Valoración deportivo-financiera de equipos de fútbol mediante aprendizaje automático

Trabajo de Fin de Grado

Carlos Gómez-Torres

8 de Septiembre, 2024

Conjunto de datos

Para los objetivos de este estudio, se pretendía construir un conjunto de datos con variables financieras y deportivas alrededor de los 200 equipos de fútbol participantes en las competiciones de primera y segunda división en España, Inglaterra, Italia, Alemania y Francia durante la campaña 2023/2024. En concreto, la información financiera (balance de situación, cuenta de pérdidas y ganancias, y transacciones de M&A) se extraería de la base de datos ORBIS de Bureau van Dijk, y la información deportiva del portal web Transfermarkt, abarcando los años en el periodo 2009-2023.

ORBIS

En un principio, se intentó un filtrado en ORBIS con palabras clave y códigos identificadores del sector. Sin embargo, el listado resultante no era completo, además de incluir equipos de otros deportes (algunos relacionados con equipos de fútbol, como el Real Madrid Baloncesto), así como otras organizaciones controladas por equipos de fútbol, como lo son las fundaciones benéficas.

De manera que, se buscó manualmente cada club en la base datos con el fin de obtener sus identificadores y filtrar en base a los mismos. En este proceso, se encontró que una porción de los clubes no estaban presentes en la base de datos. En concreto, se lograron identificar 142 equipos. No obstante, la información de las cuentas anuales de interés no estaba presente o completa para todos los años.

Con el fin de tener la mayor cantidad de datos posibles, se buscó manualmente auditorías y reportes anuales de los equipos. En el caso de los equipos ingleses, al ser sociedades limitadas, se encontraron sus auditorías en el portal web del registro mercantil del Reino Unido. Para el resto de equipos, se llevó a cabo una búsqueda exhaustiva en sus páginas web correspondientes. Este procedimiento permitió añadir datos de 30 equipos más, para un total de 172 equipos.

Transfermarkt

Por otro lado, de Transfermarkt.com se extrajo información acerca del rendimiento deportivo de los equipos (división en la que compitieron, posición conseguida, partidos ganados, goles a favor y en contra, etc) y su actividad en el mercado de transferencia.

Variables

Para el análisis, se seleccionaron variables clave del conjunto de datos de ORBIS y Transfermarkt. De las cuentas anuales, se conservaron tres variables: la cifra de inmovilizado material, los ingresos de explotación y

el beneficio antes de intereses y después de impuestos. Paralelamente, del conjunto de datos de Transfermarkt, se extrajeron para cada temporada el valor de mercado agregado de la plantilla, el gasto en fichajes, la capacidad máxima del estadio y la asistencia media al estadio. Adicionalmente, se calcularon el ranking nacional del equipo, definido como el percentil de la posición del equipo en relación con todas las posiciones entre la primera y la tercera división, y el porcentaje de victorias en liga sobre el total de partidos jugados. Ambas métricas se calcularon también como medias móviles con una ventana de 5 años.

El glosario con los nombres abreviados escogidos y las descripciones correspondientes para todas estas variables se detallan en la Tabla 1.

Tabla 1: Glosario de variables

Alias	Descripción
EQUIPO	Nombre del equipo
PAIS	País de la liga en la que compite el equipo
AÑO	Año al que corresponden los datos financieros y deportivos
MATERIAL	Cifra del inmovilizado material
INGRESOS	Cifra de ingresos de explotación
EBIT	Cifra del beneficio antes de impuestos e intereses
MV	Valor de mercado agregado de la plantilla, según Transfermarkt.com
MV5	Media del valor agregado de la plantilla en los últimos 5 años
FICHAJES	Gasto total en fichajes durante la temporada
CAPACIDAD	Capacidad máxima del estadio
ASISTENCIA	Número de espectadores medio en partidos locales durante la temporada
OCUPACION	Media del porcentaje de ocupación del estadio durante la temporada
RANKNAC	Ranking nacional del equipo (percentil de posición en liga sobre las divisiones 1-3)
RANKNAC5	Media del ranking nacional en los últimos 5 años
WINPCT	Porcentaje de partidos ganados en liga durante la temporada
WINPCT5	Media del porcentaje de partidos ganados en liga durante los últimos 5 años

Tabla 2: Nº de observaciones por país

País	N	%
Alemania	67	6.86
España	264	27.05
Francia	108	11.07
Inglaterra	320	32.79
Italia	217	22.23

Tabla 3: Nº de transacciones por país

País	N	%
Alemania	10	18.87
España	5	9.43
Francia	2	3.77
Inglaterra	9	16.98
Italia	12	22.64

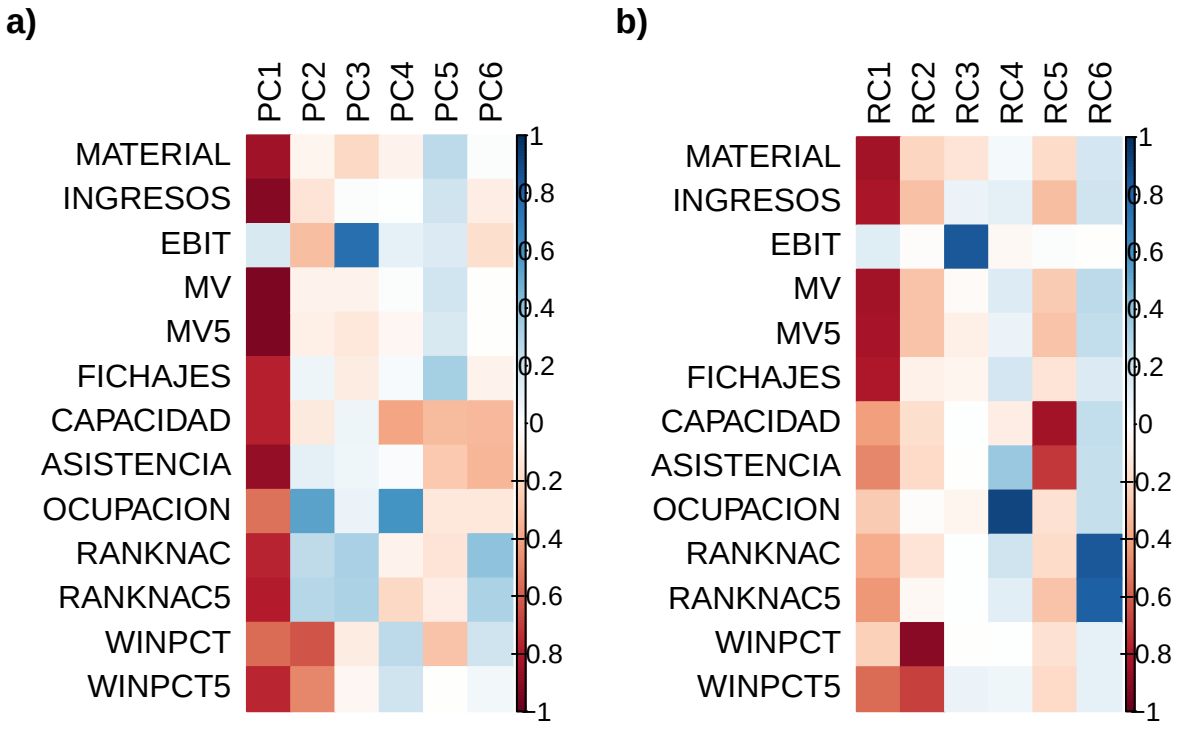
Resultados

Análisis de Componentes Principales (PCA)

La reducción de la dimensionalidad mediante la aplicación del análisis de componentes principales (PCA), permite explicar el 93,68% de la varianza del conjunto de datos originales a lo largo de 6 dimensiones. La matriz de correlación representada detalla las cargas de cada variable original sobre dichos factores, permitiendo su interpretación y etiquetado. Tras aplicar una rotación varimax sobre los componentes principales, se obtienen unos componentes rotados (RC) cuyas cargas o correlaciones con las variables originales los hacen más fáciles de interpretar. Por esta razón, se decidió utilizar la solución rotada antes que la original para entrenar la malla SOM.

El primer factor muestra una fuerte correlación negativa ($r < -0{,}8$) con los ingresos, la cifra de inmovilizado material, y el valor de mercado tanto de la plantilla actual como de la histórica reciente. Además, presenta una correlación negativa moderada ($r < -0{,}5$) con el gasto en fichajes y el porcentaje de victorias actual. En resumen, se puede interpretar que el primer componente rotado (RC1) se relaciona con el valor de los principales activos del equipo (estadio y jugadores), así como su capacidad para generar ingresos. Por su parte, el segundo componente rotado (RC2) está fuertemente correlacionado de manera negativa ($r < -0{,}9$) con el porcentaje de victorias actual, y en menor grado ($r < -0{,}6$) con la media de victorias en los últimos cinco años. El tercer componente rotado (RC3) también destaca por su fuerte correlación negativa ($r < -0{,}8$) con el EBIT generado por el club en la temporada actual. En contraste, el cuarto componente rotado (RC4) tiene una fuerte correlación positiva ($r > 0{,}8$) con el porcentaje medio de ocupación del estadio. El quinto componente rotado (RC5) está positivamente correlacionado con la capacidad total del estadio ($r > 0{,}8$) y con la cifra media de asistencia ($r > 0{,}6$). Finalmente, el sexto factor presenta una relación negativa con el ranking nacional actual y el de los últimos cinco años ($r < -0{,}8$). La tabla

Figura 1: Cargas de variables originales sobre nuevos ejes



Fuente: Elaboración propia

Tabla 4: Resultado de PCA

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.823766	1.115881	1.017037	0.8482215	0.7770867	0.7696289
Proportion of Variance	0.613360	0.095780	0.079570	0.0553400	0.0464500	0.0455600
Cumulative Proportion	0.613360	0.709140	0.788710	0.8440500	0.8905000	0.9360700

Etiquetado de los componentes rotados

RC	Etiqueta
1	ACTINGR
2	WINS
3	EBIT
4	OCUP
5	CAPAST
6	RANK

SOM

Se configuró la malla SOM con unas dimensiones de 15x15, para un total de 225 neuronas, equivalente a un 23% de las observaciones. Por su parte, para el entrenamiento se fijó un total del 10.000 iteraciones, con un learning rate de 0,01. Las Figura 3a y 3b representan los cambios en la distancia [formula] a lo largo de las iteraciones y la asignación de observaciones por neurona, respectivamente. Se observa la estabilización del modelo cerca de las 9.000 iteraciones, con una distancia media a la neurona más cercana inferior a 0,01.

Figura 2a: Evolución de la distancia entre neuronas

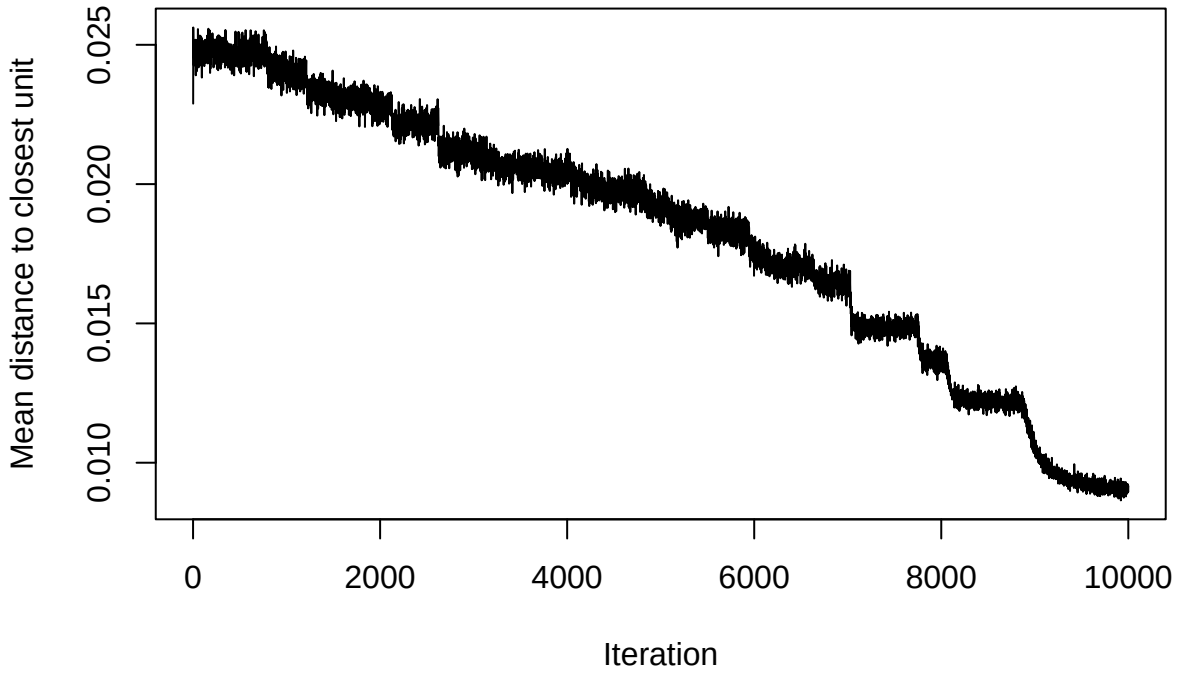
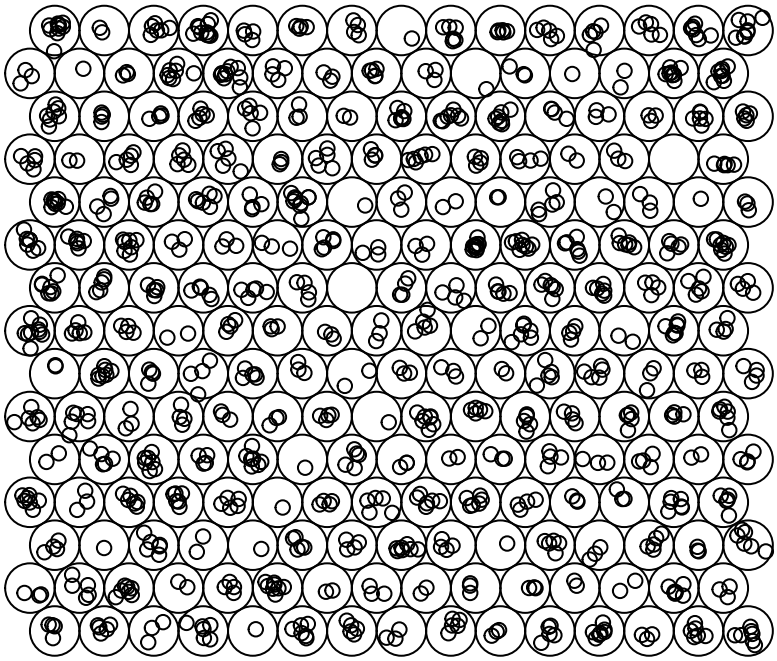
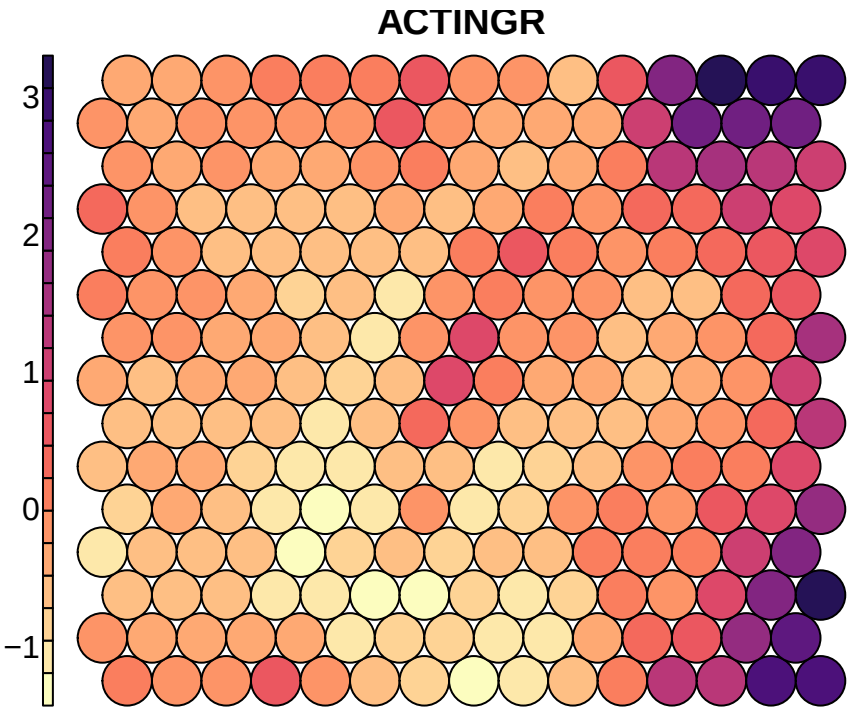
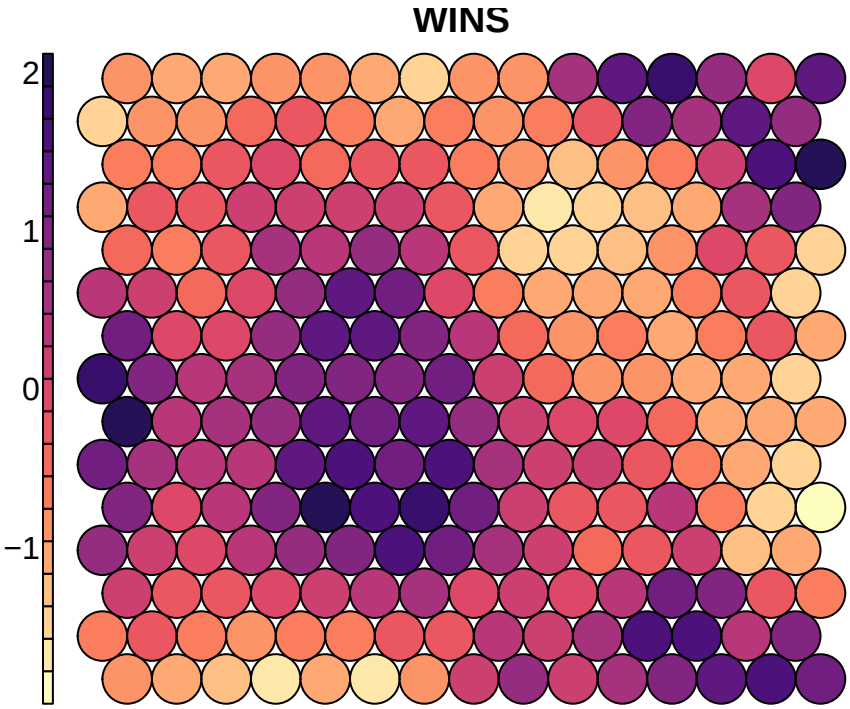
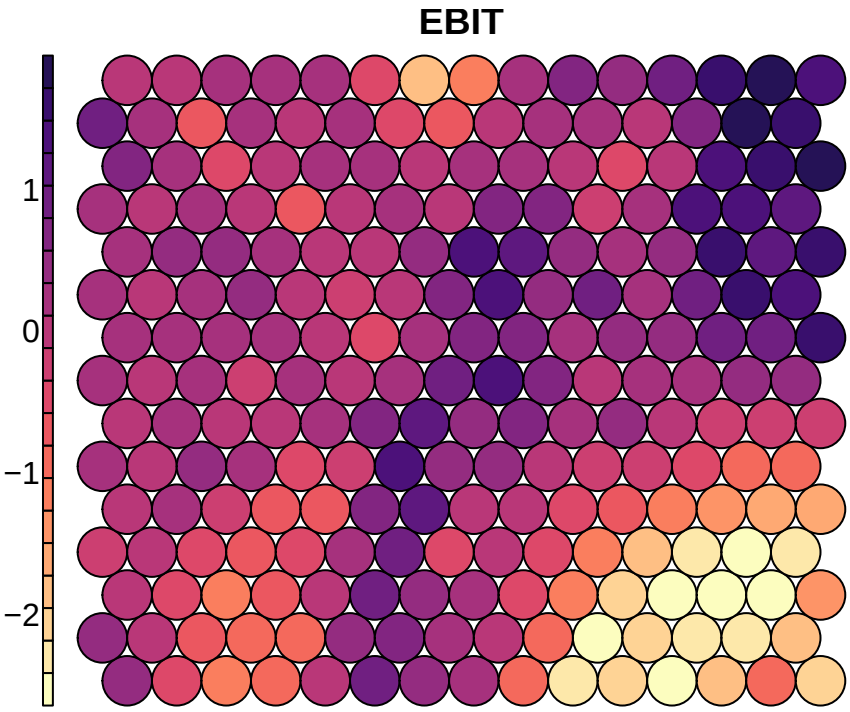


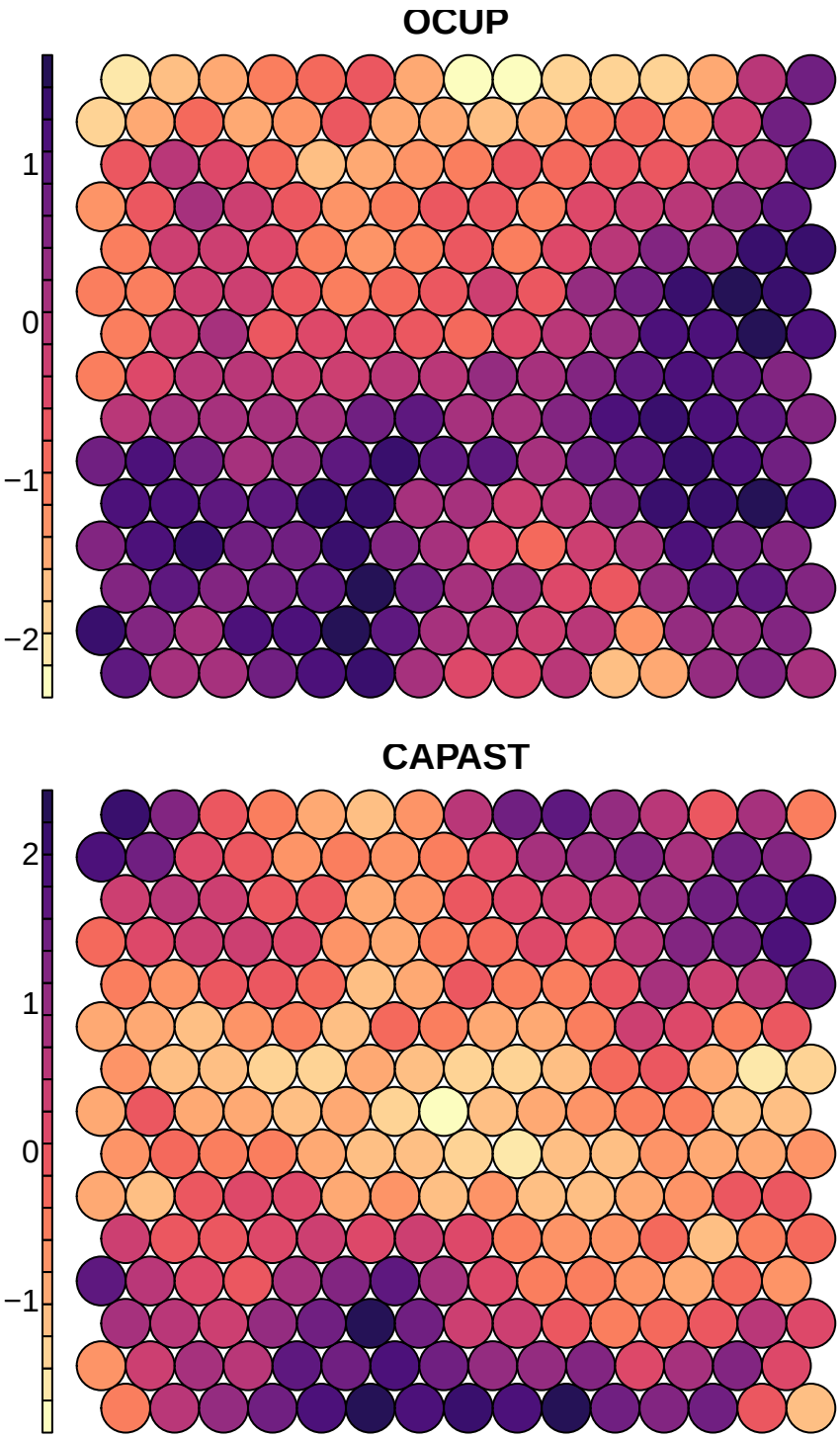
Figura 2b: Asignación de observaciones en SOM

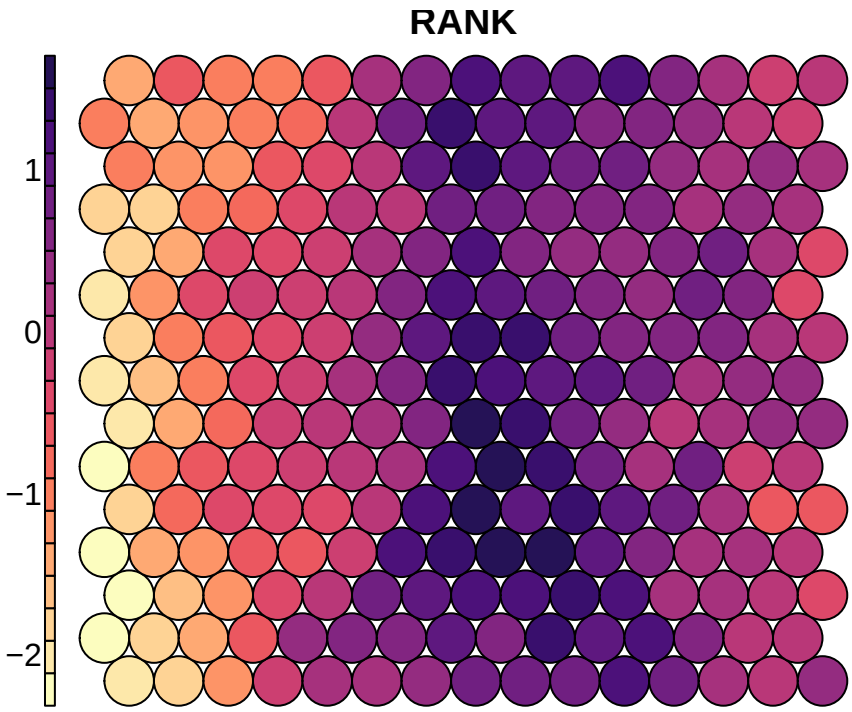





WINS
2
1
0
−1

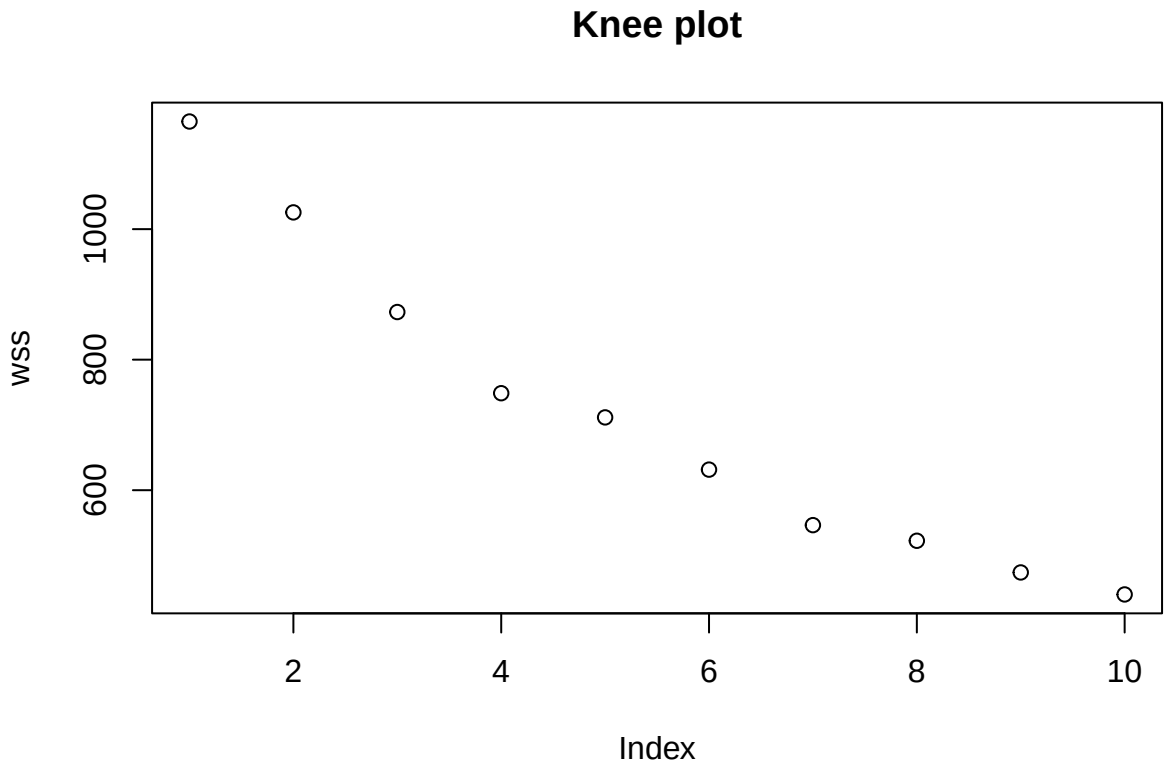

EBIT
1
0
−1
−2

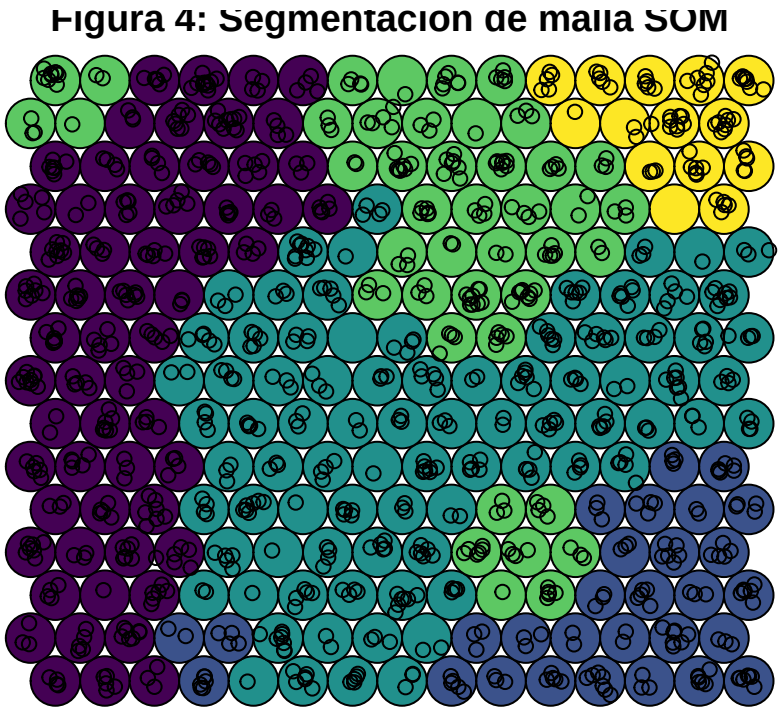

OCUP
1
0
−1
−2
CAPAST
2
1
0
−1



Clustering Jerárquico

Tras aplicar agrupamiento jerárquico con encadenamiento de Ward sobre las neuronas, se decidió segmentar el mapa en 5 clústeres.






```
## som_cluster
##  1  2  3  4  5
## 59 29 81 42 14
```

ANOVA

Tras la realización de pruebas de análisis de varianza (ANOVA), se encontraron diferencias significativas entre los clusters para cada uno de los 6 componentes principales para un nivel de significación del 1% (ver Tablas 6-11). Por lo cual, se procedió a realizar pruebas post-hoc (en concreto, TukeyHSD y Scheffe), pero sobre las variables originales relacionadas a cada uno de los componentes principales, con el fin de poder describir con mayor exactitud los clústers.

Análisis de varianza sobre ACTINGR

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cluster	4	502.7827	125.6956646	258.4625	0
Residuals	971	472.2173	0.4863206	NA	NA

Análisis de varianza sobre WINS

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cluster	4	194.6493	48.6623156	60.55112	0
Residuals	971	780.3507	0.8036568	NA	NA

Activos e ingresos (ACTINGR)

El clúster 1 presenta la mayor media de ingresos, con €523,12 millones, y es seguido por el clúster 2 con €256,95 millones, el clúster 3 con €81,01 millones, el clúster 4 con 80,69 y, finalmente, el clúster 5 con €23,31

Análisis de varianza sobre EBIT

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cluster	4	540.5179	135.1294840	301.9934	0
Residuals	971	434.4821	0.4474584	NA	NA

Análisis de varianza sobre OCUP

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cluster	4	257.62	64.4050039	87.17452	0
Residuals	971	717.38	0.7388053	NA	NA

Análisis de varianza sobre CAPAST

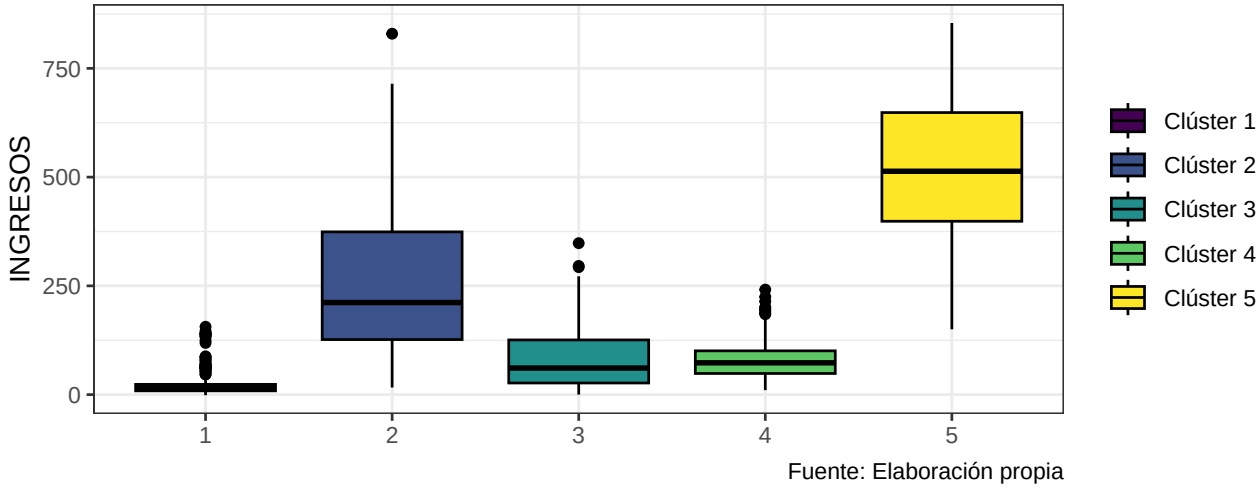
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cluster	4	118.8069	29.7017295	33.68443	0
Residuals	971	856.1931	0.8817642	NA	NA

Análisis de varianza sobre RANK

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cluster	4	550.0554	137.513858	314.2197	0
Residuals	971	424.9446	0.437636	NA	NA

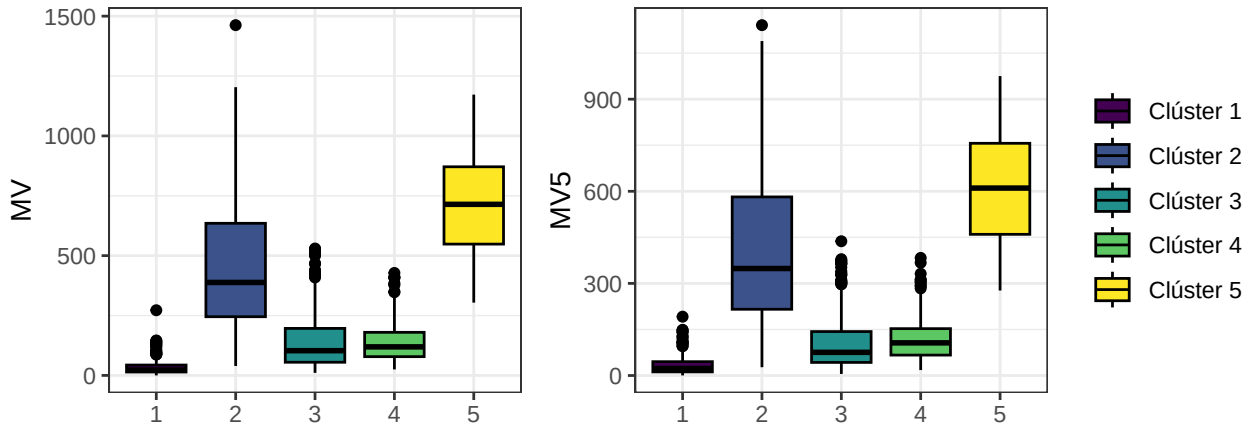
millones. Este orden se repite en las variables relacionadas con el valor de la plantilla, tanto en la valoración actual (MV) como la media de valoración de los últimos cinco años (MV5). En este sentido, el clúster 1 tiene una media actual de €714,53 millones y reciente de €601,82 millones; seguido por el clúster 2 con valores de €462,78 millones y €407,94 millones; el clúster 3 con valores de €143,28 millones y €122,49 millones; el clúster 4 con valores de €139,6 millones y €106,15 millones; y, finalmente, el clúster 5 con valores de €33,67 millones y €32,98 millones. En cuanto a la inversión en construcción de la plantilla (FICHAJES), los clústeres 1 y 2 superan los €100 millones, con medias de €133,7 millones y €101,97 millones, respectivamente. Los clústeres 3 y 4 forman un grupo sin diferencias significativas entre sí, con medias de €27,11 millones y €26,44 millones, respectivamente. Por último, el clúster 5 presenta la media más baja en fichajes, con apenas €4,32 millones. De manera similar, en lo referente al inmovilizado material (MATERIAL), el clúster 1 lidera con un valor medio de €262,21 millones, seguido del clúster 2 con €196,44 millones y el clúster 3 con €53,44 millones (sin diferencias significativas), el clúster 4 con 36,72€ millones y, finalmente, el clúster 5 con €11,33 millones.

Figura 5: Generación de ingresos por clúster



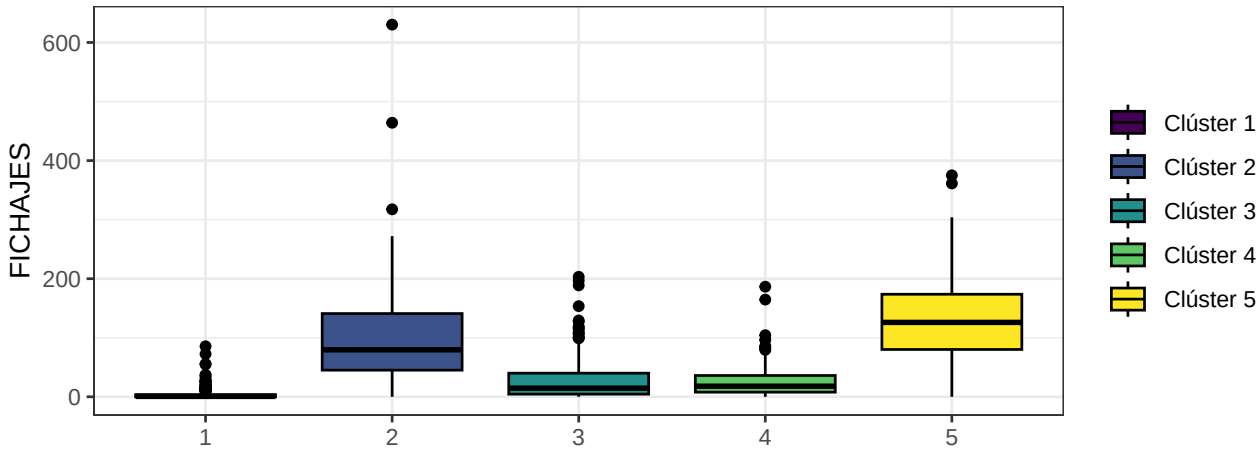
Fuente: Elaboración propia

Figura 6: Valor de mercado de plantilla por clúster



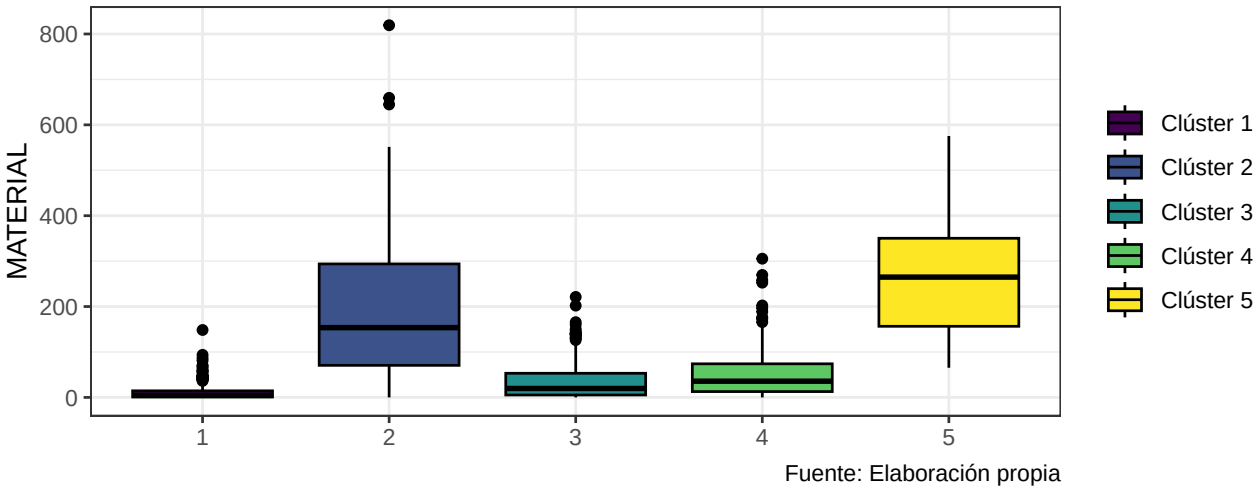
Fuente: Elaboración propia

Figura 7: Gasto en fichajes por clúster



Fuente: Elaboración propia

Figura 8: Cifra de inmovilizado material por clúster



Fuente: Elaboración propia

Grupos diferenciados por generación de ingresos

INGRESOS	GRUPO	CLUSTER
523,12	a	5
256,95	b	2
81,01	c	4
80,69	c	3
23,31	d	1

Grupos diferenciados por valor de plantilla

MV	GRUPO	CLUSTER
714,53	a	5
462,78	b	2
143,28	c	4
139,6	c	3
33,67	d	1

Grupos diferenciados por valor de plantilla (últimos 5 años)

MV5	GRUPO	CLUSTER
601,82	a	5
407,94	b	2
122,49	c	4
106,15	c	3
32,98	d	1

Porcentaje de victorias (WINS)

Asímismo, atendiendo a [] se observa un comportamiento similar respecto a las variables relacionadas con el porcentaje de victorias. En promedio, el clúster 1 gana más de la mitad de sus partidos, con un porcentaje de

Grupos diferenciados por inmovilizado material

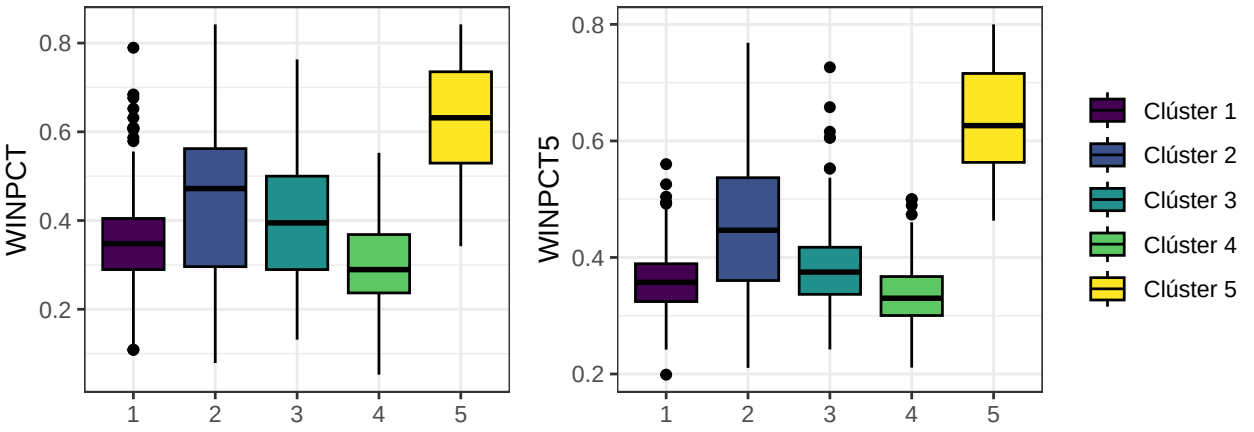
MATERIAL	GRUPO	CLUSTER
262,21	a	5
196,44	b	2
53,44	c	4
36,72	c	3
11,33	d	1

Grupos diferenciados por gasto en fichajes

FICHAJES	GRUPO	CLUSTER
133,7	a	5
101,97	b	2
27,11	c	3
26,44	c	4
4,32	d	1

victorias actual (WINPCT) del 62,4% y un porcentaje reciente (WINPCT5) del 63,56%. En segundo lugar, el clúster 2 logra un porcentaje de victorias del 44,35% en el presente y del 45,96% en el periodo reciente. En tercer lugar, los equipos del clúster 3 tienen un porcentaje de victorias promedio del 40,14% en la actualidad y del 38,33% recientemente, manteniéndose cerca de la media de la muestra.

Figura 9: Porcentaje de victorias por clúster



Fuente: Elaboración propia

Grupos diferenciados por porcentaje de victorias

WINPCT	GRUPO	CLUSTER
62,4	a	5
44,35	b	2
40,14	c	3
35,56	d	1
30,05	e	4

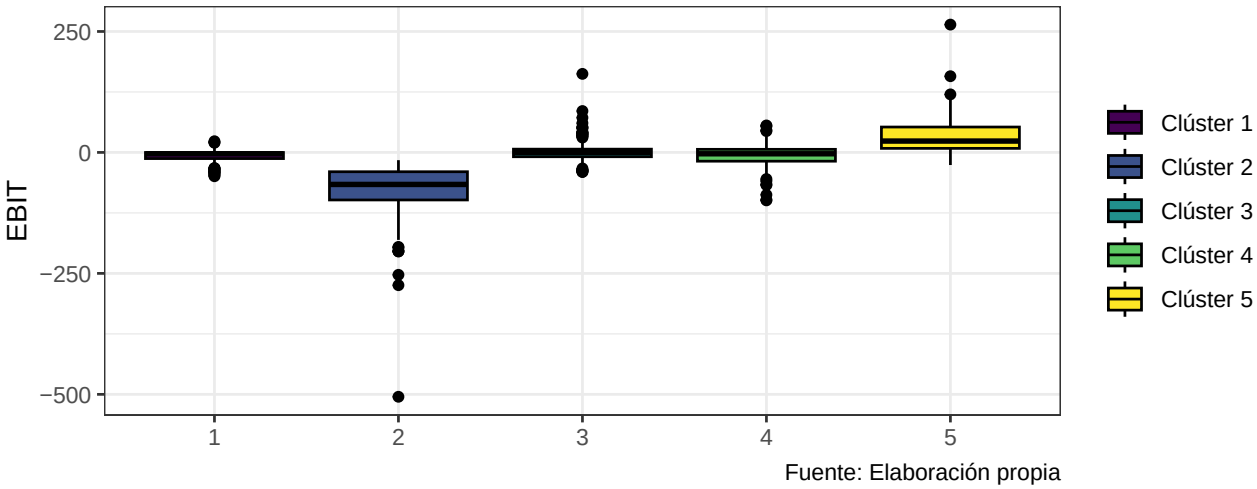
Grupos diferenciados por porcentaje de victorias (últimos 5 años)

WINPCT5	GRUPO	CLUSTER
63,56	a	5
45,96	b	2
38,33	c	3
35,99	d	1
33,91	d	4

Rentención de ingresos (EBIT)

Se observa una marcada polarización en términos del EBIT: el clúster 1 registra el EBIT más elevado, con una media de €35,63 millones de beneficio, mientras que el clúster 5 presenta, en promedio, pérdidas de €-81,57 millones. El clúster 2 también genera beneficios, aunque significativamente menores en comparación con el clúster 1, con una media de €1,01 millones. De manera similar, el clúster 4 produce pérdidas, aunque considerablemente menores respecto al clúster 5, con una media de €-7,41 millones. Finalmente, el clúster 3 se sitúa por debajo del neto cero, con una media de €-6,72 millones, sin diferencias significativas respecto a los clústeres 2 y 4 según las pruebas post-hoc.

Figura 10: Retención de beneficio (EBIT) por clúster



Fuente: Elaboración propia

Grupos diferenciados por EBIT

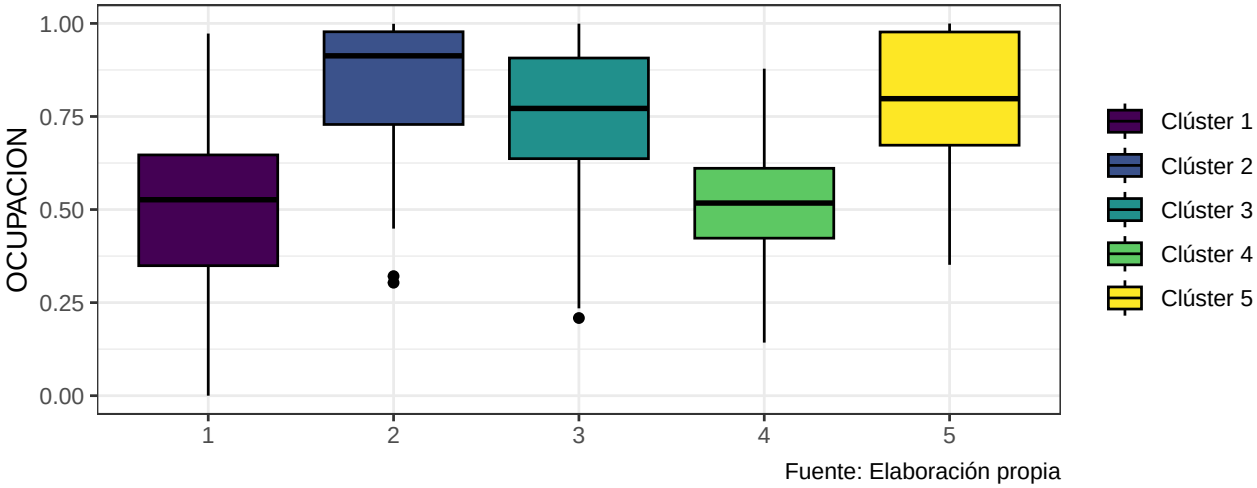
EBIT	GRUPO	CLUSTER
35,63	a	5
1,01	b	3
-6,72	bc	4
-7,41	c	1
-81,57	d	2

Ocupación del estadio (OCUP)

Los clústeres 1 y 2 conforman un mismo grupo con los más altos porcentajes de ocupación, presentando medias del 84,1% y 77,94% respectivamente. En contraste, los clústeres 4 y 5 constituyen otro grupo con los

porcentajes más bajos, con medias de 50,74% y 50,54% respectivamente. Finalmente, el clúster 3 se distingue significativamente de todos los demás en un punto intermedio, con una media del 50,54% de ocupación.

Figura 11: Porcentaje de ocupación del estadio por clúster



Fuente: Elaboración propia

Grupos diferenciados por ocupación del estadio

OCUPACION	GRUPO	CLUSTER
84,1	a	2
77,94	ab	5
74,82	b	3
50,74	c	4
50,54	c	1

Masa de espectadores (CAPAST)

En términos de masa de espectadores, el clúster 1 presenta la mayor capacidad promedio con aproximadamente 73660,22 asientos y una asistencia de 56641,12 espectadores. En un nivel intermedio, se encuentran el clúster 2 y el clúster 3, con capacidades promedio de 50046,73 y 38318,8 asientos, recibiendo a 40898,87 y 22770,54 espectadores, respectivamente. Por último, los clústeres 4 y 5 exhiben capacidades promedio considerablemente más bajas, con 29373,07 y 22816,43 asientos, y asistencias de 18084,06 y 12073,91 espectadores, respectivamente.

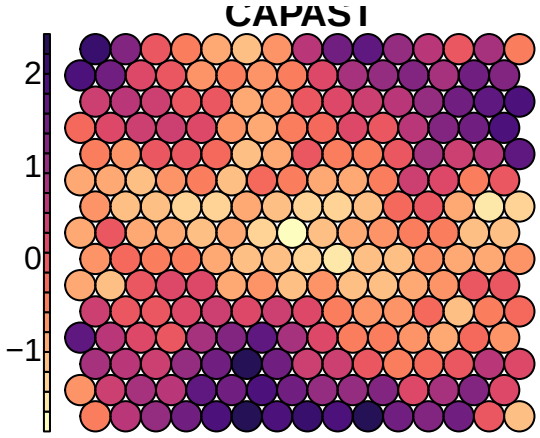
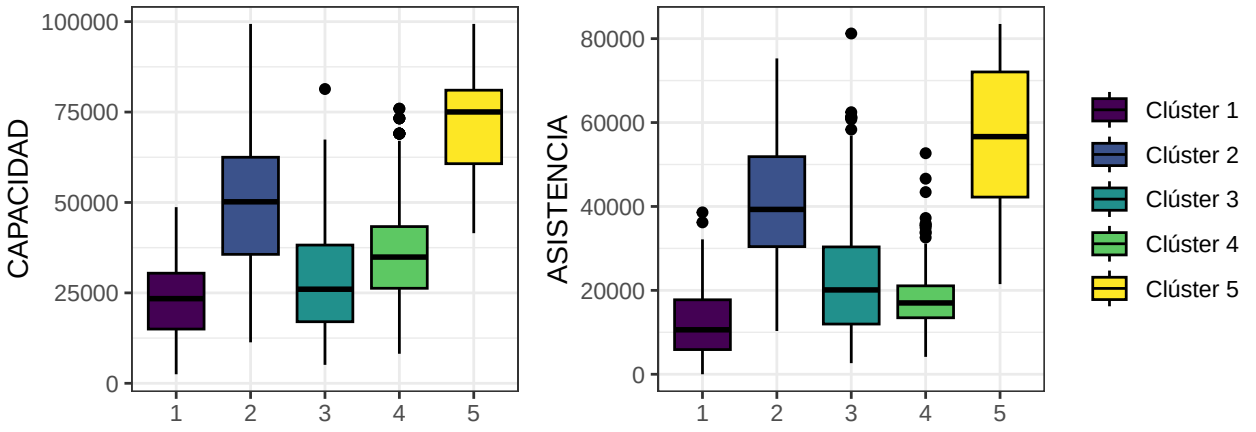


Figura 12: Capacidad total y asistencia al estadio por clúster



Fuente: Elaboración propia

Grupos diferenciados por capacidad máxima del estadio

CAPACIDAD	GRUPO	CLUSTER
73660,22	a	5
50046,73	b	2
38318,8	c	4
29373,07	d	3
22816,43	e	1

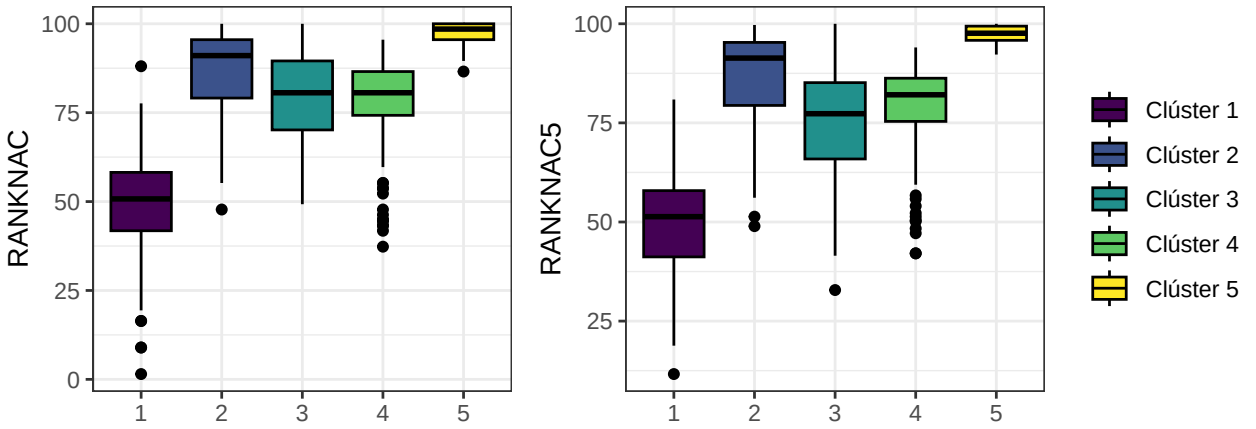
Grupos diferenciados por asistencia al estadio

ASISTENCIA	GRUPO	CLUSTER
56641,12	a	5
40898,87	b	2
22770,54	c	3
18084,06	d	4
12073,91	e	1

Ranking nacional (RANK)

Tanto para el ranking nacional (RANKNAC) actual como para el histórico reciente (RANKNAC5), se observan diferencias significativas al 5%, manteniendo los clústers el mismo orden. En primer lugar, el clúster 1 presenta las medias más altas, con un ranking nacional actual promedio de 97,21 y un histórico reciente de 97,39. A continuación, los clústers 2 y 3 forman un mismo grupo, al no mostrar diferencias significativas al 10% entre sí, con rankings actuales de aproximadamente 87,15 y 86,37, y recientes de 79,64 y 78,94, respectivamente. Posteriormente, el clúster 4 posee una media actual de 78,33 y reciente de 75,46. Finalmente, el clúster 5 muestra los peores rankings, con una media actual de 48,94 y reciente de 52,66.

Figura 13: Ranking nacional por clúster



Fuente: Elaboración propia

Grupos diferenciados por ranking nacional

RANKNAC	GRUPO	CLUSTER
97,21	a	5
87,15	b	2
79,64	c	3
78,33	c	4
48,94	d	1

Grupos diferenciados por ranking nacional (últimos 5 años)

RANKNAC5	GRUPO	CLUSTER
97,39	a	5
86,37	b	2
78,94	c	4
75,46	c	3
49,65	d	1

Descripción y etiquetado de los clústers

- **Clúster 1 - Distress B**: Este grupo incluye equipos que generalmente tienen los peores rendimientos tanto financieros como deportivos. Tienen los ingresos más bajos, con una media de €23.31 millones.

La inversión en fichajes es mínima, con un promedio de €4.32 millones, y también presentan una baja capacidad de estadio y asistencia. En términos de desempeño deportivo, muestran los peores rankings nacionales y un porcentaje de victorias bajo. Su EBIT también es negativo, indicando pérdidas consistentes.

- **Clúster 2 - Élite A-**: Este clúster agrupa a equipos de alto nivel que no alcanzan el rendimiento de los equipos de élite absoluta. Tienen altos ingresos, en promedio €256.95 millones, y una considerable inversión en fichajes de €101.97 millones. Estos equipos también disfrutan de una alta ocupación del estadio (promedio 84.1%) y un EBIT positivo. Su rendimiento deportivo es notable, con un porcentaje de victorias de alrededor de 44.35%.
- **Clúster 3 - Moderados**: Formado por equipos con un rendimiento y estabilidad medios, a menudo incluye los equipos recién ascendidos a primera división. Tienen ingresos moderados (aproximadamente €80.69 millones) y una inversión en fichajes media, unos €27.11 millones. Aunque los porcentajes de ocupación y asistencia al estadio son intermedios, estos equipos muestran una mejora en sus rankings nacionales en los últimos 5 años, sugiere que su rendimiento deportivo y financiero ha ido mejorando.
- **Clúster 4 - Distress A**: Equipos que, aunque tienen ingresos bajos-moderados (alrededor de €81.01 millones), su situación financiera y deportiva está en declive. Presentan pérdidas en EBIT, ocupación baja del estadio y niveles reducidos de asistencia. La inversión en fichajes es similar a la del clúster 3, pero su rendimiento deportivo ha disminuido en comparación con sus propias marcas históricas.
- **Clúster 5 - Élite A+**: Los mejores equipos en términos de desempeño financiero y deportivo. Tienen los ingresos más altos, promediando €523.12 millones, y una inversión significativa en fichajes de €133.7 millones. También presentan las capacidades y asistencias más altas en los estadios, con una ocupación promedio del 77.94%. Líderes en el ranking nacional, muestran un rendimiento deportivo excepcional con el mayor porcentaje de victorias (62.4%) y el EBIT más alto.

Múltiplos de ingresos

El siguiente paso consiste en la estimación de los ratios EV/Ingresos para cada clúster. Esta estimación se basa en las transacciones encontradas sobre las observaciones de los equipos. Para ello, se emplearán tres metodologías de estimación distintas: 1) la media del valor de transacción sobre ingresos por clúster, 2) la mediana, y 3) una media ponderada por distancia. Esta última técnica toma en consideración la distancia existente entre la neurona correspondiente a cada observación y la neurona asignada a cada transacción dentro del clúster. Esta triangulación permitirá una visión más precisa y robusta de la valoración económica de los equipos, proporcionando una base sólida para su valoración.

Tabla 5: Transacciones y ratios por clúster

Cluster	Nº transacciones	% media	Ratio medio	Ratio mediano
Distress B	3	90.000000	2.9718582	2.440206
Élite A-	10	52.474900	2.7938865	1.468683
Moderados	7	32.595714	0.9531937	1.119343
Distress A	6	76.655000	2.2652083	1.283890
Élite A+	12	7.685667	2.3337927	1.958711

Ratio ponderado por distancia

Figura 13c: Diagrama de caja de EV por cluster

Estimado por media ponderada por distancia de EV/Ingresos