PRÁCTICA 2:
Segmentación para Análisis Empresarial

Inteligencia de negocio

Realizado por:
Francisco Solano López Rodríguez
DNI: 20100444P
Email: fransol0728@correo.ugr.es

DOBLE GRADO EN INGENIERÍA INFORMÁTICA Y MATEMÁTICAS.
QUINTO CURSO

Índice

1. Introducción

El objetivo de esta práctica consiste en e estudio de técnicas de aprendizaje no supervisado para análisis empresarial.

El conjunto de datos del que disponemos, se corresponde con los datos publicados en el último censo de población realizado por el Instituto Nacional de Estadística en 2011. Trabajaremos con los datos relativos a la provincia de granada, con un total de 83.499 casos. El número de variables de los que disponemos es de 142. Algunas de las variables son la edad, el sexo, nacionalidad, estudios, etc.

Las variables de las que disponemos se estructuran de la siguiente forma:

- **Identificación**: donde podemos encontrar el código de la provincia (en nuestro caso Granada), y el código del municipio.
- **Datos individuales**: en este apartado tenemos datos de individuo, tales como la edad, el sexo, el nivel de estudio, estado civil, etc.
- **Datos de vivienda**: en esta sección se disponen de datos relacionados con la vivienda, como por ejemplo si posee calefacción, si tiene acceso a Internet, número de habitaciones.
- **Datos del edificio**: como por ejemplo el año de construcción, si tiene garaje, si tiene gas o si tiene ascensor.
- **Datos de parentesco**: número de familia dentro del hogar, número de núcleo dentro del hogar, etc.
- **Datos del padre**: como por ejemplo nacionalidad, estado civil, nivel de estudios, situación profesional, etc.
- **Datos de la madre**: como por ejemplo nacionalidad, estado civil, nivel de estudios, situación profesional, etc.
- **Datos del cónyuge o pareja**: como por ejemplo nacionalidad, estado civil, nivel de estudios, situación profesional, etc.
- **Datos del núcleo**: como por ejemplo tamaño del núcleo, número de hijos, tipo de pareja, etc.

Comentar que dentro del conjunto de datos del que disponemos hay muchos valores perdidos y que se encuentran en blanco, debido a esto aquellos valores en blanco han sido sustituidos por el valor 0. Esto puede alterar algo los resultados haciendo que baje la media de los valores numéricos debido a este valor 0.

Los algoritmos utilizados han sido KMeans, AgglomerativeClustering, MeanShift y MiniBatchKMeans.

Las ejecuciones han sido realizadas en un ordenador con procesador Intel Core i3, 8 GB de RAM y sistema operativo Ubuntu 16.04 LTS. Supongo que debido a mi procesador el tiempo obtenido en las ejecuciones ha sido más alto del que cabría esperar.

2. Casos de estudio considerados

2.1. Caso de estudio 1

2.1.1. Descripción del caso de estudio

En el primer caso que vamos a estudiar se centra sobre el conjunto de mujeres con edad de entre 20 y 50 años. Las características que vamos a elegir se corresponden con la edad (EDAD), el número de personas en la familia (NPFAM), el número de personas de 0 a 4 años en el hogar (HM5) y el número de personas de 5 a 15 años (H0515). El objetivo de este estudio es distinguir grupos de mujeres según el número de persona jóvenes con las que convivan, que por la edad elegida en este grupo de estudio con bastante probabilidad serán sus hijos, y en ciertos casos en los que la edad sea cercana a 20 es probable que sean hermanos. Después trataremos de sacar conclusiones sobre el perfil de mujer que parece pertenecer a cada grupo.

Este caso de estudio posee un total de 17996 ejemplos. El conjunto seleccionado ha sido obtenido mediante el siguiente código. Podemos ver que nos hemos quedado con las personas de entre 20 y 50 años de edad, y que se ha utilizado la variable categórica SEXO para fijar el estudio solo sobre las mujeres. Para el clustering como podemos ver solo hemos considerados variables numéricas.

```
mujer = 6
subset = censo.loc[(censo['EDAD']>=20) & (censo['EDAD']<=50) & (
```

```
    censo['SEXO']==mujer)]
usadas = ['EDAD', 'NPFAM', 'HM5', 'H0515']
X = subset[usadas]
```

2.1.2. Resultados de los algoritmos

En este apartado mostraremos los resultados de los algoritmos en una tabla comparativa, donde se incluirá el número de clusters, los resultados obtenidos por las dos métricas utilizadas (Calinski-Harabaz y Silhouette) y los tiempos de ejecución de los algoritmos.

Antes de mostrar la tabla comparativa, veamos que representan los valores de las métricas Calinski-Harabaz y Silhouette, para poder interpretar mejor los resultados.

Calinski-Harabaz: se define como la relación entre la dispersión dentro los clusters y la dispersión entre los clusters.

Silhouette: es una medida de como de similar es un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación), su valor varía de -1 a +1.

Veamos el código con la definición de los algoritmos utilizados.

```
k_means = KMeans(init='k-means++', n_clusters=5, n_init=5,
    random_state=random_seed)
agglo=AgglomerativeClustering(n_clusters=5,linkage="ward")
meanshift = MeanShift(bin_seeding=True)
miniBatchKMeans = MiniBatchKMeans(init='k-means++',n_clusters=4,
     n_init=5, max_no_improvement=10, verbose=0, random_state=
    random_seed)
dbscan = DBSCAN(eps=0.2)

algorithms = [(k_means, "KMeans"),
              (agglo, "AC"),
              (meanshift, "MeanShift"),
              (miniBatchKMeans, "MiniBatchKM"),
              (dbscan, "DBSCAN")]
```

Comentar que para la ejecución de los algoritmos, los datos utilizados han sido normalizados al intervalo [0,1] mediante la siguiente función:

```
def norm_to_zero_one(df):
    return (df - df.min()) * 1.0 / (df.max() - df.min())
```

A continuación mostramos la tabla comparativa de los 5 algoritmos ejecutados.

Algoritmo	**N.Clusters**	**Calinksi-Harabaz**	**Silhouette**	**Tiempo**
KMeans	5	13813.093980	0.385283	0.169730
AC	5	12821.842945	0.366786	18.260128
MeanShift	4	9369.510140	0.426968	34.642497
MiniBatchKM	4	13816.619513	0.405181	0.097726
DBSCAN	6	858.667562	0.165711	2.997457

Como podemos observar en la tabla el algoritmo DBSCAN es el que peores resultados ha obtenido y su número de clusters es de 6. Comentar que para este algoritmo he tenido que probar diferentes del parámetro eps el cual determina la distancia máxima entre dos muestras para que se consideren en el mismo vecindario, porque o bien salía demasiados clusters, donde la mayoría estaban casi vacíos o bien salían muy pocos clusters. El tamaño de cada cluster obtenido con el algoritmo DBSCAN es el siguiente:

```
0: 14625 (81.27%)
1:  2780 (15.45%)
2:   524 ( 2.91%)
-1:    26 ( 0.14%)
3:    23 ( 0.13%)
4:    18 ( 0.10%)
```

Como podemos ver el tamaño de los clusters está muy desbalanceado, teniendo casi todos los elementos de la muestra en el cluster 0, mientras que otros cluster como el 3 y el 4 están casí vacíos.

Los algoritmos restantes tienen puntuaciones similares, y algunos son mejores con la métrica CH mientras que otros son mejores con la métrica Silhoutte. Meansift por ejemplo es el que tiene peores resultados (sin tener en cuenta DBSCAN), si comparamos observado la métrica CH, mientras que vemos que es el mejor de ellos respecto a la métrica Silhoutte, también se tiene que es el que ha tomado mayor tiempo de ejecución.

El algoritmo MiniBatchKMeans es una variante de KMeans que utiliza minibatches para reducir el tiempo de cálculo. Efectivamente si comparamos los tiempos de KMeans y MiniBatchKMeans podemos comprobar que MiniBatchKM ha obtenido un tiempo de ejecución menor y con unos resultados

similares.

Como dos de los mejores algoritmos han sido KMeans y AgglomerativeClustering vamos ha variar el parámetro que fija el número de clusters para ver el efecto de esté parámetro en los valores de las métricas.

Veamos primero como afecta la variación del número de clusters en el algoritmo KMeans.

Algoritmo	**N.Clusters**	**Calinksi-Harabaz**	**Silhouette**	**Tiempo**
KMeans_5	5	13823.941361	0.382080	0.223551
KMeans_6	6	12619.974453	0.360733	0.256366
KMeans_7	7	12008.363812	0.372945	0.305757
KMeans_8	8	11454.909998	0.345979	0.340992

Vemos que los mejores resultados obtenidos han sido con el número de clusters igual a 5.

Veamos ahora los resultados obtenidos con el algoritmo AgglomerativeClustering.

Algoritmo	**N.Clusters**	**Calinksi-Harabaz**	**Silhouette**	**Tiempo**
AC_5	5	12821.842945	0.366786	19.806992
AC_6	6	11804.041236	0.373603	20.046720
AC_7	7	11106.712349	0.387814	20.121320
AC_8	8	10751.076885	0.342396	21.564551

En este caso vemos que el mejor valor con la métrica CH se obtiene con 5 clusters, mientras que con la métrica Silhouette el mejor resultado se obtiene con 7 clusters.

2.1.3. Interpretación de la segmentación

Para interpretar los resultados se van a proporcionar una tabla con las medias de cada característica dentro de cada cluster, un scatter matrix y un heatmap con los datos normalizados.

Tanto en este caso de estudio como en los posteriores, solo se mostraran los resultados respecto a scatter matrix, heatmap y tabla con las medias, del algoritmo KMeans con 5 clusters, ya que ha sido el que ha tenido mejores resultados en general (teniendo en cuenta los otros dos casos de estudio), y el dendrograma será el resultante de el algoritmo AgglomerativeClustering.

Tabla con las medias de cada característica:

CLUSTER	**EDAD**	**NPFAM**	**HM5**	**H0515**
0	33.982385	3.789973	1.199864	0.467141
1	46.242987	3.211259	0.020507	0.324241
2	33.355525	2.086479	0.000000	0.107412
3	23.625407	3.659985	0.038587	0.185167
4	40.200000	4.202357	0.045455	1.825589

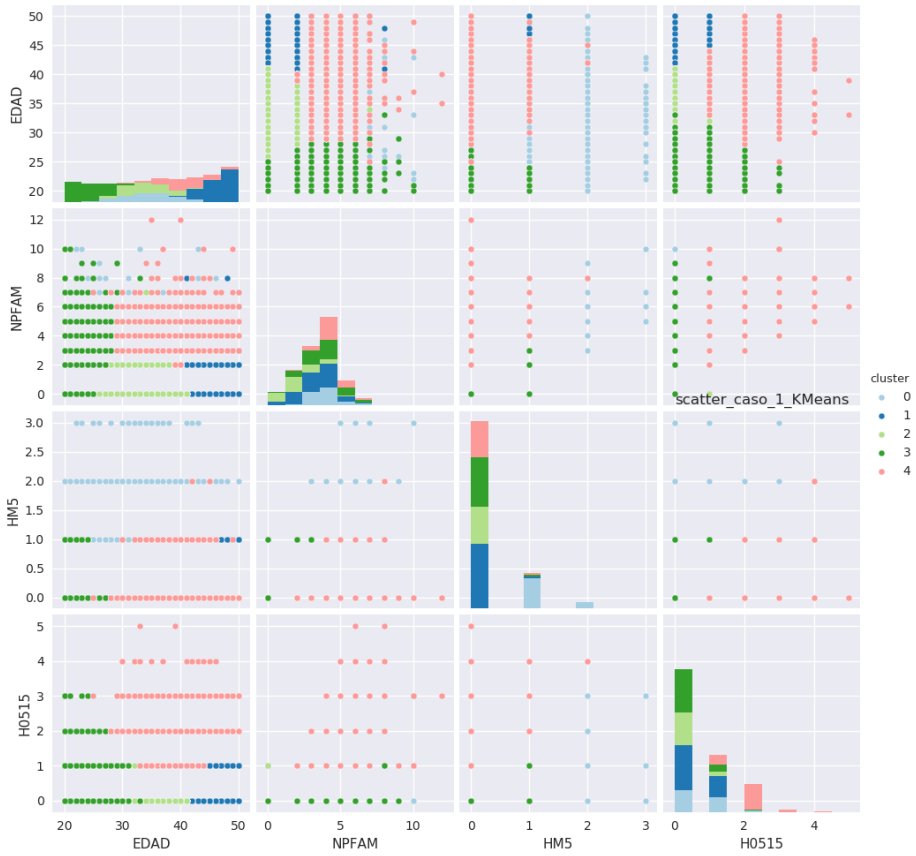

Figura 1: SccaterMatrix KMeans Caso 1

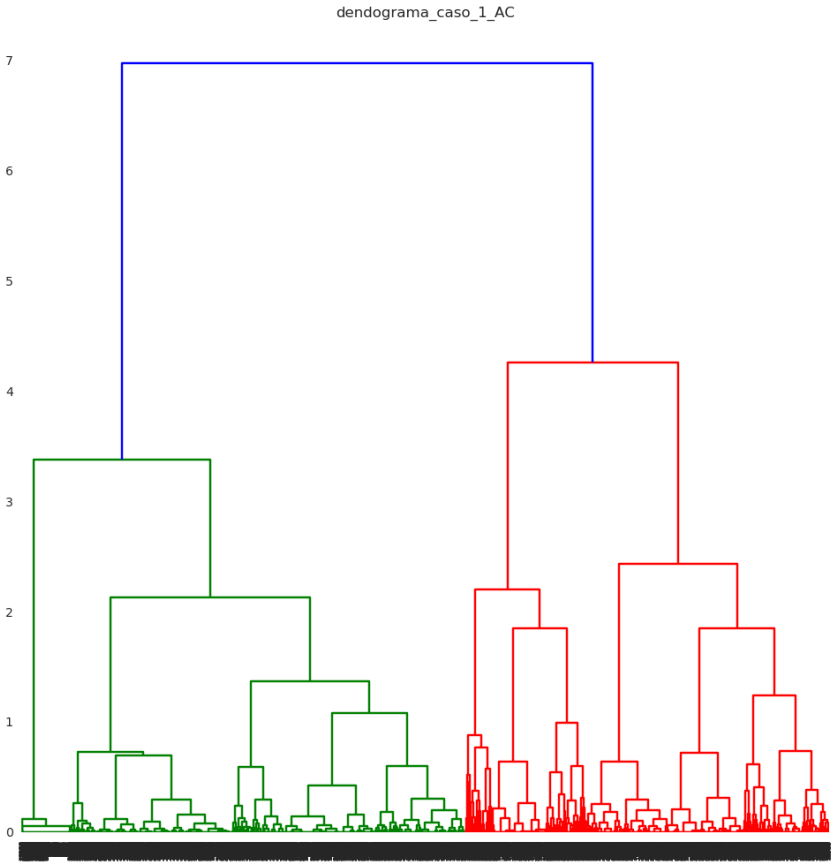

Figura 2: Dendogram AC Caso 1

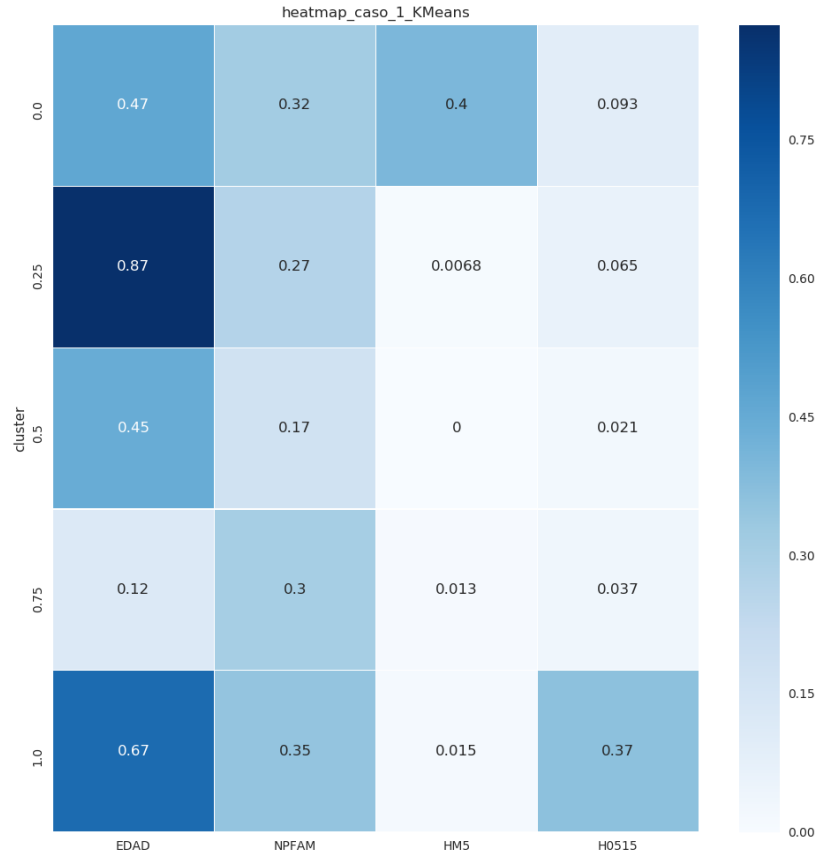

Figura 3: Heatmap KMeans Caso 1

Con la ayuda de estas visualizaciones y la tabla de medias vamos a intentar hace una descripción de cada grupo. Las conclusiones que se muestran a continuación están basadas principalmente en la observación del heatmap y la matriz de medias, al igual que las conclusiones tomadas en los posteriores casos de estudio.

En primer lugar analicemos el cluster 0. Este cluster se corresponde con mujeres de una media de edad de 33 años, donde el número de personas en la familia es de entre 3 y 4 personas de las cuales aproximadamente 1 de las personas es menor de 5 años y entre 0 y 1 persona tiene entre 5 y 15 años. Este grupo puede interpretarse como el grupo de mujeres jóvenes con pareja y que acaba de tener niños y en algunos casos estos niños ya tienen más de 5 años.

El cluster 1 está compuesto por mujeres con una media de 46 años con un número de personas en la familia de 3, en el que no hay niños menores de 5 años y es algo probable que vivan con una persona de entre 5 y 15 años. El número de personas en la casa es de 3 así que probablemente la segunda persona sea su pareja y la tercera su hijo que en algunos casos tenga entre 5 y 15 años y en otros ya sea mayor de 15 años (este último dato no lo tenemos pero se puede intuir para que cuadren las cuentas con el valor de NPFAM). También podría ser probable que en este grupo haya mujeres divorciadas que viven con sus dos hijos o tal vez cuidando de sus 2 padres.

El cluster 2 esta formado por mujeres de una media de 33 años, y conviven en el hogar 2 personas, además no hay niños menores en el hogar. Es probable que este grupo este formado en su mayoría por mujeres jóvenes con pareja, que aún no han tenido hijos.

El cluster 3 tiene la menor media edad siendo esta de 23, en la familia hay entre 3 y 4 personas, no hay niños menores de 5 y en pocas ocasiones hay un menor de entre 5 y 15 años. Este grupo parece corresponderse con el grupo de mujeres jóvenes que aún no se ha independizado y vivan con sus padres y tal vez un hermano menor de entre 5 y 15 años.

El cluster 4 esta compuesto por mujeres con 40 años de media, con 4 personas en la familia, y que conviven aproximadamente 2 menores de entre 5 y 15 años. Este grupo seguramente se corresponda con el grupo de madres que viven con su pareja y además con sus dos hijos de entre 5 y 15 años.

En el heatmap podemos apreciar algunos detalles sobre los agrupamientos como por ejemplo que en los grupos de mayor edad no suele haber menores de 5 años en la familia.

A continuación se muestra la proporción de cada cluster:

```
1:  5169 (28.72%)
3:  3991 (22.18%)
4:  2970 (16.50%)
0:  2952 (16.40%)
2:  2914 (16.19%)
```

2.2. Caso de estudio 2

2.2.1. Descripción del caso de estudio

En este caso vamos a realizar un estudio similar al primero, solo que en este caso el conjunto tomado será de hombres en lugar de mujeres, estos tendrán una edad de entre 20 y 50 años. De nuevo las características seleccionadas se corresponden con la edad (EDAD), el número de personas en la familia (NPFAM), el número de personas de 0 a 4 años en el hogar (HM5) y el número de personas de 5 a 15 años (H0515). El objetivo de este estudio es distinguir grupos de hombres según la las personas con las que convivan y sacaremos conclusiones sobre que perfil de hombre parece que pertenece a cada grupo.

El número de elementos de este conjunto de la población seleccionado cuenta con 17338 ejemplos.

```
hombre = 1
subset_2 = censo.loc[(censo['EDAD']>=20) & (censo['EDAD']<=50) &
    (censo['SEXO']==hombre)]
usadas_2 = ['EDAD', 'NPFAM', 'HM5', 'H0515']
X_2 = subset_2[usadas_2]
```

2.2.2. Resultados de los algoritmos

Al igual que hicimos en el caso 1, en este apartado mostraremos una tabla comparativa con el número de clusters, las puntuaciones obtenidas con las dos métricas utilizadas y los tiempos de ejecución.

Los algoritmos utilizados en este caso han sido los mismos que en el caso anterior.

Algoritmo	N.Clusters	Calinksi-Harabaz	Silhouette	Tiempo
KMeans	5	13224.665072	0.374337	0.168696
AC	5	12108.685870	0.362907	15.367518
MeanShift	4	8904.535099	0.430939	30.512459
MiniBatchKM	4	12955.052365	0.411698	0.173368
DBSCAN	6	745.965170	0.148728	2.977157

Los resultados obtenidos son similares al caso 1, de nuevo el que ha obtenido peores resultados es el algoritmos DBSCAN, y de nuevo MeanShift vuelve

a ser el mejor con respecto a la métrica Silhouette y el segundo peor con respecto a la métrica CH. Los dos mejores han sido KMeans y MiniBatchK-Means, así que vamos a probar a variar el parámetro que fija el número de clusters sobres estos dos algoritmos.

A continuación se muestran las tablas comparativas de ambos algoritmos.

Algoritmo	**N.Clusters**	**Calinksi-Harabaz**	**Silhouette**	**Tiempo**
KMeans_5	5	13222.474029	0.374148	0.185450
KMeans_6	6	11930.142932	0.375756	0.180921
KMeans_7	7	11225.951876	0.322783	0.288329
KMeans_8	8	11012.661337	0.339428	0.361198

Algoritmo	**N.Clusters**	**Calinksi-Harabaz**	**Silhouette**	**Tiempo**
MiniBatchKM_5	5	13129.086859	0.371412	0.088436
MiniBatchKM_6	6	11758.255179	0.367008	0.094699
MiniBatchKM_7	7	11195.628563	0.335848	0.149005
MiniBatchKM_8	8	10705.980919	0.350849	0.116380

Se puede observar que en ambos empeoran las puntuaciones obtenidas por las métricas al aumentar el número de clusters, siendo el número de clusters con mejores resultados para ambos algoritmos, igual a 5.

2.2.3. Interpretación de la segmentación

Al igual que antes vamos a mostrar una tabla con las medias de las características seleccionadas por cada algoritmo, un scatter matrix y un heatmap con los datos normalizados.

Tabla con las medias de cada característica:

CLUSTER	**EDAD**	**NPFAM**	**HM5**	**H0515**
0	33.364733	2.022401	0.000287	0.071511
1	46.023504	2.433455	0.021062	0.025336
2	35.965504	3.852161	1.210387	0.483321
3	23.784604	3.701410	0.029405	0.188382
4	43.637404	4.104017	0.038308	1.551210

Figura 4: SccaterMatrix KMeans Caso 2

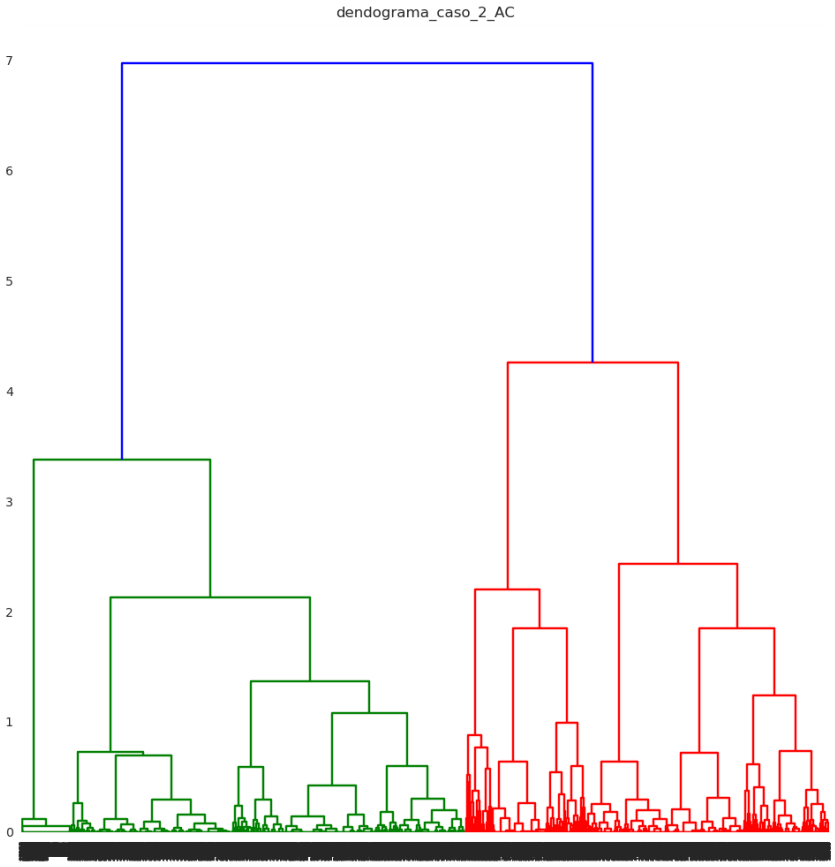

Figura 5: Dendogram AC Caso 2

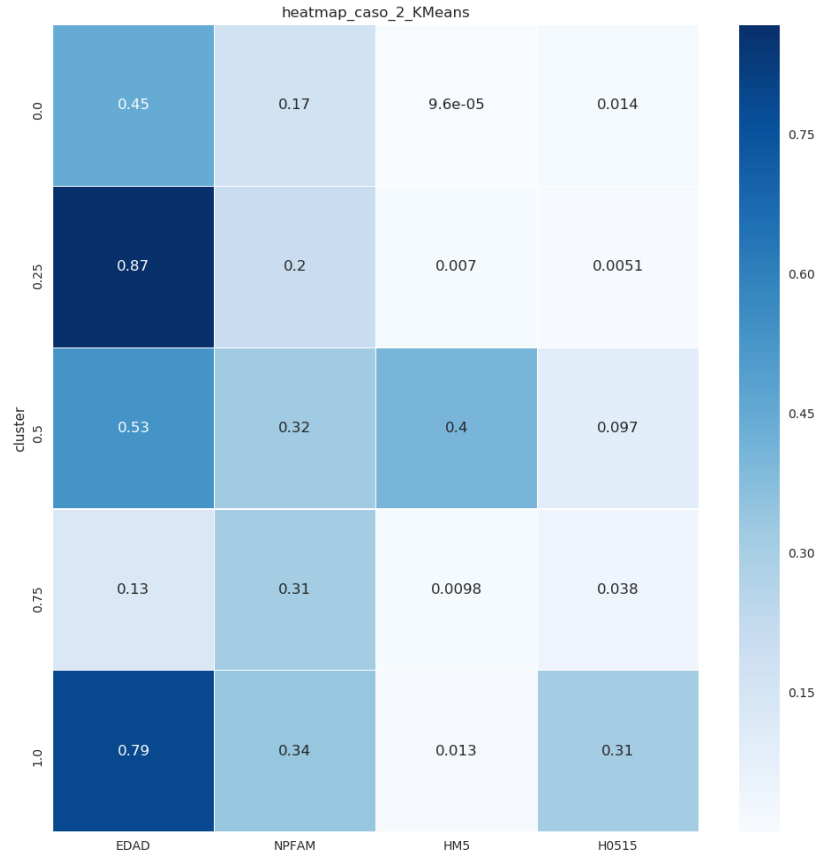

Figura 6: Heatmap KMeans Caso 2

Los resultados obtenidos han sido muy similares a los del caso 1, encontramos grupos idénticos a los de antes solo que en vez de mujeres es con hombres.

De nuevo podemos observar por ejemplo el grupo de personas jóvenes con una media de 23 años que viven con entre 3 y 4 personas, dos de las cuales es muy probable que se trate de los padre y la otra persona de un hermano de entre 5 y 15 años (cluster 3). Este grupo se puede identificar claramente con el grupo obtenido en el caso 1 correspondiente al cluster 3.

También volvemos a encontrar el grupo de personas con una media de 43 años (en el caso de las mujeres era de 40), con 4 personas en la familia, no viven personas menores de 5 años y viven entre 1 y 2 personas de entre 5 y 15 años (en el cluster 4). Este grupo parece ser el de hombres de unos 40 años que viven con sus parejas y con dos hijos adolescentes. Algo curioso es que este grupo es casi idéntico al cluster 4 del caso de estudio anterior, con la diferencia de que la media de este grupo es de 43 y en el caso anterior dicho

grupo tenía una media de 40, lo cual puede ser debido a que por lo general en una relación de pareja entre un hombre y una mujer, el hombre suele tener mayor edad.

A continuación se muestra las dos tablas de medias juntas de ambos casos (caso 1 y caso 2) para compararlas.

	Caso de estudio 1				**Caso de estudio 2**			
C	**EDAD**	**NPFAM**	**HM5**	**H0515**	**EDAD**	**NPFAM**	**HM5**	**H0515**
0	33.9823	3.7899	1.1998	0.4671	33.3647	2.0224	0.0002	0.0715
1	46.2429	3.2112	0.0205	0.3242	46.0235	2.4334	0.0210	0.0253
2	33.3555	2.0864	0.0000	0.1074	35.9655	3.8521	1.2103	0.4833
3	23.6254	3.6599	0.0385	0.1851	23.7846	3.7014	0.0294	0.1883
4	40.2000	4.2023	0.0454	1.8255	43.6374	4.1040	0.0383	1.5512

Viendo ambas tablas juntas se pueden identificar claramente los clusteres de los hombres con el de las mujeres. El cluster 0 del caso de estudio 1 se puede identificar con el cluster 2 del caso de estudio 2, el cluster 1 con el cluster 1, el cluster 2 con el 0, el cluster 3 con el 3 y cluster 4 con el 4.

Veamos ahora a proporción de tamaño de los clusters:

```
3:  4183 (24.13%)
4:  3759 (21.68%)
0:  3482 (20.08%)
1:  3276 (18.89%)
2:  2638 (15.22%)
```

El grupo mayoritario ha resultado ser el de hombres más jóvenes, mientras que el minoritario ha sido el cluster 2 formado por hombres de unos 35 años, que viven con 3 personas más, de las cuales una es un menor de menos de 5 años y en algunos casos el menor tiene entre 5 y 15 años, probablemente este sea el grupo de hombres con pareja (posiblemente casados) y con dos hijos.

2.3. Caso de estudio 3

2.3.1. Descripción del caso de estudio

En este caso de estudio de nuevo nos centramos sobre las mujeres de entre 20 y 50 años. Esta vez las características usadas son la edad (EDAD), el número de personas en la familia (NPFAM), el número de hijos (NHIJOS) y el nivel de estudios (ESREAL). Con este estudio se pretende encontrar los grupos

de mujeres según su nivel de estudios y el número de hijos que tengan. El tamaño de este conjunto es de 17996.

El código siguiente muestra como se ha seleccionado el conjunto de estudio y que características se han considerado para el clustering. La característica correspondiente al nivel de estudios, a pesar de no ser numérica, si que es ordinal, y por lo tanto tiene sentido hablar de poseer un menor o un mayor nivel de estudios, es por esto que esta característica puede utilizarse para el clustering.

```
subset_3 = censo.loc[(censo['EDAD']>=20) & (censo['EDAD']<=50) &
    (censo['SEXO']==mujer)]
usadas_3 = ['EDAD', 'NPFAM', 'NHIJOS', 'ESREAL']
X_3 = subset_3[usadas_3]
```

2.3.2. Resultados de los algoritmos

A continuación se muestran las tablas comparativas de los resultados obtenidos con los distintos algoritmos.

Algoritmo	N.Clusters	Calinksi-Harabaz	Silhouette	Tiempo
KMeans	5	12411.154979	0.331637	0.264984
AC	5	9741.022701	0.256565	19.088159
MeanShift	3	7096.641380	0.353258	33.662136
MiniBatchKM	4	12560.059256	0.350558	0.108730
DBSCAN2	3	370.173818	0.128967	1.389666

Al igual que siempre DBSCAN ha sido el que ha tenido peores resultados. Mencionar que el parámetro de este algoritmo eps ha tenido un valor diferente al de los casos anteriores, debido a que con el valor anterior de 0.2 solo obtenía un cluster en el que estaba el conjunto entero. Por ello he cambiado su valor a 0.1.

Los dos mejores han sido KMeans y MiniBatchesKMeans por lo que hemos probado diferentes valores, para el parámetro que fija el número de clusters, con estos dos algoritmos. Los resultados son los siguientes.

Algoritmo	N.Clusters	Calinksi-Harabaz	Silhouette	Tiempo
KMeans_5	5	12409.485972	0.330675	0.235832
KMeans_6	6	11355.819473	0.313344	0.386761
KMeans_7	7	10505.465223	0.271520	0.468503
KMeans_8	8	9796.676187	0.278232	0.631942

Algoritmo	N.Clusters	Calinksi-Harabaz	Silhouette	Tiempo
MiniBatchKM_5	5	12360.408603	0.327795	0.110767
MiniBatchKM_6	6	11282.780883	0.307794	0.114699
MiniBatchKM_7	7	10186.197399	0.294271	0.106650
MiniBatchKM_8	8	9605.907030	0.272033	0.206712

De nuevo podemos ver que al aumentar el número de clusters empeoran las puntuaciones de las métricas, obteniendo los mejores resultados con el número de clusters igual a 5.

2.3.3. Interpretación de la segmentación

Tabla con las medias de cada característica:

CLUSTER	EDAD	NPFAM	NHIJOS	ESREAL
0	45.578352	3.548361	1.900497	3.970209
1	28.699869	3.022534	0.289353	8.880797
2	23.824892	3.538913	0.264670	4.654416
3	35.204704	3.428018	1.406565	4.385629
4	42.804083	3.263610	1.424069	8.801576

Figura 7: SccaterMatrix KMeans Caso 3

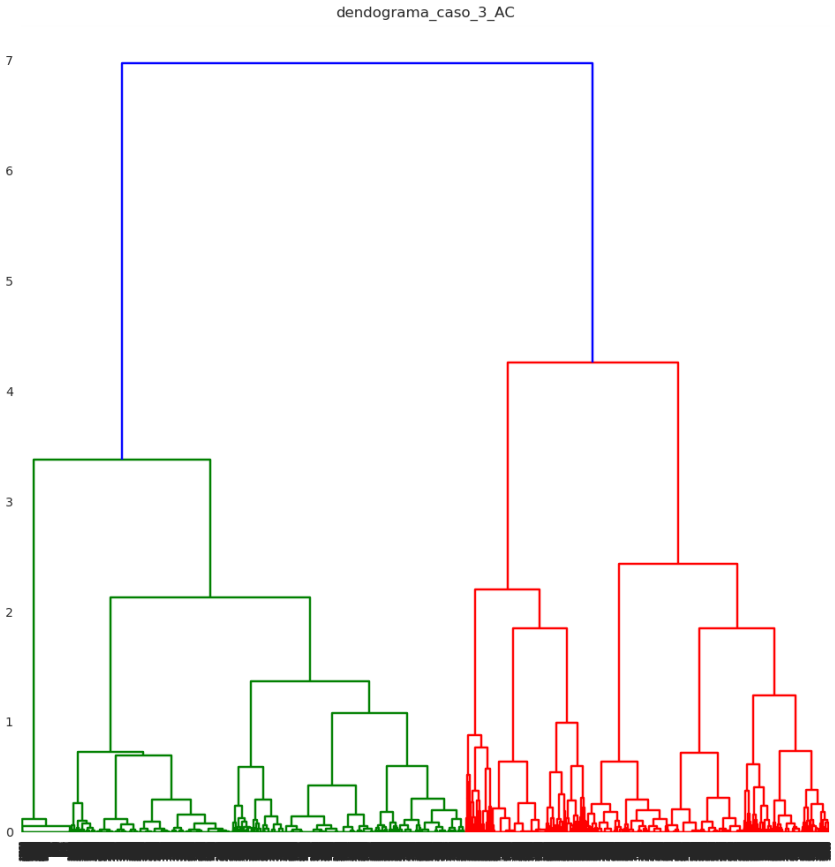

Figura 8: Dendogram AC Caso 3

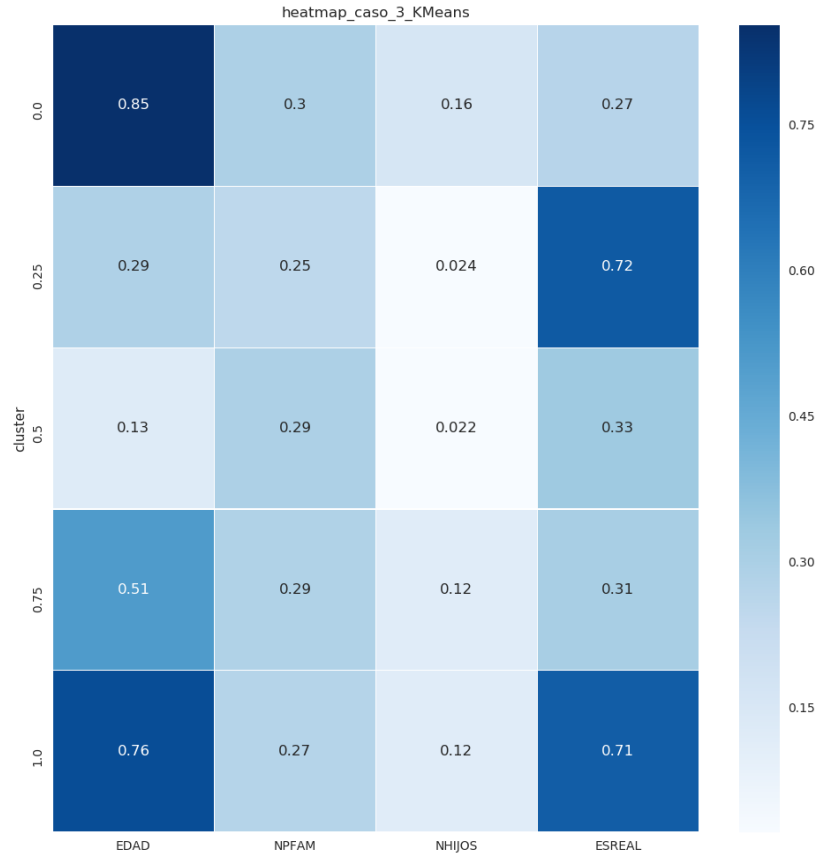

Figura 9: Heatmap KMeans Caso 3

Podemos ver que en este heatmap hay dos claros grupos con un mayor nivel de estudios, uno de ellos de mayor edad que el otro. Después al analizar las proporciones veremos que el tamaño del grupo joven con mayor nivel de estudios es mayor que el tamaño del otro grupo con nivel alto de estudios. Lo cual parece indicar que cada vez hay más mujeres con nivel alto de estudios y esto explica que el porcentaje sea mayor en las mujeres jóvenes que en las mayores.

Analicemos en primer lugar el cluster 0. La edad media de las mujeres pertenecientes a este cluster es de 45 años, el número de personas en la familia es de entre 3 y 4 y tienen en su mayoría 2 hijos. Este grupo tiene un nivel de estudios medio. Este grupo esta compuesto por madres de unos 45 años con un nivel medio de estudios y por el número de personas con las que conviven y el número de hijos que tiene, es probable que dentro de este grupo encontremos a madres que viven con su pareja, probablemente casadas, y madres

solteras que viven solo con sus hijos.

El cluster 1 esta compuesto por mujeres más jóvenes, con una media de 28 años que viven con dos personas más y en pocos casos tienen hijos. El nivel de estudios de este grupo es el mayor de todos los clusters. Posiblemente se trate de mujeres jóvenes que han terminado los estudios recientemente y que aún vivan con sus padres, y posiblemente un pequeño porcentaje de ese grupo tenga hijos y viva con su pareja (esto explicaría el 0.28 hijos y número de familiares igual a 3 sería debido a su pareja y su hijo).

El cluster 2 es el grupo más joven, con una media de edad de 23 años, el número de personas en la familia está entre 3 y 4, en pocas ocasiones tienen hijos y su nivel de estudios es medio. Pienso que este grupo está compuesto en su mayoría por jóvenes que aún están estudiando (esto explica el nivel medio de estudios) y que viven aún con sus padre y con algún hermano. Un pequeño porcentaje de este grupo puede estar compuesto por madres adolescentes que viven con su pareja.

El cluster 3 está formado por mujeres con una media de 35 años, que viven con su pareja y tiene hijos. Su nivel de estudios es medio.

El cluster 4 es similar al cluster 0, con la diferencia de que el nivel de estudios de este grupo es superior, estando formado seguramente por muchas mujeres con título universitario o incuso una titulación mayor.

La proporción de cada cluster es la siguiente:

```
0:  5035 (27.98%)
3:  3869 (21.50%)
2:  3238 (17.99%)
1:  3062 (17.01%)
4:  2792 (15.51%)
```

Algo interesante de destacar es que entre los grupos de personas mayores de 40 años, que se corresponden con los cluster 0 y 4, vemos que el cluster 0 tiene un porcentaje mayor que el cluster 4, lo cual indica que entre personas mayores de 40, es menos común ver mujeres con un nivel alto de estudios. Sin embargo observando el cluster 1 que tiene un porcentaje relativamente alto en comparación con los demás parece que entre las mujeres jóvenes es más común encontrar gente con un nivel alto de estudios.

3. Contenido adicional

4. Bibliografía

- http://scikit-learn.org/stable/modules/clustering.html
- http://www.learndatasci.com/k-means-clustering-algorithms-python-intro/
- http://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html
- https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dend
- http://www.learndatasci.com/k-means-clustering-algorithms-python-intro/
- Transparencias de clase.