Práctica 2: Clustering

Universidad de Granada

Antonio Manuel Fresneda Rodríguez

antoniomfr@correo.ugr.es

8 de diciembre de 2019

Índice

Introducción

A partir de los microdatos publicados por el Instituto Nacional de Estadística (INE) en 2018 sobre la última encuesta de fecundidad, se dispone de un conjunto de 14.556 respuestas de mujeres a una encuesta con 463 variables sobre datos personales, datos biográficos, hogar, vivienda, padres, relaciones de pareja, hijos, fecundidad, estudios, empleo y creencias. Muchas variables son categóricas como, por ejemplo, estado civil (soltera, casada, viuda, separada o divorciada). Estas variables se usarán para fijar casos de estudio donde centrar el análisis. Hay otras variables numéricas como, por ejemplo, edad.
Finalmente, también hay variables ordinales (por ejemplo, estudios alcanzados). El objetivo de la práctica es definir algunos casos de estudio de interés y aplicar distintos algoritmos de clustering.

Caso de estudio 1

Descripción del caso de uso

El primer caso de estudio se ha centrado en mujeres españolas, casadas con más de 1 hijo y las variables que se han usado para hacer clustering han sido las correspondientes a edad, número de hijos, edad a la que se tuvo el primer hijo, la edad de emancipación y los ingresos. Con esto intentaremos ver si usando estas variables podemos segmentar el conjunto de datos. El número total de datos es de 4059 ejemplos.

Resultados de los algoritmos

A continuación se muestra una tabla con los resultados de los algoritmos y sus parámetros:

Algoritmo	Tiempo(s)	Calinski	Silhouette	Clusters
K-Means	1.45	1578.10	0.26	3
Mean Shift	20.46	93.91	0.14	5
DBSCAN	0.23	29.74	0.4	2
Hierarchical Clustering	0.55	1264.14	0.21	3
BIRCH	0.08	1563.97	0.26	3

Cuadro 1: Tabla de resultados del caso 1

Los parametros que se han usado son:

- K-Means: n_clusters=3, init = k-means++
- Mean Shift: bandwith=0.3
- DBSCAN: eps=0.2
- Hierarchical clustering: n_clusters=3
- BIRCH: threshold=0.3, n_clusters=3

El tamaño de los clusters que se han encontrado son:

- **K-Means**:
  1 1311 (32 %)
  2 1262 (31 %)
  3 1486 (37 %)
- **Mean Shift**
  1 3881 (96 %)
  2 35 (0.9 %)
  3 7 (0.1 %)
  4 57 (1 %)
  5 79 (2 %)
- **DBSCAN**
  1 4032 (99 %)
  2 27 (1 %)
- **Hierarchical-Clustering**
  1 1639 (40 %)
  2 1322 (33 %)
  3 1098 (27 %)
- **BIRCH**
  1 1128 (28 %)
  2 1549 (38 %)
  3 1382 (34 %)

También se han ejecutado algunos algoritmos con distintos parámetros y los resultados han sido los siguientes (aunque los que mejor resultado han tenido son los expuestos arriba):

Algoritmo	Tiempo	Calinski	Silhouette
k-Means {'n_clusters': 4}	0.05	1420.34	0.23
k-Means {'n_clusters': 5}	0.09	1331.95	0.22
k-Means {'n_clusters': 6}	0.10	1245.82	0.22
k-Means {'n_clusters': 7}	0.12	1176.81	0.22
Birch {'n_clusters': 3, 'threshold': 0.25}	0.08	727.80	0.21
Birch {'n_clusters': 3, 'threshold': 0.2}	0.08	606.45	0.20
Birch {'n_clusters': 3, 'threshold': 0.1}	0.18	999.49	0.17
Birch {'n_clusters': 5, 'threshold': 0.25}	0.08	619.55	0.19
Birch {'n_clusters': 5, 'threshold': 0.2}	0.08	642.13	0.15
Birch {'n_clusters': 5, 'threshold': 0.1}	0.17	680.99	0.14

Cuadro 2: Algoritmos con distintos parámetros para caso 1

Como vemos, parece que ha medida que subimos en el número de clusters en k-means estamos empeorando los resultados así que hemos dejado el número de clusters de k-means en 3.
Respecto a BIRCH vemos que el aumento de clusters baja ambas métricas mientras que bajando el parámetro *threshold* notamos una bajada de las métricas y sube cuando este llega a 0.1.
Se han escogido estos dos algoritmos debido a que k-means suele ser el que mejor se ha comportado y BIRCH porque se ejecuta más rápido.
Respecto a los algoritmos DBSCAN y Mean Shift, estos tienen unos parámetros difíciles para hacer que encuentren unos clusters con unos valores para las métricas similares al resto.

Visualizaciones

En esta sección mostramos las gráficas que se han obtenido de los dos algoritmos que mejor han funcionado. El resto de gráficas están en la carpeta Ïmagenes".
En el caso de K-Means tenemos tanto un *heatmap* con los centros de los clusters y las gráficas *sparse matrix* para los dos algoritmos.

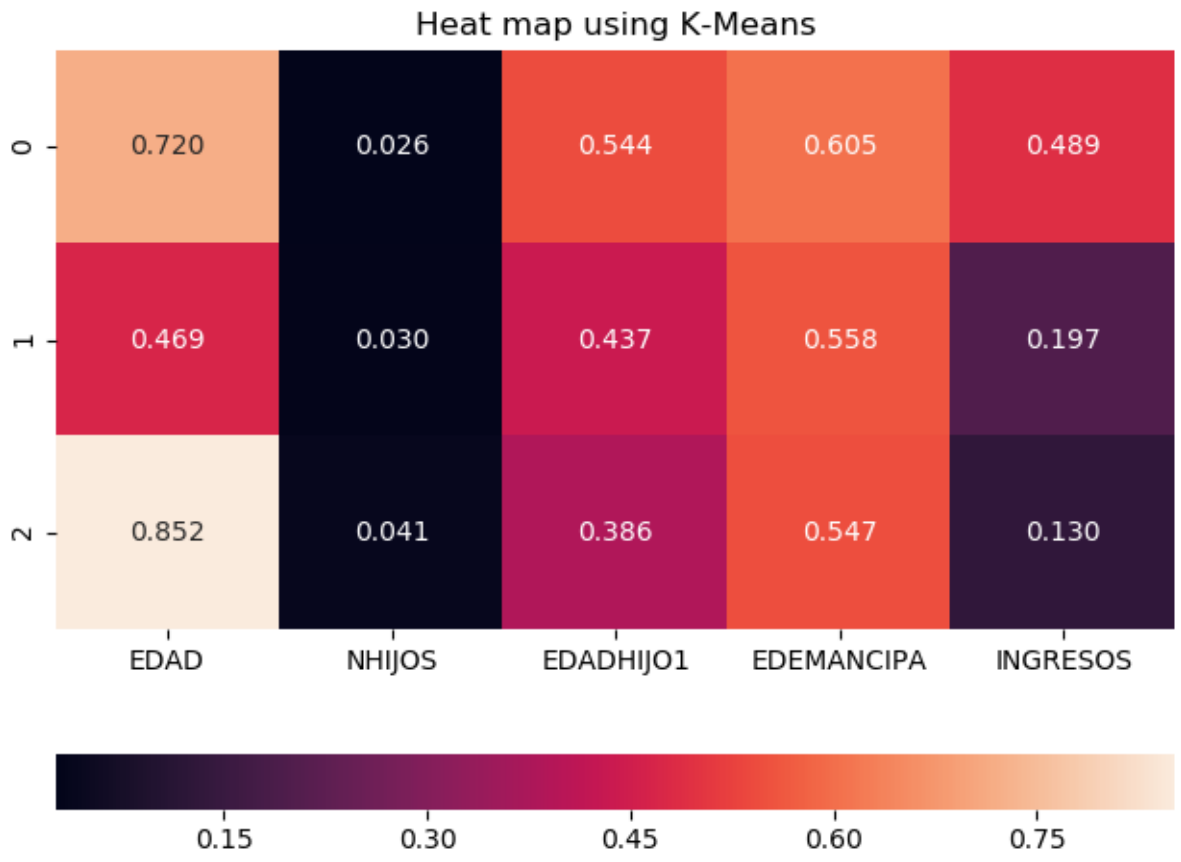

Figura 1: Heatmap de K-Means en caso 1

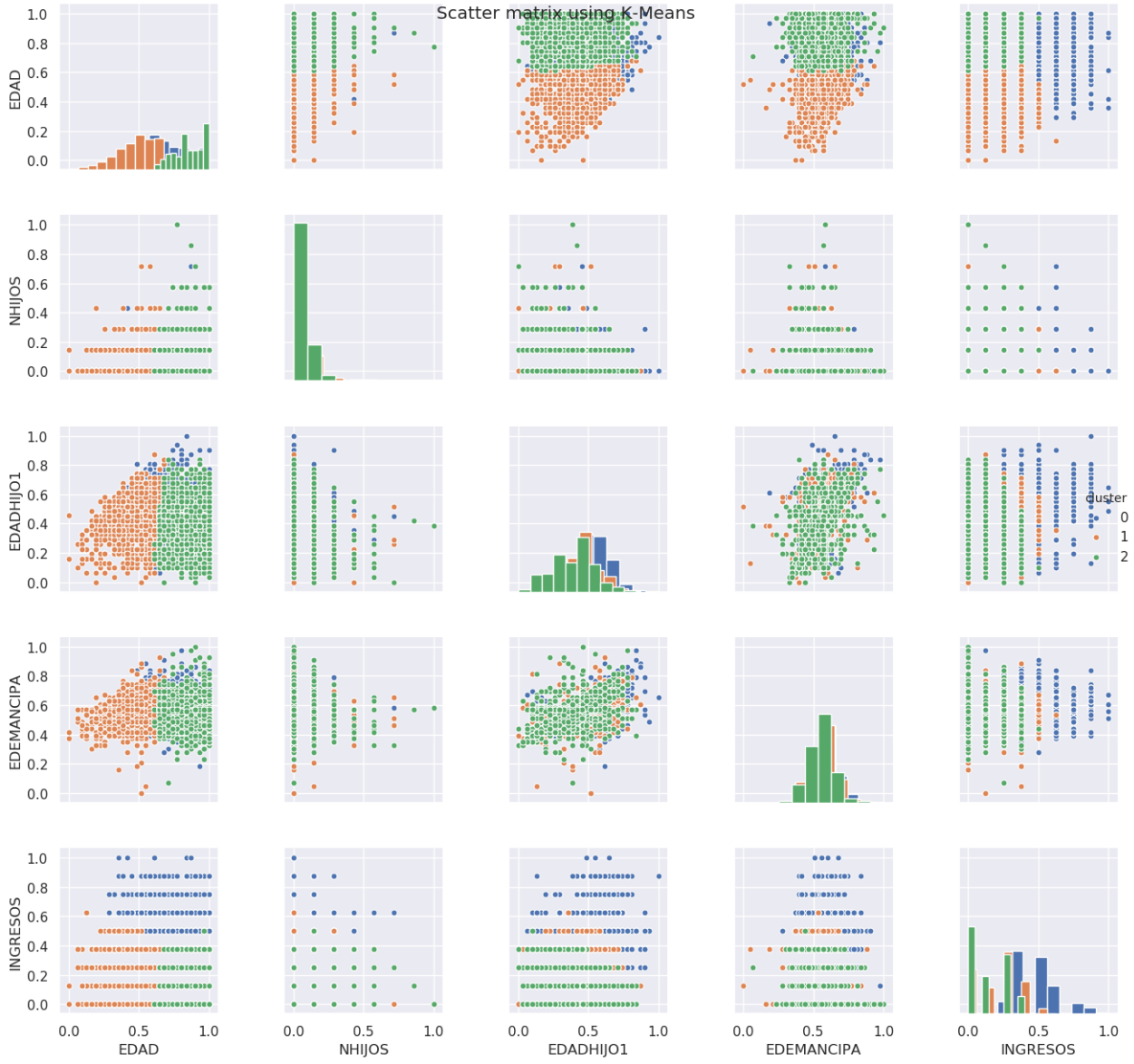

Figura 2: Sparse Matrix de K-Means en caso 1

Figura 3: Sparse Matrix de BIRCH en caso 1

Interpretación

Vamos a comentar la Figura 2. Primero recordamos que el tamaño del cluster 0 (azul) es del 28 %, el cluster 1 (naranja) es del 43 % y el del cluster 2 (verde) es de 29 %. Si nos fijamos en las variables EDAD, INGRESOS vemos como vemos tres clusters bien definidos. Comenzando por el cluster 1, vemos como casi 72 % de las mujeres con más de dos hijos tienen unos ingresos por debajo de la media (dentro del mismo caso) y que solo el 28 % está por encima de esa media. Si ahora nos fijamos en las variables EDADHIJO1 (nos señala la edad a la que tuvieron el primer hijo) e INGRESOS vamos como la mayor parte de este cluster está por encima de la edad media. De aquí podemos inferir que las madres jóvenes van a tener una fuente de ingresos menor que probablemente sea provocado por una menor experiencia.

A continuación comentamos la Figura 3.

El tamaño del cluster 0 (azul) es 28 %, el del cluster 1 (naranja) es de 38 y el del cluster 2 (verde) es de 34 %.
vemos como la divison en los casos que se han comentado anteriormente es parecida, aunque hay casos en los que usando BIRCH se puede sacar algo mas de información:
Si nos fijamos en las variables EDADHIJO1 y EDAD, vemos que el 38 % de las mujeres tiene una edad superior a la media y ha tenido el primer hijo en edades superiores a la media también. El cluster azul se refiere a madres más jóvenes.
También vemos como ese grupo de madres jovenes tiene una edad de emancipación por debajo de la media (mirando la variable EDADEMANCIPA) y tienen como se ha comentado antes, unos ingresos que están por debajo de la media.

Caso de estudio 2

Descripción del caso de uso

En este caso nos hemos centrado en las mujeres que se han intentado quedar embarazadas y se han usado las variables NDESEOHIJO, ESTUDIOSA, TEMPRELA, EDAD, MAMPRIMHIJO; que se refieren a número deseado de hijos, estudios alcanzados, tiempo de relación, edad y edad a la que la madre tuvo su primer hijo.
El total de ejemplos de este caso es de: 668 ejemplos

Resultados de los algoritmos

A continuación se muestra una tabla con los resultados de los algoritmos y sus parámetros: Los parámetros que se han usado son:

Algoritmo	Tiempo	Calinski	Silhouette	Clusters
K-Means	0.02	306.9	0.29	3
Mean Shift	1.49	134.76	0.25	3
DBSCAN	0.11	9.65	-0.3	8
Hierarchical Clustering	0.02	264.45	0.26	4
BIRCH	0.05	244.43	0.24	3

Cuadro 3: Tabla de resultados del caso 2

- K-Means: n_clusters=3, init = k-means++
- Mean Shift: bandwith=0.38
- DBSCAN: eps=0.16
- Hierarchical clustering: n_clusters=4
- BIRCH: threshold=0.1, n_clusters=3

A continuación se exponen el tamaño de los clusters encontrados:

- **K-Means**:
  1 190 (28 %)
  2 285 (43 %)
  3 193 (29 %)
- **Mean Shift**
  1º 334 (50 %)
  2º 330 (49 %)
  3º 4 (1 %)
- **DBSCAN**
  1º 245 (37 %)
  2º 300 (45 %)
  3º 25 (4 %)
  4º 72 (11 %)
  5º 11 (2 %)
  6º 5 (1 %)
  7º 5 (1 %)
  8º 5 (1 %)
- **Hierarchical-Clustering**
  1º 280 (42 %)
  2º 136 (20 %)
  3º 155 (23 %)
  4º 97 (15 %)
- **BIRCH**
  1. 349 (52 %)
  2. 208 (31 %)
  3. 111 (17 %)

En esta tabla tenemos los distintos parámetros que se han usado en clustering jerárquico y BIRCH.

Algoritmo	Tiempo	Calinski	Silhouette
Hierarchical-Clustering{'n_clusters': 3}	0.02	251.81	0.25
Hierarchical-Clustering{'n_clusters': 5}	0.02	229.71	0.20
Hierarchical-Clustering{'n_clusters': 6}	0.02	204.28	0.18
Hierarchical-Clustering{'n_clusters': 7}	0.02	186.40	0.16
BIRCH{'n_clusters': 3, 'threshold': 0.25}	0.02	205.58	0.23
BIRCH{'n_clusters': 3, 'threshold': 0.2}	0.03	218.02	0.23
BIRCH{'n_clusters': 3, 'threshold': 0.15}	0.05	245.40	0.26
BIRCH{'n_clusters': 5, 'threshold': 0.25}	0.02	201.46	0.21
BIRCH{'n_clusters': 5, 'threshold': 0.2}	0.03	222.21	0.22
BIRCH{'n_clusters': 5, 'threshold': 0.1}	0.04	202.87	0.17

En clustering jerárquico vemos como a medida que subimos el número de clusters las métricas bajan.
En BIRCH vemos como al subir el número de clusters no ha beneficiado a las métricas. En cambio la disminución del *threshold* si ha aumentado el valor de las mismas

Visualizaciones

En esta sección mostramos las gráficas que se han obtenido de los dos algoritmos que mejor han funcionado. El resto de gráficas están en la carpeta Ïmagenes".
En el caso de K-Means tenemos tanto un *heatmap* con los centros de los clusters y las gráficas *sparse matrix* para los dos algoritmos.

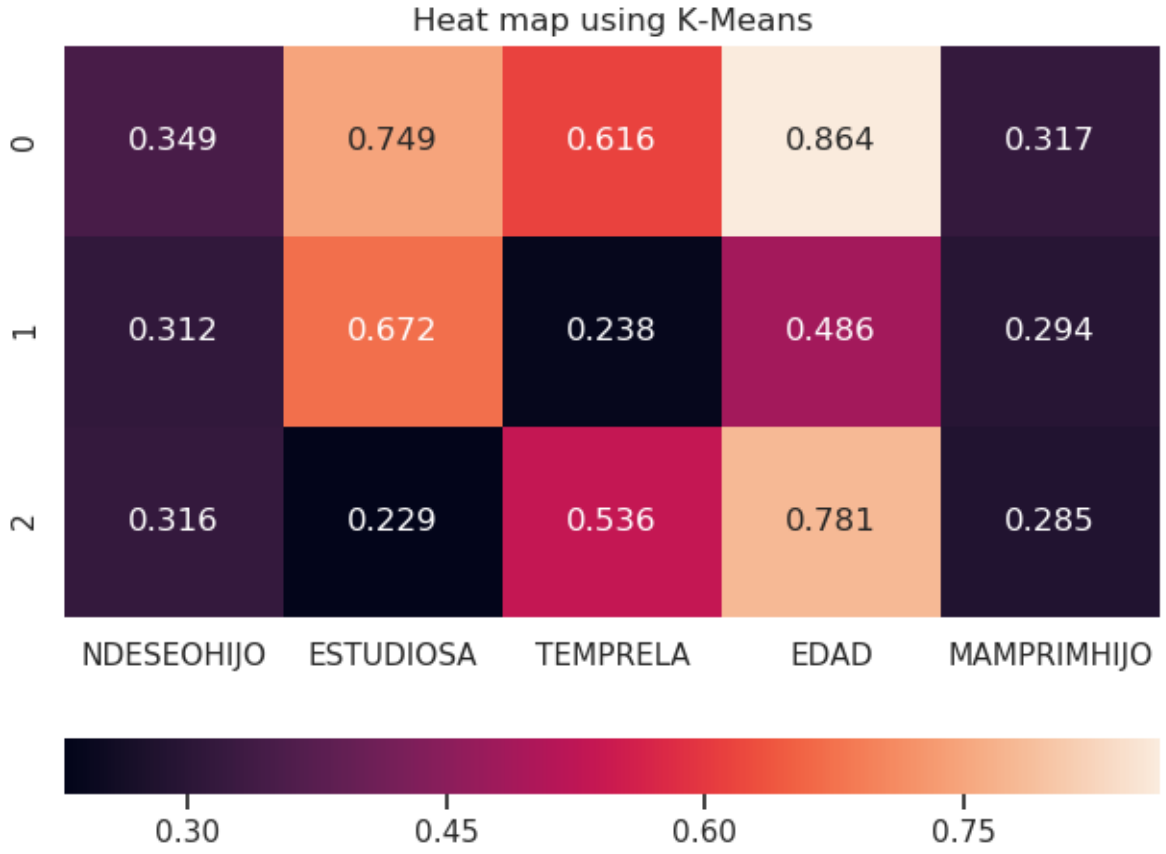

Figura 4: Heatmap de K-Means en caso 2

Figura 5: Sparse Matrix de K-Means en caso 2

Figura 6: Sparse Matrix de Hierarchical en caso 2

Interpretación

Comenzamos analizando la Figura 5. Recordamos que el tamaño de los clusters es: cluster 1 28 % (azul), cluster 2 43 % (naranja) y cluster 3 29 % (verde). Primero nos vamos a fijar en ESTUDIOSA y EDAD. Vemos como un 28 % de las mujeres que han intentado quedarse embarazadas que han alcanzado unos estudios altos tienen una edad mucho mayor que el resto y que alrededor del 29 % se han quedado por debajo de la media en los estudios alcanzados. También vemos como el 43 % de mujeres han llegado a unos estudios superiores a la media tienen relaciones de duración más corta.

Si nos fijamos en las variables ESTUDIOSA y TEMPRELA vemos como el 28 % de mujeres que han intentado quedarse embarazados que han han tenido un tiempo de relación largo tienen unos estudios mayores a la media.
También vemos como a medida que aumentamos en la edad, el tiempo de duración de la relación aumenta.

Caso de estudio 3

Descripción del caso de uso

En este último caso de uso nos vamos a centrar en mujeres menores a 30 años que no hayan tenido hijos y vamos a usar las variables NHERM, MAMPRIMHIJO, NDESEOHIJO, EDAD, ESTUDIOSA; que se refieren a número de hermanos, edad a la que la madre tuvo el primer hijo, número de hijos deseados, edad y estudios alcanzados.
El número total de datos es de: 2861

Resultados de los algoritmos

A continuación se muestra una tabla con los resultados de los algoritmos y sus parámetros: Los parámetros que se han usado son:

Algoritmo	Tiempo	Calinski	Silhouette	Clusters
K-Means	0.03	1570.29	0.25	3
Mean Shift	11	36.53	0.25	3
DBSCAN	0.12	34.55	-0.3	3
Hierarchical Clustering	0.27	1367.62	0.24	3
BIRCH	0.07	1222.96	0.27	3

Cuadro 4: Tabla de resultados del caso 3

- K-Means: n_clusters=3, init = k-means++
- Mean Shift: bandwith=0.4
- DBSCAN: eps=0.15
- Hierarchical clustering: n_clusters=3
- BIRCH: threshold=0.2, n_clusters=3

A continuación se exponen el tamaño de los clusters encontrados:

- **K-Means**:
  1 810 (28 %)
  2 922 (32 %)
  3 1129 (39 %)
- **Mean Shift**
  1º 2815 (98 %)
  2º 29 (1 %)
  3º 17 (1 %)
- **DBSCAN**
  1º 150 (5 %)
  2º 2707 (95 %)
  3º 4
- **Hierarchical-Clustering**
  1º 784 (27 %)
  2º 1225 (43 %)
  3º 852 (30 %)
- **BIRCH**
  1. 1080 (38 %)
  2. 264 (9 %)
  3. 1517 (53 %)

En esta tabla tenemos los distintos parámetros que se han usado en K-Means y BIRCH. Vemos como a medida que aumentamos el numero de clusters en ambos algoritmos hemos

Algoritmo	Tiempo	Calinski	Silhouette
k-Means{'n_clusters': 3}	0.04	1570.29	0.25
k-Means{'n_clusters': 5}	0.07	1272.99	0.23
k-Means{'n_clusters': 7}	0.07	1030.14	0.20
k-Means{'n_clusters': 9}	0.10	900.25	0.19
Birch{'n_clusters': 3, 'threshold': 0.2}	0.09	1222.96	0.27
Birch{'n_clusters': 3, 'threshold': 0.25}	0.07	851.15	0.23
Birch{'n_clusters': 3, 'threshold': 0.1}	0.15	1179.86	0.23
Birch{'n_clusters': 5, 'threshold': 0.1}	0.13	897.50	0.20

disminuido el valor de ambas métricas mientras que en BIRCH si disminuimos vemos como también perjudica.

Visualizaciones

En esta sección mostramos las gráficas que se han obtenido de los dos algoritmos que mejor han funcionado. El resto de gráficas están en la carpeta Ïmagenes".
En el caso de K-Means tenemos tanto un *heatmap* con los centros de los clusters y las gráficas *sparse matrix* para los dos algoritmos.

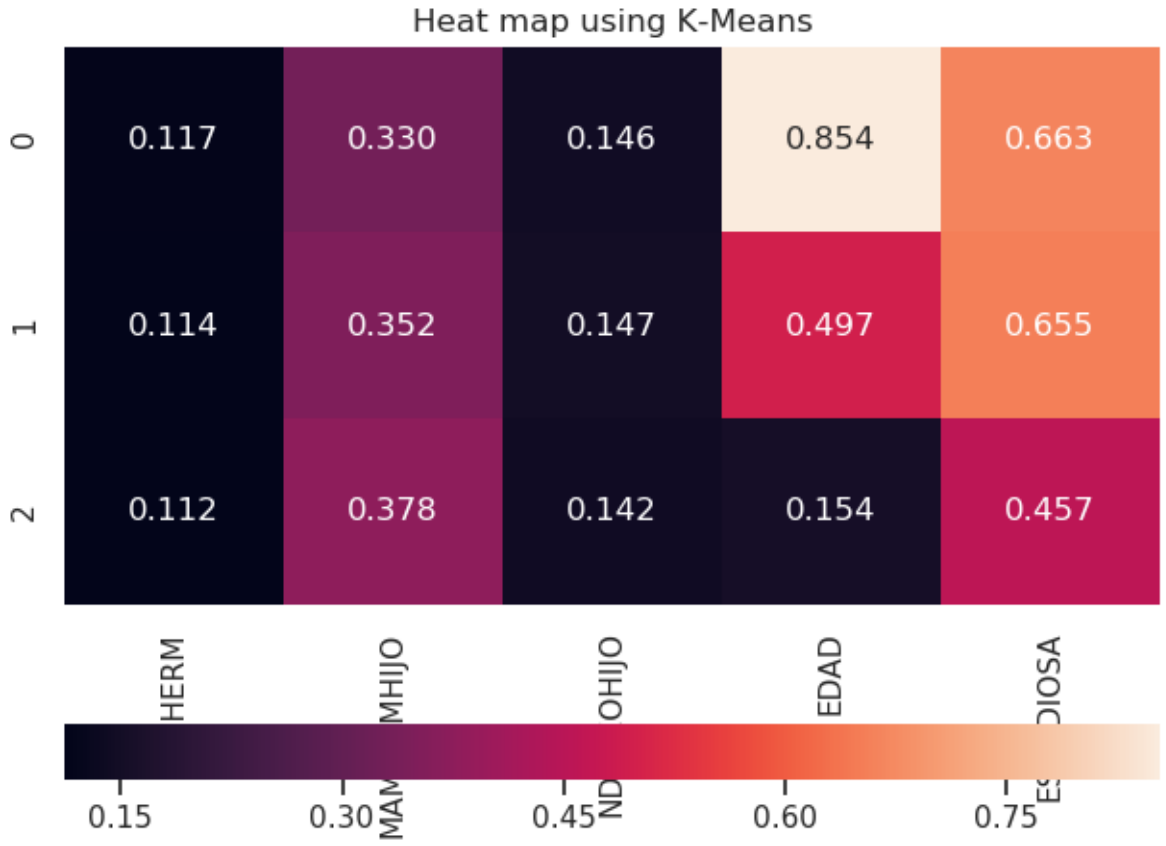

Figura 7: Heatmap de K-Means en caso 3

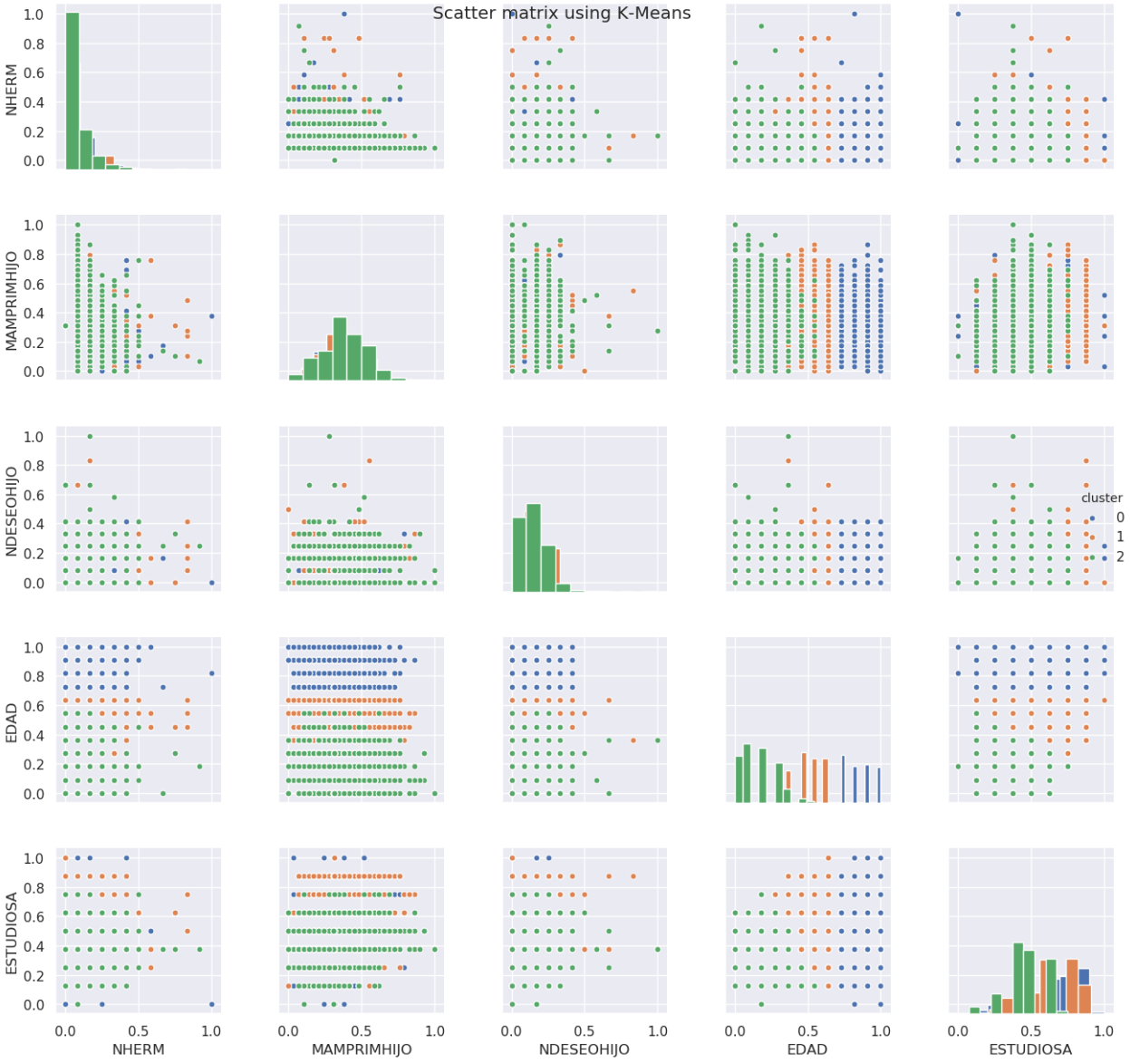

Figura 8: Sparse Matrix de K-Means en caso 3

Figura 9: Sparse Matrix de BIRCH en caso 3

Interpretación

Vamos a comentar la Figura 8. Primero recordamos que el tamaño del cluster 0 (azul) es del 28 %, el cluster 1 (naranja) es del 32 % y el del cluster 2 (verde) es de 39 %. Si vemos las variables NHERM y NDESEOHIJO vemos como tenemos un cluster (2) muy definido que se caracteriza porque tanto número de hermanos como el número de hijos deseado es menor a la media. De aquí podemos deducir que el número de hermanos puede ser un factor determinante para el número de hijos deseados.
También vemos que el numero deseado de hijos dentro de esta sección de la población permanece bajo (por debajo de 0.4) sin importar el rango edad.

Bibliografía

- Web de la asignatura.
- Documentación de Scikit Learn.