Análisis de parcelas con cabras en Teno

Enfocado en el análisis clúster

Manuel Verdejo García

18/03/2025

Índice

1. Antecedentes

Un investigador ha recolectado datos de 150 lirios, midiendo la longitud y el ancho de los pétalos, así como la longitud y el ancho de los sépalos, junto con la especie correspondiente. Previamente, realizó un análisis visual de agrupamiento y sostiene que es capaz de identificar correctamente las tres especies únicamente con la observación. Para evaluar esta hipótesis, se excluirá la variable *especie* y se realizará un análisis de clúster utilizando únicamente las variables morfológicas. Se forzará al algoritmo a identificar tres grupos y los resultados obtenidos serán comparados con la clasificación visual del investigador.

2. Objetivos

- Realizar un análisis exploratorio para determinar el número óptimo de clústers y evaluar si efectivamente existen tres grupos diferenciados en los datos.
- Llevar a cabo un análisis confirmatorio de clasificación errónea (*misclassification*) para verificar, en caso de haber tres grupos, qué tan precisa fue la clasificación realizada por el investigador.

3. Variables en el Informe

El análisis se basa en mediciones realizadas sobre 150 lirios. Las variables consideradas incluyen:

- **Longitud del pétalo**
- **Ancho del pétalo**
- **Longitud del sépalo**
- **Ancho del sépalo**
- **Especie** (no se utilizará en el análisis de clúster, según lo indicado en los antecedentes, pero sí para validar los resultados del modelo.).

4. Modelo

4.1. Distancias

Dado que se trabaja con variables cuantitativas, se estandarizan (media = 0, varianza = 1). Se aplican principalmente distancias **euclidianas** También se considera la **distancia de Mahalanobis**, que ajusta las correlaciones entre variables y mejora la detección de valores atípicos.

4.2. Algoritmo Jerárquico

Se usa el método jerárquico aglomerativo, donde cada observación comienza como un clúster individual y se fusionan sucesivamente según distintos criterios de enlace: enlace simple y método de Ward (este último es óptimo si las variables son cuantitativas). Se construyen varios dendogramas para determinar el número de clústeres mediante el mayor salto en la distancia de fusión.

4.3. Algoritmo K-Medias

Se parte de una partición aleatoria de los lirios en k = 3 clústeres. El algoritmo optimiza la asignación de individuos minimizando la distancia entre observaciones y centros de clúster, que son recalculados iterativamente hasta que la diferencia entre iteraciones sea menor que un umbral ϵ.

4.4. ANOVA

Se realiza un ANOVA para determinar si existen diferencias significativas entre las medias y significancias de las variables en los distintos clústeres generados por K-medias. El modelo es: $y_{ij} = \mu + \zeta_i + \epsilon_{ij}$

5. Estudio Piloto

No procede en este caso, ya que se supone que se ha realizado un estudio previo.

6. Estudio de la Forma y Tamaño de la Muestra:

De igual manera, no aplica en este contexto.

7. Análisis de datos

a. Tipificación

Las variables han sido tipificadas, es decir, normalizadas a una escala común, con el propósito de evitar que las componentes principales se vean influenciadas por las diferencias de escala entre las variables. Sin esta tipificación, las variables con mayores magnitudes numéricas podrían dominar el análisis, distorsionando la interpretación de las componentes principales.

b. Selección de distancia y similitud

La elección de una métrica adecuada para medir la distancia entre observaciones es fundamental en cualquier análisis de clúster. En este caso, se considerarán dos opciones relevantes:

Distancia Euclídea

Sus principales ventajas son:

- **Interpretación geométrica intuitiva**: Mide la distancia "en línea recta" entre dos puntos en el espacio de características, lo cual facilita la interpretación visual del agrupamiento.

- **Adecuada para variables en escalas similares**: Las medidas morfológicas (longitud y ancho del sépalo y del pétalo) están expresadas en las mismas unidades y poseen rangos comparables.

- **Eficiencia computacional**: Es rápida de calcular, lo que la hace ideal para métodos como *k-means* y clustering jerárquico.

Distancia de Mahalanobis

Esta métrica tiene en cuenta la estructura de covarianza de los datos, permitiendo una mejor representación de relaciones entre variables. Sus beneficios incluyen:

- **Captura correlaciones entre características**: En datos biológicos, es común encontrar relaciones entre dimensiones, como la longitud y ancho de los órganos florales. La distancia de Mahalanobis ajusta estas dependencias.

- **Normalización de escala automática**: Da un peso equitativo a todas las direcciones del espacio multivariado, compensando diferencias en varianzas.

- **Detección más precisa de valores atípicos**: Considera la forma de la distribución multivariada, ayudando a identificar observaciones inusuales.

c. Selección y aplicación de criterios de agrupación

i. Jerárquicos:

Para la generación de clústeres se emplearán tanto métodos jerárquicos como de partición, evaluando su eficacia en términos de coherencia interna y concordancia con la clasificación visual original del investigador.

Métodos Jerárquicos

Los algoritmos jerárquicos comienzan considerando cada observación como un clúster individual, y van fusionando pares sucesivos hasta formar un único grupo. La forma en que

se mide la “distancia” entre grupos se denomina criterio de enlace (*linkage*), y las variantes más comunes son:

- **Enlace sencillo (*single linkage*)**: Considera la mínima distancia entre elementos de diferentes grupos. Es útil para detectar observaciones atípicas, pero puede generar clústeres alargados o no bien definidos.

- **Enlace completo (*complete linkage*)**: Utiliza la distancia máxima entre elementos de distintos grupos. Produce clústeres compactos, pero puede ser sensible a valores extremos.

- **Enlace promedio (*average linkage*)**: Calcula la media de todas las distancias entre pares de observaciones de dos grupos. Suele dar resultados equilibrados.

- **Enlace de centroides (*centroid linkage*)**: Mide la distancia entre los centros de masa (centroides) de los grupos. No se recomienda cuando las variables son cualitativas.

- **Método de Ward (*Ward linkage*)**: Minimiza la varianza total dentro de los grupos. Tiende a producir clústeres de tamaño similar y es una de las opciones más robustas para datos cuantitativos.

La aplicación de estos métodos permitirá la generación de dendrogramas, a partir de los cuales se identificará el número óptimo de clústeres mediante la inspección visual de los saltos en las distancias de fusión.

ii. Agrupación:

Para el análisis de agrupamiento, se ha seleccionado el algoritmo **k-medias** (*k-means*) como método principal de clasificación no supervisada. Esta elección se fundamenta en varias razones que lo hacen especialmente adecuado para los objetivos y características del presente estudio:

1. Conocimiento previo del número de grupos

Una de las principales premisas del algoritmo k-medias es que requiere especificar previamente el número de clústeres (*k*) que se desea identificar. En este caso, se sabe de antemano que existen tres especies de lirios distintas en el conjunto de datos, por lo que forzar al algoritmo a formar tres grupos es coherente con el contexto del problema y permite una evaluación directa respecto a la clasificación visual propuesta por el investigador.

2. Variables numéricas y de escalas similares

El algoritmo k-medias está diseñado para trabajar eficientemente con variables continuas, cuantitativas y en escalas comparables. Dado que las variables consideradas (longitud y ancho del sépalo y del pétalo) cumplen estas condiciones y han sido recolectadas con la

misma unidad de medida, no se requiere una transformación compleja para aplicar el algoritmo correctamente.

3. Simplicidad y eficiencia computacional

K-medias es un algoritmo sencillo de implementar y computacionalmente eficiente, incluso para conjuntos de datos moderadamente grandes. Esto lo hace ideal para una exploración inicial de la estructura de agrupamiento, sin necesidad de grandes recursos ni tiempos de procesamiento prolongados.

4. Generación de clústeres bien definidos

El algoritmo tiende a formar grupos compactos y esféricos alrededor de los centroides, lo que favorece la interpretación visual de los resultados y permite evaluar fácilmente el grado de separación entre las agrupaciones. Esto es especialmente útil cuando se desea comparar los clústeres obtenidos con una clasificación visual previa.

5. Posibilidad de validación y análisis estadístico posterior

Una vez definidos los grupos con k-medias, se pueden aplicar técnicas adicionales como el análisis de varianza (ANOVA) para evaluar si existen diferencias estadísticamente significativas entre los clústeres en relación con las variables morfológicas. Esta validación cuantitativa refuerza la robustez del análisis.

d. Elección del número de grupos

Matriz de proximidades

Distancia euclídea al cuadrado

Caso	1:Case 1	2:Case 2	3:Case 3	4:Case 4	5:Case 5	6:Case 6	7:Case 7	8:Case 8	9:Case 9	10:Case 10	11:Case 11	12:Case 12	13:Case 13	14:Cas
1:Case 1	,000	16,720	10,792	18,416	11,082	,303	17,459	14,462	9,350	,758	8,003	11,120	13,965	7
2:Case 2	16,720	,000	1,179	,674	,938	18,494	,936	3,185	1,688	21,769	1,764	,986	,345	4
3:Case 3	10,792	1,179	,000	2,247	,139	12,924	1,889	2,029	1,535	15,701	,461	,515	,756	2
4:Case 4	18,416	,674	2,247	,000	2,181	20,156	,156	6,411	1,811	22,868	2,620	2,739	,438	7
5:Case 5	11,082	,938	,139	2,181	,000	12,981	2,100	2,025	1,259	15,886	,366	,201	,702	1
6:Case 6	,303	18,494	12,924	20,156	12,981	,000	19,466	16,875	10,257	,279	9,492	12,739	15,715	8
7:Case 7	17,459	,936	1,889	,156	2,100	19,466	,000	6,195	1,970	22,015	2,460	2,867	,444	7
8:Case 8	14,462	3,185	2,029	6,411	2,025	16,875	6,195	,000	5,579	20,890	3,404	1,507	4,088	1
9:Case 9	9,350	1,688	1,535	1,811	1,259	10,257	1,970	5,579	,000	12,347	,557	1,428	,856	3
10:Case 10	,758	21,769	15,701	22,868	15,886	,279	22,015	20,890	12,347	,000	11,814	15,890	18,397	11
11:Case 11	8,003	1,764	,461	2,620	,366	9,492	2,460	3,404	,557	11,814	,000	,637	,968	1
12:Case 12	11,120	,986	,515	2,739	,201	12,739	2,867	1,507	1,428	15,890	,637	,000	1,117	1
13:Case 13	13,965	,345	,756	,438	,702	15,715	,444	4,088	,856	18,397	,968	1,117	,000	4
14:Case 14	7,294	4,417	2,088	7,496	1,926	8,829	7,300	1,390	3,814	11,920	1,993	1,336	4,434	
15:Case 15	13,555	,504	,625	,734	,475	15,374	,768	3,976	,893	18,064	,769	,959	,098	4
16:Case 16	12,195	,813	1,123	2,534	,693	13,479	2,882	1,940	1,377	16,732	1,116	,213	1,209	1
17:Case 17	18,866	1,122	2,617	,113	2,514	20,652	,273	7,519	2,000	23,196	2,862	3,298	,654	8
18:Case 18	,198	13,636	8,594	15,038	8,847	,515	14,222	12,371	6,979	1,150	6,025	8,868	11,136	5
19:Case 19	7,197	2,497	1,005	4,403	,709	8,465	4,410	2,328	1,360	11,147	,464	,454	2,145	
20:Case 20	14,877	1,983	3,333	,868	3,040	15,752	1,217	8,941	,888	17,747	2,456	3,510	1,141	7
21:Case 21	27,846	8,117	9,622	4,495	9,787	30,085	4,351	19,153	8,054	31,536	9,619	12,025	6,260	19

La matriz de proximidades es una herramienta fundamental en el análisis de agrupamiento, ya que muestra las distancias entre cada par de observaciones en el conjunto de datos. En este caso, la matriz revela patrones claros de similitud y diferencia entre los casos analizados. Por ejemplo, se observan distancias mínimas entre ciertos pares de casos, como el Caso 3 y el Caso 5 (distancia de 0.139), lo que sugiere una alta similitud y la posible pertenencia al mismo grupo. Por otro lado, el Caso 21 presenta distancias

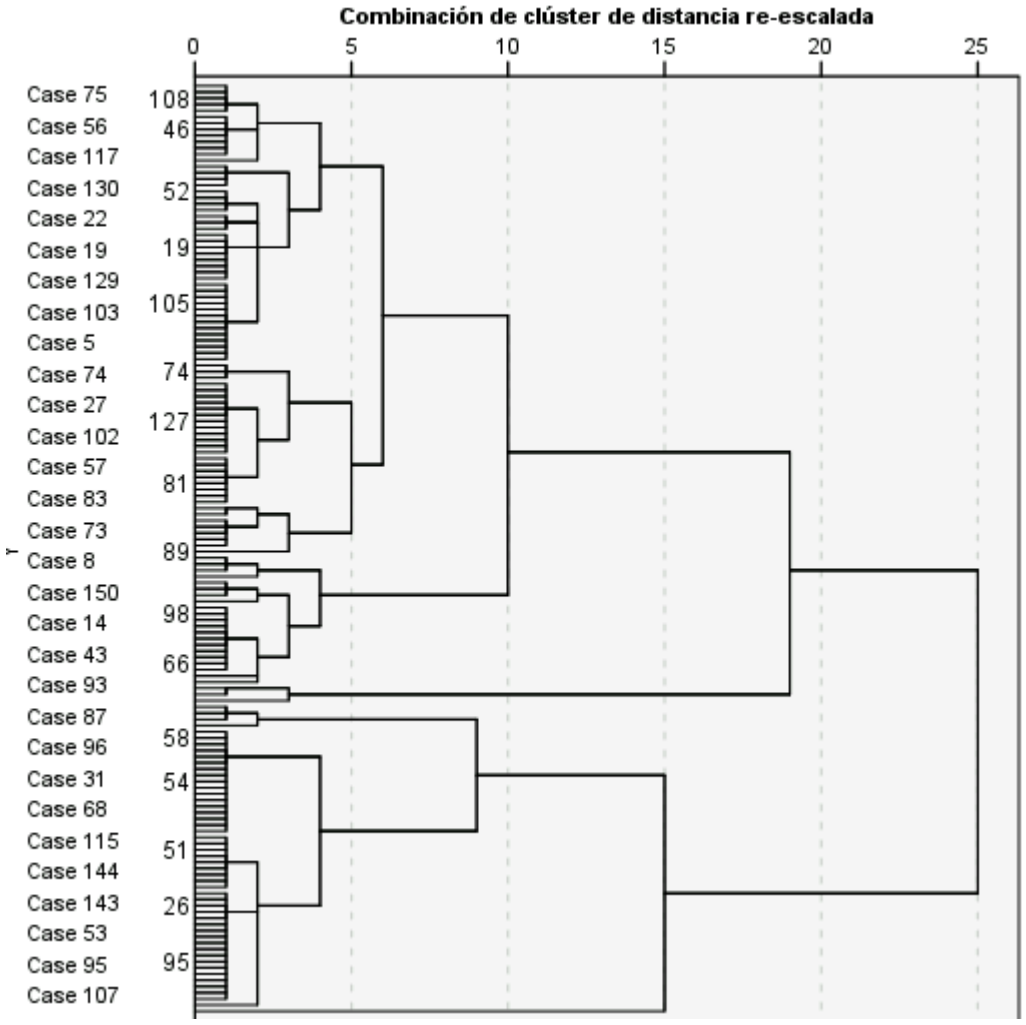
excepcionalmente altas con todos los demás casos (valores superiores a 15), lo que indica que podría tratarse de un valor atípico.

Esta matriz es esencial para identificar grupos naturales en los datos y validar la hipótesis inicial de que existen tres especies distintas de lirios. Además, permite detectar posibles anomalías en los datos, como valores extremos que podrían afectar el análisis.

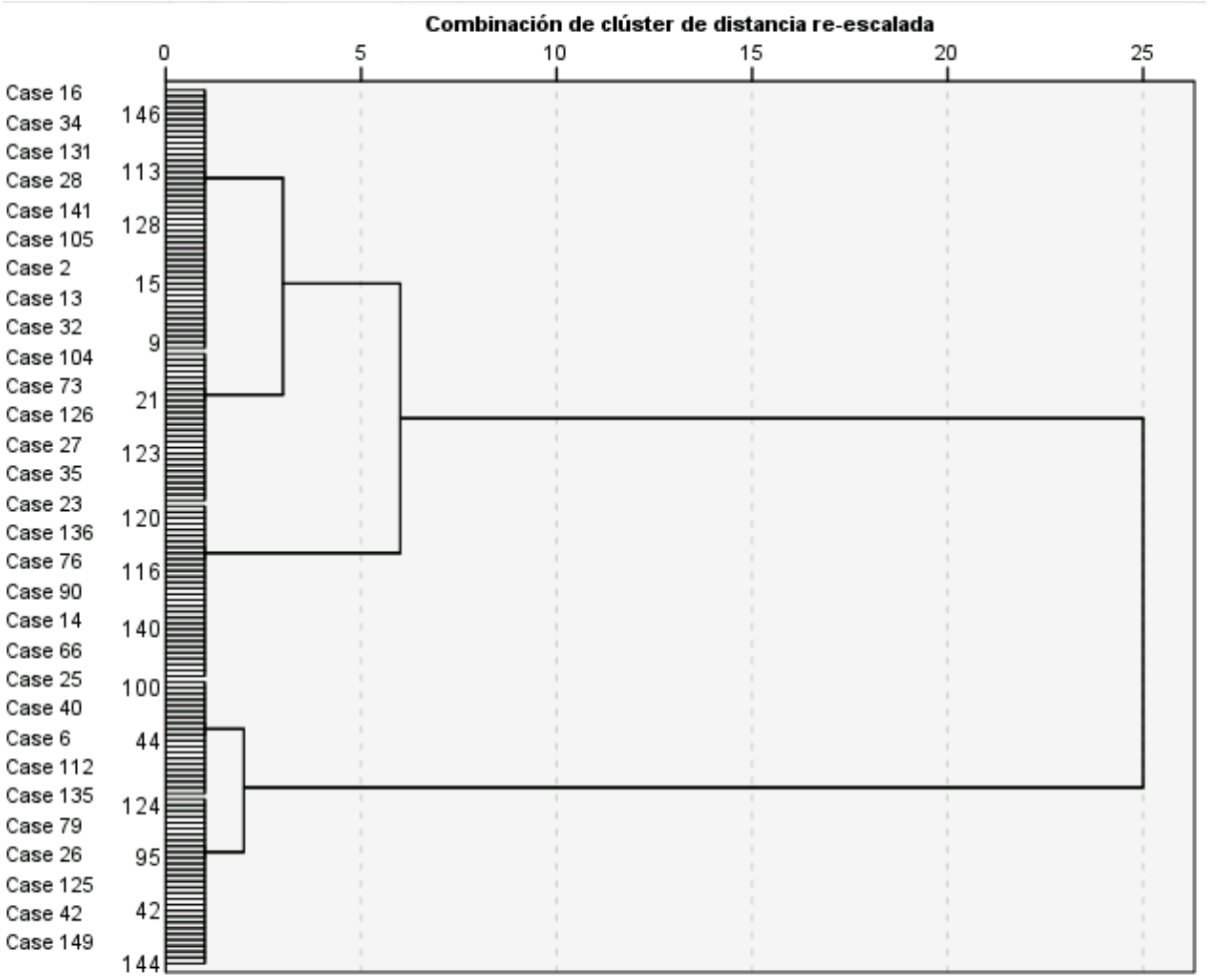
Historial de conglomeración

Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	16	75	,000	0	0	63
2	95	106	,015	0	0	19
3	115	149	,015	0	0	79
4	64	125	,017	0	0	11
5	88	112	,017	0	0	12
6	2	57	,017	0	0	92
7	52	82	,018	0	0	28
8	68	100	,018	0	0	15
9	51	139	,020	0	0	44
10	33	98	,020	0	0	45
11	64	107	,026	4	0	24
12	31	88	,026	0	5	47
13	54	63	,027	0	0	31
14	17	102	,027	0	0	58
15	47	68	,028	0	8	37
16	53	134	,029	0	0	57
17	37	72	,032	0	0	76
18	86	116	,035	0	0	72
19	36	95	,040	0	2	42
20	69	128	,045	0	0	40

El historial de conglomeración registra las fusiones sucesivas de clusters en un algoritmo jerárquico aglomerativo. Cada etapa del historial muestra qué clusters se combinan y el coeficiente de fusión correspondiente, que refleja la distancia a la que se unieron. En este análisis, las primeras etapas presentan coeficientes muy bajos (entre 0.000 y 0.020), lo que indica que se están fusionando casos muy similares. A medida que avanzan las etapas, los coeficientes aumentan (hasta 0.045), señalando que se están uniendo clusters más disímiles.



El dendrograma basado en distancia euclídea revela una estructura jerárquica clara en los datos. Algunos casos, como el 75 y el 56, se agrupan tempranamente debido a su similitud morfológica, mientras que otros, como el caso 102, requieren distancias mayores para fusionarse, lo que indica una posible separación más marcada entre grupos. Los saltos en las distancias de combinación sugieren la existencia de entre 2 y 5 clusters naturales. Aplicamos ahora la medida de Ward:



El dendrograma revela agrupaciones jerárquicas entre los lirios, con fusiones tempranas que indican alta similitud morfológica. Las últimas combinaciones ocurren a distancias mucho mayores, reflejando mayor disimilitud entre grupos. El rango de distancias sugiere una gradación clara en la similitud entre especímenes y la presencia de agrupamientos naturales. Se recomienda considerar 2 clusters principales descartando la hipótesis del biólogo de que hay 3.

Forzando a que solo sean 3, hacemos las K-Medias con 3 clústers.

e. K-Medias y misclassification

f.

K-means es un algoritmo de agrupamiento que divide los datos en k clústeres. Asigna cada punto al centro más cercano, recalcula los centros como el promedio de los puntos del grupo y repite el proceso hasta que los centros se estabilizan.

Centros de clústeres iniciales

	Clúster		
	1	2	3
SEPALLEN	6	8	5
SEPALWID	4	4	3
PETALLEN	1	7	5
PETALWID	0	2	2

La tabla de los centros iniciales indican que el algoritmo identificó en el inicio tres grupos morfológicos diferenciados entre los lirios:

- El *Cluster 1* agrupa individuos con sépalos de tamaño medio (aproximadamente 6 cm de largo por 4 cm de ancho) y pétalos pequeños (1 cm x 0 cm).

- El *Cluster 2* reúne flores con sépalos grandes (8 cm x 4 cm) y pétalos desarrollados (7 cm x 2 cm).

- El *Cluster 3* está compuesto por ejemplares con sépalos pequeños (5 cm x 3 cm) y pétalos de tamaño medio (5 cm x 2 cm).

Veamos como han ido iterando:

Historial de iteraciones[a]

	Cambiar en centros de clústeres		
Iteración	1	2	3
1	1,014	1,226	1,141
2	,000	,175	,121
3	,000	,070	,047
4	,000	,050	,033
5	,000	,000	,000

El algoritmo de k-medias alcanzó la estabilidad en cinco iteraciones. Durante la primera iteración se observaron ajustes significativos en la ubicación de los centroides (>1.0 en los tres grupos). En las siguientes iteraciones, los cambios fueron progresivamente menores (<0.2), hasta que en la quinta iteración se logró convergencia total (cambio = 0.000), lo que indica que los clusters se estabilizaron y no hubo más reasignaciones de individuos. Veamos cómo terminaron siendo los centros tras las iteraciones:

Centros de clústeres finales

	Clúster		
	1	2	3
SEPALLEN	5	7	6
SEPALWID	3	3	3
PETALLEN	1	6	4
PETALWID	0	2	1

Los centros finales del análisis de k-medias revelan tres grupos morfológicamente distintos entre sí:

- **Cluster 1 (Especie pequeña, Setosa):** Presenta sépalos de tamaño medio (5 cm de largo y 3 cm de ancho) y pétalos muy reducidos (1 cm de largo, 0 cm de ancho).
- **Cluster 2 (Especie grande, Virginic):** Agrupa flores con sépalos largos (7 cm x 3 cm) y pétalos desarrollados (6 cm x 2 cm).
- **Cluster 3 (Especie intermedia, Versicol):** Se caracteriza por sépalos medianos (6 cm x 3 cm) y pétalos de tamaño intermedio (4 cm x 1 cm).

Veamos ahora como distan unos clústers de otros.

Distancias entre centros de clústeres finales

Clúster	1	2	3
1		5,018	3,357
2	5,018		1,797
3	3,357	1,797	

La matriz de distancias entre los centros confirma una separación clara entre los grupos:

- La **mayor distancia** se observa entre el Cluster 1 y el Cluster 2 (5.018), lo que indica que son las especies más diferenciadas morfológicamente.
- La **menor distancia** se encuentra entre el Cluster 2 y el Cluster 3 (1.797), sugiriendo cierta similitud entre estos dos grupos.
- El **Cluster 1 destaca como el más diferenciado**, manteniendo distancias superiores a 3.3 respecto a los otros clusters.

Veamos ahora cómo de grande es cada clúster:

Número de casos en cada clúster

Clúster	1	50,000
	2	38,000
	3	62,000
Válidos		150,000
Perdidos		,000

Podemos observar que el clúster más grande es el 3 y el más pequeño es el dos ya que solo comprende 38 ejemplares.

					PE	
1	5	3	1	0	SETOSA	-1
2	6	3	6	2	VIRGINIC	1
3	7	3	5	2	VERSICOL	1
4	7	3	6	2	VIRGINIC	1
5	6	3	5	2	VIRGINIC	1
6	5	3	1	0	SETOSA	-1
7	7	3	5	2	VIRGINIC	1
8	6	2	5	2	VERSICOL	1
9	6	3	5	2	VERSICOL	0
10	5	4	1	0	SETOSA	-2
11	6	3	5	1	VERSICOL	0
12	6	3	5	2	VERSICOL	1
13	7	3	5	2	VIRGINIC	1

Clúster de pertenencia

Número del caso	Clúster	Distancia
1	1	,150
2	2	,561
3	3	,639
4	2	,389
5	3	,815
6	1	,415
7	2	,684
8	3	,637
9	3	,709
10	1	,640
11	3	,383
12	3	,734

Podemos observar clasos de misclassification como en la observación 5 por lo que rechazamos la hipótesis principal del biólogo.

Veamos ahora qué variables son más significativas:

g. ANOVA

Comenzamos viendo si cumple sus hipótesis:

Normalidad: aplicamos Kolmogorov Smirnov a cada variable

Prueba de Kolmogorov-Smirnov para una muestra

			SEPALLEN	SEPALWID	PETALLEN	PETALWID
N			150	150	150	150
Parámetros normales[a,b]	Media		5,84	3,06	3,76	1,20
	Desv. estándar		,828	,436	1,765	,762
Máximas diferencias extremas	Absoluta		,089	,106	,198	,173
	Positivo		,089	,106	,198	,173
	Negativo		-,049	-,068	-,148	-,119
Estadístico de prueba			,089	,106	,198	,173
Sig. asin. (bilateral)[c]			,006	<,001	<,001	<,001
Sig. Monte Carlo (bilateral)[d]	Sig.		,007	<,001	<,001	<,001
	Intervalo de confianza al 99%	Límite inferior	,005	,000	,000	,000
		Límite superior	,009	,001	,000	,000

Para todas las variables (SEPALLEN, SEPALWID, PETALLEN, PETALWID), los valores de significancia (Sig. bilateral) y Sig. Monte Carlo son inferiores a 0.01, indicando que se rechaza la hipótesis nula de normalidad en todos los casos al nivel de confianza del 99%. En este punto, en un caso real tendríamos que recurrir a estadística no paramétrica pero seguiremos ignorando este hecho. Aplicamos ANOVA y obtenemos:

ANOVA

	Clúster		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
SEPALLEN	36,888	2	,193	147	190,979	<,001
SEPALWID	6,399	2	,106	147	60,649	<,001
PETALLEN	219,109	2	,178	147	1233,690	<,001
PETALWID	38,864	2	,060	147	646,184	<,001

El análisis de varianza (ANOVA) confirma diferencias estadísticamente significativas ($p < 0.001$) entre los tres clusters para las cuatro variables morfológicas analizadas, lo que respalda la solidez del agrupamiento obtenido:

- **PETALLEN (Longitud del pétalo):** Presenta la mayor diferencia entre clusters ($F = 1233.69$), siendo la variable más útil para discriminar entre especies.
- **PETALWID (Ancho del pétalo):** También altamente discriminante, con un valor F de 646.18.
- **SEPALLEN (Longitud del sépalo):** Aunque con diferencias significativas ($F = 190.98$), su poder discriminante es menor en comparación con las variables de pétalo.
- **SEPALWID (Ancho del sépalo):** Es la variable con menor varianza entre clusters ($F = 60.65$), pero aun así aporta información relevante.

Estos resultados refuerzan la conclusión de que las variables asociadas a los pétalos son más efectivas para distinguir entre las especies de lirios.