Clustering

Curso 2019/2020

Sofía Almeida Bruno
Daniel Bolaños Martínez
José María Borrás Serrano
Fernando de la Hoz Moreno
Pedro Manuel Flores Crespo
María Victoria Granados Pozo

Índice

1. Introducción

El *clustering* consiste en agrupar objetos similares. Dos objetos se consideran similares si, considerando alguna medida de error, las observaciones podrían ser del mismo objeto. Esta clasificación ocurre constantemente en nuestra vida diaria, damos el mismo nombre a objetos que difieren en detalles insignificantes.

Si tenemos una serie de observaciones sin clasificar, el objetivo del *clustering* es agrupar los datos en clases o clústeres. Por ejemplo, en ámbitos biológicos para determinar la especie de una planta concreta. Este tipo de problema aparece cuando queremos no solo identificar especies nuevas, sino también cuando queremos establecer las relaciones entre ellas. El término *clustering* se considera sinónimo a taxonomía numérica o clasificación. En el ámbito de la ciencia de datos se considera problemas diferentes clasificación donde se conocen a priori las clases posibles (aprendizaje supervisado) y agrupamiento (aprendizaje no supervisado) donde el número de grupos es desconocido. En el primero conocemos de antemano las clases de los objetos, mientras que en el segundo, a partir de los grupos creados se inferirán las características principales de los grupos.

Utilizando el lenguaje matemático, podemos definir el problema como sigue:

> Dadas $x_1, \cdots, x_n$ medidas de p variables en n objetos considerados *heterogéneos*. El objetivo del análisis clúster es agrupar estos objetos en k clases *homogéneas*, donde k es también desconocido (aunque habitualmente se asume que es mucho menor que n).

Decimos que un grupo es *homogéneo* si sus miembros están cerca unos de otros pero los miembros de otros grupos son muy diferentes a estos. Esto lleva a definir dos métricas entre los puntos para indicar el grado de alejamiento y el de asociación o similitud. Se pueden tomar distintas distancias, creando aproximaciones diferentes al problema (trataremos este tema en más profundidad en la Sección 2).

En la Figura 1 vemos un ejemplo de agrupamiento para los datos dados. Cada punto representa un objeto x_i y los grupos encontrados se encuentran coloreados en diferentes colores.

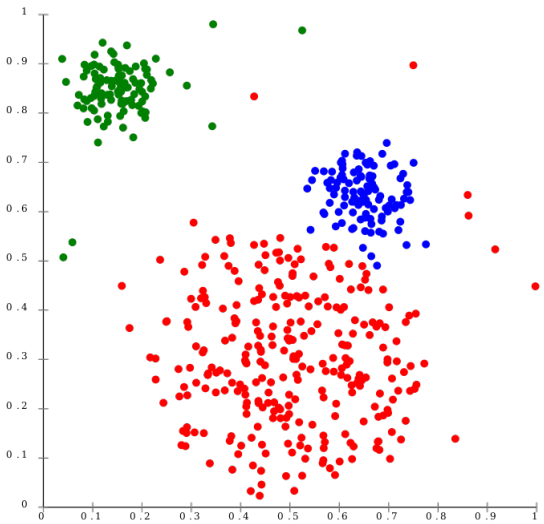

Figura 1: Ejemplo de *clustering*. [10]

El análisis clúster se aplica en numerosos campos como las ciencias naturales, médicas, económicas, *marketing*, ... En *marketing*, por ejemplo, es útil dividir a los clientes y conocer las necesidades de cada segmento de mercado para lograr alcanzar a los clientes potenciales. En psicología puede ser útil encontrar tipos de personalidad a partir de los cuestionarios realizados. En arqueología se puede aplicar esta técnica para clasificar objetos en diferentes periodos.

Para llevar a cabo un análisis clúster hay que realizar principalmente dos pasos:

1. Elegir una medida de similitud. Dependerá del objeto de estudio, tipo de medidas (nominales, ordinales, intervalos) y tipo de variables (continuas o discretas). Lo veremos en la Sección 2.
2. Elegir un algoritmo para construir los grupos. Hay dos grupos principales: de particionamiento y jerárquicos (que a su vez se dividen entre divisivos y aglomerativos). En la Sección 3 explicaremos la diferencia entre ambos grupos y veremos algunos ejemplos.

Además, en la Sección 4 trataremos el problema de decidir el número de clústeres y veremos algunos métodos utilizados para determinarlo. Por último, hemos implementado un ejemplo de agrupamiento tomando los datos de las flores de iris de Fisher, el análisis se encuentra en la Sección 5.

2. Medidas de similitud

La mayoría de los esfuerzos para producir una estructura de grupo más simple a partir de un conjunto complejo de datos requiere una medida de "cercanía" o "similitud". Habitualmente la subjetividad juega un papel importante en la elección de una medida de similitud, Algunas consideraciones importantes incluyen la naturaleza de las variables (discreta, continua, binaria), las escalas de las medidas (nominal, ordinal, intervalo) y conocimiento específico sobre el problema. En cualquier caso, los valores de las variables consideradas serán normalizados, para evitar que unas variables tengan más peso que otras a la hora de realizar el agrupamiento.

Cuando los ítems (unidades o casos) son agrupados, la proximidad se suele indicar mediante alguna medida de la distancia. En contraste, las variables se suelen agrupar según los coeficientes de correlación o medidas de asociación.

2.1. Distancias y coeficientes de similitud para parejas de ítems

Recordamos que la distancia euclídea entre dos observaciones p-dimensionales (ítems) $x' = [x_1, x_2, ..., x_p]$ e $y' = [y_1, y_2, ..., y_p]$ es

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} = \sqrt{(x-y)^T(x-y)}.$$

La distancia estadística entre las mismas dos observaciones es de la forma:

$$d(x,y) = \sqrt{(x-y)^T A(x-y)}.$$

Normalmente, $A = S^{-1}$, donde S contiene las varianzas y covarianzas de la muestra. Sin embargo, sin conocimiento previo de los distintos grupos, estas cantidades de las muestras no puede ser calculadas. Por esta razón, usualmente se prefiere la distancia euclídea para realizar *clustering*.

Otra medida de la distancia es la métrica de Minkowski, que viene dada por:

$$d(x,y) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}.$$

Para $m = 1$, $d(x,y)$ mide la distancia Manhattan entre dos puntos en p dimensiones. Para $m = 2$,$d(x,y)$ se convierta en la distancia euclídea. En general, variar m cambia el peso dado a diferencias mayores y menores.

Adicionalmente, dos medidas populares de distancia o "disimilitud" son las dadas por la métrica de Canberra y el coeficiente de Czekanowski. Ambas medidas están definidas únicamente para variables no negativas.

La métrica de Canberra se define como:

$$d(x,y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}.$$

El coeficiente de Czekanowski es:

$$d(x,y) = 1 - \frac{2\sum_{i=1}^{p} min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i}.$$

Es aconsejable que cuando sea posible se utilicen distancias "verdaderas" para *clustering*, esto es, distancias que verifiquen las propiedades de la definición matemática de distancia:

- $d(P,Q) = d(Q,P)$,
- $d(P,Q) > 0 \; si \; P \neq Q$,
- $d(P,Q) = 0 \; si \; P = Q$,
- $d(P,Q) \leq d(P,R) + d(R,Q)$,

siendo P y Q dos puntos cualquiera y R un punto intermedio.

Sin embargo, la mayoría de los algoritmos de *clustering* aceptarán números de distancia que no satisfagan la desigualdad triangular.

Cuando los ítems no puedan ser representados por medidas p-dimensionales significativas, las parejas de ítems se suelen comparar según la presencia o ausencia de ciertas características. Ítems similares tienen más características en común que ítems desemejantes. La presencia o ausencia de una característica se puede describir matemáticamente introduciendo una variable binaria, que toma el valor 1 si la característica está presente y el valor 0 cuando la característica está ausente.

Tabla 1: Ejemplo 1, binarización de variables.

Variables	1	2	3	4	5
Ítem i	1	0	0	1	1
Ítem k	1	1	0	1	0

En la Tabla 1 vemos un ejemplo (**Ejemplo 1**) en el que se utilizan 5 variables binarias para medir la distancia entre los ítem i y j. En este caso, hay dos parejas 1-1, una pareja 0-0 y dos parejas que no

coinciden.

Sea x_{ij} la puntuación (0 ó 1) de la j-ésima variable binaria en el i-ésimo ítem y sea x_{kj} la puntuación de la j-ésima variable binaria en el k-ésimo ítem, con $j = 1, 2, ..., p$. Entonces,

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{si } x_{ij} = x_{kj} \\ 3n + 1 & \text{si } x_{ij} \neq x_{kj} \end{cases}$$

y la distancia euclídea al cuadrado, $\sum_{j=1}^{p} (x_{ij} - x_{kj})^2$, proporcionan una forma de contar el número de disparidades. Una distancia grande corresponde a muchas disparidades, es decir, ítems desemejantes. En el Ejemplo 1, el cuadrado de las distancia entre los ítems i y k sería

$$\sum_{j=1}^{5} (x_{ij} - x_{kj})^2 = (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 = 2.$$

Aunque la distancia usada en el Ejemplo 1 puede ser usada para medir la similitud, tiene la desventaja de que valora por igual las parejas 1-1 y 0-0. En algunos casos, una pareja 1-1 es una indicación mayor de similitud que una pareja 0-0. Por ejemplo, al agrupar gente, que dos personas sean capaz de leer griego antiguo es una evidencia mayor de similitud que la ausencia de dicha habilidad. Para reflejar este trato diferente entre las parejas de 1-1 y 0-0, se han sugerido diversos esquemas para definir los coeficientes de similitud.
Para introducir dichos esquemas, vamos a organizar las frecuencias de las parejas que coinciden y las que no para los ítems i y k en la forma de una tabla de contingencia:

Tabla 2: Ejemplo 1, tabla de contingencia.

		Ítem k		
		1	0	Total
Ítem i	1	a	b	ab
	0	c	d	$c+d$
Total		$a+c$	$b+d$	$p = a+b+c+d$

En la Tabla 2, a representa la frecuencia de parejas 1-1, b es la frecuencia de parejas 1-0 y así sucesivamente. Dadas las anteriores cinco parejas, $a = 2$ y $b = c = d = 1$.

La Tabla 3 lista coeficientes de similitud comunes definidos en términos de frecuencias de la Tabla 2.

Tabla 3: Coeficientes de similitud para ítems *clustering*.

Coeficiente		Fundamento
1	$\frac{a+d}{p}$	Las parejas 1-1 y 0-0 ponderan lo mismo.
2	$\frac{2(a+d)}{2(a+d)+b+c}$	Las parejas 1-1 y 0-0 ponderan el doble.
3	$\frac{a+d}{a+d+2(b+c)}$	Las parejas que no coinciden ponderan el doble.
4	$\frac{a}{p}$	No hay parejas 0-0 en el numerador.
5	$\frac{a}{a+b+c}$	No hay parejas 0-0 en el numerador ni el denominador (Las parejas 0-0 son irrelevantes).
6	$\frac{2a}{2a+b+c}$	No hay parejas 0-0 en el numerador ni el denominador. Las parejas 1-1 ponderan el doble.
7	$\frac{a}{a+2(b+c)}$	No hay parejas 0-0 en el numerador ni el denominador. Las parejas que no coinciden ponderan el doble.
8	$\frac{a}{b+c}$	Proporción de parejas que coinciden (excluyendo las 0-0) en relación a las parejas que no coinciden.

Los coeficientes 1, 2 y 3 de la Tabla 3 están monotónicamente relacionados. Supongamos que el coeficiente 1 se calcula para dos tablas de contingencia: Tabla I y Tabla II. Entonces, si $(a_I + d_I)/p \geq (a_{II} + d_{II})/p$, tenemos que $2(a_I + d_I)/[2(a_I + d_I) + b_I + c_I] \geq 2(a_{II} + d_{II})/[2(a_{II} + d_{II}) + b_{II} + c_{II}]$ y el coeficiente 3 será al menos tan grande para la Tabla I como lo es para la Tabla II. Los coeficientes 5, 6 y 7 también mantienen sus órdenes relativos.

La monotonicidad es importante, ya que algunos procedimientos de *clustering* no se ven afectados si la definición de similitud se modifica de manera que los órdenes relativos de similitud no cambian. Los procedimientos jerárquicos de asociación simple y asociación completa que se verán en la Sección 3 no se ven afectados Para estos métodos, cualquier elección entre los coeficientes 1, 2 y 3 de la Tabla 3 producirá las mismas agrupaciones. Análogamente, cualquier elección entre los coefcientes 5, 6 y 7 también resultará en agrupaciones idénticas.

Veamos un ejemplo (**Ejemplo 2**) de cálculo de coeficientes de similitud. Supongamos cinco individuos con las características mostradas en la Tabla 4.

Tabla 4: Individuos y características del Ejemplo 2.

	Altura (in)	Peso (lb)	Color de ojos	Color de pelo	Mano predominante	Género
Individuo 1	68	140	Verde	Rubio	Derecha	Femenino
Individuo 2	73	185	Marrón	Moreno	Derecha	Masculino
Individuo 3	67	165	Azul	Rubio	Derecha	Masculino
Individuo 4	64	120	Marrón	Moreno	Derecha	Femenino
Individuo 5	76	210	Marrón	Moreno	Izquierda	Masculino

Definimos seis variables binarias $X_1, X_2, X_3, X_4, X_5, X_6$ como:

$$X_1 = \begin{cases} 1 & \text{si Altura} \geq 72 \\ 0 & \text{si Altura} < 72 \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{si Peso} \geq 150 \\ 0 & \text{si Peso} < 150 \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{si Ojos marrones} \\ 0 & \text{si otro color de ojos} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{si Pelo rubio} \\ 0 & \text{si otro color de pelo} \end{cases}$$

$$X_5 = \begin{cases} 1 & \text{si Diestro} \\ 0 & \text{si Zurdo} \end{cases}$$

$$X_6 = \begin{cases} 1 & \text{si Masculino} \\ 0 & \text{si Femenino} \end{cases}.$$

Las puntuaciones de los individuos 1 y 2 en las $p = 6$ variables binarias son las encontradas en la Tabla 5.

Tabla 5: Puntuaciones individuos 1 y 2.

	X_1	X_2	X_3	X_4	X_5	X_6
Individuo 1	0	0	0	1	1	1
Individuo 2	1	1	1	0	1	0

El número de parejas que coinciden y las que no se indican en la Tabla 6.

Tabla 6: Tabla de contingencias para los individuos 1 y 2 del Ejemplo 2.

		Individuo 2		
		1	0	Total
Individuo 1	1	1	2	3
	0	3	0	3
Total		4	2	6

Empleando el coeficiente de similitud 1, que pondera por igual las parejas que coinciden, calculamos

$$\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}.$$

Continuando con el coeficiente de similitud 1, calculamos los demás números de similitud para cada pareja de individuos. Mostramos los resultados en la matriz simétrica 5 x 5.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{ccccc} 1 & & & & \\ 1/6 & 1 & & & \\ 4/6 & 3/6 & 1 & & \\ 4/6 & 3/6 & 2/6 & 1 & \\ 0 & 5/6 & 2/6 & 2/6 & 1 \end{array}\right] \end{array}$$

$$\begin{array}{cc} & \text{Individuo} \\ & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \left[\begin{array}{ccccc} 1 & & & & \\ 1/6 & 1 & & & \\ 4/6 & 3/6 & 1 & & \\ 4/6 & 3/6 & 2/6 & 1 & \\ 0 & 5/6 & 2/6 & 2/6 & 1 \end{array}\right] \end{array}$$

Individuo

Basándonos en las magnitudes del coeficiente de similitud, debemos concluir que los individuos 2 y 5 son los más similares mientras que los individuos 1 y 5 son los menos similiares. Si fuéramos a dividir los individuos en dos subgrupos relativamente homogéneos basándonos en los números de similitud, formaríamos los subgrupos (1 3 4) y (2 5).
Notar que $X_3 = 0$ implica una ausencia de ojos marrones, así dos personas, una con ojos azules y otra con ojos verdes, resultarán en una pareja 0-0. Por consiguiente, puede que sea inapropiado utilizar los coeficientes de similitud 1, 2 ó 3 porque estos coeficientes ponderan lo mismo las parejas 1-1 y 0-0.

Hemos descrito la construcción de distancias y similitudes. Siempre se pueden construir similitudes a partir de distancias. Por ejemplo, podemos fijar

$$s_{ik} = \frac{1}{1 + d_{ik}},$$

donde $0 < s_{ik} \leq 1$ es la similitud entre los ítems i y k y d_{ik} es la distancia correspondiente.
Sin embargo, no siempre podemos construir distancias "verdaderas", que cumplan las propiedades para *clustering*, a partir de similitudes. Sólo se pueden construir si la matriz de similitudes es definida no negativa.
Si la matriz de similitudes fuera definida no negativa y la máxima similitud es cada de manera que $s_{ii} = 1$,

$$d_{ik} = \sqrt{2(1 - s_{ik})}$$

cumple las propiedades de una distancia.

2.2. Medidas de similitud y asociación para parejas de variables

Hasta el momento, hemos discutido medidas de similitud para los ítems. En algunas aplicaciones, son las variables, en lugar de los ítems, las que deben ser agrupadas. Las medidas de similitud para variables suelen tomar la forma de coeficientes de correlaciones muestrales. Además, en algunas aplicaciones de *clustering*, las correlaciones negativas son reemlazadas por sus valores absolutos.
Cuando las variables son binarias, los datos se pueden organizar en una tabla de contingencia. Sin embargo, esta vez las variables, en lugar de los ítems, definen las categorías. Para cada par de variables

hay n ítems categorizados en la tabla. Con la codificación usual en 0 y 1, la tabla de contingencia es como se ve en la Tabla 7. Por ejemplo, la variable i vale 1 para a ítems y la variable k vale 0 para b de los n ítems.

Tabla 7: Tabla de contingencia para un par de variables.

		Variable k		Total
		1	0	
Variable i	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	$n=a+b+c+d$

La fórmula usual del coeficiente de correlación producto-momento aplicada a las variables binarias de la tabla de contingencia nos da

$$r = \frac{ad - bc}{[(a+b)(c+d)(a+c)(b+d)]^{1/2}}.$$

Este número se puede tomar como la medida de similitud entre las dos variables.

El coeficiente de correlación anterior se relaciona con la estadística chi-cuadrado ($r^2 = \chi^2/n$) para evaluar la independencia de dos variables categóricas. Para un n fijo, una similitud (o correlación) grande es consistente con la presencia de dependencia.

Dada la Tabla 7, se pueden desarrollar medidas de asociacion (o similitud) exactamente análogas a las listadas en la Tabla 3. El único cambio que se requiere es la sustitución de n (el número de ítems) por p (el número de variables).

Para resumir esta sección, notamos que hay muchas medidas de similitud entre pares de objetos. Parece que la mayoría de profesionales usan distancias o los coeficientes de la Tabla 3 para el clúster de ítems y correlaciones para el clúster de variables. Sin embargo, en ocasiones, las entradas a los algoritmos de *clustering* pueden ser frecuencias simples.

Ejemplo 3: Midiendo las similitudes de 11 lenguajes. Los significados de las palabras cambian a lo largo de la historia. Sin embargo, el significado de los números 1,2,3,... representa una excepción llamativa. Así, una primera comparación de lenguajes puede estar basada en exclusiva en los números naturales. La Tabla 8 nos da los 10 primeros números en inglés, polaco, húngaro y otros ocho lenguajes europeos modernos. Solamente consideramos lenguajes que utilizan el alfabeto romano y omitimos acentos y signos de puntuación.

Una observación rápida de la escritura de los números sugiere que los cinco primeros lenguajes (inglés, noruego, danés, holandés y alemán) son muy parecidos. El español, francés e italiano se parecen todavía más. El húngaro y finés parecen no estar relacionados con los demás lenguajes. El polaco tiene algunas características de los lenguajes de los subgrupos más grandes.

Por porpósitos ilustrativos, comparamos los lenguajes mirando la primera letra de cada número. Las palabras para el mismo número son concordantes si tienen la misma primera letra y discordantes en caso contrario. De esta forma de la Tabla 8 obtenemos la Tabla 9. Podemos observar que el inglés y el noruego comparten la misma primera letra para 8 de las 10 parejas de números. El resto de frecuencias se ha calculado de la misma forma.
Los resultados en la Tabla 9 confirman la impresión visual inicial de la Tabla 8. Esto es, que el inglés, noruego, danés, holandés y alemán parecen formar un grupo. El francés, español, italiano y polaco

Tabla 8: Ejemplo 3. Números en 11 lenguajes.

Inglés (E)	Noruego (N)	Danés (Da)	Holandés (Du)	Alemán (G)	Francés (Fr)	Español (Sp)	Italiano (I)	Polaco (P)	Húngaro (H)	Finés (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	nelja
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

Tabla 9: Ejemplo 3. Concordancia de la primera letra para 10 números en 11 lenguajes.

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
Fr	4	4	4	1	3	10					
Sp	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
Fi	1	1	1	1	1	1	1	1	1	2	10

forman otro, mientras que el húngaro y el finés no forman parte de ninguno.

En los ejemplo hasta el momento, hemos utilizado nuestra impresión visual de las medidas de similitud o distancia para formar grupos. En la Sección 3 discutiremos formas menos subjetivas para crear los clústeres, distintos algoritmos para crear los grupos.

3. Métodos de agrupamiento

Un método de agrupamiento es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio. Estos criterios pueden ser, por ejemplo, la distancia o la similitud de una característica. La cercanía se define en términos de una determinada función de distancia como las vista en la Sección 2. Aunque la medida más utilizada para calcular la simulitud entre los casos es la matriz de correlación entre los casos, también existen muchos algoritmos que se basan en la maximización de la verosimilitud.

Los métodos de agrupamiento se pueden dividir en dos tipos: jerárquicos y no jerárquicos (o de particionamiento). La principal diferencia entre ambos métodos es que en los jerárquicos una vez que se asigna un elemento a un grupo, no se puede cambiar, mientras que en los no jerárquicos sí. Además, en los métodos no jerárquicos se necesita que el número de clústers esté fijado a priori, en cambio en el agrupamiento jerárquico el propio método va fijando el número de clústers. En la Figura 2 observamos a la izquierda la representación de un método jerárquico (las líneas horizontales indican diferentes divisiones) y a la derecha la representación del resultado de un método de particionamiento, donde cada

grupo está indicado en un color diferente).

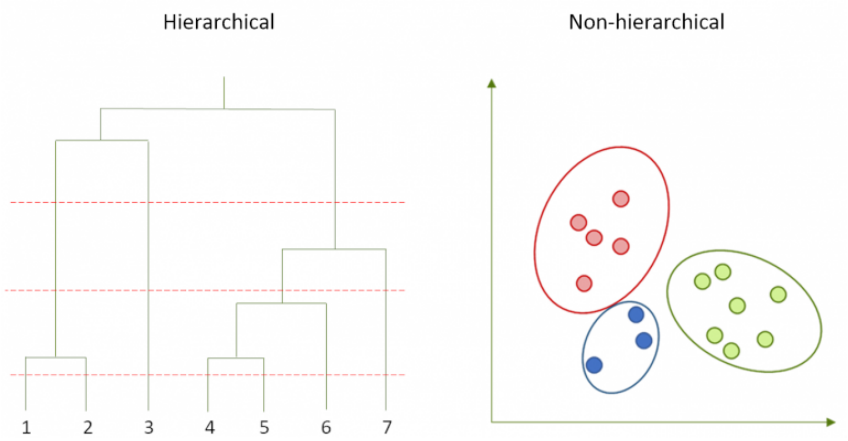

Figura 2: Comparación métodos jerárquico y no jerárquico.

3.1. Jerárquicos

El agrupamiento jerárquico es un método de análisis de grupos puntuales, que se basa en buscar una construcción de una jerarquía de grupos. Por tanto, los algoritmos jerárquicos van minimizando las distancias o, visto de otra forma, maximizando las similitudes. Es destacable que en este tipo de métodos no es necesario imponer a priori el número de grupos a formar, ya que devolverán una jerarquía de clústeres en la que según el nivel en el que nos quedemos tendremos un número u otro de grupos.

Podemos diferenciar dos técnicas para el agrupamiento jerárquico: aglomerativas y divisivas. La diferencia es que las primeras son de un acercamiento ascendente, esto es cada observación comienza en su propio grupo, y los pares de grupos se mezclan mientras uno subo en la jerarquía, mientras que las técnicas divisivas son de un acercamiento descendente, es decir, se van haciendo divisiones conforme uno baja en la jerarquía. En la Figura 3 se simbolizan ambas técnicas.

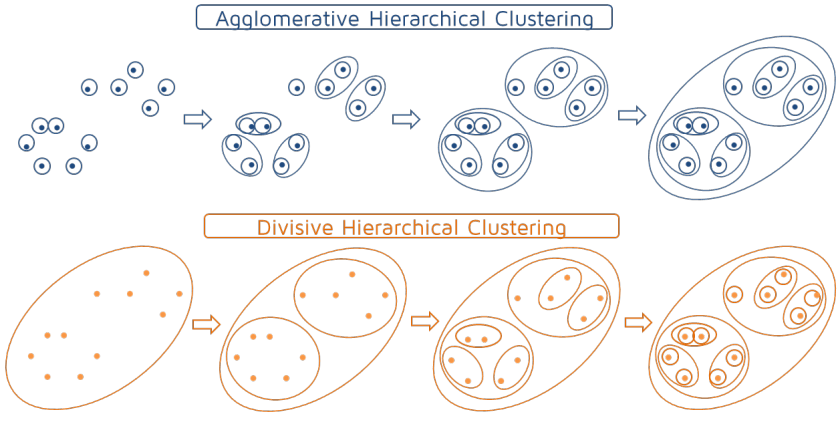

Figura 3: Comparación métodos aglomerativos y divisivos.

En general, las mezclas y divisiones son determinadas con un Algoritmo Greedy. Los resultados del agrupamiento jerárquico son usualmente presentados en un dendrograma. Un dendrograma es un diagrama de árbol que muestra los grupos que se forman al crear clústers de observación en cada paso y sus niveles de similitud.

3.1.1. Técnicas aglomerativas

Los algoritmos aglomerativos son prácticamente los más usados. Consisten en formar grupos o aglomerados según su similitud, uniendo los clúster que se encuentran a una menor distancia, habrá tantos clústeres al inicio como individuos haya. Estos algoritmos siguen los pasos que se muestran a continuación:

1. Partimos con n clústers, donde cada uno contiene un solo objeto.
2. Calcular la matriz distancia D para cada par de clúster. A cada par elegido de objetos r y s se le asocia el elemento d_{rs}.
3. Combinar r y s en un nuevo clúster (rs), reduciendo así el número de clúster y eliminando una fila y una columna para los objetos r y s de la matriz D. Ahora se calculan las disimilitudes o distancias entre (rs) y el resto de clústers, añadiendo filas y columnas a la nueva matriz D.
4. Repetir los pasos 2 y 3, $(n-1)$ veces hastas que todos los objetos estén en un único clúster.

Sean P y Q dos grupos que vamos a unir, generando un nuevo grupo $P + Q$. Ahora calculamos la distancia entre $P + Q$ y R utilizando la siguiente función (1):

$$d(R, P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 d(R,P) - d(R,Q) \tag{1}$$

Donde los δ_i son los factores de ponderación que dan lugar a los diferentes algoritmos de aglomeración, algunos se describen en la Tabla 10. Definimos n_P como el número de elementos que hay en el grupo P, de la misma forma se definen n_Q y n_R.

Tabla 10: Distancias entre grupos.

Métodos	δ_1	δ_2	δ_3	δ_4
Single linkage	1/2	1/2	0	-1/2
Complete linkage	1/2	1/2	0	1/2
Average linkage (unweighted)	1/2	1/2	0	0
Average linkage (weighted)	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	0	0
Centroid	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_Q n_P}{n_P+{n_Q}^2}$	0
Median	1/2	1/2	-1/4	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_R+n_P+n_Q}$	0

La secuencia de clúster se representan gráficamente mediante un dendrograma, esto es un digrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros, donde se ven claramente las relaciones entre los datos y los grupos.

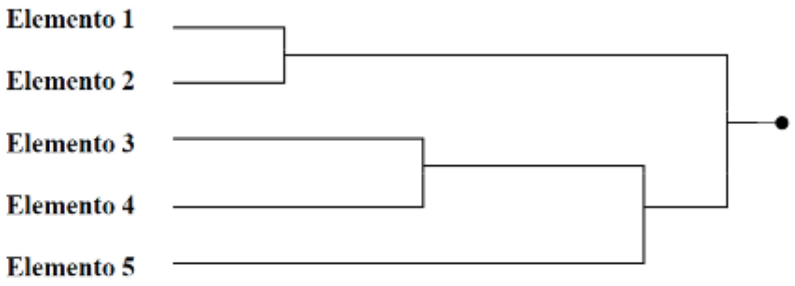

Figura 4: Ejemplo de dendrograma aglomerativo.

A continuación se muestran algunos métodos aglomerativos

3.1.1.1 Single Link Method

Para entender mejor este método se presentará un ejemplo usando *Single Link Method*. Este método es una combinación de los objetos de los clústers teniendo en cuenta las disimilitudes entre los clústers. Para realizar este ejemplo introducimos algunos términos de notación, r representa a cualquier elemento en el clúster R, $r \in R$, y s será cualquier elemento del clúster S, $s \in S$. Calculamos la distancia entre R y S mediante la fórmula 2

$$d(R,S) = min\{d_{rs} \ : \ r \in R, \ s \in S\} \tag{2}$$

Consideremos la matriz de distancias D

$$\mathcal{D} = \left[\ d_{rs} \ \right] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 2 & 4 & 7 & 9 \\ 2 & 0 & 8 & 9 & 8 \\ 4 & 8 & 0 & 3 & 7 \\ 7 & 9 & 3 & 0 & 5 \\ 9 & 8 & 7 & 5 & 0 \end{bmatrix} \end{array}$$

Si observamos la matriz D encontramos que la distancia más pequeña es 2, el elemento d_{12} esto nos quiere decir que los elementos 1 y 2 son los más cercanos o los más similares. siguiendo el algoritmo los agrupamos en un mismo clúster (12). Ahora calculamos las nuevas distancias entre el nuevo clúster y el resto de elementos.

$$d_{(12)(3)} = \min\{d_{13}, d_{23}\} = \min\{4, 8\} = 4 \; d_{(12)(4)} = \min\{d_{14}, d_{24}\} = \min\{7, 9\} = 7 \; d_{(12)(5)} = \min\{d_{15}, d_{25}\} = \min\{9, 8\} = 8$$

Borramos los elementos 1 y 2, y añadimos las nuevas filas y columnas correspodientes al nuevo clúster (12).

$$\mathcal{D}_1 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 4 & 7 & 8 \\ 4 & 0 & 3 & 7 \\ 7 & 3 & 0 & 5 \\ 8 & 7 & 5 & 0 \end{bmatrix} \end{array}$$

Volvemos a observar la matriz fijandonos en el elemento d_{34}, así los elementos que están más cerca son 3 y 4. Formamos el nuevo clúster (34). Calculamos los nuevo valores de la matriz.

$$d_{(34)(12)} = \min\{d_{(3)(12)}, d_{(4)(12)}\} = \min\{4, 7\} = 4 \; d_{(34)(5)} = \min\{d_{(3)(5)}, d_{(4)(5)}\} = \min\{7, 5\} = 5$$

La matriz quedaría

$$\mathcal{D}_2 = \begin{array}{c} \\ (12) \\ (34) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} (12) & (34) & 5 \end{array} \\ \begin{bmatrix} 0 & 4 & 8 \\ 4 & 0 & 5 \\ 8 & 5 & 0 \end{bmatrix} \end{array}$$

El valor más similar en D_2 es 4 que corresponde con los elementos (34) y (12). Calculamos la última distancia

$$d_{(12)(34)5} = \min\{d_{(12)(5)}, d_{(34)(5)}\} = \min\{8, 5\} = 5$$

Finalemente juntamos el elemento 5 con los clústers (12) y (34) formando el clúster (12345).

Así el dendograma quedaría de la siguiente forma 5

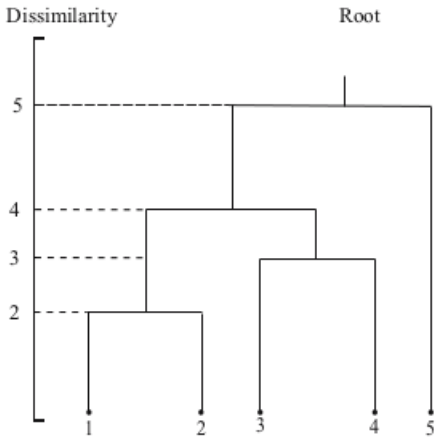

Figura 5: Dendrograma ejemplo *Single Link*.

3.1.1.2 DBSCAN

Al contrario de la estrategia seguida por K-Means, DBSCAN (*Density-Based Spatial clustering of Applications with Noise*) no presupone clústeres convexos, sino que se basa en la densidad de las muestras para identificar los clústeres. Por este motivo, los clústeres identificados por DBSCAN pueden ser de cualquier forma.

Dos parámetros importantes para definir este algoritmo son: **eps**, máxima distancia entre dos muestras para poder ser consideradas pertenecientes al mismo "vecindario", y **min_samples**, número de muestras en un vecindario para que una región pueda ser considerada densa.

Consideremos un conjunto de puntos a ser agrupados en un espacio determinado. La técnica de agrupación DBSCAN clasifica los puntos como puntos núcleo, puntos (densamente-)alcanzables, o ruido de la siguiente forma:

- Un punto p pertenece al núcleo si al menos *min_samples* puntos están a una distancia ϵ de él y esos puntos son directamente alcanzables desde p. No es posible tener puntos directamente alcanzables desde un punto distinto al núcleo.

- Un punto q es alcanzable desde p si existe una secuencia de puntos $p_1 \dots p_n$ donde $p_1 = p$ y $p_n = q$ y cada punto p_{i+1} es directamente alcanzable desde p_i.
- Un punto que no sea alcanzable desde cualquier otro se considera ruido.

Si p es un punto núcleo, este forma un clúster junto a otros puntos que sean alcanzables desde él. Cada clúster contiene al menos un punto núcleo. Los puntos no núcleos alcanzables pueden pertenecer a un clúster pero actúan como una barrera puesto que no es posible alcanzar más puntos desde estos.

Se puede observar que la relación de ser alcanzable no es simétrica. Por definición, ningún punto puede ser alcanzable desde un punto que no sea núcleo, sin importar la distancia a la que se encuentre. Por lo tanto la noción de connectividad es necesaria para definir formalmente la extensión de un clúster dada por DBSCAN. Dos puntos p y q están conectados densamente si existe un punto o tal que ambos p y q sean directamente alcanzables desde o. La relación estar densamente conectado es simétrica.

Por tanto un clúster generado por el método DBSCAN satisface dos propiedades:

- Todos los puntos de un mismo clúster están densamente conectados entre sí.
- Si un punto A es densamente alcanzable desde cualquier otro punto B del clúster, entonces A también forma parte del clúster.

3.1.1.3 Mean Shift

Mean Shift es una técnica de análisis de espacio de características no paramétrica para localizar los máximos de una función de densidad. Es un método iterativo que parte de una estimación inicial x. Dada una función núcleo $K(x_i - x)$. Esta función determina el peso de los puntos cercanos para la reestimación de la media. Normalmente se usa un núcleo gaussiano en la distancia de la estimación actual, $K(x_i - x) = e^{-c||x_i - x||^2}$. La media ponderada de la densidad en la ventana determinada por K es:

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

donde $N(x)$ es el vecindario de x, un conjunto de puntos donde $K(x_i) \neq 0$.

La diferencia $m(x) - x$ se denomina Mean Shift. El algoritmo establece $m(x) \rightarrow x$ y repite la estimación hasta que $m(x)$ converja.

Todavía no se conoce una prueba rígida de la convergencia del algoritmo utilizando un núcleo general en un espacio de alta dimensión. Aliyari Ghassabeh mostró la convergencia del algoritmo de cambio medio en una dimensión con una función de perfil diferenciable, convexa y estrictamente decreciente. Sin embargo, el caso unidimensional tiene aplicaciones limitadas en el mundo real.

Además, se ha demostrado la convergencia del algoritmo en dimensiones superiores con un número finito de los puntos estacionarios (o aislados). Sin embargo, no se han proporcionado condiciones suficientes para que una función general del núcleo tenga puntos estacionarios finitos (o aislados).

Sea un conjunto de datos finito S embebido en el espacio euclídeo n-dimensional X. Sea K un núcleo plano que tiene como función característica en X:

$$K(x) = \begin{cases} 1 \; si \; ||x|| \leq \lambda \\ 0 \; si \; ||x|| > \lambda \end{cases}$$

En cada iteración del algoritmo, se establece $m(x) \rightarrow x$ para todo $s \in S$ simultáneamente. Una de las preguntas que nos podemos hacer es, cómo estimar la función de densidad dado un conjunto escaso de muestras. Uno de los enfoques más simples es simplemente suavizar los datos, por ejemplo, convolucionándolos con un núcleo fijo de ancho h:

$$f(x) = \sum_i K(x - x_i) = \sum_i k \frac{||x - x_i||^2}{h^2}$$

donde x_i son las muestras de entrada y k la función núcleo y h es el único parámetro que toma el algoritmo que se denomina *bandwidth*. Una vez que hemos calculado $f(x)$ a partir de la ecuación anterior, podemos encontrar sus máximos locales utilizando el ascenso de gradiente o alguna otra técnica de optimización.

Usando esta aproximación por fuerza bruta hace que el problema sea computacionalmente inviable conforme aumentamos las dimensiones del problema sobre el espacio de búsqueda total. Por tanto, Mean Shift utiliza la técnica del reinicio de gradiente descendiente múltiple, la cual empieza desde un máximo local y_k y calcula su aproximación $f(x)$ y avanza en esa dirección.

3.1.2. Técnicas divisivas

Los algoritmos divisivos parten de un único clúster y en cada iteración van dividiendo clústeres según la disimilitud entre los componentes del mismo, bajando así en la jerarquía.

En cada paso del algoritmo divisivo se deben de aplicar tres opciones:

- Dividir los grupos de una forma simplificada.
- Dar una fórmula para evaluar cada una de las biparticiones consideradas.
- Dar una fórmula para determinar los niveles de nodo de la jerarquía resultante.

Como podemos ver en la figura 6 es similar al dendrograma de la figura 4, donde la diferencia es que en lugar de mirarlo de izquierda a derecha, lo miramos de derecha a izquierda.

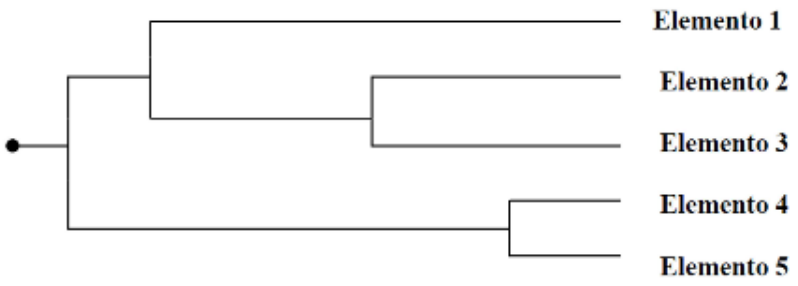

Figura 6: Ejemplo de dendrograma divisivo.

3.2. No jerárquicos

Los métodos de *clustering* no jerárquicos, también denominados de particionamiento, simplemente dividen los datos observados en un número K predeterminado de grupos o clústeres, donde no hay una

relación jerárquica entre la solución de K-clústeres y la solución de $(K+1)$-clústeres. Dado un valor K buscamos una partición de los datos en K clústeres de tal manera que los datos dentro de cada clúster sean similares entre ellos, mientras que los datos de diferentes clústeres están bastante diferenciados.

Un método de *clustering* no jerárquico probablemente involucraría como primer paso la enumeración total de todos los posibles agrupamientos de los datos. Entonces, usando algún criterio de optimización, el agrupamiento que es elegido como el mejor sería la partición que optimizara dicho criterio. Claramente, para un conjunto de datos grande, tal método se vuelve rápidamente inviable, requiriendo cantidades ingentes de tiempo de procesamiento y almacenamiento. Por ello, todas las técnicas de *clustering* disponibles son iterativas y trabajan con una cantidad de agrupamientos muy limitada. Los métodos no jerárquicos suelen ser computacionalmente más eficientes que los jerárquicos.

3.2.1. K-Medias

El algoritmo K-medias es muy popular porque es extremadamente eficiente y a menudo es usado para proyectos de *clustering* de gran escala. Hacemos la observación de que el algoritmo K-medias necesita tener acceso a los datos originales.

El algoritmo K-medias comienza o por asignación de los datos a uno de los K clústeres predeterminados y luego calculando los centroides de los K *cluster* o por pre-especificación de los centroides de los K clústeres. La pre-especificación de los centroides puede ser seleccionarlos aleatoriamente o pueden ser obtenidos cortando un dendrograma a una altura apropiada. Luego, de manera iterativa, el algoritmo busca minimizar la ESS (suma de las distancias euclídeas al cuadrado, por sus siglas en inglés *Euclidean Sum of Squares*) por reasignación de los datos a los clústeres. El proceso se detiene cuando no hay más reasignaciones que reduzcan el valor de la suma.

La solución (una configuración de los datos dentro de los K clústeres) puede no ser única. El algoritmo solo encontrará un mínimo local de la ESS. Es recomendable ejecutar el algoritmo varias veces usando una asignación aleatoria inicial diferente de los datos en los K clústeres para encontrar el mínimo más bajo de la ESS, y por tanto, la mejor solución basada en K clústeres. Los pasos del algoritmo son:

1. Entrada: $\mathbf{L} = \{\mathbf{x}_i,\ i = 1, 2, ..., n\}$, K: número de clústeres.
2. Hacer uno de los siguientes:
   - Formar una asignación aleatoria inicial de los datos en los K clústeres y, para los K clústeres, calcular su centroide, $\overline{\mathbf{x}}_k,\ k = 1, 2, ..., K$.
   - Pre-especificar los centroides de los K clústeres, $\overline{\mathbf{x}}_k,\ k = 1, 2, ..., K$.
3. Calcular la distancia euclídea al cuadrado para cada dato al centroide de su clúster actual:

$$ESS = \sum_{k=1}^{K} \sum_{c(i)=k} (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T (\mathbf{x}_i - \overline{\mathbf{x}}_k),$$

   donde $\overline{\mathbf{x}}_k$ es el centroide del k-ésimo clúster y $c(i)$ es el clúster que contiene a $\mathbf{x}_i$.
4. Reasignamos cada dato al clúster con el centroide más cercano de tal manera que ESS se reduce en magnitud. Actualizamos los centroides de los clústeres después de la reasignación de los datos.
5. Repetimos los pasos 3 y 4 hasta que no se produzcan más reasignaciones.

3.2.2. Particionamiento Alrededor de Medoides (PAM)

Este método de *clustering* es una modificación del algoritmo de *clustering* K-medoides. Aunque es similar al *clustering* de K-medias, este algoritmo busca los K "objetos representativos" (o medoides), en vez de los centroides, entre los datos del conjunto, y se usa una distancia basada en la disimilitud en lugar de la

distancia euclídea al cuadrado. Por ello, minimiza la suma de las disimilitudes en vez de la suma de las distancias euclídeas. Este método es más robusto frente a datos anómalos como valores atípicos y valores perdidos.

Este algoritmo comienza con la matriz de proximidad $\mathbf{D} = (d_{ij})$, donde $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$, ya sea dada o calculada a partir del conjunto de datos, y una configuración inicial de los datos dentro de los K clústeres. Usando $\mathbf{D}$, nosotros encontramos ese dato (llamado objeto representativo o medoide) dentro de cada clúster que minimiza la disimilitud total para todos los datos dentro de su clúster. En el algoritmo de K-medoides, los centroides de los pasos 2,3 y 4 en el algoritmo de K-medias son remplazados por los medoides, y la función objetivo ESS es sustituida por ESS_{med} (definida más adelante).

El algoritmo de particionamiento alrededor de medoides (PAM) es una modificación de el algoritmo de K-medoides que introduce una estrategia de intercambio por la cual el medoide de cada clúster es reemplazado por otro dato de ese clúster, pero solo si este intercambio reduce el valor de la función objetivo. Una desventaja para ambos algoritmos es que aunque funciona bien en pequeños conjuntos de datos, no son suficientemente eficientes para usar en *clustering* de grandes conjuntos de datos. Los pasos del algoritmo PAM son:

1. Entrada: $\mathbf{D} = (d_{ij})$, K: número de clústeres.
2. Formar una asignación inicial en los K clústeres.
3. Localizar el medoide de cada clúster. El medoide del k-ésimo clúster esta definido como ese dato en el k-ésimo clúster que minimiza la disimilitud total a los otros datos dentro del clúster, $k = 1, 2, ..., K$.

4*a*. Para el *clustering* de K-medoides:

- Para el k-ésimo clúster, reasignamos el $i_k$ésimo dato al clúster con su medoide más cercano para que la función objetivo,
$$ESS_{med} = \sum_{k=1}^{K} \sum_{c(i)=k} d_{ii_k}$$
sea reducida en magnitud, donde $c(i)$ es el clúster que contiene el $i$ésimo dato.
- Repetir el paso 3 y el paso de reasignación hasta que no se produzcan mas reasignaciones.

4*b*. Para el *clustering* de particionamiento alrededor de medoides:

- Buscamos un dato no medoide tal que al intercambiarlo por otro medoide se produce una reducción en ESS_{med}, considerando el cambio que se produce en los clústeres al cambiar de medoide (si hay varios medoides que cumplen esta condición, seleccionar el que mayor reducción produzca) y se produce el cambio.
- Repetir el proceso de intercambio hasta que no se produzca ninguna reducción del ESS_{med}.

3.2.3. Análisis difuso (fanny)

La idea detrás del *clustering* difuso es que a los datos, para ser agrupados, se les asigna probabilidades de pertenencia a cada uno de los K clústeres. Sea u_{ik} como denotamos a la fuerza de pertenencia del i-ésimo dato al k-ésimo clúster. Para el i-ésimo dato, necesitamos que $\{u_{ik}\}$ se comporte como una probabilidad. Eso es que $u_{ik} \geq 0$, para todo i y $k = 1, 2, ..., K$, y $\sum_{k=1}^{K} u_{ik} = 1$ para cada i. Esto choca con los métodos de particionamiento de K-medias o PAM, donde cada dato es asignado a un solo clúster. Dada la matriz de proximidad $D = (d_{ij})$ y un número de clústeres K, la desconocida fuerza de pertenencia, $\{u_{ik}\}$, es encontrada minimizando la función objetivo,

$$\sum_{k=1}^{K} \frac{\sum_i \sum_j u_{ik}^2 u_{jk}^2 d_{ij}}{2 \sum_l u_{lk}^2}.$$

La función objetivo se minimiza bajo las restricciones de no negatividad y unidad de la suma usando un algoritmo iterativo.

4. Número de clústeres

En los algoritmos de *clustering*, uno de los problemas principales es determinar el número idóneo de clústeres k, el cual, es distinto al propio proceso de agrupamiento. Este procedimiento conlleva tener en cuenta varios índices. Evaluar la solución obtenida para k clústeres es similar a la práctica de determinar la dimensión del espacio en el análisis de componentes principales o el orden de los factores en análisis factorial. La correcta elección de k suele ser ambigua ya que depende de las interpretaciones según la forma y la escala de de la distribución de los datos y la solución deseada. También hay que tener en cuenta que incluso con datos aleatorios se pueden detectar falsos clústeres.

En cada paso del proceso de agrupamiento se crea un nuevo clúster formado por dos observaciones, una observación y otro clúster o dos clústeres ya obtenidos. Así, el número de clústeres k decrece de n (número de observaciones) a 1. La distancia entre dos clústeres es la euclídea o la de disimilitud. Como k decrece de n a 1, el valor de la distancia debería aumentar ya que tendría que ser mayor cuando dos clústeres distintos se agrupan en uno solo.

Uno de los procedimientos para determinar el número de clústeres, por ejemplo en el algoritmo k-medias, es el denominado "método del codo" o *elbow method*. Suele ser ambiguo y no muy fiable por lo que se recomienda el uso de otras técnicas. Consiste en dibujar la gráfica de las distancia a los centros de cada clúster en función del número de clústeres. En un punto, se observará que la gráfica forma una especie de codo, de ahí nombre del método, por lo que se escoge dicho punto. Así, llamamos:

$$SSE_k = \sum_{i=1}^{n_k} = \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|^2,$$

donde $\bar{\mathbf{y}}_k$ representa el centroide del clúster C_k. Para cada número de clústeres k calculamos

$$D_k = \sum_{i=1}^{k} SSE_k,$$

y lo dibujamos en una gráfica. Si usamos el ejemplo A aportado en [8], la gráfica sería la encontrada en la Figura 7.

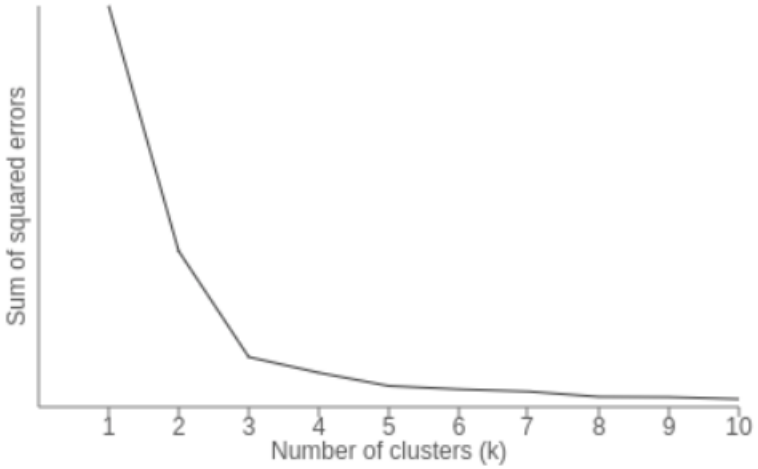

Figura 7: Ejemplo del método del codo.

Vemos que el cambio de de inclinación se produce cuando $k = 3$ por lo que ese sería el número idóneo de clústeres. Hemos construido por tanto una especie de índice de separación o test de modo que si lo dibujamos en función del número de clústeres nos ayuda a identificar visualmente el k adecuado.

Podemos construir otra serie de índices más complejos siguiendo la misma filosofía que la anterior: construir el índice de separación o una estadística de prueba y dibujarlo en una gráfica según k de tal modo que un cambio significativo en el índice nos dé una aproximación del número de clústeres adecuado. Por ejemplo, en análisis de regresión, el coeficiente de determinación, R^2, es una medida de la varianza total de la variables dependientes según las independientes. En *clustering* debemos construir un índice de R^2 que varíe en base al número de clústeres. Para n clústeres la suma total de las distancias al cuadrado es $T = \sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2$. Así, para k clústeres definimos R^2 como

$$R_k^2 = \frac{T - \sum_k SSE_k}{T}.$$

Para n clústeres $SSE_k = 0$ por lo que $R^2 = 1$. A medida que vamos realizando los agrupamientos, estos estarán más separados. Una gran disminución en R_k^2 representaría un agrupamiento diferente. También podríamos tener en cuenta el cambio en R^2 al unir los clústeres R y S como $SR^2 = R_k^2 - R_{k-1}^2$. El estadístico SR^2 representa, en función de T, la proporción de $SSE_t - (SSE_r + SSE_s)$ donde los clústeres C_R y C_S se han unido para formar el clúster C_T. Cuanto mayor sea el índice mayor será la pérdida de homogeneidad.

El objetivo del análisis clúster es encontrar el menor número de clústeres homogéneos. Para un solo clúster, la varianza agrupada para todas las variables es la media de las varianzas de cada una de las variables, es decir, $s^2 = \sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 / p(n-1)$. También podemos calcularla para un clúster C_k con n_k observaciones de la siguiente manera:

$$s^2 = \sum_{i=1}^{n_k} \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|^2 / p(n_k - 1).$$

Valores grandes de la varianza agrupada indica que los clústeres no son homogéneos. Por lo tanto, si tiende a cero para algún $k < n$ indica la formación de un clúster homogéneo.

Bajo una normal multivariante e independencia de los n vectores de dimensión p para $\Sigma = \sigma^2\mathbf{I}$, podríamos hacer una prueba para verificar que los k clústeres muestran una separación significativa usando, por ejemplo, un análisis de la varianza (ANOVA) mediante una prueba F de Fisher. También podemos comprobar si dos medias están lo suficientemente separadas en cualquier nivel usando un estadístico t, que son los más comunes en las pruebas t de Student. Como no es común que la independencia y la distribución normal multivariante se den a la vez, los estadísticos se denominan pseudo estadísticos F y t^2. El pseudo estadístico F se define como

$$F_k^* = \frac{(T - \sum_k SSE_k)/(k-1)}{\sum_k SSE_k/(n-k)}.$$

Si F_k^* disminuye con k, no deberíamos usarlo para estimar k. Sin embargo, si F_k^* disminuye con k y alcanza máximo, el valor de k en el que alcanza dicho máximo o el inmediatamente anterior el candidato del número de clústeres. Por otro lado, el pseudo estadístico t^2 se define como

$$\text{pseudo } t^2 = \frac{[SSE_t - (SSE_r + SSE_s)](n_R + n_S - 2)}{SSE_r + SSE_s},$$

para agrupar los clústeres C_R y C_S con n_R y n_S elementos respectivamente. De nuevo, uno puede construir la gráfica con los valores obtenidos y el número de clústeres. Si los valores son irregulares en cada punto de agrupamiento, no es un buen índice. Pero si la gráfica parece un palo de hockey, similar al método del codo, el valor $k+1$ que causa que la pendiente cambie es nuestro candidato a número de clústeres.

Varios estadísticos se generan por los programas que realizan el proceso de agrupamiento y son dibujados para evaluar heurísticamente el número de clústeres generados. Otras técnicas utilizadas para determinar el k óptimo son por ejemplo el el método de la silueta (*silhouette method*) o el de la brecha (*gap*).

En el *silhouette method*, se observa la similitud de cada observación con su clúster en comparación con el resto de clústeres. El índice se encuentra entre los valores -1 y 1 donde un valor próximo a 1 significa un buen agrupamiento. Definimos el índice en este método para cada observación i como

$$s(i) = \frac{b(i) - a(i)}{\text{máx}\{b(i), a(i)\}}, \quad \forall i = 1, \ldots, n$$

donde $a(i)$ es la media de la disimilitud entre i y el resto de puntos que pertenecen al mismo clúster, es decir,

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j),$$

y $b(i)$ como la menor distancia media de i a todos los puntos de los clústeres al que i no pertenece:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j).$$

Destacamos que si $|C_i| = 1$ entonces $s(i) = 0$. Así, si muchos objetos tienen un valor alto indica que el resultado obtenido es satisfactorio por lo que se escoge el k que maximice el valor medio de $s(i)$. Para ejemplificarlo, tomamos los resultados obtenidos en la referencia [5] y se obtiene la Tabla 11.

k	Silhouette coeff.
2	0.7049787496083262
3	0.5882004012129721
4	0.6505186632729437
5	0.5745566973301872
6	0.43902711183132426

Tabla 11: Ejemplo *silhouette method*.

Vemos que se obtienen los mejores resultados con 2 o 4 clústeres por lo que deberíamos decidir entre ambos según el resultado que queramos obtener.

Finalmente, el método de la brecha es parecido al método del codo y consiste en comparar la variación total dentro de un clúster para diferentes valores de k con sus valores esperados. El k elegido será aquel que maximice el valor de la brecha. Definimos el estadístico como:

$$Gap(k) = E_n^*\{\log(W_k)\} - \log(W_k).$$

En la fórmula anterior E_n^* denota la media de una de muestra de tamaño n y

$$W_k = \sum_{R=1}^{k} \frac{1}{2n_R} \sum_{ij \in C_R} d(i, j).$$

Notamos que se puede usar para cualquier método y distancia. Según [13] el número 2 en la fórmula de W_k es para que funcione correctamente. Visualmente se obtendría una gráfica como la mostrada en la Figura 8.

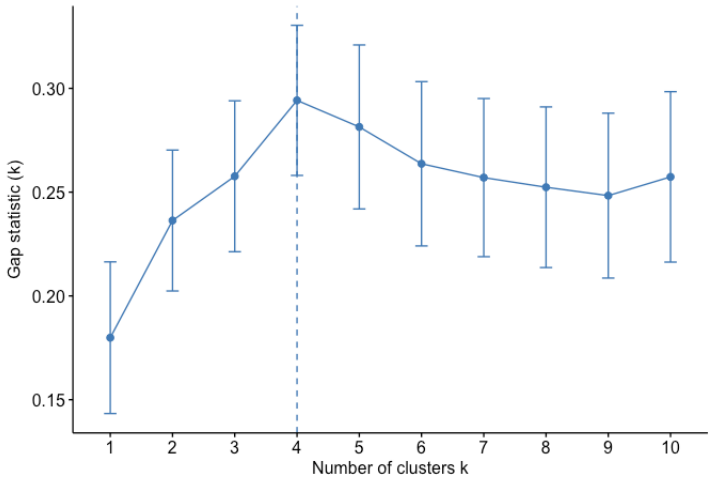

Figura 8: Ejemplo del método de la brecha [4].

Así, lo ideal sería tomar $k = 4$.

5. Parte práctica

El objetivo de este apartado consiste en el estudio de técnicas de aprendizaje no supervisado para análisis relacional mediante segmentación. Usaremos distintos algoritmos de *clustering* sobre el conjunto de datos de prueba y extraeremos conclusiones sobre los resultados obtenidos.

El conjunto de datos de prueba sobre el que realizaremos el estudio es el denominado conjunto de datos iris de **Fisher** [7] que consta de un conjunto de datos multivariante introducido por **Ronald Fisher** en su artículo de 1936 (*The use of multiple measurements in taxonomic problems* [11]) como un ejemplo de análisis discriminante lineal.

El conjunto de datos contiene 50 muestras de cada una de tres especies de flor **Iris** (setosa, virginica y versicolor). En él, se recogen las medidas de cuatro rasgos para cada muestra: el largo y ancho del sépalo y y el largo y ancho del pétalo, en centímetros. Basado en la combinación de estos cuatro rasgos, **Fisher** desarrolló un modelo discriminante lineal para distinguir entre una especie y otra. En las Figuras **??**, **??**, **??** encontramos las tres especies de iris.

Figura 9: Iris Setosa.

Figura 10: Iris virginica.

Figura 11: Iris versicolor.

A continuación, se mostrarán los resultados obtenidos por los algoritmos en este caso de estudio. Utilizaremos una tabla comparativa donde incluiremos el nombre de los algoritmos, los resultados obtenidos por las métricas utilizadas (Calinski-Harabaz y Silhouette) y los tiempos de ejecución de los mismos. Las métricas utilizadas se interpretarán de la siguiente forma:

- **Calinski-Harabaz**: se basa en el concepto de densidad y de cómo de bien están separados los clústeres. Especifica la relación entre la dispersión entre distintos clústeres y la dispersión dentro de un mismo clúster. Nos indicará si estamos usando un buen número de clústeres para un algoritmo en concreto. El número óptimo de clústeres es la solución con el valor de índice Calinski-Harabasz más alto.
- **Silhouette**: es una medida que indica cómo de similar es un objeto respecto a su propio grupo (cohesión) en comparación con otros grupos (separación). Toma valores entre -1 y +1, donde un alto valor indica que el objeto es bastante similar a su grupo y muy diferente a los de otros clúster. Si el valor es cercano a 1, la configuración de los clúster es apropiada, si no, habrá más o menos clústeres de los necesarios.

Los algoritmos elegidos han sido los siguientes: **K-Means, MeanShift, DBSCAN y Agglomerative *clustering*** (*clustering* jerárquico). A la hora de elegir las mejores versiones de cada algoritmo, nos basaremos en el coeficiente de Calinski-Harabaz (**CH**) que como hemos explicado antes, tendrá mayor valor si estamos usando el número adecuado de clústeres.

Mostramos en la Tabla 12 una comparativa que contiene las mejores versiones de cada algoritmo ejecutado. Se hará un estudio más profundo sobre aquellos dos algoritmos que obtengan un mayor valor del coeficiente de Silhouette. En este caso, **K-Means** y **Agglomerative *clustering***.

Nombre	**Nº clústeres**	**CH**	**SH**	**Tiempo (*s*)**	**Clústeres**
K-Means	3	359.845074	0.504769	0.016456	0: 61 (40.67 %) 1: 50 (33.33 %) 2: 39 (26.00 %)
DBSCAN	4	94.991819	0.306404	0.002353	0: 45 (30.00 %) 1: 39 (26.00 %) -1: 36 (24.00 %) 2: 30 (20.00 %)
AggCluster	3	349.254185	0.504800	0.019058	0: 67 (44.67 %) 1: 50 (33.33 %) 2: 33 (22.00 %)
MeanShift	3	290.470683	0.476961	0.289073	0: 81 (54.00 %) 1: 50 (33.33 %) 2: 19 (12.67 %)

Tabla 12: Tabla comparativa algoritmos.

5.1. K-Means

K-Means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en K grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster.

El algoritmo K-Means resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su clúster. Es necesario especificar el valor de K previamente antes de ejecutar el algoritmo.

Para el algoritmo K-Means se ha utilizado el siguiente código en python y se ha obtenido la siguiente agrupación de las muestras:

```
KMeans(init='k-means++', n_clusters=3, n_init=5, random_state=12345)
```

```
0:      61  (40.67%)
1:      50  (33.33%)
2:      39  (26.00%)
```

Se mostrarán 4 gráficas realizadas y se hará un estudio sobre los resultados obtenidos. Las gráficas obtenidas para este caso, serán **Scatter Matrix, Heatmap, KPlot y BoxPlot** que mostrarán la matriz de dispersión de las muestras, mapa de calor de los centroides para cada variable y distribución de las muestras según cada variable.

En la Figura 12 se representa la matriz de dispersión de las 150 muestras agrupadas en 3 clústeres: donde el clúster 0 representa las **Iris** versicolor, el clúster 1 setosa y el clúster 2 virginica. Un gráfico de matriz de dispersión es una herramienta de exploración de datos que permite buscar patrones y relaciones entre diferentes muestras en una distribución multivariante. En el gráfico se muestra el gráfico de dispersión para cada par de variables seleccionadas y una serie de histogramas en la diagonal mostrando la distribución de valores para cada una de las variables.

En la Figura 13 y 14 podemos ver la media y forma de las distribuciones de cada variable para cada clúster generado con el algoritmo K-Means. En el HeatMap, podemos observar la media de los centroides y podemos hacernos una idea de las características claves para clasificar cada muestra en uno de los tres tipos de flor de **Iris**.

Las muestras del clúster 1 (tipo setosa) se caracteriza por tener un pétalo corto y delgado, mientras que el ancho del sépalo es más ancho que la media. Las del clúster 2 (tipo virginica), se caracterizan por tener un pétalo y sépalo más grandes respecto a los otros tipos y las del clúster 3 (tipo versicolor) tienen un pétalo similar al de la virginica pero se caracteriza por tener el sépalo un poco más delgado. Podemos observar que las flores más dificiles de diferenciar serán las de los tipos virginica y versicolor puesto que tienen más características en común. Esto se puede visualizar en la matriz de dispersión y en la Figura 14.

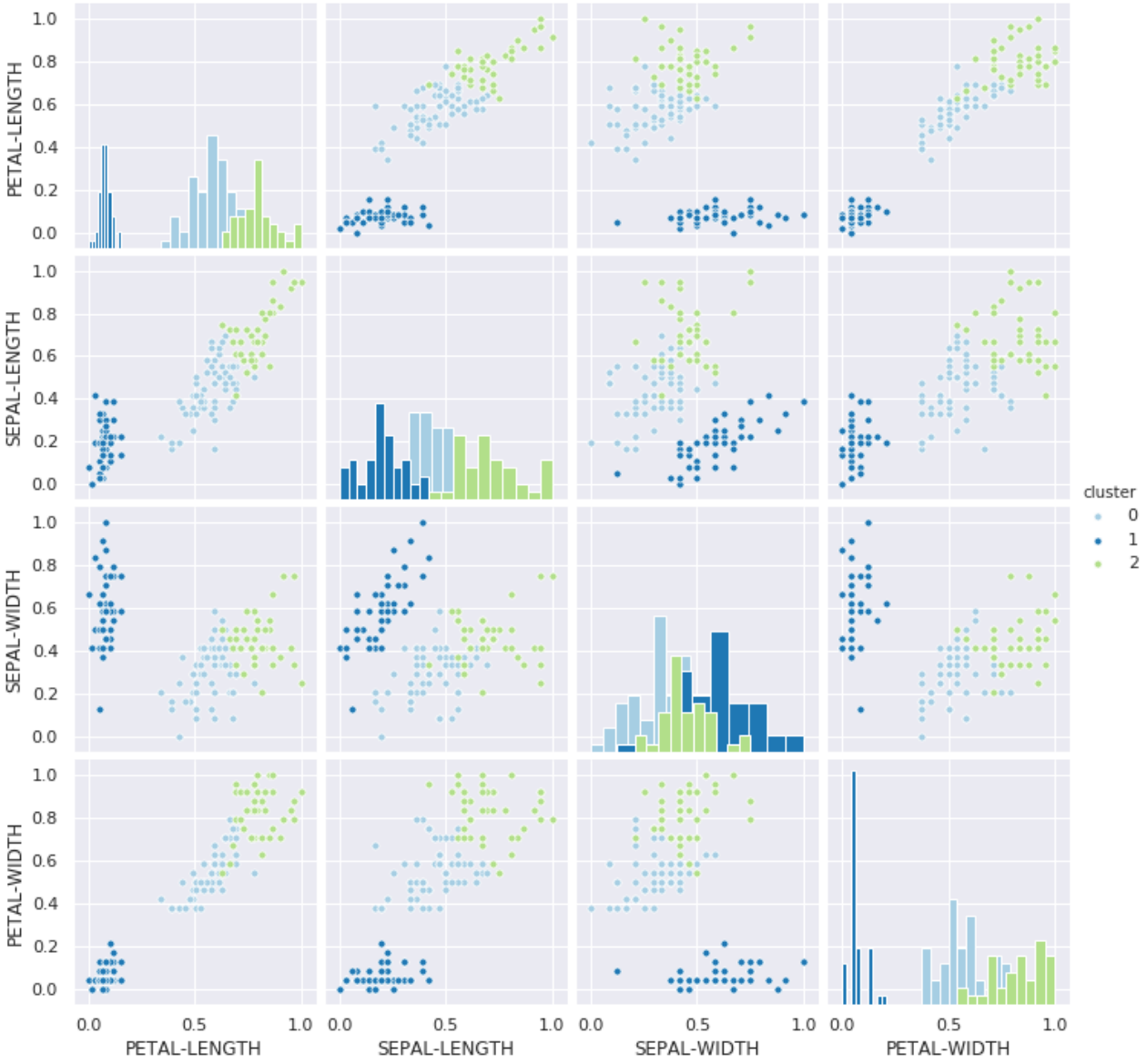

Figura 12: Scatter Matrix para K-Means con 3 clústeres.

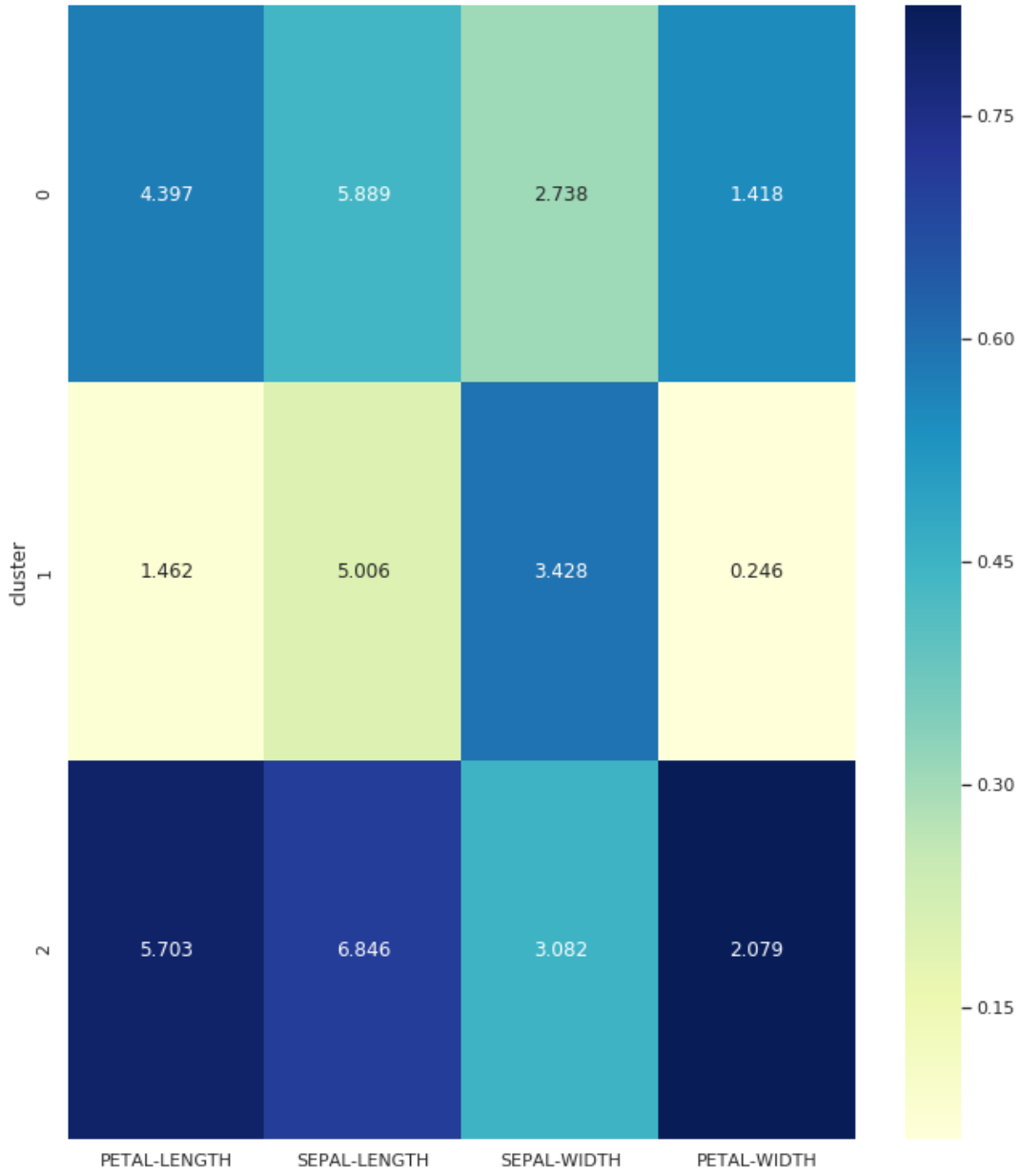

Figura 13: HeatMap para K-Means con 3 clústeres.

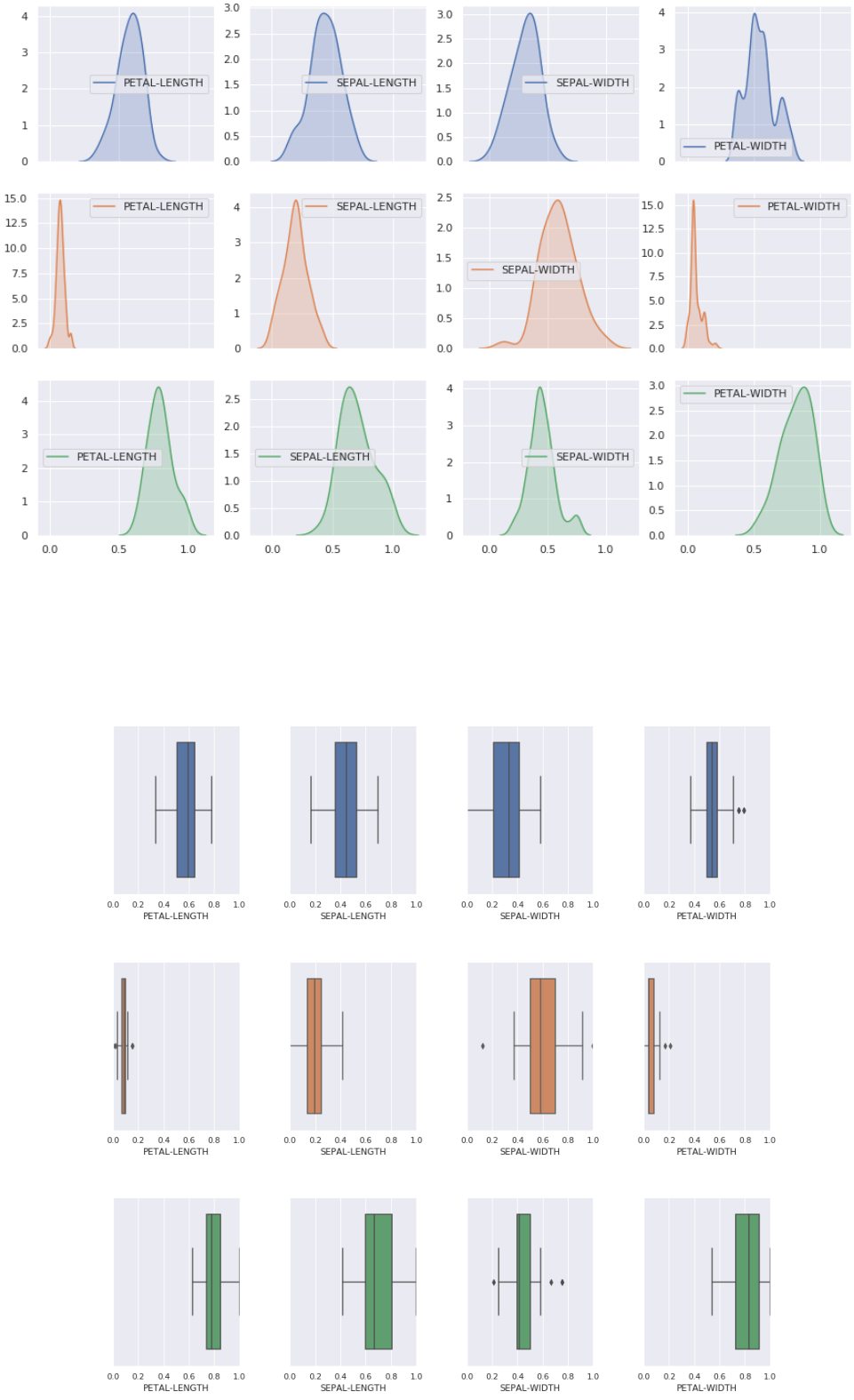

Figura 14: KPlot y BoxPlot para K-Means con 3 clústeres.

5.2. Agrupamiento Jerárquico

En orden de decidir qué grupos deberían ser combinados se requiere una medida de disimilitud entre conjuntos de observaciones. En la mayoría de los métodos de agrupamiento jerárquico, esto se consigue mediante el uso de una métrica apropiada (una medida de distancia entre pares de observaciones), y un criterio de enlace el cual determina la distancia entre conjuntos de observaciones como una función de las distancias entre observaciones dos a dos.

En nuestro caso, usaremos la estrategia de agrupamiento aglomerativo y usaremos la métrica euclídea junto con el criterio de enlace ward (el decrecimiento en la varianza para los grupos que están siendo mezclados).

Para el algoritmo jerárquico aglomerativo, se ha utilizado el siguiente código en python y se ha obtenido la siguiente agrupación de las muestras:

```
AgglomerativeClustering(n_clusters=3, linkage="ward", affinity='euclidean')
```

```
0:      67  (44.67%)
1:      50  (33.33%)
2:      33  (22.00%)
```

Se mostrarán 4 gráficas realizadas y se hará un estudio sobre los resultados obtenidos. Las gráficas obtenidas para este caso, serán la matriz de dispersión y mapa de calor también usados para el método de K-Means y ademaś se mostrarán dos dendrogramas para representar el proceso de agrupamiento seguido.

Podemos observar que la matriz de dispersión generada por este algoritmo se asemeja mucho a la obtenida por el algoritmo K-Means. Las conclusiones que podemos extraer de las Figuras 15 y 16 son similares a las que ya hemos comentado en el apartado anterior, por lo que nos limitaremos a extraer conclusiones sobre el dendrograma.

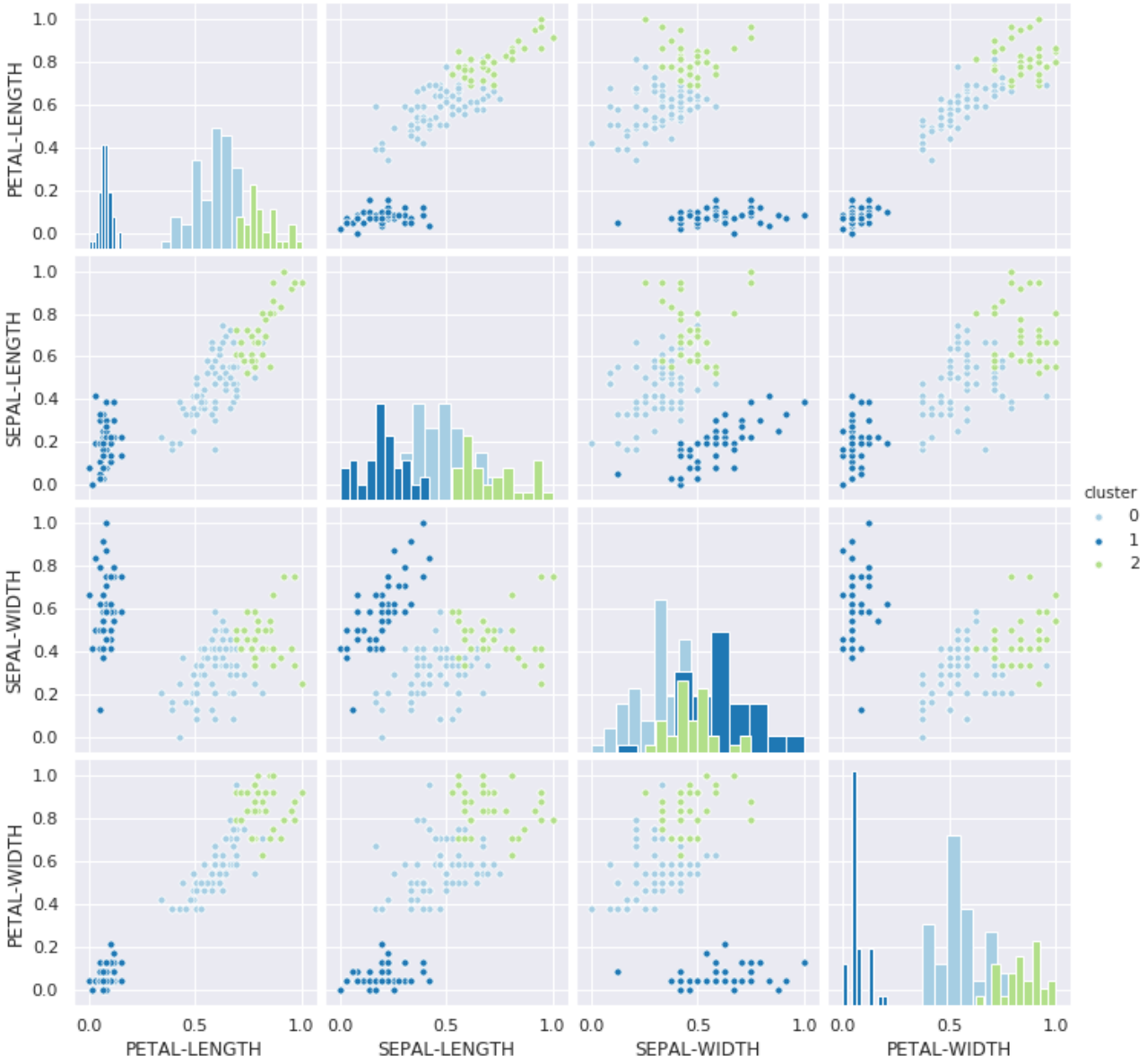

Figura 15: Scatter Matrix para Agglomerative *clustering* con 3 clústeres.

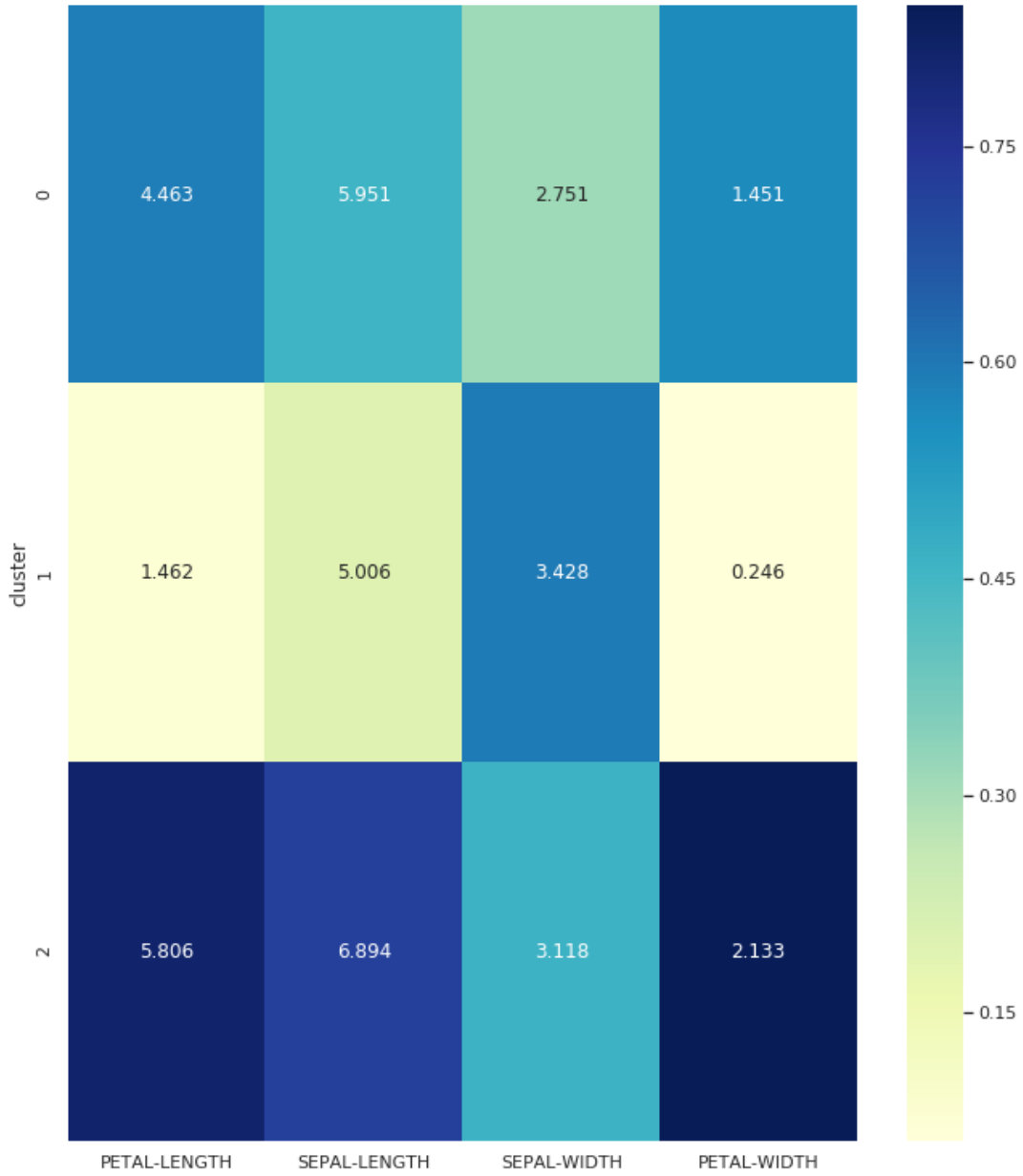

Figura 16: HeatMap para Agglomerative *clustering* con 3 clústeres.

5.2.1. Dendrogramas

Un dendrograma es un diagrama de árbol que muestra los grupos que se forman al crear clústers de observaciones en cada paso y sus niveles de similitud. En nuestro caso medimos el nivel de similitud en el eje horizontal y las diferentes muestras en el eje vertical.

Para ver los niveles de similitud, seleccionaremos una línea vertical del dendrograma. El patrón de cómo los valores de similitud/desimilitud cambian de un paso a otro será clave en la agrupación final para los datos.

La decisión acerca de la agrupación final también se conoce como cortar el dendrograma. Cortar el dendrograma es similar a trazar una línea a lo largo del dendrograma para especificar la agrupación final. También se pueden comparar diferentes agrupaciones finales en los dendrogramas para determinar cuál de ellas tiene más sentido para los datos.

En nuestro caso hemos impuesto que existan 3 clústers por lo que podríamos cortar el dendrograma a distancia 1, obteniendo el número elegido. Las muestras de color verde representarían a las **Iris** setosa y las de color rojo a las versicolor y virginica.

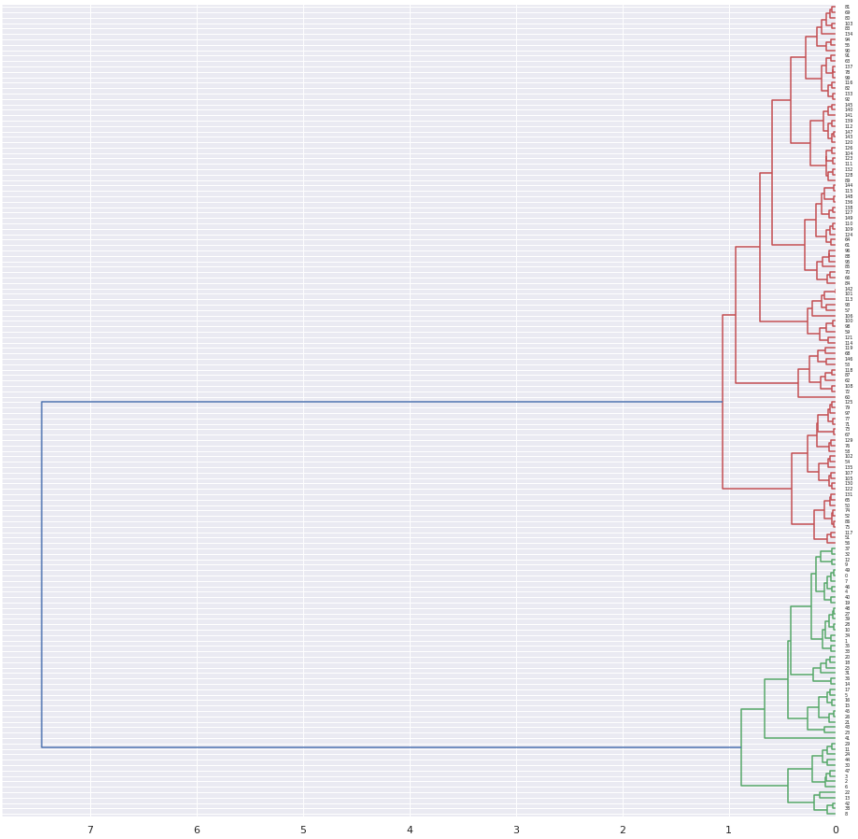

Figura 17: Dendrograma para Agglomerative *clustering* con 3 clústeres.

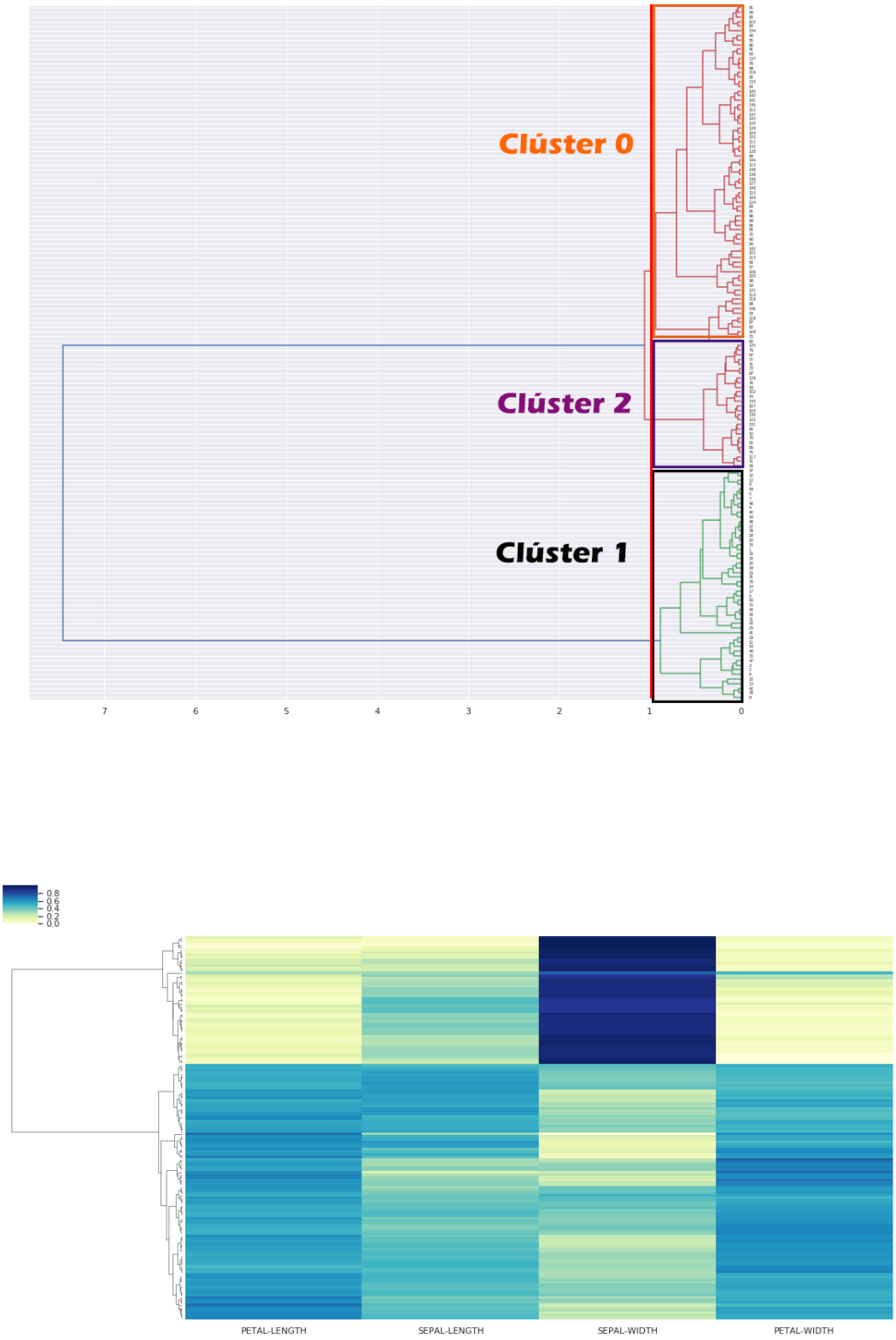

Figura 18: Dendrogramas para Agglomerative *clustering* con 3 clústeres.

5.3. DBSCAN

Para DBSCAN, se ha utilizado el siguiente código en python y se ha obtenido la siguiente agrupación de las muestras, donde el clúster -1 representa las muestras formadas por ruido:

```
DBSCAN(eps=0.12, min_samples=5)

0:      45 (30.00%)
1:      39 (26.00%)
-1:     36 (24.00%)
2:      30 (20.00%)
```

Las gráficas obtenidas para este caso, serán la matriz de dispersión (Scatter Matrix) y mapa de calor (HeatMap). En este caso podemos ver como tenemos un cuarto clúster formado por las muestras que han sido consideradas ruido"por el algoritmo DBSCAN. Este clúster se identifica con el número -1 y contiene las muestras no alcanzables para el algoritmo con los parámetros que hemos especificado.

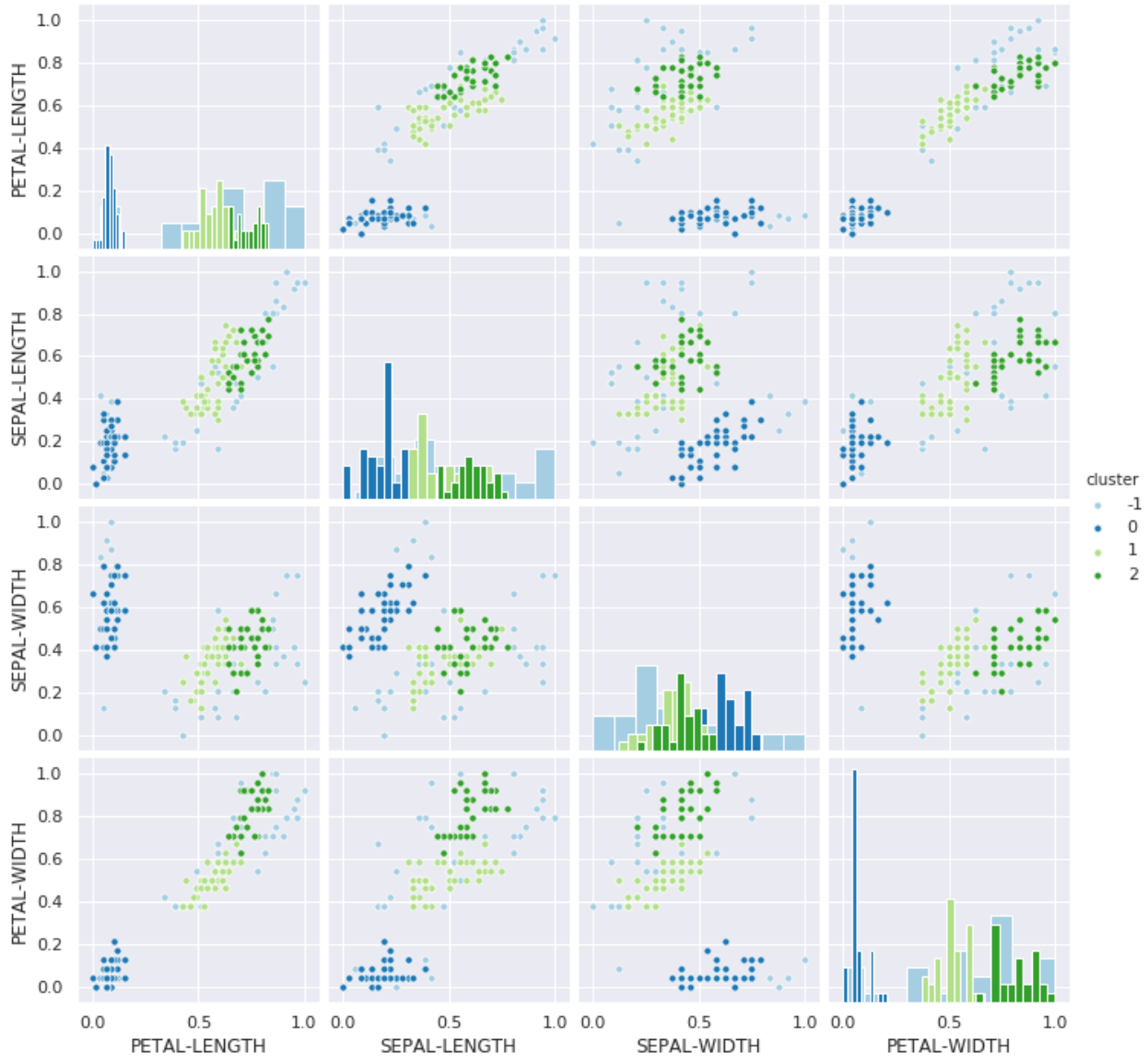

Figura 19: Scatter Matrix para DBSCAN con 4 clústeres.

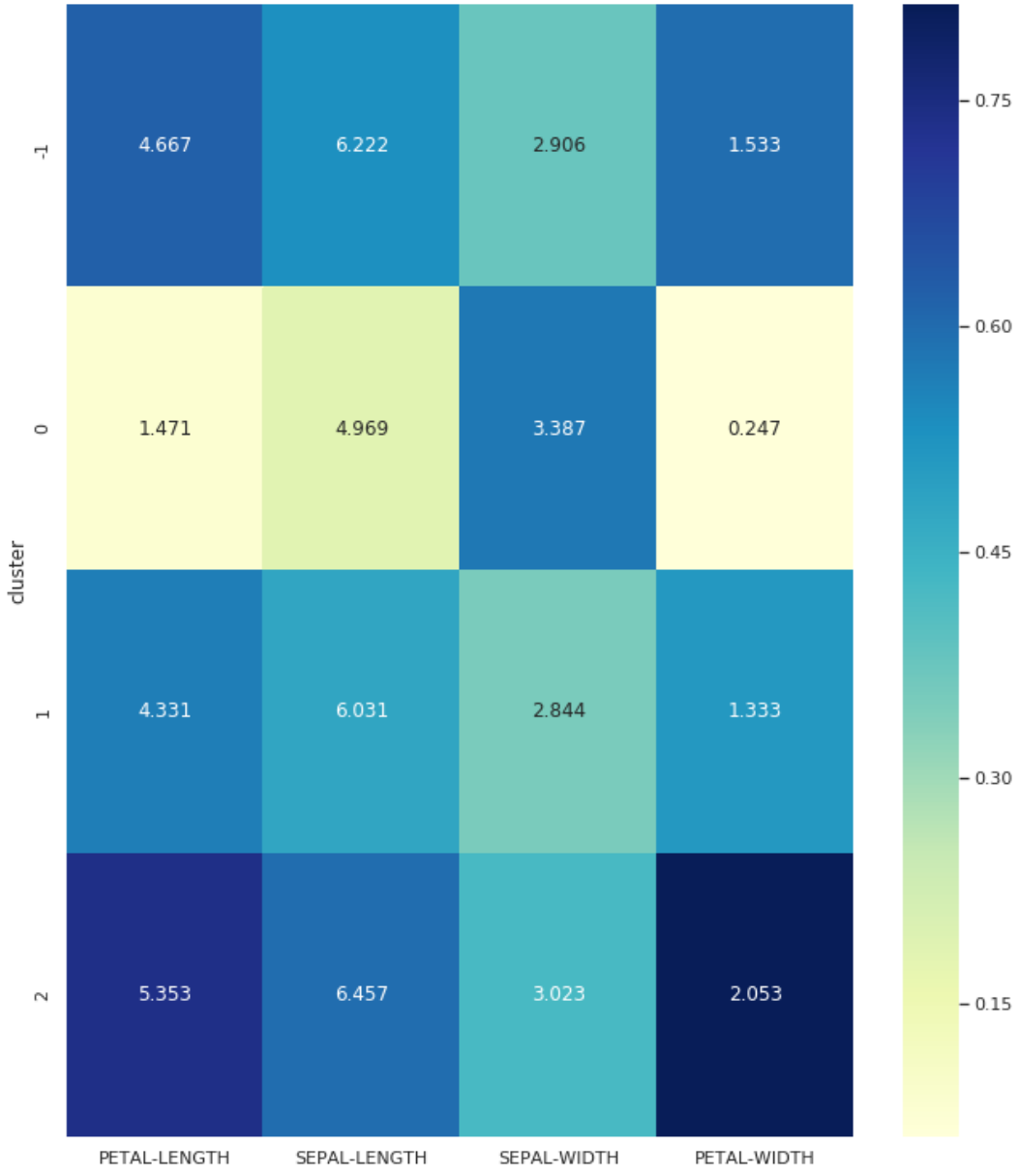

Figura 20: HeatMap para DBSCAN con 4 clústeres.

5.4. Mean Shift

Para Mean Shift, se ha utilizado el siguiente código en python y se ha obtenido la siguiente agrupación de las muestras:

```
MeanShift(bandwidth=estimate_bandwidth(X_normal, quantile=0.67,
          n_samples=400), bin_seeding=True)
```

```
0:      81  (54.00%)
1:      50  (33.33%)
2:      19  (12.67%)
```

Las gráficas obtenidas para este caso, serán la matriz de dispersión (Scatter Matrix) y mapa de calor (HeatMap).

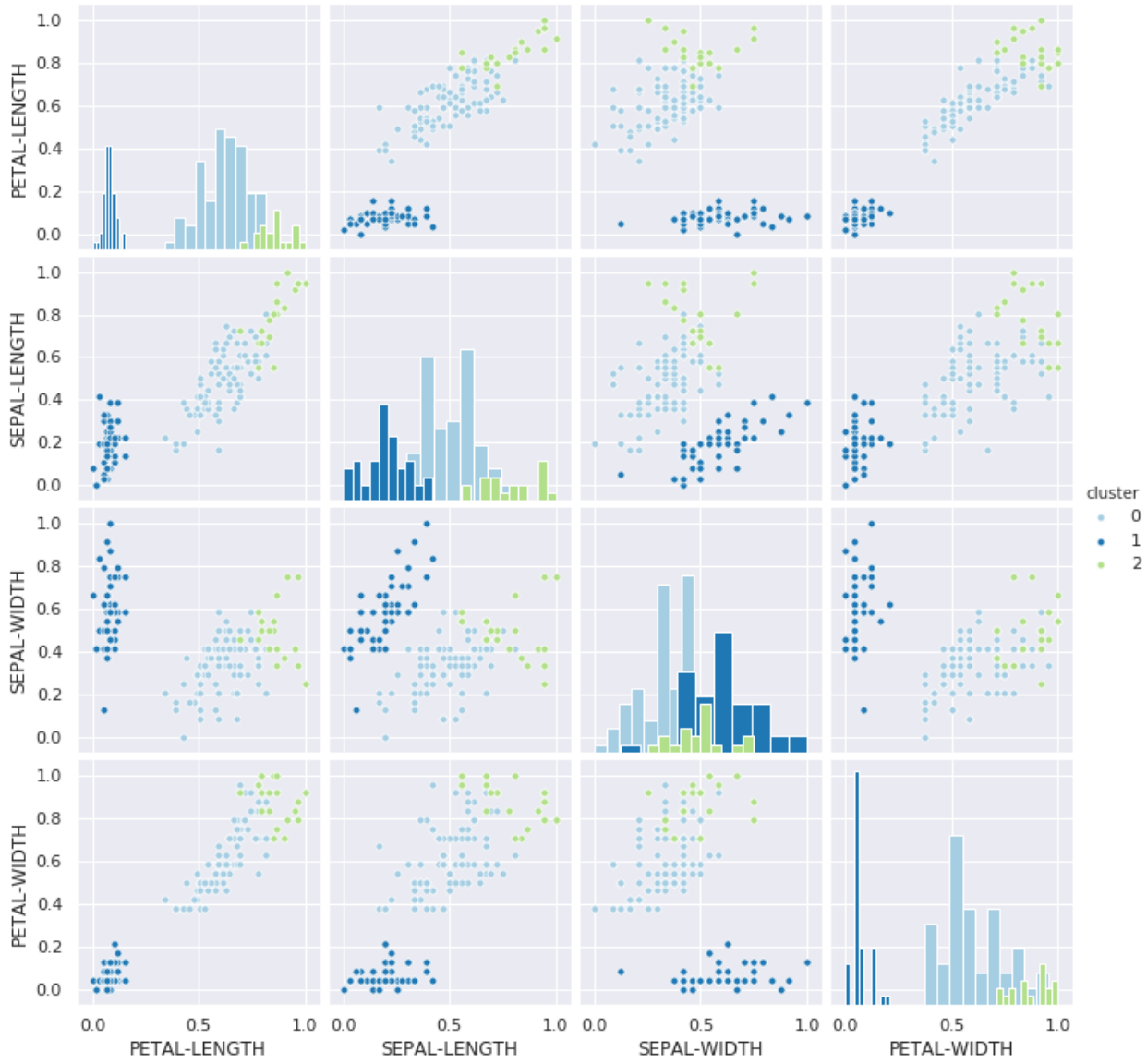

Figura 21: Scatter Matrix para Mean Shift con 3 clústeres.

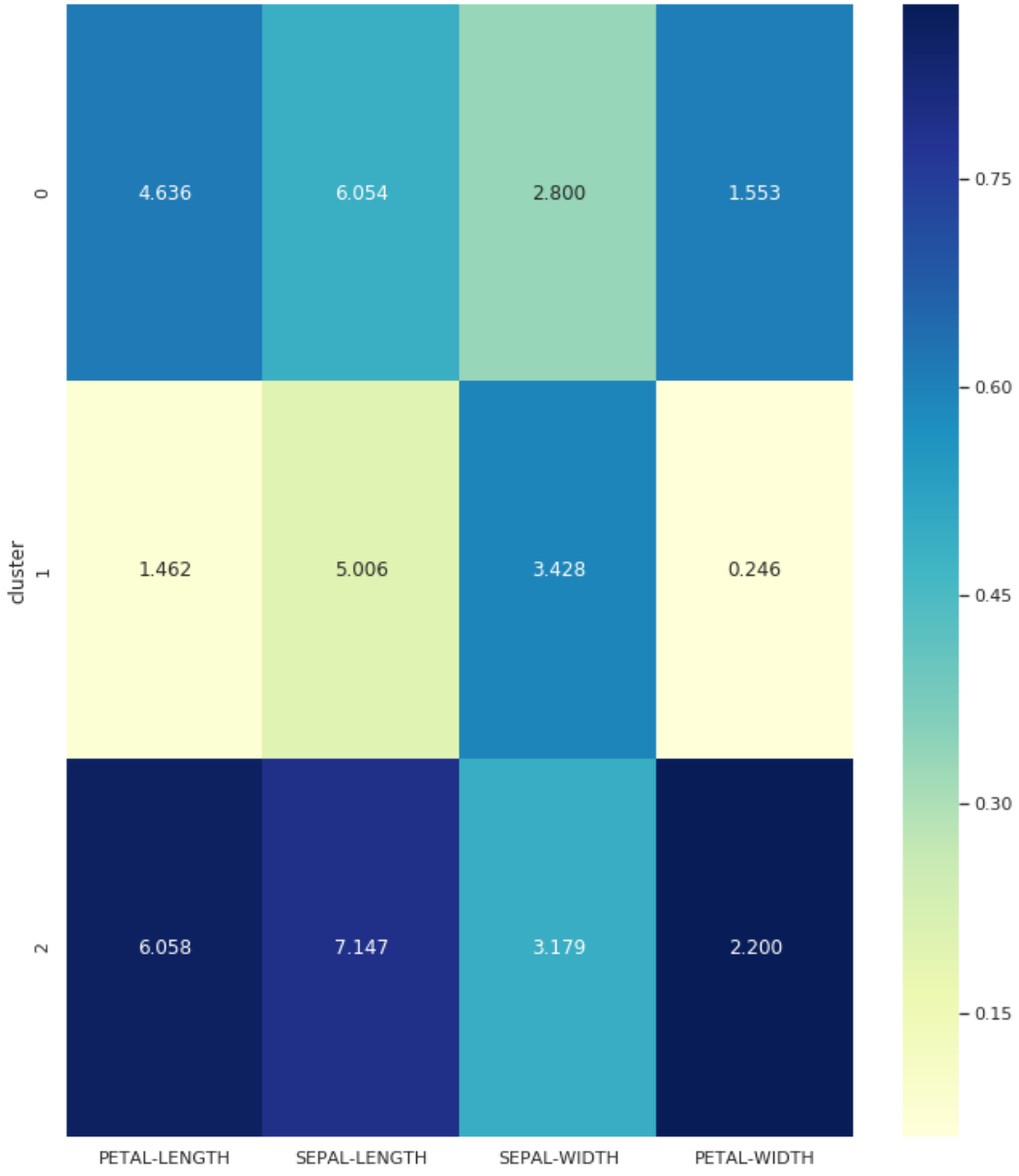

Figura 22: HeatMap para Mean Shift con 3 clústeres.

6. Conclusiones

El análisis clúster es una metodología de análisis exploratorio de datos. Trata de descubrir cómo algunos objetos se relacionan. Este tipo de análisis depende de la cantidad de ruido en los datos, la existencia de datos anómalos, las variables seleccionadas para el análisis, la medida de proximidad utilizada, las propiedades espaciales de los datos y el método de agrupamiento empleado.

Una ventaja de los métodos jerárquicos sobre los no jerárquicos es que uno no tiene que saber de antemano el número de clústeres (que hemos visto que no es fácil). Además los métodos jerárquicos muchas veces son llamados "exploratorios" mientras que los no jerárquicos se denominan "confirmatorios". Podemos ver ambos métodos como complementarios. Los métodos jerárquicos no solo dependen de la medida de proximidad utilizada, se ven afectados por el encadenamiento: los objetos tienen a unirse a un clúster existente más que a iniciar uno nuevo.

Un problema importante del *clustering* es la validación de la solución. Para ello deberemos usar criterios internos, externos y métodos de replicación o validación cruzada.

7. Bibliografía

Referencias

[1] Determining the number of clusters in a data set. https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set. Último acceso: 28/12/2019.

[2] Elbow method (clustering). https://en.wikipedia.org/wiki/Elbow_method_(clustering). Último acceso: 28/12/2019.

[3] Hierarchical grouping. https://es.wikipedia.org/wiki/Agrupamiento_jer%C3%A1rquico. Último acceso: 13/01/2020.

[4] K-means cluster analysis. https://uc-r.github.io/kmeans_clustering, note = "Último acceso: 28/12/2019".

[5] Selecting the number of clusters with silhouette analysis on kmeans clustering. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html. Último acceso: 28/12/2019.

[6] Silhouette (clustering). https://en.wikipedia.org/wiki/Silhouette_(clustering). Último acceso: 28/12/2019.

[7] UCI Machine Learning Repository: Iris Data Set.

[8] Using the elbow method to determine the optimal number of clusters for k-means clustering. https://bl.ocks.org/rpgove/0060ff3b656618e9136b, note = "Último acceso: 28/12/2019".

[9] S Aranganayagi and K Thangavel. Clustering categorical data using silhouette coefficient as a relocating measure. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, volume 2, pages 13–17. IEEE, 2007.

[10] Chire. Cluster analysis with optics on a density-based data set. https://commons.wikimedia.org/wiki/File:OPTICS-Gaussian-data.svg, October 2011.

[11] R. A. Fisher. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics*, 7(2):179–188, September 1936.

[12] Wolfgang Karl Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[13] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[14] N.H. Timm. *Applied Multivariate Analysis*, chapter 9. Springer Texts in Statistics. Springer New York, 2002.