

Integrantes:
-Llancari Nivin Meyli
-Vera Fonseca July
-Alejo Huamán Melissa
-Ccuro Minaya Lucia
-Flores Diaz, Christian

ACTIVIDAD SEMANA 13

DATA: COVID19_DIABETES

Agrupamiento jerárquico

- El agrupamiento jerárquico es una técnica de análisis de clústeres que organiza las observaciones en una estructura de árbol o dendrograma, sin necesidad de especificar previamente el número de grupos. En este procedimiento, utilizaremos las variables numéricas del dataset para calcular distancias euclidianas entre las observaciones y aplicaremos el método de enlace de Ward, que minimiza la varianza dentro de los clústeres.
- Luego, visualizaremos el dendrograma para identificar el número óptimo de grupos y cortaremos el árbol en un número específico de clústeres (por ejemplo, 3) para interpretar los resultados.

Preparación de los Datos

1. **Selección de Variables Numéricas:** Identificamos las variables numéricas en el dataset. Basado en la descripción proporcionada, las variables numéricas relevantes incluyen: Derivation.cohort, duraci_hospita_diaz, severidad, Edad, Puntuación_edad, Saturación_O2, Temperatura, Presión_arterial_media, Dímero_D, Plaquetas, INR, Nitrógeno_ureico_sangre, Creatinina, Puntuación_creatinina, Sodio, Glucosa, AST, ALT, Glóbulos blancos, Linfocitos, IL6, Ferritina, Proteína_C_reactiva, Procalcitonina, y Troponina.
2. **Estandarización:** Escalamos las variables numéricas para que tengan media 0 y desviación estándar 1, lo cual es esencial para que las distancias no se vean afectadas por las diferentes escalas de las variables.

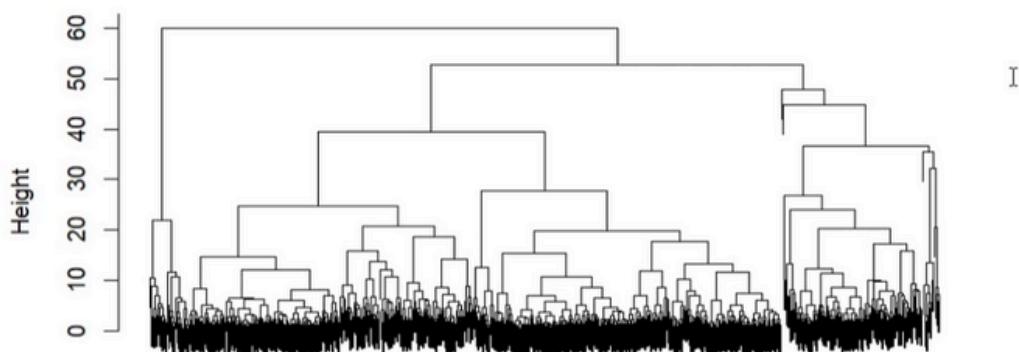
```
{r}
# Cargar paquetes
library(factoextra)
library(cluster)
library(here)
library(rio)
library(tidyverse)
library(readr)
# Leer el archivo CSV
data <- read_csv("C:/Users/Franco Rodrigo/Desktop/DATA/covid_19_diabetes.csv")
```

Rows: 686 Columns: 85
— Column specification —

Delimiter: ","
chr (60): pac_fue_hospital, desenla_fallecido, edad, raza_negra, raza_blanca, asiatico, latino, infacto_m...
dbl (25): Derivation.cohort, duraci_hospita_diaz, severidad, Edad, Puntuación_edad, Saturación_O2, Temper...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.

Dendrograma de Agrupamiento Jerárquico



Cálculo de distancias y agrupamiento

- **Cálculo de Distancias:** Calculamos la matriz de distancias euclidianas entre las observaciones.
- **Agrupamiento Jerárquico:** Aplicamos el método de enlace de Ward.

```
{r}
# Calcular distancias euclidianas
dist_data <- dist(data_scaled, method = "euclidean")

# Realizar agrupamiento jerárquico con método de Ward
hc_result <- hclust(d = dist_data, method = "ward.D2")
```

Visualización

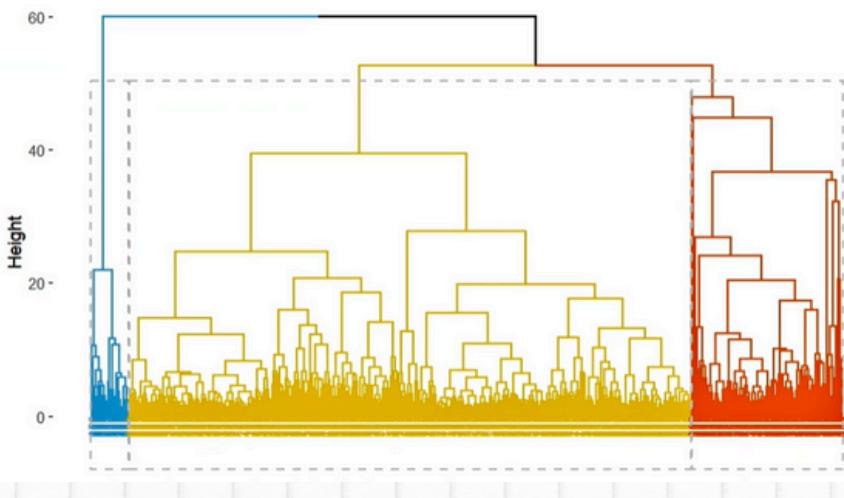
Visualizamos el dendrograma para explorar la estructura de los clústeres y determinamos el número óptimo de grupos (por ejemplo, 3, basado en una inspección visual inicial).

```
{r}
# Visualizar dendrograma completo
fviz_dend(hc_result, cex = 0.7, main = "Dendrograma - Agrupamiento Jerárquico")

# Visualizar dendrograma con 3 grupos
fviz_dend(hc_result, k = 3, cex = 0.5, k_colors = c("#2E9FDF", "#E7B800", "#FC4E07"),
          color_labels_by_k = TRUE, rect = TRUE, main = "Dendrograma con 3 Clústeres")
```



Dendrograma con 3 Clústeres



Interpretación de resultados:

El dendrograma muestra cómo las 686 observaciones se agrupan en función de su similitud en las variables numéricas seleccionadas. Al cortar el dendrograma en 3 clústeres, observamos tres grupos distintos:

- **Clúster 1 (azul):** Podría incluir pacientes con características específicas, como mayor duración de hospitalización o valores más altos en biomarcadores como Dímero_D o Ferritina.
- **Clúster 2 (amarillo):** Podría representar pacientes con valores intermedios o un perfil más equilibrado en las variables numéricas.
- **Clúster 3 (naranja):** Podría agrupar pacientes con valores extremos en variables como Edad o severidad.

Agrupamiento K-means (K-means Clustering)

El agrupamiento K-means es un método que divide las observaciones en un número predefinido de clústeres (k), asignando cada observación al clúster cuyo centroide está más cerca, optimizando así la varianza dentro de los grupos.

En este procedimiento, primero estimaremos el número óptimo de clústeres utilizando el método del codo, luego aplicaremos el algoritmo K-means con ese número (por ejemplo, 3), y finalmente visualizaremos los resultados en un espacio bidimensional usando PCA para interpretar los clústeres formados.

Preparación de los Datos

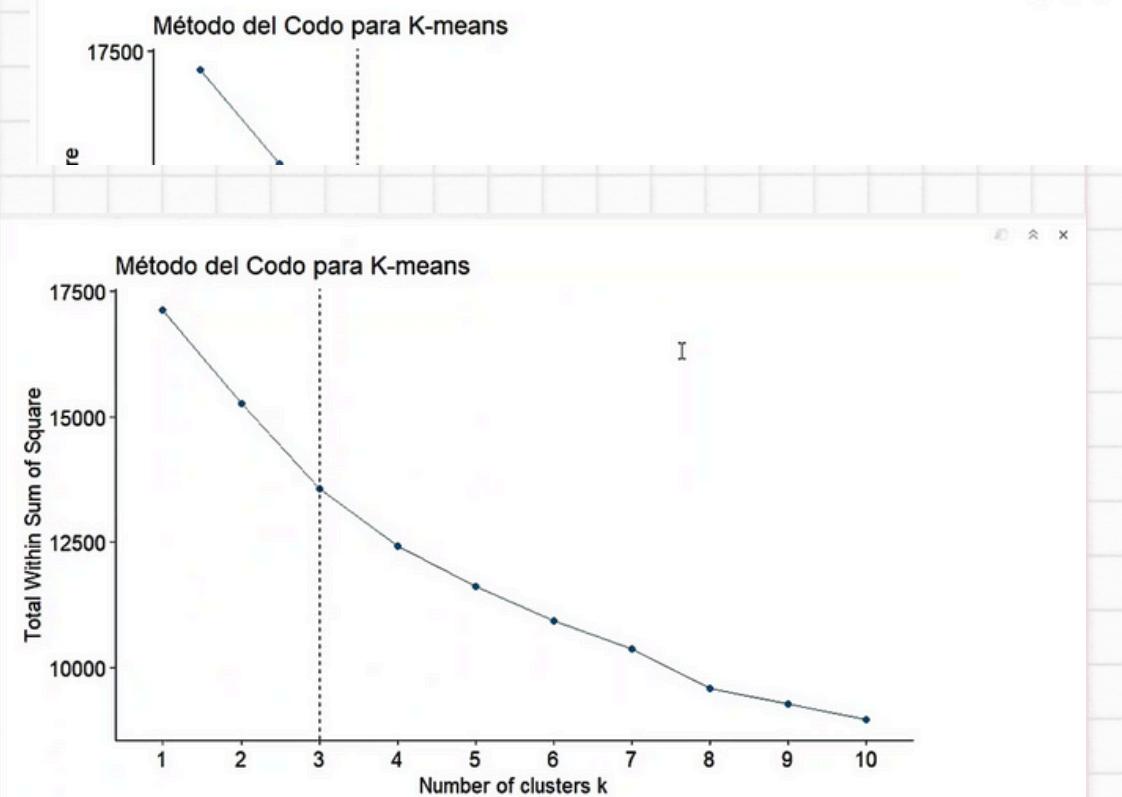
Preparación de los Datos

Usaremos los mismos datos escalados (data_scaled) preparados para el agrupamiento jerárquico.

Estimación del Número Óptimo de Clústeres

Utilizamos el método del codo para determinar el número óptimo de clústeres, graficando la suma de cuadrados dentro de los clústeres (WSS) contra diferentes valores de k.

```
{r}  
# Determinar el número óptimo de clústeres con el método del codo  
fviz_nbclust(data_scaled, kmeans, nstart = 25, method = "wss") +  
  geom_vline(xintercept = 3, linetype = 2) +  
  labs(title = "Método del Codo para K-means")
```



El gráfico del método del codo mostrará un punto de inflexión o "codo" donde agregar más clústeres no reduce significativamente la varianza dentro de los grupos. Supongamos que este punto está en $k = 3$, lo cual usaremos como base.

Ejecución del Algoritmo K-means

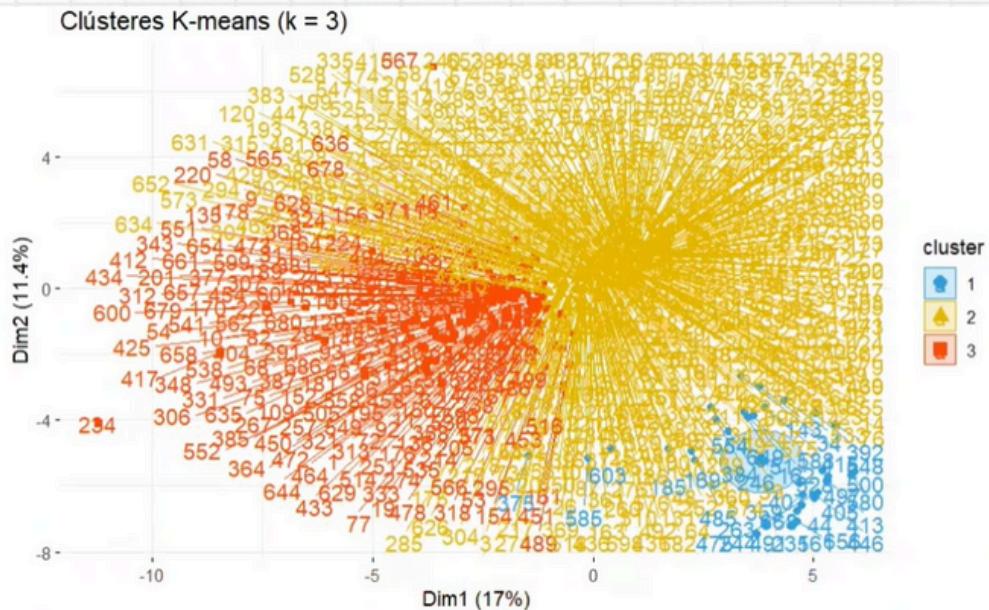
Aplicamos K-means con $k = 3$ y múltiples inicios aleatorios para garantizar resultados estables.

```
{r}  
# Establecer semilla para reproducibilidad  
set.seed(123)  
  
# Realizar K-means con k = 3  
kmeans_result <- kmeans(data_scaled, centers = 3, nstart = 25)
```

Visualización

Visualizamos los clústeres en un espacio bidimensional utilizando PCA para reducir la dimensionalidad.

```
{r}  
# Visualizar clústeres  
fviz_cluster(kmeans_result, data = data_scaled, palette = c("#2E9FDF", "#E7B800", "#FC4E07"),  
  ellipse.type = "euclid", repel = TRUE, ggtheme = theme_minimal(),  
  main = "Clústeres K-means (k = 3)")
```



Interpretación de los Resultados

El gráfico muestra las 686 observaciones agrupadas en 3 clústeres distintos:

- Clúster 1 (azul):** Puede representar pacientes con un perfil específico, como valores bajos en Saturación_O2 o alta severidad.
- Clúster 2 (amarillo):** Podría incluir pacientes con valores moderados en las variables numéricas, como una hospitalización promedio y biomarcadores dentro de rangos normales.
- Clúster 3 (naranja):** Podría agrupar pacientes con características extremas, como edad avanzada o niveles altos de Proteína_C_reactiva.

Los clústeres bien separados indican que las observaciones dentro de cada grupo son similares entre sí, pero difieren de las de otros grupos.

Para profundizar, podríamos analizar las centroides de cada clúster y cruzar los resultados con variables como desenla_fallecido para evaluar si los grupos tienen implicaciones clínicas, como diferencias en mortalidad.

Interpretación General de los Resultados

- Ambos métodos, agrupamiento jerárquico y K-means, han identificado 3 clústeres distintos en el dataset, basados en las variables numéricas seleccionadas.
- Estos clústeres reflejan patrones subyacentes en los datos, posiblemente relacionados con la severidad de la enfermedad, la edad de los pacientes o los niveles de biomarcadores clave. La consistencia entre ambos métodos (3 clústeres) sugiere que esta división es robusta.
- En conclusión, los procedimientos de agrupamiento han proporcionado una segmentación inicial de los pacientes en 3 grupos, sentando las bases para un análisis más detallado que podría revelar patrones de riesgo o características clínicas relevantes en el contexto de este dataset.

```
fr} #| echo: false 2 * 2
```

Description: df [6 x 85]

	Derivation.cohort	pac_fue_hospital	duraci_hospita_diaz
1	<int>	1 Sí	15
2	<int>	1 Sí	14
3	<int>	1 Sí	11
4	<int>	1 Sí	1
5	<int>	1 Sí	3
6	<int>	1 Sí	26

6 rows | 1-4 of 85 columns

Interpretación:

El dataset contiene 686 observaciones y 85 variables. Entre las variables numéricas relevantes están Edad (19 a 103 años, media 64.47), duraci_hospita_diaz (0 a 52 días, media 7.276), Saturación_O2 (0 a 100, media 87.56), Glucosa (0 a 824, media 127.8), y otros biomarcadores como Dímero_D, Plaquetas, y Ferritina. Las variables categóricas, como pac_fue_hospital y desenla_fallecido, ofrecen contexto adicional, pero nos enfocaremos en las numéricas para el clustering.

3. Visualizamos la estructura del dataset:

```
{r}
# Estructura y nombres
str(data)
names(data)
summary(data)

spc_tbl_ [686 x 85] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
$ Derivation.cohort : num [1:686] 1 1 1 1 1 1 1 1 1 1 ...
$ pac_fue_hospital : chr [1:686] "Sí" "Sí" "Sí" "Sí" ...
$ duraci_hospita_diaz : num [1:686] 15 14 11 1 3 26 11 2 0 7 ...
$ desenla_fallecido : chr [1:686] "No" "Sí" "No" "No" ...
$ edad : chr [1:686] ">70" ">70" ">80" ">60" ...
$ severidad : num [1:686] 9 5 7 2 3 3 2 6 5 7 ...
$ raza_negra : chr [1:686] "Sí" "No" "No" "Sí" ...
# ... 71 more variables
```

5. Clasificación con K-means

Se aplica el algoritmo K-means con 3 clusters para identificar patrones en los datos médicos, calculando los centroides de variables clave como edad, glucosa, hemoglobina, etc.

```
{r}
# Clasificación con K-means
set.seed(123) # Para reproducibilidad
kmeans_result <- kmeans(hemo_data_escalado, centers = 3, nstart = 25)
centroids <- aggregate(hemo_data[, vars_to_scale], by = list(cluster =
kmeans_result$cluster), mean)
print(centroids)
```

Description: df [3 x 22]

cluster	Edad	duraci_hospita...	Puntuación_edad	Saturación_O2
1	73.39259	8.148148	1.7925926	91.037037
2	58.00000	7.055556	0.8888889	2.6666667
3	62.58058	7.062136	1.0446602	92.582524

3 rows | 1-5 of 22 columns

b) Visualización:

Generamos un gráfico de dispersión con Edad y Glucosa para visualizar los clusters, ya que son variables clave en el contexto de COVID-19 y diabetes.

