UNIVERSIDAD
DE GRANADA

Inteligencia de Negocio
Grado en Ingeniería Informática
Universidad de Granada
Curso 2018-2019

Memoria Práctica 2. Grupo 1
Segmentación para Análisis Empresarial.

Félix Ramírez García
felixramirezgarcia@correo.ugr.es

2 de diciembre de 2018

Índice

Índice de figuras

Índice de tablas

1. Introducción.

Esta practica ha sido llevada a cabo para la asignatura Inteligencia del Negocio de cuarto curso de Ingeniería Informática de la Universidad de Granada. Veremos el uso de algoritmos de aprendizaje no supervisado de agrupamiento para el análisis empresarial.

Vamos a trabajar con el conjunto de datos publicados en el ultimo censo de población realizado por el Instituto Nacional de Estadística (INE) en 2011 [1].

Mediante las distintas variables categóricas (estado civil, sexo, ,etc..) se van a fijar tres casos de estudio donde centrar el análisis.

Para realizar la practica se ha usado el lenguaje de programación Phyton en la versión 3.7 [2] y se han utilizado los siguientes algoritmos de clustering:

K-means [3]
Spectral clustering [4]
Mean Shift [5]
Ward [6]
Birch [7]

El conjunto de datos extraídos del INE cuenta con un total de 83.499 casos ,identificados cada uno por 142 variables , del cual se seleccionan subconjuntos de datos para los casos de estudio que se presentan posteriormente.

En esta practica se abordara el problema haciendo un estudio que se reflejara en varias gráficas y tablas con datos estadísticos la mayor cantidad de información posible.

Por último se extraerán conclusiones finales apropiadas para cada caso de estudio. En particular se usaran un total de 5 algoritmos de clustering, para los cuales se analizaran varias métricas y gráficas para analizar los resultados obtenidos.

2. Casos de estudio.

Dentro de este apartado, se realizaran 3 casos de estudio, en cada uno de ellos se ejecutaran los 5 algoritmos de clustering seleccionados y se calcularan las métricas y gráficas para su posterior análisis.

Las métricas usadas para comparar los datos son: numero de clusters, indice Calinski-Harabaz, la métrica Silhouete y el tiempo de ejecución de cada algoritmo. Las gráficas son ScatterMatrix , HeatMap , Dendograma y HeatMap con Dendograma.

2.1. Primer caso de estudio: Personas que viven con personas mayores de 65 años

En este primer caso de estudio nos vamos a centrar en familias que viven en su hogar con personas mayores , analizaremos el numero de personas dentro del hogar y que rango de edades tienen. Las variables que vamos a utilizar son :

HM5 : Número de personas de 0 a 4 años en el hogar
H0515 : Número de personas de 5 a 15 años en el hogar
H1624 : Número de personas de 16 a 24 años en el hogar
H2534 : Número de personas de 25 a 34 años en el hogar
H3564 : Número de personas de 35 a 64 años en el hogar

En la siguiente tabla se muestran los datos asociados a cada algoritmo usado para este caso de estudio de personas que viven con personas mayores , datos como el numero de clusters que se han usado, la metrica Calinski-harabasz (CH) , la metrica Siljouette (SC) y el tiempo en segundos que ha tardado el algoritmo para ejecutarse . Para este caso de estudio se han contado con un total de 23897 instancias.

Algoritmo	N Clusters	HC metric	SC metric	Time
K-means	4	29697.858169	0.764292	0.100956
Birch	6	21480.901579	0.729611	0.646720
Ward	10	33361.232561	0.845975	19.899369
MeanShift	14	32215.188490	0.855566	0.784660
Spectral	3	26907.715611	0.726861	421.776694

Para la realización de las tablas del estilo de la anterior se ha usado el siguiente código python:

```
def createLatexDataFrame(data):
        my_index = list(dict(data.items()).keys())
        my_data = list(data.values())
        my_cols = list(my_data[0].keys())
        latexDF = pd.DataFrame()

        for row in range(len(my_index)):
                aux = pd.DataFrame(data=my_data[row],index=[my_index[row
                    ]],columns=my_cols)
                latexDF = pd.concat([latexDF,aux])

        return latexDF

for algorithm_name,algorithm in clustering_algorithms:
```

```
    results = dict()
    met, clusterFrame, timeAlg,cluster_predict = createPrediction(
        dataframe=X, data=X_normal, model=algorithm)
    n_clusters=len(set(cluster_predict))

    if( n_clusters > 15 ):
        X_filtrado = clusterFrame[clusterFrame.groupby('cluster').cluster
            .transform(len) > min_size]
    else:
        X_filtrado = clusterFrame

    makeScatterMatrix(data=X_filtrado,outputName="./imagenes/
        scatterMatrix_caso1_" +algorithm_name,displayOutput=False)
    makeHeatmap(data=X_filtrado,outputName="./imagenes/heatmap_caso1_"+
        algorithm_name,displayOutput=False)
    makeDendograma(data=X_filtrado,outputName="./imagenes/
        dendograma_caso1_"+algorithm_name,displayOutput=False)
    makeHeatMapConDendograma(data=X_filtrado,outputName="./imagenes/
        heatmapcondendograma_caso1_"+algorithm_name,displayOutput=False)

    results['N Clusters']=n_clusters
    results['HC metric']=met[0]
    results['SC metric']=met[1]
    results['Time']=timeAlg

    outputData[algorithm_name] = results

latexCaso1 = createLatexDataFrame(data=outputData)
```

En cada caso de estudio nos encontramos con una tabla como la anterior , que nos permitirá valorar los algoritmos y parámetros usados.

En este caso , vemos como para el indice Calinski-Harabasz hay tres algoritmos que están en cabeza (K-means, Ward, MeanShift), ya que el indice CH se define como la razón entre la dispersión interior de los clusters y la dispersión entre clusters , , el objetivo es maximizar este indice. Básicamente viene a reflejar una de las máximas de los algoritmos de clustering , y es que un buen metodo de clustering debe maximizar la similaridad intra-clusters y minimizar la similaridad inter-cluster. De esta forma podemos decir que estos tres son los que mejor comportamiento tienen. Ya que se encuentra muy lejos del algoritmo Spectral.

Para el indice Silhouette no ocurre lo mismo que para el Calinski-Harabasz, ya que todos los resultados se encuentra a la par, teniendo MeanShift la mejor métrica de los 5 algoritmos pero no por mucho. El indice Silhouette mide como de compactos y separados están los clusters, el intervalo de este indice esta entre -1 y 1 , donde los valores cercanos a -1 tienen una mala agrupacion , y los valores cercanos a 1 tienen una buena agrupación.

Quedando los algortimos Ward y MeanShift como los que mejores se comportan podríamos clonar el numero de clusters usados en ese algoritmo para ver si se producen mejoras

en las métricas de los demás algoritmos , esto lo abordaremos mas adelante.

Por ultimo cabe decir que el algoritmo Spectral es que mayor tiempo de ejecución tiene. En las siguientes subsecciones se muestran gráficas y tablas de algoritmos asociadas a cada algoritmo usado para cada caso de estudio , y al final exponemos un análisis de los resultados obtenidos.

Cabe destacar que se ha realizado la eliminación de aquellos clusters con pocos datos (ouliers) . Se ha realizado mediante un filtrado a los clusters con menos de 5 elementos.

2.1.1. Resultados algoritmo K-Means caso 1

El algoritmo K-means es uno de los mas simples de todos los algoritmos de aprendizaje no supervisado. La idea principal es definir K centroides, uno por cada cluster. Esos centroides deben ser colocados de forma astuta, ya que el alojamiento de los K centroides en diferentes posiciones nos dara diferentes resultados. El paso siguiente es colocar cada dato con su centroide mas cercano, en el siguiente punto necesitaremos recalcular los nuevos centroides como baricentros de los cluster resultantes del paso anterior. Esto se repite hasta que no haya cambios.

Ya que los fragmentos de código para generar las gráficas de cada algoritmo son los mismos, solo se van a mostrar en este primer algoritmo k-means.
Para generar la figura 2.1 se ha usado el siguiente fragmento de código:

```
def makeScatterMatrix(data,outputName=None,displayOutput=True):
        sns.set()
        variables = list(data)
        variables.remove('cluster')
        sns_plot = sns.pairplot(data, vars=variables, hue="cluster",
            palette='Paired', plot_kws={"s": 25})
        sns_plot.fig.subplots_adjust(wspace=.03, hspace=.03);

        if displayOutput:
                plt.show()

        if outputName != None:
                outputName += ".png"
                print(outputName)
                plt.savefig(outputName)
                plt.clf()
```

La figura 2.1 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas (ScatterMatrix).

Para generar la figura 2.2 se ha usado el siguiente fragmento de código (HeatMap):

Figura 2.1: Scatter Matrix usando K-means en el caso de estudio 1.

```
def makeHeatmap(data,displayOutput=True,outputName=None):
        meanDF, stdDF = createMeanClusterDF(dataFrame=data)
        meanDF = createNormalizedDF(dataFrame=meanDF)
        anotations = True
        sns.heatmap(data=meanDF, linewidths=.1, cmap='Blues_r', annot=
            anotations, xticklabels='auto')
        plt.xticks(rotation=0)

        if displayOutput:
                plt.show()

        if outputName != None:
                outputName += '.png'
                print(outputName)
                plt.savefig(outputName)
```

```
                plt.clf()
```

La figura 2.2 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo k-means para el caso 1. (HeatMap)

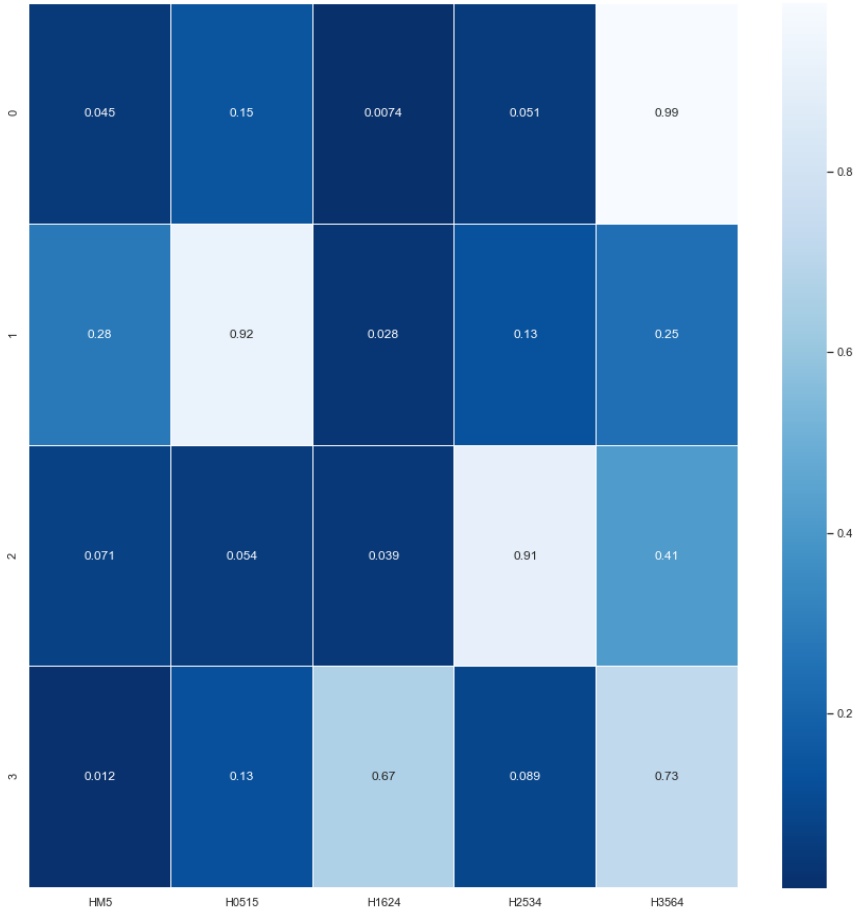

Figura 2.2: HeatMap usando K-means en el caso de estudio 1.

Para obtener los valores promedios de todas las variables sobre cada cluster, así como sus desviaciones típicas se utiliza el siguiente código:

```
def calculateMeanDictionary(cluster,cluster_col = 'cluster'):
        vars = list(cluster)
        vars.remove(cluster_col)
        return dict(np.mean(cluster[vars],axis=0))

def calculateDeviationDictionary(cluster, cluster_col = 'cluster'):
```

```
        vars = list(cluster)
        vars.remove(cluster_col)
        return dict(np.std(cluster[vars],axis=0))

def createMeanClusterDF(dataFrame, clusterCol = 'cluster'):
        n_clusters = list(set(dataFrame[clusterCol]))

        my_mean_df = pd.DataFrame()
        my_deviation_df = pd.DataFrame()

        for cluster_n in n_clusters:
                my_cluster = dataFrame[dataFrame[clusterCol] == cluster_n
                    ]
                meanDic = calculateMeanDictionary(cluster=my_cluster,
                    cluster_col = clusterCol)
                deviationDic = calculateDeviationDictionary(cluster=
                    my_cluster, cluster_col = clusterCol)
                stdDF = pd.DataFrame(deviationDic, index=[str(cluster_n)
                    ])
                auxDF = pd.DataFrame(meanDic,index=[str(cluster_n)])
                my_mean_df = pd.concat([my_mean_df,auxDF])
                my_deviation_df = pd.concat([my_deviation_df,stdDF])

        return [my_mean_df, my_deviation_df]
```

Para generar la figura 2.3 se ha usado el siguiente fragmento de código:

```
def makeDendograma(data, displayOutput=True,outputName=None):
        meanDF, stdDF = createMeanClusterDF(dataFrame=data)
        linkage_array = hierarchy.ward(meanDF)
        plt.figure()
        plt.clf()
        hierarchy.dendrogram(linkage_array)

        if displayOutput:
                plt.show()

        if outputName != None:
                outputName += '.png'
                print(outputName)
                plt.savefig(outputName)
                plt.clf()
```

La figura 2.3 es un dendograma , puede ayudar a decidir el numero de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en la que se van anidando los clusters y la medida de similitud a la cual lo hacen.

A continuación se muestra la figura 2.4 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

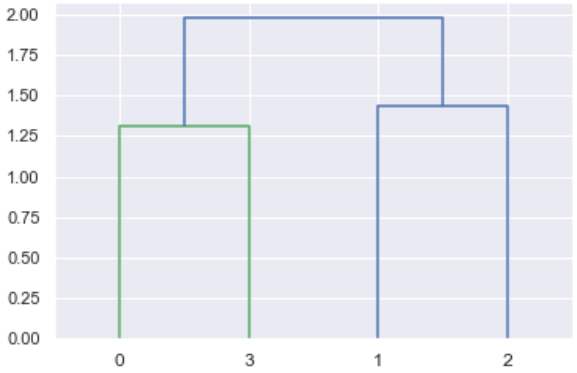

Figura 2.3: Dendograma usando K-means en el caso de estudio 1.

La figura 2.5 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

2.1.2. Resultados algoritmo Birch caso 1

Birch es un algoritmo de aprendizaje no supervisado usado para clustering jerárquico sobre grandes cantidades de datos. Una de sus principales ventajas es la capacidad para agrupar incremental y dinámicamente los clusters. En la mayoría de los casos Birch solo necesita un único escaneo de los datos.

La figura 2.6 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas. (ScatterMatrix)

La figura 2.7 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo Birch para el caso 1. (HeatMap)

La figura 2.8 es un dendograma , puede ayudar a decidir el numero de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en la que se van anidando los clusters y la medida de similitud a la cual lo hacen.

A continuación se muestra la figura 2.9 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.10 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

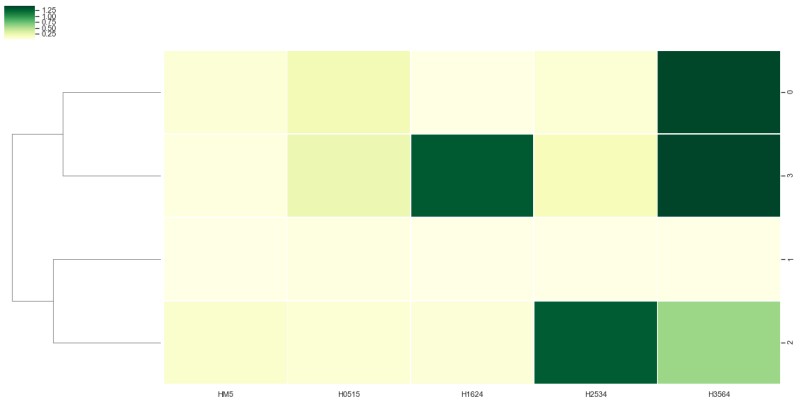

Figura 2.4: HeatMap con Dendograma usando K-means en el caso de estudio 1.

	HM5	H0515	H1624	H2534	H3564
0	0.065113	0.211557	0.010572	0.072441	1.413503
1	0.005842	0.019208	0.000594	0.002772	0.005149
2	0.101469	0.077895	0.056713	1.303041	0.594465
3	0.023174	0.248233	1.313826	0.175177	1.421838

Figura 2.5: Medias de los datos seleccionados por cluster K-means.

2.1.3. Resultados algoritmo Ward caso 1

El método de Ward es un procedimiento jerárquico en el cual, en cada etapa, se unen los dos clusters para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del cluster.

La figura 2.11 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas. (ScatterMatrix)

La figura 2.12 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo Ward para el caso 1. (HeatMap)

La figura 2.13 es un dendograma , puede ayudar a decidir el numero de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en la que se van anidando los clusters y la medida de similitud a la cual lo hacen.

A continuación se muestra la figura 2.14 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

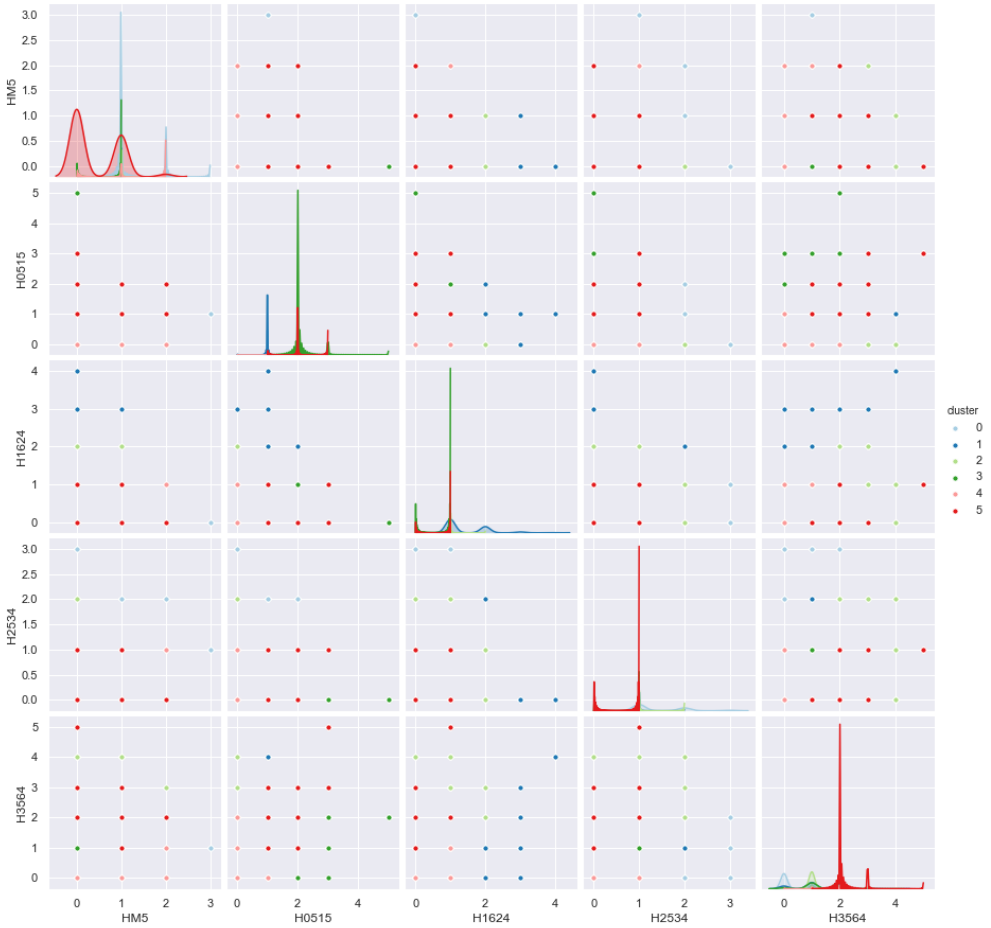

Figura 2.6: Scatter Matrix usando Birch en el caso de estudio 1.

La figura 2.15 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

2.1.4. Resultados algoritmo MeanShift caso 1

El algoritmo Mean Shift es una técnica de clustering no paramétrica que no requiere conocimiento del numero de clusters. Dado N puntos de datos, en un espacio d-dimensional, el núcleo de densidad miltivariado se estima obtener con K(x) [5]

La figura 2.16 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas. (ScatterMatrix)

La figura 2.17 representa la media normalizada de los datos totales de cada variable aso-

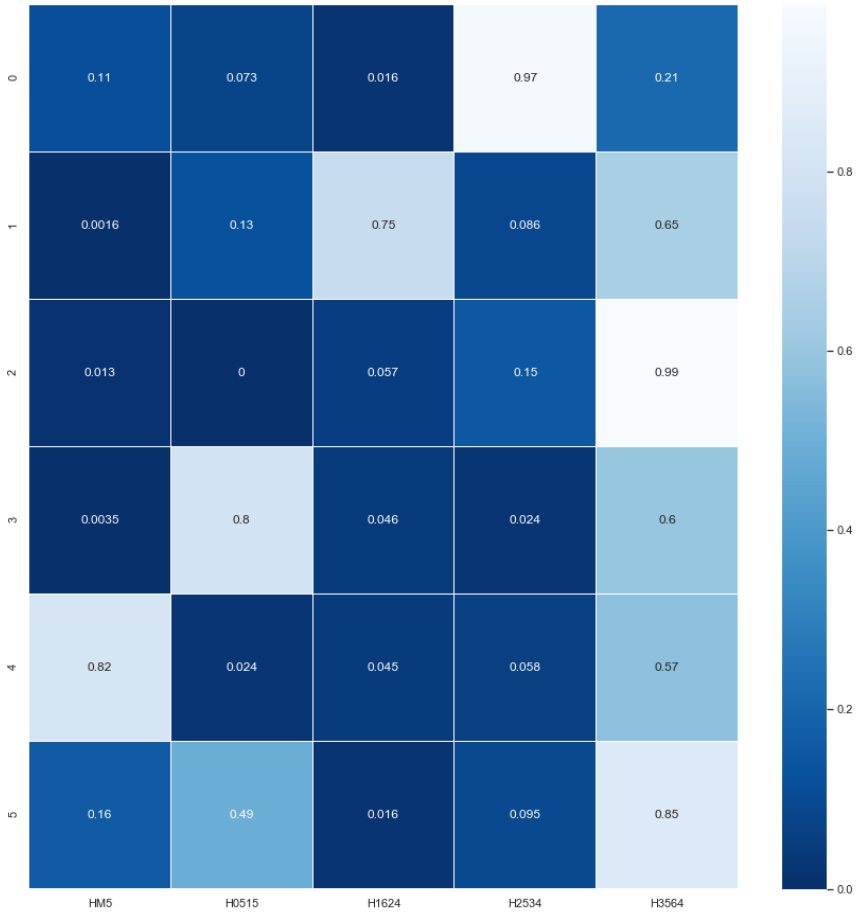

Figura 2.7: HeatMap usando Birch en el caso de estudio 1.

ciados a cada cluster usando el algoritmo MeanShift para el caso 1. (HeatMap)

La figura 2.18 es un dendograma , puede ayudar a decidir el numero de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en la que se van anidando los clusters y la medida de similitud a la cual lo hacen.

A continuación se muestra la figura 2.19 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.20 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

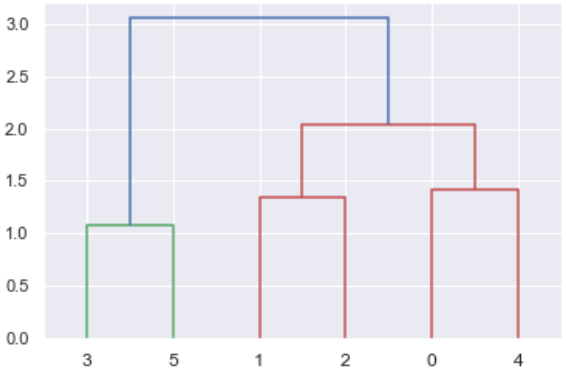

Figura 2.8: Dendograma usando Birch en el caso de estudio 1.

2.1.5. Resultados algoritmo Spectral caso 1

las técnicas de agrupamiento espectral hacen uso del espectro (valores propios) de la matriz [similitud] de los datos para realizar reducción de dimensionalidad antes de la agrupación en un menor número de dimensiones. La matriz de similitud se proporciona como una entrada y consta de una evaluación cuantitativa de la similitud relativa de cada par de puntos en el conjunto de datos. [4]

La figura 2.21 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas. (Scatter Matrix)

La figura 2.22 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo spectral para el caso 1. (HeatMap)

La figura 2.23 es un dendograma , puede ayudar a decidir el numero de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en la que se van anidando los clusters y la medida de similitud a la cual lo hacen.

A continuación se muestra la figura 2.24 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.25 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

2.1.6. Algoritmos modificados en el caso 1

En esta sección se va a exponer la modificación de los parámetros de dos algoritmos distintos y para ver sus diferencias se van a comparar los resultados de las métricas ob-

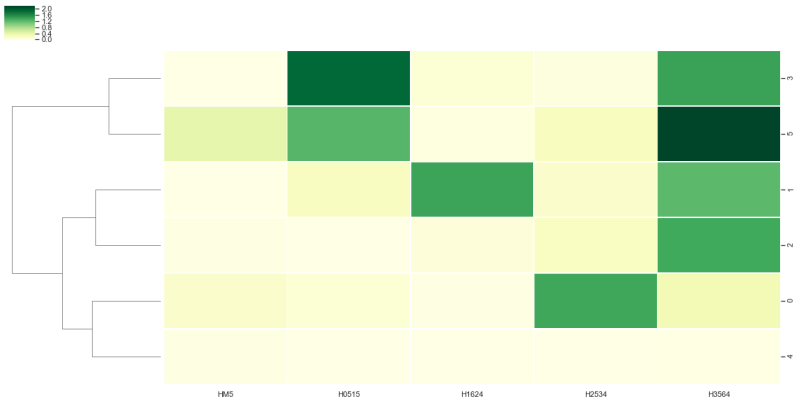

Figura 2.9: HeatMap con Dendograma usando Birch en el caso de estudio 1.

	HM5	H0515	H1624	H2534	H3564
0	0.160521	0.104664	0.022777	1.381779	0.304230
1	0.003027	0.234612	1.396569	0.160949	1.216448
2	0.017845	0.000000	0.078960	0.211920	1.360975
3	0.008152	1.879076	0.108696	0.057065	1.414402
4	0.023460	0.000690	0.001281	0.001676	0.016363
5	0.418831	1.245130	0.040584	0.240260	2.159091

Figura 2.10: Medias de los datos seleccionados por cluster Birch.

tenidas en las secciones previas.

El primer algoritmo que vamos a modificar por su pésima métrica de CH es Birch , intentando aumentar su valor disminuido el numero de clusters a 4 y el segundo algoritmo que vamos a modificar es Ward, y vamos a incrementar su numero de clusters a 35 para ver que resultados obtenemos y compararlos con la anterior ejecución.

La figura 2.26 muestra la antigua tabla pero ahora con las métricas de la ejecución de estos dos algoritmos modificados. En ella se aprecia que las modificaciones de los parámetros han sido exitosas en el caso de Ward y fatales en el caso del algoritmo Birch. En el algoritmo Birch modificado se han disminuido bastante la métrica CH y la métrica SH .En el algoritmo Ward modificado se han incrementado muchísimo las métricas CH Y SH al aumentar el numero de clusters. Por lo que podemos predecir que aumentando el numero de clusters del resto de algoritmo se obtendrían mejores resultados.

2.1.7. Algoritmo modificado Birch caso 1

La figura 2.27 representa como están distribuidos los diferentes clusters sobre las diferentes variables con el algoritmo Birch. (ScatterMatrix)

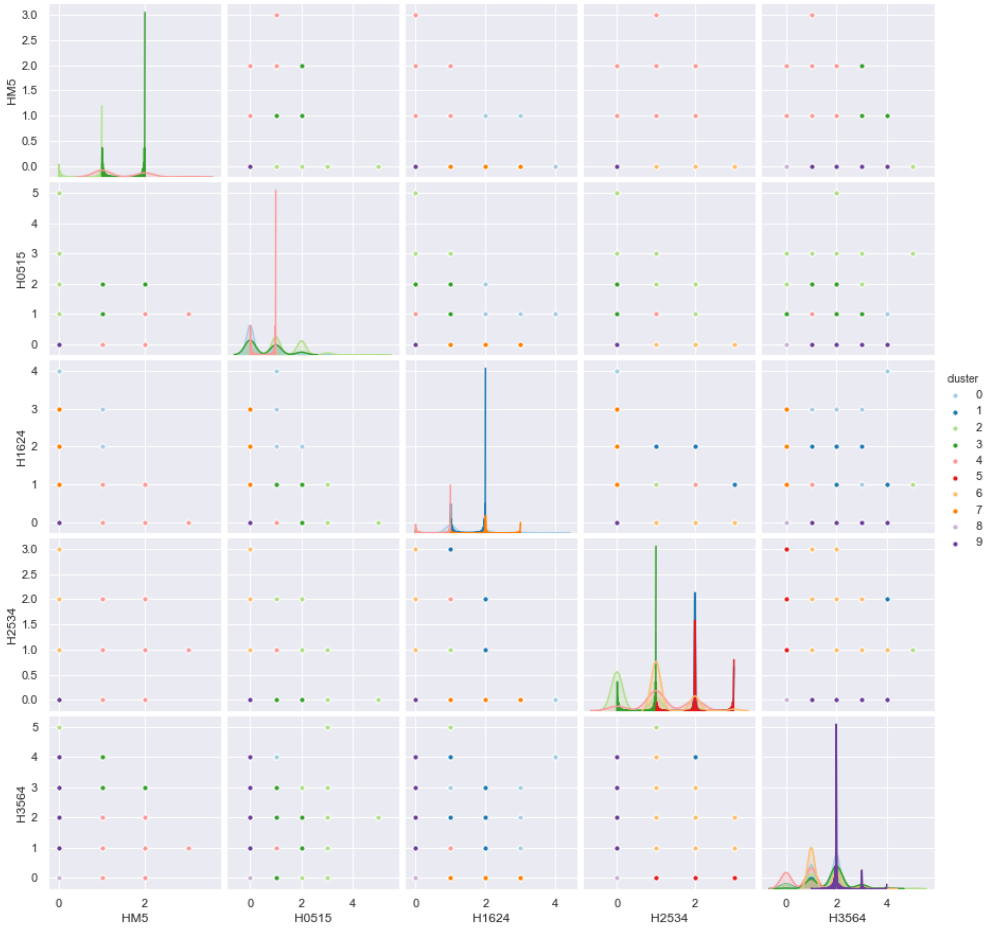

Figura 2.11: Scatter Matrix usando Ward en el caso de estudio 1.

A continuación se muestra la figura 2.28 , esta es la fusión de las gráficas de Heatmap y de Dendograma con el algoritmo Birch modificado.

2.1.8. Algoritmo modificado Ward caso 1

La figura 2.29 representa como están distribuidos los diferentes clusters sobre las diferentes variables con el algoritmo Ward. (ScatterMatrix)

A continuación se muestra la figura 2.30 , esta es la fusión de las gráficas de Heatmap y de Dendograma con el algoritmo Ward modificado.

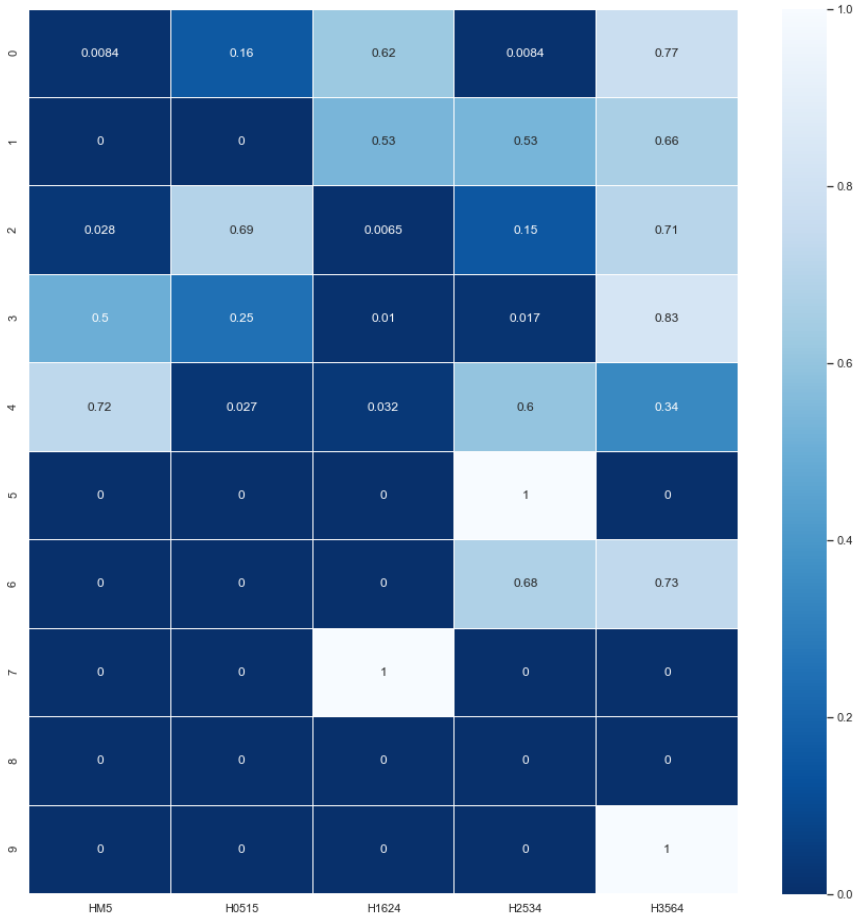

Figura 2.12: HeatMap usando Ward en el caso de estudio 1.

2.1.9. Interpretacion de la segmentacion caso 1

Para terminar con el estudio de este caso , vamos a interpretar las visualizaciones producidas por las ejecuciones de los algoritmos, recalcando de que los datos obtenidos son los de las personas con uno o mas miembros mayores de 64 años en sus familias, incluyéndose a estas.

Para el algoritmo K-Means podemos apreciar que en el cluster cero se agrupan aquellas familias con dos o mas miembros entre 35 y 64 , con 3 miembros o menos entre 5 y 15 años y en algunos casos con miembros entre 16 y 24 o menores de 15. En el cluster uno se agrupan aquellas familias con 3 miembros o mas entre 5 y 15 años , con tres miembros menores de 5 años , y uno o dos miembros entre 35 y 64 años. Puntualmente también hay miembros entre 16 y 34 años. En el cluster dos se agrupan aquellas familias con dos miembros entre 25 y 34 y varios miembros menores de 15 años y ningun miembro en el

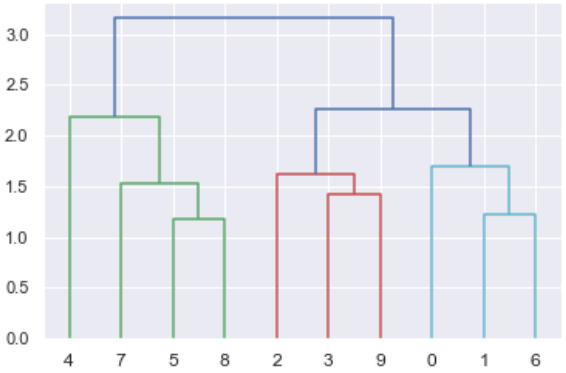

Figura 2.13: Dendograma usando Ward en el caso de estudio 1.

rango 35-64, por lo que podríamos decir que son parejas de jóvenes (24-34) que viven con sus padres y con sus hijos. En el cluster 3 se agrupan aquellas familias con algún miembro entre 0 y 15 , un miembro entre 16 y 34 años y con entre cero y dos miembros entre 35 y 64.
Para el algoritmo Birch podemos apreciar que se asemeja bastante al caso anterior de K-means , pero ahora con un cluster mas. En el cluster cero se agrupan aquellas familias con varios miembros entre 25 y 34 años y varios miembros menores de 5 años. En el cluster uno se agrupan aquellas familias con varios miembros entre 35 y 64 y varios miembros entre 16 y 24. En el cluster dos se agrupan aquellas familias con varios miembros entre 35 y 64 y algun miembro entre 25 y 34. En el cluster tres se agrupan aquellas familias varios miembros entre cinco y 15 años y varios miembros entre 35 y 64. En el cluster cuatro tenemos agrupadas a aquellas familias con varios miembros entre 35 y 64 y algun miembro menor de 15 años. En el quinto cluster tenemos aquellas familias con al menos dos menores de cinco años , un miembro entre 25 y 34 y de dos a cuatro miembros entre 35 y 64 , por lo que podríamos decir que en este tipo de familias viven personas de entre todos los rangos de edades.
Para el algoritmo Ward también se pueden apreciar similitudes respecto a los algoritmos comentados anteriormente , pero en este caso tenemos los datos agrupados en 9 clusters . En los clusters de 0 al 4 se han agrupado los datos de forma similar a los algoritmos anteriores, pero en los clusters del quinto al noveno presenciamos un agrupamiento distinto. En el cluster 5 se han agrupado aquellas familias solo con miembros entre 25 y 34 . En el cluster 6 se han agrupado aquellas familias con miembro solo entre 25 y 64. En el cluster 7 se han agrupado aquellas familias solo con miembros entre 16 y 24. En el cluster 8 se han agrupado aquellas personas que son mayores de 64 años. Y por ultimo en el cluster 9 se han agrupado aquellas familias con solo miembros (entre dos y 4) entre 35 y 64 años. Posiblemente familias de hermanos que vivan junto a sus padres.
Para el algoritmo Spectral, se puede apreciar que comparado con el resto de algoritmos los agrupamientos han sido muy heterogéneos , debido a la falta de numero de clusters.

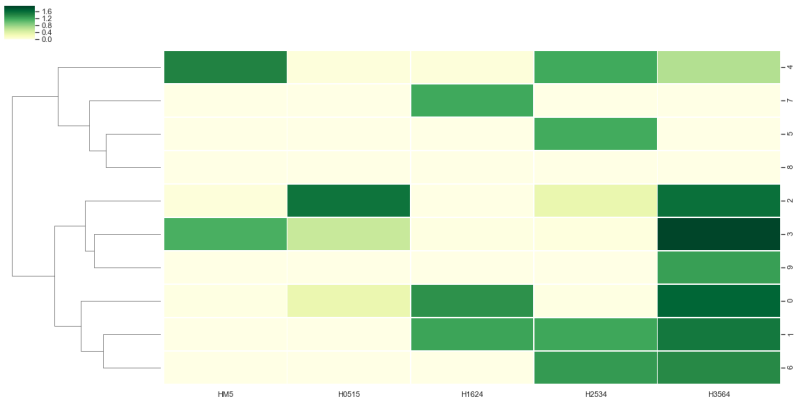

Figura 2.14: HeatMap con Dendograma usando Ward en el caso de estudio 1.

	HM5	H0515	H1624	H2534	H3564
0	0.018066	0.337407	1.333156	0.018066	1.657811
1	0.000000	0.000000	1.214286	1.202091	1.506969
2	0.061897	1.544212	0.014469	0.339228	1.574759
3	1.143508	0.560364	0.022779	0.038724	1.879271
4	1.421488	0.053719	0.061983	1.177686	0.669421
5	0.000000	0.000000	0.000000	1.170669	0.000000
6	0.000000	0.000000	0.000000	1.275629	1.376918
7	0.000000	0.000000	1.182724	0.000000	0.000000
8	0.000000	0.000000	0.000000	0.000000	0.000000
9	0.000000	0.000000	0.000000	0.000000	1.244016

Figura 2.15: Medias de los datos seleccionados por cluster Ward.

En el cluster cero se han agrupado principalmente aquellas familias con varios miembros entre 25 y 34 años. En el cluster uno se han agrupado aquellas familias con varios miembros entre 16 y 2y. En el cluster numero dos se han agrupado aquellas familias con varios miembros entre 35 y 64 . Para el caso del algoritmo Ward modificado , al aumentar el numero de clusters permitimos la segmentación en un mayor numero de grupos , pudiendo así segmentar mejor a los diferentes tipos de familias. Se puede apreciar que en el cluster 17 se agrupan aquellas personas mayores que viven solas. Así como que en el cluster 32 se agrupan aquellas familias con miembros prácticamente en todos los rangos de edades.

2.2. Segundo caso de estudio: Personas solteras que no vivan con personas mayores de 65 ni menores de 15.

En este segundo caso de estudio nos vamos a centrar en personas que no viven en su hogar con personas mayores de 65 ni menores de 15 años, analizaremos la edad de estas personas y el numero de personas y rango de edades dentro del hogar . Las variables que

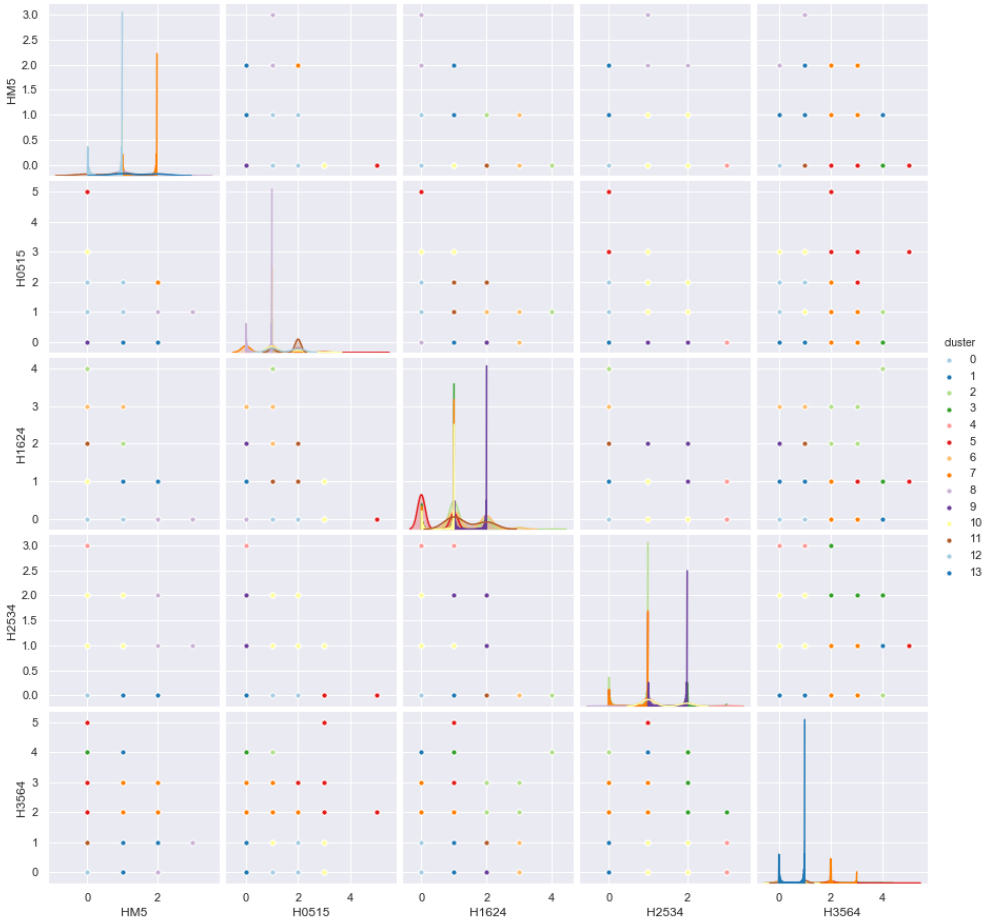

Figura 2.16: Scatter Matrix usando MeanShift en el caso de estudio 1.

vamos a utilizar son :

EDAD : Edad de la persona
TAMNUC : Tamaño del núcleo
H1624 : Número de personas de 16 a 24 años en el hogar
H2534 : Número de personas de 25 a 34 años en el hogar
H3564 : Número de personas de 35 a 64 años en el hogar

En la figura 2.31 se muestran los datos asociados a cada algoritmo usado para este caso de estudio de personas que viven con personas mayores , datos como el numero de clusters que se han usado, la métrica Calinski-harabasz (CH) , la métrica Siljouette (SC) y el tiempo en segundos que ha tardado el algoritmo para ejecutarse . Para este caso de estudio se han contado con un total de 22628 instancias.

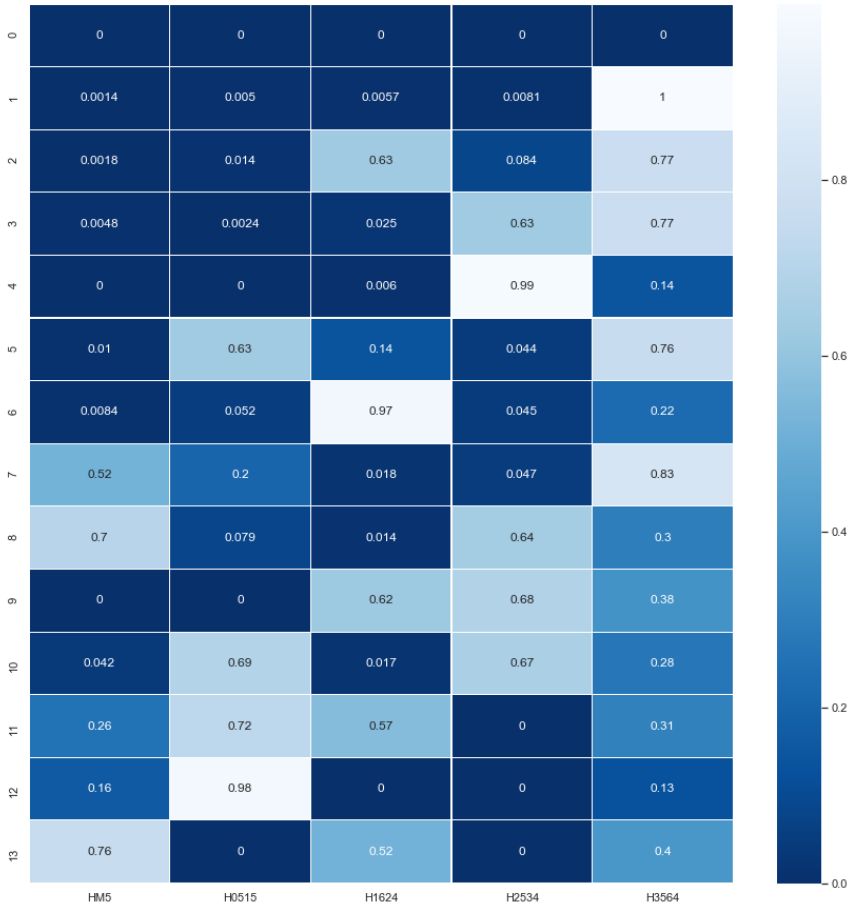

Figura 2.17: HeatMap usando MeanShift en el caso de estudio 1.

En este caso , vemos como para el indice Calinski-Harabasz hay un algoritmo que esta claramente en cabeza, K-means. Por detrás esta el algoritmo Spectal , y le sigue el algoritmo Ward. Por ultimo tenemos los pésimos resultados de los algoritmos Birch y MeanShift.

Para el indice Silhouette no ocurre lo mismo que para el Calinski-Harabasz, ya que todos los resultados se encuentra a la par, teniendo MeanShift y Birch las mejores métricas en este caso.

Quedando los algoritmos K-means y Spectral como los que mejores se comportan podríamos clonar el numero de clusters usados en ese algoritmo para ver si se producen mejoras en las métricas de los demás algoritmos.

Por ultimo cabe decir que el algoritmo Spectral es que mayor tiempo de ejecución tiene. En las siguientes subsecciones se muestran gráficas y tablas de algoritmos asociadas a

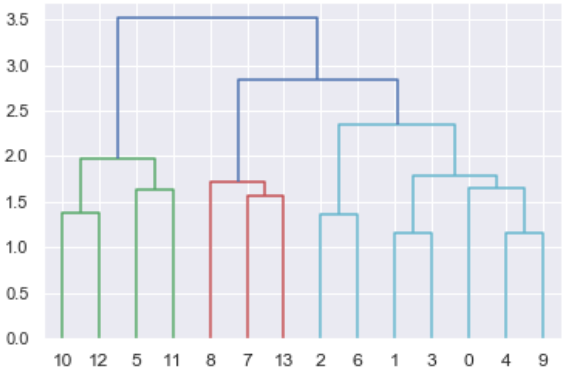

Figura 2.18: Dendograma usando MeanShift en el caso de estudio 1.

cada algoritmo usado para cada caso de estudio , y al final exponemos un análisis de los resultados obtenidos.

Cabe destacar que se ha realizado la eliminación de aquellos clusters con pocos datos (ouliers) . Se ha realizado mediante un filtrado a los clusters con menos de 5 elementos.

2.2.1. Resultados algoritmo K-Means caso 2

La figura 2.32 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas. (Scatter Matrix)

La figura 2.33 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo k-means para el caso 2. (HeatMap)

La figura 2.34 es un dendograma , puede ayudar a decidir el numero de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en la que se van anidando los clusters y la medida de similitud a la cual lo hacen.

A continuación se muestra la figura 2.35 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.36 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

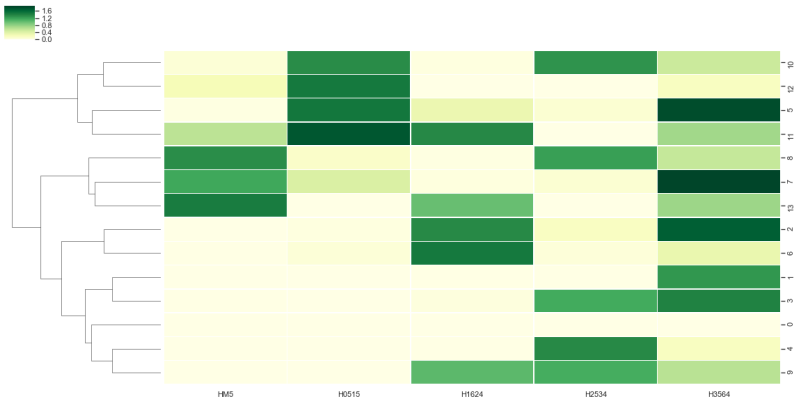

Figura 2.19: HeatMap con Dendograma usando MeanShift en el caso de estudio 1.

	HM5	H0515	H1624	H2534	H3564
0	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.001818	0.006514	0.007423	0.010453	1.291168
2	0.003981	0.030524	1.370936	0.183809	1.688122
3	0.008740	0.004370	0.045157	1.160961	1.420976
4	0.000000	0.000000	0.008270	1.365955	0.196416
5	0.024673	1.509434	0.326560	0.105225	1.809869
6	0.012931	0.079741	1.495690	0.068966	0.342672
7	1.183445	0.463087	0.040268	0.107383	1.868009
8	1.345291	0.152466	0.026906	1.237668	0.573991
9	0.000000	0.000000	1.050761	1.152284	0.639594
10	0.081522	1.353261	0.032609	1.309783	0.543478
11	0.625000	1.750000	1.375000	0.000000	0.750000
12	0.245614	1.491228	0.000000	0.000000	0.192982
13	1.461538	0.000000	1.000000	0.000000	0.769231

Figura 2.20: Medias de los datos seleccionados por cluster MeanShift.

2.2.2. Resultados algoritmo Birch caso 2

La figura 2.37 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas. (ScatterMatrix)

La figura 2.38 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo Birch para el caso 2.(HeatMap)

La figura 2.39 es un dendograma obtenido por la ejecución del algoritmo Birch para el caso 2.

A continuación se muestra la figura 2.40 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

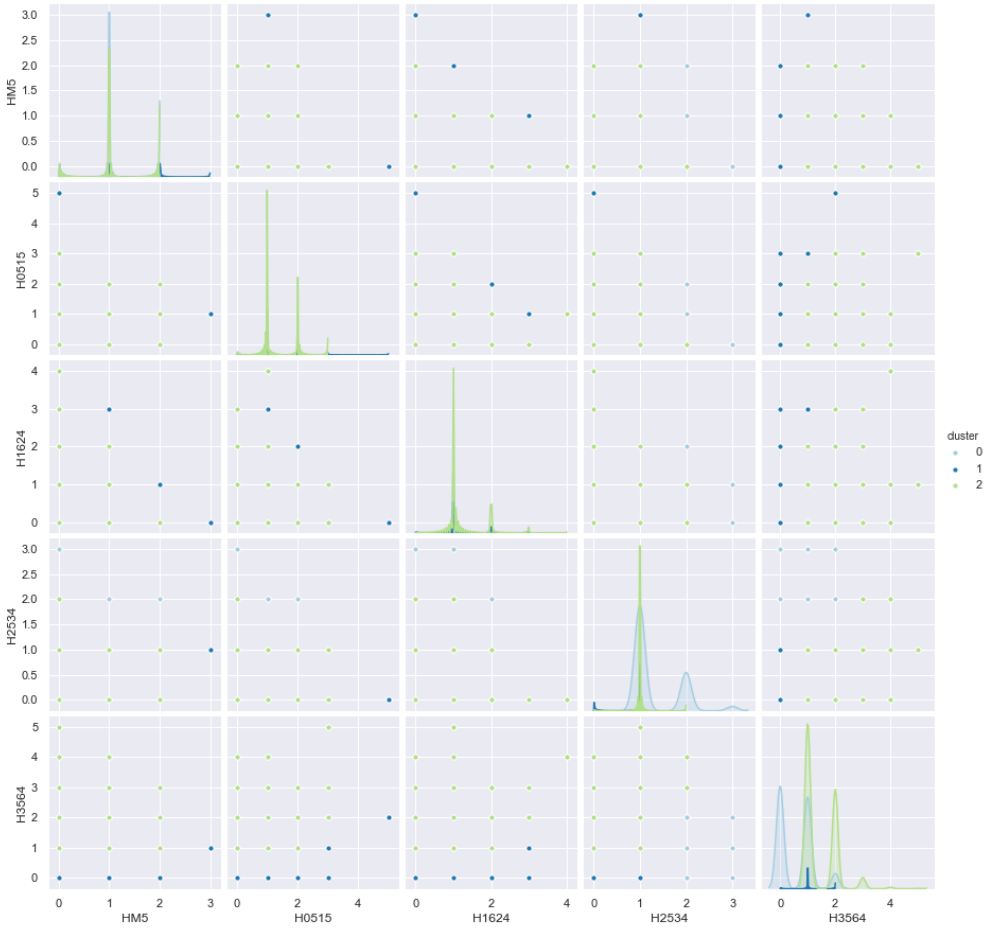

Figura 2.21: Scatter Matrix usando spectral en el caso de estudio 1.

La figura 2.41 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio 2.

2.2.3. Resultados algoritmo Ward caso 2

La figura 2.42 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas (Scatter Matrix).

La figura 2.43 representa la media normalizada de los datos totalas de cada variable asociados a cada cluster usando el algoritmo Ward para el caso 2. (HeatMap)

La figura 2.44 es un dendograma obtenido por la ejecución del algoritmo Ward para el caso 2.

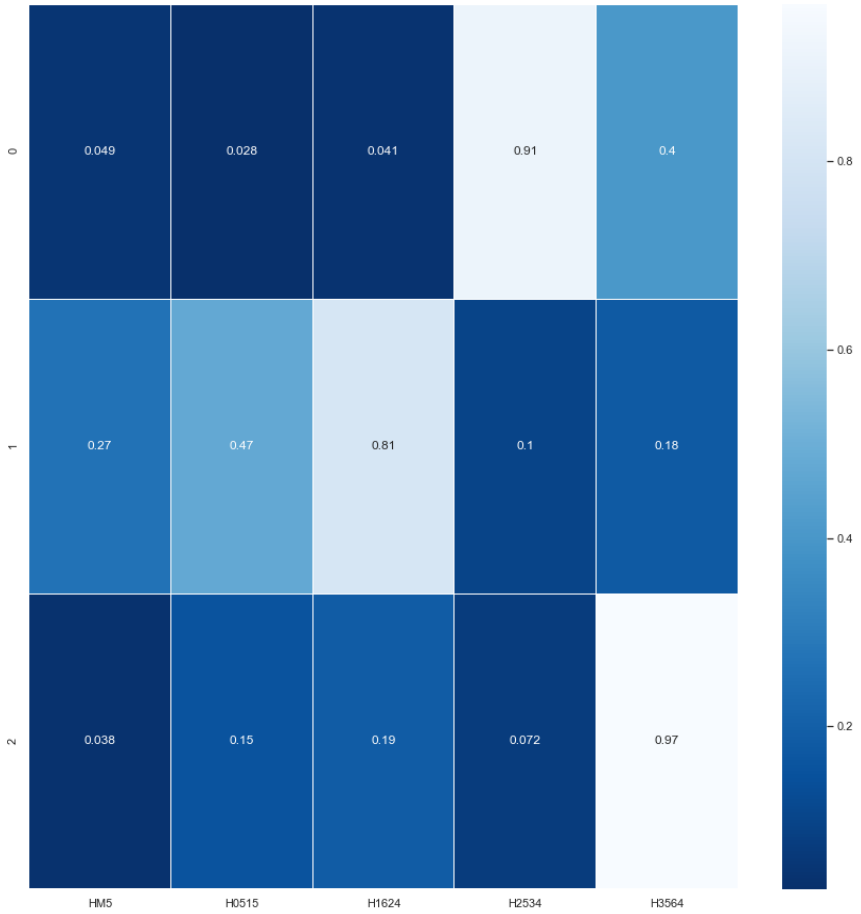

Figura 2.22: HeatMap usando spectral en el caso de estudio 1.

A continuación se muestra la figura 2.45 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.46 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio del caso 2.

2.2.4. Resultados algoritmo MeanShift caso 2

La figura 2.47 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas (Scatter Matrix).

La figura 2.48 representa la media normalizada de los datos totales de cada variable aso-

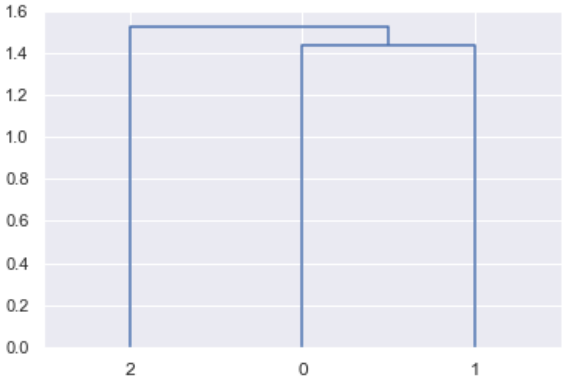

Figura 2.23: Dendograma usando spectral en el caso de estudio 1.

ciados a cada cluster usando el algoritmo MeanShift para el caso 1.

La figura 2.49 es un dendograma obtenido por la ejecución del algoritmo MeanShift para el caso 2. (HeatMap)

A continuación se muestra la figura 2.50 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.51 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

2.2.5. Resultados algoritmo Spectral caso 2

La figura 2.52 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas (Scatter Matrix).

La figura 2.53 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo spectral para el caso 2. (HeatMap)

La figura 2.54 es un dendograma obtenido por la ejecución del algoritmo MeanShift para el caso 2.

A continuación se muestra la figura 2.55 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.56 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

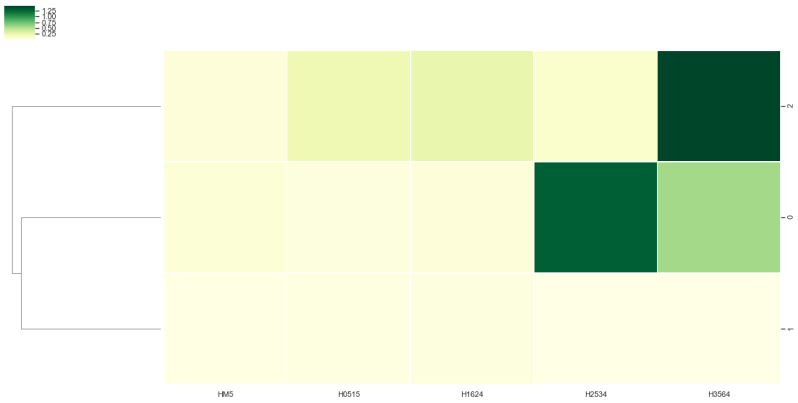

Figura 2.24: HeatMap con Dendograma usando spectral en el caso de estudio 1.

	HM5	H0515	H1624	H2534	H3564
0	0.070740	0.040014	0.059307	1.316899	0.581994
1	0.014027	0.024809	0.042462	0.005248	0.009447
2	0.057638	0.230081	0.281974	0.108307	1.455076

Figura 2.25: Medias de los datos seleccionados por cluster spectral.

2.2.6. Algoritmos modificados en el caso 2

En esta sección se va a exponer la modificación de los parámetros de dos algoritmos distintos y para ver sus diferencias se van a comparar los resultados de las métricas obtenidas en las secciones previas.

El primer algoritmo que vamos a modificar es Spectral , intentando aumentar el valor de sus métricas aumentando el numero de clusters a 5 a 6 y el segundo algoritmo que vamos a modificar es Ward, y vamos a incrementar su numero de clusters de 20 a 35 para ver que resultados obtenemos y compararlos con la anterior ejecución.

La figura 2.57 se muestra la antigua tabla pero ahora con las métricas de la ejecucion de estos dos algoritmos modificados. En ella se aprecia que las modificaciones de los parámetros no han sido exitosas en ninguno de los casos. En el algoritmo Spectral modificado han disminuido bastante la métrica CH y la métrica SH y en el algoritmo Ward modificado ha disminuido la métrica HC pero ha incrementado un poco la métrica SC.

Nombre Algoritmo	N Clusters	HC metric	SC metric	Time
K-means	4	29697.858169	0.764292	0.100956
Birch	6	21480.901579	0.729611	0.646720
Ward	10	33361.232561	0.845975	19.899369
MeanShift	14	32215.188490	0.855566	0.784660
Spectral	3	26907.715611	0.726861	421.776694
Birch_modificado	4	14704.418333	0.669432	0.687701
Ward_modificado	35	87235.116033	0.955259	11.627957

Figura 2.26: Metricas obtenidas usando algoritmos modificados en el caso de estudio 1.

2.2.7. Algoritmo modificado Spectral caso 2

La figura 2.58 representa como están distribuidos los diferentes clusters sobre las diferentes variables con el algoritmo Spectral

A continuación se muestra la figura 2.59 , esta es la fusión de las gráficas de Heatmap y de Dendograma con el algoritmo Spectral modificado.

2.2.8. Algoritmo modificado Ward caso 2

La figura 2.60 representa como están distribuidos los diferentes clusters sobre las diferentes variables con el algoritmo Ward modificado

A continuación se muestra la figura 2.61 , esta es la fusión de las gráficas de Heatmap y de Dendograma con el algoritmo Ward modificado.

2.2.9. Interpretacion de la segmentacion caso 2

Para terminar con el estudio de este caso , vamos a interpretar las visualizaciones producidas por las ejecuciones de los algoritmos.

Para este caso de estudio, en las gráficas correspondientes a los algoritmos del caso 2 podemos apreciar que la mayoría de las edades de las personas obtenidas con los filtros aplicados oscilan entre 0 y 30 años siendo entre 0 y 20 la parte mayoritaria, con un tamaño familiar de 2 a 4 personas entre las cuales suele haber un miembro entre 16 y 24 años, otro de 25 a 34 y dos miembros de 35 a 64. Son el tipo personas mayormente menores de 25 años que viven con sus dos padres y en ocasiones tienen hermanos.

En el caso del algoritmo Ward podemos ver que en los cluster 3, 6 y 8 se han agrupado aquellas personas que no tienen hermanos y solamente viven con sus padres. En los clus-

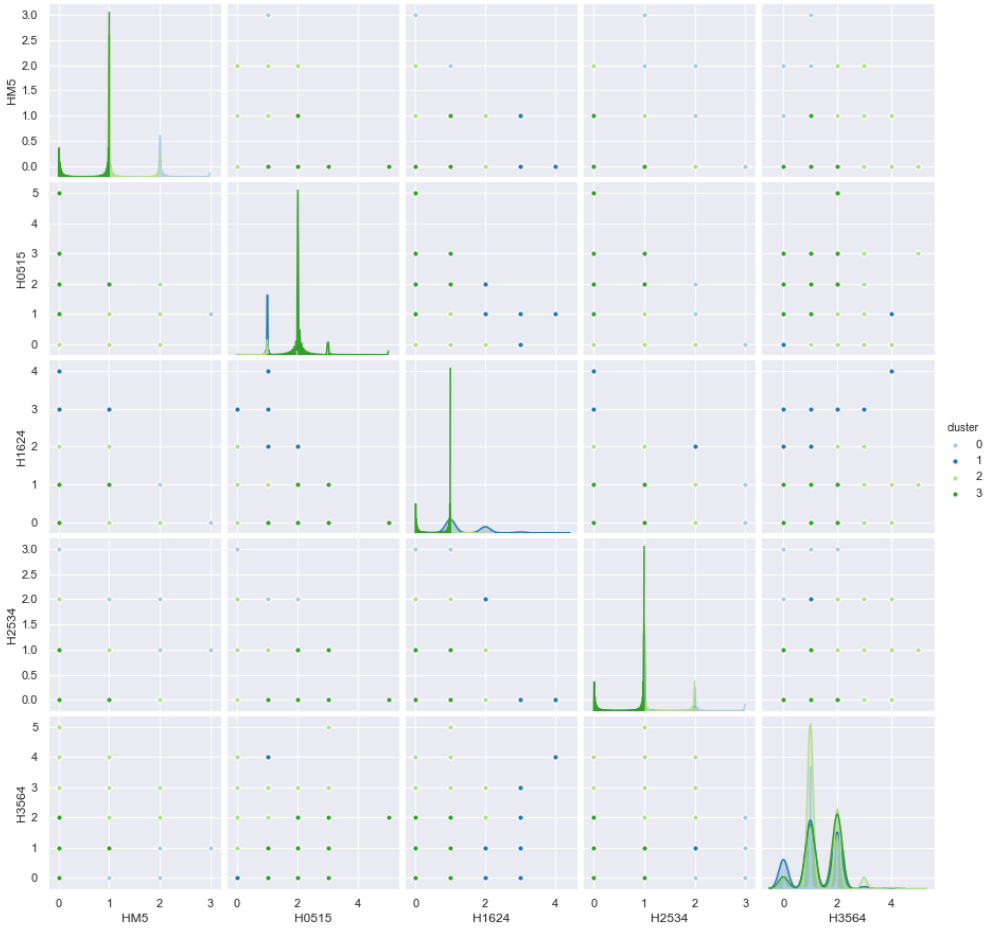

Figura 2.27: Scatter Matrix usando Birch modificado en el caso de estudio 1.

ter 1 y 18 se han agrupado las personas con algún hermano entre 16 y 24 que viven con sus padres.

Para el caso del algoritmo Spectal , en el cluster cero se agrupan principalmente aquellas personas que viven con varias personas entre 25 y 34 (posiblemente sus padres), En los clusters 3 y 4 se agrupan aquellas personas que viven con varias personas entre 35 y 64 (posiblemente sus padres) y probablemente tengan algún hermano.

2.3. Tercer caso de estudio: Personas solteras mayores de 40 años y sin hijos en el hogar

En este tercer caso de estudio nos vamos a centrar en aquellas personas solteras mayores de 40 años y sin hijos en el hogar, analizaremos su formación académica y las edades del padre o madre. Las variables que vamos a utilizar son :

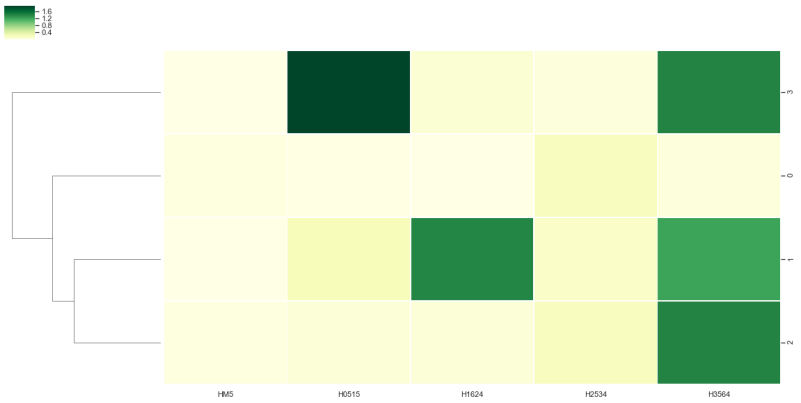

Figura 2.28: HeatMap con Dendograma usando Birch modificado en el caso de estudio 1.

TESTUD : Tipo de estudios realizados.
TAMNUC : Tamaño del núcleo familiar .
ESREAL : Estudios realizados.
EDADPAD : Edad del padre.
EDADMAD : Edad de la madre.

En la figura 2.62 se muestran los datos asociados a cada algoritmo usado para este caso de estudio de personas que viven con personas mayores , datos como el numero de clusters que se han usado, la metrica Calinski-harabasz (CH) , la metrica Siljouette (SC) y el tiempo en segundos que ha tardado el algoritmo para ejecutarse . Para este caso de estudio se han contado con un total de 22628 instancias.

En este caso , vemos como para el indice Calinski-Harabasz los algoritmos con mejor métrica son Ward y K-means, dejando muy por detrás a Birch o MeanShift.

Para el indice Silhouette no ocurre lo mismo que para el Calinski-Harabasz, ya que todos los resultados se encuentra a la par, teniendo MeanShift y Spectral las mejores métricas en este caso.

Por ultimo cabe decir que el algoritmo Spectral es que mayor tiempo de ejecucion tiene. En las siguientes subsecciones se muestran gráficas y tablas de algoritmos asociadas a cada algoritmo usado para cada caso de estudio , y al final exponemos un análisis de los resultados obtenidos.

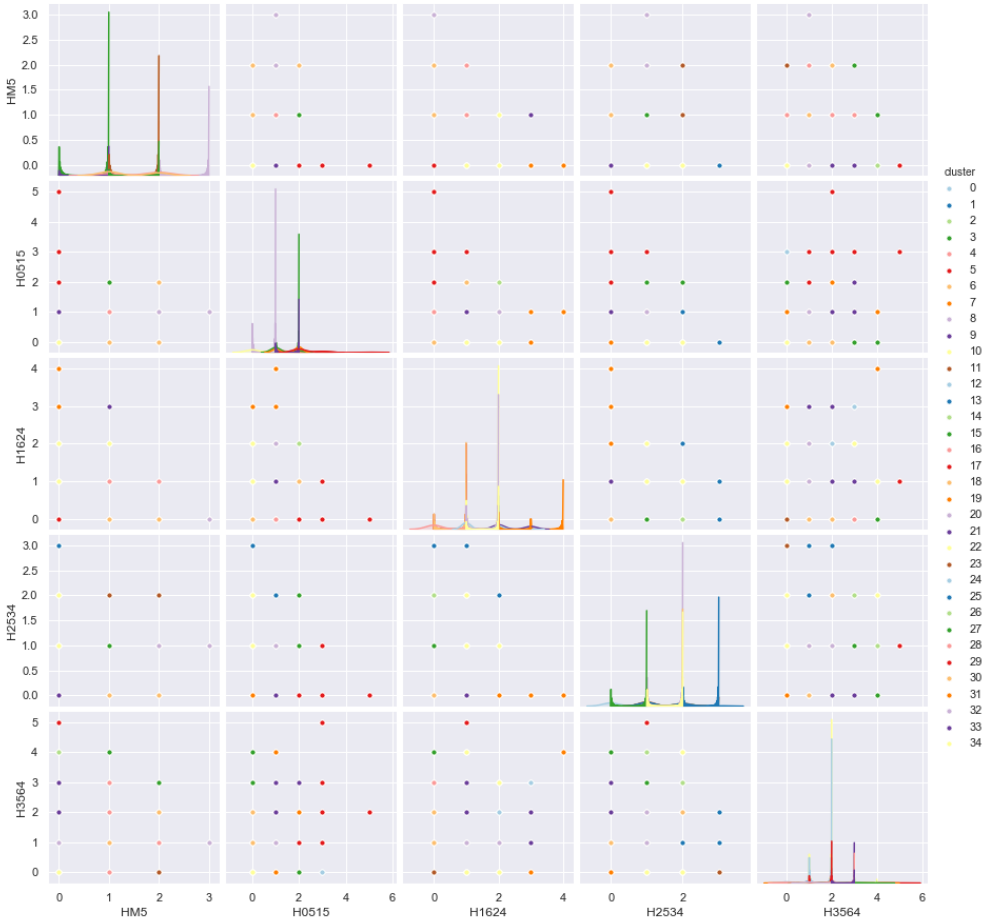

Figura 2.29: Scatter Matrix usando Ward modificado en el caso de estudio 1.

Cabe destacar que se ha realizado la eliminación de aquellos clusters con pocos datos (ouliers) . Se ha realizado mediante un filtrado a los clusters con menos de 5 elementos.

2.3.1. Resultados algoritmo K-Means caso 3

La figura 2.63 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas. (Scatter Matrix)

La figura 2.64 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo k-means para el caso 3. (HeatMap)

La figura 2.65 es un dendograma , puede ayudar a decidir el numero de grupos que podrían representar mejor la estructura de los datos teniendo en cuenta la forma en la que

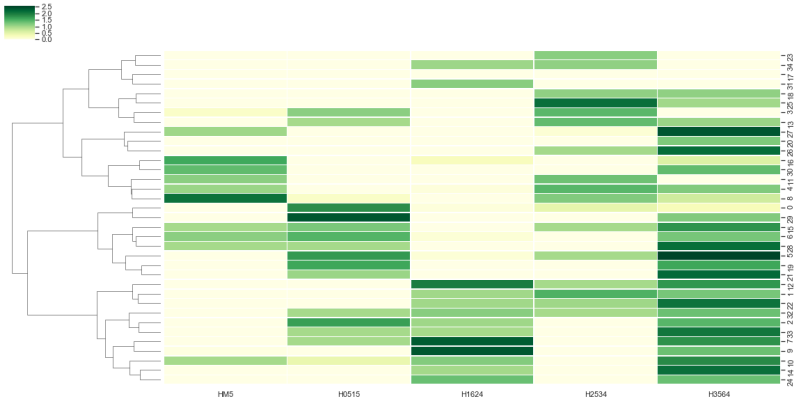

Figura 2.30: HeatMap con Dendograma usando Ward modificado en el caso de estudio 1.

Nombre Algoritmo	N Clusters	HC metric	SC metric	Time
K-means	6	18546.682880	0.357961	0.211908
Birch	6	6164.728674	0.447193	0.314865
Ward	20	10917.863379	0.329162	23.355891
MeanShift	7	2741.081833	0.445111	1.797222
Spectral	5	13462.283519	0.379498	542.808061

Figura 2.31: Resultados y caracteristicas de los algoritmos para el caso de estudio 2.

se van anidando los clusters y la medida de similitud a la cual lo hacen.

A continuacion se muestra la figura 2.66 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.67 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

2.3.2. Resultados algoritmo Birch caso 3

La figura 2.68 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas. (ScatterMatrix)

La figura 2.69 representa la media normalizada de los datos totales de cada variable aso-

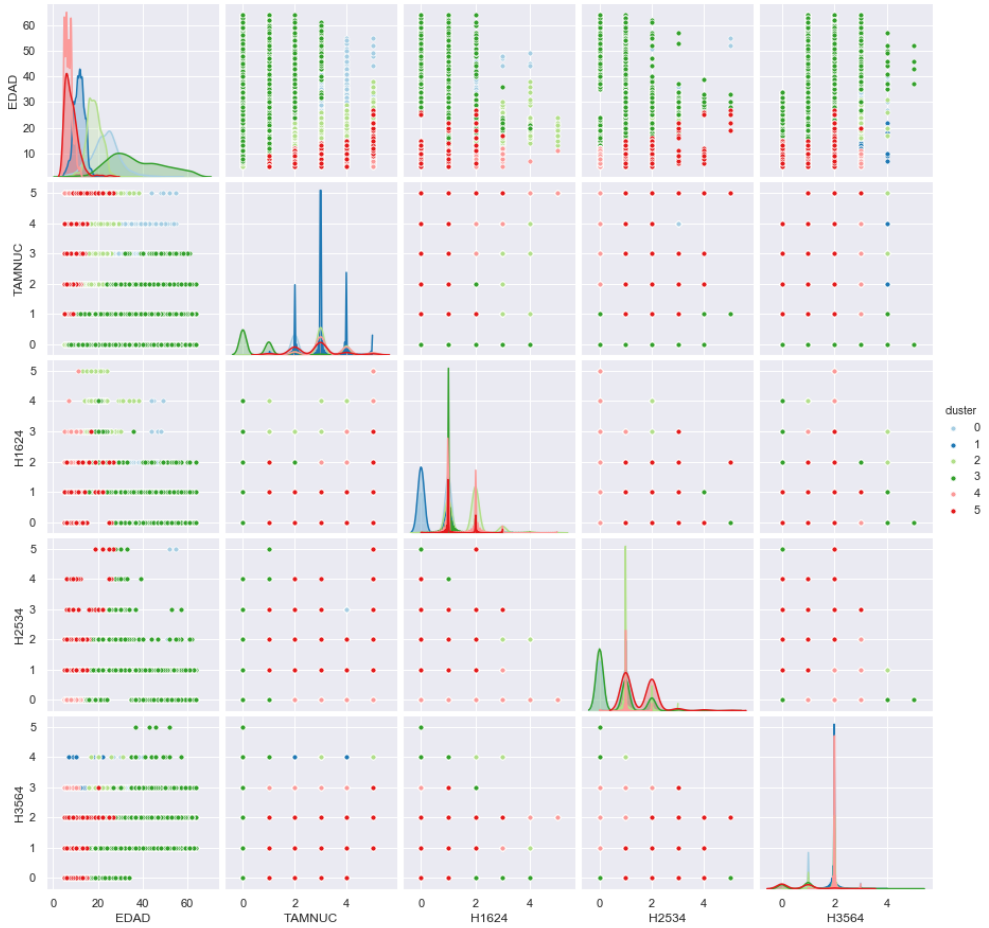

Figura 2.32: Scatter Matrix usando K-means en el caso de estudio 2.

ciados a cada cluster usando el algoritmo Birch para el caso 3.(HeatMap)

La figura 2.70 es un dendograma obtenido por la ejecucion del algoritmo Birch para el caso 3.

A continuación se muestra la figura 2.71 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.72 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio 3.

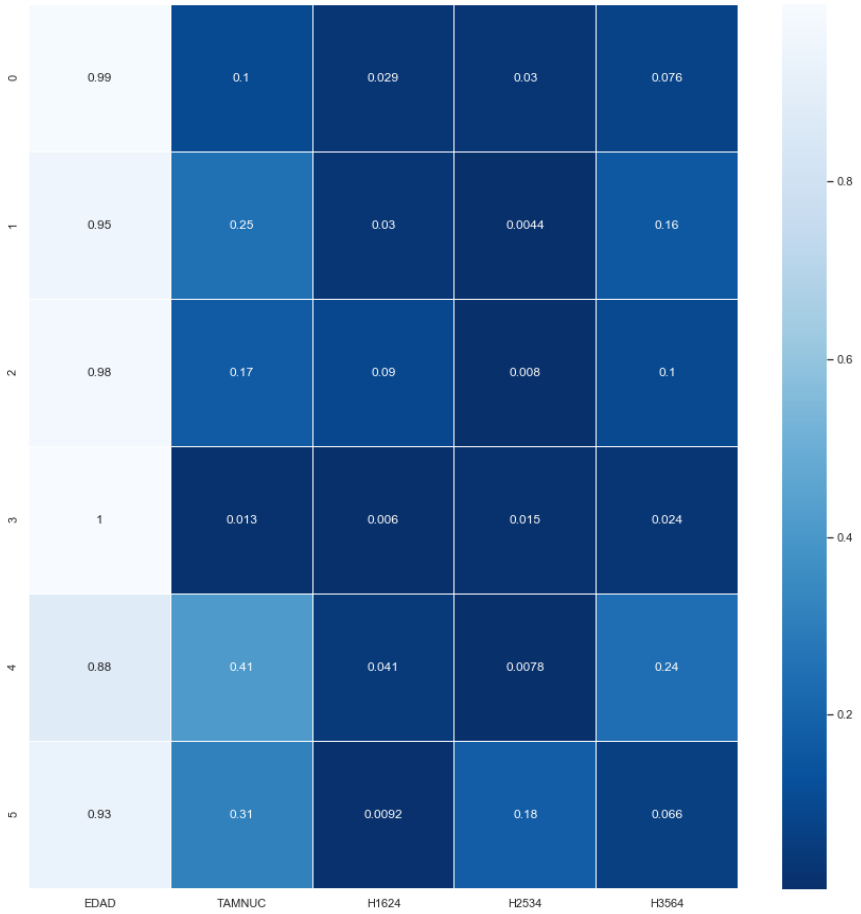

Figura 2.33: HeatMap usando K-means en el caso de estudio 2.

2.3.3. Resultados algoritmo Ward caso 3

La figura 2.73 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas (Scatter Matrix).

La figura 2.74 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo Ward para el caso 3. (HeatMap)

La figura 2.75 es un dendograma obtenido por la ejecucion del algoritmo Ward para el caso 3.

A continuación se muestra la figura 2.76 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

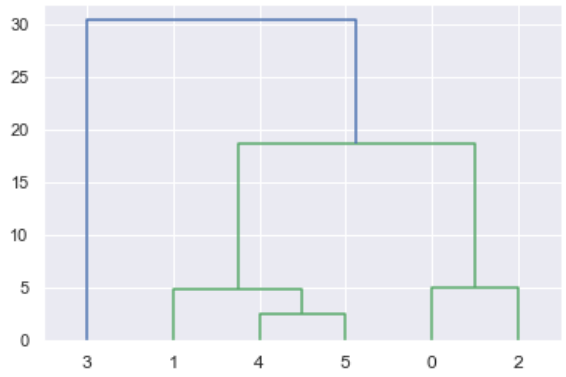

Figura 2.34: Dendograma usando K-means en el caso de estudio 2.

La figura 2.77 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio del caso 3.

2.3.4. Resultados algoritmo MeanShift caso 3

La figura 2.78 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas (Scatter Matrix).

La figura 2.79 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo MeanShift para el caso 3.

La figura 2.80 es un dendograma obtenido por la ejecucion del algoritmo MeanShift para el caso 3. (HeatMap)

A continuación se muestra la figura 2.81 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.82 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

2.3.5. Resultados algoritmo Spectral caso 3

La figura 2.83 representa como están distribuidos los diferentes clusters sobre las diferentes variables estudiadas (Scatter Matrix).

La figura 2.84 representa la media normalizada de los datos totales de cada variable asociados a cada cluster usando el algoritmo spectral para el caso 3. (HeatMap)

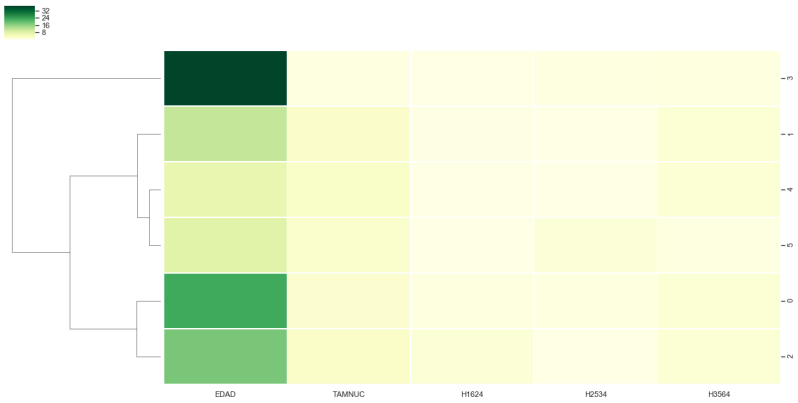

Figura 2.35: HeatMap con Dendograma usando K-means en el caso de estudio 2.

	EDAD	TAMNUC	H1624	H2534	H3564
0	23.332451	2.407407	0.685714	0.699471	1.795062
1	11.677437	3.051317	0.371436	0.053489	1.942981
2	18.563870	3.224301	1.712396	0.152494	1.904573
3	37.150435	0.499478	0.224522	0.571826	0.902609
4	6.995737	3.295981	0.325213	0.062119	1.949452
5	8.219966	2.774958	0.081218	1.554992	0.583756

Figura 2.36: Medias de los datos seleccionados por cluster K-means.

La figura 2.85 es un dendograma obtenido por la ejecucion del algoritmo MeanShift para el caso 3.

A continuación se muestra la figura 2.86 , esta es la fusión de las gráficas de Heatmap y de Dendograma.

La figura 2.87 esta compuesta por la media de los datos para cada cluster de las variables seleccionadas para el estudio.

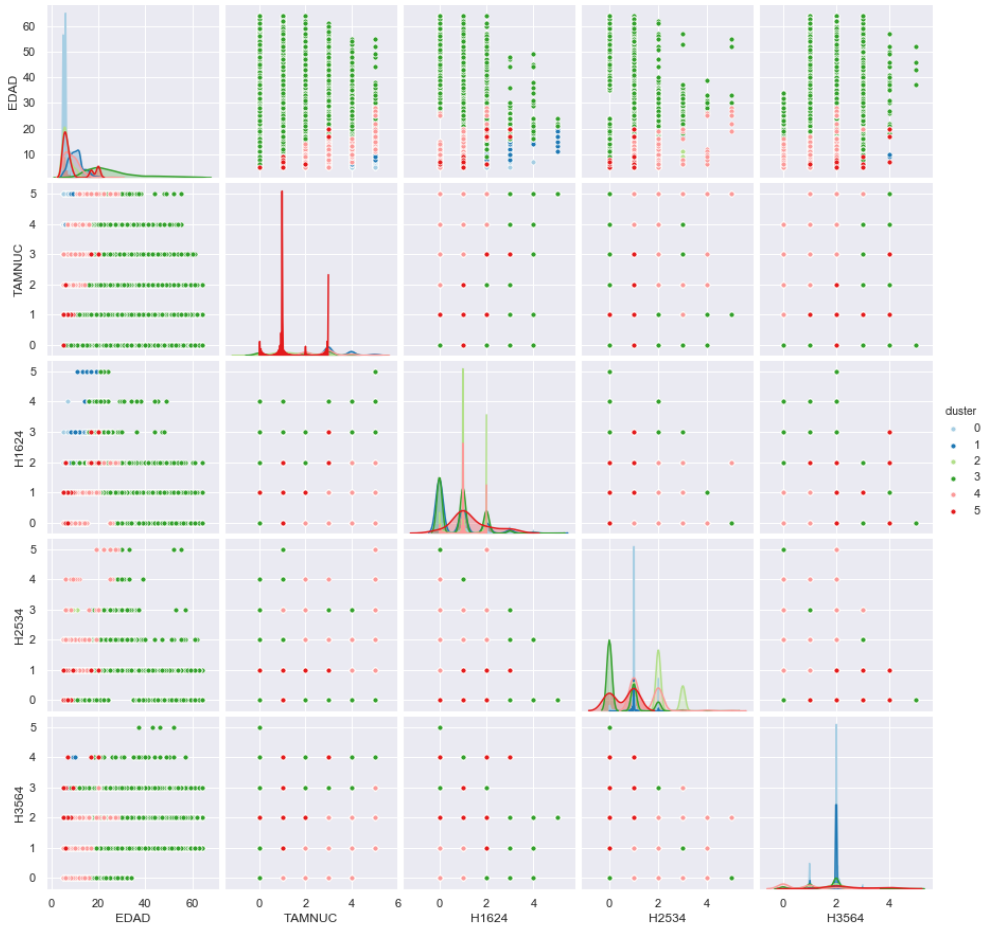

Figura 2.37: Scatter Matrix usando Birch en el caso de estudio 2.

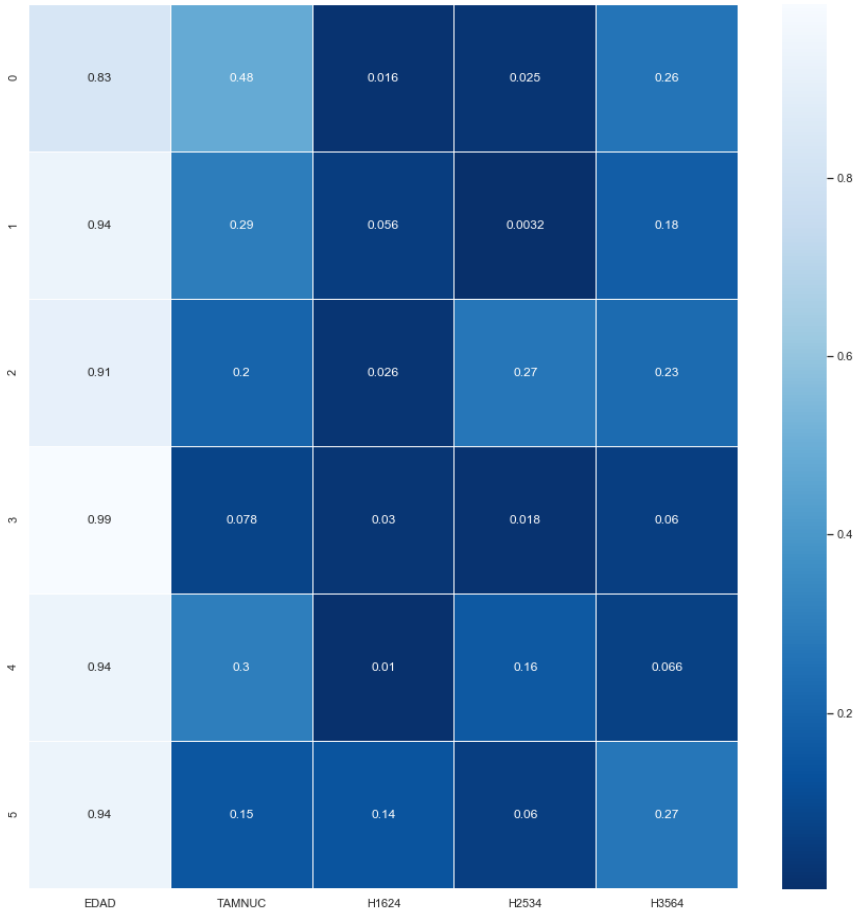

Figura 2.38: HeatMap usando Birch en el caso de estudio 2.

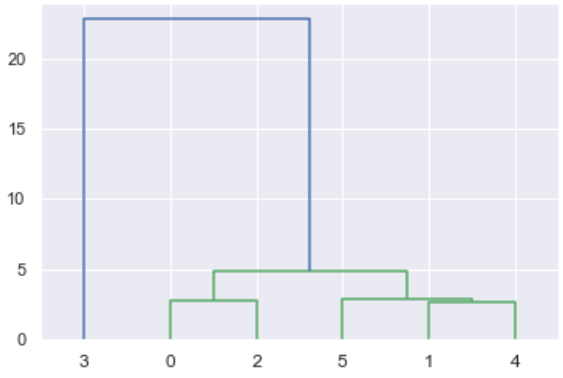

Figura 2.39: Dendograma usando Birch en el caso de estudio 2.

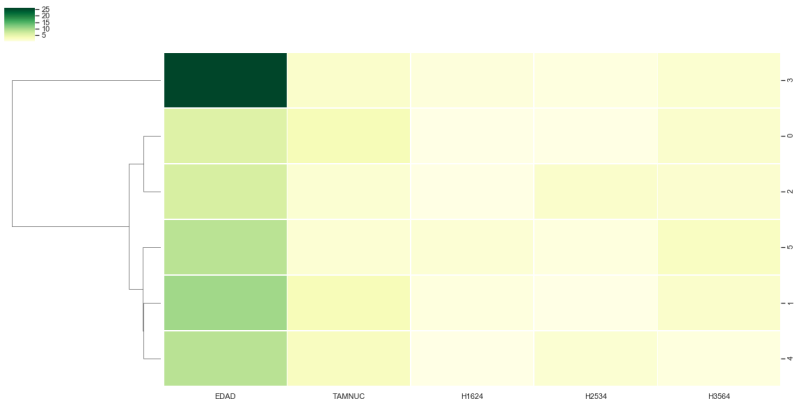

Figura 2.40: HeatMap con Dendograma usando Birch en el caso de estudio 2.

	EDAD	TAMNUC	H1624	H2534	H3564
0	5.889231	3.398462	0.112308	0.178462	1.863077
1	10.471802	3.289959	0.628886	0.035488	1.960660
2	6.625000	1.468750	0.187500	1.968750	1.656250
3	25.804270	2.030809	0.777822	0.454267	1.546356
4	8.828299	2.809221	0.093800	1.488076	0.620032
5	8.760000	1.400000	1.320000	0.560000	2.560000

Figura 2.41: Medias de los datos seleccionados por cluster Birch caso 2.

Figura 2.42: Scatter Matrix usando Ward en el caso de estudio 2.

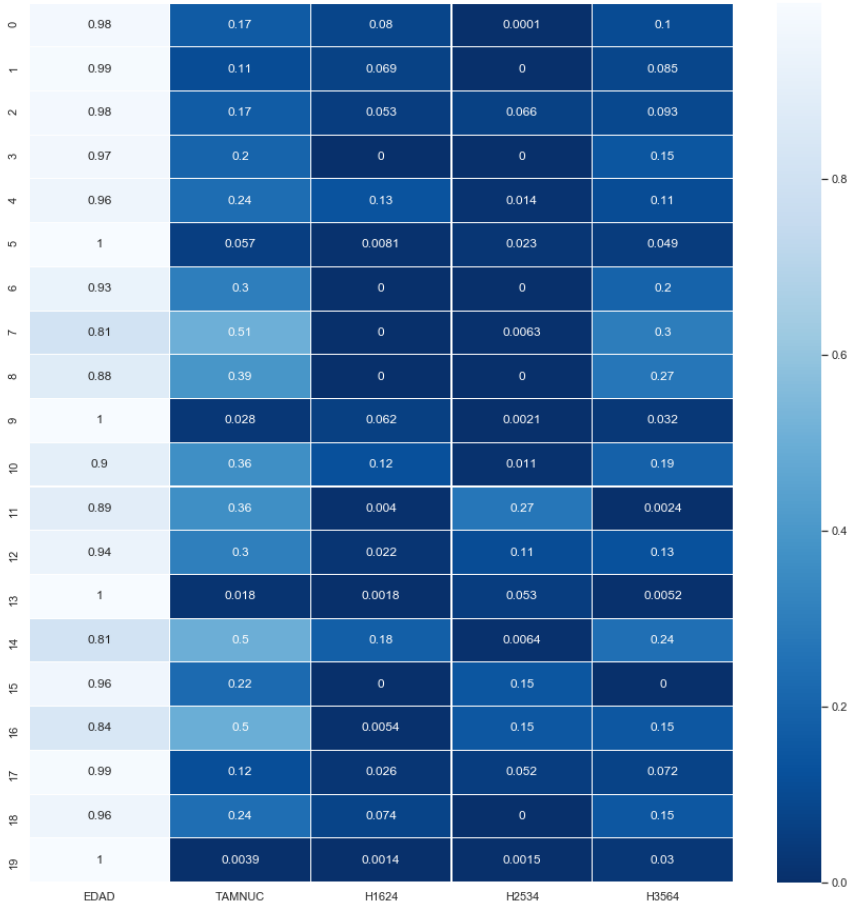

Figura 2.43: HeatMap usando Ward en el caso de estudio 2.

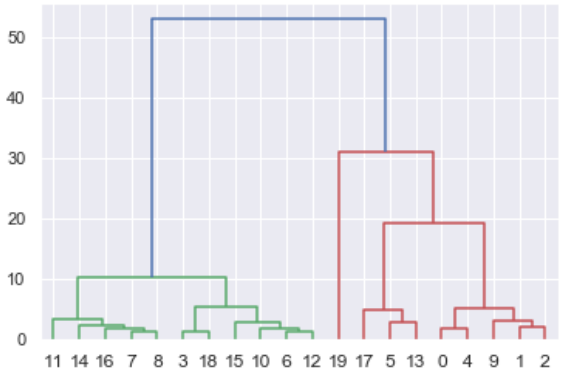

Figura 2.44: Dendograma usando Ward en el caso de estudio 2.

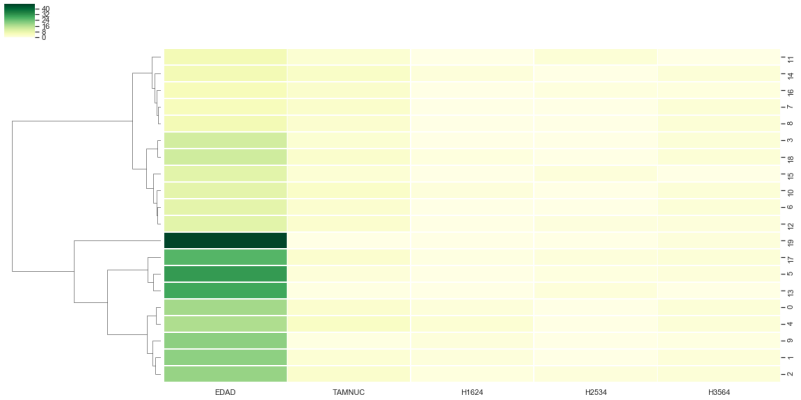

Figura 2.45: HeatMap con Dendograma usando Ward en el caso de estudio 2.

	EDAD	TAMNUC	H1624	H2534	H3564
0	18.151058	3.111289	1.489737	0.001924	1.926235
1	20.526098	2.265866	1.429591	0.000000	1.769808
2	19.634752	3.398345	1.060284	1.316785	1.873522
3	12.420742	2.562169	0.000000	0.000000	1.873895
4	16.920455	4.169318	2.367045	0.242045	1.900000
5	31.150916	1.783992	0.254098	0.732401	1.537126
6	9.368185	2.987304	0.000000	0.000000	1.961165
7	5.480769	3.426923	0.000000	0.042308	2.000000
8	6.552548	2.912420	0.000000	0.000000	2.027070
9	20.669399	0.587432	1.282787	0.043716	0.666667
10	9.387931	3.781609	1.258621	0.117816	1.951149
11	6.549708	2.678363	0.029240	2.000000	0.017544
12	9.743363	3.150442	0.227139	1.109145	1.303835
13	29.118818	0.527141	0.052473	1.556092	0.152593
14	6.532051	4.038462	1.467949	0.051282	1.961538
15	9.695122	2.256098	0.000000	1.487805	0.000000
16	5.695122	3.378049	0.036585	1.024390	1.012195
17	26.596270	3.107515	0.704882	1.386725	1.938563
18	12.880597	3.251741	1.000000	0.000000	2.004975
19	45.887335	0.178450	0.062004	0.068431	1.354631

Figura 2.46: Medias de los datos seleccionados por cluster Ward caso 2.

Figura 2.47: Scatter Matrix usando MeanShift en el caso de estudio 2.

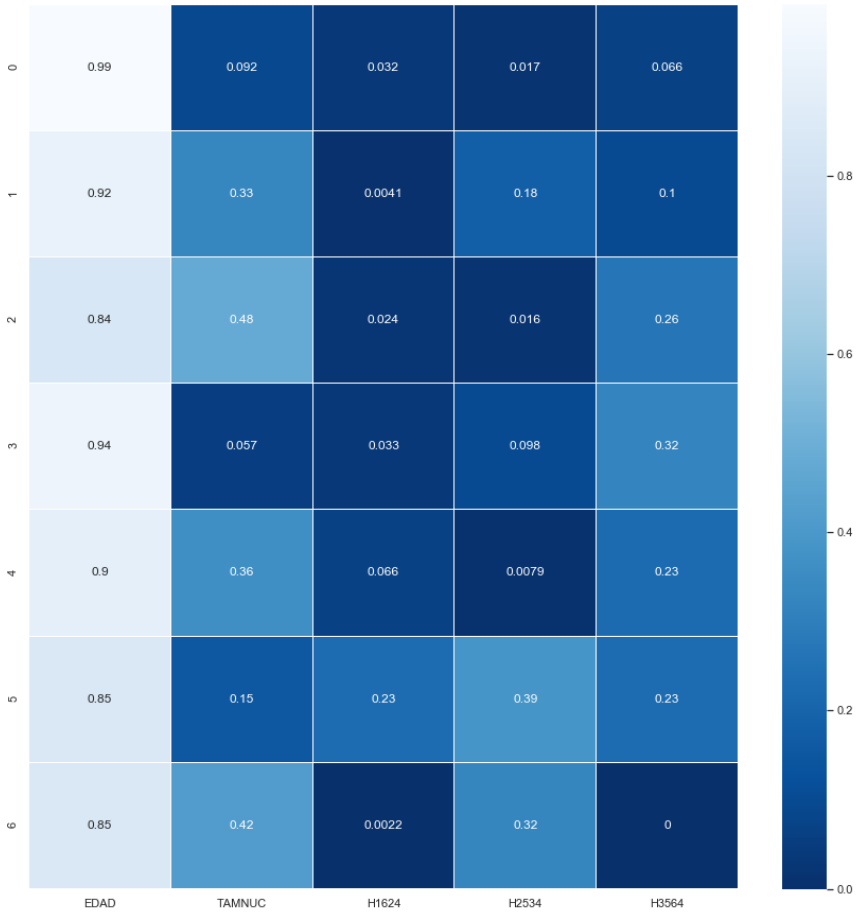

Figura 2.48: HeatMap usando MeanShift en el caso de estudio 2.

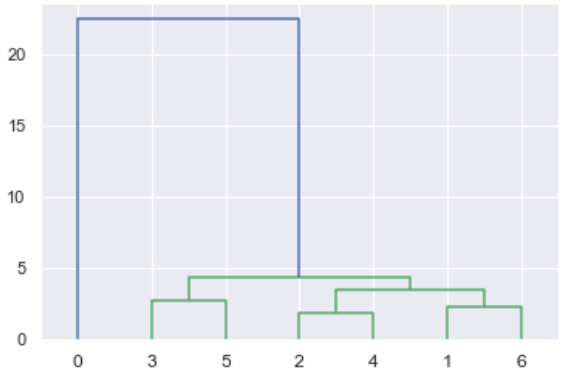

Figura 2.49: Dendograma usando MeanShift en el caso de estudio 2.

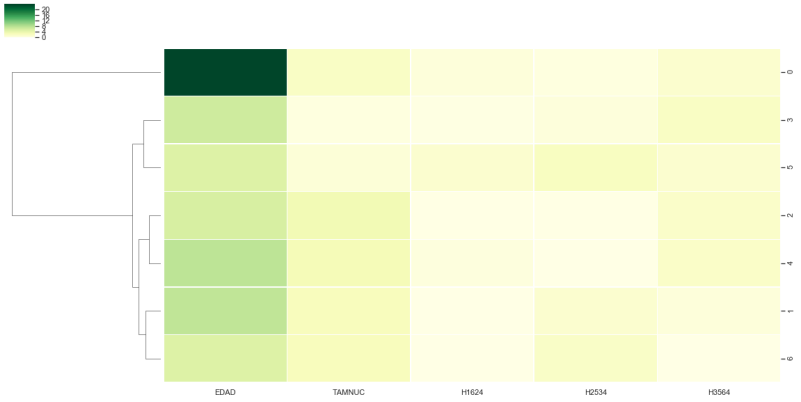

Figura 2.50: HeatMap con Dendograma usando MeanShift en el caso de estudio 2.

cluster	EDAD	TAMNUC	H1624	H2534	H3564
0	23.760192	2.205598	0.753827	0.412265	1.585937
1	7.626168	2.704050	0.034268	1.507788	0.825545
2	6.049296	3.477465	0.173239	0.112676	1.909859
3	6.764706	0.411765	0.235294	0.705882	2.294118
4	7.845905	3.176724	0.577586	0.068966	1.992457
5	5.500000	1.000000	1.500000	2.500000	1.500000
6	5.521127	2.704225	0.014085	2.112676	0.000000

Figura 2.51: Medias de los datos seleccionados por cluster MeanShift caso 2.

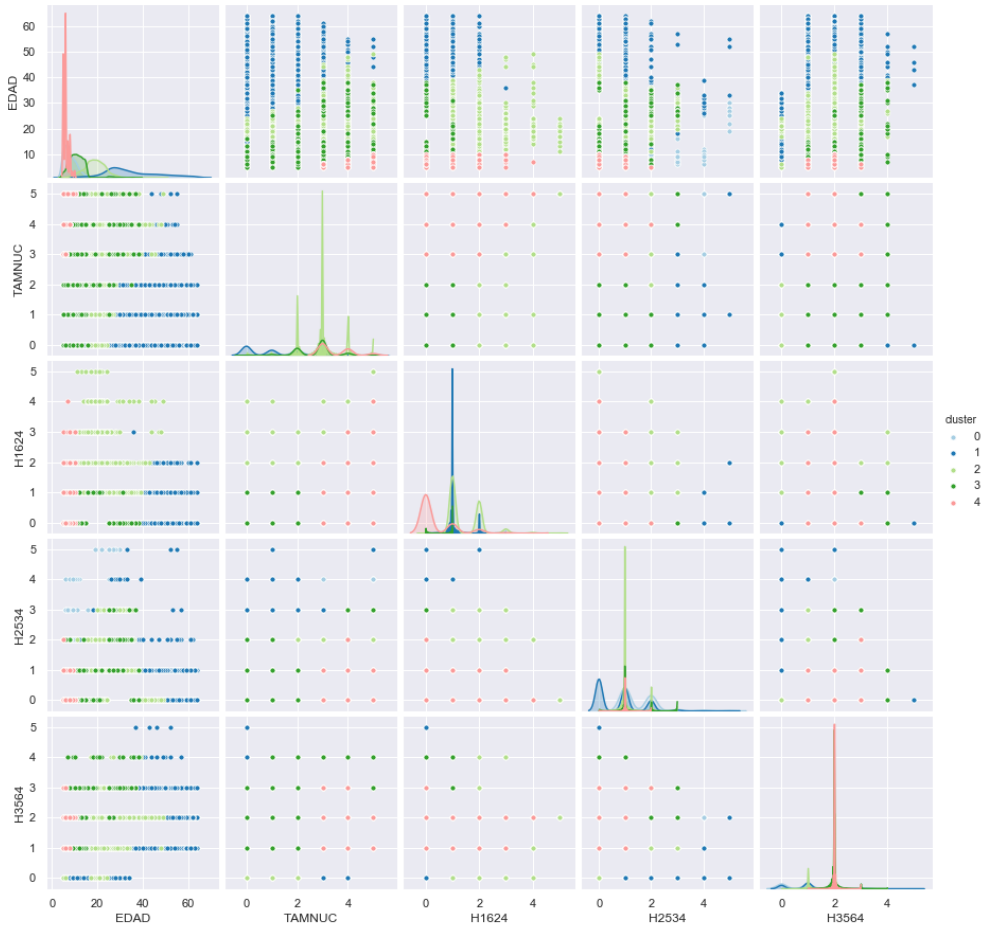

Figura 2.52: Scatter Matrix usando spectral en el caso de estudio 2.

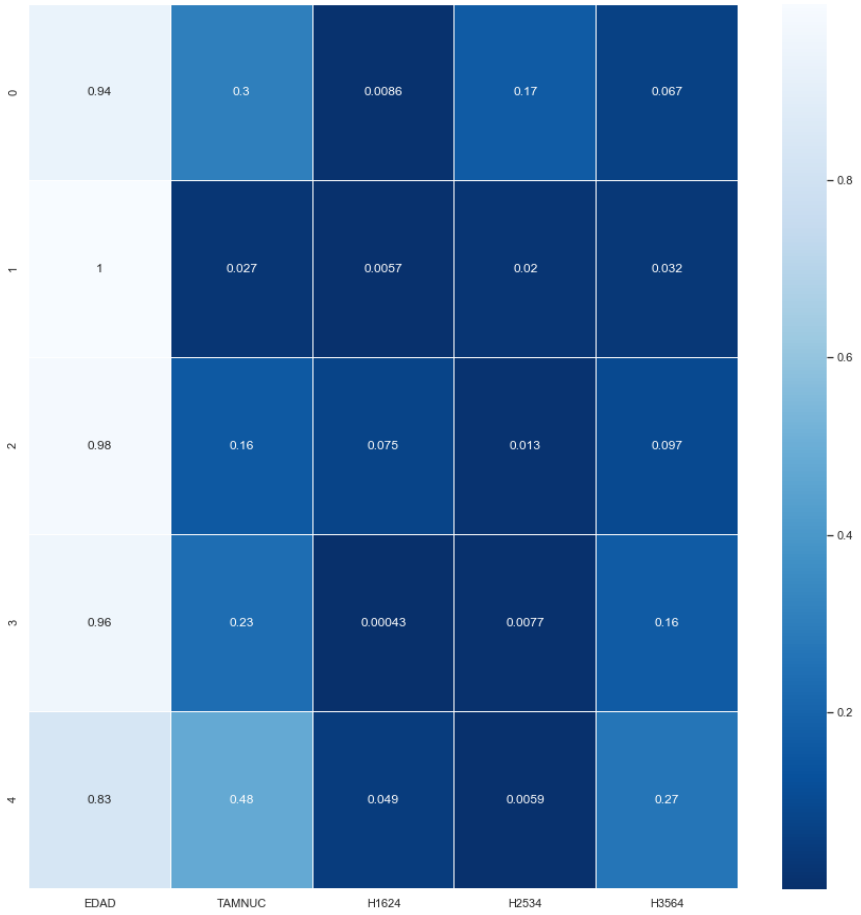

Figura 2.53: HeatMap usando spectral en el caso de estudio 2.

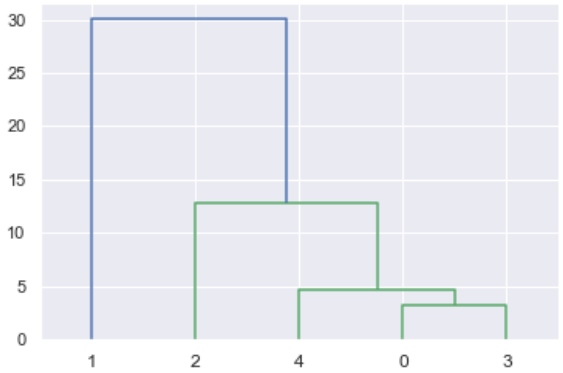

Figura 2.54: Dendograma usando spectral en el caso de estudio 2.

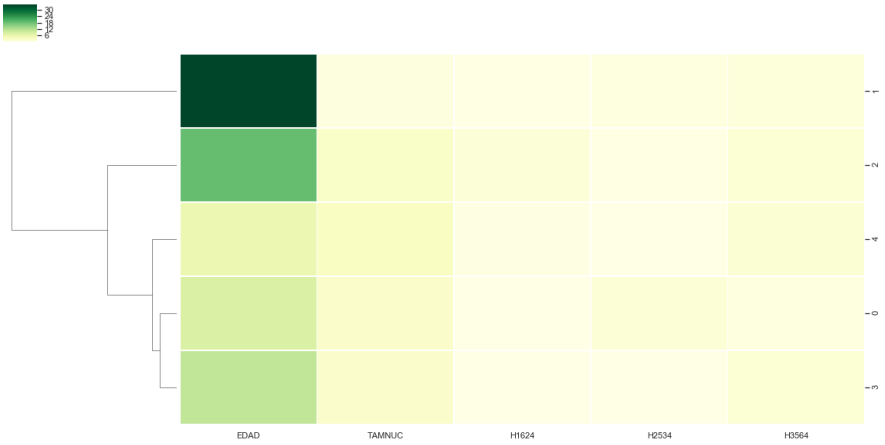

Figura 2.55: HeatMap con Dendograma usando spectral en el caso de estudio 2.

	EDAD	TAMNUC	H1624	H2534	H3564
0	8.628734	2.786629	0.079659	1.546230	0.614509
1	34.914764	0.957250	0.199894	0.708537	1.127721
2	18.933652	3.033428	1.457326	0.249238	1.881630
3	11.105749	2.717827	0.005011	0.089135	1.911392
4	6.123596	3.511236	0.362360	0.043539	1.970506

Figura 2.56: Medias de los datos seleccionados por cluster spectral caso 2.

	N Clusters	HC metric	SC metric	Time
K-means	6	18546.682880	0.357961	0.211908
Birch	6	6164.728674	0.447193	0.314865
Ward	20	10917.863379	0.329162	23.355891
MeanShift	7	2741.081833	0.445111	1.797222
Spectral	5	13462.283519	0.379498	542.808061
Spectral_modificado	6	12890.197403	0.309191	395.219940
Ward_modificado	35	10095.964535	0.388964	19.392607

Figura 2.57: Metricas obtenidas usando algoritmos modificados en el caso de estudio 2.

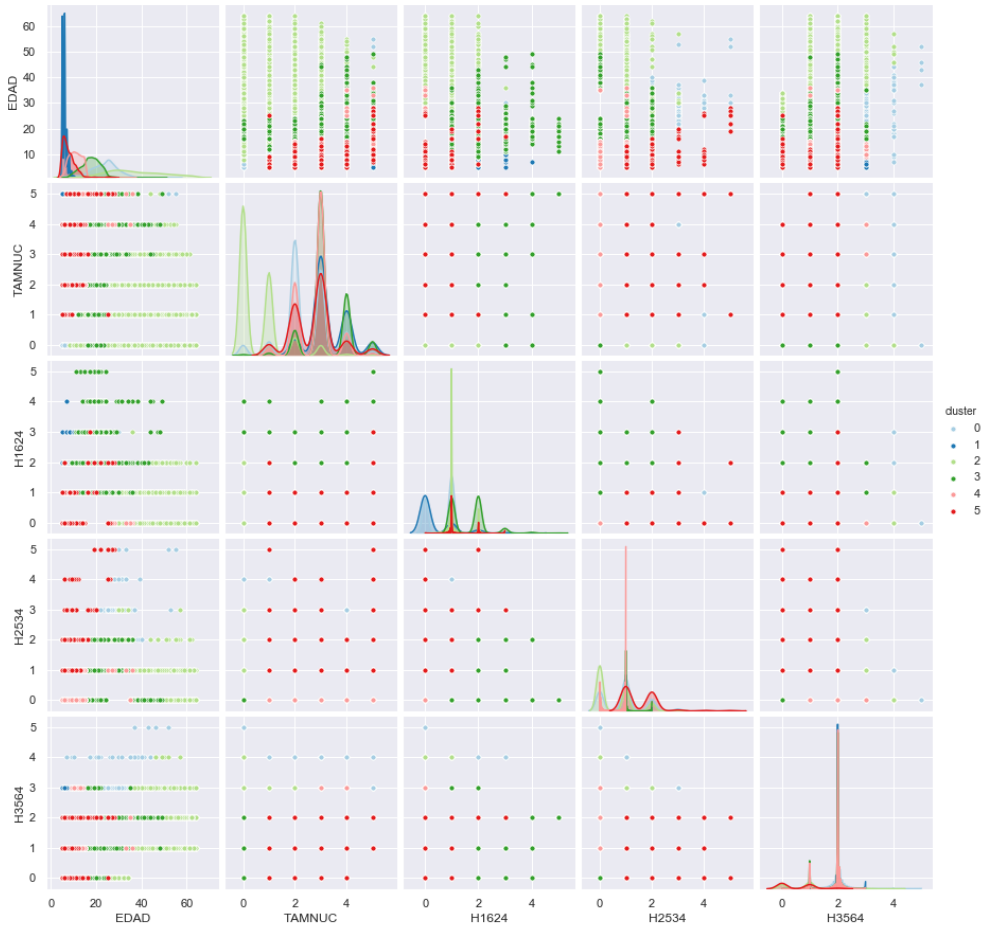

Figura 2.58: Scatter Matrix usando Spectral modificado en el caso de estudio 2.

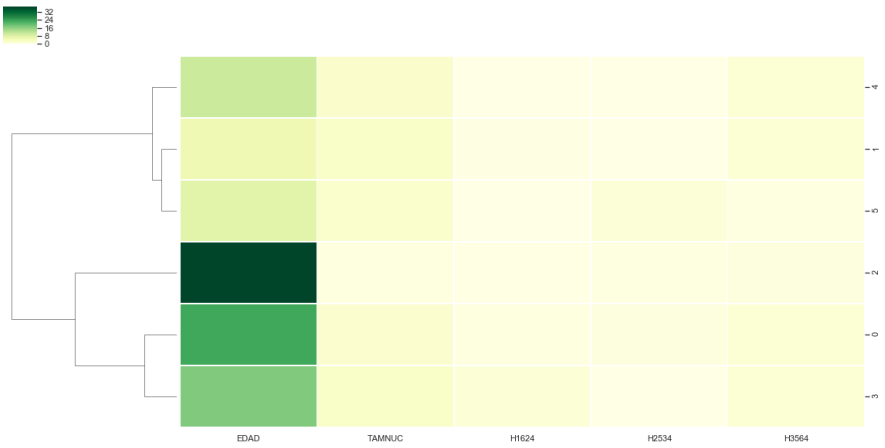

Figura 2.59: HeatMap con Dendograma usando Spectral modificado en el caso de estudio 2.

Figura 2.60: Scatter Matrix usando Ward modificado en el caso de estudio 2.

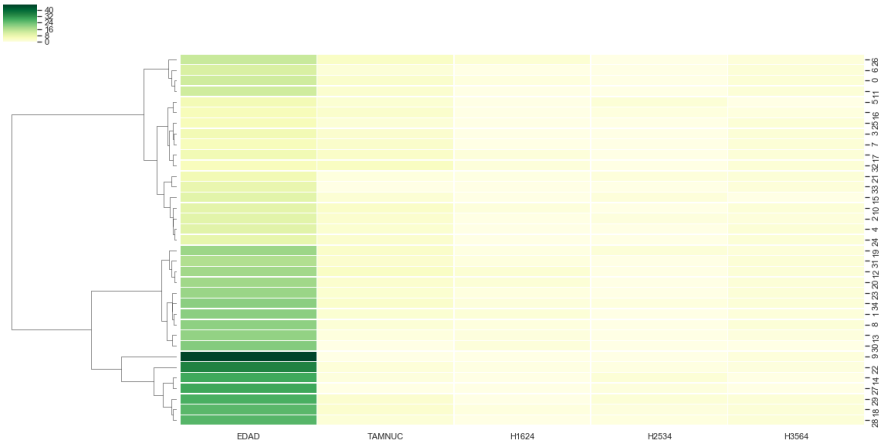

Figura 2.61: HeatMap con Dendograma usando Ward modificado en el caso de estudio 2.

Nombre Algoritmo	N Clusters	HC metric	SC metric	Time
K-means	10	7285.936286	0.634391	0.050824
Birch	10	4081.202946	0.697422	0.098930
Ward	15	7978.888763	0.696960	0.356119
MeanShift	9	3675.392770	0.714534	0.121471
Spectral	5	5824.512435	0.701737	2.284997

Figura 2.62: Resultados y caracteristicas de los algoritmos para el caso de estudio 3.

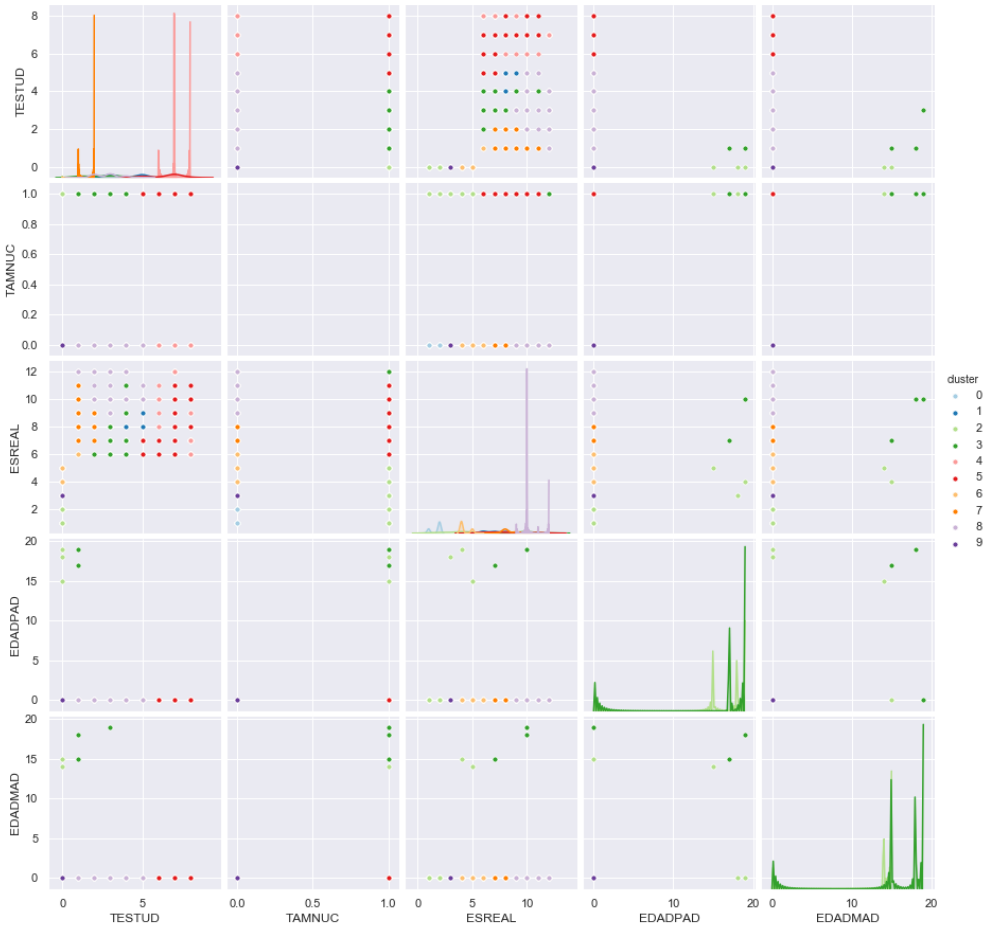

Figura 2.63: Scatter Matrix usando K-means en el caso de estudio 3.

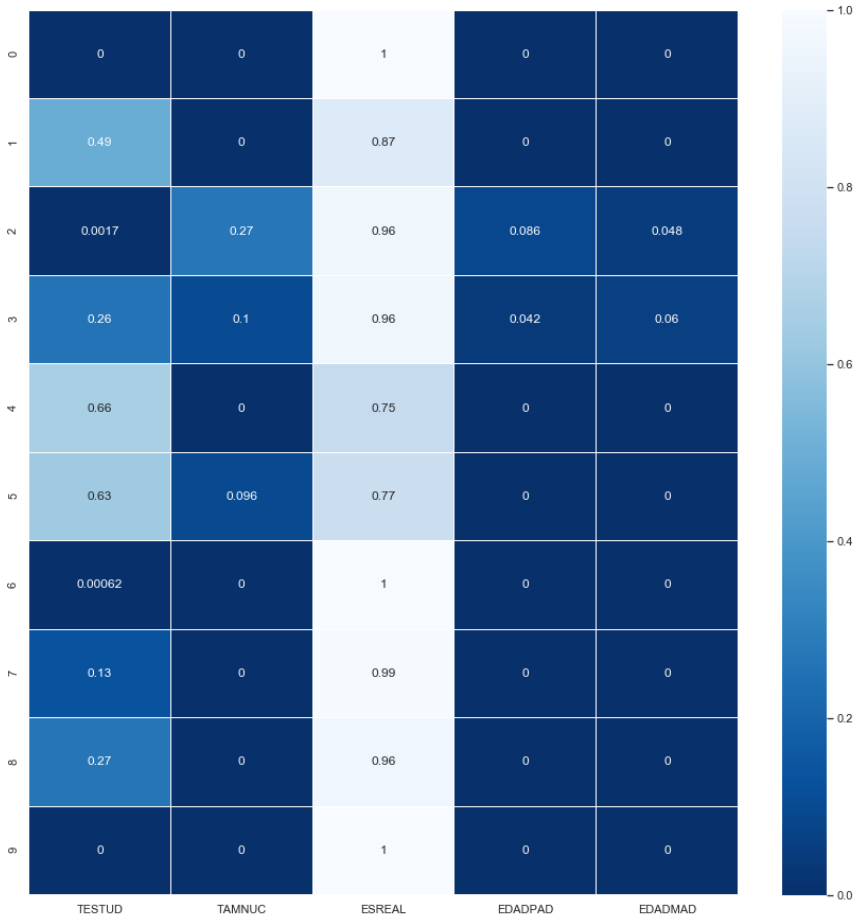

Figura 2.64: HeatMap usando K-means en el caso de estudio 3.

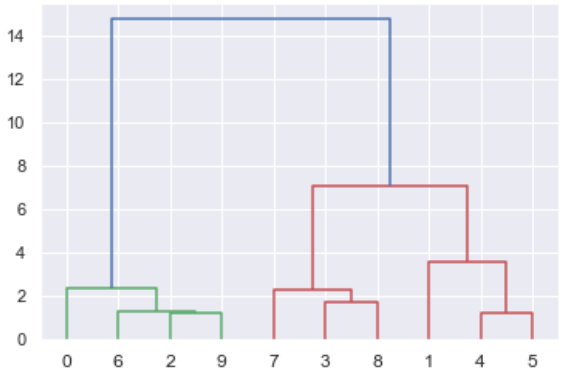

Figura 2.65: Dendograma usando K-means en el caso de estudio 3.

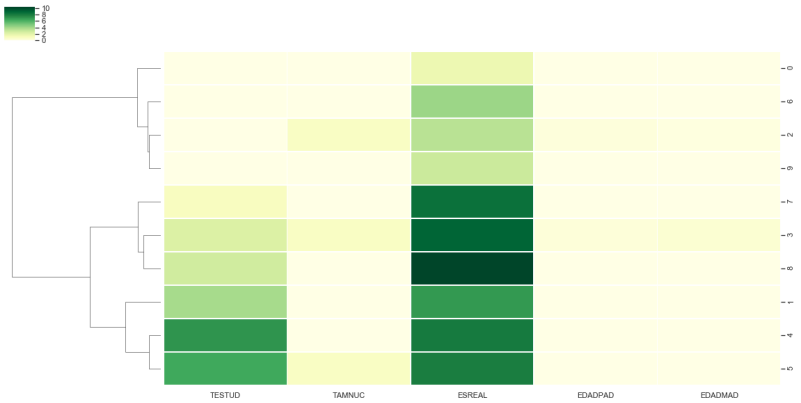

Figura 2.66: HeatMap con Dendograma usando K-means en el caso de estudio 3.

	TESTUD	TAMNUC	ESREAL	EDADPAD	EDADMAD
0	0.000000	0.0	1.721617	0.000000	0.000000
1	3.986486	0.0	7.027027	0.000000	0.000000
2	0.006098	1.0	3.518293	0.317073	0.176829
3	2.472527	1.0	9.120879	0.395604	0.571429
4	7.172093	0.0	8.186047	0.000000	0.000000
5	6.511628	1.0	8.023256	0.000000	0.000000
6	0.002635	0.0	4.268775	0.000000	0.000000
7	1.121212	0.0	8.581818	0.000000	0.000000
8	2.849315	0.0	10.301370	0.000000	0.000000
9	0.000000	0.0	3.000000	0.000000	0.000000

Figura 2.67: Medias de los datos seleccionados por cluster K-means.

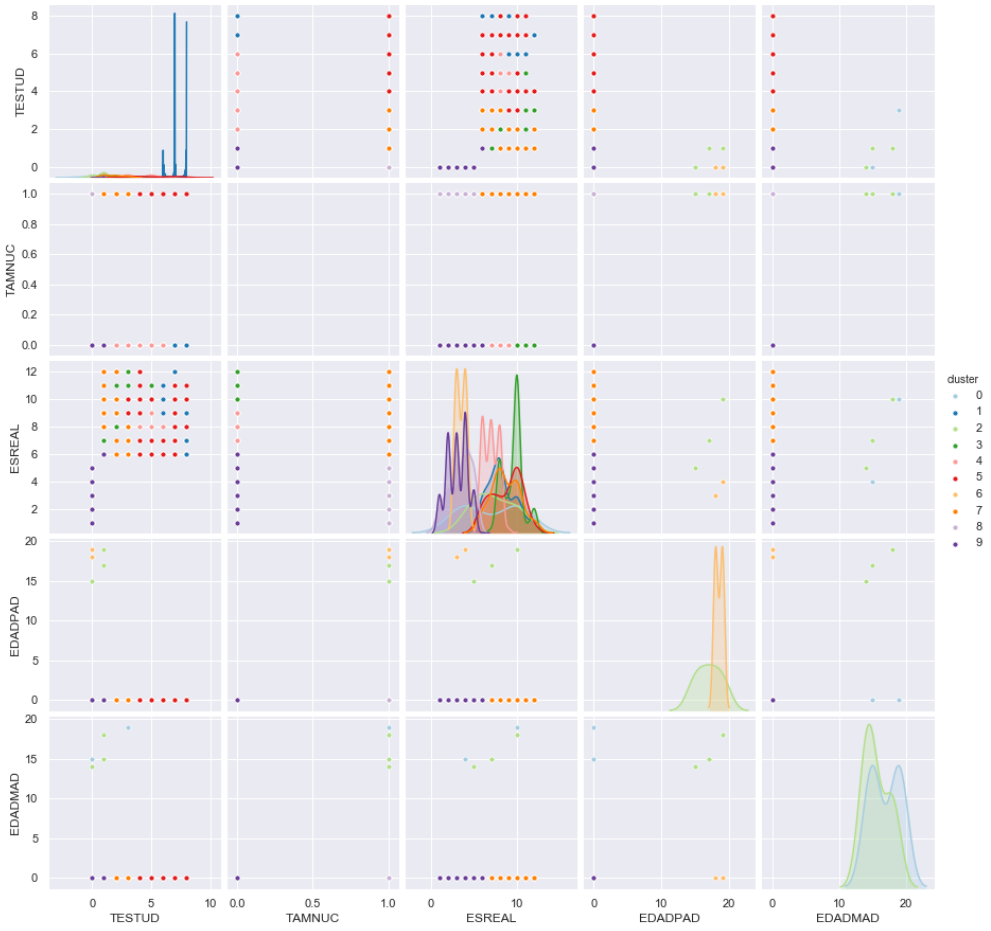

Figura 2.68: Scatter Matrix usando Birch en el caso de estudio 3.

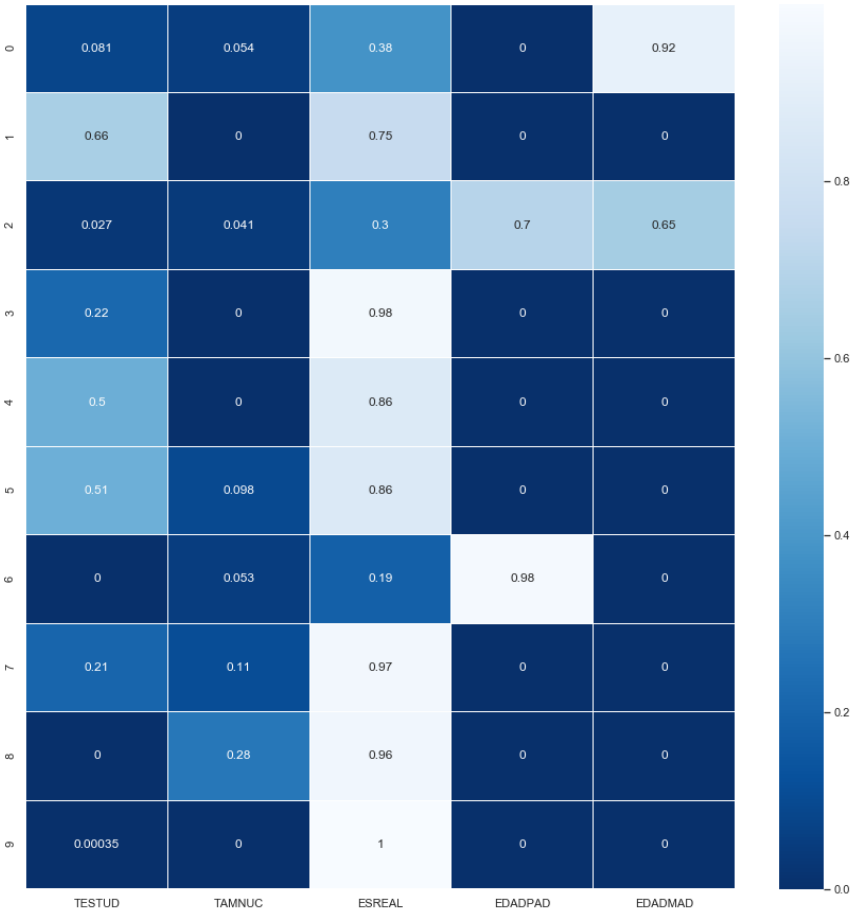

Figura 2.69: HeatMap usando Birch en el caso de estudio 3.

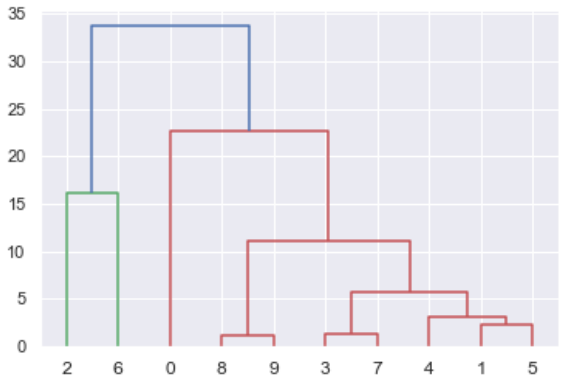

Figura 2.70: Dendograma usando Birch en el caso de estudio 3.

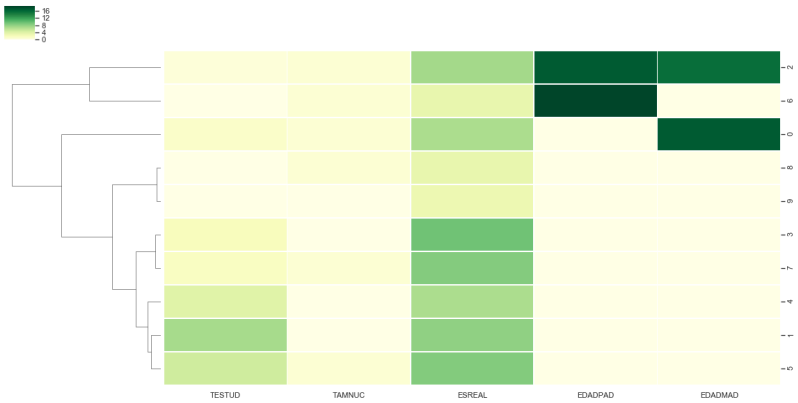

Figura 2.71: HeatMap con Dendograma usando Birch en el caso de estudio 3.

	TESTUD	TAMNUC	ESREAL	EDADPAD	EDADMAD
0	1.500000	1.0	7.000000	0.0	17.000000
1	7.217391	0.0	8.231884	0.0	0.000000
2	0.666667	1.0	7.333333	17.0	15.666667
3	2.116883	0.0	9.566234	0.0	0.000000
4	4.077419	0.0	7.000000	0.0	0.000000
5	5.194805	1.0	8.779221	0.0	0.000000
6	0.000000	1.0	3.500000	18.5	0.000000
7	1.836364	1.0	8.690909	0.0	0.000000
8	0.000000	1.0	3.490566	0.0	0.000000
9	0.001073	0.0	3.075644	0.0	0.000000

Figura 2.72: Medias de los datos seleccionados por cluster Birch caso 3.

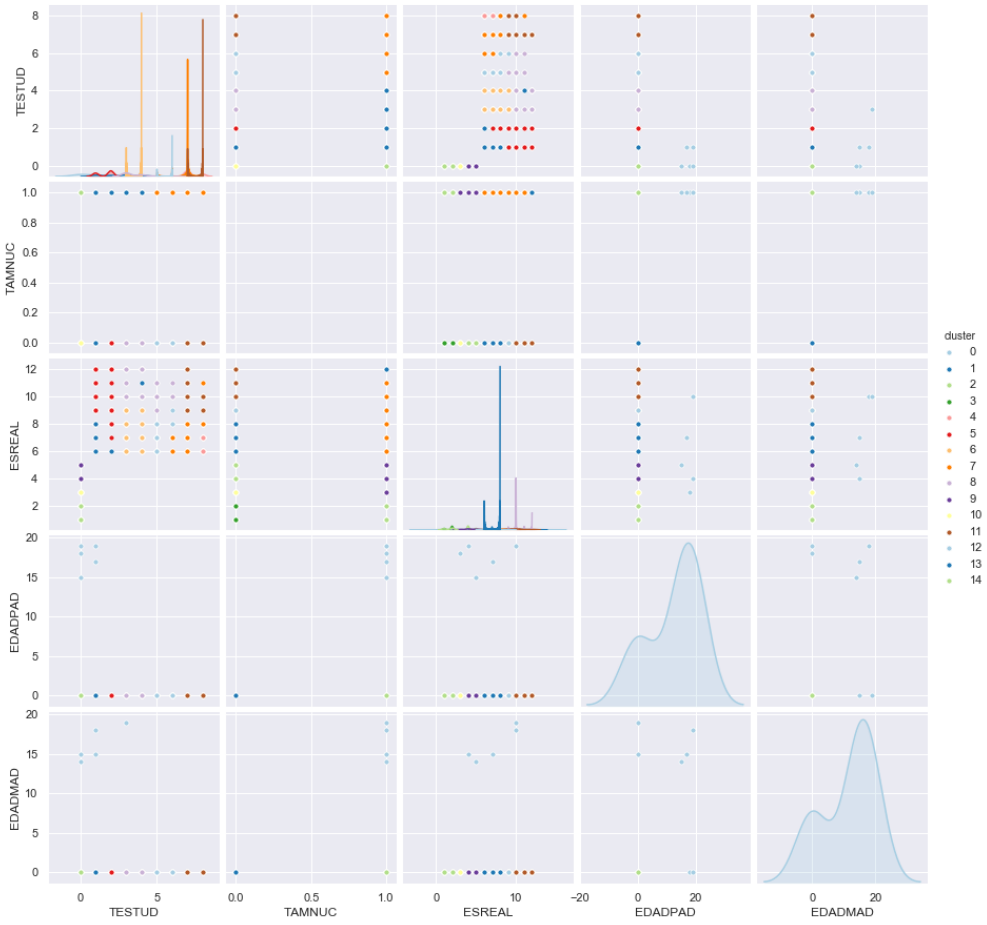

Figura 2.73: Scatter Matrix usando Ward en el caso de estudio 3.

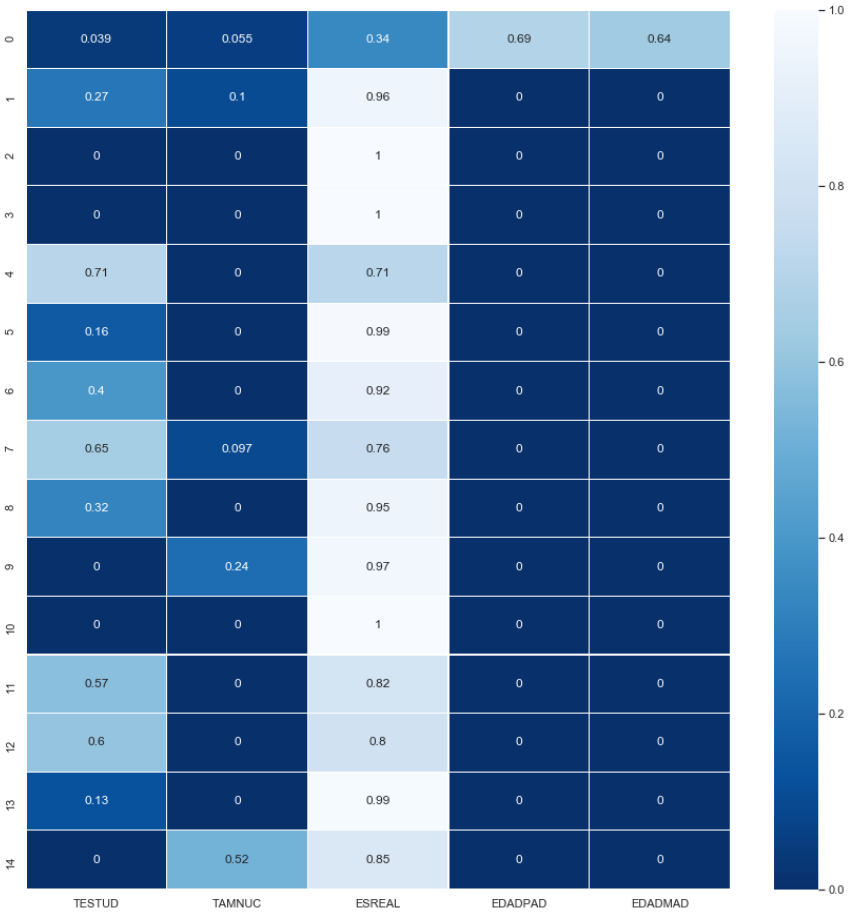

Figura 2.74: HeatMap usando Ward en el caso de estudio 3.

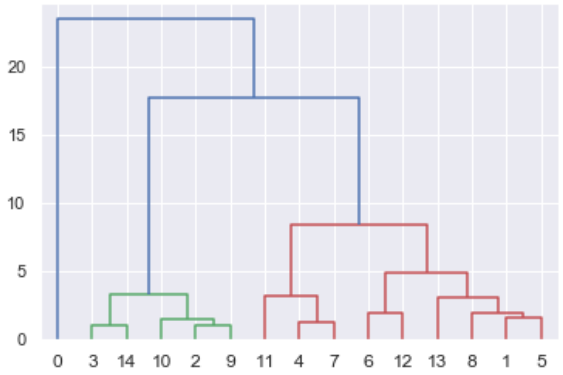

Figura 2.75: Dendograma usando Ward en el caso de estudio 3.

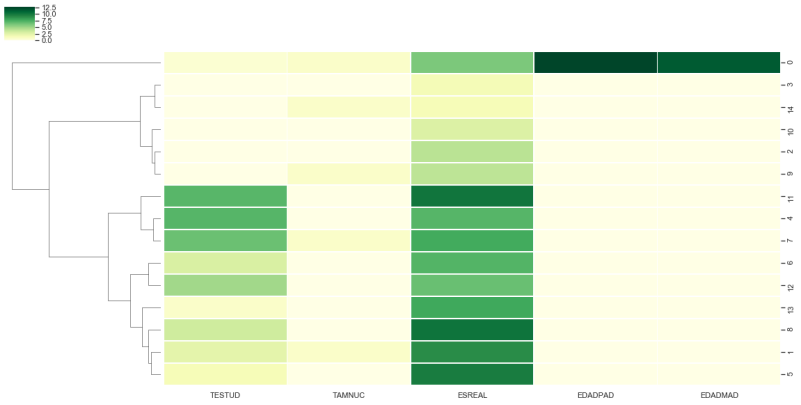

Figura 2.76: HeatMap con Dendograma usando Ward en el caso de estudio 3.

	TESTUD	TAMNUC	ESREAL	EDADPAD	EDADMAD
0	0.714286	1.0	6.142857	12.571429	11.571429
1	2.591398	1.0	9.129032	0.000000	0.000000
2	0.000000	0.0	4.264201	0.000000	0.000000
3	0.000000	0.0	1.721617	0.000000	0.000000
4	7.227586	0.0	7.241379	0.000000	0.000000
5	1.624204	0.0	9.853503	0.000000	0.000000
6	3.181818	0.0	7.363636	0.000000	0.000000
7	6.666667	1.0	7.820513	0.000000	0.000000
8	3.518248	0.0	10.328467	0.000000	0.000000
9	0.000000	1.0	4.135593	0.000000	0.000000
10	0.000000	0.0	3.000000	0.000000	0.000000
11	7.174603	0.0	10.222222	0.000000	0.000000
12	5.045455	0.0	6.727273	0.000000	0.000000
13	1.000000	0.0	7.913978	0.000000	0.000000
14	0.000000	1.0	1.634146	0.000000	0.000000

Figura 2.77: Medias de los datos seleccionados por cluster Ward caso 3.

Figura 2.78: Scatter Matrix usando MeanShift en el caso de estudio 3.

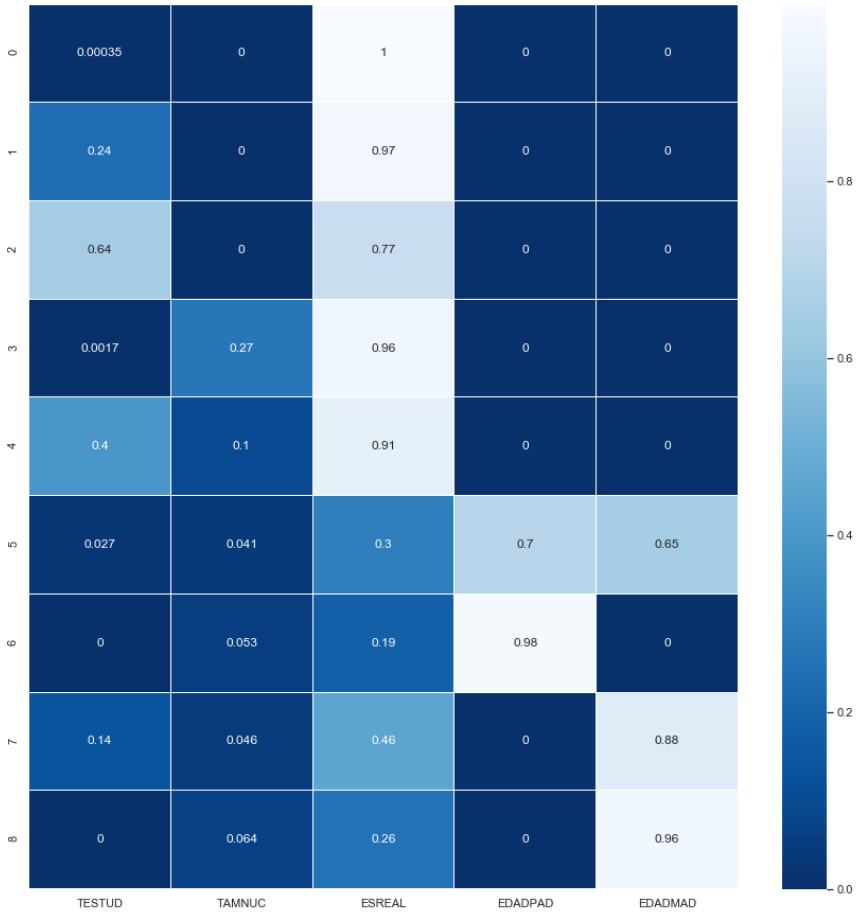

Figura 2.79: HeatMap usando MeanShift en el caso de estudio 3.

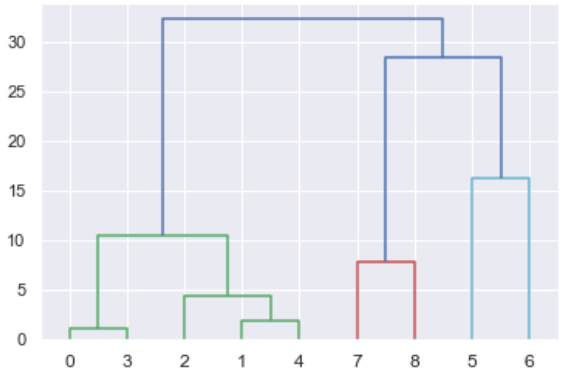

Figura 2.80: Dendograma usando MeanShift en el caso de estudio 3.

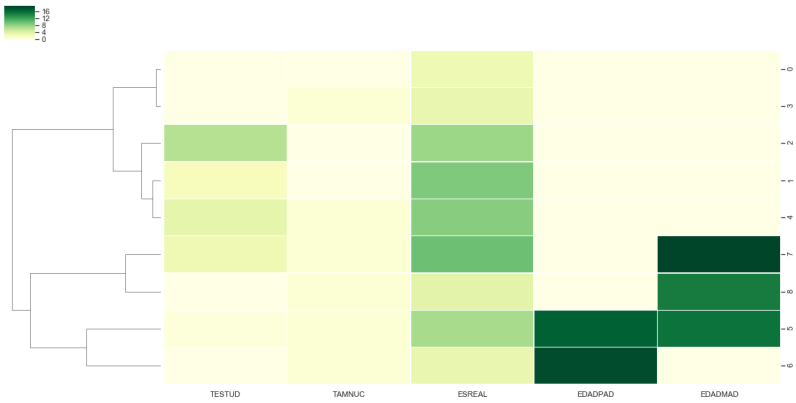

Figura 2.81: HeatMap con Dendograma usando MeanShift en el caso de estudio 3.

	TESTUD	TAMNUC	ESREAL	EDADPAD	EDADMAD
0	0.001073	0.0	3.075644	0.0	0.000000
1	2.252723	0.0	9.132898	0.0	0.000000
2	6.621528	0.0	7.916667	0.0	0.000000
3	0.006250	1.0	3.506250	0.0	0.000000
4	3.816794	1.0	8.763359	0.0	0.000000
5	0.666667	1.0	7.333333	17.0	15.666667
6	0.000000	1.0	3.500000	18.5	0.000000
7	3.000000	1.0	10.000000	0.0	19.000000
8	0.000000	1.0	4.000000	0.0	15.000000

Figura 2.82: Medias de los datos seleccionados por cluster MeanShift caso 2.

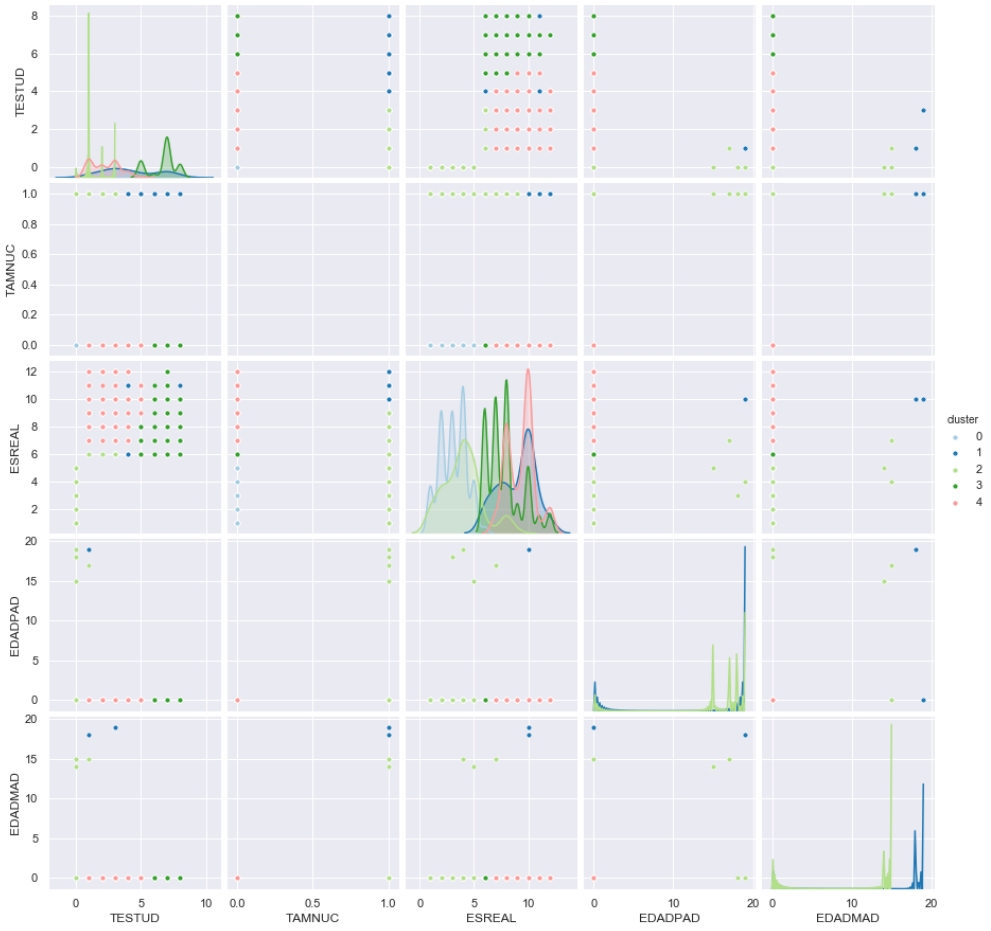

Figura 2.83: Scatter Matrix usando spectral en el caso de estudio 3.

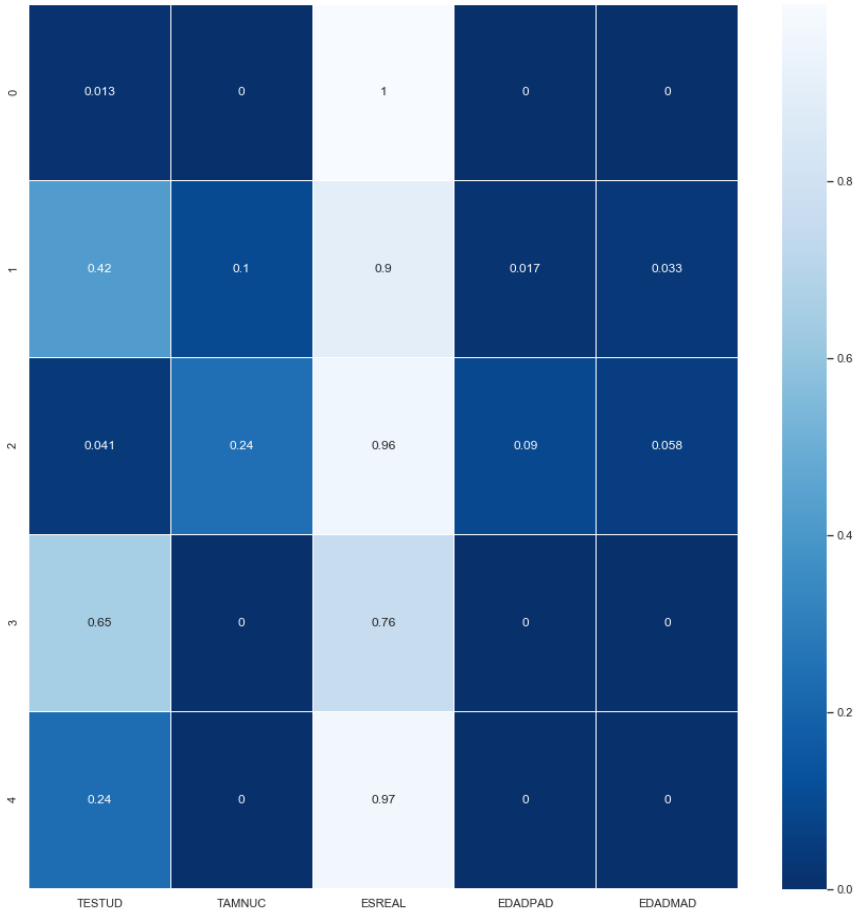

Figura 2.84: HeatMap usando spectral en el caso de estudio 3.

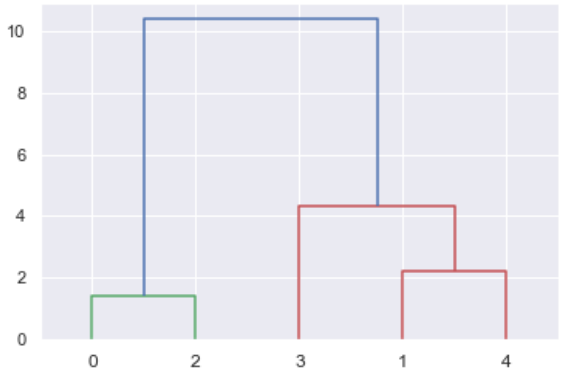

Figura 2.85: Dendograma usando spectral en el caso de estudio 3.

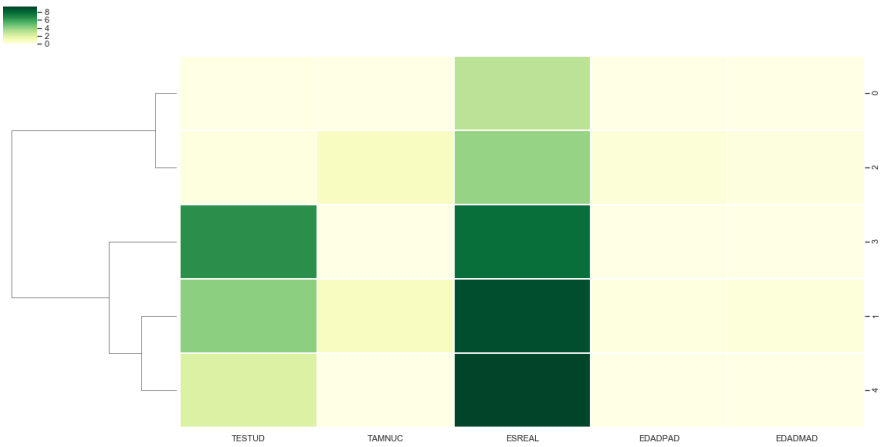

Figura 2.86: HeatMap con Dendograma usando spectral en el caso de estudio 3.

	TESTUD	TAMNUC	ESREAL	EDADPAD	EDADMAD
0	0.040276	0.0	3.111288	0.000000	0.000000
1	4.203540	1.0	9.000000	0.168142	0.327434
2	0.167568	1.0	3.972973	0.372973	0.237838
3	6.679856	0.0	7.838129	0.000000	0.000000
4	2.264574	0.0	9.316143	0.000000	0.000000

Figura 2.87: Medias de los datos seleccionados por cluster spectral caso 3.

2.3.6. Algoritmos modificados en el caso 3

En esta sección se va a exponer la modificación de los parámetros de dos algoritmos distintos y para ver sus diferencias se van a comparar los resultados de las métricas obtenidas en las secciones previas.

El primer algoritmo que vamos a modificar es Birch , intentando aumentar el valor de sus métricas reduciendo el numero de clusters a 10 a 5 y el segundo algoritmo que vamos a modificar es MeanShift, modificando el valor de quantile de 0.4 a 0.2 para ver que resultados obtenemos y compararlos con la anterior ejecucion.

La figura 2.88 se muestra la antigua tabla pero ahora con las métricas de la ejecucion de estos dos algoritmos modificados. En ella se aprecia que las modificaciones de los parámetros no han sido exitosas en el caso del algoritmo Birch, pero si que han incrementado un poco la métrica HC del algoritmo MeanShift .

	N Clusters	HC metric	SC metric	Time
K-means	10	6943.576223	0.679065	0.045921
Birch	10	4081.202946	0.697422	0.043426
Ward	15	7978.888763	0.696960	0.173204
MeanShift	9	3675.392770	0.714534	0.059399
Spectral	5	6059.545108	0.706452	1.043716
Birch_modificado	5	2615.276048	0.690900	0.042927
meanshift_modificado	10	3757.968423	0.715982	0.056404

Figura 2.88: Metricas obtenidas usando algoritmos modificados en el caso de estudio 3.

2.3.7. Algoritmo modificado Birch caso 3

La figura 2.89 representa como están distribuidos los diferentes clusters sobre las diferentes variables con el algoritmo Birch

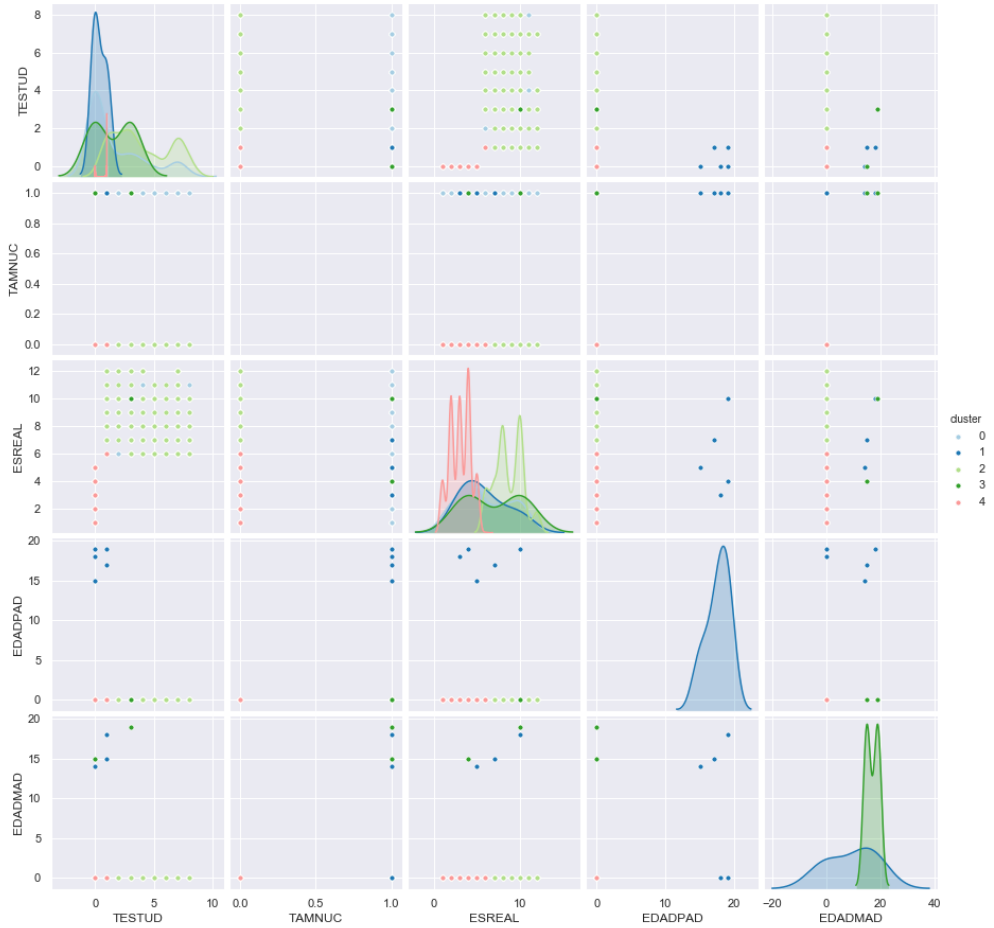

Figura 2.89: Scatter Matrix usando Birch modificado en el caso de estudio 3.

A continuación se muestra la figura 2.90 , esta es la fusión de las gráficas de Heatmap y de Dendograma con el algoritmo Birch modificado.

2.3.8. Algoritmo modificado MeanShift caso 3

La figura 2.91 representa como están distribuidos los diferentes clusters sobre las diferentes variables con el algoritmo MeanShift modificado

A continuación se muestra la figura 2.92 , esta es la fusión de las gráficas de Heatmap y de Dendograma con el algoritmo MeanShift modificado.

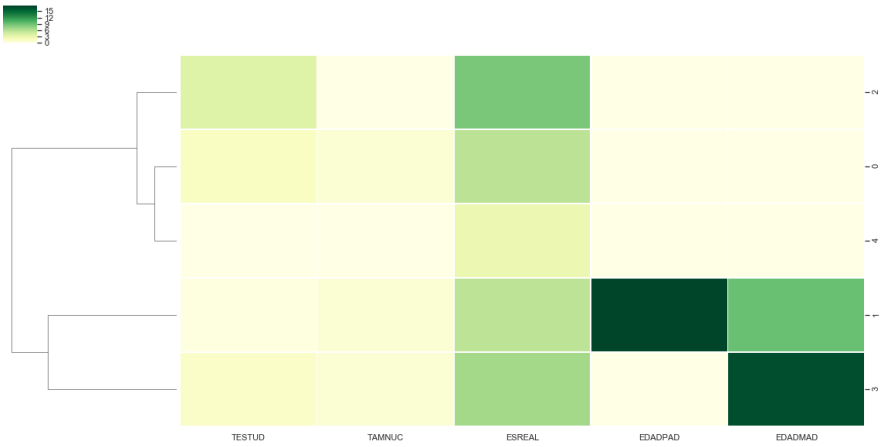

Figura 2.90: HeatMap con Dendograma usando Birch modificado en el caso de estudio 3.

2.3.9. Interpretacion de la segmentacion caso 3

Para terminar con el estudio de este caso , vamos a interpretar las visualizaciones producidas por las ejecuciones de los algoritmos.

Para este caso de estudio, en las gráficas correspondientes a los algoritmos del caso 3 podemos apreciar que las edades de los padres y las madres , como es normal , se asemejan mucho, estando en el rango 60 - 90 . También hay personas que no viven con sus padres, en el scatterMatrix tienen el valor 0 porque en el pre procesado se puso este valor a todos los huecos vacíos, que en nuestro caso son personas que no conviven con su padre o madre. El tamaño del núcleo familiar no supera las dos personas ,

En el caso del algoritmo K-means podemos ver como en los clusters 0 t 9 se han agrupado aquellas personas que tienen muchos valores 0 en sus atributos, tanto en la edad del padre y madre, que significa que no convive con ellos , como en el tipo de estudios , que quiere decir que tiene unos estudios realizados con un valor de 5 o menos , tres en nuestro caso , que significa que fueron mas de 5 años a la escuela pero no acabaron sus estudios primarios. En el cluster 8 tenemos aquellas personas con una licenciatura, arquitectura ,ingeniería o doctorado que no viven con su padre y su madre.

Figura 2.91: Scatter Matrix usando MeanShift modificado en el caso de estudio 3.

3. Bibliografía.

Referencias

[1] Instituto Nacional de Estadistica
http://www.ine.es/censos2011_datos/cen11_datos_microdatos.htm

[2] Python ORG
https://www.python.org/downloads/release/python-370/

[3] K-means
https://www.datascience.com/blog/k-means-clustering

[4] Spectral Clustering

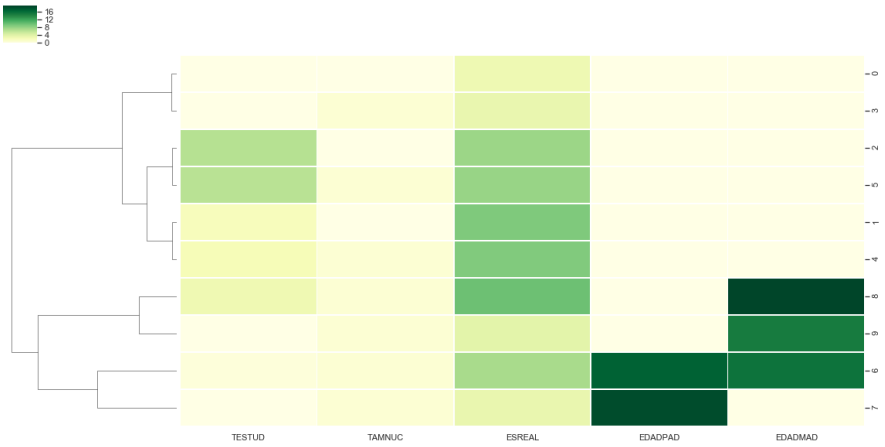

Figura 2.92: HeatMap con Dendograma usando MeanShift modificado en el caso de estudio 3.

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html

[5] Mean Shift
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html

[6] Ward
https://scikit-learn.org/stable/modules/clustering.html

[7] Birch
https://scikit-learn.org/stable/modules/clustering.html