principalmente en los tres países más grandes de Sudamérica en este deporte que son Argentina, Uruguay y Brasil.

Uno de los momentos más relevantes en la historia del fútbol ha sido la creación de la Copa del Mundo. Esta competición une a países de todo el mundo y fue en aumento desde su inicio en 1930 en Uruguay, gracias a este torneo el deporte también empezó a expandirse por más países. Desde 1930 hasta ahora la Copa del Mundo ha incrementado su importancia convirtiéndose en uno de los eventos deportivos con mayor cantidad de seguidores de todo el mundo.

En el siglo XX, el fútbol fue experimentando un gran cambio hacia la profesionalización en muchos países, llevándose a cabo la creación tanto de clubes profesionales como de las ligas en las que jugarían éstos. Esta rápida profesionalización incrementó la popularidad del deporte en todo el mundo y lo impulsó a una mayor competencia para llegar a lo más alto.

El fútbol siguió evolucionando año tras año, con los cambios de algunas reglas y la introducción de las tecnologías que han ido apareciendo a lo largo del tiempo para mejorar la justicia en el juego y también para mejorar las condiciones de los terrenos y las condiciones físicas de los jugadores. Uno de los casos del uso de nuevas tecnologías en el fútbol moderno es la aparición de reglas como la del fuera de juego, aunque ahora sea común, hace muchos años no se contaba con ella, y también el uso del VAR (*Video Assistant Referee*), que es también conocido como video arbitraje y consiste en que un grupo de árbitros que ven el partido desde una sala, le comuniquen en jugadas polémicas o en las que hubo un fallo arbitral que se puede corregir, al árbitro central y éste revise la jugada en el monitor del campo, modificando o no su primera decisión arbitral.

Actualmente, este deporte es mucho más que un simple juego, el fútbol ha evolucionado socialmente hasta el punto de convertirse en una parte integral de la cultura a nivel mundial, con una gran cantidad de aficionados apasionados y una industria multimillonaria detrás de cada club o de cada campeonato. Esta industria es tan grande que abarca desde los contratos millonarios de jugadores con renombre mundial hasta los patrocinios, sin dejar atrás a los derechos televisivos y la gran cantidad de dinero que estos mueven. El fútbol ahora mismo tiene un gran poder para unificar a las personas que va mucho más de frontera y culturas, juntando a personas de cualquier parte del mundo alrededor de un amor o una pasión que comparten entre ellas por el deporte más popular del planeta.

2.2. LA EVOLUCIÓN DEL FÚTBOL EN ESPAÑA

La evolución del fútbol a nivel español ha sido una larga historia donde el deporte ha crecido desde unos humildes comienzos hasta convertirse en uno de los países más importantes en el ámbito futbolístico a nivel europeo y mundial.

El fútbol llegó a España a finales del siglo XIX, a través de la inmigración de trabajadores y estudiantes británicos a España. En estos años se fundaría el primer club del país, que fue el Real Club Recreativo de Huelva fundado en 1889. La presencia inglesa influyó tanto en el uso y aprendizaje de las reglas como en la creación de equipos locales y de clubes, que hoy en día muchos aún siguen existiendo.

Ya en el nuevo siglo, y con muchos más equipos fundados, se creó también la Real Federación Española de Fútbol (RFEF) en 1913, convirtiéndose en el organismo rector del fútbol español. La Real Federación Española de Fútbol fue muy importante para la estandarización y regulación del fútbol en el país.

El siguiente momento importante que se vivió en el fútbol español en el siglo XX ocurrió en 1929, año en el que se fundó la Primera División de España convirtiéndose en una de las ligas más influyentes del mundo. En estas primeras ediciones ya se podían encontrar equipos como el FC Barcelona, el Atlético de Madrid, a pesar de que sufrió modificaciones en el nombre en algunos años, el Athletic Club de Bilbao y uno de los clubes más laureados del mundo, el Real Madrid. Estos clubes con el paso de los años han logrado convertirse en potencias futbolísticas, protagonizando unas grandes rivalidades hasta la actualidad.

El país sufrió una Guerra Civil en 1936, que desembocó en una dictadura de Francisco Franco desde 1939 hasta 1975, en la que el fútbol fue utilizado como un elemento para hacer propaganda política. Los equipos que representaban de algún modo a regiones sufrieron dificultades esos años, como fueron el caso del Athletic Club de Bilbao, que representaba al nacionalismo vasco, y el FC Barcelona, que representaba la identidad catalana. Pero también hubo clubes que sirvieron como propaganda política para el régimen franquista, como fue el Real Madrid, que fue considerado un símbolo y siendo uno de los clubes más laureados en esos años.

Finalizada la dictadura y con el cambio al siglo XXI, se puede hablar de la gran época dorada del fútbol español. Esto sucedió principalmente en cinco años en los que la selección española atravesó su mejor momento, logrando ser campeona de la Eurocopa de 2008 con Luis Aragonés como entrenador, siendo uno de los máximos exponentes del fútbol español tanto como jugador, como de entrenador. Siguió con el logro de la competición más importante del planeta, como es la Copa del Mundo, que ganó la selección en 2010 con Vicente del Bosque como entrenador; y se cerraría este ciclo histórico con la consagración como campeona de la Eurocopa en 2012, todavía con Vicente del Bosque en el banquillo.

En las últimas décadas se deben destacar a dos equipos que han sido los más laureados a nivel español y a nivel europeo, como han sido el Real Madrid y el FC Barcelona. Pero a quien se debe mencionar por encima de todos los demás grandes jugadores que han estado en estos clubes, como pueden ser Xavi, Iniesta, Sergio Ramos e Iker Casillas, estos cuatro ganadores del Mundial con la selección española en 2010, hay que mencionar a una de las mejores rivalidades que se ha vivido en la historia del fútbol como ha sido la de Cristiano Ronaldo, considerado uno de los mejores de la historia, y Lionel Messi, el que por muchos es considerado el mejor jugador de todos los tiempos, peleando en ese pódium con Diego Maradona y con Pelé.

También ha crecido de manera notoria el fútbol femenino en España en estas últimas décadas. La liga española femenina cada vez es más competitiva, como se pudo observar con la consagración del FC Barcelona como campeonas de la Champions League. A raíz de esta mejoría se han obtenidos éxitos a nivel de selección en torneos internacionales, consagrándose la selección como campeona de la Copa del Mundo en el año 2023.

Desde un inicio hasta hoy en día, algo que no ha cambiado y sigue vigente es la influencia del fútbol en la cultura y en la economía española, ha crecido año tras año, pudiendo ver actualmente estadios con aforos de más de cincuenta mil espectadores y que estén llenos, o el caso del Deportivo de la Coruña, que a pesar de estar jugando en 1ªRFEF sigue teniendo una cantidad de socios más común de Primera división, teniendo incluso más socios que asientos en el estadio. Esta influencia cultural hace que se mantenga las grandes rivalidades tanto en las grandes ligas como en las más bajas. Ahora mismo es muy difícil pensar en una sociedad española sin fútbol, y este está aportando crecimiento al país en el ámbito económico año tras año, gracias a la atracción de inversores y patrocinadores internacionales.

3. EXPLOTACIÓN ESTADÍSTICA Y ALMACENES DE DATOS

La explotación estadística es el proceso de analizar, resumir y presentar datos como información que resulta significativa, es decir, es la utilización de técnicas estadísticas para conseguir unas respuestas que pueden ayudar en la toma de decisiones y a la hora de hacer cualquier tipo de predicción.

La explotación tiene un proceso de elaboración para poder llevarla a cabo. Éste tiene varios pasos: el primero de ellos sería la recopilación de datos, cuando se tenga la base de datos con todos los datos recogidos, se pasa al siguiente paso que es la limpieza de los datos que se tienen. Este segundo paso consiste en corroborar y cambiar o eliminar los datos que cuenten con una falta de información, con datos erróneos o valores que no estén dentro de los parámetros que se quieren marcar. El objetivo de éste es conseguir una base en la que todos los datos sean coherentes.

El tercer y cuarto paso podría juntarse en uno solo, pero dividiéndolo en dos se podrá hacer referencia al distintivo que hay entre los dos tipos de análisis. El análisis descriptivo es el que transmite la información que contiene la base de datos y donde se puede observar que es importante o llama la atención de cada una de las variables, ya sea al realizar tablas cruzadas con otras variables o con la observación de una única variable a través de un diagrama de caja, donde poder apreciar los datos anómalos o donde se concentra la masa más grande de datos. El otro tipo de análisis que hay es el exploratorio, que es en el que se llevan a cabo las distintas técnicas estadísticas. Éstas se pueden agrupar por su función o su objetivo en distintos grupos, siendo los principales tipos: técnicas de clasificación, de predicción, de reducción de dimensiones, de comparación de medias y de regresión, entre otros.

El siguiente paso que se hará es el de interpretación de los resultados que se han obtenido en los puntos anteriores. Se tendrá que realizar una explicación estadística de los resultados que se obtengan tanto en la realización de los análisis descriptivos como de los exploratorios.

Y por último sacaremos las conclusiones pertinentes de las interpretaciones que se lleven a cabo, observando de esta forma si se cumplen los objetivos marcados inicialmente y si se ha averiguado lo que se estaba buscando.

3.1. BASES DE DATOS

Una base de datos consiste en un conjunto de datos que se encuentran almacenados de manera organizada y estructurada de cierta forma con la finalidad de permitir el acceso, la gestión y la actualización de la manera más eficiente. Éstas son imprescindibles en la informática y en la estadística y se usan para una gran cantidad de sistemas y aplicaciones en distintas áreas de trabajo, como pueden ser empresas o aplicaciones web.

Las bases de datos son esenciales en el almacenamiento de información en forma de datos en diversos tipos de aplicaciones y sectores, siendo una herramienta muy importante para organizar y acceder a información de manera inmediata.

3.2. ALMACÉN DE DATOS

El almacén de datos es un sistema de almacenamiento de datos creado para almacenar, gestionar y analizar cantidades grandes de información. El principal objetivo es proporcionar

al usuario un acceso rápido y eficiente tanto a datos históricos como actuales, permitiendo a las organizaciones realizar la toma de decisiones de manera estratégica.

El almacén de datos tiene un papel crítico en el análisis de datos y en el sector empresarial, dado que proporciona a las organizaciones una ayuda con una única verdad para tomar decisiones basadas en los datos. Al tener una fácil y veloz accesibilidad a grandes cantidades de información, éstos impulsan a las empresas a comprender mejor el rendimiento que tienen, identificar las oportunidades y las tendencias, y realizar una planificación de futuras estrategias.

3.3. MINERÍA (TÉCNICAS Y ALGORITMOS)

La minería de datos es un campo de la estadística que tiene como objetivo el descubrimiento de patrones, tendencias y resultados ocultos en grandes conjuntos de datos. Para conseguir este objetivo, se utilizan distintas técnicas y algoritmos con la intención de obtener la solución óptima al problema.

Las técnicas de datos más comunes son: las asociativas, las de agrupamiento, las de clasificación, las de regresión y las de reducción de dimensionalidad.

Las técnicas asociativas son las que se utilizan para encontrar relaciones entre las variables de un conjunto de datos o de una base de datos. El algoritmo que más se conoce de esta técnica es el algoritmo *a priori*, que habitualmente se usa en la minería de los patrones de compra de los usuarios de supermercados o tiendas.

Las técnicas de agrupamiento o *clustering* se utilizan para separar un conjunto de datos en *cluster* o en grupos, donde los elementos que se encuentran dentro de cada uno de los *cluster* son más parecidos entre sí que con los elementos que conforman los otros *cluster*. El algoritmo para la agrupación de datos más utilizado es el *k-means* o k-medias.

Las técnicas de clasificación consisten en la asignación de categorías o etiquetas a los datos del conjunto o de la base. Los algoritmos de clasificación tienen un uso común en aplicaciones como pueden ser la detección de fraudes o el filtrado de spam de correo electrónico. Los algoritmos de clasificación que más se utilizan son el árbol de clasificación y el *naive bayes*.

Las técnicas de regresión se utilizan para poder predecir valores numéricos en función de las variables independientes de la base de datos que se tenga. Éstas, por ejemplo, pueden ser útiles para predecir el precio de un jugador de fútbol en función de sus características. En este caso los algoritmos más utilizados son los que incluyen la regresión logística y la regresión lineal.

Por último, las técnicas de reducción de la dimensionalidad consisten en la reducción del número de variables o atributos de la base de datos a un mínimo de variables latentes que se intentan agrupar, que pueden ser factores o componentes principales; facilitando de esta forma un posible análisis al reducir el número de variables con las que se van a trabajar, disminuyendo, de este modo, la dimensionalidad. Las dos técnicas más populares de reducción de dimensionalidad son el análisis de componentes principales (ACP) y el análisis factorial (AF).

Los algoritmos de minería de datos que más se utilizan de las técnicas que se mencionaron antes son los que se mencionan a continuación:

Los árboles de clasificación forman parte de las técnicas de clasificación, su utilización principal es la de clasificar y tomar decisiones basadas en reglas jerárquicas; son comunes ya que es un algoritmo fácil de entender, interpretar y explicar a terceras personas.

El algoritmo *random forest* consiste en un conjunto de árboles de clasificación que se llevan a cabo con el objetivo de mejorar la precisión en los posibles problemas de clasificación y regresión que se puedan encontrar en la utilización de las técnicas que se mencionaron con anterioridad.

El algoritmo *naive bayes* se utiliza en las técnicas de clasificación y está basado principalmente en el teorema de Bayes. Este algoritmo tiene una gran eficacia en cuestiones que consisten en la resolución de problemas, de clasificación de texto y de filtrado de spam.

La regresión lineal es un algoritmo que como su nombre indica forma parte de los que se utilizan en las técnicas de regresión. Éste se emplea para la predicción de valores numéricos y para establecer relaciones entre las variables dependientes y las variables independientes.

El algoritmo *k-means* o k-medias es de tipo *clustering* o de agrupamiento que consiste en la asignación de puntos a los datos, formando grupos que están basados en la similitud de sus atributos; es utilizado principalmente en la segmentación de clientes por las empresas y en la organización de datos.

El algoritmo *a priori* forma parte de las técnicas de asociación, por lo que su utilización principal es la minería de patrones asociativos con la intención de descubrir las relaciones que hay entre los elementos de los conjuntos de datos; un ejemplo de un posible uso son los patrones de compra.

El algoritmo SVM o máquina del vector de soporte pertenece a las técnicas de clasificación y su utilización se basa en problemas de clasificación y regresión con el objetivo de encontrar un hiperplano de separación óptimo entre las diferentes clases de datos que haya dentro de la base de datos.

Estas son algunas de las técnicas más utilizadas y los algoritmos que se utilizan con mayor frecuencia en la minería de datos. La elección de cada una de las técnicas de minería de datos y de los algoritmos de manera correcta depende de los objetivos que se hayan propuesto y también depende de la naturaleza de los datos que se tienen en el conjunto a estudiar. Cada una de éstas tienen unas ventajas y desventajas dependiendo de su utilización, y seleccionar de manera correcta es importante para la obtención de unos resultados significativos.

4. TÉCNICAS UTILIZADAS

Las técnicas que se han utilizado en este trabajo han sido tres: el árbol de clasificación, el análisis de componentes principales y el análisis factorial. Lo siguiente que se hará es explicar en qué consisten cada una de estas técnicas y cómo se han de aplicar en una base de datos para que sea correcto su uso.

4.1. ÁRBOL DE CLASIFICACIÓN

La técnica estadística del árbol de clasificación es un método que se utiliza para obtener la clasificación de datos en categorías o grupos. Esta técnica está basada en la construcción de un árbol que divide de manera iterativa el conjunto de los datos en subconjuntos más reducidos; obteniendo de esta forma que los elementos que forman cada uno de los subconjuntos sean lo más homogéneos posibles, cumpliendo las características que son proporcionadas por la variable objetivo, también conocida como atributo de los datos. Cada una de las ramas que se forman a través de una de las variables significativas representa el resultado de una prueba en esa característica, y cada hoja que se forma, una etiqueta de clasificación.

Esta técnica cuenta con dos explicaciones, una relacionada con la parte teórica o matemática y otra que tiene relación con la parte lógica, en la que se trata de explicar en qué consisten los pasos de construcción de un árbol con la intención de mejorar su comprensión.

El proceso de construcción de un árbol de clasificación se realiza mediante 5 pasos, que son: la selección del nodo raíz, la división del conjunto de datos, el cálculo de impureza de los nodos hijos, la repetición del proceso y la predicción.

El primer paso es la selección del nodo raíz, que se escoge a través del cálculo de una medida de impureza o entropía, como son la entropía de Shannon o la ganancia de información, que serán mencionadas posteriormente en la parte matemática. Esta variable o atributo inicial es la característica que mejor divide el conjunto de datos.

El siguiente paso es la división del conjunto de datos en subconjuntos que estén basados en los valores de la característica que se ha seleccionado en el nodo raíz. De esta forma cada subconjunto que se va formando se convierte en un nodo hijo del nodo raíz, que se convierte en nodo padre.

El tercer paso de los cinco que hay es el cálculo de impureza de los nodos hijos de la misma forma que se realiza inicialmente para los nodos padres, para esto se utiliza de nuevo una medida de entropía.

El cuarto es la repetición del proceso que une los tres primeros pasos, para cada uno de los nodos hijo que se han generado, de manera repetida hasta que se cumpla una condición para que no se siga adelante con el proceso. Las condiciones de parada pueden ser de varios tipos, una de ellas es marcando una medida de impureza mínima, otra es marcando una cantidad mínima de muestras necesarias en cada nodo, en caso de ser inferior no se continúa dividiendo, o también podría ser al marcar inicialmente un límite en la profundidad del árbol que no pueda ser superado.

El quinto paso, y último, es el de la predicción, que consiste en predecir la etiqueta de una nueva instancia, haciéndola descender mediante el árbol de clasificación, siguiendo de este

modo las reglas de división en cada uno de los nodos que componen las ramas del árbol hasta llegar al fin de las ramas, obteniendo de este modo las hojas. La etiqueta final que se obtienen a través de las hojas es la que se toma como la predicción.

Lo siguiente que se verá es la explicación teórica detallada con las fórmulas matemáticas necesarias para realizar correctamente el árbol de clasificación. Los tres puntos que se van a explicar son la división de datos, la función de entropía de Shannon y la ganancia de información.

La división de datos es un conjunto de datos (X, Y) , donde la X contiene las variables predictoras e Y las etiquetas de clasificación o de variables de salida. De esta forma se tiene un vector que contiene las características, formado por los x_i que pertenecen a X y otro vector que está compuesto por las etiquetas de las clases correspondientes formado por los y_i pertenecientes a Y .

Lo siguiente que se tendrá en cuenta es la función de entropía de Shannon que está definida como:

$$E(S) = - \sum_{i=1}^c p_i \log_2(p_i)$$

donde se tiene: S es el conjunto de datos, c que es el número de clases que se encuentran dentro del conjunto de datos y p_i que es la proporción que tiene la clase i en el conjunto de datos S .

El uso de esta entropía consiste en la medida de la incertidumbre que se tienen en los datos que componen el conjunto.

El tercer punto que se tiene en cuenta es la ganancia de información, que se utiliza con el objetivo de medir la efectividad que tiene una característica para llevar a cabo la división de los datos. Esta ganancia se calcula haciendo la resta entre la entropía de Shannon que se ha calculado de los conjuntos de datos y la entropía ponderada de los conjuntos de datos divididos por la entropía original.

La fórmula de la ganancia de información es la siguiente:

$$IG(S, A) = E(S) - \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

donde se tiene A , que es una característica por la que se calcula la ganancia de información, S_v que consiste en los subconjuntos de S resultantes de dividir S según el valor de la característica que se tiene en A , $|S_v|$ que es el número de elementos que componen el subconjunto S_v , y $|S|$ que es el número de elementos totales que hay en el conjunto S .

En resumen, observando la parte lógica y la parte matemática, la técnica del árbol de clasificación usa medidas de impureza y ganancia de información para construir un árbol con la intención de poder predecir la variable objetivo en función de las características que se tienen. El proceso de construcción de este árbol está basado en la subdivisión de los problemas y la selección de características óptimas para dividir el conjunto de datos.

Los predictores que se obtienen son todos de tipo cualitativo mientras que las variables que se tienen como variables respuesta puede ser tanto de tipo cualitativo como de tipo cuantitativo. El árbol de clasificación está llevado a cabo por un análisis de segmentación que

se hace con los métodos de detección automática de la interacción (AID), que fue propuesto en 1963 por Morgan y Sonquist. El objetivo de estos métodos es la detección de la interacción en un modelo de predicción a través de una segmentación de los individuos en grupos homogéneos que sean excluyentes entre sí, con la intención de poder explicar la relación que hay entre los predictores y las variables respuesta sin quedar la relación enmascarada por la interacción entre las variables.

El objetivo principal de los métodos AID es la agrupación de los predictores que tienen un perfil similar para la variable dependiente. El procedimiento que usan es de tipo iterativo. Los grupos que se obtienen al final quedan agrupados después de un proceso por etapas, seleccionándose en cada una de ellas el mejor predictor posible. Dado que consiste en una estrategia de particiones de los datos sucesivamente, se denomina a estos métodos como análisis de segmentación.

El algoritmo que se utiliza comúnmente en el método general de segmentación es el algoritmo Chaid. Este algoritmo ha sido propuesto por Kass en 1980 que está basado en el test Ji-cuadrado para contrastar las independencias.

El algoritmo Chaid consiste en un algoritmo secuencial, que observando la ilustración del Trabajo de Fin de Máster, de Sergio Ramos Hernández (2015), se puede explicar en las siguientes fases:



Ilustración 1: “Fases del algoritmo Chaid”

Fase 1: En la primera etapa Kass propone la utilización del test Ji-cuadrado, cruzando la variable respuesta con los predictores, viendo de esta forma que categorías cuentan con un perfil similar con respecto a la variable respuesta, lo que quiere decir que no son significativas. En esos casos, esas categorías se unirán en agrupaciones. Kass propone cruzar cada par de categorías y realizar la fusión del par con mayor p-valor no significativo. Este proceso se llevará a cabo hasta que no se puedan formar más categorías por agrupación.

Fase 2: Al finalizar la primera fase, se continúa en la siguiente iniciándola con la selección del mejor predictor. De entre los predictores que se tienen se elige para segmentar aquella variable predictiva con un menor p-valor después de haber hecho el test Ji-cuadrado. De este modo el mejor predictor es aquel que discrimine mejor a los individuos según la variable respuesta.

Fase 3: En esta fase, cuando el p-valor ya está fijado, se dividirá el árbol en tantas ramificaciones como categorías tenga el predictor, siempre y cuando el predictor que se ha seleccionado es significativo.

Fase 4: Finalizada la segmentación de la fase 3, se realiza de nuevo el proceso en cada una de las ramificaciones hasta que todos los nodos obtenidos sean terminales, lo que quiere decir

que no haya más predictores significativos para ese nodo. En este proceso se habla de la agrupación de categorías y se ha fijado como criterio de parada la falta de variables significativas, pero esto puede traer problemas. Lo primero, no se puede realizar la agrupación de categorías sin seguir unos pasos, donde se depende de los tipos de predictores que se tienen. Lo siguiente, hay que fijar la ausencia de predictores influyentes o significativos como criterio de parada, puede llevar a la obtención de árboles con muchos nodos terminales con muy pocos individuos cada uno, por lo que se han de introducir más criterios de parada, como pueden ser un número mínimo de individuos en los nodos padre o marcar la altura máxima del árbol.

4.2. ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales es una técnica que consiste en la búsqueda de la reducción de las dimensiones de una matriz de datos inicial. Ésta se usa de manera generalizada en varios campos del conocimiento, siendo uno de los análisis más utilizados. Este análisis fue llevado a cabo principalmente por dos personas, con una diferencia de tres décadas. El primero fue Pearson en 1901 que buscaba una matriz con menor dimensión que la original, que resuma lo mejor posible la información de la base de datos inicial. El segundo, en los años treinta, fue Hotelling, que consiguió una aproximación que ofrece de manera sucesiva las combinaciones lineales de las variables de la base de datos con varianza máxima.

El análisis de componentes principales busca encontrar transformaciones ortogonales a las variables iniciales para obtener un conjunto de variables latentes que no están correlacionadas entre ellas, obteniéndose de mayor a menor importancia, que son los denominados componentes principales.

Las herramientas matemáticas que se necesitan para obtener estos componentes principales, inicialmente, son la matriz de correlaciones, la matriz de covarianzas, la descomposición espectral y la descomposición en valores singulares.

La matriz de correlaciones consiste en una matriz en la que la diagonal principal está formada por unos toda ella, ya que aporta la relación entre una variable consigo misma, mientras que en el resto de la matriz se encuentra la relación entre una variable y otra, en un rango entre menos uno y uno, siendo menos uno una correlación negativa perfecta y uno una correlación positiva perfecta. Los valores muy próximos a cero o que sean cero indican ausencia de correlación entre las variables.

La matriz de covarianzas es similar a la anterior, pero en la diagonal principal en lugar de tener unos, lo que se tienen son las varianzas de cada una de las variables. En el resto de la matriz lo que se tiene son las covarianzas que hay entre dos variables, explicando de este modo en caso de ser negativos los valores que esas variables varían en direcciones opuestas, y en caso de que los valores sean positivos ambas variables aumentan o disminuyen en la misma dirección.

La descomposición espectral consiste en la descomposición de la matriz inicial en valores y vectores propios, mediante el producto de la matriz inicial, por la izquierda con la traspuesta de una matriz ortogonal, y por la derecha con la matriz ortogonal. Entonces se tendrían tres matrices, una matriz inicial, la matriz ortogonal que se añade y la matriz final formada por los vectores propios, que cuenta con los valores propios en su diagonal principal.

La descomposición en valores singulares consiste en descomponer una matriz inicial, en la que se realiza el producto de los valores propios de ésta, por la izquierda con una matriz ortogonal y por la derecha con la traspuesta de otra matriz ortogonal, donde la matriz que multiplica por la izquierda se compone por los vectores propios izquierdos de la matriz inicial, la matriz diagonal se compone por los valores propios de la matriz inicial, y la matriz que multiplica por la derecha, que está traspuesta, se compone por los vectores propios derechos de la matriz inicial.

Existen varios métodos para poder obtener los componentes principales de una matriz inicial, los enfoques principales son cinco: la búsqueda del subespacio de mejor ajuste creada por Pearson en 1901, el método de la maximización de la varianza creada por Hotelling en 1933, el método de las regresiones alternas relacionada con los métodos *biplot*, el método de la descomposición en valores singulares de Eckart y Young y el método de las coordenadas principales de Gower.

El primero de estos métodos es el de la búsqueda del subespacio de mejor ajuste que es un precursor del análisis de componentes principales y consiste en encontrar un subconjunto con una dimensión menor del conjunto inicial con una variabilidad máxima.

El segundo es la maximización de la varianza que consiste en la reducción de la dimensionalidad con la mayor retención posible de información, a través de la técnica de análisis de componentes principales.

El tercero son las regresiones alternadas por medio de los *biplot*, que son los gráficos que ofrecen de manera simultánea la representación de las observaciones y de las variables en un espacio de dos dimensiones.

El cuarto se trata de la descomposición en valores singulares, que consiste en descomponer la matriz en tres componentes como se mencionó con anterioridad, obteniendo la matriz de vectores singulares izquierdos y derechos y la matriz diagonal que contiene los valores singulares de la matriz inicial.

El último es el método de las coordenadas principales que busca la representación de las relaciones que hay entre varias variables, siendo este método útil principalmente cuando las variables son categóricas y hay que visualizar las distancias o las similitudes entre las distintas categorías que se tengan.

Si se trata de un análisis de componentes principales con dos dimensiones, que es el más básico, se tiene en el primer componente principal la dirección de máxima variabilidad y en la segunda la dirección perpendicular donde se tiene la variabilidad restante. La dirección puede rotarse para que cuadre con los ejes principales y sea más sencilla la visualización.

La ecuación general para cada componente que se tiene para este análisis es:

$$y_j = Xv_j$$

siendo y_j uno de los componentes principales, la X la matriz con los datos de cada individuo y la v_j los coeficientes de las combinaciones lineales de esa columna.

La ecuación para todo el conjunto de datos en notación matricial es:

$$Y = XV$$

siendo Y la matriz que contiene todas las puntuaciones de cada uno de los individuos en cada componente, la X la matriz con las puntuaciones de los individuos y la V los coeficientes de las combinaciones lineales en las columnas.

El método más utilizado para la obtención de los componentes principales es a través de la maximización de la variabilidad. Lo primero que se hará en este método será la búsqueda del primer componente, el cual es el componente principal y es el componente con más varianza explicada. Como las variables que se tienen inicialmente tienen media nula, el componente principal conseguido también la tendrá, y se tendrá que calcular la varianza del componente de la siguiente forma:

$$Var(Y_1) = \frac{1}{n} y_1' y_1 = \frac{1}{n} v_1' X' X v_1 = v_1' S v_1$$

cómo se depende del vector v_1 si se aumenta el módulo del vector se puede alcanzar una varianza máxima sin límite, por lo que existe un factor de escala arbitraria ofreciendo de este modo infinitas soluciones con una misma dirección. Al ser vectores directores lo más común es que sean de módulo uno porque son con los que más fácil se trabaja, y se deben tener varias restricciones para que la varianza del primer componente tenga solución, éstas son:

$$\begin{aligned} v_1' v_1 &= 1 \\ V' V &= I \end{aligned}$$

esto quiere decir que en la dirección del primer componente principal se toma un vector unitario con la misma dirección que el componente.

Con las restricciones anteriores, ahora se busca el máximo de la siguiente ecuación:

$$L(v_1) = v_1' S v_1$$

si a esto se le introduce el multiplicador de Lagrange, que se utiliza para obtener los puntos críticos de una función con una o más restricciones y se añaden para incluir esas restricciones a la optimización. La ecuación queda así:

$$L(v_1) = v_1' S v_1 - \lambda(v_1' v_1 - 1)$$

que al derivar e igualar a cero se llega a la igualdad siguiente:

$$S v_1 = \lambda v_1$$

en esta igualdad se observa que v_1 debe ser un vector propio de la matriz de covarianzas, que es S , que tiene tantos valores λ como variables tenga, que son distintos y están ordenados de manera decreciente, siendo el mayor el primero y reduciendo el valor sucesivamente hasta ser igual o muy próximo a cero.

Si se tiene en cuenta la siguiente ecuación:

$$Var(X v_1) = v_1' S v_1 = v_1' \lambda v_1 = \lambda$$

se observa que λ tiene que ser λ_1 el mayor valor propio y v_1 el vector propio que se asocia a ese valor propio, de este modo se entiende que el primer valor propio es la varianza muestral del primer componente principal que se obtiene.

Del mismo modo se halla el segundo componente principal, pero añadiendo una nueva restricción que es que el primer componente principal (Y_1) y el segundo componente principal (Y_2) sean incorreladas.

Por lo tanto, se puede llegar a la conclusión de que los componentes principales se consiguen de la descomposición en valores y vectores propios de la matriz de covarianzas, siendo la ecuación:

$$S = V'DV$$

siendo S la matriz de covarianzas, D es la matriz diagonal con los valores propios y V es la matriz de vectores propios.

Por lo tanto, la ecuación que se tiene finalmente será:

$$Y_q = XV_q$$

donde se tiene las primeras q columnas de V y la matriz centrada X.

A través de los valores propios que se obtienen, se pueden calcular tanto la proporción de varianza absorbida por cada componente y la bondad de ajuste de los q componentes.

$$\text{Proporción de varianza absorbida por cada componente} = \frac{\lambda_j}{\sum_{j=1}^p \lambda_j} \times 100$$

$$\text{Bondad de ajuste de las q componentes} = \frac{\sum_{j=1}^q \lambda_j}{\sum_{i=1}^p \lambda_i} \times 100$$

Los objetivos que se tienen de las variables que se obtendrán son que sean incorreladas y se obtengan en orden decreciente, se espera que unas pocas variables recojan la mayor parte de la información que tienen los datos iniciales, el espacio que se genera por los nuevos componentes es un subespacio de las variables iniciales, por lo que el análisis de componentes principales se puede entender como la búsqueda del subespacio con mayor explicación y mejor ajuste a raíz de los datos iniciales.

Lo siguiente que se trata es la cantidad de componentes que se deben retener, y esto puede realizarse de cuatro formas: la primera consiste en seleccionar todos los valores propios que sean mayores que uno; la segunda es el *Scree Plot*, que se basa en seleccionar los valores propios de la tabla hasta que ocurra un descenso brusco; la tercera reside en fijarse en el porcentaje de varianza explicada y escoger los valores propios que superen un determinado porcentaje, habitualmente el 75%; y la última radica en asignar el número de componentes previamente.

En la interpretación de los ejes factoriales o de los componentes principales, hay que fijarse en las cargas. Las variables que cuentan con cargas altas son la que tienen más importancia en ese componente principal, destacando aún más las variables con altas cargas en un componente y bajas cargas para los demás.

4.3. ANÁLISIS FACTORIAL

El análisis factorial aparece en el siglo pasado con la intención de comprender las dimensiones que puede llegar a contener la inteligencia humana. Los orígenes de esta técnica están influenciados por cuatro personas que fueron Spearman en 1904, Pearson y Hotelling en 1933, y por último Thurstone en 1947. El campo en el que más se avanzó en esta técnica fue en el campo de la psicometría.

Este método intenta dar una explicación del conjunto de variables que han sido observadas mediante un número menor de variables latentes que no pueden ser observadas que se denominan factores. Este modelo está referenciado por dos tipos: los comunes que unen el conjunto de las variables y los únicos, habiendo uno de estos para cada variable observada.

El análisis que se trata tiene el objetivo de reducir la dimensionalidad del conjunto de datos hasta un mínimo relativo que sea más fácil de visualizar o de explicar, obteniendo los factores que son estructuras latentes subyacentes.

Una variable latente o un constructo consiste en un concepto que es difícil de definir dentro de una teoría científica, siendo algo de lo que se sabe su existencia pero que es difícil de definir o de decir que es.

El análisis estadístico es la técnica de análisis no supervisada que ha sido diseñada para analizar y explicar las relaciones que hay entre las variables originales, explicadas por un nuevo conjunto de variables latentes comunes y un conjunto de factores únicos.

Este análisis utiliza la matriz de correlaciones y la matriz de correlaciones reducida entre las variables que se tienen para conseguir e interpretar los factores comunes. Para ello se buscan unos factores hipotéticos que expliquen de la mejor manera las variables originales.

El modelo factorial lineal consiste en que, si se tiene un conjunto de variables aleatorias sobre una población, se trate de encontrar los factores, que son las nuevas variables, que sean tanto factores comunes como factores únicos; siendo F los factores comunes y U los factores únicos. La ecuación general que tenemos para el conjunto, que luego se dividirá uno a uno, es:

$$X = AF + DU$$

siendo la primera de la X_1 la siguiente ecuación, y como esta todas las demás para cada X :

$$X_1 = a_{11}F_1 + \dots + a_{1k}F_k + d_1U_1$$

de este modo, las a son las saturaciones de la variable X que se trate, en el factor correspondiente. Las d son las unicidades, que consisten en la contribución del factor único U a la respectiva variable.

En los supuestos que se pueden encontrar del modelo, hay que tener en cuenta que los factores que se obtienen de las variables iniciales siempre son menos que la cantidad de variables, por ser el análisis factorial una técnica de reducción de la dimensionalidad. Los factores que se obtienen son variables no correlacionadas entre ellas, es decir, entre los factores comunes y los factores únicos. Que las variables sean no correlacionadas quiere decir que los factores comunes y únicos son incorrelacionados dos a dos, y tampoco hay correlación entre los factores comunes y los factores únicos. Para la realización de todo esto se tiene en cuenta que las variables observadas y los factores obtenidos estén centrados y estandarizados, lo que quiere decir que la media es cero y la varianza uno.

Lo siguiente que se trata son las saturaciones, las comunalidades y las unicidades que ya han sido mencionadas con anterioridad. Las saturaciones son los coeficientes de a_{ij} , tratando de este modo la variable X_i en el factor F_j . De estas saturaciones, elevándolas al cuadrado se obtienen las contribuciones de cada uno de los factores, siendo la contribución del factor correspondiente de la variable X que se esté tratando. También se obtiene la unicidad del mismo modo, la unicidad es la contribución del factor único y consiste en la elevación al cuadrado de la d del factor correspondiente.

La comunalidad indica la varianza explicada de una variable por el factor, mientras que la unicidad es la cantidad de varianza de una variable que no depende del factor.

La comunalidad es la suma de las contribuciones de todos los factores comunes, esto es:

$$h_i^2 = a_{i1}^2 + \dots + a_{im}^2$$

como se ha dicho previamente las variables y los factores están estandarizados por lo tanto la varianza de todos los factores tanto comunes como únicos, es igual a uno. Entonces la suma de la comunalidad y de la unicidad, de cada ítem, es igual a uno, la fórmula es la siguiente:

$$h_j^2 + d_j^2 = 1$$

lo siguiente que se trata es la búsqueda de la matriz factorial A que es el principal objetivo del análisis factorial, siendo encontrar e interpretar la matriz. Este análisis obtiene e interpreta los factores comunes a partir de la matriz de correlaciones entre las variables. Para la búsqueda se utiliza el teorema de Thurstone, que permite relacionar la matriz factorial con la matriz de correlaciones reducida, que consiste en la matriz de correlaciones con la sustitución de los unos de la diagonal por las comunalidades.

El problema factorial consiste en la descomposición de una matriz de correlaciones en otra factorial A, que reproduzca la matriz de correlaciones reducida al hacer el producto de la matriz A por su traspuesta.

El método del factor principal es el método de obtención de la matriz factorial con la intención de que los factores expliquen sean incorrelacionados y la varianza sea máxima. Por lo que consiste en la diagonalización de la matriz de correlaciones reducidas R^* en un proceso iterativo en el que se reestiman las comunalidades, que son estimadas *a priori*.

Este proceso cuenta con cinco pasos, consistiendo el primero de ellos con la estimación previa de las comunalidades, la estimación *a priori* puede ser de dos formas, la primera de ellas consiste en estimar las comunalidades con el valor uno, y la segunda de ellas consiste en la estimación de las comunalidades a partir del cuadrado del coeficiente de correlación múltiple de la variable con las demás. El segundo paso consiste en el cálculo de la matriz reducida. En el tercer paso se diagonaliza la matriz de correlaciones reducida y se estima la matriz factorial. Lo siguiente que se hará es el cálculo de las nuevas comunalidades, en caso de que sean iguales a las que se tienen previamente se finaliza, en caso de ser distintas se vuelve al segundo paso.

En el modelo factorial con un solo factor, las saturaciones que se tienen son iguales a las correlaciones de cada variable con el factor. El valor que se obtenga puede ser tanto negativo como positivo, en caso de ser próximo a cero o cero, no se tiene una relación entre la variable y el factor que se obtiene, mientras que unas saturaciones muy extremas quieren decir que hay una relación muy fuerte entre las variables y el factor.

Y cuando se tiene un modelo factorial con más de un factor se obtiene la varianza explicada por cada uno de los factores realizando la suma de las saturaciones en ese factor al cuadrado.

Otro punto importante para la realización del análisis factorial es la selección del número de factores. Para la selección hay varias formas o métodos para determinar cuál es el número óptimo de factores que se deben retener. La primera de las cuatro formas que se tienen es la de retener todos los factores que tengan el valor propio superior a uno. La segunda consiste en seleccionar los factores que acumulen el 75% de la varianza. La tercera opción es un método visual que consiste en visualizar el "Scree Plot" y realizar la regla del codo, que consiste en observar el factor en el que la variabilidad o el valor propio se mantiene más o menos

constante, sin cambios bruscos. Y la última opción es la especificación del usuario, que consiste en que el usuario escoge la cantidad de factores que quiere retener.

En la interpretación de los factores se tiene varias reglas, ya que no tiene por qué ser sencilla la interpretación de éstos por no mostrar con claridad lo que se mide. Para ello se cuenta con tres reglas.

La primera de ellas consiste en analizar las saturaciones que se tienen, en valor absoluto, observando que las variables que tienen unas saturaciones altas son las más importantes en el momento de interpretar el eje.

La segunda consiste en observar las variables que cuentan con un valor alto en uno de los factores, siendo relevantes en él, y que tienen un valor bajo en los demás factores. Las variables se unen en los factores, que indican las variables que hay en cada uno de ellos y para interpretar el factor hay que observar la relación entre las variables altas de ese.

La tercera regla, y última, consiste en la búsqueda de una mejor interpretabilidad a través de la rotación de los ejes factoriales, consiguiendo que aumenten las saturaciones de algunas variables y disminuyan las de otras.

De esta tercera regla, de la rotación de los ejes factoriales se tiene el principio de estructura simple en la matriz factorial. Esto consiste en la búsqueda de una mayor sencillez en la interpretación de los factores, teniendo unas variables fuertemente relacionadas en el factor y el resto poco o nada relacionadas. Para ello habrá en cada una de las columnas de la matriz valores que sean nulos o prácticamente nulos. En cada fila habrá más de un cero y para cada par de columnas se tendrán pocos elementos que no sean nulos en ambas.

Después de obtener una solución inicial, se realiza la rotación de los ejes factoriales para conseguir de este modo con la estructura más simple posible, minimizando de este modo el número de variables con altas saturaciones en los factores y muchas de las otras variables son nulas o prácticamente nulas. Tenemos dos tipos de rotaciones, las ortogonales y las oblicuas. Los principales tipos que hay de rotaciones ortogonales son la rotación varimax, la quartimax y la equamax, al ser ortogonales los nuevos factores que se obtienen siguen sin tener correlación entre ellos. Por otro lado, las principales rotaciones que hay son la cuartimin y la oblimin, en estas rotaciones los factores están relacionados, es decir, no respetan la ortogonalidad entre los factores.

5. METODOLOGÍA

En la metodología se trata la base de datos, como está formada, los datos que se han suprimido y las transformaciones que se han realizado sobre la base. También se abordan los *softwares* que se han utilizado tanto para almacenar los datos y crear la base de datos, como es Excel, como los que se han utilizado para realizar los análisis estadísticos, que son RStudio y SPSS, y con los que se han generado las gráficas de los estadísticos descriptivos.

5.1. BASE DE DATOS

La base de datos que se utiliza en este trabajo se ha realizado con estadísticas de las cinco grandes ligas europeas, que son la Premier League inglesa, LaLiga española, la Ligue 1 de Francia, la Bundesliga alemana y la Serie A italiana. Las estadísticas de cada uno de los jugadores se han recogido del desempeño de cada uno en su respectiva liga en la temporada 21/22, estas estadísticas se han obtenido de la página web de *Whoscored* y de *Transfermarkt*.

Si se tratan las variables que forman parte de la base de datos, las seis primeras variables fueron obtenidas a través de la página web de *Transfermarkt*. Las variables que se obtuvieron son:

1. “Jugador”, esta variable contiene el nombre de los jugadores, ya sea el nombre con el que es conocido en caso de ser un apodo, el nombre si se le conoce por él y no es necesario su apellido, el apellido únicamente si es conocido por él y no es utilizado el nombre de pila, por último, ambas cosas, tanto el nombre como el apellido, para que no pueda haber ningún tipo de confusión entre un jugador y otro.
2. “Posición”, esta variable está conformada por varias opciones, la posición de cada jugador es en la que desempeñan desde un inicio de su carrera, y no en la que juegan este año o en este mismo momento. Las posiciones con las que cuenta la base son la de portero, defensa central, lateral, mediocentro, extremo, delantero. En esta base se han unido a los media puntas a la posición de delantero y a los mediocentros ofensivos se les ha marcado como mediocentros.
3. “Edad”, esta variable se ha tomado en el año de referencia de la temporada por lo que se han tomado en el inicio de las ligas en 2021.
4. “Valor de mercado en millones”, esta variable se ha tomado en el inicio de la temporada también, en todos los casos que se podía, dado que hay jugadores que no tenían un valor inicial, por lo tanto, se intentó mirar el primer valor de mercado posible, y en caso de no poder encontrarlo, asignarles un valor similar a otros jugadores con su edad y de la misma liga.
5. “Equipo”, esta variable es el equipo en el que juega ese jugador en la temporada 21/22.
6. “Liga”, consiste en la liga a la que forma parte cada uno de los equipos, es decir, la liga en la que compite cada equipo.

A mayores de las seis variables que se han obtenido de la página web de *Transfermarkt*, hay que añadir las variables que se han obtenido de la página web *Whoscored*, en las que encontramos las estadísticas de los jugadores separadas por la característica de cada una, ya fueran características generales, como son las once primeras, las características defensivas que son seis, las características ofensivas que son cinco y, por último, las características asociadas al pase que son cuatro.

Las características generales son:

1. “Partidos jugados”, esta variable consiste en la cantidad de partidos en los que ha participado el jugador en cuestión a lo largo de toda la temporada, sumando tanto los partidos que ha jugado de titular como los que ha jugado de suplente.
2. “Minutos”, la variable siguiente está referida a la cantidad de minutos que ha participado el jugador en cuestión en toda la liga, sumando los que ha jugado tanto de titular como de suplente.
3. “Goles”, esta variable se refiere a la cantidad de goles que ha anotado un jugador a lo largo de la liga.
4. “Asistencias”, consiste en la cantidad de pases de gol, que han acabado con un gol anotado por el equipo durante toda la liga.
5. “Tarjeta amarilla”, está referida al número de tarjetas amarillas recibidas por un jugador en la liga.
6. “Tarjeta roja”, trata de la cantidad de tarjetas rojas que le han sacado a un jugador en esa temporada.
7. “Tiros por partido”, esta variable habla de la cantidad de veces que tira un jugador por partido en la competición.
8. “Acierto pase %”, es el porcentaje de pases acertados que tiene cada jugador a lo largo de la competición liguera.
9. “Duelos aéreos ganados por partido”, esta variable consiste en la cantidad de duelos aéreos que gana el jugador de media por partido.
10. “MVP”, *most valued player* o también conocido como hombre del partido, consiste en las veces que el jugador en cuestión ha recibido el premio a mejor jugador del partido a lo largo de la campaña liguera.
11. “Valoración media”, esta variable trata de la media de la valoración que ha obtenido el jugador en cada partido jugado a lo largo de la temporada.

Las características que tenemos de carácter defensivo son:

1. “Entradas por partido”, esta variable consiste en la cantidad de entradas defensivas que hace el jugador en cuestión de media por partido a lo largo de toda la competición de liga.
2. “Intercepciones por partido”, la variable en cuestión cuenta la cantidad de veces que un jugador intercepta el balón al rival, de media, durante toda la temporada.
3. “Faltas por partido”, consiste en la cantidad de faltas que un jugador realiza sobre un rival a lo largo de la competición.
4. “Fuera de juegos ganados por partido”, esta variable consiste en la cantidad de veces que el jugador en cuestión deja a un rival en posición de fuera de juego por un movimiento del defensor.
5. “Despejes por partido”, la variable trata de la cantidad de despejes que realiza un jugador por partido en esa temporada.
6. “Regateado por partido”, esta variable consiste en las veces que este jugador ha sido regateado por un rival de media por partido a lo largo de la competición liguera.

Observando el carácter ofensivo, las variables con las que cuenta la base de datos son:

1. “Pases clave por partido”, es la variable en la que se recogen la cantidad de pases clave que realiza un jugador por partido, lo que quiere contabilizar los últimos pases previos a un posible gol, que puede que hayan terminado entrando en la portería o puede que no hayan entrado.

2. “Regates por partido”, es la cantidad de regates que realiza un jugador a un rival de media a lo largo de un partido en la competición liguera.
3. “Faltas a favor por partido”, esta variable cuenta la cantidad de faltas que provoca el jugador en cuestión, de promedio a lo largo de toda la temporada.
4. “Fuera de juegos por partido”, esta variable consiste en la cantidad de fueros de juego que le pitan en contra por una posición antirreglamentaria del jugador que se habla a lo largo de un partido.
5. “Controles malos por partido”, es la cantidad de veces que un jugador falla en un control lo que propicia una pérdida de balón ante el rival, de media por partido a lo largo de toda la temporada.

Por último, las variables que tratan sobre las cualidades de pase, que son:

1. “Pases por partido”, esta variable consiste en la cantidad de pases que realiza un jugador de promedio en los noventa minutos de un partido de fútbol.
2. “Centros por partido”, es la cantidad de veces que un jugador centra el balón de media en un partido.
3. “Balones largos por partido”, se trata del número de veces que un jugador golpea un balón en largo de media por partido en la competición.
4. “Pases al hueco por partido”, esta variable consiste en la cantidad de pases por partido que realiza un jugador al hueco con la intención de encontrar a un compañero desmarcándose.

Esta base de datos ha sufrido unas modificaciones ya que no todos los jugadores que aparecen en la página web de *Transfermarkt* han disputado algún minuto con su equipo, por lo que han tenido que ser suprimidos porque no aparecían en la página web de *Whosocred*.

El resto de los jugadores que han disputado algún minuto se han mantenido en la base de datos, sin aplicarles ningún otro tipo de requisito, como podría ser un número mínimo de partidos o de minutos jugados, o una valoración mínima o máxima, suprimiendo de esta forma los valores atípicos de la base de datos.

La base de datos cuenta con una cantidad de jugadores inicialmente de tres mil doscientos diecisiete individuos de las cinco grandes ligas europeas. Después de realizar la limpieza de la base de datos, eliminando principalmente a los jugadores que no habían contado con minutos en el equipo de la correspondiente liga, en la temporada 21/22. Hay que tener en cuenta que no se han suprimido los jugadores que han jugado en más de un equipo algún partido porque interesa la cantidad de datos y no el nombre de cada uno de los jugadores, y un jugador que haya jugado esa temporada en dos equipos puede tener datos interesantes con ambos equipos.

Después de realizarle la limpieza correspondiente la base de datos se ha reducido desde tres mil doscientos diecisiete jugadores de las cinco grandes ligas a dos mil novecientos diecinueve jugadores, es decir, una bajada de casi trescientos jugadores que no han disputado ningún minuto a lo largo de toda la competición liguera. Es importante el matiz de la competición liguera, porque puede ser que haya jugadores que hayan contado con minutos en las competiciones coperas, pero esos datos no están incluidos en la base de datos, porque en cada país la copa es distinta y por el motivo de que la Premier League inglesa cuenta con un número mayor de copas que las otras cuatro ligas.

5.2. SOFTWARE R

La aplicación R es un entorno de *software* y un lenguaje de programación que está diseñado de manera específica para el análisis de datos y para la estadística. Fue creado por Robert Gentleman y Ross Ihaka en 1933 en Nueva Zelanda, concretamente en la Universidad de Auckland. Desde ese momento, R se ha convertido en una de las herramientas más importantes en la comunidad del análisis de datos y de las estadísticas porque presenta una facilidad a los usuarios por su flexibilidad, su capacidad para modificar y visualizar los datos, y su gran cantidad de desarrolladores y usuarios que aportan tanto bibliotecas como paquetes, que pueden facilitar su uso y la obtención de resultados.

Una de las cosas que nos aporta este *software* es su código abierto y que se distribuye bajo la Licencia Pública General de GNU (GPL). Esto quiere decir que es gratuito y que la comunidad de desarrolladores puede crear paquetes adicionales para el *software* y también pueden contribuir al desarrollo de éste.

Otra de las características importantes de R es la visualización de los datos pudiendo crear gráficos o ver los datos con una gran calidad. En concreto hay una biblioteca que es de las más utilizadas que tiene el nombre de “ggplot2”, que fue creada por Hadley Wickhman y que es la más utilizada para la visualización de datos y creación de gráficos.

Como se mencionó con anterioridad las bibliotecas y los paquetes también aportan mucho al mundo de R, ya que son llevados a cabo por desarrolladores que van actualizando cada uno de ellos para que sea lo mejor posible en todo momento. La principal fuente de paquetes de R es la función “CRAN” (*Comprehensive R Archive Network*) que alberga una gran cantidad de paquetes que incluyen distintas necesidades de análisis de datos.

El programa de R también contiene un gran conjunto de herramientas para el análisis estadístico, contando con pruebas de hipótesis, regresiones y más. Los paquetes que se utilizan principalmente para el análisis estadístico son el de “stats” y “lme4”.

Al contar R con un *script* donde poder escribir y realizar funciones personalizadas pudiendo automatizar las tareas y la creación de flujos de trabajo propios para cada caso.

5.3. SOFTWARE SPSS

El *software* SPSS (*Statistical Package for the Social Sciences*), también conocido como IBM SPSS *Statistics*, es un *software* que se utiliza mucho en el campo del análisis de datos y de la estadística. Fue desarrollado en la década de 1960 por Norman H. Nie, C. Hadlai “Tex” Hull, y Dale H. Bent y desde entonces ha ido evolucionado como una muy buena herramienta de análisis de datos utilizada en diversas disciplinas, como pueden ser el deporte, la educación, la salud, la psicología y muchos más.

Este *software* destaca porque los usuarios pueden llevar a cabo todos los análisis estadísticos sin necesidad de programar, por lo que es una opción muy sencilla para los usuarios que no son programadores.

Además, SPSS tiene muchas técnicas estadísticas, que incluyen análisis descriptivos, análisis de la varianza, regresiones, análisis factorial y más técnicas. También cuenta con una fácil manipulación de los datos ya sea por la inclusión de hojas de cálculo, bases de datos y más

formatos comunes. Los usuarios pueden manipular y hacer transformaciones de datos, recodificar variables y limpiar datos con facilidad.

Las herramientas de visualización que nos aporta este *software* son fáciles de entender, incluyendo gráficos de barras, gráficos de dispersión, histogramas, entre otros. También tiene la facilidad de personalizar gráficos para remarcar las celdas más significativas o que nos puedan llamar más la atención.

Este *software* es uno de los más utilizados para realizar los distintos análisis estadístico que se precisen. La facilidad de los botones y que no sea necesario para el usuario el saber programar, ayuda a que sea utilizado en muchos ámbitos externos a los que en un principio deberían saber programar. Se utiliza para predecir tendencias, para estudios económicos de los mercados, y sobre todo en la medicina para obtener los resultados en las investigaciones.

5.4. SOFTWARE EXCEL

Excel es un programa de hojas de cálculo que ha desarrollado Microsoft. Es una herramienta que permite al usuario crear, organizar y analizar los datos de las hojas de cálculo.

Las características principales y más importantes de Excel son:

1. Las hojas de cálculo consisten en la organización de información en una cuadrícula formada por filas y columnas. Cada una de las celdas de la hoja de cálculo tienen una única dirección que está basada en la intersección que ocurre entre la fila y la columna en las que se encuentre.
2. Una de las características más importantes del Excel son las capacidades de cálculo con las fórmulas y las funciones. Gracias a estas fórmulas y funciones se pueden realizar cálculos de todo tipo relacionados con las matemáticas, la estadística, las finanzas y la lógica.
3. Otra característica importante son los gráficos que se pueden crear en base a los datos que hay en la hoja de cálculo. La importancia de los distintos tipos de gráficos, como pueden ser los de barras, los histogramas y los gráficos de sectores, entre otros, es que éstos facilitan la visualización de datos y la identificación de tendencias o de patrones.
4. La característica del formato condicional ayuda al aplicar las reglas de formato para poder resaltar automáticamente los valores que no cumplan o que cumplan ciertas condiciones marcadas, facilitando la identificación de datos dentro de un grupo o excluyéndolos.
5. El análisis de datos en Excel se encuentra proporcionado por varias herramientas como son las tablas dinámicas y el *solver*, que ayudan en la realización de tendencias y escenarios hipotéticos.
6. La automatización con macros que se usan para la automatización de tareas repetitivas, son utilizados habitualmente por usuarios que buscan acciones más específicas como podría ser un árbol de clasificación.
7. Y la última característica es la posibilidad de importar y exportar los datos desde diversas fuentes o hacia diversas fuentes, como pueden ser bases de datos, archivos de texto y otros.

En conclusión, Excel es una herramienta que aporta una gran cantidad de funciones para realizar sobre una base de datos plasmada en una hoja de cálculo.

6. ANÁLISIS ESTADÍSTICO

Primeramente, se ha realizado una matriz de correlaciones para saber la relación que hay entre las variables de la base de datos y saber de esta forma cuales van a ser los estadísticos descriptivos que más expliquen la base de datos.

Con las técnicas estadísticas se pretenden dos cosas, la primera es obtener que variables son más significativas a la hora de clasificar en grupos la base de datos siendo la variable de influencia la “liga” en la que se juegue a través del árbol de clasificación, y la segunda es realizar una reducción de la dimensionalidad de nuestra base de datos con la intención de obtener tanto los componentes como los factores, dependiendo del método, que mejor representen y expliquen la base de datos.

6.1. ESTADÍSTICOS DESCRIPTIVOS

Con el siguiente gráfico de barras se intentará obtener una distribución de los jugadores respecto de cada una de las posiciones y a su mismo tiempo, dentro de las posiciones, realizar una división de los jugadores por la liga en la que compitieran ese año.

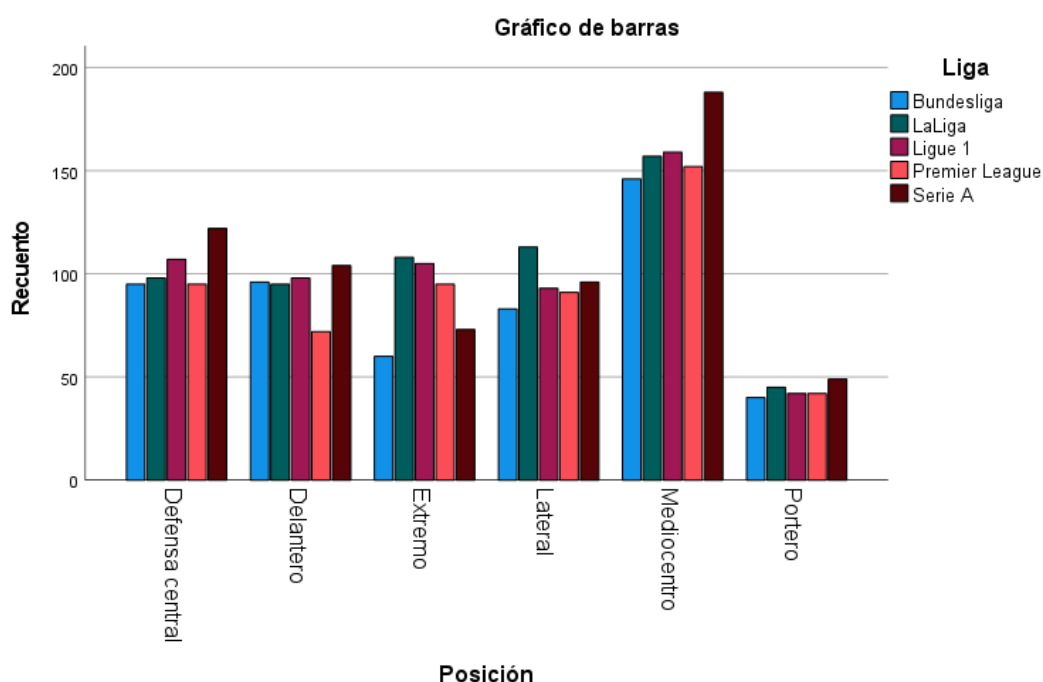


Ilustración 2: “Gráfica de barras agrupadas de las posiciones separadas por las “ligas”.”

Observando la gráfica anterior, que tiene el recuento de los jugadores en cada una de las “posiciones” respecto de cada “liga”, se observa que la posición que contiene más jugadores es la de mediocentro, que las posiciones de defensa central, de delantero, de lateral y de extremo están más o menos en una misma cantidad de jugadores, mientras que la posición de portero es la menos poblada, como cabía esperar.

Yendo posición por posición, desde el posicionamiento en el terreno de juego, desde lo más defensivo hasta lo más ofensivo, siendo el primero grupo los porteros, donde ninguna de las

ligas llega a tener cincuenta porteros, habiendo entre las cinco hay doscientos dieciocho jugadores. La Serie A es la que más tiene contando con cuarenta y nueve, mientras que la que menos tiene es la Bundesliga con cuarenta. En el medio de ambas ligas se pueden encontrar con cuarenta y dos la Ligue 1 y la Bundesliga, y LaLiga tiene cuarenta y cinco.

Si se habla de los defensas centrales que hay en las cinco grandes ligas europeas se estaría hablando de quinientos diecisiete jugadores, se puede observar que la liga donde más hay es en la Serie A con ciento veintidós, seguida de por los ciento siete jugadores que juegan de defensa central en la Ligue 1, ya por debajo de los cien jugadores se encuentran LaLiga con noventa y ocho, la Bundesliga y la Premier League que cuentan ambas con noventa y cinco jugadores en esa posición.

La siguiente posición en el campo es el lateral, en la base de datos hay cuatrocientos setenta y seis jugadores en esta posición. La liga que más destaca en esta posición es LaLiga con ciento trece jugadores, mientras que el resto están por debajo de los cien ya, apareciendo la Serie A, la Ligue 1 y la Premier League, con noventa y seis, noventa y tres y noventa y uno, respectivamente. En último lugar se encuentra la Bundesliga con solamente ochenta y tres jugadores que ocupen la posición de lateral en toda la liga.

En la posición de mediocentro cuenta con ochocientos dos jugadores, siendo la posición que cuenta con más jugadores. La liga que más jugadores tiene en esta posición es la Serie A con ciento ochenta y ocho, seguida de la Ligue 1 con ciento cincuenta y nueve, de LaLiga con ciento cincuenta y siete, de la Premier League con ciento cincuenta y dos y de la Bundesliga con ciento cuarenta y seis mediocentros.

La siguiente posición es la de extremo, en la que se tienen datos muy dispares por las ligas, se puede observar que de los cuatrocientos cuarenta y un jugadores que hay, la Bundesliga solamente tiene sesenta y la Serie A setenta y tres, mientras las otras tres ligas están próximas a los cien jugadores, contando con noventa y cinco la Premier League, ciento cinco la Ligue 1 y ciento ocho LaLiga.

La última posición es la de delantero, en la que solo se tiene a la liga italiana por encima de la marca de los cien jugadores, mientras que las ligas francesas, españolas y alemanas están próximas a los cien jugadores. La liga que más destaca por falta de delanteros es la liga inglesa, que es la que cuenta con menos en esta posición.

Lo siguiente que se realizará es un diagrama de cajas con la intención de observar la distribución del “valor de mercado” en cada una de las “ligas” en cuestión, pudiendo entender que ligas tienen jugadores con mayor valor de mercado, quienes son los valores anómalos.

Diagrama de cajas Simple de Valor de mercado (millones) por Liga

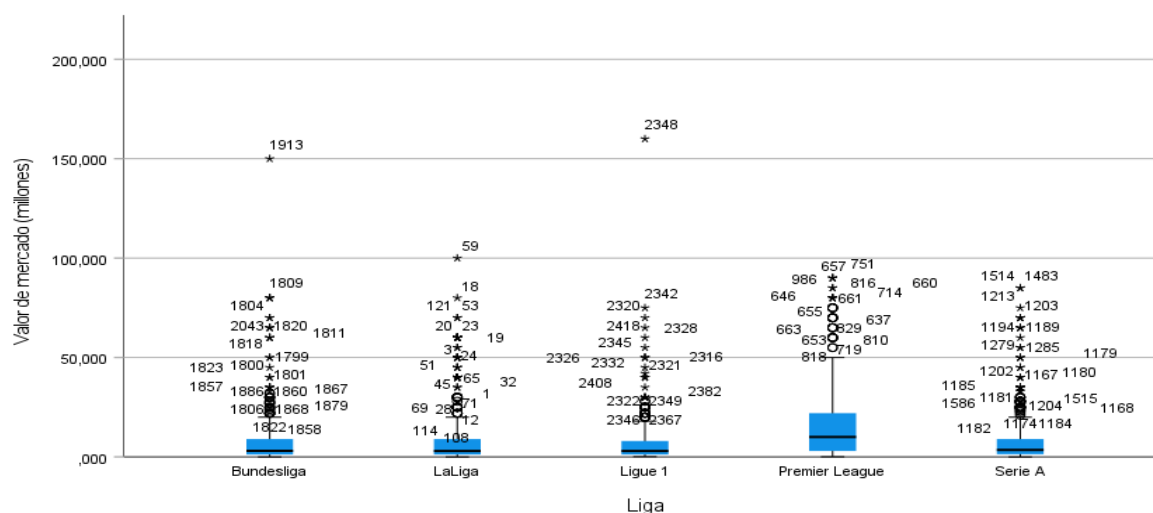


Ilustración 3: “Box-Plot del “valor de mercado en millones” en las “ligas”.”

Observando los “Box-plot” en general, se puede apreciar que hay dos jugadores que están muy por encima del resto, que son Haaland y Mbappé, ambos en el intervalo de ciento cincuenta millones y doscientos millones de euros. También se observa que hay muchos jugadores que están por encima de los límites de cada una de las ligas. Se puede ver que prácticamente la Bundesliga, LaLiga, la Ligue 1 y la Serie A no llegan a tener un valor de los jugadores próximo a los cincuenta millones, pero que si tienen una gran cantidad de valores atípicos que llegan a esos valores y hasta lo superan. Por otro lado, la Premier League tiene el final del bigote superior en los cincuenta millones, lo que viene precedido porque tiene un mayor rango intercuartílico que las otras ligas, esto también puede estar motivado porque es probablemente la mejor de las cinco ligas y que cuenta con jugadores más valiosos que las otras.

También se puede observar que las cuatro mencionadas primeramente tienen una asimetría positiva, esto quiere decir que la parte grande de la caja es la parte entre la mediana y el tercer cuartil, concentrándose los datos principalmente en la parte inferior de la distribución y teniendo una media superior a la mediana.

También es curioso que la Premier League sea la única que llega con los bigotes a los cincuenta millones de euros, pero que también sea una de las dos ligas que no tiene ningún jugador con un valor de mercado igual o mayor a los cincuenta millones.

6.2. TÉCNICAS ESTADÍSTICAS

Las técnicas que se han realizado sobre la base de datos han sido tres, dos de ellas en RStudio y la otra en SPSS. Las técnicas en cuestión son el análisis de componentes principales (ACP) y el análisis factorial (AF), realizadas con RStudio y el árbol de clasificación realizado con SPSS.

6.2.1. ÁRBOL DE CLASIFICACIÓN

La primera técnica que se va a tratar de manera práctica será el árbol de clasificación, pudiendo observar varias imágenes con la distribución del árbol y como se forma nodo a nodo.

Lo que se quiere obtener con esta técnica son unos grupos finales que engloben a los jugadores de la base de datos, a través de las variables predictivas que van segmentando el árbol.

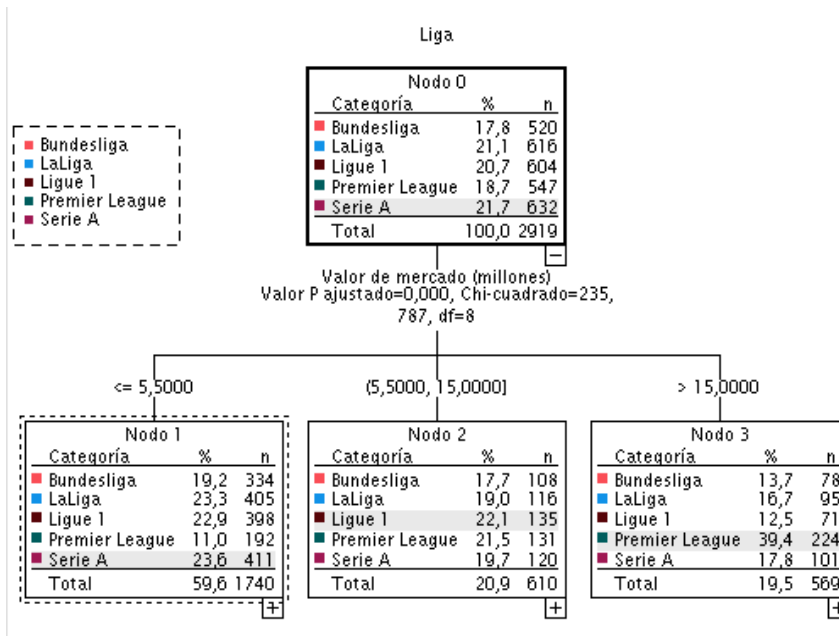


Ilustración 4: “Árbol de clasificación del nodo cero al nodo tres”

El primer nodo que hay en el árbol es el nodo cero, que está dividido en las “ligas”, pudiendo observar que se tienen dos mil novecientos diecinueve jugadores, de los cuales la liga que más tiene es la Serie A italiana, con seiscientos treinta y dos, y la liga que menos tiene es la Bundesliga alemana con quinientos veinte, el resto de las ligas se mueven entre estas dos cifras. Hay que tener en cuenta que los valores próximos a 20% cumplen los requisitos y estarían bien clasificados ya que tenemos cinco ligas y si todas las clasificaciones que se hicieran fueran perfectas, el porcentaje perfecto de jugadores en cada liga sería del 20%.

Se puede observar que, del primer nodo, el nodo cero, se obtienen tres nodos hijos, que están clasificado por la variable más significativa, que mejor nos clasifica los datos, que es el “valor de mercado en millones”, que clasifica en tres grupos a los jugadores, los que valen hasta 5,5 millones de euros, los que valen entre 5,5 millones hasta 15 millones, incluyendo sólo el segundo valor, y el tercer grupo que son los que valen más de 15 millones.

El nodo uno está formado por los jugadores con un “valor de mercado” de 5,5 millones de euros o inferior, en este grupo hay mil setecientos cuarenta jugadores, que es el 59,6% de la base de datos. Las dos ligas más numerosas de este nodo son la Serie A y LaLiga, mientras que la que menos destaca es la Premier League. Los valores que más destacan son los que sean inferiores a 17,5% y superiores a 22,5%, entonces la Premier League destaca por tener pocos jugadores que cumplan la característica, y la Serie A y LaLiga por tener más del 22,5% de jugadores con esta característica.

En el siguiente nodo, el nodo dos, cuenta con seiscientos diez jugadores formado por los jugadores que tienen un “valor de mercado” entre 5,5 millones y 15 millones, únicamente incluido el superior en el grupo. Las ligas que más destacan son la Ligue 1 con un 22,1% y la Bundesliga con un 17,7%. Se puede observar que ninguna liga supera el 22,5% y el 17,5% por lo que ninguna destaca en el número de jugadores que forman parte de este nodo.

El nodo tres está formado por quinientos sesenta y nueve jugadores que tienen un “valor de mercado” por encima de los 15 millones, donde la liga que más destaca es la Premier League con 39,4%, que equivale a doscientos veinticuatro jugadores, siendo una cantidad de jugadores muy por encima del 22,5%, siendo la liga más importante del nodo. Si se aprecian los porcentajes inferiores, se puede observar que hay tres ligas inferiores al 17,5%, que son de menos a más jugadores, la Ligue 1, la Bundesliga y LaLiga.

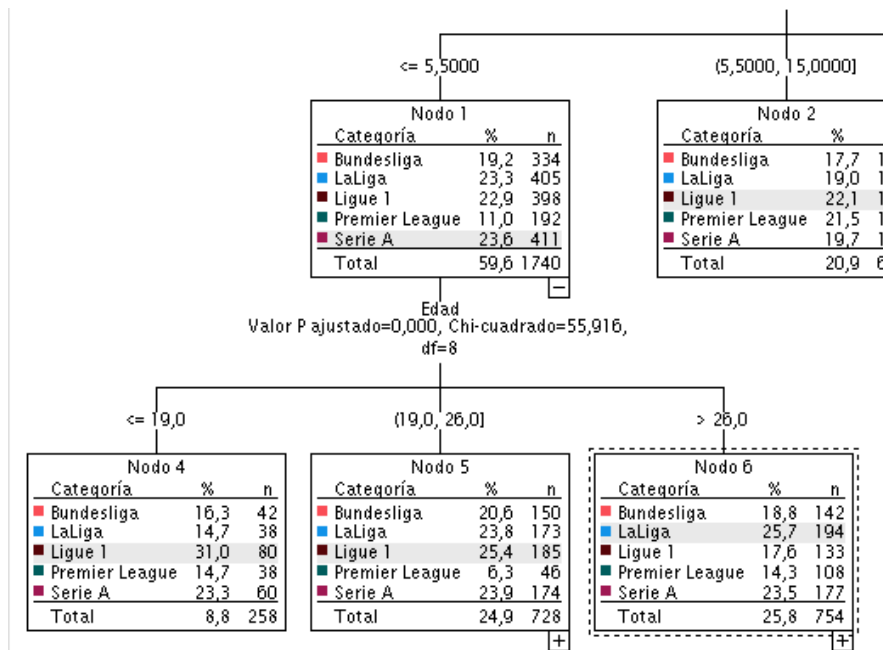


Ilustración 5: “Árbol de clasificación del nodo cuatro al seis”

Del nodo uno que está formado por los jugadores que tienen un “valor de mercado” igual o inferior a 5,5 millones de euros, que contiene mil setecientos cuarenta jugadores, se divide en tres nuevos nodos hijos, mediante la variable significativa “edad”, que lo divide en un grupo de menores de diecinueve años, los jugadores que tienen entre diecinueve y veintiséis años, incluido el último año, y el nodo seis contiene a los jugadores mayores de veintiséis años.

El nodo cuatro, está formado por los jugadores que tienen un “valor de mercado” inferior o igual a 5,5 millones con una “edad” menor a diecinueve años, en el cual se tienen doscientos cincuenta y ocho jugadores siendo la que destaca inferiormente LaLiga y la Premier League con un 14,7%, y destaca superiormente la Ligue 1 con un 33%. También destaca la Bundesliga con un 16,3% ya que tiene un porcentaje menor de jugadores que el 2,5% que se deja de cortesía, y la Serie A que tiene un 23,3% que supera el 2,5% que se deja de cortesía superiormente. Por lo tanto, todas las ligas están fuera del rango entre 17,5% y 22,5%.

El siguiente nodo es el 5, que contiene a los jugadores que tienen una “edad” de veinte a veintiséis años con un “valor de mercado” inferior o igual a 5,5 millones, donde hay setecientos veintiocho jugadores, destacando de inferiormente la Premier League con un

6,3%, mientras que LaLiga, la Serie A y la Ligue 1 que superan el 22,5%, teniendo la que menos ciento setenta y tres jugadores.

El último de estos tres nodos es el nodo número seis, que contiene a setecientos cincuenta y cuatro jugadores con más de veintiséis años y con un “valor de mercado” inferior a 5,5 millones. La liga inferior es la Premier League con un 14,3%, mientras que se tiene LaLiga con un 25,7% y la Serie A con 23,5%, siendo las dos ligas que superan el límite marcado y las que más destacan en este nodo.

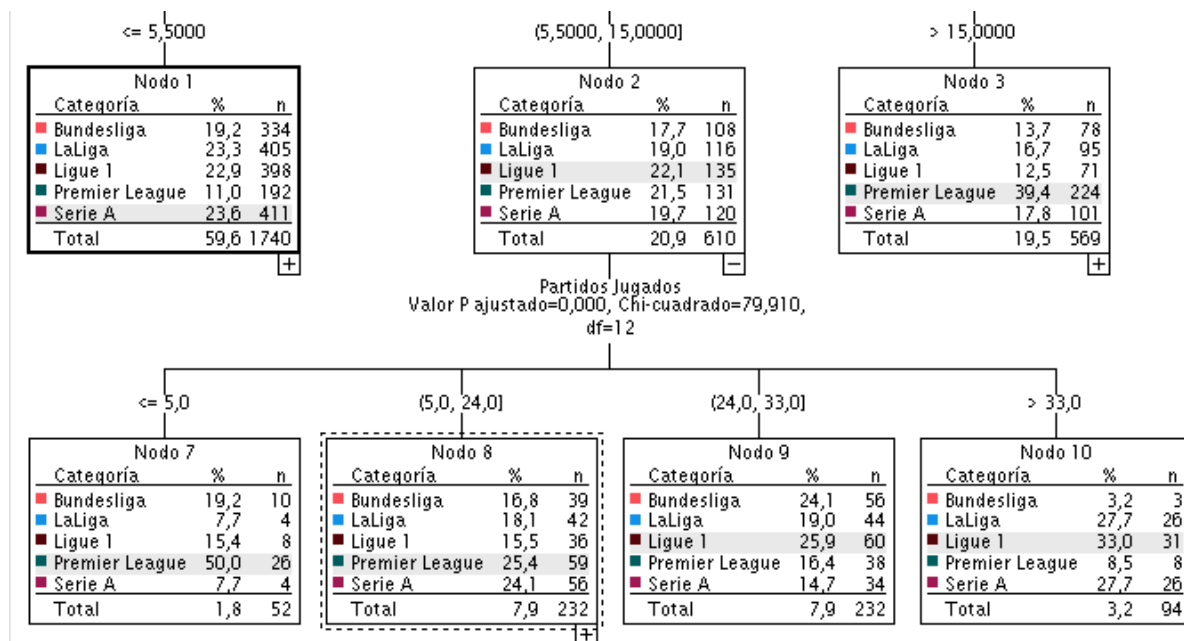


Ilustración 6: “Árbol de clasificación del nodo siete al diez”

Del nodo padre dos, que tiene como característica el “valor de mercado” de los jugadores, que tienen entre 5,5 millones y 15 millones de euros de “valor de mercado”, se obtienen cuatro nodos hijos, del nodo siete al nodo diez, que se dividen por la “cantidad de partidos jugados”.

El primero de estos nodos, el nodo siete, contiene cincuenta y dos jugadores de los seiscientos diez del nodo padre, que cumplen la característica del nodo padre y que han jugado cinco o menos partidos en esa temporada. Esos cincuenta y dos jugadores se dividen en las distintas ligas de manera asimétrica, la liga que tiene más jugadores es la Premier League, que cuenta con un 50% de los jugadores, que son veintiséis jugadores, mientras que si se habla de las otras ligas, tres de ellas están fuera de los límites marcados, que son LaLiga con cuatro jugadores que equivale a 7,7%, la misma cantidad de jugadores es con la que cuenta la Serie A, y la tercera de estas ligas es la Ligue 1 que cuenta con 15,4% del total del nodo, que es igual a ocho jugadores. La Bundesliga es la liga que se encuentra dentro de la normalidad, contando con diez jugadores de los cincuenta y dos del nodo.

El siguiente nodo es el ocho, que cuenta con doscientos treinta y dos jugadores con el “valor de mercado” determinado con anterioridad y que han jugado entre seis y veinticuatro partidos, ambos incluidos. Este nodo solo cuenta con una liga dentro de los límites, que cuenta con cuarenta y dos jugadores, que es LaLiga, mientras que en el resto de las ligas se puede apreciar que dos son superiores y otras dos son inferiores. Las dos inferiores son la Ligue 1 con

treinta y seis jugadores y la Bundesliga que tiene treinta y nueve. Las dos ligas superiores son la Serie A y la Premier League, conteniendo cincuenta y seis y cincuenta y nueve jugadores, respectivamente.

El nodo nueve cuenta también con doscientos treinta y dos jugadores que han jugado entre veinticinco partidos y treinta y tres, cumpliendo con la cualidad anterior. Tanto la Bundesliga como la Ligue 1 cuentan con mayor porcentaje que el límite que hemos marcado, llegando de este modo a contar con cincuenta y seis y sesenta jugadores cada una de ellas, mientras que, observando el límite inferior, otras dos ligas están con una cantidad de jugadores que tienen un porcentaje inferior al 17,5%, que son la Serie A con treinta y cuatro jugadores y la Premier League con treinta y ocho.

El último nodo hijo del nodo dos, es el nodo diez, que contiene noventa y cuatro jugadores que han jugado treinta y cuatro o más partidos, siendo el segundo nodo con menos jugadores de estos cuatro. Se puede apreciar una gran disparidad de porcentajes entre las cinco ligas observadas, teniendo un 3,2% de jugadores en la Bundesliga y un 8,5% en la Premier League, que son las dos ligas con menos jugadores. Mientras que, si se presta atención a las otras tres ligas, la que más porcentaje de jugadores que hay en una liga en este nodo, es la Ligue 1 con un 33%.

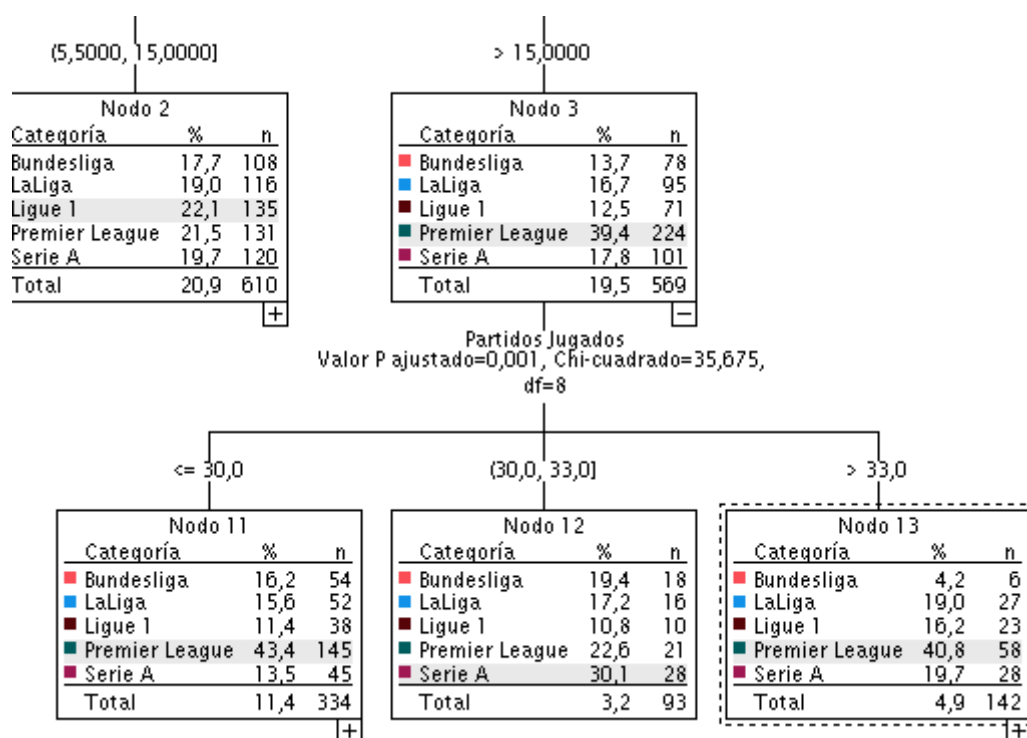


Ilustración 7: "Árbol de clasificación del nodo once al trece"

Los siguientes tres nodos que se obtienen provienen del nodo tres, que son los jugadores que tienen un valor superior a 15 millones, en el que destaca la Premier League con doscientos veinticuatro jugadores de los quinientos sesenta y nueve del nodo padre. La variable significativa que divide a estos jugadores en grupos vuelve a ser la "cantidad de partidos jugados".

El primero de estos tres nodos es el nodo once que cuenta con trescientos treinta y cuatro jugadores, que jugaron treinta o menos partidos. Teniendo en cuenta los porcentajes de cada

una de las ligas en el nodo tres, se puede decir que sigue una distribución similar, la Premier League sigue siendo la más numerosa con ciento cuarenta y cinco jugadores, seguida de lejos por el resto de las ligas, en las que se tienen cincuenta y cuatro jugadores en la Bundesliga, cincuenta y dos en LaLiga, treinta y ocho en la Ligue 1 y cuarenta y cinco en la Serie A. Como se puede apreciar la cantidad de jugadores de las cuatro ligas es muy bajo, estando estos cuatro porcentajes muy cerca entre ellos, con 16,2% la que más y 11,4% la que menos, que son la Bundesliga y la Ligue 1, respectivamente.

El nodo doce cuenta con solamente noventa y tres jugadores, en los que se agrupan los jugadores con un valor mayor a 15 millones y con treinta y uno, treinta y dos o treinta y tres partidos jugados en sus ligas. Se puede observar que la liga más numerosa no es la Premier League como cabía esperar, sino que es la Serie A con 30,1% que equivale a veintiocho jugadores. La Premier League es la otra liga que supera los límites, con veintiún jugadores. La única liga centrada es la Bundesliga con dieciocho jugadores y un porcentaje de 19,4%. Mientras que las otras dos ligas son inferiores al límite con un 17,2% LaLiga y 10,8% la Ligue 1.

El último nodo hijo del nodo tres es el nodo trece, que cuenta con ciento cuarenta y dos jugadores que cumplan con las características del nodo tres y que hayan jugado treinta y cuatro o más partidos. Este nodo tiene una distribución parecida a la del nodo tres, con un 40,8% de jugadores formando parte de la Premier League, un 19,7% y un 19% en la Serie A y en LaLiga, respectivamente. Las otras dos ligas cuentan con 16,2% la Ligue 1 que es equivalente a veintitrés jugadores, y la última liga es la Bundesliga con solamente seis jugadores, que es únicamente un 4,2% de los ciento cuarenta y dos jugadores del nodo.

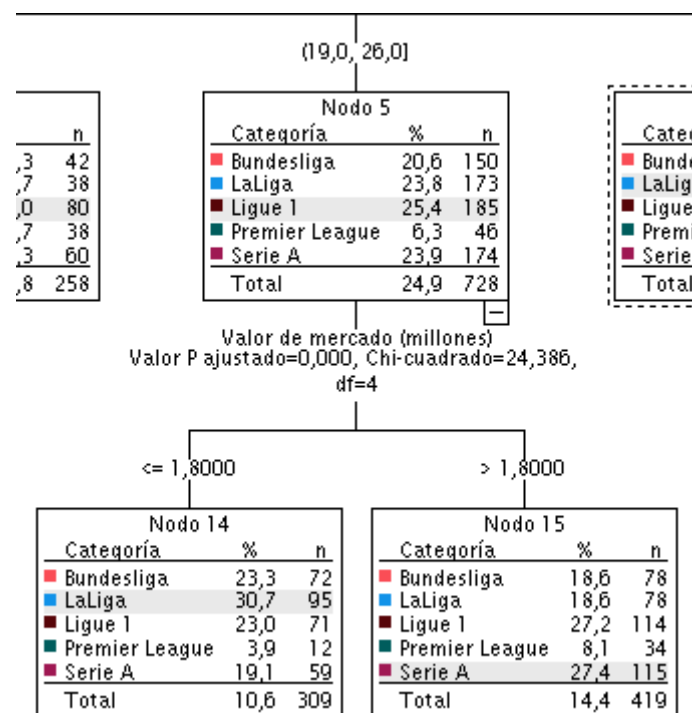


Ilustración 8: “Árbol de clasificación del nodo catorce al quince”

Los siguientes dos nodos que se obtienen, el catorce y el quince, son los que se forman a través del nodo cinco, que es nodo hijo del nodo uno. Esta secuencia de nodos indica que es un grupo que tiene jugadores con un “valor de mercado” de hasta 5,5 millones de euros, con jugadores

con una “edad” entre veinte y veintiséis años, y que se divide en un grupo de jugadores con un “valor de mercado” inferior o igual a 1,8 millones y otro con los superiores a 1,8 millones de euros, ya que la variable significativa que divide el grupo es el “valor de mercado”.

El nodo catorce que cuenta con trescientos nueve jugadores, en el que hay dos ligas principales, una por lo bajo y otra por lo alto. La liga que destaca por la poca cantidad de jugadores, con solo doce jugadores, que equivale a un 3,9%, es la Premier League. Esto es normal, ya que en el nodo cinco únicamente cuenta con cuarenta y seis jugadores de los setecientos veintiocho. Mientras que la liga que más jugadores tiene es LaLiga con noventa y cinco jugadores, superando el 30%.

El nodo quince tiene cuatrocientos diecinueve jugadores que tienen de “valor de mercado” entre 1,8 millones y 5,5 millones de euro, ya que el valor superior lo marca el nodo uno. Prácticamente cuenta con tres grupos, uno de ellos formado por la Premier League con 8,1% de los jugadores, el segundo de estos está formado por dos ligas que cuentan con un 18,6%, por lo que están dentro de los límites marcados, que son la Bundesliga y LaLiga, y las dos ligas que quedan son la Ligue 1 y la Serie A que están formadas por el 27,2% y el 27,4% cada una de ellas, lo que equivale a ciento catorce y ciento quince jugadores.

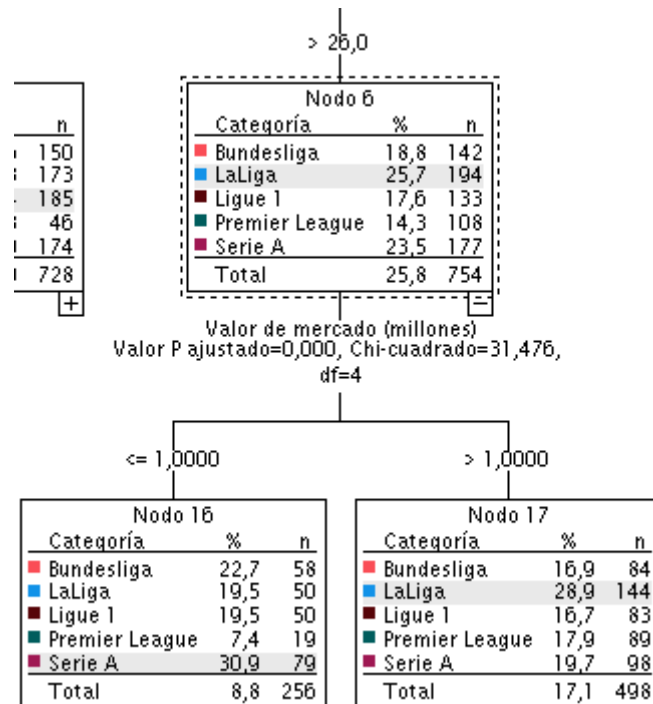


Ilustración 9: “Árbol de clasificación del nodo dieciséis al diecisiete”

Los siguientes nodos que se crean son el dieciséis y el diecisiete, formado por los jugadores que se encuentran en el nodo seis, es decir, que tienen más de veintiséis años y un “valor de mercado” de hasta 5,5 millones. Estos nodos también tienen como variable significativa el “valor de mercado”, dividiéndose en los que valen hasta 1 millón de euros y los que valen más.

En el nodo dieciséis, donde se encuentran los jugadores que tienen un “valor de mercado” de hasta un millón, hay doscientos cincuenta y seis jugadores que se dividen en 30,90% en la Serie A; mientras que LaLiga, la Ligue 1 y la Bundesliga muy próximas a los límites, las dos primeras con cincuenta jugadores y la tercera con cincuenta y ocho. Y la liga que queda, que

es la Premier League que tiene únicamente 7,4% que son diecinueve jugadores, siendo la que menos jugadores tiene con estas características.

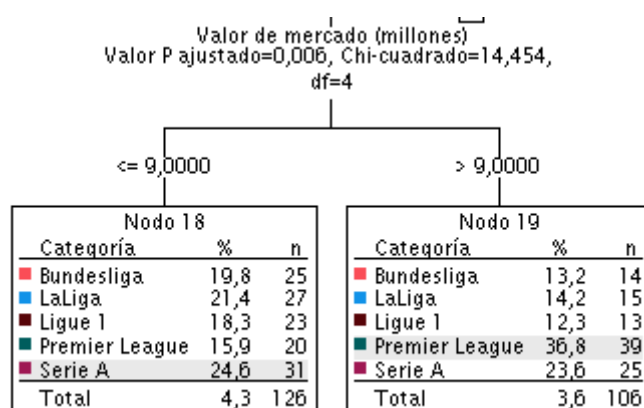


Ilustración 10: “Árbol de clasificación del nodo dieciocho al diecinueve”

Los dos siguientes nodos son el dieciocho y el diecinueve, que provienen de los nodos dos y ocho, el primero de estos contiene a los jugadores que tiene un “valor de mercado” desde 5,5 millones hasta 9 millones de euros, con más de cinco “partidos jugados” hasta veintinueve “partidos jugados”, mientras que el segundo cuenta con los jugadores que cuentan con el mismo intervalo de “partidos jugados” que el nodo anterior, pero con un “valor de mercado” entre 9 millones a 15 millones de euros.

Observando el nodo dieciocho se sabe que tiene ciento veintiséis jugadores, de los cuales el 24,6% forman parte de la Serie A y el 15,9% forman parte de la Premier League, siendo estas dos ligas las únicas que están fuera de los límites marcados. El resto de las ligas están entre el 18,3% y el 21,4% de jugadores del nodo, esto quiere decir que tienen entre veintitrés y veintisiete jugadores.

En el diecinueve, están ciento seis jugadores que cumplen las condiciones mencionadas con anterioridad, dividiéndose estos de manera más notoria entre la Premier League y la Serie A, teniendo treinta y nueve jugadores la primera y veinticinco la segunda. Mientras que, si se presta atención en la Ligue 1, la Bundesliga y LaLiga, tienen trece, catorce y quince, respectivamente.

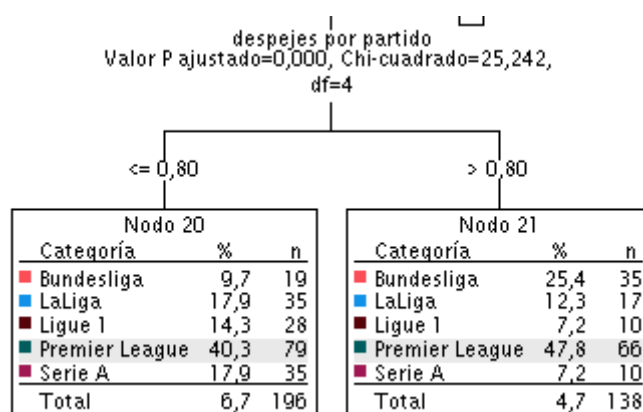


Ilustración 11: “Árbol de clasificación del nodo veinte al veintiuno”

Los dos siguientes nodos son nodos hijos del nodo once que a su vez es nodo hijo del nodo tres, esto quiere decir que inicialmente ambos tienen las características que están marcadas en el nodo once, que son las de las variables “valor de mercado”, que es el grupo que cuenta con los jugadores que tienen un “valor de mercado” mayor a los 15 millones de euros y que han jugado hasta treinta partidos en la liga. Estos trescientos treinta y cuatro jugadores que se tienen en el nodo once se dividen en dos grupos mediante la variable significativa que es la de “despejes por partido”.

El nodo veinte es el más numeroso de los dos y cuenta con ciento noventa y seis jugadores que realizan hasta 0,8 “despejes por partido”. El 40,3% que equivale a setenta y nueve jugadores, forman parte de la Premier League, mientras que observando las ligas que son menos numerosas son la Bundesliga con diecinueve jugadores y la Ligue 1 con veintiocho jugadores. Las otras dos ligas cuentan ambas con treinta y cinco jugadores, es decir un 17,9%.

El nodo veintiuno contiene a los jugadores que realizan más de 0,8 “despejes por partido”. Al igual que ocurre en el nodo veinte, la liga que más destaca es la Premier League con un 47,8% de los ciento treinta y ocho jugadores que forman el nodo, lo que quiere decir que tiene sesenta y seis jugadores, mientras que hay tres ligas que son mucho menores en cantidad de jugadores como son LaLiga con solo diecisiete jugadores y la Ligue 1 y la Serie A que cuentan únicamente con diez jugadores en este grupo.

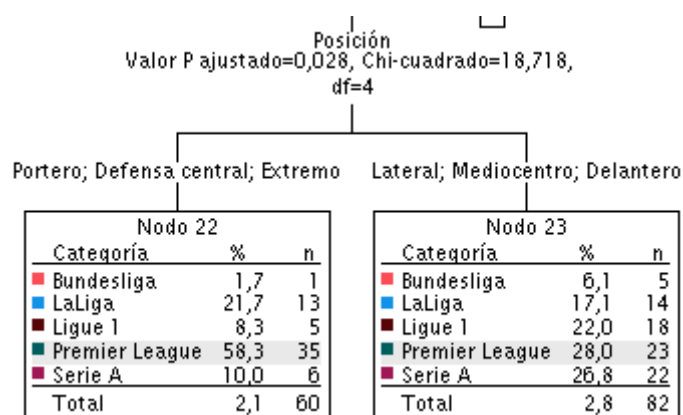


Ilustración 12: “Árbol de clasificación del nodo veintidós al veintitrés”

Los dos últimos nodos que hay en el árbol son los nodos veintidós y veintitrés que tienen como variable significativa la “posición” de los jugadores en el terreno de juego, pero teniendo en cuenta que provienen de los nodos tres y trece, diciendo estos que son jugadores con un “valor de mercado” mayor que 15 millones y más de 33 “partidos jugados”.

El nodo veintidós cuenta con sesenta jugadores que han jugado más de treinta y tres partidos y que valen más de 15 millones, y a mayores juegan en las “posiciones” de portero, defensa central y extremo, en la que destaca principalmente en cantidad la Premier League, ya que es la liga que cuenta con más jugadores en el nodo padre, tiene un 58,3% de los jugadores del nodo veintidós.

Por otra parte, en el nodo veintidós se puede observar que la Bundesliga tiene escasamente un jugador, la Ligue 1 tiene cinco, la Serie A tiene seis, y LaLiga cuenta con trece jugadores.

En el último nodo, el nodo veintitrés, se encuentran ochenta y dos jugadores que juegan de laterales, mediocentros o delanteros, que superaron los treinta y tres “partidos jugados” y que tienen un “valor de mercado” mayor a los 15 millones de euros. La liga que más destaca en este nodo es la Bundesliga, pero destaca por demérito contando únicamente con cinco jugadores, mientras que LaLiga, la Serie A y la Ligue 1 están más centrados, con catorce, veintidós y dieciocho jugadores respectivamente. Como en el anterior nodo, la liga que más destaca es la Premier League con veintitrés jugadores que es equivalente al 28% de los ochenta y dos jugadores.

6.2.2. ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales (ACP) se centra en la representación de la variabilidad de los datos en componentes principales propiciando así la reducción de las variables, en unas variables que no estén correlacionadas entre ellas.

Primeramente, se realiza la prueba de esfericidad de Bartlett para comprobar si la matriz de correlación de las variables de la base de datos es similar a una matriz identidad y comprobando las relaciones lineales entre ellas. Para comprobarlo hay que fijarse en el p-valor y corroborar que tiene un p-valor significativo, en caso de ser menor de 0,05, y cuanto más pequeño sea el p-valor más significativa serán las relaciones lineales entre las variables.

TEST DE ESFERICIDAD DE BARTLETT	
Ji-cuadrado de Bartlett	851017
Grados de libertad	29
p-valor	$<2,2 \cdot 10^{-16}$

Tabla 1: “Test de Esfericidad de Bartlett”

Se puede observar en la tabla anterior relacionada con el test de esfericidad de Bartlett que el p-valor que trata las relaciones entre las variables de la base de datos es altamente significativo, ya que está muy por debajo del 0,05, de este modo se sabe que es correcta la utilización del análisis de componentes principales para la base de datos con la que se trabaja, pudiendo hallar así un mínimo de componentes que definan la base de datos a través de las relaciones que se produzcan en cada uno de ellos.

Lo siguiente que se realiza es el análisis de componentes principales en el que se hallaron únicamente cinco componentes que se pueden apreciar en la siguiente tabla.

En la siguiente tabla se tienen las treinta variables que forman la base de datos, con sus coeficientes en cada uno de los componentes principales. En la tabla están marcadas las variables que son más significativas en cada uno de los componentes.

VARIABLE	CP1	CP2	CP3	CP4	CP5
Posición	0,07	0,34	-0,19	0,18	-0,07
Edad	0,06	-0,10	0,23	0,02	-0,46
Valor de mercado (millones)	0,21	0,04	0,18	-0,01	0,38
Liga	-0,01	0,01	-0,04	-0,06	0,13
Partidos jugados	0,26	-0,09	0,14	0,04	-0,28
Minutos	0,26	-0,16	0,19	0,02	-0,21
Goles	0,23	0,17	0,22	0,22	0,07
Asistencias	0,25	0,12	0,15	-0,16	0,01
Tarjeta amarilla	0,21	-0,16	-0,09	0,07	-0,30
Tarjeta roja	0,06	-0,09	-0,02	0,07	-0,21
Tiros por partido	0,25	0,20	0,06	0,20	0,07
Acierto pase %	0,07	-0,08	-0,11	-0,12	0,22
Duelos aéreos ganados por partido	0,12	-0,20	-0,03	0,44	0,11
MVP	0,24	0,07	0,25	0,02	0,09
Valoración media	0,27	-0,08	0,15	-0,04	0,19
Entradas por partido	0,17	-0,21	-0,36	-0,12	0,05
Intercepciones por partido	0,13	-0,31	-0,23	0,00	0,12
Faltas por partido	0,19	-0,08	-0,34	0,16	-0,17
Fuera de juegos ganados por partido	0,03	-0,30	-0,02	0,17	0,27
Despejes por partido	0,05	-0,36	0,01	0,20	0,15
Regateado por partido	0,19	-0,05	-0,34	-0,19	-0,15
Pases clave por partido	0,26	0,14	0,00	-0,27	0,01
Regates por partido	0,22	0,14	-0,15	-0,13	0,13
Faltas a favor por partido	0,23	0,06	-0,20	0,04	-0,06
Fuera de juego por partido	0,12	0,21	0,07	0,36	-0,02
Controles malos por partido	0,23	0,21	-0,16	0,15	-0,01
Pases por partido	0,18	-0,30	0,08	-0,13	0,11
Centros por partido	0,17	0,08	-0,03	-0,39	-0,08
Balones largos por partido	0,03	-0,26	0,36	-0,17	-0,12
Pases al hueco por partido	0,14	0,08	0,05	-0,21	0,19

Tabla 2: "Componentes del Análisis de Componentes Principales"

Como se puede observar en la tabla anterior, hay cinco componentes que sean obtenido de la realización del análisis en el que hay que centrarse en las variables que más influyen en cada uno de ellos, ya sea con valores bajos de manera negativa o con valores altos de manera positiva.

Se puede ver en la tabla que variables están marcadas en cada uno de los componentes ya que son las que más les influyen, siendo marcadas las que tengan valores superiores a 0,25 o valores inferiores a -0,25.

En el primer componente, que tiene una desviación estándar de 2,85, se observa que las variables que más sobresalen son los "partidos jugados", los "minutos", las "asistencias", los "tiros por partido", la "valoración media", los "pases clave por partido". Estas seis variables son las que más influyen en este componente, todas ellas de manera directamente proporcional, pudiéndose observar que la "cantidad de partidos jugados" y los "minutos" son

importantes y tienen relación tanto con el número de “asistencias” y con los “tiros por partido”. También se observa la relación con la “valoración media”, juntando de esta forma cinco variables relacionadas con las características generales de los jugadores y a mayores los “pases clave por partido”, que es normal, ya que una de las otras variables significativas son las “asistencias”.

Por lo tanto, este primer componente puede estar referida al tiempo jugado y a la aparición ofensiva del jugador en el último tercio del campo, ya que una buena aparición ofensiva y la cantidad de tiempo jugado repercute en la “valoración media”.

En el segundo componente, que cuenta con una desviación estándar de 2,25, se observa que las variables que más influyen son, de manera directamente proporcional, la “posición”, y de manera inversamente proporcional, las “intercepciones por partido”, los “fuera de juegos ganados por partido”, los “despejes por partido”, los “pases por partido” y los “balones largos por partido”. Se puede pensar que todas las variables que son inversamente proporcionales son de un carácter defensivo, ya que todas ellas forman parte de las características defensivas, pero al ser inversamente proporcionales es más probable que se estén adecuando a jugadores con un espacio de juego más ofensivo, que suelen ser los que menos realicen esas acciones y a las que se le suman los “balones largos por partido” y los “pases por partido”, ya que los jugadores más ofensivos suelen ser los que menos contacto tengan con la pelota.

Por lo tanto, este segundo componente puede estar referido a jugadores ofensivos, como pueden ser los delanteros o los extremos, que tienen un menor trabajo defensivo por lo que es normal que las variables sean inversamente proporcionales.

En el tercer componente, que tiene una desviación estándar de 1,47, se tiene cinco variables influyentes, que son el “MVP” y los “balones largos por partido” de manera directamente proporcional, y las “entradas por partido”, las “faltas por partido” y las “veces que es regateado por partido” de manera inversamente proporcional. Se observa que los jugadores que menos realizan entradas por partido, faltas o son regateados son los jugadores más defensivos ya que dada la zona en la que se encuentran no pueden realizar acciones arriesgadas para el juego. También es una posición en las que se realizan muchos balones largos por partidos ya que muchos equipos tienen un juego directo, que se inicia con el desplazamiento de balón desde la línea defensiva.

Por lo tanto, este tercer componente puede referirse a los defensas centrales de los equipos, ya que no suelen realizar una gran cantidad de entradas ni de faltas por ser los últimos jugadores previos al portero por lo que acarrearían una tarjeta roja muchas de esas faltas. Y dado que muchos equipos tienen un juego directo, destacan sus centrales con los desplazamientos de balón en largo.

En el cuarto componente, que cuenta con una desviación estándar de 1,35, que tiene cuatro variables significativas, dos directamente proporcionales y dos inversamente proporcionales. Las dos directamente proporcionales son los “duelos aéreos ganados por partido” y los “fuera de juego por partido”, mientras que los indirectamente proporcionales son los “pases clave por partido” y los “centros por partido”. Observando las cuatro variables, se puede apreciar que tienen algo en común muy claro, un jugador que gana muchos duelos aéreos por partido no puede ser el que realice los pases clave o el que realice los centros al área, porque será habitualmente el que los remate. Añadiendo a esto que la variable que se añade de manera directamente proporcional es la de los “fuera de juegos por partido” que le pitan en contra, quiere decir esto que el propio jugador se encuentre en posición antirreglamentaria.

Por lo tanto, este componente puede referirse a los jugadores ofensivos de los equipos, más concretamente a los delanteros centros de los equipos, ya que suelen ser los que más caen en fuera de juego por partido, los que más duelos aéreos ganan por partido, y los que menos pases clave y centros hacen por partido.

En el quinto y último componente, que tiene una desviación estándar de 1,13, se tienen cinco variables significativas, tres de ellas inversamente proporcionales y las otras dos directamente proporcionales. Las dos directamente proporcionales son los “fuera de juegos ganados por partido” y el “valor de mercado”, mientras que las inversamente proporcionales son los “partidos jugados”, la “edad” y las “tarjetas amarillas”. Este componente se puede ver cómo es un poco más confuso ya que las variables que son directamente proporcionales son los “fuera de juegos ganados por partido” y el “valor de mercado”, podría pensarse que tendrían una relación con los jugadores más defensivos, como serían los laterales y los propios defensas, pero si se presta atención a las variables inversamente proporcionales aparecen la “edad”, los “partidos jugados” y las “tarjetas amarillas”, por lo que tendrían que ser jugadores que tengan una edad baja, con pocos partidos jugados y con pocas tarjetas amarilla, pero cumpliendo también lo anterior.

Por lo tanto, este componente puede estar referido a los jugadores que ocupan la posición de lateral o de defensa central suplente, que sean jóvenes y tengan una gran proyección dado que el “valor de mercado” es una variable directamente proporcional, y al jugar pocos partidos es más fácil que reciban menos tarjetas amarillas.

6.2.3. ANÁLISIS FACTORIAL

La siguiente técnica que se trata es el análisis factorial, que consiste en la reducción de la dimensionalidad de la base de datos en ciertos factores que explican las correlaciones que hay entre las variables iniciales, facilitando de esta forma tanto el análisis de los datos como la interpretación de estos, ya que se pueden observar patrones y relaciones lineales que hay entre las variables iniciales.

Para saber si se debe usar el análisis factorial es necesario realizar el test de Kaiser-Meyer-Olkin, también conocido como KMO, que indica la proporción de varianza compartida entre las variables. Cuanto más cerca esté el valor del KMO a uno es más adecuado hacer el Análisis Factorial.

TEST DE KAISER-MEYER-OLKIN (KMO)	
KMO	0,89

Tabla 3: “Test de Kaiser-Meyer-Olkin”

Como se puede observar en la tabla anterior, la prueba de Kaiser-Meyer-Olkin tiene un valor de 0,89, lo cual indica que es muy adecuada la realización del Análisis Factorial para la reducción de la dimensionalidad de la base de datos, uniendo en factores a las variables que están relacionadas e influyen de mayor manera en cada uno de los factores.

VARIABLE	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Posición	0,21		0,80		-0,25
Edad	0,13		-0,16	0,21	0,14
Valor de mercado (millones)	0,57			0,18	0,19
Liga					
Partidos jugados	0,34	0,23		0,86	
Minutos	0,36	0,25	-0,12	0,86	0,21
Goles	0,63	-0,16	0,43	0,27	0,24
Asistencias	0,73	0,11	0,12	0,23	
Tarjeta amarilla	0,14	0,46		0,52	0,18
Tarjeta roja		0,14		0,17	0,11
Tiros por partido	0,67		0,55	0,15	0,20
Acierto pase %		0,23			
Duelos aéreos ganados por partido		0,25		0,16	0,72
MVP	0,67		0,11	0,27	0,19
Valoración media	0,66	0,27		0,26	0,34
Entradas por partido		0,82		0,10	0,15
Intercepciones por partido		0,67	-0,22	0,13	0,40
Faltas por partido	0,11	0,58	0,30	0,16	0,19
Fuera de juegos ganados por partido	-0,14	0,27	-0,30		0,55
Despejes por partido	-0,17	0,33	-0,37	0,16	0,68
Regateado por partido	0,23	0,67	0,11	0,12	-0,10
Pases clave por partido	0,81	0,29	0,14		-0,21
Regates por partido	0,54	0,32	0,30		-0,13
Faltas a favor por partido	0,39	0,40	0,33	0,19	
Fuera de juegos por partido	0,31	-0,17	0,52		0,14
Controles malos por partido	0,47	0,20	0,62	0,16	
Pases por partido	0,28	0,45	-0,48	0,29	0,38
Centros por partido	0,58	0,28			-0,30
Balones largos por partido			-0,66	0,31	0,22
Pases al hueco por partido	0,47				

Tabla 4: "Factores del Análisis Factorial"

En el caso del análisis factorial también se han cogido cinco factores en los que se busca una reducción de la dimensionalidad en búsqueda de explicar nuestra base de datos en estos factores. El motivo por el que se han tomado cinco factores es porque la varianza acumulada supera el 55% y realizándolo con más factores la proporción de varianza que aportaba era mínima.

Las variables que se toman como significativas son aquellas que tienen un valor inferior a -0,5 o superior a 0,5, y que están marcadas en rojo en la tabla de factores obtenidos.

El primer factor cuenta con nueve variables significativas que son directamente proporcionales y como mencionamos con anterioridad, mayores de 0,5. Las variables son el "valor de mercado", los "goles", las "asistencias", los "tiros por partido", los "MVP", la "valoración media", los "pases clave por partido", los "regates por partido" y los "centros por

partido". Se puede apreciar que las que más destacan de las nueve son los "pases clave por partido", las "asistencias", los "tiros por partido", los "MVP", la "valoración media" y los "goles", pudiendo sacar como conclusión que puede tratarse de un jugador ofensivo, que suelen ser los que estén más relacionados con estas cualidades, y también por el motivo de que un jugador con "goles" y "asistencias" suele tener un "valoración media" y una cantidad de "MVP" mayor que otros jugadores, y a raíz de esto también aumenta su "valor de mercado", al fin y al cabo, el deporte consiste en el meter más goles que el rival, por lo que un jugador que destaque en estas cualidades tiene un mayor "valor de mercado".

Por lo tanto, el primer factor trata de un jugador ofensivo y al tener también como variable significativa los "centros por partido" y los "regates por partido", se puede pensar que es un jugador de banda, posiblemente un extremo.

El segundo factor cuenta con cuatro variables significativas, todas ellas directamente proporcionales, siendo las "entradas por partido", las "intercepciones por partido", las "faltas por partido" y las "veces que es regateado por partido". Se puede observar que todas ellas tienen un carácter más bien defensivo que ofensivo, tratándose todas ellas de acciones sin pelota. Las "entradas por partido" suelen ser realizadas por los centrales, los laterales y los mediocentros, también en las "intercepciones por partido" suelen destacar más los centrales, los laterales y los mediocentros. En el caso de las "faltas por partido" todos los jugadores pueden estar muy similar, porque al fin y al cabo todos realizan faltas, pero podemos pensar que los jugadores que más las pueden realizar son los mediocentros. Las "veces que son regateados por partido" puede ser más alta en jugadores del mediocentro o laterales, dado que pueden ser lugares donde se encaren más.

Por lo tanto, uniendo lo que se sabe de las variables, se podría pensar que la posición que mejor está definida con estas variables es la de mediocentro.

El tercer factor cuenta con cinco variables significativas que son cuatro directamente proporcionales y una inversamente proporcional. Las cuatro que son directamente proporcionales son la "posición", "tiros por partido", "fuera de juego por partido", "controles malos por partido" e indirectamente proporcional los "balones largos por partido". Si se presta atención a tres de las cinco variables, se puede observar que los jugadores que realizan pocos "balones largos por partido" son los jugadores ofensivos que no tienen a nadie por delante, también son estos los que realizan más "tiros por partido" y a los que les pitan más "fuera de juego". Y si nos centramos en la característica de los "controles malos por partido", suelen ocurrir más en jugadores ofensivos que defensivos por el hecho de la situación en el campo es más lejana a la portería propia.

Por lo tanto, echando un vistazo a las variables significativas del tercer factor se puede pensar que se esté hablando de jugadores con un carácter más ofensivo por lo que se está a tratar con jugadores que tengan la posición de extremos o de delanteros.

El cuarto factor tiene tres variables significativas, que son la "cantidad de partidos jugados", la "cantidad de minutos jugados" y las "tarjetas amarillas". En este caso la única característica que podría decantar hacia una posición es la de las "tarjetas amarillas" que suelen sacárselas más a los jugadores defensivos, como defensas centrales, laterales o mediocentros, pero no se puede saber nada más sobre las posiciones, porque los "minutos" y los "partidos jugados" no son diferenciales entre unas posiciones y otras.

Por lo tanto, en el cuarto factor no se puede pensar acerca de una única posición, así que denominaremos que es un factor que trata a los jugadores del mediocentro hacia atrás, que son el mediocentro, los laterales y los defensas centrales.

El quinto y último factor tiene como variables significativas los “duelos aéreos ganados por partido”, los “fuera de juego ganados por partido” y los “despejes por partido”. Todas estas variables son de jugadores con una posición más bien defensiva, obviando los “duelos aéreos ganados por partido” que también puede ser una característica de los delanteros. Pero observando las tres en conjunto, y que todas ellas son directamente proporcionales, se observa que está muy relacionado con los jugadores defensivos.

Por lo tanto, el quinto factor se puede decir que trata de los jugadores que compiten en la posición de defensa, ya que son los que más duelos aéreos suelen ganar por partido, son los que más intercepciones realizan por partido y son los que realizan más despejes por partido.

7. CONCLUSIONES

Como se pudo observar a lo largo del trabajo, la minería de datos y los deportes han comenzado a ir de la mano desde hace varias décadas, y esta relación parece que va a ser duradera y prolífera para ambas.

La base de datos obtenida ha terminado siendo muy completa, contando con una gran cantidad de jugadores que formaban parte de las cinco ligas mencionadas. También estaba formada por muchas variables que ayudaron a realizar mejor el análisis estadístico.

El árbol de clasificación cuenta con siete variables predictoras, que son: el “valor de mercado”, la “edad”, los “partidos jugados”, “despejes por partido” y la “posición”. En éste se han obtenido quince nodos terminales que dividiremos por la variable “liga”, destacando cada una por mérito o por demérito.

- La liga inglesa o Premier League ha destacado en los nodos siete, diecinueve, veinte, veintiuno, veintidós y veintitrés. Se aprecia que los nodos en los que más destaca son los que en la variable “valor de mercado” pertenecen al grupo mayor.
- La liga francesa o Ligue 1 es referencia de los nodos cuatro, nueve y quince. Esta liga tiene dos nodos que tienen relación, que son el cuatro y el quince, dado que ambos cuentan con jugadores con un “valor de mercado” inferior o igual a los 5,5 millones.
- La liga alemana o Bundesliga sobresale únicamente en el nodo diez. Esta mención es por demérito ya que cuenta con un número muy bajo de jugadores que cumplan las variables del nodo.
- La liga italiana o Serie A se distingue en los nodos doce, quince, dieciséis, dieciocho y veintitrés. Esta liga tiene una peculiaridad en cuestión a sus nodos ya que destaca en cuatro rangos distintos de la variable “valor de mercado”.
- La liga española o LaLiga despunta en los nodos catorce y diecisiete. Éstos tienen en común que comparten el nodo padre uno, y que las variables que les afectan son la “edad” y el “valor de mercado”, pero en rangos diferentes.

El método que se observará a continuación es el análisis de componentes principales, del que se han obtenido cinco componentes principales. De éstas obtendremos una respuesta o varias para la variable “posición”, dependiendo de las variables que formen cada uno de los componentes.

- Los delanteros son representados por el segundo y el cuarto componente principal.
- Los extremos están descritos en el primero y en el segundo componente.
- Los mediocentros ven plasmadas sus cualidades en el primer componente.
- Los laterales están representados por el quinto componente.
- Los defensas centrales son descritos por el tercer y el quinto componente.
- Los porteros son la única posición que no está descrita en ningún componente, o que puede estarlo, pero no es representativo.

El método que se trata de seguido es el análisis factorial, donde se obtuvieron cinco factores. De éstos se adquieren respuestas para la variable “posición”, determinadas por las variables que afectan a los factores.

- Los delanteros están representados en el tercer factor.
- Los extremos están plasmados en el primero y en el tercer factor.
- Los mediocentros se encuentran descritos en el segundo y cuarto factor.

- Los laterales se describen en el cuarto y el quinto factor.
- Los defensas centrales se encuentran representados en el cuarto y quinto factor.
- Los porteros no están descritos en ningún factor, o puede que sí pero no sean estas variables significativas para su posición.

Una vez finalizado el trabajo elaborado anteriormente, se puede abstraer que el deporte, ya sea particularmente el fútbol o sea otro, ha comenzado a ir de la mano con el mundo de los datos y de la estadística. Pudiendo observar como por medio de variables y distintos análisis se pueden describir las distintas posiciones de juego o las distintas características comunes de los jugadores que forman parte de una liga.

8. BIBLIOGRAFÍA

8.1. LIBROS Y ARTÍCULOS

- Aguilar, L. J. (2014). Análisis de grandes volúmenes de datos en organizaciones. Alfaomega Grupo Editor.
- Benedek, E. (2006). FÚTBOL INFANTIL. Editorial Paidotribo.
- Bueno, J. A., & Mateo, M. Á. (2012). Historia del fútbol. EDAF.
- Fernández, J., & Bornn, L. (2021). SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer (Vol. 12461, pp. 491-506). https://doi.org/10.1007/978-3-030-67670-4_30
- Ferri Ramírez, C., Ramírez Quintana, M. J., & Hernández Orallo, J. (2004). Introducción a la minería de datos. Pearson Madrid.
- García, N. G., Londoño, A. T., Villardón, M. P. G., & Librero, A. B. N. (2015). "ANÁLISIS DE COMPONENTES PRINCIPALES SPARSE.
- Goldblatt, D. (2007). The Ball is Round: A Global History of Football. Penguin UK.
- González Campos, G., Campos Mesa, M. D. C., & Romero Granados, S. (2015). Análisis de la influencia de la evaluación del rendimiento en jugadores de un equipo de fútbol (Influence of the performance evaluation analysis of the football team players). Retos, 25, 85-89. <https://doi.org/10.47197/retos.v0i25.34486>
- Jariego, I. M., & Gómez-Millán, M. R. B. (2013). Análisis de las redes de distribución de balón en fútbol: pases de juego y pases de adaptación. Redes. Revista hispana para el análisis de redes sociales, 24(2), 135-155. <https://raco.cat/index.php/Redes/article/view/274762>
- López-Roldán, P., & Fachelli, S. (2015). METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA.
- Martínez, D. P. (2021). Aplicación de la estadística en el ámbito deportivo.
- Richard's, M. M., Solanas, A., & Ledesma, R. D. (2008). Técnicas estadísticas de clasificación: un estudio comparativo y aplicado. Psicothema, 20(4), 863-871.
- Sumpter, D. (2016). Soccermatics: Mathematical Adventures in the Beautiful Game. Bloomsbury Publishing Plc.
- Wilson, J. (2013). Inverting the pyramid: the history of soccer tactics. Bold Type Books.

8.2. PÁGINAS WEB Y PDF

- arbolesdecision.pdf. (s. f.). <https://www.uv.es/mlejarza/actuariales/tam/arbolesdecision.pdf>
- Así es el Big Data del Liverpool, el equipo que lo importó de los Red Sox. (2019, octubre 10). Estadio Deportivo. <https://www.estadiodeportivo.com/futbol-internacional/2019/10/10/big-data-liverpool-equipo-importo/246525.html>
- Cuando el fútbol es matemático. (2018, noviembre 14). <https://marcetfootball.com/es/futbol-y-estadistica/>
- Estadísticas de Fútbol. (s. f.). <https://es.whoscored.com/statistics>
- Fichajes de fútbol, rumores, valores de mercado y noticias. (s. f.). https://www.transfermarkt.es/investigacion_estadistica.pdf. (2021). https://revista.giepm.com/wp-content/uploads/investigacion_estadistica.pdf
- Revista-APF35-Completa.pdf. (2020). <https://futbolpf.org/wp-content/uploads/2020/12/Revista-APF35-Completa.pdf#page=32>

Salgueiros, C. C. (2023, noviembre 8). Historia del fútbol: sus orígenes y actualidad. Unisport.
<https://unisport.es/futbol-historia/>
TFM_MAADM_Ramos_Hernandez_Sergio.pdf. (2015)
https://gredos.usal.es/bitstream/handle/10366/128233/TFM_MAADM_Ramos_Hernandez_Sergio.pdf?sequence=4