Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

Uso de técnicas de clustering para encontrar perfiles de jugadores en una competición de fútbol profesional

Autor: Mario Garrido Tapias

Tutor: José Belarmino Pulido Junquera

Curso 2021-22

Resumen

El Big Data se ha ido haciendo un hueco en el mundo del fútbol, convirtiendose en un fijo en el personal de la mayoria de equipos, siendo fundamental en el análisis del rendimiento de los jugadores tanto en los partidos como en el mercado de traspasos.

En el presente trabajo se estudiará una de las cinco grandes ligas mundiales, esta no es otra que la liga nacional española, más concretamente nos centraremos en los jugadores que jugaron en dicha competición durante los años 2017 y 2020, comprendiendo entre ellos 3 temporadas. Se extraerán y analizarán las estadísticas más importantes, teniendo en cuenta las diferentes acciones importantes en el juego.

Sobre estos jugadores se buscarán varios *clusters*, analizando los perfiles resultantes y estudiando la posibilidad de creación de una herramienta que aporte ayuda a la hora de elegir nuevas incorporaciones para un equipo.

Abstract

Big Data has been making its way into the world of football, becoming a permanent fixture in the staff of most teams, being fundamental in the analysis of the performance of players both in matches and in the transfer market.

In this paper we will study one of the five major world leagues, this is none other than the Spanish national league, more specifically we will focus on the players who played in that competition during the years 2017 and 2020, comprising between them 3 seasons. The most important statistics will be extracted and analysed, taking into account the different important actions in the game.

Several clusterings will be carried out on these players, analysing the resulting profiles and studying the possibility of creating a tool that will help when choosing new additions to a team.

Agradecimientos

Antes que nada reconocer el esfuerzo de todos los profesores implicados en mi formación durante estos cinco años. Y en particular, al tutor de este trabajo, Belarmino, por la función de apoyo desempeñada y por darme la oportunidad de trabajar sobre un campo que ha originado en mi un interés alto a través del desempeño de este proyecto.

Una mención especial a mis compañeros, por hacer más fácil y amena esta formación.

Por último, pero no menos importante, a ese círculo más cerrado y familiar que me ha animado en cada momento y ha confiado ciegamente en mí.

Índice general

Resumen **I**

Abstract **II**

Agradecimientos **III**

Índice de tablas **VI**

Índice de figuras **VIII**

1. Introducción **1**
1.1. Funcionamiento de LaLiga 1
1.2. Trabajos relacionados 3
1.3. Objetivos 3

2. Marco teórico **5**
2.1. Análisis de componentes principales (PCA) 5
2.2. Técnicas de aprendizaje 6
2.2.1. Clustering 6

3. Búsqueda de datos **11**
3.1. Bases de datos disponibles 11

4. Exploración de datos **15**
4.1. Descripción del conjunto de datos 15
4.2. Preparación de datos 16
4.2.1. Búsqueda de valores ausentes 16
4.2.2. Creación del conjunto de datos a estudiar 16
4.2.3. Tratamiento de jugadores que juegan en 2 equipos en una misma temporada 17
4.2.4. Creación de variables 19
4.3. Análisis univariante 20
4.3.1. Estadísticas físicas del jugador 20
4.3.2. Estadísticas relacionadas con los pases 21
4.3.3. Estadísticas relacionadas con el tiro y el gol 28

4.3.4. Estadísticas relacionadas con el ámbito ofensivo 34
4.3.5. Estadísticas centradas en el ámbito defensivo 42
4.3.6. Estadísticas centradas en porteros 49
4.4. Análisis bivariante 53
4.4.1. Relación entre variables acumuladas y porcentajes de éxito 53
4.4.2. Promedio de estadísticas 59
4.4.3. Análisis de correlaciones altas 60
4.4.4. Relaciones de pares interesantes 61
4.4.5. Análisis de correlaciones 65
4.5. Análisis multivariante 67
4.5.1. Estandarización de los datos 67
4.5.2. Análisis de componentes principales 67
4.6. Creación del conjunto de entrenamiento y test 69

5. Uso de técnicas de aprendizaje **73**
5.1. Búsqueda del k óptimo 73
5.2. Resultados 74
5.2.1. K-medias 74
5.2.2. Jerárquico 80
5.2.3. DBSCAN 82
5.3. Clustering en nuevo conjunto 83
5.3.1. Búsqueda del k óptimo 84
5.3.2. K-medias 85
5.4. Validación 88
5.4.1. Evaluación interna 88
5.4.2. Evaluación externa 89

6. Conclusiones **91**
6.1. Conclusiones del trabajo 91

7. Trabajos futuros **93**
7.1. Implementación de la aplicación 93
7.1.1. Requisitos funcionales 93
7.1.2. Casos de uso 94
7.2. Mejoras de la aplicación 94

Bibliografía **95**

A. Código utilizado **99**
A.1. Estructura 99

Índice de tablas

4.1. Jugadores que jugaron en varios clubs en una misma temporada 17
4.2. Comparación de Nolito en la temporada 19-20 18
4.3. Comparación de Youssef En Nesyrl en la temporada 19-20 18
4.4. Jugadores que jugaron en varios clubs en una misma temporada 19
4.5. Estadísticos variables físicas 21
4.6. Estadísticos variable xA 22
4.7. Estadísticos variable passes_completed 22
4.8. Estadísticos variable passes 23
4.9. Estadísticos variable passes_pct 23
4.10. Estadísticos variable passes_total_distance 24
4.11. Estadísticos variable assisted_shots 25
4.12. Estadísticos variable passes_switches 26
4.13. Estadísticos variable pass_targets 26
4.14. Estadísticos variable passes_received 26
4.15. Estadísticos variable passes_received_pct 27
4.16. Estadísticos variable xg 28
4.17. Estadísticos variable npxg 28
4.18. Estadísticos variable shots_total 29
4.19. Estadísticos variable shots_on_target 30
4.20. Estadísticos variable shots_on_target_pct 30
4.21. Estadísticos variable goals_per_shot 31
4.22. Estadísticos variable pens_made 31
4.23. Estadísticos variable pens_att 32
4.24. Estadísticos variable pens_made_pct 32
4.25. Estadísticos variable goals_assists_per90 33
4.26. Estadísticos variable dribbles 34
4.27. Estadísticos variable dribbles_completed 35
4.28. Estadísticos variable dribbles_completed_pct 35
4.29. Estadísticos variable sca_dribbles 36
4.30. Estadísticos variable gca_dribbles 37
4.31. Estadísticos variable sca_passes_live 37
4.32. Estadísticos variable gca_passes_live 38
4.33. Estadísticos variable sca_passes_dead 39
4.34. Estadísticos variable gca_passes_dead 40

4.35. Estadísticos variable sca_fouled 41
4.36. Estadísticos variable gca_fouled 41
4.37. Estadísticos variable passes_intercepted 42
4.38. Estadísticos variable ball_recoveries 43
4.39. Estadísticos variable pressure_regains 44
4.40. Estadísticos variable fouls 44
4.41. Estadísticos variable tackles_won 45
4.42. Estadísticos variable tackles 46
4.43. Estadísticos variable tackles_pct 46
4.44. Estadísticos variable aerials_won 47
4.45. Estadísticos variable aerials_contested 48
4.46. Estadísticos variable aerials_won_pct 48
4.47. Estadísticos variable goals_against_per90_gk 49
4.48. Estadísticos variable pens_saved 50
4.49. Estadísticos variable pens_played 50
4.50. Estadísticos variable pens_saved_pct 50
4.51. Estadísticos variable shots_on_target_against 51
4.52. Estadísticos variable saves 51
4.53. Estadísticos variable save_pct 52
4.54. Número de jugadores por posición en cada temporada 70
4.55. Proporción de jugadores diferenciados por posición en cada conjunto 70

5.1. Tabla resumen para $k = 6$ 75
5.2. Tabla resumen para $k = 7$ 78
5.3. Coeficientes de silhouette por *cluster* en los últimos 3 modelos . . 88

7.1. Casos del uso de la aplicación 94

Índice de figuras

2.1. Análisis de componentes principales explicado [13] 6
2.2. Ejemplo de un dendograma [14] 8

4.1. Distribuciones de los atributos físicos de los jugadores de nuestro conjunto de datos 21
4.2. Asistencias esperadas en las temporadas 22
4.3. Pases con éxito en las temporadas 23
4.4. Pases totales en las temporadas 23
4.5. Porcentaje de éxito en el pase en las distintas temporadas 24
4.6. Distancia recorrida por sus pases en las temporadas 24
4.7. Distribución pases que conllevan un tiro en las temporadas . . . 25
4.8. Pases que conllevan un tiro en las temporadas 25
4.9. Pases que cambian la orientación del juego en las temporadas . . 26
4.10. Pases con objetivo cada jugador en las 3 temporadas 27
4.11. Pases totales que recibe el jugador en las 3 temporadas 27
4.12. Porcentaje de pases recibidos con éxito por temporadas 27
4.13. Goles esperados en las temporadas 28
4.14. Goles esperados no de penalti en las temporadas 29
4.15. Tiros totales intentados en las 3 temporadas 29
4.16. Tiros a puerta en las 3 temporadas 29
4.17. Porcentaje de éxito en tiros a puerta en las temporadas 30
4.18. Ratio goles por tiro en las temporadas 31
4.19. Penaltis convertidos en las 3 temporadas 32
4.20. Penaltis intentados en las 3 temporadas 32
4.21. Porcentaje de éxito en el lanzamiento de penaltis en las temporadas 33
4.22. Goles + Asistencias en 90 mins en las temporadas 34
4.23. Regates intentados en las 3 temporadas 34
4.24. Regates conseguidos en las 3 temporadas 34
4.25. Porcentaje de regates exitosos por temporada 35
4.26. Acciones de creación de tiro tras regate exitoso por temporada . 36
4.27. Acciones de creación de gol tras regate exitoso por temporada . . 37
4.28. Acciones de creación de tiro con el balón en juego 38
4.29. Acciones de creación de gol con el balón en juego 39
4.30. Acciones de creación de tiro a balon parado en las temporadas . 39

4.31. Acciones de creación de gol en las temporadas 40
4.32. Acciones crean un tiro después de una falta realizada en las temporadas 41
4.33. Acciones crean un gol después de una falta realizada en las temporadas 42
4.34. Pases interceptados en las temporadas 42
4.35. Balones recuperados en las temporadas 43
4.36. Ocasiones en las que el jugador recupera la posesión mediante una presión por temporadas 44
4.37. Faltas en las temporadas 45
4.38. Entradas exitosas en las 3 temporadas 46
4.39. Entradas realizadas en las 3 temporadas 46
4.40. Porcentaje de éxito en entradas defensivas por temporada 47
4.41. Balones aéreos ganados en las 3 temporadas 47
4.42. Balones aéreos disputados en las 3 temporadas 47
4.43. Porcentaje de éxito en balones aéreos disputados por temporadas 48
4.44. Goles en contra por 90' en las temporadas 49
4.45. Penaltis atajados en las 3 temporadas 50
4.46. Penaltis disputados en las 3 temporadas 50
4.47. Porcentaje de éxito parando penaltis en las temporadas 51
4.48. Tiros a puerta recibidos en las 3 temporadas 52
4.49. Paradas realizadas en las 3 temporadas 52
4.50. Porcentaje de paradas en las temporadas 52
4.51. Correlaciones entre las variables que indican los pases realizados 54
4.52. Correlaciones entre las variables que indican los pases dirigidos hacia el jugador 54
4.53. Correlaciones entre las variables que indican los tiros a puerta realizados 55
4.54. Correlaciones entre las variables que indican los balones aéreos disputados 55
4.55. Correlaciones entre las variables que indican los regates intentados 56
4.56. Correlaciones entre las variables que indican las ocasiones que el rival te tira a puerta 57
4.57. Correlaciones entre las variables que los lanzamientos desde los 11 metros para un jugador 57
4.58. Correlaciones entre las variables que indican las entradas realizadas para un jugador 58
4.59. Correlaciones entre las variables que indican las ocasiones en las que un portero se enfrenta a un penalti 58
4.60. Correlaciones entre las variables promediadas que indican información sobre los pases realizados 59
4.61. Correlaciones entre las variables promediadas que indican información sobre los balones aéros disputados 60
4.62. Correlaciones que superan el valor de 0.9 entre distintas variables 61
4.63. Relación entre presiones y pases interceptados 62
4.64. Relación exito en el *tackling* y faltas realizadas 63

4.65. Relación entre las acciones de tiro desde balón parado y los cambios de orientación del juego 64
4.66. Relación entre los regates intentados y la cantidad de veces que intentan pasarte el balón 64
4.67. Matriz de correlaciones de *Pearson* 65
4.68. Matriz de correlaciones de *Spearman* 66
4.69. Contribución a la varianza explicada por cada componente principal 67
4.70. Representación 2D de las dos componentes principales 68
4.71. Contribución de cada variables respecto a las dos primeras componentes principales 69
4.72. Distribución por posiciones en cada temporada para el conjunto de entrenamiento 71
4.73. Distribución por posiciones en cada temporada para el conjunto de test 71

5.1. Método del codo hasta 20 *clusters* 73
5.2. 6-medias 74
5.3. 6 *clusters individualizados* 76
5.4. 7-medias 77
5.5. 7 *clusters individualizados* 79
5.6. Jerárquico con 5 grupos 80
5.7. Jerárquico *clusters individualizados* 81
5.8. División del cluster morado en 2 82
5.9. Distribución de distancias medias para 5 vecinos más próximos . 82
5.10. DBSCAN con $\epsilon = 2.8$ y minPts $= 3$ 83
5.11. Método del codo hasta 20 *clusters* 84
5.12. Resultados clustering 5-medias 85
5.13. Resultados clustering 6-medias 86
5.14. Resultados clustering 7-medias 87
5.15. Coeficiente de silhouete por individuo de cada cluster 89
5.16. Conjunto de test en el modelo de la figura 5.14 90

Capítulo 1

Introducción

1.1. Funcionamiento de LaLiga

En la actualidad, el *Big Data* ha causado una gran renovación dentro del sector deportivo profesional, dado que permite a los equipos fundamentar su toma de decisiones, generando una superioridad deportiva, respecto a su rival, en muchas secciones como la estrategia, la prevención de lesiones o la captación de talento. Tanto es así que incluso los jugadores acuden al análisis de datos, así como para controlar su rendimiento en el campo o para elegir el equipo al que ser traspasado, como es el caso de Héctor Bellerín, que eligió al Betis de entre sus candidatos, dada la influencia que el *Big Data* predecía que iba a tener en el campo, por estilo de juego del equipo y del entrenador [1].

Cada vez son más las empresas que se dedican a la obtención de datos en el deporte, y en nuestro caso, en el fútbol dos de las más utilizadas en España son *StatsBomb* y *Microsoft*, con *Beyond Stats*, que genera estadísticas en tiempo real en las retransmisiones de los partidos de LaLiga. Pero, a pesar de que estas compañias están bastante cerradas en el mercado, ya que su acceso se restringe a los equipos de fútbol, medios de comunicación y casas de apuestas, existen bastantes conjuntos de datos gratuitos que nos dan la oportunidad de generar valor con ellos. Es por esto que surge la idea de poder sacar potencial a los datos desde el punto de vista de los traspasos, pudiendo dar lugar a una herramienta del tipo *general manager*, que con una serie de filtros ayude a los equipos a seleccionar mejor al jugador que buscan.

Es un torneo de fútbol organizado y regulado por la Liga Nacional de Fútbol Profesional [2], cuyos participantes son los propios clubes miembros. Es una competición disputada anualmente, teniendo comienzo alrededor de finales del mes de agosto o principios de septiembre, y llegando a su fin entre los meses de mayo y junio del año posterior, en función de los calendarios de las competiciones internacionales de selecciones.

La Primera División consta de un grupo único integrado por veinte equipos, pertenecientes a clubes de fútbol o sociedades anónimas deportivas (S.A.D.).

Sigue un sistema de liga, donde todos los equipos se enfrentan contra todos en dos ocasiones, una en campo propio y otra en campo del rival, completando así un total de 38 jornadas. El orden de dichos encuentros se determina por sorteo antes de comenzar con la competición.

La clasificación final se obtiene mediante un recuento de puntos totales, conseguidos mediante el siguiente sistema de puntuación: tres puntos por cada partido ganado, un punto por partido empatado y ningún punto por los partidos perdidos. La Liga no es un sistema cerrado de equipos, ya que cada año hay 2 o 3 equipos (según el sistema instaurado en cada temporada) que desciende a Segunda División, y estos son sustituidos por los mejores durante la temporada de la segunda liga española.

Mediante estos métodos se ha permitido participar en la Primera División a sesenta y tres equipos diferentes, siendo solamente tres los que han permanecido siempre en la primera categoría desde su primera edición: Athletic Club, F. C. Barcelona y Real Madrid C. F.. [3]

Cada equipo puede contar con 25 jugadores, como máximo, en su plantilla. A los cuales puede sacar su máximo rendimiento analizando distintos marcadores que les aportan las nuevas metodologías como el uso de GPS, drones o modelos predictivos. Entre las métricas más actuales y utilizadas encontramos:

- Goles esperados (xG) Se basa en la cuantificación de una ocasión de gol creada, entre 0 y 1, dependiendo esta de distintos factores como el número de defensas en la trayectoria del disparo, la distancia entre el jugador que tira y la portería rival o con que parte del cuerpo chuta a puerta. El resultado es interpretable como un porcentaje de conseguir dicho gol.

- Asistencias esperadas (xA) Similar a la anterior, pero, esta vez midiendo la probabilidad de que el pase se transforme en asistencia de gol. Existe un pequeño matiz, que la hace distinta a la anterior métrica, y es que el valor asignado no es condicionado por si su compañero consigue rematar a portería. Al igual que la anterior, las variables que predicen este valor, depende del modelo que aplique la compañia.

  Ambas métricas son bastante interesantes a la hora de medir el rendimiento de un jugador, ya que si para un jugador su número de goles marcados o su número de asistencias aportadas es mayor que su xG o xA respectivamente, estaremos antes un jugador que o bien esta rindiendo por encima de lo que se espera o que simplemente está teniendo suerte.

- Pases permitidos por acción defensiva (PPDA) Está métrica permite evaluar la efectividad de la presión sometida al rival. Consiste en calcular la cantidad de pases que el rival da en los primeros 3/5 del campo, tomando como referencia su portería, y a este número dividirle por la cantidad de acciones defensivas realizadas exitosamente en esa misma zona del campo. Por tanto, cuanto mayor sea el resultado, peor le habrá resultado a nuestro equipo realizar una presión alta.

Alguna métricas más se puede encontrar en [4].

1.2. Trabajos relacionados

Tras la búsqueda previa realizada para la selección de un conjunto de datos interesante se encontraron distintos informes cuyo foco era similar al nuestro. En primer lugar, Jason Zivkovic, presentó un análisis en *Kaggle* [5] donde, a través de un conjunto de datos basado en las valoraciones del videojuego FIFA 19, intentó medir el potencial de los distintos jugadores, indagando sobre los jugadores con mayores potenciales, tanto jovenes como veteranos, y cuyo objetivo final era, encontrar la posibilidad de descubrir "gangas", es decir, jugadores con un potencial similar al de los jugadores más buscados del mercado pero, con un valor de mercado bastante inferior. Después, Pablo Reyes en su blog [6], realizó un análisis sobre un conjunto de estadísticas futbolísticas, a los cuales aplicó agrupamiento con ayuda de Python. Previo a esto, realizó un estudio de correlaciones junto a una reducción de la dimensionalidad para mejorar la interpretabilidad de su trabajo. Consiguió separar a los jugadores en 5 grupos, parámetro elegido objetivamente con el método del codo.

Respecto al ámbito universitario se han revisado varios trabajos finales de algunos antiguos alumnos de la UVa. Entre ellos están el de Ramiro Gómez Nuño [7], consistente en la elaboración minuciosa de un paquete en R para el análisis de futbolistas y de distintos eventos de un partido fútbol. También se exploró el trabajo de Jorge San José Lorza [8], el cual utilizaba unos datos de eventos ocurridos durante los partidos de la Copa del Mundo de la FIFA del 2018, y con ellos creo un modelo de goles esperados utilizando el algoritmo *Gradient Boosting*. En ambos, su foco estaba en el análisis de datos del mundo del fútbol, siendo más escalable el de Ramiro, ya que se puede aplicar a distintas competiciones, y el de Jorge se enfocaba en una competición en concreto. Nuestro trabajo busca aportar una visión mejorada a la hora de buscar traspasos para un equipo, para ello exploraremos distintos conjuntos de datos disponibles gratuitamente con el objetivo de encontrar los que mejor se ajuste a este fin principal.

1.3. Objetivos

El principal cometido de este trabajo será utilizar técnicas de aprendizaje no supervisado para encontrar diferentes perfiles de jugadores, que puedan ayudar a hipotéticos directivos de distintos clubes de fútbol a la hora de enfocar un traspaso para un tipo de jugador determinado.

Dentro de este enfoque, se deberá analizar los conjuntos de datos disponibles en la web de manera gratuita, en busca de los que mejor se ajusten a los requisitos del trabajo.

Sobre el conjunto de datos seleccionado [9], se realizará un análisis exploratorio, de alto nivel, de sus distintas variables. Donde se buscará aplicar distintas técnicas de análisis, que nos aporten explicaciones y razones para descartar o escoger variables, para así determinar el conjunto de estadísticas, más reducido, que nos permitan realizar una mejor clasificación.

Por último, en base a los resultados obtenidos, se estudiará la posibilidad de generar una herramienta, destinada a los "*general manager*"de los distintos equipos.

Capítulo 2

Marco teórico

2.1. Análisis de componentes principales (PCA)

Existen multitud de técnicas cuyo objetivo es disminuir el número de variables aleatorias observadas. Estas surgieron a partir de buscar una solución para la **maldición de la dimensionalidad**. Entre ellas se encuentran el análisis factorial [10], el análisis de componentes independientes [11] o el análisis de componentes principales [12], ... Siendo esta última la más conocida y la que usaremos nosotros como primera aproximación.

Es un método que se comprende dentro de los procedimientos lineales, que consiste en la tranformación ortogonal de las n dimensiones de partida de nuestro conjunto de datos a un nuevo conjunto de menor dimensión conocido como **componentes principales**. Esta transformación tiene como objetivo reconocer el hiperplano que es más afín a los datos, denominado primera componente principal. Éste será el que mayor suma varianza explique, de los datos destinados al entrenamiento.

El procedimiento consiste en ir encontrando ejes ortogonales a la anterior componente principal, siendo estos los que mejor guarden la variabilidad de los datos, de manera que la siguiente componente debe recoger la máxima variabilidad no recogida por las anteriores calculadas. Cada una de ellos corresponde con una combinación lineal de los atributos originales, siendo independientes entre sí.

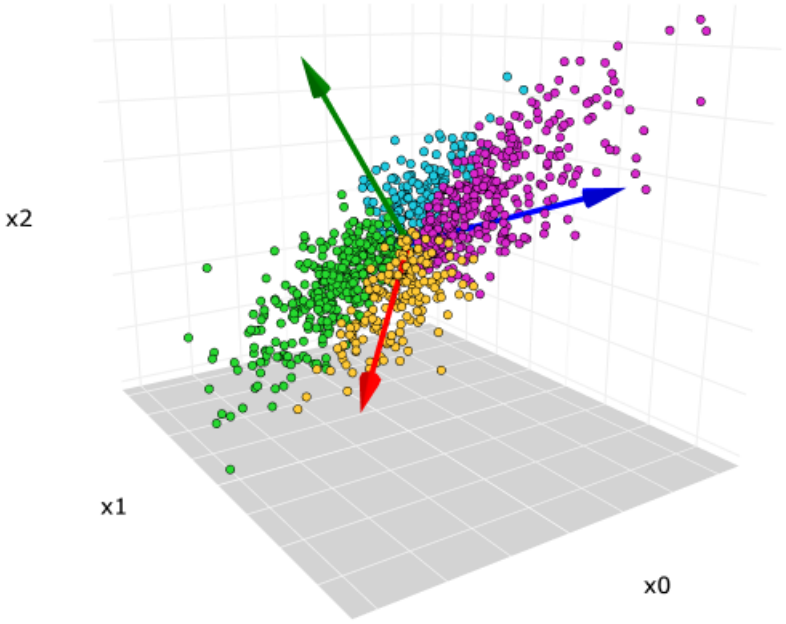


Figura 2.1: Análisis de componentes principales explicado [13]

Cabe mencionar que, antes de efectuar esta alteración de los datos, es conveniente normalizarlos/estandarizarlos, debido a la sensibilidad que tiene sobre la escala relativa de las columnas originales.

Además, dicha transformación hace que se pierda la interpretabilidad de nuestros datos, por lo que no está recomendado en problemas en los que esta sea importante.

2.2. Técnicas de aprendizaje

2.2.1. Clustering

El clustering consiste en la agrupación automática de datos. Está comprendido entre los algoritmos de aprendizaje no supervisado, esto quiere decir que a priori no conocemos la variable respuesta para cada una de nuestras muestras disponibles.

K-medias

Es el algoritmo de agrupamiento más usado. Una de sus virtudes es su gran escalabilidad. Para utilizarlo debemos especificar antes el número de grupos que deseamos encontrar, k. Las distintas etapas de este método son:

1. Elección de los k **centroides** de manera aleatoria.
2. Asignación de cada muestra al centroide que se encuentra más cerca de ésta.
3. Actualización de los k centroides al valor medio de entre todos los puntos que conforman el *cluster*.

Se repite los dos ultimos pasos hasta que no haya diferencia con la iteración anterior.

Jerárquico

Este algoritmo agrupa en base a la distancia que existe entre cada uno de los individuos, buscando que exista la mayor similitud posible dentro de los *clusters* formados. Una diferencia respecto al algoritmo k-medias es que no es necesario indicar antes la cantidad de grupos a dividir.

En función de la dirección de agrupamiento que realizamos se pueden distinguir dos tipos:

- **Aglomerativo**

  Partimos de n *clusters*, es decir, cada muestra sola forma un grupo distinto. En cada etapa, fusionamos aquellos son más cercanos. Estas uniones se siguen sucediendo hasta conseguir el número de grupos que necesitemos. En la etapa final todas las muestras formarían un único *cluster*.

- **Divisivo**

  El punto de partida es un solo *cluster* conformado por todas las muestras, en cada iteración iremos dividiendo en agrupamientos más pequeños.

También podemos diferenciar 4 modelos en función de la manera que elegimos para evaluar la distancia entre *clusters*:

- Conexión completa → buscamos la distancia más larga entre los puntos de cada *cluster*.
- Conexión simple → buscamos la distancia más corta entre los puntos de cada *cluster*.
- Distancia entre promedio de enlaces → tomamos la distancia media de cada punto de un *cluster* a cada uno de los puntos de otro cluster.
- Distancia entre enlaces centroide → tenemos en cuenta el centroide de cada *cluster* para medir las distancias.

Una representación muy interesante de este tipo de algoritmos es el **dendograma**. En la figura 2.2 podemos ver un ejemplo real, en el observamos líneas verticales, que representan la división o la fusión dependiendo de si la formación de los conglomerados se realizó mediante el método divisivo o aglomerativo, respectivamente.

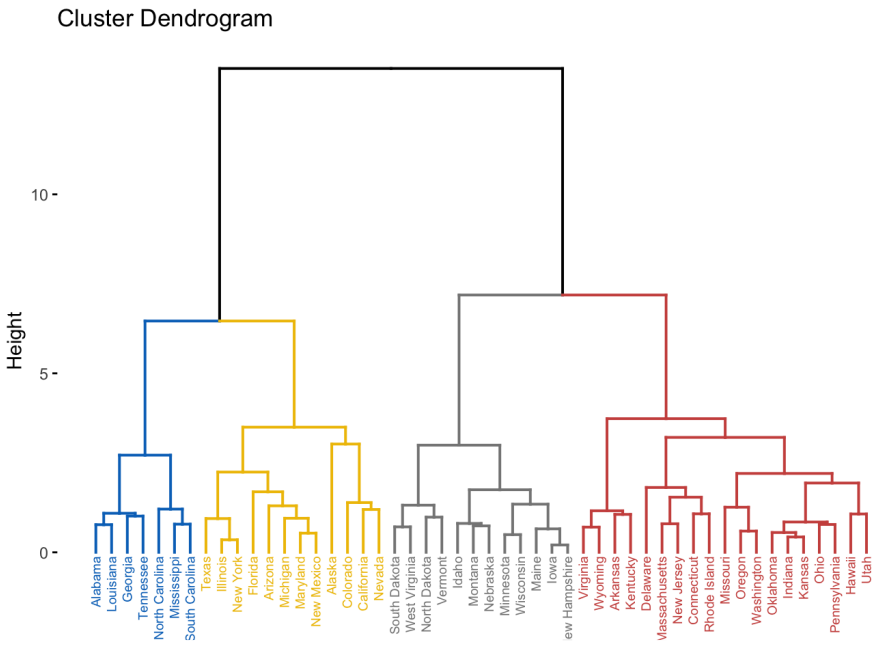


Figura 2.2: Ejemplo de un dendograma [14]

Cabe indicar, que no es un método adecuado para *datasets* de gran tamaño.

***Density-based spatial clustering of applications with noise* (DBSCAN)**

Es un método de agrupamiento que se basa en la idea de densidad, es decir, reunir los datos que se encuentren más concentrados. Como su nombre indica, es un algoritmo ideal para conjuntos de datos con ruido y valores atípicos. Este algoritmo requiere de 2 parámetros:

- **Épsilon** (ϵ): Es la distancia máxima para considerar a dos puntos vecinos.
- **Puntos mínimos** (minPts): Es el número mínimo de puntos que forman una región densa.

Con este tipo de algoritmo podemos caracterizar nuestras muestras en 3 tipos: **punto núcleo**, si tiene más de el número especificado de puntos mínimos en su radio de vecindad, **punto de borde**, si en su radio de vecindad se

encuentran menos puntos que la cantidad especificada para *minPts*, y **punto de ruido**, que es cualquier punto que no cumple ninguna de las características anteriores.

El procedimiento a seguir consiste en elegir un punto aleatorio entre los que no se han visitado aún. Si este contiene puntos dentro del vecindario ϵ, comenzamos la formación de un nuevo cluster, en caso contrario, se etiqueta como punto de ruido. En el caso de que la cantidad de vecinos sea igual o mayor al número fijado como mínimo para la formación de un cluster pasaría a ser un punto núcleo, además este cluster lo conformarán sus vecinos y los puntos de la vencidad de estos, en el caso de que ellos mismos sean puntos núcleo también. El proceso continua hasta encontrar el agrupamiento relacionado a la densidad, y se retomaría con un nuevo punto.

Sus ventajas son que puede encontrar *cluster* cuya forma no es circular, es muy útil para diferenciar grupos de gran densidad frente grupos de baja densidad y no es necesario especificarle el número de agrupaciones previamente. Entre sus debilidades se encuentran la gran sensibilidad a los hiperparámetros y la agrupación de baja calidad que hace en *datasets* con grupos de densidad variable.

Podemos leer explicaciones más detalladas de estos algoritmos en [15], [16] y [17], respectivamente. Encontrando una referencia más técnica en [18], siendo bibliografía básica de una de las asignaturas del grado.

Recomendaciones y preprocesamientos previos

A continuación, se explicarán de forma breve una serie de sugerencias a realizar, antes de aplicar un algoritmo de aprendizaje no supervisado, junto a las razones que las hacen de vital importancia a la hora de maximizar el rendimiento del clasificador.

- **Normalización**

  Corresponde con realizar una transformación que convierta la escala de la distribución de la variable. Con esto conseguimos proporciones adimensionales o invariantes de escala. Este preproceso tiene gran importancia a la hora de generar los *cluster*, ya que estos se forman a partir de las distancias.

- **Selección de variables**

  Debido a la naturaleza de los problemas de aprendizaje no supervisado, es de gran importancia reconocer los atributos que mejor clasificación generen sobre la distinción que queremos hacer. Por esto, será una parte en la que pondremos bastante énfasis en nuestro trabajo, dedicando bastante tiempo a detectar las estadísticas más relevantes pero teniendo en cuenta la dimensionalidad.

- **Selección del número de *clusters***

Dado que es un parámetro que debemos indicar antes de realizar el algoritmo, es vital conocer una estimación objetiva sobre el mejor número para nuestros datos, existen dos métodos que pueden proporcionarnos esto:

- Método del codo

  Consiste en la representación de la variación explicada en función del número de grupos, siendo el K que se situe en el codo de nuestra gráfica, el parámetro más recomendado para nuestros datos.

- Método de la silueta

  Responde a la medida de similitud de cada una de nuestras muestras respecto con el *cluster* que le ha sido asignado. Esta medida responde al rango [-1, 1], siendo los valores cercanos al límite superior los que indican mayor semejanza a su grupo y, por tanto, mayor distancia respecto a los demás. Es una representación que nos permite determinar la calidad de nuestras agrupaciones.

- Índice Davies-Bouldin (DB) Se define como:

$$DB = \frac{1}{k} \sum_{i=1, i \neq j}^{k} max(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)})$$

  Valores pequeños para este índice avisa de *clusters* compactos, es decir, que los centroides de cada uno de ellos se encuentra bien separado del resto. El parámetro óptimo de k, nos lo indicará el que minimice el valor del índice.

Estas y otras médidas útiles para la validación de técnicas de clustering se encuentran explicadas por Elizabeth León Guzmán en [19].

Capítulo 3

Búsqueda de datos

3.1. Bases de datos disponibles

La búsqueda se hizo con el objetivo de encontrar proveedores de datos interesantes para la liga española en particular, debido a que los objetivos del trabajo no se encontraban todavía muy especificados, en el momento de la investigación. Se incluyeron conjuntos de datos tanto de estadísticas de jugadores como de equipos en general. Tras la exploración en la web, como en trabajos similares, estos fueron las páginas más atractivas:

- **Football-Data.co.uk**

  Web gratuita con datos, desde la temporada 1993/94 hasta la temporada actual 2021/22, accesibles mediante descarga directa [20] y en formato *.csv*. Los datos contienen información, de todos los encuentros de la temporada, tales como: la cantidad de corners efectuados, los tiros totales realizados, la cantidad de ellos que vieron puerta o incluso cuantos de ellos se chocaron con el palo. Estas variables se podían encontrar tanto para el equipo local como para el visitante. Además podías encontrar información sobre las cuotas para distintas apuestas y para distintas casas de juegos.

  Estos conjuntos de datos tiene una gran facilidad descarga, teniendo información sobre casi 30 temporadas, pero si descartabamos los datos referidos a las apuestas, ya que a pesar de ser interesantes no eran el perfil que buscábamos, y nos quedamos con los que aportaban información sobre partidos, quizás se quedaban algo cortos, por lo que decidimos prescindir de ellos.

- **Footballdatabase.eu**

  Web completa [21], donde puedes encontrar información general para los jugadores de una plantilla en concreto, además de ojear bastantes gráficos, como comparaciones con el nivel de forma medio de su liga respecto al suyo en cada jornada, evolución de su posición en la clasificación junto

con la cantidad de goles encajados o conseguidos en cada minuto de los distintos partidos, etc. Respecto a la información de un partido en concreto puedes encontrar las alineaciones iniciales, los cambios realizados en ese partido o incluso los cambios con la alineación del anterior partido, momentos clave del partido, comparaciones varias entre los dos equipo o incluso información histórica sobre encuentros entre esos dos equipos.

El aporte de información de esta página era alto, pero la necesidad de tener que hacer *scraping*, sumado a la dificultad que tenía éste, al ser multidud de información presentada de diferente forma, y a que este procedimiento no se encontraba entre los objetivos del trabajo, fueron los motivos de descarte de este proveedor.

- **Resultados-futbol**

  Web con información básica [22] de cada partido como las alineaciones, las sustituciones realizadas o la comparativa entre equipos al final del partido. Respecto a los jugadores aporta pocos datos de valor, reducidos a: encuentros en los que ha realizado algún gol o asistencia, goles y tarjetas amarillas hasta el momento o información personal del jugador.

  De nuevo la necesidad de hacer *scraping* junto con que el aporte de información era reducido, nos llevo a no tener en cuenta esta página.

- **StatsBomb**

  Servicio interesante que proporciona una *API* tanto para los lenguajes *Python* como *R* [23]. Esta utilidad es de pago, teniendo accesos a una porción de datos gratuitos, accesibles desde su *GitHub*, los referidos a La Liga, son de las temporadas 2004-2005 a la 2020-2021, aunque los datos no son completos para ninguna de ellas, siendo el máximo de partidos recogidos 34, de 380 posibles, en alguna de las temporadas y todos son referidos a partidos que enfrentaban al FC Barcelona contra otro equipo, los datos se encuentran en formato JSON.

  El motivo principal de descarte de este proveedor obviamente fue la necesidad de pagar para adquirir sus servicios, ya que los conjuntos que aportaban gratuitamente no nos abrían la posibilidad de analizar más que un equipo en profundidad.

- **BDFutbol**

  Servicio que proporciona conjuntos de datos por rangos de temporadas, diferencia entre conjuntos sobre: clasificaciones, resultados y plantillas [24].

  Como el suministrador anterior, la necesidad de pagar fue la que nos llevo a descartar sus servicios. Además, tras analizar los ejemplos disponibles para los tres tipos de conjuntos, la información que aportaban no nos brindaba tanto potencial como para realizar dichos pagos, ya que tenía un carácter básico.

- **UnderStat**

  Web bastante sencilla pero con información bien organizada tanto para estadísticas por temporadas sobre equipos como para los jugadores de sus plantillas. La mayoría de estas variables tenían que ver con la métrica de goles esperados (xG), muy utilizada ultimamente, ya que cuantifica la calidad de las ocasiones de gol generadas. Cada página utiliza un modelo distinto, en este caso, estaba basado en una red neuronal entrenada con un gran conjunto de datos, más de 100.000 tiros, teniendo en cuenta en cada uno de ellos hasta 10 parámetros.

  Debido al beneficio de esta métrica para el análisis del rendimiento en el fútbol, se buscó una manera sencilla de obtener dicha información. Se encontraron varios paquetes creados por diferentes autores y disponibles de manera gratuita, que nos permitían hacer *scraping* de una forma bastante sencilla mediante el uso de diferentes funciones donde, indicando como parámetros los distintos filtros, conseguíamos el conjunto de datos deseados. La documentación de la biblioteca sobre la que se estudió más se encuentra en [25].

  Tras el estudio, realizamos una valoración de los tipos de conjuntos que nos permitía conseguir, donde destacaba una gran información sobre estadísticas de jugadores, la cual nos describía el comportamiento de cada jugador en diferentes aspectos del juego como: las posiciones donde había jugador durante la temporada, en diferentes tramos del partido, en diferentes alineaciones habituales o incluso con las diferentes partes de golpeo. El principal motivo de descarte fue que dichas estadísticas daban principalmente información de caracter ofensivo, además de que dichos conjuntos nos llevaban a un proyecto sobre previsión de apuestas relacionadas con goles en diferentes etapas del juego.

- **Kaggle** Filtrando por fútbol en la multitud de conjuntos de datos disponibles en dicha plataforma encontramos 2 que podrían encajar con nuestros requisitos:

  - *European Soccer Database*
    Conjuntos de datos que comprendían más de 25.000 partidos de 11 países europeos comprendidos entre los años 2008 y 2016 [26]. Incluyendo información de atributos de equipos y jugadores basados en el videojuego *EA Sports' FIFA*, hecho que nos hizo descartar dicho conjunto ya que queríamos variables que reflejaran la realidad al 100 %.
  - *Soccer players values and their statistics*
    Datos combinados de transfermarkt.de y fbref.com mediante *scraping* para las 5 grandes ligas europeas. Accesible desde [9].

  Este último, fue finalmente el conjunto elegido, ya que presentaba aproxidamente 200 variables interesantes, relacionadas con aspectos tanto defensivos como ofensivos, y teniendo en cuenta atributos para los porteros.

Capítulo 4

Exploración de datos

4.1. Descripción del conjunto de datos

El conjunto de datos finalmente elegido pertenece a la plataforma *Kaggle*, una plataforma web que cuyos usuarios son principalmente expertos o aficionados del mundo de la Ciencia de Datos, en la cual podemos encontrar más de un millón de consumidores con el objetivo de publicar, analizar y modelar conjuntos de datos. El *dataset* de trabajo se llama "***Soccer players values and their statistics***", disponible en [9]. Este comprende información de jugadores de fútbol para las 5 grandes ligas (*Premier League*, La Liga, *Bundesliga*, Serie A y *Ligue 1*) recogida mediante *scrapping* de dos web: transfermarket.de [27] y fbref.com [28]. Estos datos fueron recaudados por *Rafał Stępień* para su tesis "***Modelling Football Players Values and Their Determinants on a Transfer Market using Robust Regression Models***"[29]. Este trabajo comprende, desde la obtención de los datos de manera propia haciendo *scraping*, hasta la elaboración de 4 modelos logarítmicos lineales con mínimos cuadrados, uno para cada posición genérica del campo. Con ellos predice el valor de un futbolista a través de una serie de estadísticas y rasgos físicos. El mejor modelo conseguido por Rafał es el de los porteros, siendo peores los de los mediocentros y delanteros debido a la presencia de jugadores de alto rendimiento que se interpretan como *outliers*. Algunas conclusiones interesantes acerca de las variables más importantes por posición son: los defensas con mayor implicación ofensiva tienden a aumentar su valor de mercado, para los mediocentros no solo su capacidad de creación de juego fue importante sino que también fue decisiva su habilidad en las entradas y, finalmente, para los delanteros su habilidad en el regate y su eficacia de cara a gol fueron evidentemente muy importantes.

Podemos encontrar hasta 6 trabajos recientes [30] que han utilizado nuestro conjunto de datos, siendo su principal foco el valor de mercado de los distintos jugadores de las 5 grandes ligas. En ellos se analiza la distribución de estos en función de la edad, pierna dominante, pais de procedencia o equipo al que pertenecen. El objetivo final más destacado de estos trabajos es la predicción

del precio de los jugadores de estas ligas. A la vista de este breve resumen de los propósitos de estos proyectos, podemos concluir que nuestro trabajo presenta un enfoque distinto y novedoso, ya que se centra en encontrar agrupaciones de jugadores con habilidades parecidas y analizar dichas formaciones.

Comprende las estadísticas acumuladas para las temporadas 2017-18, 2018-19 y 2019-20, por lo que tenemos 3 *datasets* distintos, y al enfocarnos unicamente en la liga española, disponemos de 478, 456 y 538 muestras respectivamente. Cada uno de ellos contiene 400 atributos o variables, 12 de ellos con información sobre el jugador, como su altura, su equipo, su nacionalidad, etc. Otros 19 representan datos globales sobre sus equipos como los goles esperados o la cantidad de puntos conseguidos en la temporada. Después, existen 186 variables que modelan el rendimiento del jugador, siendo el resto de columnas, los valores por minuto de la mayoría de estas estadísticas.

4.2. Preparación de datos

4.2.1. Búsqueda de valores ausentes

Uno de los pasos en la limpieza de los datos es el análisis de posibles datos perdidos en alguna de nuestras variables. Con ayuda del lenguaje elegido, realizamos una inspección a cada una de estas variables para cada uno de los 3 datasets que tenemos, siendo el código resultante para el primer *dataset* el siguiente:

```
na.var <- function(variable){
  sum(is.na(variable) + 0)
}
nas <- apply(laLigaPlayers.1718, 2, na.var)
prop.nas <- round(nas/dim(laLigaPlayers.1718)[1], 2)
dataframe.na.1718 <- data.frame("No␣NAs" =  dim(laLigaPlayers.1718)[1] - nas,
                                "NAs" = nas, "Porcentaje␣NAs" = 100*prop.nas)
```

Tras aplicar dicho código, encontramos valores ausentes, para el dataset de la temporada 2017-18, en todas las variables que indican las estadísticas por minuto de los jugadores. Para estas variables tenemos un total de 23 valores ausentes, lo que implica aproximadamente un 5 % del total. Para el último *dataset*, el de la temporada 2019-20, encontramos 78 valores ausentes para la variable que indica si el jugador fue el máximo goleador de la *Champions League*, lo que representa un 14 % del total.

Dado que los atributos que presentan datos perdidos no son objeto de nuestro estudio, no aplicaremos ninguna operación a estas variables ni intentaremos completarlas con sus valores reales.

4.2.2. Creación del conjunto de datos a estudiar

Debido al enfoque del trabajo, tiene más sentido usar únicamente los datos para los jugadores que jugaron en la liga para las tres temporadas de estudio. Por tanto, se ha creado un conjunto de datos mediante operación *inner_join()* del paquete *dplyr* de R. Como resultado obtenemos 194 muestras, pero teniendo 179 jugadores que permanecieron en la máxima competición española durante

los años 2017 y 2020. Las 15 muestras de más, corresponden a 15 jugadores que cambiaron de equipo, a un club de La Liga, durante una misma temporada. Estos jugadores junto a los equipos involucrados en su traspaso se muestran en la siguiente tabla:

Nombre jugador	Temporada	Equipo 1	Equipo 2
John Guidetti	17-18	Celta Vigo	Alavés
Víctor Machín "Vitolo"	17-18	Las Palmas	Atlético Madrid
Ibai Gómez	18-19	Alavés	Athletic Club
Munir El Haddadi	18-19	Barcelona	Sevilla
Ruben Sobrino	18-19	Alavés	Valencia
Jeison Murillo	18-19	Valencia	Barcelona
Facundo Roncaglia	18-19	Celta Vigo	Valencia
Takashi Inui	18-19	Betis	Alavés
Rúben Vezo	18-19	Valencia	Levante
Youssef En Nesyrl	19-20	Leganés	Sevilla
Erick Cabaco	19-20	Levante	Getafe
Kévin Rodrigues	19-20	Real Sociedad	Leganés
Bruno González	19-20	Getafe	Levante
Manuel Agudo "Nolito"	19-20	Sevilla	Celta Vigo
Leandro Cabrera	19-20	Getafe	Espanyol

Tabla 4.1: Jugadores que jugaron en varios clubs en una misma temporada

La mayoría corresponden con fichajes en el mercado de invierno para los equipos correspondientes a la columna "Equipo 2". Este periodo de posibles traspasos, tiene lugar en Enero del segundo año que comprende la temporada, teniendo fin el 31 de ese mismo mes a las 23:59h, permitiendo a los equipos reforzarse de cara a la segunda vuelta de la competición.

4.2.3. Tratamiento de jugadores que juegan en 2 equipos en una misma temporada

De cara a analizar el posible procedimiento a seguir con esos jugadores que presentan 2 muestras para alguna de las temporadas, contemplamos dos opciones posibles:

- Agrupar en una única muestra, acumulando las estadisticas que indiquen el recuento total de una acción del juego, y recalculando las variables que indiquen los porcentajes de éxito para diferentes registros.
- Mantener las 2 muestras, debido a que su rendimiento puede verse afectado por distintas variables no contempladas en el estudio, como el correcto ajuste al estilo de juego de cada uno de los entrenadores, la química con sus compañeros, los minutos que se le brindan o su estado físico respecto a lesiones.

Como la primera suposición es de carácter fuerte, decidimos hacer una comparación de algunos de estos jugadores, para medir las diferencias de rendimiento en los dos equipos. Los jugadores elegidos son Nolito y Youssef En Nesyrl .

Equipo 1	Estadística	Equipo 2
Sevilla		Celta de Vigo
15	Partidos jugados	7
12	Partidos de titular	1
898	Minutos jugados	324
2.1	Goles esperados (xG)	2.0
0.7	Asistencias esperadas (xA)	0.4
0.3	Goles + Asist por 90'	0.83
76.2 %	Porcentaje de pases con éxito	70.1 %
70 %	Porcentaje de pases recibidos	79.2 %
42.9 %	Porcentaje de tiros a puerta	66.7 %
13	Asistencias que generan un tiro	5
56	Balones recuperados	29

Tabla 4.2: Comparación de Nolito en la temporada 19-20

Podemos observar que a pesar de tener un protagonismo menor en su nuevo equipo, Nolito presentó un rendimiento similar e incluso en cierta acciones mejor, como en la recepción de pases o la calidad de sus tiros.

Equipo 1	Estadística	Equipo 2
Leganés		Sevilla
18	Partidos jugados	18
15	Partidos de titular	7
1385	Minutos jugados	796
3.9	Goles esperados (xG)	3.6
0.8	Asistencias esperadas (xA)	0.2
0.39	Goles + Asist. por 90'	0.45
61.3 %	Porcentaje de pases con éxito	65.6 %
45 %	Porcentaje de pases recibidos	54.5 %
46.2 %	Porcentaje de tiros a puerta	38.1 %
7	Asistencias que generan un tiro	4
61	Balones recuperados	27

Tabla 4.3: Comparación de Youssef En Nesyrl en la temporada 19-20

El rendimiento de En Nesyrl se ve similar, a pesar de verse reducidos los minutos que está en el campo, en algunos campos es peor como en el porcentaje de tiros a puerta o los balones recuperados.

Por último cabe mencionar, que al intentar realizar una comparación similar con Munir El Haddadi, se ha detectado un error en el scrapping hecho para este

jugador en sus partidos jugados para el Sevilla, ya que, para dicha muestra se repiten las mismas estadísticas que cuando jugó en el Barcelona en la 2018-19. Después de esto, se ha realizado una revisión para el resto de casos y se ha encontrado el mismo error para los jugadores que cambiaron de equipo en esa temporada, no ocurriendo para los que la hicieron en la 17-18 o en la 19-20. Por lo que se ha decidido eliminar dichas muestras, ya que el *scraping* no entraba dentro del objetivo del estudio.

Jugador	Muestra válida	Muestra errónea
Ibai Gómez	Alavés	Athletic Club
Munir El Haddadi	Barcelona	Sevilla
Ruben Sobrino	Alavés	Valencia
Jeison Murillo	Valencia	Barcelona
Facundo Roncaglia	Celta Vigo	Valencia
Takashi Inui	Betis	Alavés
Rúben Vezo	Valencia	Levante

Tabla 4.4: Jugadores que jugaron en varios clubs en una misma temporada

4.2.4. Creación de variables

De cara a tener un criterio común en la mayoría de estadísticas, se ha decidido completar el conjunto de datos con la creación de nuevas variables a partir de las que tenemos. El objetivo principal es tener, para la mayoría de acciones del juego que contemplemos, 3 atributos: la cantidad de acciones intentadas, la cantidad que el jugador consiguió con éxito y la eficacia del jugador en dicha accción.

Más adelante, analizaremos con cuál de estas variables nos quedamos, para la formación de nuestro modelo.

Las variables en las cuales ha sido necesaria dicha creación han sido:

- Penalties ejecutados

  Se contaba con la información para los penaltis intentados y cuántos de ellos había conseguido marcar cada jugador. Por lo que se creó una variable que indicaba el porcentaje de acierto en el tiro desde los 11 metros.

- Penalties atajados

  Parecido a la variable anterior, pero ahora contabamos con los penaltis encajados y parados para cada jugador, aunque como es normal esta variable solo era no nula cuando se trataba de un portero. Por tanto, creamos las variables: cantidad de penaltis disputados (sumando las dos anteriores) y la eficacia en estos (dividiendo los atajados entre los jugados).

- Entradas

Variable defensiva, de la que se contaba de partida con la información de las entradas totales y cuántas de estas fueron exitosas. La nueva variable indica la eficacia en el *tackling*, dividiendo la segunda entre la primera.

- Balones aéreos disputados:

  Disponíamos de los balones aéreos que ganaba y de los que perdía, por lo que fue necesaria la creación de la suma de ambas, para prescindir de la que nos indicaba la cantidad de duelos perdidos. La eficacia en esta acción del juego ya la contemplaba el conjunto de datos.

4.3. Análisis univariante

Se ha realizado una selección de variables, basándonos en estudios previos junto al conocimiento del estudiante sobre el fútbol. Se ha escogido una porción de las casi 200 variables interesantes, repartidas en función de los distintos aspectos a comparar de un jugador, y teniendo en cuenta las diferentes posiciones existentes en el fútbol actual. El conjunto elegido de variables lo componen 51 estadísticas, siendo este número revisable a través de los distintos procedimientos que se estudiarán en las siguientes secciones de este capítulo. A continuación, se describe el significado de cada uno de estos atributos, acompañando con gráficos que presentan sus distintas distribuciones y estadísticos importantes.

Antes de comenzar, las estadísticas se analizarán por temporada, y es por ello que se debe mencionar que la última temporada, la disputada entre 2019 y 2020, fue especial. Se la ha denominado "La liga del COVID", y es que esta se vió suspendida el 12 de marzo por la situación epidémica del mundo, para después reanudarse el 11 de junio con la intención de completar las más de 10 jornadas que quedaban pendientes. Esto provocó que hubiese partidos de la liga doméstica durante todos los días hasta el 13 de julio. Esto es un hecho importante a la hora de interpretar los gráficos y tablas que se presentarán seguidamente, ya que el reducido tiempo de descanso que vivieron los distintos jugadores de los equipos puede afectar a su rendimiento.

4.3.1. Estadísticas físicas del jugador

Variable age

Indica la edad del jugador en la temporada indicada.

Variable height

Altura del jugador medida en centímetros (cm).

Variable foot

Indicador de la pierna buena del jugador, con 3 categorías posibles: diestro (*right*), zurdo (*left*), ambidiestro (*both*).

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
	2017-18	19	23	26	25.78	28	36	3.59	0
Edad	2018-19	20	24	27	26.78	29	37	3.59	0
	2019-20	21	25	28	27.78	30	38	3.59	0
Altura (cm)	-	163	176	182	181	186	196	6.34	0

Tabla 4.5: Estadísticos variables físicas

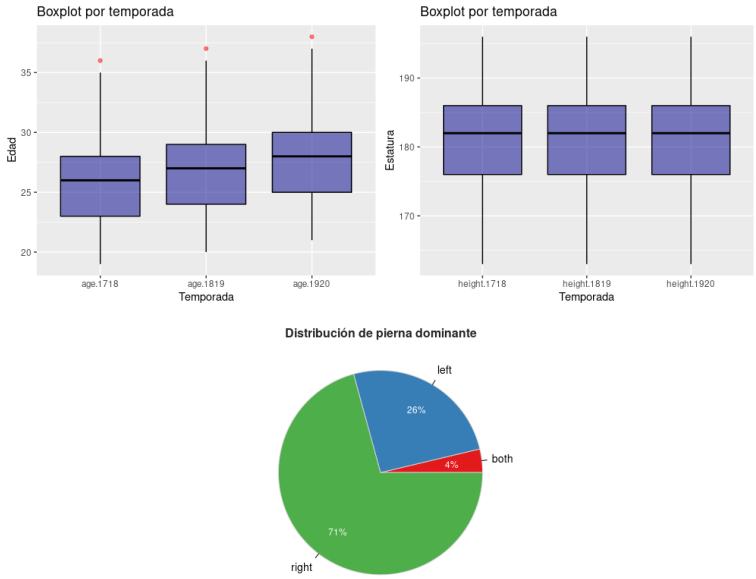


Figura 4.1: Distribuciones de los atributos físicos de los jugadores de nuestro conjunto de datos

En el *box plot* que describe la variable edad, podemos ver un "casi" *outlier*, esto es porque se encuentra muy cercano al valor resultante de $\mu \pm 3\sigma$, y si buscamos en dicha columna, el jugador más veterano corresponde con Aritz Aduriz, un reconocido delantero centro, que la mayoría de su trayectoria militó en las filas del Athletic Club, retirandosé a final de la temporada 2019-20, coincidiendo con el final de nuestro estudio. También podemos indicar, que la mayoría de los futbolistas que jugaron en la liga durante las 3 temporadas contempladas, presentan un estatura de 1,81 metros. Siendo abundantes los jugadores cuya pierna dominante es la derecha.

4.3.2. Estadísticas relacionadas con los pases

Variable xa

Asistencias esperadas, siendo ésta la suma de la variable xG de los goles generados por los jugadores a los que asiste.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Asistencias esperadas	2017-18	0.0	0.3	1.1	1.87	3.03	13.6	2.1	28
	2018-19	0.0	0.2	1.1	1.71	2.93	13.5	1.92	23
	2019-20	0.0	0.1	0.7	1.43	2.2	15.3	1.93	37

Tabla 4.6: Estadísticos variable xA

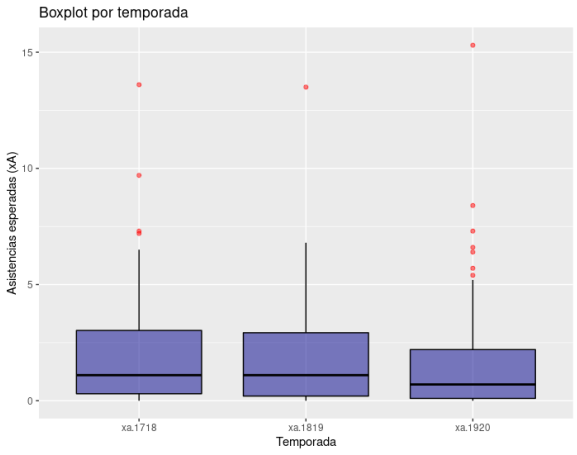


Figura 4.2: Asistencias esperadas en las temporadas

El promedio de asistencias esperadas para un jugador, no sobrepasa los 2 pases de gol, en ninguna de las temporadas. Observamos un *outlier* en todas ellas, que corresponde con Lionel Messi, del cual se espera que aporte casi 15 goles en forma de asistencia.

Variables passes_completed

Cantidad de pases realizados con éxito.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Pases realizados	2017-18	3	343.5	624.5	706.93	979.5	2157	480.66	0
	2018-19	12	303.75	629.5	706.01	986.25	2417	505.96	0
	2019-20	0	236.75	490	634.33	871.5	2469	532.11	1

Tabla 4.7: Estadísticos variable passes_completed

La media de pases con éxito para las 3 temporadas ronda los 700, siendo algo menor para la última temporada. Destacamos la existencia de jugadores capaces de dar más de 2000 pases buenos a sus compañeros.

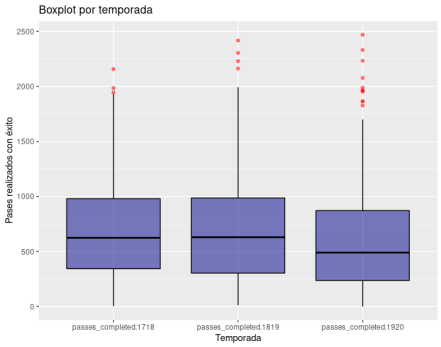


Figura 4.3: Pases con éxito en las temporadas

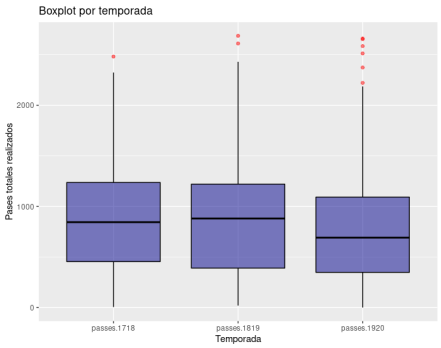


Figura 4.4: Pases totales en las temporadas

Variable passes

Cantidad de pases totales realizados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Pases totales	2017-18	6	454.75	844.5	912.2	1236.75	2481	562.06	0
	2018-19	20	390.5	880.5	897.9	1220.25	2686	590.93	0
	2019-20	1	348.25	691.5	803.24	1092.25	2659	614.46	0

Tabla 4.8: Estadísticos variable passes

Un jugador promedio, en una temporada da entre 800 y 900 pases, estando en el rango menor la temporada 19-20.

Variable passes_pct

Porcentaje de pases completados con éxito.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje de pases	2017-18	43.3	69.78	76.5	75.13	81.13	93.1	9.29	0
	2018-19	42.4	71.48	77.6	76.67	83.7	94.1	9.6	0
	2019-20	0.0	70.08	76.8	75.49	82.6	94.1	11.18	1

Tabla 4.9: Estadísticos variable passes_pct

El 50 % de los jugadores de nuestro estudio cumplen en el 75 % de las ocasiones que realizan un pase. En cuanto a los promedios, observamos una tendencia diferente a la que venimos observando, y es que la primera temporada fue en la que peores porcentajes de éxito se observaron. El *outlier* presente en la última temporada, corresponde con Vitorino Antunes, jugador del Getafe, que volvió de una rotura de ligamento cruzado el 21/11/2019 [31], lo que pudo provocarle que unicamente jugara 6 minutos en dicha campaña.

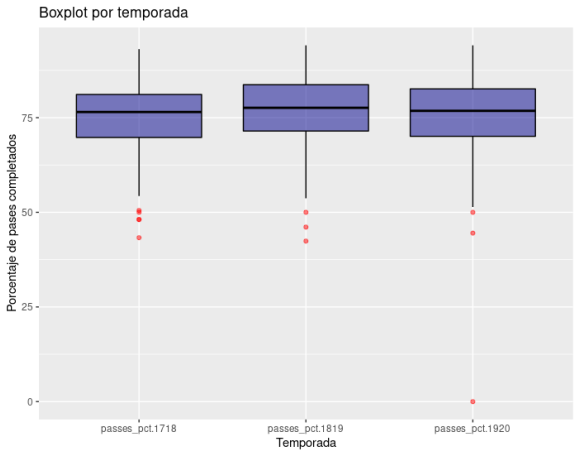


Figura 4.5: Porcentaje de éxito en el pase en las distintas temporadas

Variable passes_total_distance

Distancia total recorrida por sus pases completados en cualquier dirección, medida en yardas (yd).

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Distancia	2017-18	43.0	6535.5	12003.50	14035.8	20824.75	42311.0	9610.76	0
que recorres	2018-19	225.0	5523.5	12410.00	14049.02	19966.75	47146.0	10277.27	0
sus pases	2019-20	0.0	4581.0	10012.00	12783.27	17951.25	53752.0	10935.26	1

Tabla 4.10: Estadísticos variable passes_total_distance

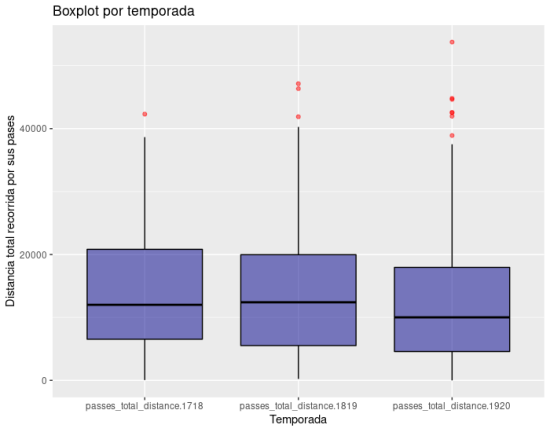


Figura 4.6: Distancia recorrida por sus pases en las temporadas

Variable assisted_shots

Número de pases que directamente conllevan a un disparo.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Tiros asistidos	2017-18	0	3	11.5	18.04	30	84	17.78	18
	2018-19	0	3	10.5	16.43	27	91	17.12	15
	2019-20	0	2	8	13.90	19	86	16.48	25

Tabla 4.11: Estadísticos variable assisted_shots

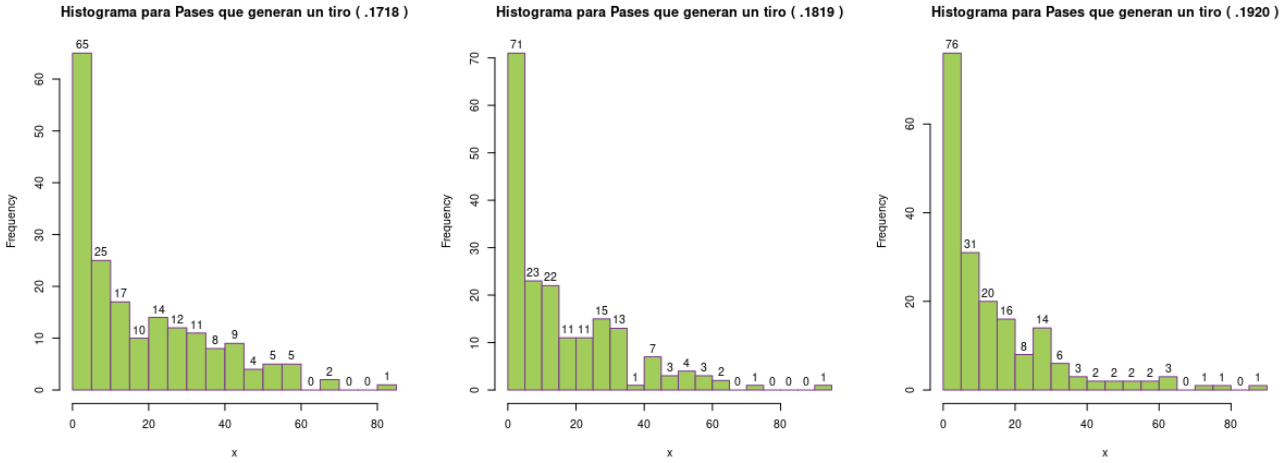


Figura 4.7: Distribución pases que conllevan un tiro en las temporadas

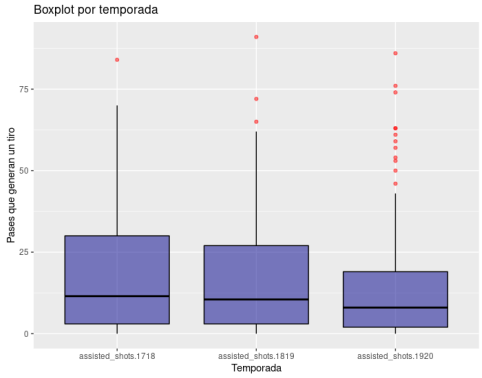


Figura 4.8: Pases que conllevan un tiro en las temporadas

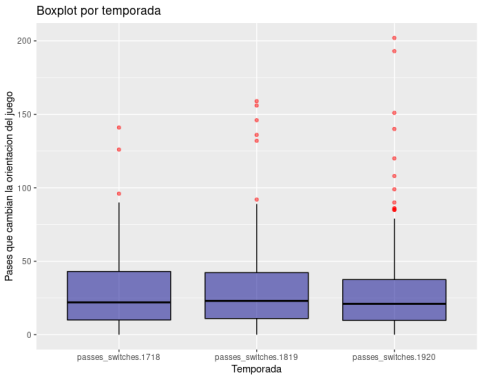


Figura 4.9: Pases que cambian la orientación del juego en las temporadas

Variable passes_switches

Número de pases que cambian la orientación del juego.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Cambios de juego	2017-18	0	10	22	28.8	43	141	24.69	7
	2018-19	0	11	23	30.69	42.25	159	29.25	5
	2019-20	0	9.75	21	29.38	37.5	202	31.64	12

Tabla 4.12: Estadísticos variable passes_switches

Variable pass_targets

Cantidad de ocasiones en las que el jugador fue objetivo de un pase.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Pases dirigidos hacia el jugador	2017-18	3	431	773	827.15	1228.5	2445	513.77	0
	2018-19	9	340	808	851.94	1239.5	2437	560.27	0
	2019-20	1	299.75	682	745.31	1002	2616	567.1	0

Tabla 4.13: Estadísticos variable pass_targets

Variable passes_received

Cantidad de ocasiones en las que el jugador recibió con éxito un pase.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Pases recibidos correctamente	2017-18	2	363	625	704.81	973.5	2051	453.08	0
	2018-19	9	300.25	647	709.48	962.75	2208	482.03	0
	2019-20	1	254.25	540	634.12	828	2229	510.82	0

Tabla 4.14: Estadísticos variable passes_received

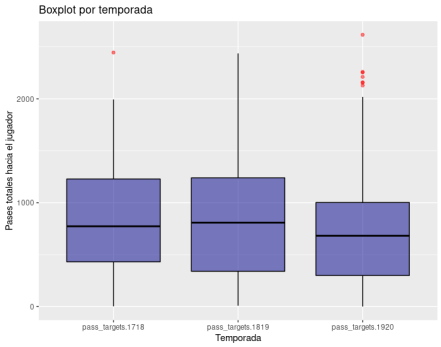


Figura 4.10: Pases con objetivo cada jugador en las 3 temporadas

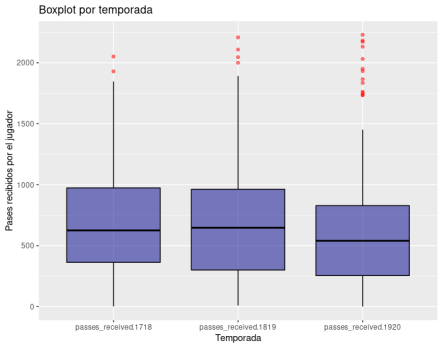


Figura 4.11: Pases totales que recibe el jugador en las 3 temporadas

Variable passes_received_pct

Porcentaje de pases, con destino el jugador, que recibe con éxito.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje de	2017-18	52.1	79.38	91.1	86.33	97.53	100	13.43	0
pases recibidos	2018-19	49.1	76.75	90	84.71	95.95	100.0	14.79	0
correctamente	2019-20	45	76.3	90.15	84.9	96.83	100.0	14.57	0

Tabla 4.15: Estadísticos variable passes_received_pct

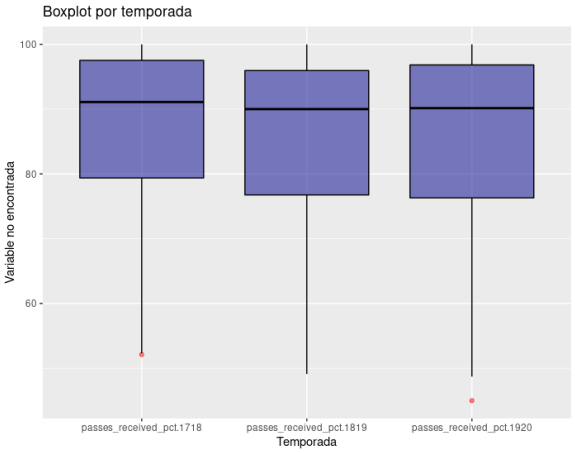


Figura 4.12: Porcentaje de pases recibidos con éxito por temporadas

4.3.3. Estadísticas relacionadas con el tiro y el gol

Variable xg

Goles esperados incluyendo los tiros realizados desde el punto de penalti.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Goles esperados	2017-18	0.0	0.400	1.300	2.864	3.53	25.30	4.09	21
	2018-19	0.0	0.475	1.15	2.55	2.33	23.3	3.87	21
	2019-20	0.0	0.2	1.0	2.03	2.33	19.2	3.1	31

Tabla 4.16: Estadísticos variable xg

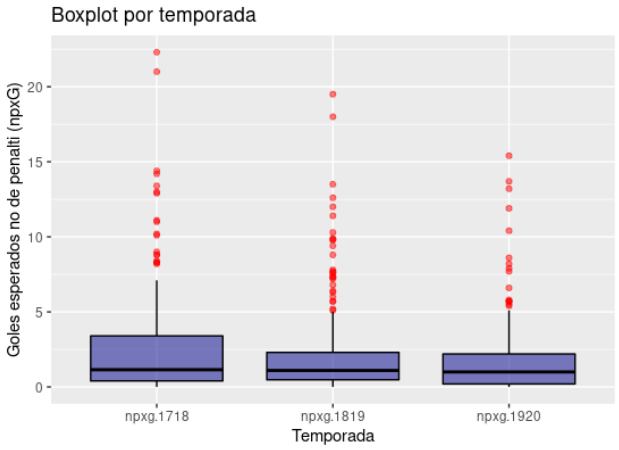


Figura 4.13: Goles esperados en las temporadas

Variable npxg

Goles esperados excluyendo los tiros de penalti.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Goles esperados (sin penaltis)	2017-18	0.0	0.4	1.15	2.6	3.4	22.3	3.64	21
	2018-19	0.0	0.475	1.1	2.28	2.3	19.5	3.27	21
	2019-20	0.0	0.2	1	1.79	2.2	15.4	2.57	31

Tabla 4.17: Estadísticos variable npxg

Como vemos Q_3 se situa en torno a 3 goles esperados por temporada, en este caso excluyendo los tiros realizados desde los 11 metros. Por tanto, la mayoría de los jugadores a estudiar no son goleadores, bien por su posición o por su estilo de juego, de hecho para más de 20 de ellos ni siquiera se espera que metan un gol. Podemos ver en la figura 4.14 como se distribuyen esos jugadores con mayor capacidad para el gol, llegando incluso a más de 22 para la primera temporada.

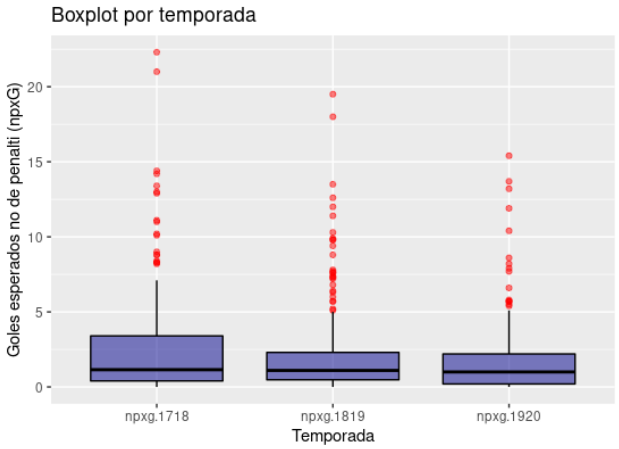


Figura 4.14: Goles esperados no de penalti en las temporadas

Variables shots_total

Cantidad de tiros totales realizados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Tiros totales	2017-18	0.0	6	17	24.83	34	194	27	19
	2018-19	0.0	6	14	22.56	30	167	25.11	19
	2019-20	0.0	3	10	18.26	25	154	21.42	26

Tabla 4.18: Estadísticos variable shots_total

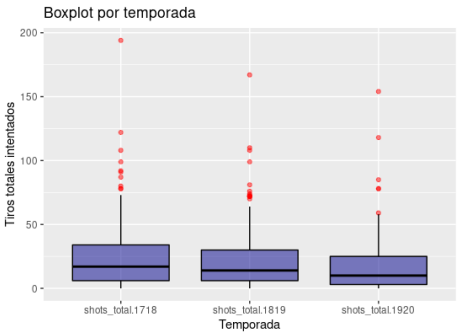


Figura 4.15: Tiros totales intentados en las 3 temporadas

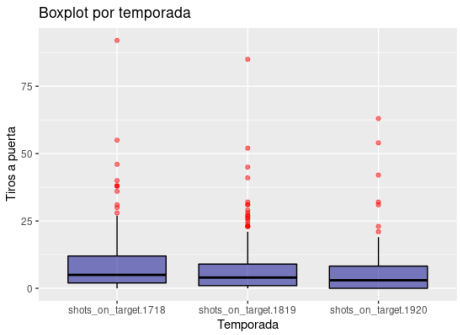


Figura 4.16: Tiros a puerta en las 3 temporadas

Variable shots_on_target

Cantidad de tiros a puerta realizados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Tiros a	2017-18	0.0	2	5	8.9	12	92	11.64	34
puerta	2018-19	0.0	1	4	7.85	9	85	10.76	30
totales	2019-20	0.0	0	3	6.28	8.25	63	8.58	48

Tabla 4.19: Estadísticos variable shots_on_target

Se puede apreciar una disminución tanto en la cantidad de tiros realizados como en los que van a puerta. Además, fijándonos en la última columna de la tabla 4.19 se puede ver que la cantidad de jugadores que no consiguen un tiro a puerta en toda la temporada es algo menos del doble de la que no consigue ni siquiera ejecutar un tiro.

Variable shots_on_target_pct

Porcentaje de tiros a puerta, de esta variable se excluyen los tiros de penalti.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje	2017-18	0.0	16.7	30.5	28.91	41.55	80	18.67	34
tiros a	2018-19	0.0	18.05	30	29.51	40.7	100.00	20.47	30
puerta	2019-20	0.0	0.0	30	26.96	40	100	20.65	48

Tabla 4.20: Estadísticos variable shots_on_target_pct

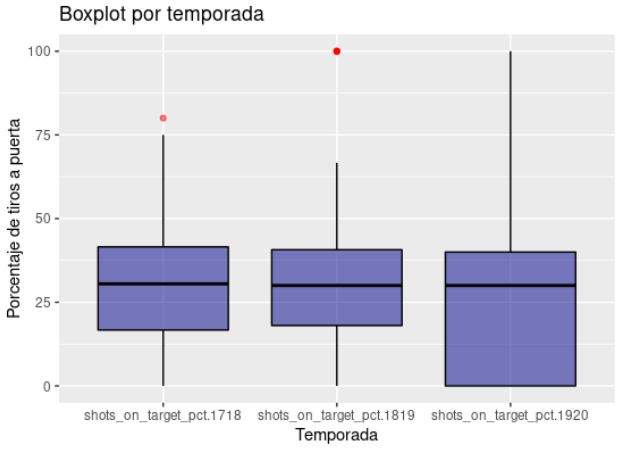


Figura 4.17: Porcentaje de éxito en tiros a puerta en las temporadas

El 50 % de los jugadores está por debajo del 31 % de acierto en las 3 temporadas. Destacamos negativamente el bajo acierto de algunos jugadores en la

temporada 19-20, ya que Q_1 se encuentra en 0 %, también se puede apreciar visualmente en el *boxplot*.

Variable goals_per_shot

Ratio de goles por tiros ejecutados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Ratio goles/asistencias	2017-18	0.0	0.0	0.065	0.076	0.13	0.38	0.079	71
	2018-19	0.0	0.0	0.05	0.081	0.13	1.0	0.12	77
	2019-20	0.0	0.0	0.05	0.079	0.13	0.38	0.09	79

Tabla 4.21: Estadísticos variable goals_per_shot

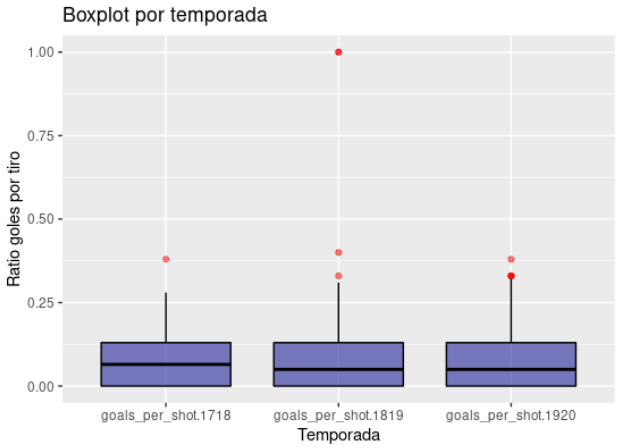


Figura 4.18: Ratio goles por tiro en las temporadas

Distribuciones bastante parecidas para las 3 temporadas. Observamos 2 jugadores con un ratio de 1 en la temporada 18-19 y observando ese conjunto de datos, vemos que son dos defensas, Sergio Postigo y Jesus Vallejo, ambos realizando 1 tiro en toda la temporada pero ambos convirtiendo en el gol esa oportunidad.

Variable pens_made

Cantidad de tiros de penalti ejecutados con éxito.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Penaltis anotados	2017-18	0.0	0.000	0.000	0.24468085	0.0000	5.00	0.72	162
	2018-19	0.0	0.000	0.000	0.30851064	0.0000	6.00	1	166
	2019-20	0.0	0.000	0.000	0.25000000	0.0000	6.00	0.93	170

Tabla 4.22: Estadísticos variable pens_made

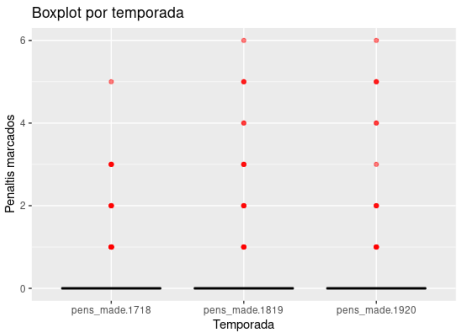


Figura 4.19: Penaltis convertidos en las 3 temporadas

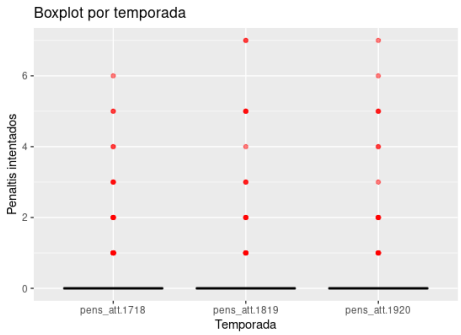


Figura 4.20: Penaltis intentados en las 3 temporadas

Variable pens_att

Cantidad de tiros de penalti lanzados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Penaltis intentados	2017-18	0.0	0.0	0.0	0.35	0.0	6	0.97	156
	2018-19	0.0	0.000	0.0	0.35	0.0	7	1.15	164
	2019-20	0.0	0.0	0.0	0.32	0.0	7	1.07	165

Tabla 4.23: Estadísticos variable pens_att

Vemos como el lanzamiento de penaltis es común que lo realicen unicamente unos pocos jugadores de la plantilla, ya que en torno a 160 jugadores de nuestro conjunto de datos no son los elegidos para tirar un penalti. Los estadísticos y gráficos de estas variables son menos interpretables por dicho motivo.

Variable pens_made_pct

Porcentaje de penalties exitosos.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje de penaltis anotados	2017-18	0.0	0.0	0.0	11.53	0.0	100	29.98	162
	2018-19	0.0	0.0	0.0	11.16	0.0	100	30.94	166
	2019-20	0.0	0.0	0.0	8.21	0.0	100	26	170

Tabla 4.24: Estadísticos variable pens_made_pct

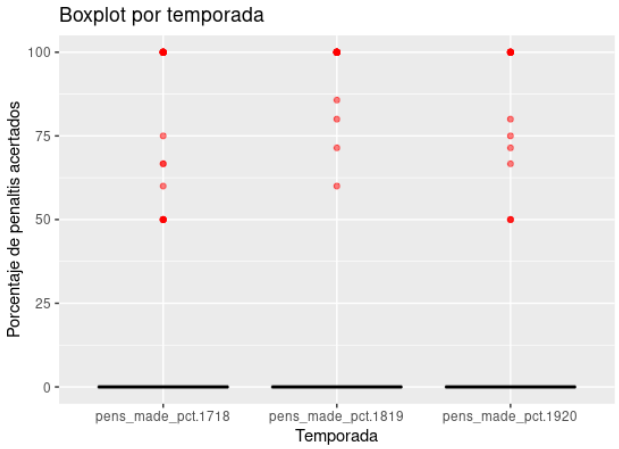


Figura 4.21: Porcentaje de éxito en el lanzamiento de penaltis en las temporadas

Lo más interpretable de esta varible es la figura 4.21, en la cual podemos observar que para las 3 temporadas existen varios jugadores infalibles desde los 11 metros.

Variable goals_assists_per90

Indica la suma de goles más asistencias promediados en 90 minutos.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Goles +	2017-18	0.0	0.03	0.15	0.23	0.37	1.38	0.26	46
Asist. en 90	2018-19	0.0	0.0	0.12	0.19	0.28	1.63	0.23	53
min.	2019-20	0.0	0.0	0.13	0.19	0.3	1.44	0.22	52

Tabla 4.25: Estadísticos variable goals_assists_per90

Existen al menos 3 jugadores en cada temporada, que por partido promedian al menos un gol o asistencia, futbolistas muy fiables y que probablemente den muchos puntos a sus equipos con sus acciones.

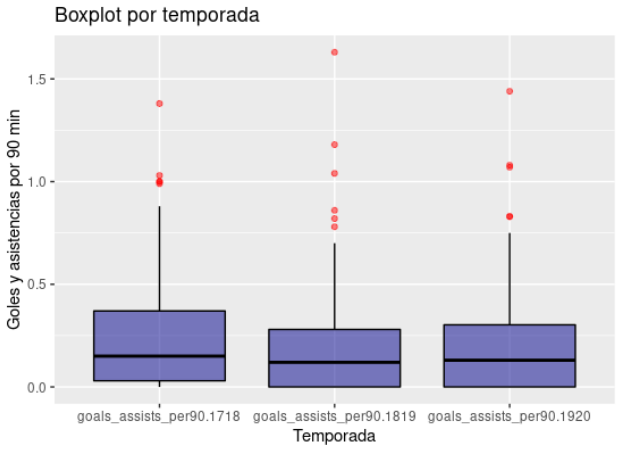


Figura 4.22: Goles + Asistencias en 90 mins en las temporadas

4.3.4. Estadísticas relacionadas con el ámbito ofensivo

Variables dribbles

Cantidad de regates totales intentados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Regates intentados	2017-18	0	6	21	30.24	46.25	256	33.32	14
	2018-19	0	5	21	33.59	53	243	37.41	15
	2019-20	0	4	15	24.9	34	274	31.61	21

Tabla 4.26: Estadísticos variable dribbles

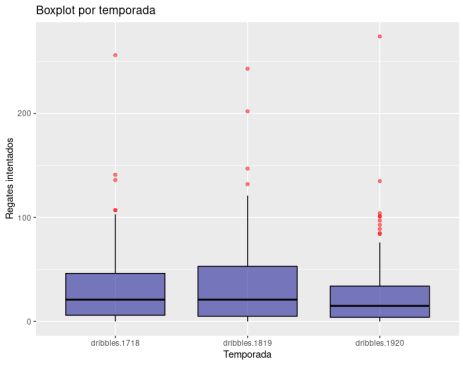


Figura 4.23: Regates intentados en las 3 temporadas

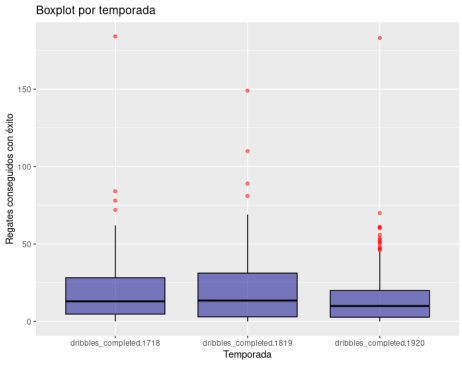


Figura 4.24: Regates conseguidos en las 3 temporadas

Para ambas variables, las dos primeras temporadas poseen distribuciones similares, siendo la temporada del COVID donde menos regates se intentaron y consiguieron. Podemos observar un *outlier* bastante claro, en las temporadas de los extremos sobre todo. Este no es otro que Lionel Messi, consiguiendo 184 y 183 regates respectivamente, es decir, más de 100 regates con éxito que el límite que pondríamos siguiendo el criterio de $\mu \pm 3\sigma$.

Variables dribbles_completed

Cantidad de regates conseguidos con éxito.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Regates conseguidos	2017-18	0	4.75	13	18.98	28.25	184	20.85	14
	2018-19	0	3	13.5	20.6	31.25	149	22.53	16
	2019-20	0	2.75	10	15.12	20	183	19.72	26

Tabla 4.27: Estadísticos variable dribbles_completed

Variable dribbles_completed_pct

Porcentaje de regates realizados con éxito.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje de regates conseguidos	2017-18	0	55.15	66.2	65.66	83.3	100	25.49	14
	2018-19	0	50	61.6	60.22	76.43	100	25.34	16
	2019-20	0	47.93	60.75	56.46	75	100	28.12	26

Tabla 4.28: Estadísticos variable dribbles_completed_pct

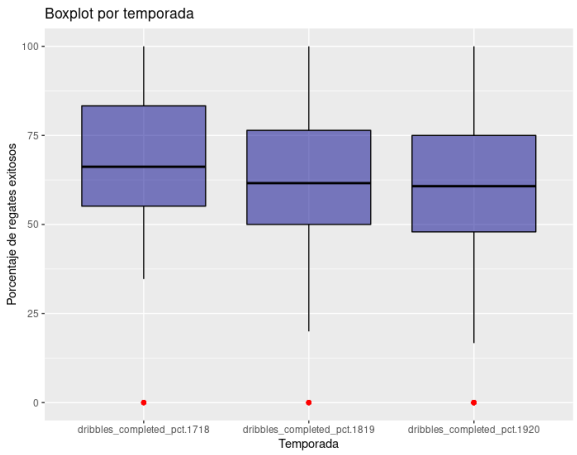


Figura 4.25: Porcentaje de regates exitosos por temporada

Si investigamos los *outliers* corresponden todos con jugadores que ocupan la posición de portero, alguno de ellos son: Antonio Sivera, Jasper Cillessen o Rubén Blanco. Cabe mencionar que, para la temporada 19-20, existe algún jugador que no juega en la posición de guardameta y tiene un 0 % en los regates que intentó.

Variable sca_dribbles

Cantidad de regates exitosos que generan un disparo.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Tiros	2017-18	0	0	1	2.45	3	39	4.5	82
generados	2018-19	0	0	1	2.44	3	35	4.31	80
por regate	2019-20	0	0	0	1.9	3	37	3.66	99

Tabla 4.29: Estadísticos variable sca_dribbles

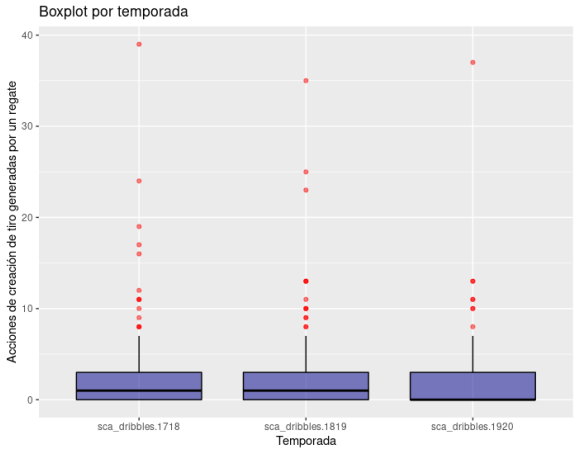


Figura 4.26: Acciones de creación de tiro tras regate exitoso por temporada

Vemos que la mayoría de jugadores son capaces de provocar como mucho tres tiros tras un quiebro exitoso al rival, durante toda al temporada. Este dato se explica en que la habilidad de regatear no es común en todos los futbolistas, y menos cuanto más alejado del área rival juegue este. Cabe destacar, que en todas temporadas, al menos existen 10 jugadores, que rebasan a sus rivales con mayor peligrosidad que el resto. Entre ellos los que sobrepasen las 16 acciones de creación de tiro con un regate son considerados *outliers*.

Variable gca_dribbles

Cantidad de regates exitosos que generan un gol.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Goles	2017-18	0	0	0	0.33	0	7	0.99	155
generados	2018-19	0	0	0	0.32	0	5	0.76	145
por regate	2019-20	0	0	0	0.29	0	5	0.78	157

Tabla 4.30: Estadísticos variable gca_dribbles

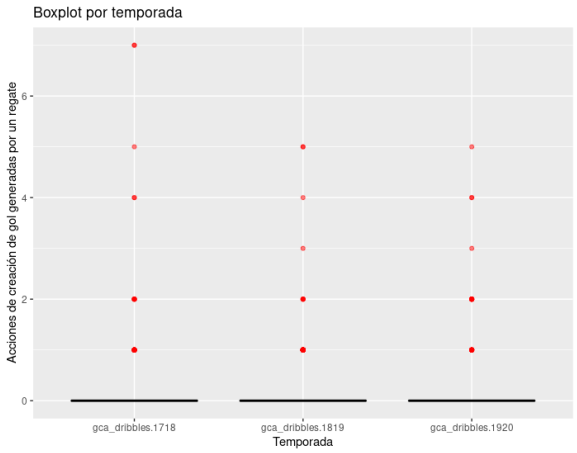


Figura 4.27: Acciones de creación de gol tras regate exitoso por temporada

El 75 % de los jugadores no genera ningún gol mediante un regate al rival, y la media del conjunto completo para cada una de las temporadas es de aproximadamente 0.3 regates existosos que generan un gol. Lo cual nos indica la dificultad de dicha acción, y del valor que tiene un jugador que es capaz de generar goles con esos movimientos.

Variable sca_passes_live

Cantidad de pases mientras se está en juego que generan un disparo.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Tiros	2017-18	0	8	19.5	27.69	45	134	24.47	12
generados por	2018-19	0	7	18.5	25.48	42	130	23.07	7
pase en juego	2019-20	0	4	15	21.25	31.75	126	22.09	14

Tabla 4.31: Estadísticos variable sca_passes_live

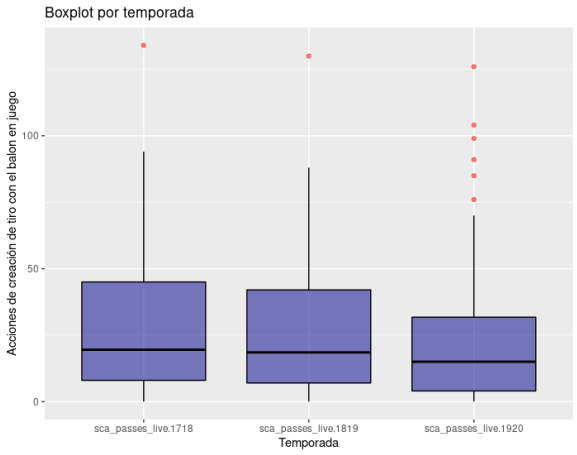


Figura 4.28: Acciones de creación de tiro con el balón en juego

Se puede notar un aumento en la cantidad de acciones en este tipo de jugadas, lo cual nos hace indicar que existe mayor facilidad a la hora de generar ocasiones, ligado a esto se entiende que la mayoría del tiempo el balón está moviendose por el campo.

Variable gca_passes_live

Cantidad de pases mientras se está en juego que generan un gol.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Goles	2017-18	0	0	2	3.46	6	24	3.95	48
generados por	2018-19	0	0	2	2.85	5	20	3.34	55
pase en juego	2019-20	0	0	1	2.58	4	22	3.28	62

Tabla 4.32: Estadísticos variable gca_passes_live

Podemos ver como los números se reducen cuando se trata de crear goles con el balon en movimiento. Destacamos la temporada 17-18 como la mejor en cuanto a rendimiento para esta estadística.

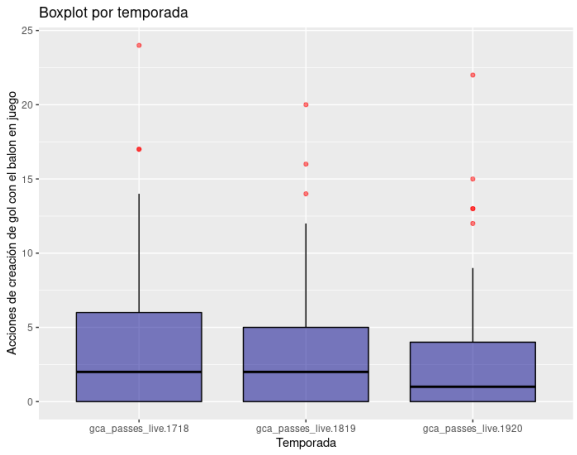


Figura 4.29: Acciones de creación de gol con el balón en juego

Variable sca_passes_dead

Cantidad de acciones a balón parado que generan un disparo, se incluyen los tiros libres, saques de esquina, saques de banda o saques de portería.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Tiros	2017-18	0	0	1	3.63	3	37	7.1	92
generados a	2018-19	0	0	0	3.18	2.25	52	7.59	99
balón parado	2019-20	0	0	0.5	3.04	3	37	6.26	94

Tabla 4.33: Estadísticos variable sca_passes_dead

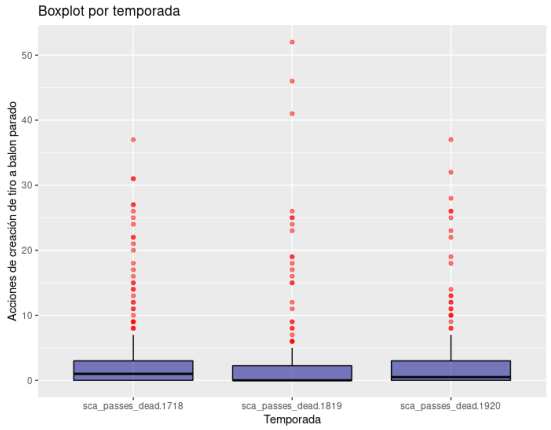


Figura 4.30: Acciones de creación de tiro a balon parado en las temporadas

Gracias al gráfico de cajas podemos ver que existe un gran grupo de jugadores que sobresalen respecto al resto. Probablemente, la principal razón de esto sea que los libres indirectos suelen ser ejecutados por un número reducido de jugadores en un equipo, por lo que estos serán los que tengan un mayor número de oportunidades de generar acciones de tiro, aunque también depende de su calidad. Esto lo podemos corroborar con la última columna de la tabla 4.33.

Variable gca_passes_dead

Cantidad de acciones a balón parado que generan un gol, se incluyen los tiros libres, saques de esquina, saques de banda o saques de portería.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Goles	2017-18	0	0	0	0.31	0	4	0.77	153
generados a	2018-19	0	0	0	0.27	0	7	0.86	162
balón parado	2019-20	0	0	0	0.23	0	5	0.71	164

Tabla 4.34: Estadísticos variable gca_passes_dead

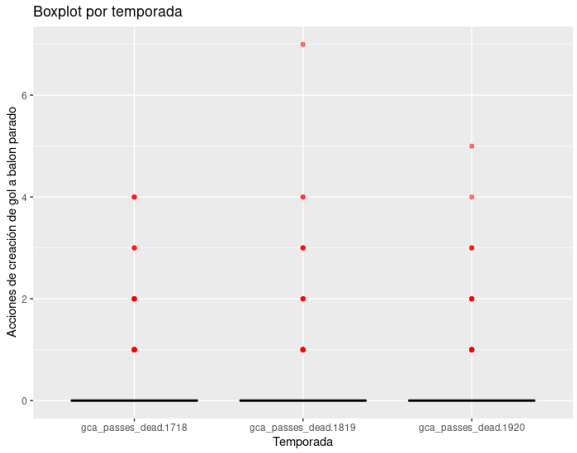


Figura 4.31: Acciones de creación de gol en las temporadas

Este es un gráfico parecido al de la variable *gca_passes_dribbles*, en él el 75 % de los futbolistas no crean goles desde falta indirecta.

Variable sca_fouled

Cantidad de faltas cometidas que provocan un intento de tiro.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Tiros	2017-18	0	0	1	2.64	4	32	3.82	64
generados	2018-19	0	0	1	2.34	3.25	28	3.5	76
por falta	2019-20	0	0	1	1.88	3	27	2.92	80

Tabla 4.35: Estadísticos variable sca_fouled

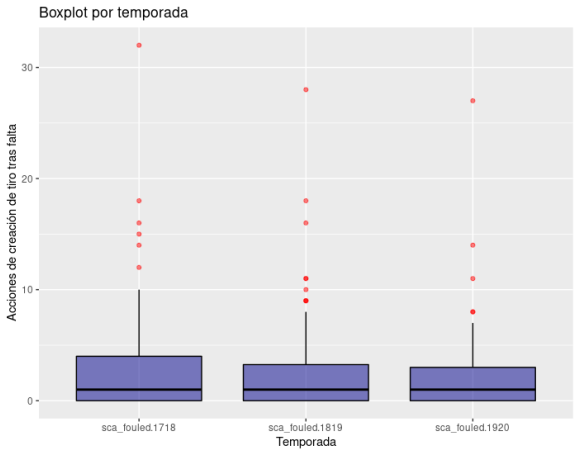


Figura 4.32: Acciones crean un tiro después de una falta realizada en las temporadas

Variable gca_fouled

Cantidad de faltas cometidas que conducen a un gol.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Goles	2017-18	0	0	0	0.39	1	4	0.73	134
generados	2018-19	0	0	0	0.38	0	5	0.82	143
por falta	2019-20	0	0	0	0.34	0	3	0.65	142

Tabla 4.36: Estadísticos variable gca_fouled

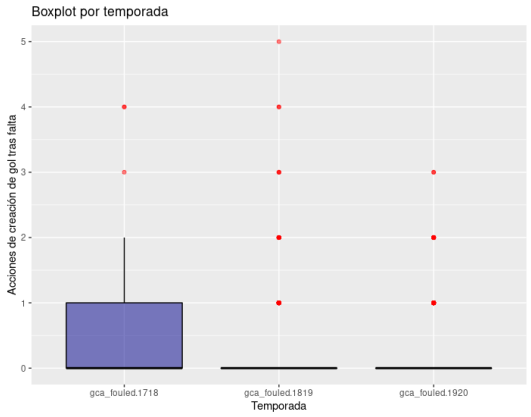


Figura 4.33: Acciones crean un gol después de una falta realizada en las temporadas

4.3.5. Estadísticas centradas en el ámbito defensivo

Variable passes_intercepted

Cantidad de pases del rival interceptados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Pases interceptados	2017-18	0	7	15	16.68	25	62	11.71	7
	2018-19	0	6	14	15.68	23	66	12.04	7
	2019-20	0	3	8	10.29	15	61	10.03	20

Tabla 4.37: Estadísticos variable passes_intercepted

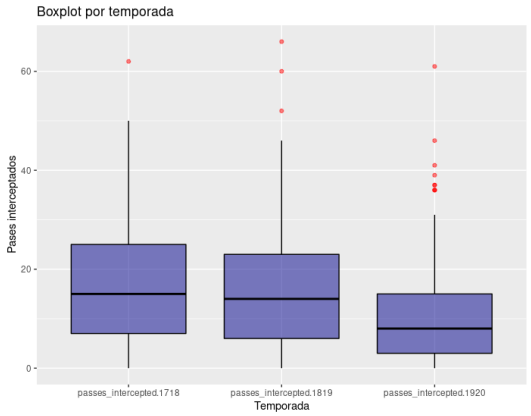


Figura 4.34: Pases interceptados en las temporadas

Se experimenta un menor rendimiendo defensivo para la temporada 2019-20, respecto a las otras dos, que parecen bastante similares. Sorprende que el jugador que más balones intercepta sea, Leo Messi, ya que se trata de un jugador más ofensivo, aún así destaca de nuevo en otra estadística.

Variable ball_recoveries

Cantidad de ocasiones en el que el jugador recupera la posesión para su equipo.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Balones recuperados	2017-18	0	94	152	175.16	248.25	513	112.72	2
	2018-19	3	77	157	166.20	239.75	502	110.57	0
	2019-20	0	61	129	146.14	205	490	107.98	2

Tabla 4.38: Estadísticos variable ball_recoveries

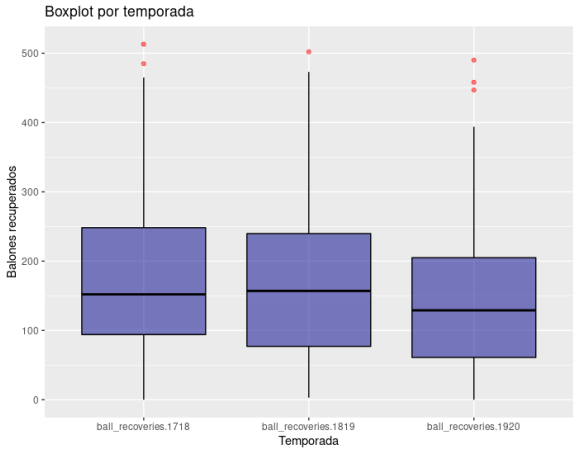


Figura 4.35: Balones recuperados en las temporadas

La mediana para esta variables se encuentra en torno a 150 balones recuperados para las 2 primeras temporadas y para la última es de 20 recuperaciones menos. Los jugadores más destacables en esta labor defensiva son: Djené y Casemiro.

Variable pressure_regains

Cantidad de ocasiones en las que el jugador recupera la posesión tras al menos 5 segundos de presión al rival.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Presiones exitosas	2017-18	0	37	79.5	80.73	116.25	213	52.48	15
	2018-19	0	35.25	69.5	75.78	110.25	217	53.32	11
	2019-20	0	21.75	58.5	62.49	93	219	46.91	9

Tabla 4.39: Estadísticos variable pressure_regains

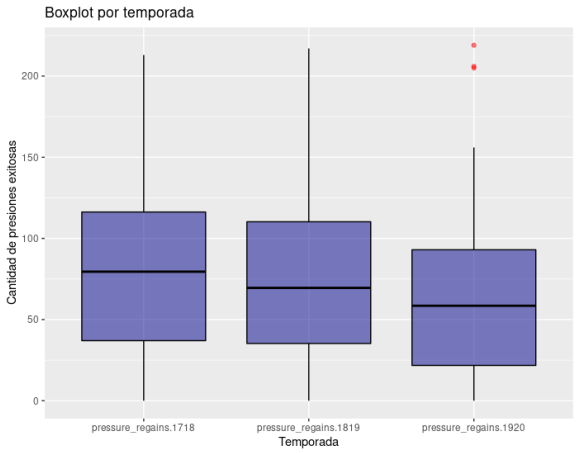


Figura 4.36: Ocasiones en las que el jugador recupera la posesión mediante una presión por temporadas

Volvemos a observar como las dos primeras temporadas se parecen, mientras que la última tiene menor rendimiento, aunque existen ciertos jugadores que presentan los rendimientos más altos de las anteriores. Destacamos que la temporada 17-18, siendo la mejor, es la que más jugadores tiene sin una presión exitosa en toda la campaña.

Variable fouls

Cantidad de faltas cometidas.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Faltas	2017-18	0	13	21	23.14	32.25	70	15.39	10
	2018-19	0	10	22	23.55	34	70	16.95	15
	2019-20	0	7.75	20	20.77	30.25	87	16.25	16

Tabla 4.40: Estadísticos variable fouls

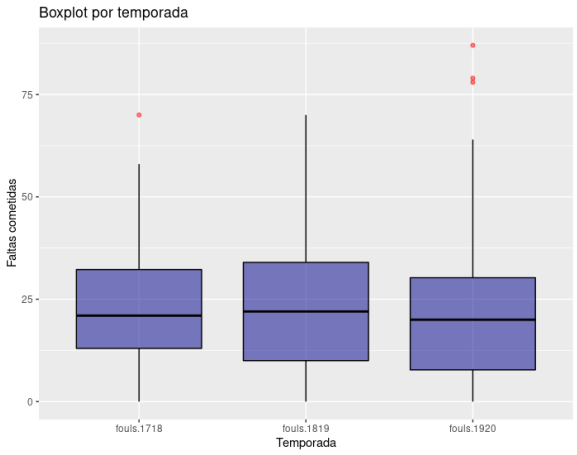


Figura 4.37: Faltas en las temporadas

El jugador que más faltas cometió en la temporada 19-20, fue Casemiro, con 87 faltas, máximo global de las 3 temporadas. Después, le siguen Mauro Arambarri y Marc Roca con 79 y 78 respectivamente.

Variable tackles_won

Cantidad de entradas realizadas con éxito, es decir, entradas que recuperan la posesión sin hacer falta o cortan la jugada del rival de manera limpia.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Entradas exitosas	2017-18	0	7.75	15.5	17.74	24.25	68	13.85	17
	2018-19	0	7.75	16	19.8	31	67	15.73	17
	2019-20	0	4	13	15.28	21.25	75	14.04	22

Tabla 4.41: Estadísticos variable tackles_won

Distribuciones muy parecidas para ambas variables. Como vimos en la figura 4.37, la temporada 18-19 era la que más jugadores realizaban más faltas, esto también lo podemos ver reflejado en estos gráficos, donde vemos que la desviación estándar para dicha temporada es algo mayor respecto a las otras dos.

Variable tackles

Cantidad de entradas realizadas.

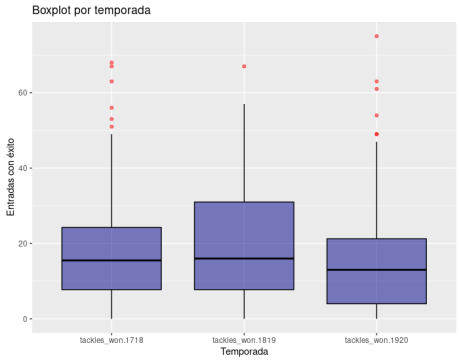


Figura 4.38: Entradas exitosas en las 3 temporadas

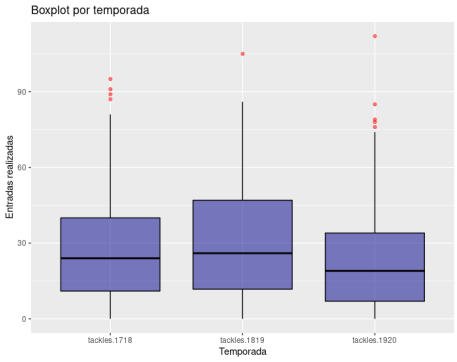


Figura 4.39: Entradas realizadas en las 3 temporadas

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Entradas realizadas	2017-18	0	11	24	27.27	40	95	20.58	17
	2018-19	0	11.75	26	30.1	47	105	23.16	16
	2019-20	0	7	19	24.12	34	112	21.19	19

Tabla 4.42: Estadísticos variable tackles

Variables tackles_pct

Porcentaje de entradas exitosas.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje de entradas con éxito	2017-18	0	55.13	64.71	59.83	72.73	100	22.17	17
	2018-19	0	52.40	63.53	60.09	73.72	100	23.13	17
	2019-20	0	51.81	61.29	56.36	69.78	100	23.7	22

Tabla 4.43: Estadísticos variable tackles_pct

Distribuciones similares para todas las temporadas. Teniendo un porcentaje de acierto de casi el 60 % a la hora de realizar un *tackling*. Después de analizar que jugadores eran más efectivos por temporada, estos son: Sergi Enrich, Munir El Haddadi y Aritz Aduriz, pero, ninguno de los 3 se encuentra dentro de los que más entradas realizan o consiguen, sino que simplemente han sido muy eficaces en ellas.

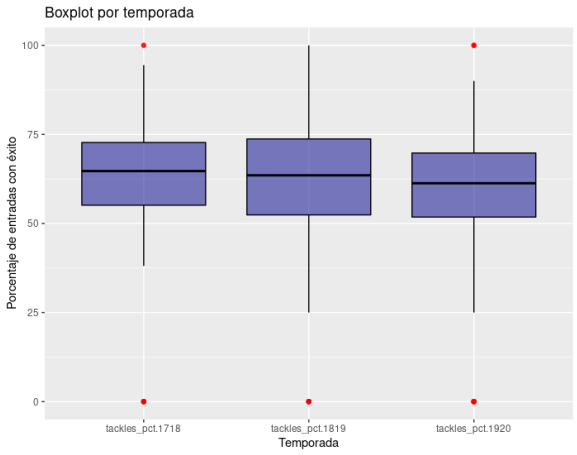


Figura 4.40: Porcentaje de éxito en entradas defensivas por temporada

Variable aerials_won

Cantidad de balones aéreos ganados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Balones	2017-18	0	4	11	16.51	24.25	73	16.04	23
aéreos	2018-19	0	5	17	25.94	36.25	156	28.8	22
ganados	2019-20	0	6.75	18	27.02	39.25	128	27.33	20

Tabla 4.44: Estadísticos variable aerials_won

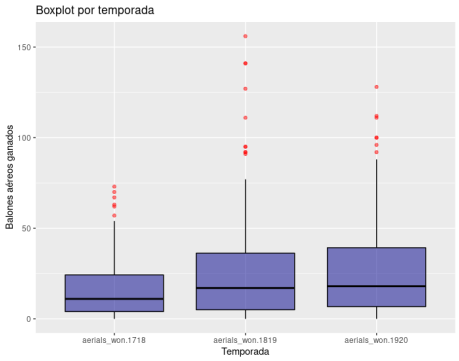


Figura 4.41: Balones aéreos ganados en las 3 temporadas

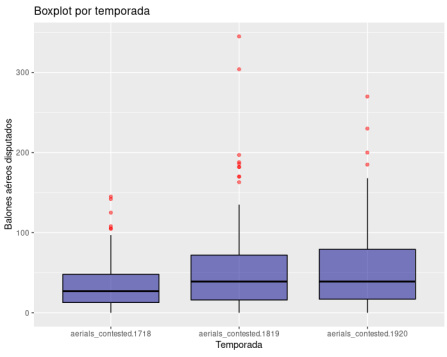


Figura 4.42: Balones aéreos disputados en las 3 temporadas

En ambos gráficos podemos ver como en cada temporada existen unos cuan-

tos jugadores que destacan sobre el resto en sus duelos aéreos.

Variable aerials_contested

Cantidad de balones aéreos disputados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Balones	2017-18	0	12.75	27	32.75	48	145	29	20
aéreos	2018-19	0	16	39	51.23	72	345	51.14	18
disputados	2019-20	0	17	39	53.57	79.25	270	49.49	19

Tabla 4.45: Estadísticos variable aerials_contested

Variable aerials_won_pct

Porcentaje de balones aéreos ganados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje	2017-18	0	25	44.7	42.07	60.08	100	22.96	23
de balones	2018-19	0	31.25	45.15	41.54	55.95	80	21.06	22
aéreos	2019-20	0	33.3	46.6	43.58	59.2	100	21.11	20

Tabla 4.46: Estadísticos variable aerials_won_pct

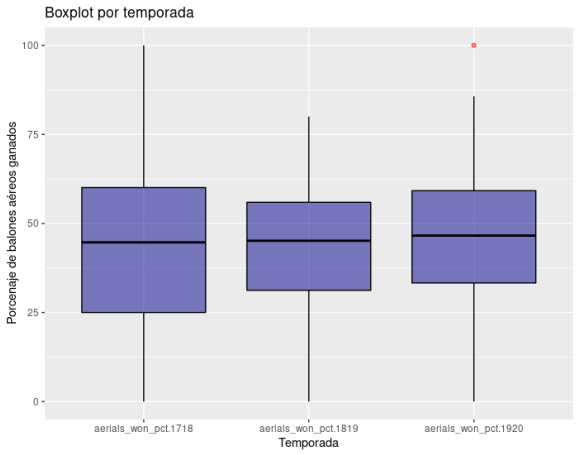


Figura 4.43: Porcentaje de éxito en balones aéreos disputados por temporadas

La media de nuestros duelos aéreos ganados se encuentra en torno al 45 % para las 3 temporadas, siendo un 60 % el máximo presentando en estos 3 años.

4.3.6. Estadísticas centradas en porteros

La interpretación de los análisis realizados sobre estas variables serán poco interpretables debido a que están hechos sobre todo el conjunto de datos, por lo que la mayoría de valores para estas estadísticas serán 0, al ser específicas de un guardameta.

Variable goals_against_per90_gk

Goles en contra promediados en 90 minutos.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Goles en	2017-18	0	0	0	0.11	0	2.4	0.39	172
contra por	2018-19	0	0	0	0.11	0	2.0	0.37	172
90 min.	2019-20	0	0	0	0.09	0	2.0	0.34	174

Tabla 4.47: Estadísticos variable goals_against_per90_gk

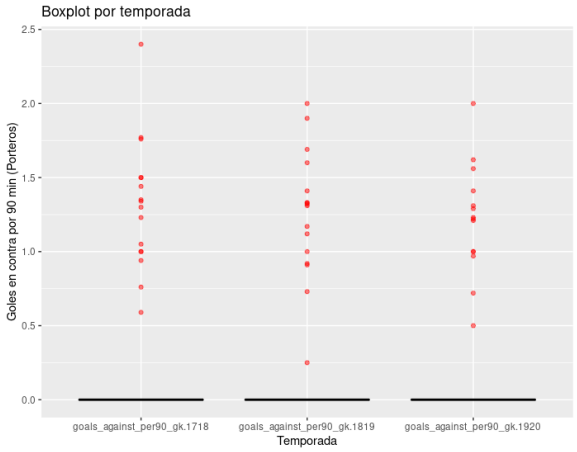


Figura 4.44: Goles en contra por 90' en las temporadas

Observamos que pocos porteros bajan de 1 gol por partido. Un dato relevante es que para la temporada 19-20 2 porteros presentan 0 goles por partido. Estos fueron Antonio Adán, y Oier Olazábal, que jugaron 1 y 2 partidos respectivamente.

Variable pens_saved

Cantidad de penalties atajados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Penaltis parados	2017-18	0	0	0	0.064	0	3.0	0.37	181
	2018-19	0	0	0	0.032	0	2.0	0.23	184
	2019-20	0	0	0	0.053	0	3.0	0.31	181

Tabla 4.48: Estadísticos variable pens_saved

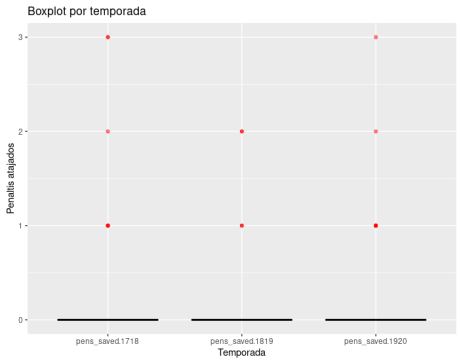


Figura 4.45: Penaltis atajados en las 3 temporadas

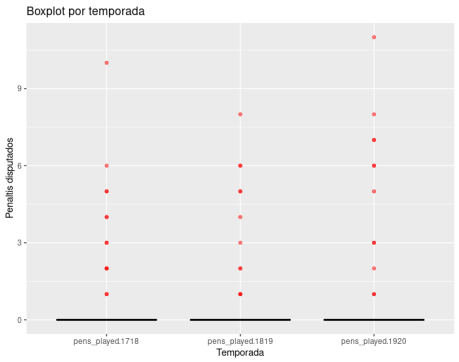


Figura 4.46: Penaltis disputados en las 3 temporadas

Variable pens_played

Cantidad de penalties disputados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Penaltis disputados (por.)	2017-18	0	0	0	0.26	0	10.0	1.13	175
	2018-19	0	0	0	0.23	0	8.0	1.06	176
	2019-20	0	0	0	0.32	0	11.0	1.43	176

Tabla 4.49: Estadísticos variable pens_played

Variable pens_saved_pct

Porcentaje de éxito parando penaltis.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje de penaltis parados	2017-18	0	0	0	1.44	0	75	8.09	181
	2018-19	0	0	0	0.59	0	33.33	4.12	184
	2019-20	0	0	0	0.87	0	50	5.01	181

Tabla 4.50: Estadísticos variable pens_saved_pct

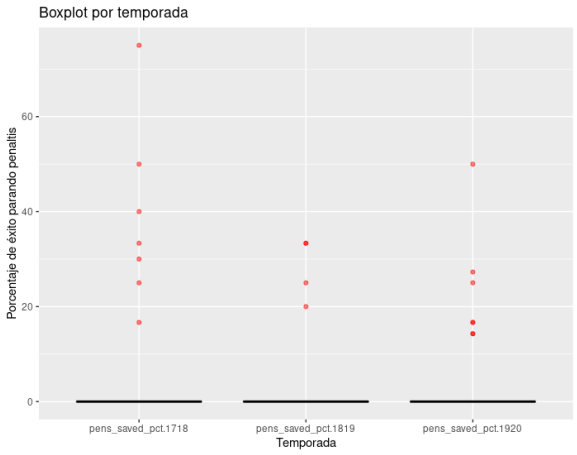


Figura 4.47: Porcentaje de éxito parando penaltis en las temporadas

Existen pocos porteros que lleguen al 50 % de penaltis parados, lo que nos indica la dificultad que tiene y que quizás sea una ocasión más favorable para los tiradores.

Variable shots_on_target_against

Cantidad de tiros a puerta en contra.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Tiros a puerta en contra	2017-18	0	0	0	6.62	0	175.0	26.74	172
	2018-19	0	0	0	7.13	0	164.0	28.73	172
	2019-20	0	0	0	6.34	0	153	25.69	172

Tabla 4.51: Estadísticos variable shots_on_target_against

Variable saves

Cantidad de tiros parados.

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Paradas realizadas	2017-18	0	0	0	4.89	0	134	19.92	172
	2018-19	0	0	0	5.26	0	124	21.34	172
	2019-20	0	0	0	4.51	0	107	18.25	172

Tabla 4.52: Estadísticos variable saves

Variable save_pct

Porcentaje de ocasiones paradas.

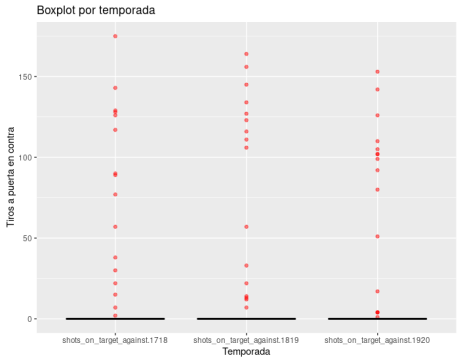


Figura 4.48: Tiros a puerta recibidos en las 3 temporadas

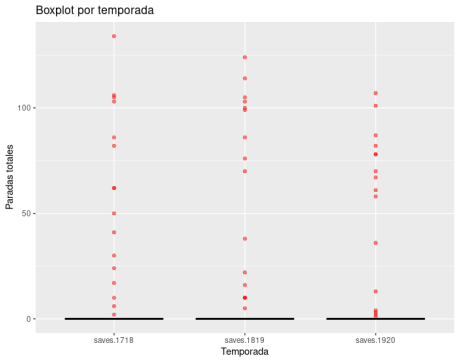


Figura 4.49: Paradas realizadas en las 3 temporadas

	Temporada	Mínimo	1er Cuantil	Mediana	Media	3er Cuantil	Máximo	Desv. Típica	Zero inflated
Porcentaje	2017-18	0	0	0	6.12	0	86.0	20.27	172
de paradas	2018-19	0	0	0	6.24	0	91.7	20.58	172
realizadas	2019-20	0	0	0	6.27	0	100	20.88	172

Tabla 4.53: Estadísticos variable save_pct

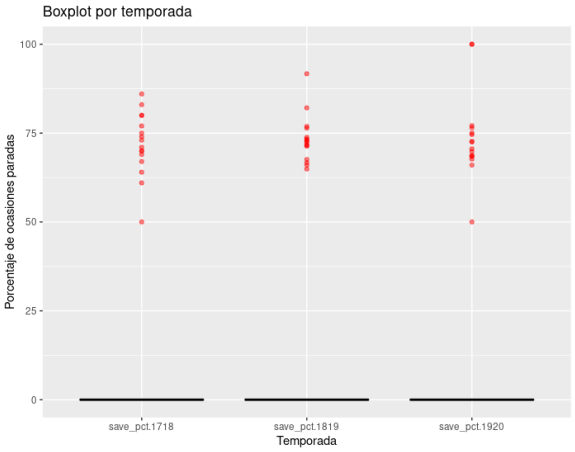


Figura 4.50: Porcentaje de paradas en las temporadas

La mayoría de porteros atajan entre el 70 y 80 por ciento de las tiros que les lanzan.

4.4. Análisis bivariante

A continuación, se examinarán pares de variables, siendo estos parejas bastante correlacionadas o que en el mundo futbolístico puedan tener relación alguna.

El objetivo de esta exploración será indagar en la posibilidad de reducir el conjunto de variables empleado. Además intentaremos encontrar las relaciones existentes entre los parejas más atractivas.

4.4.1. Relación entre variables acumuladas y porcentajes de éxito

De cara a seleccionar las variables finales del modelo, y más concretamente, decidir si escogemos las variables que indican las estadísticas acumuladas para distintas acciones de juego, o si por el contrario podemos elegir las variables que representan el porcentaje de éxito para dichas acciones, hemos realizado gráficos de puntos, enfrentando por pares las variables que implicaban a cada acción de juego. En concreto, para cada evento, tendremos 3 variables que representarán: la cantidad total intentada o realizada, la cantidad total realizada con éxito y el porcentaje de consecución, respectivamente.

Las estadísticas en las que debatimos esta decisión son:

- Pases realizados

  En la figura 4.51 se pueden observar las relaciones entre los pases realizados, completados y su porcentaje de éxito.

- Pases dirigidos hacia el jugador

  En la figura 4.52 se encuentran las relaciones entre los pases que tenían como objetivo al jugador, cuantos de estos recibió con éxito y su porcentaje de éxito en la recepción del balón.

- Tiros a puerta realizados

  En la figura 4.53 se encuentran las relaciones entre los tiros realizados, la cantidad de estos que fueron a portería y el porcentaje de éxito en rematar entre los tres palos para todos los jugadores.

- Cantidad de balones aéreos ganados frente a los intentados

  La figura 4.54 muestra las relaciones entre los balones aéreos disputados, cuantos de ellos salió victorioso y la proporción de ellos que consiguió ganar.

- Candidad de regates conseguidos frente a los intentados

  La figura 4.55 nos da la información sobre el vínculo entre los regates intentados, los que se completaron y el porcentaje de consecución.

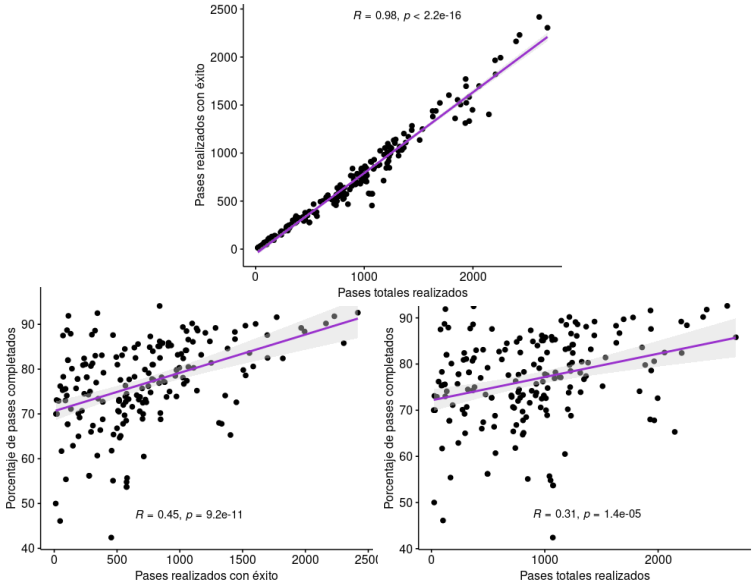


Figura 4.51: Correlaciones entre las variables que indican los pases realizados

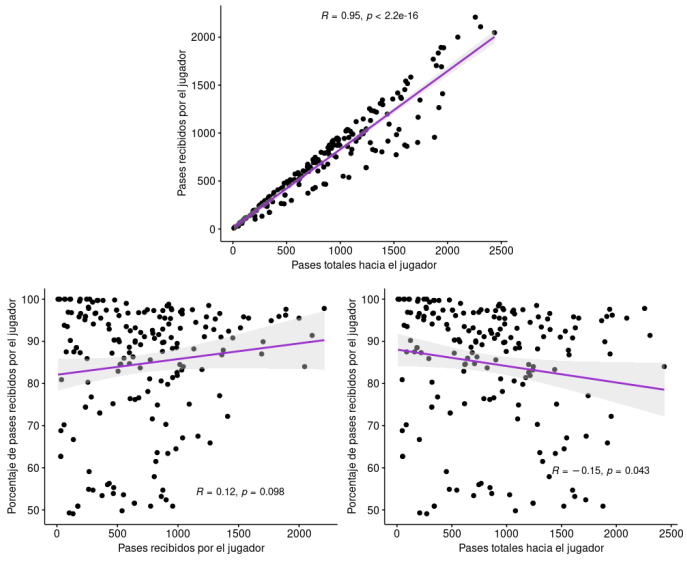


Figura 4.52: Correlaciones entre las variables que indican los pases dirigidos hacia el jugador

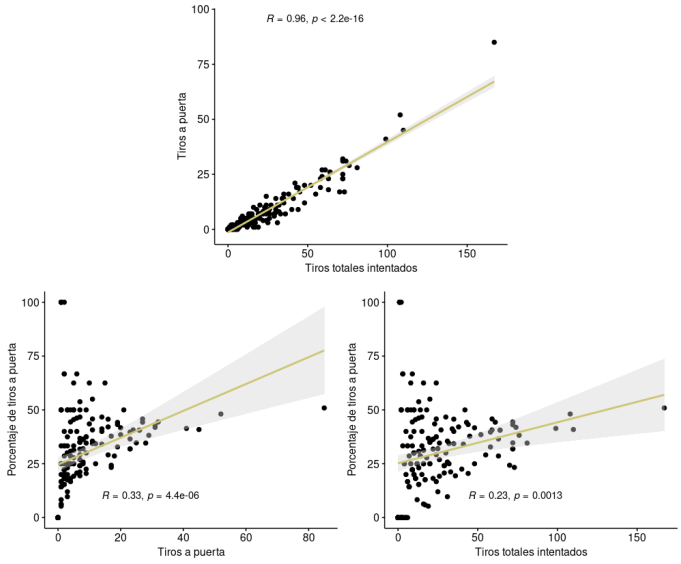


Figura 4.53: Correlaciones entre las variables que indican los tiros a puerta realizados

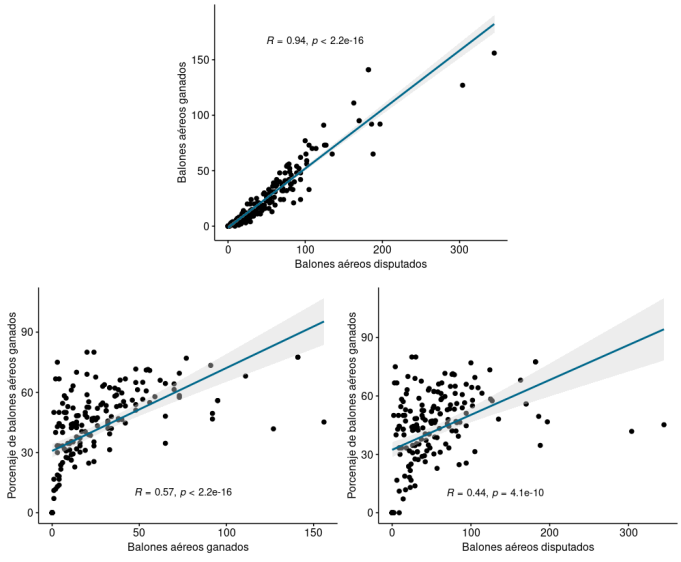


Figura 4.54: Correlaciones entre las variables que indican los balones aéreos disputados

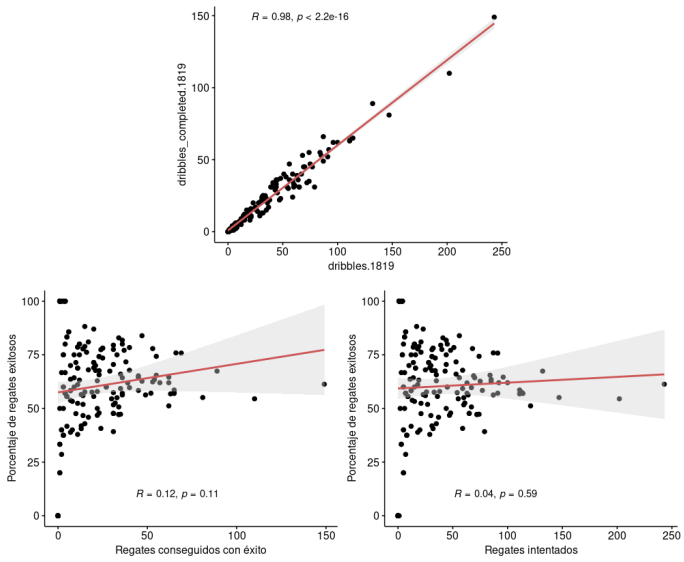


Figura 4.55: Correlaciones entre las variables que indican los regates intentados

- Tiros a puerta en contra y cantidad de ocasiones paradas

  La figura 4.56 muestra la correlación existente entre los tiros a puerta sufridos, cuantas paradas se realizaron y el porcentaje de paradas existosas.

- Lanzamiento de penaltis

  En la figura 4.57 podemos ver las relaciones entre los penaltis intentados, los que se convirtieron en gol y el porcentaje de éxito.

- Entradas

  La figura 4.58 muestra la relación que existe entre las entradas realizadas, las que fueron exitosas y el porcentaje de ellas que lograron cortar la jugada.

- Lanzamiento de penaltis (portero)

  En la figura 4.59 podemos ver las relaciones existentes entre los penaltis disputados, los parados y el porcentaje de éxito en este tipo de acciones para los porteros.

Se puede observar una correlación bastante alta entre las 2 variables que indican los acumulados. En la mayoría de casos es mayor a 0.95, por lo que para la clasificación podríamos prescindir de una de ellas.

En nuestro caso hemos decido descartar las variables cuya información era la cantidad de acciones realizadas con éxito, quedándonos con las variables que

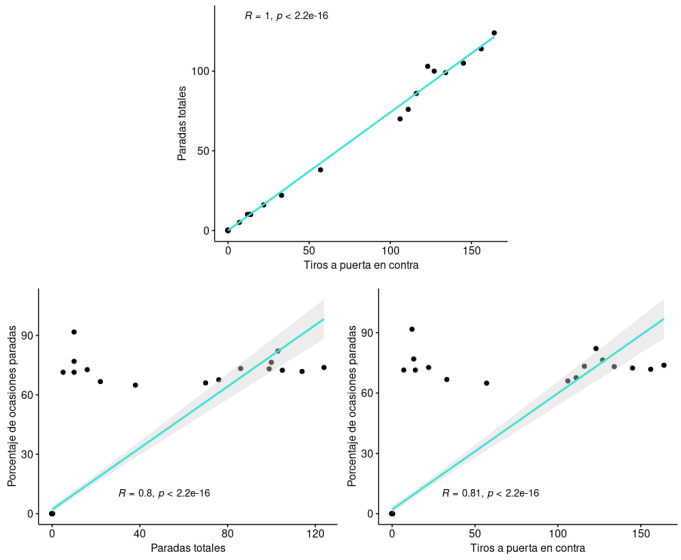


Figura 4.56: Correlaciones entre las variables que indican las ocasiones que el rival te tira a puerta

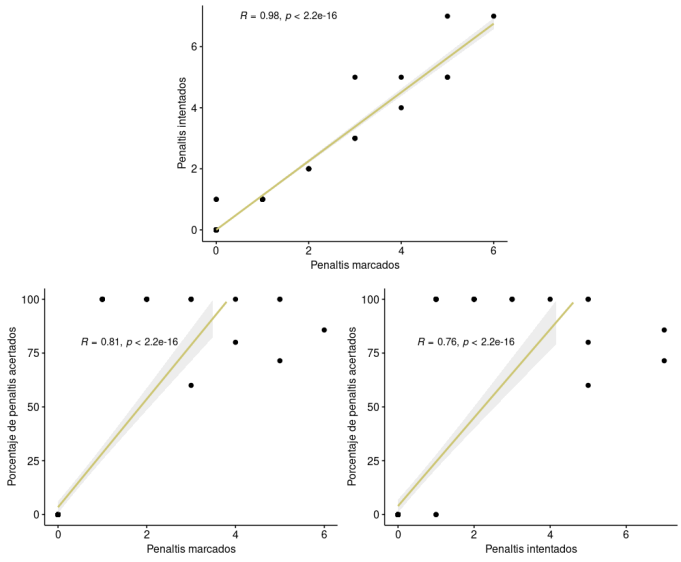


Figura 4.57: Correlaciones entre las variables que los lanzamientos desde los 11 metros para un jugador

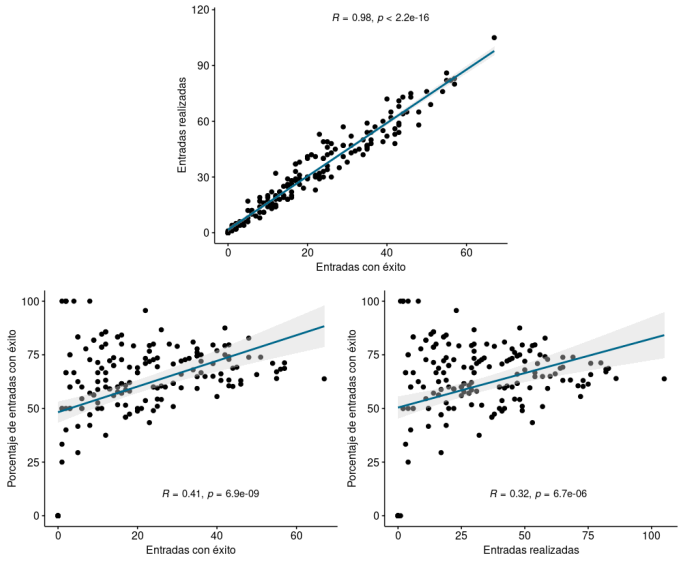


Figura 4.58: Correlaciones entre las variables que indican las entradas realizadas para un jugador

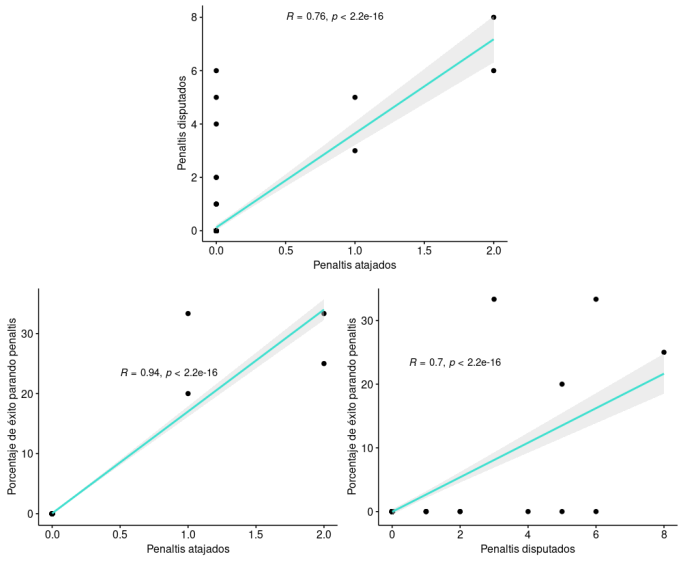


Figura 4.59: Correlaciones entre las variables que indican las ocasiones en las que un portero se enfrenta a un penalti

indican la cantidad total intentada, para cada acción, junto con sus porcentajes de éxito. Esto es así porque con dicha combinación de estadísticas podemos calcular tanto la variable eliminada, como la cantidad de intentos fallidos o incluso su porcentaje de fallo.

4.4.2. Promedio de estadísticas

En nuestro conjunto de datos tenemos 3 columnas para cada una de las estadísticas comentadas en el análisis univariante, dando lugar a un *dataset* bastante grande. Por tanto, una opción viable, debido a que el tiempo de estudio del conjunto de datos es pequeño, es promediar los valores para las tres temporadas y trabajar únicamente con esas nuevas variables.

La posibilidad de hacer esto se basa en que la carrera de un futbolista profesional, dura en torno a 8 años [32], siendo cada vez mayor debido al avance de conocimiento en áreas como la medicina, la nutrición o la prevención de lesiones, sumado a que ahora cada vez existen más casos de jugadores que debutan más jovenes. Por tanto, el rendimiento en la mayoría de los jugadores de nuestro estudio se debería conservar.

Para ello realizamos un estudio de cómo se mantienen las relaciones entre las distintas variables, una vez fusionadas.

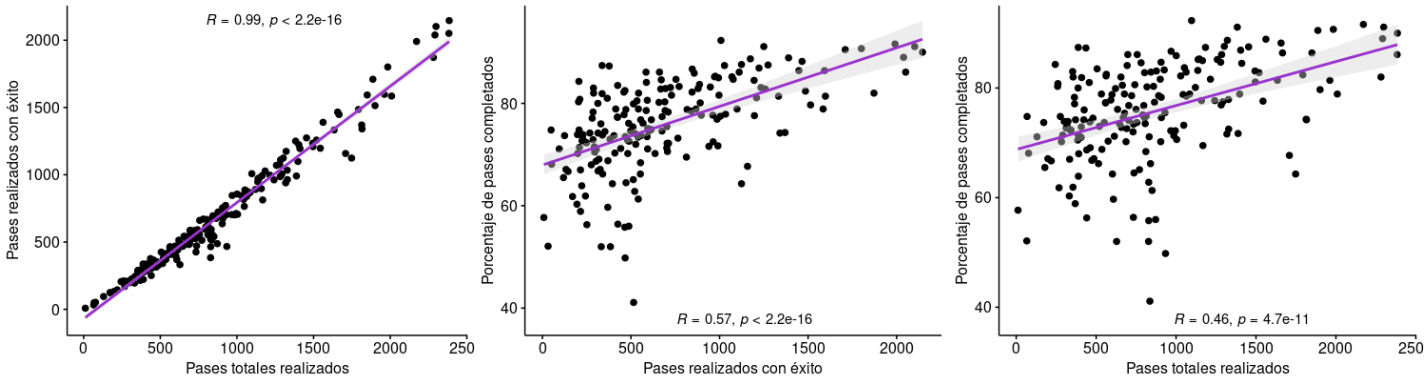


Figura 4.60: Correlaciones entre las variables promediadas que indican información sobre los pases realizados

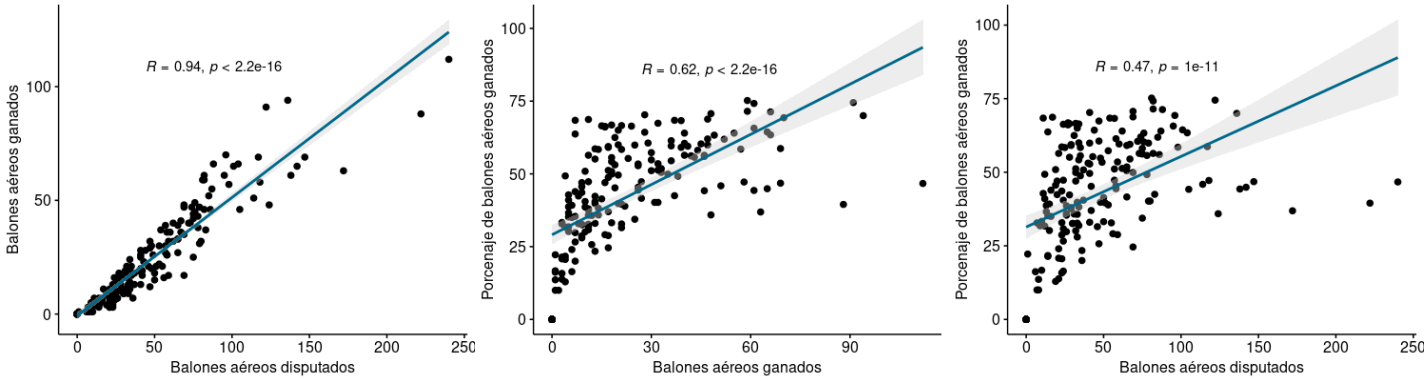


Figura 4.61: Correlaciones entre las variables promediadas que indican información sobre los balones aéros disputados

Como podemos observar la correlaciones y los p-valores se mantienen. En los gráficos de las variables promedidadas se presentan los datos con mayor agrupación, es decir, no se encuentran igual de dispersos que en 4.51 y 4.54 respectivamente.

4.4.3. Análisis de correlaciones altas

De cara a reducir el número de variables del modelo, se ha realizado una búsqueda de los distintos atributos con un nivel de correlación mayor. Después de analizar la matriz de correlaciones, calculada para los coeficientes de *Pearson*, se ha extraído las estadísticas altamente correlacionadas para mejorar la legibilidad, debido al número tan alto de variables, ya que en este momento contemplamos 42 (incluyendo las físicas).

Correlación > 0.9

Como podemos observar en la figura 4.62 hemos colocado 5 pares distintos de variables, que se encuentran altamente correlados, al menos en un 95 %, todas ellas con tendencias crecientes. Eso significa que podemos prescindir de una variable por pareja. Teniendo en cuenta el objetivo final del trabajo y la utilidad de cada estadística decidimos eliminar:

- passes_total_distance
  Se la consideró en un principio, debido a la importancia que tiene en el fútbol acercate lo más posible al área rival, ya que cuanto más cerca estés, más presión generaras. Debido a que lo que queremos conseguir es una herramienta de comparación en los traspasos, se ha creído que es una variable más prescindible que la cantidad de pases que realiza un jugador.

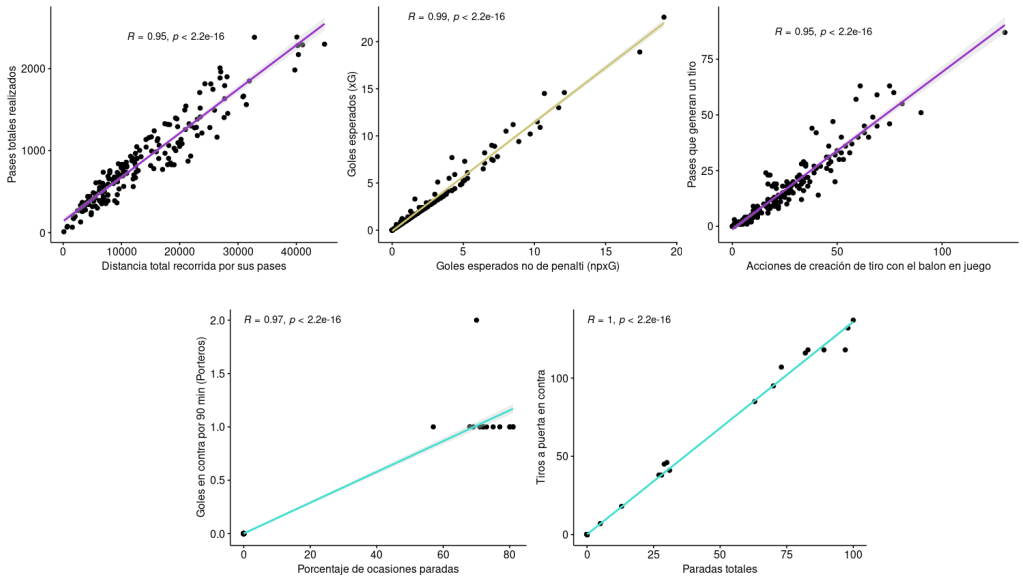


Figura 4.62: Correlaciones que superan el valor de 0.9 entre distintas variables

- xG
  Cuantifica la calidad de las ocasiones de un jugador, pero incluye los lanzamientos de penalti, aspecto que ya controlamos con los penaltis lanzados y el éxito en ellos, por lo que preferimos quedarnos con npxG, que nos aporta un comparación entre jugadores goleadores.

- assisted_shots
  Fue considerada en un principio, al no incluir la cantidad de acciones con el balon en juego que generan un tiro, pero de cara a poder filtrar por los 4 tipos de acciones de generación de tiro y gol decidimos prescindir de ella ya que es la suma de nuestros 4 "scas", acciones que generan un golpeo.

- goals_against_per90_gk
  A la hora de comparar un portero, consideramos más importante el porcentaje de ocasiones que logra atajar, ya que los goles que recibe no depende unicamente de él, sino de también de sus defensas.

- shots_on_target
  Se ha prescindido de ella por una razón parecida a la anterior.

4.4.4. Relaciones de pares interesantes

A continuación, estudiaremos la relación entre pares de variables atractivas, es decir, variables que tiene un alto nivel de correlación, cuyo interés en el ámbito

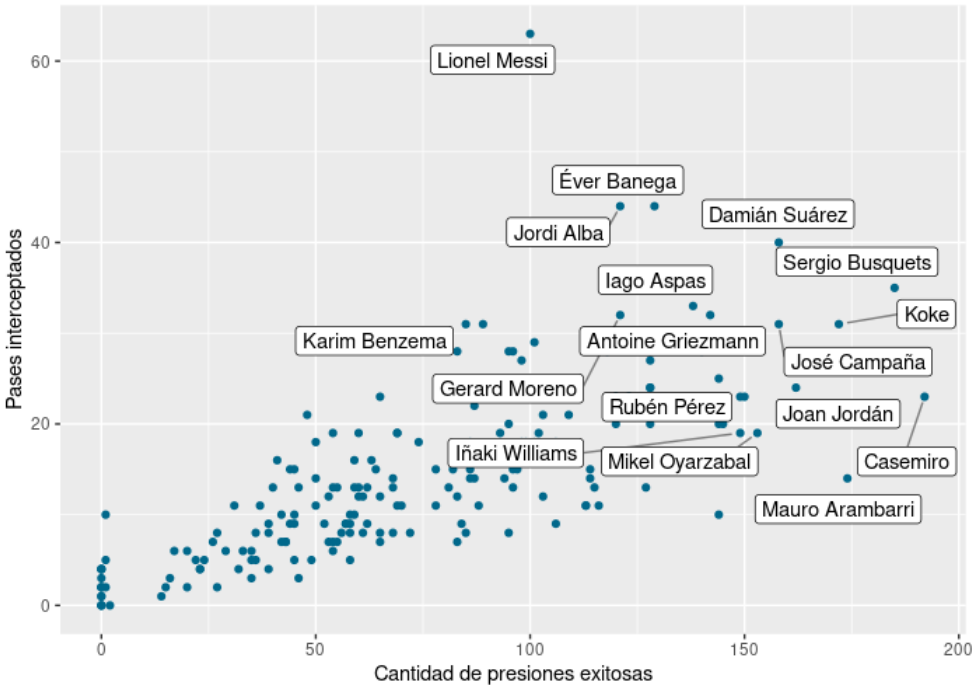


Figura 4.63: Relación entre presiones y pases interceptados

futbolístico es interesante, o ambas.

pressure_regains ~ passes_intercepted

Los nombres indicados en cada uno de los siguientes gráficos de puntos corresponden con los diez mejores de cada una de las variables estudiadas, pudiendo coincidir alguno de ellos.

En la figura 4.63 vemos que el grupo que más destaca en estas estadísticas corresponde principalmente a delanteros o mediocentros muy destacables defensivamente como Casemiro o Sergio Busquets. La presión se hace sobre todo en campo rival y cuando este se encuentra dando pases hacia atrás o de manera horizontal, y principalmente se ejecuta por los jugadores que forman el ataque.

tackles_pct ~ fouls

Los posiciones de los jugadores más destacados en estas dos estadísticas no es tan claro como antes, ya que podemos encontrar defensas, laterales, pivotes e incluso algún delantero. Esto nos indica que no es una característica específica de un tipo de jugadores dado que saber realizar la entrada a tiempo sin hacer falta al rival es una habilidad que debe ser trabajada.

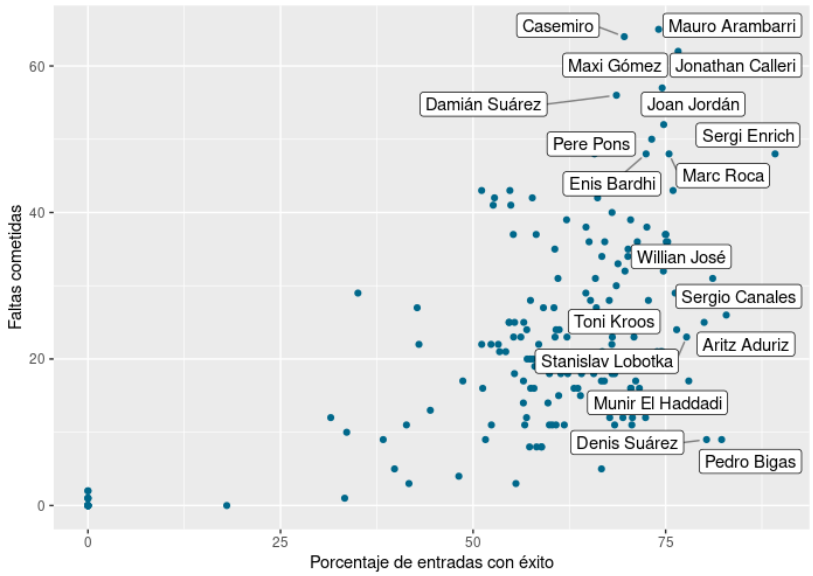


Figura 4.64: Relación exito en el *tackling* y faltas realizadas

En base a la figura 4.64, se podría decir que los jugadores más limpios, pero a la vez más efectivos en su elección en las entradas, se encuentran en la esquina inferior derecha.

sca_passes_dead ~ passes_switches

Destacable la visión de juego y la calidad a balón parado de Toni Kroos y Éver Banega, dos jugadores clave en sus equipos, en aspectos como colocar el balón en el área desde una falta o saber desahogar el juego por otra parte del campo. En la figura 4.65 destacamos el rendimiento de 2 laterales como son José Ángel y Damían Suarez, o como un joven Joan Jordán ya despuntaba, talento que no dejo escapar el Sevilla FC que le contrató en la última temporada de nuestro estudio.

dribbles ~ pass_targets

Se puede ver una gran diferencia entre Messi y el resto. La conclusión que sacamos de la figura 4.66 sobre la relación aparente entre estas variables, es que tener en tu equipo a un jugador desequilibrante puede provocar, que sea más buscado por el resto de sus compañeros en casos en los que el ataque este atascado.

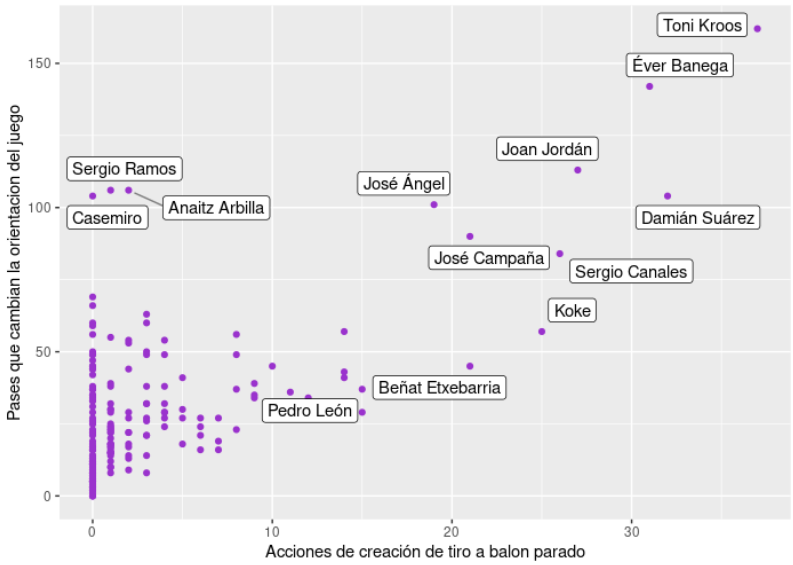


Figura 4.65: Relación entre las acciones de tiro desde balón parado y los cambios de orientación del juego

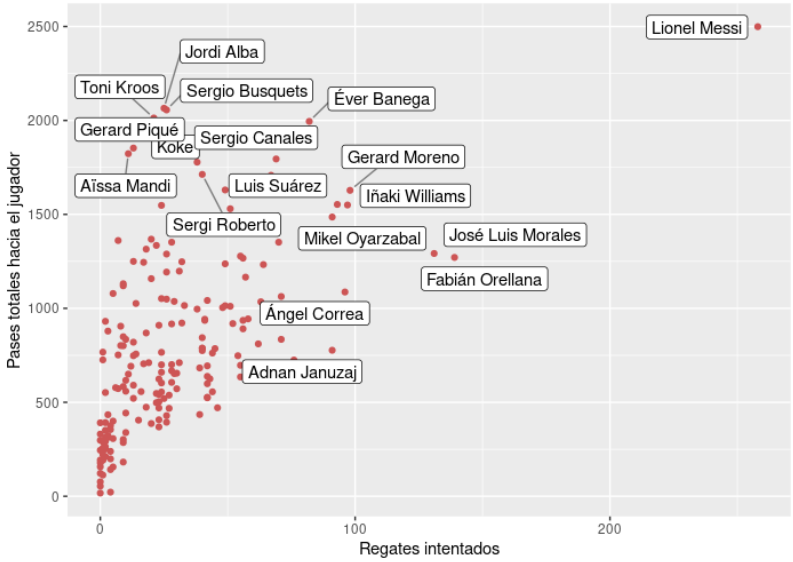


Figura 4.66: Relación entre los regates intentados y la cantidad de veces que intentan pasarte el balón

4.4.5. Análisis de correlaciones

En esta sección podremos ver las matrices de correlación para nuestras 37 variables finales, calculadas tanto para los coeficientes de *Pearson* como de *Spearman*. Ambas matrices se muestras en las figuras 4.67 y 4.68. No observamos gran diferencia entre las dos, por lo que no será necesario fijarnos en una en concreto para sacar conclusiones.

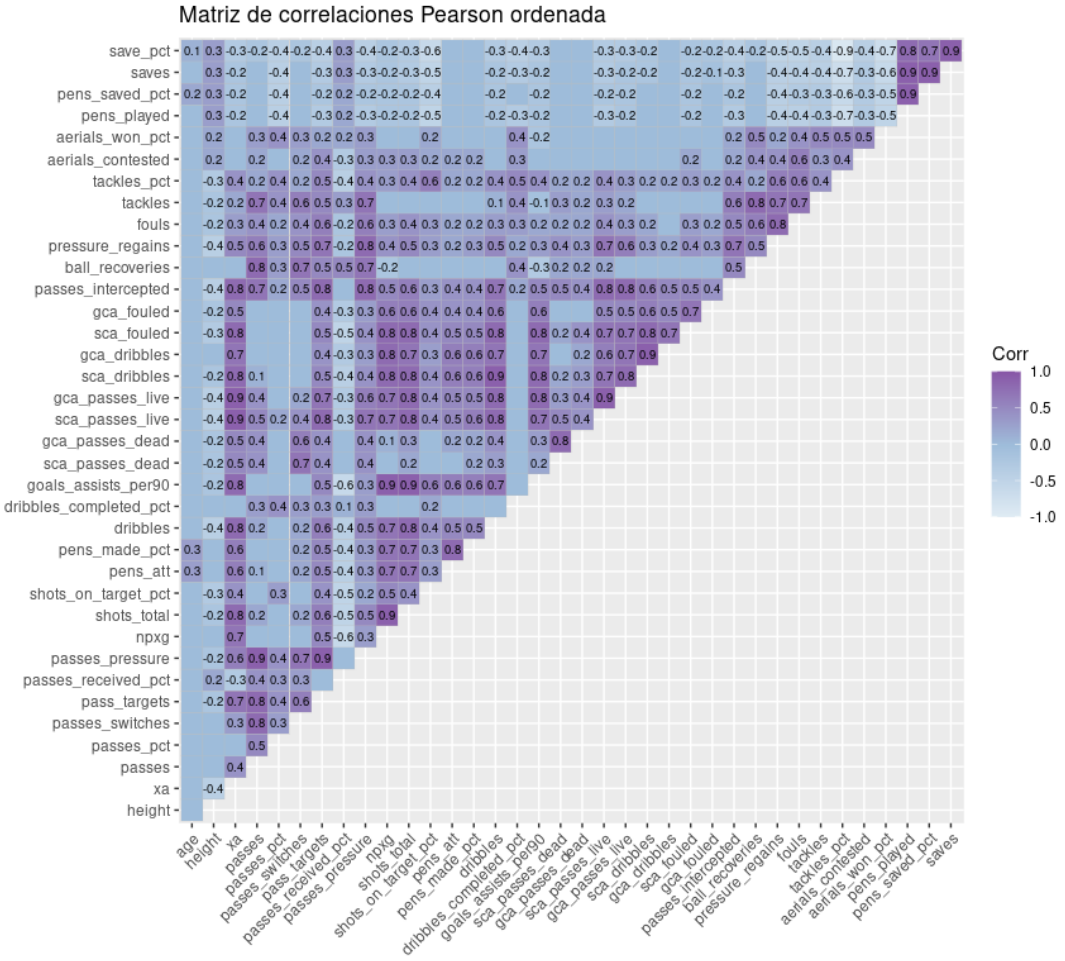


Figura 4.67: Matriz de correlaciones de *Pearson*

Se observan aún correlaciones altas, cercanas a 0.9 y 0.8, que quizás podrían indicarnos que algunas de las variables incluidas explica lo mismo que otras. Para la figuras 4.67 y 4.68, la variable *passes* tiene correlaciones altas con la cantidad de cambios de orientación en el juego, la cantidad de pases que se dirigen hacia el jugador, las ocasiones con el balon en juego que generan tanto tiros como goles

o la cantidad de balones recuperados. Las variables *xa* y *shots_ total* presentan varias correlaciones altas similares. Entre las estadísticas cuya correlación de *Pearson* es alta encontramos las ocasiones que generan tanto tiros como goles, la cantidad de pases interceptados, la suma de goles y asistencias en 90 minutos o los regates intentados. Esta última estadística también tiene correlación alta con la mayoría de estadísticas *sca* y *gca*, es decir, las relacionadas con la creación de acciones de tiro y gol respectivamente. Respecto a las variables relacionadas con los guardametas existe una alta correlación cruzada entre el porcentaje de penaltis parados, la cantidad de paradas y el porcentaje de ocasiones paradas.

Pero la razón de no seguir eliminando variables se basó en mantener la posibilidad de luego filtrar por algunas de ellas a la hora de buscar los mejores fichajes en un perfil de jugador concreto.

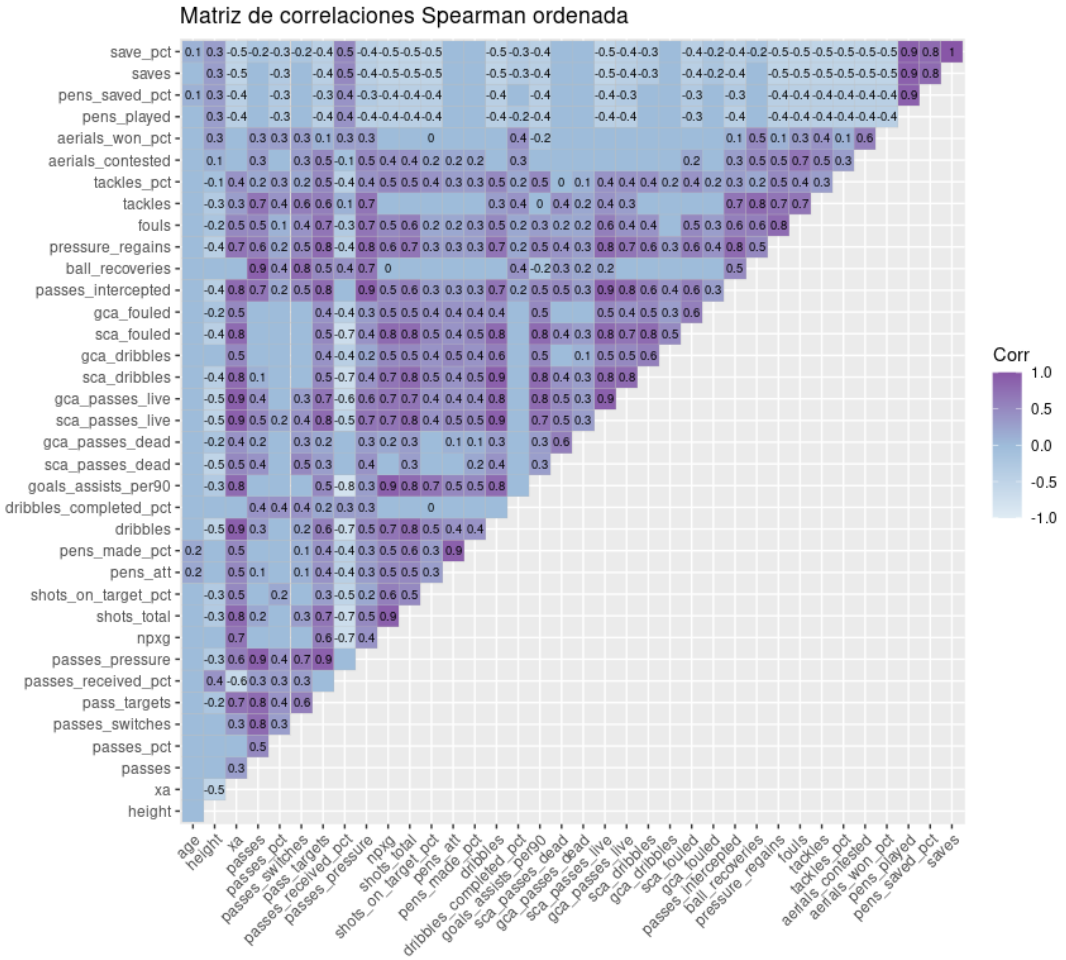


Figura 4.68: Matriz de correlaciones de *Spearman*

4.5. Análisis multivariante

Previo a realizar las distintas transformaciones de esta sección, debemos preparar los datos.

4.5.1. Estandarización de los datos

Dada la existencia de multiples escalas entre nuestras variables, realizamos una estandarización, es decir, forzar a que su media sea igual a 0 y su desviación típica a 1. Si no realizaramos dicha transformación las variables que presentaban mayor varianza tendrían mayor influencia en nuestra clasificación.

4.5.2. Análisis de componentes principales

Scree plot

Una vez aplicada la transformación debemos decidir cuántas componentes principales es adecuado escoger. Un gráfico que nos puede ayudar se muestra en la figura 4.69, llamado *scree plot*, el cual nos indica la proporción de variabilidad explicada por cada una de las componentes principales. Si escogierámos unicamente las dos primeras, sólo cubriríamos algo más del 54 % de la variabilidad, por tanto, hemos optado por recoger hasta un 73 % de variabilidad, seleccionando las 4 primeras componentes principales.

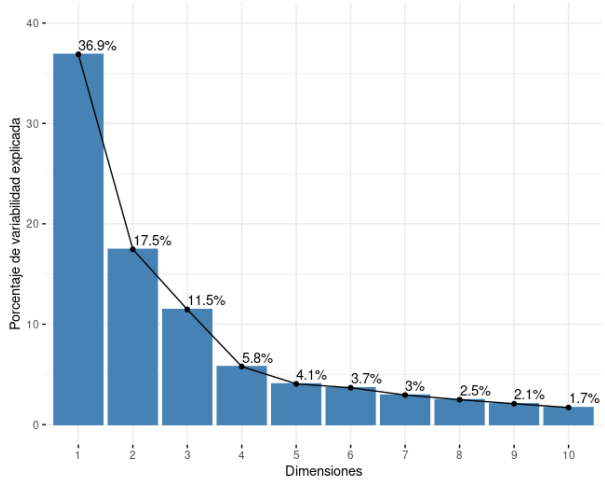


Figura 4.69: Contribución a la varianza explicada por cada componente principal

De cara mejorar la interpretación del análisis de las primeras componentes principales, trabajaremos con las dos primeras para ver una primera representación de los jugadores a entrenar y la influencia de cada una de las estadísticas

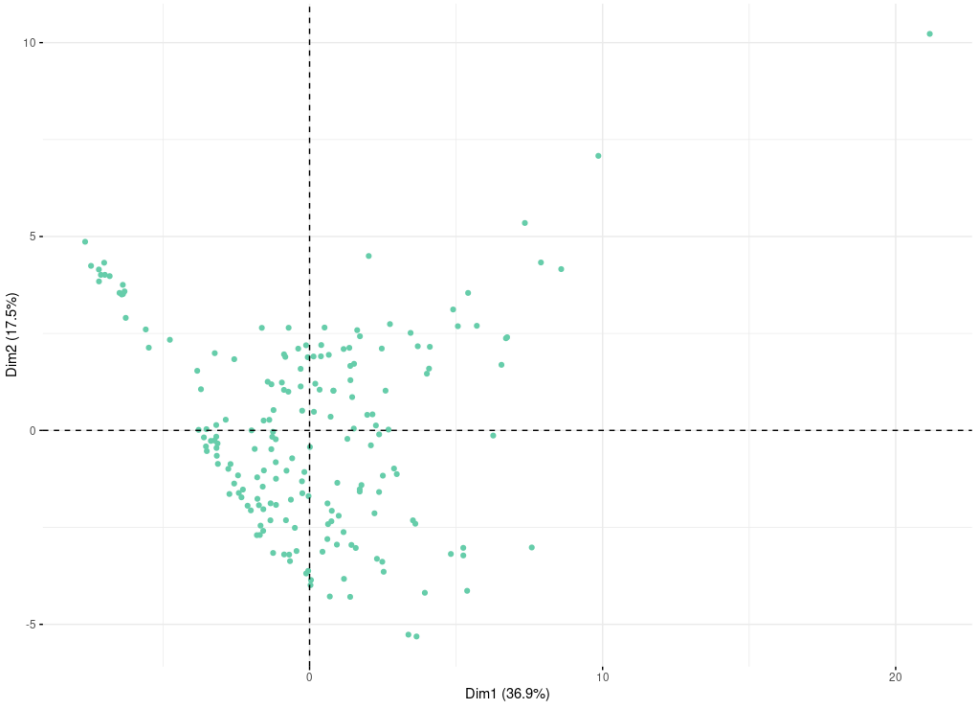


Figura 4.70: Representación 2D de las dos componentes principales

seleccionadas. Escogiendo los dos primeras componentes principales, ambas forman un plano, que constituye la mejor representación de nuestros datos en 2D. La representación de los 188 jugadores en base a sus coordenadas para esas componentes se muestra en la figura 4.70.

A partir de la figura 4.71 podemos sacar bastantes conclusiones sobre la figura 4.70. Comenzando por las variables que más peso tienen en estas componentes principales son las que aparecen con un color verde fuerte, algunas de ellas son: goles y asistencias esperadas, tiros totales, regates o los pases que generan ocasiones de tiro o gol. En el segundo cuadrante aparecen las estadísticas especificas de los porteros con una dirección similar, lo cual nos hace indicar la zona que ocuparán estos jugadores. En el cuarto cuadrante predominan las estadísticas defensivas, como recuperaciones o entradas realizadas, y cuanto más nos acercamos al primer cuadrante estas son más ofensivas, muchas de ellas relacionadas con los pases, ya que es en este cuadrante donde tendremos a los jugadores con mayor rendimiento en estadísticas de ataque.

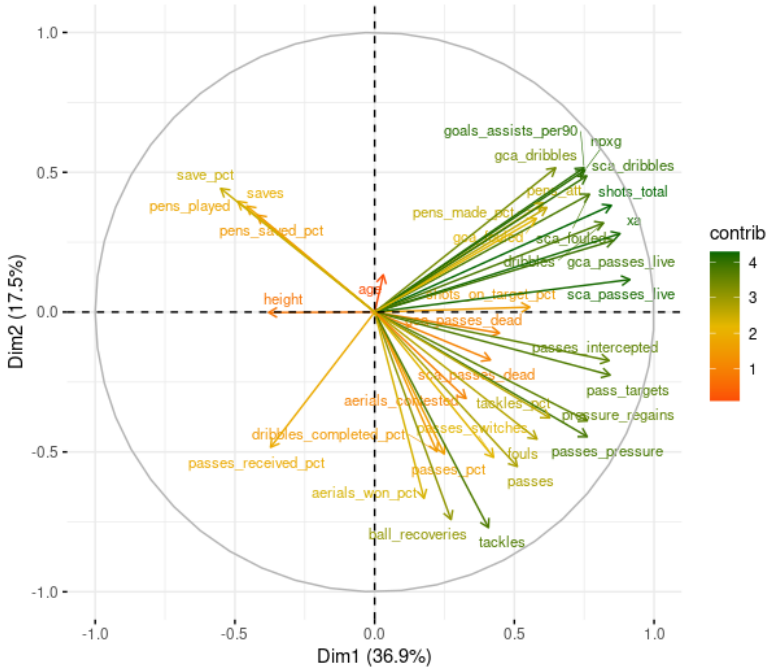


Figura 4.71: Contribución de cada variables respecto a las dos primeras componentes principales

4.6. Creación del conjunto de entrenamiento y test

Con el objetivo de realizar un análisis a posteriori sobre los *cluster* resultantes se ha decidido reservar un 20 % de los 188 jugadores que participaron en las tres temporadas en estudio, es decir, nuestro conjunto de prueba estará formado por 35 jugadores y el 80 % restante será el conjunto de aprendizaje con 153 jugadores. Se tendrá especial cuidado en que los 9 jugadores que cambiaron de equipo en una misma temporada sus dos muestras caigan en el conjunto de entrenamiento.

Veamos la distribución de posiciones de los jugadores de nuestro conjunto de datos:

Temporada 2017-18				
Posición	Porteros	Defensas	Mediocentros	Delanteros
Cantidad	16	75	79	70
Temporada 2018-19				
Posición	Porteros	Defensas	Mediocentros	Delanteros
Cantidad	16	78	73	64
Temporada 2019-20				
Posición	Porteros	Defensas	Mediocentros	Delanteros
Cantidad	16	75	76	64

Tabla 4.54: Número de jugadores por posición en cada temporada

Hay que indicar que la suma total de la fila, que indica la cantidad de jugadores que juegan en esa posición para cada temporada, resulta mayor que 188, ya que en el fútbol es bastante común los jugadores que pueden jugar en distintas posiciones del juego, a los que se denomina como polivalentes. Además las cantidades no coinciden entre temporadas debido a que también es habitual cambiar tu posición a lo largo de tu carrera o incluso ampliar tu rango de acción en el campo. Para los porteros suele ser más dificil su reconversión por eso se mantiene el número constante durante las 3 temporadas.

Estos conjuntos se han escogido de manera aleatoria estableciendo la semilla 5682, mediante la función *set_ seed()*. Las distribuciones para las 4 posiciones generales que tenemos se presentan en las figuras 4.72 y 4.73.

Las proporciones de los jugadores en función de su posición para los dos conjuntos son las siguientes:

	Porteros	Defensas	Mediocentros	Delanteros
Entrenamiento	6.35 %	31.75 %	32.8 %	29.1 %
Validación	9.52 %	35.71 %	33.33 %	21.43 %

Tabla 4.55: Proporción de jugadores diferenciados por posición en cada conjunto

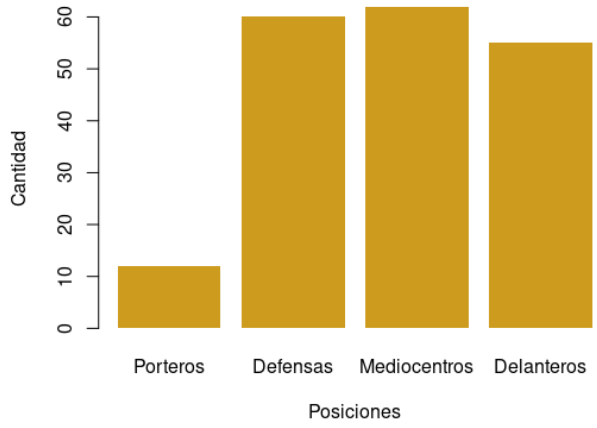


Figura 4.72: Distribución por posiciones en cada temporada para el conjunto de entrenamiento

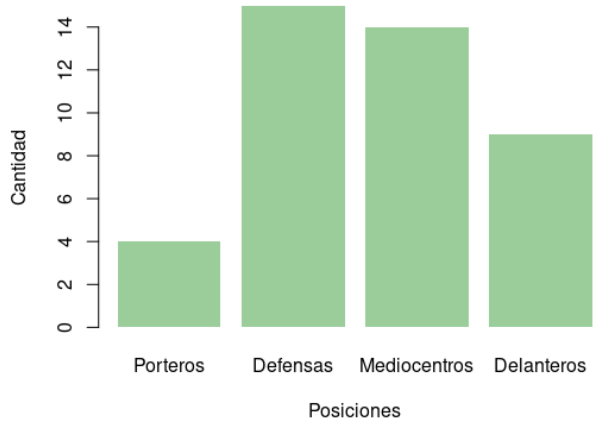


Figura 4.73: Distribución por posiciones en cada temporada para el conjunto de test

Capítulo 5

Uso de técnicas de aprendizaje

5.1. Búsqueda del k óptimo

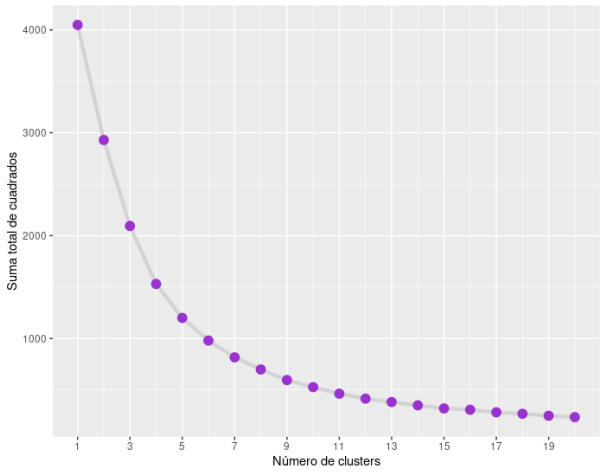


Figura 5.1: Método del codo hasta 20 *clusters*

Basandonos en la figura 5.1 podemos decir que el número de grupos en nuestro conjunto de datos es 6, ya que es ahí donde se situa el çodo", y es a partir de aquí donde la suma de cuadrados en el cluster comienza a decrecer linealmente.

Dividir en 7 *cluster* también parece coherente, y dado que de partida tenemos 4 posiciones generales, tiene sentido poder buscar cuantos más perfiles posibles. Por tanto, asimismo probaremos con $k = 7$.

5.2. Resultados

5.2.1. K-medias

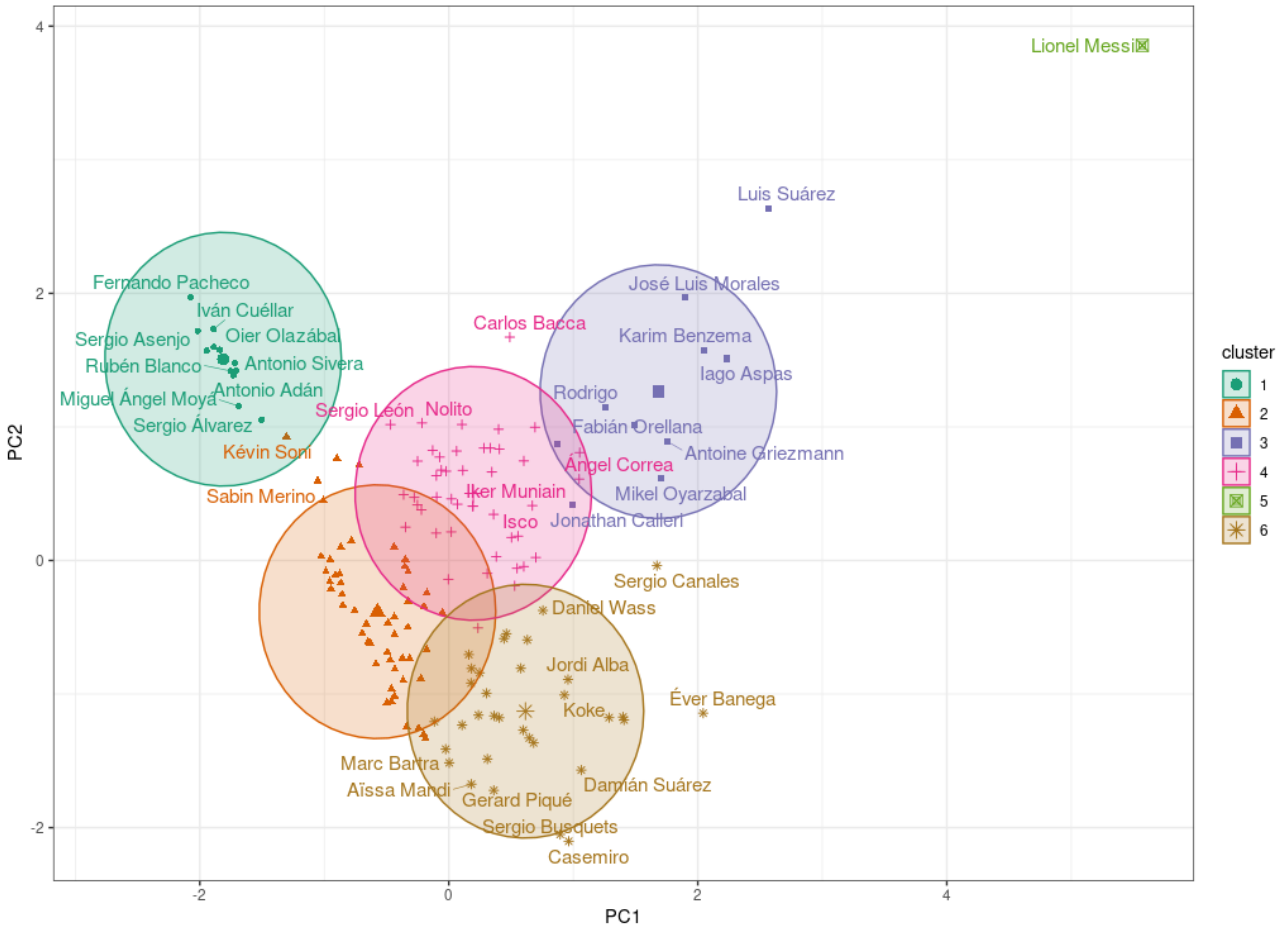


Figura 5.2: 6-medias

A primera vista podemos identificar un cluster formado por una única muestra, no sorprende que ese jugador tan lejano al resto sea Lionel Messi. Luego podemos ver que el primer cluster lo forman unicamente porteros. Después el cluster 3 le constituyen principalmente delanteros de alto nivel como Luis Suarez y Karim Benzema, o de muy alta influencia en sus equipos como Iago Aspas o Rodrigo. En el último cluster podemos ver una mezcla de jugadores con gran actitud defensiva como Piqué, Busquets o Casemiro, pero tambien tenemos jugadores con gran creación de juego y grandes asistentes como Koke, Jordi Alba

o Ever Banega.

Dado el alto solapamiento de nombres de la figura 5.2, en la siguiente página presentaremos los *clusters* de manera individual para ver más concretamente que jugadores forman cada grupo, y así mejorar la interpretación que podamos dar a cada uno de ellos.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
12	52	10	45	1	33

Tabla 5.1: Tabla resumen para $k = 6$

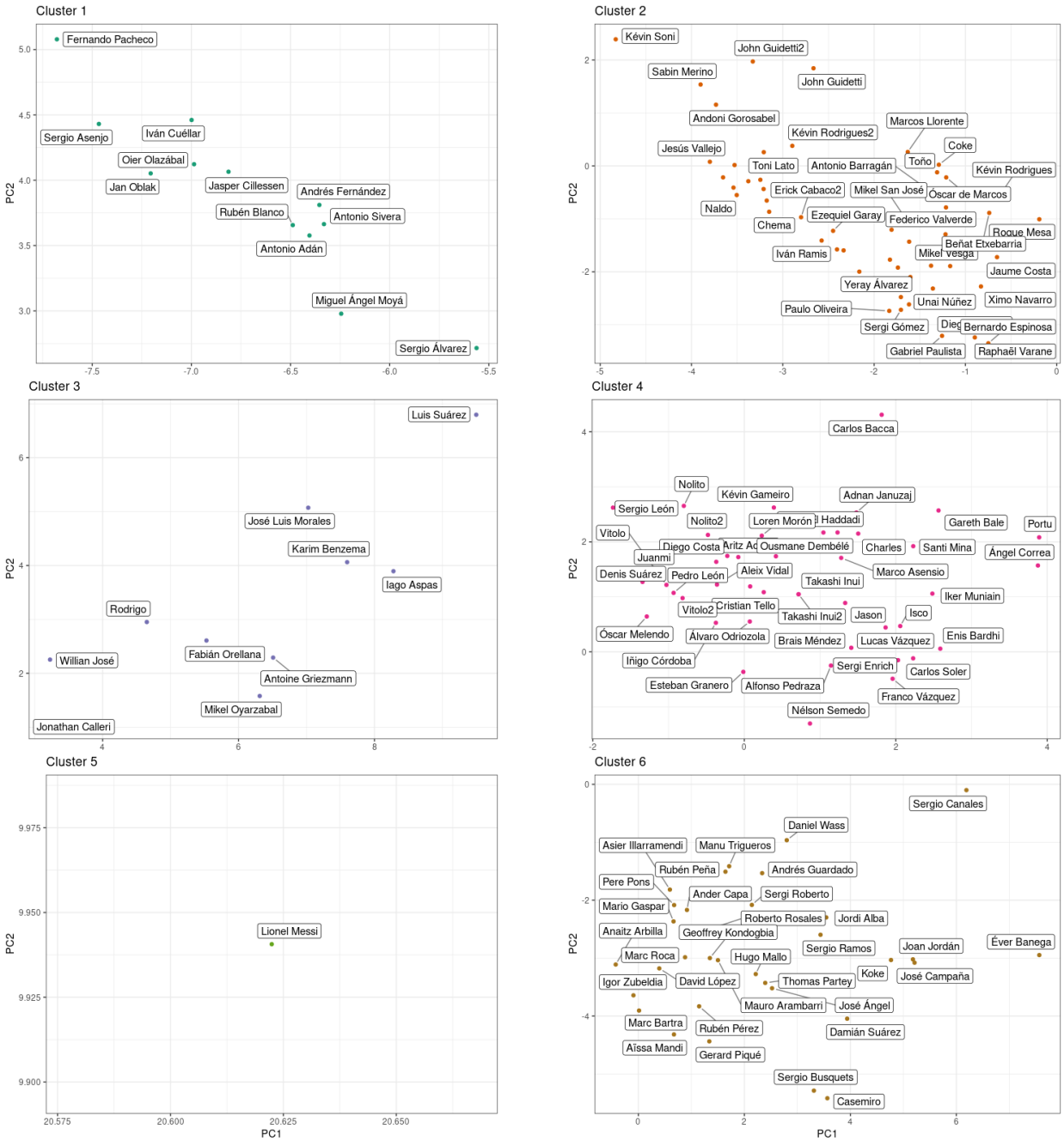


Figura 5.3: 6 *clusters individualizados*

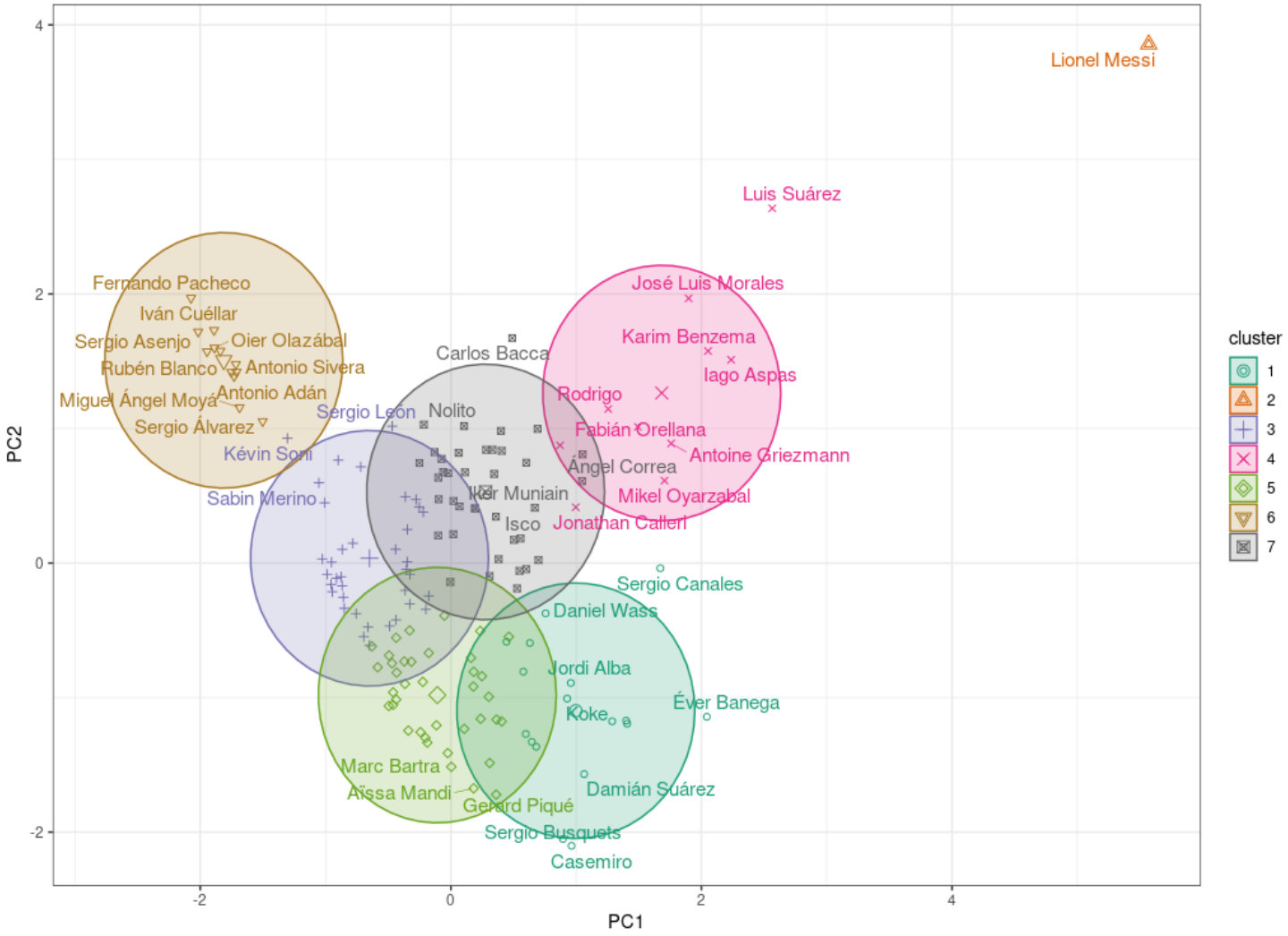


Figura 5.4: 7-medias

No apreciamos grandes en cambios salvo en el cluster 6 anterior, que ahora lo podemos encontrar dividido en 2. Conseguimos que esa mezcla de jugadores ofensivos y defensivos se vea eliminada, teniendo ahora a los defensores en el cluster 5 y a los asistentes y creadores de juego en el cluster 1.

Comparando con el *clustering* anterior, Messi se mantiene en un cluster diferenciado, los porteros siguen formando un único grupo y el tercer cluster anterior es idéntido al cuarto cluster de la figura 5.5. Las diferencias entre cluster 4 obtenido en la figura 5.2 y el cluster 7 de ahora, son mínimas, Nélson Semedo,

ahora forma parte del cluster más defensivo, el 5, estando cerca de pivotes como Illarramendi o Pere Pons. El portugués del Barcelona, que militó justo durante las 3 temporadas de estudio, destacaba por su proyección ofensiva y desborde, ya que su fichaje era con la intención de suplir al laureado Dani Alves. También 6 jugadores han dejado de formar parte del cluster 4 para irse al nuevo cluster 3, estos son: Vitolo (con ambas muestras), Pedro León, Oscar Melendo, Sergio León y Denis Suarez. Estos nombres son nombres de jugadores con calidad, la cual se ha visto en alguna de sus temporadas en la Liga, pero han sido muy efímeros. Además si nos fijamos en Pedro León, uno de los jugadores culpables del gran rendimiento del Eibar en la temporada 16-17, vemos como dos lesiones, una en la temporada 17-18 y otra en la temporada 18-19, le llevaron a perderse 25 y 26 partidos respectivamente, haciendo temporadas mediocres. Vitolo sufrió situaciones parecidas llegandose a perder casi 30 partidos en las 3 temporadas de estudio.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
17	1	37	10	38	12	38

Tabla 5.2: Tabla resumen para $k = 7$

En la figura 5.5, se ha omitido el primer *cluster*, ya que es el que forma unicamente Messi y no aportaba más información.

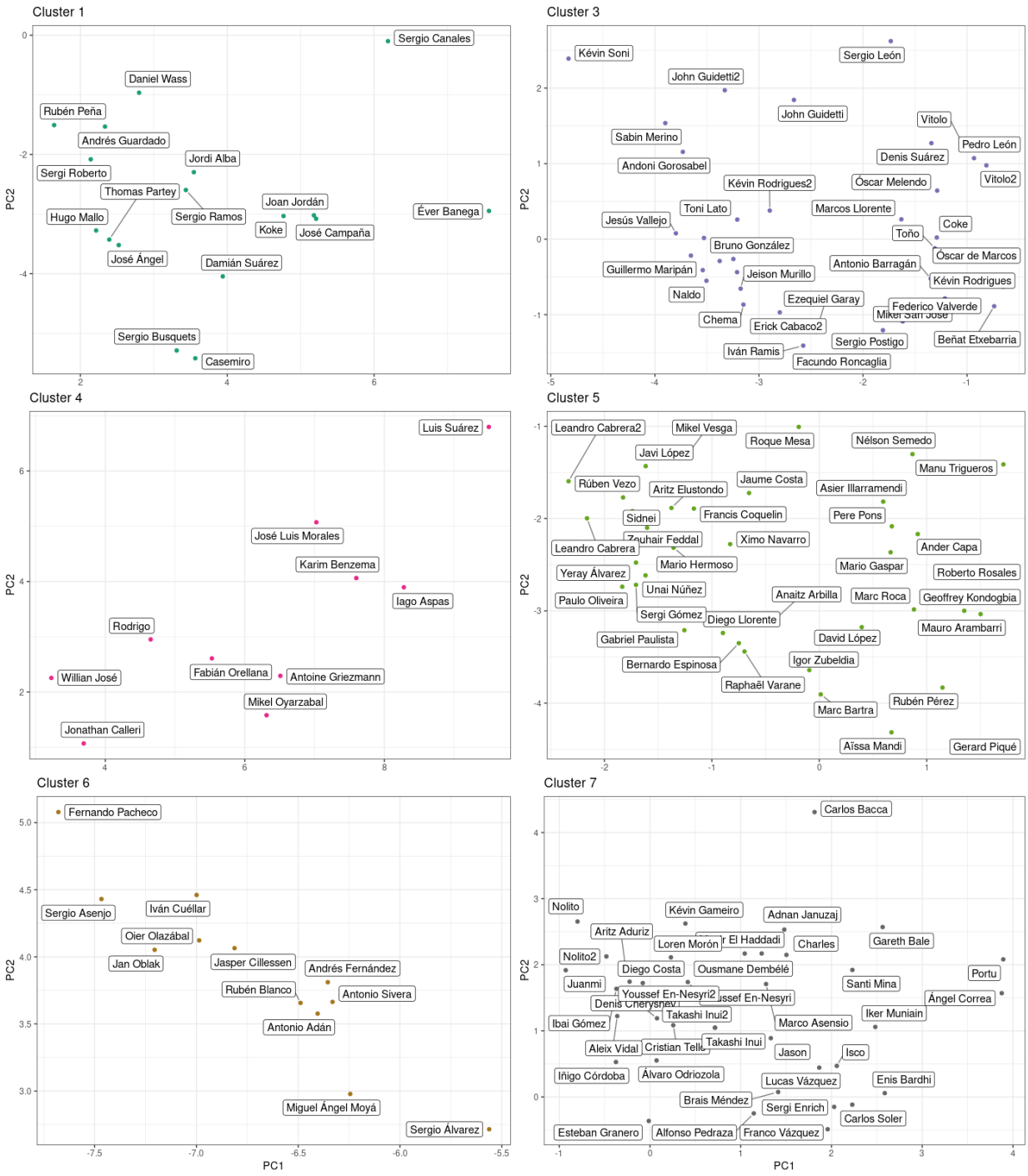


Figura 5.5: 7 *clusters individualizados*

5.2.2. Jerárquico

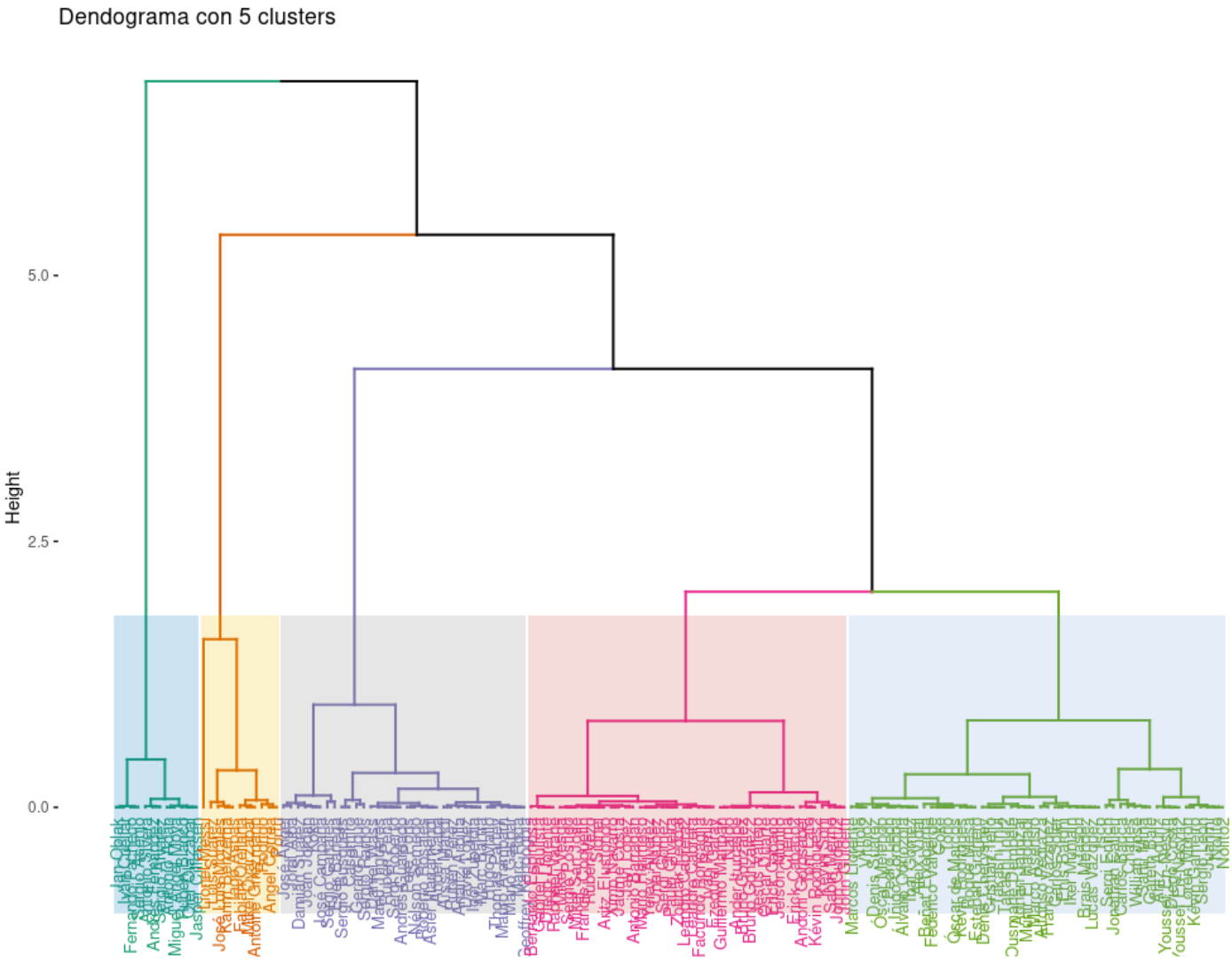


Figura 5.6: Jerárquico con 5 grupos

La elección de 5 *clusters*, se debe a que con 6 la única diferencia era que Leo Messi pasaba a ser un *cluster*, por tanto, con esta división vemos que a jugadores se asemeja más Messi, y como era de esperar se encuentra en el grupo de delanteros con un gran rendimiento como son Iago Aspas, Portu, Luis Suarez, entre otros.

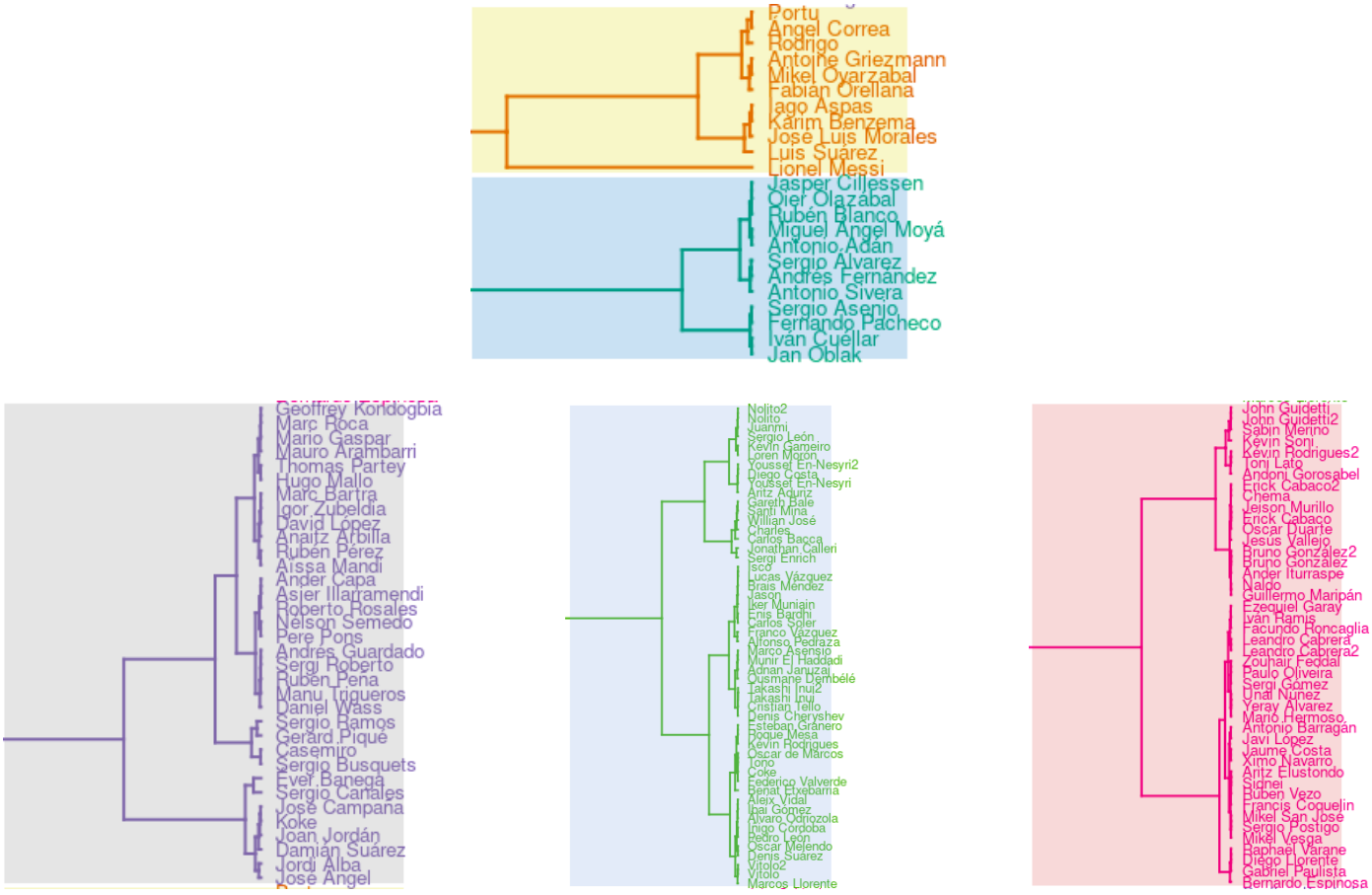


Figura 5.7: Jerárquico *clusters individualizados*

Si pasamos a $k = 7$, el cluster que se divide es el morado, algo que podiamos intuir si observamos el dendograma de la figura 5.6, ya que el siguiente corte después del de Messi del *cluster* naranja es el situado en torno a la altura de 1. La división sería la que se muestra en la figura 5.8.

En el *cluster* rosa tenemos a los creadores de juego que se ubicaban en la parte inferior y ligeramente a la derecha de la figura 5.2, destaca que entre ellos se encuentren 2 laterales, Damián Suarez y Jose Ángel, del Getafe y Eibar, respectivamente, pero es que ambos promediaron casi 4 asistencias en los 3 años, siendo más regular el del conjunto madrileño. A diferencia del *cluster* 7 obtenido por el algoritmo k-media con $k = 7$, ubicamos en un grupo de jugadores destacados en el ámbito defensivo a Sergio Ramos, Hugo Mallo, Andrés Guardado o Sergio Busquets entre otros.

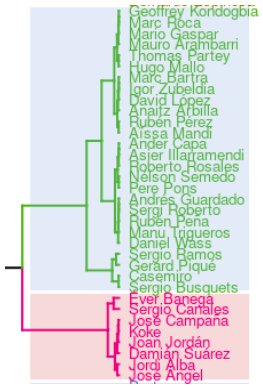


Figura 5.8: División del cluster morado en 2

5.2.3. DBSCAN

El método propuesto para ajustar el parámetro ϵ, consiste en calcular las distancias de los k vecinos más cercanos en una matriz de puntos. Nos hemos apoyado en la función *kNNdistplot()* de la librería *dbscan*, que calcula la media de las distancias de cada punto con sus k vecinos, argumento especificado por el usuario, que corresponde con el otro parámetro a ajustar, el mínimo número de puntos que forman un cluster. Como la dimensión de nuestros datos es 4, el número mínimo de puntos que pueden conformar un cluster debe ser mayor o igual al nº de dimensiones + 1, en nuestro caso, será 5. Por tanto, si nos fijamos en la figura 5.9, y seguimos un método similar al del codo, el valor óptimo de ϵ sera cercano a 2.8. Este es el punto de inflexión, aproximado visualmente, de la curva pintada.

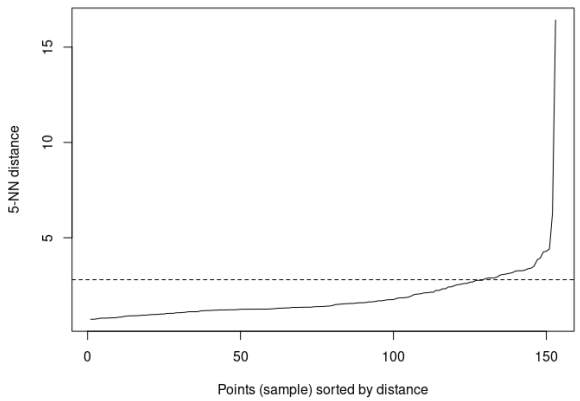


Figura 5.9: Distribución de distancias medias para 5 vecinos más próximos

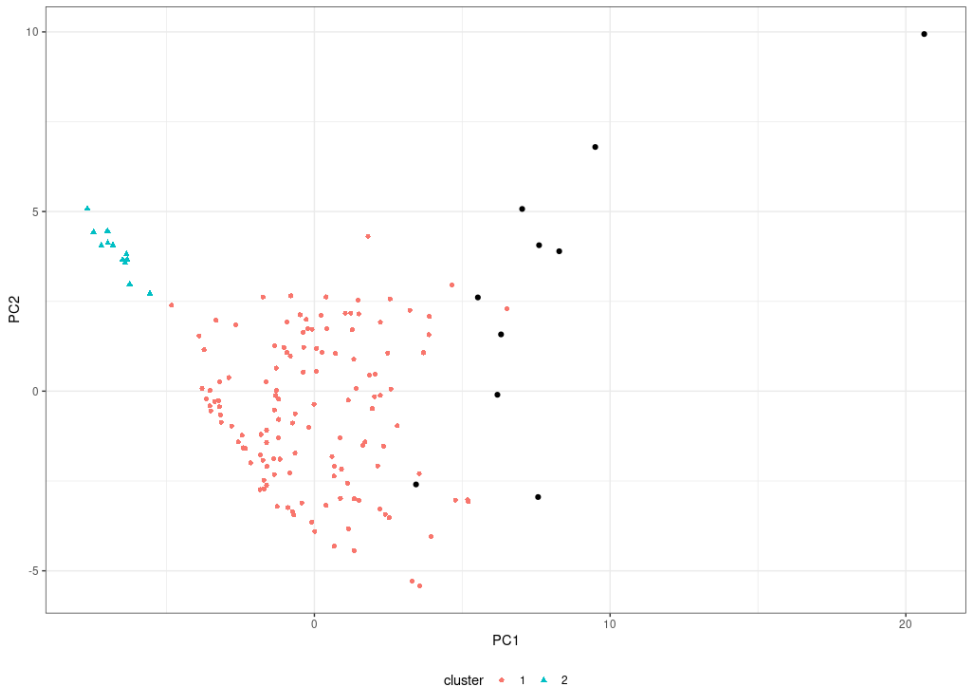


Figura 5.10: DBSCAN con $\epsilon = 2.8$ y minPts $= 3$

Los resultados no son nada buenos, obtenemos unicamente 2 *clusters*. Al menos consigue diferenciar los porteros del resto de jugadores, clasificando 10 jugadores como **puntos de ruido** (los jugadores que estan pintados con color negro), esto es una ventaja de este algoritmo, que jugadores como Leo Messi que nos obligaba a fijar un grupo para él solo podemos clasificarlos como ruido. Dentro de este grupo de jugadores estan: Luis Suares, Jose Luis Morales, Karim Benzema, Éver Banega, Sergio Canales, Fabián Orellana, Mikel Oyarzabal, Iago Aspas y Sergio Ramos.

El motivo del mal rendimiento de dicho algoritmo puede encontrarse en su mal funcionamiento con grupos de densidades variables como es el caso de nuestro conjunto de datos.

5.3. Clustering en nuevo conjunto

Durante este apartado se probará a aplicar los algoritmos más favorables en la sección anterior a un conjunto de entrenamiento más reducido. Las muestras eliminadas son la de Lionel Messi, por el hecho de que en los resultados conforme él solo un *cluster*, y las respectivas a los 12 porteros que teníamos, ya que para

ningún algoritmo se ha observado un comportamiento diferente referente a estas muestras, de forma que estos siempre han formado su cluster personalizado.

Por tanto, el conjunto de datos de entrenamiento lo conforman 140 futbolistas, y el conjunto de validación 31 jugadores, de los cuales analizaremos un conjunto de 33 estadísticas al quitar las especificas de porteros explicadas en la sección 4.3.6.

5.3.1. Búsqueda del *k* óptimo

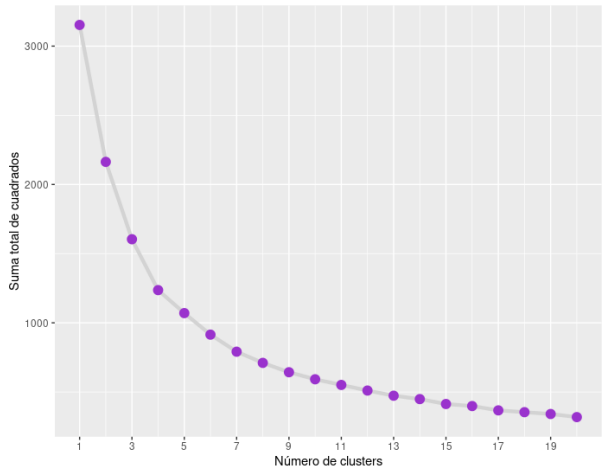


Figura 5.11: Método del codo hasta 20 *clusters*

En la figura 5.11 vemos el método del codo para nuestro nuevo conjunto de datos reducido de nuevo a las 4 primeras componentes principales. En esta ocasión el codo no es tan claro, quizás el k que parece más óptimo sería 6, teniendo en cuenta que el mayor k probado en la sección anterior fue 7 y que hemos retirado 2 clusters de nuestro conjunto de datos tiene sentido que probemos con k mayor que 5, para así evitar obtener *clusters* similares a los anteriores. Finalmente se ha decidido probar en un rango de k igual a [5,7], de este modo comprobaremos la semejanza con los anteriores clusters obtenidos y estudiaremos las ventajas que nos ofrece eliminar dichas muestras para buscar más perfiles.

5.3.2. K-medias

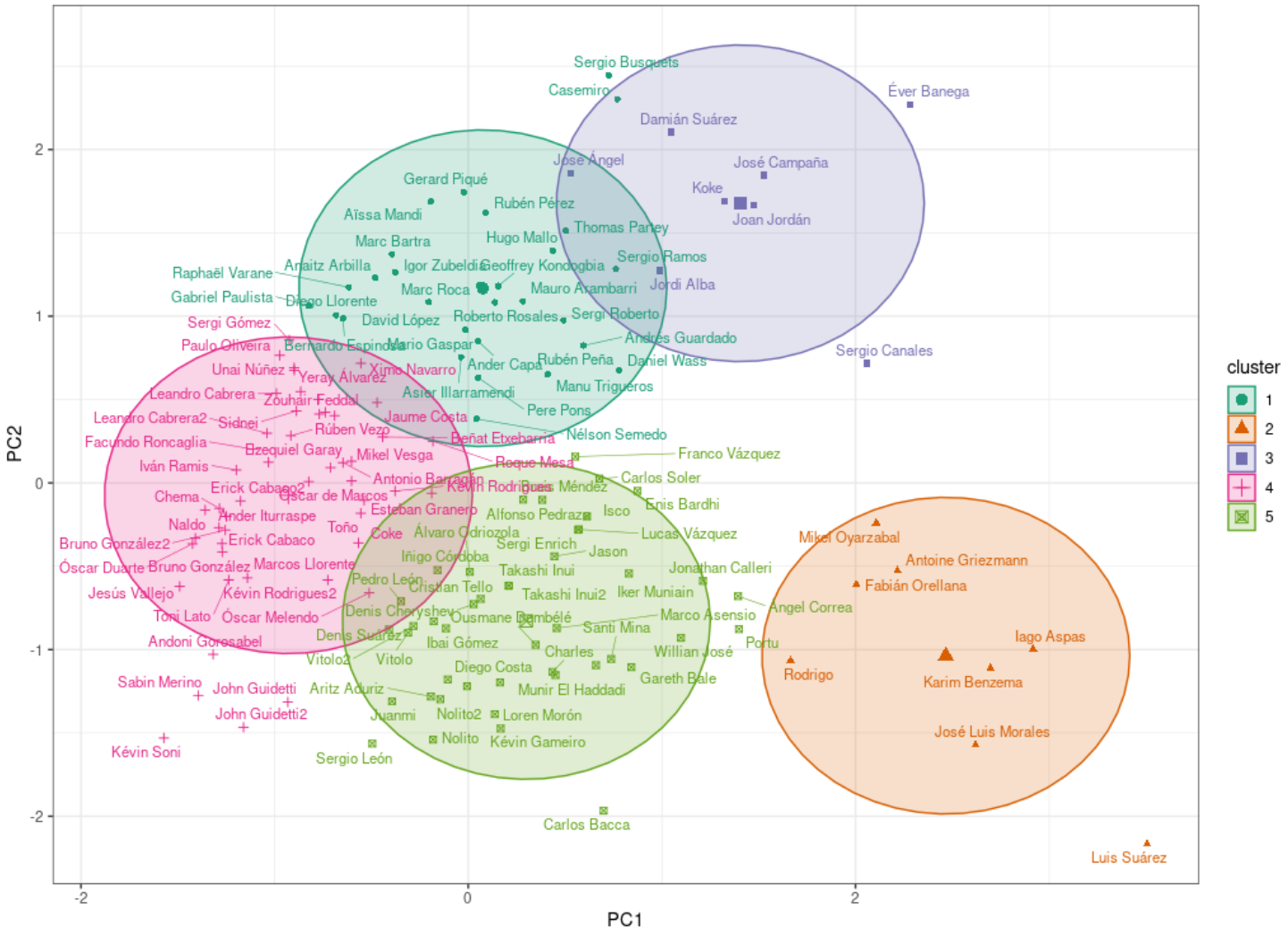


Figura 5.12: Resultados clustering 5-medias

Existen pocos cambios respecto al clustering presentado en la figura 5.4. Tenemos un cluster de jugadores con gran rendimiento ofensivo, en este caso pintado de naranja, respecto al obtenido en el *cluster* 4 de la figura 5.5 se caen William José y Jonathan Calleri, que ahora forman parte del *cluster* 5. De este grupo de jugadores de gran calidad vuelven a formar parte los que pasaron al *cluster* 3 de la figura 5.4. El *cluster* 3 de la figura 5.12, el cual

consideramos que agrupa jugadores con gran capacidad de creación de juego y con gran influencia en sus equipos es ahora más reducido, ya que sacamos a jugadores como Busquets, Casemiro o Daniel Wass, que pasan a formar parte de un cluster más defensivo, el grupo 1 de la figura 5.12.

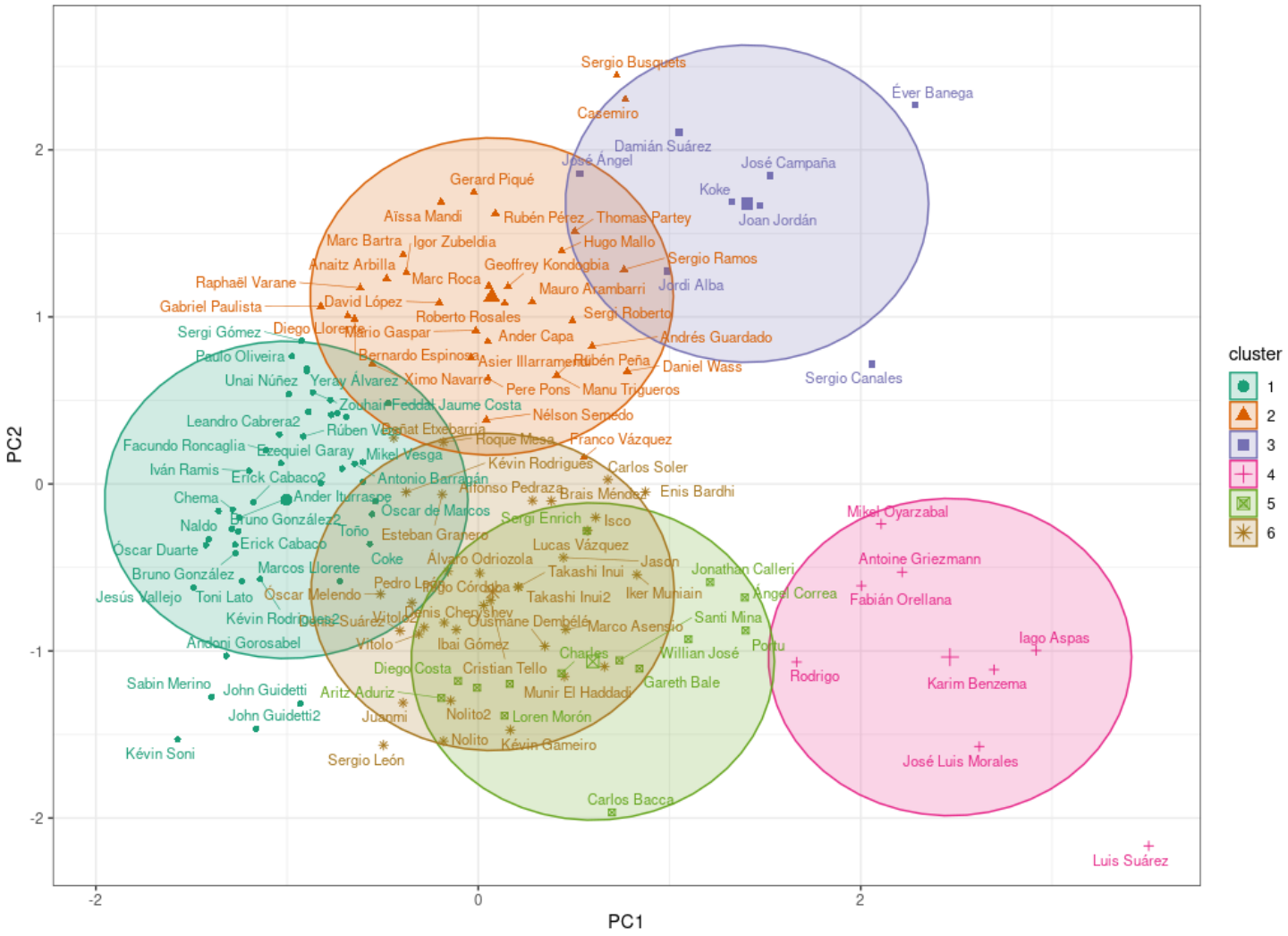


Figura 5.13: Resultados clustering 6-medias

Un cambio interesante se produce en la figura 5.13, se divide el *cluster* 5 anterior en 2, como resultado obtenemos un grupo marrón que lo forman principalmente mediocentros ofensivos como Marco Asensio, Tello, Muniain o Carlos

Soler, y un grupo ver en el encontramos sobre todo delanteros muy conocidos como Diego Costa, Carlos Bacca, Portu, o Charles. Del nuevo *cluster* 6 incluimos a Esteban Granero y Oscar Melendo que habían pasado a formar parte del grupo menos destacado, el rosa de la figura 5.12.

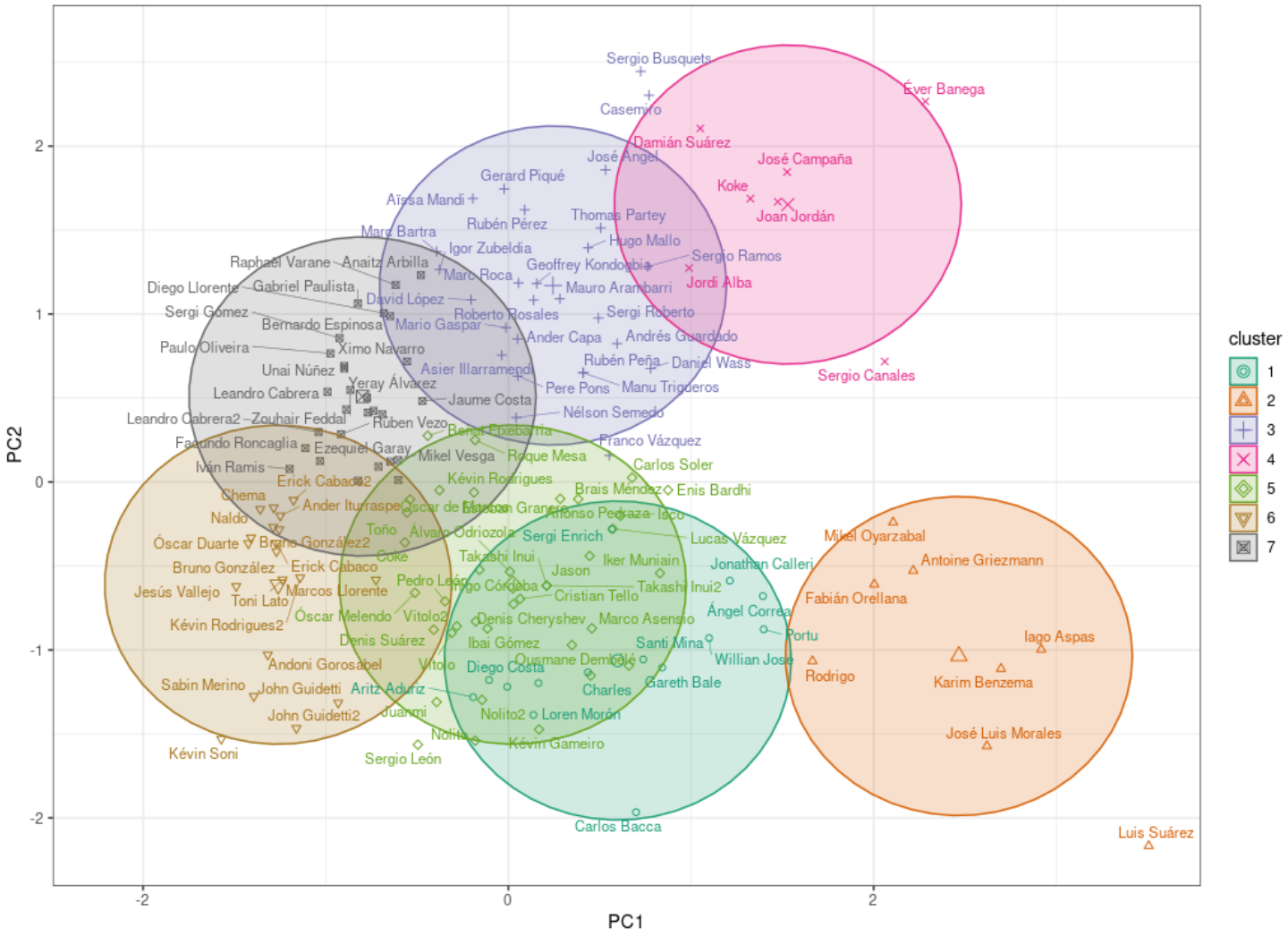


Figura 5.14: Resultados clustering 7-medias

En la figura 5.14 también se presenta un cambio interesante, y es que se produce una división del grupo de jugadores con coordenadas más bajas en las primeras componentes principales, y es la primera división que se produce sobre

este. Obtenemos un *cluster* 6 en el cual ubicamos jugadores defensivos en su parte inferior como Óscar Duarte, Erick Cabaco o Iturraspe, lo cual concuerda con que el nuevo grupo 7 lo consituyen principalmente defensas centrales como Garay, Yeray o Ruben Vezo. El resto de *clusters* obtenidos los forman los mismos jugadores que hemos ido comentando sobre las figuras 5.12 y 5.13.

Algunas representaciones 3D de estos modelos finales se pueden encontrar en `https://chart-studio.plotly.com/~mariogt09/#/`.

5.4. Validación

En esta sección vamos a validar el modelo de k-medias con k = 7, ajustado para el conjunto de datos sin los porteros y sin Messi. Para esto vamos a proceder de dos formas diferentes, evaluando internamente cada *cluster*, mediante el coeficiente de *silhouette*, y externamente introduciendo las muestras que reservamos para test en la sección 4.5.1. Para este último método hay que mencionar que es necesario conocer información previa sobre las características de los jugadores del conjunto de validación, para así entender y valorar el ajuste que haga nuestro modelo sobre ellos. Es por esto, que el estudiante valorará de una manera objetiva pero con falta de información al no ser un experto.

5.4.1. Evaluación interna

En la figura 5.15 observamos el coeficiente de *silhouette* para cada unos de los 7 clusters formados, encontramos más parecidos entre los jugadores que forman los *clusters* 2 y 7, es decir, entre el grupo que hemos detectado como defensas centrales puros y los delanteros con mayor rendimiento ofensivo durante los 3 temporadas de estudio. La media de dicho coeficiente sobre todos los grupos formados es 0.46, valor que quizás sería pequeño en otro tipo de estudio, donde se busque semejanza entre miembros del propio *cluster* y diferencias con el resto de grupos, pero en nuestro caso es un valor correcto que nos permite sacar bastantes perfiles entre los jugadores estudiados.

Modelo	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Media
k = 5	0.42	0.66	0.54	0.59	0.33	-	-	0.47
k = 6	0.55	0.36	0.54	0.62	0.27	0.53	-	0.48
k = 7	0.29	0.62	0.26	0.55	0.44	0.57	0.62	0.46

Tabla 5.3: Coeficientes de silhouette por *cluster* en los últimos 3 modelos

En la tabla 5.3 se presenta una compación de los coeficientes de silhouette de los 3 últimos modelos aplicando *k-medias* con difrentes parámetros de *k*. Vemos que los ajustes son similares ya que sus coeficiente de *silhoette* medio es casi igual. Por tanto, ya que el objetivo del trabajo se centra en buscar perfiles de jugadores vamos fijarnos en el modelo de $k = 7$ que tendrá mayor interes a la hora de evaluar el conjunto de test sobre él.

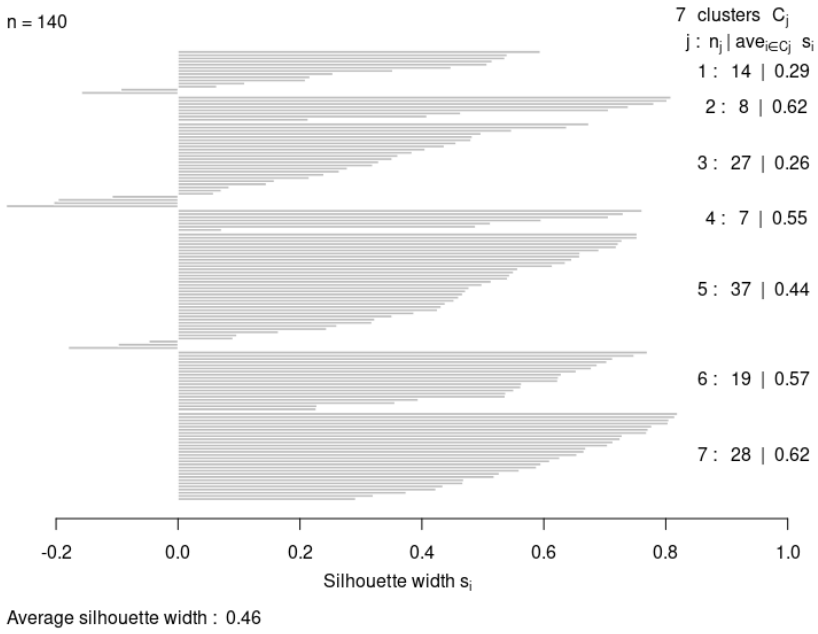


Figura 5.15: Coeficiente de silhouete por individuo de cada cluster

5.4.2. Evaluación externa

Después de calcular los valores para las dos primeras componentes principales para nuestro conjunto de test y calcular con ellas el cluster al que se encuentran más cercanos, el resultado es el mostrado en la figura 5.16. La asignación es bastante coherente, tenemos a dos laterales con gran capacidad ofensiva como son Marcelo y Jesús Navas, junto a un jugador con gran responsabilidad en el juego de su equipo, como es Toni Kroos, formando parte del cluster 4 el cual lo rellenaban jugadores con influencia en el campo y con visión de juego. Djené y Lenglet, dos centrales con muy buen rendimiento defensivo se colocan en el cluster 3, cercanos a jugadores como Gerard Piqué. En el grupo de los defensas centrales puros, el *cluster* 7, tenemos a jugadores como Umtiti, Sergio Escudero o Raul Navas. Respecto a los grupos más ofensivos, tenemos a Borja Mayoral, Piatti o Pione Sisto en el *cluster* 1 que lo formaban delanteros de gran renombre. Y finalmente, en el grupo naranja, formando parte de los jugadores con gran rendimiento ofensivo son: Gerard Moreno, Iñaki Williams, Jorge Molina o Maxi Gómez.

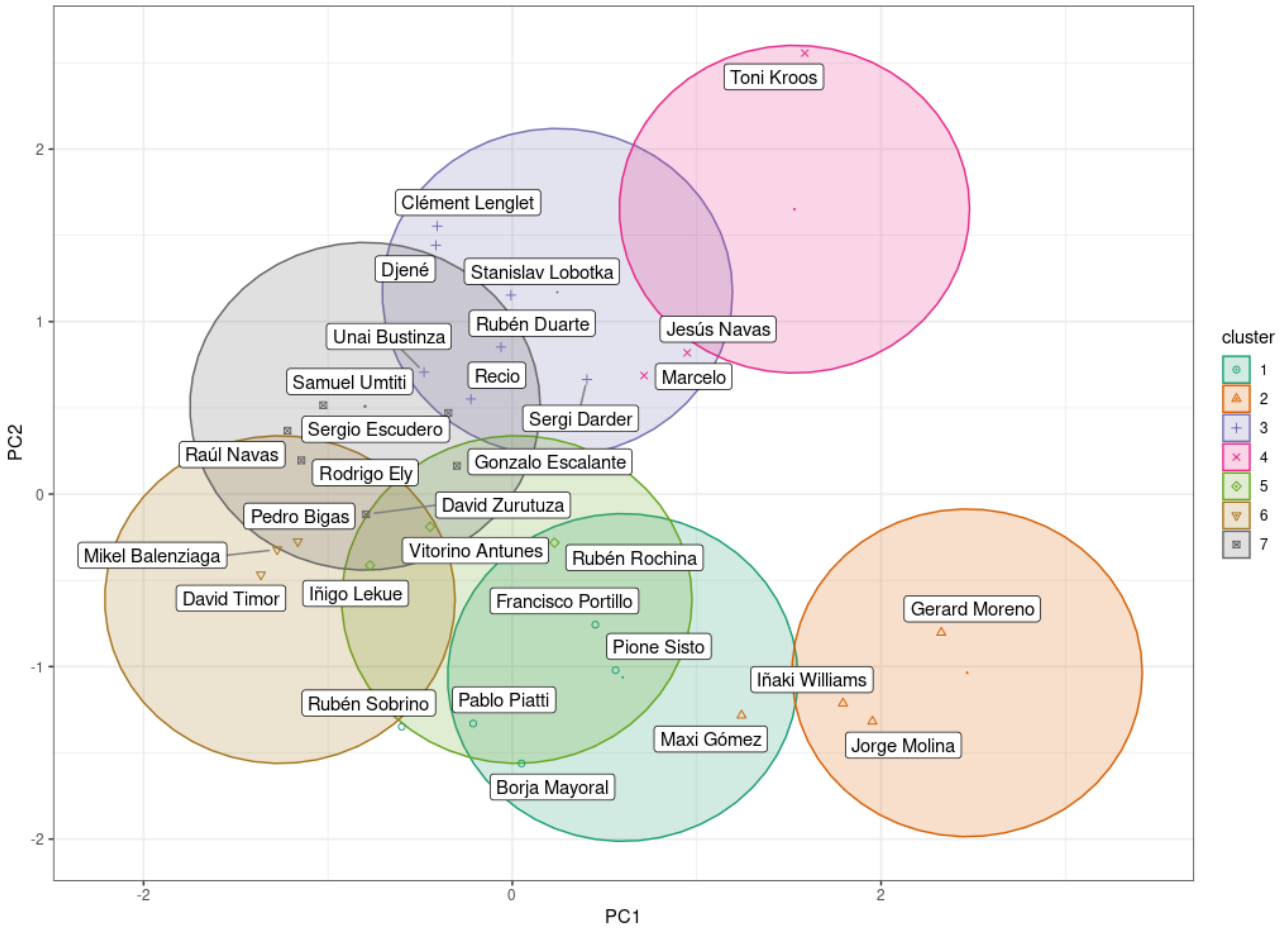


Figura 5.16: Conjunto de test en el modelo de la figura 5.14

Capítulo 6

Conclusiones

6.1. Conclusiones del trabajo

En este Trabajo de Fin de Grado se ha buscado perfilar a los diferentes jugadores de La Liga en las temporadas 17-18 hasta la 19-20, utilizando para ello los datos reclutados en Kaggle y disponibles en [9], elegidos después de estudiar las diferentes alternativas del mercado, cumpliendo uno de los objetivos fijados al comienzo del trabajo.

Sobre este *dataset* se ha realizado un análisis exhaustivo de varias de las variables que lo formaban, siendo éstas elegidas con ayuda de trabajos previos, para después discriminar cuántas de ellas eran útiles para el cometido del trabajo, mediante distintas técnicas estudiadas, logrando así el siguiente objetivo propuesto.

A la vista de los resultados comentados durante la sección anterior, se ha generado distintos perfiles de jugadores en base a sus estadísticas de rendimiento, siendo estos un punto de partida para la herramienta que se propondrá en el siguiente capítulo, alcanzando el último objetivo y dando un cierre lógico al trabajo. Estos perfiles han sido analizados por el estudiante, si este fuera un trabajo real y con gran amplitud sería necesario recurrir a un experto que estudiara más en concreto cada perfil obtenido.

Existen varias conclusiones que podemos sacar respecto a los resultados que hemos obtenido, una de ellas es sobre la influencia que tiene el parámetro k en el algoritmo k-medias, ya que cuanto menor era el valor de éste los cluster obtenidos se acercaban más hacia las posiciones más generales en el mundo del fútbol que son: portero, defensa, mediocentro y delantero. En cambio, cuanto más grande ha sido k nos ha permitido ver como se dividian grupos que parecían fijos en un primer momento. Por tanto, respecto a este parámetro podemos concluir que cuánto menor sea k mayor será el número de habilidades o aspectos del juego en los que destacarán los jugadores que formen los *clusters*. Mientras que cuánto mayor sea el número de grupos preseleccionado lograremos *clusters* más específicos, alejándonos de la idea de jugadores "multiposición".

Respecto al peso de las estadísticas actualizadas podemos destacar la cantidad de entradas realizadas, de tiros lanzados, de balones recuperados, de pases que sus compañeros les envían, los modelos de goles y asistencias esperados, o la cantidad de goles más asistencias por partido.

Capítulo 7

Trabajos futuros

7.1. Implementación de la aplicación

En esta sección se presentará una propuesta de aplicación, que podría realizarse partiendo de este trabajo, de cara a poder servir como guía por si alguien decidiera realizar su implementación.

7.1.1. Requisitos funcionales

A continuación, se lista una serie de requisitos para la consecucción de un versión básica de la aplicación:

- El sistema debera identificar al usuario.
- El sistema deberá permitir cargar diferentes conjuntos de datos de estadísticas acumuladas por temporada de La Liga.
- El sistema deberá ser fácil de utilizar por todos los usuarios.
- El sistema deberá dar a elegir entre distintos perfiles de jugadores.
- El sistema deberá permitir filtrar por una serie de estadísticas como: edad, porcentajes de éxito o ratios en diferentes acciones del juego, en función del perfil/es seleccionados.
- El sistema deberá permitir al usuario interactuar con el gráfico resultante.

7.1.2. Casos de uso

ID	Nombre	Descripción
UC01	Registro	El usuario deberá introducir sus credenciales para que el sistema reconozca a que equipo pertenece.
UC02	Selección de un perfil	El sistema deberá dar a elegir al usuario los distintos perfiles que quiere que abarque su análisis.
UC03	Filtrado	En función del perfil o los perfiles seleccionados el sistema dará la opción al usuario de filtrar por un grupo de estadísticas más importantes en dicho perfil.
UC04	Análisis	En función de todos los parámetros seleccionados el sistema mostrará distintas visualizaciones que ayuden al usuario a realizar un estudio del mercado de jugadores.

Tabla 7.1: Casos del uso de la aplicación

7.2. Mejoras de la aplicación

Tras una pequeña introducción de los usos generales de aplicación, podrían aplicarse mejoras que ofrecieran un mayor abanico de posibilidades al igual que pudieran aumentar el rango de análisis del mundo del fútbol. En los siguientes párrafos se detallan algunas de las propuestas de progreso, que han ido surgiendo durante el transcurso del trabajo.

Para este trabajo se ha seleccionado unicamente los jugadores de la liga española, pero el *dataset* a estudiar también contiene las mismas estadísticas para otras 4 ligas. Esto proporcionaría un mayor campo de análisis a la aplicación y sería una funcionalidad interesante para los "general manager"de todos los equipos.

Durante la primera selección de variables, se tuvo en cuenta distintos trabajos similares y la opinión del estudiante para elegir las estadísticas más relevantes. Pero una vez dentro del mundo del análisis de datos deportivos, se ha podido observar que existen otras posibilidades interesantes como las estadísticas por partido, los ratios o incluso se podría llegar a crear estadísticas propias, que aportan valor a la hora de comparar jugadores.

Mientras se hacia la elaboración de los *clusters*, se decidió eliminar del conjunto de datos a los jugadores que jugaban en la posición de portero. Con un conjunto de mayor tamaño del que se disponía (16 porteros que jugaron en La Liga durante las 3 temporadas de estudio), podría realizarse un *clustering*, unicamente a jugadores de dicha demarcación, y ver si es capaz de sacar distintos perfiles de guardametas, como quizás los denominados "para-penaltis", porteros que dominan el juego con los pies o los que son dueños de los balones aéreos.

Bibliografía

[1] Miguel Á. Morán. *Cómo Bellerín utilizó el 'big data' para decantarse por la oferta del Betis*. Accedido el 24/06/2022. URL: https://www.marca.com/futbol/betis/2022/03/23/623b07c8ca4741b44f8b45d0.html.

[2] *Página oficial de la Real Federación Española de Fútbol*. Accedido el 13/06/2022. URL: https://www.rfef.es/federacion/ligas-comisiones/liga-futbol-profesional.

[3] I. Trujillo. «Estos son los 63 equipos que han pasado, al menos una temporada, por la Primera División española». En: *Diario La Razón* (). Accedido el 27/02/2022. URL: https://www.larazon.es/deportes/futbol/20220126/tp3tgzae6zevbnccxkepljtntq.html.

[4] Adrián Martín Castellanos. *Métricas en el fútbol moderno*. Accedido el 17/06/2022. URL: https://objetivoanalista.com/metricas-en-el-futbol-moderno/.

[5] Jason Zivkovic. *Clustering to help club managers*. Accedido el 13/06/2022. URL: https://www.kaggle.com/code/jaseziv83/clustering-to-help-club-managers.

[6] Pablo Reyes. *Analizando jugadores de La Liga con machine learning*. Accedido el 13/06/2022. URL: https://blog.pabloreyes.es/laliga-jugadores-machine-learning/.

[7] Ramiro Gómez Nuño. *Nuevo paquete para el análisis de datos de juego y jugadores de fútbol: GASB*. Accedido el 17/06/2022. URL: https://uvadoc.uva.es/handle/10324/50487.

[8] Jorge San José Lorza. *Construcción de un modelo de goles esperados para los partidos de la Copa Mundial de la FIFA del año 2018*. Accedido el 17/06/2022. URL: https://uvadoc.uva.es/handle/10324/50507.

[9] Rafał Stępień. *Soccer players values and their statistics*. Accedido el 11/06/2022. URL: https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics.

[10] Daniel Peña. *Análisis de Datos Multivariantes*. Capítulo 12 págs. 355-389. Ed. Mc Graw Hill, 2002.

[11] Aapo Hyvärinen, Juha Karhunen y Erkki Oja. *Independent Component Analysis*. Ed. Wiley Interscience, 2001.

[12] Tomàs Aluja Banet y Alain Morineau. *Aprender de los datos: el análisis de componentes principales*. Ed. EUB S.L., 1999.

[13] Casey Cheng. *Análisis de componentes explicado*. Accedido el 16/06/2022. URL: https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d.

[14] Ignacio Cassol. *Clustering*. Accedido el 10/07/2022. URL: https://icassol.github.io/CursoR/Mod8.html.

[15] AprendeIA. *Teoría KMeans*. Accedido el 23/06/2022. URL: https://aprendeia.com/algoritmo-kmeans-clustering-machine-learning/.

[16] AprendeIA. *Teoría Clustering Jerárquico*. Accedido el 23/06/2022. URL: https://aprendeia.com/algoritmo-agrupamiento-jerarquico-teoria/.

[17] AprendeIA. *Teoría DBSCAN*. Accedido el 23/06/2022. URL: https://aprendeia.com/dbscan-teoria/.

[18] Ian H. Witten, Eibe Frank y Mark A. Hall. *Data Mining: practical machine learning tools and techniques (third Edition)*. Ed. Morgan Kaufmann, 2011.

[19] Elizabeth León Guzmán. *Técnicas de validación en algoritmos de clustering*. Accedido el 23/06/2022. URL: https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf.

[20] Football Data. *Football-Data*. Accedido el 11/06/2022. URL: https://www.football-data.co.uk/spainm.php.

[21] *Footballdatabase.eu*. Accedido el 11/06/2022. URL: https://www.footballdatabase.eu/es/.

[22] *Resultados Futbol.com*. Accedido el 11/06/2022. URL: https://www.resultados-futbol.com/primera.

[23] *StatsBomb*. Accedido el 11/06/2022. URL: https://statsbomb.com/es/.

[24] *BDFutbol archivos*. Accedido el 11/06/2022. URL: https://www.bdfutbol.com/es/c/archives.html.

[25] *Scraping Understat*. Accedido el 13/06/2022. URL: https://understat.readthedocs.io/en/latest/classes/understat.html.

[26] *European Soccer Database*. Accedido el 11/06/2022. URL: https://www.kaggle.com/datasets/hugomathien/soccer.

[27] *Transfermarkt.es*. Accedido el 13/06/2022. URL: https://www.transfermarkt.es/.

[28] *FBREF*. Accedido el 13/06/2022. URL: https://fbref.com/es/.

[29] *GitHub - Rafał Stępień*. Accedido el 13/06/2022. URL: https://github.com/RSKriegs/Modelling-Football-Players-Values-on-Transfer-Market-and-Their-Determinants-using-Robust-Regression.

[30] *Proyectos que usan los datos utilizados*. Accedido el 11/07/2022. URL: https://www.kaggle.com/code/ubiratanfilho/soccer-transfer-market-prediction.

[31] *Historial lesiones - Antunes*. Accedido el 13/06/2022. URL: https://www.transfermarkt.es/vitorino-antunes/verletzungen/spieler/44699.

[32] *Duración media de la carrera de un futbolista*. Accedido el 16/06/2022. URL: https://penaltyfile.com/average-career-length-of-soccer-player/.

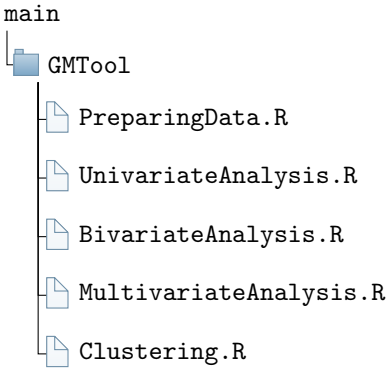
[33] Mario Garrido Tapias. *Código creado para el trabajo*. Accedido el 12/07/2022. URL: https://gitlab.inf.uva.es/margarr/gmtool.git.

Apéndice A

Código utilizado

A.1. Estructura

El conjunto de programas creados se reparte en función de la etapa en la que aplican sus operaciones, análisis o distintas funciones propias o de las bibliotecas utilizadas. A continuación, se nombra los programas utilizados junto con una breve explicación de su cometido dentro del trabajo:



- **PreparingData.R**

  Programa creado para la lectura y preparación del conjunto de datos, en el se realizan la creación de nuevas variables, el tratamiento de los jugadores que jugaron para dos clubes en la misma temporada o el análisis de los mismos.

- **UnivariateAnalysis.R**

  En él podemos encontrar distintas funciones propias que crean diferentes tipos de gráficos con la intención de realizar un análisis univariante de las diferentes estadísticas seleccionadas. Se usan bibliotecas como *tidyverse* o *tidyr* para operar y filtrar con los diferentes conjuntos de datos. Ade-

más para la parte gráfica nos apoyamos en *ggplot2* y *ggpubr* para tener visualizaciones multiples en la misma ventana.

- **BivariateAnalysis.R**

  Programa encargado de la parte bivariante del análisis, aquí se encuentran las funciones para crear los gráficos de la sección 4.4. A las bibliotecas mencionadas antes se suman *ggcorplot* y *correlation* para la creación de las matrices de correlación.

- **MultivariateAnalysis.R**

  Los tratamiento y gráficos creados en la sección 4.5 se crean desde este programa. La creación de los conjunto de entrenamiento y test junto con el análisis de componentes principales son las funciones básicas de este archivo. Las bibliotecas usadas son *stats* y *factoextra*, este último muy útil para la visualización de gráficos destinados a PCA y *clustering*.

- **Clustering.R**

  Finalmente este archivo recoge las componentes principales seleccionadas y aplica los algoritmos de aprendizaje no supervisado comentados en el capítulo 5. Las bibliotecas utilizadas son *stats*, para k-medias, *FactoMineR*, para jerárquico, y *dbscan*, para el algoritmo que lleva su nombre. En este archivo también podemos encontrar la validación de nuestro último modelo aplicado.

Estos programas son accesibles en el siguiente repositorio público de GitLab [33].