UNIVERSIDAD
DE GRANADA

Práctica 2

Análisis Relacional
mediante Segmentación

Mauricio Luque Jiménez

78004003D

maulujim@correo.ugr.es

Subgrupo de prácticas A1

14 de diciembre de 2025

Inteligencia de Negocio

Índice

Introducción

A partir del conjunto de datos descrito en el guión, el objetivo de esta práctica es aprovechar ese volumen de información sobre condiciones de vida para identificar tipologías de hogares que no aparecen explícitamente en los datos. En lugar de predecir una variable concreta, se busca descubrir grupos relativamente homogéneos de hogares que compartan patrones similares en términos de renta, estructura familiar y situación reciente, usando técnicas de aprendizaje no supervisado basadas en clustering.

El enfoque general consiste en seleccionar, para cada caso de estudio, un subconjunto de hogares definido por alguna condición clara (por ejemplo, un tipo de hogar concreto o la presencia de menores), elegir un conjunto coherente de variables numéricas y ordinales, aplicar diferentes algoritmos de agrupamiento y comparar las soluciones obtenidas mediante índices de calidad como el coeficiente de Silhouette y el índice de Calinski-Harabasz, junto con medidas simples de tiempo de ejecución. En concreto, los algoritmos que se van a utilizar en esta práctica son K-Means, clustering jerárquico aglomerativo (se va a calcular varianza mínima Ward), Birch, DBSCAN y MeanShift. A partir de ahí, el trabajo se centra en la interpretación sustantiva de los clusters. Eso sí, antes de realizar operaciones con algoritmos de clustering, se aplicará una normalización a todos los datos, así como una imputación de aquellos valores ausentes que sean útiles para cada caso de estudio.

En esta práctica se han elegido tres casos de estudio que explotan bloques distintos del cuestionario. El primero se centrará en los cambios recientes en los ingresos del hogar y las expectativas económicas de futuro, combinando la evolución declarada de los ingresos y las expectativas para los próximos doce meses. El segundo analizará la estructura familiar y los niveles de bienestar económico, poniendo el foco en tipos de hogar concretos (como parejas con hijos o hogares monoparentales) para estudiar cómo se organizan internamente en función de su renta y/o su situación de pobreza. El tercer caso restringe la atención a hogares con menores de 16 años y utiliza principalmente variables de carencias infantiles para identificar perfiles de privación y bienestar de los niños en relación con la renta disponible y otros indicadores del hogar.

Primer caso de estudio

Caso analizado y motivación

El primer caso de estudio se va a centrar en los ingresos más recientes de cada hogar y su expectativa económica para los próximos meses. Después de unos años de bonanza a principios de siglo, la crisis económica del año 2008 dejó a muchos hogares en una situación económica complicada, una situación de la cual muchas familias no se han recuperado. Si bien en años como 2017 o 2018 había una relativa suficiencia económica a nivel general, la crisis derivada de la pandemia, el aumento del coste de la vida y el estancamiento de los salarios no auguran un futuro alentador para la mayoría de hogares españoles. En ese contexto, en el que el poder adquisitivo de los españoles es cada vez menor, puede ser interesante encontrar patrones y relacionar las condiciones de la mayoría de hogares con su perspectiva económica a futuro.

Para realizar este análisis, se intentará extraer conclusiones de la salud económica y de la deriva que afrontan los hogares españoles. Para ello, se tendrán en cuenta datos como la renta antes y después de ayudas sociales, la posibilidad de irse de vacaciones, el acceso a comida de calidad, la capacidad para llegar a fin de mes y, por último, la expectativa dentro del propio hogar de cara al futuro.

Resultados obtenidos y análisis de parámetros

El primer paso, una vez leídos los datos, aprovechando que hay una variable que indica si se ha recogido información sobre un cambio de ingresos en los últimos meses (HI010), se van a seleccionar únicamente aquellas instancias con respuesta registrada, sean cambios positivos o negativos. De las 29781 instancias totales que tiene el conjunto de datos, el número de filas con las que se va a trabajar es 29714. Al haber muy pocas instancias que quedan excluidas, se puede proceder al estudio sin tener estos valores en cuenta, aunque también se podría hacer una imputación (como se va a hacer para el resto de atributos que se van a analizar para obtener resultados y sacar conclusiones). Ahora que ya tenemos el conjunto de instancias, vamos a seleccionar los atributos que vamos a consultar. Concretamente, en este caso de estudio los atributos elegidos son:

- HY020 (renta disponible total del hogar)
- HY022 (renta antes de transferencias sociales excepto prestaciones)
- HY023 (rentas antes de transferencias sociales incluyendo prestaciones)
- vhRentaa (renta total del hogar, incluyendo esquemas privados de pensiones)
- HS040 (posibilidad de vacaciones fuera de casa al menos una semana al año)
- HS050 (comida de calidad: acceso a carne, pollo o pescado cada dos días)
- HS060 (capacidad para afrontar gastos imprevistos)
- HS120 (capacidad del para llegar a fin de mes)
- HI040 (expectativa de ingresos en el próximo año)

Respectivamente, estas variables van a renombrarse de las siguiente manera: *renta_disp, renta_antes_trans_no_pensiones, renta_antes_todas_trans, renta_equiv, vacaciones, comida_calidad, gasto_imprevisto, fin_de_mes, exp_ingresos.* Después de normalizar todos los datos en un rango entre 0 y 1, se va a realizar una imputación de valores perdidos sobre la matriz ya normalizada mediante KNNI, al que le vamos a indicar que se fije en los 5 vecinos más cercanos.

```
NaN por columna antes de imputar:
renta_disp                          0
renta_antes_trans_no_pensiones      0
renta_antes_todas_trans             0
renta_equiv                         0
vacaciones                         62
comida_calidad                     16
gasto_imprevisto                   37
fin_de_mes                         56
exp_ingresos                     3346
dtype: int64
```

```
NaN por columna después de imputar:
renta_disp                         0
renta_antes_trans_no_pensiones     0
renta_antes_todas_trans            0
renta_equiv                        0
vacaciones                         0
comida_calidad                     0
gasto_imprevisto                   0
fin_de_mes                         0
exp_ingresos                       0
dtype: int64
```

Ahora que ya tenemos los datos normalizados y sin valores ausentes, podemos aplicar los algoritmos de clustering como tal. Para cada algoritmo de clustering mencionado anteriormente en la introducción de la memoria se van a probar ejecuciones con distintos parámetros. Para K-Means, clustering aglomerativo y Birch, que necesitan definir de inicio el número de clusters, se van a probar ejecuciones con valores entre 2 y 8. Para DBSCAN, se van a probar valores entre 0'2 y 0'5 para decidir el radio de alcance de cada punto, y se van a probar tamaños mínimos de 20 y 50 puntos. Por último, para Birch se van a probar varios hiperparámetros para estimar el radio ideal.

Para probar todas las diferentes ejecuciones, se va a ejecutar cada algoritmo dentro de un bucle en el que, además de probar cada algoritmo con sus propios hiperparámetros, se medirá el tiempo de ejecución, De esta manera, una vez se ejecuten todos los algoritmos (cada resultado se va almacenando), tenemos una tabla global con todos los resultados obtenidos. A continuación se recoge una parte de los resultados, eligiendo para cada algoritmo el conjunto de parámetros que obtienen el mejor coeficiente de Silhouette.

	K-Means	Aglomerativo	Birch	DBSCAN	Mean Shift
Clusters	5	5	5	8	8
Silhouette	0.631740	0.628068	0.632444	0.637683	0.640083
Calinski Harabatz	30955.570385	3040.463760	30934.853818	20278.776104	2079.762232
Tiempo	0.069453	0.127964	0.179733	5.651010	0.345910
K	5.0	5.0	5.0	-	-
Treshold (Birch)	-	-	0.5	-	-
Radio (DBSCAN)	-	-	-	0.4	-
Tamaño mínimo (DBSCAN)	-	-	-	20	-
Radio (Mean Shift)	-	-	-	-	0.621552

En los algoritmos donde k es un parámetro explícito (K-Means, aglomerativo y Birch) las métricas de calidad mejoran al aumentar k hasta valores en torno a 4–5 clusters, y a partir de ahí tanto Silhouette como CH tienden a empeorar, lo que sugiere que un número intermedio de grupos describe mejor la estructura de los datos que particiones demasiado gruesas o demasiado finas. De hecho, los tres algoritmos coinciden en que obtienen su mejor coeficiente de Silhouette cuando forman 5 clusters. Por otra parte, en DBSCAN y MeanShift, donde el número de clusters es un resultado emergente, las mejores métricas se obtienen también con un número más alto de clusters (en torno a 8–10 para DBSCAN y 8–9 para MeanShift) mientras que configuraciones que inducen muy pocos o demasiados clusters producen segmentaciones de menor calidad. Igual que ocurre con los tres primeros algoritmos, tanto DBSCAN como Mean Shift obtienen su mejor coeficiente de Silhouette cuando forman el mismo número de clusters, en este caso 8. A continuación se muestran cinco tablas, una para cada algoritmo con todas las combinaciones de hiperparámetros que se han probado.

K-Means			
K (clusters)	Silhouette	Calinski Harabatz	Tiempo
2	0.531659	30355.257557	0.077926
3	0.539624	26868.973380	0.060825
4	0.607464	29978.155706	0.071889
5	0.631740	30955.570385	0.069453
6	0.541719	30427.739110	0.083260
7	0.460589	27663.477402	0.084242
8	0.519981	28518.963981	0.110399

Clustering jerárquico (aglomerativo)			
K (clusters)	Silhouette	Calinski Harabatz	Tiempo
2	0.516734	2867.800064	0.131400
3	0.527267	2532.313897	0.127235
4	0.591328	2665.438180	0.126449
5	0.628068	3040.463760	0.127964
6	0.538840	2912.793033	0.127606
7	0.473985	2775.261206	0.157774
8	0.482924	2701.001037	0.129678

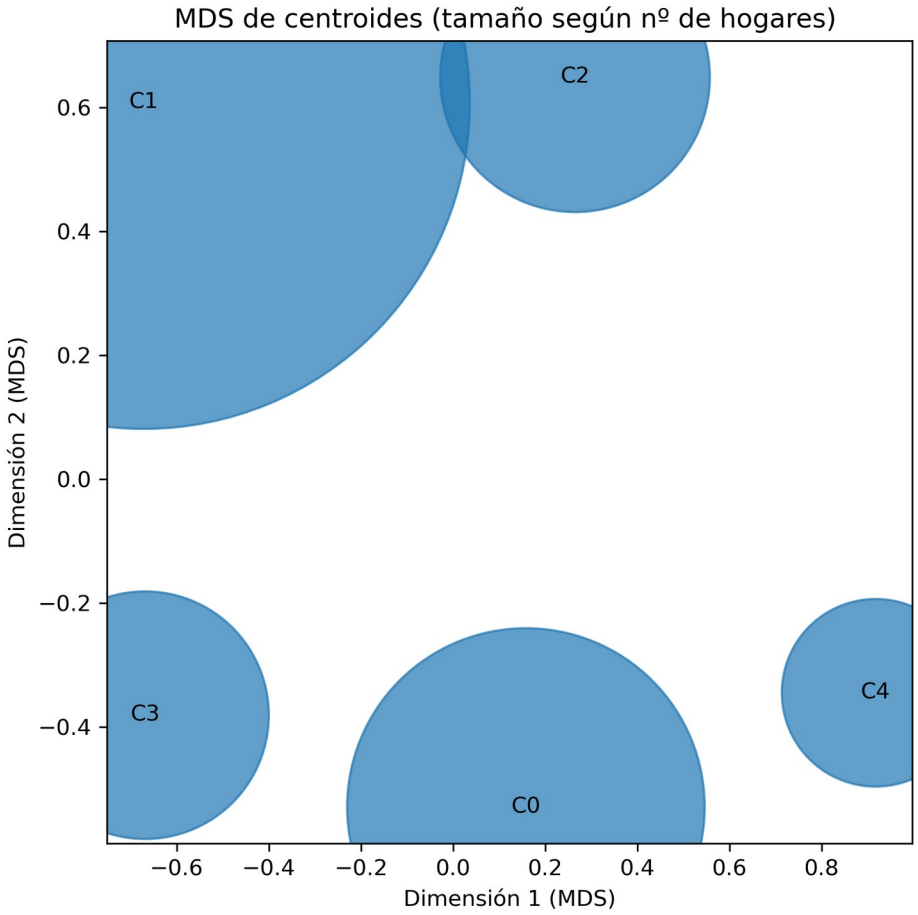
Birch				
K (clusters)	Treshold	Silhouette	Calinski Harabatz	Tiempo
2	0.5	0.539087	30202.584291	0.157453
3	0.5	0.517782	24386.336738	0.162318
4	0.5	0.602286	29328.500441	0.154870
5	0.5	0.632444	30934.853818	0.179733

DBSCAN					
Clusters	Radio	Número de puntos	Silhouette	Calinski Harabatz	Tiempo
33	0.2	20	0.315720	6848.969360	4.246177
23	0.2	50	0.407829	12284.611918	4.114599
10	0.3	20	0.636051	15883.582973	5.693637
8	0.3	50	0.637395	20258.908172	5.682886
8	0.4	20	0.637683	20278.776104	5.651010
8	0.4	50	0.637445	20308.240506	6.021660
4	0.5	20	0.205111	292.397413	6.085154
4	0.5	50	0.195023	290.960436	5.712194

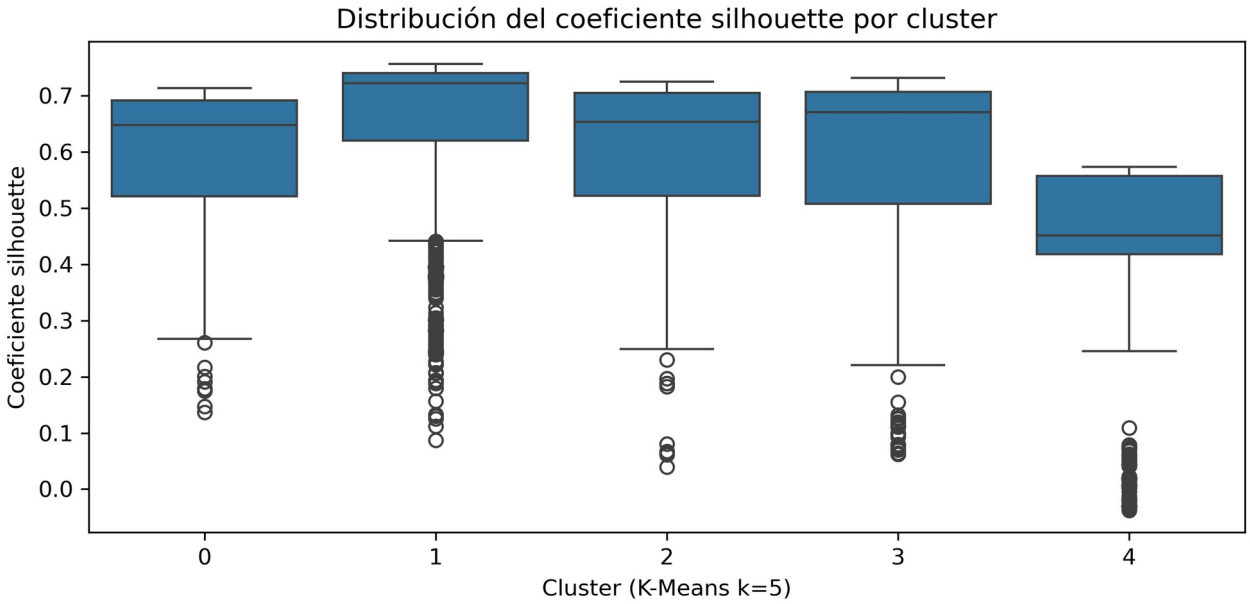
Mean Shift				
Clusters	Radio	Silhouette	Calinski Harabatz	Tiempo
16	0.423747	0.545043	1690.077701	0.345910
8	0.621552	0.640083	2079.762232	0.400162

Interpretación de la segmentación

Para analizar los resultados y mostrar gráficas que acompañen el análisis, se va a tomar como referencia la segmentación realizada con K-Means con un valor K=5. Para empezar, como se puede comprobar en el gráfico de burbujas con escalado multidimensional (MDS), se puede comprobar que el reparto de clusters es bastante razonable, ya que los tamaños de cada cluster (directamente proporcional al número de hogares que forman cada cluster) es bastante equilibrado: no todos los clusters tienen el mismo tamaño, pero tampoco hay unos excesivamente más grandes que otros. Además, están bastante separados entre ellos, con una ligera excepción de algunos casos entre dos clusters.



Además, otra manera de comprobar que K=5 es un parámetro válido para realizar la segregación es comprobar la distribución de coeficientes de Silhouette, que se mantiene de manera uniforme y con valores que, si bien no son brillantes, son razonables.

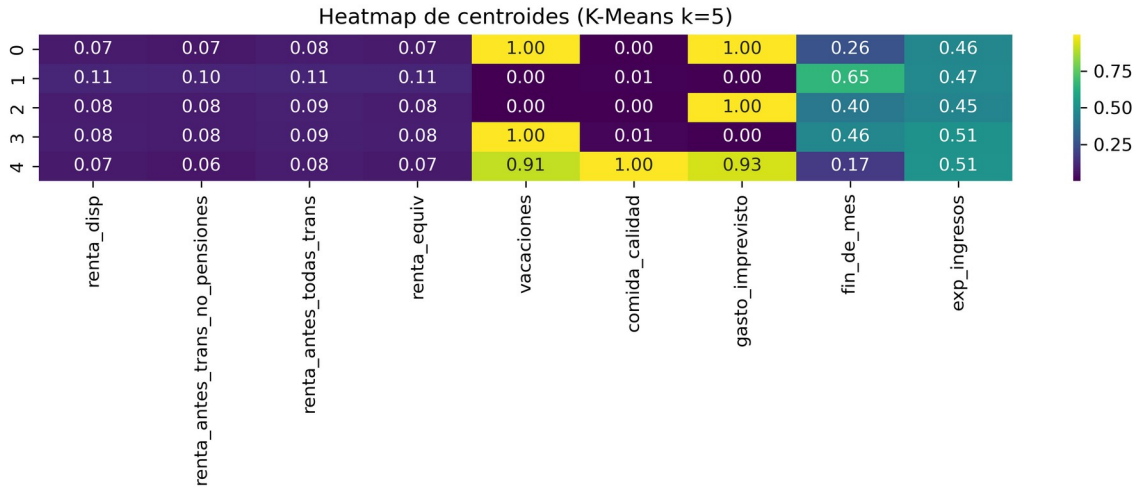


Como se puede ver en el heatmap que se obtiene al realizar esta ejecución, hay tres variables que condicionan altamente la segmentación de los hogares: la capacidad de tomarse unas vacaciones fuera de casa, el acceso a comida de calidad (pollo, carne y pescado) cada dos días y la capacidad de responder a un gasto imprevisto. Para esas tres variables, hay que tener en cuenta los posibles valores que se podían obtener: originalmente, se podía obtener un 1 o un 2 (ya que el 0 estaba reservado para otras opciones), de manera que, al normalizar los datos, el 0 significa que la respuesta a la pregunta en cuestión es positiva, mientras que si el valor tiende a 1 la respuesta es negativa. De esta manera, tenemos cinco clusters:

- Hogares que no pueden permitirse tanto las vacaciones, ni la comida de calidad ni un gasto imprevisto. Es el grupo con mayor dificultad para llegar a fin de mes (su código de respuesta, TH120, indica que los valores más cercanos a 1, o 0 normalizado, representan “mucha dificultad” para llegar a fin de mes). Además, tienen la valoración más alta en la expectativa a futuro, lo que en este caso significa que esperan empeorar su situación, algo lógico si se tienen en cuenta todas las dificultades comentadas.
- Hogares que no pueden permitirse vacaciones y tampoco un gasto imprevisto, pero sí la comida de calidad. Son el siguiente grupo con mayor dificultad para llegar a fin de mes, algo lógico porque son los segundos que menos cosas pueden permitirse, aunque su expectativa a futuro es mejor. Siendo el segundo cluster de mayor tamaño, este grupo de hogares es una buena representación parcial de la población.

- Hogares que no pueden permitirse vacaciones, pero sí la comida de calidad y el gasto imprevisto. Su capacidad para llegar a fin de mes es buena, pero no tienen una gran expectativa a futuro. Esto indica que es un hogar en el que se cubren necesidades básicas y tienen cierto margen, pero no pueden permitirse muchas alegrías.
- Hogares que pueden permitirse vacaciones y comida de calidad, pero no un gasto imprevisto. Esto puede suponer que las vacaciones forman parte de los planes que realiza el hogar, por lo que entra dentro de los gastos cubiertos, pero no tienen mucho margen si ocurre una incidencia. Esto puede suponer que, igual que los grupos anteriores, tengan que cancelar dichas vacaciones si surge un problema que requiere un gasto extra.
- Hogares que pueden permitírselo todo. Evidentemente son los que tienen más solvencia para llegar a fin de mes, aunque tampoco tienen una gran expectativa de futuro a nivel económico. Este es el grupo más grande y más representativo de toda la población estudiada: una economía de clase media, que puede vivir sin preocupaciones serias para llegar a fin de mes y que de vez en cuando puede permitirse algún extra como unas vacaciones, pero no demasiados y sin una sensación de ir a más.

Que este grupo sea el más representativo muestra el estado actual de la población española: una economía de primer punto sostenida con alfileres y sin expectativa de poder mejorar su situación actual, lo que habla a las claras de un estancamiento problemático a medio y largo plazo.



Segundo caso de estudio

Caso analizado y motivación

El segundo caso de estudio se va a centrar en analizar la estructura familiar de cada hogar y si tiene relación con el bienestar económico del núcleo familiar. Durante los años 60 del siglo pasado tuvo lugar el *baby boom* en el que era normal encontrar familias con gran cantidad de hijos, lo que dio como resultado unos años 80 con una gran cantidad de españoles entre los 20 y 40 años, que permitía mucha actividad en el mercado laboral, una gran cantidad de bienes y activos en circulación y una aportación general a la seguridad social muy positiva. Sin embargo, después de esa época con familias muy numerosas, en la actualidad el número de dichas familias se ha reducido bastante y cada vez los hogares tienden más a parejas sin hijos o directamente hogares unipersonales. Afectados también por la situación económica general analizada en el primer caso de estudio, puede ser interesante analizar la relación entre los distintos núcleos familiares y su situación económica, a fin de poder entender qué se debe mejorar para permitir familias más numerosas y evitar problemas futuros como la sostenibilidad de un sistema de pensiones que con los años va teniendo más personas, los niños del *baby boom*, mientras el número de contribuyentes es menor y exige un mayor esfuerzo a cada persona activa.

Para realizar este análisis, primeramente se diferenciarán los distintos tipos de hogares en función del número de miembros, y posteriormente se analizarán atributos que muestren tanto capacidad económica como calidad de vida: renta, capacidad para llegar a fin de mes, acceso a vacaciones o comida de calidad, índice de pobreza o capacidad material.

Resultados obtenidos y análisis de parámetros

Para este estudio, vamos a fijarnos en el atributo HX060, que indica el número de miembros por hogar. Vamos a seleccionarlos prácticamente todos, quitando los casos en los que se indican “otros hogares” con o sin niños dependientes. De esta manera, de las 29781 instancias iniciales, nos vamos a quedar con 25296 instancias, lo cual nos deja un conjunto suficiente para trabajar. Ahora que ya tenemos el conjunto de instancias, vamos a seleccionar los atributos que vamos a consultar. Concretamente, en este caso de estudio los atributos elegidos son:

- HX040 (número de miembros dentro del hogar)
- HX060 (tipo de hogar)
- vhRentaa (renta total del hogar, incluyendo esquemas privados de pensiones)
- HS040 (posibilidad de vacaciones fuera de casa al menos una semana al año)
- HS050 (comida de calidad: acceso a carne, pollo o pescado cada dos días)
- HS060 (capacidad para afrontar gastos imprevistos)
- HS120 (capacidad del para llegar a fin de mes)
- vhPobreza (hogar en riesgo de pobreza)
- vhMATDEP (hogar en carencia material severa)

Respectivamente, estas variables van a renombrarse de las siguiente manera: *n_miembros, tipo_hogar, renta_equiv, vacaciones, comida_calidad, gasto_imprevisto, fin_de_mes, pobreza, carencia_material.* Después de normalizar todos los datos en un rango entre 0 y 1, se va a realizar una imputación de valores perdidos sobre la matriz ya normalizada mediante KNNI, al que le vamos a indicar que se fije en los 5 vecinos más cercanos. En este caso, nos encontramos con muy pocos valores ausentes para estas variables, lo cual es fácilmente corregible con KNNI sin crear demasiados datos sintéticos.

```
NaN por columna antes de imputar:
n_miembros            0
hay_menores           0
monoparental          0
pareja_hijos          0
unipersonal           0
renta_equiv           0
fin_de_mes           63
gasto_imprevisto     42
vacaciones           61
comida_calidad       19
pobreza               0
carencia_material     0
dtype: int64
```

```
NaN por columna después de imputar:
n_miembros           0
hay_menores          0
monoparental         0
pareja_hijos         0
unipersonal          0
renta_equiv          0
fin_de_mes           0
gasto_imprevisto     0
vacaciones           0
comida_calidad       0
pobreza              0
carencia_material    0
dtype: int64
```

Ahora que ya tenemos los datos normalizados y sin valores ausentes, podemos aplicar los algoritmos de clustering como tal. Vamos a realizar las mismas pruebas que en el anterior caso de estudio. Se han realizado las mismas pruebas que en el primer caso, de manera que a continuación se recoge una parte de los resultados, eligiendo para cada algoritmo el conjunto de parámetros que obtienen el mejor coeficiente de Silhouette.

	K-Means	Aglomerativo	Birch	DBSCAN	Mean Shift
Clusters	8	8	6	45	3
Silhouette	0.466994	0.461003	0.402761	0.779489	0.773873
Calinski Harabatz	9341.200177	1044.788634	7235.603783	20076.095981	1505.679201
Tiempo	0.089542	0.109449	0.232350	2.206766	0.293812
K	8	8	6	-	-
Treshold (Birch)	-	-		-	-
Radio (DBSCAN)	-	-	-	0.3	-
Tamaño mínimo (DBSCAN)	-	-	-	50	-
Radio (Mean Shift)	-	-	-	-	1.407030

En este caso, en algoritmos que fijan los clusters inicialmente, se forman valores más altos que en el primer caso, lo cual indica que los datos son más dispersos y hay patrones menos reconocibles. Esto se nota especialmente en DBSCAN, que no fija el número de clusters inicial y genera 45 clusters, lo que indica que es difícil agrupar los datos. En cambio, Mean Shift sí consigue generar tres clusters, aunque para ello necesite fijar un radio muy amplio, lo cual limita la homogeneidad de los datos y los hace menos interpretables. Cabe destacar también que estos ejemplos se han escogido utilizando como único criterio el coeficiente de Silhouette para simplificar la muestra, pero sería erróneo fijarse únicamente en eso para decidir qué atributo es mejor para cada algoritmo. De hecho, esto se comprueba revisando el coeficiente Calinski Harabatz, que no ofrece muy buenos resultados en aquellas pruebas en las que se ha obtenido mejor coeficiente de Silhouette.

Para mostrar mejor los diferentes resultados obtenidos, a continuación se muestran cinco tablas, una para cada algoritmo con todas las combinaciones de hiperparámetros que se han probado.

K-Means			
K (clusters)	Silhouette	Calinski Harabatz	Tiempo
2	0.353162	11092.471465	0.074675
3	0.375016	10779.357797	0.061547
4	0.376944	10022.250416	0.065448
5	0.411993	10150.802873	0.075208
6	0.444920	10144.696383	0.075856
7	0.457584	9629.215215	0.126899
8	0.466994	9341.200177	0.089542

Clustering jerárquico (aglomerativo)			
K (clusters)	Silhouette	Calinski Harabatz	Tiempo
2	0.363033	1367.224345	0.110662
3	0.333345	1172.488721	0.108790
4	0.349121	1035.176031	0.110352
5	0.391691	986.861444	0.121640
6	0.421303	991.061003	0.111823
7	0.434691	994.298155	0.116691
8	0.462238	1008.149311	0.109449

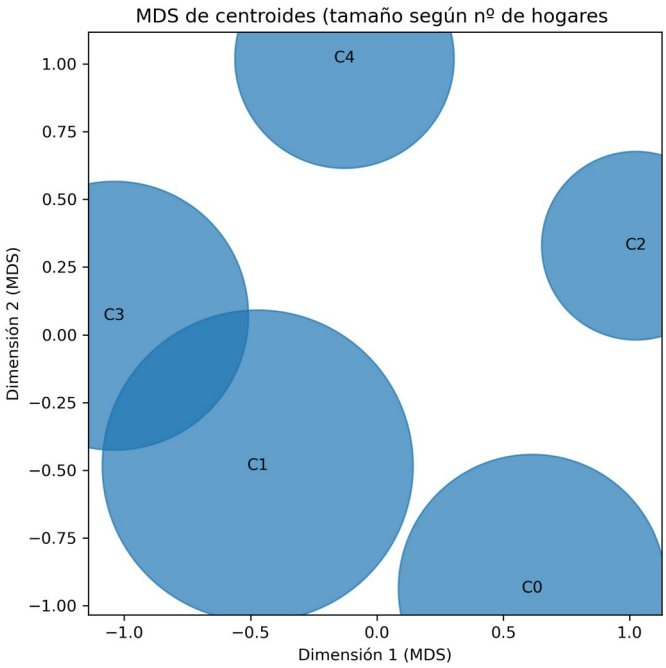
Birch				
K (clusters)	Treshold	Silhouette	Calinski Harabatz	Tiempo
2	0.5	0.333179	9696.119011	0.249805
3	0.5	0.359842	8336.478353	0.306064
4	0.5	0.388640	7214.585463	0.245237
5	0.5	0.376392	5920.867353	0.249552
6	0.5	0.402761	7235.603783	0.248912
7	0.5	0.385330	6742.092803	0.294338
8	0.5	0.387750	5891.014679	0.248087

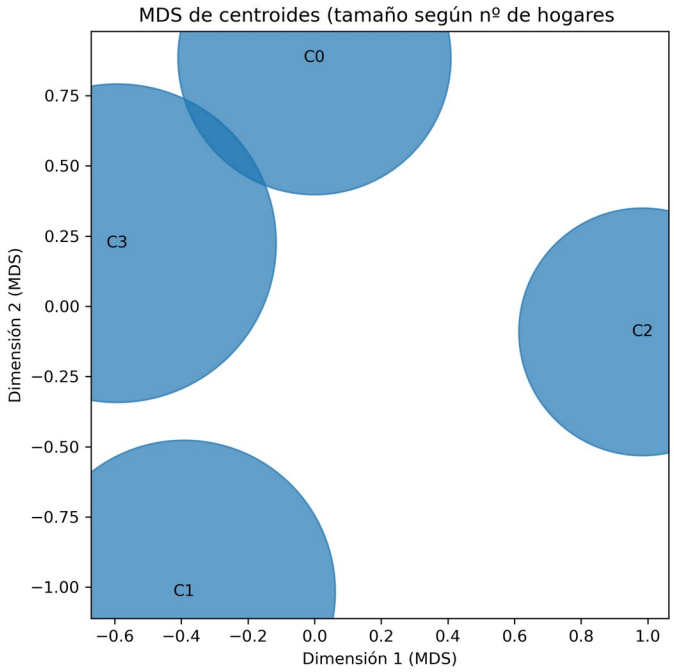
DBSCAN					
Clusters	Radio	Número de puntos	Silhouette	Calinski Harabatz	Tiempo
111	0.2	20	0.532729	17448.620921	2.367772
65	0.2	50	0.551692	25133.425809	1.771122
57	0.3	20	0.778440	16167.958093	2.157638
45	0.3	50	0.779489	20076.095981	2.048607
57	0.4	20	0.777415	16126.651485	2.150110
45	0.4	50	0.776411	20014.060331	2.151688
44	0.5	20	0.421829	2152.206979	2.083435
32	0.5	50	0.440781	2852.070604	2.253663

Mean Shift				
Clusters	Radio	Silhouette	Calinski Harabatz	Tiempo
75	0.754077	0.768252	1400.269402	0.306598
6	1.138542	0.407122	950.319005	0.429328
3	1.394246	0.383401	921.102364	0.502157

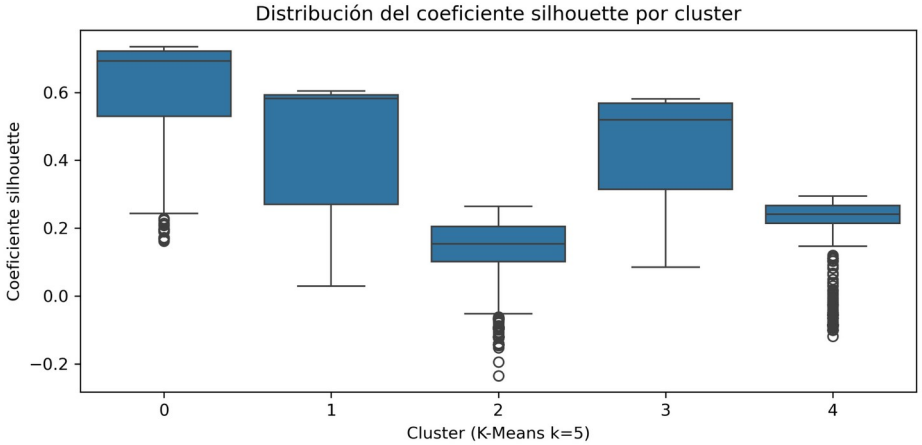
Interpretación de la segmentación

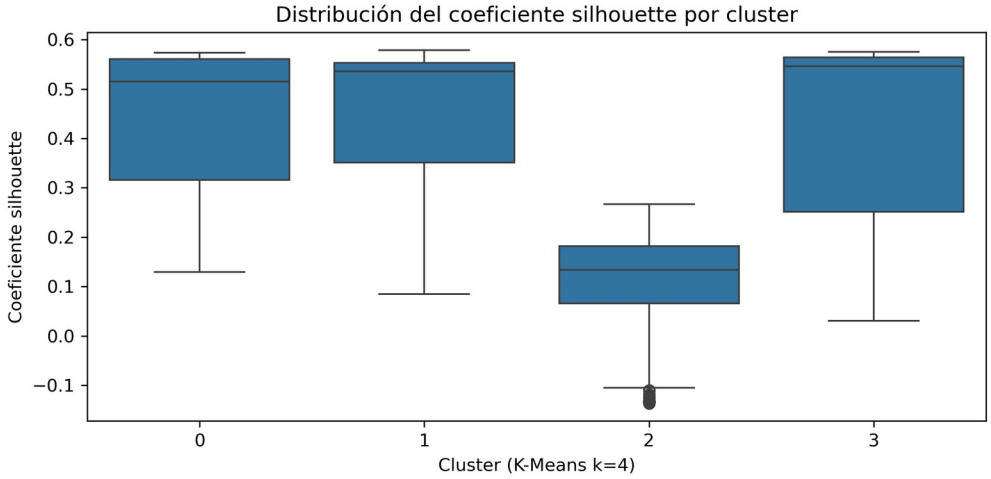
Para poder realizar una interpretación razonable de la segmentación, vamos a elegir una segmentación realizada aplicando el algoritmo K-Means con K=4. Los motivos para esta configuración son los siguientes: mantener una muestra de resultados parecida al caso anterior y hacerla lo más legible posible sin perder calidad en la segmentación. Para valores muy altos de K, aunque se gane mayor precisión métrica, se pierde interpretabilidad, lo cual no es de mucho interés para esta práctica.





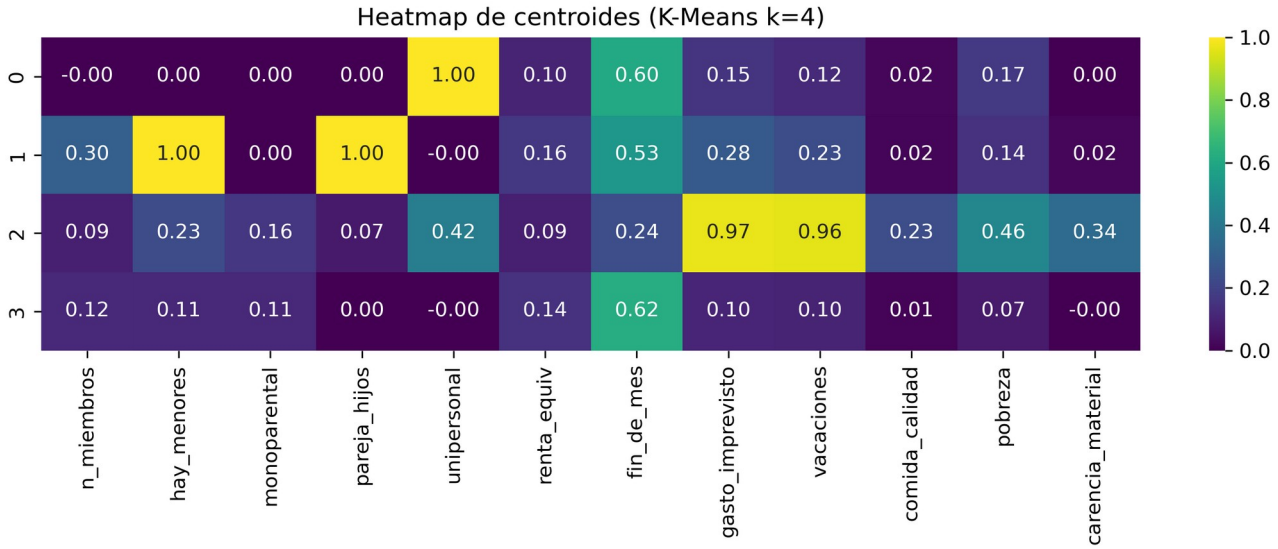
Por ejemplo, en este caso, si repetimos K=5 como en el primer caso, vemos dos clusters con muchos elementos en la frontera de dos clusters, lo que se podría entender como un cluster que se ha separado forzosamente para llegar a 5 clusters. En cambio, con un valor ligeralmente menor, K=4, la separación de los datos es más notoria, con menos elementos que podrían estar en otro grupo, y mantiene valores de Silhouette y Calinski-Harabatz medianamente decente. Se ve más claro con la comparación de la distribución del coeficiente Silhouette, donde se obtienen valores claramente mejores con K=4.





Centrados en K=4, si analizamos el heatmap de centroides, vemos que una de las principales variables de interés para segmentar es la que originalmente conocíamos como tipo de hogar, además de variables que ya hemos examinado como la facilidad para llegar a fin de mes, afrontar gastos imprevistos o tomarse unas vacaciones, aparte de obtener comida de calidad o índice de pobreza. Aquí encontramos cuatro grupos:

- Un primer grupo en el que en el hogar habita únicamente una persona, sin más adultos ni niños. Este grupo destaca por tener gran solvencia para llegar a fin de mes, y por no llegar ni a la pobreza ni a la carencia material, lo cual está relacionado con lo que comentamos en el primer caso: la mayoría de la población es económicamente y la natalidad ha disminuido respecto a hace un par de décadas, lo que disminuye la contribución a las arcas públicas y congestiona servicios como la seguridad social y el sistema de pensiones.
- Un segundo grupo que en principio representa la familia tradicional: una pareja con hijos, principalmente menores. Este grupo de familias, relacionándolo con el caso anterior, no tiene mucha solvencia económica, si bien están lejos de situaciones de pobreza. Además, la carencia material y de comida de calidad es muy baja, lo que indica que en el día no pasan apuros, aunque no contemplan vacaciones tanto como los que no tienen hijos y tienen menos capacidad de reacción a un gasto inesperado. En general, una familia de clase media-baja que vive con lo justo.
- Un tercer grupo con familias no tradicionales, donde hay más familias monoparentales que formadas por una pareja con hijos, o directamente hogares unipersonales. Aquí lo que destaca es la dificultad de llegar a fin de mes, la imposibilidad de plantearse unas vacaciones o de afrontar un gasto inesperado, y unos valores más preocupantes relacionados con comida de calidad, pobreza y escasez material. Este grupo apunta directamente a familias vulnerables, en las que la situación económica del adulto responsable (haya hijos de por medio o no) no es suficiente para llevar el día a día con suficiencia.
- Un último grupo que descarta completamente las parejas con hijos o los hogares unipersonales, lo que nos deja un caso de familias con un sólo adulto responsable a cargo de un menor, pero con bastante solvencia económica a diferencia del grupo anterior.


Heatmap de centroides (K-Means k=4)
0
1
2
3
-0.00 0.00 0.00 0.00 1.00 0.10 0.60 0.15 0.12 0.02 0.17 0.00
0.30 1.00 0.00 1.00 -0.00 0.16 0.53 0.28 0.23 0.02 0.14 0.02
0.09 0.23 0.16 0.07 0.42 0.09 0.24 0.97 0.96 0.23 0.46 0.34
0.12 0.11 0.11 0.00 -0.00 0.14 0.62 0.10 0.10 0.01 0.07 -0.00
n_miembros
hay_menores
monoparental
pareja_hijos
unipersonal
renta_equiv
fin_de_mes
gasto_imprevisto
vacaciones
comida_calidad
pobreza
carencia_material
1.0
0.8
0.6
0.4
0.2
0.0

Tercer caso de estudio

Caso analizado y motivación

El tercer caso de estudio se va a centrar en la situación de los menores de 16 años. Después de haber analizado la bonanza económica en general y su relación con cada estructura familiar, es interesante comprobar qué impacto tienen en los jóvenes, que han nacido en un contexto condicionado primero por la crisis financiera de 2008 y posteriormente por la pandemia.

Para realizar este análisis, primeramente se diferenciarán los distintos tipos de hogares en función del número de miembros, y posteriormente se analizarán atributos que muestren tanto capacidad económica como calidad de vida: renta, capacidad para llegar a fin de mes, acceso a vacaciones o comida de calidad, índice de pobreza o capacidad material.

Resultados obtenidos y análisis de parámetros

Para este caso de estudio, una vez leídos los datos, vamos a filtrar por la misma variable que en el caso anterior: el tipo de hogar. En este caso nos interesan todos los hogares en los que habiten menores, por lo que vamos a elegir todas las columnas que incluyan menores y vamos a descartar las demás.

Código	Descripción
	No consta
1	Una persona: hombre de menos de 30 años
2	Una persona: hombre de entre 30 y 64 años
3	Una persona: hombre de 65 o más años
4	Una persona: mujer de menos de 30 años
5	Una persona: mujer de entre 30 y 64 años
6	Una persona: mujer de 65 o más años
7	2 adultos sin niños dependientes económicamente, al menos...
8	2 adultos sin niños dependientes económicamente, teniendo...
9	Otros hogares sin niños dependientes económicamente
10	Un adulto con al menos un niño dependiente
11	Dos adultos con un niño dependiente
12	Dos adultos con dos niños dependientes
13	Dos adultos con tres o más niños dependientes
14	Otros hogares con niños dependientes

```
# ------------------------------------------------------------
# 2) Selección de segmento: Hogares según tipo y número de miembros
# ------------------------------------------------------------
datos["HX060"] = pd.to_numeric(datos["HX060"], errors = "coerce")

# Elegimos todos los posibles valores de tipo de hogar donde pueda haber menores
codigos = [10,11,12,13, 14]
subset = datos[datos["HX060"].isin(codigos)].copy()
```

De esta manera, el conjunto de variables sobre las que vamos a trabajar es el siguiente:

- HCH010 (asistencia médica no recibida)
- HCH030 (asistencia dental no recibida)
- HD100 (tenencia de ropa nueva)
- HD110 (tenencia de dos pares de zapatos adecuados)

- HD120 (fruta fresca y verdura al menos una vez al día)
- HD140 (carne, pollo o pescado al menos una vez al día)
- HD150 (libros adecuados para su edad)
- HD160 (equipos de ocio al aire libre)
- HD170 (juguetes)
- HD180 (actividades de ocio)
- HD190 (celebración de ocasiones especiales)
- HD200 (frecuencia de reuniones con amigos)
- HD210 (participación en viajes o acontecimientos escolares)
- HD220 (lugar de estudio adecuado)
- HD240 (vacaciones fuera de casa al menos una vez al día)

Respectivamente, estas variables van a renombrarse de la siguiente manera. *medico, dentista, ropa_nueva, pares_zapatos, comida_fresca, proteinas, libros, equipos_aire_libre, juguetes, ocio, ocasiones_especiales, reuniones_amigos, excursiones, lugar_estudio, vacaciones.*

El siguiente paso, una vez se han normalizado los datos, es imputar valores perdidos. En este caso, posiblemente al tratarse de datos sobre menores, la información está mucho más restringida y hay muchos más valores ausentes, lo cual provoca que la imputación sea más importante que en casos anteriores. Después de esto, podemos aplicar los algoritmos de clustering como tal. Vamos a realizar las mismas pruebas que en los anteriores casos de estudio. Se han realizado las mismas pruebas que hasta ahora, de manera que a continuación se recoge una parte de los resultados, eligiendo para cada algoritmo el conjunto de parámetros que obtienen el mejor coeficiente de Silhouette.

	K-Means	Aglomerativo	Birch	DBSCAN	Mean Shift
Clusters	2	5	5	7	278
Silhouette	0.674595	0.658334	0.662875	0.801669	0.725310
Calinski Harabatz	2553.470095	399.226548	962.765521	3044.124026	293.826614
Tiempo	0.067089	0.091324	0.105645	2.011064	0.468679
K	2	5	5	-	-
Treshold (Birch)	-	-		-	-
Radio (DBSCAN)	-	-	-	0.3	-
Tamaño mínimo (DBSCAN)	-	-	-	50	-
Radio (Mean Shift)	-	-	-	-	0.244967

En este caso, no parece haber una progresión clara en cuanto al coeficiente de Silhouette se refiere. En casos como K-Means puede parecer hasta aleatoria, aunque también se puede decir que es capaz de hacer unas primeras segmentaciones muy precisas para un número muy reducido de grupos (como 2), que empieza a tener complicaciones cuando tiene que generar algunos grupos más, pero que posteriormente y con más clusters es capaz de generar mejores resultados. Por otra parte, la progresión en Calinski-Harabatz es mucho más clara: cuanto más clusters se forman, peor es el resultado de la métrica. A continuación se muestran cinco tablas, una para cada algoritmo con todas las combinaciones de hiperparámetros que se han probado.

K-Means			
K (clusters)	Silhouette	Calinski Harabatz	Tiempo
2	0.674595	2553.470095	0.067089
3	0.569537	2096.668313	0.051740
4	0.590235	1999.816844	0.057531
5	0.600786	1858.070176	0.071589
6	0.564654	1670.761954	0.063428
7	0.621534	1817.363317	0.067107
8	0.646048	1746.942356	0.059810

Clustering jerárquico (aglomerativo)			
K (clusters)	Silhouette	Calinski Harabatz	Tiempo
2	0.642804	536.254449	0.091255
3	0.645604	459.806816	0.090541
4	0.655714	410.848542	0.092923
5	0.658334	399.226548	0.091324
6	0.591212	400.079680	0.092089
7	0.601190	401.357276	0.089551
8	0.609707	405.659214	0.091244

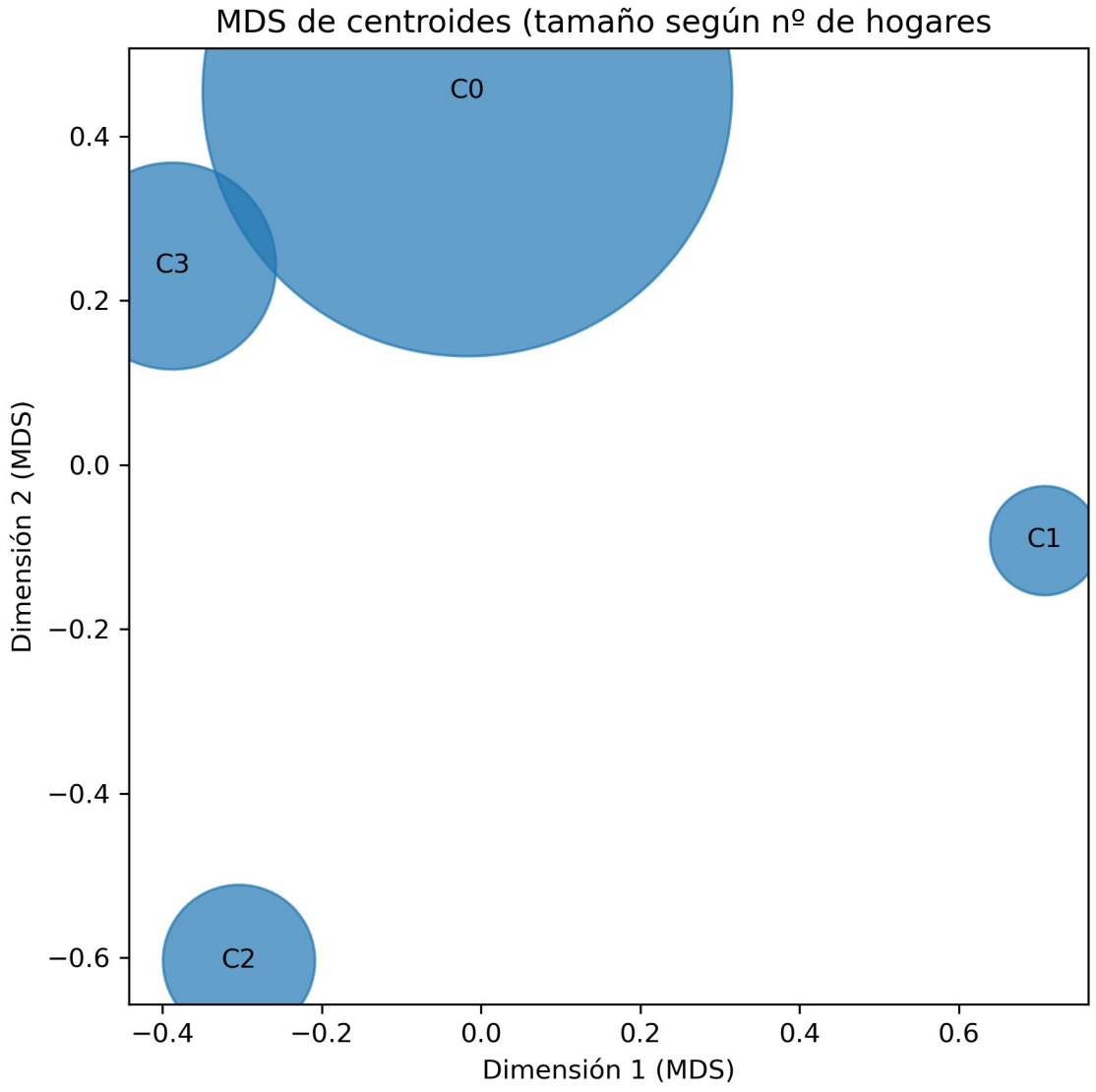
Birch				
K (clusters)	Treshold	Silhouette	Calinski Harabatz	Tiempo
2	0.5	0.654190	1263.915730	0.104502
3	0.5	0.641477	914.414660	0.110130
4	0.5	0.655123	1015.095079	0.106391
5	0.5	0.662875	962.765521	0.105645
6	0.5	0.659032	905.697808	0.105846
7	0.5	0.660079	825.014670	0.104784
8	0.5	0.634734	795.737309	0.104457

DBSCAN					
Clusters	Radio	Número de puntos	Silhouette	Calinski Harabatz	Tiempo
21	0.2	20	0.777887	4445.466172	1.793931
8	0.2	50	0.776693	7915.088051	1.821770
18	0.3	20	0.676407	1760.050627	2.108873
7	0.3	50	0.801669	3044.124026	2.011064
18	0.4	20	0.736654	1661.506439	2.024158
8	0.4	50	0.764657	2418.403362	2.007913
3	0.5	20	0.679955	748.148812	2.135935
3	0.5	50	0.697376	734.628920	2.356000

Mean Shift				
Clusters	Radio	Silhouette	Calinski Harabatz	Tiempo
278	0.244967	0.725310	293.826614	0.468679
278	0.246027	0.725310	293.826614	0.640968
278	0.246027	0.725310	293.826614	1.023183

Interpretación de la segmentación

Para estudiar la segmentación realizada en este caso de estudio, vamos a aplicar la misma segmentación que en el caso anterior. K-Means con K=4. Esta decisión tiene dos motivos principales: mantener la consistencia con los otros dos casos de estudio, para realizar un análisis más uniforme, y equilibrar los coeficientes de Silhouette (se manejaba mejor con muchos o muy pocos clusters) y Calinski-Harabatz (cuantos más clusters, peor). No se ha elegido, por ejemplo, K=2 porque, si bien es interesante obtener buenas métricas de rendimiento, también debe buscarse una segmentación real de los datos para obtener patrones interesantes. En cambio, si sólo se generaran dos grupos, se diferenciarían probablemente por una única variable y tenderían a generalizar u obviar muchos otros atributos. En la gráfica inferior se puede apreciar una segmentación fácil de identificar y que realmente divide a los datos.



Si analizamos los grupos formados por el heatmap de centroides, encontramos cuatro grupos muy diferenciados:

- Un primer grupo sin problemas a la hora de recibir asistencia médica y dental, y en general sin grandes problemas de bienestar. Este grupo representa a la mayoría de la población: tal y como se esperaría de un país de primer mundo con una amplia mayoría de la clase media, los niños pueden crecer en situaciones estables. Esto se puede entender de manera que aquellas familias económicamente solventes son precisamente las que se pueden permitir tener hijos y cuidar de menores, si bien evidentemente no todo el mundo lo hace con la misma suficiencia financiera.
- De hecho, un segundo grupo tiene valores bastante similares en casi todos los atributos salvo en el de la comida fresca, por el que se puede intuir que los menores comerán a menudo en comedores escolares o similares.
- El tercer grupo es el que presenta más dificultades económicas, careciendo en ocasiones de asistencia dental, privándose de reuniones con amigos, excursiones y vacaciones. Se puede interpretar como familias más vulnerables o en peligro de exclusión social, por ejemplo.
- Por último, un grupo de relativa bonanza pero que rehúye de tener vacaciones fuera de casa. Puede casar con grupos vistos en el anterior caso de estudio de familias que sobrellevaban el día a día pero no tenían capacidad de afrontar situaciones extraordinarias.