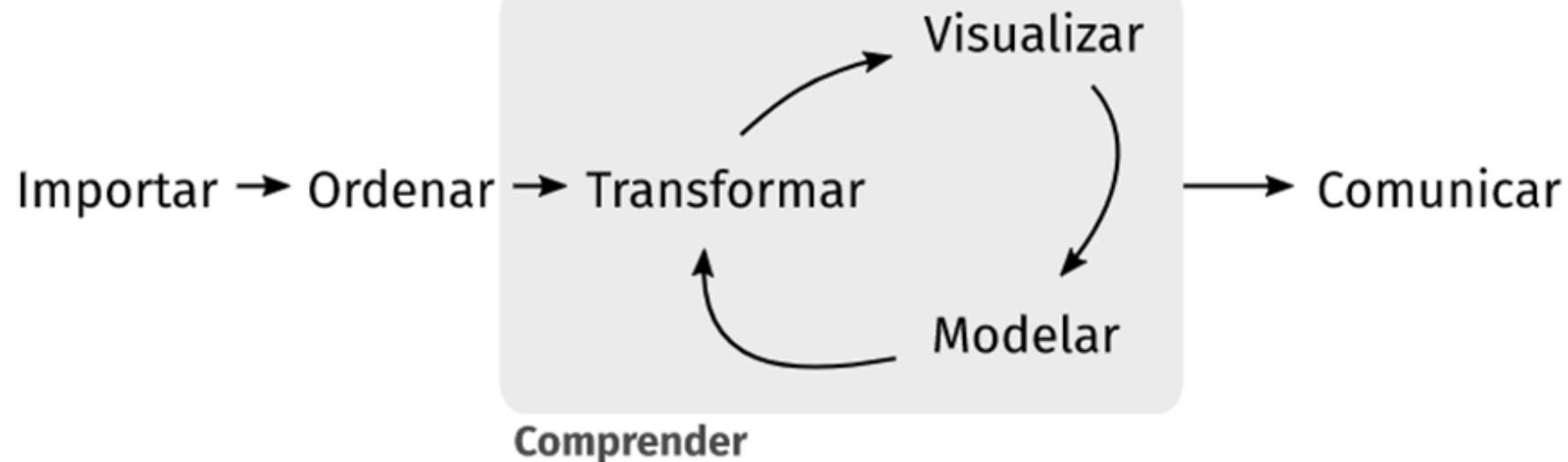


Analítica de datos aplicada a estudios sobre desarrollo

Giovany Babativa, PhD

Proceso de analítica



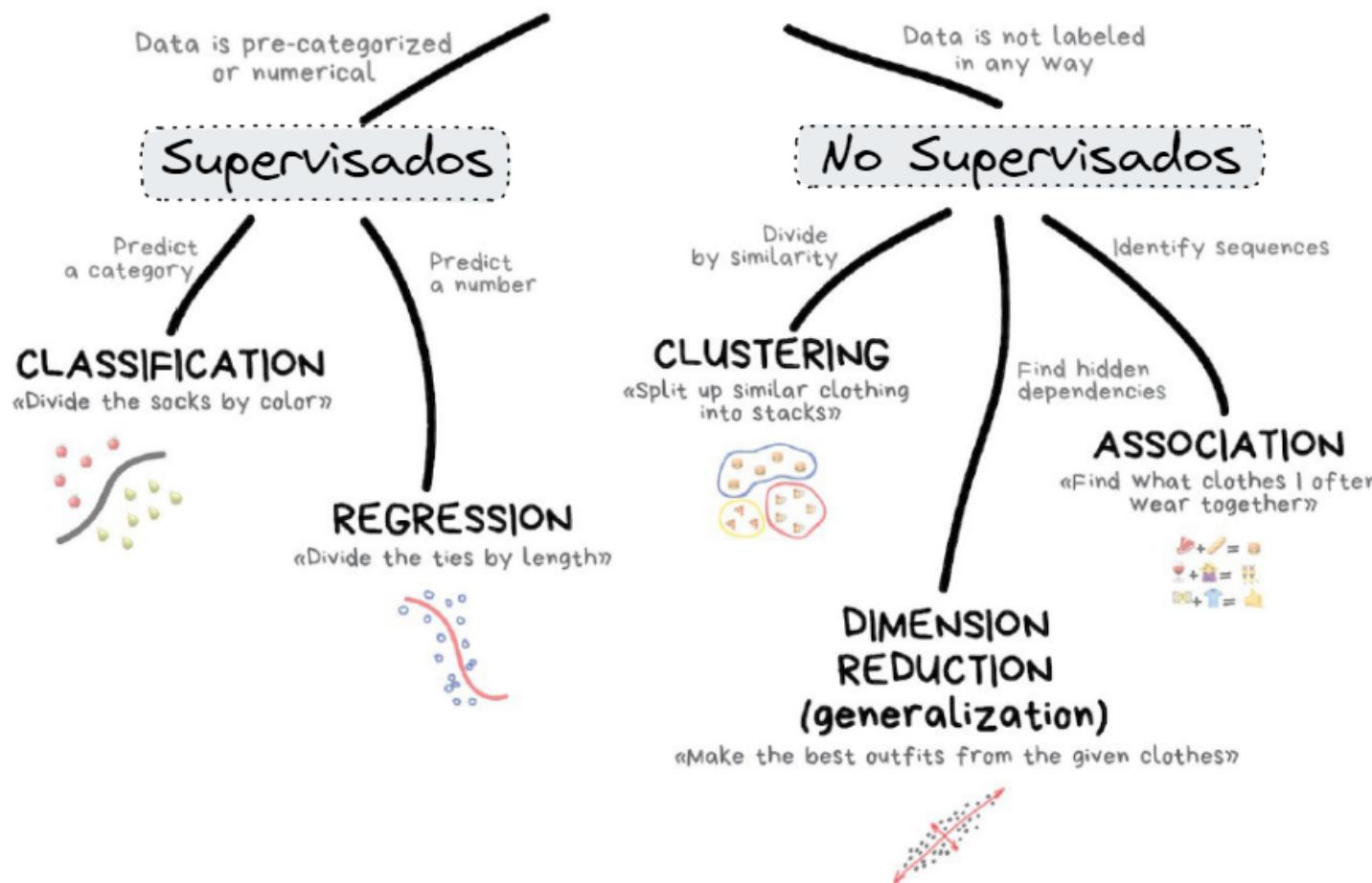
Programar

Wickham, H. y otros (2023)

MÉTODOS MULTIVARIANTES

Modelos de analítica

Modelos Multivariantes



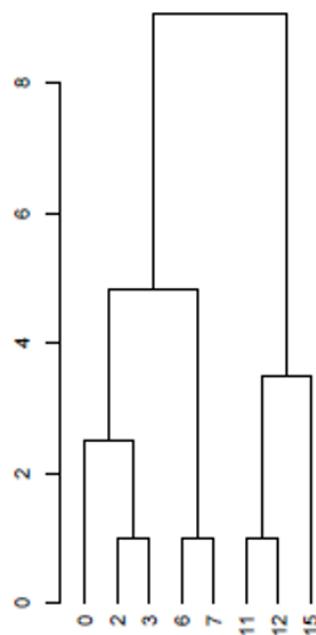
Fuente: Machine Learning for Everyone

ANÁLISIS DE CONGLOMERADOS

Análisis clúster

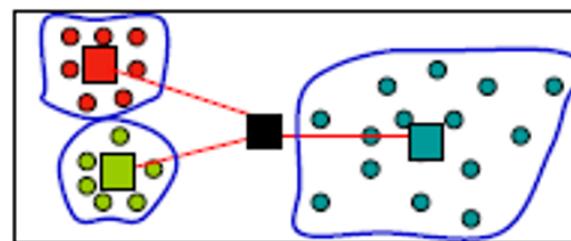
Métodos para realizar la agrupación de individuos con base en la similaridad que tienen en un vector de variables $\mathbf{x}' = (x_1, \dots, x_p)$.

Jerárquicos

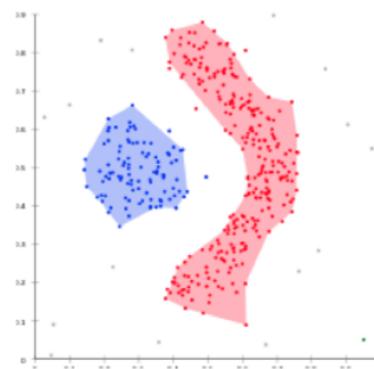


Costosos computacionalmente

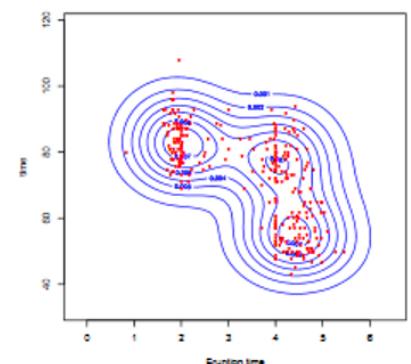
No Jerárquicos o de partición



Density Based Clustering



Model Based Clustering



Qué es el análisis de clúster

Es una técnica para combinar observaciones en grupos o clúster de forma que:

1. Cada grupo o clúster sea lo más homogéneo con respecto a las características de análisis. Es decir las observaciones dentro de cada grupo deben ser similares.
2. Cada grupo debe diferenciarse de los otros grupos respecto a las características que se midieron.

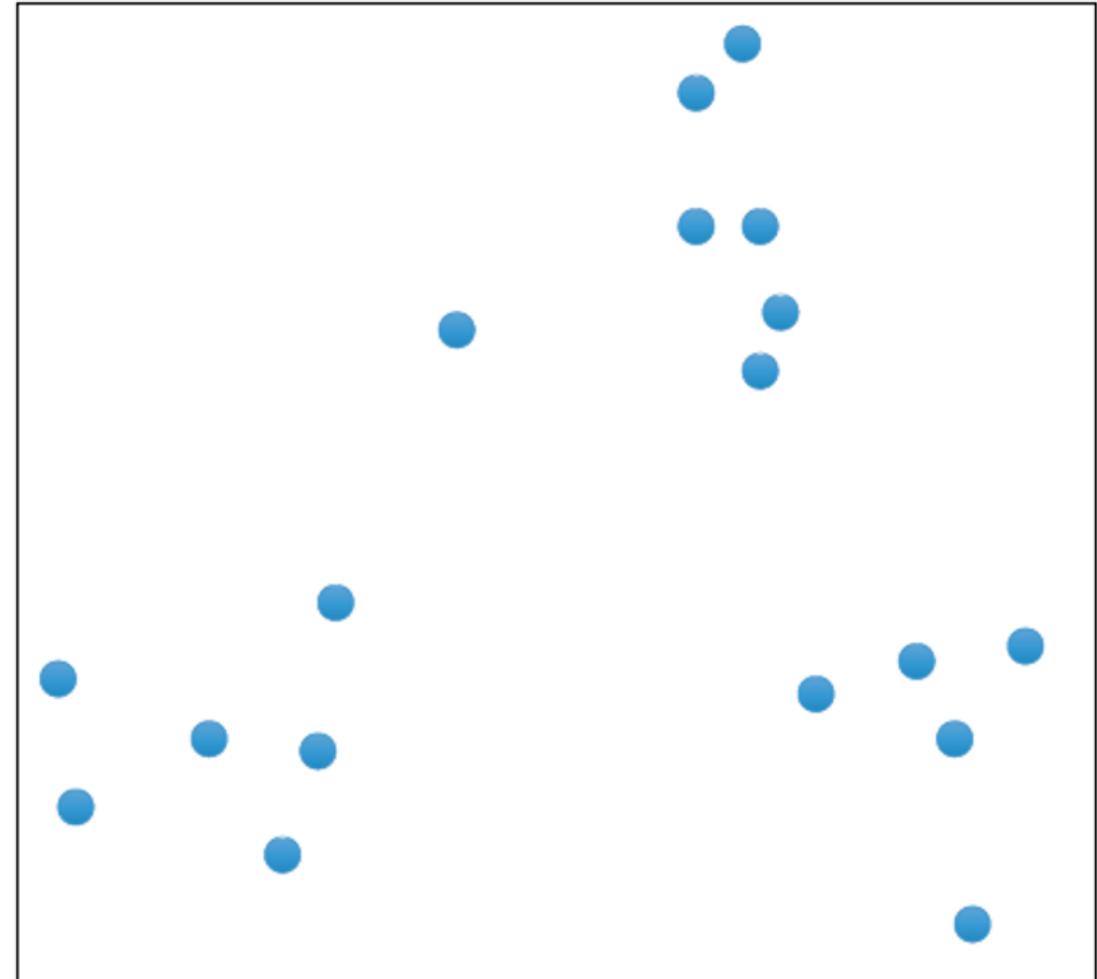
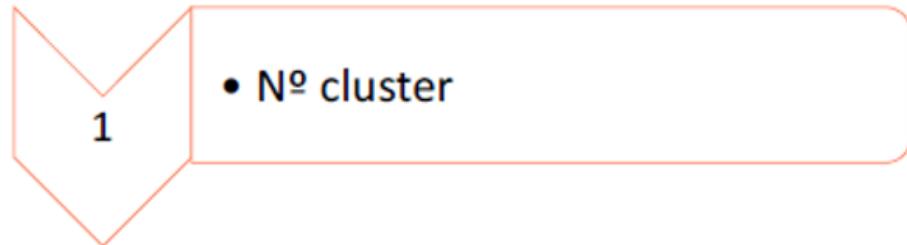
Tipos de Análisis de clúster

1. Jerárquicos: Consisten en agrupar los individuos o grupos más similares a partir de algún criterio de aglomeración.
2. No jerárquicos: Consiste en dividir el conjunto de objetos o individuos en un número de grupos prefijado y aplicar un algoritmo para obtener las agrupaciones.

Métodos no jerárquicos

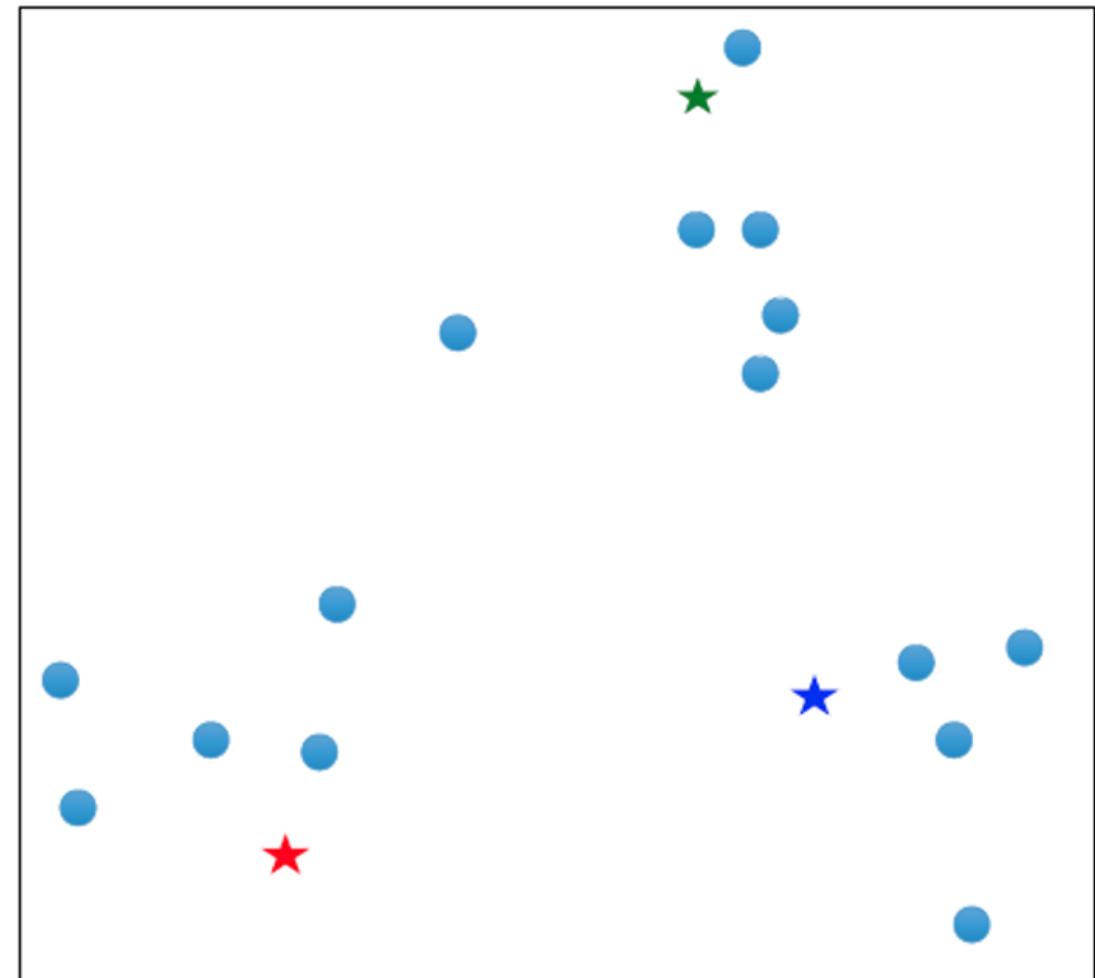
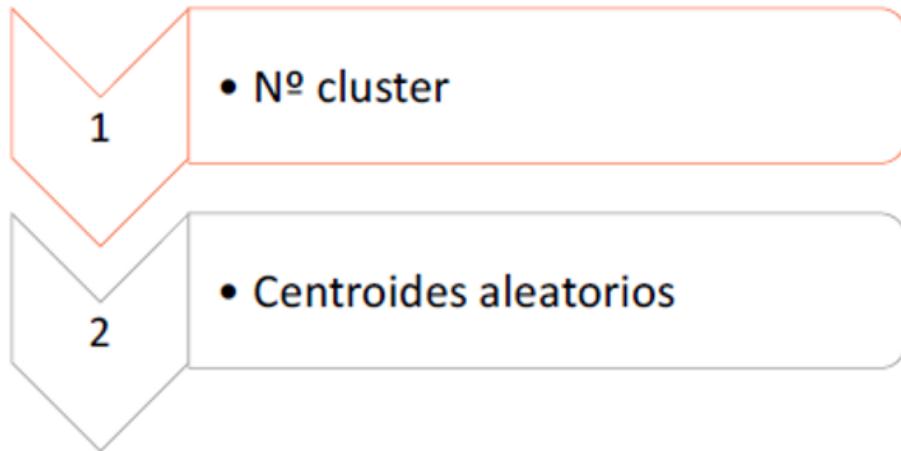
Algoritmo de las K -medias

Forgy (1965)



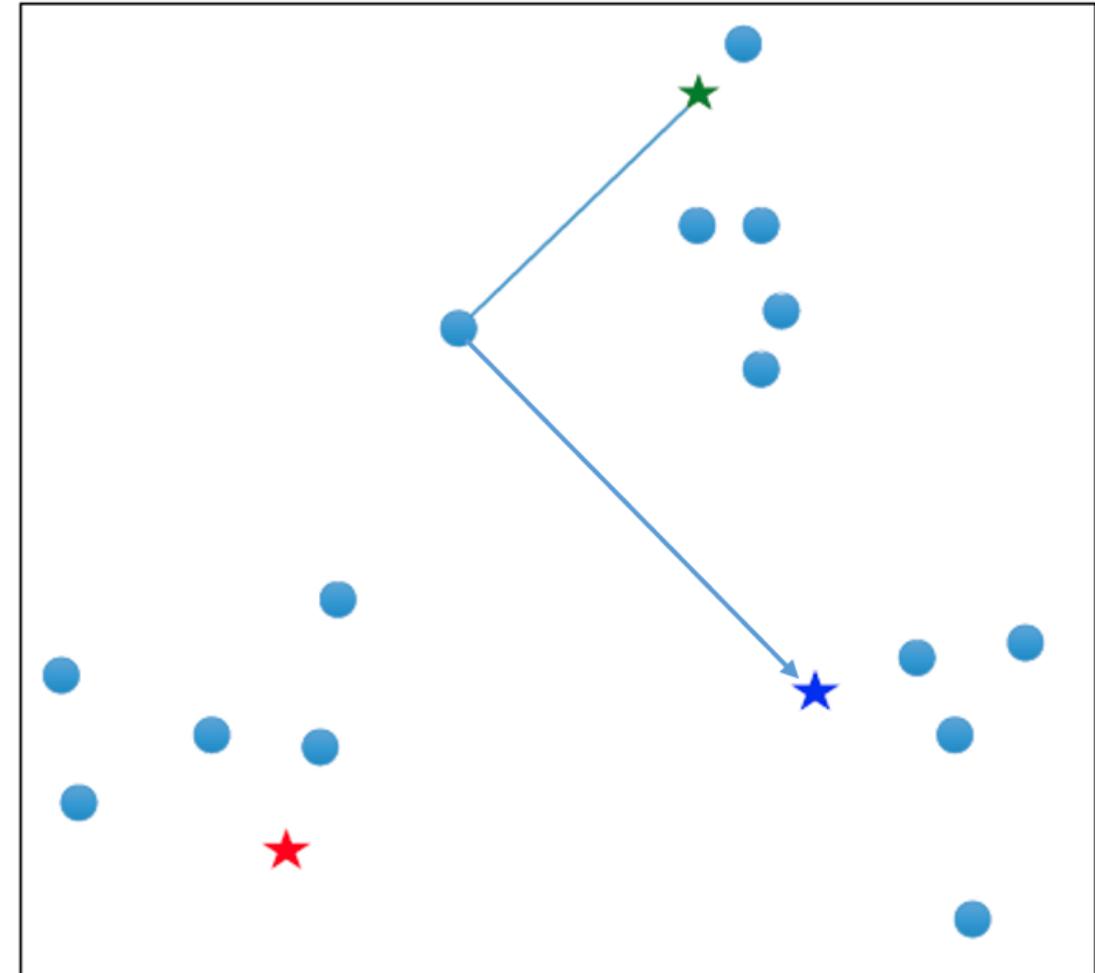
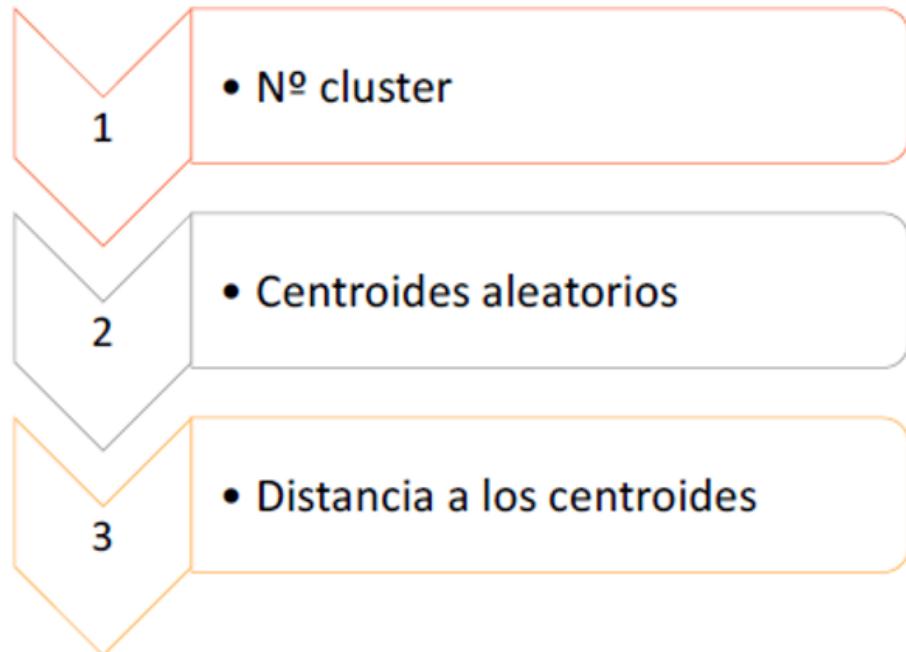
Algoritmo de las K -medias

Forgy (1965)



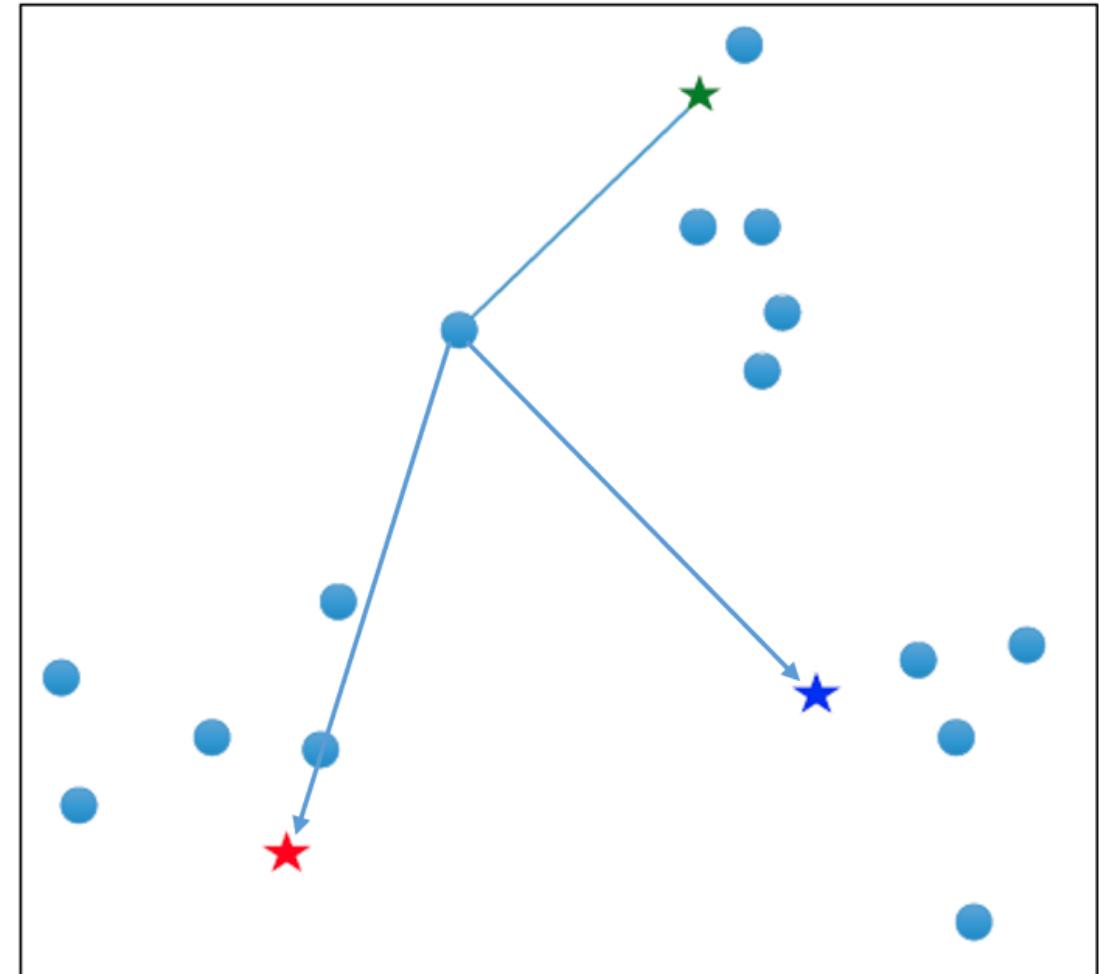
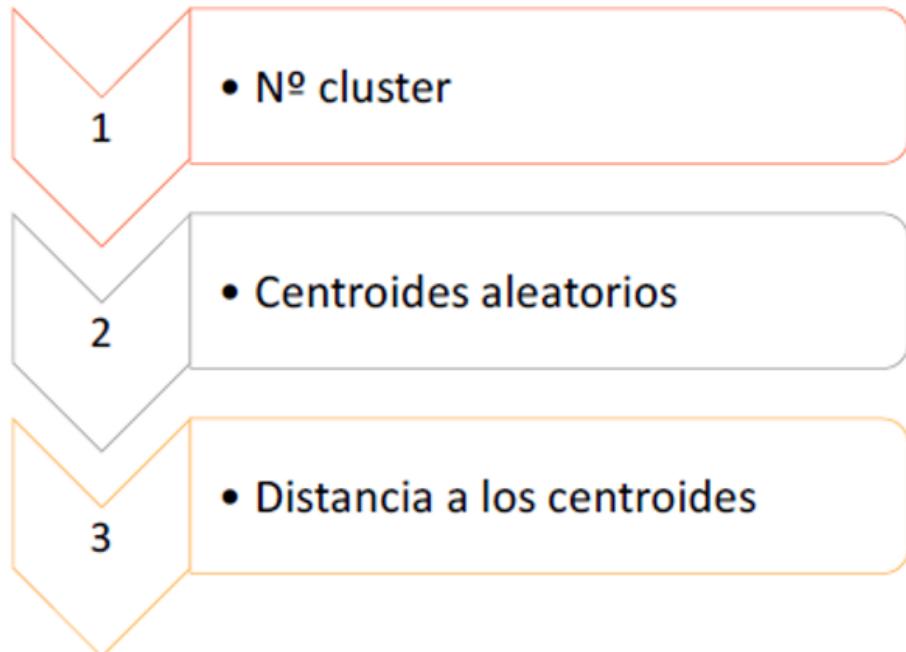
Algoritmo de las K -medias

Forgy (1965)



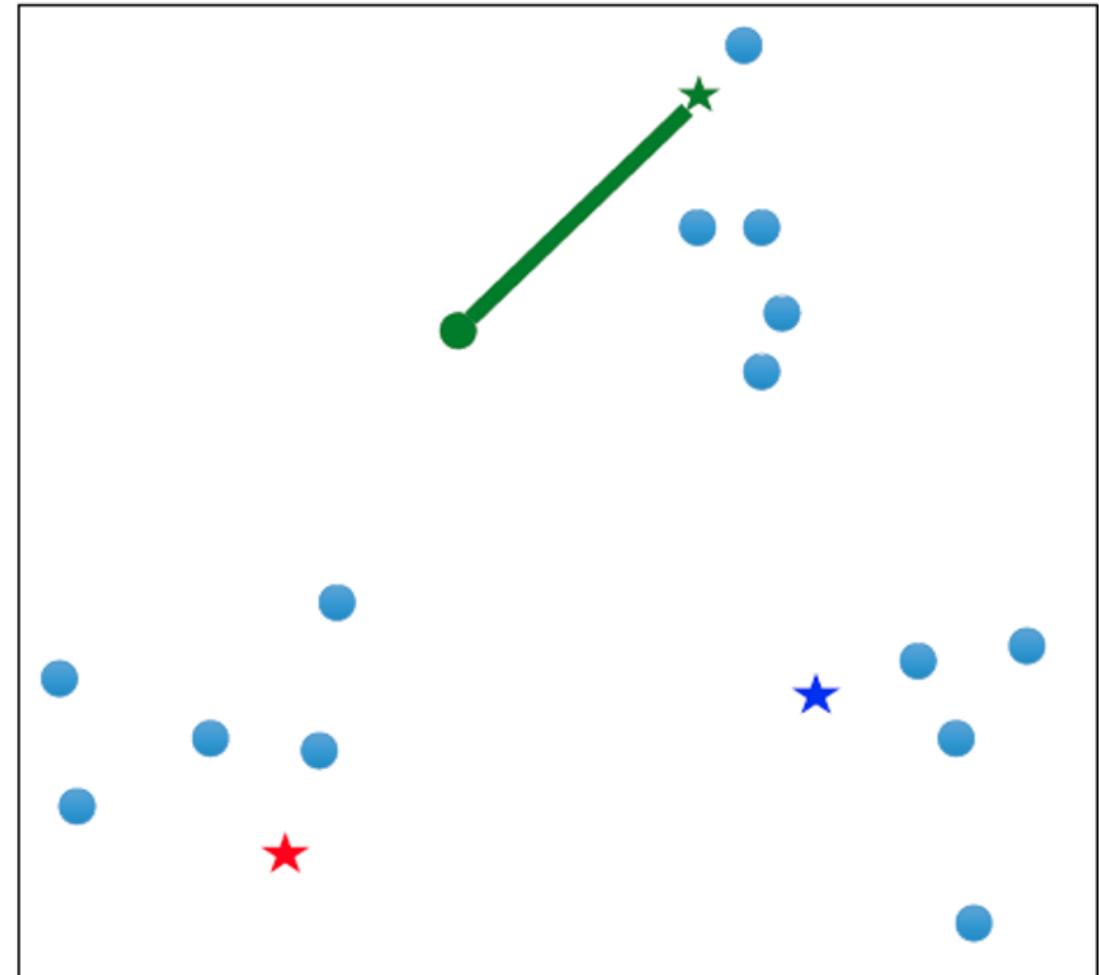
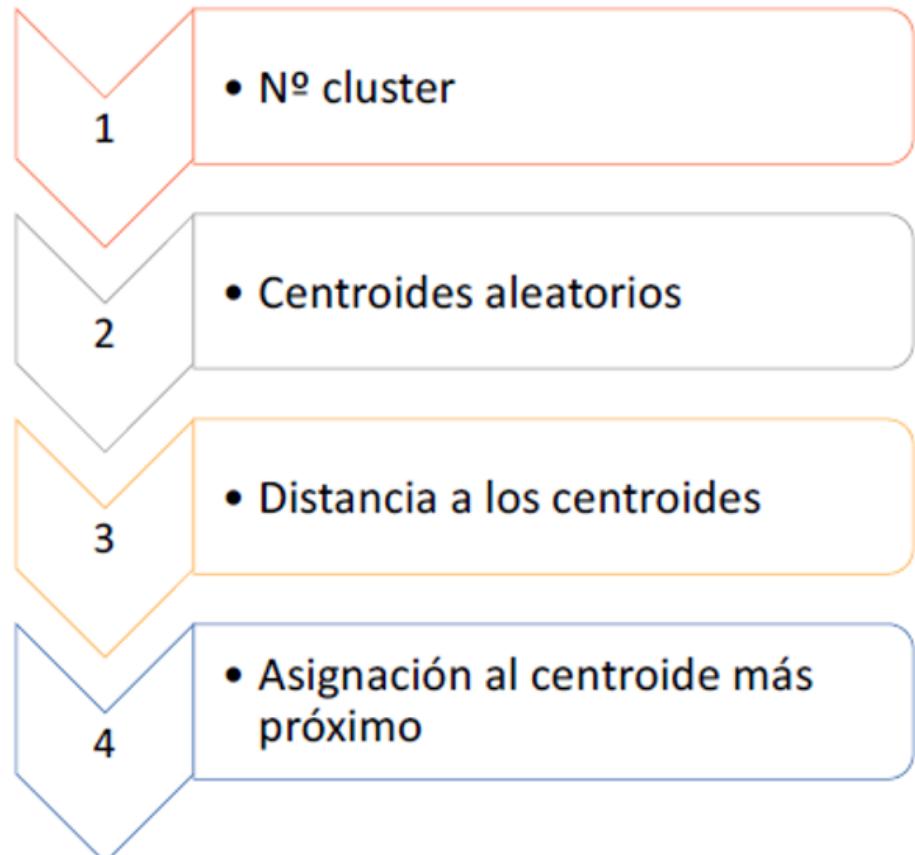
Algoritmo de las K -medias

Forgy (1965)



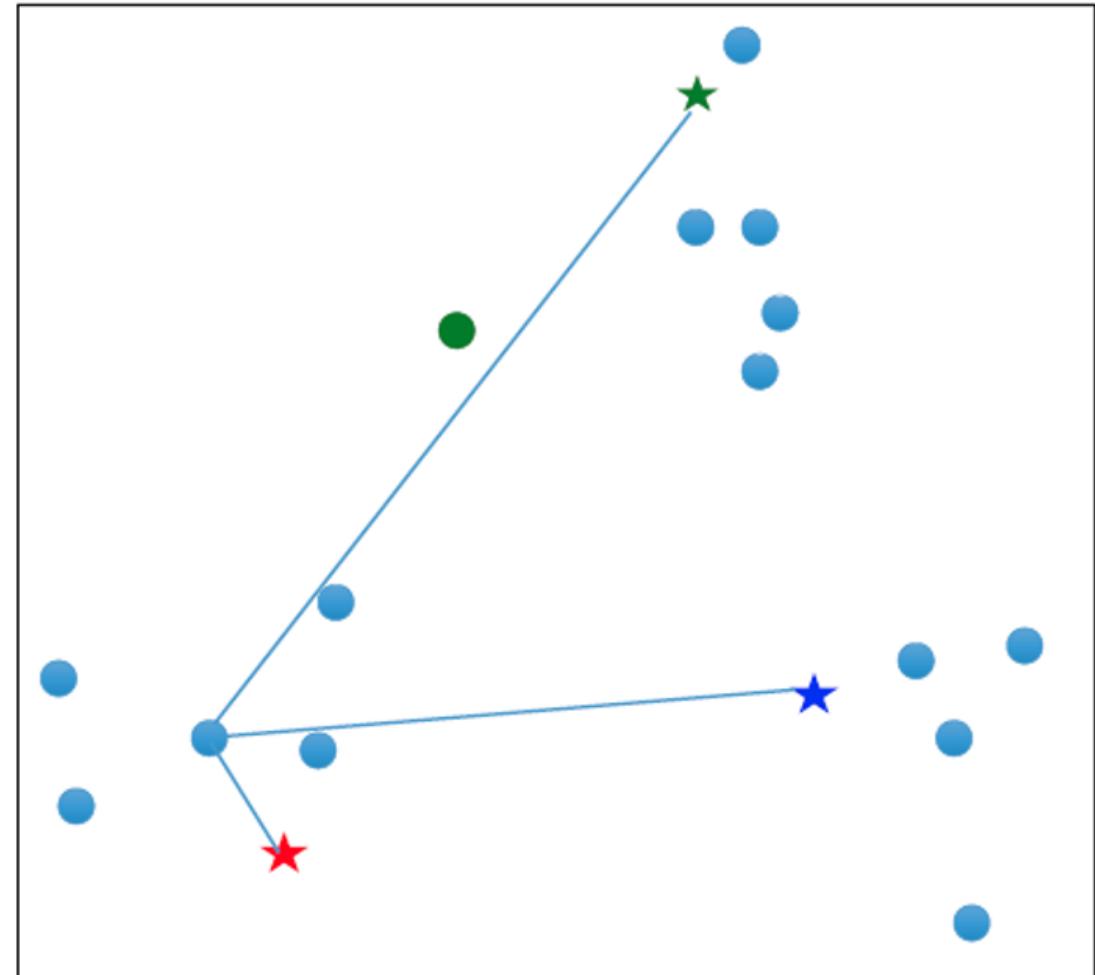
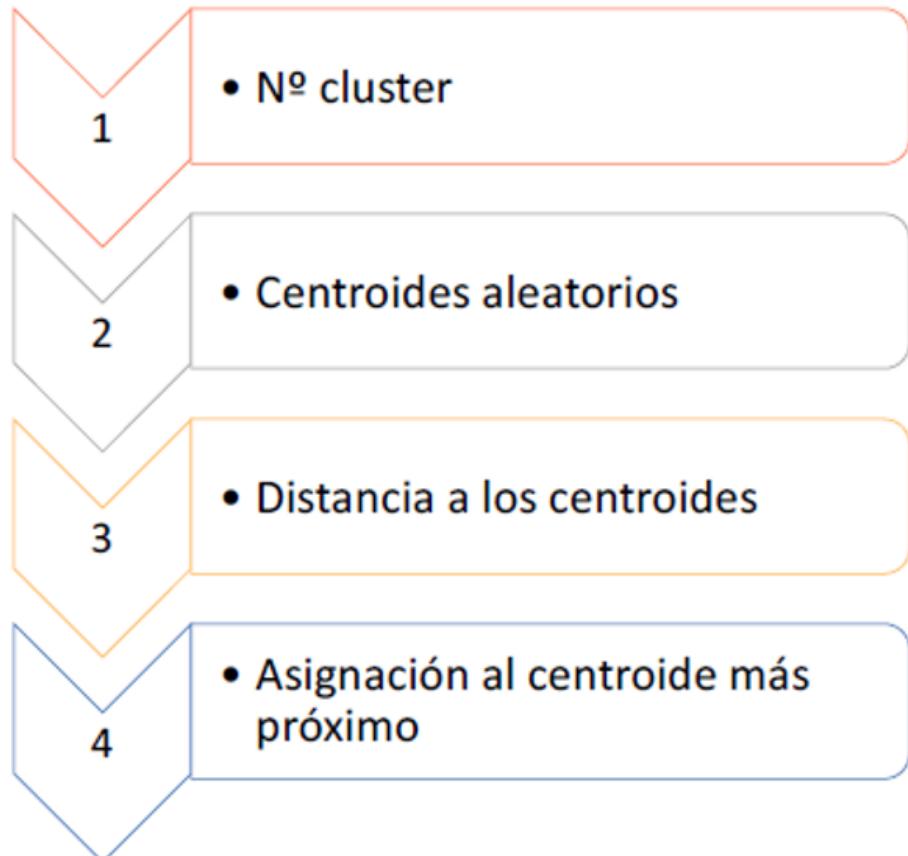
Algoritmo de las K -medias

Forgy (1965)



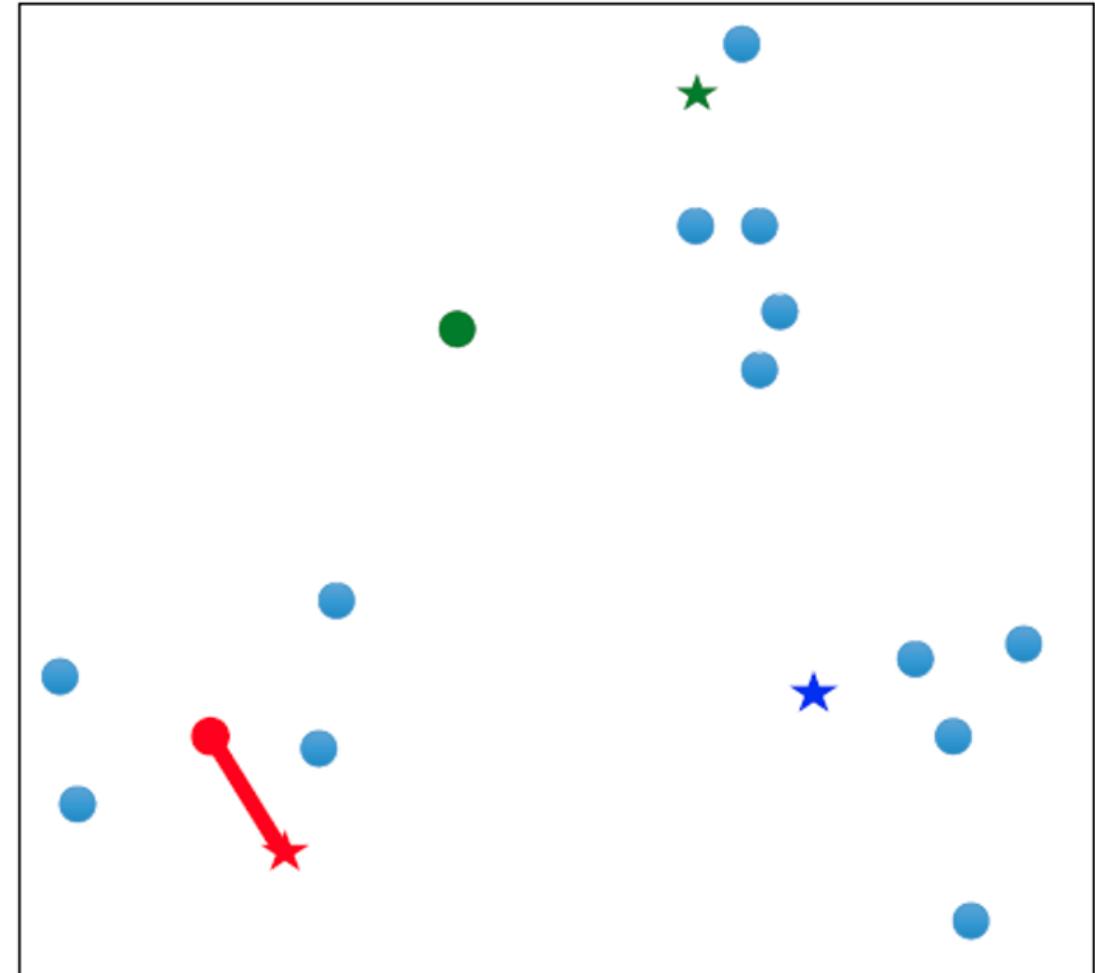
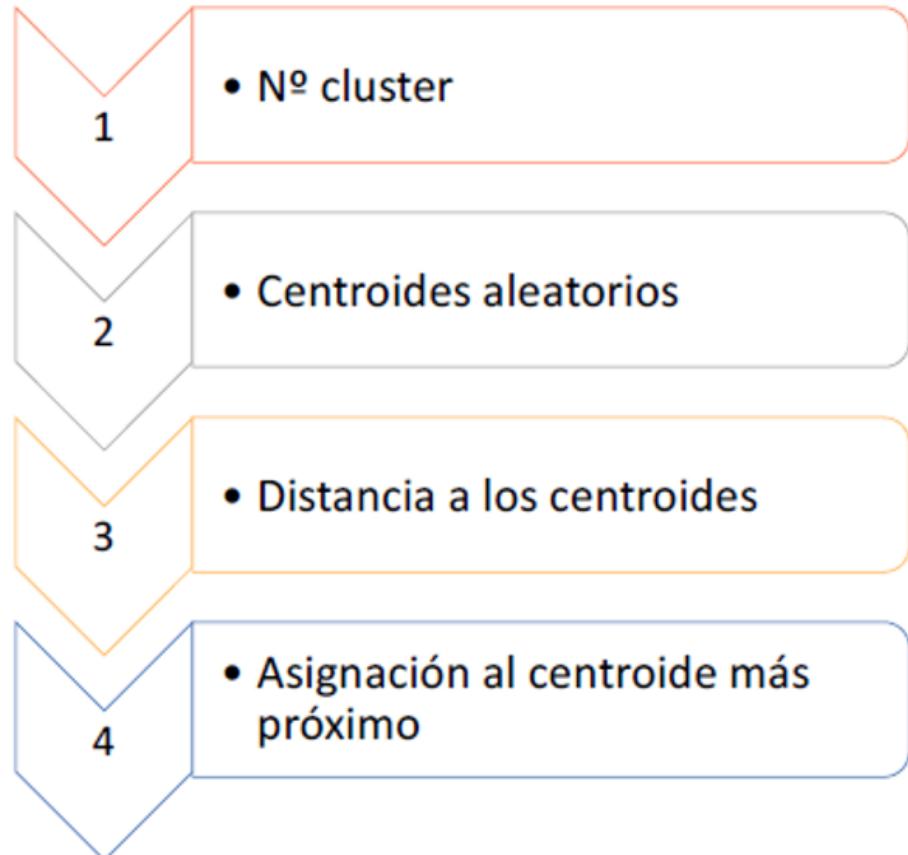
Algoritmo de las K -medias

Forgy (1965)



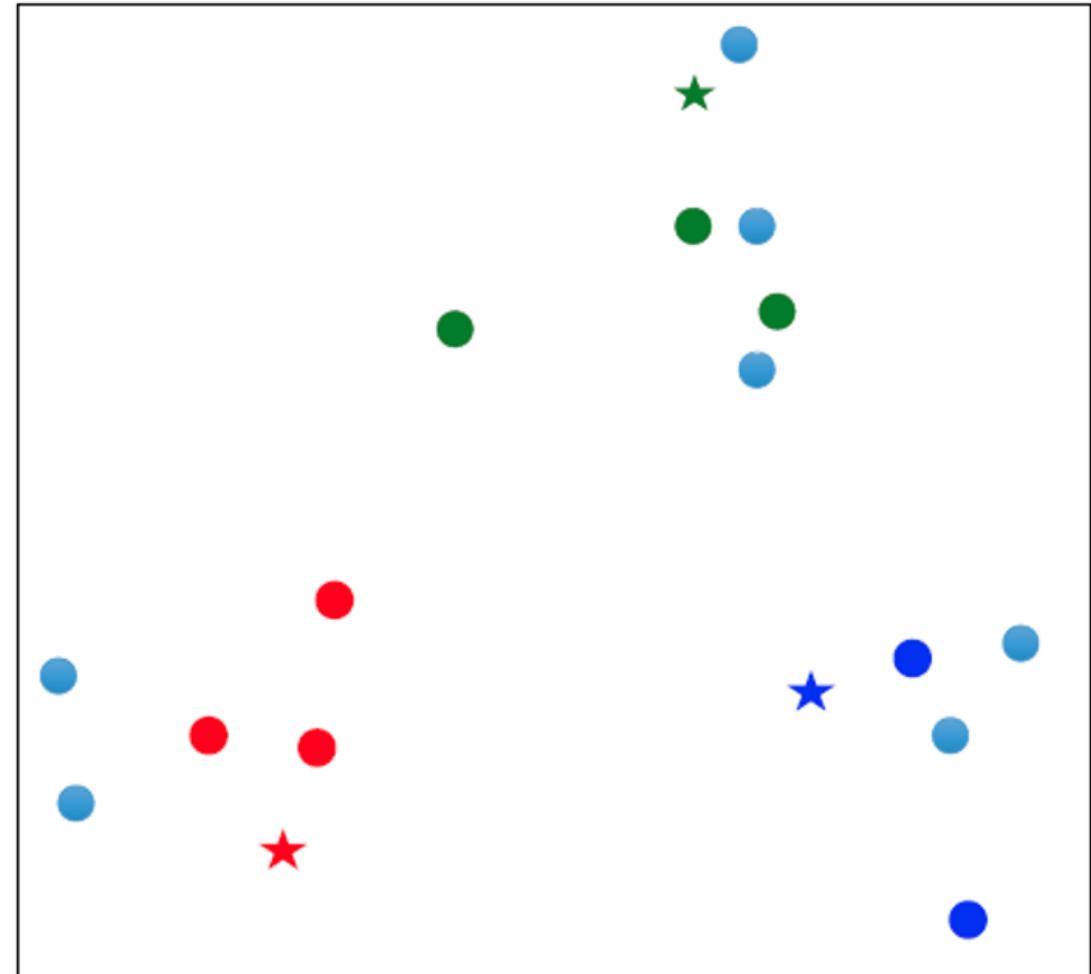
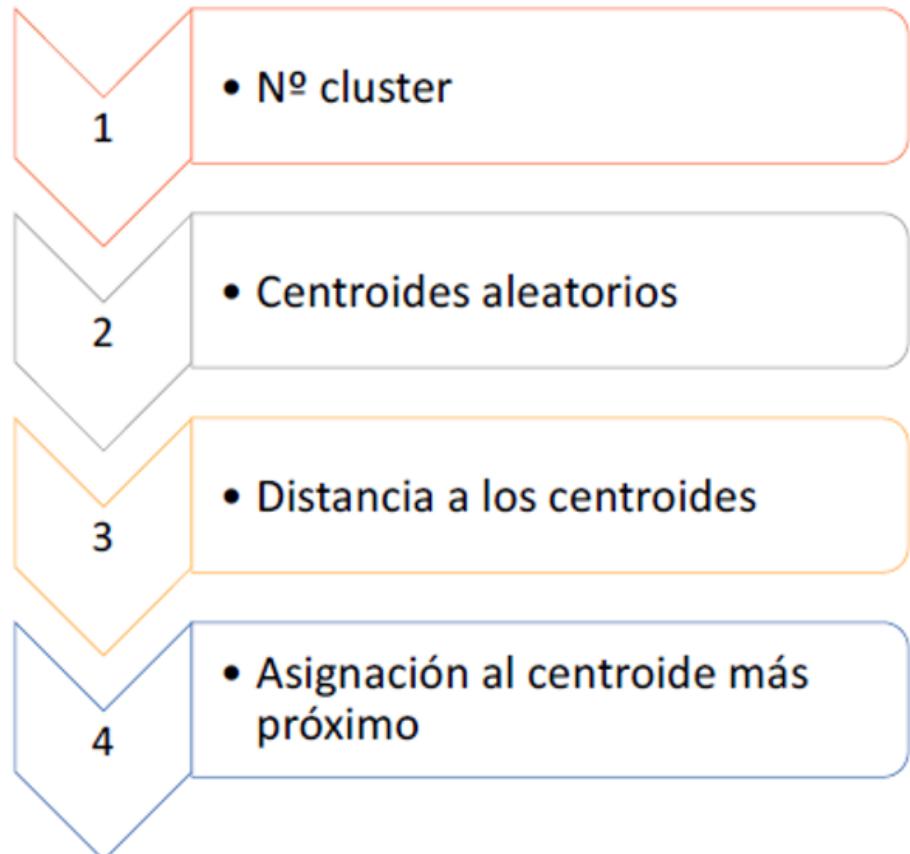
Algoritmo de las K -medias

Forgy (1965)



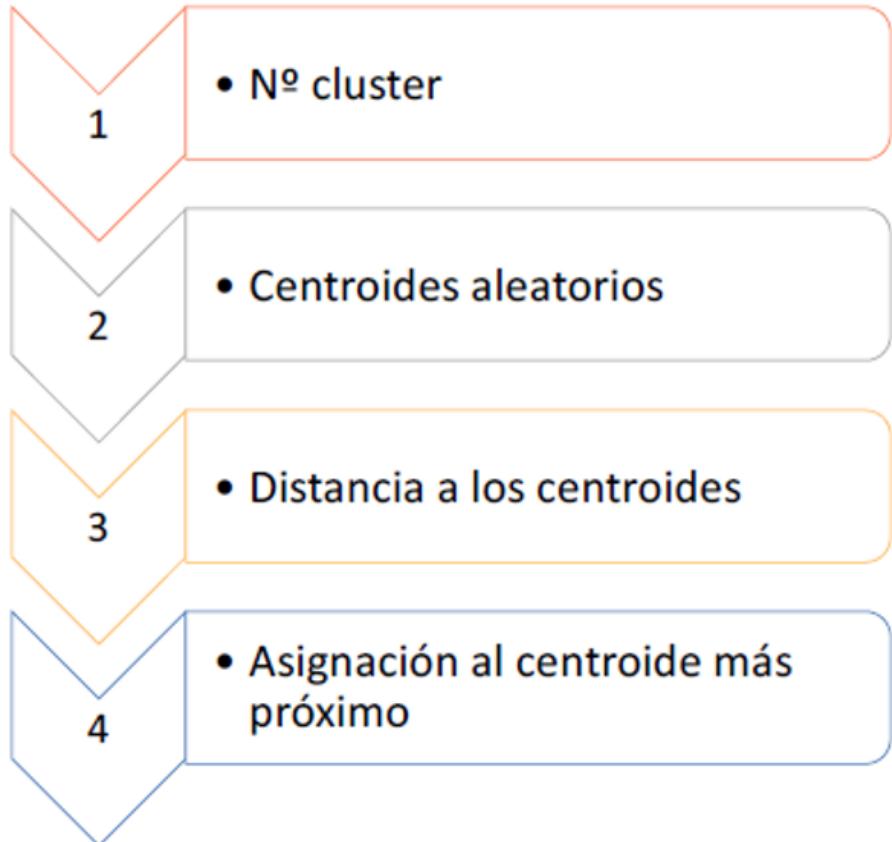
Algoritmo de las K -medias

Forgy (1965)

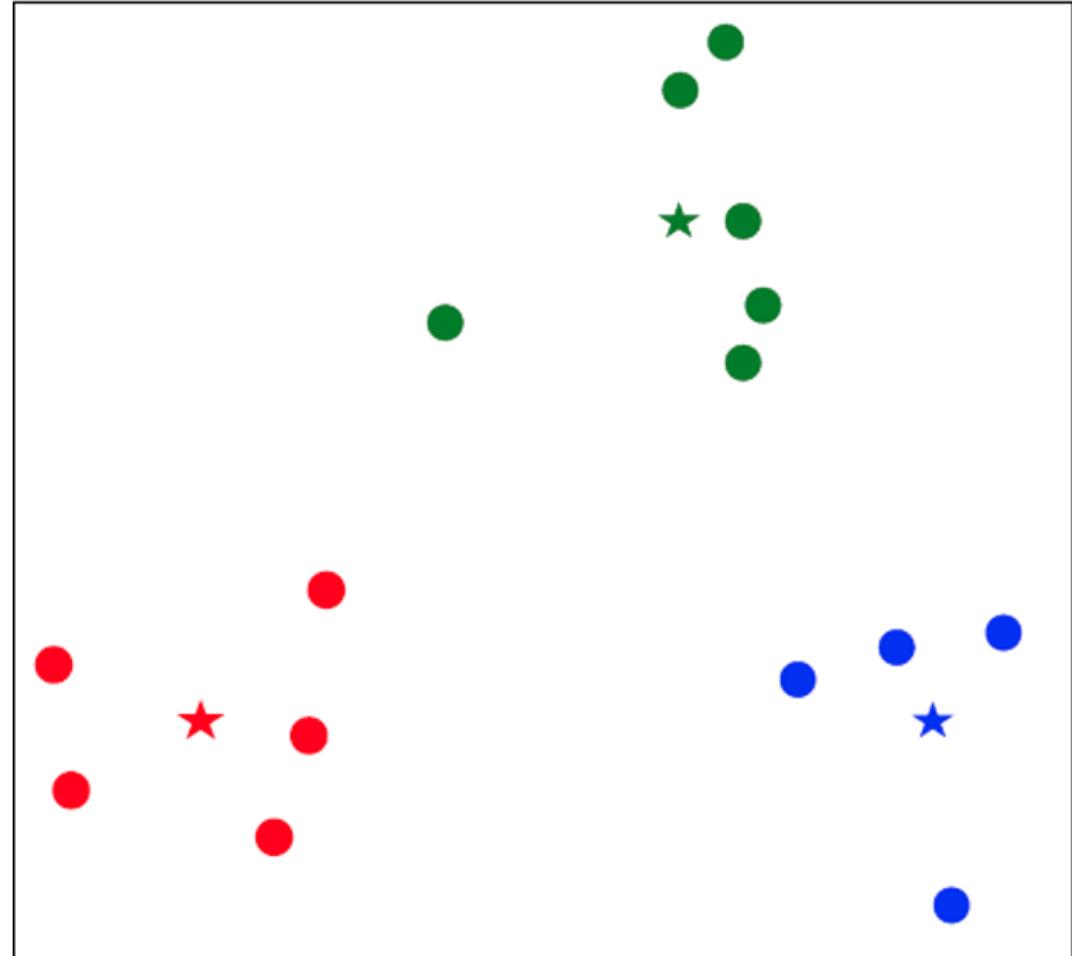


Algoritmo de las K -medias

Forgy (1965)



El centroide se calcula como el vector de medias de los puntos que pertenecen al clúster.



Comparación de algoritmos

K-MEANS



Algoritmo K-Means (Mcqueen, 1967)

Parte de centroides aleatorios, asigna una de las observaciones al cluster del centroide más cercano. Tras esto, recalcula los centroides (media) y asigna la siguiente observación al cluster del centroide más cercano. Y así sucesivamente.

Algoritmo K-Means (Forgy, 1965)

Partiendo de unos centroides aleatorios, asigna una de las observaciones al cluster del centroide más cercano. Tras haber asignado todas las observaciones a los diferentes cluster, recalcula los centroides de cada agrupación.

COMPUTACIONALMENTE MÁS COSTOSO

¡EN AMBOS ES NECESARIO INDICAR EL NÚMERO DE CLUSTERS K!

ALTERNATIVAS



Familia del K-Means

Fuzzy C-means (Dunn, 1973 ; Bezdek, Ehrlich & Full, 1984)

K-medoids (PAM) (Kaufman & Rousseeuw, 1990)

K-modes (Chaturvedi, Green, Carroll, 2001)

Kernel K-Means (Dhillon, Guan & Kulis, 2004)

K-means ++ (Arthur & Vassilvitskii, 2007)

MINI-BATCH K-means (Sculley, 2010)

Spherical K-Means (Hornik, Feinerer, Kober & Buchta, 2012)

Distributed K-means (Oliva, Setola & Hadjicostis, 2014)

Online K-Means (Sequential K-Means)

(Liberty, Sriharsha & Sviridenko, 2015)

Otro Tipo de Cluster

CLARA

(Kaufman & Rousseeuw, 1990)

DBSCAN

(Ester, Kriegel, Sander & Xy, 1996)

QT Clustering

(Heyer, Kruglyak, & Yoosheph 1999)

CLARANS

(Ng & Han, 2002)

HDBSCAN

(Campello, Moulavi & Sanders, 2013)

Métodos jerárquicos

Algoritmo para los métodos de aglomeración

1. Comenzar con tantas clases como elementos o individuos se tenga.
Las distancias entre clases son las distancias entre elementos originales.
2. Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos un grupo.
3. Sustituir los dos elementos utilizados en (2) para definir el grupo en (2), por un nuevo elemento que la represente.
4. Volver a (2) y repetir (2) y (3) hasta que tengamos todos los elementos quedan agrupados en un solo grupo.

EJEMPLOS

Consumo de proteínas

Considere el conjunto de datos del taller sobre el consumo de proteínas en algunos países de Europa. Realice un análisis clúster a partir de:

1. Un algoritmo de K -medias para agrupar a los países en 3 grupos.
2. Un algoritmo de aglomeración jerárquica. Defina el número de grupos apropiado.

Paso 1: Importar el conjunto de datos

```
1 options(scipen = 999)
2 library(pacman)
3
4 p_load(tidyverse, janitor, haven,
5         FactoMineR, factoextra, cluster)
6
7 url <- "https://github.com/jgbabativotvam/AnaDatos/raw/main/datos/PaisesProteinas.sav"
8
9 datos <- read_sav(url)
```

Paso 2: Preparar los datos

En el algoritmo de las K -medias es indispensable que las variables de análisis sea de tipo cuantitativo. Además, las variables son estandarizadas para evitar el efecto de la escala, de manera que:

$$z_i = \frac{x_i - \bar{x}_i}{s_i}, i = 1, \dots, p$$

```
1 datos <- datos |>
2   column_to_rrownames(var = "Pais") |>
3   scale()
```

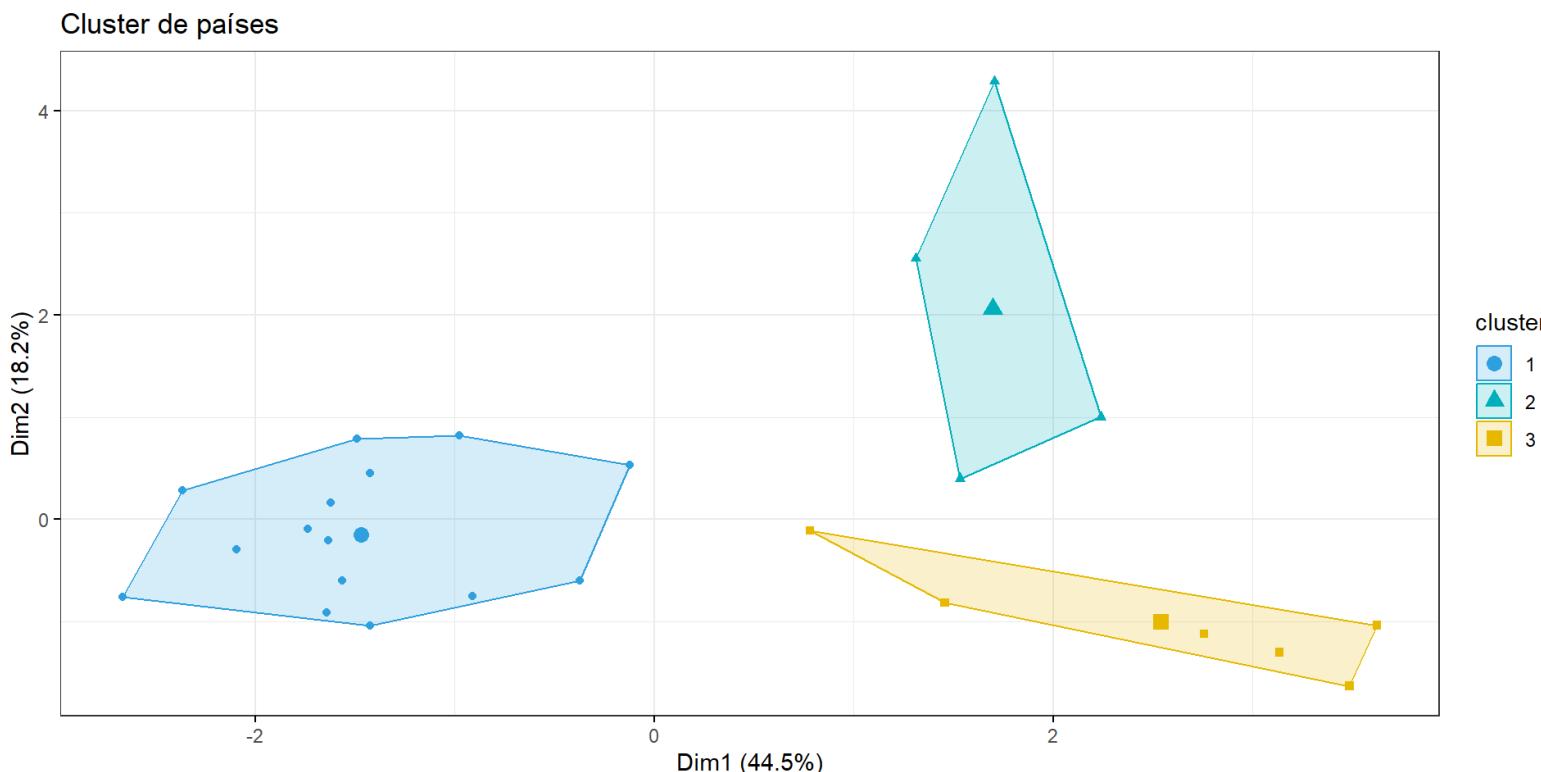
Algoritmo de K -medias

Paso 3: Aplicar el algoritmo

```
1 set.seed(26052013)
2
3 res <- kmeans(datos, centers = 3)
4
5 datos.clus <- data.frame(datos, cluster = res$cluster)
```

Paso 4: Visualización

```
1 fviz_cluster(res, data = datos,
2                         palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
3                         geom = "point",
4                         ellipse.type = "convex",
5                         main = "Cluster de países",
6                         ggtheme = theme_bw()
7 )
```



Paso 5: Validación

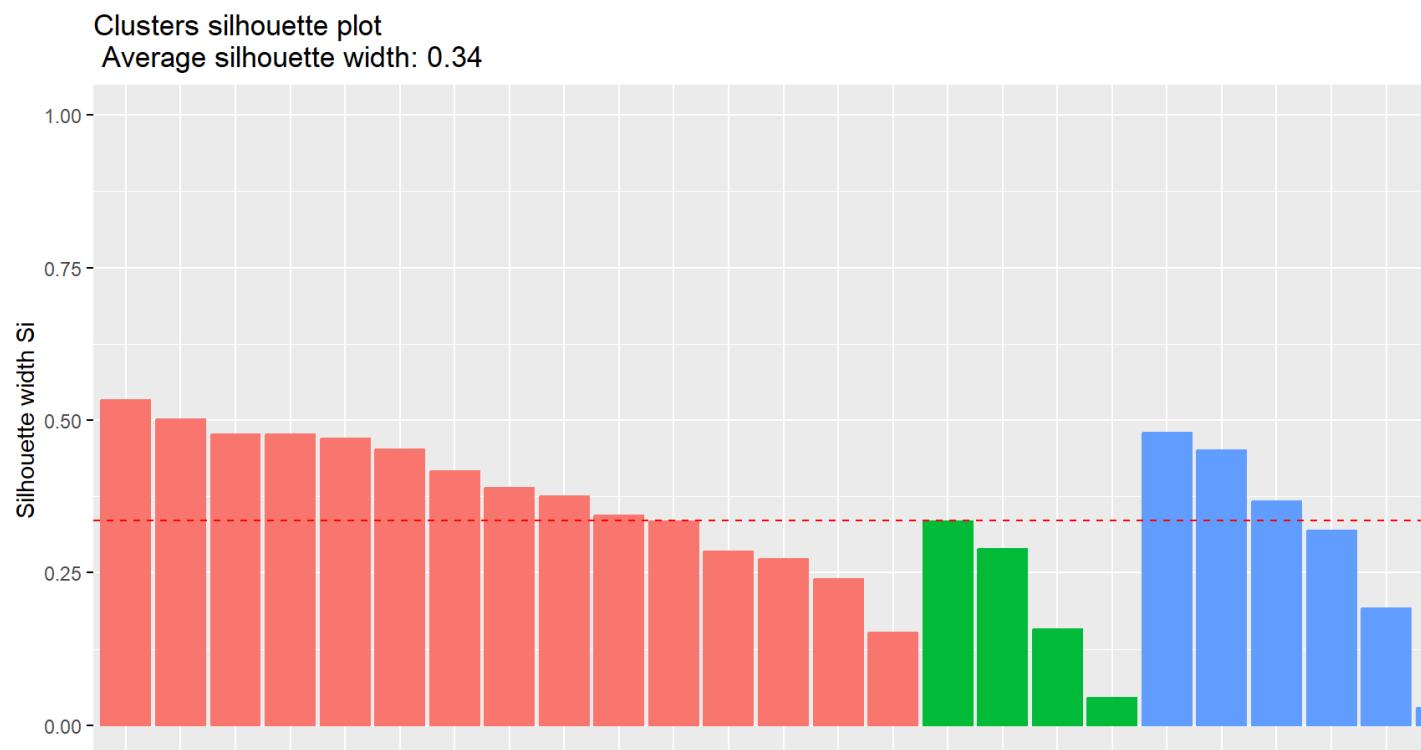
El gráfico de silueta es una herramienta útil para evaluar la calidad de las agrupaciones resultantes del algoritmo k-medias. Se debe revisar:

1. Barra alta: Si un punto de datos tiene una barra alta en el gráfico de silueta, significa que está bien asignado al grupo y está lejos de los puntos de otros grupos, esto indica una buena calidad del agrupamiento.
2. Valor medio de silueta: El valor medio de la silueta es la medida agregada de la calidad del método de clúster, en general se espera que esté en el rango de -1 a 1.

Paso 5: Validación

```
1 g1 <- silhouette(res$cluster, dist(datos))  
2 fviz_silhouette(g1)
```

	cluster	size	ave.sil.width
1	1	15	0.38
2	2	4	0.21
3	3	6	0.31



Paso 6: Interpretación

```
1 datos.clus |>
2   group_by(cluster) |>
3   summarise(across(where(is.numeric), ~mean(., na.rm=T)) )
```



```
# A tibble: 3 × 10
  cluster CarneRoja CarneBlanca Huevos Leche Pescado Cereales Féculas
    <int>     <dbl>      <dbl>   <dbl>   <dbl>   <dbl>     <dbl>     <dbl>
1       1      0.452     0.506   0.576   0.584   0.118   -0.610    0.353
2       2     -0.509    -1.11   -0.412  -0.832    0.982    0.130   -0.184
3       3     -0.790    -0.527  -1.17   -0.905  -0.950    1.44    -0.760
# i 2 more variables: Frutossecos <dbl>, Frutosyvegetales <dbl>
```

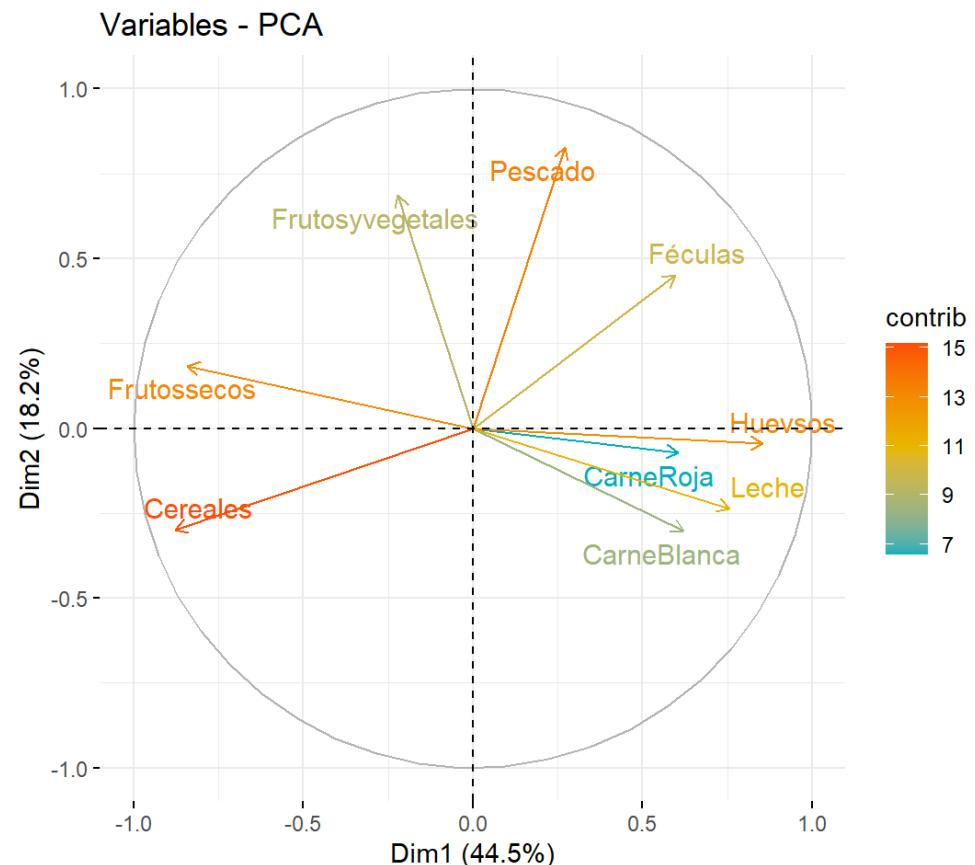
Algoritmo de jerárquico de Ward

Paso 3: Aplicar PCA

```

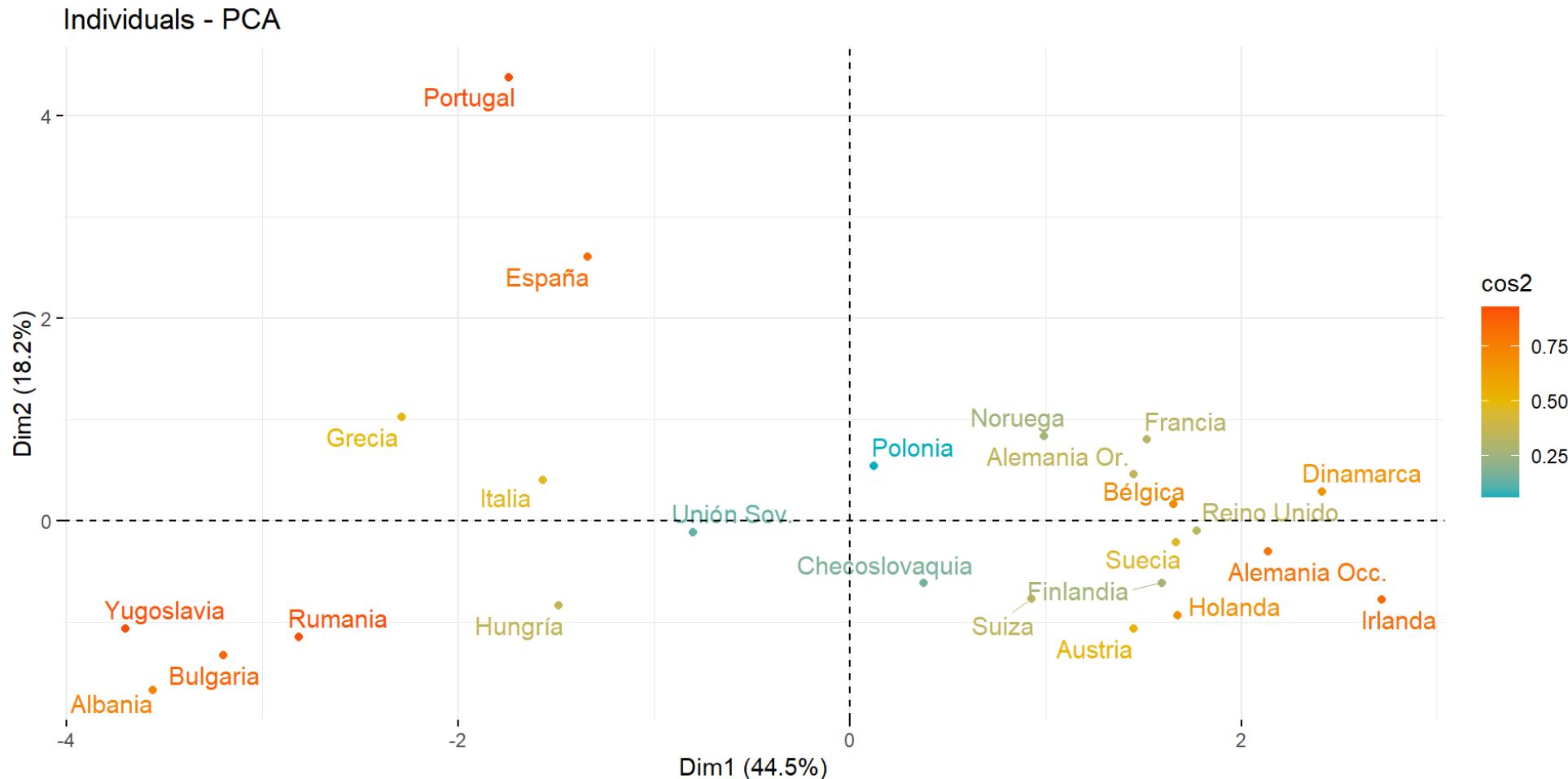
1 res.pca <- PCA(datos, graph = F)
2 fviz_pca_var(res.pca,
3               col.var="contrib",
4               gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
5               repel = TRUE)

```



Paso 3: Aplicar PCA

```
1 fviz_pca_ind(res.pca, col.ind = "cos2",
2                 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
3                 repel = TRUE)
```



Paso 4: Aplicar clúster jerárquico

Defina primero el número de clúster que se debería usar basado en el criterio de los índices de nivel o inercia iterclases.

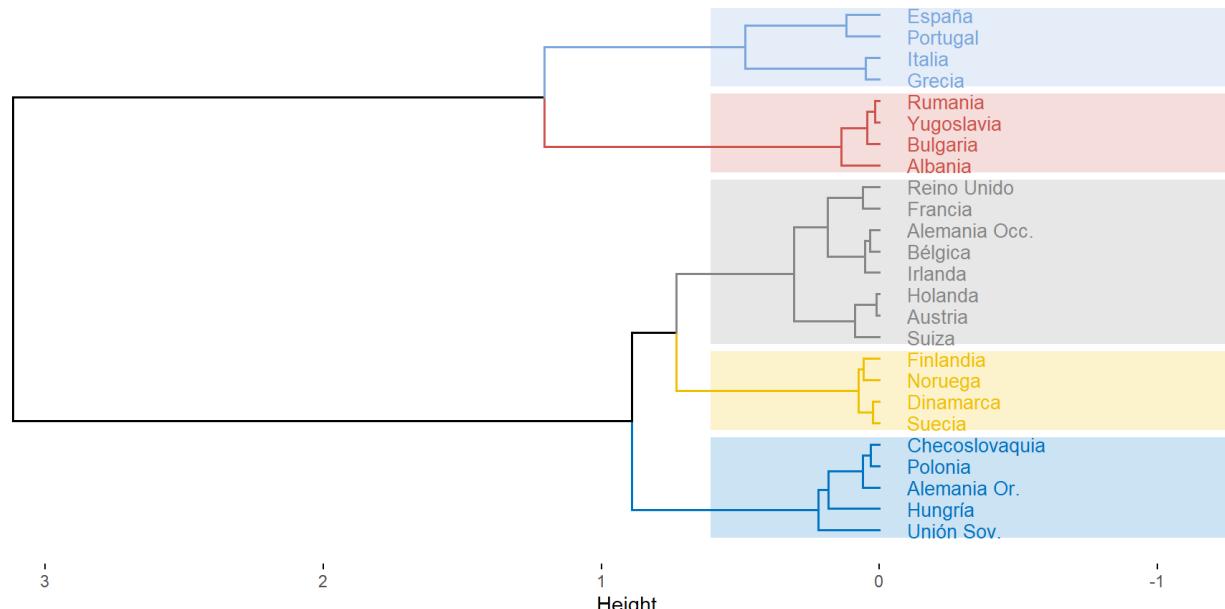
```
1 res.clus <- HCPC(res.pca)
```

```
1 res.clus <- HCPC(res.pca, nb.clust = 5, graph = FALSE)
```

Paso 5: Visualización del dendrograma

```
1 fviz_dend(res.clus,
2           cex = 0.7,
3           palette = "jco",
4           rect = TRUE, rect_fill = TRUE,
5           horiz = TRUE,
6           rect_border = "jco",
7           labels_track_height = 0.8,
8           main = "Cluster de paises"
9 )
```

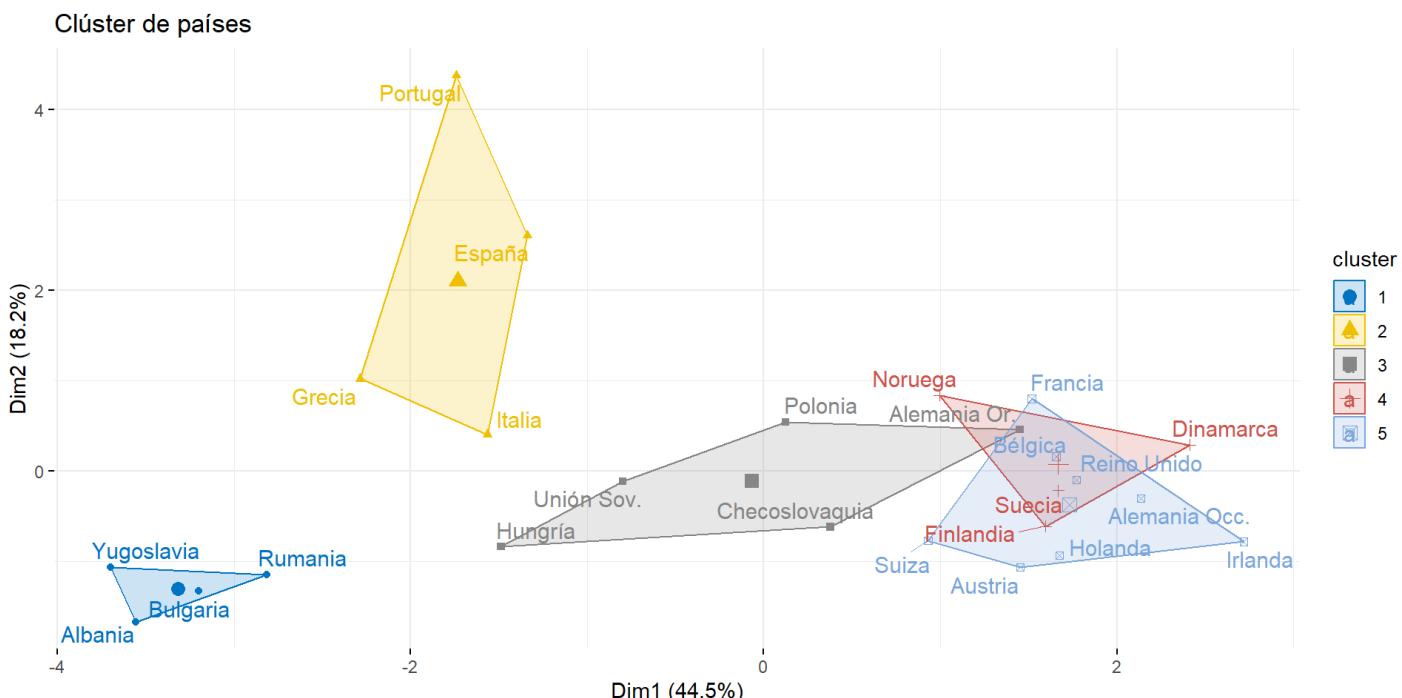
Cluster de paises



Paso 5: Visualización del plano de países

```
1 fviz_cluster(res.clus,
2             repel = TRUE,
3             show.clust.cent = TRUE,
4             palette = "jco",
5             ggtheme = theme_minimal(),
6             main = "Clúster de países"
7 )
```

```
1 # Puede usar el comando `res.clus$desc.var$quanti` para analizar la caracterización de
```



Ejemplo Electoral

Considere los datos artificiales `Bogota.sav` que simulan el resultado de la percepción de 350 encuestados. El ejercicio consiste en que a cada encuestado se le da una tarjeta con los nombres de los candidatos, posteriormente se leen algunas frases o se le mencionan algunas cualidades y deberá asociarlo con el candidato que considere que mejor la cumple.

Realice un análisis de correspondencias y posteriormente un análisis clúster para concluir sobre el perfil de los candidatos.

Paso 1: Importar el conjunto de datos

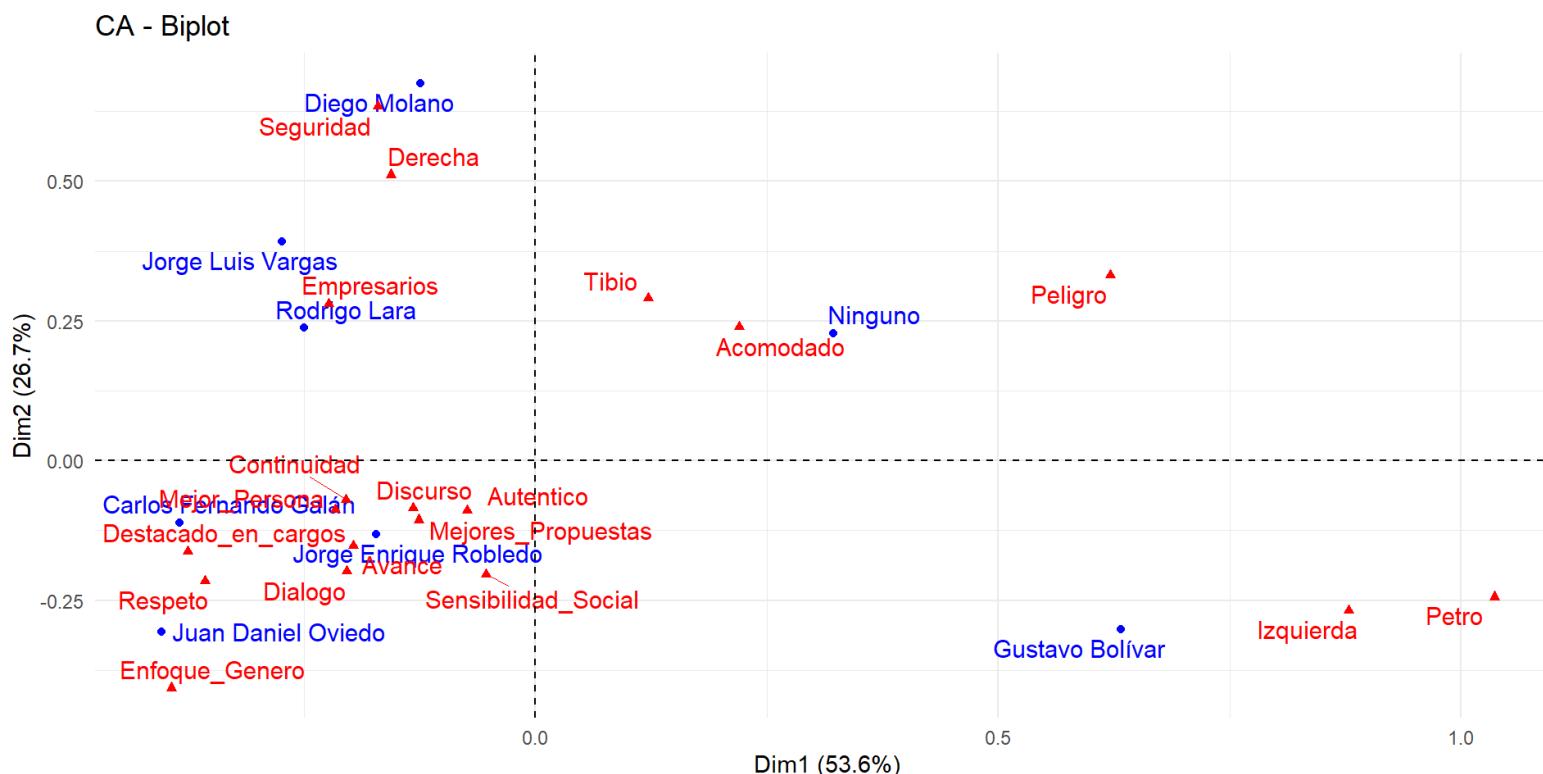
```
1 library(pacman)
2
3 p_load(tidyverse, janitor, haven, readxl,
4         FactoMineR, factoextra, cluster)
5
6 url <- "https://github.com/jgbabativam/AnaDatos/raw/main/datos/Bogota.sav"
7
8 datos <- read_sav(url)
```

Paso 2: Preparar los datos

```
1 datos <- datos |>  
2     column_to_rownames(var = "Candidatos")
```

Paso 3: Análisis de correspondencias

```
1 res.ac <- CA(datos, graph = F)
2
3 fviz_ca_biplot(res.ac,
4                   col.row="blue",
5                   col.col = "red",
6                   repel = TRUE) +
7   theme_minimal()
```



Paso 4: Clasificación de candidatos

Defina primero el número de clúster que se debería usar basado en el criterio de los índices de nivel o inercia iterclases.

```
1 res.clus <- HCPC(res.ac)
```

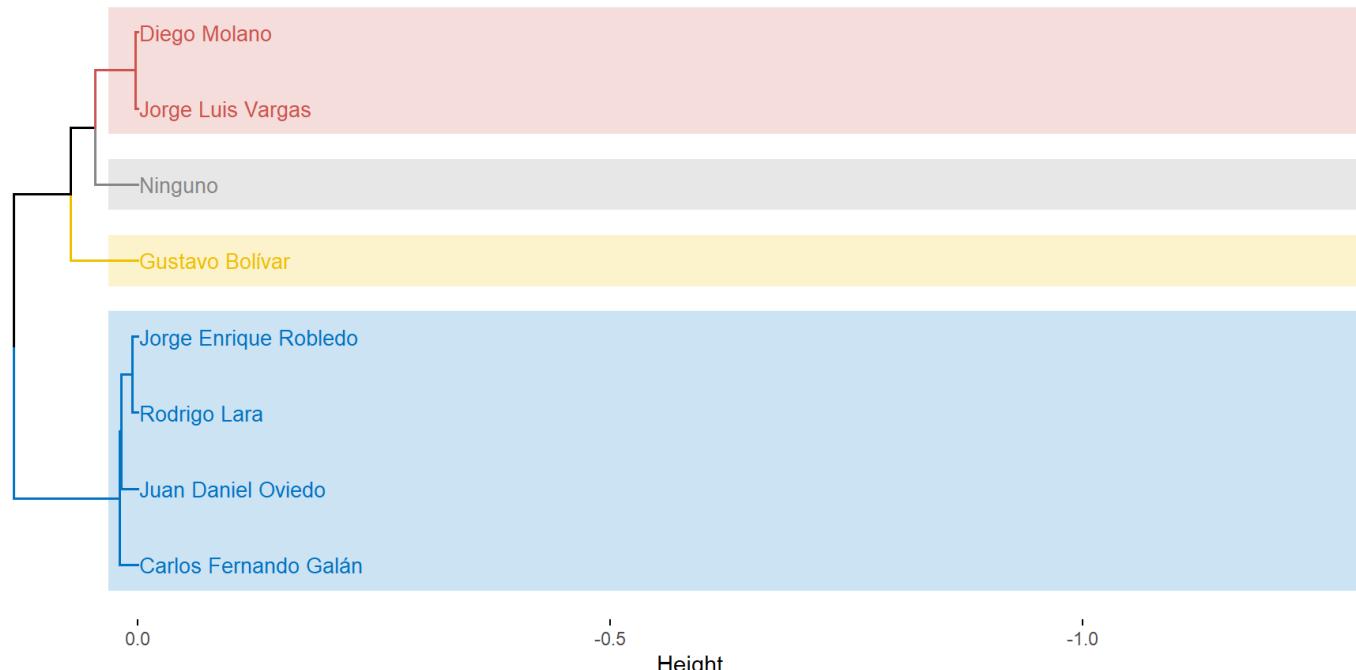
```
1 res.clus <- HCPC(res.ac, nb.clust = 4, graph = FALSE)
```

Paso 5: Visualización del dendrograma

```
1 fviz_dend(res.clus,
2           cex = 0.7, palette = "jco",
3           rect = TRUE, rect_fill = TRUE, horiz = TRUE,
4           rect_border = "jco", labels_track_height = 0.8,
5           main = "Cluster de candidatos"
6         )
```

```
1 #Puede usar el comando `res.clus$desc.var` para analizar la caracterización de cada cl
```

Cluster de candidatos



GRACIAS!

Referencias

- Husson, F., Lê, S., & Pagès, J. (2017). Exploratory multivariate analysis by example using R. CRC press.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). Multivariate data analysis 6th Edition.
<https://doi.org/10.1201/9780367409913>
- Aldás Manzano, J., & Uriel Jiménez, E. (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo, SA.

