Inteligencia de Negocio: Práctica 2

Análisis Relacional mediante Segmentación

José Antonio Álvarez Ocete - 77553417Q
joseantonioao@correo.ugr.es

6 de diciembre de 2019

Índice

1. Introducción **3**

2. Casos de estudio **3**

2.1. Caso 1: Estudio de los hijos deseados 3

2.1.1. Análisis general 4

2.1.2. Análisis específico: algoritmo *KMeans* 7

2.2. Caso 2: Análisis jerárquico sobre tratamientos de reproducción asistida. 10

2.2.1. Análisis general 10

2.2.2. Análisis específico: algoritmo *Ward* 11

2.2.3. Análisis jerárquico de clustering aglomerativo. 13

2.3. Caso 3: Análisis del reparto de tareas familiares en parejas heterosexuales y homosexuales. 15

2.3.1. Análisis general 16

2.3.2. Análisis específico: algoritmo *KMeans* en parejas heterosexuales 19

2.3.3. Comparativa con parejas homosexuales 22

3. Bibliografía **23**

1. Introducción

En esta práctica veremos el uso técnicas de segmentación para realizar análisis relacional sobre un conjunto de datos dado. En nuestro caso utilizaremos los datos de una encuesta sobre fecundidad [1] realizada por el Instituo Nacional de Estadística (INE). En ella se estudian distintas variables de la vida de una persona que pueden haber tenido influencia en la fecundidad de la misma.

En particular, 14556 personas respondieron a esta encuesta con un total de 483 variables. Cabe destacar que este es el primer año en el que la encuesta es realizada también por hombres.

Realizaremos tres casos de estudio. En cada uno nos reduciremos a un subconjunto de las variables proporcionadas por la encuesta, al mismo tiempo que realizamos un filtrado sobre la población. Por un lado esto se debe a que la capacidad de cómputo disponible es reducida. Por otro, la información que podemos obtener a partir de la interpretación de los resultados es considerablemente menor si el número de variables es alto.

Para esta práctica se han seleccionado 5 algoritmos que ejecutaremos en cada caso de estudio. Detallamos a continuación los parámetros con los que se ejecuta cada algoritmo en el análisis general de cada caso de estudio, a no ser que se especifique lo contrario:

- **K-means**: 5 clusters (*n_clusters=5*) y ejecutado 5 veces con distintos centroides iniciales, quedándonos con el mejor resultado (*n_init=5*).
- **MeanShift**: no forzamos a que se añadan todos los elementos a un cluster (*all_cluster=False*) e imponemos que el número de elementos de las seeds iniciales sea al menos 3 para reducir el tiempo de cómputo (*min_bin_freq=3*).
- **Ward**: utilizamos el algoritmo *AgglomerativeClustering* con parámetro *linkage='ward'* como implementación del algoritmo Ward. Adicionalmente fijamos el número de clusters a 5 (*n_clusters=5*) .
- **DBScan**: Fijamos *épsilon* a 0.35 (*eps=0.35*) y el número mínimo de elementos en el vecindario para que un punto pueda ser centroide a 5 (*min_samples=5*).
- **Birch**: Número de clusters fijado a 5 (*n_clusters=5*) y el *threshold* a 0.1 (*threshold=0.1*).

2. Casos de estudio

2.1. Caso 1: Estudio de los hijos deseados

En este primer caso estudiaremos principalmente el número de hijos del encuestado en función de las siguientes variables:

- **NHBIOADOP**: Número de hijos que ha tenido el encuestado, ya sean biológicos o adoptivos.
- **EDAD**: Edad del encuestado.
- **NTRABA**: Tiempo que lleva el encuestado en su empleo actual, en años.
- **TEMPRELA**: Tiempo de relación con la pareja actual, en años.
- **NHOGAR**: Tiempo que el encuestado lleva viviendo en el hogar actual, en años.

Adicionalmente realizamos un filtrado de la población para estudiar unicamente a las personas actualmente empleadas (utilizando **TRABAJAACT**) para que la variable **NTRABA** sea lo más relevante posible. Filtramos también aquellos objetos con más de 10 hijos o más de 7 años en el mismo hogar. Estos elementos eran principalmente *outlayers* que complicaban la visuzalización. Con ello reducimos nuestra población a un total de 9161 individuos.

2.1.1. Análisis general

En primer lugar ejecutamos todos los algoritmos sobre nuestro conjunto de datos. Presentamos a continuación un resumen de los resultados obtenidos.

Tabla 1: Caso 1: Resultados generales

Algoritmo	Calinski-Harabasz	Silh	Tiempo	Número de clusters
KMeans	3871.90	0.25	0.40	5
MeanShift	660.81	0.19	186.14	2
Ward	2979.72	0.17	2.50	5
DBScan	11.74	0.49	1.29	2
Birch	3026.56	0.17	0.44	5

Antes de analizar los clusters obtenidos miramos los tiempos de ejecución de cada algoritmo para observar como el tiempo de ejecución del algorimto *MeanShift* es absurdamente alto. Ya que no proporcionamos valores para el *bandwidth* ni para las seeds del algoritmo, éste las estima, disparándose el tiempo de ejecución.

A continuación miramos los resultados obtenidos por el *DBScan*. A pesar de que obtiene el menor valor del coeficiente *Calinski-Harabasz* entre todos los algoritmos, el coeficiente Silhouette es es el mejor de todos, también con diferencia. Este nos hace pensar que el agrupamiento realizado puede ser desigual. Para comprobar esta hipótesis representamos los tamaños de los clusters (en tanto por ciento) para cada algoritmo.

Observamos en la figura 1 la desigualdad en la distribución comentada con anterioridad. En el caso de *DBScan*, practicamente todos los datos (un 99.98 %) se agrupan en un único cluster, resultando en un alto valor del coeficiente *Silhouette* y un terrible valor para *Calinski-Harabasz*.

Los resultados de *MeanShift* son relativamente parecidos. Al no forzar la inclusión de todos los elementos en algún cluster, el algoritmo detecta un 37.49 % de elementos que no sabe clasificar. Estos elementos son el cluster número 1.

Respecto a los otros tres algoritmos, obtenemos resultados similares, tanto para el coeficiente *Calinski-Harabasz* como para el *Silhouette*, obteniendo *KMeans* valores algo mejores en ambos. Obviamente el número de clusters obtenido es 5, ya que habíamos fijado tal número como parámetro.

Realizamos un estudio paramétrico sobre los distintos algoritmos, excluyendo a *MeanShift* debido al alto tiempo de cómputo. En el caso de *DBScan* alteraremos el *épsilon*, mientras que para los otros tres algoritmos modificamos el números clusters. En la figura 2 representamos las distintas ejecuciones realizadas, coeficiente *Calinski-Harabasz* respecto al *Silhouette*. Finalmente, el tamaño de cada burbuja indica el valor del parámetro seleccionado en cada caso. En la tabla 2 se encuentran los resultados obtenidos.

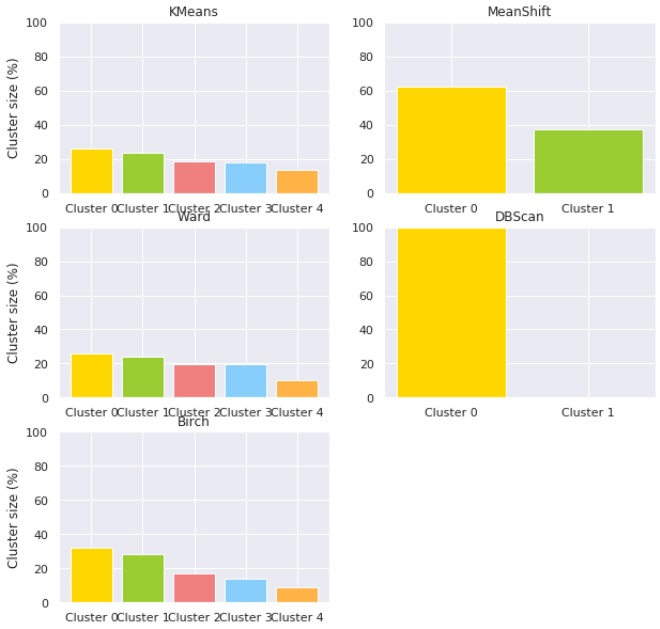

Figura 1: Caso 1: Tamaños de clusters

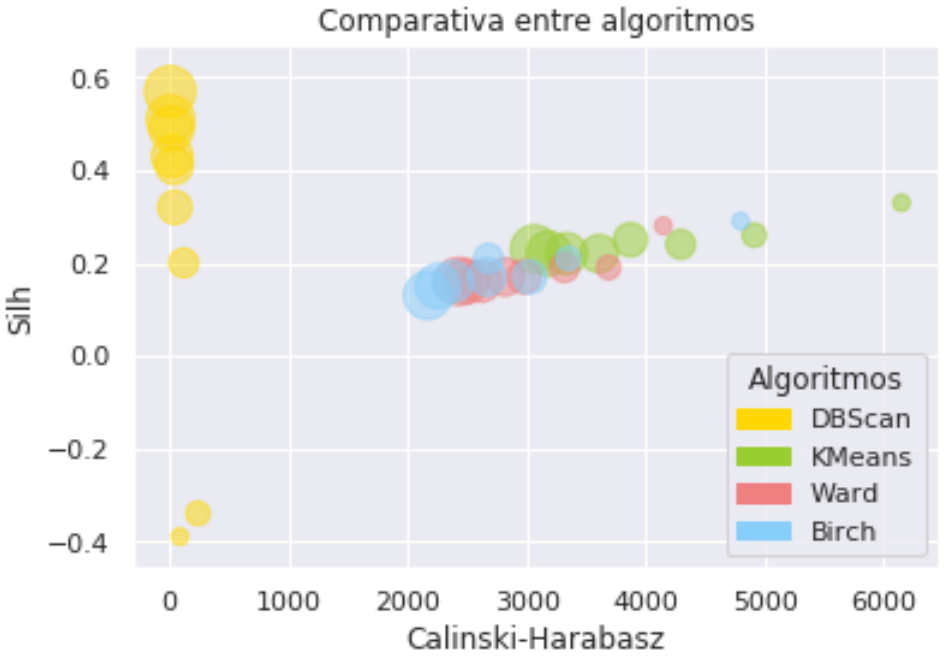

Figura 2: Caso 1: Comparativa de parámetros

Tabla 2: Caso 1: Resultados de ajuste de parámetros

Algoritmo	Silh	Calinski-Harabasz	Valor del parámetro	N. clusters
DBScan	-0.39	90.24	0.05	16
DBScan	-0.34	241.96	0.10	20
DBScan	0.20	122.09	0.15	2
DBScan	0.32	48.07	0.20	2
DBScan	0.41	43.20	0.25	2
DBScan	0.43	26.39	0.30	2
DBScan	0.49	19.40	0.35	2
DBScan	0.51	11.53	0.40	2
DBScan	0.57	9.16	0.45	2
KMeans	0.33	6148.06	2.00	2
KMeans	0.26	4911.49	3.00	3
KMeans	0.24	4293.59	4.00	4
KMeans	0.25	3871.87	5.00	5
KMeans	0.22	3609.15	6.00	6
KMeans	0.22	3355.72	7.00	7
KMeans	0.22	3187.80	8.00	8
KMeans	0.23	3067.40	9.00	9
Ward	0.28	4147.42	2.00	2
Ward	0.19	3688.62	3.00	3
Ward	0.19	3317.32	4.00	4
Ward	0.17	2979.72	5.00	5
Ward	0.17	2819.62	6.00	6
Ward	0.16	2615.95	7.00	7
Ward	0.16	2493.78	8.00	8
Ward	0.16	2429.35	9.00	9
Birch	0.29	4796.80	2.00	2
Birch	0.21	3350.47	3.00	3
Birch	0.21	2680.54	4.00	4
Birch	0.17	3026.56	5.00	5
Birch	0.17	2668.37	6.00	6
Birch	0.16	2381.54	7.00	7
Birch	0.15	2255.30	8.00	8
Birch	0.13	2170.71	9.00	9

Por un lado, el algoritmo *DBScan* obtiene resultados similares independientemente del valor *épsilon*, agrupando los datos a partir de *épsilon=0.15* en dos únicos clusters. Para valores menores que 0.15 se obtienen más clusters, sin llegar a obtener un coeficiente Calinski-Harabasz cercano al del resto de algoritmos. Concluimos por tanto que o bien este algoritmo no es nada adecuado para este subconjunto de individuos debido a la distribución de los mismos, o bien no hemos sabido ajustar los parámetros lo suficiente.

Para el resto de casos, los resultados son consistentes: *KMeans* es el mejor algoritmos para cada valor del parámetro, sin presentar claras mejoras salvo para *n_clusters=0.2*. Es por ello que procedemos a estudiarlo de forma detallada.

2.1.2. Análisis específico: algoritmo *KMeans*

Fijado el algorimto *KMeans* con 5 clusters, representamos el *Scatter Matrix* 3 de los resultados obtenidos, así como el *Heatmap* de centroides asociado 4. Procedemos a analizar la distribución en clusters obtenida.

Por un lado, los dos clusters cuyos encuestados apenas tienen hijos son el 0 y el 3. Además, podemos diferenciar estos claramente en función del resto de variables. En el primero los encuestados tienen entre 20 y 40 años y han estado con su pareja entre 0 y 18 años (10, en media). Además, la mayor parte de este grupo lleva muy poco tiempo en su trabajo, como podemos apreciar en la casilla central de la figura 3. Entendemos entonces que este cluster esta formado por gente joven, con relaciones relativamente cortas.

El cluster 3 es prácticamente opuesto. La edad de sus individuos oscila entre 35 y 60 años, y sus relaciones han sido mucho más duraderas: de 18 a 40 años. Aunque el rango temporal en el que han permanecido en el mismo trabajo es similar al del grupo anterior, de 0 a 18 años, la media es considerablemente mayor, 10.39 años frente a 2.64. Podemos afirmar por lo tanto que este grupo se compone de personas mayores, con relaciones mucho más duraderas y que apenas han tenido hijos.

Llegado este punto cabe destacar como la variable **NHOGAR** apenas es significativa, como se observa en la figura 4. Los clusters toman valores parecidos en dicha variable, resultando curioso que el cluster recién descrito es aquel con menor media (1.75).

Comparemos a continuación los clusters 1 y 4. Ambos tienen media de edad parecida y bastante alta, entorno a los 50 años. De la misma forma, ambos grupos llevan viviendo entorno a 3.5 años en su hogar y han tenido, en media, un número de hijos parecido. El factor determinante para diferenciarlos es el tiempo de permanencia en su empleo actual. El cluster 4 tiene una media de 25 años, donde ninguno de sus individuos baja de 17. Estos son los valores más altos de toda la población. Por otro lado, el cluster 2 tiene una media de 6.6 años, tomando valores distribuidos pero aún así claramente diferenciados de los del cluster anterior.

Caracterizamos por lo tanto ambos clusters por componerse de gente mayor que han tenido entorno a uno o dos hijos. Diferenciándose entre si por el tiempo de permanencia en su empleo actual.

Finalmente, podemos distinguir fácilmente el cluster 2 mirándo unicamente a las dos primeras variables. El número de hijos es relativamente elevado, diferenciándose del primer subgrupo (los clusters 0 y 3), pero cuya edad es, en media, diez años inferior a la del segundo subgrupo (los clusters 1 y 4).

Figura 3: Caso 1: *Scatter Matrix* de *KMeans*

Figura 4: Caso 1: *Heatmap* de *KMeans*

Antes de terminar con este análisis podemos observar en la figura 1 como la segmentación de la población obtenida es relativamente heterogénea, teniendo el cluster 0 casi el doble de individuos que el 4.

2.2. Caso 2: Análisis jerárquico sobre tratamientos de reproducción asistida.

En este segundo caso de estudio buscamos realizar un análisis de los métodos jerárquicos. En particular analizaremos un algoritmo de cluster aglomerativo: **WARD**. Para ello tomaremos un subconjunto de objetos mucho más reducido que en el caso anterior. Las variables utilizadas han sido las siguientes:

- **NHIJOS**: Número de hijos del encuestado o su pareja.
- **TRAREPRO**: Tipo de tratamiento al que se ha sometido el encuestado. Los posibles valores son:
  - 1. Coito programado.
  - 2. Inseminación artificial.
  - 3. Fecundación in vitro (FIV) o inyección intracitoplasmática (ICSI).
  - 4. Gestación subrogada.
  - 5. Otros tratamientos médicos.
- **NMESESTRAREPRO**: Número de meses en tratamiento.
- **NEMBTRAREPRO**: Número de embarazos conseguidos.

De cara a obtener información relevante en el caso de estudio decidimos estudiar unicamente a aquellos encuestados que se hayan sometido a algún tratamiento artificial usando la variable **TRAREPRO**. Adicionalmente eliminamos aquellos objetos con un número de embarazos conseguidos mayor a 6 para visualizar mejor la información. Con ello nos restringimos a 807 individuos. Un número tan bajo nos permitirá visualizar mejor los dendogramas asociados al algoritmo de clustering aglomerativo.

2.2.1. Análisis general

En primer lugar ejecutamos todos los algoritmos sobre nuestro conjunto de datos. Presentamos a continuación un resumen de los resultados obtenidos.

Tabla 3: Caso 2: Resultados generales

Algoritmo	Calinski-Harabasz	Silh	Tiempo	Número de clusters
KMeans	316.74	0.27	0.02	5
MeanShift	92.54	0.29	3.36	6
Ward	276.24	0.25	0.03	4
DBScan	15.93	0.43	0.02	2
Birch	199.59	0.23	0.09	5

Comenzamos observando los tiempos obtenidos. De nuevo, el algoritmo *MeanShift* obtiene un tiempo muy superior a los demás, siendo muy inferior al del caso anterior (186 segundos).

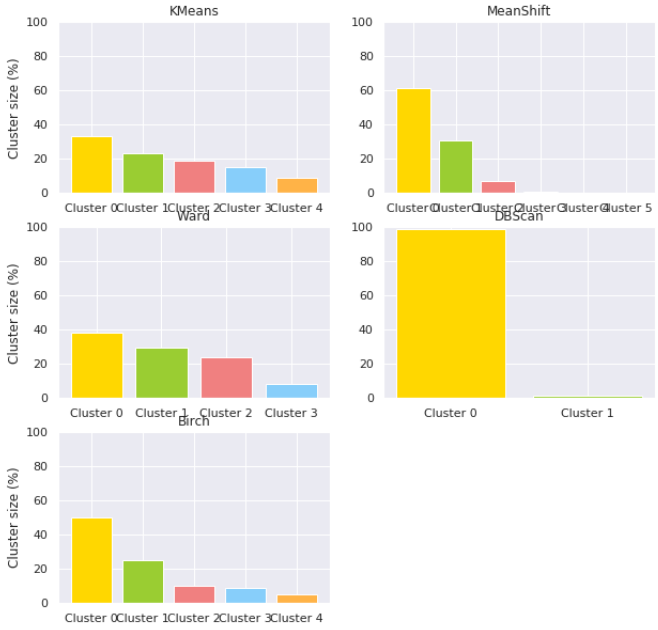

Figura 5: Caso 2: Tamaños de clusters

Esto se debe lógicamente a la diferencia en el tamaño de la población.

Para comprender mejor la distribución en clustering representamos de nuevo el tamaño de los mismos en la figura 5. Observamos que la distribución es similar a la obtenida en el caso anterior (1). Los algoritmos obtienen valores parecidos en el coeficiente Silhouette mientras que respecto al Calinski-Harabasz podemos diferenciar dos subgrupos: valores bajos (*MeanShift* y *DBScan*) frente a valores altos (*KMeans*, *Ward* y *Birch*). Esto encaja con la distribución en clustering representada en 5, donde *MeanShift* y *DBScan* agrupan la mayoría de los datos en un único cluster. Sin embargo, hemos obtenido valores mucho menores del coeficiente *Calinski-Harabasz* que los que obtuvimos en el primer caso de estudio (1). Esto podría deberse a que todas la variables utilizadas en este caso de estudio son discretas. Refutaremos esta hipótesis en el tercer caso de estudio.

Pasamos ahora a estudiar los resultados obtenidos por el algoritmo *Ward* en profundidad.

2.2.2. Análisis específico: algoritmo *Ward*

Representamos para cada cluster obtenido un *BoxPlot* de la distribución de cada variable 6. Visualizamos también un *HeatMap* con los centroides de los distintos clusters 7. Debido a que las variables utilizados son todas discretas, la *Scatter Matrix* en este caso de estudio no nos proporciona información relevante. Es por ello que se ha omitido.

Los clusters obtenidos son sencillos de analizar mirando la distribución de centroides obtenida en la figura 7. Comenzamos fijándonos en el tipo de tratamiento realizado (**TIPOTRAREPRO**), donde el cluster 2 obtiene un alto valor para el centroide mientras que el resto de clusters tienen valores similares entre si. Comprendemos mejor aún este comportamiento observando la correspondiente columna en el *BoxPlot* 6. Es claro que se ha agrupado en el segundo cluster aquellos

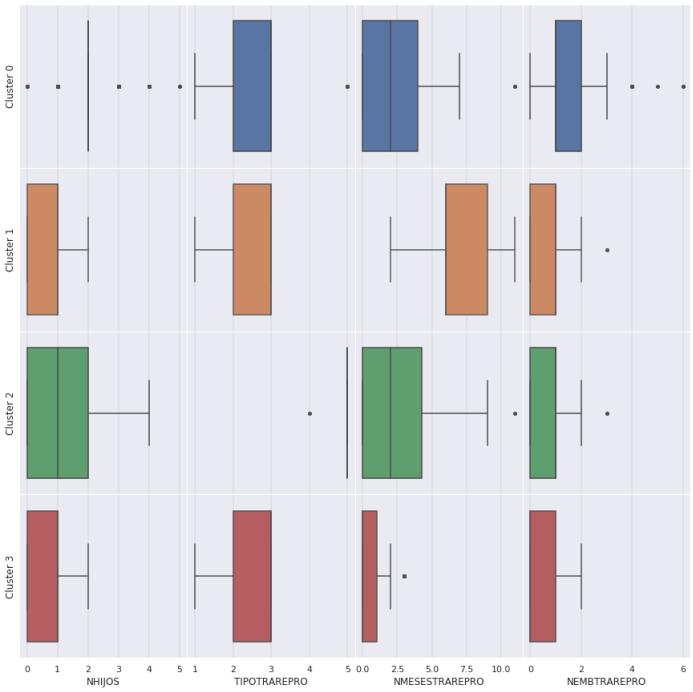

Figura 6: Caso 2: Distribución de variables en *BoxPlot* para cada cluster

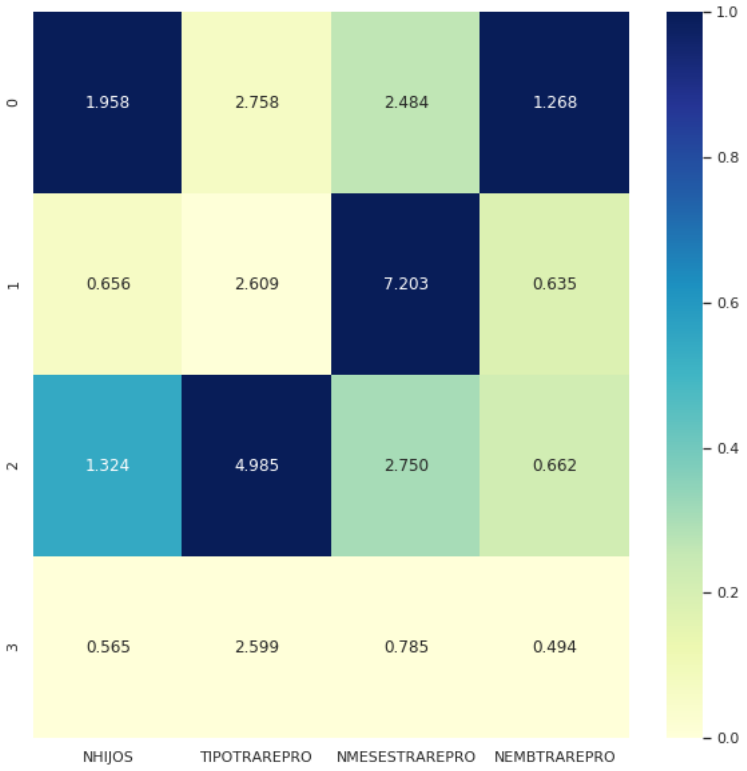

Figura 7: Caso 2: *HeatMap* de centroides.

encuestados que respondieron 'Otros tratamientos' a esta pregunta. El resto de clusters toman los valores 'Inseminación artificial' y 'Fecundación in vitro (FIV) o inyección intracitoplasmática (ICSI)', con algunas respuestas de tipo 'Coito programado'. Es posible que esta segmentación fuese distinta si la respuesta 'Otros tratamientos' estuviese asociada al valor 4 en vez de al 5, puesto que no distaría tanto numéricamente de las demás.

El cluster 0 se caracteriza por tener altos valores en número de hijos y número de embarazos conseguidos. Si bien estas variables están claramente relacionadas, estudiando ambas buscaba encontrar subgrupos de la población con desbalanceos en las mismas. Por ejemplo, un alto número de hijos y pocos embarazos conseguidos (mostrando parejas que han utiliado métodos de inseminación asistida y han tenido hijos de otras formas), o bajo número de hijos y alto número de embarazos conseguidos, mostrando un alto índice de aborto o situaciones similares. Esto se ha manifestado parcialmente en el cluster dos, dándose el primer ejemplo mencionado.

El cluster 1 se caracteriza por presentar un alto valor en meses de tratamiento. Aqui la diferencia es sustancial: siete meses respecto a los uno o dos obtenidos en el resto de clusters. Sin embargo esto tampoco repercute de forma especialmente positiva en el número de embarazos conseguido. Además, apreciamos en 5 que este cluster representa a un 24 % de la población. Es decir, alrededor de un cuarto de los encuestados requirieron tratamientos notablemente más largo, que el resto, de aproximadamente 5 meses más.

Finalmente llegamos al cluster 3, con un 10 % de la población, obteniendo valores bajos en todas las variables. Caracterizaremos este cluster principalmente por la corta duración de sus tratamientos, apenas un mes en la mayoría de los casos (6). No todos los encuestados de este subgrupo obtuvieron resultados positivos, ya que el número de embarazos conseguidos es, en media, alrededor de un medio.

2.2.3. Análisis jerárquico de clustering aglomerativo.

Pasamos a estudiar un dendograma con un *HeatMap* asociado generado utilizando la librería *Seaborn* (8).

He descartado comparar la asociación de clusters estudiada en el apartado anterior añadiendo una columna al dendograma con dicha asociación pues esta no aportaba ninguna información relevante, o al menos que haya podido inferir de la misma.

Estudiaremos el dendograma de forma escalonada, viendo las divisiones realizadas de izquierda a derecha, como es natural.

Observando la primera división podemos suponer que esta se hace principalmente debido a los altos valores de la variable **TIPOTRAREPRO** y los bajos valores de **NMESESTRAREPRO**. Esto es, el tipo de tratamiendo es 'Otros tratamientos' y los tratamientos fueron de corta duración. Podemos relacionar esta segmentación con el cluster número dos del apartado anterior. Denotaremos a este cluster por **cluster superior**.

La segunda división del dendograma divide a la población basándose en la duración del tratamiento. El **cluster central** agrupa largas duraciones de tratamiento mientras que el **cluster inferior** contiene duraciones medias: no tan altas como las del cluster central ni tan altas como las del cluster superior.

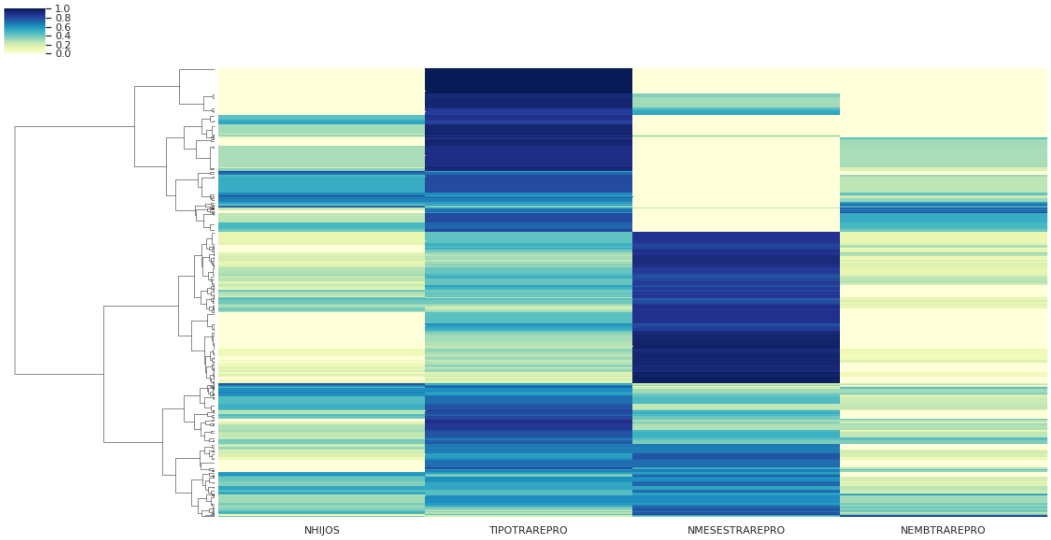

Figura 8: Caso 2: Dendograma obtenido con la librería *Seaborn*.

Hasta ahora las divisiones realizadas unicamente se refieren a la duración del tratamiento y al tipo del mismo. Ante esta segregación podemos suponer que el tipo de tratamiento utilizado por aquellos encuestados que marcaron 'Otros tratamientos' repercute en una poca duración del mismo, ya sea debido a que el tratamiento tiene dicha duración o porque los encuestados deciden dejar de realizarlo. Esta hipótesis dista notablemente de las conclusiones obtenidas en el apartado anterior, donde el cluster 2 recogía a las personas con dicho tipo de tratamiento pero cuyas duraciones eran bajas o medias, no altas (6).

Volviendo al dendograma, veamos como los tres clusters comentados se subdividen. Por un lado, el cluster superior se segmenta en basándose en el número de hijos: valores muy pequeños frente a valores intermedios. Llama la antención en esta subdivisión el cluster con encuestados sin hijos y sin embarazos conseguidos. Éste último a su vez se divide dependiendo de la duración de los tratamientos, pero manifiesta esencialmente procedimientos abandonados o fallidos.

Caben destacar dos subclusters adicionales del cluster superior. El primero por encajar con una de las conclusiones realizadas en el ejemplo anterior: encuestados con un número de hijos visiblemente superior a número de embarazos conseguidos mediantes tratamientos de reproducción asistida. El segundo por una configuración curiosa de las respuestas obtenidas: Un subgrupo cuyo embarazos conseguidos es superior al del resto de la población y cuya duración de tratamiento es ridiculamente baja. Este cluster se segmenta en otros dos, basado al mismo tiempo en el número de hijos y el tipo de tratamiento.

Las divisiones más relevantes en los clusters central e inferior se basan en el número de hijos y la duración del tratamiento respectivamente. El único subgrupo relevante que he podido apreciar está en el cluster central. Se trata de un grupo de encuestados sin hijos y sin embarazos conseguidos cuyo duración de tratamiento es bastante alta.

2.3. Caso 3: Análisis del reparto de tareas familiares en parejas heterosexuales y homosexuales.

En el último caso de estudio estudiaremos como afecta el reparto de tareas de casa relacionadas con los hijos en las parejas heterosexuales frente a como lo hace en las homosexuales. Comenzamos estudiando unicamente al primer subgrupo. Manteniendo en nuestro conjunto de datos a aquellas parejas Hombre - Mujer (utilizando las variables **SEXO** y **SEXOPAR**) que además conviven con al menos un hijo de menos de catorce años (con **CONVIVEH14**) nos restringimos a un total de 4845 objetos.

Cabe destacar que en esta subpoblación, todos los encuestados son mujeres. Aunque esto no era un factor decisivo al decidir el filtrado nos facilitará notablemente el posterior análisis.

En la encuesta realizada hay de diez preguntas relacionadas con las tareas de casa que nos interesan particularmente. Son las siguientes:

- **VESTIR**: Quién viste a los niños.
- **BANAR**: Quién baña a los niños.
- **ACOSTAR**: Quién acuesta a los niños.
- **COMIDAS**: Quién decide la comida de los niños.
- **ENFERMOS**: Quién se queda con los niños cuando estan enfermos.
- **JUGAR**: Quién juega con los niños.
- **DEBERES**: Quién ayuda a los niños con los deberes.
- **COLEGIO**: Quién lleva a los niños al colegio.
- **ROPA**: Quién le compra ropa a los niños.
- **ELIGEEXTRAESC**: Quién elige las actividades extraescolares de los niños.

A todas estas preguntas los encuestados podían responder una de las siguientes opciones:

- 1. Entrevistado.
- 2. Pareja.
- 3. Entrevistado y pareja por igual.
- 4. Abuelos.
- 5. Otra persona del hogar.
- 6. Otra persona de fuera del hogar.
- 7. Los niños lo hacen por sí mismos.

Sintetizaremos la información de todas estas preguntas en un único coeficiente, un valor entre -1 y 1, que denotaremos por **COEFTAREAS**. Por cada tarea que realice la mujer de la pareja restaremos *0.1* mientras que para cada tarea que realice el hombre sumaremos *0.1*. De cara a entender esta métrica, valores cercanos a cero muestra un equilibro en el repartor de las tareas mientras que valores lejanos denotan un desbalanceo en el mismo. Si el valor es negativo, la mujer está realizando más tareas mientras que si es positivo, lo hace el hombre.

Además de este coeficiente utilizaremos las siguientes variables:

- **EDAD**: Edad del entrevistado.
- **NHIJOSCONV**: Número de hijos con los que convive. A pesar de que me habría gustado utilizar **NHIJOSDESEO** (número de hijos deseado) en su lugar, los 4845 encuestados a los que nos hemos restringido tienen valor *NaN* en dicha variable.
- **ESTUDIOSA**: Nivel de estudios del encuestado. Toma valor en una escala de uno a nueve, donde uno es 'menos que primaria' y nueve, 'enseñanzas de doctorado'.
- **SATISFACENINOS**: Satisfacción del encuestado respecto al reparto de tareas en casa. Toma valor en un escala de cero al diez, donde cero es totalmente insatisfecho y 10, totalmente satisfecho.

2.3.1. Análisis general

En primer lugar ejecutamos todos los algoritmos sobre nuestro conjunto de datos, buscando esta vez seis clusters (*n_clusters=6* para aquellos algoritmos que admiten dicho parámetro). Presentamos a continuación un resumen de los resultados obtenidos.

Tabla 4: Caso 3: Resultados generales

Algoritmo	CH	Silh	Tiempo	Número de clusters
KMeans	1409.53	0.20	0.18	6
MeanShift	971.95	0.24	44.94	2
Ward	1128.03	0.15	0.62	6
DBScan	8.11	0.41	0.46	2
Birch	1024.14	0.15	0.28	6

Observando los tiempos obtenidos vemos como el algoritmo *MeanShift* vuelve a obtener un tiempo absurdarmente alto en comparación con el resto de algoritmos. Uniendo esto a la descomposición de los clusters observada en la figura 9 concluimos que el algoritmo está obviamente mal configurado. Podríamos decir lo mismo sobre *DBScan*.

Utilizamos la misma estrategia que en el primer caso de estudio (2) para buscar una reconfiguración de los parámetros. Encontramos el resultado en la figura 10 y la tabla 5. Esencialmente se repite el escenario del primer caso de estudio: *DBScan* no genera buenos resultados para ninguno de los valores del parámetro utilizados mientras que los coeficientes Calinski-Harabasz y Silhouette obtenidos por el resto de algoritmos están relativamente agrupados, siendo *KMeans* consistentemente mejor fijando un valor del parámetro. Es por ello que en el análisis específico volveremos a estudiar este algoritmo.

Antes de pasar a dicho análisis volvemos a la tabla 4 para descartar una hipótesis formulada previamente. En el segundo caso de estudio atribuimos los bajos valores del coeficiente Calinski-Harabasz proporcionados por todos los algoritmos a la naturaleza de las variables escogidas: todas categóricas. Los valores obtenidos en este caso junto con las variables escogidas nos hace pensar que esto no es cierto. Observando la forma en la que se calcula este coeficiente nos damos cuenta de que depende directamente del número de objetos en estudio, que en el segundo caso era especialmente bajo. Debido a esto obtenemos en este caso de estudio valores intermedios frente a los altos valores del primer caso y los bajos valores del segundo, cuadrando a la perfección con los tamaños de las poblaciones estudiadas en cada caso.

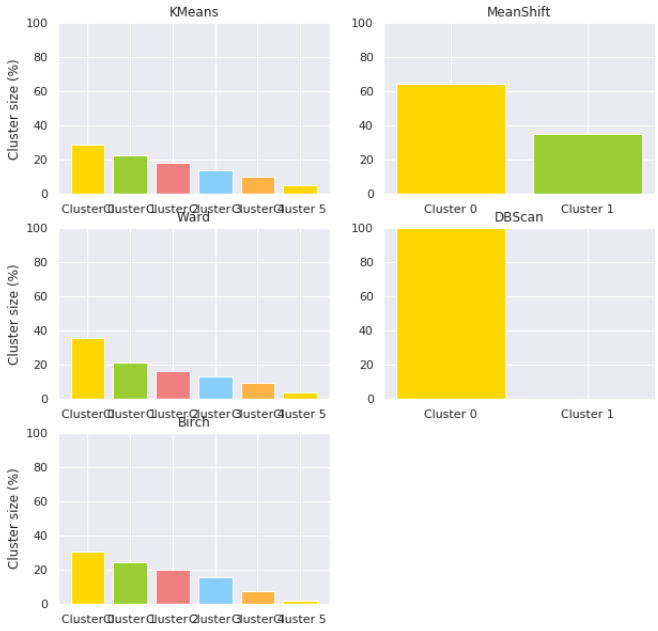

Figura 9: Caso 3: Tamaños de clusters para parejas heterosexuales.

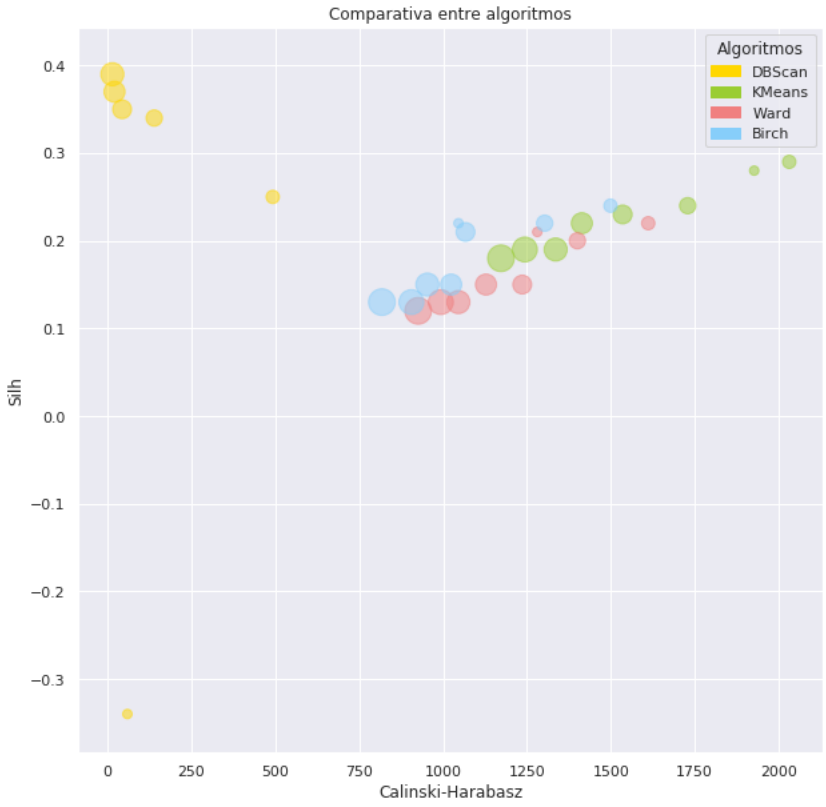

Figura 10: Caso 3: Comparativa de parámetros en parejas heterosexuales.

Tabla 5: Caso 3: Resultados de ajuste de parámetros para parejas heterosexuales

Algoritmo	Silhouette	Calinski-Harabasz	Parámetro	N. clusters
DBScan	-0.34	59.19	0.10	13
DBScan	0.25	492.42	0.15	2
DBScan	0.34	139.16	0.20	2
DBScan	0.35	43.38	0.25	2
DBScan	0.37	20.70	0.30	2
DBScan	0.39	14.29	0.35	2
KMeans	0.28	1927.36	2.00	2
KMeans	0.29	2031.81	3.00	3
KMeans	0.25	1734.67	4.00	4
KMeans	0.22	1535.91	5.00	5
KMeans	0.20	1409.53	6.00	6
KMeans	0.19	1335.62	7.00	7
KMeans	0.19	1235.59	8.00	8
KMeans	0.18	1168.57	9.00	9
Ward	0.21	1280.85	2.00	2
Ward	0.22	1611.55	3.00	3
Ward	0.20	1400.32	4.00	4
Ward	0.15	1236.06	5.00	5
Ward	0.15	1128.03	6.00	6
Ward	0.13	1045.59	7.00	7
Ward	0.13	993.35	8.00	8
Ward	0.12	925.92	9.00	9
Birch	0.22	1045.99	2.00	2
Birch	0.24	1499.04	3.00	3
Birch	0.22	1302.90	4.00	4
Birch	0.21	1066.88	5.00	5
Birch	0.15	1024.14	6.00	6
Birch	0.15	953.29	7.00	7
Birch	0.13	905.71	8.00	8
Birch	0.13	817.72	9.00	9

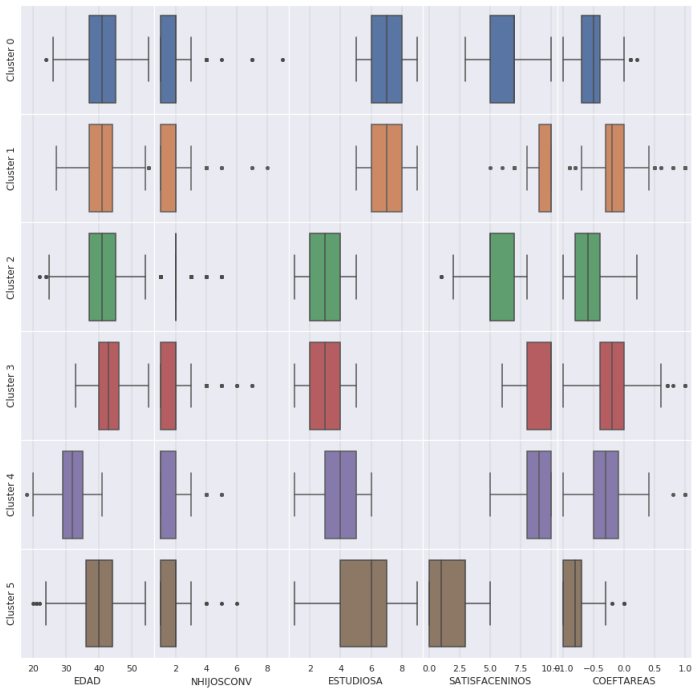

Figura 11: Caso 3: Distribución de variables en *BoxPlot* para cada cluster, para parejas heterosexuales.

2.3.2. Análisis específico: algoritmo *KMeans* en parejas heterosexuales

Para este análisis de los clusters obtenidos por el algoritmo *KMeans* generamos una *ScatterMatrix* (13), un *HeatMap* (12) y las distribuciones de las distintas variables para cada cluster usando *BoxPlot* 11.

En primer lugar nos fijamos en el coeficiente cálculado sobre el reparto de tareas en la pareja (**COEFTAREAS**) y la satisfacción de la respectiva encuestada respecto a dicho reparto (**SATISFACENINOS**). Como es natural podemos observar cierta relación entre las variables. Observamos que en todos los clusters la mujer realiza, en media, más tareas que el hombre. Si bien en algunos clusters esta media es muy cercana a cero y en la figura 11 podemos observar como en algunas parejas si que existe el desbalanceo contrario.

En función a estos dos parámetros podemos clasificar los clusters en tres grupos: clusters con bajo (0, 4 y 5), medio (1 y 3) y gran desbalanceo (2). Estudiemos por separado estos grupos.

En primer lugar, el cluster 2 es el que más desequilibrio presenta en el reparto. De hecho, observando de nuevo la figura 11, presenta valores muy agrupados respecto al coeficiente de tareas. La media de edad se situa entorno a 40 años, conviviendo con casi 1.8 hijos y un nivel de estudios medio/alto. Vale la pena remarcar este último hecho: el cluster donde encontramos más descompensación, lo que podría ser signo de un menor nivel de educación, presenta un nivel de estudios relativamente alto.

Pasamos a estudiar los clusters con un desbalanceo intermedio (1 y 3). Podemos distinguirlos

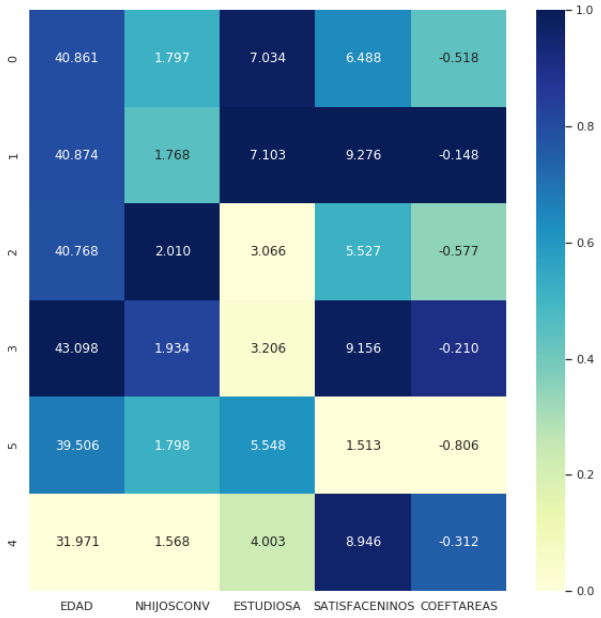

Figura 12: Caso 3: *HeatMap* de centroides para parejas heterosexuales.

entre si debido a la abismal diferencia respecto a los estudios alcanzados. El rango de edad y los hijos apenas varían. De hecho, estas dos variables nos dan información relevante en contadas ocasiones de cara a este caso de estudio. Observando 11, la distribución del número de hijos es practicamente igual en todos los clusters, encontrando más agrupación entorno a dos hijos en el cluster 1.

Estudiando finalmente el último subgrupo con un mayor equilibrio en el reparto de tareas (clusters 0, 4 y 5) llegamos al único cluster, el 4, con una diferencia significativa en la edad de los encuestados, entorno a 32 años. También sostiene la menor media en número de hijos, entorno a 1.5, aunque esta varía menos respecto a los demás clusters. Sin embargo, este es uno de los clusters más pequeños (9), con unicamente un 10.3 % de los encuestados. Es decir, la mayor parte la subpoblación escogida se sitúa entorno a los cuarenta años.

Finalmente diferenciamos los cluster 0 y 5 a partir del nivel de estudios, de igual forma que hicimos con los clusters 1 y 3. Cabe destacar la diferencia de tamaño de los clusters, donde el cluster 0, compuesto por gente con un alto nivel de estudios y un reparto equilibrado, representa el 29 % de la población total mientras que los otros dos clusters con reparto equilibrado suman algo más de un 15 %. Podemos ver como afecta el nivel de estudios a este reparto a partir de estos clusters.

A pesar de ello no he obtenido alguno de los resultados que esperaba. En particular esperaba revelar una diferencia entre el reparto de tareas y la percepción del mismo para valores bajos de nivel de estudios. Desconozco hasta qué punto el orden de las preguntas en la encuesta podría haber sugestionado de alguna forma este resultado, y si la respuesta a la satisfacción sobre el reparto de tareas sería la misma si no le precediesen veinte detalladas preguntas sobre el mismo.

En esta ocasión de la *Scatter Matrix* 13 podemos sacar poca información. Debido a la for-

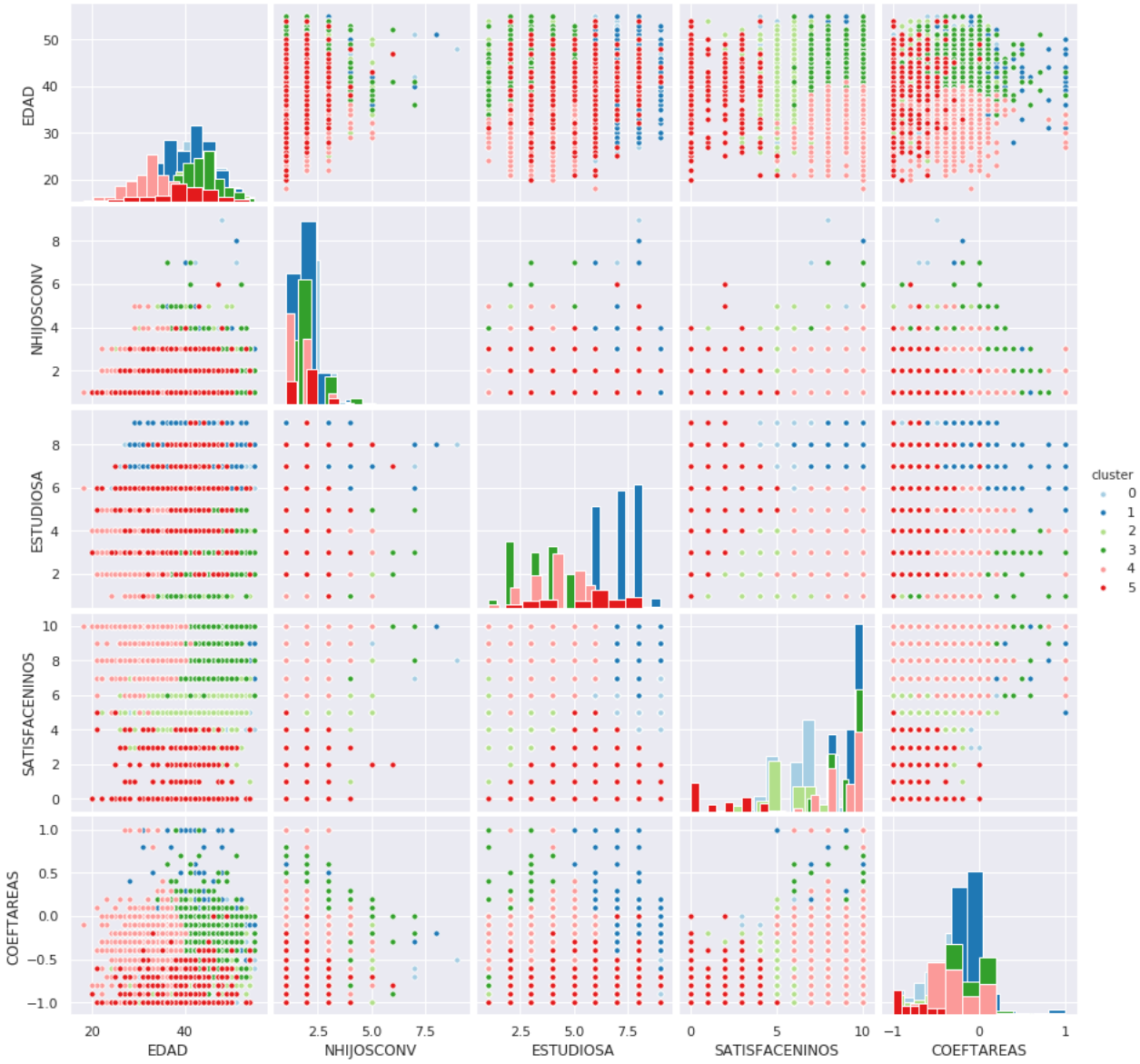

Figura 13: Caso 3: *ScatterMatrix* para parejas heterosexuales.

ma en la que está implementada la librería *Matplotlib*, los clusters con mayor número se pinta encima de aquellos con menor número. Es por ello que en dicha figura se observan muchos más valores de los clusters tres, cuatro y cinco a pesar de ser los más pequeños. Esto significa que cuantos más clusters obtengamos, más complicada será la visualización de los mismo utilizando este tipo de gráfica. La única representación que me gustaría descatar en esta *ScatterMatrix* es la representación de **ESTUDIOSA** frente a **SATISFACENINOS** (lo que sería la posición *(2,3)* en la matriz), donde vemos una clara segmentación de los seis clusters con las dos variables que más relevantes han sido durante todo el análisis.

2.3.3. Comparativa con parejas homosexuales

Realizamos a continuación el mismo análisis que en el apartado anterior para parejas homosexuales. En este caso cambiaremos la forma de calcular el coeficiente de tareas, ya que no tiene sentido estudiar relación Mujer - Hombre en dicho reparto. Para ello realizaremos la misma suma que en el apartado anterior, añadiendo *-0.1* si el encuestado hace la tarea y *0.1* si la hace su pareja. Aplicaremos valor absoluto y obtenemos así un valor entre 0 y 1 donde valores cercanos a 1 representan un desequilibrio en el resparto y valores cercanos a 0, un equilibrio del mismo. De esta forma unicamente tenemos en cuenta el equilibrio en el reparto añadiendo la simetría de este caso de estudio.

Nos restringimos por tanto a los 30 encuestados cuyo sexo coincide con el de su pareja y viven al menos con un hijo menor de catorce años. Como hicimos anteriormente, representaremos los resultados del algoritmo *KMeans* mediante una *ScatterMatrix* (17), un *HeatMap* (16), las distribuciones de las distintas variables para cada cluster usando *BoxPlot* 14 y los tamaños de los distintos clusters 15. Pasemos a comentar los resultados.

En primer lugar, al tratarse de una subpoblación tan sumamente pequeña las conclusiones obtenidas no serán muy representativas de la totalidad población española. Sin embargo, puede servirnos para una comparación con el caso heterosexual. Segmentaremos en cuatro grupos y a pesar del reducido número de objetos obtenemos clusters claramente diferenciados.

Podemos apreciar en la figura 16 como los clusters 1 y 3 presentan un gran equilibrio en el reparto de tareas frente al inmenso desequilibro del cluster 2. De hecho vemos una gran concentración de los valores de este último cluster para las variables **SATISFACENINOS** y **COEFTAREAS**.

Respecto a los clusters 1 y 3, el factor discriminante vuelve a ser el nivel de estudios aunque no de forma tan decisiva como en el caso heterosexual. Ambos clusters se diferencian también en edad y número de hijos.

Finalmente el cluster cero, con un 40 % de la población, aglomera parejas con un reparto relativamente desequilibrado (tomando el coeficiente de tareas valores entre 0.25 y 0.75), pero cuya satisfacción al respecto es notablemente alta, concentrándose la gran mayoría de valores en el 8 (14).

Con tan pocos objetos es complicado realizar una comparación eficaz con la segmentación presentada anteriormente. Aunque se realiza un agrupamiento principalmente basado en el reparto de tareas (como era de esperar, ya que dos de las cinco variables están relacionadas directamente con éste), no se aprecia una clara segunda segmentación basada en el nivel de estudios como ocurría para parejas heterosexuales. Si bien el número de clusters es menor y esto permite en menor medida una segunda segmentación, aumentar el número de clusters de-

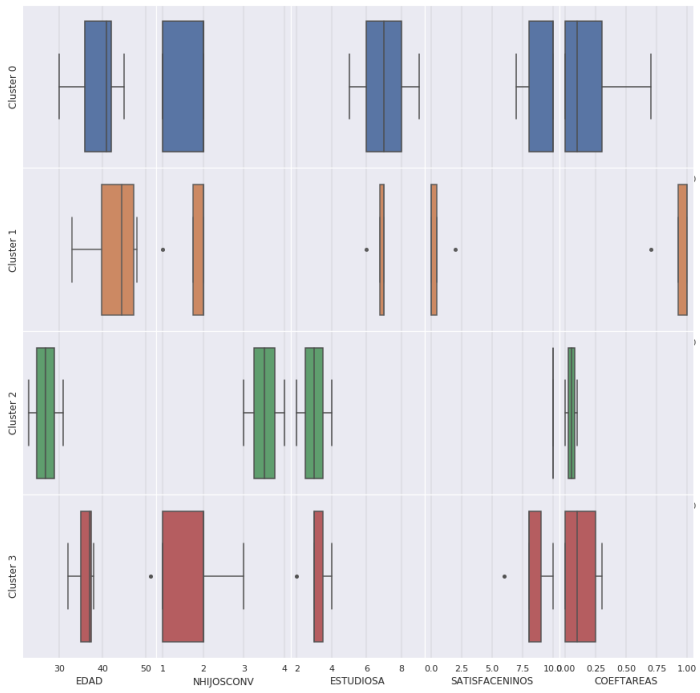

Figura 14: Caso 3: Distribución de variables en *BoxPlot* para cada cluster, para parejas homosexuales.

generaba en grupos demasiado reducidos para ser relevantes.

3. Bibliografía

Referencias

[1] Encuesta realizada por el Instituto Nacional de Estadística (INE) de donde obtenemos los datos. www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177006&menu=ultiDatos&idp=1254735573002

[2] Matplotlib API documentation. matplotlib.org/contents.html

[3] Scikit Sklearn API documentation. scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster

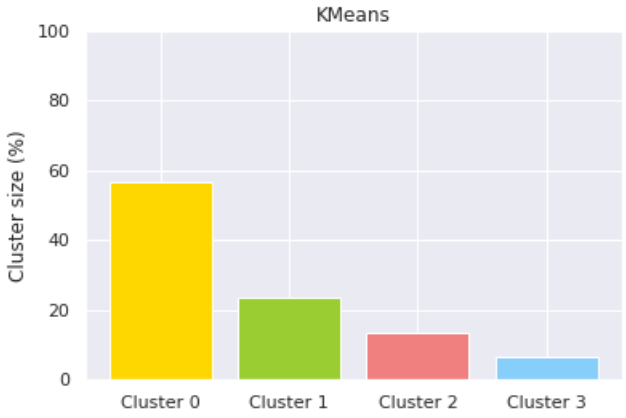

Figura 15: Caso 3: Tamaños de clusters para parejas homosexuales, fijado el algoritmo *KMeans*

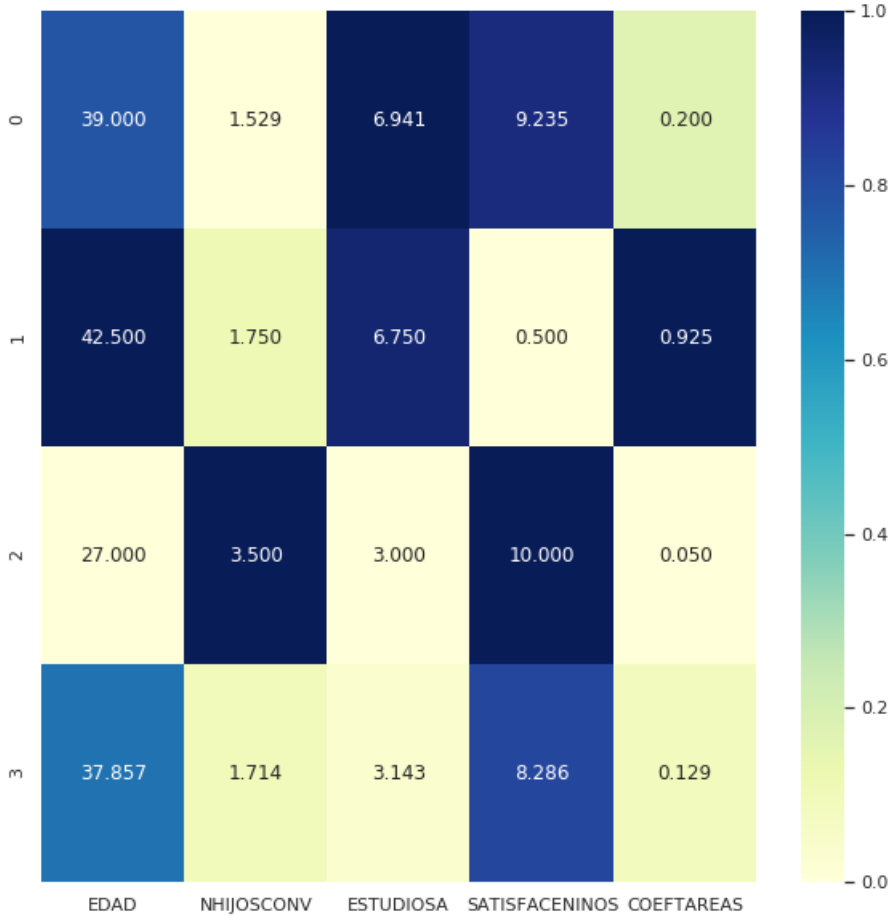

Figura 16: Caso 3: *HeatMap* de centroides para parejas homosexuales.

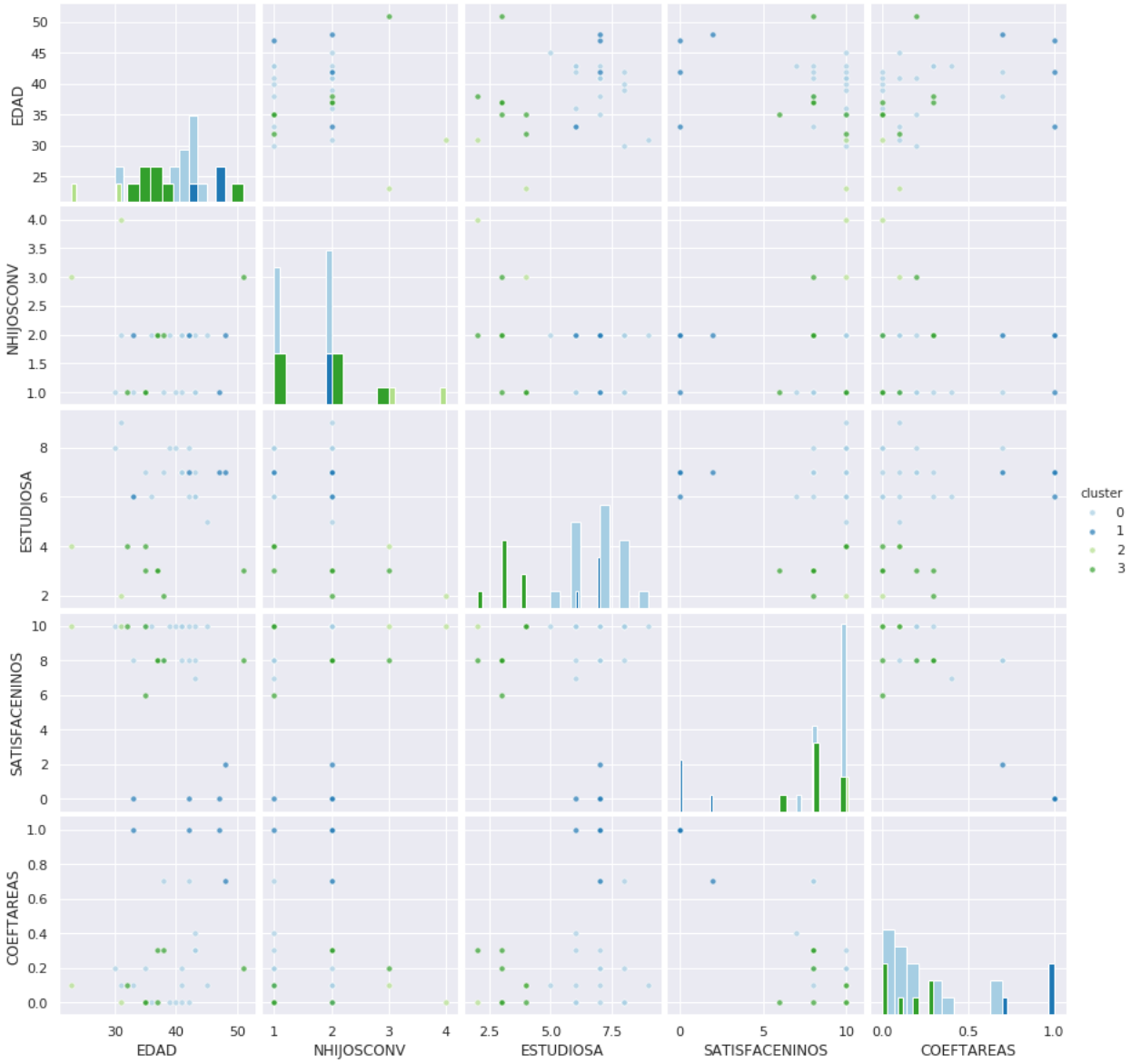

Figura 17: Caso 3: *ScatterMatrix* para parejas homosexuales.