

Grau en Estadística

Títol: Disseny i aplicació de tècniques d'aprenentatge automàtic en el futbol femení

Autor: Montse Garcia Fernandez

Director: Francisco Javier Sierra Martinez

Departament: Dep. d'Econometria, Estadística i Economia Aplicada

Convocatòria: Gener 2024



UNIVERSITAT DE
BARCELONA



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística

ÍNDEX DE CONTINGUT

Introducció	8
1. Big Data	9
2. Knowledge Discovery in Databases (KDD)	11
3. Aprenentatge automàtic	13
3.1. Aprenentatge Supervisat	14
- Veí més proper (KNN)	14
3.2. Aprenentatge No Supervisat	14
3.2.1. Anàlisi factorial	15
- Anàlisi de Components Principals (ACP)	15
3.2.2. Anàlisi d'agrupament (Clustering)	17
4. Aplicacions informàtiques	18
4.1. R	18
4.2. Microsoft Excel	18
4.3. Microsoft Power BI	19
5. El futbol femení	20
5.1. La Liga F	21
5.2. Women's Super League	22
5.3. Frauen-Bundesliga	23
6. Base de dades	24
6.1. Descripció de les variables	25
1. Característiques personals i disciplina	25
2. Variables rellevants en la posició de portera (PO)	26
3. Variables rellevants en les posicions de defensa (DEF)	26
4. Variables rellevants en les posicions del mig del camp (MED)	27
5. Variables rellevants en les posicions avançades o d'atacant (DEL)	28
6.2. Estadística descriptiva i preprocessament	29
6.2.1. NAs	31
6.2.2. Variables categòriques	32
6.2.3. Variables numèriques	36
7. Implementació dels algoritmes	38
7.1. ACP	38
7.2. Clustering	43
- Posició de PORTERA	43

- Posició de DEFENSA.....	45
- Posició de MIGCAMPISTA	47
- Posició d' ATACANT	49
8. Cas pràctic	51
- Substitució per lesió	51
9. Conclusions	54
Bibliografia	55
Annex	1
METADATA	1

ÍNDIX DE FIGURES

Figura 1: Esquema dels diferents passos del procés KDD.	12
Figura 2: Top 3 en la classificació UEFA de la temporada 2022/23.	20
Figura 3: Pantalla inicial de la web fbref.com.	24
Figura 4: Distribució de la variable Lliga.	32
Figura 5: Mapa coroplètic on es pot veure la freqüència de la variable País.	34
Figura 6: Distribució de la variable Posició.	35
Figura 7: Representació de la formació 4-4-3.	35
Figura 8: Boxplot any de naixement per cada lliga.	36
Figura 9: Boxplots del nombre de partits jugats i del nombre de minuts jugats per cada lliga. .	36
Figura 10: Gols en contra i Xuts en contra de les porteres de cada lliga.	37
Figura 11: Regatejadors abordats i intents perduts de les defenses per cada lliga.	37
Figura 12: Representació dels valors de les correlacions entre variables.	38
Figura 13: Quantitat d'inèrcia acumulada per cada component.	39
Figura 14: Dendrograma inicial Portera	43
Figura 15: Visualització del nombre de clústers òptim Portera.	43
Figura 16: Dendrograma final Portera.	44
Figura 17: Dendrograma ampliat del clúster 2 Portera.	44
Figura 18: Dendrograma inicial Defensa.	45
Figura 19: Visualització del nombre de clústers òptim Defensa.	45
Figura 20: Dendrograma final Defensa.	46
Figura 21: Dendrograma ampliat del clúster 4 Defensa.	46
Figura 22: Dendrograma inicial Migcampista.	47
Figura 23: Visualització del nombre de clústers òptim Migcampista.	47
Figura 24: Dendrograma final Migcampista.	48
Figura 25: Dendrograma ampliat del clúster 6 Migcampista.	48

Figura 26: Dendrograma inicial Atacant.	49
Figura 27: Visualització del nombre de clústers òptim Atacant.....	49
Figura 28: Dendrograma final Atacant	50
Figura 29: Dendrograma ampliat del clúster 4 Atacant.	50
Figura 30: Gràfic de radar Leah Williamson.	51
Figura 31: Dendrograma general Defensa.....	52
Figura 32: Dendrograma clúster 5.	52
Figura 33: Branca dendrograma Leah Williamson.....	52
Figura 34: Gràfic de radar per comparar propostes Clustering.	53
Figura 35: Gràfic de radar per comparar propostes PCA.....	53

ÍNDIX DE TAULES

Taula 1: Equips de La Liga F (temporada 22/23).....	21
Taula 2: Equips de la Women's Super League (temporada 22/23).....	22
Taula 3: Equips de la Frauen-Bundesliga (temporada 22/23).	23
Taula 4: Variables de la BBDD, descripció, classe, valors absents i percentatges que suposen. .	29
Taula 5: Valors absents per fila (jugadora).	31
Taula 6: Primeres variables de la jugadora Vanessa Fischer, amb el valor més alt de missings..	31
Taula 7: Comprovació de que no queda cap valor faltant a les variables numèriques després d'aplicar KNN.	31
Taula 8: Distribució del nombre de jugadores que hi ha a cada equip	33
Taula 9: Freqüències de la variable País i els percentatges que suposen.....	34
Taula 10: Percentatges d'inèrcia i inèrcia acumulada per cada variable.....	39
Taula 11: Nombre d'observacions per cada clúster Portera.....	44
Taula 12: Nombre d'observacions per cada clúster Defensa.....	46
Taula 13: Nombre d'observacions per cada clúster Migcampista	48
Taula 14: Nombre d'observacions per cada clúster Atacant.....	50
Taula 15: Proposta de substitucions a través de PCA.....	53

Resum

L'objectiu d'aquest projecte de final de grau és explorar un conjunt de dades relacionades amb el futbol femení; recollir les dades, visualitzar-les i aplicar algoritmes d'aprenentatge automàtic amb la finalitat de classificar les jugadores per similitud.

La contextualització i el marc teòric es poden dividir en dues parts. La primera es centra en introduir conceptes més informàtics i/o estadístics: *Big Data* o aprenentatge automàtic, per exemple. La segona part introdueix el futbol femení i les tres lligues les quals fan referència les dades tractades.

Totes les tècniques utilitzades han sigut implementades amb el llenguatge de programació R, llenguatge utilitzat durant el transcurs de tot el grau. En concret, les tècniques utilitzades, a més de les descriptives (gràfics i taules), han sigut dues: Anàlisi de Components Principals i *Clustering*.

Per acabar, s'han plantejat un parell de casos pràctics, situacions que ens podríem trobar en la vida real i on que ens podria ser d'utilitat l'aplicació d'aquests algoritmes. S'ha pogut concloure que els algoritmes són útils per determinar jugadores amb característiques semblants.

Abstract

This report focuses on analysing data related to women's football, utilizing methods such as data collection, visualization, and machine learning for player categorization.

The aim of the research paper is to explore the results using statistical multivariate techniques of analysis to identify behavioural patterns.

After the analysis, it is concluded that the algorithms are useful in order to determinate players with similars characteristics.

Introducció

L'estadística és la disciplina científica que s'ocupa de l'obtenció, l'ordre i l'anàlisi d'un conjunt de dades amb el propòsit d'obtenir explicacions i prediccions sobre els fenòmens observats, però, realment totes les dades que es recullen, s'analitzen? De què serveix emmagatzemar grans volums de dades si al final queden en l'oblit?

En l'últim segle, els avenços tecnològics han accelerat exponencialment la creació de dades en tots els sectors, el futbol ha sigut un d'aquests. A poc a poc s'han anat recollint dades cada vegada més sofisticades i, actualment, fins i tot la mateixa LaLiga¹ ofereix un portal anomenat “*Beyond Stats*” que, juntament amb Microsoft, pretenen extreure coneixement de les dades recollides i, així, poder-les compartir amb els aficionats i amb tots els equips que així ho vulguin. (Microsoft News, 2021)

Per altra banda, el futbol femení està fent créixer la seva audiència ràpidament els últims anys, jugadores com l'Alèxia o l'Aitana² i el fet que la selecció Espanyola guanyes la Copa del Món Femenina de Futbol aquest darrer any 2023, han fet que es fes més visible aquesta àrea del futbol. Tot i així, encara és difícil trobar grans bases de dades, si més no amb la facilitat amb la qual es troben les de la secció masculina.

Aquest treball pretén agrupar, ordenar i analitzar un conjunt de dades que fan referència al futbol femení de tres grans lligues europees; la lliga espanyola, la lliga anglesa i la lliga alemanya.

Els objectius principals del treball són:

- Cerca i agrupament d'una base extensa de dades del futbol femení
- Aplicar tècniques estadístiques de classificació
- Plantejament de possibles casos en la vida real i anàlisis de similituds entre jugadores

El treball es distribueix en 4 grans blocs:

- Apartats de l'1 al 5: Contextualització i marc teòric
- Apartat 6: Descripció de la base de dades
- Apartat 7 i 8: Implementació dels algorismes i casos pràctics

¹ Lliga masculina del futbol Espanyol.

² Alèxia Putelles i Aitana Bonmatí, jugadores del FCBarcelona i guanyadores de les últimes 3 pilotes d'or (2 i 1, respectivament)

1. Big Data

Les dades massives, “*macrodatos*” en castellà o “*big data*” en anglès, són un conjunt de dades que, a causa del seu enorme volum, han de ser processades per mitjà de tecnologies modernes especials. Dins aquestes tecnologies podem trobar l'emmagatzematge, la classificació, l'anàlisi, el filtratge o la visualització, entre altres. I, a través de la mineria de dades converteixen les dades en informació significativa. (enciclopèdia.cat)

Els avenços tecnològics han permès, a través dels ordinadors i les xarxes informàtiques, que tinguem accés a una quantitat molt més gran de dades i, a la vegada, han permès evolucionar algunes tècniques de processament que ja existien i que ara, gràcies a la tecnologia, siguin molt més ràpides d'utilitzar.

L'augment de facilitats per analitzar dades també ha suposat un canvi en el món esportiu, tot i que encara queda molt recorregut per explorar en aquest àmbit ja en són molts els clubs que contracten un grup d'experts assignats a analitzar les dades de tots els sectors que poden intervenir en cada equip i extreure conclusions per millorar. En el cas del futbol, més enllà d'usar les anàlisis per futurs fitxatges o per “descobrir” aquella peça en brut de les categories inferiors, també podem trobar casos d'aplicacions del *Big Data* per millorar l'assistència dels socis al camp. En el cas del València, per exemple, va augmentar un 20% els assistents als partits analitzant dades dels seguidors que els van permetre crear esdeveniments i experiències més gratificants i atractives per ells. (Microsoft News, El Valencia Club de Fútbol incrementa un 20% el número de asistentes a sus partidos, gracias al análisis de datos, 2022)

No es pot deixar de banda que tot i que hi ha moltes possibles variables a analitzar, hi ha una certa part molt complicada de mesurar i, per tant, poder predir o classificar, per exemple. Les dades d'un jugador poden ser molt bones i, tot i això, no demostrar el rendiment esperat, ja sigui perquè no s'ha adaptat bé al grup, no té la motivació suficient o simplement per alguna lesió. Ho podem veure en el cas de l'*Ajax*³ masculí en aquest inici de temporada, tot i que van invertir un gran nombre de diners i fan fer diversos fitxatges que, segons el *Big Data*, havien de suposar una millora enorme a l'equip, no ha sortit com esperaven i ha fet que tinguin un dels seus pitjor inicis de temporada de la seva història. Ha fallat el *Big Data*? Ha fallat l'anàlisi? A falta de molts partits per acabar la temporada, és d'hora per extreure conclusions?. (Torres, 2023)

³ L'Amsterdamsche Football Club Ajax, és un equip de futbol neerlandès de la ciutat d'Amsterdam.

Tot i el fàcil accés a les dades i a tecnologies ràpides, un mal ús pot portar a extreure conclusions errònies o manipulades, per això, és important que les persones encarregades de gestionar-ho tinguin els coneixements necessaris.

«L'experiència ha ensenyat als estadístics que les dades poden ser enganyoses i, pitjor encara, aportar una sensació errònia d'objectivitat». (Rossell, 2014)

2. Knowledge Discovery in Databases (*KDD*)

El descobriment de coneixement en bases de dades (*KDD*, de l'anglès *Knowledge Discovery in Databases* és el procés per descobrir patrons en grans volums de dades i així poder resoldre els problemes plantejats.

Aquest procés es pot considerar una col·lecció de diferents tècniques que serveixen per induir coneixement i informació de manera estructurada a un conjunt de dades. (Vilches Gonzáles & Escobar Broitman, 2007)

Des dels anys seixanta, els estadístics utilitzaven termes com *data fishing* o *data mining* amb la idea de trobar correlacions sense hipòtesi prèvia en bases de dades amb soroll, és a dir, bases de dades extenses amb més informació de la necessària per poder extreure patrons. Més endavant, als anys vuitanta, aquests termes acabarien esdevenint a la mineria de dades que coneixem actualment tot i que, realment, la mineria de dades és només un dels passos que formen el procés *KDD*.

Veiem quin és el conjunt de passos que formen el *KDD* (Pérez López, 2007):

1. Plantejar el problema que es vol resoldre i definir dels objectius. Aquest punt és important per poder determinar quins processos seran necessaris aplicar i quines dades són necessàries per poder completar el propòsit de l'anàlisi.
2. Seleccionar o recollir les dades necessàries i preparar-les per l'anàlisi. Aquest pas, juntament amb l'anterior, són els més importants del procés i els que s'hauria de dedicar més temps tot i que no sempre és així, és un error comú pensar que la fase important és la que ve a continuació, la de crear models, però si no tens una base de dades correctament depurada, pot portar-te a confondre en els resultats.
3. Minería de dades: Estudiar els patrons i crear models. En aquest moment del procés s'apliquen tècniques que provenen de l'estadística i/o de la intel·ligència artificial, algorismes més o menys sofisticats que ens permeten classificar, predir o detectar anomalies, entre altres.

4. Avaluar els resultats, obtenir coneixement i aplicar-lo. En l'últim pas es comparen les diferents tècniques o models aplicats i es busca aquell que millor s'ajusta al problema. Un cop aconseguit un resultat, es valora si compleix amb tots els objectius del procés, en cas afirmatiu, s'implementa el coneixement o millora i, en cas negatiu, s'itera el procediment variant alguna de les parts anteriors per poder assolir resultats diferents.

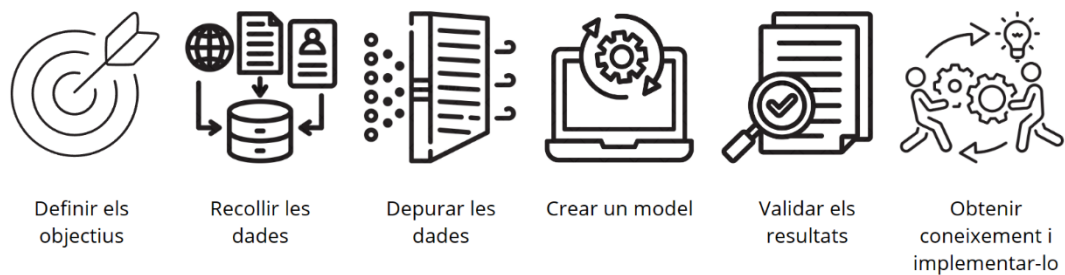


Figura 1: Esquema dels diferents passos del procés KDD.

Aplicacions a la vida real

Tot i que hi ha infinites àrees i casos on es pot aplicar el procés *KDD*, aquests són alguns dels exemples:

- Àrees de comerç i banca. Es pot utilitzar per fer previsions de vendes, segmentacions de clients o per detectar patrons en el cistell de consum, per exemple.
- Medicina. Pot ser útil en un diagnòstic més precís d'algunes malalties o per identificar riscos en certs segments de poblacions, entre altres.
- Mitjans digitals. És possible emmagatzemar i analitzar informació anònima en temps real de les visualitzacions de canals per tal de poder conèixer els interessos i les activitats dels oients i, així, poder crear emissions personalitzades, recomanacions i, també, és un punt important pels anunciants, ja que poden tenir més precisió a l'hora d'escollir els clients potencials.
- Ciències ambientals. Es fa servir per identificar models de comportament dels ecosistemes i, així, millorar la gestió i el control d'aquests.

3. Aprenentatge automàtic

L'aprenentatge automàtic també conegut com a Machine Learning, és el procés pel qual els algoritmes milloren la seva pròpia capacitat per dur a terme tasques analitzant noves dades sense la necessitat que les persones li donin instruccions en la forma de programar. (Cambridge Dictionary, 2020)

El Machine Learning, a partir d'ara abreviat com a ML, és una part molt important en la intel·ligència artificial i les seves tècniques ajuden a estudiar els patrons i crear models, és a dir, formen part del punt 3 en els passos, exposats anteriorment, de la mineria de dades.

Els seus inicis es remunten a la dècada de 1940, Alan Turing⁴ va formar les bases de la teoria de la computació tot i que no va ser fins uns vint anys després, el 1959, que es va introduir el terme d'aprenentatge automàtic.

Als anys vuitanta es van fer avenços destacats en les xarxes neuronals que va permetre aplicacions de ML més sofisticats i efectius. I tot i que durant els anys l'aprenentatge ha anat avançant, no va ser fins a la dècada del 2010 que aquests canvis van poder ser més notoris i més ràpids gràcies a la introducció de més potència computacional a través d'ordinadors, que també van permetre emmagatzemar i disposar de grans conjunts de dades.

Aplicacions a la vida real

Les noves tecnologies evolucionen d'una manera molt accelerada i això fa que les tècniques d'aprenentatge automàtic cada vegada siguin més sofisticades i s'apliquin a més sectors. Aquests són alguns exemples d'on es pot aplicar: (IBM)

- Atenció al client automatitzat. Permet a les empreses crear un diàleg en temps real amb el client a través dels *Chatbots*⁵. Aquests reben una pregunta per part del client i, gràcies a l'entrenament amb preguntes freqüents que ha tingut prèviament, crea una resposta.
- Tendències de consum. Permet desenvolupar estratègies de venda creuada més eficaces i, aprenent de les compres anteriors, recomana productes que poden interessar al nou comprador. Com més entrenat, més bones recomanacions farà.
- Detecció de frau. Les institucions financeres i/o els bancs poden utilitzar-ho per detectar possibles transaccions fraudulentament. El model estudia transaccions ja conegudes, extreu possibles situacions que es repeteixen i pot classificar les noves per similitud.

⁴ Alan Mathison Turing. 1912-1954. Va ser un científic, matemàtic, lògic i biomatemàtic britànic que va treballar en camps com la informàtica teòrica o la intel·ligència artificial. És considerat el pare de la informàtica moderna.

⁵ Chatbot: programari basat en Intel·ligència Artificial capaç de mantenir una conversació en temps real per text o veu.

Tot i que també hi pot haver una barreja entre els dos (aprenentatge semisupervisat), existeixen dos subgrups d'aprenentatge; el supervisat i el no supervisat:

3.1. Aprenentatge Supervisat

L'aprenentatge supervisat utilitza conjunts de dades etiquetades per entrenar algoritmes.

Per exemple, en el cas de voler classificar un conjunt de registres, aquests estaran etiquetats i, per tant, sabrem a quin grup pertanyen. Això ens permetrà, a través de grups d'entrenament, avaluar i ensenyar al model les característiques de cada dada i així, quan introduïm un nou registre, aquest serà classificat automàticament en el grup corresponent.

Aquest seria el cas d'algoritmes de classificació com *Naive Bayes* o de regressió com les regressions lineals, per exemple. També podem trobar algoritmes que són a la vegada de classificació i de regressió com serien els arbres de decisió o les xarxes neuronals. (IBM)

- Veí més proper (KNN)

L'algoritme KNN, de l'anglès "*k-nearest neighbors*", és un mètode de classificació supervisada que permet, entre altres coses, imputar valors absents a les bases de dades. El model intenta identificar els *k* casos més similars al valor faltant en particular, per fer-ho, es considera cada individu com un vector i s'utilitza una mesura de distància (la de Gower⁶ per exemple) per mesurar la diferència entre els valors de la mateixa posició del vector, és a dir, la mateixa variable. Com més petita sigui la distància entre els dos casos, més semblants es podran considerar. Un cop detectats els *k* valors més propers, s'assigna el valor corresponent a la mesura faltant. (Kowarik & Templ, 2016)

3.2. Aprenentatge No Supervisat

A diferència de l'anterior, l'aprenentatge no supervisat fa servir conjunts de dades no etiquetades. El mateix model serà l'encarregat de detectar les possibles etiquetes.

Per exemple, igual que abans, es vol classificar un conjunt de registres, però aquest cop no estan etiquetats i, per tant, no tenen cap variable que els adjudiqui en un grup en concret. El mateix model serà l'encarregat de buscar les característiques que formen un patró o detectar aquelles variables que tenen més similituds a d'altres i, al contrari, aquelles que són més diferents. Una vegada creats els grups o clústers, en introduir un registre nou, el mateix model serà capaç de classificar-lo correctament. Sovint s'utilitzen en regles d'associacions, reduccions de la dimensionalitat o agrupacions en clústers. (IBM)

⁶ Distància de Gower. 1971. Mètrica per calcular les similituds entre parells d'observacions.

3.2.1. Anàlisi factorial

L'anàlisi factorial s'utilitza principalment per reduir la dimensió, eliminant variables redundants, o per detectar estructures imperceptibles que estan presents en les dades.

- Anàlisi de Components Principals (ACP)

L'anàlisi de components principals (*Karl Pearson*⁷, 1901), també anomenat PCA de l'anglès *Principal Component Analysis*, forma part del grup d'aprenentatges no supervisats que s'encarreguen de reduir la dimensionalitat, és a dir, l'objectiu principal de l'ACP és reduir les dimensions d'un conjunt de dades conservant la informació important d'aquest. (Novales, 2016)

Això implica transformacions lineals entre variables que tenen una major variància i així, obtenir el menor nombre de variables necessàries, anomenades components, ordenar-les de major a menor. Les components s'ordenen de manera que la primera component és la que més explica les dades i, successivament, fins a l'última component.

Model matemàtic del càlcul de les components principals:

Sigui $X = (X_1, \dots, X_p)'$ un vector aleatori de llargada p amb vector de mitjanes $\mu = E(X)$ i matriu de covariàncies $\Sigma = D(X) = Cov(X, X)$.

- Es defineix la Primera Component Principal de X com una variable aleatòria Z_1 tal que:

$$Z_1 = v_{11}X_1 + \dots + v_{p1}X_p = v_1'X \quad \text{amb} \quad v_1 = (v_{11}, \dots, v_{p1})' \in \mathbb{R}^p$$

serà una combinació lineal normalitzada de les variables de X i, entre totes les combinacions lineals normalitzades, ens quedarem amb la que aconseguim:

$$Var(Z_1) = \max \{ Var(v'X) : v \in \mathbb{R}^p, \quad v'v = 1 \}$$

o, equivalentment:

$$\begin{aligned} & \text{Max} \quad v'\Sigma v \\ & \text{s. a.:} \quad v'v = 1 \end{aligned}$$

és a dir, la que té una variància major. Per tal de garantir una solució, és necessari normalitzar el vector de càrregues de la següent manera: $\|v\| = \sqrt{v'v} = 1$

Per portar a terme l'elecció de v_1 , és comú utilitzar el mètode dels multiplicadors de *Lagrange*⁸ de manera que obtenim: $Var(Z_1) = \lambda_1$

⁷ Karl Pearson. 1857-1936. Va ser un científic, matemàtic i pensador britànic molt influent, considerat el fundador de la bioestadística.

⁸ Mètode dels multiplicadors de *Lagrange*: és un mètode per trobar l'extrem d'una funció de diverses variables subjecte a una o més restriccions; és l'eina bàsica en l'optimització no lineal amb restriccions.

De la mateixa manera;

- Es defineix la Segona Component Principal de X com una variable aleatòria Z_2 tal que:

$$Z_2 = v_{12}X_1 + \dots + v_{p2}X_p = v_2'X \quad \text{amb} \quad v_2 = (v_{12}, \dots, v_{p2})' \in \mathbb{R}^p$$

$$Var(Z_2) = \max \{ Var(v'X) : v \in \mathbb{R}^p, \quad v'v = 1, \quad v'v_1 = 0 \}$$

o, equivalentment:

$$\begin{aligned} & \text{Max} \quad v' \Sigma v \\ & \text{s. a.:} \quad v'v = 1 \\ & \quad \quad v'_1 v = 0 \end{aligned}$$

Aquest cop haurem de tenir en compte una condició més, la covariància entre les components ha de ser igual a 0, és a dir:

$$Cov(v'X, Z_1) = Cov(v'X, v_1'X) = v' \Sigma v_1 = 0$$

La segona component principal és una altra combinació lineal de les variables de X i, entre totes les combinacions formades pels vectors unitaris ortogonals a v_1 , és la que té major variància.

De nou, amb el mètode dels multiplicadors de *Lagrange*, obtindrem: $Var(Z_2) = \lambda_2$

Aquest cop, el valor de λ_2 serà menor que λ_1 , però major a tots els valors de λ_j que vindran consecutivament.

Finalment,

Es defineixen les p Components Principals de X com les variables aleatòries $(Z_1, \dots, Z_p)$ tals que:

$$Z_j = v_{1j}X_1 + \dots + v_{pj}X_p = v_j'X, \quad j \in \{1, \dots, p\} \quad \text{amb} \quad (v_{1j}, \dots, v_{pj}) \in \mathbb{R}^p$$

$$Var(Z_j) = \max \{ Var(v'X) : v \in \mathbb{R}^p, \quad v'v = 1, \quad v'v_1 = 0, \dots, v'v_{j-1} = 0 \}$$

o, equivalentment:

$$\begin{aligned} & \text{Max} \quad v' \Sigma v \\ & \text{s. a.:} \quad v'v = 1 \\ & \quad \quad v'_{j-1} v = 0 \end{aligned}$$

cada una de les components haurà de tenir covariància igual a zero amb la que la precedeix.

De forma general, la solució de cada mètode dels multiplicadors de *Lagrange* serà: $Var(Z_j) = \lambda_j$

de forma que els valors de λ estaran ordenats de major a menor variància ($\lambda_1 \geq \dots \geq \lambda_p$).

3.2.2. Anàlisi d'agrupament (Clustering)

L'anàlisi d'agrupament s'utilitzen per trobar estructures o patrons a col·leccions de dades que no estan categoritzades (aprenentatge no supervisat), per formar diferents subgrups anomenats clústers. Les observacions de cada grups seran molt semblants entre si i, a més, seran tant diferents com sigui possible amb les dels altres grups.

- Procediment jeràrquic aglomeratiu

El procediment jeràrquic aglomeratiu és el procés iteratiu per identificar grups relativament homogenis que s'utilitza quan, a priori, no es coneix el número de clústers òptim. Es parteix de la suposició que hi ha tants grups com observacions, és a dir, cada observació forma un grup, seguidament es comença a iterar unint els dos grups que siguin més semblants. Tenint en compte que un cop s'han unit dues observacions ja no es separen en tot el procés es segueix iterant fins a ubicar cada registre en un grup.

- Mètode de Ward

El mètode de Ward és un procediment jeràrquic aglomerat on s'escull el pròxim grup a unir-se agafant aquell que té la mínima variància, això implica que maximitzarà la homogeneïtat dins el propi grup i, consegüentment, farà que hi hagi la màxima diferència entre la resta dels grups.

El procediment que segueix és el següent:

- 1r. Es calcula la distància euclidiana al quadrat entre cada individu i el seu centroides per cada possible subgrup d'individus, és a dir, hi haurà $\binom{i}{2}$ ⁹ possibles clústers. S'escull aquell clúster que fa mínima la suma de quadrats dels errors.
- 2n. Es torna a fer el 1r pas, aquest cop, hi haurà $\binom{i-1}{2}$ possibles clústers. Es sumen els quadrats dels errors de cada un i es calcula l'increment respecte al valor anterior. S'escull aquell que implica una variació més baixa.
- 3r. Es segueix iterant de la mateixa manera fins a obtenir un únic clúster que integra totes les observacions de la mostra.

Aquest procediment ens permetrà crear un dendrograma i, així, veure les diferents branques que es formen.

⁹ i indica el nombre d'observacions que té la mostra.

4. Aplicacions informàtiques

Les aplicacions informàtiques utilitzades per a la realització d'aquest treball han estat:

- R, en el seu format de R-Studio; utilitzat per implementar els algorismes.
- Microsoft Excel; utilitzat per a crear la selecció de la base de dades.
- Microsoft PowerBI; utilitzat per la visualització de les dades en alguns gràfics.

4.1. R

R és un llenguatge i un entorn per la computació estadística. Està disponible com a Software Lliure i permet manipular dades, calcular i visualitzar gràficament. (r-project.org)

Les funcions disponibles es poden ampliar fàcilment a través de *paquets*, alguns d'ells són propis i altres externs. Aquests últims passen una rigorosa revisió a través de la xarxa *CRAN*¹⁰ abans d'estar disponibles a l'entorn de treball. A més, R té disponible un entorn de desenvolupament integrat (**Rstudio**), és a dir, un programa per poder utilitzar R d'una manera més còmode.

El fet que sigui un programari obert, lliure i gratuït i, a més, ofereixi entorns més intuïtius fa que sigui molt utilitzat per persones que s'inicien en la programació estadística i, consegüentment, s'utilitzi força en l'àmbit acadèmic. Tot i això, també el fan servir sovint empreses privades, investigadors científics, analistes financers o economistes, per exemple.

En aquest treball s'ha emprat per a l'aplicació dels algorismes i, també, per la realització de la major part dels gràfics.

4.2. Microsoft Excel

Excel és un editor de fulls de càlcul creat per Microsoft que està disponible en sistemes operatius d'ordinadors com Windows o Mac i també, des de fa uns anys, estan disponibles les versions de telèfon mòbil que operen amb sistemes d'*Android* o *iOS*. (Microsoft, 2021)

Excel va ser creat els anys vuitanta i, des d'aleshores, no ha deixat d'introduir noves funcionalitats, permet importar o exportar dades, crear taules, ordenar i filtrar variables, crear gràfiques, calcular fórmules matemàtiques i, fins i tot, automatitzar processos a través de la programació de macros. Aquestes funcions són un petit recull de tots els procediments que permet l'aplicació.

En aquest treball, s'ha emprat per a poder unir diferents taules i poder crear el conjunt de dades que s'han analitzat.

¹⁰ The Comprehensive R Archive Network: Xarxa en la que s'arxivien les versions de R base així com els paquets. S'encarreguen d'assegurar el correcte funcionament.

4.3. Microsoft Power BI

Power BI és una col·lecció de serveis de programari, aplicacions i connectors que funcionen per convertir dades sense relació entre si en informació coherent, interactiva i visualment atractiva. (Microsoft Learn, 2024)

Aquest servei, que ofereix Microsoft, es va donar a conèixer pel públic en general l'any 2015 i, tot i no tenir ni deu anys d'antiguitat, ja és una eina de visualització molt important en les empreses, ja que permet fàcilment veure els *KPI*¹¹ en temps real i d'una forma molt amena.

En aquest treball, s'ha emprat per poder crear un mapa coroplètic¹² (gràfic ) amb la col·laboració del navegador Bing, *Power BI* permet ubicar fàcilment els països sobre el mapa detectant les inicials de cada un d'ells.

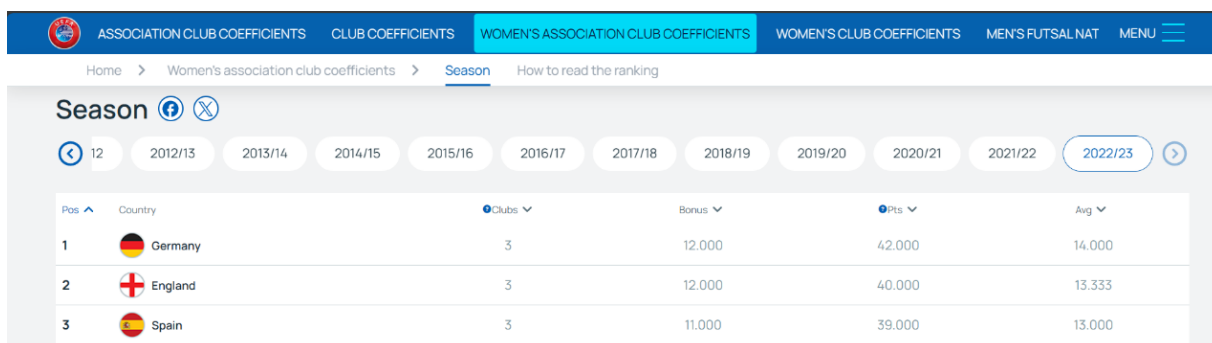
¹¹ KPI: Indicadors claus de rendiment (de l'anglès: Key Performance Indicator).

¹² Mapa coroplètic: és un mapa temàtic que s'utilitza per representar dades estadístiques mitjançant la tècnica de simbologia de mapes de colors.

5. El futbol femení

Tot i que sempre han existit equips femenins, els últims anys, aquests han començat a agafar més importància, augmentant l'assistència als camps <<El Camp Nou ho torna a fer i bat un nou rècord mundial d'assistència amb 91.648 assistents!¹³>>, informava el FCBarcelona l'abril de 2022 i, també l'audiència des de casa <<La Liga F casi triplica su audiencia en el arranque de la temporada 2023-2024¹⁴>>, titulava el diari Palco23 el setembre de 2023.

La Unió d'Associacions Europees de Futbol, *UEFA* a partir d'ara, puntua i classifica les lligues dels països europeus segons el resultat que obtenen els equips en competicions europees. Per exemple, aconsegueixen dos punts per cada victòria d'un equip a partir de la fase de grups o un punt per cada ronda que aconsegueixin després de fase de grups. La temporada 2022/23 les lligues que van assolir més punts van ser l'alemanya (*Frauen-Bundesliga*), l'anglesa (*Women's Super League*) i l'espanyola (*LaLiga F*).



The screenshot shows the UEFA website's 'WOMEN'S ASSOCIATION CLUB COEFFICIENTS' section. It features a navigation bar with tabs for 'ASSOCIATION CLUB COEFFICIENTS', 'CLUB COEFFICIENTS', 'WOMEN'S ASSOCIATION CLUB COEFFICIENTS' (selected), 'WOMEN'S CLUB COEFFICIENTS', and 'MEN'S FUTSAL NAT'. Below the navigation bar, there's a breadcrumb trail: 'Home > Women's association club coefficients > Season'. The 'Season' tab is active, showing a list of seasons from 2012/13 to 2022/23, with 2022/23 selected. Below the season list, there's a table with columns: Pos, Country, Clubs, Bonus, Pts, and Avg. The table shows the top 3 countries for the 2022/23 season.

Pos	Country	Clubs	Bonus	Pts	Avg
1	Germany	3	12.000	42.000	14.000
2	England	3	12.000	40.000	13.333
3	Spain	3	11.000	39.000	13.000

Figura 2: Top 3 en la classificació UEFA de la temporada 2022/23.

¹³ Van aconseguir el rècord d'assistència en el partit de semifinals de *Champions* contra el *Wolfsburg*. (fcbarcelona.cat)

¹⁴ En la segona jornada va augmentar la mitjana d'audiència un 127% respecte tota la temporada anterior. (Palco23)

5.1. La Liga F



Història

La Primera Divisió Femenina d'Espanya o Liga F es juga sense interrupció des de l'any 1988 i és organitzada per la Lliga Professional de Futbol Femení (LPFF). És considerada la tercera millor lliga Europea segons la classificació de la UEFA en els últims cinc anys. (UEFA, 2023)

L'any 2021 va deixar de ser una lliga semiprofessional i va passar a considerar-se una lliga professional, això va suposar un gran avenç pel futbol femení, ja que, en l'àmbit nacional, únicament tres lligues posseeixen aquest estatus, la Lliga masculina de Bàsquet, la Primera i la Segona Divisió de futbol masculí.

L'equip amb més títols al llarg de la història és el FC Barcelona amb un total de vuit campionats guanyats, seguit de l'Athletic Club de Bilbao, que ha sigut vencedor en un total de cinc ocasions.

Sistema de competició

Des de l'any 2012, la màxima categoria de la competició és disputa amb un únic grup de setze equips, jugant tots contra tots amb un partit com a local i un com a visitant. Els equips puntuen segons els seus resultats; tres punts per partit guanyat i un punt per partit empatat mentre que els partits perduts no suposen puntuació. Aquell equip amb més puntuació és proclama campió de la temporada i obté una plaça per la major competició Europea (*UEFA Women's Champions League*) i, contràriament, aquells dos equips que obtenen menys puntuació descendeixen a la segona categoria nacional (Primera Federació Fufem).

Els 16 equips que formaven part de *La Liga F* durant la temporada 22/23: (La Liga F)

	FC Barcelona		Madrid CFF		VCF Femenino		Sporting Club Huelva
	Real Madrid CF		Costa Adeje Tenerife		Athletic Club		Villarreal CF
	Levante UD		Sevilla FC		FC Levante Las Planas		Alhama CF ElPozo
	Atlético de Madrid		Real Sociedad		R. Betis Femeninas		Deportivo Alavés

Taula 1: Equips de La Liga F (temporada 22/23)

5.2. Women's Super League



Història

La *Women's Super League* (WSL) és la lliga més alta del futbol a Anglaterra es va crear el 2011 substituint a la *FA Women's Premier League* que va passar a ser la segona categoria nacional. La lliga és organitzada per la FA¹⁵ i no va ser fins a la temporada 2017-18 que va passar a considerar-se una lliga completament professional. La lliga femenina anglesa és la quarta en la classificació general de la *UEFA*. (UEFA, 2023)

L'equip amb més títols és el *Chelsea* amb un total de sis campionats guanyats, la meitat dels títols jugats des de 2011. El segueixen l'*Arsenal*, amb tres victòries, el *Liverpool* amb dues i finalment, el *Manchester City* que posseeix un títol de lliga.

Sistema de competició

Al llarg dels anys ha anat variant el format de la lliga i del nombre d'equips que hi participaven, inicialment la lliga era de vuit equips i, progressivament, es va anar augmentant fins a obtenir, l'any 2019, el format de l'actual lliga, una lliga de dotze equips. Els dotze clubs juguen tots contra tots amb un total de dos partits contra el mateix contrincant i aconseguint tres punts per cada partit guanyat, un punt per cada partit empatat i cap punt pels partits perduts. En aquest cas, els tres millors equips classificats assoleixen una plaça a la major competició Europea (*UEFA Women's Champions League*) i un únic equip, aquell que menys punts obté, descendeix a la Segona Divisió Anglesa (*Women's Championship*).

Els 12 equips que formaven la *Women's Super League* durant la temporada 22/23: (The FA Player)

	Chelsea		Aston Villa		Tottenham Hotspur
	Manchester United		Everton		Leicester City
	Arsenal		Liverpool		Brighton and Hove Albion
	Manchester City		West Ham United		Reading FC

Taula 2: Equips de la *Women's Super League* (temporada 22/23).

¹⁵ Football Association: Organisme rector del futbol a Anglaterra i les dependències de la Corona de Jersey, Guernsey i Illa de Man.

5.3. Frauen-Bundesliga



Història

La Frauen-Bundesliga o Lliga Femenina Alemanya va ser creada el 1990 per la Federació Alemanya de Futbol (DFB), la qual continua dirigint la lliga avui en dia. Tot i considerar-se la segona millor lliga segons la classificació general *UEFA*, el futbol femení Alemany encara no és considerat futbol professional. (UEFA, 2023)

A l'alt de la classificació de clubs amb més títols, amb un total de set campionats guanyats cada un, trobem al *Eintracht Frankfurt* i al *VfL Wolfsburg*.

Sistema de competició

La competició consta de dues rondes on els dotze equips juguen tots contra tots, un partit com a local i un com a visitant. Actualment, igual que en les competicions anteriors, per cada victòria, els clubs s'emporten tres punts, per cada empat n'és un i cap punt per les derrotes, tot i que, en el passat, fins a la temporada 1995/96, les victòries únicament sumaven dos punts. El conjunt amb més punts a final de temporada es proclama vencedor i obté una plaça per la major competició Europea (*UEFA Women's Champions League*) i, aquells dos equips amb menys punts descendeixen a la segona divisió alemanya (2. Frauen-Bundesliga). (DFB, Informació de la lliga)

Els 12 equips que formaven part de la *Frauen-Bundesliga* durant la temporada 22/23: (DFB)

	Bayern München		Bayer 04 Leverkusen		1. FC Köln
	VfL Wolfsburg		SC Freiburg		MSV Duisburg
	Eintracht Frankfurt		SGS Essen		SV Meppen
	TSG Hoffenheim		SV Werder Bremen		Turbine Potsdam

Taula 3: Equips de la Frauen-Bundesliga (temporada 22/23).

6. Base de dades

Per desenvolupar els punts que vindran a continuació s'ha recollit al voltant de 130 variables i s'ha fet una selecció d'aquelles més rellevants per, finalment, aconseguir una base de dades d'un total de 1.020 registres i 57 variables. Els registres fan referència a estadístiques de les jugadores de tres grans lligues europees ([LaLiga F](#), [Women's Super League](#) i [Frauen-Bundesliga](#)) durant la temporada 22/23.

Les dades han estat extretes de la plataforma *FBref* ([Estadísticas e Historia del Fútbol | FBref.com](#)). FBref és la secció, creada el 2018 pel grup *Sports Reference*, dedicada al seguiment d'estadístiques d'equips de futbol i dels seus jugadors. Actualment, aquesta secció té una cobertura de 47 països i un total de 145 competicions que implica a un total de 202.375 jugadors i jugadores. Les dades són proporcionades per *Data Sports Group* i les estadístiques avançades provenen del grup *Opta Sports*¹⁶, experts en anàlisi de dades en els esports.

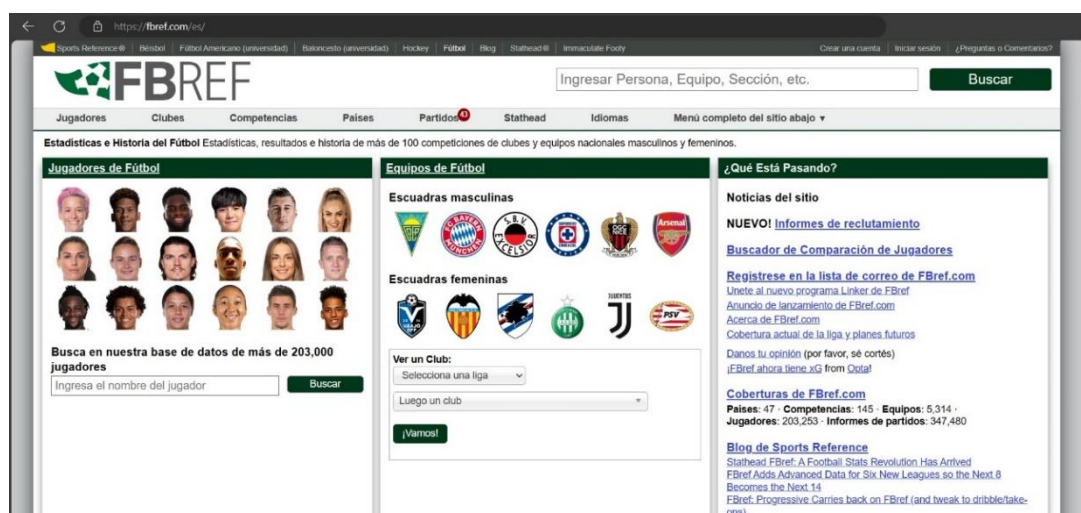


Figura 3: Pantalla inicial de la web fbref.com.

¹⁶ Opta Sports actualment forma part de *Perform Group*, empresa dedicada a la recollida de dades esportives i a l'anàlisi d'aquestes. (Stats Perform, 2021)

6.1. Descripció de les variables

Per poder conèixer millor les variables que integren la base de dades s'han classificat en els següents punts:

1. Característiques personals i disciplina
2. Variables rellevants en la posició de portera (PO)
3. Variables rellevants en les posicions de defensa (DEF)
4. Variables rellevants en les posicions del mig del camp (MED)
5. Variables rellevants en les posicions avançades o d'atacant (DEL)

Veiem ara quines variables inclouen cada un dels punts anteriors i la seva descripció¹⁷:

1. Característiques personals i disciplina

- **Jugadora:** Nom i cognom de la jugadora
- **Lliga:** Inicials del país de la lliga on juga
- **País:** País on s'ha nacionalitzat la jugadora
- **Posició:** Rol natural que desenvolupa la jugadora al camp, com a primera opció
- **Equip:** Club al qual pertanyia la jugadora durant la temporada 22-23
- **Any:** Any de naixement
- **Partits jugats:** Nombre total de partits jugats
- **Titular:** Nombre de partits on ha començat el partit en l'11 inicial
- **Minuts:** Nombre total de minuts disputats
- **Partits:** Nombre de minuts jugats entre 90, nombre de partits sencers
- **Faltes comeses:** Nombre d'infraccions comeses per la futbolista
- **Faltes rebudes:** Nombre d'infraccions comeses pel rival sobre la futbolista
- **Targetes grogues:** Nombre de sancions que ha aplicat l'àrbitre a cada jugadora
- **Targetes vermelles:** Nombre d'expulsions que ha aplicat l'àrbitre a cada jugadora

¹⁷ Veure metadata de l'annex. ([METADATA](#))

2. Variables rellevants en la posició de portera (PO)

- **Gols en contra:** Nombre de gols que ha rebut per part dels equips contraris
- **Aturades:** Nombre d'intercepcions de la pilota per evitar un gol
- **Percentatge Aturades:** Proporció d'aturades exitoses entre aturades intentades
- **Porteries a 0:** Quantitat de partits jugats on no ha rebut cap gol contrari
- **Percentatge de porteries a 0:** Percentatge de partits jugats amb cap gol rebut
- **Xuts en contra:** Nombre de xuts efectuats pels equips contraris entre els tres pals
- **Gols en contra esperats:** Nombre de gols que s'esperen de l'equip contrari després del xut. Dada calculada per *Opta*
- **Gols en contra esperats per partit:** Gols que s'esperen de l'equip contrari després del xut menys gols permesos entre 90 minuts. Dada calculada per *Opta*

3. Variables rellevants en les posicions de defensa (DEF)

- **Duels aeris guanyats:** Quantitat total de duels aeris guanyats
- **Duels aeris perduts:** Quantitat total de duels aeris perduts
- **Regatejadors abordats:** Nombre de dribleadors abordats
- **Regatejadors no abordats:** Nombre de dribleadors abordats sense èxit
- **Desafiaments intentats:** Nombre total d'enfrontaments intentats (suma dels regatejadors abordats amb èxit i els no abordats amb èxit)
- **Xuts bloquejats:** Quantitat de xuts que el jugador han bloquejat posant-se al mig
- **Passades bloquejades:** Quantitat de passades que s'han bloquejat posant-se al mig
- **Bloquejos:** Nombre de vegades que un jugador bloqueja la pilota posant-se al mig (és la suma dels bloquejos de xuts i els bloquejos de passades)
- **Rebuig:** Quantitat de vegades que s'ha allunyat la pilota de la zona pròxima a la porteria per neutralitzar una acció ofensiva de l'equip atacant

4. Variables rellevants en les posicions del mig del camp (MED)

- **Tocs:** Número de vegades que una jugadora toca la pilota. Rebre, driblar i passar-la a una companya conta com a 1 toc.
- **Distància total:** Distància total que una jugadora mou la pilota mentre està als seus peus.
- **Distància progressiva:** Distància total que una jugadora mou la pilota en direcció a la porteria contrària.
- **Rebudes:** Quantitat de passades que una jugadora rep, amb èxit, d'una companya
- **Recuperacions:** Nombre de pilotes soltes que recupera i, per tant, suposa tenir la possessió de la pilota altre cop
- **Controls progressius:** Nombre de vegades que es mou la pilota direcció a la porteria contrària
- **Passades progressives:** Quantitat de passades completades en direcció a la porteria contrària
- **Bloquejats:** Nombre de jugadors bloquejats, abordats.
- **Bloquejos aconseguits:** Quantitat de vegades que s'ha aconseguit un abordatge i ha permès recuperar la possessió de la pilota
- **Intercepcions:** Nombre de cops que s'impedeix que la pilota llençada pel contrari arribi al destí, tallant o desviant la seva trajectòria

5. Variables rellevants en les posicions avançades o d'atacant (DEL)

- **Enfrontaments:** Nombre d'intents d'enfrontament a un defensor contrari mentre intenta avançar, regatejant, cap a la porteria
- **Enfrontaments exitosos:** Nombre de jugadors enfrontats amb èxit
- **Rebudes progressives:** Quantitat de passades rebudes quan la jugadora es movia direcció a la porteria contrària
- **Fores de joc:** Vegades que la jugadora es troba en posició avançada. En el moment del llançament de la pilota, la jugadora es troba més a prop de la porteria contrària que qualsevol jugadora de l'altre equip (a excepció de la portera) i també més a prop que la pilota
- **Xuts a porteria:** Quantitat de vegades que la jugadora tira a porteria contrària
- **Xuts entre els 3 pals:** Nombre de vegades que la jugadora tira la pilota entre els 3 pals, contenen els gols i els xuts que són parats per alguna jugadora de l'equip contrari
- **Xuts a porteria per partit:** Nombre de llançaments efectuats repartit entre 90, els minuts d'un partit
- **Gols:** Quantitat de gols marcats
- **Assistències:** Nombre de passades que han implicat que alguna companya d'equip marqui un gol en porteria contrària
- **Gols esperats:** Nombre de gols que s'esperen. Dada calculada per *Opta*
- **Assistències esperades:** Nombre d'assistències que s'esperen al llarg de la temporada. Dada calculada per *Opta*
- **Gols per partit:** Quantitat de gols marcats per partit
- **Assistències per partit:** Nombre de passades que han implicat que alguna companya d'equip marqui un gol en porteria contrària per partit
- **Gols esperats per partit:** Nombre de gols que s'esperen per partit. Dada calculada per *Opta*
- **Assistències esperades per partit:** Nombre d'assistències que s'esperen per partit. Dada calculada per *Opta*

6.2. Estadística descriptiva i preprocessament

El conjunt de dades hi ha total de 1020 registres i 57 variables, 5 d'elles són categòriques i la resta, 52, són numèriques.

Primer de tot, comprovarem si cada variable té assignada la classe correcta i localitzem els valors absents (*missings*), per això, la següent taula ens mostra el nom de la variable amb una petita descripció del que mesuren o classifiquen, la seva categoria, la quantitat de valors faltant que tenen i la proporció que aquests suposen. Amb el fons gris, veiem aquells valors que hem de revisar i, si cal, modificar:

Taula 4: Variables de la BBDD, descripció, classe, valors absents i percentatges que suposen.

Variable	Descripció	Classe	NAs	propNA
Jugadora	Nom i cognom	character	0	0%
Lliga	Lliga	character	0	0%
País	País de nacionalització	character	0	0%
Posc1	Posició natural	character	0	0%
Equip	Equip	character	0	0%
Any	Any de naixement	numeric	0	0%
PJ	Partits jugats	numeric	0	0%
Titular	Titular	numeric	0	0%
Min	Minuts jugats	numeric	0	0%
90s	Partits sencers	numeric	0	0%
FC	Faltes comeses	numeric	0	0.1%
FR	Faltes rebudes	numeric	0	0.1%
TG	Targetes grogues	numeric	0	0%
TV	Targetes vermelles	numeric	0	0%
GC	Gols en contra	numeric	0	0%
DaPC	Xuts en contra	numeric	1	0.1%
Parades	Aturades	numeric	1	0.1%
Parades%	% Aturades	numeric	2	0.2%
PaC	Porteries a 0	numeric	0	0%
PaC%	% Porteries a 0	numeric	3	0.29%
PSxG	Gols esperats en contra	numeric	1	0.1%
PSxG/90	Gols esperats en contra per partit	numeric	1	0.1%
Ganados	Duels aeris guanyats	numeric	1	0.1%

Variable	Descripció	Classe	NAs	propNA
Perdidos	Duels aeris perduts	numeric	1	0.1%
TklDf	Regatejadors abordats	numeric	1	0.1%
AttDf	Desafiaments intentats	numeric	1	0.1%
Perdida	Desafiaments perduts	numeric	1	0.1%
Bloqueos	Bloquejos de pilota	numeric	1	0.1%
Dis	Xuts bloquejats	numeric	1	0.1%
Pases	Passades bloquejades	numeric	1	0.1%
Desp	Rebuig	numeric	1	0.1%
Toc	Tocs	numeric	1	0.1%
Dist.tot	Distància total	numeric	1	0.1%
Dist.prg	Distància progressiva	numeric	1	0.1%
Rec	Passades rebudes	numeric	1	0.1%
Recup	Recuperacions	numeric	1	0.1%
PrgC	Controls progressius	numeric	1	0.1%
PrgP	Passades progressives	numeric	1	0.1%
Tkl	Bloquejats	numeric	1	0.1%
TklG	Bloquejos aconseguits	numeric	1	0.1%
Int	Intercepcions	numeric	1	0.1%
Att	Intents d'enfrontament	numeric	1	0.1%
Succ	Defenses enfrontats amb èxit	numeric	1	0.1%
PrgR	Passades progressives rebudes	numeric	1	0.1%
PA	Posicions avançades	numeric	1	0.1%
DisC	Xuts a porteria	numeric	1	0.1%
DaP	Xuts entre els 3 pals	numeric	1	0.1%
TalArc/90	Xuts a porteria per cada 90 min	numeric	1	0.1%
Gls	Gols	numeric	0	0%
Ass	Assistències	numeric	0	0%
xG	Gols esperats	numeric	1	0.1%
xAG	Assistències esperades	numeric	1	0.1%
G-xG	Diferència entre gols i gols esperats	numeric	1	0.1%
Gls/90	Gols en 90 min	numeric	0	0%
Ass/90	Assistències en 90 min	numeric	0	0%
xG/90	Gols esperats en 90 min	numeric	1	0.1%
xAG/90	Assistències esperades en 90 min	numeric	1	0.1%

6.2.1. NAs

Analitzem els valors absents de les dades, a la següent taula podem veure a quines files es concentren els valors faltants:

Fila	Quantitat NAs
313	1
496	5
497	33
606	2
631	1

Taula 5: Valors absents per fila (jugadora).

La majoria d'aquests valors es troben a la fila 497 que correspon a la portera de la lliga alemanya:

Jugadora <chr>	Lliga <chr>	País <chr>	Posc1 <chr>	Equip <chr>	Any <dbl>	PJ <dbl>	Titular <dbl>	Min <dbl>	90s <dbl>	...
Vanessa Fischer	GER	DEU	PO	Turbine Potsdam	1997	3	3	270	3	...

Taula 6: Primeres variables de la jugadora Vanessa Fischer, amb el valor més alt de missings.

Com que aquest registre no aportaria informació significativa a 33 variables, d'un total de 57, s'opta per obviar-la i, per tant, la base de dades constarà de 57 columnes i 1.019 files.

Pel que fa a la resta de *missings*, com que son variables numèriques, podem aplicar l'algoritme de kNN per imputar, aquest algoritme té en compte els *k* veïns més propers al valor faltant.

Comprovem que ara ja no queda cap valor faltant per poder seguir amb els següents punts:

Any	PJ	Titular	Min	90s	FC	FR	TG	TV
0	0	0	0	0	0	0	0	0
GC	DaPC	Parades	PaC	PaC%	PSxG	PSxG/90	Ganados	Perdidos
0	0	0	0	0	0	0	0	0
Tk1Df	AttDf	Perdida	Bloqueos	Dis	Pases	Desp	Toc	Dist.tot
0	0	0	0	0	0	0	0	0
Dist.prg	Rec	Recup	PrgC	PrgP	Tk1	Tk1G	Int	Att
0	0	0	0	0	0	0	0	0
Succ	PrgR	PA	DisC	DaP	TalArc/90	Gls	Ass	xG
0	0	0	0	0	0	0	0	0
xAG	G-xG	Gls/90	Ass/90	xG/90	xAG/90			
0	0	0	0	0	0			

Taula 7: Comprovació de que no queda cap valor faltant a les variables numèriques després d'aplicar KNN.

6.2.2. Variables categòriques

A la columna *Classe* tenim 4 caselles marcades, corresponen a *Lliga*, *Pais*, *Posc1* i *Equip*, aquestes variables són categòriques amb una quantitat finita d'opcions, per això, canviarem la categoria de *character* a *factor*.

- Lliga

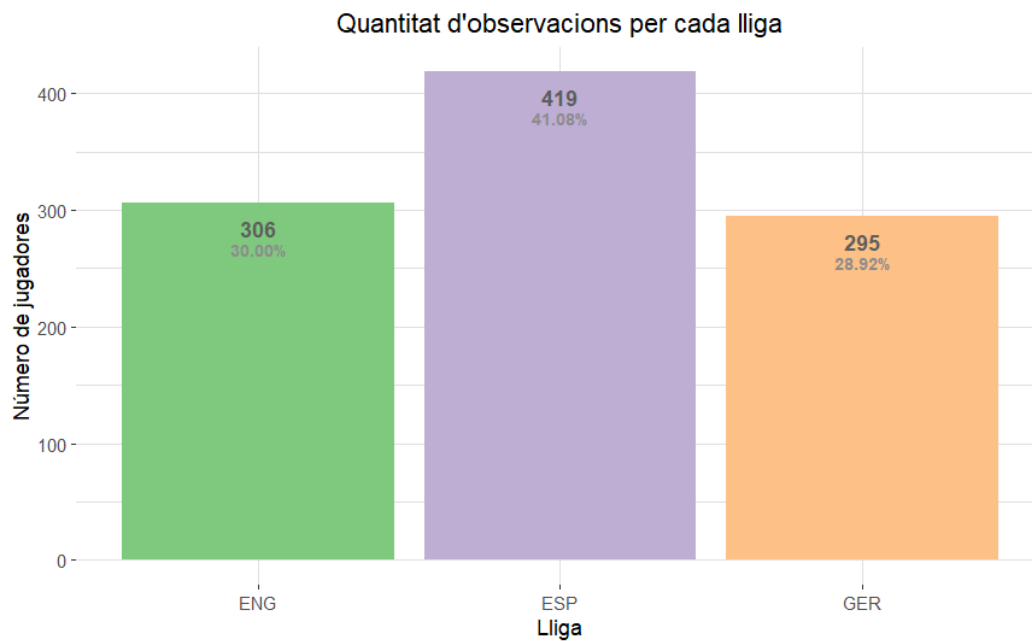


Figura 4: Distribució de la variable Lliga.

La lliga espanyola (color lila) és la que concentra el nombre més gran de jugadores, amb un 41,08%, això és degut al fet que hi ha un nombre més gran d'equips (16 equips a diferència de les altres dues lligues que només formen la lliga 12 equips, cada una). La lliga anglesa (color verd) concentra un 30% de les jugadores i, la lliga alemanya (color taronja) recull el 28,92% restant.

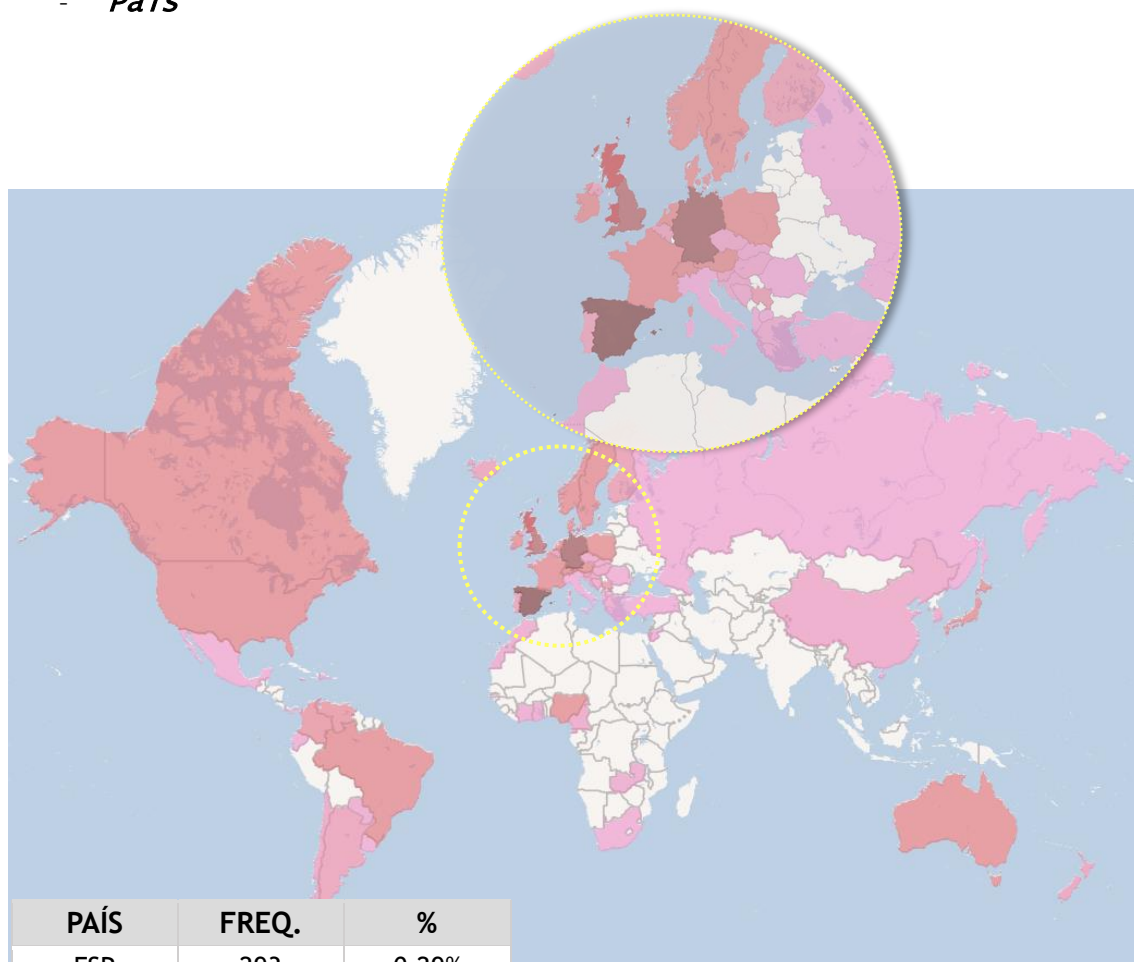
- **Equip**

UDG Tenerife	Barcelona	Turbine Potsdam	Leicester City	Levante Planas	
33	31	31	30	30	
Valencia	Villarreal	Arsenal	Liverpool	Real Betis	
29	28	27	27	27	
Werder Bremen	Aston Villa	Leverkusen	MSV Duisburg	Tottenham	
27	26	26	26	26	
West Ham	Athama	Athletic Club	Bayern Munich	Chelsea	
26	25	25	25	25	
Levante	Real Madrid	Atlético Madrid	Brighton	Freiburg	
25	25	24	24	24	
Hoffenheim	Manchester City	Manchester Utd	Reading	Real Sociedad	
24	24	24	24	24	
Sevilla	Sporting Huelva	Wolfsburg	Eint Frankfurt	Everton	
24	24	24	23	23	ESP
Madrid CFF	Alavés	Essen	SV Meppen	Köln	ENG
23	22	22	22	21	GER

Taula 8: Distribució del nombre de jugadores que hi ha a cada equip

L'equip de Tenerife, també conegut com a "Granadilla", és l'equip que més jugadores (33) ha tingut durant la temporada 22/23, el Barcelona i el Turbine Potsdam el segueixen amb un total de 31 jugadores. A l'altra part de la taula trobem al Köln amb 21 jugadores, la quantitat mínima, i, seguidament, amb un conjunt de 22 jugadores inscrites, hi ha l'Alavés, l'Essen i el SV Meppen.

- **País**



PAÍS	FREQ.	%
ESP	293	0.29%
GER	206	0.20%
ENG	145	0.14%
NOR	19	0.02%
AUT	18	0.02%
SUI	18	0.02%
BRA	17	0.02%
SWE	17	0.02%
USA	17	0.02%
FRA	16	0.02%
Altres (56)	254	0.25%
TOTAL	1.020	100%

Figura 5: Mapa coroplètic on es pot veure la freqüència de la variable País.

Un color més fort implica més freqüència.

Hi ha un total de 66 nacionalitats diferents i, com podíem esperar, les 3 nacionalitats més freqüents són l'espanyola (293), l'anglesa (206) i l'alemanya (145), això és degut al fet que la majoria d'equips es formen amb jugadores del mateix país. Les segueixen la nacionalitat noruega (19), austríaca (18) i suïssa (18) però ja amb menys freqüència.

Taula 9: Freqüències de la variable País i els percentatges que suposen.

- Posició

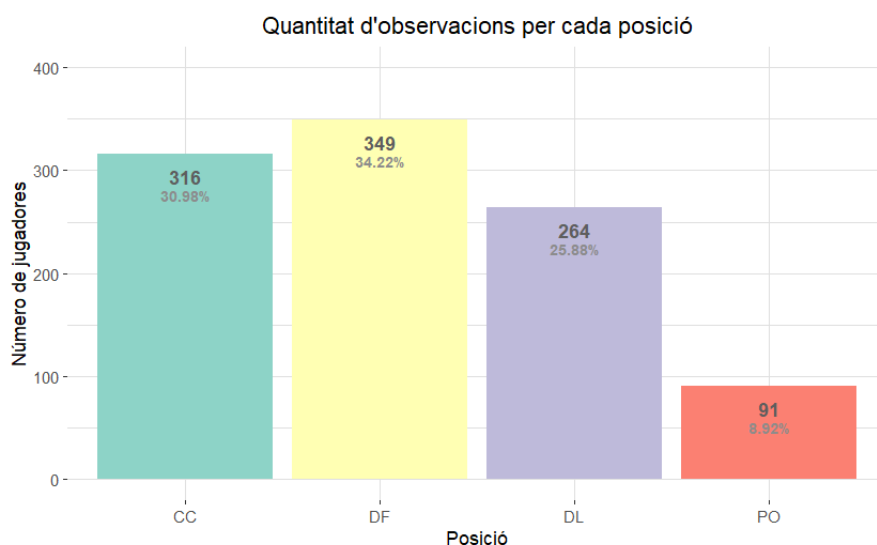


Figura 6: Distribució de la variable Posició.

La posició natural de defensa és un 34% del total de les jugadores, seguit de les migcampistes, amb un 31%, després són les davanteres i, com era d'esperar, darrerament es troben les porteres, que són prop d'un 9%.

El gràfic anterior es pot entendre fàcilment tenint en compte que una de les formacions més clàssiques i conegudes al món és la 4-4-3, és a dir, en l'onze inicial, es col·loquen quatre defenses, quatre centrecampistes i dues davanteres.

Aquesta formació proporciona un bon equilibri entre la defensa i l'atac i, per això es considera una de les millors alineacions actualment. (Ertheo, 2023)

Aplicar aquesta formació fa que els equips tinguin la necessitat de tenir més jugadores en les posicions defensives i centrals respecte a les davanteres.

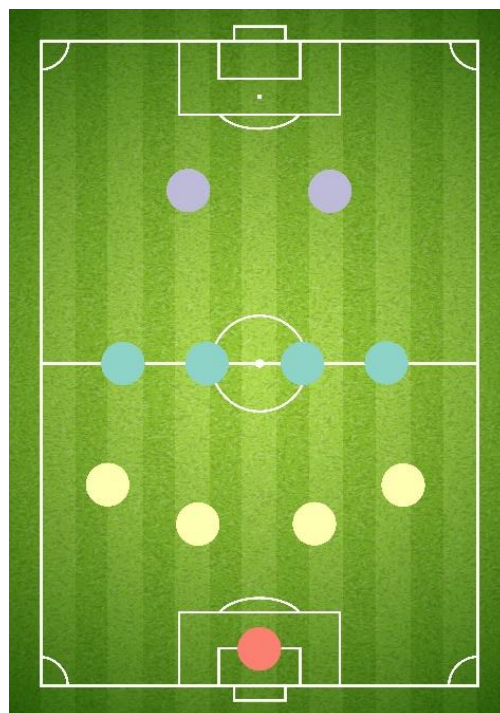


Figura 7: Representació de la formació 4-4-3.
Fons de la imatge: Freepik.com

6.2.3. Variables numèriques

Al *Metadata* de l'annex s'hi pot trobar un recull de les variables i, per cada una que és numèrica hi consta el valor mínim, el màxim, la mitjana i la desviació típica.

A continuació, es destaquen alguns gràfics per poder comparar les tres lligues:

1. Any de naixement

Les jugadores més joves es troben a la lliga alemanya, amb una mitjana es troba en l'any 1998. Tot i això, la diferencia amb la lliga anglesa, la que té una mitjana més gran, és només d'un any.

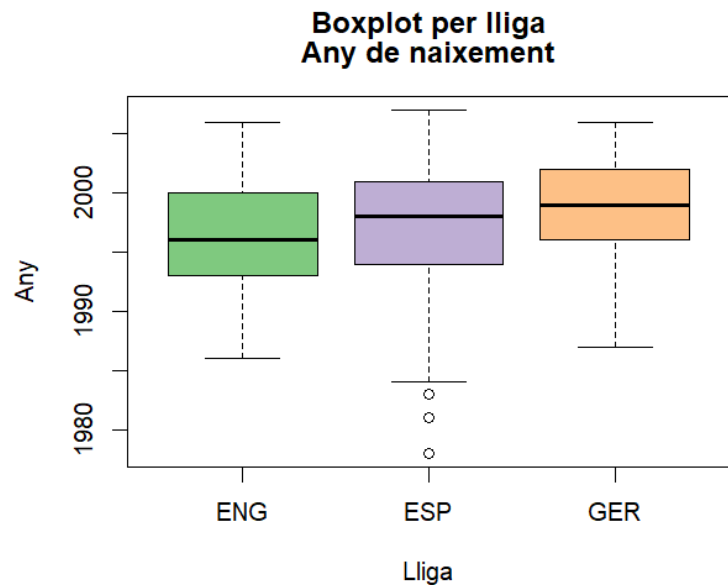


Figura 8: Boxplot any de naixement per cada lliga.

2. Partits i minuts jugats

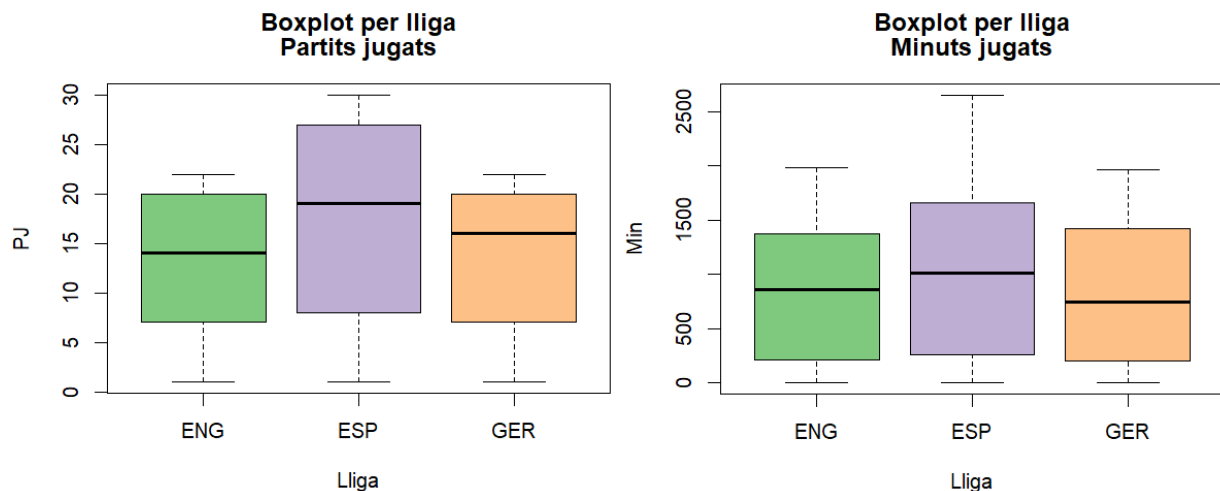


Figura 9: Boxplots del nombre de partits jugats i del nombre de minuts jugats per cada lliga.

Com és d'esperar, la lliga espanyola obté valors més elevats en el nombre de partits jugats i, conseqüentment en el nombre de minuts jugats, ja que la seva competició és més extensa que les altres. Tot i això, es pot destacar que els percentils inferiors es troben en els mateixos valors en les tres lligues.

Pels següents gràfics s'han agafat cada posició per separat, d'aquesta manera podrem veure les diferències entre lligues de les variables rellevant per la posició de portera, la de defensa, la de migcampista i la de davantera sense que interfereixin els valors de les altres posicions.

3. Portera: Gols en contra i Xuts en contra

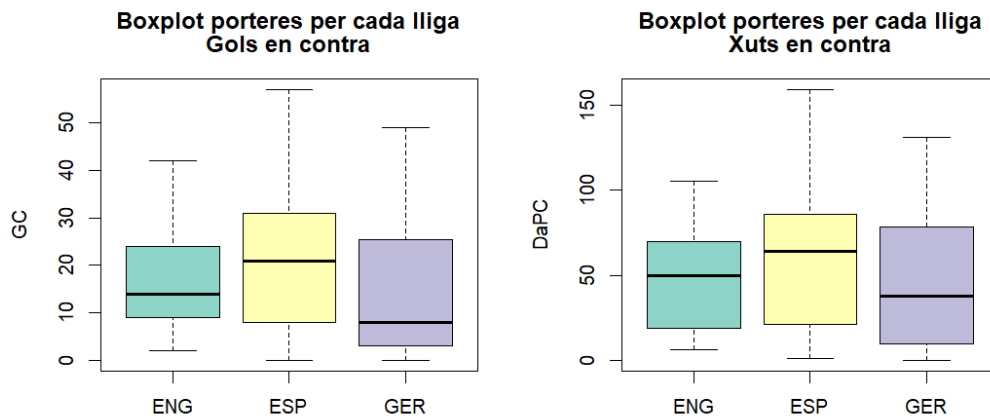


Figura 10: Gols en contra i Xuts en contra de les porteres de cada lliga.

La lliga alemanya és la que, de mitjana rep menys gols en contra, tot i així, és la lliga que també rep menys xuts de mitjana i, al contrari, la lliga espanyola és la que rep més gols de mitjana i també més xuts.

4. Defensa: Desafiaments aconseguits i perduts

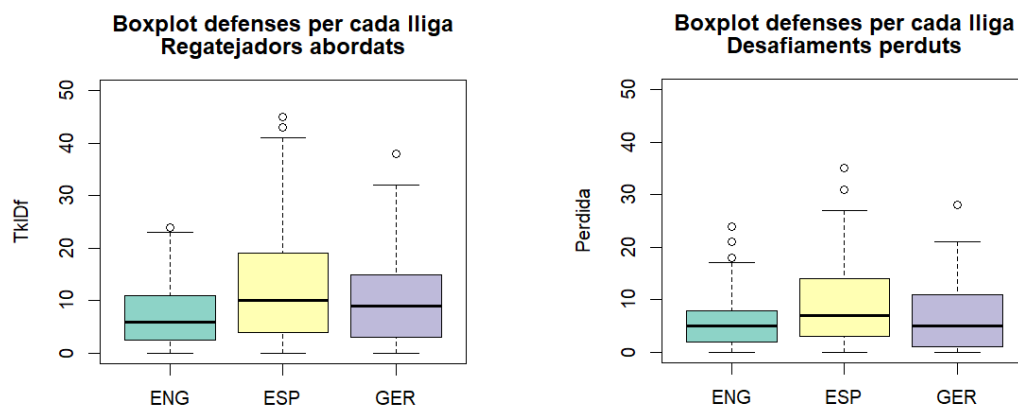


Figura 11: Regatejadors abordats i intents perduts de les defenses per cada lliga.

Podem veure que l'allargada entre quartils és major en els intents aconseguits (regatejadors abordats), les mitjanes en aquests *boxplots* també són majors, tot i així, no es veu gran diferència entre lligues.

7. Implementació dels algoritmes

7.1. ACP

Abans de començar a aplicar l'algoritme de PCA hem d'assegurar-nos que no existeixen valors faltants a les variables que tindrem en compte, aquestes variables únicament podran ser numèriques. Això implica un total de 52 variables i un total de 1020 registres.

A continuació podem veure gràficament els nivells de correlacions entre variables:

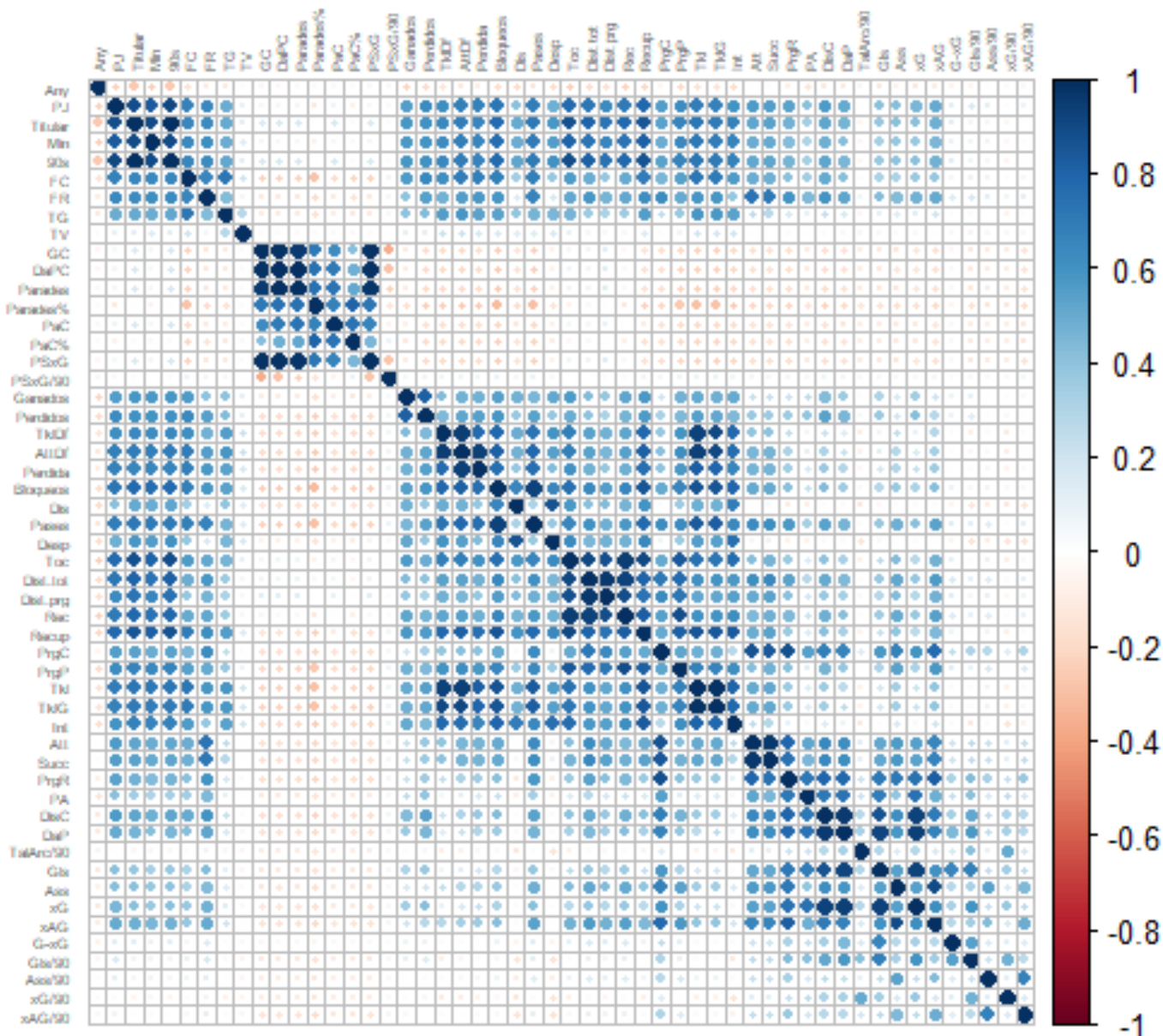


Figura 12: Representació dels valors de les correlacions entre variables.

Aquelles variables que tenen un color més fort i un símbol més gran tenen una alta correlació entre si. A més, aquells símbols que són de color blau, tenen una correlació positiva mentre que si el color és vermell, la correlació és negativa. La matriu és simètrica, per tant, els valors que hi ha a un costat i a l'altre de la diagonal són iguals.

Una vegada comprovat que hi ha variables correlacionades i, per tant, és adient reduir la dimensionalitat de la base de dades, calculem els valors de les components.

Elecció de les components

Per saber el nombre mínim de components que hem d'escollir per poder tenir suficient informació explicada, utilitzem la inèrcia acumulada. Ens quedarem amb aquelles components que expliquen un mínim del 80%.

Aquesta quantitat se supera a la component 9, com podem veure en el següent gràfic:

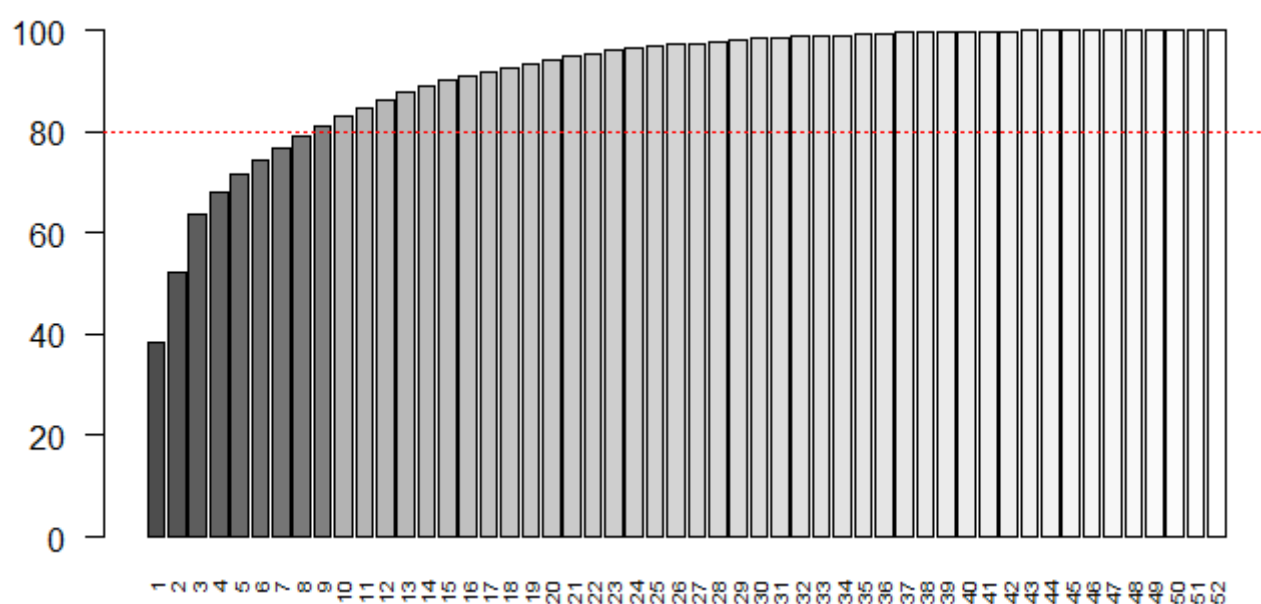


Figura 13: Quantitat d'inèrcia acumulada per cada component.

Inèrcia	C1	C2	C3	C4	C5	C6	C7	C8	C9
Simple	38,40%	13,79%	11,31%	4,46%	3,70%	2,72%	2,48%	2,35%	2,05%
Acumulada	38,40%	52,20%	63,51%	67,97%	71,66%	74,39%	76,86%	79,21%	81,26%

Taula 10: Percentatges d'inèrcia i inèrcia acumulada per cada variable.

Veiem que la primera component aporta un 38,4% de variància, la segona i la tercera ho fan amb un 13,8% i un 11,3%, respectivament i de la quarta a la novena, per arribar al mínim de 80% acumulat, ho fan amb valors entre 2% i 5%.

Veiem ara com estaran explicades cada una de les components. Aquests seran els vectors propis de la primera component:

- **Primera Component:**

$$\begin{aligned}
 Y_1 = & (-0,0478)Any + (0,192)PJ + (0,1952)Titular + (0,1883)Min + \\
 & (0,1954)90s + (0,1723)FC + (0,1671)FR + (0,1261)TG + (0,0301)TV + \\
 & (-0,0469)GC + (-0,0484)DaPC + (-0,0479)Parades + (-0,0699)Parades\% + \\
 & (-0,0401)PaC + (-0,0514)PaC\% + (-0,0476)PSxG + (0,0187)PSxG/90 + \\
 & (0,1376)Ganados + (0,1494)Perdidos + (0,1675)TklDf + (0,1793)AttDf + \\
 & (0,1708)Perdida + (0,1955)Bloqueos + (0,1044)Dis + (0,1924)Pases + (0,1157)Desp + \\
 & (0,1902)Toc + (0,1841)Dist.tot + (0,1607)Dist.prg + (0,1812)Rec + (0,2013)Recup + \\
 & (0,1636)PrgC + (0,1773)PrgP + (0,19)Tkl + (0,1873)TklG + (0,1593)Int + \\
 & (0,1582)Att + (0,1602)Succ + (0,1479)PrgR + (0,0989)PA + (0,1541)DisC + \\
 & (0,1383)DaP + (0,0124)TalArc/90 + (0,1178)Gls + (0,1312)Ass + (0,1251)xG + \\
 & (0,1486)xAG + (0,0451)G - xG + (0,0482)Gls/90 + \\
 & (0,0427)Ass/90 + (0,0128)xG/90 + (0,037)xAG/90
 \end{aligned}$$

Els primer valors fan referència a les variables de característiques personals i disciplina, les variables marcades amb el requadre vermell fan referència a aquelles rellevants en la posició de portera, les variables en el requadre groc son les rellevants en la posició de defensa, les del requadre verd son les rellevants en la posició de migcampista i, les últimes, referents a la posició d'atacant.

El grup de variables que més aporta en aquesta component són les referents al mig del camp, concretament, la que més informació aporta és la variable *Recup*, aquesta variable emmagatzema el nombre de pilotes soltes recuperades per la jugadora. A més, aquest conjunt de variables té càrregues positives, per tant, una jugadora tindrà puntuacions baixes en aquesta component quan tingui valors baixos en aquestes variables i, al contrari, puntuacions altes quan els valors siguin elevats.

Per tant, podríem dir que Y_1 avalua el nivell de les jugadores del centre del camp, que tant han de tenir actituds defensives com ofensives.

- **Segona Component:**

$$\begin{aligned}
 Y_2 = & (-0,0404)Any + (0,0512)PJ + (0,1038)Titular + (0,0944)Min + \\
 & (0,1084)90s + (0,0293)FC + (-0,049)FR + (0,1007)TG + (0,069)TV + \\
 & (0,136)GC + (0,1379)DaPC + (0,138)Parades + (0,1151)Parades\% + \\
 & (0,1169)PaC + (0,0893)PaC\% + (0,1357)PSxG + (-0,0364)PSxG/90 + \\
 & (0,0625)Ganados + (0,0116)Perdidos + (0,1431)TkIDf + (0,124)AttDf + \\
 & (0,0874)Perdida + (0,0844)Bloqueos + (0,1911)Dis + (0,0095)Pases + (0,2111)Desp + \\
 & (0,1286)Toc + (0,0442)Dist.tot + (0,0722)Dist.prg + (0,0565)Rec + (0,1179)Recup + \\
 & (-0,1518)PrgC + (0,0324)PrgP + (0,1079)Tkl + (0,1038)TklG + (0,1814)Int + \\
 & (-0,1376)Att + (-0,1254)Succ + (-0,2211)PrgR + (-0,2078)PA + (-0,2117)DisC + \\
 & (-0,235)DaP + (-0,1615)TalArc/90 + (-0,2441)Gls + (-0,1654)Ass + (-0,2343)xG + \\
 & (-0,1776)xAG + (-0,1442)G - xG + (-0,2321)Gls/90 + \\
 & (-0,1396)Ass/90 + (-0,1726)xG/90 + (-0,1698)xAG/90
 \end{aligned}$$

Els colors dels requadres segueixen fent referencia igual que abans; característiques personals i disciplina les primeres variables, segueixen les variables destacades en les porteres, després les rellevants del centre del camp i, finalment, les que fan destacar a les davanteres.

Destaquen totes les càrregues negatives del grup final, són les que més informació aporten, tot i això, la variable *Desp* és la que individualment més aporta a la segona component. Els valors negatius de les variables finals ens indiquen que com més alt sigui el valor d'aquestes variables més baixa serà la puntuació de la component i, contràriament, com més baix sigui el valor, més puntuarà. A més, les variables defensives també aporten un gran valor a la component.

Així doncs, podem dir que Y_2 ens indicarà el nivell ofensiu de la jugadora.

- Tercera Component:

$$\begin{aligned}
 Y_3 = & (0,0398)Any + (-0,0876)PJ + (-0,1257)Titular + (-0,0735)Min + \\
 & (-0,1273)90s + (0,0584)FC + (-0,0342)FR + (0,0585)TG + (0,0426)TV + \\
 & (-0,3432)GC + (-0,3541)DaPC + (-0,3544)Parades + (-0,3)Parades\% + \\
 & (-0,2981)PaC + (-0,2346)PaC\% + (-0,3483)PSxG + (0,0939)PSxG/90 + \\
 & (0,0175)Ganados + (0,0359)Perdidos + (0,1002)Tkldf + (0,0932)AttDf + \\
 & (0,074)Perdida + (0,0777)Bloqueos + (0,1037)Dis + (0,0452)Pases + (0,0753)Desp + \\
 & (-0,0698)Toc + (-0,1184)Dist.tot + (-0,1273)Dist.prg + (-0,0683)Rec + \\
 & (0,0283)Recup + (-0,0694)PrgC + (0,0193)PrgP + (0,091)Tkl + \\
 & (0,0891)TklG + (0,1106)Int + \\
 & (-0,0547)Att + (-0,0541)Succ + (-0,0892)PrgR + (-0,088)PA + (-0,0872)DisC + \\
 & (-0,1024)DaP + (-0,0137)TalArc/90 + (-0,1124)Gls + (-0,0771)Ass + \\
 & (-0,1039)xG + (-0,0794)xAG + (-0,0746)G - xG + (-0,0628)Gls/90 + \\
 & (-0,0253)Ass/90 + (-0,0199)xG/90 + (-0,0126)xAG/90
 \end{aligned}$$

Altres cop, les primeres variables són les referents a característiques personals i disciplina, en color vermell les referents a les porteres, en groc a les defenses, en verd a les mig campistes i l'últim grup a les atacants.

Veiem que la majoria dels vectors propis són amb càrrega negativa i, els que no tenen valors força petits. Això implica que com més alts siguin els valors de les variables, més baixa serà la puntuació i al revés.

Les variables que més informació aporten a aquesta component són les del requadre vermell, és a dir, les rellevants en les porteres, concretament, la que més aporta és la variable *Parades*, que recull el nombre de parades que ha realitzat una portera

En aquest cas, la variable Y_3 ens indicarà les característiques de les porteres però recollirà la puntuació al revés, és a dir, com més baixes siguin les variables, més alta serà la puntuació.

7.2. Clustering

Separem les jugadores per cada posició i, per tant, hi haurà un *clustering* per cada una, 4 en total:

- Posició de PORTERA

Depèn per on fem el tall en el dendrograma tindrem unes solucions o altres. La línia blava ens donarà de resultat 3 clústers, la vermella en seran 4 i, la violeta, ens en donarà 5.

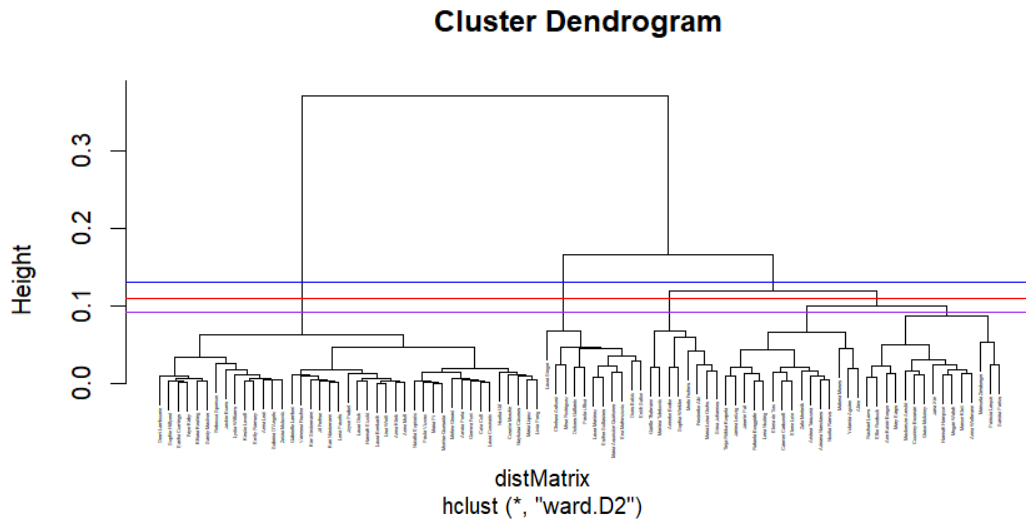


Figura 14: Dendrograma inicial Portera

Veiem el número de clústers òptim segons la suma de la distància al quadrat entre els punts dins del grup i el centroide d'aquest (wss¹⁸):

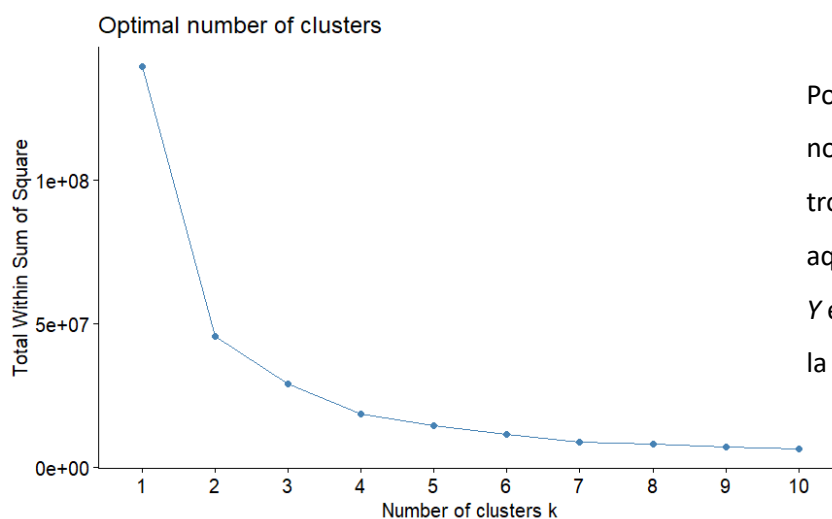


Figura 15: Visualització del nombre de clústers òptim Portera.

Podem veure al gràfic que el nombre òptim de clústers es trobaria entre el 4 i el 5, ja que, en aquells punts, la diferencia en l'eix Y és menor, és a dir, la pendent de la corba és més petita.

¹⁸ De l'anglès: *Within-Sums of Squares*

Finalment, s'ha escollit k igual a 5. El nombre de porteres en cada clúster serà el següent:

Taula 11: Nombre d'observacions per cada clúster Portera

Clúster	C1	C2	C3	C4	C5	TOTAL PO
Nombre d'observacions	15	11	41	8	15	90

Veiem ara l'estructura que queda al dendrograma amb els grups marcats en colors:

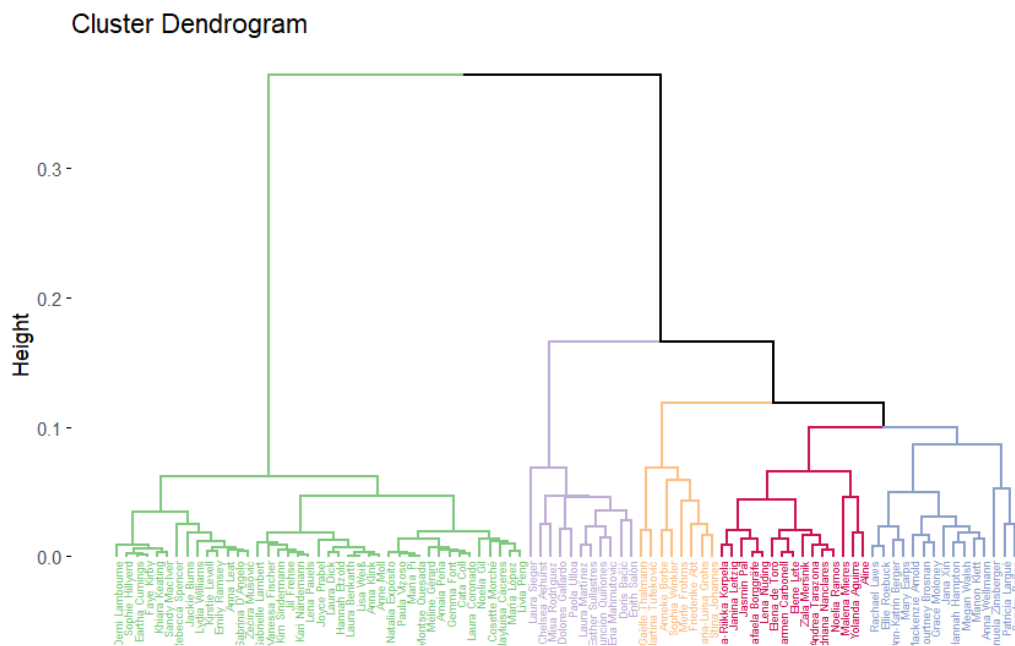


Figura 16: Dendrograma final Portera

Ampliant un dels grups podem veure els noms de totes les jugadores que l'integren, aquestes seran les més properes pel que fa a les variables estudiades.

Subtree 5

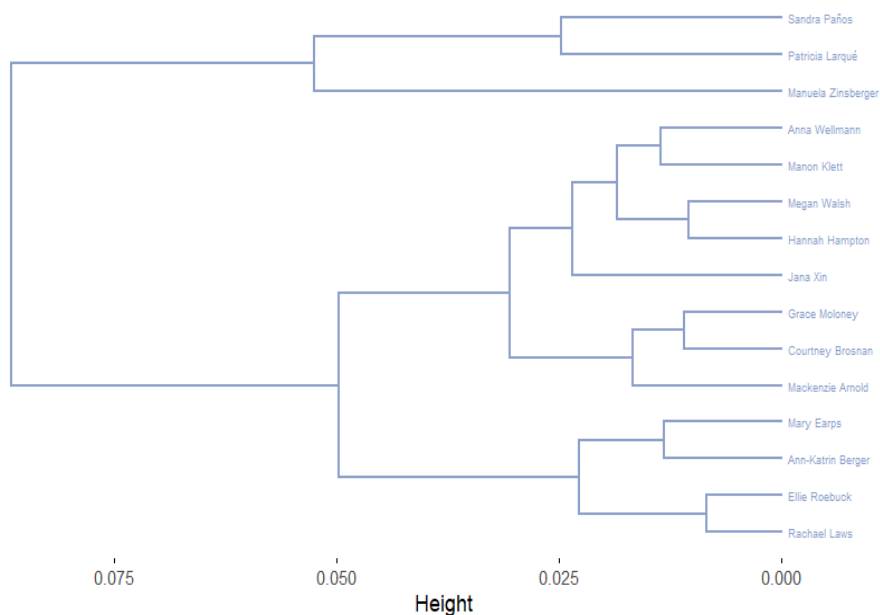


Figura 17: Dendrograma ampliat del clúster 2 Portera.

- Posició de DEFENSA

Segons on fem el tall del dendrograma obtindrem un nombre o altre de clústers. A primera vista, en aquest cas, podríem en 3 clústers en el cas de la línia blava, 4 en la vermella, 5 en la violeta i fins a 7 amb la de color verd.

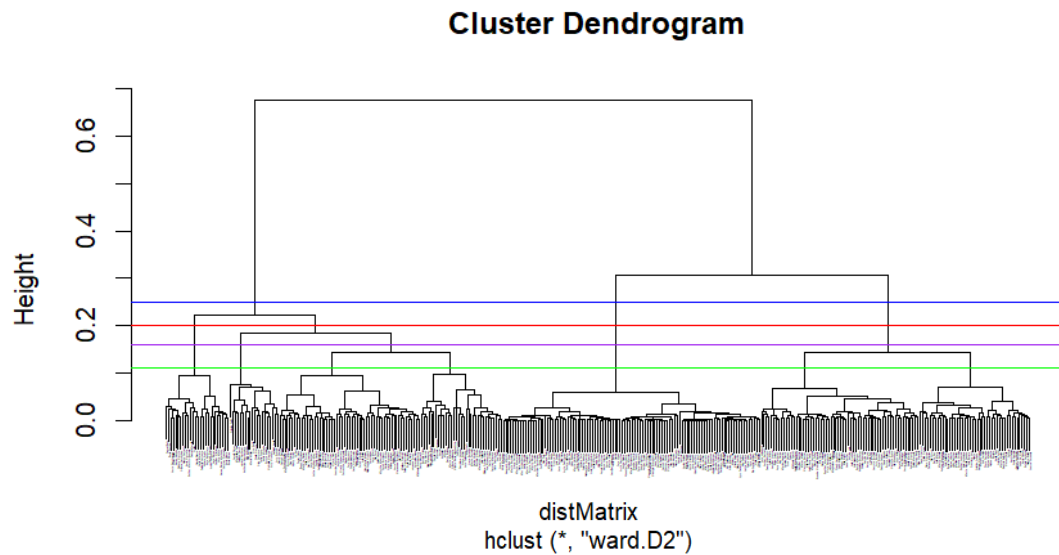


Figura 18: Dendrograma inicial Defensa.

Veiem el número de clústers òptim segons la suma de la distància al quadrat, igual que en el cas anterior:

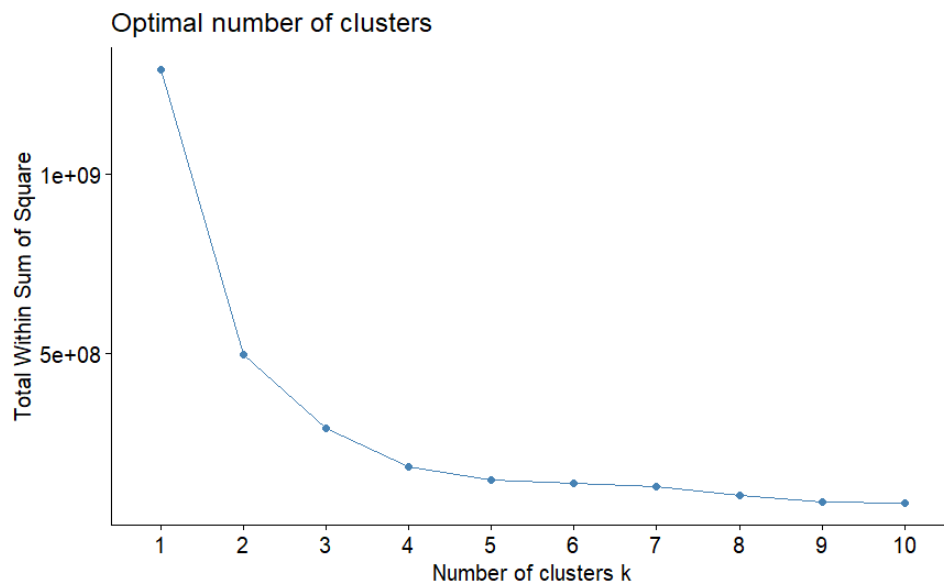


Figura 19: Visualització del nombre de clústers òptim Defensa.

El nombre òptim també es trobaria entre el 4 i el 5, ens quedem amb 5 grups.

El nombre de jugadores de posició defensa es reparteixen en cada clúster de la següent manera:

Taula 12: Nombre d'observacions per cada clúster Defensa

Clúster	C1	C2	C3	C4	C5	TOTAL DF
Nombre d'observacions	104	109	90	26	20	349

Veiem la distribució gràficament:

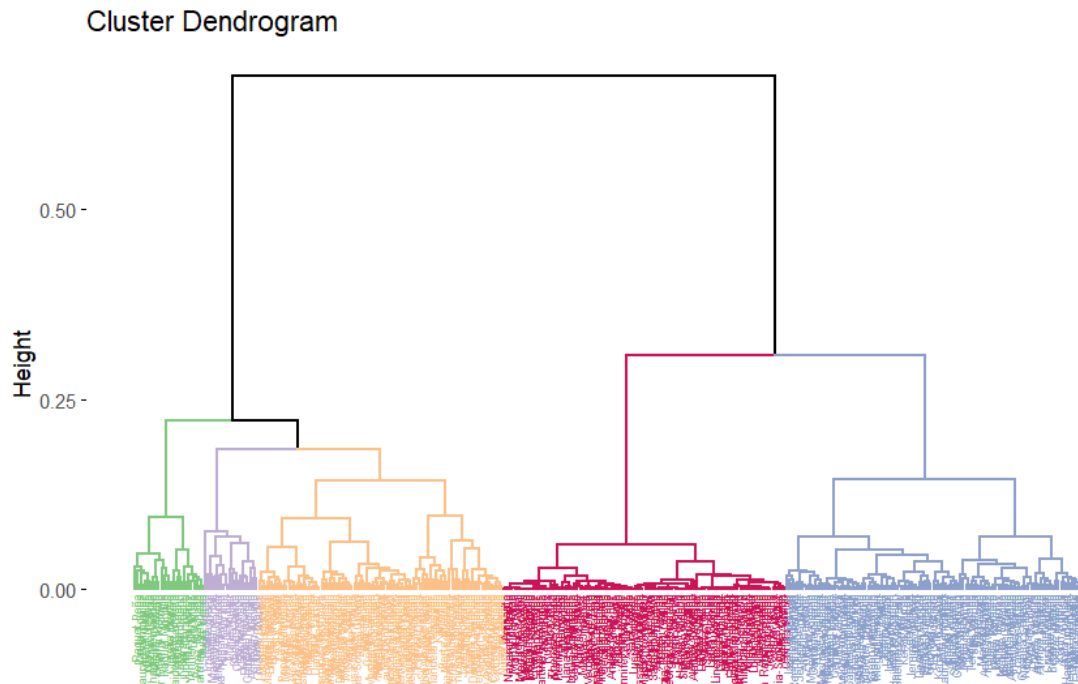


Figura 20: Dendrograma final Defensa

Aquest cop, com que hi ha moltes més jugadores, no es poden distingir els noms en el gràfic general, veiem un dels clústers de prop, si les podem distingir:

Subtree 1

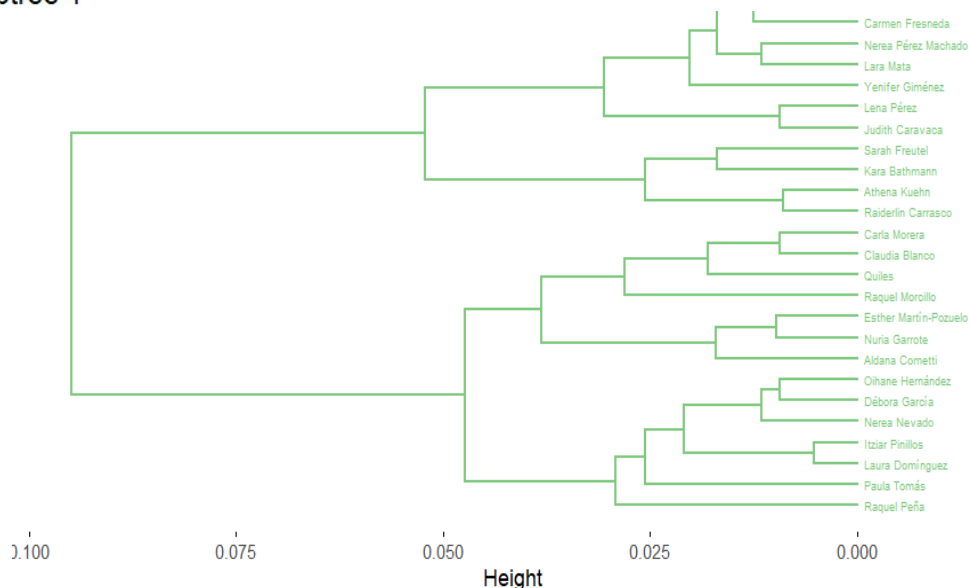


Figura 21: Dendrograma ampliat del clúster 4 Defensa.

- Posició de MIGCAMPISTA

Segons el tall on es faci del dendrograma obtindrem diferents opcions de clústers, en aquest cas podem tenir 4 clústers en el tall de la línia blava, tot i que es vegi molt just, 5 en la vermella, i fins a 6 en la violeta, podrien ser més si anéssim baixant les línies però ja no es veuen tant clar.

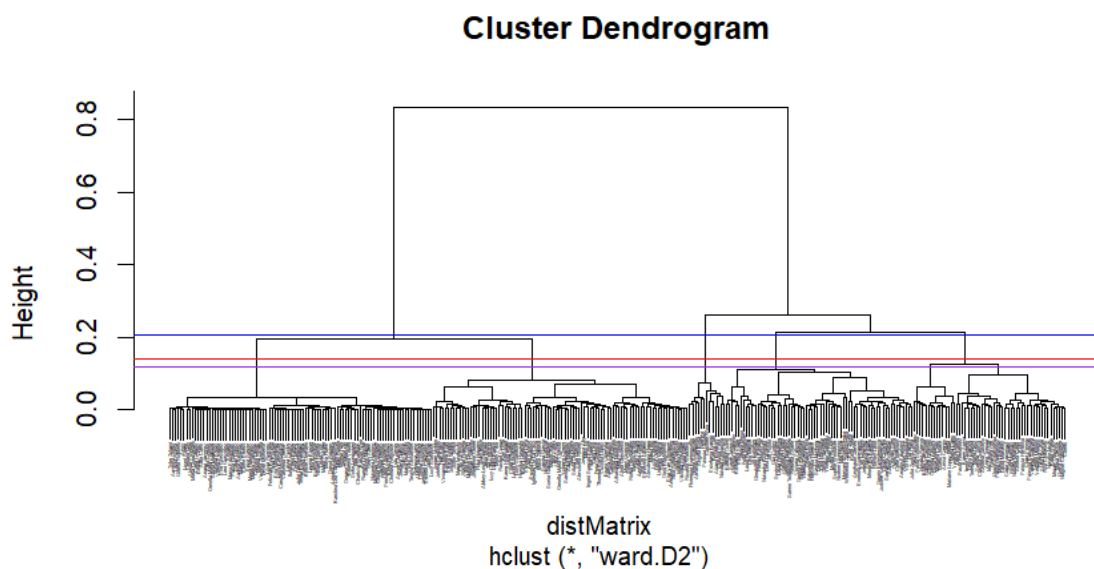


Figura 22: Dendrograma inicial Migcampista

Segons la suma de la distància al quadrat, el nombre òptim de clústers es trobaria en 5 o 6, veiem-ho en el següent gràfic:

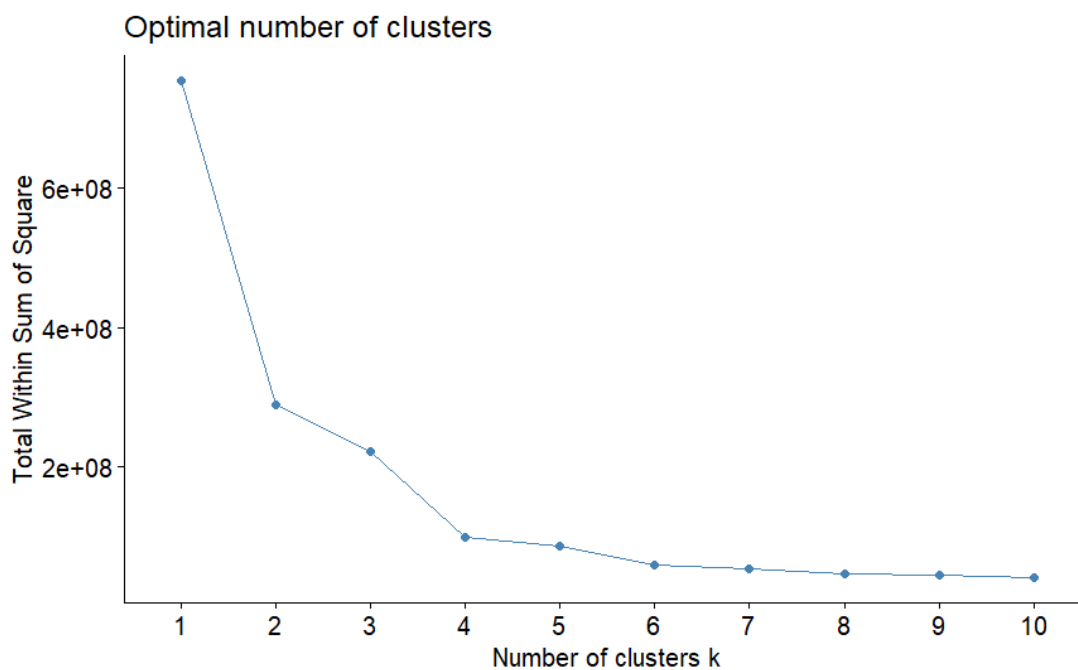


Figura 23: Visualització del nombre de clústers òptim Migcampista.

Escollim el nombre òptim de 6 grups, com que tenim un gran nombre de jugadors es podran diferenciar més entre si.

El nombre de jugadores de posició el mig del camp es reparteixen en cada clúster així:

Taula 13: Nombre d'observacions per cada clúster Migcampista

Clúster	C1	C2	C3	C4	C5	C6	TOTAL CC
Nombre d'observacions	39	90	13	93	15	66	316

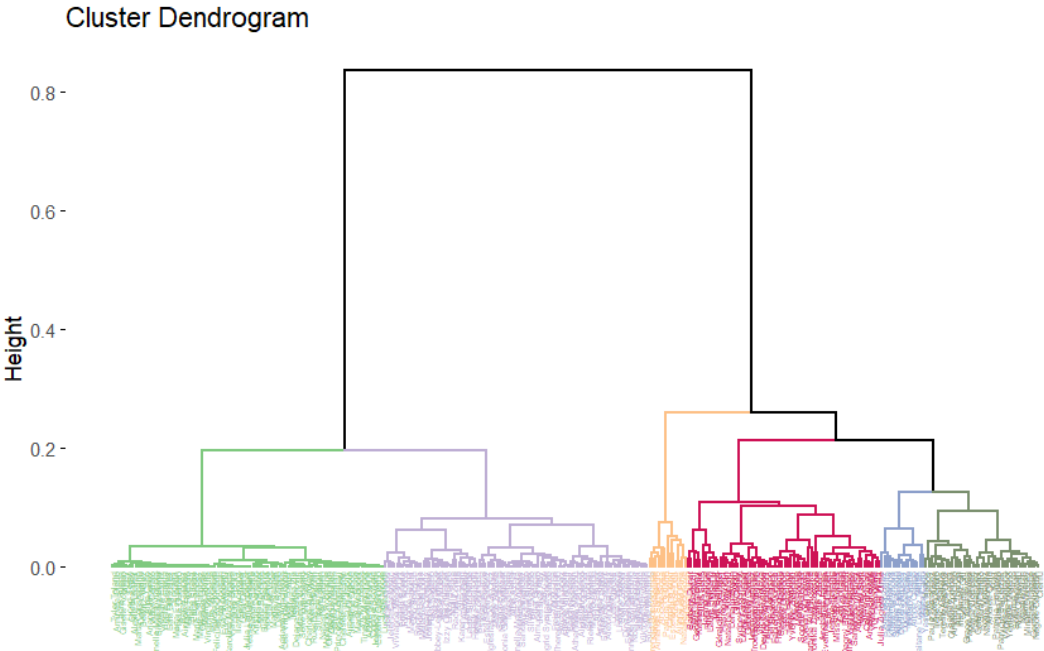


Figura 24: Dendrograma final Migcampista

Com ha passat en les jugadores defensives, és impossible veure els noms de les jugadores que integren cada grup, veiem un clúster de més a prop:

Subtree 6

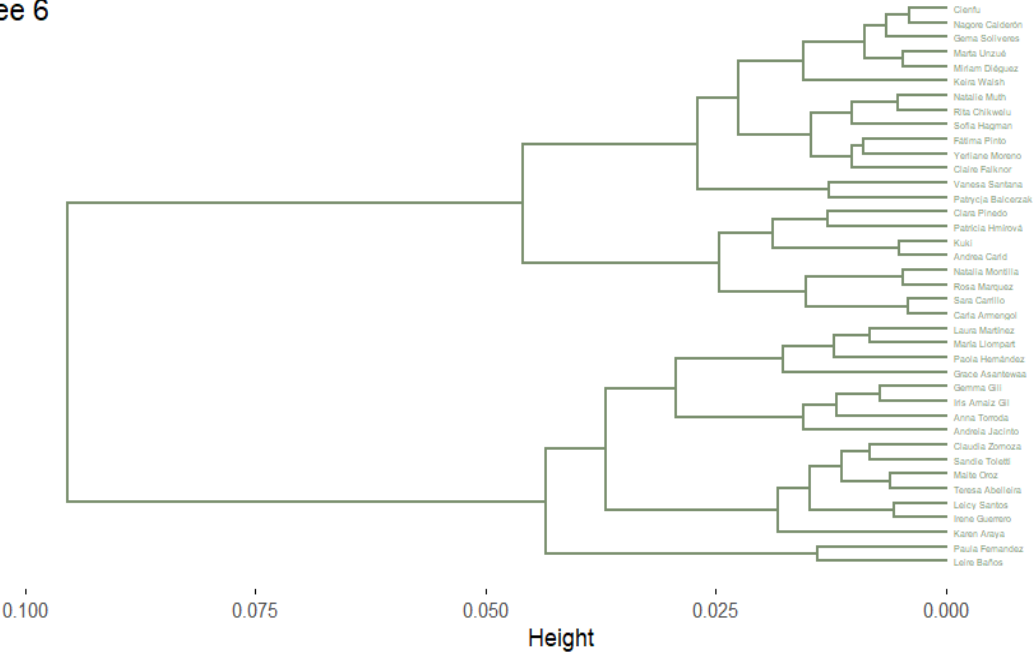


Figura 25: Dendrograma ampliat del clúster 6 Migcampista.

- Posició d' ATACANT

Veiem possibles talls que podríem fer al dendrograma per aconseguir el número òptim de clústers. El tall on hi ha la línia blava ens donaria 4 grups, si la fem una mica més a avall, on hi ha la línia vermella, en serien 5 i si el tall es formes sobre la línia violeta, els grups serien 6, per exemple.

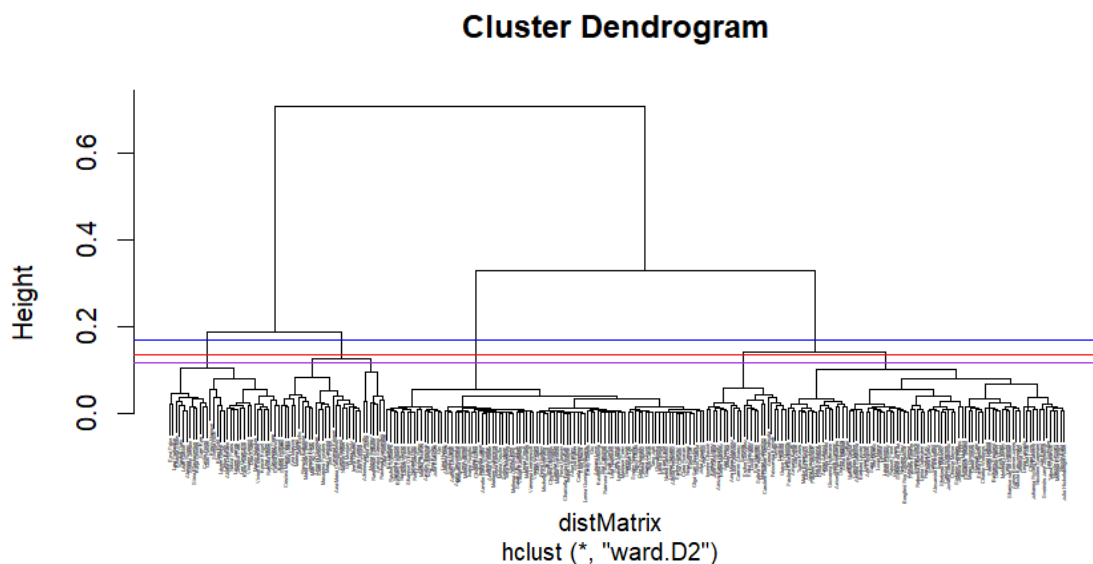


Figura 26: Dendrograma inicial Atacant.

Veiem-ho ara a través de la suma de la distància al quadrat:

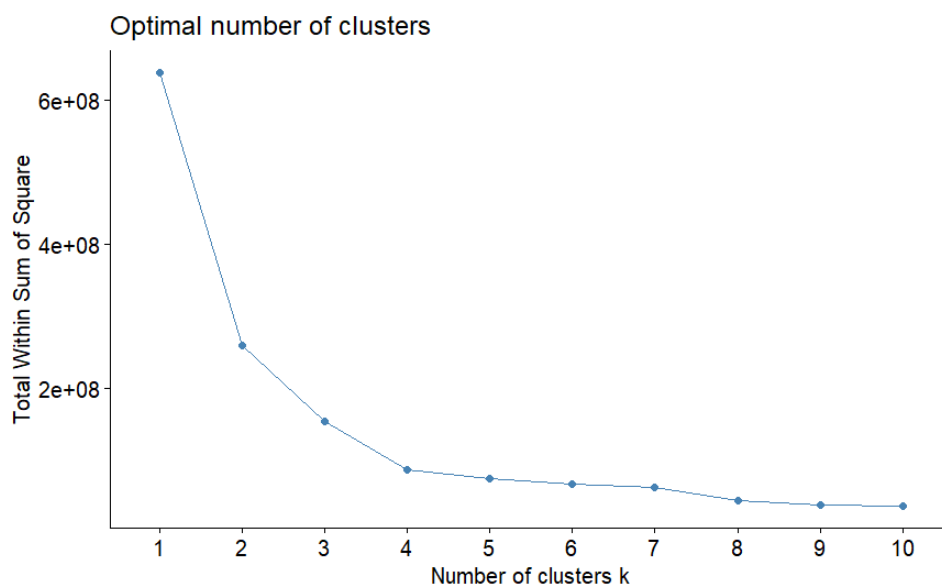


Figura 27: Visualització del nombre de clústers òptim Atacant.

El nombre òptim de clústers es podria trobar en el 4 o en el 5, fins i tot, podríem pensar en el 6 o el 7. Fent una comparació en com es distribueixen el nombre de jugadors per cada clúster sembla que el més òptim és escollir 5 grups.

El nombre de jugadores atacants que hi ha a cada clúster és el següent:

Taula 14: Nombre d'observacions per cada clúster Atacant

Clúster	C1	C2	C3	C4	C5	TOTAL DL
Nombre d'observacions	94	31	82	24	33	264

Gràficament es veu d'aquesta manera:

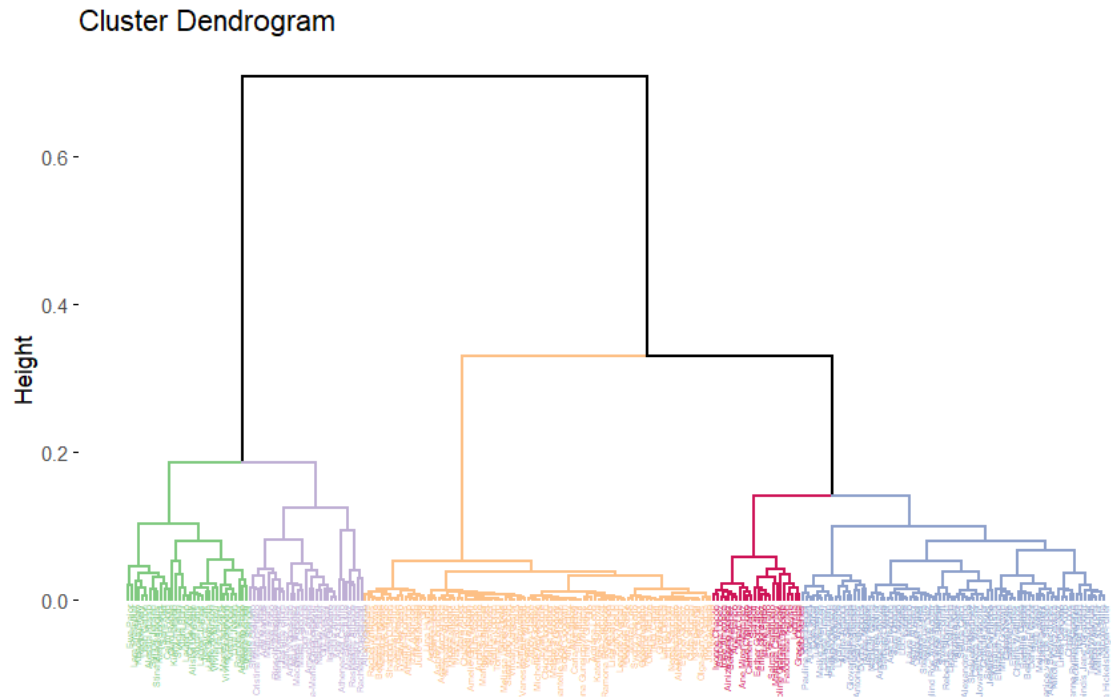


Figura 28: Dendrograma final Atacant

Ens fixem en el grup 4 de més a prop, per exemple:

Subtree 4

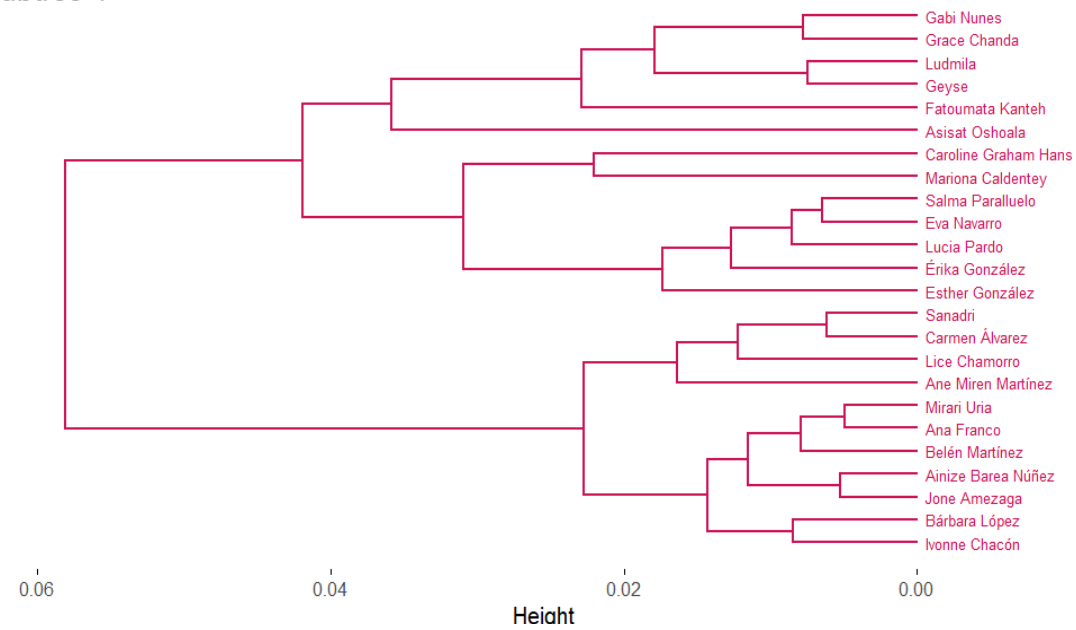


Figura 29: Dendrograma ampliat del clúster 4 Atacant.

8. Cas pràctic

Un cop aplicats els algorismes, veiem un cas on aquests resultats es podrien aplicar a la vida real en un equip.

- Substitució per lesió

Situació:

S'acosta el final de temporada, jornada 18 de la Women's Super League, juga Arsenal contra Manchester United, partit important, segon contra tercer, guanyar aquest partit suposa un pas més a prop de la victòria del títol o, pel contrari, posar més difícil l'objectiu d'aconseguir-lo. Fa poc que ha començar el partit, és el minut 15 i criden a les assistències, la defensa i capitana de les "gunners"¹⁹ es veu obligada a abandonar el terreny de joc per una lesió al genoll. (DAZN Noticias, 2023)

Després de varies proves es confirma, Leah Williamson pateix un trencament del lligament creuat anterior del genoll dret. El pronòstic mèdic preveu uns vuit o nou mesos de recuperació després de l'operació. Això implica que es perdrà el que queda de temporada i una bona part de la temporada següent.

Objectiu:

Necessitem buscar possibles jugadores per la posició que ocupa Leah Williamson durant el temps que estigui de baixa.

Perfil de la jugadora a substituir:

Nom: Leah Williamson
Lliga: ENG
Equip: Arsenal
Posició: DF
Any de naixement: 1997
Partits jugats: 12
Minuts jugats: 820

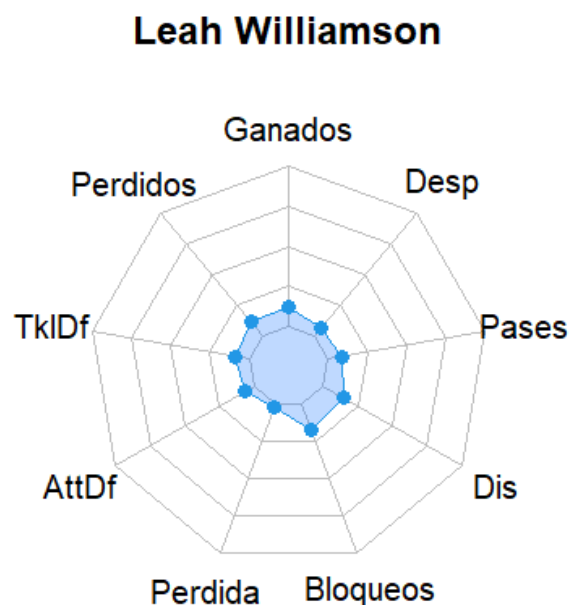


Figura 30: Gràfic de radar Leah Williamson.

¹⁹ "Gunnners", en català "Artillers", és el sobrenom al qual es fa referència a les jugadores i als jugadors de l'Arsenal FC.

Primer de tot, veiem en quin clúster esta ubicada la jugadora i quines jugadores té al costat.

Recordem el dendrograma realitzat anteriorment de les jugadores defensives, la Leah es troba en l'últim clúster, el de color blau, ampliem el gràfic fins a obtenir les últimes branques i així poder distingir els noms d'algunes jugadores:

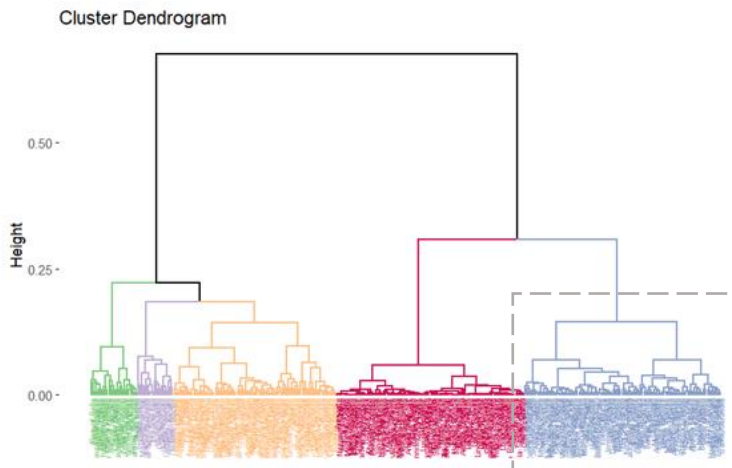


Figura 31: Dendrograma general Defensa

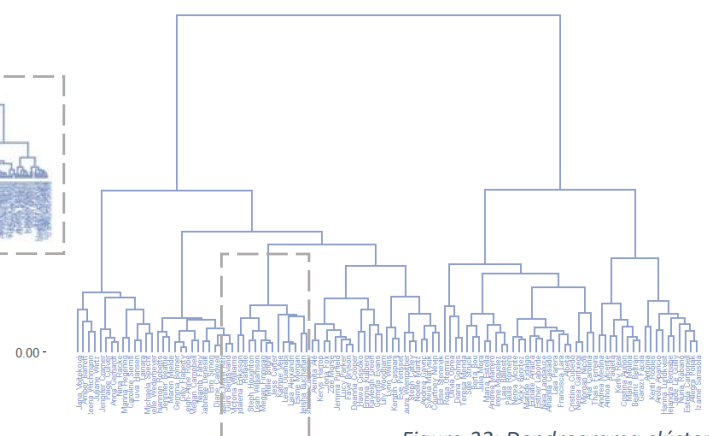


Figura 32: Dendrograma clúster 5.

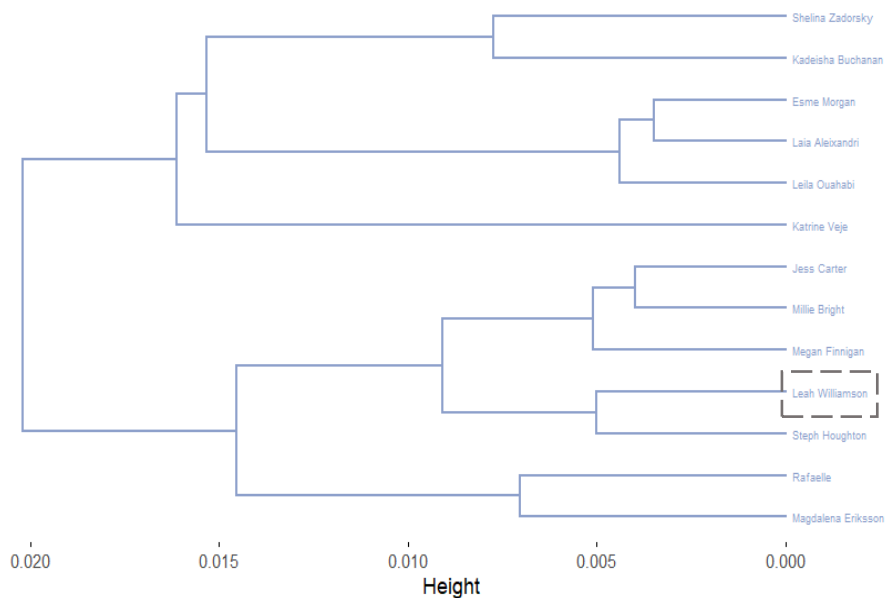
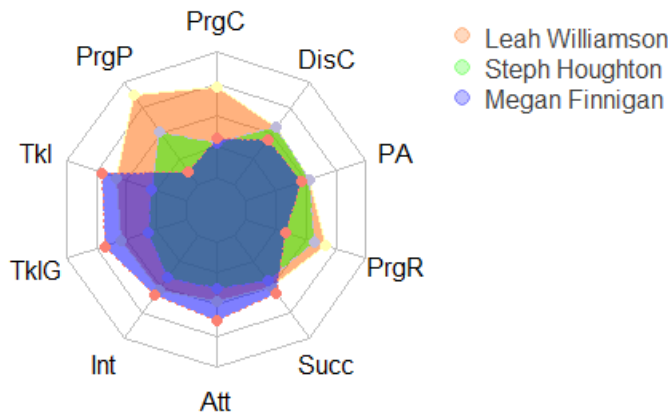


Figura 33: Branca dendrograma Leah Williamson.

Propostes per l'algoritme de *Clustering*:

- Steph Houghton
- Megan Finnigan
-
- Millie Bright
- Jess Carter

Comparativa variables DF



Veiem gràficament, a través d'aquest gràfic de radar, com queden les dades de 2 de les jugadores proposades anteriorment, juntament amb les de la jugadora que hem de substituir.

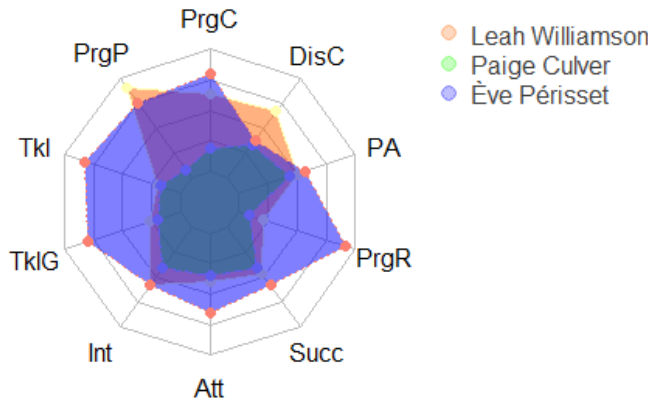
Figura 34: Gràfic de radar per comparar propostes Clustering.

Agafem el clúster número 5, el qual pertany la Leah, i veiem les propostes que ens ofereix l'algoritme de PCA són les següents:

Taula 15: Proposta de substitucions a través de PCA

Jugadora	Valor PCA	Clúster
Allegra Poljak	-870.9281	5
Paige Culver	-781.2275	5
Leah Williamson	-751.9598	5
Ève Périsset	-749.4026	5
Tainara	-737.8172	5

Comparativa variables DF



A simple vista, sembla que les variables en el gràfic anterior s'adaptin més als valors de la jugadora Leah Williamson.

Figura 35: Gràfic de radar per comparar propostes PCA.

9. Conclusions

L'objectiu del d'aquest treball era recollir i manipular un conjunt de dades, a més d'aplicar tècniques estadístiques de mineria de dades per analitzar-les. Tots aquest punts s'han complert, s'ha obtingut una extensa base de dades, s'han aplicat tècniques de reducció de la dimensionalitat i tècniques de classificació.

A més, s'ha introduït la situació actual de les dades en el futbol en general i, per tant, també en el futbol femení. A dia d'avui, són temes que estan en gran creixement.

Els món de l'aprenentatge automàtic és molt extens, existeixen moltes tècniques i sectors aplicables, en aquest treball únicament s'ha tractat una petita part. I s'ha pogut veure un exemple de com utilitzar els algoritmes per trobar substitucions a altres i/o a trobar aquella que és millor en la seva posició.

Futures línies d'investigació proposades:

- Seleccionar únicament una lliga i introduir dades de temporades anteriors per poder analitzar millor els perfils de les jugadores.
- Analitzar una base de dades d'una única posició amb totes les jugadores a nivell europeu possibles.
- Aplicar els algoritmes en altres casos pràctics.

Les tècniques esposades son perfectament aplicables a la vida real, com s'ha vist en l'últim punt, tot i això, aquest treball només és un plantejament introductori a l'anàlisi, per poder utilitzar-ho en la presa de decisions importants s'hauria d'ampliar l'estudi.

Bibliografia

- Cambridge Dictionary. (2020). *Machine Learning*. Recollit de <https://dictionary.cambridge.org/dictionary/english/machine-learning>
- DFB. (2023). *Tabelle*. Recollit de https://www.dfb.de/google-pixel-frauen-bundesliga/spieltagtable/?no_cache=1&spieledb_path=%2Fdatencenter%2Fgoogle-pixel-frauen-bundesliga%2F2022-23%2Fcurrent
- DFB. (sense data). *Informació de la lliga*. Recollit de Google Pixel Frauen-Bundesliga: <https://www.dfb.de/google-pixel-frauen-bundesliga/ligainfos/modus/>
- enciclopèdia.cat. (sense data). *Big Data*. Recollit de <https://www.enciclopedia.cat/gran-enciclopedia-catalana/big-data>
- Ertheo. (27 / 09 / 2023). *Los mejores sistemas de juego en el fútbol: alineaciones y formaciones*. Recollit de <https://www.ertheo.com/blog/futbol-mejores-sistemas-de-juego-alineaciones-formaciones/>
- fcbarcelona.cat. (22 / Abril / 2022). *Notícies*. Recollit de <https://www.fcbarcelona.cat/ca/noticies/2585994/el-camp-nou-ho-torna-a-fer-i-bat-un-nou-record-mundial-dassistencia-amb-91648-assistents>
- Hermana, D. (21 / Abril / 2023). *DAZN Noticias*. Recollit de <https://www.dazn.com/es-ES/news/f%C3%BAtbol/leah-williamson-lesionada-que-pasa-que-lesion-tiene-cuanto-tiempo-estara-baja-que-partidos-pierde/142korwdehek910bma0oyiagf4>
- IBM. (sense data). *¿Qué es el aprendizaje no supervisado?* Recollit de <https://www.ibm.com/es-es/topics/unsupervised-learning>
- IBM. (sense data). *¿Qué es el aprendizaje supervisado?* Recollit de <https://www.ibm.com/es-es/topics/supervised-learning>
- IBM. (sense data). *¿Qué es machine learning?* Recollit de <https://www.ibm.com/es-es/topics/machine-learning>
- kaspersky. (sense data). *¿Qué es la minería de datos y por qué es importante?* Recollit de <https://latam.kaspersky.com/resource-center/definitions/data-mining>
- Kowarik, A., & Templ, M. (2016). *Imputation with the R Package VIM*. Journal of Statistical Software, 74(7), 1–16. Recollit de <https://doi.org/10.18637/jss.v074.i07>
- La Liga F. (2023). *Clasificación Liga F*. Recollit de https://ligaf.es/clasificacion/primera_division_femenina/30/2023
- Microsoft Learn. (2024). Recollit de <https://learn.microsoft.com/es-es/power-bi/fundamentals/power-bi-overview>
- Microsoft News. (6 / Octubre / 2021). *LaLiga y Microsoft presentan Beyond Stats, un proyecto de análisis futbolístico avanzado que profundiza en el juego de cada equipo*. Recollit de <https://news.microsoft.com/es-es/2021/10/06/laliga-y-microsoft-presentan-beyond-stats-un-proyecto-de-analisis-futbolistico-avanzado-que-profundiza-en-el-juego-de-cada-equipo/>
- Microsoft News. (21 / Septiembre / 2022). *El Valencia Club de Fútbol incrementa un 20% el número de asistentes a sus partidos, gracias al análisis de datos*. Recollit de <https://news.microsoft.com/es-es/2022/09/21/el-valencia-club-de-futbol-incrementa-un-20-el-numero-de-asistentes-a-sus-partidos-gracias-al-analisis-de-datos/>

- Microsoft, S. t. (2021). *Tareas básicas en Excel*. Recollit de <https://support.microsoft.com/es-es/office/tareas-b%C3%A1sicas-en-excel-dc775dd1-fa52-430f-9c3c-d998d1735fca>
- Novales, A. (Novembre / 2016). *Modelos Factoriales*. Recollit de <https://www.ucm.es/data/cont/media/www/pag-41460/Reduccion%20dimensionalidad.pdf>
- Palco23. (19 / Setembre / 2023). Recollit de <https://www.palco23.com/media/la-liga-f-casi-triplica-su-audiencia-en-el-arranque-de-la-temporada-2023-2024>
- Pérez López, C. (2007). *Minería de datos: técnicas y herramientas*. Paraninfo.
- Real Academia Española. (sense data). *Macrodatos*. Recollit de <https://dle.rae.es/macrodatos>
- Rossell, D. (20 / Novembre / 2014). *Dades massives i estadística. La perspectiva d'un estadístic*. Recollit de <https://metode.cat/revistes-metode/monografics/dades-massives-i-estadistica.html>
- r-project.org. (sense data). *What is R?* Recollit de <https://www.r-project.org/about.html>
- Sánchez Sellero, C. (2008-2009). *Análisis de componentes principales*. Recollit de Máster en Técnicas Estadísticas: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP/MATERIALESMATER/Mat_14_master0809multi-tema5.pdf
- Stats Perform. (2021). *About Stats Perform*. Recollit de <https://www.statsperform.com/about/>
- TERMCAT, C. d. (sense data). *DADES MASSIVES, MACRODADES O BIG DATA?* Recollit de <https://www.termcat.cat/ca/diccionaris-en-linia/187/fitxa/MjkwNjY2Mg%3D%3D>
- The FA Player. (2023). *Tables*. Recollit de <https://faplayer.thefa.com/scores/womens-super-league?tab=tables&season=2022-2023#>
- Torres, D. (30 / Setembre / 2023). El big data hunde al Ajax. *El País*. Recollit de <https://elpais.com/deportes/2023-09-30/el-big-data-hunde-al-ajax.html>
- UEFA. (2023). *Women's association club coefficients*. Recollit de <https://www.uefa.com/nationalassociations/uefarankings/womenscountry/%23/yr/2023>
- Unayta. (28 / Gener / 2019). Recollit de <https://unayta.es/data-mining-big-data/>
- Vilches Gonzáles, E., & Escobar Broitman, I. A. (2007). *Minería de datos*. Ipmagazine. Recollit de https://www.researchgate.net/publication/267711514_Mineria_de_Datos
- Wikipedia. (2023). *Frauen-Bundesliga*. Recollit de <https://en.wikipedia.org/wiki/Frauen-Bundesliga>
- Wikipedia. (2024). *Primera División Femenina de España*. Recollit de https://es.wikipedia.org/wiki/Primera_Divisi%C3%B3n_Femenina_de_Espa%C3%B1a
- Wikipedia. (2024). *Women's Super League*. Recollit de https://en.wikipedia.org/wiki/Women%27s_Super_League

Annex
METADATA

NOM VAR. A LA BBDD	DESCRIPCIÓ	TIPUS	CATEGORIES		
Jugadora	Nom i cognom	character	Cadena de caràcters		
Lliga	Lliga	character	ESP Espanya GER Alemanya ENG Anglaterra		
País	País de nacionalització	character	ALB - Albània	RUS - Federació Russa	NOR - Noruega
			DEU - Alemanya	FIN - Finlàndia	NZL - Nova Zelanda
			AND - Andorra	FRA - França	NLD - Països Baixos
			ARG - Argentina	WAL - Gales	PAN - Panamà
			AU - Austràlia	GMB - Gàmbia	PRY - Paraguai
			AUT - Àustria	GHA - Ghana	PER - Perú
			BEL - Bèlgica	GRC - Grècia	POL - Polònia
			BIH - Bòsnia i Hercegovina	GNQ - Guinea Equatorial	DOM - República Dominicana
			BRA - Brasil	HUN - Hongria	PRT - Portugal
			CMR - Camerun	IRL - Irlanda	GBR - Regne Unit
			CAN - Canadà	NIR - Irlanda del Nord	CZE - República Txeca
			COL - Colòmbia	ISL - Islàndia	ROU - Romania
			KOR - Corea del Sud	ISR - Israel	SER - Sèrbia

			CIV - Costa de Vori HRV - Croàcia DNK - Dinamarca ECU - Equador SCO - Escòcia SVK - Eslovàquia SVN - Eslovènia ESP - Espanya USA - Estats Units	ITA - Itàlia JAM - Jamaica JPN - Japó JOR - Jordània KSV - Kosovo MKD - Macedònia MA - Marroc MEX - Mèxic NGA - Nigèria	RSA - Sud Àfrica SWE - Suècia CHE - Suïssa TUR - Turquia URY - Uruguai VEN - Veneçuela CHL - Xile CHN - Xina ZMB - Zàmbia	
Posc1	Posició natural	character	PO Portera DF Posició defensiva CC Posició central DL Posició atacant			
Equip	Equip	character	Alavés Alhama Arsenal Aston Villa Athletic Club Atlético Madrid Barcelona Bayern Munich Brighton Chelsea	Eint Frankfurt Essen Everton Freiburg Hoffenheim Köln Leicester City Levante Levante Planas Leverkusen	Liverpool Madrid CFF Manchester City Manchester Utd MSV Duisburg Reading Real Betis Real Madrid Real Sociedad Sevilla	Sporting Huelva SV Meppen Tottenham Turbine Potsdam UDG Tenerife Valencia Villarreal Werder Bremen West Ham Wolfsburg

NOM DE LA VARIABLE A LA BBDD	DESCRIPCIÓ	TIPUS	ESTADÍSTICS DESCRIPTIUS			
			MIN	MAX	MEAN	SD
Any	Any de naixement	numeric	1978	2007	1997,59	4,6
PJ	Partits jugats	numeric	1	30	14,94	8,59
Titular	Partits jugats des de l'inici	numeric	0	30	10,88	8,48
Min	Minuts jugats	numeric	1	2648	923,92	701,45
90s	Partits sencers	numeric	0	29,4	10,87	7,98
FC	Faltes comeses	numeric	0	72	9,56	9,8
FR	Faltes rebudes	numeric	0	105	9,08	10,57
TG	Targetes grogues	numeric	0	14	1,36	1,88
TV	Targetes vermelles	numeric	0	1	0,05	0,21
GC	Gols en contra	numeric	0	57	1,63	6,63
DaPC	Xuts en contra	numeric	0	159	4,62	19,16
Parades	Aturades	numeric	0	115	3,05	12,75
Parades%	% Aturades	numeric	0	100	6,05	19,95
PaC	Porteries a 0	numeric	0	15	0,29	1,41
PaC%	% Porteries a 0	numeric	0	100	2,43	10,96
PSxG	Gols esperats en contra	numeric	0	54,4	1,35	5,68
PSxG/90	Gols esperats en contra per partit	numeric	-1,5	3,46	-0,02	0,18
Ganados	Duels aeris guanyats	numeric	0	87	8,89	11,41
Perdidos	Duels aeris perduts	numeric	0	89	8,87	10,35
Tk1Df	Regatejadors abordats	numeric	0	45	6,89	7,85
AttDf	Desafiaments intentats	numeric	0	78	13,44	13,96
Perdida	Desafiaments perduts	numeric	0	38	6,55	6,95
Bloqueos	Bloquejos de pilota	numeric	0	51	11,61	10,95
Dis	Xuts bloquejats	numeric	0	34	2,8	4,4
Pases	Passades bloquejades	numeric	0	46	8,81	8,74
Desp	Rebuig	numeric	0	192	16,23	22,26
Toc	Tocs	numeric	0	2247	547,58	460,27

NOM DE LA VARIABLE A LA BBDD	DESCRIPCIÓ	TIPUS	ESTADÍSTICS DESCRIPTIUS			
			MIN	MAX	MEAN	SD
Dist.tot	Distància total (en iardes ²⁰)	numeric	0	7038	1408,83	1289,66
Dist.prg	Distància progressiva (en iardes)	numeric	0	4552	668,79	683,83
Rec	Passades rebudes	numeric	0	1760	314,73	290,41
Recup	Recuperacions	numeric	0	229	59,33	52,11
PrgC	Controls progressius	numeric	0	124	13,2	17,88
PrgP	Passades progressives	numeric	0	273	33,87	37,56
Tkl	Bloquejats	numeric	0	100	18,44	18,36
TklG	Bloquejos aconseguits	numeric	0	66	12,3	12,39
Int	Intercepcions	numeric	0	74	12,09	13,71
Att	Intents d'enfrontament	numeric	0	203	13,56	18,74
Succ	Defenses enfrontats amb èxit	numeric	0	102	6,56	9,19
PrgR	Passades progressives rebudes	numeric	0	285	33,4	45,71
PA	Posicions avançades	numeric	0	37	2,43	4,92
DisC	Xuts a porteria	numeric	0	132	12,42	16,77
DaP	Xuts entre els 3 pals	numeric	0	63	4,53	7,05
TalArc/90	Xuts a porteria per cada 90 min	numeric	0	22,5	0,45	1,02
Gls	Gols	numeric	0	27	1,54	2,96
Ass	Assistències	numeric	0	13	1,07	1,87
xG	Gols esperats	numeric	0	23,2	1,36	2,37
xAG	Assistències esperades	numeric	0	10,6	0,93	1,39
G-xG	Diferència entre gols i gols esperats	numeric	-3,1	9,3	0,19	1,16
Gls/90	Gols en 90 min	numeric	0	2,37	0,13	0,25
Ass/90	Assistències en 90 min	numeric	0	4,74	0,09	0,21
xG/90	Gols esperats en 90 min	numeric	0	6,58	0,14	0,31
xAG/90	Assistències esperades en 90 min	numeric	0	2,06	0,09	0,14

²⁰ Conversió: 1 iarda = 0.9144 metres

CODI

El codi utilitzat en aquest treball es pot trobar en la següent carpeta de Goggle Drive:

[Codi utilitzat i base de dades](#)

També es pot trobar en aquest mateix enllaç el document Excel amb les dades utilitzades.