VNiVERSiDAD D SALAMANCA

GRADO EN ESTADÍSTICA

EXPLOTACIÓN ESTADÍSTICA DE ALMACENES DE DATOS DE FÚTBOL

Trabajo Fin de Grado

Autor: Roberto Marcos Rodríguez

Tutora: María Teresa Cabero Morán

Salamanca, 2021

VNiVERSiDAD D SALAMANCA

GRADO EN ESTADÍSTICA

EXPLOTACIÓN ESTADÍSTICA DE ALMACENES DE DATOS DE FÚTBOL

Trabajo Fin de Grado

Firmado por CABERO MORAN MARIA TERESA - 07867178M el día 07/07/2021 con un certificado emitido por AC FNMT Usuarios

Fdo.: Roberto Marcos Rodríguez

Fdo.: María Teresa Cabero Morán

Salamanca, 2021

Abstract

Every year, the technological advance arrives and expands to more areas. We can see this advance if we take a look around us, we are surrounded by technology. The sciencie is in all parts of our lives especially with the arrival of the internet, probably one of the most sofisticated tool that ever created by the humanity. Because of this, every person use this tool every day of his life and because of this, the people have started to be treated like data by the companies, including the goverment, in order to understand the attitude and the behavior of all of us.

One of this areas is the sports. The sport is a big business around the Earth, it is global, and there are so much money behind this. Also the sport is starting to get more competitive tan ever, so the coaches and the players of any sport are transforming his perfomance in data that it can be measure and analyzed to improve the perfomance and the money benefits in the sports.

In this paper we can see the perfomance of football player in a season like data and we analyzed it with the general objective of get value out of this data and get closer to the way of working of profesional analysts in football.

To be more specific and understand this paper we will try to meet certain specific objectives:

- Understand and organise the databases to start analyzing the data.
- Perform a filter to work more efficient.
- Descriptive analysis of data.
- Perform a hierarchical *cluster*ing to reprsent the features of the players.
- Perform a PCA to understand the variables in data.
- With the help of PCA, *cluster*ing the data observation.
- Perform a regression analysis to predict the number of goals of each player by the creation of models.
- Finding out if it is true that lefty players have more precision than righty players using a hypothesis test.
- Process the results obtained by the techniques in order to extract knowledge.
- Interpret and draw conclusions from the results of the analyzes.

In this paper there are two similar analysis. One of them is the general study of the offensive perform of the player based in the data, buy the analist and the clubs not always need to improve his teams with the best offensive players. So one of the analysis made has a real application in football because it had helped a real football analyst.

This study includes and focuses in the players with the characteristic of been a *BoxToBox*, this is a player with the hability and capacity of go all over the field form his area to opponent´s área. But let,s give context to this, we are going to explain the born in football of this type of player.

Context of the *Box to Box* player: the system

Since the beginning of this decade that we left behind and the end of the previous one, back in the years 2009, 2010 and 2011 as the beginning, football evolved to use the 4-3-3 system as a rule, used by Guardiola in Barcelona and Vicente del Bosque in the Spanish National Team, two teams with a lot of similarity. Motivated by the results obtained by them, most of the teams have adapted and built their squad with the idea of playing this system, which is still used today by the most powerful teams in recent years. Zidane and Ancelotti's Real Madrid and Klopp's Liverpool are a clear example of this and a relay in terms of the transformation that the system itself has undergone.

Guardiola's Barcelona and Del Bosque's National Team

The first to use this formation in a winning way was Guardiola's Barcelona in 2009. The year of the treble. When we think of this team, the players that come to mind are Xavi, Iniesta, Busquets and Messi. Players with such high quality and with such good game control that they rarely had to worry about defending. Xavi and Iniesta, the interiors of this team, will not be remembered as *Box to Box* midfielders, because very few times they had to make an exceptional physical display to stand out, in fact, when Barcelona suffered the most at that time it was against the cons. When there were many meters to run, rival players took advantage of those spaces and Barcelona suffered.

The Vicente Del Bosque´s national team of Spain was a mirror of Barça, in this case, Xabi Alonso entered the 3-means system, leaving Iniesta more freedom in the face of attack. With Xabi Alonso, being more positional and a better defender than Iniesta, leaving him in more offensive and less defensive areas, a more absolute control of the game was achieved, more than Guardiola's own team and without counting on the most decisive player in history, Messi.

Beginnings of the *Box to Box* profile

After the absolute dominance of Barcelona, Real Madrid got down to work and signed Mourinho as coach. Mou was Guardiola's nemesis those years. To beat him, his game consisted of creating a strong midfield and playing against him.

One of the first *Box to Box* in modern times was Khedira, Mourinho himself requested his signing as part of his plan to create that strong midfield. While Xabi Alonso was the positional midfielder, Khedira had to be strong enough to support him in defensive tasks and loosen up in attack so as not to leave Özil and Cristiano alone.

The current importance of the *Box to Box*

As previously mentioned, the idea of the *Box to Box* was perfected using a 4-3-3 with at least one middle center of this profile, it has been seen in Ancelotti's Real Madrid with Di Maria and later with Modric in Real Madrid by Zidane. The leap in quality that Modric made from his early years in Madrid until today has been partly due to his conversion into this player profile, accompanied by Casemiro, who is dedicated to defensive tasks, Kroos whose role is to be the first Support line with the team's ball, leaves Modric with the need to support both his teammates in the center of the field, as well as to loosen up in attack to support the forwards.

But the total perfection of this profile is Klopp's Liverpool, since all the midfielders he has are capable of destroying as well as building and finishing. Doing it at a level that has rarely been seen on a soccer team.

The process of the analysis

We have the data, there are many statistics technics to extract knowledge as we seen before, but we need a powerful tool or software that provide us this technics. The program that is used throughout the process of this paper is the program named *R* and his enviroment *RStudio*, wich is based on statistical computing and, with the enviroment, make very sophisticated graphics to help us to interpret the results.

Databases used in the paper

The databases of this work have been collected from the WyScout software.

Specifically, to carry out our analyzes they have been based on different csv files, to be more exact three databases have been used. Each one with the same variables as the rest but with different observations. These databases are the ones that have been called:

- **Delanteros_19-20:** This database contains 111 variables related to the game and the performance of each player (observations), which are given by 3102 rows. These observations correspond to the players of the Spanish Second Division B, SmartBank League and Ledman Liga Pro, all of them have been collected during the 2019-2020 season. As a note, this base has been called "Delanteros_19-20" because the analysis has taken into account, within this base, the players who fulfilled their position as forward.

- **JUGADORES:** Like the previous one, this base contains the same 111 variables but its observations correspond to the 532 players who participated in the Spanish First Division of soccer during the 2018-2019 season.

- **SEGUNDAB:** Same variables. His observations correspond to all the players who officially participated in the Spanish Second Division B during the 2018-2019 season, a total of 2000 players.

Obviusly this datasets have been filtered to help to understand and make better techniques. In this paper there are two parallel analysis, in one hand we have got an analysis of the offensive perfomance of the player of *LaLiga* in the season 18/19, and in the other hand we have got a analyse the best *BoxToBox* players in *Segunda División B* in the season 19/20. This last study has served to help a real analyst of professional football team in the spanish third division, in other words, a team wich is in *Segunda División B.*

Conclusions of the results

1. The first result of this study is the hierarchical *cluster*. This technique show us that the player are very good grouped in base of their characteristics. This is a great fisrt analysis because we can begin to develope the idea of type of offensive players are in case of the first study made or the types of *Box to Box* players there are in case of our second analysis. This have a important disadvantage, we have the type of players but we have not his concrete characteristics.

2. The second result of the research is applied the PCA (Principal Component Analysis) to prepare the data to applied a future k-means *cluster* algorithm.This is very important part of the study, in this way we have built the axes of the future graphic of k-means.

3. Once we have made the ACP, the k-means algorithm is applied to the diferent databases. Before its application, an Elbow Method algorithm was carried out in order to separate the players in some groups or *cluster*s. This algorithm told us that the base should be divided into 5 *cluster*s in the two studies. In this paper there are programming an interactive 3D graphic of the kmeans, but it can´t put this in this type of file, so we put photos of it.

4. In case of the study with the offensive players, the k-means algorithm show the different types of players in LaLiga. For example we have a *cluster* (*cluster* 5) with the players that his teams based his attack in them like Ben Yedder in Sevilla F.C., Benzema in Real Madrid, Griezmann in Atlético de Madrid… we also have a *cluster* (*cluster* 3) with players who give a lot of depth to their teams, either through passes or offensive actions who open defenses through their vision of the game such as Banega, Canales or Cazorla and very vertical players like Navas, Jordi Alba or Yuri who are the ones who give the offensive style to their teams of *LaLiga.* We also have the players (in *cluster* 4) who are the classical number 9 in football.

5. In case of the study that contains the *Box to Box* players of *Segunda División B*, we can conclude that the players closest to being a *Box to Box* are the *cluster* 5, and this players are found in the space formed between the (0, -6) of PC1, players with better offensive capabilities, (0, 2) of PC2, players with greater duels defensive players and (0, -2) from PC3, players with the highest % of defensive duels won. In this analysis, the ideal player for a team looking for a *Box to Box* would be Álvaro Fidalgo, while those close to him would be players of the same profile that we are looking for.

6. We also do four differents models that pretend to predict the number of goals that all the offensive players in database *Delanteros19_20* has scored in the season. This models are:

- Model 1 wich is formed by the variables: market value, minutes played, xG, attempts and shots on goal. This model has significatives the variables minutes played, xG, and shots on goal.

- Model 2 wich is formed by the variables offensive actions, to lose the mark, touch in the box, matches played, aereal duels. This model has significatives all the variables except aereal duels.

- Model 3, wich use all of the variables selected in the dataset. This model has significatives the variables minutes played, expected goals, offensive actions, header goals, attempts, goals maded, offensive duels gains and touches in the penalty area.

- StepAIC model, this model uses an algorithm in which it begins by using all the variables on the model and gradually eliminates those that are not significant until a model is found in which all its variables are.

In conclusión we have that the variables in StepAIC model is the best model to predict the goals because it was the best perfomance to predict the goal of the players.

7. The final analysis in this study explore the affirmation that the lefty players are more precises that the righty players. We use the *ANOVA* test to compare the lefty and the righty players based in his success rate of passing and shots on goal. We observe that there aren´t differences in the percentage in shots on goal by lefty and righty players but we observe differences in the rate of passing being righty players better than lefty players.

Índice general

Glosario

- **Minería de datos:** Es el proceso mediante el cual se recoge información a partir de un almacén de datos.
- **KDD:** Por sus siglas en inglés *Knowledge Discovery in Databases*, hace referencia al proceso automático de recoger información a partir de los datos con el fin de analizarla.
- **Inteligencia artificial:** Hace referencia a la capacidad que tiene una máquina de hacer análisis complejos.
- **Almacén de datos:** Conjunto de datos recogidos de forma integrada y compleja que no varía con el tiempo y que se usa para ayudar en optimizar las decisiones tomadas por el organismo que lo analiza.
- **Patrones:** En estadística hace referencia a un comportamiento específico de los datos.
- **Data mining:** Conjunto de algoritmos, técnicas y ciencias que ayudan a explotar un almacén de datos.
- **Regresión Lineal:** Modelo matemático que mide la dependencia entre una variable y otras**.**
- **Redes neuronales:** Sistemas de computación vagamente inspirados en las redes neuronales que constituyen el cerebro animal.
- **Árboles de decisión:** Un árbol de decisión es una herramienta de apoyo que usa un modelo de decisión caracterizado como un árbol y sus posibles consecuencias, incluido la probabilidad que ocurran diversos eventos, minimizar costes y su utilidad.
- **Reglas de asociación:** Técnica estadística que se basa en revelar hechos que se dan de forma recurrente en un almacén de datos
- **Campos:** Hace referencia a las variables/columnas de un almacén de datos.
- **Registros:** Hace referencia a los individuos/filas de un almacén de datos.
- **Opta:** Empresa especializada en la recogida y el análisis de datos deportivos.
- **Plantilla:** Todos los jugadores que forman parte de un equipo de fútbol.
- **Robos:** Acción en la que un jugador le quita el balón a otro.
- **WyScout:** Empresa especializada en la recogida de datos de la mayor parte de las ligas en el mundo.

- **Temporada:** Periodo de tiempo en el que tienen lugar el inicio y la finalización de las competiciones oficiales**.**
- **R:** Programa especializado en el análisis estadístico.
- **Primera División:** Máxima división española de fútbol.
- **Liga Smartbank:** Segunda división española de fútbol
- **Ledman Liga Pro:** Segunda división portuguesa de fútbol.
- **Segunda División B:** Tercera división española de fútbol.
- **Portero:** Posición de fútbol cuya función es evitar que el balón entre la portería propia usando todas las partes de su cuerpo para ello.
- **Lateral:** Posición que se encuentra en las zonas de banda defensivas, esto es, más cerca de la portería propia.
- **Defensa central:** Posición que se encuentra en el eje central en la zona defensiva, más cercana al portero propio.
- **Centrocampista:** Posición que se encuentra en la zona central del campo.
- **Extremo:** Posición que se encuentra en las zonas de banda ofensivas, esto es, más cerca de la portería contraria.
- **Delantero:** Posición más cercana a la portería contraria.
- **Gol:** Acción en la que un equipo introduce el balón en la portería contraria.
- **Asistencia:** Contribuir mediante un pase a que un jugador logre realizar un gol.
- **Mundial de Fútbol:** Torneo internacional que reúne a las 32 mejores selecciones del mundo y que se disputa cada 4 años.
- **Goles a favor:** Goles marcados por un equipo a lo largo de una competición.
- **Goles en contra:** Goles recibidos por un equipo a lo largo de una competición.
- **Sistema:** Formación en la que se distribuyen los jugadores de un equipo durante el partido.
- **Fichar:** Dar de alta en tu plantilla a un jugador que estaba en otro equipo.
- **Modelo de juego:** Acciones de un equipo durante el juego y que da como resultado forma de jugar que tiene. También se puede llamar estilo de juego.
- ***Merchandasing*:** Productos relacionados con el equipo que saca a la venta.
- ***Machine Learning*:** Capacidad que tiene una máquina de aprender a razonar y, hacer análisis y tomar decisiones por sí misma.
- **Visión de juego:** Capacidad que tiene un jugador de optimizar sus decisiones en el juego.

- **Contraataque:** Acción de robar el balón al equipo contrario y acto seguido ir al ataque.
- **Acciones ofensivas:** Capacidad de un jugador de realizar acciones que lleven a su equipo a intentar marcar gol.
- **Desmarques:** Capacidad de un jugador de recibir el balón alejado de los jugadores rivales.
- **Ataque en profundidad:** Acción por la cual el balón se acerca a la portería contraria con intención de marcar gol.
- **Duelos defensivos:** Acción en la que un jugador disputa el balón que se encuentra en posesión del equipo contrario.
- **Duelos aéreos:** Acción en la que un jugador disputa el balón que se encuentra en el aire.
- ***Box to Box*:** Centrocampista con características ofensivas y defensivas, capaz de llegar a las dos porterías y con un gran despliegue físico durante el partido.
- **Profundidad:** Característica basada en tomar acciones siempre en dirección e intención de acercarse a la portería contraria.
- **xG o goles esperados:** Predicción de goles que hará un jugador durante la temporada en base a la posición (tanto suya como de sus compañeros y rivales), trayectoria y calidad en base a sus tiros.
- **xA o asistencias esperadas:** Predicción de asistencias que hará un jugador durante la temporada en base a la posición (tanto suya como de sus compañeros y rivales), trayectoria y calidad en base a sus pases.

Introducción

Estado de la cuestión

La Minería de Datos es un proceso mediante el cual, a partir de grandes volúmenes de datos, extraemos información relativa a ellos para obtener beneficio.

El concepto surgió en los años 60 con la idea de buscar patrones y correlaciones de las bases de datos obtenidas, pero no fue hasta finales de los 80 cuando se empezó a profundizar más en el *Data Mining* gracias en parte, a la creación de Internet y el desarrollo de los ordenadores.

La minería de datos forma parte de un conjunto de técnicas estadísticas que, agrupadas y ordenadas de cara a analizar una base de datos, forman el concepto denominado *KDD* o *Knowledge Discovery in Databases* cuya traducción libre sería la materia que se dedica a sacar información de una base de datos en bruto. La minería de datos es la etapa de análisis del *KDD*, busca patrones en grandes volúmenes de almacenes de datos mediante procesos de inteligencia artificial y se apoya en la estadística para poder dar una estructura a los datos, así sacar información íntegra de estos y utilizarla en nuestro beneficio. Es la parte estadística de todo el estudio.

Todo este proceso se puede dividir en una serie de pasos generales:

- Selección de un conjunto de datos
- Análisis descriptivo de los datos
- Preparación de los datos para su posterior análisis
- Aplicar las técnicas adecuadas para sacar patrones o información
- Procesar los resultados obtenidos por las técnicas para así extraer el conocimiento
- Interpretación de los resultados
- Implementación de los patrones

Para poder indagar en este concepto, es necesario tener conocimiento en estadística y entender las técnicas más utilizadas del *Data Mining*. Una de las técnicas base que usa esta rama es la regresión lineal, que mide la correlación de los datos entre dos variables y que, debido a esto, puede ser no óptima de cara al estudio de más de dos variables.

Las redes neuronales es otra de las técnicas más utilizadas en los procesos de minería de datos ya que tienen multitud de aplicaciones en el mundo real. Es un sistema basado en las interconexiones neuronales que se dan en el propio cerebro humano y que tiene como finalidad el reconocimiento de la información para transformarla en valores que podemos interpretar. La mayor ventaja de esta técnica reside en que es de aprendizaje automatizado; es decir, que cuanta más información previa haya trabajado la red neuronal, mayor será su nivel de precisión en el reconocimiento de la información.

La predicción es un concepto muy demandado en el mundo real y es uno de los objetivos de la minería de datos (MD) que usa los árboles de decisión como técnica estadística para solucionar problemas que se suelen dar de forma repetida.

Los modelos lineales son otra de las técnicas usadas en la MD y que tiene como finalidad resaltar cuales son los factores que tienen más relevancia y que modifican la variable respuesta.

El agrupamiento de la información es una técnica muy utilizada también en esta rama, ya que es una manera muy visual de dar forma a nuestra base de datos dividiéndola en diferentes *cluster*, esta técnica es llamada *clustering.*

Finalmente, hablamos de reglas de asociación como una de las técnicas que permiten ver que asociaciones o patrones tienen en común los datos medidos.

En la actualidad, el *data mining* es una de las ciencias que están en alza y aunque sea utilizada en multitud de ámbitos y en muchos es clave, en otros se está empezando a mirar con buenos ojos y multitud de proyectos se están desarrollando teniéndola como clave.

Es innegable que, actualmente, vivimos en una sociedad cada vez más marcada por el mundo digital. Las redes sociales van teniendo más peso en nuestro día a día y con el desarrollo de Internet en lo que va de siglo, se crean cada vez más datos individualizados de cada ser humano. Todos estos son de capital importancia para las empresas e incluso los Gobiernos y cada vez son mayores las inversiones, tanto económicas como en desarrollo para poder dar forma y entender los datos medidos.

Una empresa que tenga en su mano muchos datos no sería competitiva si no tiene una estructura interna dedicada a darles valor, por lo que, una empresa con menos datos, pero a los que le saca el óptimo provecho, generaría una desigualdad competitiva inmensa. Como se ha expuesto anteriormente, no son solo las empresas las que usan los datos para ser más competitivas, sino que los propios gobiernos aprovechan los datos para dar un salto cualitativo en la vida de los ciudadanos, incluso, en nuestra vida diaria tienen una importancia mucho más vital de la que creemos, por ejemplo, para mejorar la fluidez del tráfico, para combatir el fraude de tarjetas de crédito, para predecir desastres naturales e incluso para combatir el terrorismo.

Uno de los ámbitos donde la minería de datos se encuentra en un estado embrionario es el fútbol.

En el fútbol, las opiniones basadas en la estadística muchas veces son vistas con malos ojos, pues sin duda alguna este es el deporte con mayor cantidad de escépticos y donde existen multitud de opiniones subjetivas debido a la importancia emocional que tiene en la sociedad; es decir, que debido a las pasiones de cada persona se genera una disociación a la hora de tener una idea objetiva dentro del juego. Por ejemplo, teniendo los datos de los campeones de la liga española hasta la fecha, se puede decir, que, según el número de trofeos, objetivamente el Real Madrid es el mejor equipo nacional de la historia, afirmación que, por ejemplo, un aficionado del Barcelona o del Atlético de Madrid rechazará y que utilizará argumentos más bien subjetivos para afirmar que esto no es así.

En la actualidad, las llamadas estadísticas tienen menor calado en los aficionados; sin embargo, en los clubes es diferente. Cada vez son muchos más los equipos que se suman a desarrollar infraestructuras internas orientadas a la toma de decisión o a organizarse internamente apoyándose lo más objetivamente posible en ellas.

Al igual que en el mundo real, en el mundo del fútbol, la minería de datos tiene multitud de aplicaciones.

Uno de los ámbitos donde más se usa es a la hora de fichar jugadores, gracias a programas y empresas como *Opta* o *WyScout* que miden de forma individualizada las acciones de cada jugador en cada partido y que los transforman en multitud de variables medibles. Son muchos los equipos que dejan en mano de sus expertos en estadística las posibles soluciones para reforzar el equipo tras la pérdida de jugadores clave por el aumento de las diferencias económicas entre las élites y el resto de clubes de fútbol. Uno de los casos más recientes se dio en la liga española hace escasamente dos años, Fabián Ruiz, por entonces jugador del Betis, completó una de sus mejores temporadas como profesional y llegó a ser una pieza fundamental para el equipo, esto hizo que equipos de mayor envergadura que el Betis se fijaran en él, lo que desembocó en la marcha del jugador hacia el Napoli, club histórico y referencia en la liga italiana. Tras su marcha, el Betis rastreó el mercado y analizó estadísticamente a cada jugador, dando como resultado que el que tenía atributos más similares a Fabián era el argentino Lo Celso. El Betis apostó por el futbolista argentino y tras una temporada en el club, se convirtió en el referente de su equipo. La historia además es similar a la de Fabián, Lo Celso acabó siendo vendido al Tottenham inglés por una cantidad superior al pagada por su fichaje, por lo que el Betis salió ganando con las operaciones incluso a nivel económico. Esta es una de las maneras que tienen los clubes con menos recursos de seguir siendo competitivos.

Un caso similar es el del Liverpool, campeón de Europa hace dos temporadas, que utilizó la MD incluso para desarrollar un modelo de juego ganador, además de construir una de las mejores plantillas de Europa de los últimos años.

El uso de *data mining* en el fútbol, no se utiliza solo para organizar los clubes y sacar rendimiento en el juego, también es muy importante en lo relativo a la salud de los jugadores. A través de monitorizar a los jugadores en cada entrenamiento y en cada partido, los entrenadores, preparadores físicos e incluso los médicos tienen información relevante del estado muscular de cada jugador, disminuyendo o aumentado la carga de trabajo de forma individual para así optimizar el entrenamiento y la preparación de los partidos.

El referente en el uso de técnicas de la minería de datos en el deporte es sin duda la *NBA*. El fútbol se ha nutrido de la información y de la forma de trabajar en baloncesto para desarrollarse más pese a que el fútbol, al ser de baja anotación y en el que participan más jugadores, es más difícil de medir.

Uno de los ejemplos del uso de técnicas de minería de datos más interesante que ha usado la *NBA* y que en el fútbol aún está por desarrollar es el descubrimiento de futuras estrellas o dar un valor potencial a los jugadores con menos años. El ejemplo real de esto se tiene en que se ha descubierto que los jugadores jóvenes que participan en las ligas universitarias de Estados Unidos que promedian un gran número de robos, acaban siendo los jugadores más valorados de la *NBA* en el futuro; es decir, que gracias a los datos, descubrieron un patrón que puede llegar a predecir una futura estrella de la *NBA*.

Objetivos

El objetivo en resumen de este estudio es obtener valor de los datos medidos a partir de la realidad, para estudiar de manera general todos los jugadores, con la posibilidad de individualizar alguno que destaque en una característica concreta.

Las conclusiones sacadas están orientadas a que un directivo de un club pueda hacer uso de ellas para maximizar el rendimiento de su equipo, ya sea fichando jugadores de unas características concretas, o reforzándolo con jugadores muy destacables de manera cuantitativa que, quizá, no se vea tan fácil en la realidad. Además, siempre podrá optimizar sus decisiones intentando buscar jugadores más rentables en cuanto a la relación calidad/precio.

Para ser más concreto, este estudio ha tenido tres objetivos principales utilizando para ello tres bases de datos similares en cuanto a campos pero con diferentes registros. Estos primordiales son:

- Plasmar los datos de manera que se puedan ver los distintos tipos de perfil de jugador existentes en el almacén.
- A partir de estos perfiles, poder determinar un jugador que cumpla unas características concretas, buscando el mejor con dichas características.
- Realizar un modelo de predicción de goles que ayude a determinar, a partir de los datos, el número de goles que marcará el jugador durante la temporada.
- Saber si es cierto lo que se dice de que los zurdos tienen más calidad y precisión que los diestros mediante un contraste de hipótesis.

Capítulos y estructura

Este TFG contiene cinco capítulos fundamentales, los cuales tratan sobre los temas de este trabajo, la explicación de los algoritmos y técnicas utilizadas y la metodología y resultados obtenidos.

El primer capítulo habla del tema primordial a tratar, el fútbol, en él se verán los comienzos del deporte, se explicará en que consiste, cómo se medían las primeras estadísticas y cómo han llegado al fútbol actual para quedarse, desarrollándolo aún más al tratarlo como una ciencia medible y objetiva.

En el segundo capítulo, hablaremos de la explotación de datos. Cómo surgió y por qué, el estado actual del tema en cuestión y las técnicas más vistas y utilizadas relacionándolas con el deporte además de otros ámbitos.

En el tercer capítulo, se hablará de las técnicas utilizadas en este trabajo, es decir, se verá el porqué de utilizarlas y cómo funcionan matemáticamente.

El cuarto capítulo tratará de explicar los programas utilizados en este proyecto. En este caso, hablaremos del lenguaje de programación R.

El quinto y último capítulo trata de la metodología y los resultados. Es el más práctico de todos, pues se verán los análisis estadísticos hechos buscando dar una explicación científica de lo que pasa en el propio juego a través de los datos.

Resultados fundamentales

Tras realizar los análisis y las técnicas estadísticas con el fin de explotar los alamacenes de datos de este estudio, se han tenido como resultados fundamentales los siguientes puntos:

- Clasificación de los jugadores según sus características medidas en los atributos de la base de datos. Se agrupan de esta manera, los jugadores similares en las ramas contiguas del algoritmo de clasificación *cluster* jerárquico. Este algoritmo se usó para la base llamada "JUGADORES".

- Clasificación mediante el método k-means de las bases de datos JUGADORES y SEGUNDAB, la cual nos da una información más profunda que el cluster jerárquico ya que se tienen en cuenta un conjunto de variables seleccionadas con anterioridad y que ayuda a visualizar el por qué los jugadores forman parte de un grupo.
- Se realizaron cuatro modelos de predicción del número de goles para la base llamada "Delanteros_19-20" teniendo como conclusión final que el mejor algoritmo se llevó a cabo gracias a la realización de la técnica llamada StepAIC, la cual indicó las variables más significativas a la hora de predecir el número de goles marcados.
- Otro resultado fundamental fue el de contrastar la creencia popular de que los zurdos son mejores y más precisos que los diestros. Así pues, se realizó un contraste ANOVA que midió si había diferencias entre ellos en dos variables: Precisión en el pase y precisión en el tiro, teniendo como conclusión final que sí había diferencias significativas en el pase a favor de los diestros. En cuanto a la precisión en el tiro no se encontraron diferencias significativas.

Almacenes de datos

Las bases de datos de este trabajo han sido recolectadas a partir del programa informático *WyScout.*

En concreto para realizar nuestros análisis se han basado en diferentes archivos *csv*, para ser más exactos se han utilizado tres bases de datos. Cada una con las mismas variables que el resto, pero con diferentes observaciones. Después de limpiezas y transformaciones se han convertido en almacenes y se han llamado:

- **Delanteros_19-20:** Contiene 111 variables relativas al juego y al desempeño de cada jugador (observaciones), los cuales vienen dadas por 3102 filas. Dichas observaciones corresponden a los jugadores de Segunda División B española, Liga *SmartBank* y *Ledman Liga Pro*, todas ellas han sido recogidas durante la temporada 2019-2020. Como apunte, este almacén se ha llamado "Delanteros_19-20" porque en el análisis se han tenido en cuenta los jugadores cuya posición es la de delantero.
- **JUGADORES:** Al igual que el anterior este almacén contiene las mismas 111 variables pero sus observaciones se corresponden con los 532 jugadores que participaron en la Primera División española de fútbol durante la temporada 2018-2019.
- **SEGUNDAB:** Mismas variables. Sus observaciones se corresponden con todos los jugadores que participaron oficialmente en Segunda División B española durante la temporada 2018-2019, un total de 2000 jugadores.

Como se ha explicado para construir el primer almacén de datos, para hacer los análisis se ha requerido de un filtrado. Con este tratamiento conseguimos que el almacén sea más objetivo y se adapte mejor a lo que se quiere transmitir en este proyecto. En capítulos posteriores, se verán estos filtros y se explicará de manera más desarrollada el porqué de los mismos.

Capítulo 1: Fútbol

1.1. Historia y comienzos del fútbol

El fútbol es un deporte inventado por los ingleses en la segunda parte del siglo XIX. Si bien es cierto que en siglos pasados existían juegos similares, no fue hasta 1864 cuando, en la Universidad de Cambridge, se crearon lo que serían las reglas del fútbol moderno. Como curiosidad, a la hora de discutir las reglas había gente que quería que el juego fuera más agresivo y hubiera contacto, por lo que los defensores de este tipo de forma de juego crearon su propio deporte, hoy en día es lo que conocemos como el rugby.

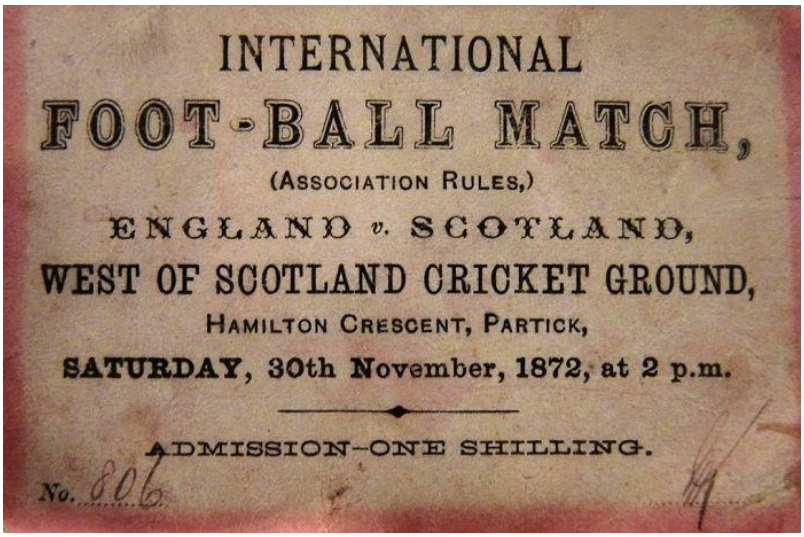

Ilustración 1. Cártel del primer partido internacional de fútbol de la historia. Fuente: futbolday.com

Este deporte comenzó rápidamente a ganar mucha popularidad entre la mayoría de la población. Esto se explica que, a diferencia de otros deportes en alza de la época, no se necesitaba de ningún tipo de material para practicarlo, así que caló profundamente entre las clases más bajas de la sociedad, las cuales eran mayoría. Rápidamente, viendo el potencial que tenía para atraer a las masas, muchos barrios de las ciudades empezaron a fundar sus propios equipos, los cuales estaban formados por gente corriente que hacía su vida dentro de la propia comunidad. Fue una excusa perfecta para competir contra otros barrios con los que se tenían algún tipo de rivalidad. Todo esto supuso un auge a nivel económico y social y, en este punto, se empezaron a organizar ligas que agrupaban equipos ya no solo de la misma ciudad, si no equipos de todo el país por todo el continente europeo. Esto provocó construcción de nuevos estadios, un punto de reunión para la gente los días de descanso del trabajo y aumento de consumo en los comercios que rodeaban las sedes y los estadios de los equipos de fútbol.

Llegados casi al primer tercio de siglo, la II Guerra Mundial supuso un parón muy importante para este deporte, el cual ya contaba con una competición internacional oficializada como el Mundial de Fútbol, el cual agrupaba a selecciones de jugadores de un mismo país que competían por ver cuál se llevaba el trofeo de campeón. En ese momento, se habían disputado tres Copas del Mundo: en 1930, con sede en Uruguay, país que logró alcanzar el éxito en esta primera edición; en 1934, se repitió la misma historia que la edición anterior, solo que esta vez el campeón y la sede fue Italia (muchos

artículos periodísticos e historiadores comentan que Mussolini hizo lo posible por organizar el Mundial, para así sabotearlo y utilizar la victoria italiana a su favor para fines políticos); finalmente, antes del parón, el último Mundial disputado fue en 1938, cuya sede fue Francia y en el que Italia repitió como campeón.

Ilustración 2. Selección italiana, campeona del mundo en 1938. Fuente: as.com

Tras la finalización de la Guerra, se puso en marcha el inicio de las competiciones oficiales y, en el año 1955, la revista francesa *France Football* en colaboración con Santiago Bernabéu, presidente del Real Madrid, crearon la Copa de Europa, competición que haría enfrentarse a los campeones de la Primera División de cada país europeo. Como bien se puede apreciar en la actualidad, este torneo es, sin duda, el que más éxito cosechó en Europa a nivel de seguimiento y calidad en el juego.

Gracias al seguimiento de este deporte en todo el mundo (especialmente Europa, África y América del Sur) el fútbol se fue desarrollando tanto a nivel organizativo, mediático e incluso científico. En este sentido, este deporte no fue el único que tuvo este auge, deportes como el baloncesto o el béisbol, los más seguidos en EE. UU., tuvieron un gran desarrollo en todos estos niveles. Gracias a esto, Europa aprendió de los métodos revolucionarios que llegaban desde el otro lado del charco y empezó a tener más en cuenta la ciencia a la hora de desarrollar tácticas y metodologías modernas que supusieron un aumento en ingresos y éxitos a nivel deportivos en los clubes que las integraron.

1.2. Acercamiento de la estadística al fútbol

En el fútbol del siglo XX, el primer acercamiento que había a la estadística por parte de este deporte no era más que el conteo de goles (a favor y en contra) de los equipos en la tabla de clasificación y los puntos. Además de medir la cantidad de goles individuales que metía un jugador durante el torneo. Estas estadísticas, que si bien son las más importantes, por sí solas son imposibles de predecir o de buscar algún tipo de relación con otras variables que pueden darse en el juego.

En 1984, ya entrados aún más en el fútbol moderno se creó la *IFFHS* (*International Federation of Football History and Statistics*) por sus siglas en inglés. Este organismo tiene como objetivo recopilar información estadística, científica y cronológica cuyo objetivo es recoger los récords y eventos notables que se den en el fútbol a nivel mundial.

Si bien es cierto, que en el siglo XX, el fútbol, tanto para expertos como para aficionados, era más emocional que racional, no es hasta la segunda década del siglo XXI cuando los datos han llegado a formar parte del fútbol y a ser considerados una herramienta de apoyo, tanto de clubes, como de analistas e incluso jugadores.

En este trabajo final, se ha tenido como objetivo realizar un análisis descriptivo, análisis no supervisados y la creación de modelos y contrastes de hipótesis que ayuden a explicar el rendimiento de los jugadores dentro del terreno de juego a partir de estas técnicas.

Por ejemplo, es bien sabido que, incluso entre dentro de las posiciones de un sistema, hay jugadores que, aun estando en las mismas posiciones, aportan cosas diferentes a un equipo.

Si se le pregunta a una persona con nula relación con el mundo de la estadística que qué es lo que piensa cuando le preguntamos por ella, seguramente en su mente lo primero que ocurra es pensar en el concepto de porcentajes o probabilidades.

La probabilidad es algo inherente al ser humano por el hecho de que la usamos para llegar a un fin siempre que tenemos que elegir tomar una decisión de cualquier tipo. No solo en el ser humano se ven la utilización de probabilidades. Por ejemplo, cuando se enseña a un perro a sentarse, generalmente, reforzamos su comportamiento y acción de sentarse cuando se le dice un comando con la recompensa de comida una vez ha hecho lo que se buscaba. En este punto, el perro ve ante él dos opciones, hacer caso a lo que le dice su dueño y llevarse la recompensa en forma de comida o no hacer caso y no conseguir el premio. Inconscientemente ha usado probabilidades, hace lo que le dice el comando que le haya enseñado su dueño, porque sabe que es la acción que hará que, con mayor probabilidad, consiga el premio que busca.

Si se lleva lo anterior al ámbito que estamos tratando, en este TFG tenemos muchos ejemplos similares. Un futbolista cuando recibe el balón en sus pies se le abren multitud de opciones de cara a qué decisión tomar o acción a llevar a cabo y, siempre, intentará buscar la que haga que su equipo sea más eficiente sobre el terreno de juego. Estas decisiones no siempre son fáciles de ver. En una misma jugada se pueden tener dos opciones igual de válidas e intrascendentes para el fin de marcar un gol. Por ejemplo, la elección de un pase de un portero hacia su central derecho o izquierdo parece intranscendente de cara a lograr este objetivo, pero, poco a poco y según se avanza sobre el campo al acercarse a la portería rival, este tipo de decisiones se hacen cada vez más evidentes con el fin de marcar gol.

1.3. Uso de la estadística en el fútbol actual

Anteriormente, ya se han explicado las variables de las bases de datos que se han usado en este estudio. Estas son las que los expertos de *WyScout* analizan porque creen que son los factores o variables que hacen que un futbolista sea considerado buen jugador y, por tanto, domine mejor que los demás la toma de decisiones sobre el terreno de juego.

Al igual que un jugador tiene muchas decisiones que tomar en un partido ya sea, para atacar o defender, con la estadística pasa lo mismo. Esta herramienta es usada de manera muy variada por los analistas dentro del fútbol. Existen en la realidad multitud de ejemplos de este estilo. La figura del

entrenador es, sin duda, la persona que aprovecha esta herramienta al máximo en este deporte. Dentro del fútbol, la estadística basa su evidencia en la realidad a partir de acciones pasadas que ayuden a predecir el futuro de algo concreto. Se ha visto sobre todo la implementación de esta ciencia a la hora de tomar decisiones para fichar a un jugador u otro.

Como se ha comentado anteriormente, empresas como *Opta*, *StatsBomb* o *WyScout* miden, lo que para ellos son las variables que mayor incidencia tienen durante el juego. Esto converge a una recogida de datos a nivel internacional, por lo que las bases de datos con las que se trabajan pueden llegar a tener tamaños tan grandes que requieren de técnicas y equipos de trabajo dentro del club especializados en el *Big Data*. Así pues, gracias a la estadística, un analista que busque un jugador con características concretas puede filtrar la información que proporcionan estas bases de datos y empezar a buscar jugadores de manera más eficiente a que si se busca dentro de toda la información proporcionada. De esta manera, se puede llegar a la idea de la calidad bruta que tiene un jugador, además de ahorrarse tiempo y dinero en ir a ver a tantos jugadores en el mundo real. Así pues, realiza un filtro y va a ver solamente a un número asequible de jugadores que le encajan mediante el tamiz realizado.

Además, si un equipo pierde un jugador importante porque lo ha fichado un equipo más grande, lo primero que hace el club para reemplazarlo es rastrear el mercado para dar con un otro con características similares al perdido. Gracias a esta forma de fichar, muchos clubes salen ganando no solo a nivel deportivo, sino económico, ya que aprovechan los ingresos generados por la venta del jugador para fichar a alguien, posiblemente más joven y barato, que en años posteriores será vendido por una cantidad más alta que la invertida por él en su fichaje, lo que hace que se optimice la pérdida de los mejores jugadores que pertenecen a clubes más humildes y que puedan permitirse luchar en la élite contra equipos con recursos mucho más sofisticados y caros. No solo en cuanto a los fichajes se puede sacar beneficio a partir de la estadística, también se puede generar un modelo de juego que ayude a que los jugadores en el campo se compenetren mejor y aumenten su rendimiento al jugar juntos. Por ejemplo, estudios que ha hecho LaLiga con su propio organismo dedicado al *Big Data*, llegaron a la conclusión de que todos los equipos que contaban con centrales titulares, cuya velocidad es inferior a 30 km/h acababan descendiendo de Primera División a Segunda División. De esta manera, un entrenador podría enfocar la evidencia de estos datos a un perfil concreto de defensa central, dando como resultado una afinación para con su modelo de juego.

Finalmente, la estadística también se usa a nivel profesional para trazar un plan físico durante la temporada, haciendo que las plantillas lleguen más frescas a los meses del calendario donde más exigencia encuentran. Muchos médicos utilizan estas evidencias para predecir el estado de los músculos de los jugadores para optimizar su descanso y así prevenir lesiones.

Pese al desarrollo de todo lo que se acaba de comentar, la estadística en este deporte está muy lejos aún del nivel que tiene en otras disciplinas, como por ejemplo la *NBA*. Además, el fútbol, a diferencia de otros deportes donde la estadística es más evidente, es un deporte de baja anotación entre los que incluso hay partidos que pueden acabar sin marcar goles. Todo ello conlleva a que mucha gente opte por seguir viendo el fútbol de manera más tradicional y ve con recelo el uso de la ciencia en un deporte tan pasional cómo es este. Obviamente, las máquinas generan un gran apoyo para desarrollar la calidad de este deporte y todo lo que le rodea, pero es innegable que sin trabajo humano detrás, los datos de por sí no nos aportan ningún valor. Estos datos requieren tratamiento y es lo que se verá en el capítulo siguiente.

Para acabar este capítulo y, como curiosidad, hay una película que explica de manera detallada la primera persona que llevó estas técnicas al mundo del deporte profesional para transformar todos los análisis científicos llevados a cabo en éxitos deportivos. El ejemplo del que se está haciendo referencia

es la historia de la que habla la película de *Moneyball*. La cual trata de cómo, en el béisbol, el entrenador de los Atléticos de *Oakland*, aprovechó sus estudios, al igual que se ha tratado en párrafos anteriores, para convertir la marcha de sus estrellas a equipos más grandes en una solución. Aprovechó los ingresos generados por las ventas para construir una plantilla, que, pese a no tener ningún nombre destacado, más competitiva y compenetrada respecto a la del año anterior y así sacar rendimiento de todos los elementos de los que disponía en el equipo, acabando finalmente como campeón de su respectiva liga.

Con el desarrollo de la recogida de estadísticas en el fútbol, surgieron videojuegos de simulación basados en ellas. El *Football Manager* o el *PC Fútbol* fueron videojuegos pioneros en cuanto al uso de la estadística en el fútbol, ya que se rigen en su totalidad en aprovechar las estadísticas de cada futbolista a tu favor para sacar el mayor rendimiento deportivo durante el juego. Estos videojuegos están tan bien construidos que incluso pueden servir de predicción en cuanto a los resultados cosechados por los equipos durante la temporada en el mundo real, y, si bien, no tienen un acierto del 100%, las predicciones que hacen suelen estar muy asemejadas con la realidad e incluso, con el paso de los años dentro del juego, grandes promesas se acaban convirtiendo en estrellas mundiales que acaban por materializarse en el fútbol real en el futuro

Ilustración 3. Perfil con los atributos de Cristiano Ronaldo en el famoso juego *Football Manager*. Fuente: cristianoronaldoinstagramnews.blogspot.com

Capítulo 2: Explotación de datos

2.1. Historia y comienzos de la explotación de datos

La explotación estadística basa su forma de proceder a partir de bases o almacenes de datos. Estas bases no son más que información guardada que ahora se recoge de manera automática. Es interesante que en los comienzos del desarrollo de esta ciencia y lo que le rodea, la recogida de datos se hacía toda a mano, generando un problema de tiempo, además de ser un trabajo muy costoso. A día de hoy, ya hay máquinas o programas que recogen la información de manera directa y automatizada, ahora bien, para entender mejor el concepto de información según la RAE se define como "Comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada".

Con el avance tecnológico que se materializó con las nuevas generaciones de ordenadores, empezaron a desarrollarse programas únicamente perfilados para almacenar toda esta información. De esta manera, nacieron programas como Excel, lanzado por Microsoft en 1985, y que poco a poco se convirtió en las hojas de cálculo que conocemos hoy en día. Los archivos sobre los que trabaja este *software* son los denominados XML, que tienen una estructura muy parecida a los ficheros de texto básicos que se pueden crear en los ordenadores a partir del Bloc de Notas.

De esta manera, se llegó a tener la información muy bien guardada y almacenada, pero ocurría un problema: ante la ingente cantidad de datos que se tenían a disposición, encontrar un dato concreto era una tarea muy complicada, así no había ninguna fórmula que permitiera filtrar la información que se buscaba para así, ahorrarnos tiempo y costes en encontrarla.

Es por ello, que se empezó a trabajar en desarrollar archivos de acceso directo o archivos secuenciales indexados, ISAM, por sus siglas en inglés que servían para acceder a la información mucho más rápida y de manera mucho más eficiente a través de índices creados sobre los ficheros. Lo índices es una manera de organizar la información mediante una estructura de árbol, ordenando los valores de manera que se visualiza mejor dónde se puede encontrar un valor determinado en un campo o variable. Además, estos índices tienen como características principales que se pueden crear varios en un mismo fichero y puede haber índices únicos o repetidos.

Fue en los años 60 donde estos programas tuvieron un gran alcance a nivel comercial e industrial, pero no fue hasta la década siguiente cuando se empezó a pensar qué se podía hacer a partir de toda esta información recogida. De esta manera, los nuevos sistemas utilizarán almacenes de datos ya recogidos para poder empezar a sacar conclusiones a partir de los mismos. Estas conclusiones dan respuesta a un gran espectro de preguntas de gestión y administración.

Así pues, llegados a este punto, se permitió diferenciar entre datos: hechos aislados; e información: datos procesados, organizados y resumidos.

2.2. ¿En qué consiste una base de datos?

Una base de datos está formada por registros, también llamados tuplas o individuos, los cuales son los componentes de las filas de una tabla. Todos los datos que aparecen en un mismo registro hacen referencia a un mismo individuo. Por otra parte, tenemos los campos, también llamados variables o atributos, los cuales son las columnas de una tabla. A estos campos se les pueden asignar propiedades especiales, como por ejemplo ser definido como índice o clave. Finalmente tenemos los datos, los cuales son la intersección entre un registro y un campo.

Los datos pueden ser de muchos tipos:

- **Texto:** Textos de longitud de 255 caracteres.
- **Fecha**
- **Numérico (entero corto, entero largo, decimal o doble)**
- **Booleano:** Valores binomiales
- **MEMO:** Textos de longitud ilimitada.
- **Moneda**
- **Objeto OLE:** Introducen fotos, gráficos e incluso hojas de cálculo.
- **Hipervínculo:** Enlace a una página web
- **Asistente para búsquedas:** Crea un campo que permite elegir el valor de otra tabla.

2.3. Explotación de los datos en este trabajo

En el caso de este trabajo final, se tiene un almacén de datos, definido como un conjunto de datos orientados al fútbol, integrados, no volátiles y variables en el tiempo, ya que las estadísticas que miden a cada jugador, se va actualizando en función del desempeño de estos en la realidad y a tiempo real.

Para poder extraer todo el conocimiento posible que nos brinda esta base de datos se lleva a cabo la extracción de información, tanto bruta como escondida de los datos mediante la explotación estadística. Para llevar a cabo esta acción, se puede explotar de varias formas, ya sea mediante la utilización de filtros a lo que ya se han hecho referencia anteriormente, consultas directas del almacén de datos, aplicando técnicas de minería de datos o un poco de todo lo dicho.

Un filtro resulta muy útil porque se seleccionan datos que tienen unas características que se buscan de antemano, lo que hace ahorrar mucho tiempo mediante las llamadas consultas. Estas consultas son operaciones sencillas que permiten explotar en primera instancia la información. Además, es el análisis más simple que se puede aplicar a un almacén de datos. Hay programas enteros dedicados a realizar consultas, como por ejemplo *SQL*.

Para sacar información de manera más sofisticada se aplican técnicas de minería de datos, estas técnicas no difieren demasiado de otras que ya se aplicaban antes; sin embargo, se crean en busca de solucionar las necesidades que requieren las nuevas tecnologías y negocios. Gracias a esto, se crea una nueva visión de la estadística, más aplicada a la realidad y que permite generar valor a partir de datos aislados. Estas nuevas necesidades a las que se han hecho referencia no son más que la necesidad de sacar información a volúmenes de datos, que cada vez son más amplios, para entender los datos que se tienen en el presente e intentar predecir comportamientos futuros de estos con el fin de aprovecharlos y estar preparados para solventar futuros problemas o planificar de manera más sofisticada un plan de actuación.

Estas técnicas, generalmente, buscan sacar patrones a partir de los datos de los almacenes; es decir, es extraído al completo a partir de la herramienta, no se basa en ningún modelo, si no que permite crear modelos propios tales como árboles de decisión, redes neuronales, regresiones lineales, regresiones logísticas, grafos, etc. que permitirán empezar a manipular los datos para lograr un fin.

Gracias a estas técnicas, se puede sacar qué factores son más significativos a la hora de explicar el comportamiento de una variable. Además, toda esta información ha de ser comprensible para el ser humano.

Antes de llegar al final del estudio de un almacén de datos, los seres humanos hemos creado diversas técnicas de recogida de información. Un ejemplo de esto se puede explicar fácilmente gracias a *WyScout*, empresa de la que se ha extraído la base de datos de este estudio. Esta empresa ha diseñado un *software*, que mediante la implementación de un video en él, es capaz de medir, de manera automática y a tiempo real, todas las acciones que lleva un jugador durante un partido. Esto se realiza durante todos los partidos que se juegan a nivel profesional en muchos países del mundo. En resumen, gracias a este desarrollo se tienen información de, prácticamente, cada jugador de cada equipo de cada país. Una cantidad de datos ingente que requiere tratamientos a priori y a posteriori para sacarle el máximo partido a la información.

WyScout no es la única empresa dedicada a la recogida de datos, hay empresas que destacan a nivel mundial, como por ejemplo *Opta* y *StatsBomb*, que trabajan con multitud de equipos y federaciones para ayudar a los equipos a trabajar con sus datos.

En el fútbol moderno, tal y como se ha explicado en el capítulo anterior, mucha gente aún no es consciente de lo que supone ayudarse de esta gran herramienta y ven con recelo meter ciencia en el deporte.

El método de trabajo que tienen los clubes en este ámbito, generalmente, se bifurca hacia dos opciones. La primera de estas opciones es contratar empresas externas especializadas en el *Big Data*, lo que ahorra mucho trabajo a los clubes a corto plazo, pero a la vez es más caro, es por esto que los equipos que optan por la otra opción, la cual es la de desarrollar una estructura interna especializada en este tema, son equipos con recursos más limitados con peores resultados a corto plazo, ya que requiere mucho tiempo formar y empezar a trabajar con personas que empiezan siendo inexpertas para, a la larga, acabar siendo expertos en el tema y a partir de ahí, dar una nueva dimensión al club a la hora de aprovechar esta herramienta que, al mismo tiempo, ayuda a las demás áreas dentro del club a tomar decisiones. El trabajo que requieren estos expertos no solo influye el ámbito deportivo-económico, sino también en lo social. De esta manera, al igual que una empresa, los clubes de fútbol tienen sus propias áreas de trabajo dedicadas al *marketing* y a la relación con sus aficionados. Así pues, un club puede permitirse estudiar las relaciones que tienen sus fans en redes sociales con el propio club y entre ellos y, a partir de esta información, lograr un beneficio que se traduce en una mejor imagen para los

aficionados y para la gente corriente, lo que ayuda a expandir la marca y a generar ingresos extras por la comercialización de productos u ofertas derivadas del *merchandising* del club que los propios aficionados demandan.

La mayoría de estas estrategias están basadas en el denominado aprendizaje automático, el cual está basado en una regresión o tendencia, cuya finalidad es la de predecir el comportamiento de una variable a partir de otras que contiene nuestro almacén de datos.

Estas estrategias a su vez se dividen en dos grandes grupos, por un lado, se tienen las estadísticas descriptivas (técnicas no supervisadas), mientras que por el otro lado están las predictivas (técnicas supervisadas). A su vez, dentro de las descriptivas se dividen, de forma general, en algoritmos de asociación y *clustering*. Las predictivas se agrupan en algoritmos de clasificación y de predicción.

En este estudio, se han realizado técnicas de estos dos grandes grupos. Se ha realizado un algoritmo de *clustering* basado en el K-medias, lo que entraría a formar parte de las técnicas no supervisadas, mientras que dentro de las técnicas supervisadas se han realizado modelos de regresión con el fin de predecir el comportamiento de alguna variable.

También en este trabajo se han realizado contrastes de hipótesis, lo que corresponde con la conocida inferencia estadística. En capítulos posteriores, veremos de manera mucho más profunda y de manera matemática el desarrollo que han tenido estas técnicas en nuestro trabajo y cómo pueden aplicarse en la realidad, ya sea para explicarla o para predecirla, pudiendo servir en el proceso a mejorar el entendimiento del juego mediante la clasificación de nuestras observaciones (jugadores) y la predicción de variables, para poder comparar a nuestros jugadores y medir su calidad de manera más objetiva a que si se fía esta medición a métodos únicamente subjetivos.

En este trabajo, no se ha dejado de lado el *Machine Learning* mencionado con anterioridad y, se ha querido explicar mediante modelos de regresión la variable que indica el número de goles que marcará un jugador durante la temporada.

Aún con todo, estas herramientas no son las única para sacar información de nuestra base. Al igual que se han realizado estas técnicas, este trabajo podría ampliarse mediante alguna más, como por ejemplo los árboles de decisión y algoritmos de asociación.

Un árbol de decisión toma este nombre porque la clasificación que realiza se trata de forma esquemática en forma de árbol. Es una representación muy simple del conocimiento obtenido, lo que lo convierte en un procedimiento muy simple. El primer sistema que utilizó los árboles de decisión fue CLS de Hunt en el año 1959, pero no fue hasta 1979 el año en el que Quinlan desarrolla la versión definitiva de este árbol denominado el sistema C4.5. El C4.5 se usa generalmente para clasificaciones, por lo que a menudo se ha descrito como un clasificador estadístico propio.

Las técnicas de asociación son descriptivas ene la Minería de Datos. Su objetivo es describir los datos sobre los que partimos y que permiten la búsqueda automática de reglas y patrones que relacionan el conjunto de las variables entre sí. Estas técnicas, como se puede intuir, son no supervisadas, por lo que no existen relaciones conocidas a priori con las que se puedan llegar a contrastar la validez de los resultados obtenidos.

Por lo tanto, usándolas se puede establecer relaciones o correlaciones entre dos situaciones que en apariencia son independientes.

Este tipo de técnicas tiene multitud de aplicaciones prácticas, como en el ámbito comercial para así conocer hábitos de compra de los clientes y a partir de ahí planificar ofertas personalizadas para sacar el mayor beneficio posible de los productos. También tiene una importancia capital en el ámbito sanitario, pues gracias a ellas se pueden prever e identificar factores de riesgo que produzcan la complicación de enfermedades. Otro ámbito al que llega esta herramienta es la minería web, de manera que se estudia el comportamiento de los internautas en su navegación por Internet para conocer qué les inquieta y, de esta manera, aprovechar la información resultante para estructurar la página web con el fin de optimizar el beneficio.

Uno de estos algoritmos más conocidos es el A Priori. Fue propuesto por Agrawal & Srikant en 1994, siendo el más simple, popular y utilizado desde entonces hasta nuestros días.

Capítulo 3: R y RStudio

3.1. Historia de R y RStudio

Con la llegada a la sociedad de las nuevas tecnologías, el aumento de la recogida de datos o información que deja cada usuario en la red en cada movimiento que realiza se ha maximizado. Las empresas se han visto obligadas a recolectar esta información con el objetivo de conocer mejor a sus potenciales clientes y, así, ofrecerles un servicio personalizado dependiendo de sus preferencias o gustos personales.

Para lograr este objetivo se han desarrollado, de manera general, dos tipos de herramientas complementarias.

Por un lado, se tienen herramientas dedicadas a la recogida de información. Por ejemplo, los datos que se analizan en este proyecto se han recogido mediante el *software* especializado WyScout, el cual basa su recogida en el análisis de vídeos de partidos de fútbol. Mediante redes neuronales, el programa fue desarrollado de tal manera, que recoge las acciones de cada jugador durante un partido sin necesidad de utilizar seres humanos en esta recogida, de esta manera se ha conseguido una herramienta muy potente de recogida de información, ya que se realiza de forma automática y al instante, ahorrando mucho tiempo y maximizando la cantidad de información que se puede recoger en contraparte a que se hiciera a mano mediante personas dedicadas a esto.

Una vez se han recogido los datos gracias a estas herramientas, se necesita otra que sea capaz de tratar estos datos y analizarlos para poder obtener verdadero valor de esta información. Así es como nacieron programas dedicados al *Big Data* y análisis de datos, como por ejemplo el *software* R y su interfaz gráfico más utilizado, RStudio.

El programa estadístico R se basa en un lenguaje de programación similar llamado S, desarrollado este último principalmente por Rick Becker y John Chambers a finales de la década de 1970. R por su parte es desarrollado por dos profesores del Departamento de Estadística de la Universidad de Auckland, Robert Gentlemen y Ross Ihaka en 1993, combinando la idea del ya mencionado S con el lenguaje Scheme, cuyo lenguaje es en el que está basado R.

3.2. Características principales

Este programa se caracteriza principalmente por su gran capacidad analítica en estadística, esto es que es capaz de desarrollar y analizar modelos de regresión lineales y no lineales, algoritmos de clasificación, como, por ejemplo, árboles de clasificación. Es capaz también de desarrollar series

temporales y, además, es muy eficiente a la hora de realizar contrastes de hipótesis. En el plano gráfico, R es capaz de generar sus propios gráficos de alta calidad basados en el programa LaTex. Es gracias a esto que todas sus gráficas tienen un grado de personalización muy alta según lo que requiera el usuario.

R con el tiempo ha tenido la capacidad de adaptarse a programas que también son herramientas muy utilizadas por los usuarios, tales como *Matlab*, *GNU Octave* o *Weka*. Gracias a su potente capacidad adaptativa es posible trabajar en R con las herramientas que proporcionan estos programas, minería de datos en caso de *Weka* y cálculo numérico por parte de *Matlab*. De esta manera R se ha convertido en una herramienta muy versátil y, por lo tanto, una de las preferidas por los usuarios a la hora de analizar o explotar almacenes de datos.

R es un proyecto a nivel mundial, es un lenguaje de código abierto, por lo que los propios usuarios pueden desarrollar sus propios paquetes e integrarlos en la aplicación. De esta manera, R tiene una gran capacidad de transformación y adaptación hacia las nuevas tecnologías o algoritmos que van surgiendo, ya que gracias a esto es muy complicado que quede obsoleto. Está en constante cambio y modernización por lo que siempre va a ser una herramienta muy útil para analistas.

Capítulo 4: Técnicas utilizadas

4.1. *Cluster* jerárquico

Los algoritmos basados en *clustering* juegan un papel fundamental en la extracción de información útil en grandes bases de datos. El objetivo final del *clustering* es agrupar las N observaciones de nuestro almacén de datos en un número k de *cluster*. Estas agrupaciones se dan debido a la similitud de las observaciones en cuanto a sus características.

4.1.1. Formulación

Este algoritmo empieza con un grupo que contiene todas las observaciones de nuestra base. Entonces, este *cluster* converge en dos o más *cluster*, los cuales tienen mayores diferencias entre ellos. El número de grupos a dividir viene dado o bien por el número de observaciones que haya o bien puede ser especificada por el usuario.

Pasos a realizar:

1. Empieza con un *cluster* que contiene todas las observaciones.
2. Calcula el diámetro de cada *cluster*. Este diámetro es la distancia máxima entre las observaciones dentro del *cluster*. Elige el *cluster* teniendo en cuenta este diámetro entre todos los *cluster* por los que lo ha dividido.
3. Encuentra el individuo **x** más diferente del *cluster* formado **C**. Coge este **x** y lo separa del *cluster* **C** original para formar un nuevo *cluster* independiente **N**. Todos los miembros del *cluster* **C** son asignados como $\mathbf{M_C}$.
4. Repite estos pasos hasta que los miembros de **C** y **N** no cambian.
5. Calcula la similitud de cada miembro de $\mathbf{M_C}$ al *cluster* **C** y **N**, y hace que el propio miembro con más similitud en $\mathbf{M_C}$ se mueva hacia su *cluster* más similar, **C** o **N**, actualizando los miembros de cada *cluster* en el proceso.
6. Repite los pasos 2, 3, 4 y 5 hasta que el número de *cluster* empiezan a ser los especificados por el usuario.

4.2. *Cluster* no jerárquico

El análisis de *cluster* no jerárquico tiene como objetivo encontrar una agrupación de objetos que maximice o minimice algún criterio de evaluación. Muchos de estos algoritmos asignarán objetos de forma iterativa a diferentes grupos mientras buscan algún valor óptimo del criterio.

En este trabajo se ha llevado a cabo el desarrollo de un *cluster* no jerárquico basado en el algoritmo *k-medias*.

4.2.1.K-medias

Cluster de K-medias es una técnica de agrupación en grupos particional basada en prototipos que intenta encontrar un número de *cluster* (k) especificado por el usuario, que están representados por sus centroides.

4.2.1.1. Procedimiento del algoritmo k-medias

Primero, elegimos k centroides iniciales, donde k es un parámetro especificado por el usuario; es decir, el número de agrupaciones deseadas. Luego, cada punto se asigna al centroide más cercano, y cada conjunto de puntos asignados a un centroide se denomina grupo. El centroide de cada grupo se actualiza luego en función de los puntos asignados al grupo. Repetimos la asignación y actualizamos los pasos hasta que ningún punto cambie de grupo, o de manera similar, hasta que los centroides permanezcan iguales.

Se consideran los datos cuya medida de proximidad es la distancia euclídea. Para nuestra función objetivo, que mide la calidad de una agrupación, se usa la suma del error al cuadrado (SSE), que también se conoce como dispersión.

En otras palabras, se calcula el error de cada punto de datos; es decir, su distancia euclídea al centroide más cercano, y luego se calcula la suma total de los errores al cuadrado. Dados dos conjuntos diferentes de conglomerados que son producidos por dos iteraciones diferentes de K-medias, es preferible el que tiene el error al cuadrado más pequeño, ya que esto significa que los prototipos (centroides) de este conglomerado son una mejor representación de los puntos en su conglomerado.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

Para ilustrar que K-medias no está restringido a los datos en el espacio euclídeo, consideramos los datos del documento y la medida de similitud del coseno:

$$similarity = \cos(\theta) = \frac{A * B}{||A|| * ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Para elegir el número óptimo de grupos a dividir la base, se realiza el denominado *Elbow Method*.

4.2.2.Elbow Method

Este criterio basa su idea en elegir el número según el cual la curva del gráfico empieza a aplanarse, de esta manera la varianza explicada varía muy poco de un número de grupos a dividir y su siguiente, sugiriendo que es en este punto, conocido como *Elbow Point*, el óptimo para realizar la partición. El método del codo se expresa mediante la suma del error al cuadrado:

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in S_k} ||X_i - C_k||_2^2$$

Se determina el punto central del *cluster* al azar. Cada centroide inicial de cada *cluster*, se calcula al azar gracias a los objetos disponibles de cada *cluster* k. Para calcular el próximo centroide del siguiente *cluster*, se usa la expresión siguiente:

$$v = \frac{\sum_{i=1}^{n} x_i}{n}; i = 1,2,3, \dots, n$$

Para calcular la distancia de cada objeto al centroide, se aplica la fórmula relativa a la Distancia Euclídea.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Por lo tanto, coloca cada objeto dentro del *cluster* con el centroide más cercano, de esta manera, mediante el proceso iterativo, va calculando nuevos centroides y añadiendo las observaciones dentro del *cluster* con el centroide más cercano a cada una hasta que todas las observaciones están agrupadas.

4.3. ACP

El Análisis de Componentes Principales o ACP, es una técnica cuyo fin es el de la reducción de la dimensión de un conjunto de datos.

4.3.1.Formulación

Hay varias maneras de aplicar esta técnica, pero en este trabajo se ha basado en el método que usa la matriz de correlaciones que viene dada por:

$$r_{ij} = \frac{cov(F_i, F_j)}{\sqrt{var(F_i)var(F_j)}}$$

La matriz de correlaciones es siempre diagonalizable, ya que es simétrica. El cálculo de sus valores propios, viene dado por:

$$\sum_{i=1}^{m} \lambda_i = m$$

Tiene como resultado *m*, los cuales hacen referencia a cada componente principal creado. Se identifican matemáticamente mediante los vectores propios de la matriz de correlaciones.

4.4. Modelos de Regresión

La regresión es un tipo de problema de aprendizaje automático supervisado, bajo el cual se establece la relación entre las características en el espacio de datos y el resultado, para predecir los valores del resultado, que son continuos por naturaleza.

El nombre en sí proporciona una explicación del concepto detrás de la regresión lineal; específicamente, asume una relación lineal entre las variables de entrada (x) y las variables de salida (y). La regresión lineal tiende a establecer una relación entre ellos al formular una ecuación que describe el resultado (y) como una combinación lineal de las variables de entrada (multiplicadas por las variables correspondientes aprendidas por el modelo del entrenamiento).

Entre los modelos de regresión se encuentran principalmente tres tipos de modelos:

- Regresión lineal simple
- Regresión logística
- Regresión lineal múltiple

En este trabajo, se ha desarrollado un análisis de regresión múltiple, con el fin de predecir una variable a partir del resto. Este modelo sigue la siguiente ecuación:

$$\mathbf{modelo} = \mathbf{y} = \boldsymbol{b_0} + \boldsymbol{b_1}\mathbf{x_1} + \boldsymbol{b_2}\mathbf{x_2} + \cdots + \boldsymbol{b_n}\mathbf{x_n} + \boldsymbol{\varepsilon}$$

donde *y* es la variable objetivo, $x_1, \ldots, x_n$ son las variables explicativas o independientes, mientras que desde $b_1, \ldots, b_n$ son los valores que describen la relación lineal entre cada variable explicativa y la objetivo.

b_0 es el denominado *intercept value*, que se define como el valor esperado de *y* cuando todas las variables explicativas toman el valor 0.

4.5. *ANOVA*

El análisis de la varianza o *ANOVA* es una técnica estadística que permite ver el efecto de un factor sobre la media de una variable continua. En este estudio, se ha realizado un *ANOVA* de un factor, por lo que en este apartado de procederá a explicar el desarrollo matemático de dicho análisis.

Los pasos a seguir en el *ANOVA* son:

1. Decidir la hipótesis nula y alternativa.
2. Dar un nivel de significación (generalmente es un valor de 0,05)
3. Calcular la media total, así como la de cada una de los factores medidos.

$$\overline{\boldsymbol{x}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{\boldsymbol{x_1} + \boldsymbol{x_2} + \ldots + \boldsymbol{x_n}}{\boldsymbol{n}}$$

4. Calcular la media de la suma de cuadrados total:

$$\widehat{\boldsymbol{S}}_{\boldsymbol{T}}^{\boldsymbol{2}} = \frac{\boldsymbol{TSS}}{\boldsymbol{N-1}}$$

5. Calcular la media de la suma de cuadrados del factor:

$$\widehat{\boldsymbol{S}}_{\boldsymbol{t}}^{\boldsymbol{2}} = \frac{\boldsymbol{SST}}{\boldsymbol{k-1}}$$

6. Calcular la suma de cuadrados del error:

$$\widehat{\boldsymbol{S}}_{\boldsymbol{E}}^{\boldsymbol{2}} = \frac{\boldsymbol{SSE}}{\boldsymbol{N-k}}$$

7. Calcular los grados de libertad de cada suma de cuadrados. En cuanto a la suma total los grados de libertad, se calculan como **N-1**, los grados correspondientes a la suma de cuadrados del factor son **k-1** y finalmente los grados de libertad de la suma de la suma de cuadrados del factor son **N-k**.
8. Calcular el valor **F**:

$$\boldsymbol{F_{ratio}} = \frac{\boldsymbol{Cuadrados\ Medios\ del\ Factor}}{\boldsymbol{Cuadrados\ Medios\ del\ Error}} = \frac{\widehat{\boldsymbol{S}}_{\boldsymbol{t}}^{\boldsymbol{2}}}{\widehat{\boldsymbol{S}}_{\boldsymbol{E}}^{\boldsymbol{2}}} = \frac{\boldsymbol{Intervarianza}}{\boldsymbol{Intravarianza}} \sim \boldsymbol{F_{k-1,N-k}}$$

9. Usar ese valor **F** dentro de la tabla de la distribución **F** para calcular el p-valor correspondiente y así rechazar o aceptar la hipótesis nula.

Además, para poder realizar un *ANOVA* se tienen que dar una serie de condiciones:

- Los datos deben ser normales (aunque también es una técnica bastante robusta ante datos extremadamente no normales).
- Los *outliers* pueden provocar errores a la hora de tomar el análisis y ser inválido. Por lo tanto, es recomendable tratar estos datos antes de realizar este análisis.
- Tiene que haber homocedasticidad entre los grupos, ya que esto supone que los datos han sido tomados de la misma población y, por tanto, tienen la misma media y varianza.
- Cuanto menor es el tamaño de los factores, más importancia tiene la homocedasticidad.
- Si cada factor tiene el mismo número de observaciones, pese a no haber homocedasticidad, el *ANOVA* se puede tomar como una técnica válida.
- Si alguna de estas condiciones hacen inviable el análisis *ANOVA*, se utilizaría la corrección de Welch (*Welch test*) sobre los datos con el fin de poder aplicar la técnica.

Capítulo 5: Metodología y Resultados

Partiendo de las bases de datos, mencionadas en capítulos anteriores, extraídas del programa especializado en la recogida de datos denominado *WyScout* se ha desarrollado un estudio que pretende imitar a los realizados por los analistas profesionales de fútbol. Obviamente hay ciertas limitaciones y mediciones subjetivas de las que no se pueden escapar que ya se verán más adelante.

En este trabajo, se han realizado dos estudios paralelos muy generales, que abarca sobre todo la caracterización de los jugadores de las bases de datos usadas, utilizando en el proceso técnicas que ya se han explicado con anterioridad, como un *cluster* jerárquico, no jerárquico, un análisis de componentes y finalmente un algoritmo k-medias.

5.1. Estudio general de 1ª División

Para el primer estudio, se ha utilizado el almacén llamado *JUGADORES*, el cual contiene los jugadores participantes de la primera división española de fútbol durante la temporada 2018-2019. Este estudio se ha basado en caracterizar las aportaciones ofensivas de los jugadores a sus equipos.

Tras un primer vistazo a la base, el primer paso que se realizó fue el de filtrarla para así poder trabajar más cómodamente con los datos, ganando la capacidad de no tener análisis sesgados. Este filtro consistió en eliminar los porteros, cuyas variables son totalmente diferentes a las de los jugadores de campo. Otro filtro fue el de quedarse con los jugadores que habían tenido una participación en la liga de, al menos 2700 minutos, cifra bastante significativa que indica que un jugador ha tenido un peso importante en el equipo a lo largo de la temporada. Tras acabar el filtro relativo a los jugadores, también se realizó uno respecto a las variables, ya que nuestra base se constituye de 111, por lo que muchas de estas variables no van a aportar nada al estudio. La manera de seleccionar las variables importantes fue totalmente subjetiva, basada en la idea de quedarse con las de carácter ofensivo. De esta manera, la base sobre la que se realizó este estudio quedó compuesta por 79 observaciones y 36 variables (estas variables pueden verse en el Anexo en el código correspondiente a este estudio). Durante la implementación de estos filtros surgió un problema importante, había jugadores que tenían el mismo nombre, por lo que el programa R no fue capaz de implementarlos. La manera de arreglar este problema fue la de añadir junto al nombre, el equipo del que ese jugador "repetido" forma parte.

Una vez realizado el filtro se procedió a realizar el siguiente *cluster* jerárquico:

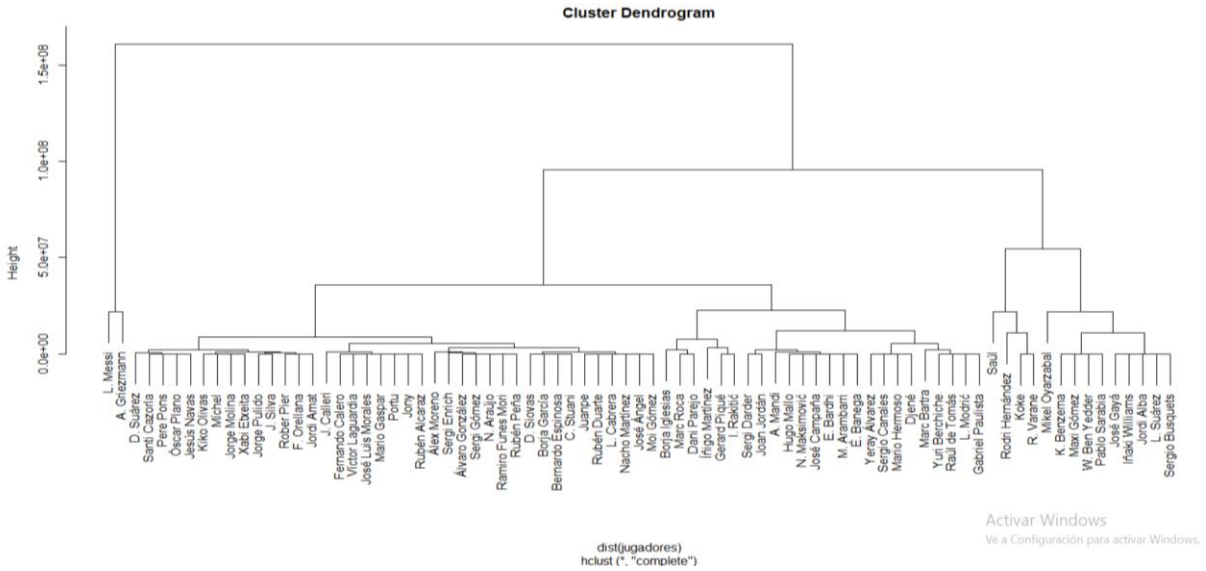

Ilustración 4. Representación en un *cluster* jerárquico de la base llamada *JUGADORES.*

Tras realizar el *cluster* jerárquico se puede concluir que existen 5 grupos diferenciados según su rendimiento ofensivo, destacando a dos jugadores por encima del resto, Lionel Messi y Antoine Griezmann, casualmente estos dos jugadores son probablemente los dos mejores jugadores ofensivos con mejor rendimiento durante esa temporada.

Tras el *cluster* jerárquico, el siguiente paso, la realización del *cluster* no jerárquico basado en k-medias. Este algoritmo requiere indicarle el número de centroides; es decir, grupos, a dividir nuestra base de datos. Para poder calcular el número de *cluster* se realizó el *Elbow Method.*

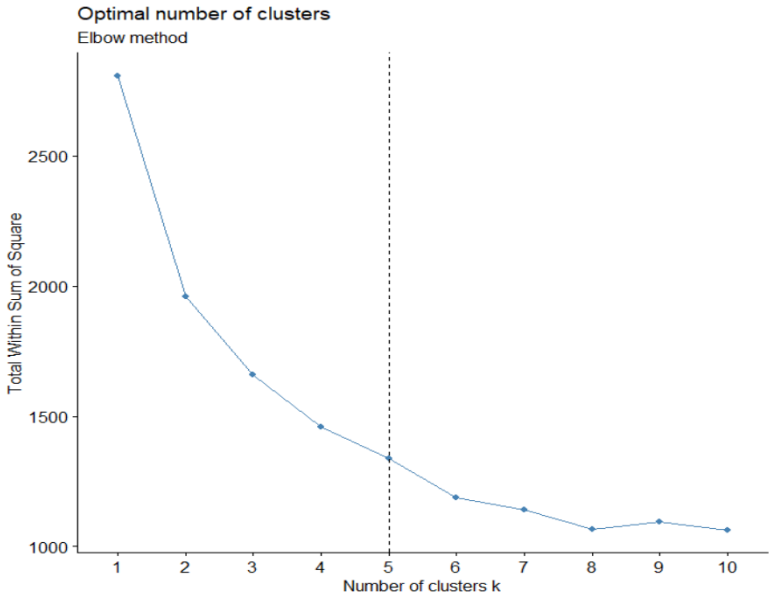

Ilustración 5. *Elbow Method* sobre la base *JUGADORES* con el fin de encontrar el número óptimo de *cluster* a dividir la base en el k-medias.

La interpretación de este gráfico sugiere que el conjunto de datos sea dividido en 5 grupos, así pues, el siguiente paso fue realizar el k-medias en 2D con 5 centroides.

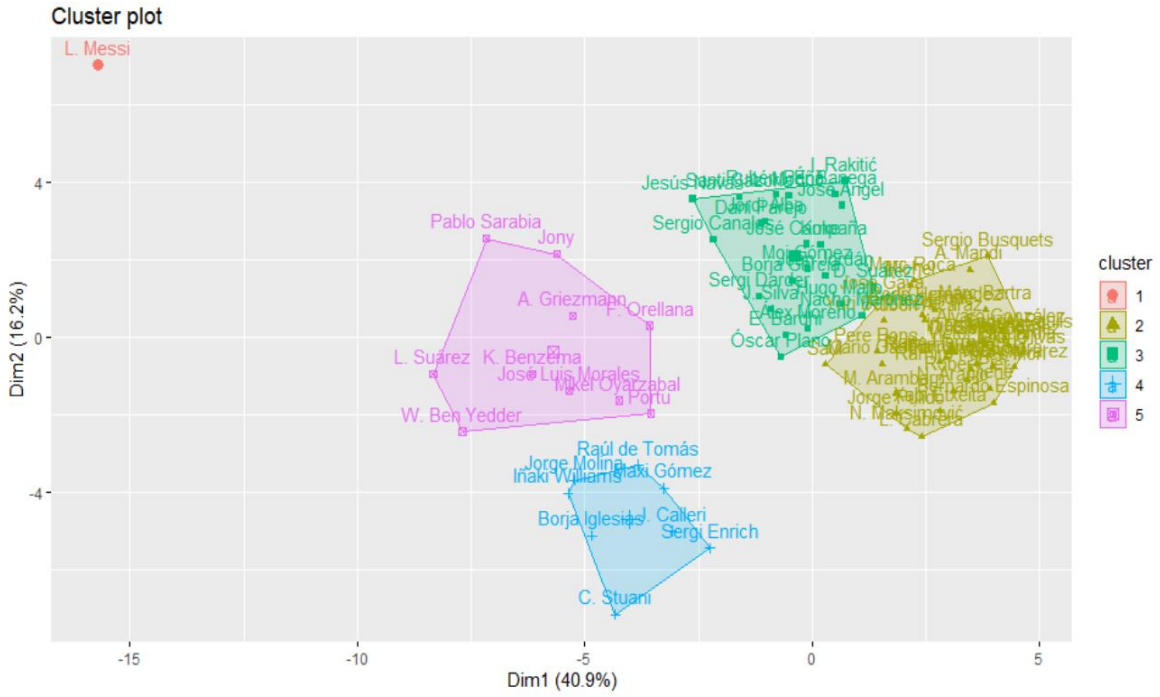

Ilustración 6. Algoritmo *k-medias* sobre la base llamada *JUGADORES*.

Probablemente, este estudio sea el más interesante de todo el trabajo, ya que se tiene la división de los jugadores en función de su rendimiento ofensivo. Es increíble ver como Leo Messi es tan bueno y tiene unas estadísticas tan excepcionales que se crea un *cluster* solo para él, además de estar bastante alejado de cualquier jugador, lo que indica que no hay absolutamente nadie que se le asemeje.

En el resto de *cluster*, se agruparon jugadores que generalmente tienen la misma función en sus equipos; así pues, si nos fijamos en el grupo número 5 (el morado en la gráfica), se tienen a los jugadores con mejor bagaje ofensivo de LaLiga si obviamos a Leo Messi. En este *cluster*, tenemos a jugadores diferenciales como Benzemá para el Real Madrid, Luis Suárez en el Barcelona, Griezmann en el Atlético de Madrid o Ben Yedder para el Sevilla.

En el *cluster* número 4 (azul en nuestro dibujo), se tienen delanteros que absorben mucho juego en sus equipos; es decir, la mayoría de estos equipos tienen un estilo directo que aprovechan la fortaleza física de sus delanteros y su buen juego aéreo para buscarles con pases por alto y así avanzar en el campo, con el fin de acercarse a la portería contraria y marcar gol. Absolutamente todos los jugadores que se encuentran aquí basan su juego en aguantar los balones que le llegan o rematarlos, excepto Williams, que no los aguanta o remata, si no que recibe los balones en largo aprovechando su velocidad para sobrepasar a las defensas contrarias.

En el grupo número 3 (el señalado cómo verde en el gráfico), se encuentran los principales generadores ofensivos de LaLiga; esto es, los encargados de empezar a crear los ataques de sus equipos y sobre los que se aprovechan los jugadores que hemos visto en los otros grupos. Este *cluster* está formado por este tipo de jugadores, pero también por jugadores que dan mucha profundidad a sus equipos como es el caso de Jesús Navas en el Sevilla, Jordi Alba en el Barcelona o José Ángel en el

Éibar, capaces de generar una gran cantidad de ocasiones de ataque por las bandas y realizar muchos centros que aprovechan los delanteros. Es por esto que este grupo ha sido caracterizado como el que forman los generadores ofensivos de los equipos.

En el *cluster* número 2, tenemos a los jugadores defensivos con poca aportación ofensiva en el juego.

Ya se tiene interpretado en k-medias en 2D o dos dimensiones, pero para dar sentido matemático a la explicación de por qué los jugadores se han colocado de esta manera en el espacio, se realizó un biplot que permitió ver las características concretas de cada *cluster*.

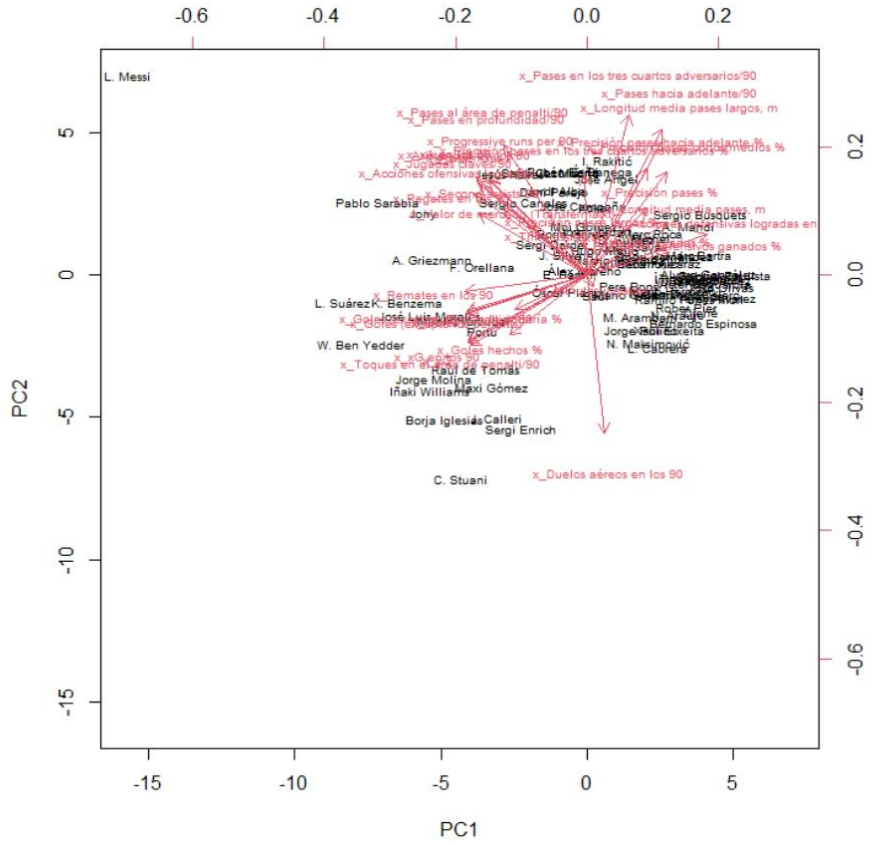

Ilustración 7. *Biplot* sobre la base llamada *JUGADORES*.

Pese a que se han eliminado mediante los filtros un cantidad considerable de observaciones y variables, aún hay un número significativo de ellas que dificultan un poco la visualización de esta técnica, pero si se hace un esfuerzo se pueden discernir las variables que hacen que cada jugador se coloque en un aparte u otra y, es evidente que la explicación dada anteriormente coincide con el peso de las variables de los jugadores para cada *cluster*.

Tras realizar el k-medias en 2D se tiene que las dos dimensiones forman un total de 57,1% de la varianza explicada respecto a los datos originales. Es por esto que se realizó un Análisis de Componentes Principales, que ayudó a construir 3 ejes que serán la base del gráfico en 3D del mismo k-medias que hemos visto.

La aportación de varianza explicada de cada dimensión mediante el ACP viene dada por la función en R llamada *fviz_screeplot* del paquete *FactoMineR*.

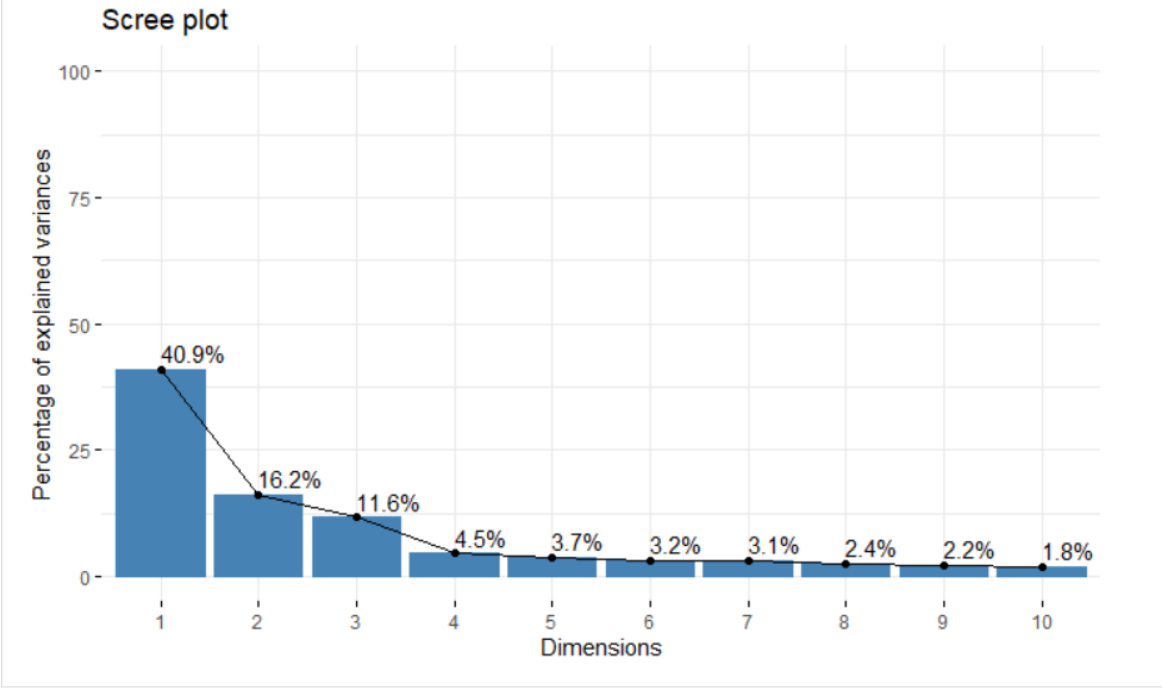

Ilustración 8. Representación de la varianza que explica cada dimensión del ACP.

Como podemos ver, el gráfico en tres dimensiones tendrá una varianza explicada del 40,9 + 16,2 + 11,6 = 68,7 %. Por lo que se tiene un 11,6% de varianza explicada respecto al gráfico en 2D.

Ahora veremos mediante la función en R denominada *corrplot*, qué variables forman cada dimensión.

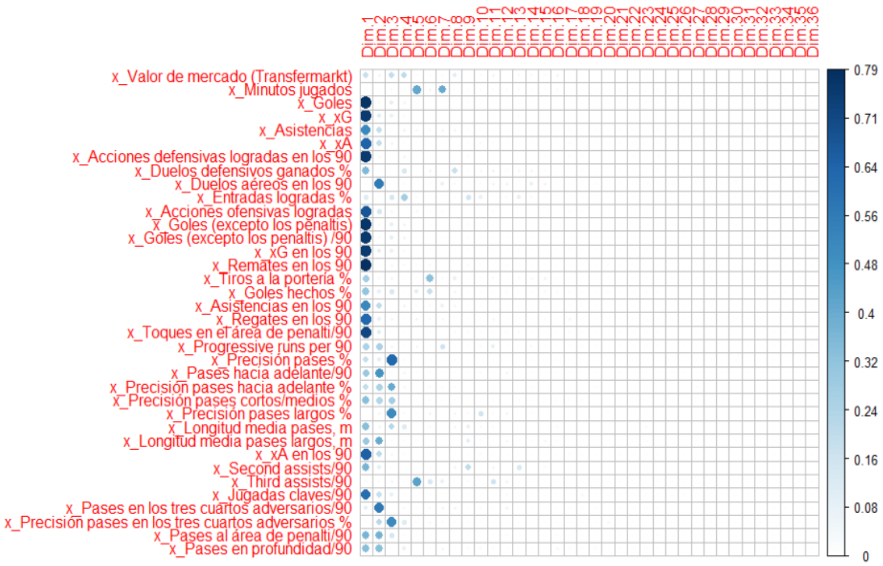

Ilustración 9. Representación del peso de las variables en cada dimensión creada en el ACP.

Finalmente, se realizó el gráfico en 3D. En este gráfico surge un problema, pues es interactivo y Word no permite integrar este tipo de archivos. Pese a ello mediante fotos del gráfico podemos tener una idea de los *cluster* de manera más amplia que en el de dos dimensiones.

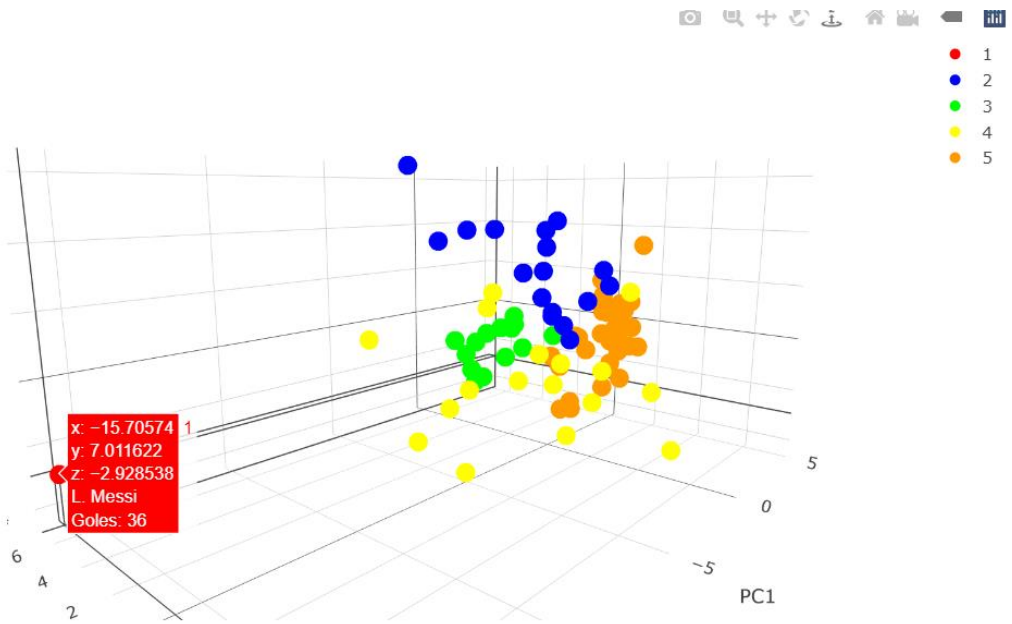

Ilustración 10. Representación de los *cluster* en 3 dimensiones.

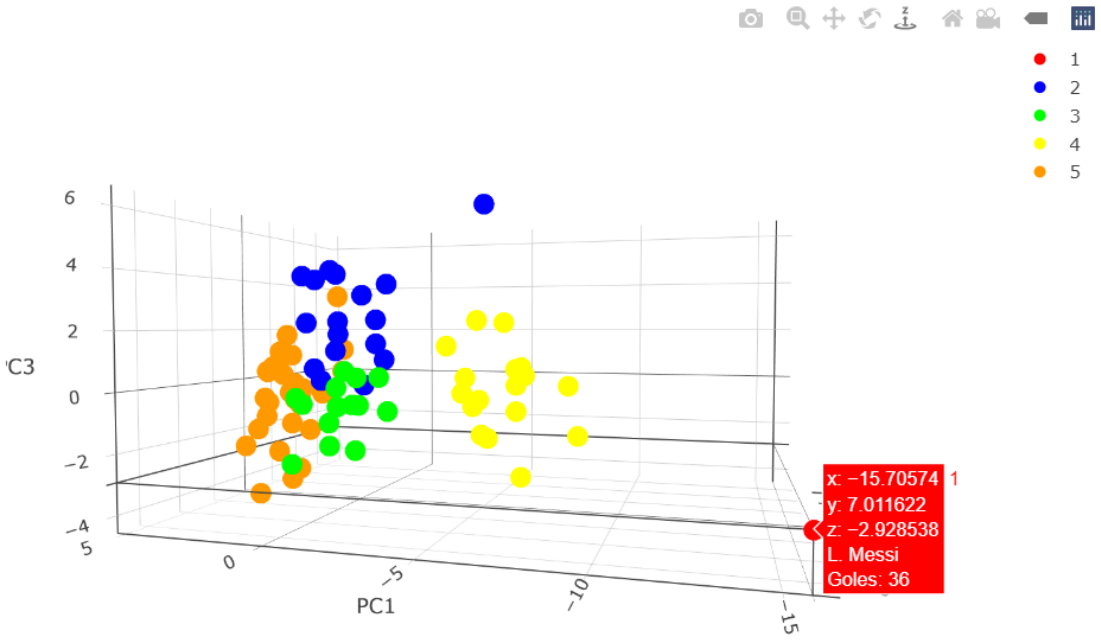

Ilustración 11. Representación de los *cluster* en 3 dimensiones.

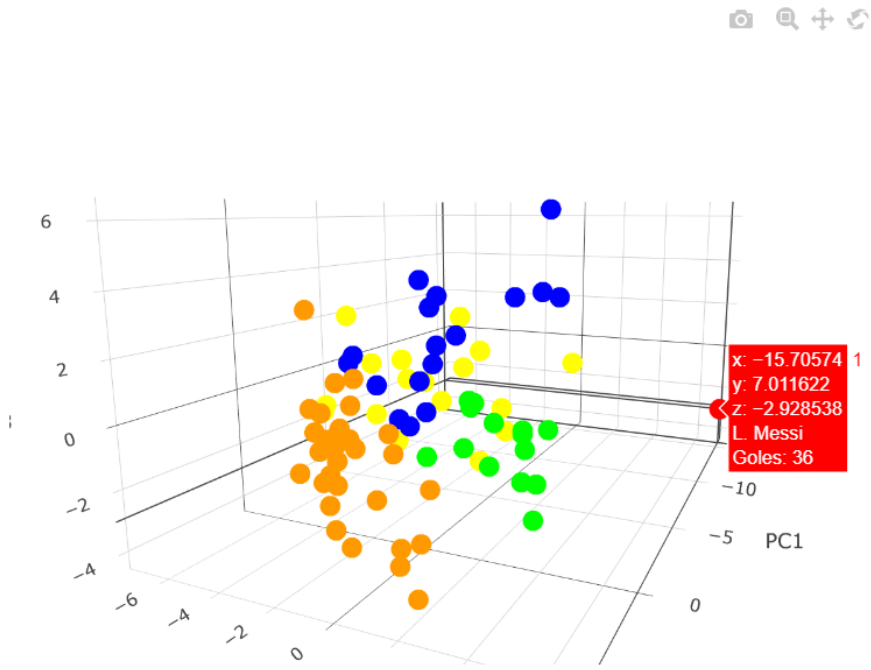

Ilustración 12. Representación de los *cluster* en 3 dimensiones.

Este gráfico ha sido integrado en *plotly.com* con el fin de poder visualizarlo de manera interactiva. Aquí se encuentra el enlace que contiene el gráfico en la página web.

5.2. Estudio *Box to Box* 2ªB

Como se ha venido apuntando en todo este trabajo, se ha realizado un estudio paralelo con la base de datos denominada *SEGUNDAB.*

Al igual que el anterior estudio, se ha filtrado esta base de datos, la cual contiene todos los jugadores que han participado en la Segunda División B española durante la temporada 2019/2020. Los filtros realizados han sido, por parte de las observaciones, quedarse con los centrocampistas y por parte de los campos, quedarse con las nueve variables que más caracterizan a un medio centro *Box to Box*. En este estudio se hizo directamente un k-medias en 2D previo paso por un *Elbow Method* que nos indique en cuántos grupos dividir la base.

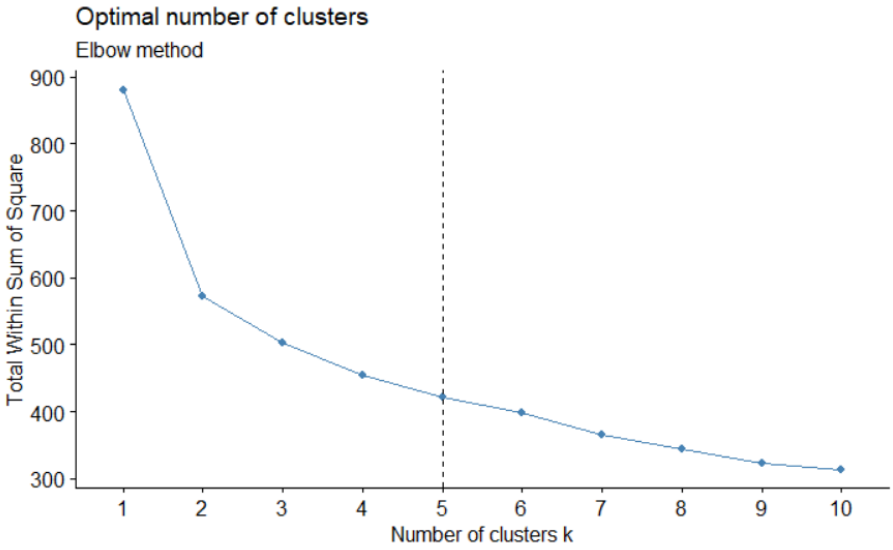

Ilustración 13. *Elbow Method* para la base llamada *SEGUNDAB.*

Al igual que el anterior estudio de Primera División, los grupos óptimos a dividir son 5.

Así pues se realizó el gráfico de *cluster* basado en el algoritmo de k-medias.

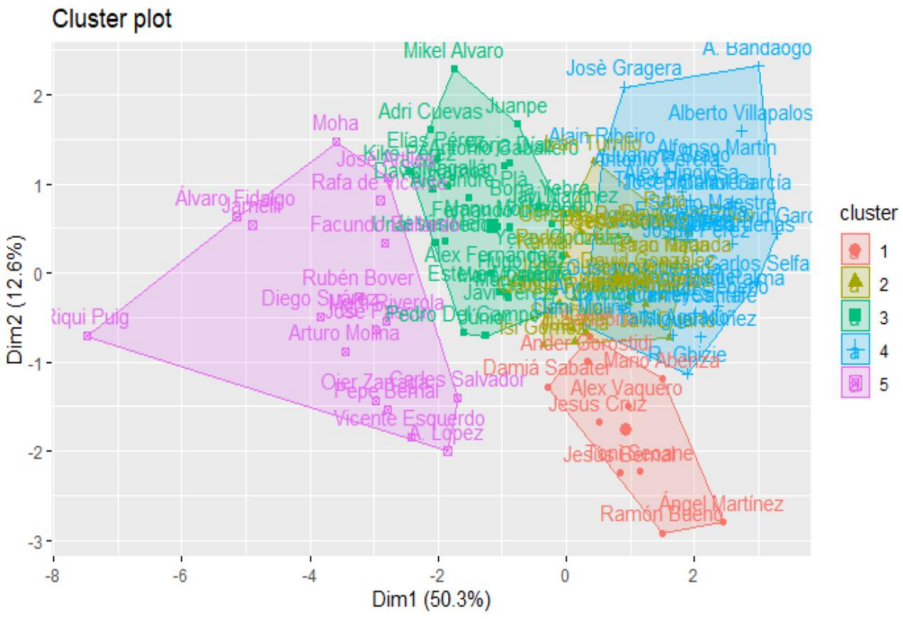

Ilustración 14. Visualización de los *cluster* de la base llamada *SEGUNDAB*

Una vez divididos los centrocampistas hay que ver qué características tiene cada *cluster* mediante la realización de un *Biplot.*

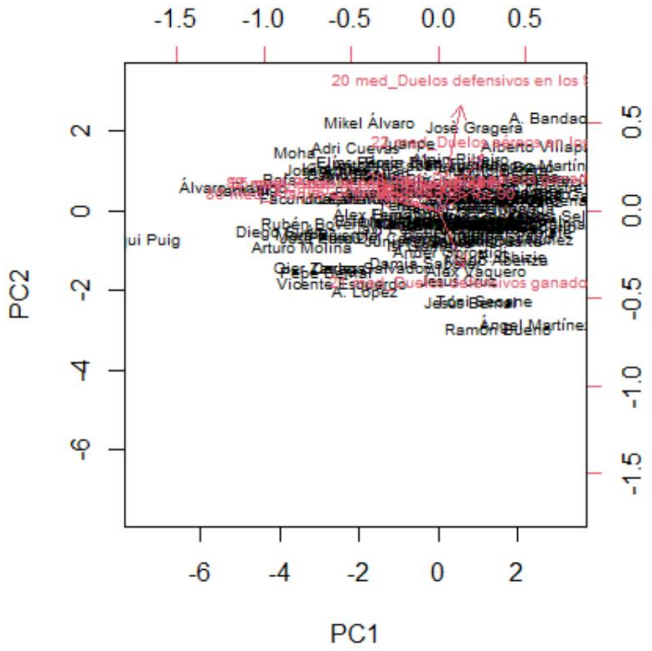

Ilustración 15. *Biplot* de la base *SEGUNDAB.*

Como este Biplot es muy difícil de interpretar, debido a que la cantidad de observaciones que no dejan ver bien las variables, se realizó un *Loading Plot* que permita ver las flechas del gráfico y su peso en la colocación de las observaciones, así pues, un jugador con un alto valor en la variable de una flecha, seguirá la dirección de esta.

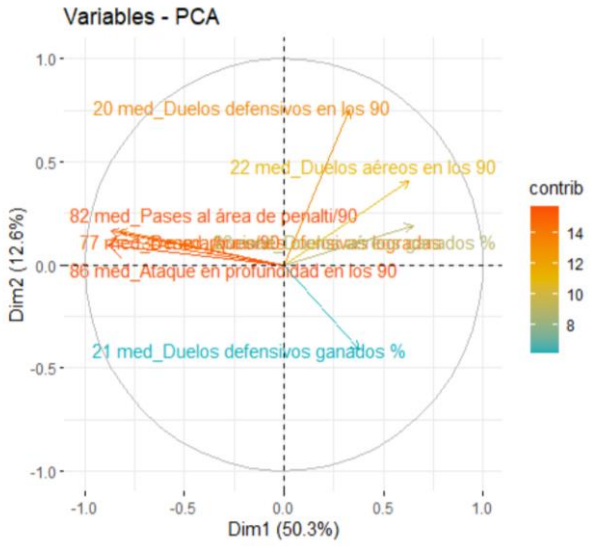

Ilustración 16. *Loading Plot* del *Biplot* creado.

Ya se tiene el motivo por el que los jugadores se sitúan en un lugar del gráfico u otro.

El siguiente paso fue realizar el *screeplot* y el peso de las variables que forman cada dimensión.

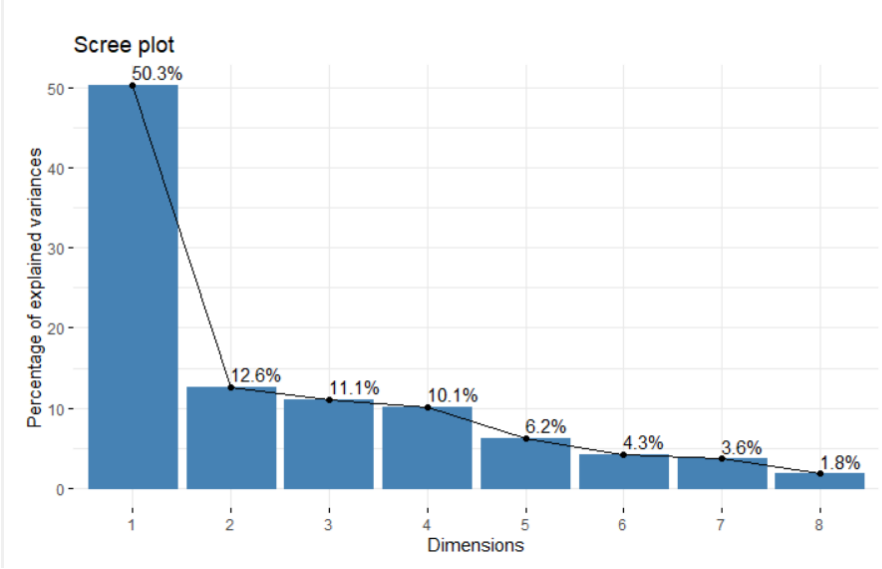

Ilustración 17. Visualización de la varianza que explica cada dimensión del ACP.

Ilustración 18. Representación del peso que tienen las variables sobre las dimensiones del ACP.

Como se pudo apreciar en este gráfico, se tiene la dimensión 5 formada por las dos variables relativas al juego aéreo, por lo que esta dimensión se escaló con valores del 1 al 100 para caracterizar el nivel en el juego aéreo de cada jugador. Así que, además de usar las 3 primeras dimensiones como ejes del gráfico, se usó la dimensión 5, teniendo un total de 74% de varianza explicada

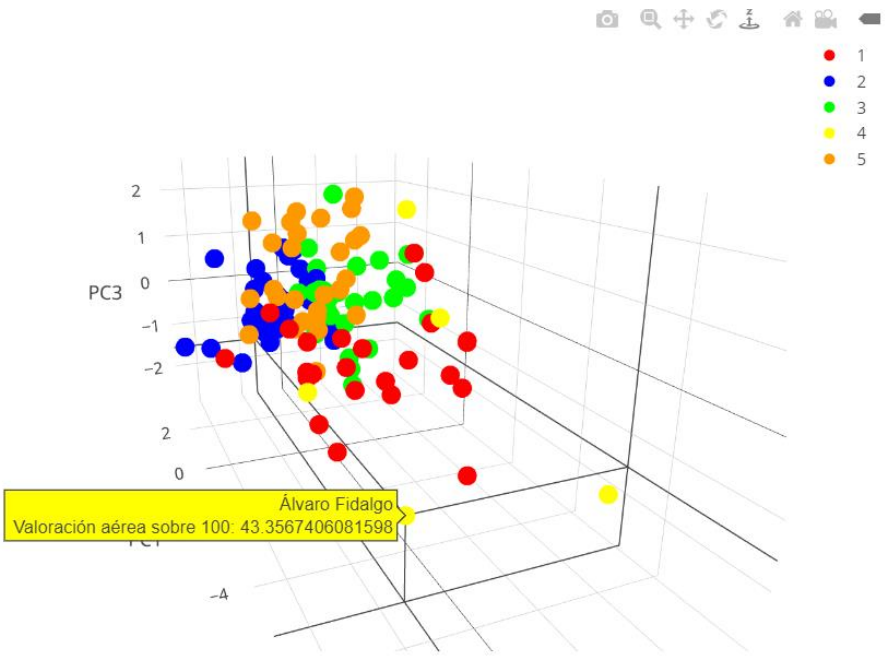

Ilustración 19. Representación en 3 dimensiones de los *cluster*

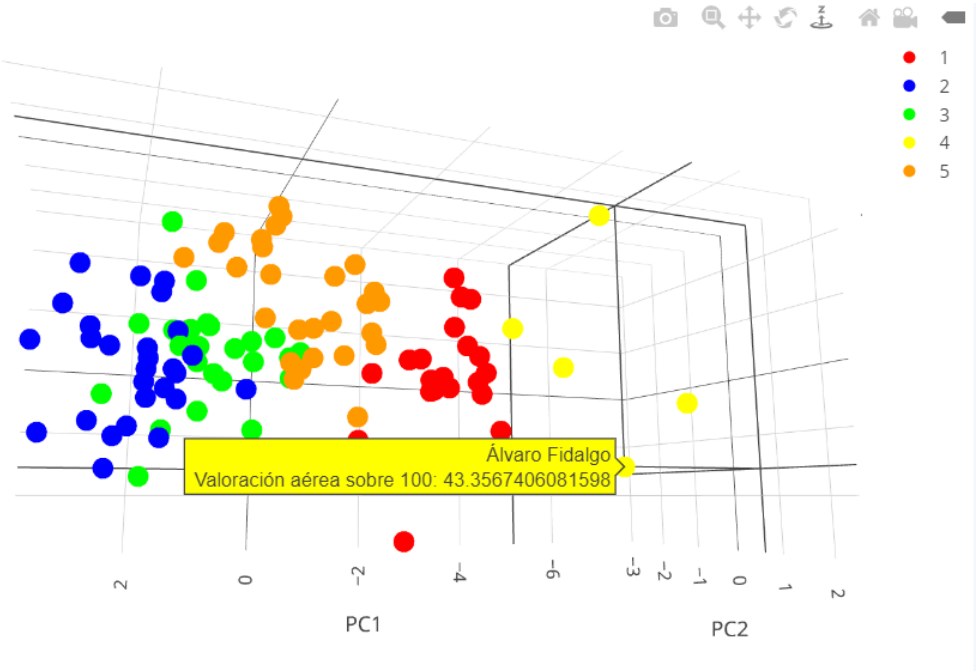

Ilustración 20. Representación en 3 dimensiones de los *cluster*

Al igual que el anterior gráfico, en este enlace se podrá visualizarlo de manera interactiva.

Como conclusión final del estudio *Box to Box*, se concluyó que los jugadores situados en el espacio Una vez realizado el gráfico podemos concluir que los jugadores más cercanos a ser un *Box to Box* son los que se encuentran entre (0, -6) de PC1, , (0, 2) de PC2 y (0,-2) de PC3, ya que este espacio sitúa a los jugadores en función de sus buenas capacidades ofensivas que además tienen un gran número de

duelos defensivos y un gran porcentaje de dichos duelos ganados. Cómo conclusión se tuvo que Álvaro Fidalgo es probablemente el jugador que mejor encarna las características de este perfil de medio centro.

5.3. Predicción de goles del almacén de datos llamado *Delanteros_19-20*

5.3.1. Modelos construidos

Se importa la base de datos con 3102 observaciones y 111 columnas. Cada observación se refiere a un jugador, mientras que las columnas dan las características de los mismos. Este estudio se centra en predecir los goles de los delanteros. Por tanto, se filtró la base de datos para tener en cuenta solo los delanteros que superen la media de minutos jugados. La base final tiene 543 observaciones y 22 variables características de los delanteros.

Antes de crear un modelo hay que dividir la base en dos. Una parte llamada *train* que va a contener el 80% de las observaciones y otra parte llamada *test* tendrá el 20% restante.

En este estudio se han construido cuatro modelos de regresión múltiple con el fin de predecir el número de goles de cada jugador. Los tres primeros están formados por las siguientes variables:

- **Modelo 1**: Valor de mercado, Minutos jugados, xG, Remates y tiros a portería.
- **Modelo 2**: Acciones ofensivas, desmarques, toques en el área de penalti, partidos jugados y duelos aéreos.
- **Modelo 3**: Utiliza todas las variables disponibles
- **Modelo *StepAIC*:** Este modelo utiliza un algoritmo en el que empieza utilizando todas las variables sobre el modelo y va eliminando mediante pasos las que no son significativas hasta dar con un modelo en el que todas sus variables lo son.

5.3.1.1. Modelo 1

```
Call:
lm(formula = x_Goles ~ ., data = input1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8358 -1.0455 -0.1147  1.0409  9.0301

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -3.540e+00  6.017e-01  -5.883 8.14e-09 ***
x_Valor.de.mercado..Transfermarkt. -7.517e-08  1.467e-07  -0.512 0.608575
x_Minutos.jugados                9.733e-04  2.489e-04   3.911 0.000107 ***
x_xG                             8.601e-01  5.696e-02  15.102  < 2e-16 ***
x_Remates.en.los.90              2.857e-01  1.926e-01   1.483 0.138734
x_Tiros.a.la.portería..          6.408e-02  8.504e-03   7.535 2.93e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.818 on 428 degrees of freedom
Multiple R-squared:  0.697,     Adjusted R-squared:  0.6934
F-statistic: 196.9 on 5 and 428 DF,  p-value: < 2.2e-16
```

Tabla 1. Salida en pantalla del modelo 1 creado para la predicción de goles de la base llamada *Delanteros_19-20.*

Como se puede apreciar en la imagen, este modelo tiene como variables los minutos jugados, xG y tiros a puerta.

Este modelo fue entrenado con la parte llamada *train* y se usó la parte llamada *test* para comparar la predicción del número de goles de cada jugador con las reales que tuvo en esa temporada. El resultado de esta predicción fue que el modelo 1 tuvo un 24,77% de acierto.

5.3.1.2. Modelo 2

```
Call:
lm(formula = x_Goles ~ ., data = input2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2285 -1.9525 -0.4101  1.3964 16.5820

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -4.596235   1.026368  -4.478 9.67e-06 ***
x_Acciones.ofensivas.logradas  -0.273384   0.107556  -2.542   0.0114 *
x_Desmarques.90                 0.560572   0.334774   1.674   0.0948 .
x_Toques.en.el.área.de.penalti.90  1.471685   0.179385   8.204 2.75e-15 ***
x_Partidos.jugados              0.262808   0.037703   6.970 1.20e-11 ***
x_Duelos.aéreos.en.los.90      -0.007943   0.046471  -0.171   0.8644
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.863 on 428 degrees of freedom
Multiple R-squared:  0.2489,    Adjusted R-squared:  0.2401
F-statistic: 28.37 on 5 and 428 DF,  p-value: < 2.2e-16
```

Tabla 2 . Salida en pantalla del modelo 2 creado para la predicción de goles de la base llamada Delanteros_19-20.

En este modelo las variables significativas fueron todas excepto la variable que mide la cantidad de duelos aéreos que realiza un jugador por partido. Este modelo tuvo un 16,51% de acierto en las predicciones para los jugadores que forman parte de la base *test*.

5.3.1.3. Modelo 3

```
Call:
lm(formula = x_Goles ~ ., data = input3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5676 -0.6988 -0.0400  0.5160  5.0212

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -2.372e+00  8.436e-01  -2.812 0.005158 **
x_Edad                            -3.340e-04  1.398e-02  -0.024 0.980950
x_Valor.de.mercado..Transfermarkt.  1.636e-08  1.017e-07   0.161 0.872229
x_Partidos.jugados                 2.179e-02  2.542e-02   0.857 0.391978
x_Minutos.jugados                  1.355e-03  2.612e-04   5.186 3.37e-07 ***
x_xG                               5.809e-01  4.671e-02  12.436  < 2e-16 ***
x_Asistencias                      1.691e-02  5.041e-02   0.335 0.737533
x_xA                              -1.018e-01  8.561e-02  -1.189 0.235190
x_Piederecho                       9.366e-02  1.770e-01   0.529 0.596904
x_Pieizquierdo                     1.687e-01  2.117e-01   0.797 0.425890
x_Duelos.aéreos.en.los.90         -3.676e-02  2.275e-02  -1.615 0.106994
x_Duelos.aéreos.ganados..         -6.205e-03  6.461e-03  -0.960 0.337439
x_Acciones.ofensivas.logradas      1.922e-01  8.315e-02   2.312 0.021259 *
x_Goles.de.cabeza.90               4.672e+00  1.193e+00   3.915 0.000106 ***
x_Remates.en.los.90                1.099e+00  1.489e-01   7.382 8.65e-13 ***
x_Tiros.a.la.portería..           -6.602e-03  6.568e-03  -1.005 0.315382
x_Goles.hechos..                   1.969e-01  9.263e-03  21.261  < 2e-16 ***
x_Duelos.ofensivos.en.los.90      -2.664e-02  2.767e-02  -0.963 0.336308
x_Duelos.ofensivos.ganados..      -3.888e-02  1.142e-02  -3.404 0.000730 ***
x_Toques.en.el.área.de.penalti.90 -4.304e-01  9.817e-02  -4.385 1.48e-05 ***
x_Desmarques.90                   -1.281e-01  1.513e-01  -0.847 0.397638
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.179 on 413 degrees of freedom
Multiple R-squared:  0.877,     Adjusted R-squared:  0.8711
F-statistic: 147.2 on 20 and 413 DF,  p-value: < 2.2e-16
```

Tabla 3. Salida en pantalla del modelo 3 creado para la predicción de goles de la base llamada Delanteros_19-20.

En este modelo las variables significativas son los minutos jugados, xG, acciones ofensivas por partido, goles de cabeza por partido, remates por partido, porcentaje de goles hechos, toques en el área por partido y duelos ofensivos ganados por partido. Este modelo tuvo un 42,21% de acierto en las predicciones para los jugadores que forman parte de la base *test*.

5.3.1.4. Modelo StepAIC

Este modelo empezó con todas las variables disponibles y en su inicialización es igual que el modelo 3 que ya se ha visto. El proceso iterativo, en el cual se van eliminando las variables no significativas del modelo hasta quedarse con un modelo óptimo en el que todas sus variables lo son. Dio como resultado el siguiente modelo:

```
Call:
lm(formula = x_Goles ~ x_Minutos.jugados + x_xG + x_Duelos.aéreos.en.los.90 +
    x_Acciones.ofensivas.logradas + x_Goles.de.cabeza.90 + x_Remates.en.los.90 +
    x_Goles.hechos.. + x_Duelos.ofensivos.ganados.. + x_Toques.en.el.área.de.penalti.90,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6294 -0.7322 -0.0396  0.4869  5.4040

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.5974634  0.5676802  -4.576 6.25e-06 ***
x_Minutos.jugados                0.0014153  0.0001713   8.262 1.85e-15 ***
x_xG                             0.5781255  0.0447836  12.909  < 2e-16 ***
x_Duelos.aéreos.en.los.90       -0.0463007  0.0193230  -2.396 0.017001 *
x_Acciones.ofensivas.logradas    0.1066585  0.0511752   2.084 0.037742 *
x_Goles.de.cabeza.90             4.8277839  1.1491851   4.201 3.24e-05 ***
x_Remates.en.los.90              1.0838413  0.1406478   7.706 9.29e-14 ***
x_Goles.hechos..                 0.1923038  0.0082145  23.410  < 2e-16 ***
x_Duelos.ofensivos.ganados..    -0.0361307  0.0102489  -3.525 0.000469 ***
x_Toques.en.el.área.de.penalti.90 -0.4098573  0.0934545  -4.386 1.46e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.173 on 424 degrees of freedom
Multiple R-squared:  0.8752,    Adjusted R-squared:  0.8725
F-statistic: 330.4 on 9 and 424 DF,  p-value: < 2.2e-16
```

Tabla 4. . Salida en pantalla del modelo *StepAIC* creado para la predicción de goles de la base llamada Delanteros_19-20.

Este fue el mejor de los construidos, ya que tuvo un 50,45% de acierto en las predicciones para los jugadores que forman parte de la base *test*.

5.4. Contrastes de hipótesis: *ANOVA*

Siempre se ha dicho en el fútbol que los zurdos tienen mejor pie que los diestros, esto quiere decir que son más hábiles a la hora de pasar o tirar a puerta. Es por esto que el siguiente paso en el estudio fue el de realizar dos contrastes de hipótesis, por un lado veremos si hay diferencias significativas entre zurdos o diestros en cuanto a la precisión de pase en nuestra base de datos llamada *JUGADORES*, y por otra parte veremos el mismo contraste para el porcentaje de tiros a puerta.

Al igual que en el resto de estudios, las filas se filtraron conforme a jugadores que superan los 2700 minutos de juego durante el año, eliminando al igual que antes a los porteros.

En este caso el contraste usado es el *ANOVA*, ya que permite comparar una variable categórica con una numérica.

5.4.1. Precisión en el pase: zurdos vs diestros

Antes de realizar el contraste se calculó la media de la precisión en el pase tanto para zurdos como para diestros.

```
derecho izquierdo
82.50421  78.60455
```

Tabla 5. Media de la precisión en el pase de los zurdos y diestros de la base llamada *JUGADORES*.

Así pues, vemos que los diestros tienen más precisión de media que los zurdos, pero para saber si estas medias son significativas realizamos un contraste *ANOVA*:

```
             Df Sum Sq Mean Sq F value Pr(>F)
jugadores$Pie  1    241  241.39   4.912 0.0296 *
Residuals     77   3784   49.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 6. Salida en pantalla de los resultados del *ANOVA*.

Tras realizar el contraste *ANOVA*, vemos que el p-valor es significativo al ser menor que 0,05, así pues podemos decir que hay diferencias en la precisión en el pase entre los zurdos y los diestros, siendo estos últimos más precisos.

5.4.2. Precisión en el tiro: zurdos vs diestros

Antes de realizar el *ANOVA*, al igual que en el contraste anterior, se calculó la media de la precisión en el tiro tanto para zurdos como para diestros, teniendo como resultado:

```
derecho izquierdo
34.95702  34.00409
```

Tabla 7. Media de la precisión en el tiro de los zurdos y diestros de la base llamada *JUGADORES*.

A primera vista, la precisión es bastante similar, así que es muy posible que no hubiera diferencias significativas entre los diestros y los zurdos. Se realiza el contraste *ANOVA* para confirmarlo:

```
              Df Sum Sq Mean Sq F value Pr(>F)
jugadores$Pie  1     14   14.41   0.092  0.763
Residuals     77  12076  156.84
```

Tabla 8. Salida en pantalla de los resultados del *ANOVA*.

Como el p-valor es superior a 0,05 se puede concluir que no hay diferencias significativas en la precisión de tiros a puerta entre zurdos y diestros.

Conclusiones

Se ha realizado un estudio de minería de datos sobre un almacén de fútbol. Este proceso ha abarcado desde la recopilación y edición de las bases de datos generadas por *WyScout*, hasta técnicas sofisticadas de la minería con el fin de sacar un valor de los datos. De hecho, una parte de este estudio ha ayudado a un analista de datos profesional en su labor como tal en un club de la Segunda División B española. Estos análisis han servido para analizar bases de diferentes ligas profesionales, por lo que se podrán realizar análisis con los datos de temporadas más recientes y actualizadas hasta el día de hoy. Todos los análisis se han realizado mediante programación basada en R para llegar a los objetivos que se marcaron al inicio del trabajo. Todo el código realizado se ha desarrollado con el fin de conseguir realizar dichos objetivos, pero, además, se ha creado de tal manera que es servible para otras bases de datos si se realizan pequeñas modificaciones al código.

Este trabajo ha conseguido transformar los números y valores recogidos mediante variables de una base de datos en información real y en la caracterización de los jugadores de fútbol estudiados sin necesidad de verles jugar. De esta manera, un ojeador podría filtrar un tipo concreto de jugador de entre miles de ellos, ahorrando muchos recursos en su estudio y siendo mucho más incisivo a la hora de encontrar al jugador que requiera su club.

A continuación se muestran las conclusiones principales de este trabajo:

1. El primer resultado de este estudio es el *cluster* jerárquico. Esta técnica nos muestra que los jugadores están agrupados según sus características. Este es un gran primer análisis porque podemos comenzar a desarrollar la idea del tipo de jugadores ofensivos que se tienen en el caso del primer estudio realizado o los tipos de jugadores *Box to Box* que hay en el caso de nuestro segundo análisis. Esto tiene una desventaja importante, tenemos el tipo de jugadores, pero no tenemos sus características respecto a las variables del *dataset* concretas.

2. En el segundo análisis de la investigación se aplica el ACP (Análisis de Componentes Principales) preparando los datos para aplicar un futuro algoritmo de *cluster* basado en el k-medias. Esta es una parte muy importante del estudio, de esta manera hemos construido los ejes de la gráfica del futuro k-medias.

3. Una vez realizado el ACP, se llevó a cabo un algoritmo denominado *Elbow Method* con el fin de concretar el número de grupos que habrá en nuestra base. Este algoritmo dio como resultado que el almacén debería dividirse en 5 grupos en los dos estudios. En este trabajo, se programó un gráfico 3D interactivo del k-medias, pero no se puede integrar en este tipo de archivo de manera interactiva, por lo que ponemos fotos del mismo.

4. En el caso del estudio con los jugadores ofensivos, el algoritmo k-medias muestra los diferentes tipos de jugadores en LaLiga. Por ejemplo tenemos un *cluster* (*cluster* 5) con los jugadores en los que sus equipos basaron sus ataques como Ben Yedder en el Sevilla FC, Benzema en el Real Madrid, Griezmann en el Atlético de Madrid… también se tiene un *cluster* (*cluster* 3) con jugadores que dan mucha profundidad a sus equipos, ya sea mediante pases o acciones ofensivas que abren defensas a través de su visión del juego como Banega, Canales o Cazorla o jugadores muy verticales como Navas, Jordi Alba o Yuri que son los que dan la estilo ofensivo a sus equipos de LaLiga. También tenemos a los jugadores (en el grupo 4) que son el clásico número 9 en el fútbol.

5. En el caso del estudio que contiene los jugadores *Box to Box* de Segunda División B, podemos concluir que los jugadores más cercanos a ser *Box to Box* son los que se encuentran en el *cluster* 5. Estos jugadores se encuentran en el espacio formado entre los (0 , -6) de PC1, jugadores con mejores capacidades ofensivas, (0, 2) de PC2, jugadores con mayores duelos defensivos y (0, -2) de PC3, jugadores con mayor % de duelos defensivos ganados. En este análisis, el jugador ideal para un equipo que busca un *Box to Box* sería Álvaro Fidalgo, mientras que los cercanos a él serían jugadores del mismo perfil que él.

6. También se han realizado cuatro modelos diferentes que pretenden predecir el número de goles que todos los jugadores ofensivos de la base de datos *Delanteros19_20* han marcado en la temporada. Estos modelos son:

    - Modelo 1 que está formado por las variables: valor de mercado, minutos jugados, xG, remates y tiros a puerta. Este modelo tiene significativas las variables minutos jugados, xG y tiros a puerta.

    - Modelo 2 que está formado por las variables acciones ofensivas, desmarques, toque en el área, partidos jugados, duelos aéreos. Este modelo tiene significativas todas las variables excepto los duelos aéreos.

    - Modelo 3 que utiliza todas las variables seleccionadas en el conjunto de datos. Este modelo tiene significativas las variables minutos jugados, goles esperados, acciones ofensivas, goles de cabeza, remates, goles marcados, duelos ofensivos ganados y toques en el área penalti.

    - Modelo *StepAIC*, que utiliza un algoritmo en el que comienza utilizando todas las variables del modelo y se eliminan gradualmente las que no son significativas hasta que se encuentra un modelo en el que todas sus variables lo son

    El modelo *StepAIC* es el mejor modelo para predecir los goles porque tuvo el mayor porcentaje de aciertos respecto a los demás.

7. Se ha descubierto mediante el contraste de hipótesis *ANOVA* que los diestros son más precisos en el pase que los zurdos, mientras que no se encontraron diferencias en cuanto a la precisión en el tiro.

Bibliografía

Douglas H Fisher. “Knowledge acquisition via incremental conceptual clustering”. En: Machine learning 2.2 (1987), págs. 139-172.

Miguel Garre y col. “Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software”. En: REICIS. Revista Española de Innovación, Calidad e Ingeniería del Software 3.1 (2007), págs. 6-22.

James MacQueen y col. “Some methods for classification and analysis of multivariate observations”. En: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. 14. Oakland, CA, USA. 1967, págs. 281-297.

Ryszard S Michalski y Robert E Stepp. “Learning from observation: Conceptual clustering”. En: Machine learning. Springer, 1983, págs. 331-363.

J Ross Quinlan. “C4. 5: Programming for machine learning”. En: Morgan Kauffmann 38 (1993), pág. 48.

Theodoridis, S., Pikrakis, A., Koutroumbas, K., & Cavouras, D. (2010). *Introduction to pattern recognition: a matlab approach*. Academic Press.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-medias clustering method and elbow method for identification of the best customer profile cluster. In *IOP Conference Series: Materials Science and Engineering* (Vol. 336, No. 1, p. 012017). IOP Publishing.

Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of sports sciences*, *32*(20), 1881-1887.

Severini, T. A. (2020). *Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports*. Crc Press.

García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., & Luengo-Sánchez, S. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching*, *16*(1), 148-157.

Jackson, P. R. (1972). Football Statistics. *Mathematical Spectrum*.

Jang, J., Kim, H. J., Lim, S., Ryoo, H., Jung, T. Y., & Suh, S. H. (2020). Discovering primary indicators for evaluating defender's technical performance using multivariate statistics in football games. *International Journal of Applied Sports Sciences*, *32*(1).

Fútbol. (2021, 3 de junio). *Wikipedia, La enciclopedia libre*. Fecha de consulta: junio 6, 2021 desde https://es.wikipedia.org/w/index.php?title=F%C3%BAtbol&oldid=136062079.

Clayfield, B. (s.f). *Are Left Footed Soccer Players Better?* Yoursoccerhome. https://yoursoccerhome.com/are-left-footed-soccer-players-better/

IFFHS. (2021, 28 de junio). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 16:00, junio 30, 2021 desde https://es.wikipedia.org/w/index.php?title=IFFHS&oldid=136654626.

Anexo

El Anexo de este trabajo estará formado por el código en R desarrollado en los análisis expuestos anteriormente. A continuación se detallarán uno a uno dichos códigos.

ESTUDIO PRIMERA DIVISIÓN

```
library(foreign)
library(readxl)

datos <- read_excel("JUGADORES.xlsx")

duplicated(datos$Jugador)

datos[duplicated(datos$Jugador),] #VEMOS CUALES SON LOS DUPLICADOS

#VAMOS A ELIMINAR LOS PORTEROS PRIMERO

datos[datos$`Posición específica`== "GK",]

jugadores <- datos[datos$`Posición específica`!= "GK",] #Nueva df sin port
eros

#DIFERENCIAMOS LOS DUPLICADOS

jugadores <- as.data.frame(jugadores) #de tibble a df

row.names(jugadores) <- ifelse(duplicated(jugadores$Jugador), yes = paste0
(jugadores$Jugador, "_", jugadores$`Mi Base de datos`), no= jugadores$Juga
dor
 ) #nombres como columna de id, los duplicados tienen su equipo en el id p
ara diferenciarlos

jugadores <- jugadores[jugadores$`x_Minutos jugados` >= 2700,]
str(jugadores)

library(factoextra)

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.4

## Welcome! Want to learn more? See two factoextra-related books at https:
//goo.gl/ve3WBa

library(stats)

clusterJer <- hclust(dist(jugadores))

plot(clusterJer)
```

```r
#CHUNK PARA REALIZAR EL KMEANS CON LAS VARIABLES IMPORTANTES

names(jugadores)

#Creamos la base con las variables importantes

importantes <- jugadores[,grepl(pattern = "x_",x = names(jugadores))]

impescalados <- scale(importantes)

#ESTE CHUNK PREPARARÁ LA BASE "JUGADORES" PARA KMEANS CON TODAS LAS VARIAB
LES QUITANDO LAS CATEGÓRICAS

jugadores <- jugadores[,sapply(X = jugadores, is.numeric)]

#ESTE CHUNK ESTARÁ DESTINADO A CALCULAR EL NÚMERO ÓPTIMO DE CLUSTERS A REA
LIZAR MEDIANTE EL MÉTODO WSS

library("factoextra")
library("NbClust")

jugescalados <- scale(jugadores) #Escalamos los datos para optimizar los c
álculos



jugescalados <- jugescalados[ , colSums(is.na(jugescalados)) == 0] #Quitan
do columnas con algún Na, quita muchas variables
impescalados <- impescalados[ , colSums(is.na(impescalados)) == 0] #Quitan
do columnas con algún Na, quita muchas variables


#RESULTADOS
##Con todas las variables


fviz_nbclust(jugescalados,kmeans, method = "wss" ) +
    geom_vline(xintercept = 5, linetype = 2)+
  labs(subtitle = "Elbow method")

##Con las variables que consideramos importantes
fviz_nbclust(impescalados,kmeans, method = "wss" ) +
    geom_vline(xintercept = 5, linetype = 2)+
  labs(subtitle = "Elbow method")

#ESTE CHUNK ESTÁ DESARROLLADO PARA REALIZAR EL ALGORITMO KMEANS UNA VEZ SA
BEMOS QUE K=5 POR EL MÉTODO ELBOW

##todas las variables
juguescalados <- as.data.frame(jugescalados)
cluster1 <- kmeans(jugescalados,centers = 5)
fviz_cluster(cluster1,data = juguescalados)

##variables importantes
impescalados <- as.data.frame(impescalados)
```

```
cluster2 <- kmeans(impescalados,centers = 5)
fviz_cluster(cluster2,data = impescalados)

# PARA REALIZAR UNA ACP HACE FALTA ESCALAR LOS DATOS, TRABAJAREMOS SOBRE L
A BASE "impescalados"

acp <- prcomp(impescalados)

biplot(acp, scale = 0, cex = 0.6) #BIPLOT SOBRE LA BASE

library("FactoMineR")

fviz_screeplot(acp, addlabels = TRUE, ylim = c(0,100), cex = 0.6)

library("rgl")
library("plotly")

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.0.4

## corrplot 0.84 loaded

impescalados <- scale(impescalados)

var <- get_pca_var(acp)
corrplot(var$cos2, is.corr=FALSE)

library(plotly)

a <- data.frame(acp$x)
a$goleador <- ifelse(jugadores$x_Goles>20, 'goleador', 'paquete')

b <- kmeans(jugescalados, centers=5)
fig <- plot_ly(a, x=~PC1, y=~PC2, z=~PC3, text=~paste0(rownames(a), "\nGol
es: ", jugadores$x_Goles), color=~factor(b$cluster), colors=c('#FF0000', '
#0000FF', '#00FF00', 'yellow', '#FF9900'))
fig
```

ESTUDIO *BOX TO BOX* 2ªB

```
library("readxl")
library("dplyr")

datos <- read_excel("SEGUNDAB.xlsx")

datos <- as.data.frame(datos)

for (i in 1:length(colnames(datos))){
  colnames(datos)[i] <- paste0(i, ' ', colnames(datos)[i])
}

medios_base <- datos[grepl(pattern = "DM|CM", datos$`3 Posición específica
`),] #Filtro para elegir a los medios

medios <- medios_base[!grepl(pattern = "AMF|CF|CB", medios_base$`3 Posició
n específica`),]

row.names(medios)  <- ifelse(duplicated(medios$`1 Jugador`), yes = paste0(
medios$`1 Jugador`, "_", medios$`2 Mi Base de datos`), no= medios$`1 Jugad
or`)

medios <- medios[medios$`8 med_Minutos jugados` >= 1710,]


medios <- as.data.frame(medios) #De tibble a df

#NUEVO FILTRADO

filtro_variables <- c(2,35,77,82,86,
                      20,21,22,23)

medios <- medios[,filtro_variables]

#PRIMERO HAY QUE DECIDIR CUÁNTOS GRUPOS SE CREARÁN MEDIANTE EL ELBOW METHO
D


library("factoextra")

library("NbClust")

importantes <- medios[,grepl(pattern = "med_",x = names(medios))] #Filtram
os la base para estudiar las columnas importantes en un medio centro Box t
o Box

impescalados <- scale(importantes) #Escalamos los datos para realizar el a
lgoritmo KMEANS
#impescalados <- scale(medios) #Escalamos los datos para realizar el algor
itmo KMEANS
impescalados <- impescalados[ , colSums(is.na(impescalados)) == 0] #Quitan
do columnas con algún Na, quita muchas variables
```

```
#RESULTADOS

#Con las variables que consideramos importantes

fviz_nbclust(impescalados,kmeans, method = "wss" ) +
    geom_vline(xintercept = 5, linetype = 2)+
  labs(subtitle = "Elbow method")

#ESTE CHUNK ESTÁ DESARROLLADO PARA REALIZAR EL ALGORITMO KMEANS UNA VEZ SA
BEMOS QUE K=5 POR EL MÉTODO ELBOW


impescalados <- as.data.frame(impescalados)
cluster2 <- kmeans(impescalados,centers = 5)
fviz_cluster(cluster2,data = impescalados)

# PARA REALIZAR UNA ACP HACE FALTA ESCALAR LOS DATOS, TRABAJAREMOS SOBRE L
A BASE "impescalados"

acp <- prcomp(impescalados)

biplot(acp, scale = 0, cex = 0.6) #BIPLOT SOBRE LA BASE

fviz_pca_var(acp,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )

library("FactoMineR")

fviz_screeplot(acp, addlabels = TRUE, cex = 0.6)

#COMPONENTES



library("rgl")
library("plotly")

library(corrplot)

impescalados <- scale(impescalados)

var <- get_pca_var(acp)
corrplot(var$cos2, is.corr=FALSE)

a <- data.frame(acp$x)
b <- kmeans(impescalados, centers=5)


library(scales)
library(kableExtra)
```

```
a$PC5 <- rescale(a$PC5, to = c(0,100))

#PREPARAR LA GRÁFICA PARA HACER UN RANKING RESPECTO A SU CAPACIDAD AÉREA
a$PC5

boxtobox <- plot_ly(a, x=~PC1, y=~PC2, z=~PC3, text=~paste0(rownames(a),"\
n Valoración aérea sobre 100: ",a$PC5 ), color=~factor(b$cluster), colors=
c('#FF0000', '#0000FF', '#00FF00', 'yellow', '#FF9900'), hoverinfo = "text
")
boxtobox

#PC1 EXPLICA VARIABLES OFENSIVOS NEGATIVO
#PC2 EXPLICA DUELOS DEFENSIVOS POSITIVO
#PC3 EXPLICA DUELOS DEF GANADOS NEGATIVO
```

PREDICCIÓN DE GOLES PARA LOS DELANTEROS

```
library("MASS")

library(corrplot)

datos <- read.csv2("Delanteros_19-20.csv")
delanteros <- datos[grepl(pattern = "CF|AMF|RWF|LWF", datos$Posición.espec
ífica),]

duplicados <- delanteros[duplicated(delanteros$x_Jugador),]
rownames(delanteros) <- ifelse(duplicated(delanteros$x_Jugador), yes = pas
te0(delanteros$x_Jugador, "_", delanteros$x_Equipo), no = delanteros$x_Jug
ador)

delanteros$x_Minutos.jugados <- as.numeric(delanteros$x_Minutos.jugados)
delanteros <- delanteros[delanteros$x_Minutos.jugados > mean(delanteros$x_
Minutos.jugados),] # Delanteros que superen la media de minutos jugados

# NOMBRES COMO ROWNAMES Y ELEGIMOS LAS VARIABLES BUENAS PARA DELANTEROS

delanteros <- delanteros[,grepl(pattern = "x_", colnames(delanteros))]
str(delanteros)

set.seed(4)
# PARTIMOS LA BASE EN DOS

## PRIMERO CREAMOS LA BASE QUE UTILIZAREMOS PARA ENTRENAR EL MODELO

train <- delanteros[sample(nrow(delanteros), 0.8*nrow(delanteros)),] # Bas
e que servirá para entrenar el modelo, es el 80% del total
train <- as.data.frame(train)

## SEGUNDO CREAMOS LA BASE QUE UTILIZAREMOS PARA PREDECIR EL NÚMERO DE GOL
ES

test <- delanteros[setdiff(rownames(delanteros), rownames(train)),]

delanteros$x_Pie <- as.factor(delanteros$x_Pie)
```

```
full.model <- lm(x_Goles~.-x_Jugador-x_Equipo, data = train)
summary(full.model)

test_predFULL <- test[,names(test)!="x_Goles"] # Eliminamos la columna que
queremos predecir
prediccionFULL <- predict(modback, test_predFULL)
prediccionFULL <- round(prediccionFULL ,digits = 0)
prediccionFULL <- ifelse(as.integer(prediccionFULL) < 0, yes = as.integer(
prediccionFULL) == 0, no = as.integer(prediccionFULL))
# View(prediccionFULL)

test_predFULL <- cbind(test_predFULL, prediccionFULL)

# View(test_predFULL$prediccionFULL)

colnames(test_predFULL)[22] <- "Predicción de goles"

calidad_prediccionFULL <- ifelse(test$x_Goles == test_predFULL$`Predicción
de goles` , yes = "BUENA", no = "MALA" )
calidad_prediccionFULL <- as.data.frame(calidad_prediccionFULL)

xFULL <- calidad_prediccionFULL[calidad_prediccionFULL == "BUENA",]
xFULL <- as.data.frame(xFULL)
xnumFULL <- nrow(xFULL)
xnumFULL

ynumFULL <- nrow(calidad_prediccionFULL)
ynumFULL

bien_clasificadosFULL <- xnumFULL/ynumFULL*100
bien_clasificadosFULL

t.test(test$x_Goles, test_predFULL$`Predicción de goles`)

"cor(delanteros)"

## [1] "cor(delanteros)"

test_pred1 <- test[,names(test)!="x_Goles"] # Eliminamos la columna que qu
eremos predecir

test_pred1 <- as.data.frame(test_pred1)

input1 <- train[,c("x_Valor.de.mercado..Transfermarkt.","x_Minutos.jugados
","x_xG","x_Remates.en.los.90","x_Tiros.a.la.portería..","x_Goles")]

modelo1 <- lm(x_Goles~., data=input1)
summary(modelo1)

prediccion1 <- predict(modelo1, test_pred1)

test_pred1 <- cbind(test_pred1, ifelse(as.integer(prediccion1) < 0, yes =
as.integer(prediccion1) == 0, no = as.integer(prediccion1)))
test_pred1 <- as.data.frame(test_pred1)
colnames(test_pred1)[22] <- "Predicción de goles"
```

```
calidad_prediccion1 <- ifelse(test$x_Goles == test_pred1$`Predicción de goles` , yes = "BUENA", no = "MALA" )
calidad_prediccion1 <- as.data.frame(calidad_prediccion1)

x1 <- calidad_prediccion1[calidad_prediccion1 == "BUENA",]
x1 <- as.data.frame(x1)
xnum1 <- nrow(x1)
xnum1

ynum1 <- nrow(calidad_prediccion1)
ynum1

bien_clasificados1 <- xnum1/ynum1*100
bien_clasificados1

test_pred2 <- test[,names(test)!="x_Goles"] # Eliminamos la columna que queremos predecir
test_pred2 <- as.data.frame(test_pred2)

input2 <- train[,c("x_Acciones.ofensivas.logradas","x_Desmarques.90","x_Toques.en.el.área.de.penalti.90","x_Partidos.jugados","x_Duelos.aéreos.en.los.90",
                   "x_Goles")]

modelo2 <- lm(x_Goles~., data=input2)
summary(modelo2)

prediccion2 <- predict(modelo2, test_pred2)

test_pred2 <- cbind(test_pred2, ifelse(as.integer(prediccion2) < 0, yes = as.integer(prediccion2) == 0, no = as.integer(prediccion2)))

test_pred2 <- as.data.frame(test_pred2)

colnames(test_pred2)[22] <- "Predicción de goles"

calidad_prediccion2 <- ifelse(test$x_Goles == test_pred2$`Predicción de goles` , yes = "BUENA", no = "MALA" )
calidad_prediccion2 <- as.data.frame(calidad_prediccion2)

x2 <- calidad_prediccion2[calidad_prediccion2 == "BUENA",]
x2 <- as.data.frame(x2)

xnum2 <- nrow(x2)
xnum2

ynum2 <- nrow(calidad_prediccion2)
ynum2

bien_clasificados2 <- xnum2/ynum2*100
bien_clasificados2

test_pred3 <- test[,names(test)!="x_Goles"] # Eliminamos la columna que queremos predecir
test_pred3 <- as.data.frame(test_pred3)
```

```
input3 <- train[,colnames(train)!=c("x_Jugador","x_Equipo")]

modelo3 <- lm(x_Goles~., data=input3)
summary(modelo3)

prediccion3 <- predict(modelo3, test_pred3)
test_pred3 <- cbind(test_pred3, ifelse(as.integer(prediccion3) < 0, yes =
as.integer(prediccion3) == 0, no = as.integer(prediccion3)))
test_pred3 <- as.data.frame(test_pred3)
colnames(test_pred3)[22] <- "Predicción de goles"

calidad_prediccion3 <- ifelse(test$x_Goles == test_pred3$`Predicción de go
les` , yes = "BUENA", no = "MALA" )
calidad_prediccion3 <- as.data.frame(calidad_prediccion3)

x3 <- calidad_prediccion3[calidad_prediccion3 == "BUENA",]
x3 <- as.data.frame(x3)
xnum3 <- nrow(x3)
xnum3

ynum3 <- nrow(calidad_prediccion3)
ynum3

bien_clasificados3 <- xnum3/ynum3*100
bien_clasificados3
```

CONTRASTE *ANOVA*

```
library(foreign)
library(readxl)

datos <- read_excel("JUGADORES.xlsx")

#VAMOS A ELIMINAR LOS PORTEROS PRIMERO

datos[datos$`Posición específica`== "GK",]

jugadores <- datos[datos$`Posición específica`!= "GK",] #Nueva df sin port
eros

#DIFERENCIAMOS LOS DUPLICADOS

jugadores <- as.data.frame(jugadores) #de tibble a df

row.names(jugadores) <- ifelse(duplicated(jugadores$Jugador), yes = paste0
(jugadores$Jugador, "_", jugadores$`Mi Base de datos`), no= jugadores$Juga
dor
 ) #nombres como columna de id, los duplicados tienen su equipo en el id p
ara diferenciarlos

jugadores <- jugadores[jugadores$`x_Minutos jugados` >= 2700,]

hist(jugadores$`x_Precisión pases %`)
```

```
hist(jugadores$`x_Tiros a la portería %`)

jugadores <- jugadores[!is.na(jugadores$`x_Precisión pases %`),]
jugadores <- jugadores[!is.na(jugadores$`x_Tiros a la portería %`),]
jugadores <- jugadores[!is.na(jugadores$Pie),]


jugadores$Pie <- as.factor(jugadores$Pie)

chisq.test(jugadores$Pie, jugadores$`x_Precisión pases %`)

chisq.test(jugadores$Pie, jugadores$`x_Tiros a la portería %`)

library(data.table)

tapply(jugadores$`x_Precisión pases %`, jugadores$Pie, mean)

summary(aov(jugadores$`x_Precisión pases %`~jugadores$Pie))

tapply(jugadores$`x_Tiros a la portería %`, jugadores$Pie, mean)

summary(aov(jugadores$`x_Tiros a la portería %`~jugadores$Pie))
```