Cluster-Discriminante

Santos G

Tabla de contenidos

1 Contexto del proyecto

La finalidad de esta guía es mostrar un flujo reproducible y profesional para realizar análisis de agrupamiento (clustering) sobre variables continuas en un contexto ambiental/biológico. Aprenderás a preparar los datos, establecer la tendencia a agruparse, seleccionar el número óptimo de clústeres, ejecutar clustering jerárquico y particional (k-means), evaluar la calidad del agrupamiento y describir los grupos desde el punto de vista estadístico y práctico.

2 Carga de librerías y dataset

```
# Librerías necesarias
library(tidyverse)    # manipulación de datos y ggplot2
library(cluster)      # silhouette, pam, etc.
library(factoextra)   # funciones para visualizar clustering y gap statistic
library(NbClust)      # selección de k por múltiples índices
library(clusterCrit)  # índices de validación (opcional)
library(dendextend)   # manipular dendrogramas
library(knitr)        # kable()
library(kableExtra)   # estilo de tablas
```

```r
library(clustertend)  # get_clust_tendency (Hopkins)

# Cargar dataset y preparar
data("USArrests")
A_raw <- USArrests %>% as_tibble(rownames = "region")

# Guardar copia original para auditoría
A_orig <- A_raw

# Selección de variables numéricas y escalado (estandarizar)
num_vars <- A_raw %>% select(Murder, Assault, UrbanPop, Rape)
num_scaled <- scale(num_vars) %>% as_tibble() %>% setNames(colnames(num_vars))
```

3 Preparación de la matriz para clustering

```r
summary_tbl <- num_vars %>%
summarise(across(everything(), list(mean = ~mean(.), sd = ~sd(.),
min = ~min(.), max = ~max(.)))) %>%
pivot_longer(everything(), names_to = c("variable","stat"), names_sep = "_") %>%
pivot_wider(names_from = stat, values_from = value) %>%
select(variable, mean, sd, min, max)

knitr::kable(summary_tbl, digits = 3)
```

Tabla 1: Resumen numérico de variables para clustering

variable	mean	sd	min	max
Murder	7.788	4.356	0.8	17.4
Assault	170.760	83.338	45.0	337.0
UrbanPop	65.540	14.475	32.0	91.0
Rape	21.232	9.366	7.3	46.0

El conjunto de variables seleccionadas para el análisis de agrupamiento incluye indicadores sociales y de criminalidad: tasas de homicidios (Murder), asaltos (Assault), porcentaje de población urbana (UrbanPop) y delitos sexuales (Rape).

De acuerdo con el resumen estadístico (ver Tabla 1), se observa una variabilidad considerable entre estados:

- Murder presenta una media de 7.79 delitos con valores entre 0.8 y 17.4 , lo que evidencia fuertes contrastes en las tasas de homicidios a nivel estatal.
- Assault (asaltos) muestra la mayor dispersión (SD = 83.3), con un rango amplio entre 45 y 337 incidentes, indicando marcadas diferencias en la incidencia de violencia física.
- UrbanPop, porcentaje de población urbana, promedia 65.5 % (rango: 32 – 91), mostrando una distribución relativamente homogénea, aunque con algunos estados menos urbanizados.

- Finalmente, Rape presenta un promedio de 21.2 casos por 100.000 habitantes y una desviación estándar de 9.4 , lo que sugiere heterogeneidad en la prevalencia de este tipo de delitos.

En conjunto, la amplitud de los rangos y desviaciones indica una base de datos con suficiente variabilidad para aplicar técnicas de agrupamiento y explorar posibles perfiles regionales de criminalidad y urbanización.

```
# Hopkins
set.seed(42)
hopkins_stat <- get_clust_tendency(num_scaled, n = nrow(num_scaled) - 1,
                                   graph =FALSE)$hopkins_stat

hopkins_res <- tibble(test = "Hopkins", value = round(hopkins_stat, 3))
knitr::kable(hopkins_res, caption = "")
```

Tabla 2: Test de tendencia al agrupamiento (Hopkins)

test	value
Hopkins	0.608

El test de tendencia al agrupamiento de Hopkins (ver Tabla 2) arrojó un valor de 0.608 . Este estadístico evalúa si los datos presentan una estructura de agrupamiento no aleatoria. Donde, valores cercanos a 0 indican una fuerte tendencia a agruparse, mientras que valores iguales o mayores oa 0.5 sugieren una distribución aleatoria. En este caso, el valor obtenido (0.608) se aproxima a 0.5, lo que sugiere una estructura de agrupamiento débil en el conjunto de datos. Esto implica que los estados de EE.UU. no se diferencian en grupos muy definidos según las variables de criminalidad y urbanización, aunque podrían existir subgrupos moderadamente diferenciados que justifiquen la aplicación de métodos de clustering exploratorio.

4 Elección del número de clusters (validación)

```
# Distancia Euclidiana y Ward.D2
dist_mat <- dist(num_scaled, method = "euclidean")
hc <- hclust(dist_mat, method = "ward.D2")

# Silhouette promedio para k = 2..6 (kmeans)
sil_vals <- tibble(k = 2:6, sil_avg = NA_real_)
for (k in 2:6) {
km <- kmeans(num_scaled, centers = k, nstart = 50)
sil <- silhouette(km$cluster, dist_mat)
sil_vals$sil_avg[sil_vals$k == k] <- mean(sil[, 3])
}
knitr::kable(sil_vals, digits = 3)
```

Tabla 3: Silhouette promedio por número de clusters (k)

k	sil_avg
2	0.408
3	0.309
4	0.340
5	0.303
6	0.286

De acuerdo con los resultados del índice de Silhouette (ver Tabla 3), el valor más alto se obtuvo para k = 2, con un promedio de 0.408, lo que indica que esta partición presenta la mejor cohesión interna y separación entre grupos. A medida que aumenta el número de clústeres, el valor de la silueta disminuye progresivamente (por ejemplo, 0.309 para $k = 3$ y 0.34 para $k = 4$), reflejando una menor nitidez en la estructura del agrupamiento.

```
# Gap statistic
set.seed(42)
gap <- clusGap(num_scaled, FUN = kmeans, K.max = 6, B = 50)
fviz_gap_stat(gap)  # figura
```

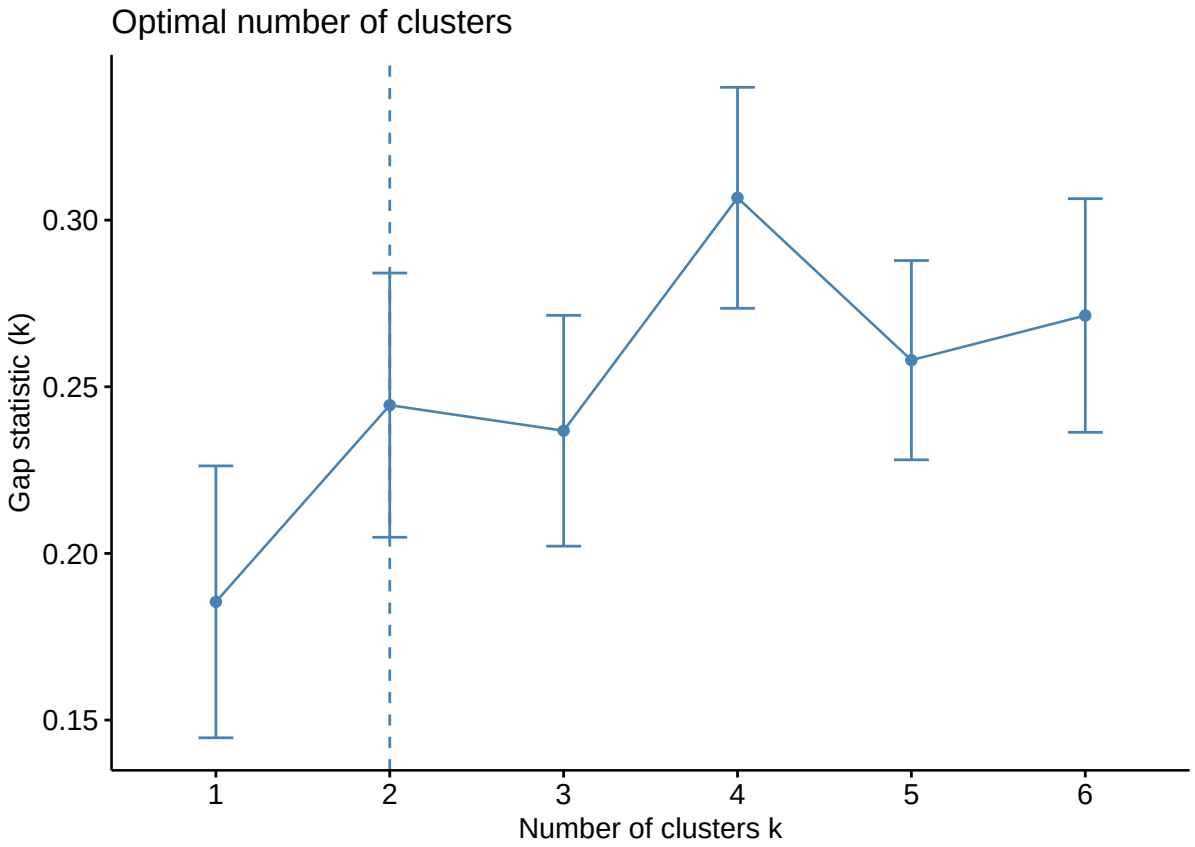

Figura 1: Análisis del Gap statistic

El análisis del *Gap Statistic* (ver Figura 1) mostró un patrón coherente, con el valor máximo en torno a dos clústeres, lo que respalda una estructura de dos grupos bien diferenciados en el espacio morfológico.

5 K-means

```r
# Ejecutar k-means con el k elegido (aquí usamos el k con mayor sil_avg)
k_opt <- sil_vals$k[which.max(sil_vals$sil_avg)]
set.seed(42)
km_res <- kmeans(num_scaled, centers = k_opt, nstart = 50)

# Añadir cluster al df original
df_clust <- A_raw %>% mutate(cluster_k = factor(km_res$cluster))

# Visualizar clusters sobre PC1-PC2 (representación)
fviz_cluster(km_res, data = num_scaled, geom = "point", ellipse.type = "norm",
palette = "jco", ggtheme = theme_minimal())
```

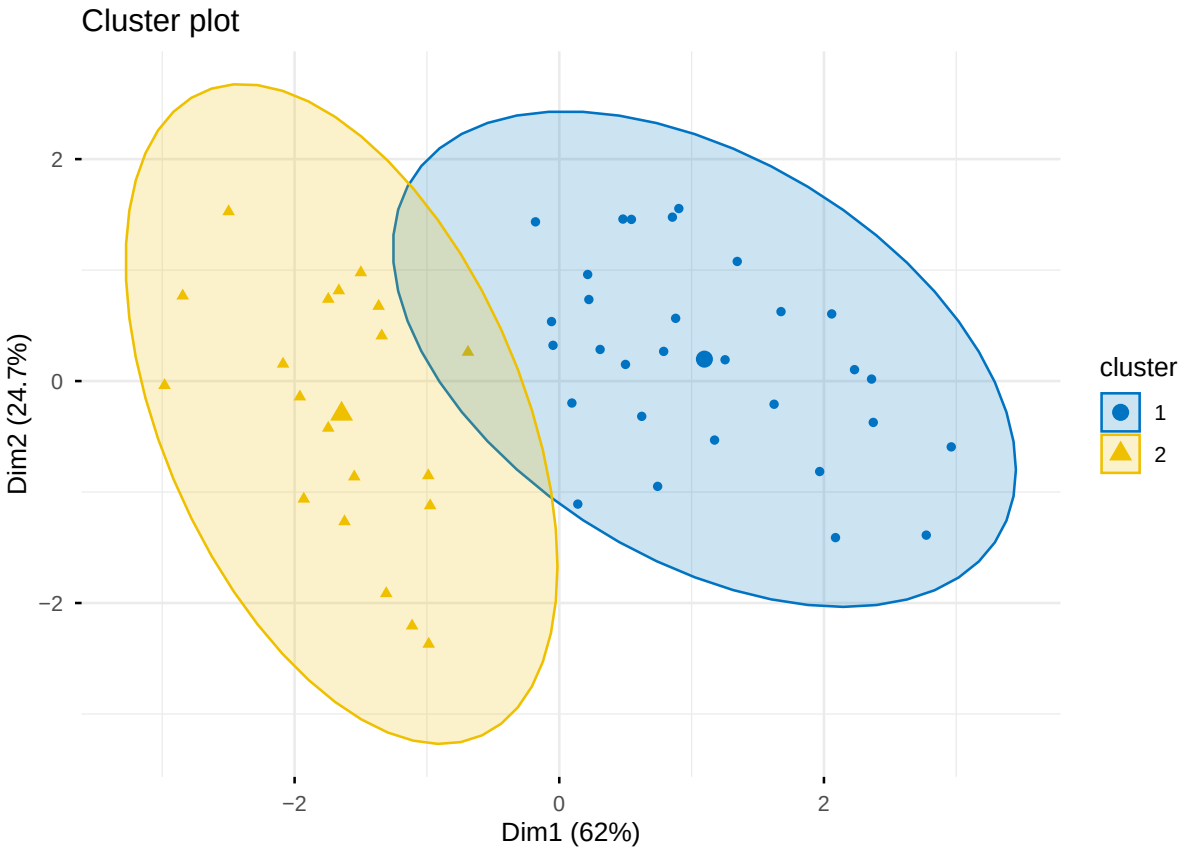

Figura 2: K-means (k = 2) y representación en PC1-PC2

El gráfico de agrupamiento obtenido mediante el algoritmo k-means (ver Figura 2) muestra la distribución de las observaciones en el espacio definido por las dos primeras componentes principales (Dim1 y Dim2), las cuales explican conjuntamente una proporción considerable de la variabilidad total de los datos (aproximadamente el 87%).

Con k = 2 clústeres, se distinguen dos grupos principales bien definidos, representados por las elipses de color azul (Cluster 1) y amarillo (Cluster 2). El Cluster 1 agrupa observaciones con valores más altos en la primera dimensión (Dim1), lo que sugiere un patrón particular en las variables originales que contribuyen positivamente a ese componente. Por su parte, el Cluster 2 se concentra hacia el extremo

opuesto de Dim1, reuniendo observaciones con valores más bajos en dicha dimensión, lo cual indica diferencias sistemáticas respecto al grupo azul.

La ligera superposición entre ambas elipses evidencia cierta similitud parcial entre algunos estados de los dos grupos, aunque la separación global es clara, lo que respalda la validez de la partición obtenida mediante k-means.

```r
# Calcular medias por grupo en variables escaladas
cluster_summary <- num_scaled %>%
  mutate(cluster = km_res$cluster) %>%
  pivot_longer(-cluster, names_to = "variable", values_to = "valor") %>%
  group_by(cluster, variable) %>%
  summarise(mean = mean(valor), .groups = "drop") %>%
  pivot_wider(names_from = variable, values_from = mean)

# Reordenar etiquetas de clúster según la media de Murder
cluster_order <- cluster_summary %>%
  mutate(avg_violence = Murder) %>%
  arrange(avg_violence) %>%
  mutate(new_cluster = row_number())

# Crear un vector de equivalencia
mapping <- cluster_order$new_cluster
names(mapping) <- cluster_order$cluster

# Reetiquetar clusters en los datos originales
df_clust <- A_raw %>%
  mutate(cluster_k = factor(mapping[km_res$cluster]))

# Actualizar cluster_summary con nuevas etiquetas
cluster_summary <- cluster_summary %>%
  mutate(cluster = mapping[cluster]) %>%
  arrange(cluster)

# Tabla con medias reordenadas
knitr::kable(cluster_summary, digits = 2)
```

Tabla 4: Resumen comparativo de variables por clúster

cluster	Assault	Murder	Rape	UrbanPop
1	-0.68	-0.67	-0.56	-0.13
2	1.01	1.00	0.85	0.20

La Tabla 4 presenta las medias estandarizadas de cada variable en los grupos identificados mediante k-means (k = 2).

El Cluster 1 muestra valores negativos en todas las variables (por ejemplo, *Assault* = -0.68 , *Murder* = -0.67 , *Rape* = -0.56), lo que indica estados con niveles de criminalidad inferiores al promedio nacional. Su valor en *UrbanPop* (-0.13) también se sitúa ligeramente por debajo de la media, lo que sugiere una menor proporción de población urbana.

En contraste, el Cluster 2 presenta valores positivos en todas las variables (por ejemplo, *Assault* = 1.01 , *Murder* = 1 , *Rape* = 0.85), indicando estados con mayores tasas de criminalidad. Su leve incremento en *UrbanPop* (0.2) apunta a una tendencia hacia contextos más urbanizados, aunque esta diferencia es menos marcada.

En conjunto, los resultados evidencian que la separación entre grupos responde principalmente a un eje de intensidad delictiva, más que a diferencias en urbanización o tamaño poblacional.

6 Clustering jerárquico (HC)

```r
# Cortar el dendrograma en 2 grupos
grupos_hc <- cutree(hc, k = 2)

# Calcular medias estandarizadas por grupo
hc_summary <- num_scaled %>%
  mutate(cluster = grupos_hc) %>%
  pivot_longer(-cluster, names_to = "variable", values_to = "valor") %>%
  group_by(cluster, variable) %>%
  summarise(mean = mean(valor), .groups = "drop") %>%
  pivot_wider(names_from = variable, values_from = mean)

# Mostrar tabla resumen
knitr::kable(hc_summary, digits = 2)
```

Tabla 5: Resumen comparativo de variables por clúster (clustering jerárquico)

cluster	Assault	Murder	Rape	UrbanPop
1	1.06	1.04	0.85	0.19
2	-0.65	-0.64	-0.52	-0.12

La Tabla 5 resume las medias estandarizadas de las variables consideradas en el análisis jerárquico.

El Cluster 1 presenta valores positivos en todas las variables (*Assault* = 1.06 , *Murder* = 1.04 , *Rape* = 0.85 , *UrbanPop* = 0.19), lo que indica un grupo de estados con niveles relativamente altos de criminalidad y una proporción de población urbana ligeramente superior a la media. Este conjunto representa los contextos con mayor incidencia de delitos violentos dentro del país.

Por el contrario, el Cluster 2 muestra valores negativos en todas las dimensiones (*Assault* = -0.65 , *Murder* = -0.64 , *Rape* = -0.52 , *UrbanPop* = -0.12), lo que sugiere estados caracterizados por menores tasas de homicidios, asaltos y delitos sexuales, así como una urbanización ligeramente inferior al promedio nacional.

En conjunto, el patrón obtenido revela un contraste claro entre regiones de alta y baja criminalidad, coherente con la segmentación previamente identificada mediante el algoritmo k-means.

```r
# Asignar colores según interpretación previa
colores <- c("firebrick", "forestgreen")

# Graficar dendrograma
```

```r
plot(hc, hang = -1, labels = A_raw$region,
     main = "Dendrograma - Ward.D2",
     xlab = "", sub = "")

# Dibujar rectángulos de los dos clústeres
rect.hclust(hc, k = 2, border = colores)

# Agregar leyenda
legend("topright",
       legend = c("Cluster 1: alta criminalidad",
                  "Cluster 2: baja criminalidad"),
       col = colores,
       lwd = 3, cex = 0.9, box.lwd = 0.8, bg = "white")
```

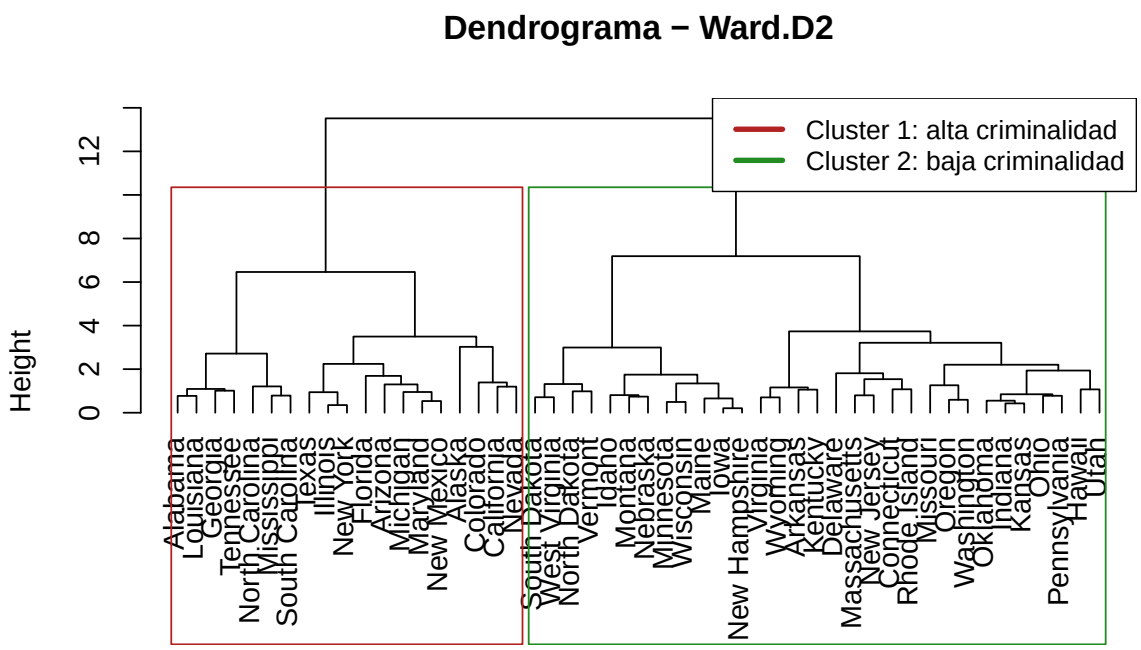

Figura 3: Dendrograma (clustering jerárquico, método Ward.D2 sobre variables estandarizadas)

El dendrograma obtenido refleja la estructura jerárquica de similitud entre los estados, basado en las tasas de criminalidad y urbanización, utilizando el método de enlace Ward.D2 (ver Figura 3).

El Cluster 1 reúne estados del Sur y Sureste de Estados Unidos, como *Alabama, Louisiana, Georgia, Texas* y *Mississippi*, además de grandes áreas urbanas como *California, New York* e *Illinois*. Estos presentan valores positivos en todas las variables, indicando una mayor intensidad de criminalidad y un grado de urbanización ligeramente superior.

Por otro lado, el Cluster 2 grupa principalmente estados del Norte, Centro-Norte y Noreste, como *Montana, Dakota del Norte, Dakota del Sur, Iowa* y *Minnesota*, junto con otros de menor población y

criminalidad relativa. Este grupo se caracteriza por valores negativos en Murder, Assault y Rape, lo que refleja una menor incidencia delictiva y contextos más seguros.

En conjunto, los resultados muestran una división clara entre estados con baja criminalidad y aquellos con niveles más altos, que coincide parcialmente con un gradiente geográfico norte–sur.

7 Análisis discriminante (LDA)

8 Conclusiones (Clúster / LDA)

El análisis de agrupamiento permitió identificar dos grupos bien diferenciados de estados en el conjunto *USArrests*, basados en variables de criminalidad y urbanización. Un grupo esta conformado por estados con niveles relativamente bajos de homicidios, asaltos y delitos sexuales, junto con una urbanización ligeramente menor. Estos pueden considerarse estados de baja criminalidad, donde los indicadores sociales y demográficos tienden a reflejar mayor estabilidad.

En cambio, el otro grupo está compuesto por estados con valores claramente superiores al promedio en todas las dimensiones de violencia, sugiriendo contextos de alta criminalidad. Este grupo también presenta una leve mayor proporción de población urbana, aunque esta diferencia no parece ser el principal factor de separación entre los grupos.

En conjunto, ambos métodos de agrupamiento (k-means y jerárquico) revelan un contraste consistente entre regiones de alta y baja criminalidad, evidenciando patrones geográficos y socioeconómicos subyacentes en los datos.