Ciberseguridad e inteligencia de datos

Práctica de agrupamiento

Extracción de Conocimiento en Bases de Datos

JORGE CABRERA RODRÍGUEZ
10-24-2023

Introducción al Dataset

En la base de datos *mallCustomers.csv* se recogen los datos de 200 clientes de un centro comercial. El fichero consta de 5 variables, incluyendo el identificador de cada cliente (**CustomerID**).

Las variables en cuestión son las siguientes:

Variable	Descripción	Valores
Customer ID	identificador del cliente	Numérico
Gender	género del cliente	Male / Female
Age	edad del cliente	Numérico
Annual.income	ingresos anuales del cliente	Numérico
Spending.score	puntuación dada al cliente por el supermercado	Numérico

Ilustración 1. Variables del conjunto de datos

Ya que el ID de los usuarios es irrelevante a la hora de realizar el agrupamiento, se ha decidido eliminar esta columna del dataset.

Construcción de agrupamientos

Agrupamiento con k-medias *(k-means)*

Elección de subconjunto de datos

Lo primero que podemos hacer para visualizar los datos es representarlos mediante nubes de puntos, una por cada par de variables. De esta forma, podemos hacernos una idea de cómo se distribuyen los datos y si existen agrupamientos claros.

Para ello, vamos a graficar las siguientes variables:

- *Age* y *Annual.income*
- *Age* y *Spending.score*
- *Annual.income* y *Spending.score*

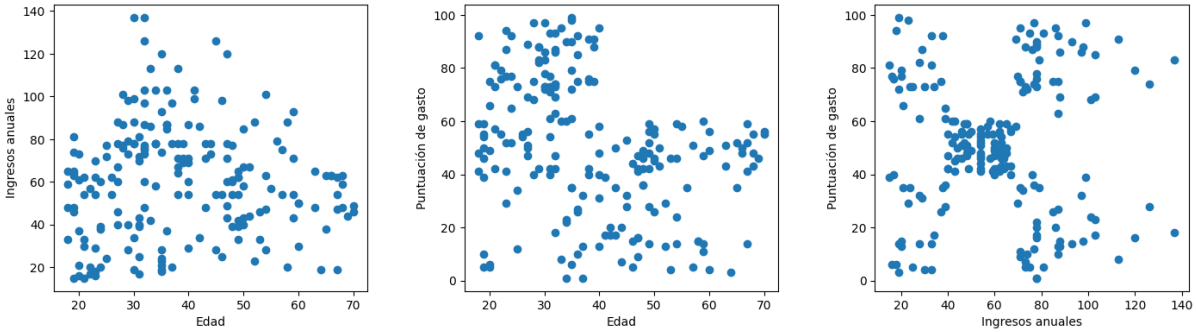

Ilustración 2. Nubes de puntos de cada par de variables

Podemos observar una clara agrupación entre las variables *Spending Score* y *Annual Income*, por lo que podría ser interesante comenzar por ese par de variables para realizar el agrupamiento.

Comenzaremos con un *clustering* de 5 grupos, para probar el funcionamiento del algoritmo de forma primitiva. Más adelante, calcularemos el número de grupos acorde al criterio del codo (*elbow method*).

Creación del modelo

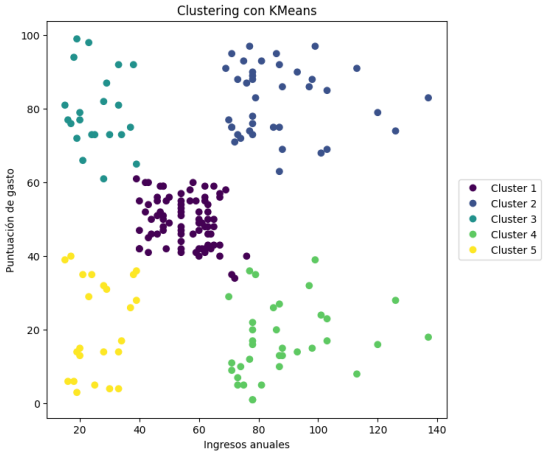

Ilustración 3. Agrupamiento k-means general

Podemos ver un agrupamiento con 5 grupos bastante bien distanciados y definidos. Ahora calcularemos el valor de información retenida de cada posible valor de *K* para ver cuál es el número óptimo de grupos.

Para ello, graficaremos la información retenida, y para aquel valor de *K* tal que el crecimiento a partir del mismo sea menor que un determinado umbral, determinaremos ese punto como valor óptimo de *K*. En este caso, el valor elegido para el umbral es de *0.2*.

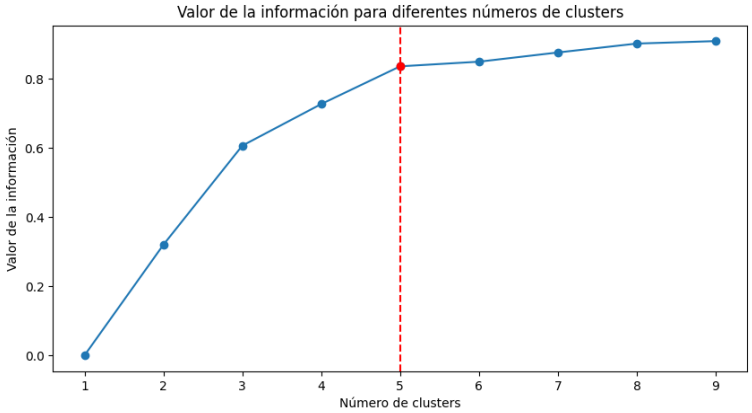

Ilustración 4. Cálculo del elbow method en k-means general

Como podemos ver, el valor óptimo de *K* es 5, que es el valor que habíamos elegido anteriormente.

Evaluación del modelo

La información retenida para este valor de *K* es la siguiente:

```
Información retenida para 5 clusters: 0.84
```

Ilustración 5. Información retenida para k-means general

Si quisieramos calcular otra medida de rendimiento del agrupamiento, podríamos aplicar el método de la silueta. Para ello, calcularemos el valor de la silueta para cada punto, y graficaremos dichas siluetas.

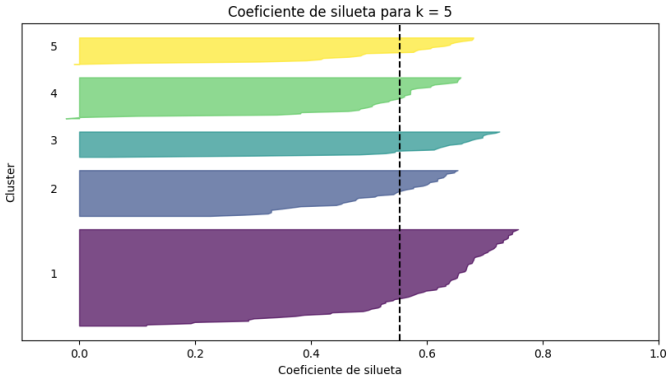

Ilustración 6. Coeficiente de silueta de k-means general

El coeficiente medio de silueta de este agrupamiento es de 0.55, lo que indica que el agrupamiento es adecuado. Podemos ver que el *cluster* 1 es el que parece tener una silueta más extrema: Cuenta con los puntos con silueta más alta (rozando el 0.8), pero también con los puntos con silueta más baja (rozando el 0.2).

El resto de *clusters* mantienen un valor de silueta más regular, con valores entre 0.3 y 0.7 aproximadamente.

Agrupamiento jerárquico

Elección del subconjunto de datos

Para el agrupamiento jerárquico, vamos a utilizar las mismas variables que para el agrupamiento con k-medias.

Creación del modelo

Para este modelo de agrupamiento jerárquico, vamos a comparar el subconjunto de variables *Annual Income* y *Spending Score*. En este caso calcularemos los diferentes coeficientes cofenéticos y dendrogramas para los métodos de:

- Enlace *single*
- Enlace *centroid*
- Enlace *ward*

Para poder realizar el _clustering_ jerárquico, primero debemos normalizar los datos. Posteriormente calcularemos las matrices de distancia y los dendrogramas para los diferentes métodos de enlace. De esta forma, podremos elegir aquel método cuyo coeficiente cofenético sea mayor.

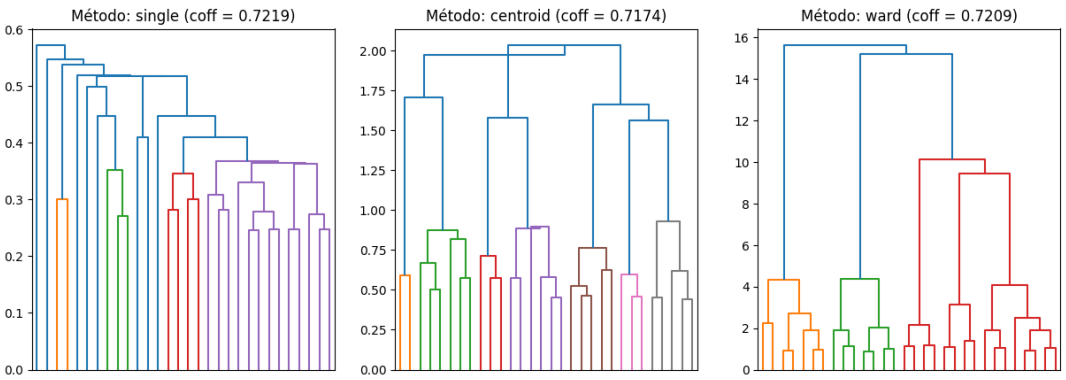

Ilustración 7. Cálculo cofenético para agrupamiento jerárquico

Podemos observar que el método *single* es el que obtiene un coeficiente cofenético mayor (*0.7219*), superando en una milésima al método *ward* (*0.7209*). Es, por tanto, el método que mejor conserva las distancias originales del subconjunto de datos.

Para comparar el rendimiento de los tres métodos, los graficaremos para ver los agrupamientos generados en el plano.

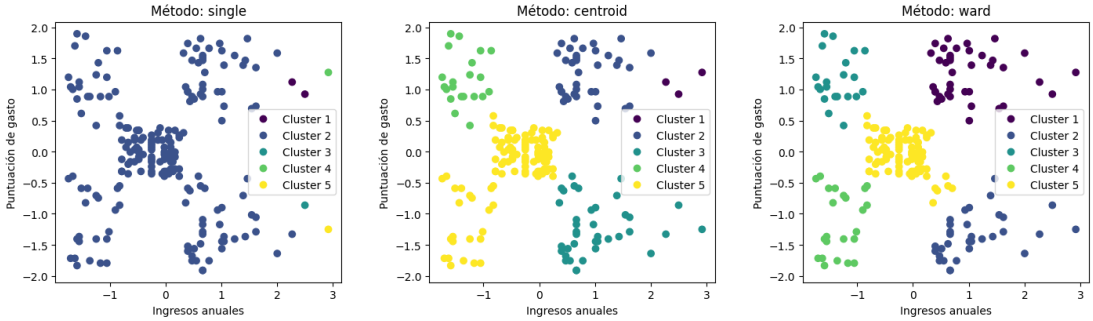

Ilustración 8. Diagramas de puntos de agrupamiento jerárquico

Podemos observar que, si bien el método *single* supera ligeramente el resto de métodos, los grupos generados son muy dispares y no parecen dividirse de forma correcta: Genera un macrogrupo con la mayoría de puntos, y el resto de grupos con un único punto.

Sin embargo, las divisiones de *centroid* y *ward* parecen más adecuadas, pues dividen cada grupo de forma más homogénea. Por tanto, se utilizará el segundo método con el coeficiente cofenético más alto: el método *ward*.

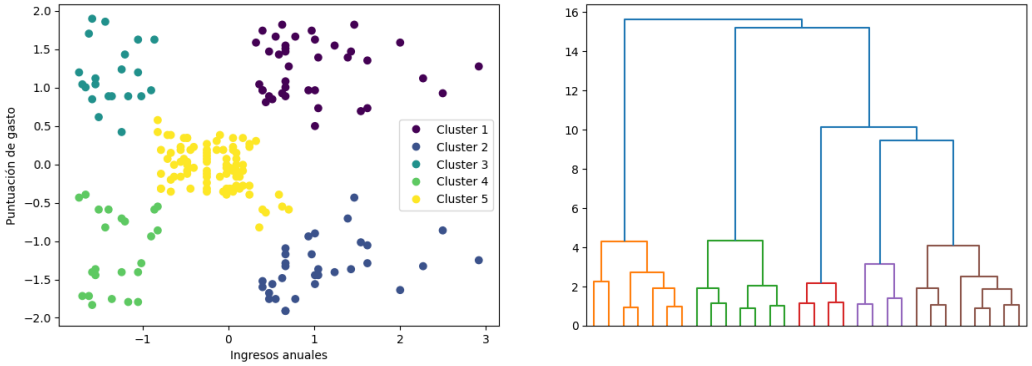

Ilustración 9. Agrupamiento jerárquico con método 'ward'

Construcción de agrupamientos considerando el género

Agrupamiento con k-medias *(k-means)*

Elección del conjunto de datos

En este paso vamos a dividir el conjunto de datos original según su campo *Genre*, para poder realizar un agrupamiento por separado para cada género y observar sus diferencias.

	Genre	Age	Annual Income	Spending Score
0	Male	19	15	39
1	Male	21	15	81
8	Male	64	19	3
10	Male	67	19	14
14	Male	37	20	13

Ilustración 10. Subconjunto de datos de género masculino

	Genre	Age	Annual Income	Spending Score
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40
5	Female	22	17	76
6	Female	35	18	6

Ilustración 11. Subconjunto de datos de género femenino

A continuación vamos a mostrar las mismas gráficas de puntos que realizamos en el primer agrupamiento por *k-means*, pero teniendo en cuenta los dos subconjuntos de datos.

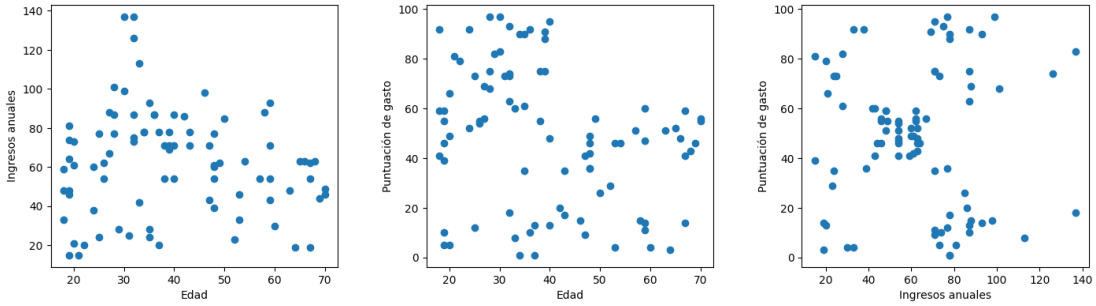

Ilustración 12. Nube de puntos de subconjunto de datos masculino

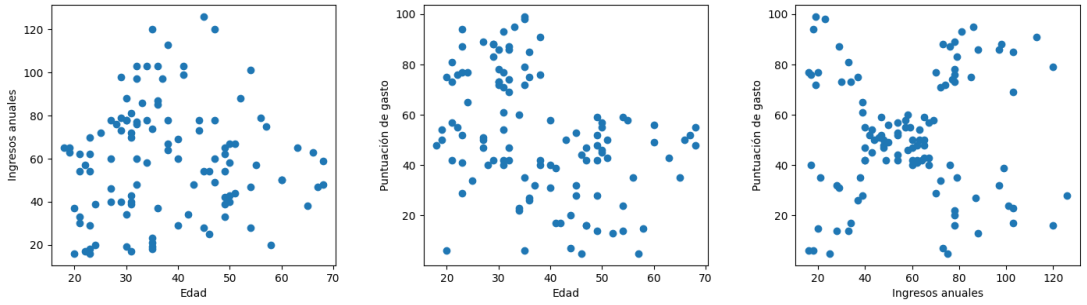

Ilustración 13. Nube de puntos de subconjunto de datos femenino

Observamos que las distribuciones de datos son similares a la del conjunto de datos original, por lo que podemos esperar un resultado similar al del agrupamiento anterior.

Creación del agrupamiento

Para este par de agrupamientos *k-means*, calcularemos directamente el valor de *K* óptimo para cada subconjunto de datos.

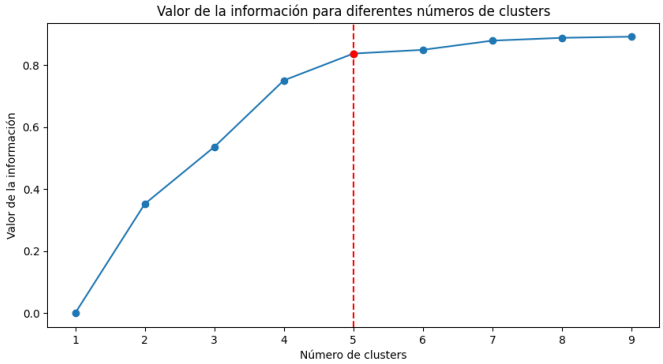

Ilustración 14. Cálculo del elbow method de subconjunto masculino

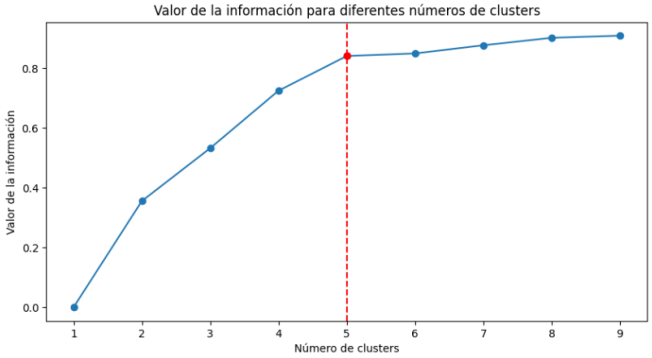

Ilustración 15. Cálculo del elbow method de subconjunto femenino

Podemos observar que en ambos subconjuntos el *K-valor* en el cual la información retenida parece estancarse (el codo en cuestión) es *K=6* en ambos subconjuntos.

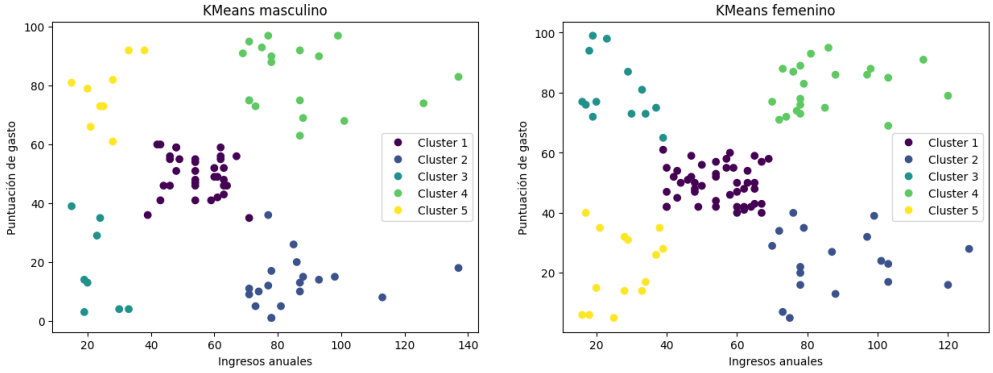

Ilustración 16. Comparativa entre k-means masculino (izquierda) y femenino (derecha)

Para evaluar ambos agrupamientos, graficaremos sus coeficientes de silueta como en el primer agrupamiento K-means.

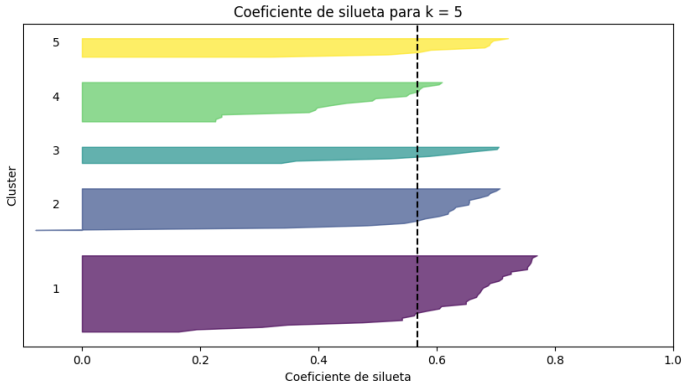

Ilustración 17. Coeficiente de silueta de k-means masculino

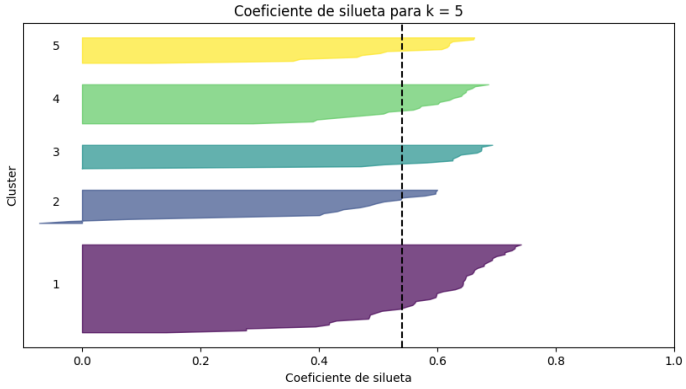

Ilustración 18. Coeficiente de silueta de k-means femenino

Se puede observar una distribución de siluetas bastante similar a la del primer agrupamiento k-means. En el caso de los subconjuntos, los valores medios de silueta son los siguientes:

- 0.56 para el subconjunto masculino
- 0.54 para el subconjunto femenino

Estos valores son muy similares al valor de 0.55 del primer agrupamiento k-means, además de que sus siluetas son muy similares entre sí.

Evaluación del agrupamiento

Agrupamiento jerárquico

Elección del subconjunto de datos

Para el agrupamiento jerárquico, vamos a utilizar las mismas variables que para el agrupamiento con k-medias.

Creación del modelo

Realizaremos la misma operación que hicimos en el agrupamiento jerárquico anterior, pero esta vez para cada subconjunto de datos.

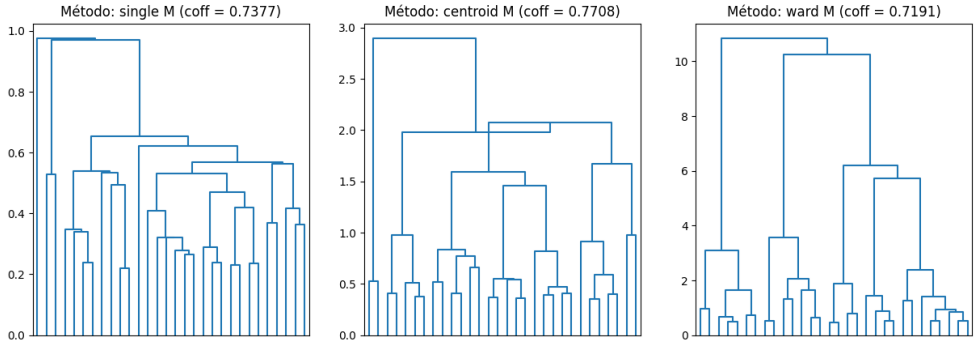

Ilustración 19. Cálculo cofenético para agrupamiento jerárquico masculino

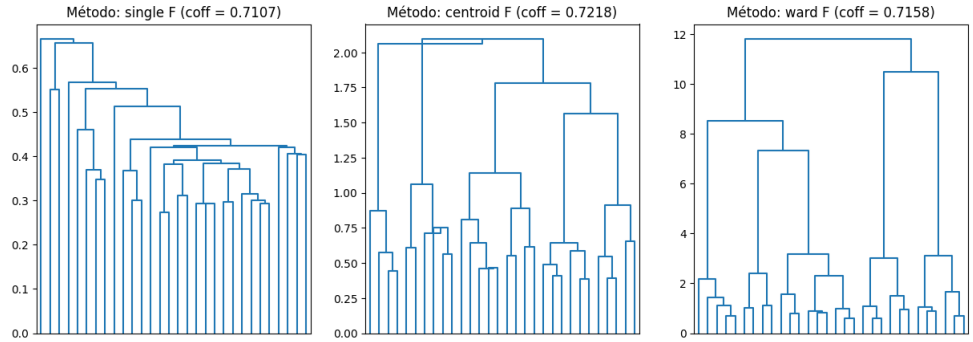

Ilustración 20. Cálculo cofenético para agrupamiento jerárquico femenino

Evaluación del modelo

Como podemos observar en los gráficos anteriores, tanto el subconjunto de hombres como el de mujeres tienen su coeficiente cofenético máximo para el método de enlace *centroid*, siendo estos de:

- *0.77* para los hombres
- *0.72* para las mujeres

Como hicimos en el primer agrupamiento jerárquico, graficaremos los tres métodos para cada subconjunto de datos para observar cómo se comportan. Comenzaremos por el agrupamiento para el subconjunto masculino:

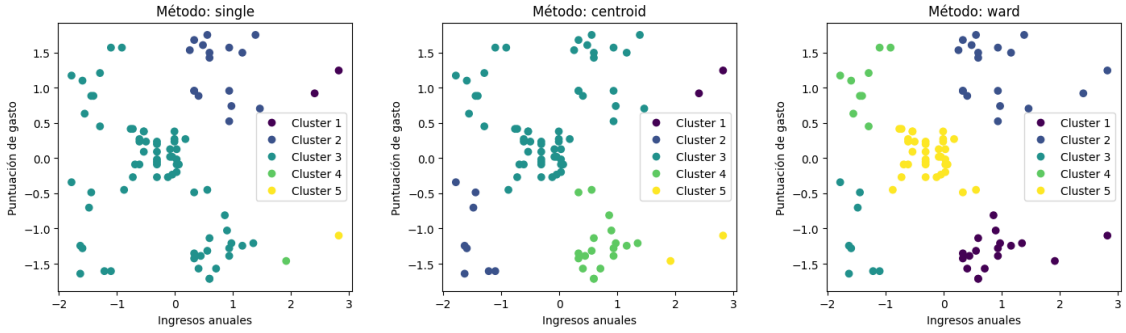

Ilustración 21. Diagramas de puntos de agrupamiento jerárquico masculino

Vemos que el agrupamiento por *centroid* da un agrupamiento aparentemente correcto, pero con ciertos grupos que parecen estar mal definidos. Por otro lado, el agrupamiento por *ward* parece ser el que mejor define los grupos, pero con un coeficiente cofenético menor.

Como en el primer agrupamiento jerárquico, nos quedaremos con el método *ward*.

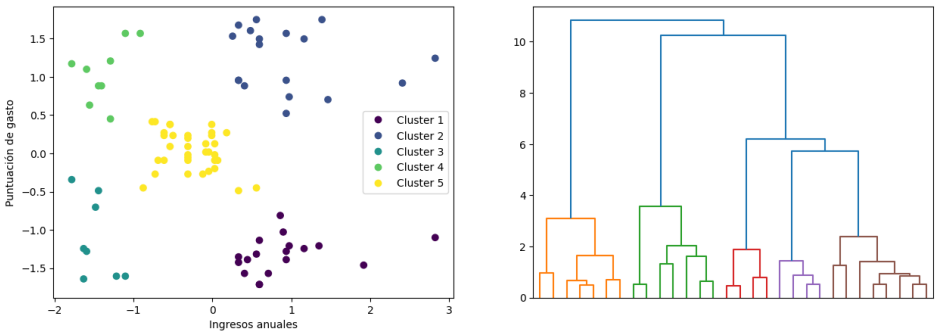

Ilustración 22. Agrupamiento jerárquico mascilino con método 'ward'

Continuamos el análisis con el subconjunto de datos femenino:

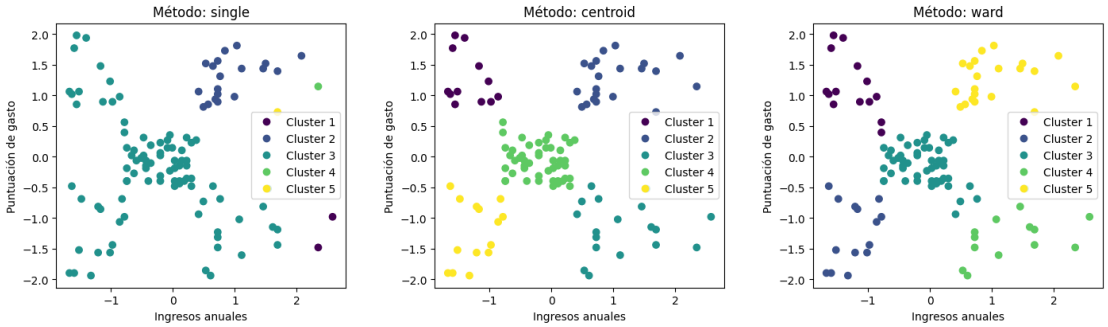

Ilustración 23. Diagramas de puntos de agrupamiento jerárquico femenino

Vemos que al contrario que en el caso anterior y en el caso del agrupamiento genérico, el método con mayor coeficiente cofenético coincide con el método que en la práctica produce los grupos más correctos: El *centroid*. Por tanto, nos quedaremos con este método.

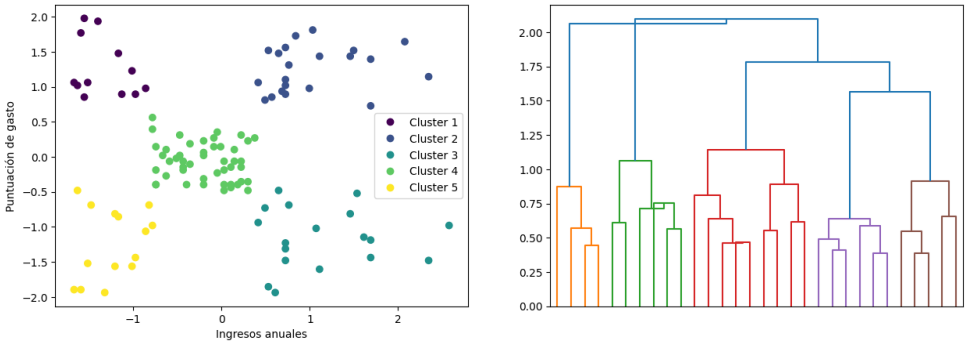

Ilustración 24. Agrupamiento jerárquico femenino con método 'centroid'