

UNIVERSIDAD DE GRANADA



**UNIVERSIDAD
DE GRANADA**

Carlos de Alonso Andrés, Grupo 1
Doble Grado en Ingeniería Informática y Matemáticas
e.carlosdealonso@go.ugr.es

Inteligencia de Negocio

Práctica 2:
Análisis Relacional Mediante Segmentación

Curso 2022-2023

Índice

1.	Introducción	3
	Conjunto de datos.....	3
	Algoritmos utilizados.....	4
	Métricas utilizadas	4
2.	Caso de estudio 1	5
	K-means.....	6
	Meanshift	11
	DBSCAN	15
	Agglomerative Clustering.....	19
	Birch	21
	Interpretación de la segmentación.....	21
3.	Caso de estudio 2	25
	K-means.....	26
	Meanshift	31
	DBSCAN	35
	Agglomerative Clustering.....	39
	Birch	41
	Interpretación de la segmentación.....	41
4.	Caso de estudio 3	45
	K-means.....	46
	Meanshift	48
	DBSCAN	55
	Agglomerative Clustering.....	59
	Birch	61
	Interpretación de la segmentación.....	61
5.	Contenido adicional	65
6.	Bibliografía	65

1. Introducción

En esta segunda práctica se utilizará la metodología de aprendizaje no supervisado mediante el uso de algoritmos de clustering, que realizarán un análisis relacional mediante segmentación de los datos proporcionados por *Kaggle* de incendios ocurridos en Toronto entre 2011 y 2018, consta de 30 variables y un total de 11.214 incendios. El objetivo es definir tres casos de estudio de interés con el fin de encontrar relaciones entre distintas variables mediante estos algoritmos.

Para cada caso se deciden las variables a estudiar, se aplican diversos filtros para obtener un subconjunto de datos nuevo y se obtienen resultados por cada algoritmo utilizado. Además, se realizará una interpretación global de los resultados obtenidos con el fin de conseguir una conclusión clara del problema. Para los algoritmos K-means y DBSCAN realizaremos una comparación de su ejecución en función de los parámetros que hay que establecer *a priori*.

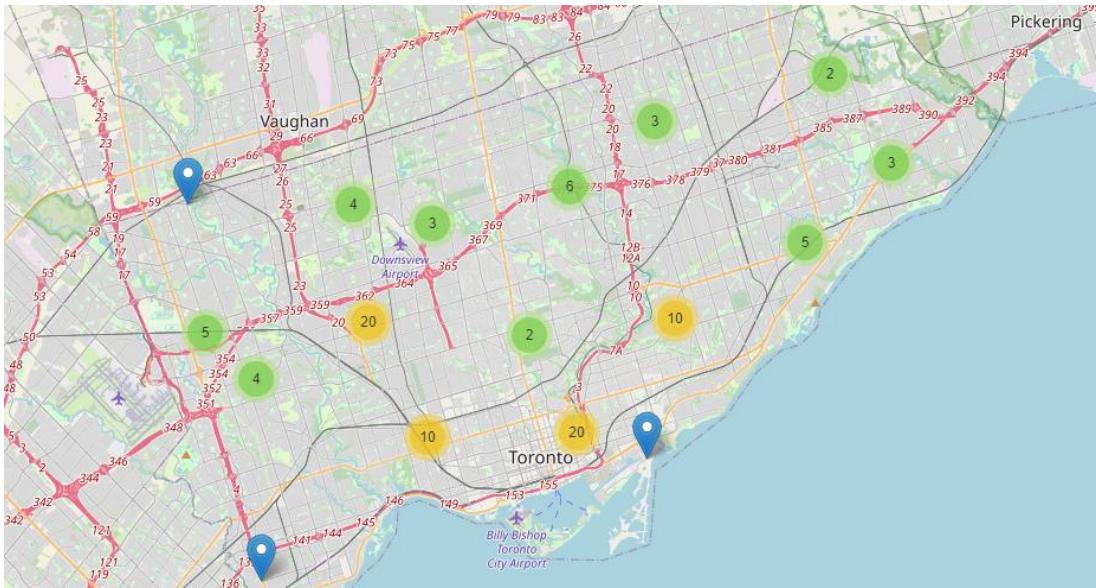
NOTA: El código tiene unas variables booleanas que en caso de ser *True* ejecutarán el todo el proceso del caso al que correspondan. También ocurre que el código tarda mucho en crear algunas gráficas como las de Meanshift, si solo quieras ver clusters y medidas comente la línea *visualizar()* en *FuncionesP2.py*.

Conjunto de datos

Area_of_Origin	Zona de origen del incendio
Business_Impact	Impacto a nivel de negocio
Civilian_Casualties	Víctimas encontradas en la escena
Count_of_Persons_Rescued	Número de personas rescatadas
Estimated_Dollar_Loss	Coste estimado
Estimated_Number_Of_Persons_Displaced	Número estimado de personas desplazadas
Ext_agent_app_or_defer_time	
Extent_Of_Fire	Alcance del incendio
Fire_Alarm_System_Impact_on_Evacuation	Impacto de la alarma en la evacuación
Fire_Alarm_System_Operation	Funcionamiento del sistema de alarma
Fire_Alarm_System_Presence	Presencia de sistema de alarma
Fire_Under_Control_Time	Fecha y hora de control del incendio
Ignition_Source	Origen del incendio
Incident_Station_Area	Área de la estación del incidente
Incident_Ward	
Last_TFS_Unit_Clear_Time	Ultimo instante de tiempo de extinción
Latitude	Latitud
Longitude	Longitud
Material_First_Ignited	Material que inició el incendio
Method.Of_Fire_Control	Método de control del incendio
Possible_Cause	Possible causa del incendio
Property_Use	Uso de la propiedad
Smoke_Alarm_at_Fire-Origin_Alarm_Failure	Causa del fallo en el sistema de alarma en el origen del incendio
Smoke_Alarm_at_Fire-Origin_Alarm_Type	Tipo de sistema de alarma contra incendios
Status_of_Fire_On_Arrival	Estado del incendio en
TFS_Alarm_Time	Instante de tiempo que sonó la alarma
TFS_Arrival_Time	Instante de tiempo de llegada de servicios
Arrival_Time	Tiempo que tardaron en llegar
Week	Día de la semana que ocurrió
Month	Mes del año que ocurrió

Tabla 1: Variables de los datos

Podemos observar como una forma de segmentación que nos proporciona la página de Toronto es a partir de los lugares donde ocurrió el incendio, INSERTAR. Nosotros compararemos los clusters generados tanto en el conjunto escogido como en el complementario, para ver que conclusiones podemos sacar del filtrado escogido para los datos.



Algoritmos utilizados

- **K-means:** es un algoritmo que trata de separar los ejemplos en un número prefijado de clusters. A partir de uno centros iniciales aleatorios calcula los ejemplos más cercanos y los vuelve a recalcular, este proceso se va repitiendo hasta que no hay más cambios.
- **Meanshift:** Este algoritmo no utiliza centroides, sino que utiliza el concepto de densidad, a partir de un radio prefijado, determina el número de clusters y desplaza los centros hacia regiones con mayor densidad.
- **DBSCAN:** Del mismo modo que el algoritmo anterior, utiliza dos parámetros para asignar los clusters. El primero es ϵ psilon, que determina cuando un objeto es alcanzable a partir de otro, y el segundo es el número mínimo de elementos por los que debemos alcanzar a otro para que pertenezcan al mismo cluster.
- **Agglomerative Clustering:** Es un algoritmo jerárquico de abajo hacia arriba, utilizaremos el *linkage=Ward* que va agrupando los clusters dos clusters en cada paso, tratando de minimizar la varianza.
- **Birch:** es un algoritmo de clustering incremental, crea un árbol con las características de los clusters guardando únicamente la información necesaria para separarlos. De cada nodo cuelgan ciertos subclústers, cuyo número es acotado por un factor de ramificación y un umbral (que determina si una instancia está lo suficientemente cerca del centro).

Métricas utilizadas

- Tiempo de ejecución: tiempo en segundos de ejecución de cada algoritmo.
- Número de Clusters: útil para aquellos algoritmos que no lo tienen fijado *a priori*.
- Coeficiente de Silhouette: mide como de parecidos son los objetos de un mismo cluster, en comparación con los demás clusters. Toma valores entre [-1,1], donde un valor alto indica que el objeto está bien emparejado con los demás del cluster y mal emparejado con los otros clusters.
- Índice de Calinski-Harabasz: Es un cociente entre la dispersión intra-cluster e inter-cluster. Un valor alto nos indica que los grupos están bien separados.

2. Caso de estudio 1

En el primer caso se estudiarán aquellos incendios que ocurrieron en los meses de verano y que fueron apagados por un departamento de bomberos, así compararemos con el resto del año. Puede ser interesante ver si existe alguna característica que los agrupó dependiendo de cuando ocurre el incendio. Se han escogido un total de 10980 ejemplos donde los meses van desde junio hasta septiembre y la variable *Method_Of_Fire_Control* coincide con *Extinguished by fire department*. Por otro lado las variables estudiadas serán:

- *Civilian_Casualties* -> Muertos
- *Count_Of_Person_Rescued* -> Rescatados
- *Estimated_Dollar_Loss* -> Coste
- *Estimated_Number_Of_Person_Displaced* -> Desplazados
- *Arrival_Time* -> Tiempo

Para la ejecución de este caso se han obtenido las siguientes medidas, donde el caso de estudio se puede observar una ventaja en el algoritmo de Birch tanto en el coeficiente de Silhouette como en el índice de Calinski-Harabasz. Por otro lado en el conjunto complementario el que parece que ha obtenido mejores resultados es DBSCAN, aunque el índice de Calinski no es muy alto, lo que puede significar que las agrupaciones no están bien separadas.

	Silhouette	Calinski-Harabasz	Tº ejecución	Número de Clusters
Kmeans	0.86732	2770.079	0.04	2
Meanshift	0.33107	384.090	0.45	28
DBSCAN	0.86992	1570.883	0.19	2
Birch	0.87002	2768.048	0.04	2
AC			0.03	2

Tabla 2: Métricas Caso 1

	Silhouette	Calinski-Harabasz	Tº ejecución	Número de Clusters
Kmeans	0.73798	1697.073	0.02	2
Meanshift	0.53373	588.104	0.55	28
DBSCAN	0.87811	140.003	0.26	1
Birch	0.79254	660.742	0.04	2
AC			0.02	3

Tabla 3: Métricas Complementario Caso 1

K-means

En el algoritmo de K-means podemos apreciar como conforme el número de clusters aumenta, el coeficiente de Silhouette y el índice de Calinski-Harabasz va aumentando a partir de una cota superior que encontramos en la configuración de dos clusters.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
2	0: 2150 (97.91%) 1: 46 (2.09%)	0.86732	2770.079	0.04
3	0: 1319 (60.06%) 2: 831 (37.84%) 1: 46 (2.09%)	0.47026	2830.636	0.04
4	1: 1002 (45.63%) 0: 760 (34.61%) 3: 388 (17.67%) 2: 46 (2.09%)	0.39547	2406.200	0.05
5	0: 966 (43.99%) 2: 815 (37.11%) 3: 368 (16.76%) 1: 44 (2.00%) 4: 3 (0.14%)	0.40152	2265.712	0.05
6	0: 950 (43.26%) 3: 900 (40.98%) 2: 271 (12.34%) 1: 46 (2.09%) 5: 26 (1.18%) 4: 3 (0.14%)	0.42752	2294.536	0.06
7	1: 872 (39.71%) 6: 657 (29.92%) 3: 304 (13.84%) 0: 197 (8.97%) 4: 117 (5.33%) 2: 46 (2.09%) 5: 3 (0.14%)	0.44598	2363.129	0.06
8	2: 868 (39.53%) 0: 613 (27.91%) 6: 380 (17.30%) 4: 168 (7.65%) 3: 116 (5.28%) 1: 46 (2.09%) 5: 3 (0.14%) 7: 2 (0.09%)	0.44502	2402.641	0.08
9	0: 771 (35.11%) 7: 602 (27.41%) 4: 330 (15.03%) 6: 284 (12.93%) 3: 115 (5.24%) 2: 45 (2.05%) 1: 44 (2.00%) 5: 3 (0.14%) 8: 2 (0.09%)	0.44322	2418.664	0.09
10	0: 776 (35.34%) 4: 616 (28.05%) 2: 330 (15.03%) 6: 278 (12.66%) 3: 116 (5.28%) 1: 46 (2.09%) 8: 28 (1.28%) 5: 3 (0.14%) 9: 2 (0.09%) 7: 1 (0.05%)	0.45072	2501.035	0.10

Tabla 4: Comparación K-means Caso 1

Algo similar ocurre en el caso complementario, dónde las mejores medidas las obtenemos para 3 clusters, aunque después va aumentando hasta que alcanza un máximo loca y comienza a decrecer ambas medidas

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
2	1: 2534 (90.53%) 0: 265 (9.47%)	0.73798	1697.073	0.02
3	0: 2516 (89.89%) 1: 265 (9.47%) 2: 18 (0.64%)	0.75859	2481.575	0.02
4	1: 1554 (55.52%) 3: 962 (34.37%) 2: 265 (9.47%) 0: 18 (0.64%)	0.52619	2571.836	0.04
5	4: 1532 (54.73%) 0: 984 (35.16%) 3: 232 (8.29%) 1: 33 (1.18%) 2: 18 (0.64%)	0.53609	2686.196	0.04
6	1: 1556 (55.59%) 3: 958 (34.23%) 0: 232 (8.29%) 4: 33 (1.18%) 2: 17 (0.61%) 5: 3 (0.11%)	0.53867	2827.346	0.11
7	2: 1555 (55.56%) 0: 958 (34.23%) 6: 128 (4.57%) 5: 104 (3.72%) 1: 33 (1.18%) 3: 18 (0.64%) 4: 3 (0.11%)	0.52194	2458.487	0.07
8	0: 1190 (42.52%) 6: 831 (29.69%) 3: 491 (17.54%) 2: 232 (8.29%) 4: 33 (1.18%) 1: 18 (0.64%) 5: 3 (0.11%) 7: 1 (0.04%)	0.50216	2863.473	0.08
9	0: 1192 (42.59%) 7: 831 (29.69%) 3: 490 (17.51%) 2: 232 (8.29%) 4: 33 (1.18%) 1: 17 (0.61%) 5: 2 (0.07%) 8: 1 (0.04%) 6: 1 (0.04%)	0.50102	2920.052	0.08
10	1: 1060 (37.87%) 4: 684 (24.44%) 6: 657 (23.47%) 5: 140 (5.00%) 9: 111 (3.97%) 0: 92 (3.29%) 3: 33 (1.18%) 2: 18 (0.64%) 7: 3 (0.11%) 8: 1 (0.04%)	0.48532	2813.139	0.09

Tabla 5: Comparación K-means Complementario Caso 1

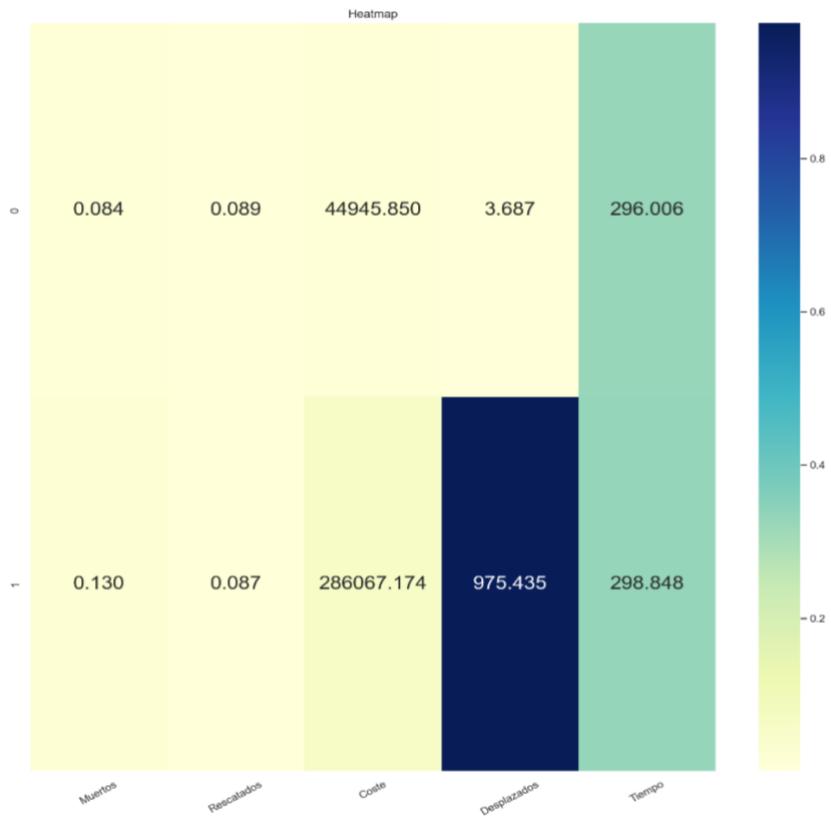


Figura 1: Heatmap K-means Caso 1

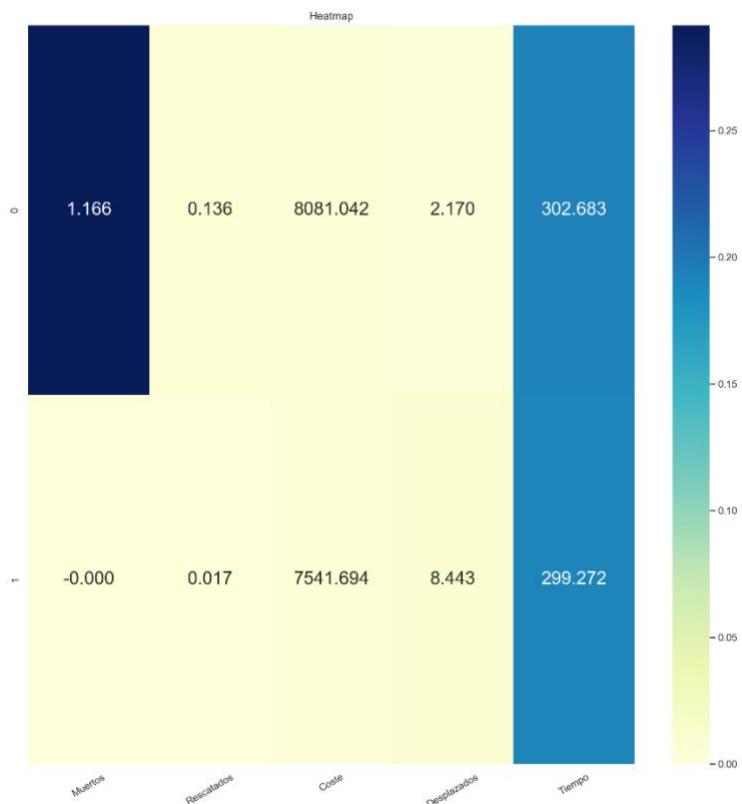


Figura 2: Heatmap K-means Complementario Caso 1

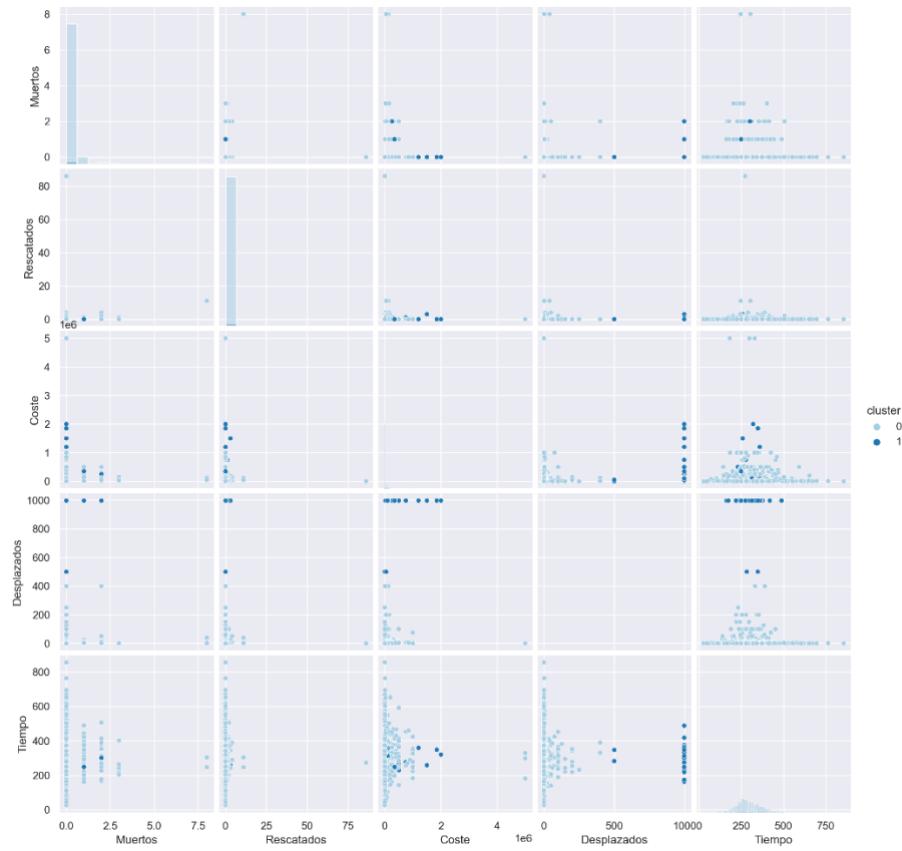


Figura 3: ScatterMatrix K-means Caso 1

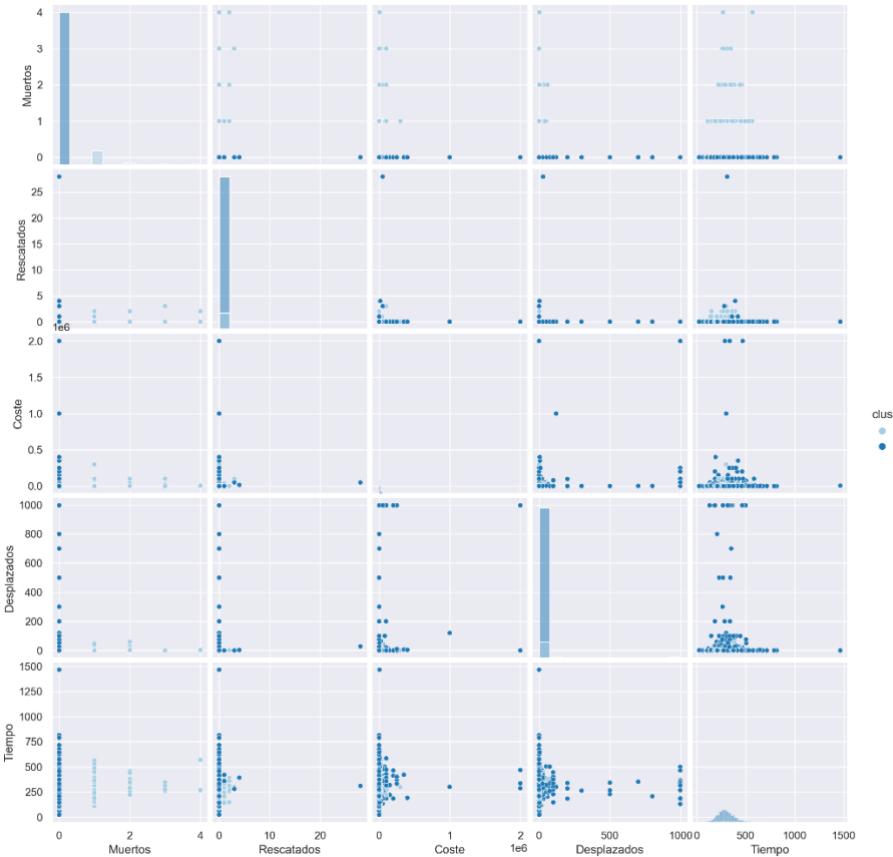


Figura 4: ScatterMatrix K-means Complementario Caso 1

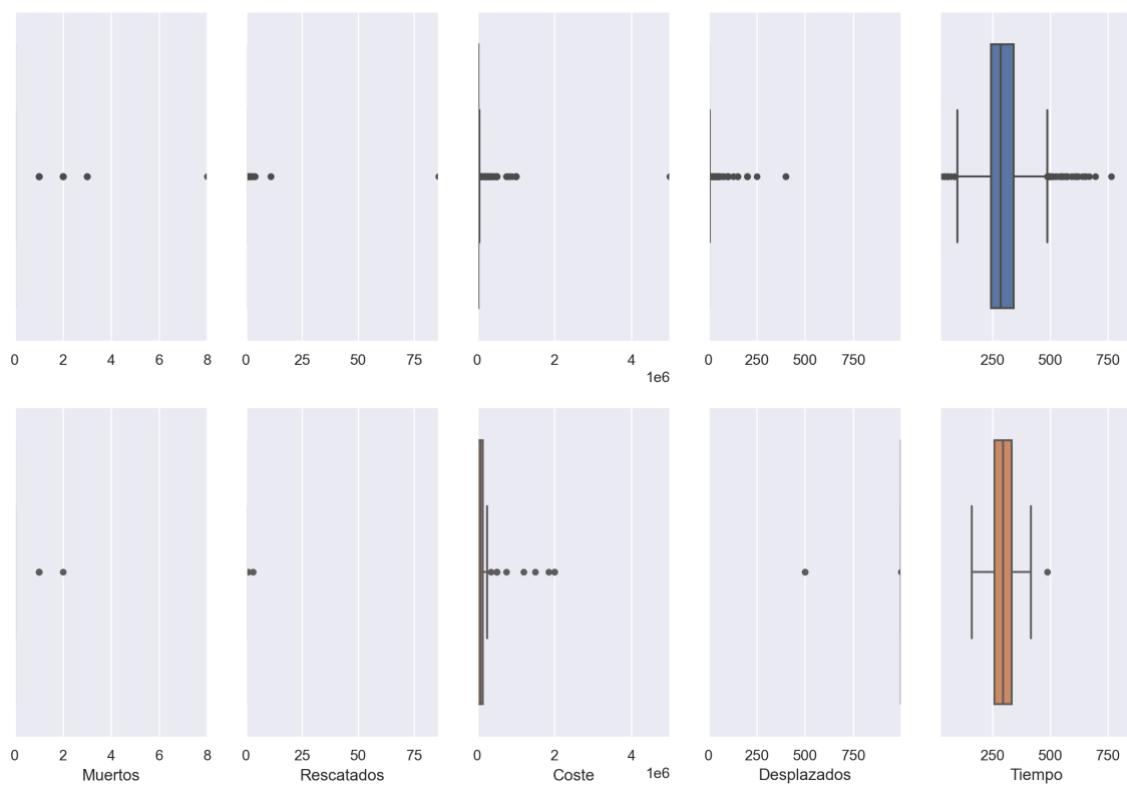


Figura 5: BoxPlot K-means Caso 1

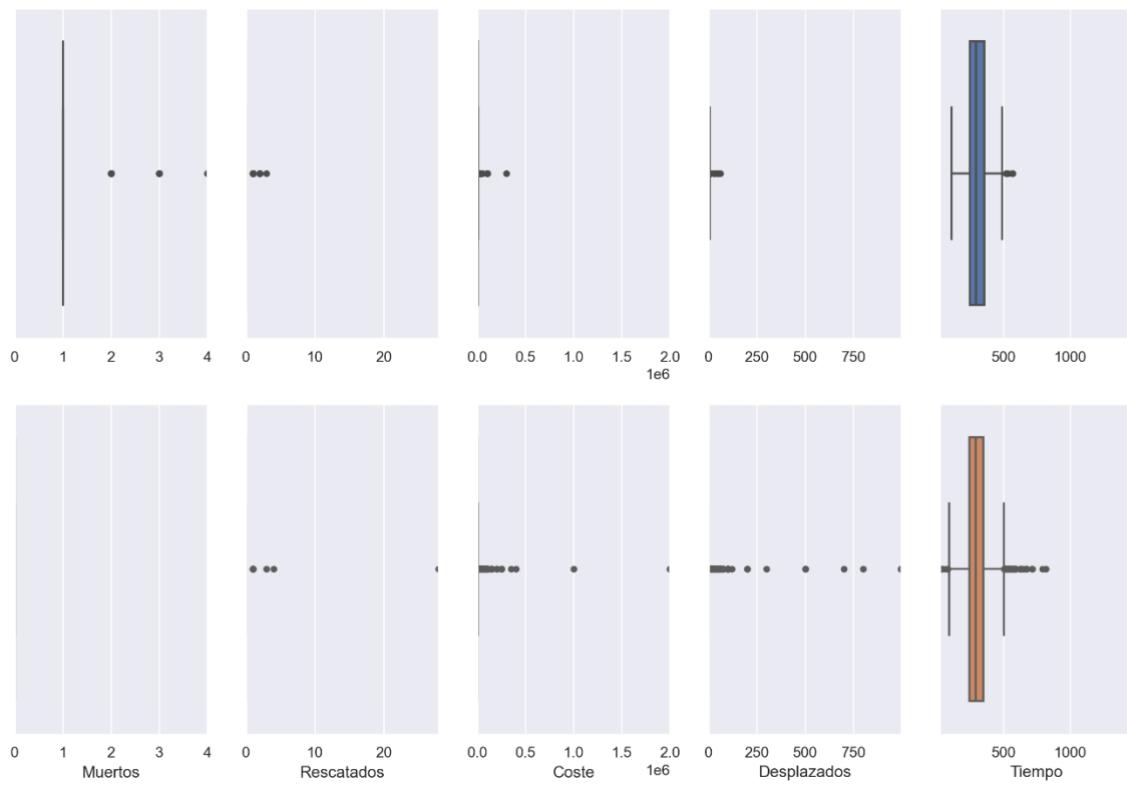


Figura 6: BoxPlot K-means Complementario Caso 1

Meanshift

Para el algoritmo meanshift se puede apreciar que se han obtenido mejores medidas para el caso complementario que para el que teníamos interés en estudiar.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
28	0: 1833 (83.47%) 1: 100 (4.55%) 26: 70 (3.19%) 12: 69 (3.14%) 2: 26 (1.18%) 4: 17 (0.77%) 18: 13 (0.59%) 3: 12 (0.55%) 14: 11 (0.50%) 6: 11 (0.50%) 5: 5 (0.23%) 9: 4 (0.18%) 10: 4 (0.18%) 7: 3 (0.14%) 8: 3 (0.14%) 11: 2 (0.09%) 13: 2 (0.09%) 17: 1 (0.05%) 19: 1 (0.05%) 20: 1 (0.05%) 24: 1 (0.05%) 22: 1 (0.05%) 25: 1 (0.05%) 21: 1 (0.05%) 15: 1 (0.05%) 23: 1 (0.05%) 16: 1 (0.05%) 27: 1 (0.05%)	0.33107	384.090	0.45

Tabla 6: Métricas Meanshift Caso 1

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
28	0: 2444 (87.32%) 1: 231 (8.25%) 5: 53 (1.89%) 2: 19 (0.68%) 11: 9 (0.32%) 3: 6 (0.21%) 4: 6 (0.21%) 8: 4 (0.14%) 7: 3 (0.11%) 6: 3 (0.11%) 9: 2 (0.07%) 10: 2 (0.07%) 22: 2 (0.07%) 21: 1 (0.04%) 27: 1 (0.04%) 18: 1 (0.04%) 26: 1 (0.04%) 20: 1 (0.04%) 17: 1 (0.04%) 24: 1 (0.04%) 25: 1 (0.04%) 13: 1 (0.04%) 23: 1 (0.04%) 16: 1 (0.04%) 12: 1 (0.04%) 15: 1 (0.04%) 14: 1 (0.04%) 19: 1 (0.04%)	0.53373	588.104	0.55

Tabla 7: Métricas Meanshift Complementario Caso 1

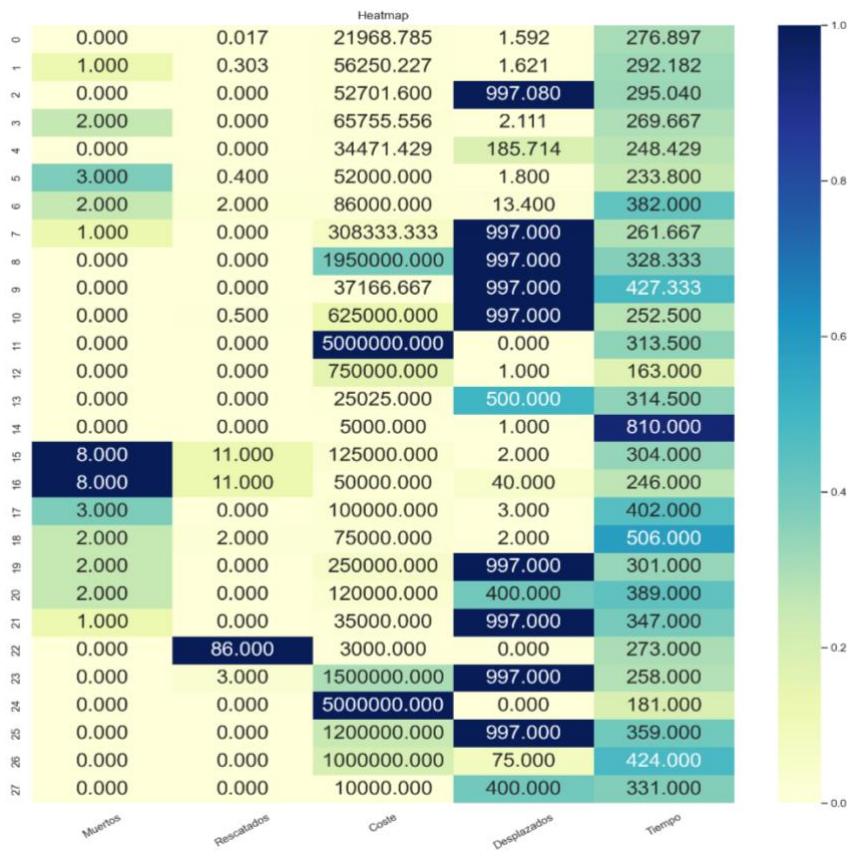


Figura 7: Heatmap Meanshift Caso 1

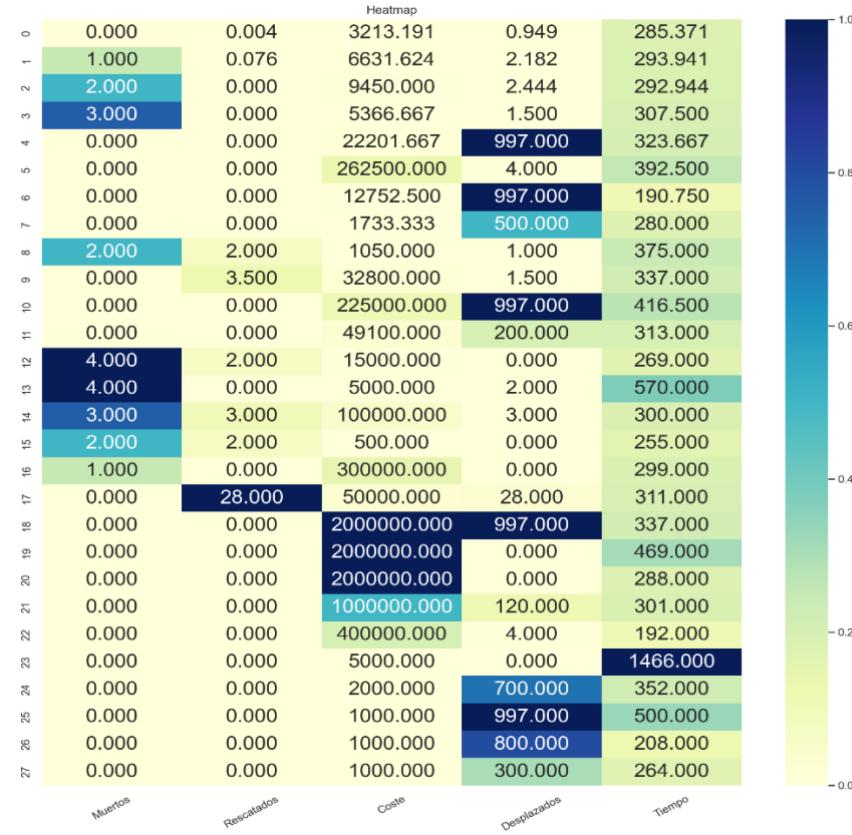


Figura 8: Heatmap Meanshift Complementario Caso 1



Figura 9: ScatterMatrix Meanshift Caso 1



Figura 10: ScatterMatrix Meanshift Complementario Caso 1

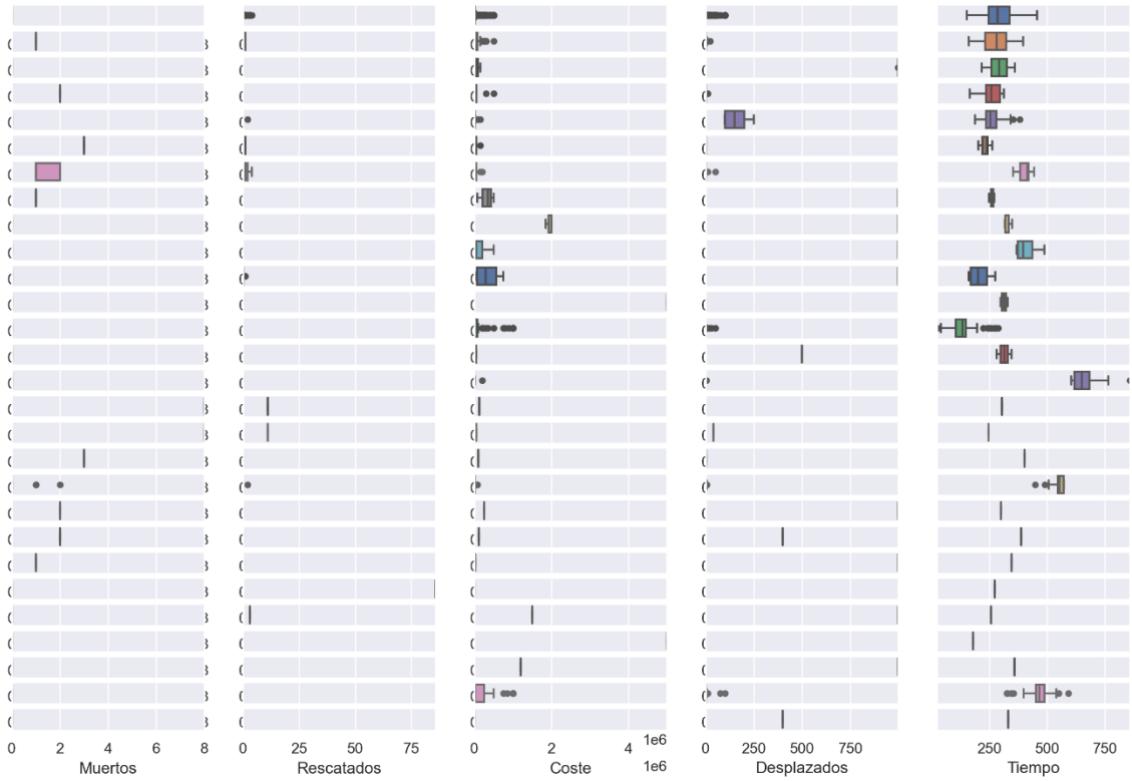


Figura 11: BoxPlot Meanshift Caso 1

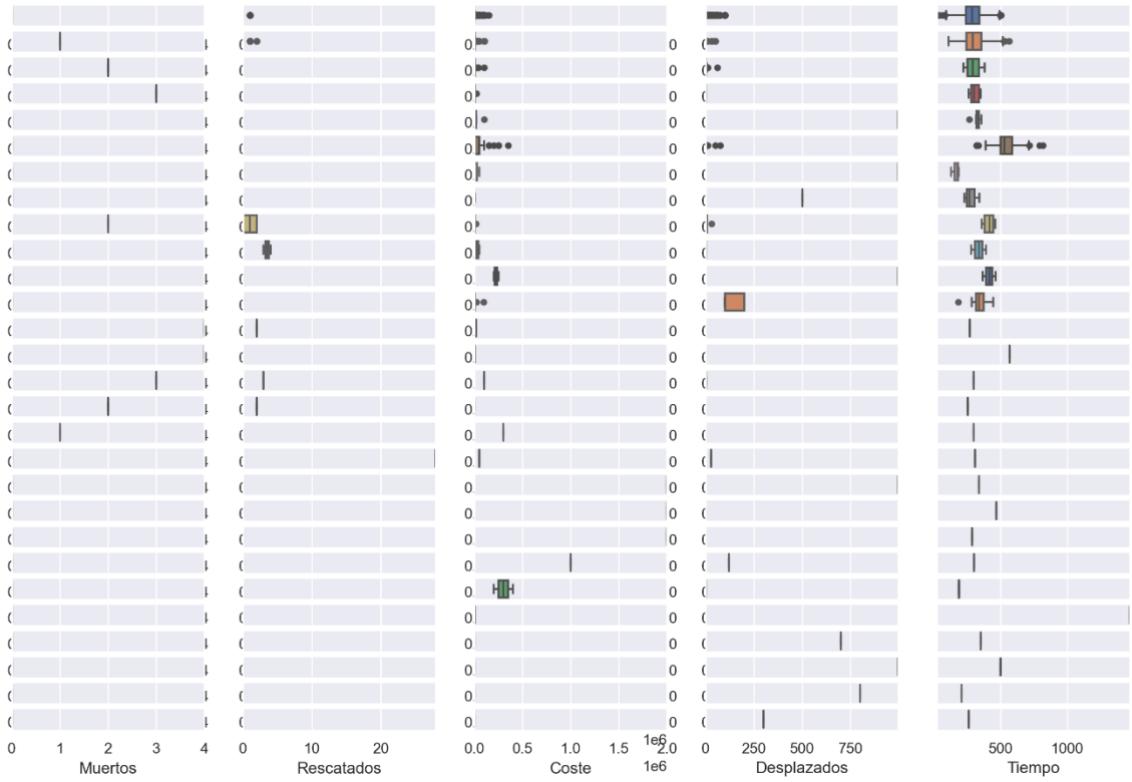


Figura 12: Boxplot Meanshift Complementario Caso 1

DBSCAN

En este apartado se obtienen métricas muy buenas tanto para el caso de interés como para el complementario, consiguiendo mejores segmentaciones conforme el valor de épsilon crece y manteniendo los mismos tiempos de ejecución.

Epsilon	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
0.15	0: 2145 (97.68%) 1: 44 (2.00%) -1: 7 (0.32%)	0.83838	1525.777	0.20
0.2	0: 2146 (97.72%) 1: 44 (2.00%) -1: 6 (0.27%)	0.84615	1531.634	0.16
0.25	0: 2145 (97.68%) 1: 44 (2.00%) -1: 7 (0.32%)	0.86041	1557.008	0.19
0.3	0: 2146 (97.72%) 1: 44 (2.00%) -1: 6 (0.27%)	0.86992	1570.883	0.19
0.35	0: 2146 (97.72%) 1: 44 (2.00%) -1: 6 (0.27%)	0.86992	1570.883	0.16
0.4	0: 2146 (97.72%) 1: 44 (2.00%) -1: 6 (0.27%)	0.86992	1570.883	0.17

Tabla 8: Comparación DBSCAN Caso 1

Epsilon	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
0.15	0: 2511 (89.71%) 3: 232 (8.29%) 2: 24 (0.86%) -1: 13 (0.46%) 4: 12 (0.43%) 1: 7 (0.25%)	0.74520	1205.139	0.22
0.2	0: 2511 (89.71%) 3: 232 (8.29%) 2: 24 (0.86%) 4: 13 (0.46%) -1: 12 (0.43%) 1: 7 (0.25%)	0.74565	1241.558	0.22
0.25	0: 2784 (99.46%) -1: 8 (0.29%) 1: 7 (0.25%)	0.82891	176.509	0.23
0.3	0: 2793 (99.79%) -1: 6 (0.21%)	0.87811	140.003	0.26
0.35	0: 2794 (99.82%) -1: 5 (0.18%)	0.88773	129.633	0.25
0.4	0: 2794 (99.82%) -1: 5 (0.18%)	0.88773	129.633	0.22

Tabla 9: Comparación DBSCAN Complementario Caso 1

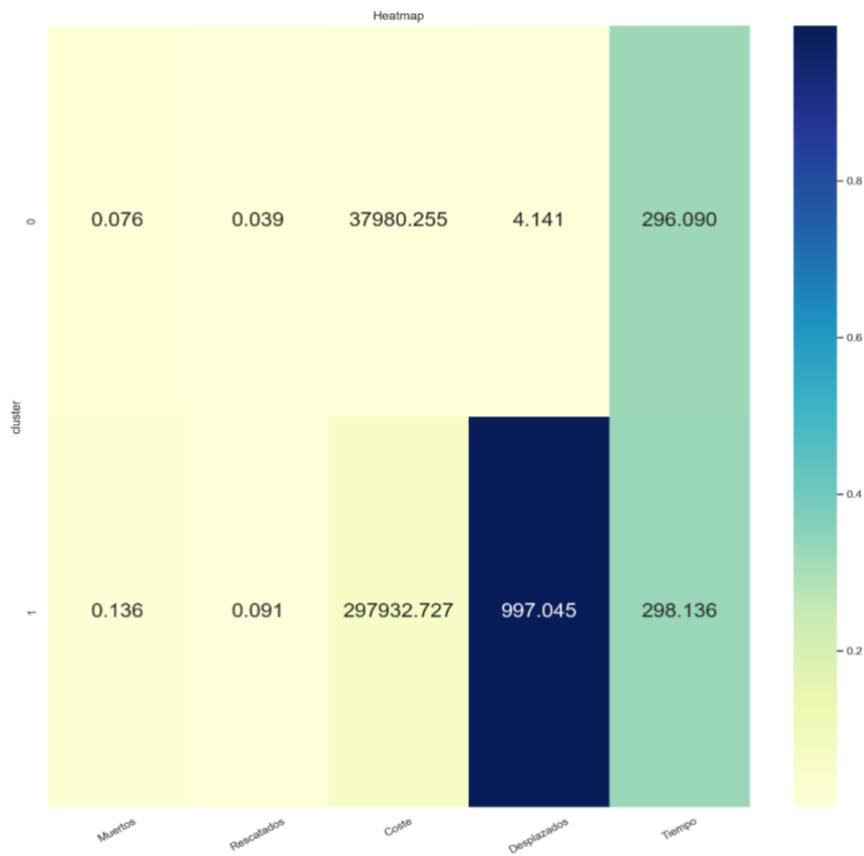


Figura 13: Heatmap DBSCAN Caso 1

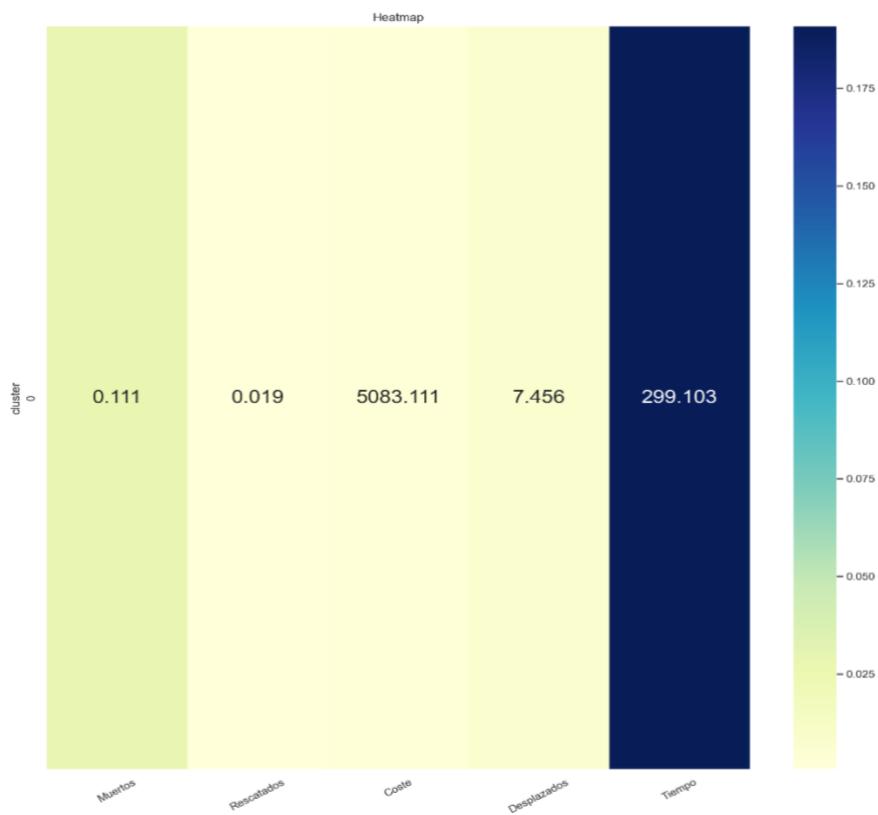


Figura 14: Heatmap DBSCAN Complementario Caso 1

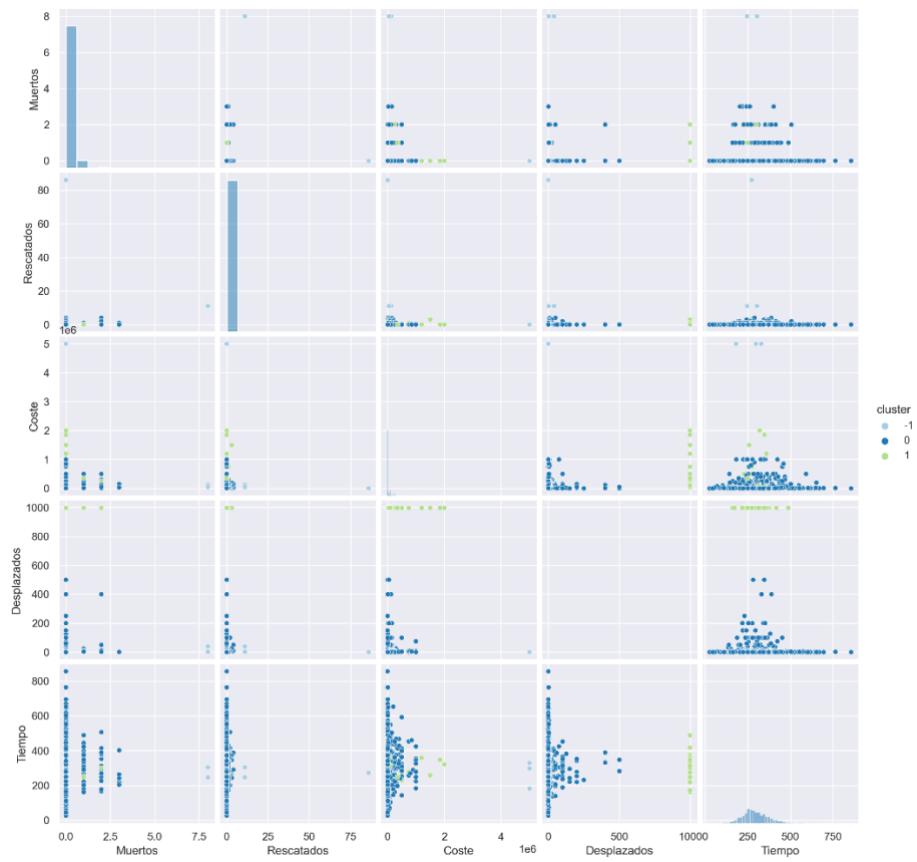


Figura 15: ScatterMatrix DBSCAN Caso 1



Figura 16: ScatterMatrix DBSCAN Complementario Caso 1

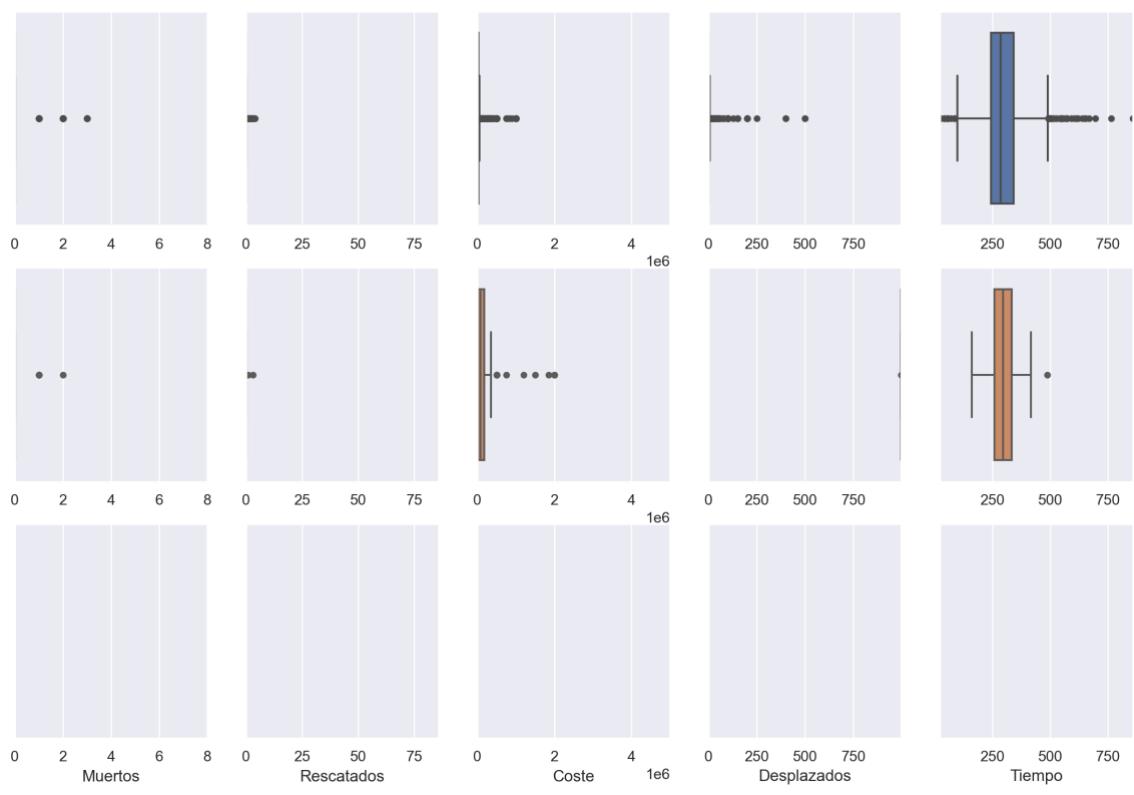


Figura 17: BoxPlot DBSCAN Caso 1

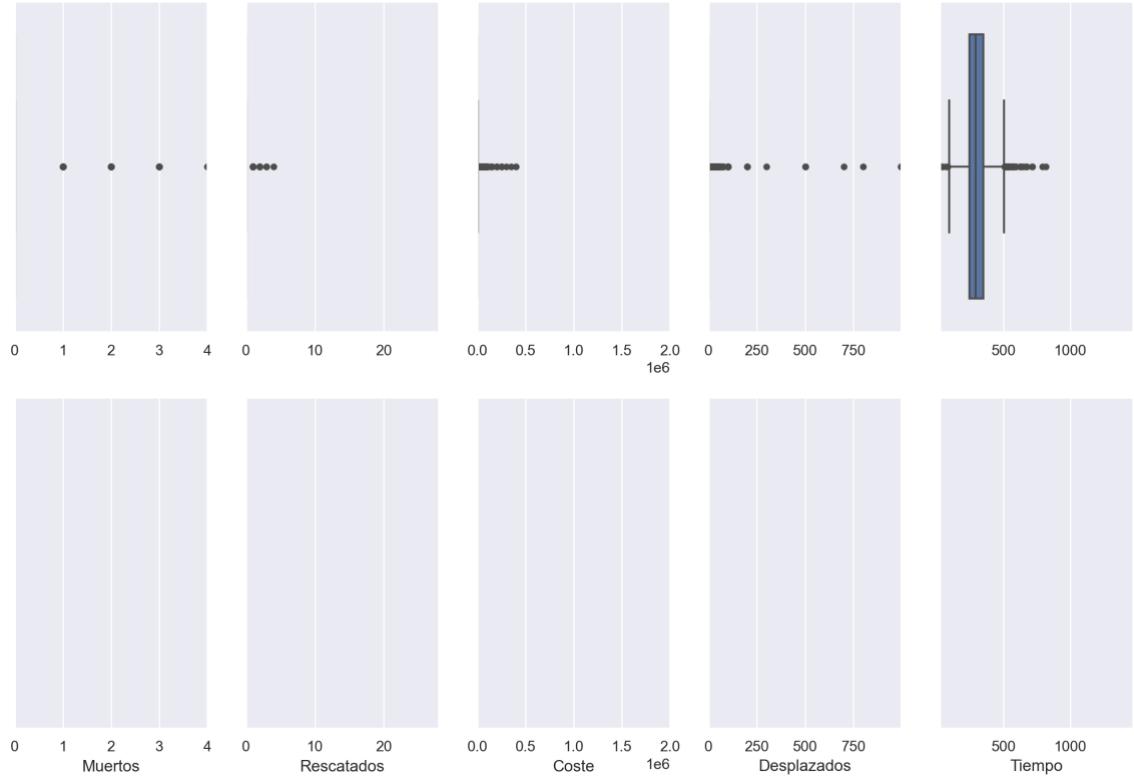


Figura 18: BoxPlot DBSCAN Complementario Caso 1

Agglomerative Clustering

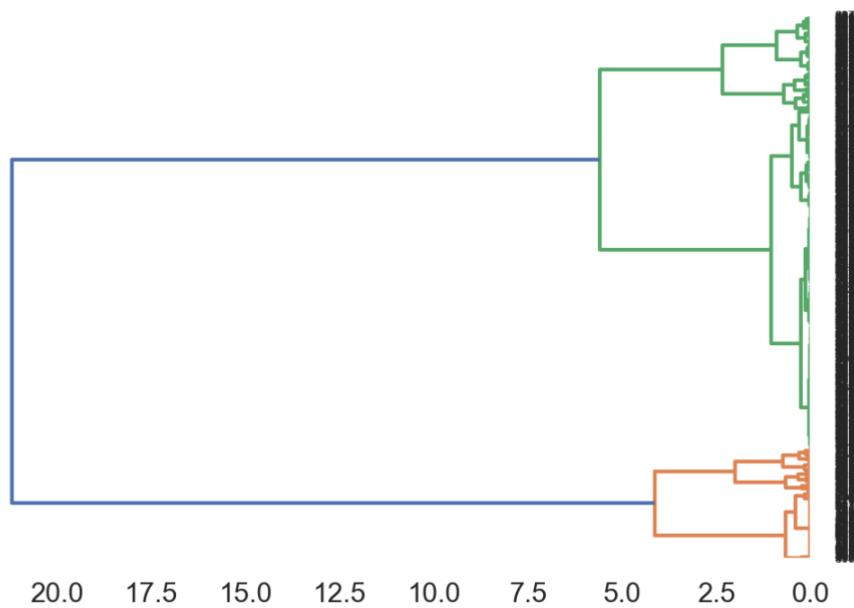


Figura 19: Dendograma Ward Caso 1

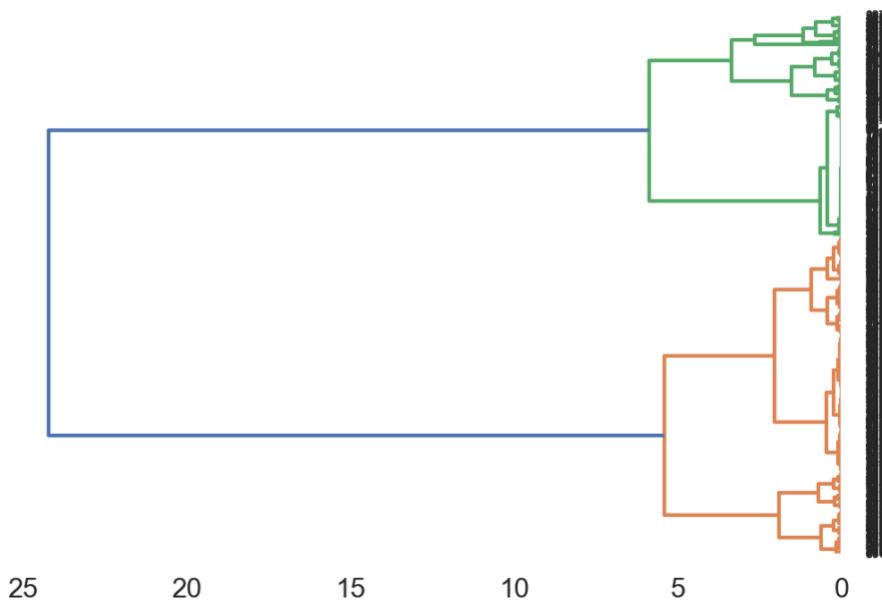


Figura 20: Dendograma Ward Complementario Caso 1

Dentro del algoritmo de clustering jerárquico podemos observar que primero se realiza una gran división a partir de la variable coste, tanto en el caso de estudio como en el complementario. Después la otra variable que también va generando las divisiones es el tiempo que tardan en llegar al incendio.

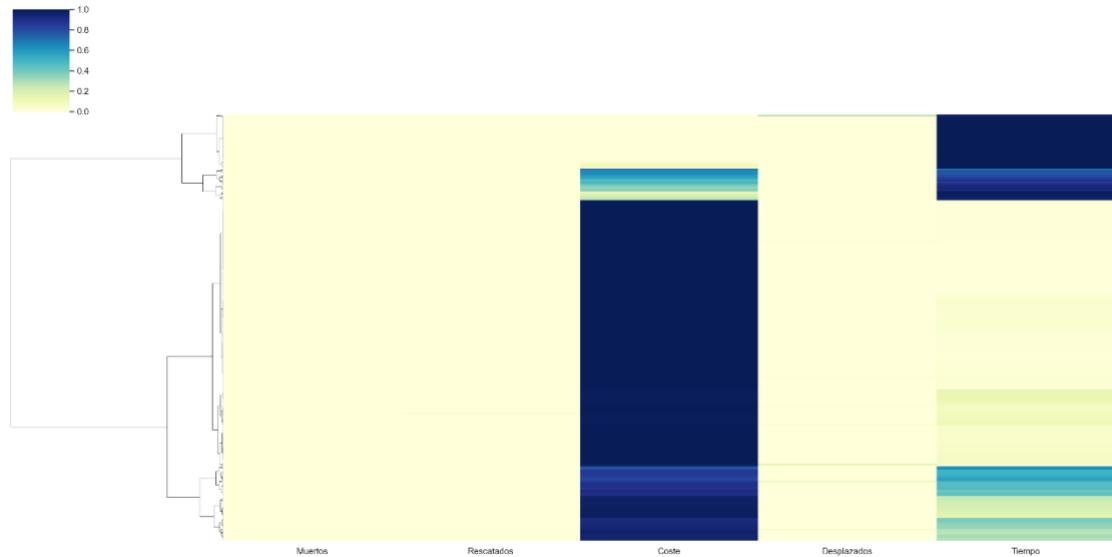


Figura 21: Dendrogram Heatmap Ward Caso 1

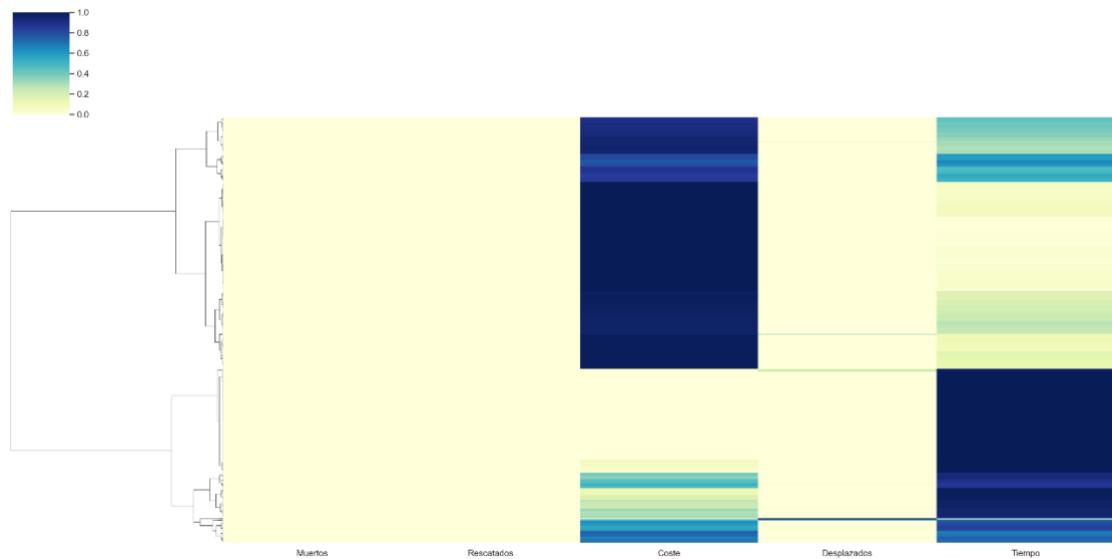


Figura 22: Dendrogram Heatmap Ward Complementario Caso 1

Birch

En este último algoritmo conseguimos el mejor valor de medidas en el caso de estudio de interés, con solo dos clusters obtenemos los valores más altos del Silhouette y Calinski-harabasz.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
2	0: 2152 (98.00%) 1: 44 (2.00%)	0.87002	2768.048	0.04

Tabla 10: Métricas Birch Caso 1

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
2	0: 2766 (98.82%) 1: 33 (1.18%)	0.79254	660.742	0.04

Tabla 11: Métricas Birch Complementario Caso 1

Interpretación de la segmentación

Observando las diferentes gráficas y métricas obtenidas vemos que los resultados distan mucho de unos algoritmos y otros, por lo que es complicado sacar una conclusión fiable y de peso a partir de estos datos. Aunque se mencionarán algunos detalles como puede ser que el tiempo de llegada es importante a la hora de segmentar nuestro conjunto.

Por otro lado, otras variables que parecen ser importantes a la hora de agrupar elementos son el coste estimado para la extinción de los incendios y el número estimado de personas que han sido desplazadas para que no reciban daños.

Entonces podemos sacar ideas como, a mayor tiempo mayor será el coste de diversas reparaciones o que también el tiempo está directamente correlacionado con el número de personas desplazadas de la zona del incendio.

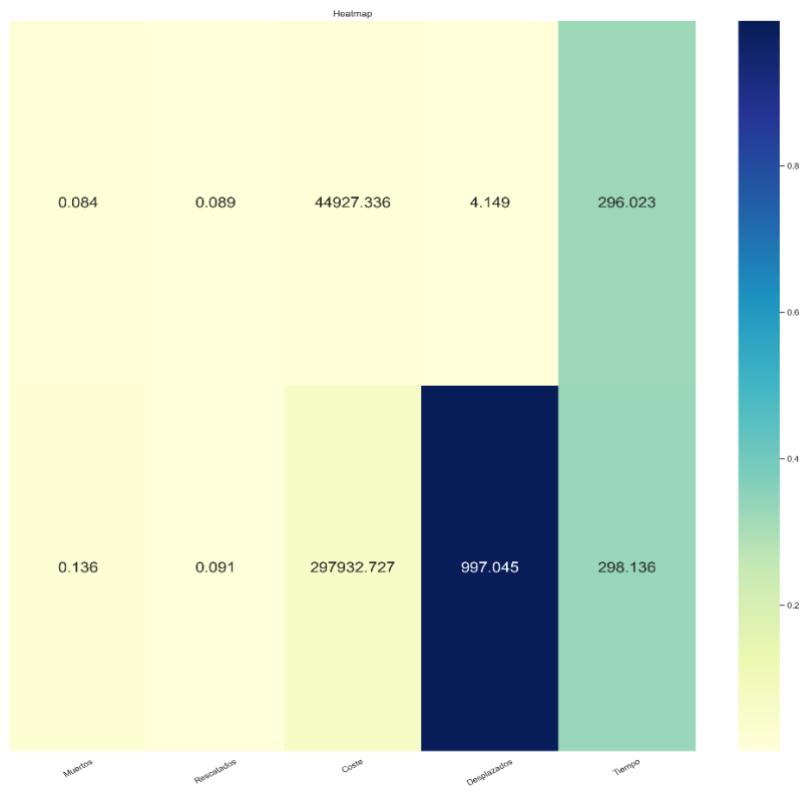


Figura 23: Heatmap Birch Caso 1

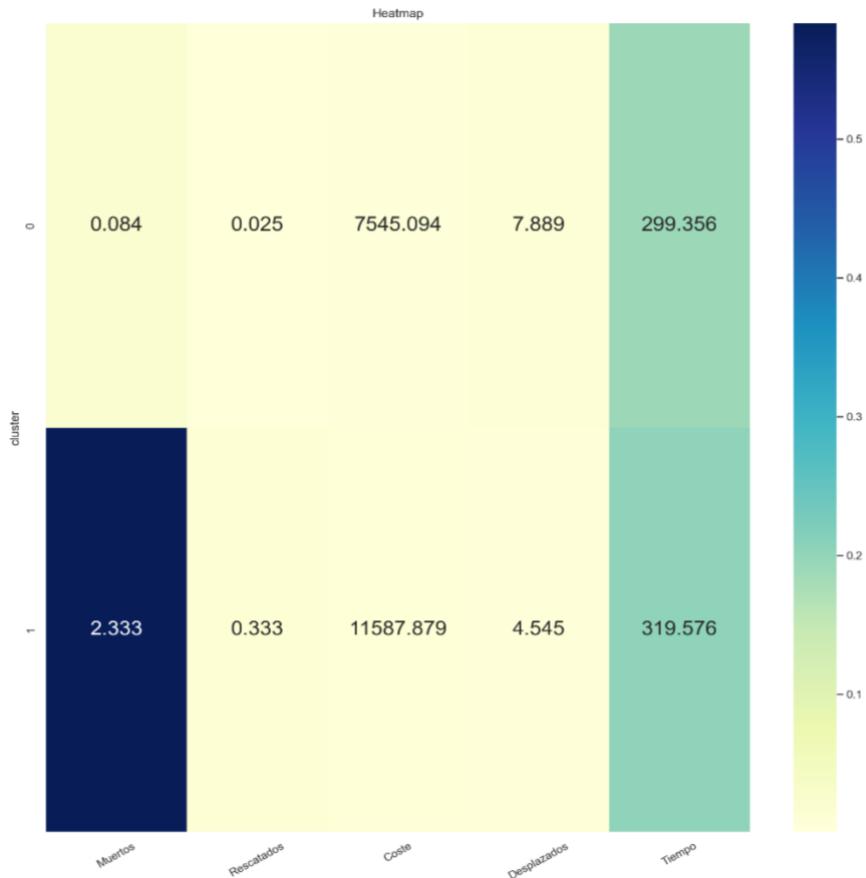


Figura 24: Heatmap Birch Complementario Caso 1

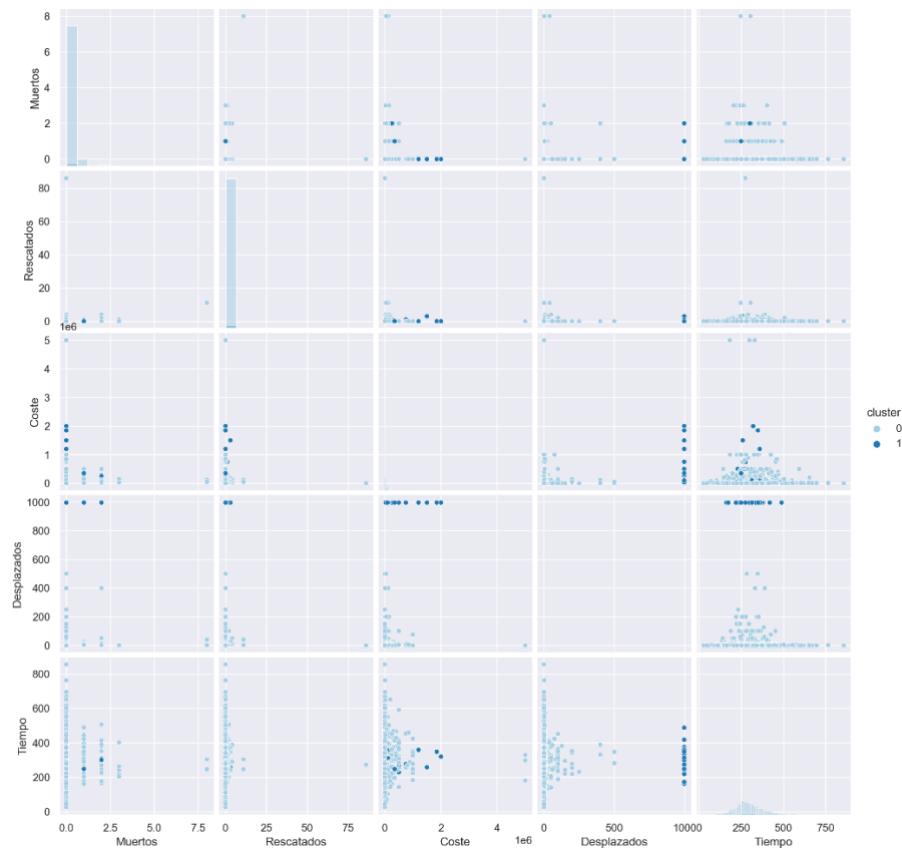


Figura 25: ScatterMatrix Birch Caso 1

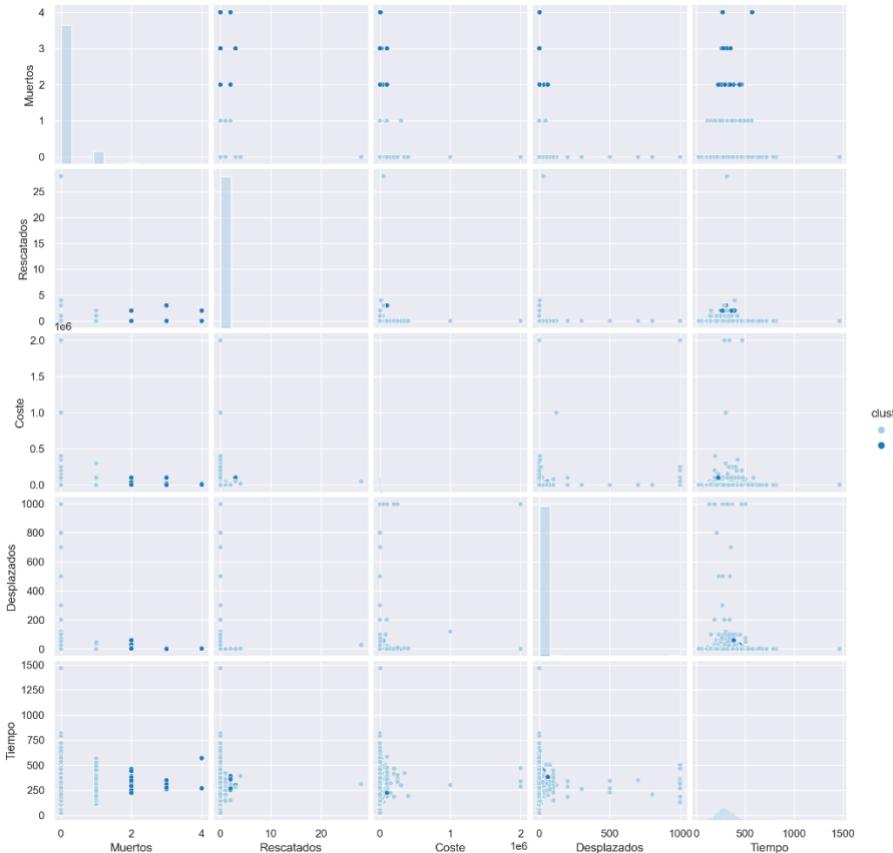


Figura 26: ScatterMatrix Birch Complementario Caso 1

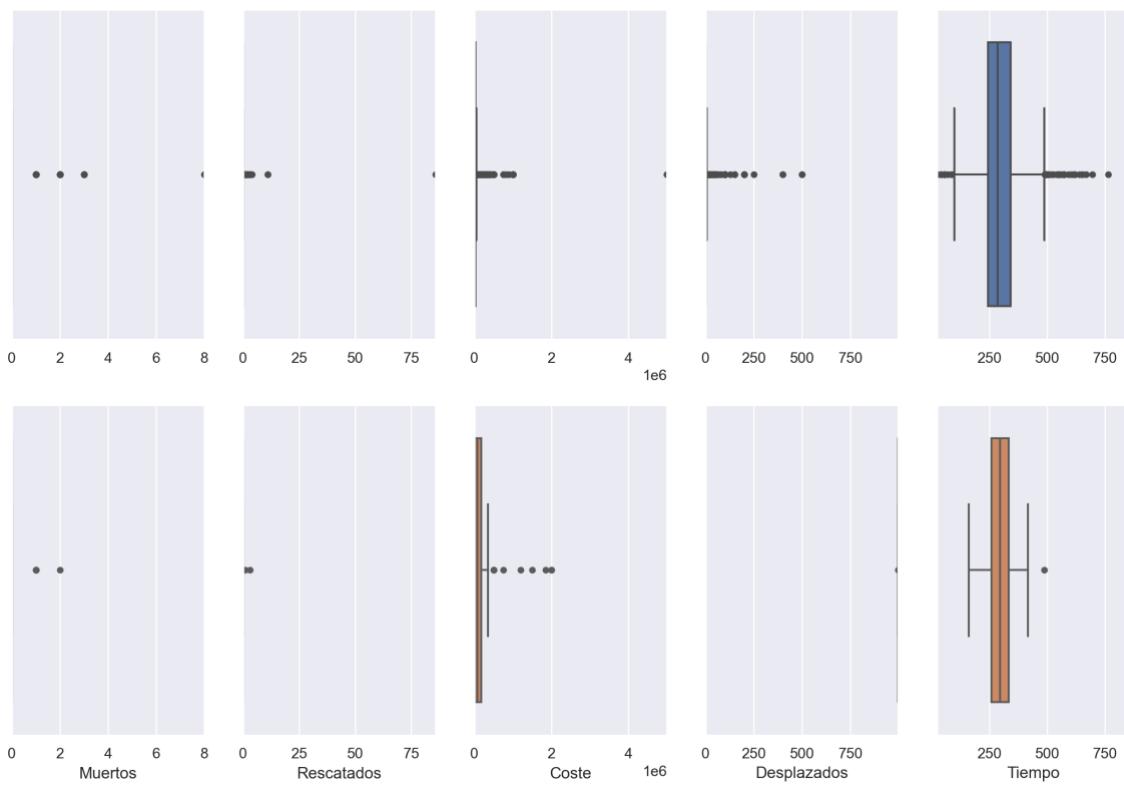


Figura 27: BoxPlot Birch Caso 1

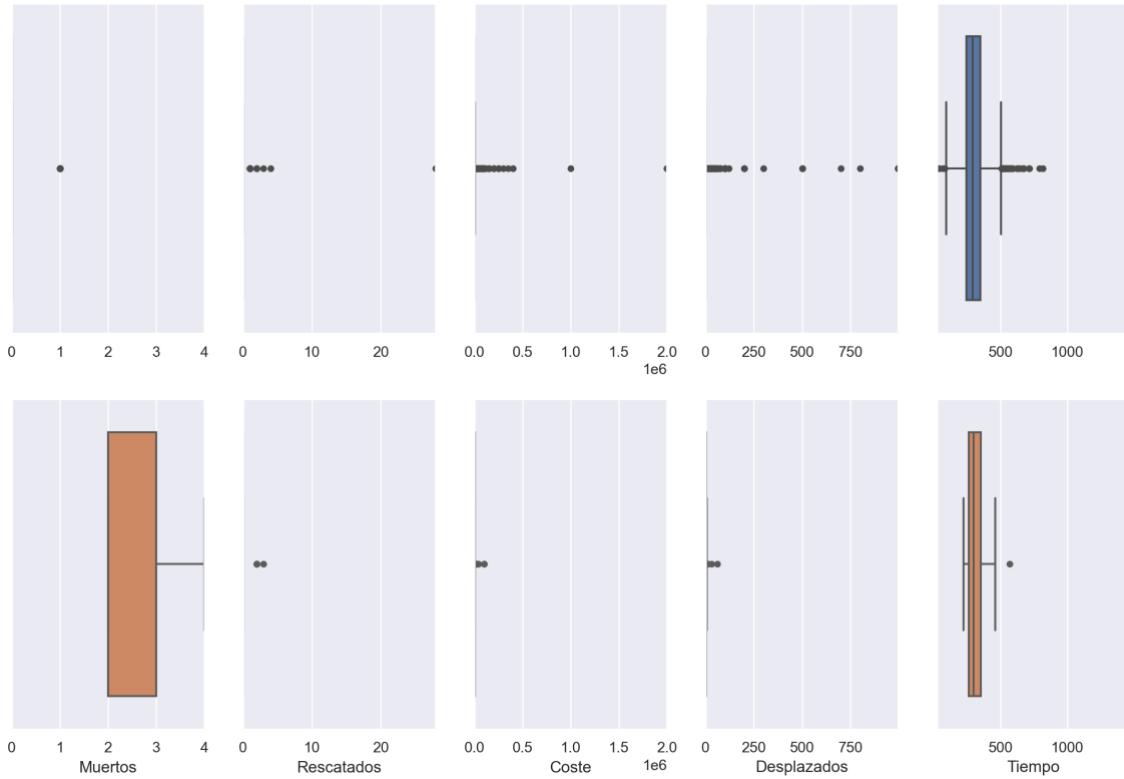


Figura 28: Boxplot Birch Complementario Caso 1

3. Caso de estudio 2

En este segundo caso se estudiarán aquellos incendios donde hubo alguna víctima, añadiendo además, que el sistema de alarma funcionara y alertara a todos o a alguna de las personas que fueron evacuadas. El objetivo es ver el impacto de una alarma a la hora de iniciarse el incendio y la actuación de las personas ante dicho evento y si existe alguna tendencia que los agrupe.

Se han escogido 1735 ejemplos donde se ha exigido que las variables *Civilian_Casualties* sea mayor que 0, *Fire_Alarm_System_Operation* coincida con *Fire alarm system operated* y *Fire_Alarm_System_Impact_On_Evacuation* coincida con *Some persons (at risk) evacuated as a result of hearing fire alarm system* o *All persons (at risk of injury) evacuated as a result of hearing fire alarm system*. Las variables utilizadas serán las mismas:

- *Civilian_Casualties* -> Muertos
- *Count_Of_Person_Rescued* -> Rescatados
- *Estimated_Dollar_Loss* -> Coste
- *Estimated_Number_Of_Person_Displaced* -> Desplazados
- *Arrival_Time* -> Tiempo

Para la ejecución de este caso se han obtenido las siguientes métricas, donde el caso que ha obtenido mejor coeficiente de Silhouette es DBSCAN seguido por Birch. Aunque cabe mencionar que no han obtenido muy buenos índices de Calinski-Harabasz por lo que es posible que los datos estén muy juntos o que las agrupaciones estén mal establecidas.

	Silhouette	Calinski-Harabasz	Tº ejecución	Número de Clusters
Kmeans	0.39203	168.142	0.02	3
Meanshift	0.25538	68.940	0.32	19
DBSCAN	0.81200	91.269	0.01	1
Birch	0.80843	59.385	0.01	3
AC		0.00		3

Tabla 12: Métricas Caso 2

	Silhouette	Calinski-Harabasz	Tº ejecución	Número de Clusters
Kmeans	0.46710	16223.030	0.05	3
Meanshift	0.39607	2498.971	1.19	34
DBSCAN	0.89852	49.553	1.53	1
Birch	0.91904	10449.524	0.12	
AC		0.02		3

Tabla 13: Métricas Complementario Caso 2

K-means

El algoritmo de K-means ha obtenido valores bastante pobres oscilando alrededor de sus máximos locales, como pueden ser en los números de clusters 3 u 8. De hecho se obtienen mejores medidas en el caso complementario y que también van oscilando de la misma manera.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
2	0: 192 (55.33%) 1: 155 (44.67%)	0.36552	139.413	0.02
3	0: 206 (59.37%) 1: 138 (39.77%) 2: 3 (0.86%)	0.39203	168.142	0.02
4	1: 146 (42.07%) 0: 110 (31.70%) 2: 88 (25.36%) 3: 3 (0.86%)	0.31835	157.313	0.02
5	4: 146 (42.07%) 1: 109 (31.41%) 0: 86 (24.78%) 2: 4 (1.15%) 3: 2 (0.58%)	0.32678	148.839	0.02
6	3: 145 (41.79%) 0: 109 (31.41%) 1: 85 (24.50%) 4: 3 (0.86%) 5: 3 (0.86%) 2: 2 (0.58%)	0.33332	146.474	0.02
7	1: 138 (39.77%) 2: 107 (30.84%) 3: 78 (22.48%) 4: 18 (5.19%) 5: 3 (0.86%) 0: 2 (0.58%) 6: 1 (0.29%)	0.34182	146.787	0.02
8	5: 145 (41.79%) 0: 111 (31.99%) 1: 78 (22.48%) 2: 5 (1.44%) 3: 3 (0.86%) 6: 2 (0.58%) 7: 2 (0.58%) 4: 1 (0.29%)	0.34825	149.041	0.02
9	0: 126 (36.31%) 3: 98 (28.24%) 8: 73 (21.04%) 1: 37 (10.66%) 7: 5 (1.44%) 4: 3 (0.86%) 2: 2 (0.58%) 6: 2 (0.58%) 5: 1 (0.29%)	0.30295	149.808	0.03
10	6: 115 (33.14%) 0: 83 (23.92%) 1: 78 (22.48%) 7: 30 (8.65%) 5: 28 (8.07%) 8: 6 (1.73%) 3: 3 (0.86%) 2: 2 (0.58%) 9: 1 (0.29%) 4: 1 (0.29%)	0.31468	148.510	0.02

Tabla 14: Comparación K-means Caso 2

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
2	0: 6346 (97.95%) 1: 133 (2.05%)	0.91913	20234.912	0.03
3	2: 4006 (61.83%) 0: 2340 (36.12%) 1: 133 (2.05%)	0.46710	16223.030	0.04
4	2: 3744 (57.79%) 3: 2142 (33.06%) 0: 460 (7.10%) 1: 133 (2.05%)	0.54167	17428.341	0.07
5	3: 2647 (40.86%) 2: 2529 (39.03%) 4: 714 (11.02%) 0: 457 (7.05%) 1: 132 (2.04%)	0.50780	16611.151	0.08
6	4: 2721 (42.00%) 2: 2337 (36.07%) 0: 832 (12.84%) 3: 436 (6.73%) 1: 132 (2.04%) 5: 21 (0.32%)	0.50954	16127.381	0.13
7	4: 2504 (38.65%) 2: 1640 (25.31%) 0: 1416 (21.86%) 3: 436 (6.73%) 5: 330 (5.09%) 1: 132 (2.04%) 6: 21 (0.32%)	0.50391	15857.120	0.13
8	5: 2010 (31.02%) 2: 1849 (28.54%) 7: 1022 (15.77%) 4: 876 (13.52%) 0: 437 (6.74%) 1: 132 (2.04%) 3: 132 (2.04%) 6: 21 (0.32%)	0.49785	15129.707	0.16
9	6: 2365 (36.50%) 4: 1698 (26.21%) 2: 1178 (18.18%) 0: 610 (9.42%) 3: 436 (6.73%) 1: 124 (1.91%) 8: 38 (0.59%) 5: 21 (0.32%) 7: 9 (0.14%)	0.50324	14919.728	0.18
10	7: 1730 (26.70%) 4: 1717 (26.50%) 0: 1287 (19.86%) 2: 649 (10.02%) 5: 472 (7.29%) 3: 436 (6.73%) 1: 131 (2.02%) 8: 35 (0.54%) 6: 21 (0.32%) 9: 1 (0.02%)	0.49837	14621.630	0.21

Tabla 15: Comparación K-means Complementario Caso 2

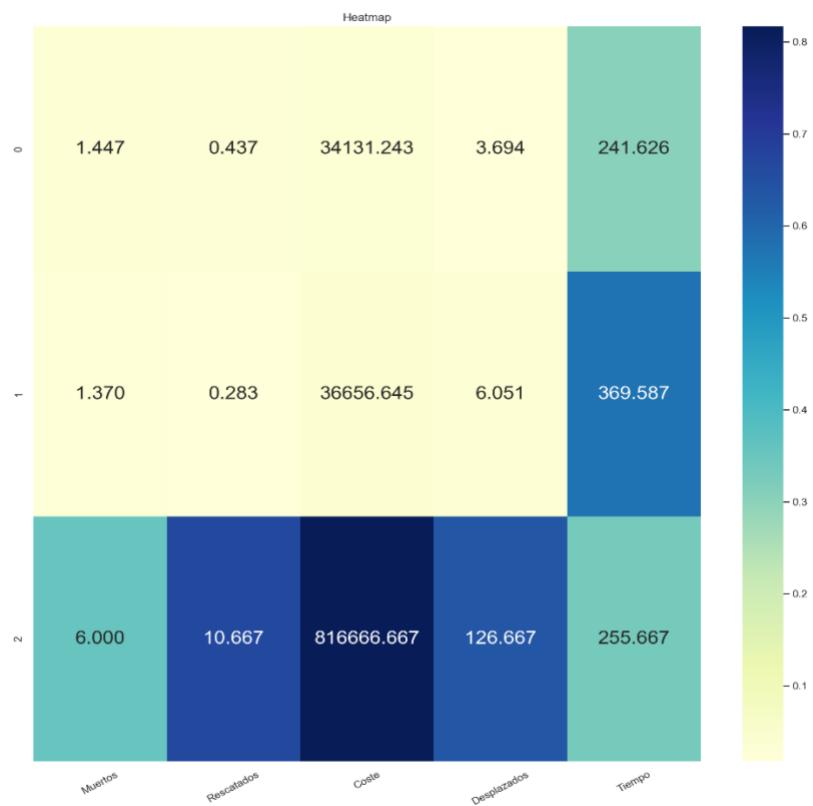


Figura 29: Heatmap K-means Caso 2

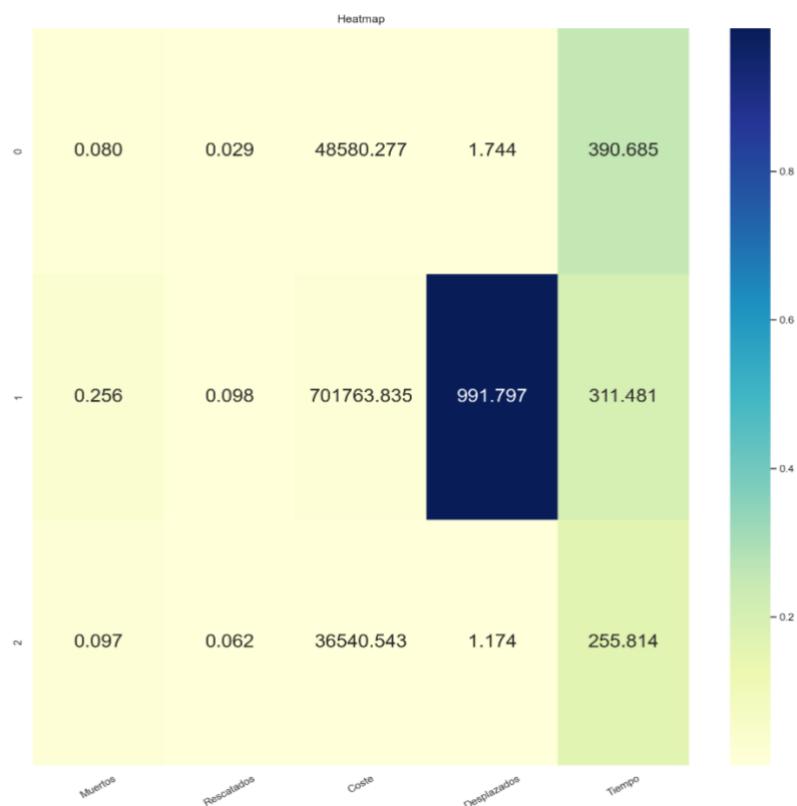


Figura 30: Heatmap K-means Complementario Caso 2



Figura 31: ScatterMatrix K-means Caso 2

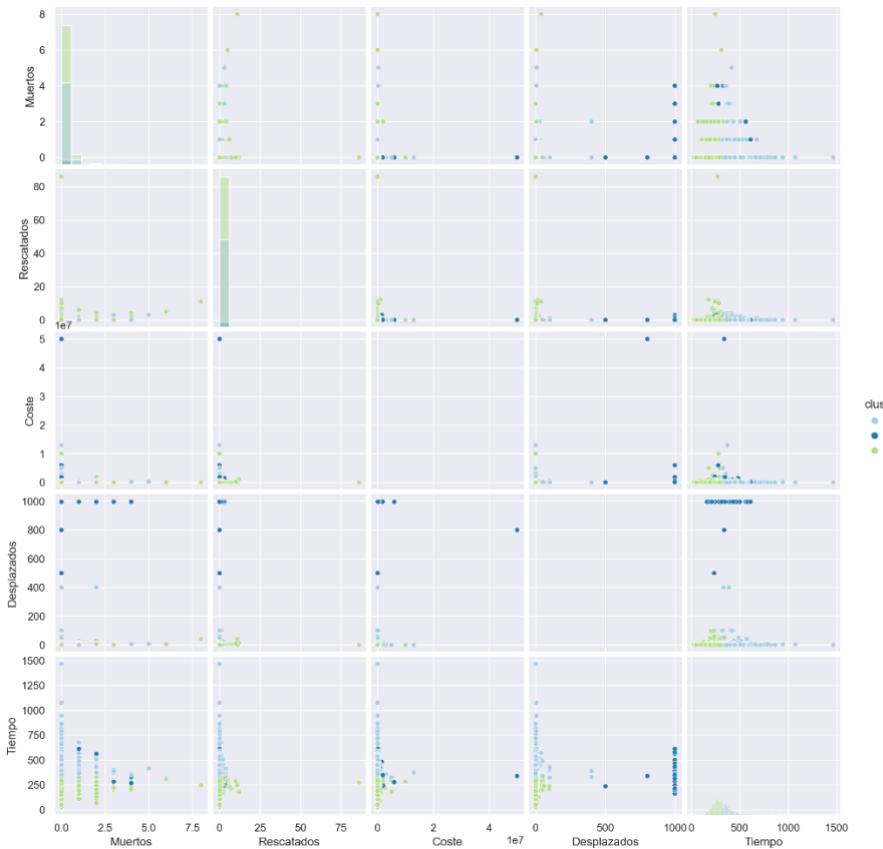


Figura 32: ScatterMatrix K-means Complementario Caso 2

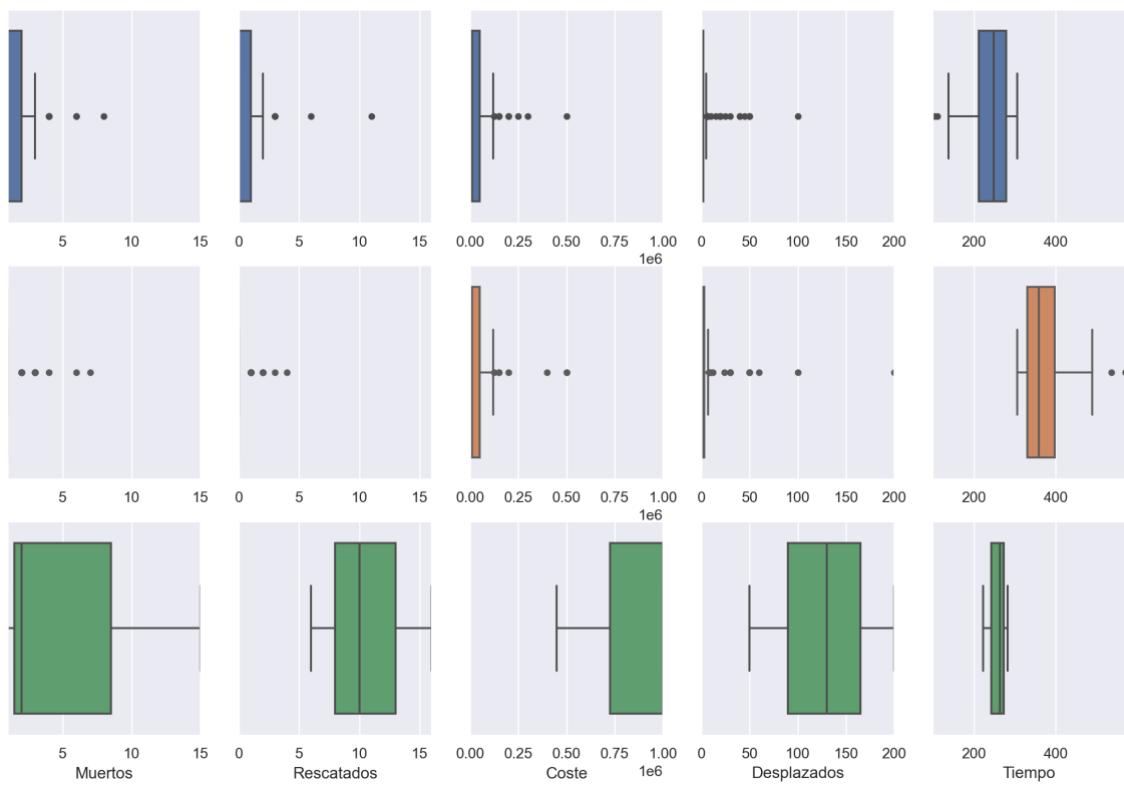


Figura 33: BoxPlot K-means Caso 2

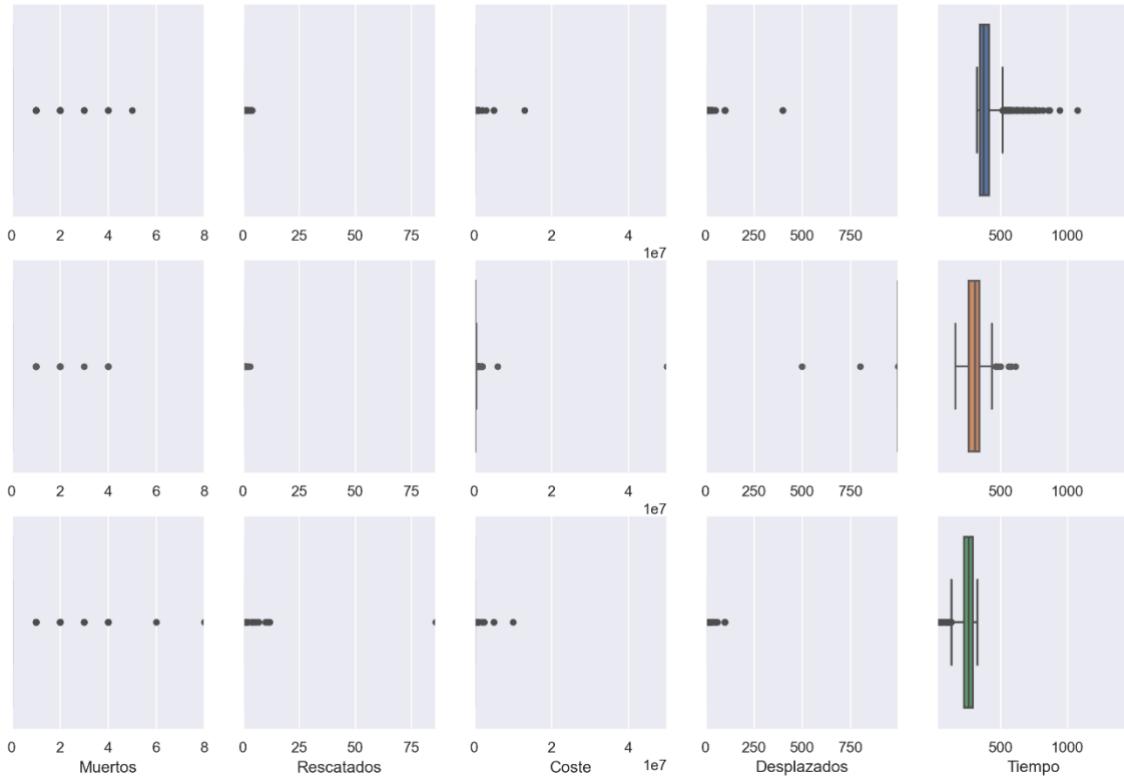


Figura 34: BoxPlot K-means Complementario Caso 2

Meanshift

En el algoritmo de meanshift hemos obtenido el peor valor en este caso de estudio y de toda la práctica. Esto puede significar que las agrupaciones en formas convexas no funcionan bien con el conjunto de datos escogido en este caso, ya que tampoco se consigue un buen resultado ni en el complementario ni en K-means.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
19	0: 221 (63.69%) 1: 68 (19.60%) 5: 25 (7.20%) 10: 6 (1.73%) 2: 5 (1.44%) 3: 4 (1.15%) 4: 3 (0.86%) 8: 3 (0.86%) 16: 2 (0.58%) 11: 1 (0.29%) 6: 1 (0.29%) 18: 1 (0.29%) 15: 1 (0.29%) 14: 1 (0.29%) 7: 1 (0.29%) 13: 1 (0.29%) 9: 1 (0.29%) 12: 1 (0.29%) 17: 1 (0.29%)	0.25538	68.940	0.22

Tabla 16: Métricas Meanshift Caso 2

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
34	0: 5369 (82.87%) 1: 388 (5.99%) 14: 383 (5.91%) 2: 111 (1.71%) 29: 74 (1.14%) 3: 40 (0.62%) 21: 34 (0.52%) 10: 15 (0.23%) 13: 9 (0.14%) 5: 8 (0.12%) 6: 6 (0.09%) 4: 6 (0.09%) 7: 5 (0.08%) 9: 4 (0.06%) 12: 3 (0.05%) 8: 3 (0.05%) 11: 3 (0.05%) 22: 2 (0.03%) 33: 1 (0.02%) 31: 1 (0.02%) 19: 1 (0.02%) 18: 1 (0.02%) 26: 1 (0.02%) 27: 1 (0.02%) 28: 1 (0.02%) 23: 1 (0.02%) 32: 1 (0.02%) 25: 1 (0.02%) 20: 1 (0.02%) 24: 1 (0.02%) 17: 1 (0.02%) 16: 1 (0.02%) 15: 1 (0.02%) 30: 1 (0.02%)	0.39607	2498.971	1.19

Tabla 17: Métricas Meanshift Complementario Caso 2



Figura 35: Heatmap Meanshift Caso 2

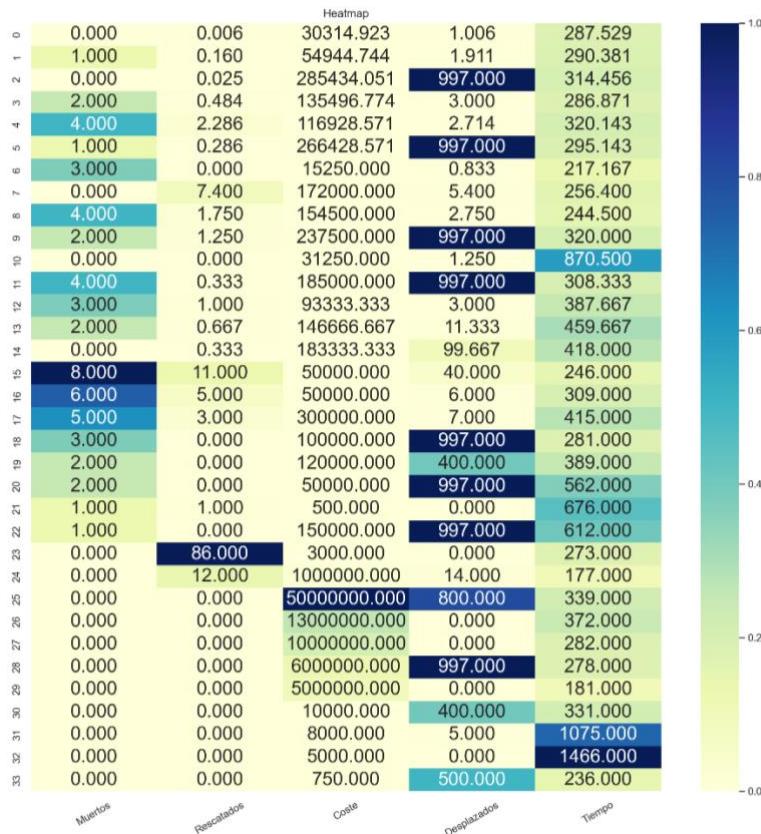


Figura 36: Heatmap Meanshift Complementario Caso 2



Figura 37: ScatterMatrix Meanshift Caso 2

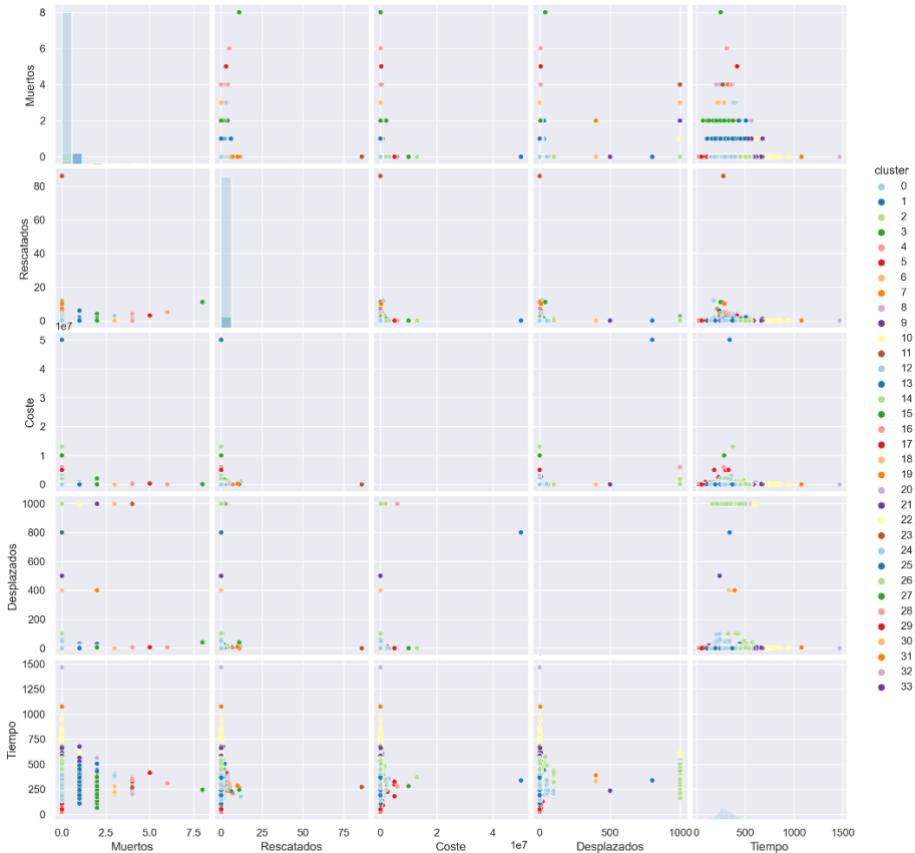


Figura 38: ScatterMatrix Meanshift Complementario Caso 2

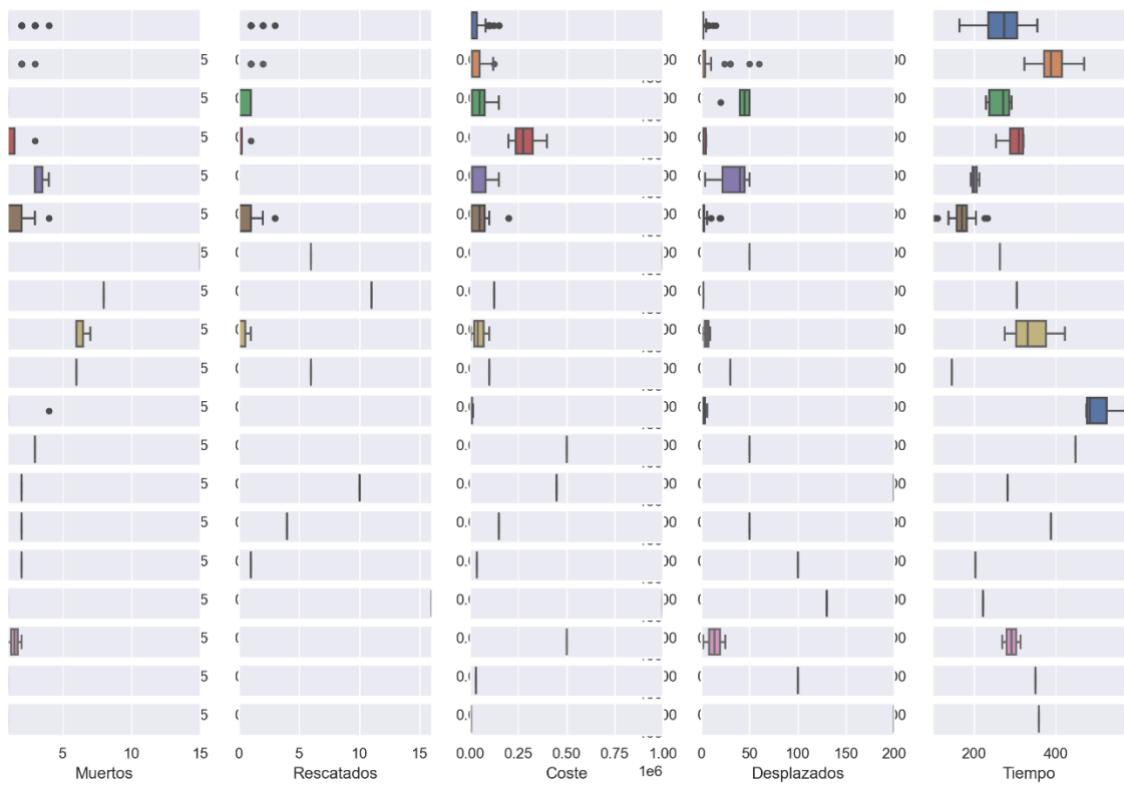


Figura 39: BoxPlot Meanshift Caso 2

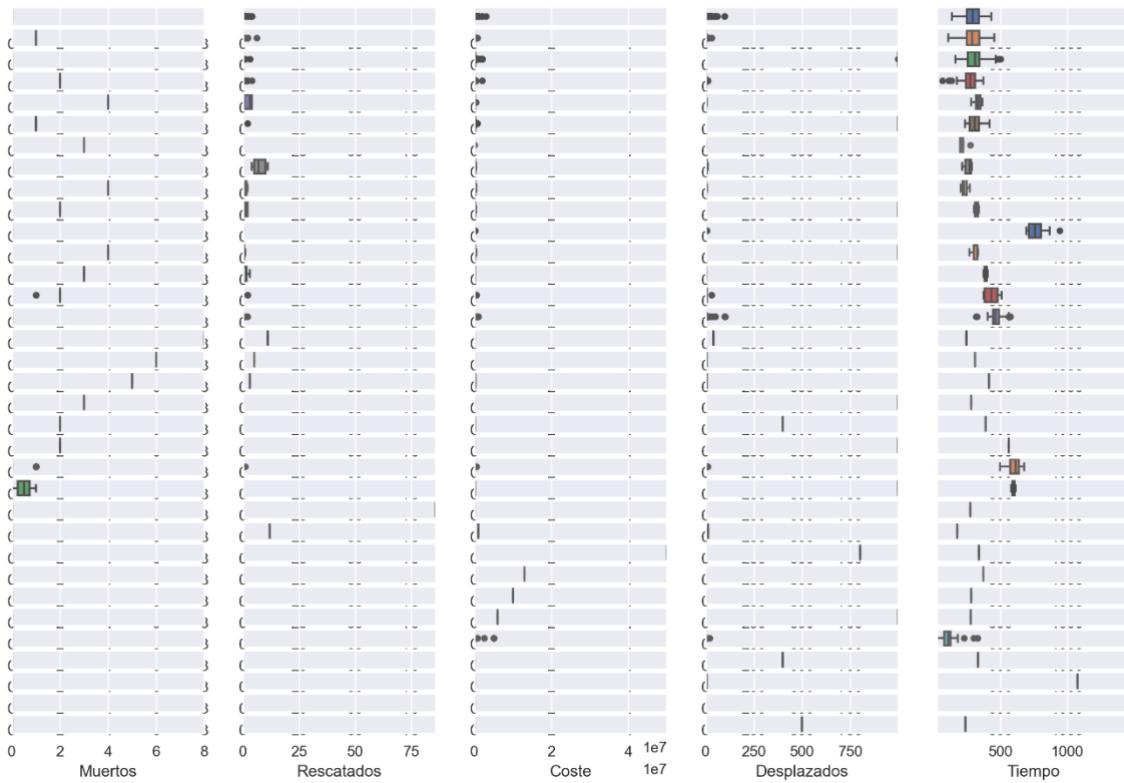


Figura 40: BoxPlot Meanshift Complementario Caso 2

DBSCAN

En DBSCAN se consigue las mejores medidas para el valor más alto de epsilon calculado. Aquí ya se mejora mucho en el conjunto de datos complementario los resultados. Aunque es un único clusters según las medidas esa agrupación es bastante fiable.

Epsilon	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
0.15	0: 326 (93.95%) -1: 21 (6.05%)	0.60513	63.616	0.00
0.2	0: 333 (95.97%) -1: 14 (4.03%)	0.67300	71.166	0.01
0.25	0: 339 (97.69%) -1: 8 (2.31%)	0.73666	84.773	0.01
0.3	0: 340 (97.98%) -1: 7 (2.02%)	0.75168	86.750	0.01
0.35	0: 340 (97.98%) -1: 7 (2.02%)	0.75168	86.750	0.01
0.4	0: 342 (98.56%) -1: 5 (1.44%)	0.78442	89.186	0.01
0.45	0: 342 (98.56%) -1: 5 (1.44%)	0.78442	89.186	0.01
0.5	0: 342 (98.56%) -1: 5 (1.44%)	0.78442	89.186	0.01
0.55	0: 344 (99.14%) -1: 3 (0.86%)	0.81200	91.269	0.01

Tabla 18: Comparación DBSCAN Caso 2

Epsilon	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
0.15	0: 6341 (97.87%) 1: 131 (2.02%) -1: 7 (0.11%)	0.89704	10326.479	1.51
0.2	0: 6341 (97.87%) 1: 131 (2.02%) -1: 7 (0.11%)	0.89704	10326.479	1.68
0.25	0: 6341 (97.87%) 1: 131 (2.02%) -1: 7 (0.11%)	0.89704	10326.479	1.60
0.3	0: 6343 (97.90%) 1: 131 (2.02%) -1: 5 (0.08%)	0.89054	10334.992	1.67
0.35	0: 6346 (97.95%) 1: 131 (2.02%) -1: 2 (0.03%)	0.91934	10345.765	1.68
0.4	0: 6346 (97.95%) 1: 131 (2.02%) -1: 2 (0.03%)	0.91934	10345.765	1.69
0.45	0: 6346 (97.95%) 1: 131 (2.02%) -1: 2 (0.03%)	0.91934	10345.765	1.54
0.5	0: 6477 (99.97%) -1: 2 (0.03%)	0.89852	49.553	1.54
0.55	0: 6477 (99.97%) -1: 2 (0.03%)	0.89852	49.553	1.57

Tabla 19: Comparación DBSCAN Complementario Caso 2

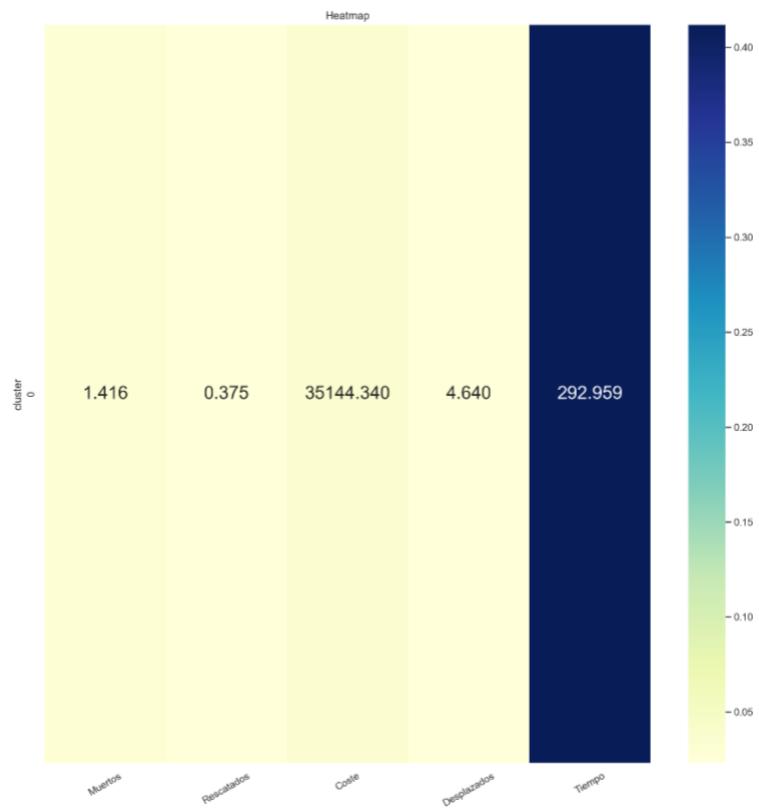


Figura 41: Heatmap DBSCAN Caso 2

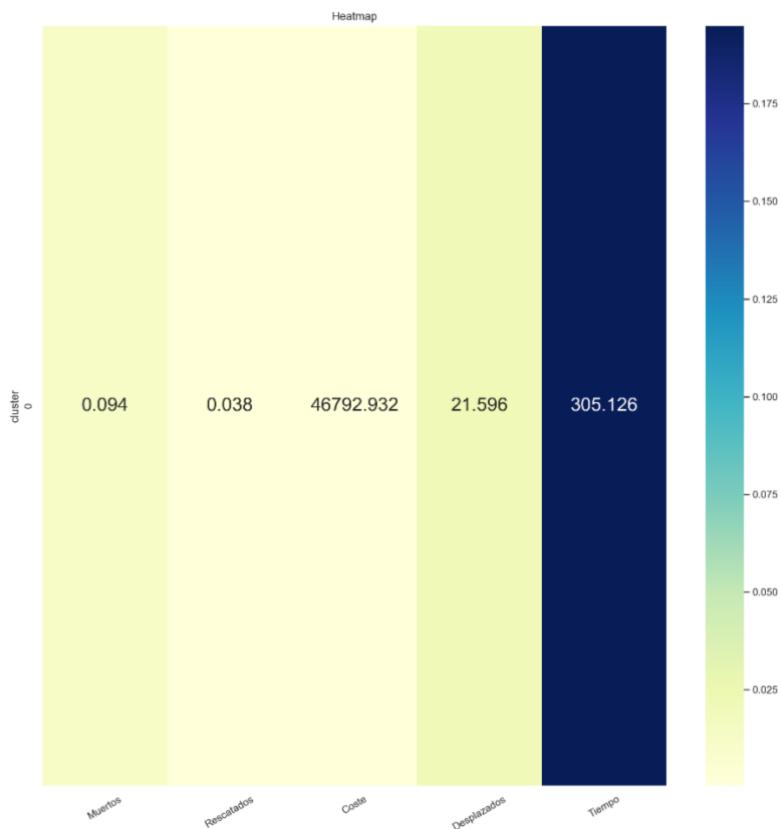


Figura 42: Heatmap DBSCAN Complementario Caso 2

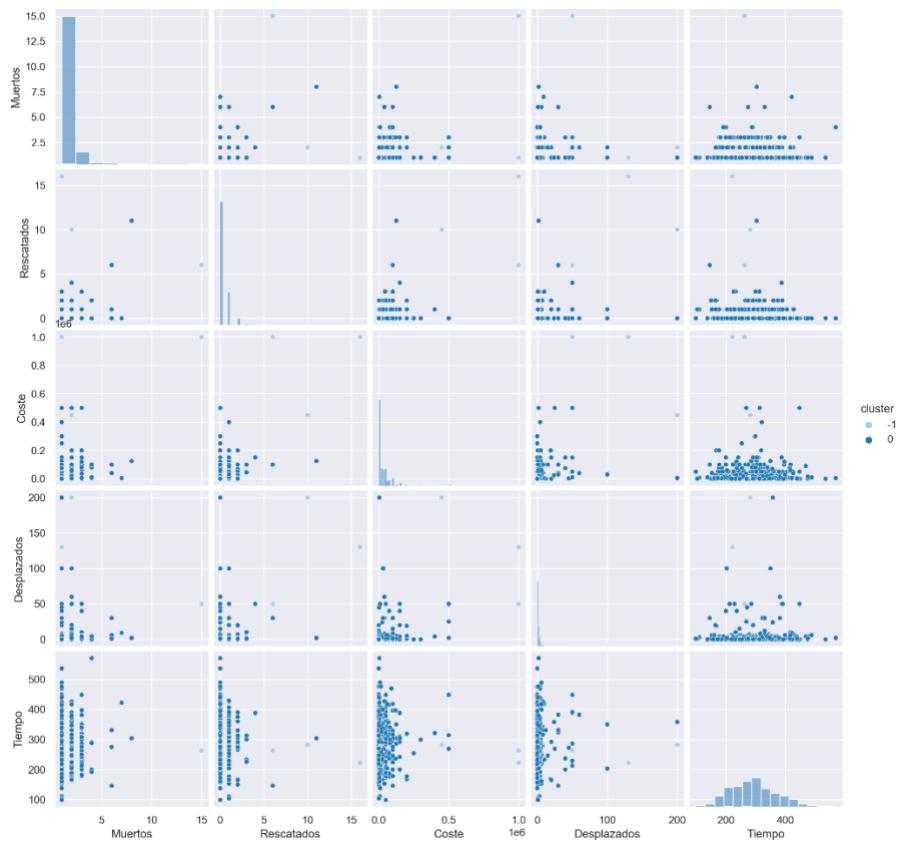


Figura 43: ScatterMatrix DBSCAN Caso 2

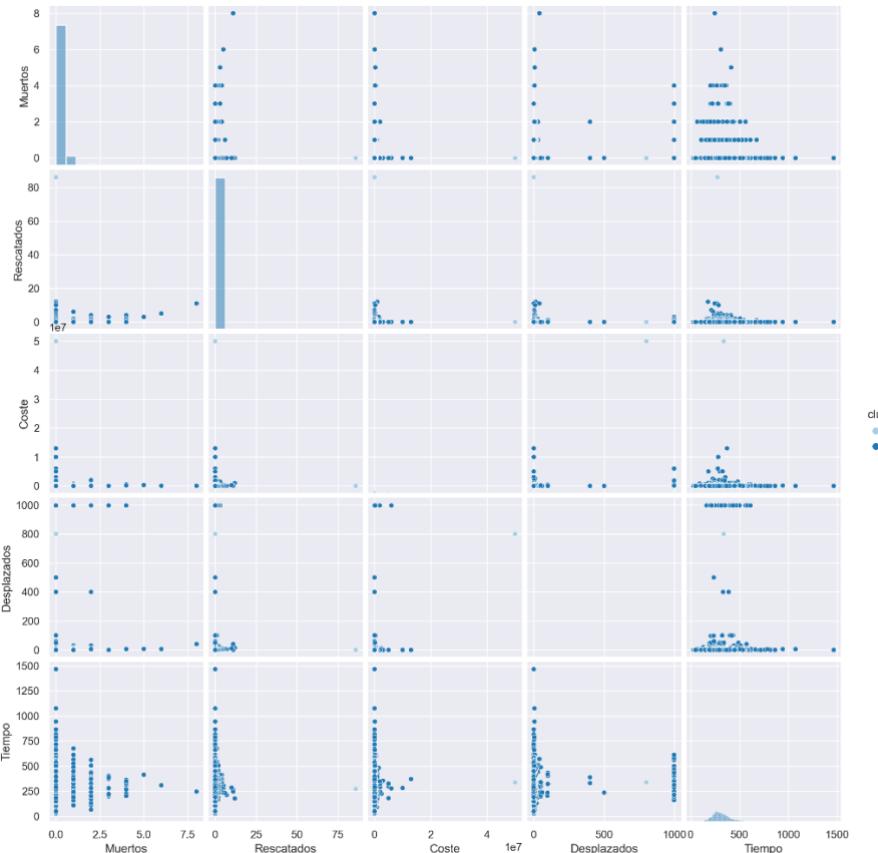


Figura 44: ScatterMatrix DBSCAN Complementario Caso 2

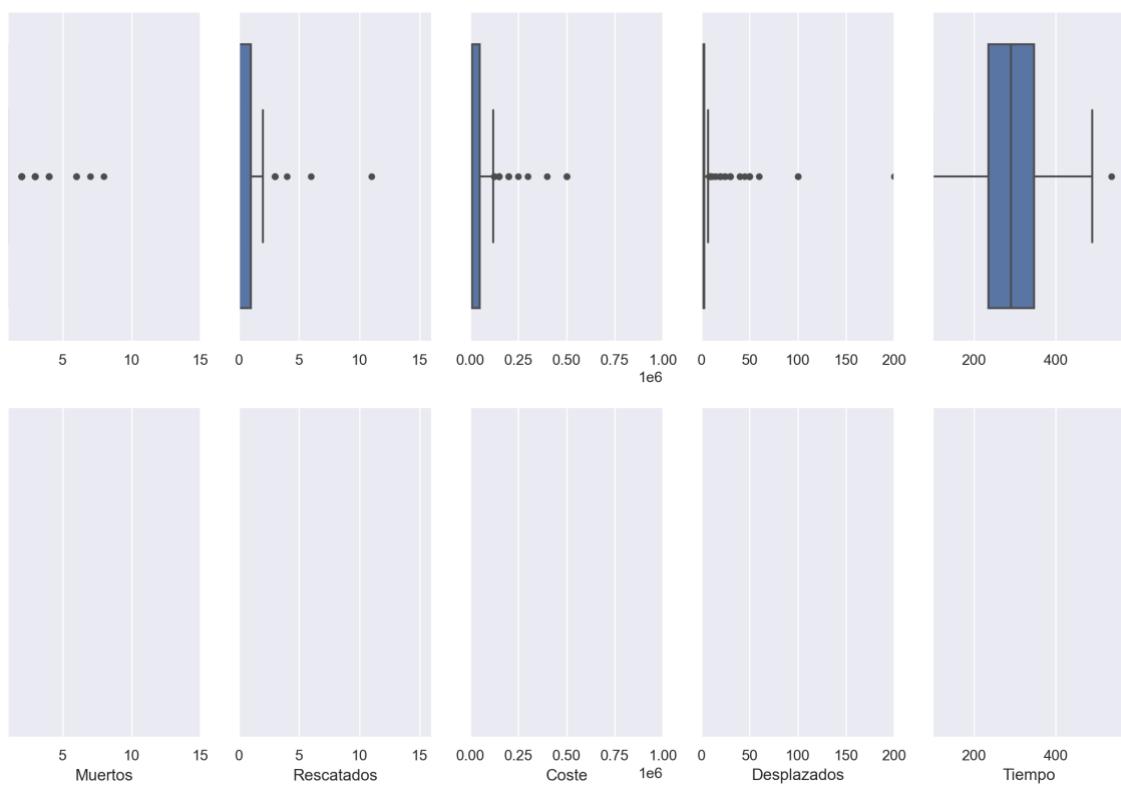


Figura 45: BoxPlot DBSCAN Caso 2

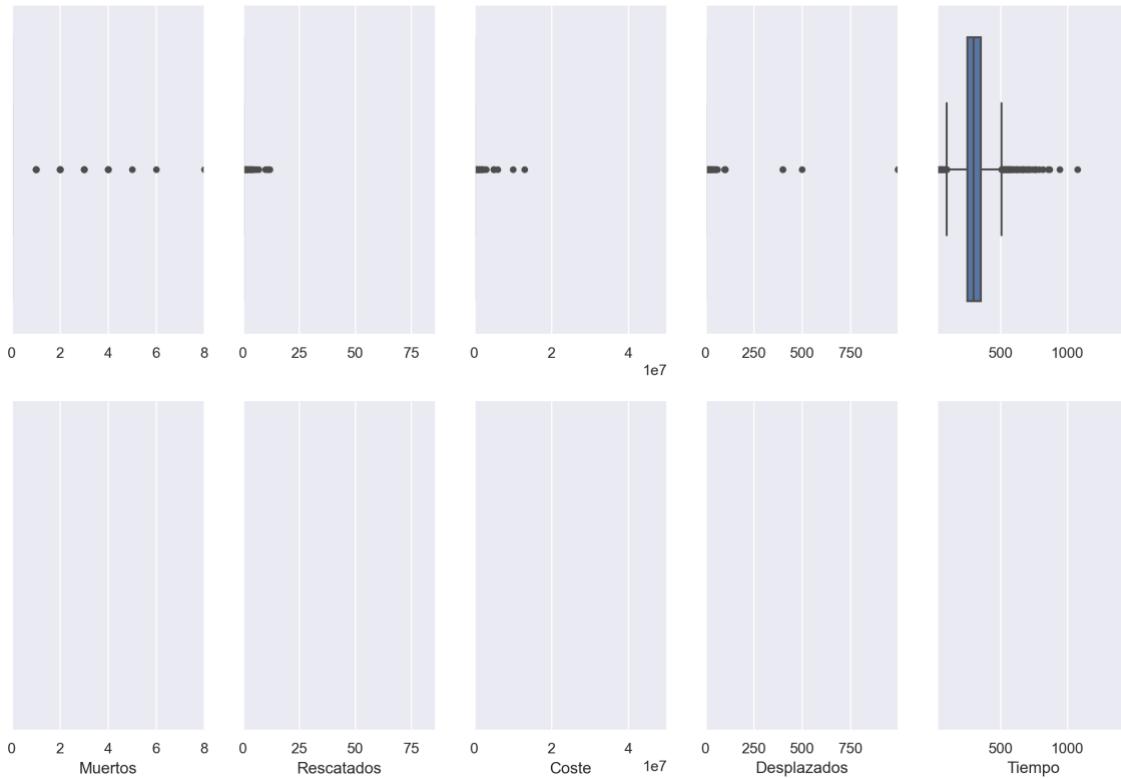


Figura 46: BoxPlot DBSCAN Complementario Caso 2

Agglomerative Clustering

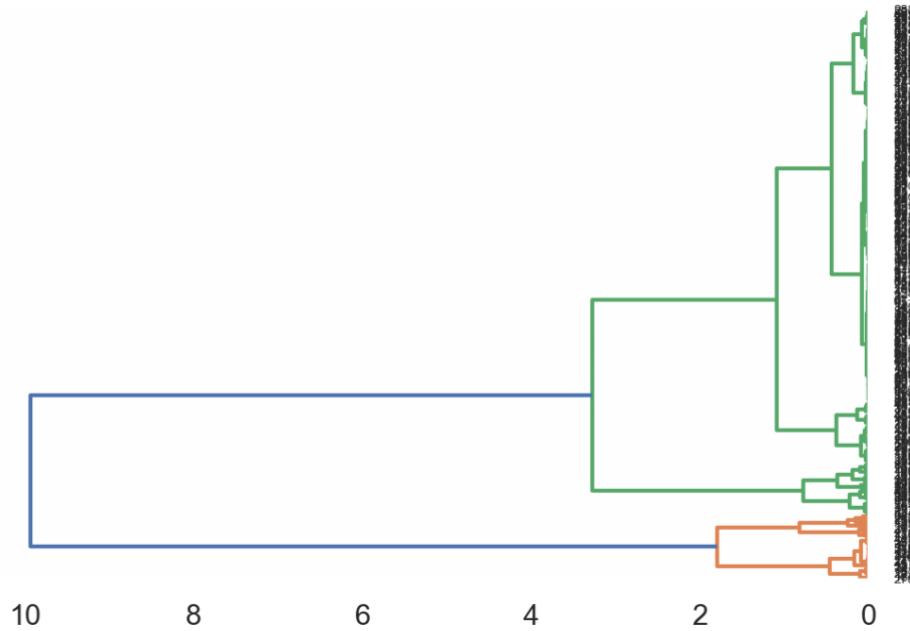


Figura 47: Dendograma Ward Caso 2

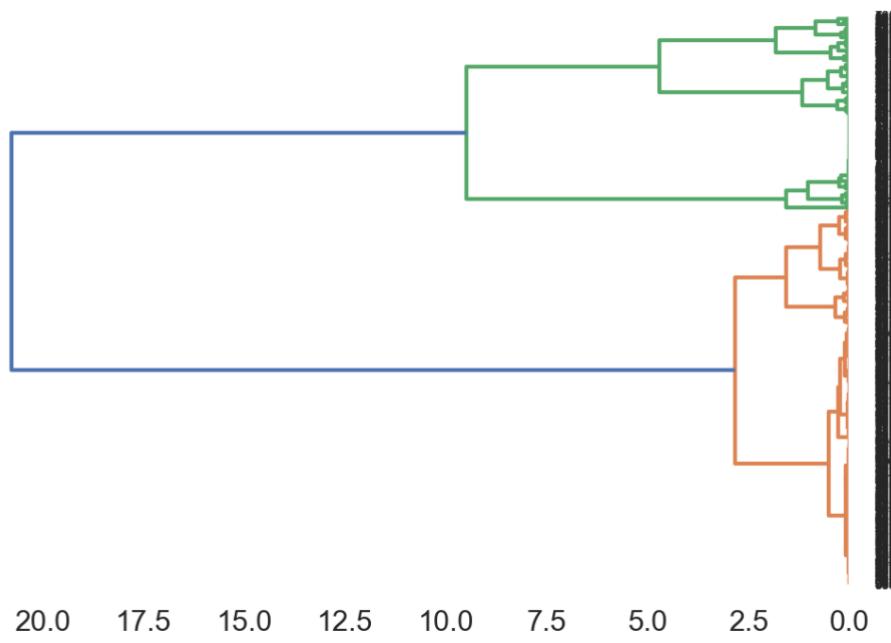


Figura 48: Dendograma Ward Complementario Caso 2

En esta caso de nuevo vemos que el clustering jerárquico ward realiza divisiones en función del coste y del tiempo como ocurre en los demás caso.

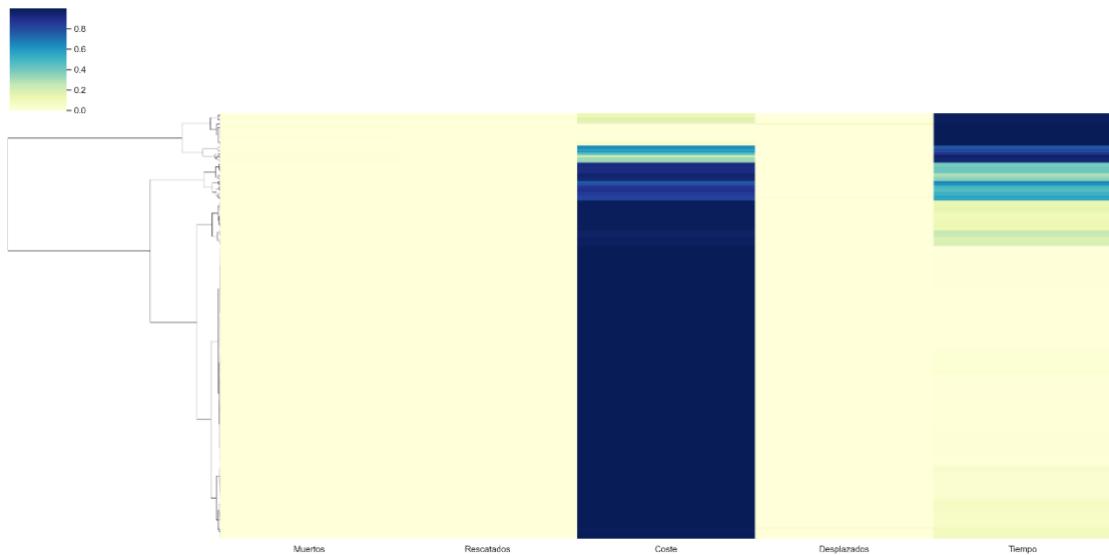


Figura 49: Dendrograma Heatmap Ward Caso 2

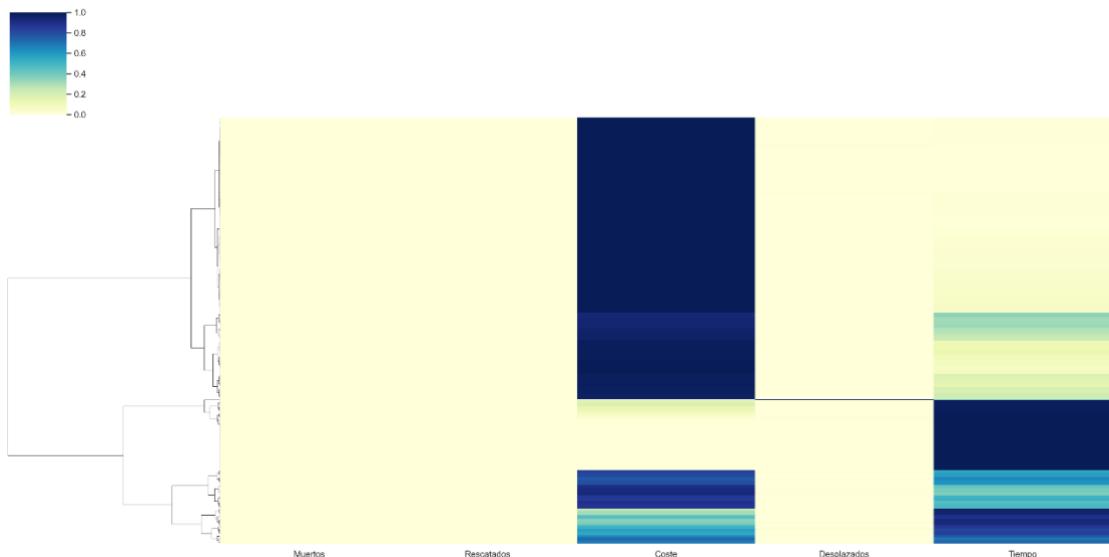


Figura 50: Dendrograma Heatmap Ward Complementario Caso 2

Birch

En Birch se consiguen también muy buenos resultados tanto para el caso de interés con para el complementario. De hecho en las figuras podemos observar como se consiguen muchas correlaciones entre las variables y los clusters generados.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
3	0: 344 (99.14%) 1: 2 (0.58%) 2: 1 (0.29%)	0.80843	59.385	0.02

Tabla 20: Métricas Birch Caso 2

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
3	1: 6345 (97.93%) 0: 133 (2.05%) 2: 1 (0.02%)	0.91904	10449.524	0.12

Tabla 21: Métricas Birch Complementario Caso 2

Interpretación de la segmentación

En este caso si se han obtenido algunas similitudes en algoritmos como K-means y Birch. Dónde por ejemplo en el tercer cluster se han correlacionado todas las variables.

En este estudio si que hemos podido conseguir resultados más fáciles de interpretar, por ejemplo en las divisiones que se observan por la leyenda de color en las ScatterMatrix de algunos algoritmos.

Algunas conclusiones que se pueden sacar son que los incendios en los que ha habido al menos una víctima provocan que las demás variables también crezcan y por lo tanto eso significa que el incendio sería complicado de extinguir y los costes aumenten, los desplazados y rescatados también.

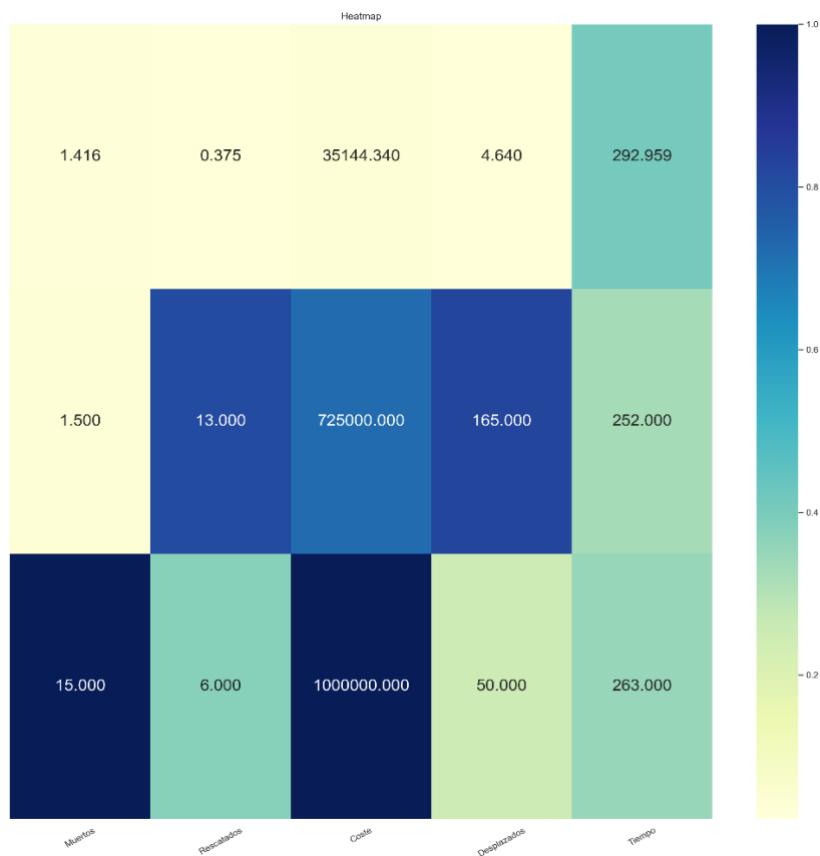


Figura 51: Heatmap Birch Caso 2

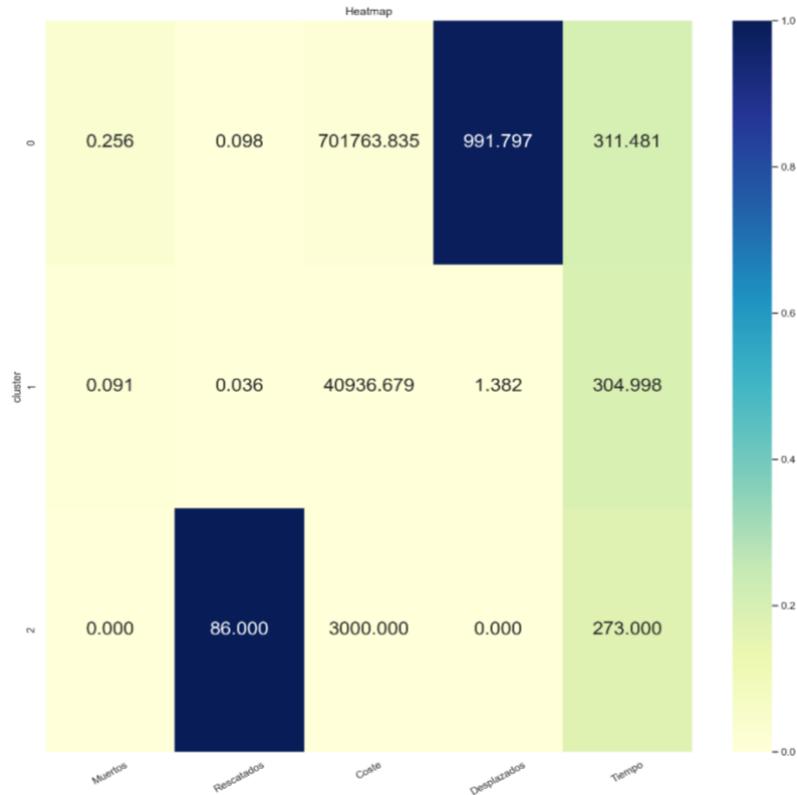


Figura 52: Heatmap Birch Complementario Caso 2

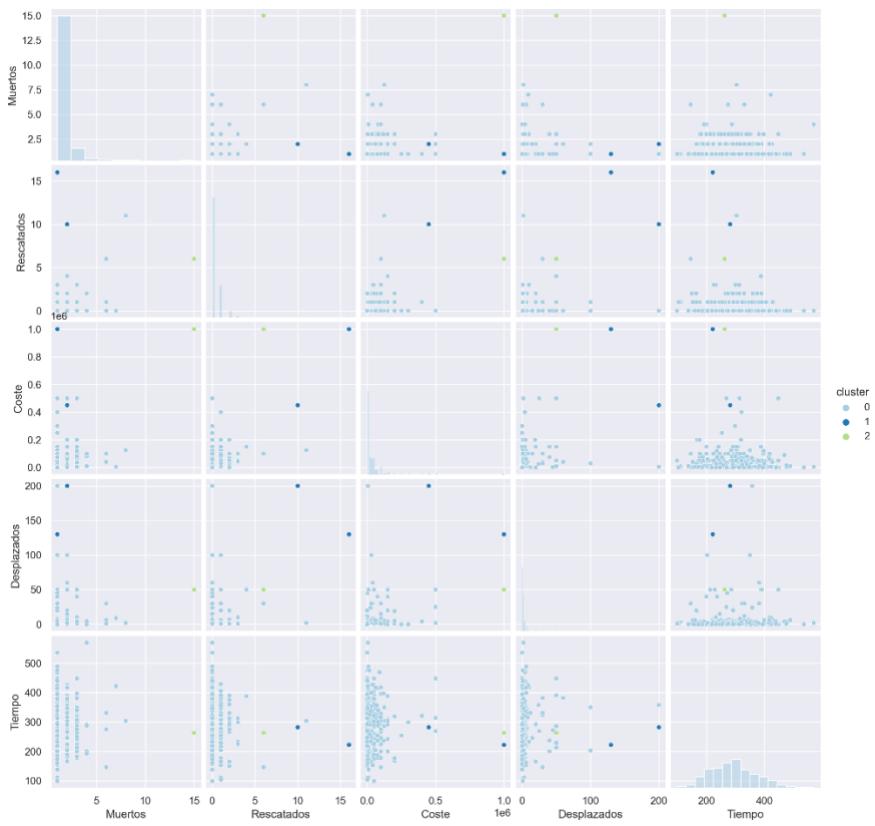


Figura 53: ScatterMatrix Birch Caso 2

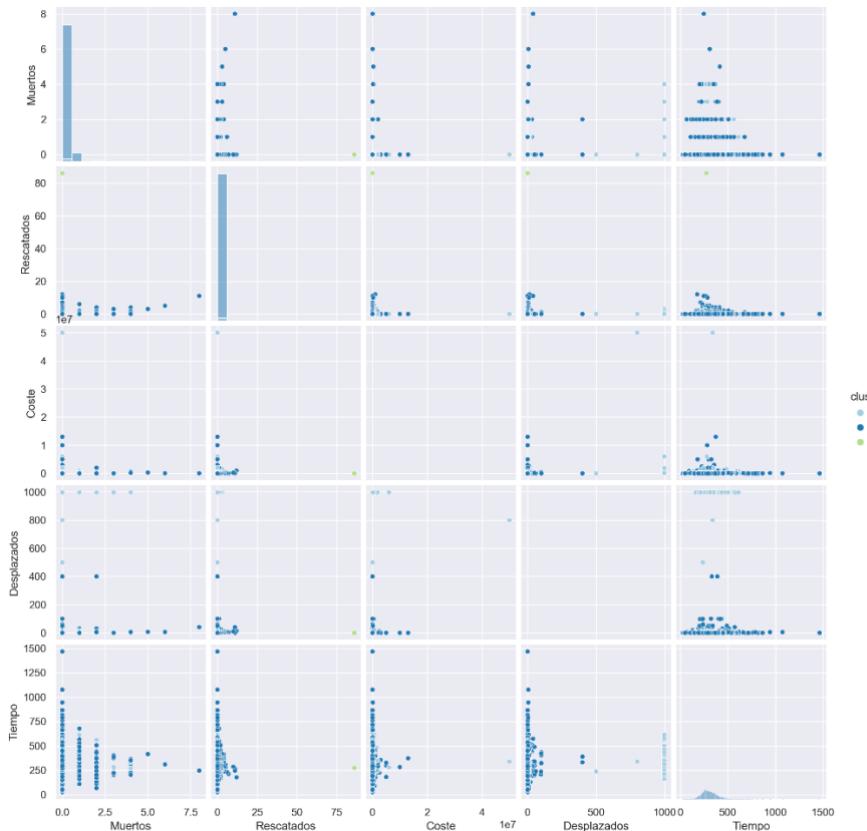


Figura 54: ScatterMatrix Birch Complementario Caso 2

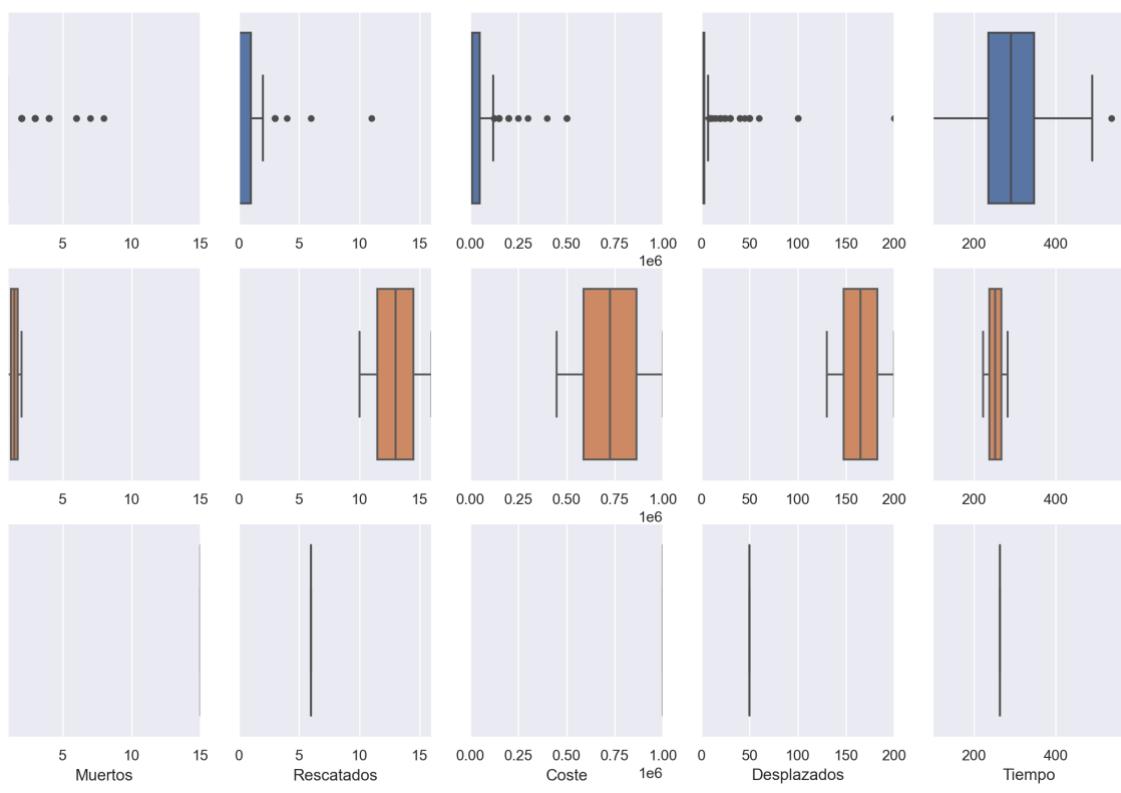


Figura 55: BoxPlot Birch Caso 2

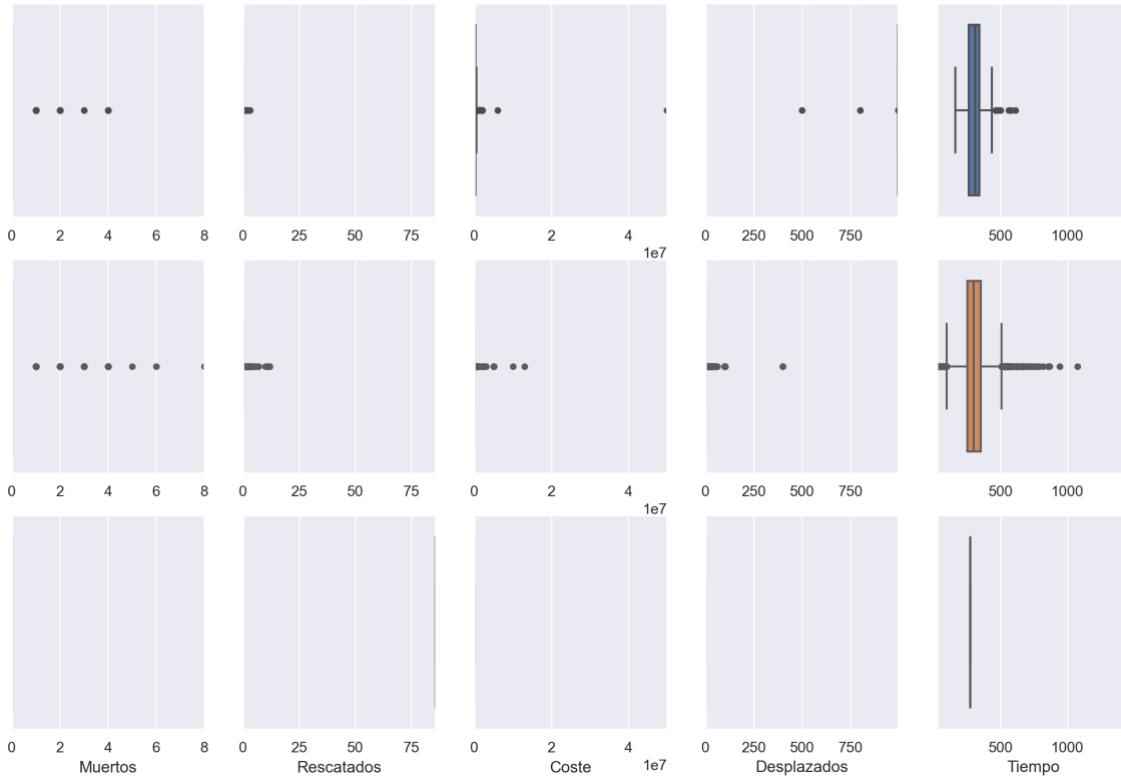


Figura 56: BoxPlot Birch Complementario Caso 2

4. Caso de estudio 3

En este último caso se van a estudiar los incendios originados por artículos de fumadores, como pueden ser colillas, cerillas... ocurridos los fines de semana de los meses de verano. Puede tener alguna característica distintiva que consigan agrupar con éxito dichos incendios con los originados por otros medios.

Se han escogido un total de 1125 ejemplos donde los días van de viernes a domingo y los meses de junio a septiembre y la variable *Ignition_Source* coincide con *Smoker's Articles* (eg. *cigarettes, cigars, pipes already ignited*). De nuevo las variables escogidas son:

- *Civilian_Casualties* -> Muertos
- *Count_Of_Person_Rescued* -> Rescatados
- *Estimated_Dollar_Loss* -> Coste
- *Estimated_Number_Of_Person_Displaced* -> Desplazados
- *Arrival_Time* -> Tiempo

Para este último caso se han obtenido métricas un tanto pobres para el caso de interés y buenas en algunos algoritmos del conjunto de datos complementario. Por un lado el mejor algoritmo en base a Silhouette es Birch mientras que si tomamos el coeficiente de Calinski-Harabasz K-means estaría en la cima. Veremos a continuación resultados más específicos en casa uno.

	Silhouette	Calinski-Harabasz	Tº ejecución	Número de Clusters
Kmeans	0.55193	207.868	0.02	4
Meanshift	0.62545	82.316	0.06	8
DBSCAN	0.64328	104.173	0.00	3
Birch	0.75310	97.203	0.01	4
AC		0.00		4

Tabla 22: Métricas Caso 3

	Silhouette	Calinski-Harabasz	Tº ejecución	Número de Clusters
Kmeans	0.50188	4272.699	0.05	4
Meanshift	0.40528	345.144	0.63	45
DBSCAN	0.84234	38.222	0.43	1
Birch	0.82193	1541.259	0.07	4
AC		0.02		4

Tabla 23: Métricas Complementario Caso 3

K-means

De nuevo ocurre que obtenemos unas medidas muy altas para el primer valor de clusters, pero que después decrece mucho y ya va aumentando hasta que llega a otro máximo local. En este caso utilizaremos ese máximo para ver cómo actúa el algoritmo.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
2	0: 217 (96.44%) 1: 8 (3.56%)	0.78620	170.389	0.01
3	0: 150 (66.67%) 2: 67 (29.78%) 1: 8 (3.56%)	0.46422	167.023	0.01
4	1: 150 (66.67%) 3: 56 (24.89%) 0: 11 (4.89%) 2: 8 (3.56%)	0.55193	207.868	0.02
5	0: 141 (62.67%) 2: 64 (28.44%) 3: 11 (4.89%) 1: 8 (3.56%) 4: 1 (0.44%)	0.53706	258.231	0.01
6	0: 103 (45.78%) 1: 62 (27.56%) 5: 40 (17.78%) 3: 11 (4.89%) 2: 8 (3.56%) 4: 1 (0.44%)	0.48642	296.521	0.02
7	6: 102 (45.33%) 0: 61 (27.11%) 1: 40 (17.78%) 3: 11 (4.89%) 2: 8 (3.56%) 4: 2 (0.89%) 5: 1 (0.44%)	0.50381	324.336	0.02
8	0: 87 (38.67%) 5: 51 (22.67%) 7: 47 (20.89%) 3: 18 (8.00%) 1: 11 (4.89%) 2: 8 (3.56%) 6: 2 (0.89%) 4: 1 (0.44%)	0.48071	353.415	0.02
9	0: 89 (39.56%) 5: 51 (22.67%) 6: 48 (21.33%) 2: 15 (6.67%) 4: 11 (4.89%) 1: 7 (3.11%) 7: 2 (0.89%) 8: 1 (0.44%) 3: 1 (0.44%)	0.47944	363.151	0.02
10	9: 88 (39.11%) 0: 50 (22.22%) 1: 47 (20.89%) 4: 18 (8.00%) 3: 10 (4.44%) 2: 7 (3.11%) 6: 2 (0.89%) 7: 1 (0.44%) 8: 1 (0.44%) 5: 1 (0.44%)	0.47633	388.555	0.03

Tabla 24: Comparación K-means Caso 3

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
2	0: 3701 (98.20%) 1: 68 (1.80%)	0.86819	4062.507	0.02
3	2: 2191 (58.13%) 0: 1510 (40.06%) 1: 68 (1.80%)	0.43586	3795.603	0.04
4	2: 2039 (54.10%) 0: 1339 (35.53%) 3: 323 (8.57%) 1: 68 (1.80%)	0.50188	4272.699	0.05
5	0: 1624 (43.09%) 2: 1163 (30.86%) 4: 596 (15.81%) 3: 318 (8.44%) 1: 68 (1.80%)	0.47722	4171.934	0.06
6	5: 1627 (43.17%) 1: 1176 (31.20%) 0: 579 (15.36%) 4: 257 (6.82%) 3: 68 (1.80%) 2: 62 (1.64%)	0.48903	4059.441	0.08
7	3: 1624 (43.09%) 0: 1145 (30.38%) 2: 609 (16.16%) 4: 295 (7.83%) 1: 68 (1.80%) 6: 22 (0.58%) 5: 6 (0.16%)	0.49191	4031.773	0.09
8	7: 1420 (37.68%) 2: 879 (23.32%) 0: 878 (23.30%) 3: 295 (7.83%) 6: 201 (5.33%) 1: 68 (1.80%) 5: 22 (0.58%) 4: 6 (0.16%)	0.48413	4101.335	0.09
9	7: 1072 (28.44%) 0: 1057 (28.04%) 8: 587 (15.57%) 4: 539 (14.30%) 6: 259 (6.87%) 3: 119 (3.16%) 1: 68 (1.80%) 2: 61 (1.62%) 5: 7 (0.19%)	0.46849	4004.156	0.10
10	0: 1072 (28.44%) 9: 1057 (28.04%) 3: 589 (15.63%) 4: 538 (14.27%) 5: 172 (4.56%) 2: 130 (3.45%) 6: 114 (3.02%) 1: 68 (1.80%) 7: 22 (0.58%) 8: 7 (0.19%)	0.46826	3966.432	0.12

Tabla 25: Comparación K-means Complementario Caso 3

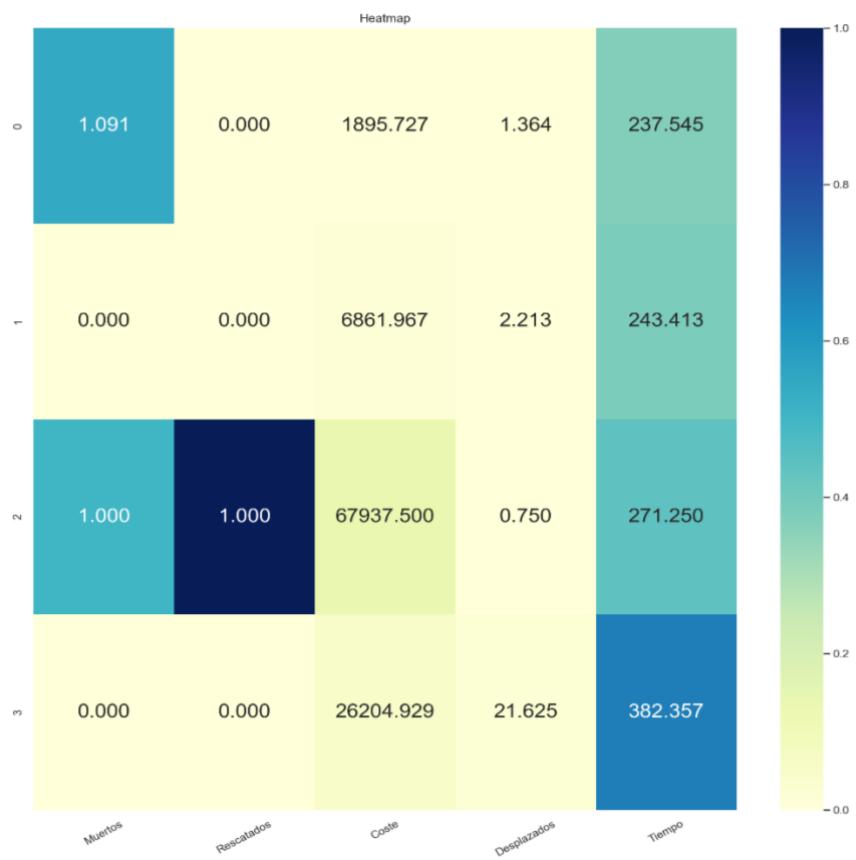


Figura 57: Heatmap K-means Caso 3

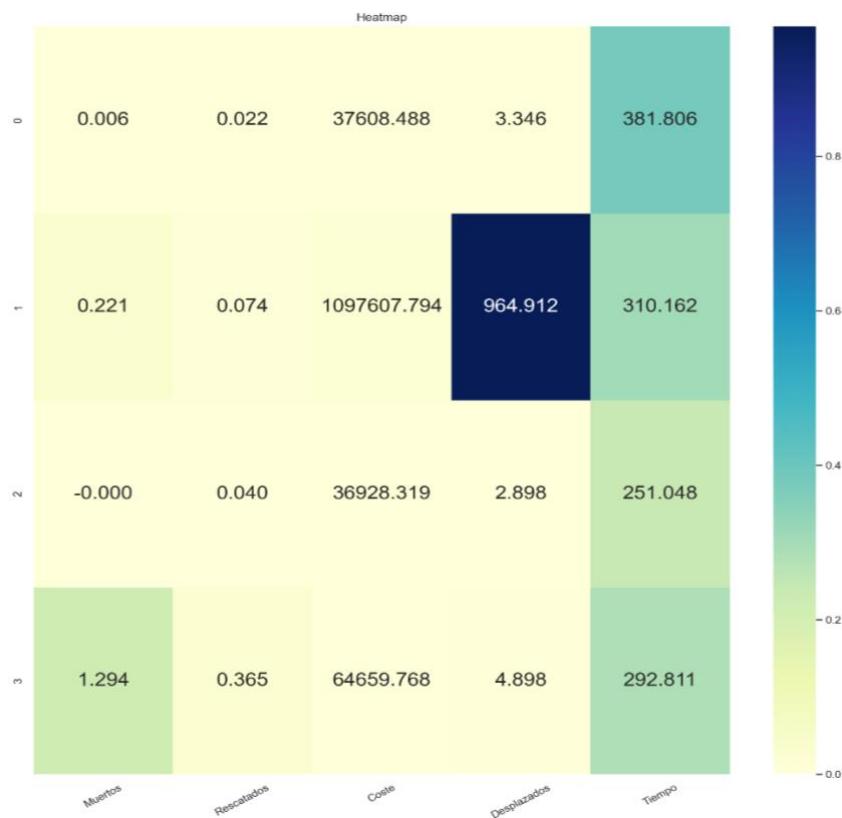


Figura 58: Heatmap K-means Complementario Caso 3

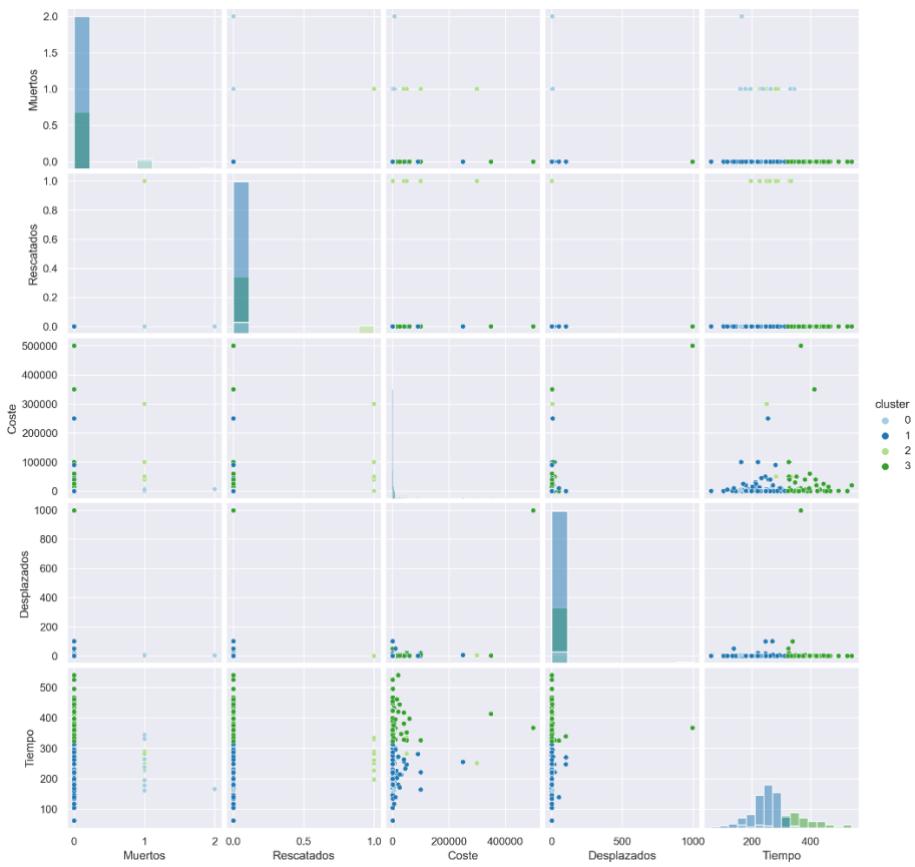


Figura 59: ScatterMatrix K-means Caso 3

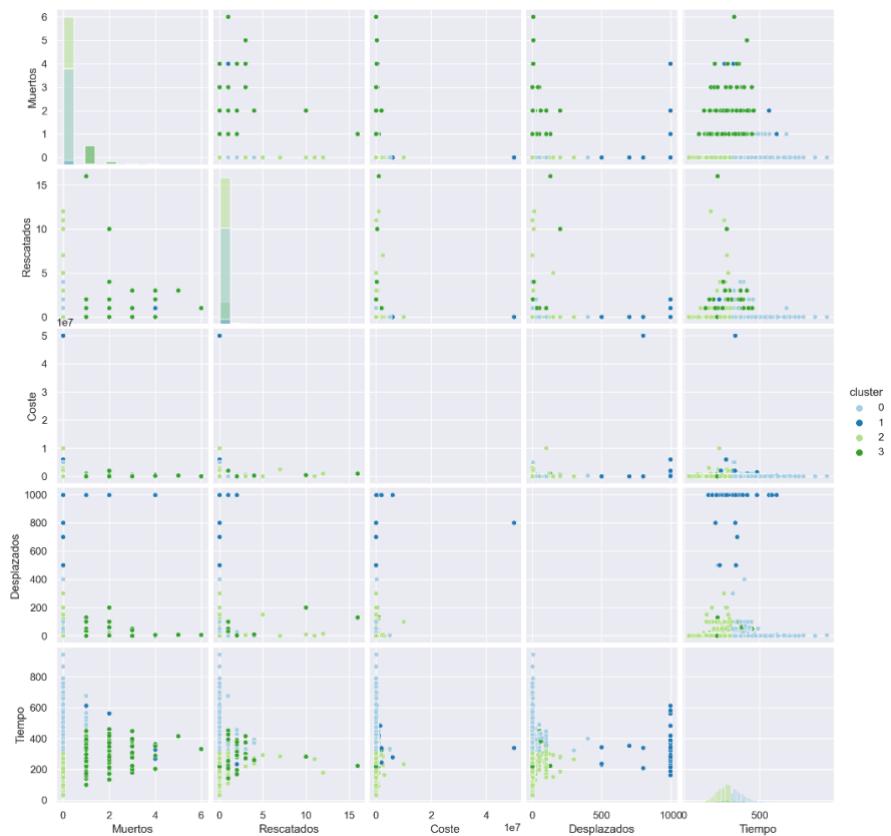


Figura 60: ScatterMatrix K-means Complementario Caso 3

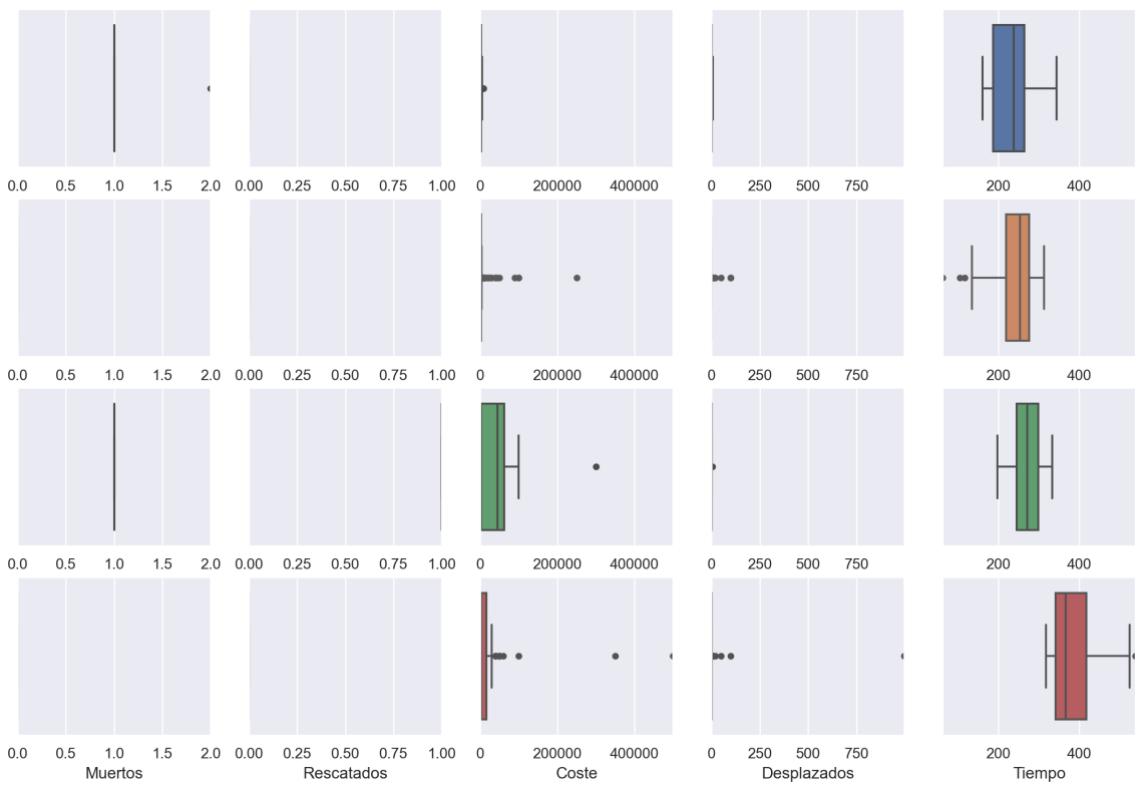


Figura 61: BoxPlot K-means Caso 3

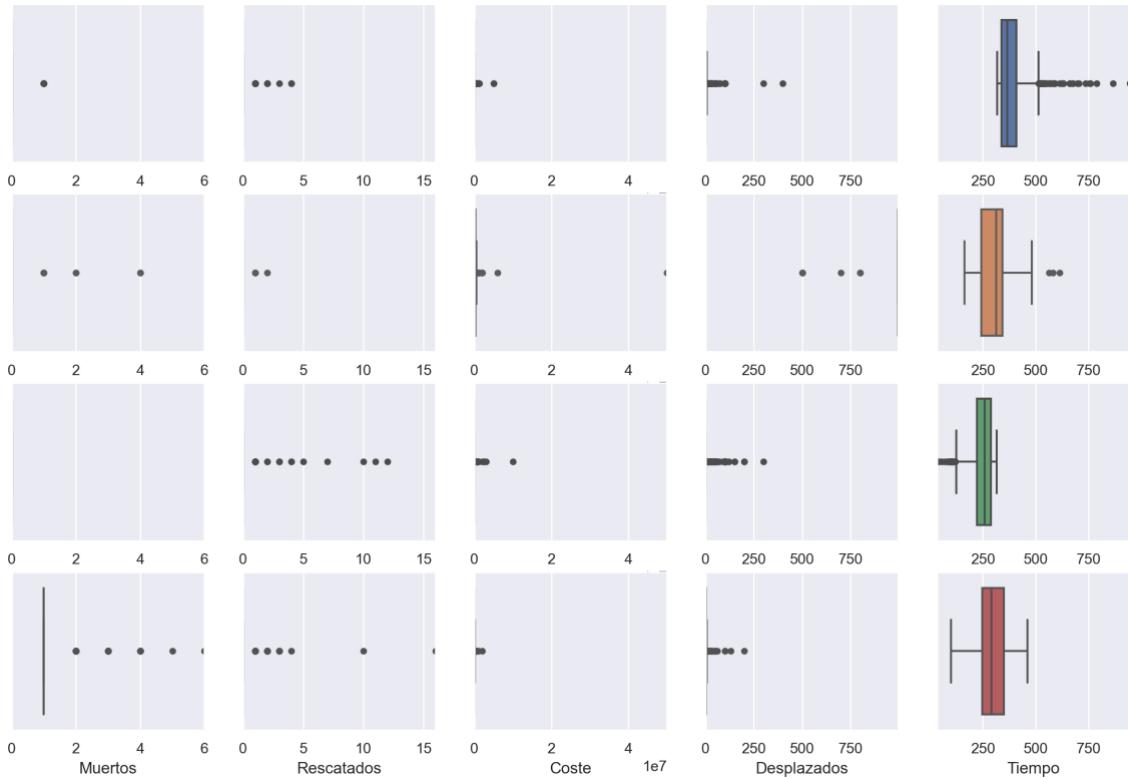


Figura 62: BoxPlot K-means Complementario Caso 3

Meanshift

En este caso, meanshift obtiene medidas mejores en el caso de estudio que en el complementario, aunque no muy buenas como ya se ha comentado.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
8	0: 203 (90.22%) 1: 10 (4.44%) 2: 7 (3.11%) 4: 1 (0.44%) 7: 1 (0.44%) 6: 1 (0.44%) 3: 1 (0.44%) 5: 1 (0.44%)	0.62545	82.316	0.05

Tabla 26: Métricas Meanshift Caso 3

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
45	0: 3256 (86.39%) 1: 257 (6.82%) 5: 54 (1.43%) 2: 53 (1.41%) 18: 36 (0.96%) 3: 17 (0.45%) 4: 16 (0.42%) 7: 7 (0.19%) 9: 6 (0.16%) 8: 6 (0.16%) 32: 5 (0.13%) 6: 5 (0.13%) 13: 5 (0.13%) 17: 4 (0.11%) 11: 4 (0.11%) 10: 3 (0.08%) 14: 2 (0.05%) 41: 2 (0.05%) 15: 2 (0.05%) 20: 2 (0.05%) 16: 2 (0.05%) 19: 2 (0.05%) 44: 1 (0.03%) 43: 1 (0.03%) 39: 1 (0.03%) 33: 1 (0.03%) 24: 1 (0.03%) 42: 1 (0.03%) 38: 1 (0.03%) 30: 1 (0.03%) 21: 1 (0.03%) 29: 1 (0.03%) 40: 1 (0.03%) 25: 1 (0.03%) 23: 1 (0.03%) 35: 1 (0.03%) 28: 1 (0.03%) 12: 1 (0.03%) 31: 1 (0.03%) 22: 1 (0.03%) 27: 1 (0.03%) 37: 1 (0.03%) 36: 1 (0.03%) 26: 1 (0.03%) 34: 1 (0.03%)	0.40528	345.144	0.63

Tabla 27: Métricas Meanshift Complementario Caso 3



Figura 63: Heatmap Meanshift Caso 3

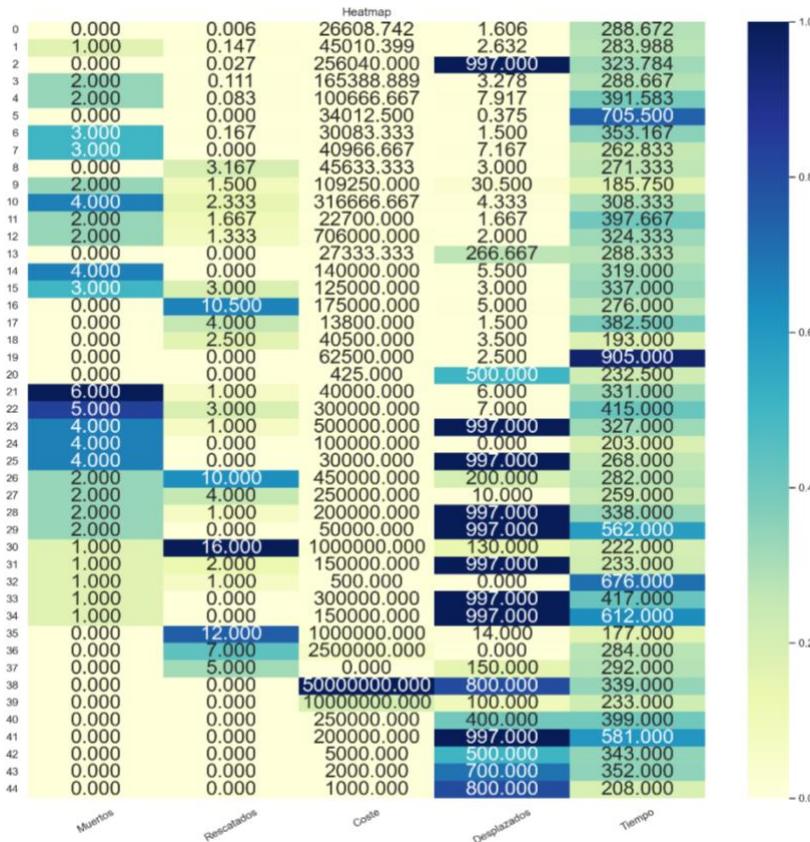


Figura 64: Heatmap Meanshift Complementario Caso 3

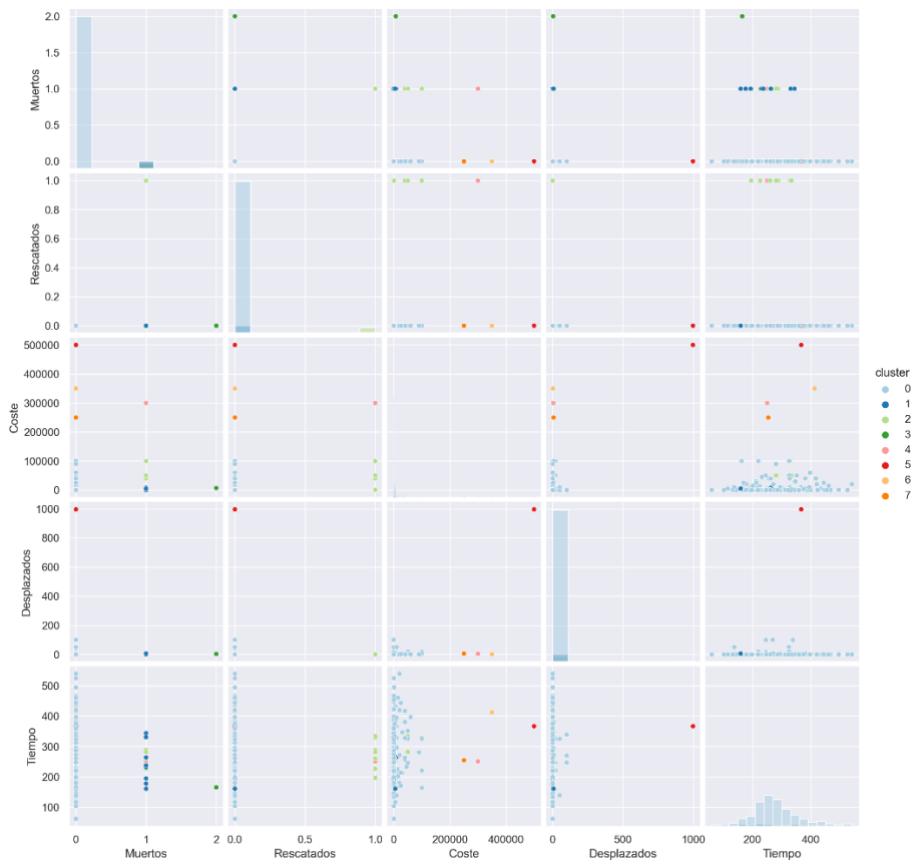


Figura 65: ScatterMatrix Meanshift Caso 3

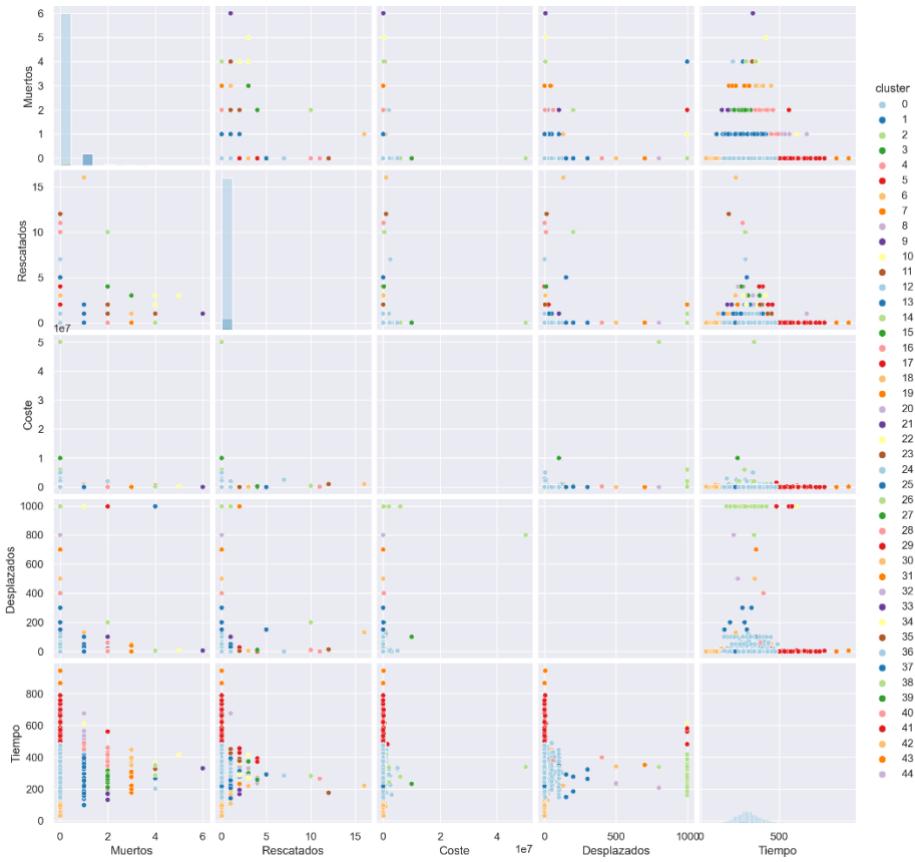


Figura 66: ScatterMatrix Meanshift Complementario Caso 3

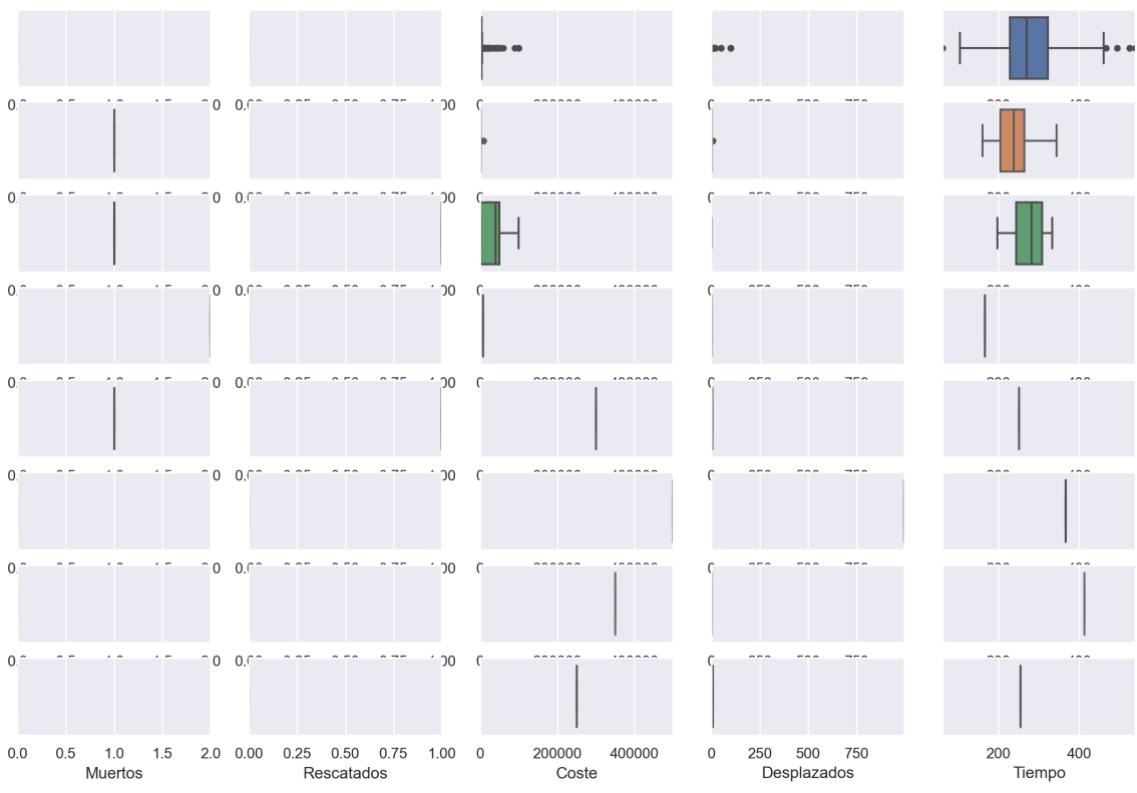


Figura 67: BoxPlot Meanshift Caso 3

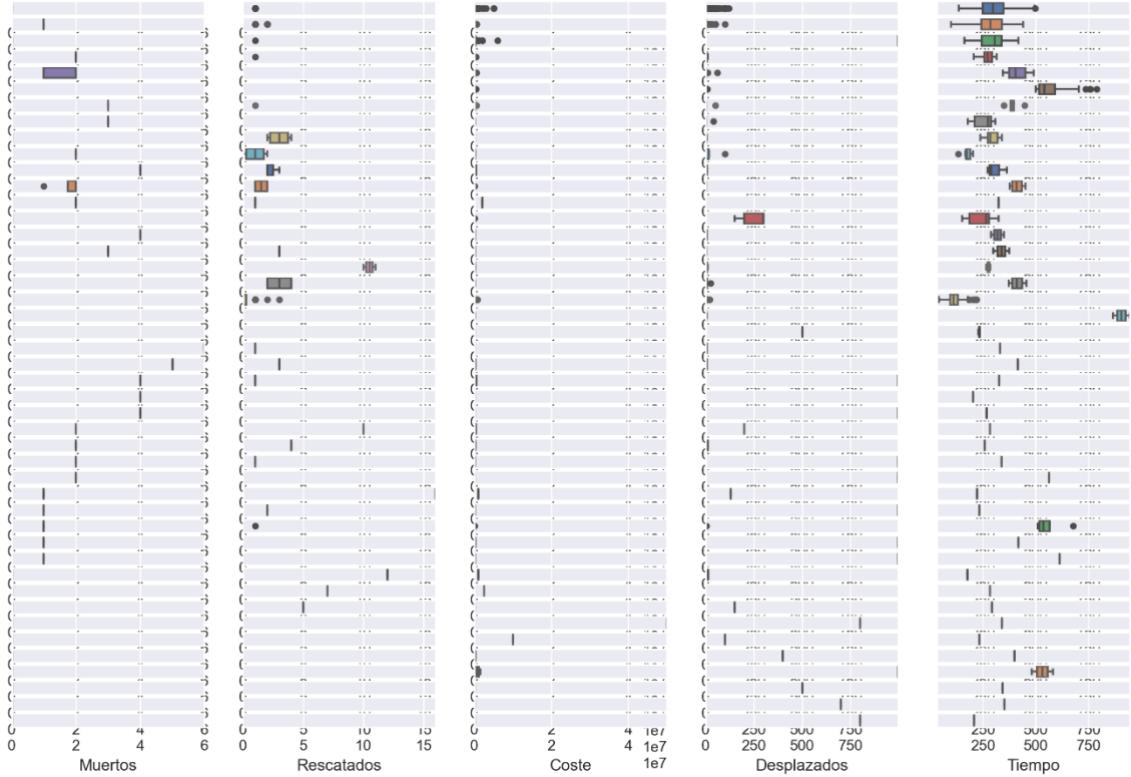


Figura 68: BoxPlot Meanshift Complementario Caso 3

DBSCAN

En el algoritmo de DBSCAN se observa que más métricas van aumentando conforme aumenta el valor épsilon hasta llegar a un máximo local para el valor de 0.35. Sin embargo, en el caso complementario ese máximo de encuentra antes y no coincide con el que vamos a observar.

Epsilon	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
0.15	1: 203 (90.22%) 2: 9 (4.00%) -1: 7 (3.11%) 0: 6 (2.67%)	0.63142	84.338	0.01
0.2	1: 203 (90.22%) 2: 10 (4.44%) 0: 7 (3.11%) -1: 5 (2.22%)	0.63142	104.471	0.01
0.25	1: 203 (90.22%) 2: 10 (4.44%) 0: 7 (3.11%) -1: 5 (2.22%)	0.64370	104.471	0.00
0.3	1: 203 (90.22%) 2: 10 (4.44%) 0: 7 (3.11%) -1: 5 (2.22%)	0.64370	104.471	0.00
0.35	1: 204 (90.67%) 2: 10 (4.44%) 0: 7 (3.11%) -1: 4 (1.78%)	0.64328	104.173	0.00
0.45	1: 205 (91.11%) 2: 10 (4.44%) 0: 7 (3.11%) -1: 3 (1.33%)	0.64037	102.383	0.00

Tabla 28: Comparación DBSCAN Caso 3

Epsilon	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
0.15	0: 3363 (89.23%) 1: 268 (7.11%) 5: 55 (1.46%) 2: 39 (1.03%) -1: 25 (0.66%) 4: 13 (0.34%) 3: 6 (0.16%)	0.51353	1153.303	0.34
0.2	0: 3695 (98.04%) 1: 59 (1.57%) -1: 15 (0.40%)	0.85142	2028.314	0.38
0.25	0: 3762 (99.81%) -1: 7 (0.19%)	0.83976	104.826	0.45
0.3	0: 3763 (99.84%) -1: 6 (0.16%)	0.84569	102.806	0.49
0.35	0: 3767 (99.95%) -1: 2 (0.05%)	0.84234	38.222	0.48
0.45	0: 3768 (99.97%) -1: 1 (0.03%)	0.87282	51.338	0.48

Tabla 29: Comparación DBSCAN Complementario Caso 3

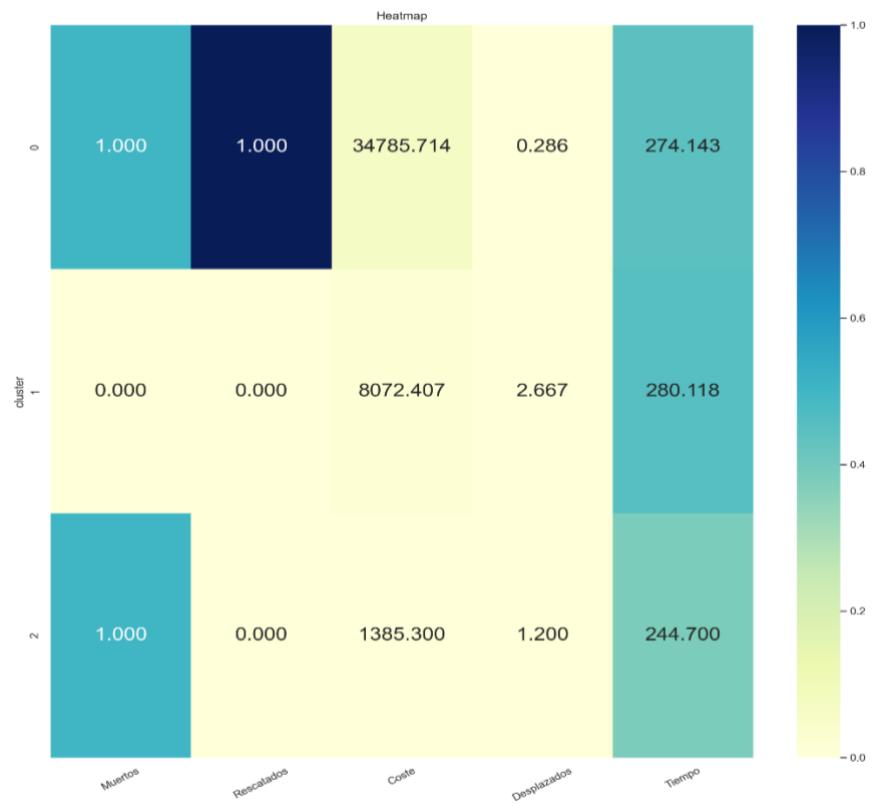


Figura 69: Heatmap DBSCAN Caso 3

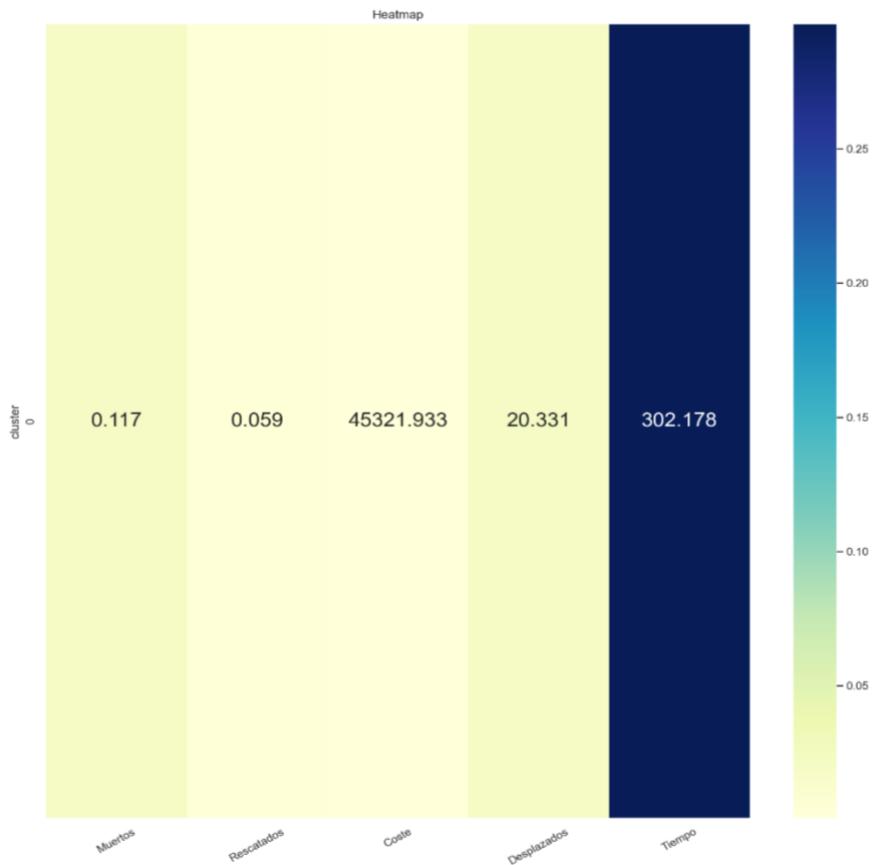


Figura 70: Heatmap DBSCAN Complementario Caso 3

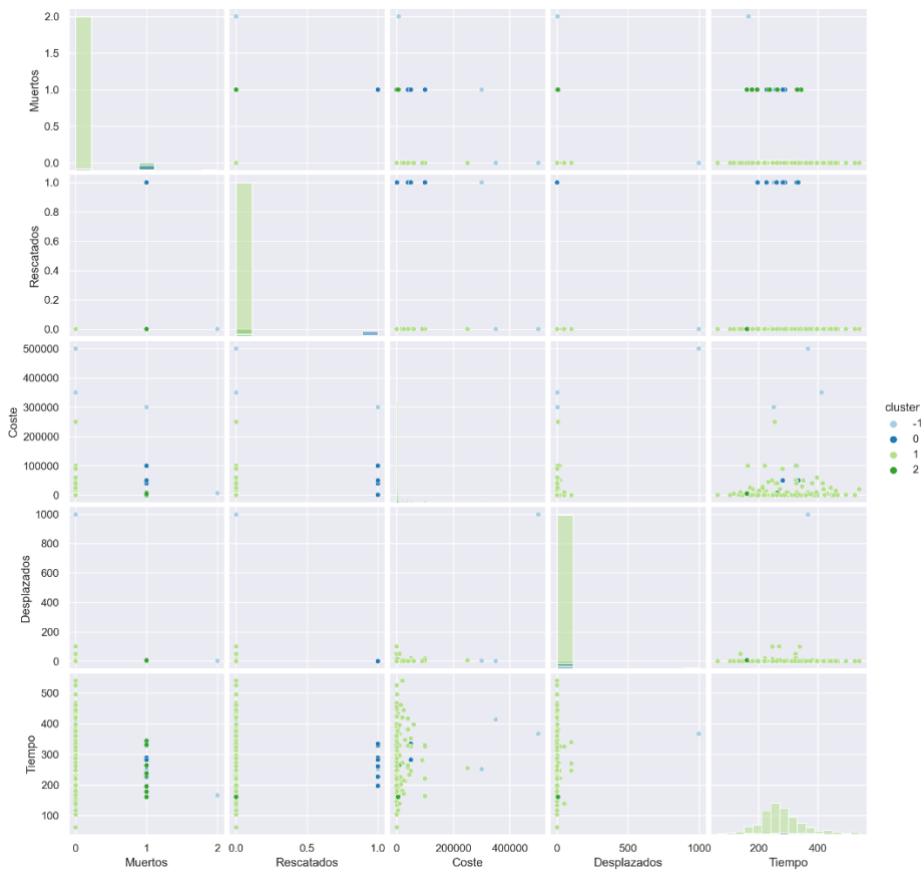


Figura 71: ScatterMatrix DBSCAN Caso 3

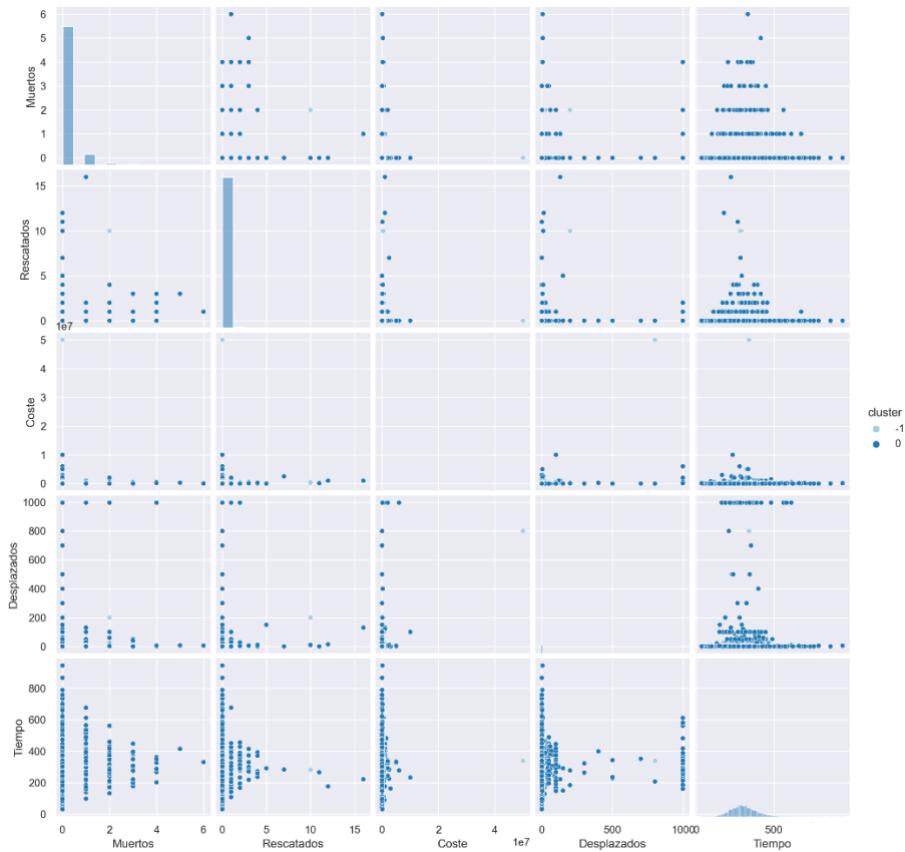


Figura 72: ScatterMatrix DBSCAN Complementario Caso 3

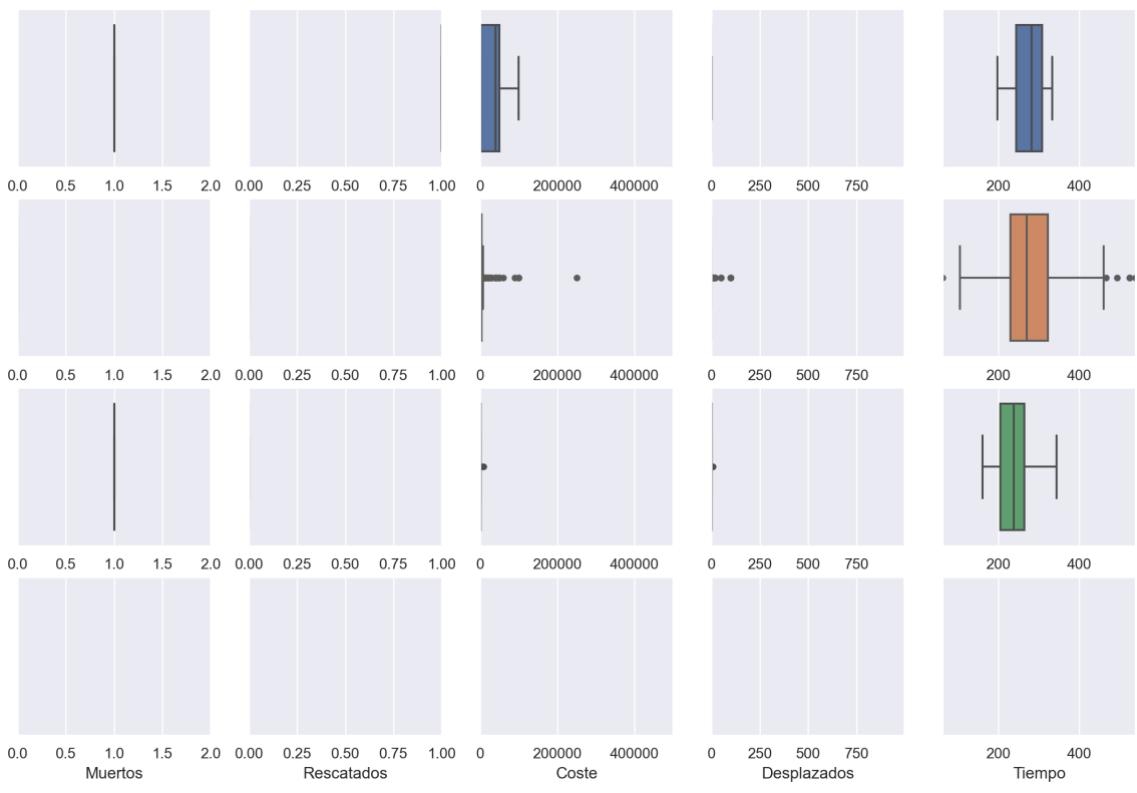


Figura 73: BoxPlot DBSCAN Caso 3

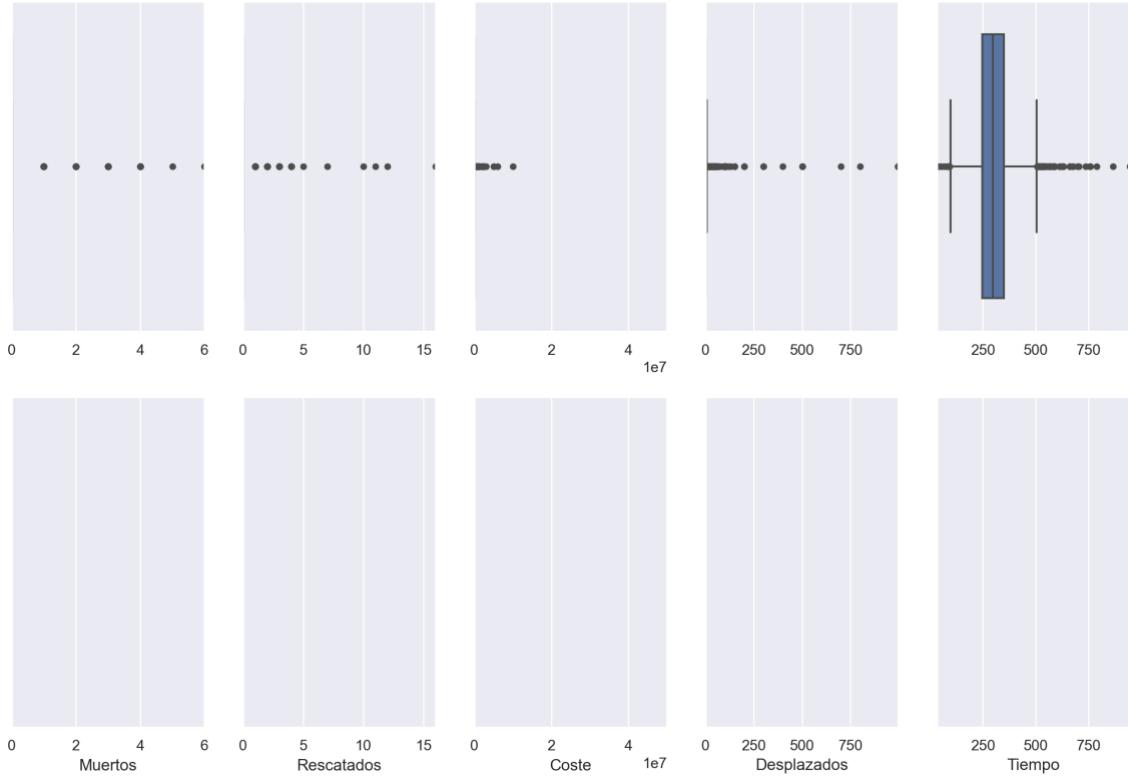


Figura 74: BoxPlot DBSCAN Complementario Caso 3

Agglomerative Clustering

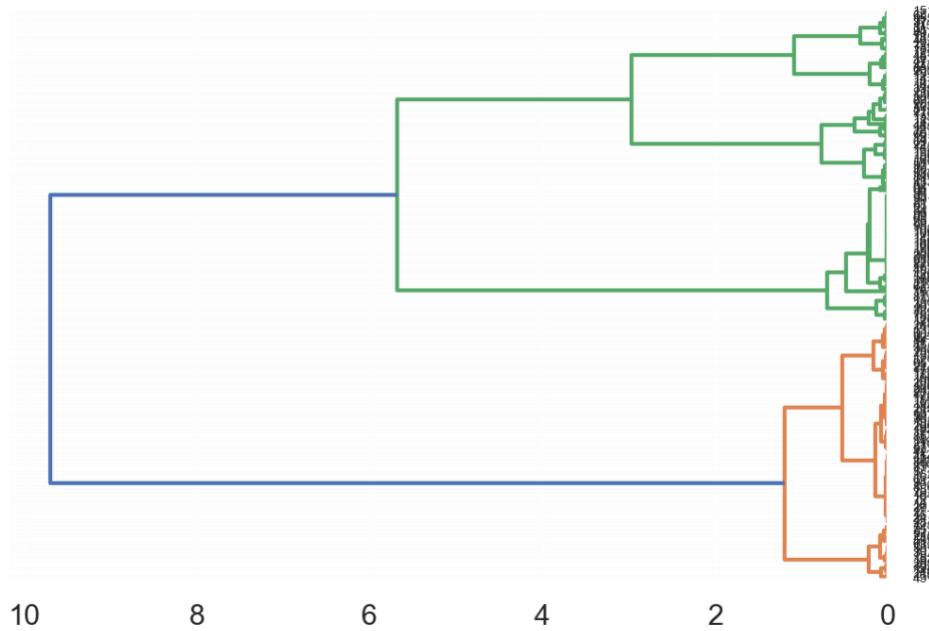


Figura 75: Dendograma Ward Caso 3

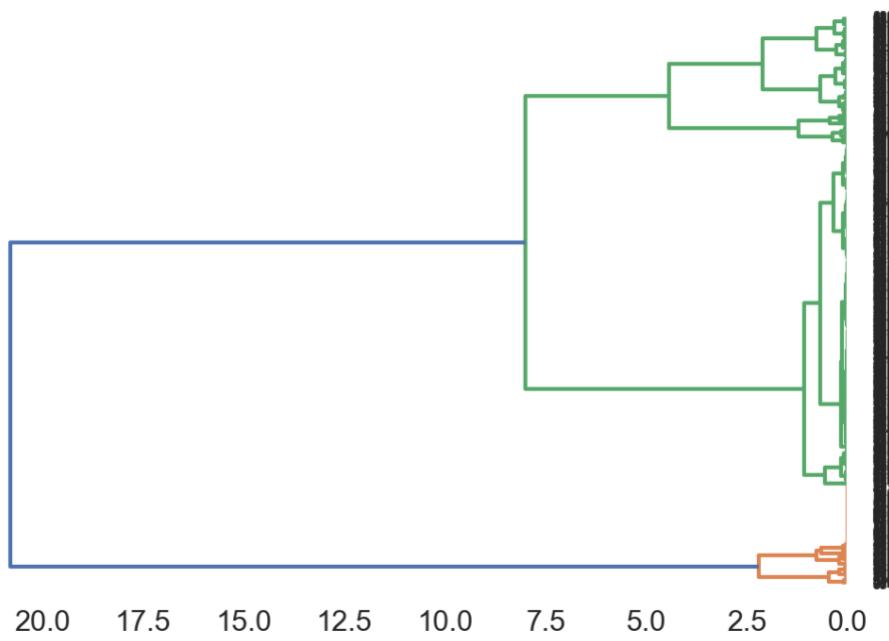


Figura 76: Dendograma Ward Complementario Caso 3

Para el algoritmo de clustering jerárquico volvemos a obtener dendogramas similares a los que vimos en el primer caso. Dónde las divisiones estaban altamente influenciadas por las variables coste y tiempo, tanto en el caso de estudio con en el complementario, aunque en este último el tiempo no ha sido tan relevante.

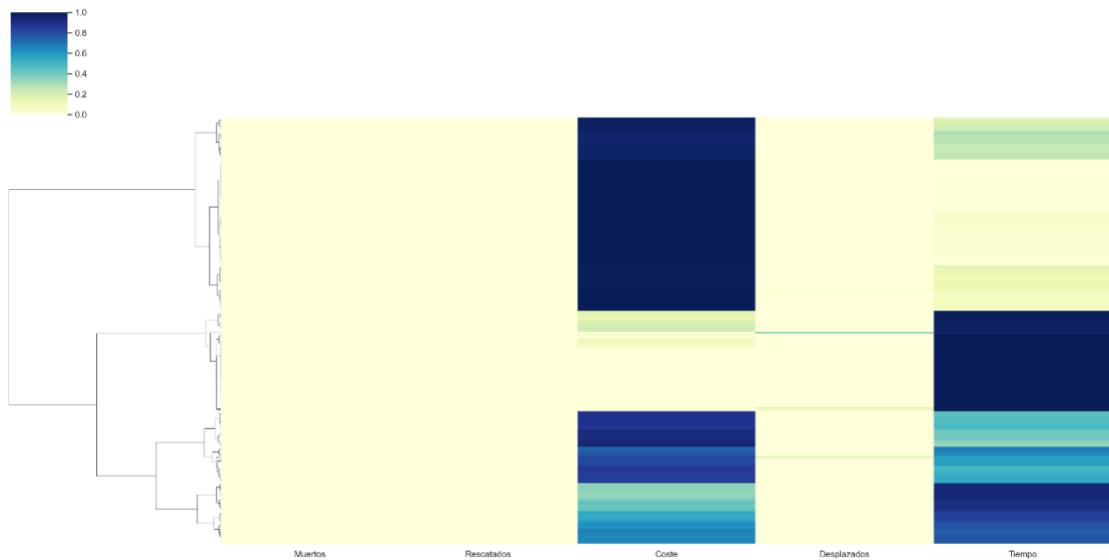


Figura 77: Dendrograma Heatmap Ward Caso 3

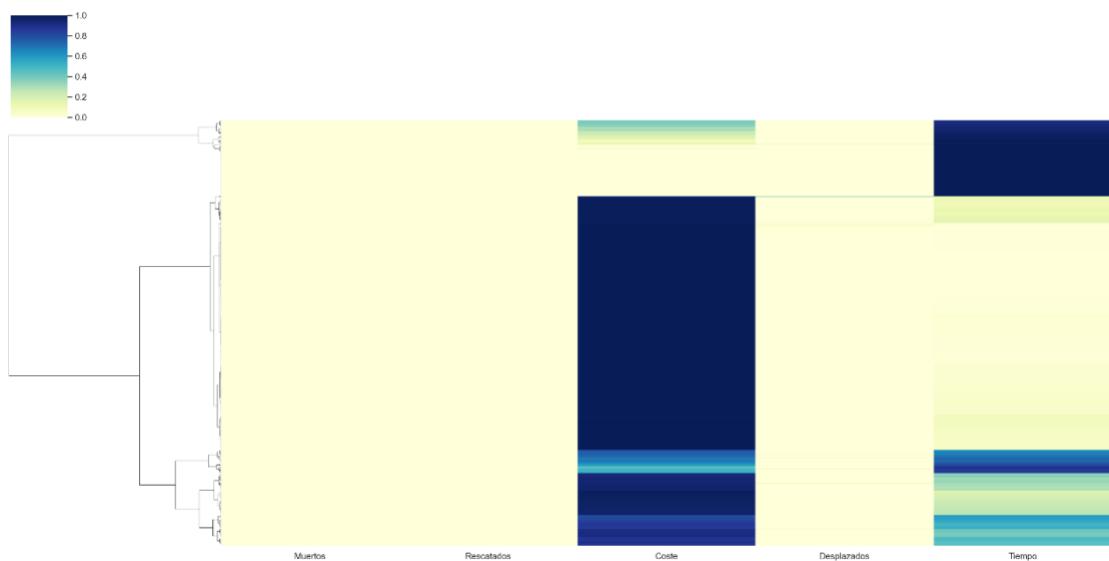


Figura 78: Dendrograma Heatmap Ward Complementario Caso 3

Birch

Por último, el algoritmo de Birch obtiene los mejores resultados como hemos podido observar en varios casos. De hecho obtiene una mejor segmentación para el conjunto complementario y donde se supera con creces el coeficiente de Calinski-Harabasz.

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
4	0: 215 (95.56%) 2: 8 (3.56%) 3: 1 (0.44%) 1: 1 (0.44%)	0.75310	97.203	0.01

Tabla 30: Métricas Birch Caso 3

Número de Clusters	Tamaño de cada cluster	Silhouette	Calinski-Harabasz	Tº ejecución
4	0: 3695 (98.04%) 1: 67 (1.78%) 3: 6 (0.16%) 2: 1 (0.03%)	0.82193	1541.259	0.07

Tabla 31: Métricas Birch Complementario Caso 3

Interpretación de la segmentación

En este último caso, vuelve a ocurrir que algunos algoritmos difieren de otros en el clustering. Pero se vuelve a ver qué las variables tiempo y coste están relacionadas, además de variables como muertos y rescatados.

Es importante entender que los artículos de fumadores son utensilios que cualquiera puede tener, como pueden ser unas cerillas o un mechero. En Canadá, como en muchos países, existe una gran adicción al tabaco, de hecho, estudios avalan que la mayoría de fumadores intentan al menos dejar de fumar treinta veces.

Por eso, como es una actividad social podemos sacar una idea de que esos incendios son provocados en lugares cerrados con gente en los alrededores que tiene que ser puesta a salvo por eso que existe esa relación entre rescatados y muertos.

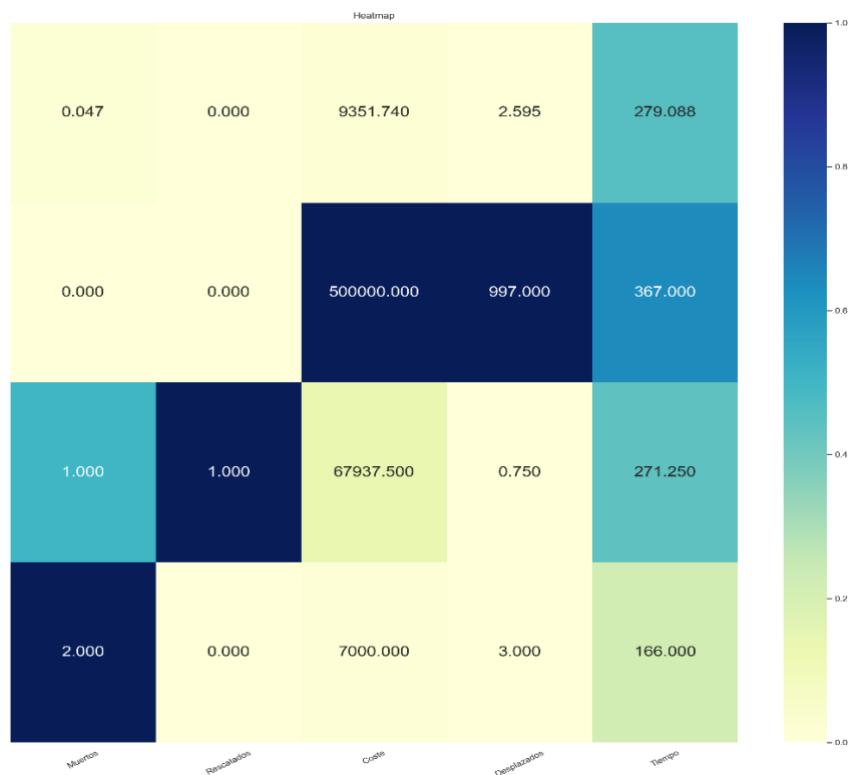


Figura 79: Heatmap Birch Caso 3

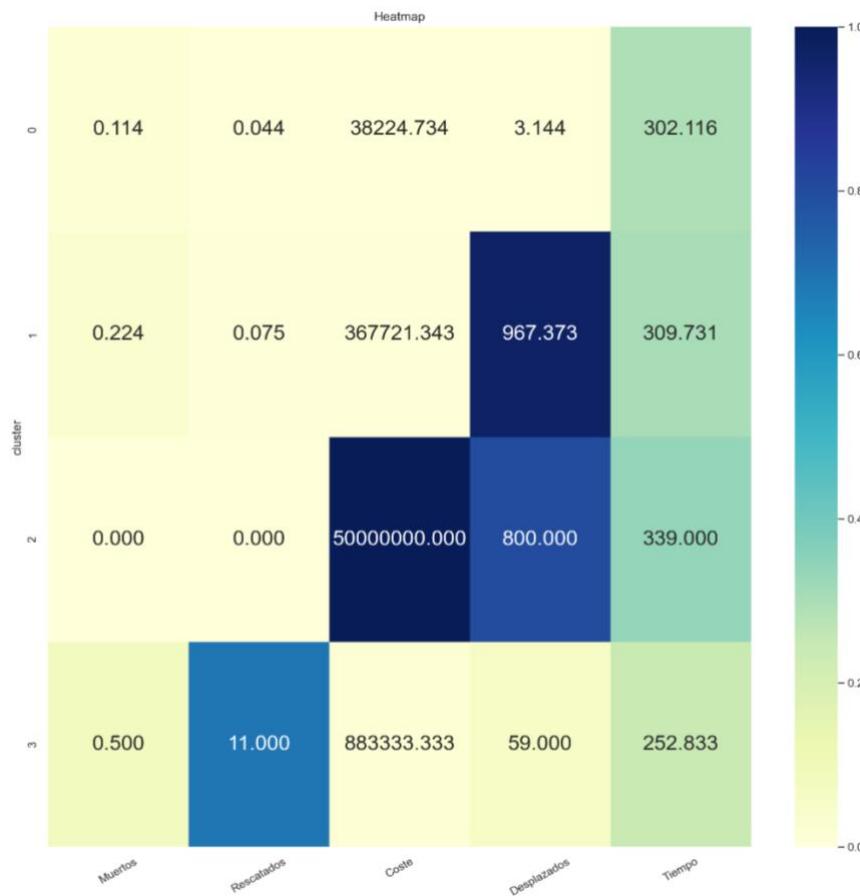


Figura 80: Heatmap Birch Complementario Caso 3

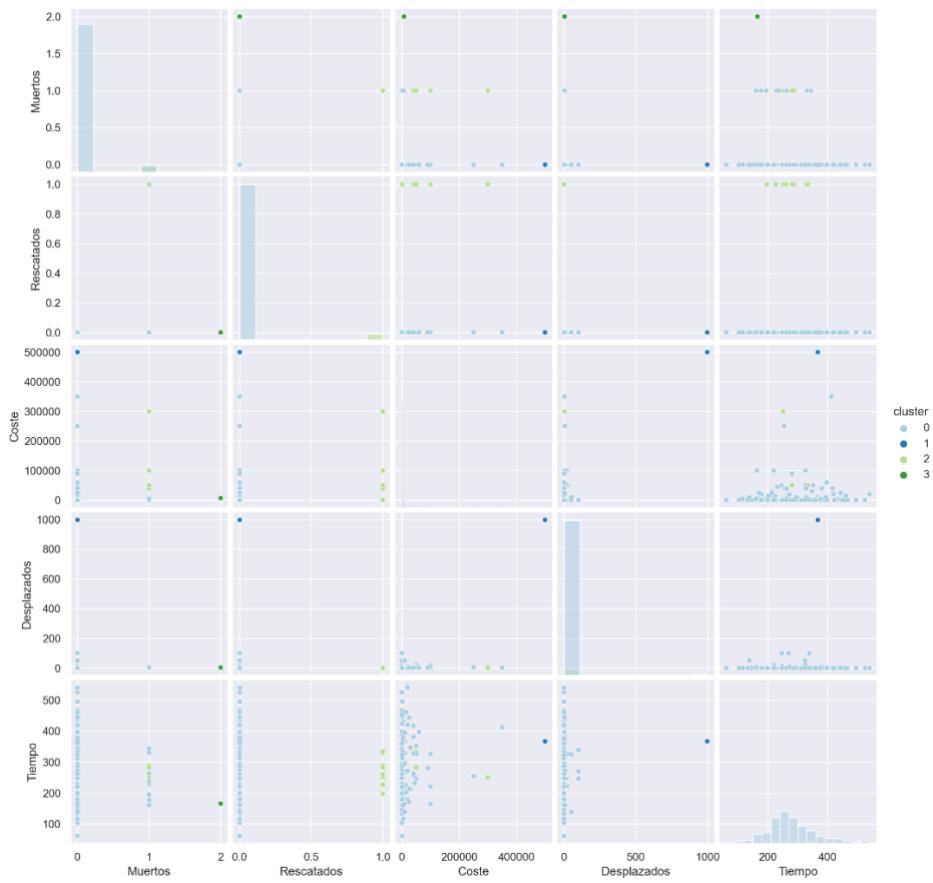


Figura 81: ScatterMatrix Birch Caso 3

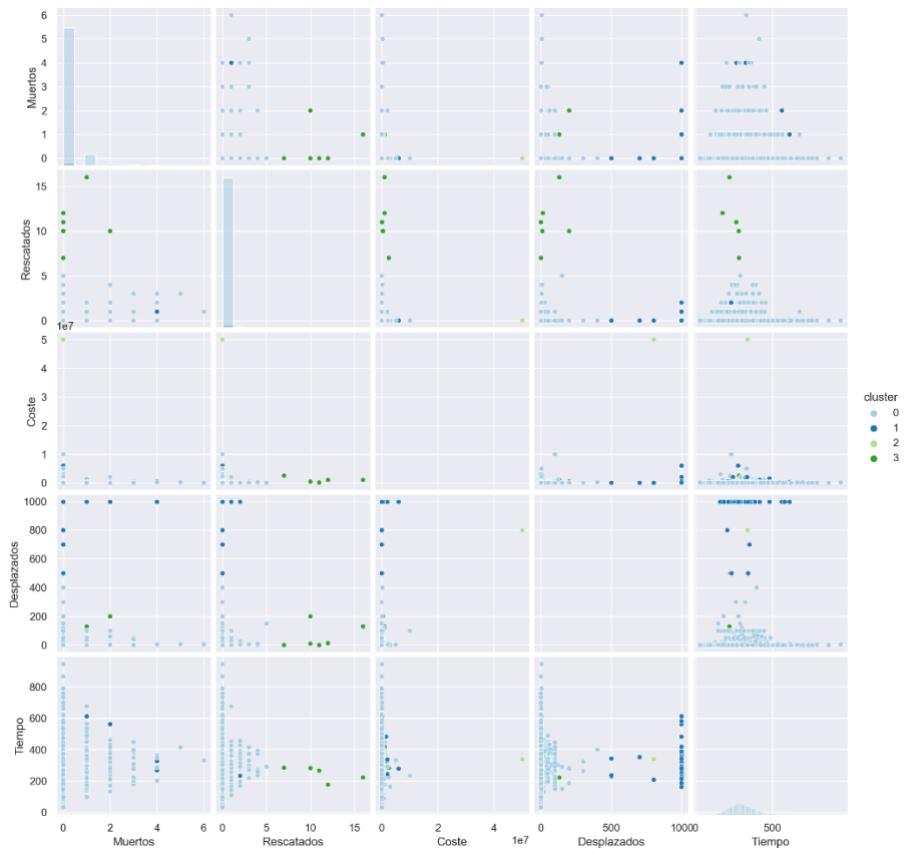


Figura 82: ScatterMatrix Birch Complementario Caso 3

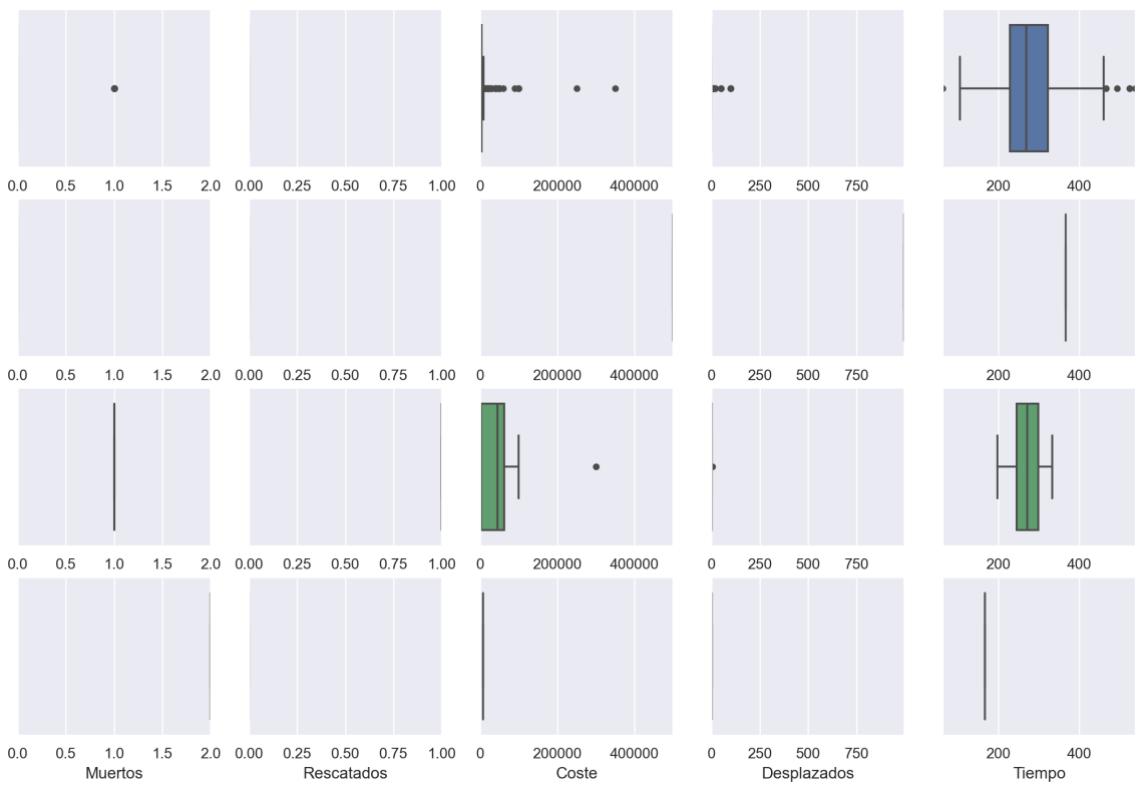


Figura 83: BoxPlot Birch Caso 3

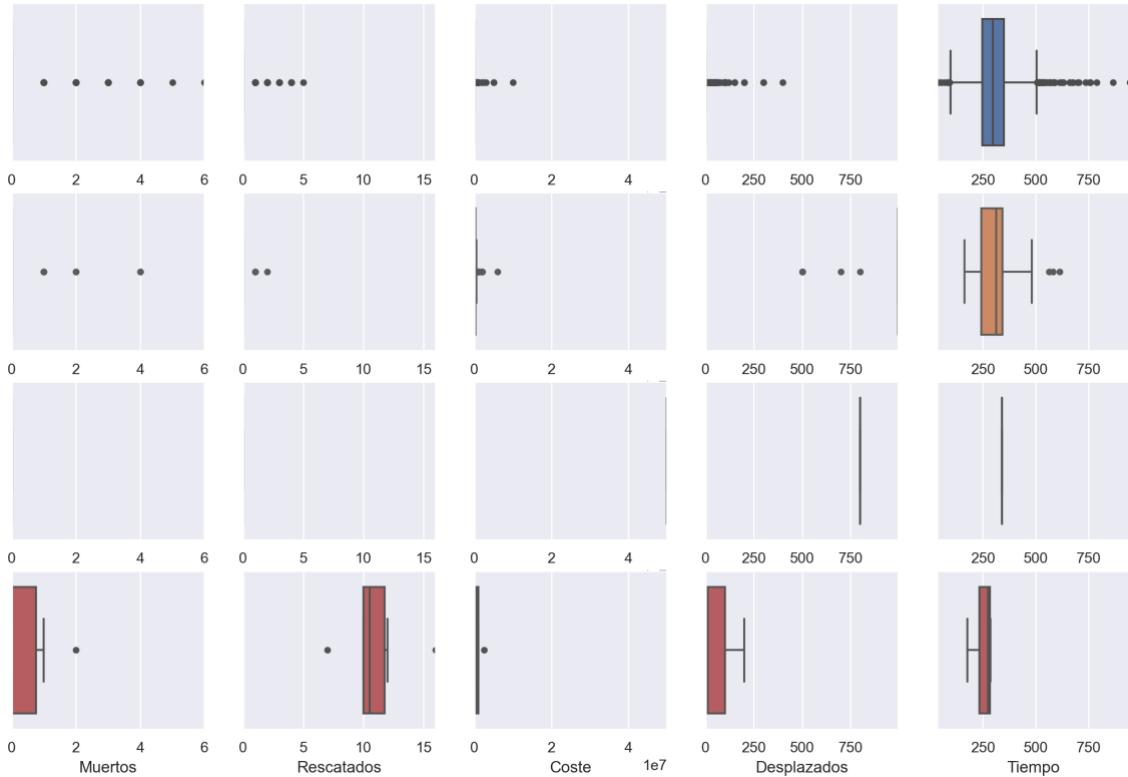


Figura 84: BoxPlot Birch Complementario Caso 3

5. Contenido adicional

6. Bibliografía

<https://open.toronto.ca/dataset/fire-incidents/>

<https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/fire-incidents/resource/fa5c7de5-10f8-41cf-883a-9b30a67c7b56/view/3259d54e-4dad-4c88-9abf-7cb75001dcda>

<https://www.kaggle.com/datasets/reihanenamdar/fire-incidents>

<https://scikit-learn.org/stable/modules/clustering.html>

https://pandas.pydata.org/docs/reference/general_functions.html

<https://pandas.pydata.org/docs/reference/frame.html>