Universidad de Granada

Inteligencia de Negocio

Práctica 2: Segmentación para análisis empresarial.

Miguel Ángel Torres López (Grupo 1)

torresma1996@correo.ugr.es

December 2, 2018

Contents

List of Figures

1 Introducción

1.1 Problema abordado

En esta práctica vamos a estudiar la distribución de la población frente a ciertos parámetros. Como datos, se ha extraído una muestra de la población de la provincia de Granada de 83499 personas en el ao 2011. Los datos pueden descargarse de la página del Instituto Nacional de Estadística (http://www.ine.es/censos2011_datos/cen11_datos_microdatos.htm)

En el conjunto de datos se recogen 141 variables de carácter general sobre la vida social de los individuos. Para poder hacer contrastes de manera ms sencilla y hallar perfiles ms generales, vamos a trabajar con subconjuntos de estas variables. Para cada conjunto realizaremos un caso, en el que extraeremos una parte de la población con alguna cualidad interesante de estudio.

1.2 Valores perdidos

En los datos hay algunos campos que se desconocen para ciertos individuos. Bien porque no se pudieron conseguir o bien porque el individuo no quiso facilitarlos. Para poder tratar bien los datos con los algoritmos, vamos a sustituir en todas las filas el valor perdido por un 0. Esto va a provocar ruido en los resultados que obtengamos, reuniendo a todos estos individuos en la catergora 0, aunque en realidad puede que no tengan el mismo perfil.

Por este motivo, en todos los casos de estudio sucesivos tenemos que tener cuidado con las grandes aglomeraciones de individuos alrededor de la categoría 0, ya que podría un cluster formado por los individuos con valores perdidos, y no un perfil de individuo real.

1.3 Inicialización de los algoritmos y ejecución

Al estar trabajando en Python, para inicializar los algoritmos y poder ejecutar todo de forma secuencial y rápida, he realizado un clase que crea un vector con las funciones inicializadas y sus respectivos nombres.

```
def getAlgorithms(bandwidth,n_cluster,connectivity,eps):

 ms = cluster.MeanShift(bandwidth=bandwidth, bin_seeding=False)
 two_means = cluster.MiniBatchKMeans(n_clusters=n_cluster)
  ward = cluster.AgglomerativeClustering(n_clusters=n_cluster,
      linkage='ward',connectivity=connectivity)
  dbscan = cluster.DBSCAN(eps=eps)
  average_linkage = cluster.AgglomerativeClustering(linkage="average",
      affinity="cityblock",n_clusters=n_cluster, connectivity=connectivity)

  clustering_algorithms = (
    ('DBSCAN', dbscan),
    ('Ward', ward),
    ('AgglomerativeClustering', average_linkage),
    ('MiniBatchKMeans', two_means),
    ('MeanShift', ms)
  )
  return clustering_algorithms
```

Para ejecutar todos los algoritmos basta ahora recorrer dicho vector con un bucle y ejecutar cada elemento. Utilizaremos el nombre del algoritmo para distinguir las gráficas que debemos hacer con cada resultado.

Así por ejemplo, el algoritmo *MiniBatchKMeans* tiene, por su naturaleza, calculados los centros de los clusters, y podemos realizar un *Heatmap* de forma inmediata. Sin embargo, para otros algoritmos como *DBSCAN* será necesario calcular manualmente el punto central de cada cluster e, incluso, en otros no será posible hacer un *Heatmap* propiamente dicho.

2 Casos de estudio

2.1 Caso 1: Compañía de los hombres en una vivienda

2.1.1 Ejecución

El primer caso que proponemos tiene como objetivo estudiar con quienes comparten vivienda los hombres. Notar que, a priori, un hombre mayor o de mediana edad vive, en mayor probabilidad, con otra persona de su edad que será su pareja. En el segundo caso haremos lo mismo para las mujeres y comparemos sendos casos.

Vamos a extraer de la población del censo a los varones. Y vamos a seleccionar 4 atributos que consideramos relevantes de los originales 141. Estos son edad (EDAD), número de personas en casa (NPFAM), el número de personas entre 65 y 84 años con las que habita (H6584) y el número de personas mayores de 85 con las que habita (H85M).

Por motivos de computación y de nuestro conocimiento previo, vamos a seleccionar de entre todos los varones, solo aquellos que no estén casados. Esto nos restringe a la mitad de los varones, aproximadamente, y adems nos va a filtrar un cluster bastante amplio y con edades y ncleos familiares bastante heterogéneos.

Tras la ejecución de distintos algoritmos obtenemos la siguiente tabla de resultados:

Algoritmo	Calinksi-Harabaz	Silhouette	Clusters	Tiempo
KMeans	14391.2	0.514758	5	0.23189
MiniBatchKMeans	13891.5	0.387741	6	0.0745771
MeanShift	4044.87	0.548458	26	81.2256
DBSCAN	3957.13	0.501685	16	2.17222
Ward	13441.7	0.515223	5	119.665
AgglomerativeClustering	5012.57	0.536569	5	53.1029

Figure 1: Resultados de algoritmos en el caso 1.

Lo primero que observamos es que *MeanShift* y *DBSCAN* no tiene un buen resultado. Sendos algoritmos funcionan bien cuando hay varias aglomeraciones de puntos densas. Como no las hay, y no tienen limite de clusters, salen muchos de ellos.

Aunque no se muestra en la memoria por motivos de dimensionalidad, muchos de los clusters realizados por *MeanShift* y *DBSCAN* tienen menos de 20 individuos. Como nuestro objetivo es generalizar grupos y no entrar en casos demasiado concretos, podemos despreciar los resultados de estos

dos algoritmos.

Los mejores resultados los obtenemos al ejecutar los algoritmos *KMeans* y *Ward*. Por tanto, vamos a intentar afinar sus puntuaciones ejecutándolos con otros parámetros. En la figura 2 se muestra una tabla con las distintas pruebas para cada algoritmo.

Algoritmo	Calinksi-Harabaz	Silhouette	Clusters	Tiempo
KMeans (3 clusters, 0.00001 tolerancia)	14313.6	0.488144	3	0.102738
KMeans (5 clusters, 0.00001 tolerancia)	14391.2	0.514758	5	0.23189
KMeans (6 clusters, 0.00001 tolerancia)	13965.9	0.380025	6	0.185994
KMeans (7 clusters, 0.00001 tolerancia)	14144.5	0.403189	7	0.18931
Ward (3 clusters, 3 vecinos)	12866.4	0.496029	3	84.4439
Ward (5 clusters, 2 vecinos)	13441.7	0.515223	5	119.665
Ward (6 clusters, 5 vecinos)	12960.7	0.531288	6	48.9716
Ward (7 clusters, 2 vecinos)	12506.8	0.493931	7	120.782

Figure 2: Alteración de parámetros de algoritmos en el caso 1.

A la vista de los resultados, se intuye que la mejor división viene dada por el algoritmo *KMeans* para 5 clusters.

2.1.2 Interpretaciones gráficas

Para la mejor puntuación de cada algoritmo, se muestra en las figuras siguientes una *ScatterMatrix* con las distribuciones de unos atributos frente a otros con las respresentaciones de los individuos por colores según el cluster en el que se hayan clasificado.

Figure 3: ScatterMatrix para el algoritmo KMeans aplicado al caso 1.

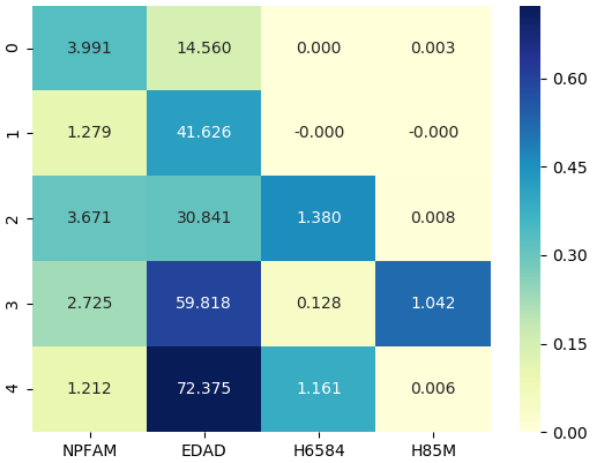

Figure 4: HeatMap para el algoritmo KMeans aplicado al caso 1.

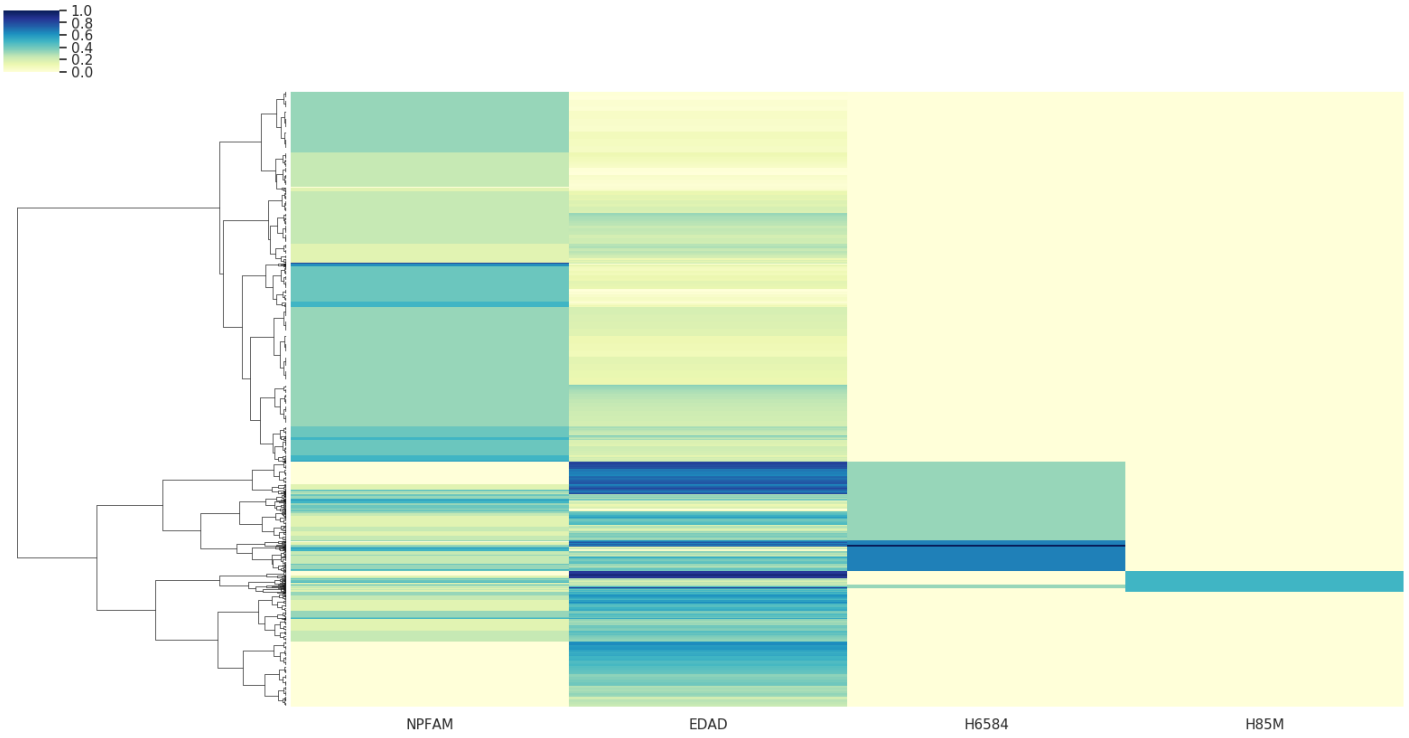

Figure 5: Dendrograma para el algoritmo Wards aplicado al caso 1.

En la figura 4, ayudados por los gráficos de la diagonal de la *ScatterMatrix*, podemos observar claramente como se distribuyen los clusters. Vamos a describir los grupos y vamos a intentar aventurar que tipo de individuos lo representan. La descripción se hace por orden decreciente de la cantidad de miembros de los mismos.

- En primer lugar hay que destacar el grupo 1. Lo componen hombres jóvenes, que conviven en casa con 4 personas y muy rara vez alguna de estas 4 personas es mayor. Este es el grupo de los niños que viven en familias.
- En el grupo 0, encontramos hombres no casados de unos 40 años, que viven solos o con otra personas menor de 65 años.
- Con 2304 individuos tenemos el grupo 2. Formado por hombres de alrededor de 31 años que conviven con otras 3 o 4 personas en su casa. En muchas ocasiones, esas personas están situadas en la franja de los 65-84 años de edad. Es el grupo de los solteros o divorciados que viven con gente mayor, probablemente sus padres.
- Vamos a descartar el interés sobre el grupo 4. Lo componen personas de 72 años de media que, en su mayoría, viven solos. Por tanto, este grupo es el grupo de los hombres mayores no casados, es decir, son viudos, solteros o divorciados.
- Al grupo 3 pertenecen individuos de unos 60. La mayoría de las veces conviven con alguien de más de 85 años y, probablemente, con una personas más situada en la franja de los 65-84. Es el grupo de las personas mayores que viven juntas sin ser pareja, es el grupo más minoritario.

2.1.3 Conclusión

En este estudio sacamos una relación con bastante seguridad. Hay muchos hombres solteros o divorciados de mediana edad que viven con sus padres. Para ilustrar esta idea, vamos a incidir en el número de individuos que hay en cada grupo anteriormente mencionados.

Longitud del cluster 0: 4260 (Solteros solitarios)
Longitud del cluster 1: 12091 (Niños)
Longitud del cluster 2: 2304 (Solteros con padres)
Longitud del cluster 3: 665 (Mayores con compañía)
Longitud del cluster 4: 1413 (Mayores solitarios)

Si quitamos del recuento a los niños, podemos sugerir que aproximadamente un 25% de los hombres solteros o divorciados de mediana edad viven con gente mayor, probablemente sus padres.

2.2 Caso 2: Compañía de las mujeres en una vivienda.

2.2.1 Ejecución

Al igual que en el caso de los hombres, una gran parte de las mujeres mayores y de mediana edad viven con su pareja. Esto nos hace restringir el estudio a mujeres no casadas, ahorrando tiempo computacional y limpiando la muestra de un grupo bastante heterogéneo.

Vamos a extraer de la población del censo a los muejeres no casadas. Y vamos a seleccionar los mismos 4 atributos que consideramos relevantes para los hombres. Estos son edad (EDAD), número de personas en casa (NPFAM), el número de personas entre 65 y 84 años con las que habita (H6584) y el número de personas mayores de 85 con las que habita (H85M).

Nuestro objetivo es comparar los distintos grupos de mujeres que obtengamos con los grupos que salieron en el caso anterior. A priori, es de esperar que también encontremos un cluster formado por niñas y otro formado por ancianas que tendrían unas características similares a los cluster homónimos de los varones.

Tras la realización de la ejecución se obtienen los siguientes resultados:

Algoritmo	Calinksi-Harabaz	Silhouette	Clusters	Tiempo
KMeans	21691.8	0.519276	5	0.177271
MiniBatchKMeans	20187.8	0.39013	6	0.150428
MeanShift	4296.29	0.420158	16	80.0409
DBSCAN	3062.41	0.426256	13	2.85126
Ward	19595.5	0.495119	5	151.12
AgglomerativeClustering	10361.1	0.507974	5	66.0226

Figure 6: Resultados de algoritmos en el caso 2.

Para los dos mejores algoritmos, vamos a probar varias configuraciones de sus parámetros. Aunque, como los dos mejores algoritmos vuelven a ser *KMeans* y *Ward*, por simetría hay bastantes posibilidades de que los parámetros óptimos vuelvan a ser los mismos que en el caso de los hombres.

No obstante, vamos a probar con distintos valores, los resultados pueden verse en la figura 7.

Algoritmo	Calinksi-Harabaz	Silhouette	Clusters	Tiempo
KMeans (3 clusters, 0.0001 tolerancia)	14313.6	0.488144	3	0.107802
KMeans (5 clusters, 0.0001 tolerancia)	21691.8	0.519276	5	0.177271
KMeans (6 clusters, 0.0001 tolerancia)	13966.6	0.385358	6	0.282528
KMeans (7 clusters, 0.0001 tolerancia)	14144.5	0.403189	7	0.20618
Ward (3 clusters, 2 vecinos)	12943	0.491885	3	121.675
Ward (5 clusters, 2 vecinos)	19595.5	0.495119	5	151.12
Ward (5 clusters, 5 vecinos)	13315.3	0.513287	5	49.3796
Ward (7 clusters, 2 vecinos)	12506.8	0.493931	7	123.149

Figure 7: Alteración de parámetros de algoritmos en el caso 2.

Efectivamente, los valores óptimos para los parámetros en estos algoritmos se alcanzan cuando realizamos una partición de 5 clusters, y el mejor algoritmo de los dos es el *KMeans*

2.2.2 Interpretación gráfica

Para ilustrar los resultados, mostraremos varias figuras. En la figura 10 se muestra el dendrograma realizado por el algoritmo Ward. En la figura 9 se muestran las relaciones dos a dos de los atributos de estudio y los individuos coloreados con el color de su cluster. Por último, en la figura 8 se muestra el *Heatmap* de las agrupaciones realizadas por la mejor ejecución de KMeans.

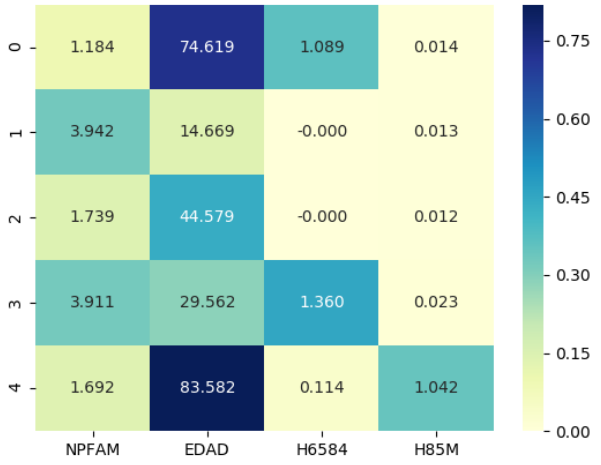

Figure 8: HeatMap para el algoritmo KMeans aplicado al caso 2.

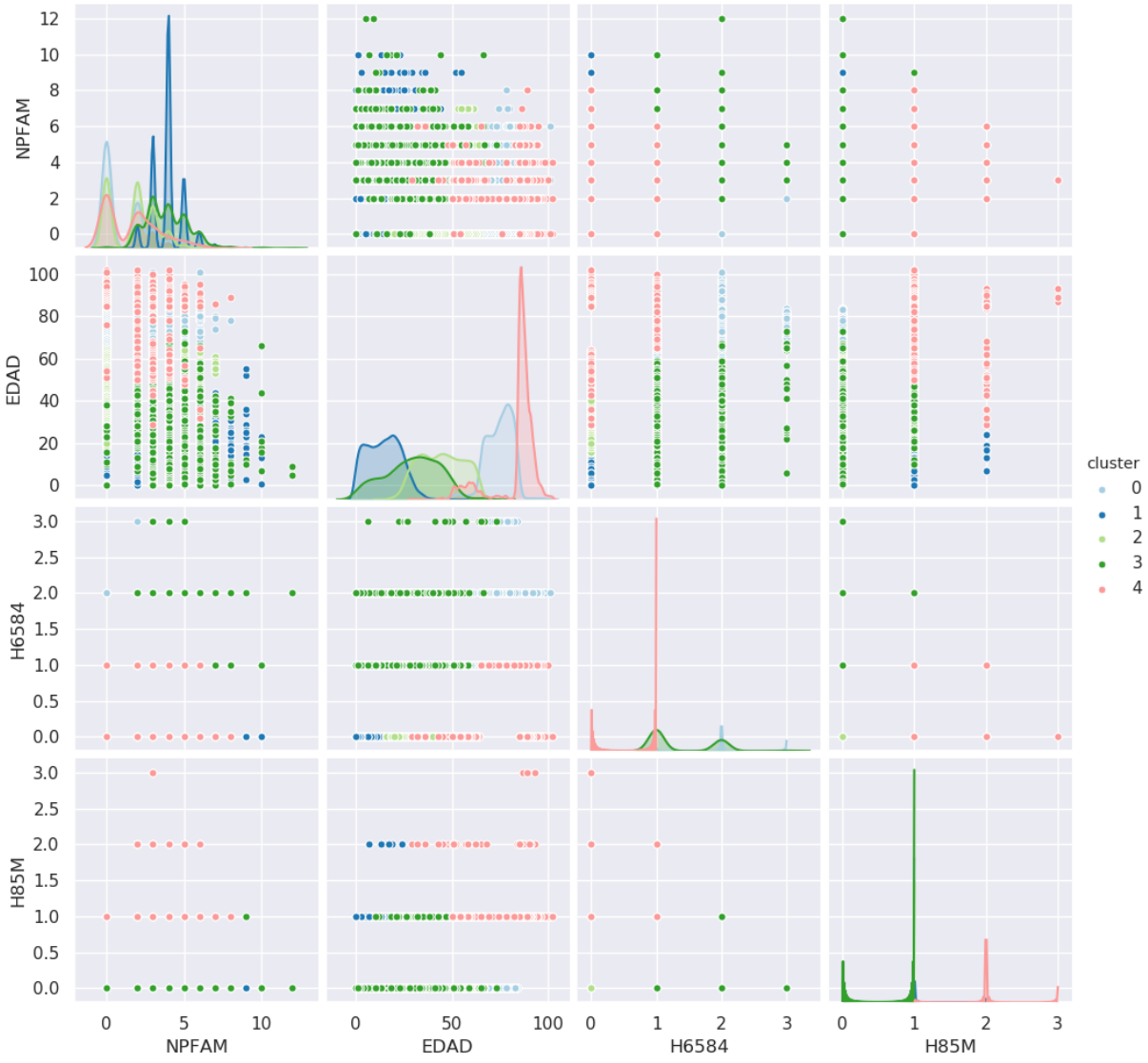

Figure 9: ScatterMatrix para el algoritmo KMeans aplicado al caso 2.

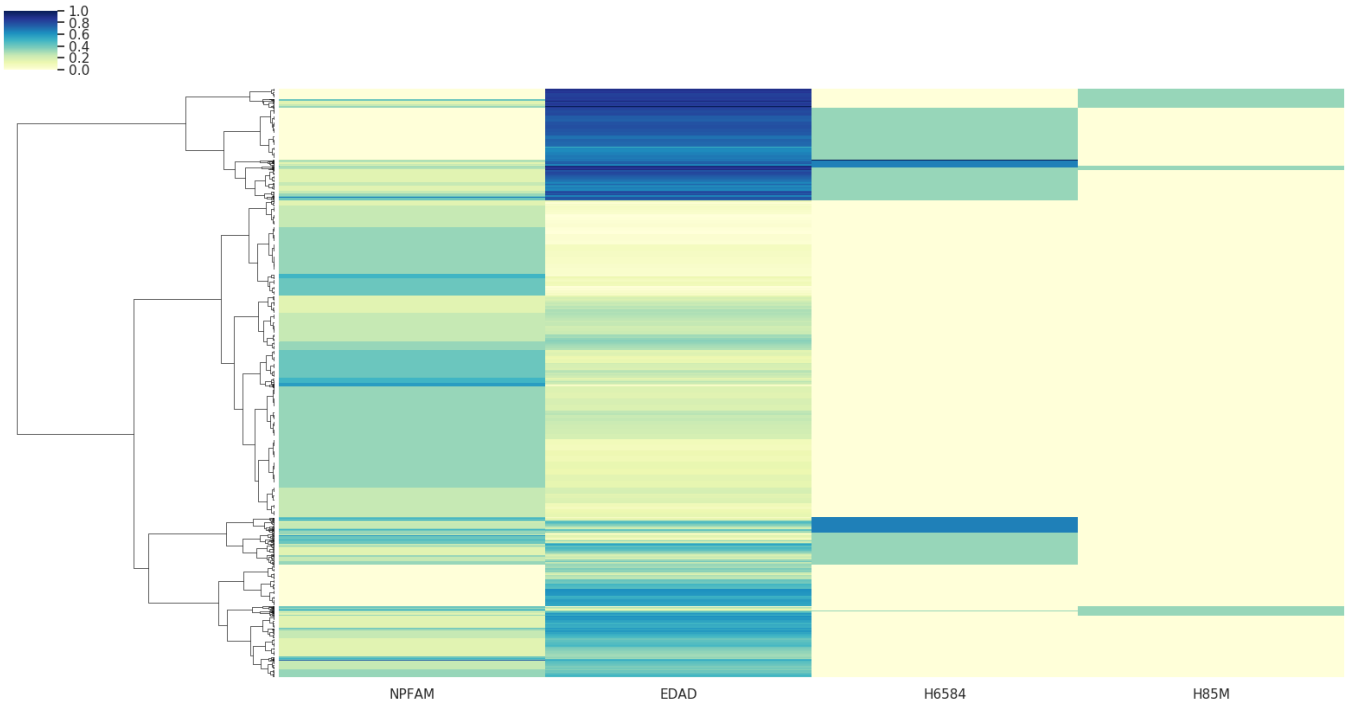

Figure 10: Dendrograma para el algoritmo Wards aplicado al caso 2.

En la figura 8, ayudados por los gráficos de la diagonal de la *ScatterMatrix*, podemos observar claramente como se distribuyen los clusters. Vamos a describir los grupos y vamos a intentar aventurar que tipo de individuos lo representan. Una vez más, la descripción se hace por orden decreciente de la cantidad de miembros de los mismos.

- El grupo 1 es, como se esperaba, el grupo de mujeres jóvenes, en casas de unas 4 personas con menos de 65 años. Es por tanto, el grupo de las niñas que viven con sus familias.
- En el grupo 2, encontramos mujeres no casadas de unos 44 años que comparte o no vivienda con otra persona de menos de 65 años. Es por tanto un grupo heterogéneo de mujeres que viven con alguien joven, previsiblemente su pareja (sin matrimonio), o que viven solas.
- Vamos a descartar el interés sobre el grupo 0. Lo componen personas de 75 años de media que, en su mayoría, viven solos. Por tanto, este grupo es el grupo de las mujeres no casadas, es decir, son viudas, solteras o divorciadas que viven solas.
- Al grupo 3 pertenecen individuos de unos 30 años. La mayoría de las veces conviven con alguien de entre 65 y 84, en ocasiones hasta con

dos personas de esta franja. Estamos hablando del grupo de mujeres no casadas que viven con gente mayor, probablemente sus padres.

- El grupo más minoritario es el grupo 4. Corresponde a mujeres con una media de 84 años, que viven con alguien más de edad superior a 85 y ocasionalmente con alguien de 65 a 84. Es por tanto, el grupo de mujeres que viven con alguien de su misma edad en casa que no es su pareja.

2.2.3 Conclusión

Obtenemos una división parecida a la que apareció en el caso 1. Aunque con algunas diferencias. La primera, es la cantidad de miembros de cada cluster:

Longitud del cluster 0: 3567 (Mayores solitarias)
Longitud del cluster 1: 11917 (Niñas)
Longitud del cluster 2: 4537 (Solteras con compañía)
Longitud del cluster 3: 1821 (Solteras con padres)
Longitud del cluster 4: 977 (Mayores con compañí)

Otra diferencia bastante notable es que las mujeres no casadas que no viven con sus padres, las del grupo 2, viven con mayor probabilidad con alguien que en el caso de los hombres no casados que no viven con sus padres. Podemos observar que en hombres este grupo tiene de media un 1.279, mientras que en las mujeres alcanza un 1.739 de media.

Otra diferencia algo más sutil es la edad media de las muejeres que viven con sus padres, una año inferior que la de los hombres. También vemos una variación de 0.3 en la cantidad de personas que viven en la casa. No sabemos si es un hecho casual o tiene verdadera significancia. Pero podemos aventurar que en familias con una hija, bien esta suele ser la menor entre los hijos o bien existe una relación que hace que los descendientes abandonen un poco antes la casa de sus padres.

Se puede hacer una última observación que no era objetivo inicial de nuestro estudio pero que sale a la vista. Las mujeres tienen en los grupos de mayor edad una media más alta de este parámetro. Puede deberse a dos razones.

La primera sería que hay mayor número de hombres de 60-70 años que mujeres de la misma franja, produciendo que baje la media. Esta es poco probable, pues el sexo de nacimiento es un 50-50. La segunda sería que las mujeres son más longevas que los hombres.

Pero, como ya se ha observado en muchos estudios, la hipótesis correcta es la segunda.

2.3 Caso 3: Vivienda de los extranjeros

2.3.1 Ejecución

En primera instancia, el objetivo de está sección era hacer un estudio similar a los dos anteriores pero restringiendo a los individuos que no tuviesen la nacionalidad española. A priori, esperaba encontrarme unas agrupaciones similares a la de los casos anteriores. Aunque en los clusters de solteros suponía que no vivirían con sus padres, pues serían jóvenes desplazados por motivos laborales y probablemente no vendrían con sus padres.

Consecuente a esto, también esperaba encontrar familias inmigrantes afincadas en España. Y esto justifica la elección de dos atributos de estudio adicionales a los ya mencionados en apartados anteriores.

Vamos a seleccionar los mismos 4 atributos que consideramos relevantes para los casos anteriores. Estos son edad (EDAD), número de personas en casa (NPFAM), el número de personas entre 65 y 84 años con las que habita (H6584) y el número de personas mayores de 85 con las que habita (H85M). Pero vamos a considerar, en adición a estos, el número de personas menores de 5 años en la familia y el número de personas entre 5 y 15.

Tras realizar algunas pruebas para decidir que parámetros eran los adecuados y qué atributos tenían mas interés, me di cuenta de que el número de personas mayores extranjeras es bastante reducido y, además, suelen vivir en compañía de otra persona de avanzada edad. Esto puede observarse en la figura 11, en concreto, se trata del *Heatmap* realizado por una de las mejores ejecuciones del algoritmo *KMeans*, probado con múltiples parámetros.

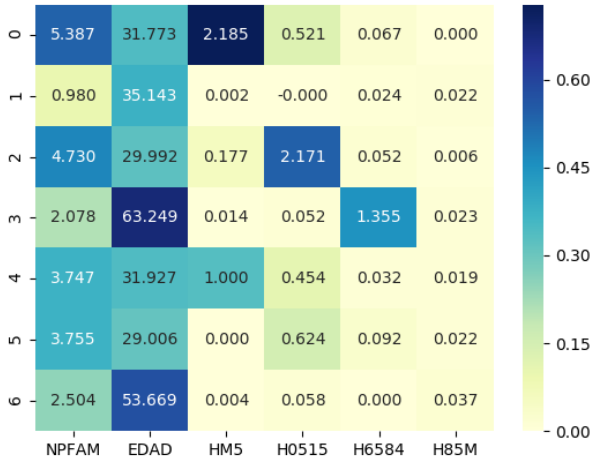

Figure 11: Situación de los extranjeros de avanzada edad.

En la figura se observa como no hay ningún grupo cuya edad supere los 65 de media. Y tampoco hay, en la mayoría de los casos, nadie que viva con personas mayores de 85. Por este motivo, he preferido suprimir este atributo del estudio y añadir un atributo que aporte cierta información del tipo de vivienda en el que viven los individuos, la superficie útil de la misma (SUP).

Con estos atributos, hemos realizado algunas ejecuciones con parámetros distintos para los algoritmos. Se muestra la tabla de los mejores resultados encontrados.

Algoritmo	Calinksi-Harabaz	Silhouette	Clusters	Tiempo
KMeans	1583.92	0.314047	5	0.0610268
MiniBatchKMeans	1469.56	0.279695	5	0.0173154
MeanShift	345.519	0.245393	6	12.5434
DBSCAN	397.482	0.321096	24	0.105961
Ward	1234.7	0.254704	6	6.54167
AgglomerativeClustering	635.254	0.285523	5	5.43273

Figure 12: Resultados de algoritmos en el caso 3.

Para los dos mejores algoritmos, *KMeans* y *MiniBatchKMeans*, vamos a probar una conjunto de parámetros más extenso. La tabla resumen puede verse en la figura 13. Cabe destacara, que la diferencia entre estos dos algoritmos es muy poca. Como puede verse en `https://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html`, las agrupaciones son prácticamente iguales, a excepción de un poco de eficiencia que sacrifica *KMeans* para dar un resultado un poco mejor.

Algoritmo	Calinksi-Harabaz	Silhouette	Clusters	Tiempo
KMeans (3 clusters, 0.0001 tolerancia)	1471.63	0.302005	3	0.0543618
KMeans (5 clusters, 0.0001 tolerancia	1583.92	0.314047	5	0.0674241
KMeans (7 clusters, 0.0001 tolerancia)	1345.73	0.298894	7	0.091398
KMeans (9 clusters, 0.0001 tolerancia)	1286.46	0.314416	9	0.108686
MiniBatchKMeans (3 clusters)	1337.13	0.299712	3	0.0324848
MiniBatchKMeans (5 clusters)	1530.9	0.306902	5	0.0163782
MiniBatchKMeans (7 clusters)	1343.72	0.286639	7	0.0214183
MiniBatchKMeans (9 clusters)	1256.44	0.31393	9	0.0259318

Figure 13: Alteración de parámetros de algoritmos en el caso 3.

2.3.2 Interpretación gráfica

Para el mejor algoritmo con los mejores parámetros, de nuevo *KMeans*, mostramos la matriz *ScatterMatrix* y su correspondiente *Heatmap*.

Figure 14: ScatterMatrix para el algoritmo KMeans aplicado al caso 3.

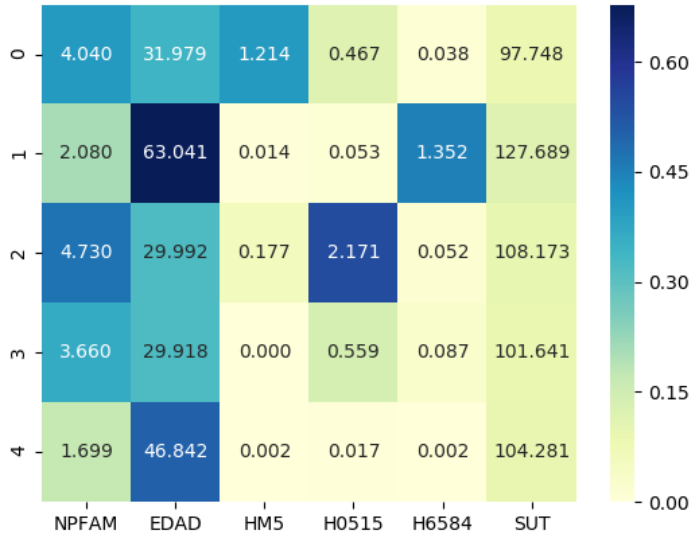

Figure 15: HeatMap para el algoritmo KMeans aplicado al caso 3.

Una vez más, vamos a proceder a desglosar los clusters, de mayor a menor número de individuos, y comentar alguna de sus propiedades características.

- El grupo 4 es el más numeroso, formado por personas de 47 años de media que viven solas o acompañadas por otra persona de entre 15 y 65 años. Es un grupo, probablemente, de parejas sin hijos y solteros. En cuanto a las dimensiones de la vivienda, nos encontramos en la media del resto de grupos.

- En el grupo 3 encontramos gente de unos 30 años de edad, que vive en casas con 3 o 4 personas. A priori se podría pensar que son familias, pero en la mayoría de los casos no viven niños en la vivienda. Podemos deducir que se trata de gente joven que comparte piso y emigró por trabajo, ya que no tienen niños.

  Hay que tomar esta deducción con cierta precaución. Dado que la edad está muy centrada y la media de personas entre 5 y 15 en casa, podría tratarse de un grupo escoba. Es decir, en este grupo se encontrarían individuos que no encajan en el resto de grupos. Esto explicaría la baja puntuación obtenida en las métricas.

- El grupo 0 lo componen personas de 32 años de media que viven con alrededor de 4 personas. Si nos fijamos en que el valor de HM3 es 1.214 que, de forma bastante acertada, nos dice que la familia tiene un hijo nacido en España. Por tanto, este grupo es el grupo de las familias

inmigrantes con hijos nacidos en territorio español. Hay que destacar que este grupo tiene los hogares con menos superficie útil de todos, por lo que podemos presuponer que son familias pobres que emigraron para buscar una mejor situación para sus hijos.

- Al grupo 2 pertenecen individuos de unos 30 años. La mayoría de las veces conviven con dos personas de entre 5 y 15, en ocasiones hasta con tres menores de 15 años. Estamos hablando del cluster que agrupa a las familias con hijos que, probablemente, no nacieron en España, ya que en ese caso, sus padres habrían tenido que venir demasiado jóvenes.

- El grupo más minoritario es el grupo 1. Corresponde a las personas mayores, situadas alrededor de los 63. No viven con niños y en la mayoría de las ocasiones viven con una persona de su misma edad. Tienen los hogares más grandes. Se trata, con bastante probabilidad, del grupo formado por parejas acomodadas y jubiladas o cercanas a la jubilación.

2.3.3 Conclusión

Obtenemos una división algo distinta a los casos 1 y 2, algo que parece obvio porque el caso se compone de gente en situaciones diferentes. Mostramos los clusters en relación a su longitud y su supuesta índole:

Longitud del cluster 0: 658 (Familias con hijos españoles)
Longitud del cluster 1: 489 (Mayores acomodados)
Longitud del cluster 2: 496 (Familias con hijos no españoles)
Longitud del cluster 3: 957 (Inmigrantes que comparten vivienda)
Longitud del cluster 4: 1140 (Inmigrantes solitarios)

Como ya hemos mencionado antes, el cluster asociado a los jóvenes que viven con sus padres ha desaparecido. Tampoco se ve que la gente joven habite con gente mayor, en su lugar, los mayores son un grupo minoritario acomodado que viven en parejas. Estas son las dos diferencias más notorias en este sesgo de la población inmigrante. No obstante, existen otras posibles características algo menos seguras.

Si hacemos el contraste de las dimensiones de las viviendas de los grupos 0 y 2, podemos intuir que las familias inmigrantes con hijos de mayor edad tiene más dinero. Aunque no es una afirmación nada segura, ya que podría deberse al lugar donde viven. Normalmente en las zonas urbanas las vivendas son de menor dimensión que en las zonas rurales. Aun siendo poco probable, la gente con hijos mayores puede que esté más orientada a vivir en nucleos rurales.

3 Bibliografía

- Documentació de Scikit-learn (https://scikit-learn.org/stable/)
- Comparación MiniBatchKMeans y KMeans (https://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html)
- Documentación de Pandas (http://pandas.pydata.org/pandas-docs/stable/)
- Documentación de Seaborn (https://seaborn.pydata.org/tutorial.html)