Vamos a empezar aplicando técnicas de clustering a la base de datos espalda.csv

Sigue los pasos de la hoja de trabajo y trata de ser crítico con los resultados que obtienes.

La práctica hace el maestro ☺

IMPORTA LOS DATOS ESPALDA.CSV

Lee lo datos que has descargado junto a esta hoja de trabajo.

Una vez leída.

Escoge las siguientes características para aplicar las recetas de código del clustering.

Escoge estas variables:

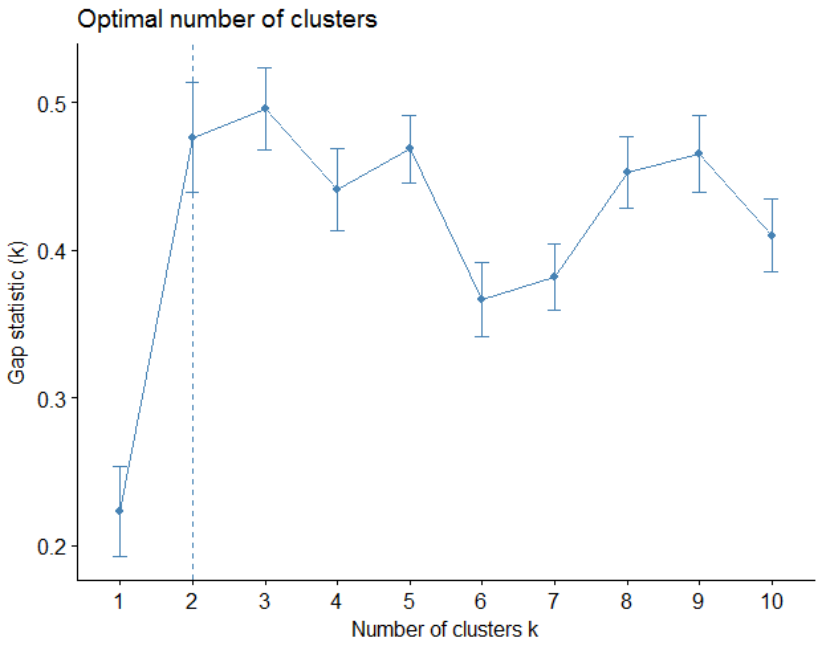
- Edad
- Peso
- Altura
- ODI mes0
- ODI mes 1

CALCULA EL NÚMERO ÓPTIMO DE CLUSTERS

Utiliza una de las recetas de código de R para saber el número de clusters óptimo.

La línea de código es: fviz_nbclust(my_data, kmeans, method = "gap_stat")

En el caso de la figura de abajo es 2.



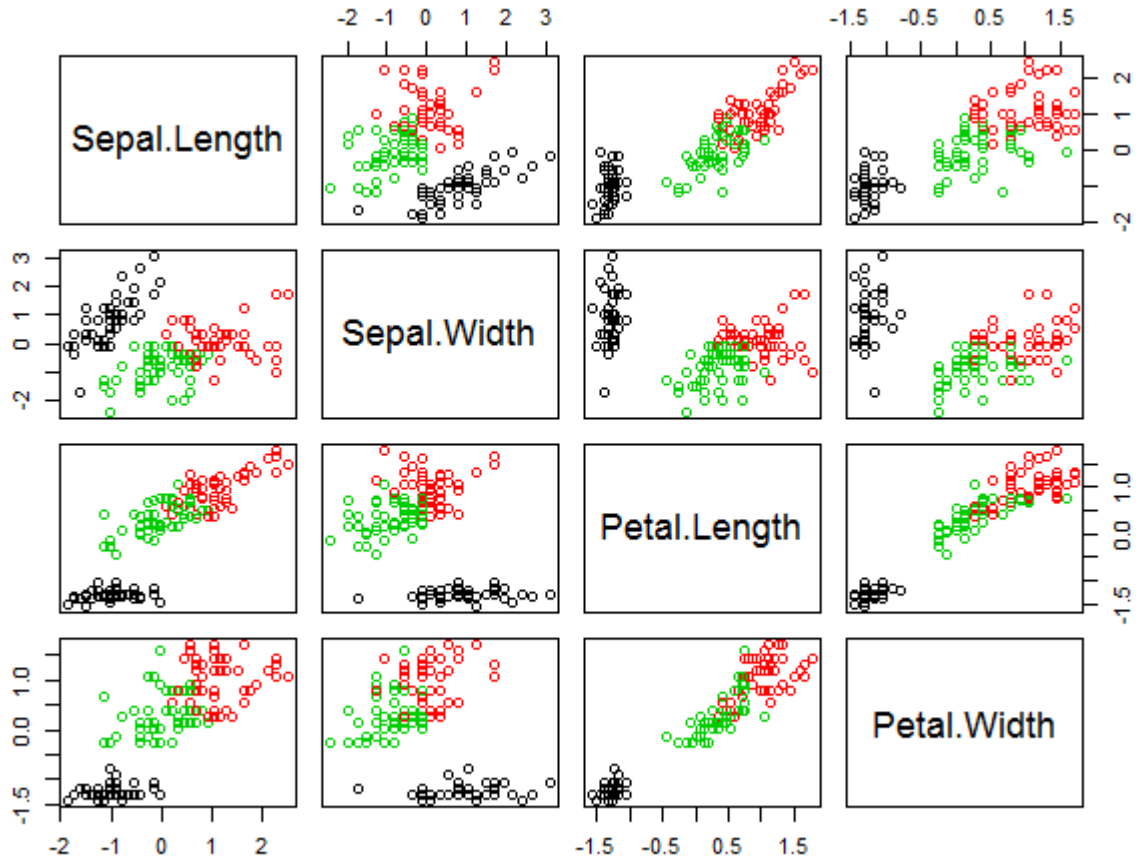
APLICA LOS ALGORTIMOS DE CLUSTERING

Aplica y compara los algoritmos de clustering que has visto:

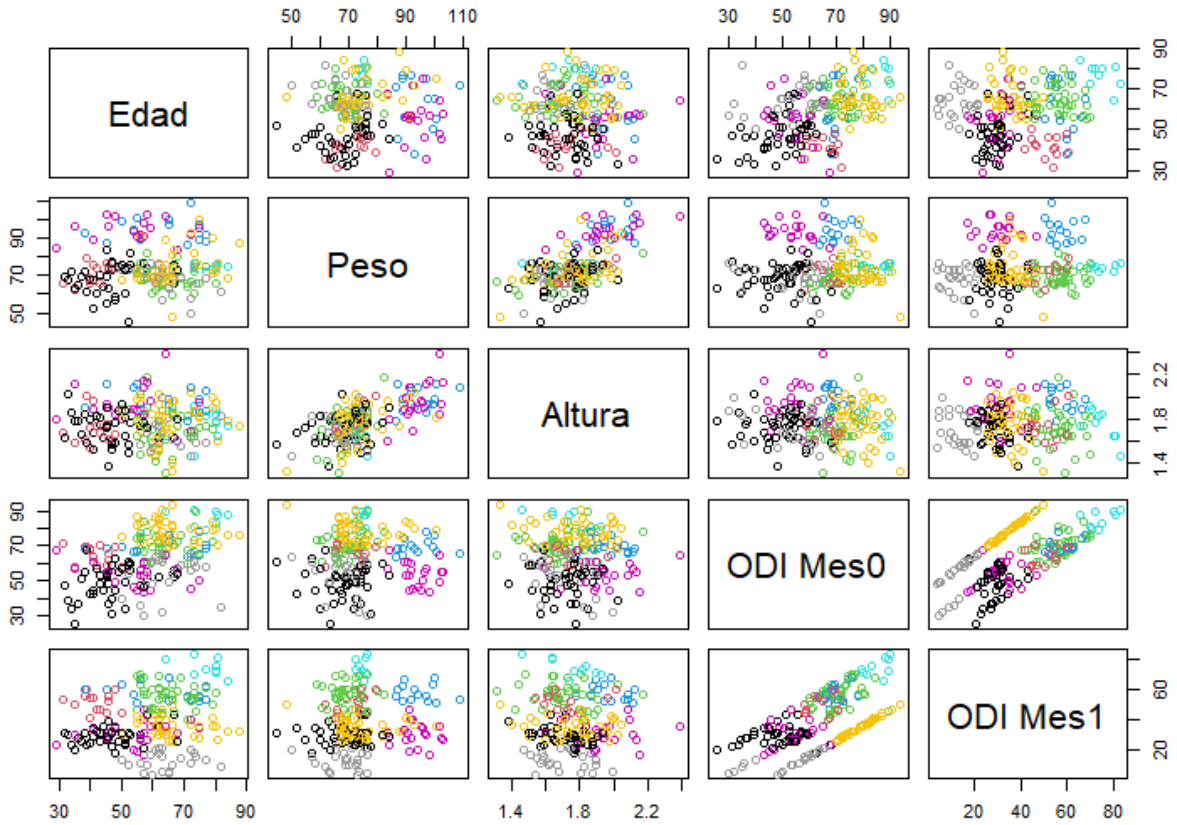
- Hierarchical
- K-means
- GMM – gaussian mixture models

Dibuja los clusters en un matrixplot para las tres técnicas:

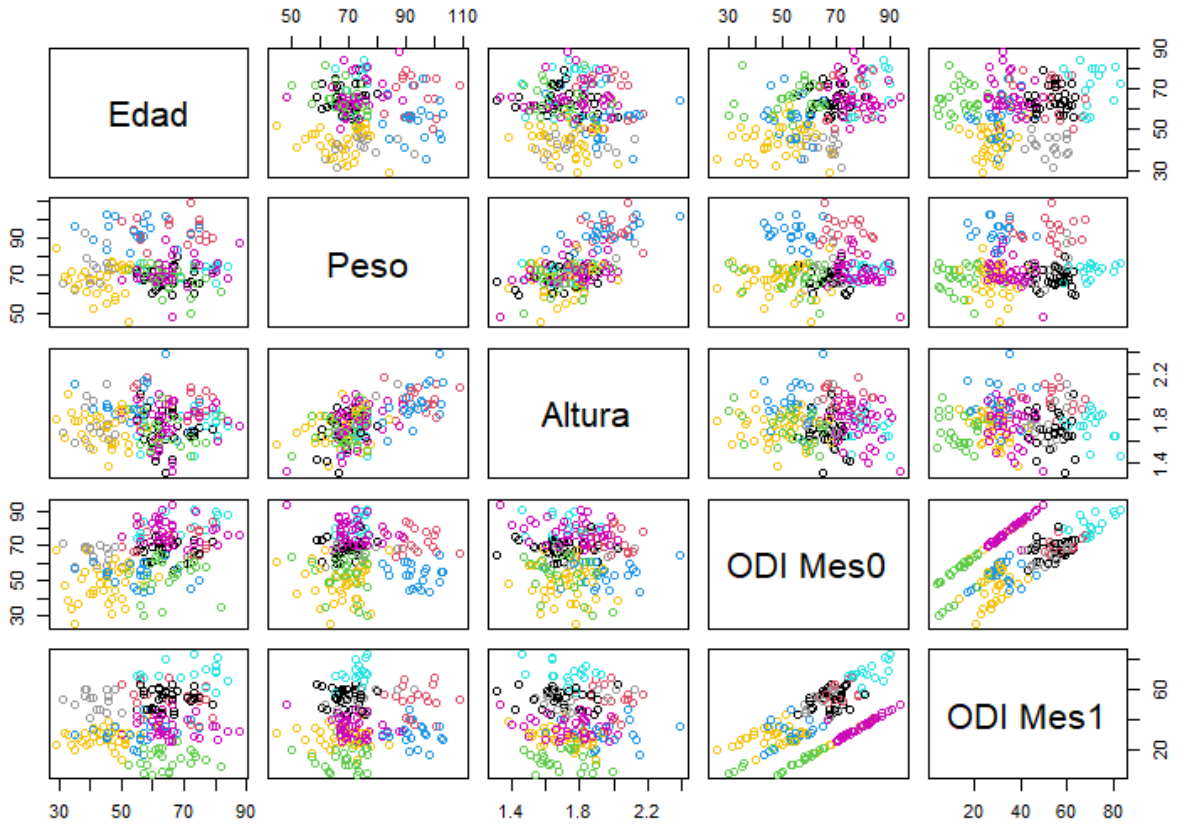
Ejemplo:



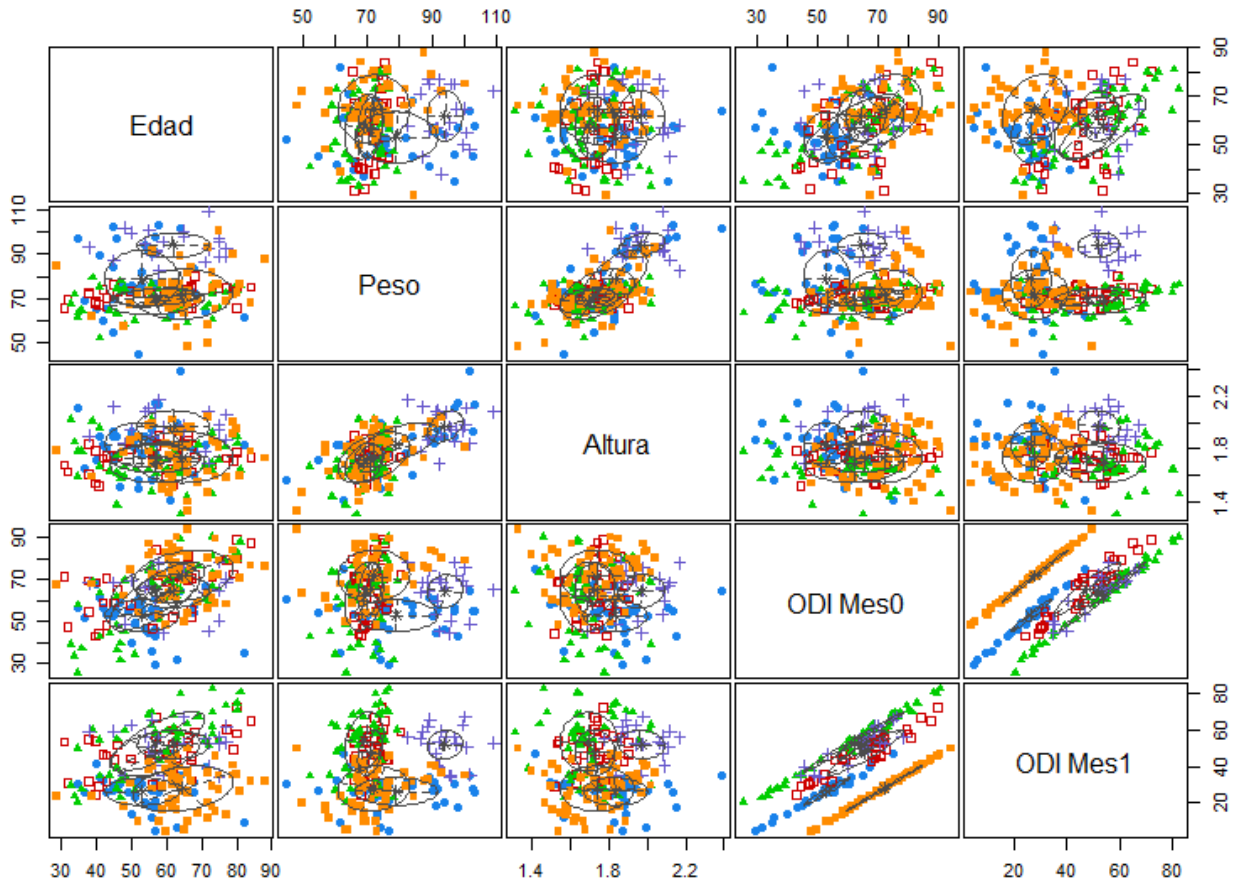
Jerárquico:
8 clusters



K-Means
8 clusters



GMM


Edad
Peso
Altura
ODI Mes0
ODI Mes1
50 70 90 110
30 50 70 90
50 70 90 110
30 50 70 90
50 70 90
1.4 1.8 2.2
30 50 70 90
20 40 60 80
30 40 50 60 70 80 90
1.4 1.8 2.2
20 40 60 80

COMPARA LOS RESULTADOS

¿Describe las diferencias que ves en cada una de las técnicas?

Se ven pocas diferencias entre el Jerárquico y el Kmeans, puede que porque el número de clusters sea alto (8) y se mezclan mucho visualmente. En cambio, en el GMM, que entiendo que utilizó 5 clusters, se pueden apreciar más la diferencia de los grupos.

Es difícil ver las diferencias con tantos grupos. Quizá algo que se ve muy claro es que en las distribuciones según Diff ODI 0 y 1, en donde se aprecian claramente los grupos, el Jerárquico y el Kmeans dividieron los grupos lineales superiores e inferiores en 2 partes, los más cercanos a cero y los más lejanos, en cambio el GMM como usa menos clusters solo los dividió en las líneas claras que se aprecian.

¿Según tu opinión que técnicas está separando mejor los datos?

Creo que los 3 modelos separan bien los grupos, aunque basados más que nada en las distribuciones comparativas de DIff-ODI 0 y 1, porque en estas distribuciones los grupos se ven muy claros. En este sentido creo que el GMM los separa en grupos más coherentes, por lo que me decanto por este método..