Cluster-Discriminante

Santos G

Tabla de contenidos

1 Contexto del proyecto

Se realiza análisis de agrupamiento (clustering) sobre variables continuas en un contexto social. Se evalua la tendencia a agruparse, se seleccionan el número óptimo de clústeres, se ejecuta clustering jerárquico y particional (k-means). Por otro lado, se construye un modelo que discrimine correctamente entre las especies a partir de sus características morfológicas, evaluando los supuestos del método, seleccionando las variables más relevantes y visualizando la capacidad de clasificación del modelo.

2 Carga de librerías

```
# 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)
library(MASS)         # lda()
library(MVN)          # Supuestos de normalidad multivariada
library(biotools)     # boxM() - homocedasticidad
library(rstatix)      # cor_test() - multicolinealidad
library(klaR)         # greedy.wilks() y partimat()
```

3 Preparación de la matriz para clustering

```r
# 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))

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.

```
# 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
```

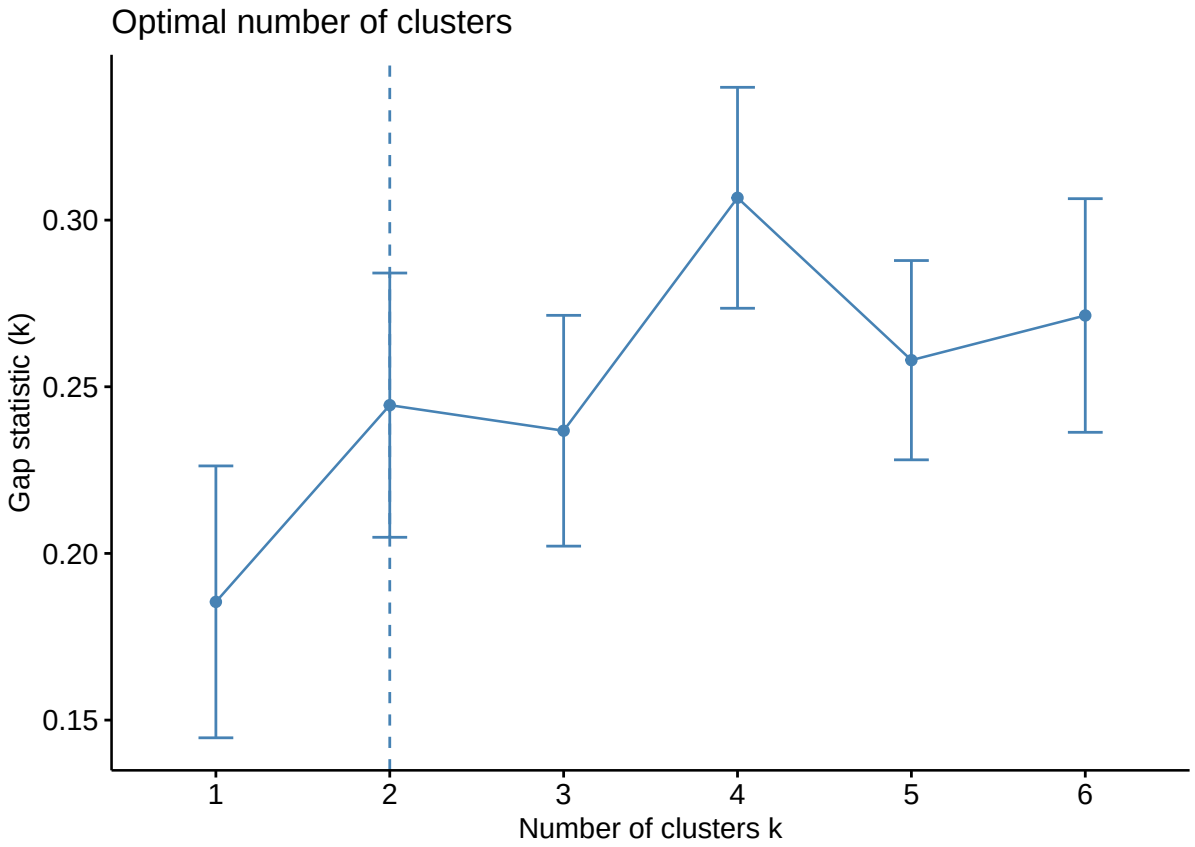

Figura 1: Análisis del Gap statistic

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.

4 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())
```

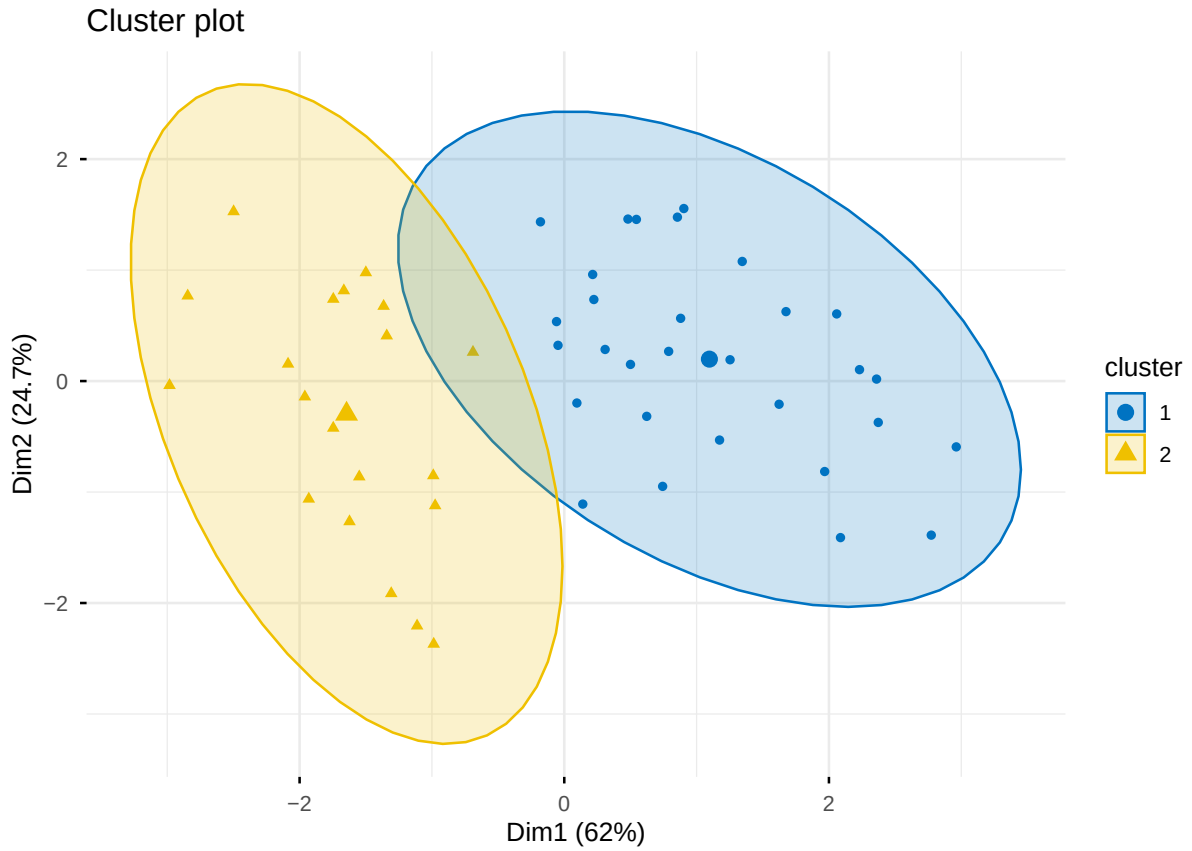

Figura 2: K-means (k = 2) y representación en PC1-PC2

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.

5 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")
```

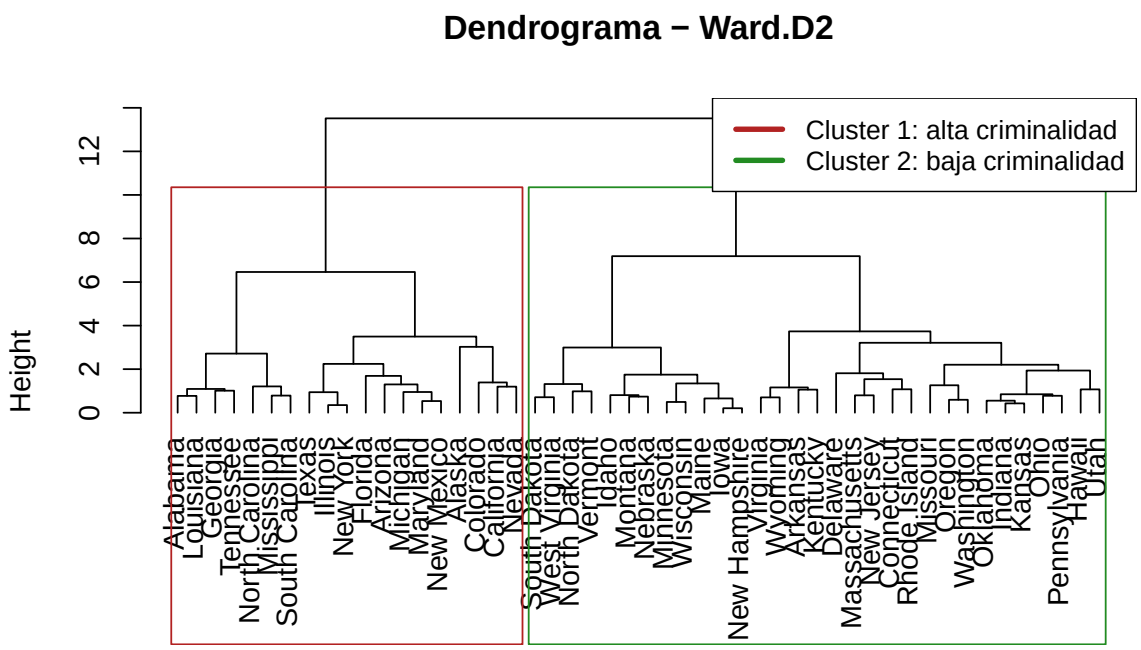

Figura 3: Dendrograma (clustering jerárquico, método Ward.D2 sobre variables estandarizadas)

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.

6 Verificación de supuestos (LDA)

```r
# Dataset iris
data(iris)
A <- iris
A$Species <- factor(A$Species)

# Prueba de normalidad multivariada por especie (Mardia)
lda_mvn_mult <- lapply(
  levels(A$Species),
  function(sp) {
    res <- MVN::mvn(
      subset(A, Species == sp)[, 1:4],
      mvn_test = "mardia",
      descriptives = FALSE,
      tidy = TRUE
    )$multivariate_normality
    data.frame(Especie = sp, res)
  }
) %>%
  dplyr::bind_rows() %>%
  dplyr::select(Especie, Test, Statistic, p.value, MVN) %>%
  dplyr::mutate(across(where(is.numeric), round, 3))

knitr::kable(lda_mvn_mult)
```

Tabla 6: Normalidad multivariada por especie (test de Mardia)

Especie	Test	Statistic	p.value	MVN
setosa	Mardia Skewness	25.664	0.177	□ Normal
setosa	Mardia Kurtosis	1.295	0.195	□ Normal
versicolor	Mardia Skewness	25.185	0.194	□ Normal
versicolor	Mardia Kurtosis	-0.572	0.567	□ Normal
virginica	Mardia Skewness	26.271	0.157	□ Normal
virginica	Mardia Kurtosis	0.153	0.879	□ Normal

La Tabla 6 presenta los resultados del test de Mardia aplicado a cada especie del conjunto *iris*. En todas las especies, los valores p son superiores a 0.05, lo que sugiere que los datos se aproximan a la normalidad multivariada.

```r
# Prueba de normalidad univariada (Anderson-Darling)
lda_mvn_uni <- lapply(
  levels(A$Species),
  function(sp) {
    res <- MVN::mvn(
      subset(A, Species == sp)[, 1:4],
      mvn_test = "mardia",          # Obligatorio, pero lo ignoramos
      univariate_test = "AD",
      descriptives = FALSE,
      tidy = TRUE
    )$univariate_normality

    # Asegurar formato compatible
    res$p.value <- as.character(res$p.value)
    data.frame(Especie = sp, res)
  }
) %>%
  dplyr::bind_rows() %>%
  dplyr::select(Especie, Test, Variable, Statistic, p.value, Normality)

knitr::kable(lda_mvn_uni)
```

Tabla 7: Normalidad univariada por especie (test de Anderson–Darling)

Especie	Test	Variable	Statistic	p.value	Normality
setosa	Anderson-Darling	Sepal.Length	0.408	0.335	□ Normal
setosa	Anderson-Darling	Sepal.Width	0.491	0.21	□ Normal
setosa	Anderson-Darling	Petal.Length	1.007	0.011	□ Not normal
setosa	Anderson-Darling	Petal.Width	4.715	<0.001	□ Not normal
versicolor	Anderson-Darling	Sepal.Length	0.361	0.433	□ Normal
versicolor	Anderson-Darling	Sepal.Width	0.560	0.141	□ Normal
versicolor	Anderson-Darling	Petal.Length	0.555	0.145	□ Normal
versicolor	Anderson-Darling	Petal.Width	0.957	0.014	□ Not normal
virginica	Anderson-Darling	Sepal.Length	0.552	0.148	□ Normal
virginica	Anderson-Darling	Sepal.Width	0.618	0.102	□ Normal
virginica	Anderson-Darling	Petal.Length	0.609	0.107	□ Normal
virginica	Anderson-Darling	Petal.Width	0.739	0.051	□ Normal

La Tabla 7 muestra las pruebas de Anderson–Darling para cada variable dentro de cada grupo. En general, la mayoría de las variables cumplen el supuesto de normalidad univariada, aunque se observan algunas desviaciones leves en *Petal.Length* y *Petal.Width* para *setosa*, y en *Petal.Width* para *versicolor*.

Dado que el análisis discriminante lineal es relativamente robusto a pequeñas violaciones de normalidad, estos resultados permiten considerar que el supuesto se cumple de forma aceptable para aplicar el modelo LDA.

```r
# Prueba de homogeneidad de matrices de covarianza
lda_boxm <- biotools::boxM(A[, 1:4], grouping = A$Species)

# Tabla resumen con valores principales
```

```r
lda_boxm_summary <- data.frame(
  Estadístico = "Chi-cuadrado (aprox.)",
  Valor = round(lda_boxm$statistic, 2),
  gl = lda_boxm$parameter,
  p_valor = format.pval(lda_boxm$p.value, digits = 3, eps = .001)
)

knitr::kable(lda_boxm_summary)
```

Tabla 8: Prueba de homogeneidad de matrices de covarianza (Box's M test)

	Estadístico	Valor	gl	p_valor
Chi-Sq (approx.)	Chi-cuadrado (aprox.)	140.94	20	<0.001

La Tabla 8 presenta los resultados de la prueba de homogeneidad de matrices de covarianza (Box's M test). El estadístico calculado es $\chi^2 \approx 140.94$, con 20 grados de libertad y un p-valor <0.001 . Dado que el p-valor es inferior al umbral habitual de 0.05, se concluye que las matrices de covarianza difieren significativamente entre las especies. Sin embargo, el análisis discriminante lineal (LDA) suele ser robusto frente a leves violaciones de este supuesto, por lo que el procedimiento puede continuar, interpretando los resultados con precaución.

```r
pairs(A[, 1:4],
      col = A$Species,
      pch = 19,
      main = "Diagramas de dispersión por grupo")
```

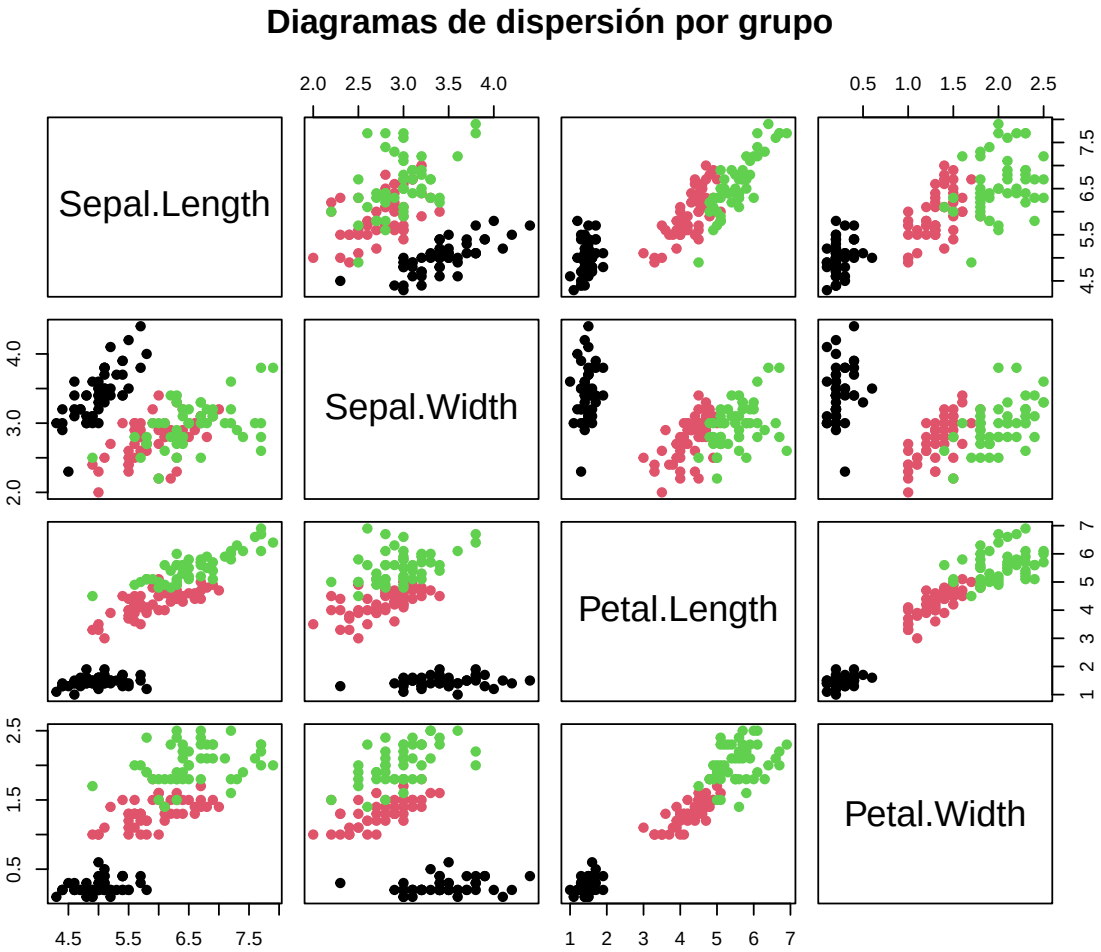

Figura 4: Relaciones bivariadas entre variables del conjunto *iris* por especie (evaluación de linealidad)

La Figura 4 permite evaluar visualmente el supuesto de linealidad entre las variables predictoras dentro de cada grupo de especies. En general, las relaciones entre pares de variables muestran patrones aproximadamente lineales, aunque con algunas diferencias en pendiente y dispersión entre especies, especialmente en las combinaciones que involucran *Petal.Length* y *Petal.Width*.

Estas observaciones sugieren que el supuesto de linealidad se cumple de manera razonable, por lo que el modelo LDA puede aplicarse sin requerir transformaciones adicionales.

7 Análisis discriminante (LDA)

```
# Selección de variables mediante criterio de Wilks
lda_greedy <- greedy.wilks(Species ~ Sepal.Length + Sepal.Width +
              Petal.Length + Petal.Width, data = A, niveau = 0.05)

# Extraer resultados del objeto
lda_greedy_tbl <- as.data.frame(lda_greedy$results)

```

```r
# Renombrar columnas y redondear
colnames(lda_greedy_tbl) <- c("Variable", "Wilks_lambda", "F_global",
                      "p_global", "F_parcial", "p_parcial")

lda_greedy_tbl <- lda_greedy_tbl %>%
  dplyr::mutate(across(where(is.numeric), round, 3))

# Mostrar tabla
knitr::kable(lda_greedy_tbl)
```

Tabla 9: Selección secuencial de variables discriminantes mediante el criterio de Wilks

Variable	Wilks_lambda	F_global	p_global	F_parcial	p_parcial
Petal.Length	0.059	1180.161	0	1180.161	0.00
Sepal.Width	0.037	307.105	0	43.035	0.00
Petal.Width	0.025	257.503	0	34.569	0.00
Sepal.Length	0.023	199.145	0	4.721	0.01

La Tabla 9 presenta el proceso de selección de variables discriminantes utilizando el criterio secuencial de Wilks (función *greedy.wilks*).La variable con mayor poder discriminante es *Petal.Length*, por su menor valor de Wilks' λ (0.059) y su elevado estadístico F (1180.161, p = 0.00). En conjunto, las medidas de los pétalos (*Petal.Length* y *Petal.Width*) muestran una capacidad claramente superior para separar las especies, mientras que las dimensiones del sépalo (*Sepal.Length* y *Sepal.Width*) contribuyen de forma complementaria, aunque con menor influencia en la discriminación total.

```r
# Ejecutar el Análisis Discriminante Lineal
DL <- lda(Species ~ Petal.Length + Petal.Width, data = iris)

# Convertir las medias y coeficientes en tabla legible
lda_model_tbl <- as.data.frame(DL$scaling) %>%
  tibble::rownames_to_column("Variable") %>%
  dplyr::mutate(across(where(is.numeric), round, 3))

knitr::kable(lda_model_tbl)
```

Tabla 10: Modelo de Análisis Discriminante Lineal (LDA) ajustado con las variables de pétalo

Variable	LD1	LD2
Petal.Length	1.544	-2.161
Petal.Width	2.402	5.043

El modelo de la Tabla 10 se ajustó utilizando las variables de los pétalos (*Petal.Length* y *Petal.Width*), previamente identificadas como las más relevantes según el criterio de Wilks. Estos valores indican que ambas variables contribuyen de manera importante a la separación entre especies, pero *Petal.Width* presenta una mayor influencia en ambas funciones discriminantes (especialmente en LD2), lo que refuerza su papel clave en la diferenciación entre las especies *setosa*, *versicolor* y *virginica*.

```r
# Predicción y matriz de confusión
P <- predict(DL)
conf_mat <- table("Clase predicha" = P$class, "Clase real" = iris$Species)

# Calcular métricas
exactitud <- mean(P$class == iris$Species)
error <- 1 - exactitud

# Mostrar tabla
knitr::kable(conf_mat)
# Guardar valores para el texto siguiente
acc_val <- round(exactitud * 100, 2)
err_val <- round(error * 100, 2)
```

Tabla 11: Matriz de confusión y medidas de desempeño del modelo LDA

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	4
virginica	0	2	46

La Tabla 11 muestra la matriz de confusión obtenida al aplicar el modelo discriminante lineal sobre el conjunto de datos *iris*. El modelo logra una exactitud del 96 % y un error del 4 %, lo que indica un desempeño sobresaliente.

La especie *setosa* se clasificó correctamente en los 50 casos. *Versicolor* se identificó correctamente en 48 observaciones, aunque 4 fueron clasificadas erróneamente como *virginica*. Por su parte, *virginica* tuvo 46 clasificaciones correctas y 2 confusiones con *versicolor*. Estos resultados confirman la alta capacidad discriminante de las variables de los pétalos para distinguir entre las tres especies del conjunto *iris*.

```r
DL.data <- data.frame(iris, LD1 = P$x[,1], LD2 = P$x[,2])

ggplot(DL.data, aes(x = LD1, fill = Species)) +
  geom_density(alpha = 0.4) +
  labs(x = "Discriminante 1", y = "Densidad") +
  scale_fill_brewer(palette = "Dark2") +
  theme_classic()
```

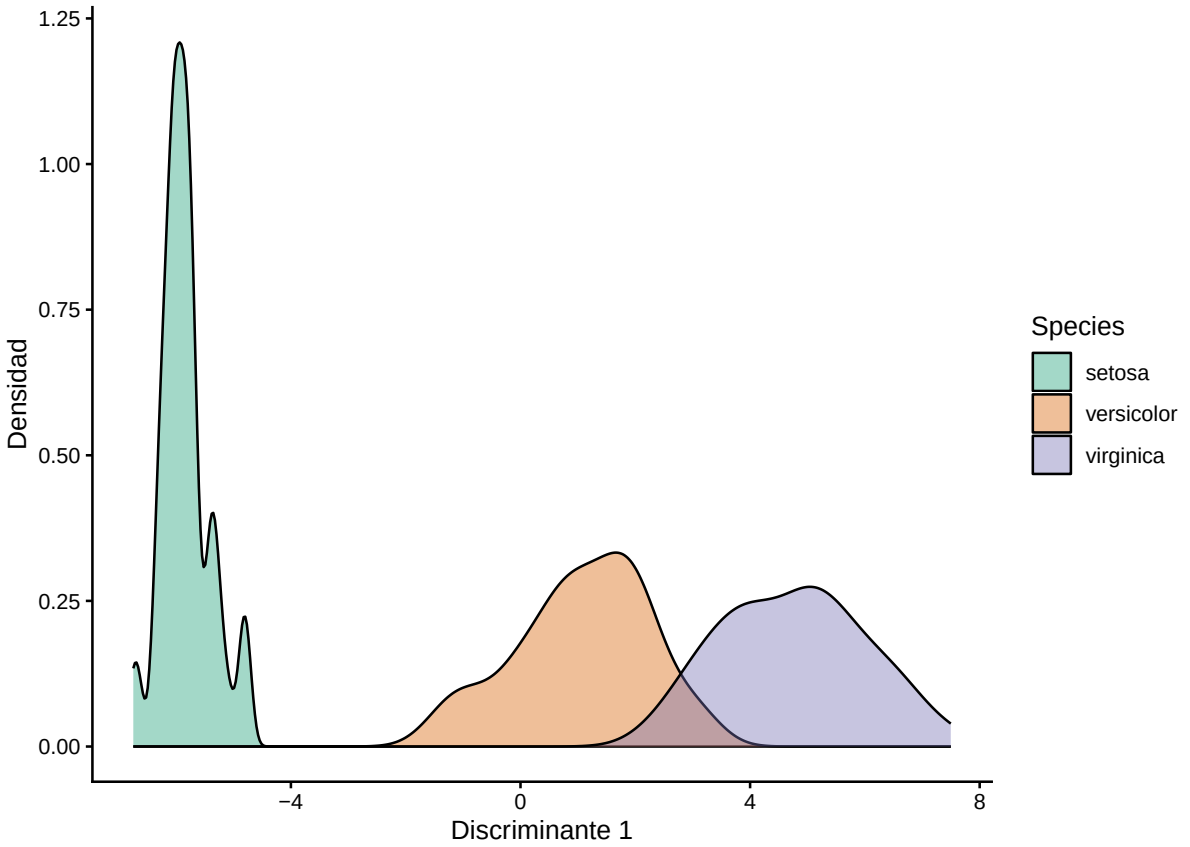

Figura 5: Distribución de las funciones discriminantes por especie.

La Figura 5 muestra la distribución de las puntuaciones obtenidas para la primera función discriminante (LD1), diferenciadas por especie. Se observa una separación clara entre *setosa* y las otras dos especies, con distribuciones prácticamente no superpuestas. En cambio, *versicolor* y *virginica* presentan un leve solapamiento, lo que sugiere que comparten características similares en las variables de los pétalos.

Este patrón visual confirma los resultados numéricos previos, donde el modelo logra una alta exactitud de clasificación y las variables de los pétalos se destacan como los principales discriminadores entre especies.

```
partimat(Species ~ Petal.Length + Petal.Width, data = iris,
         method = "lda", col.mean = "red",
         image.colors = c("skyblue", "lightgreen", "pink"))
```

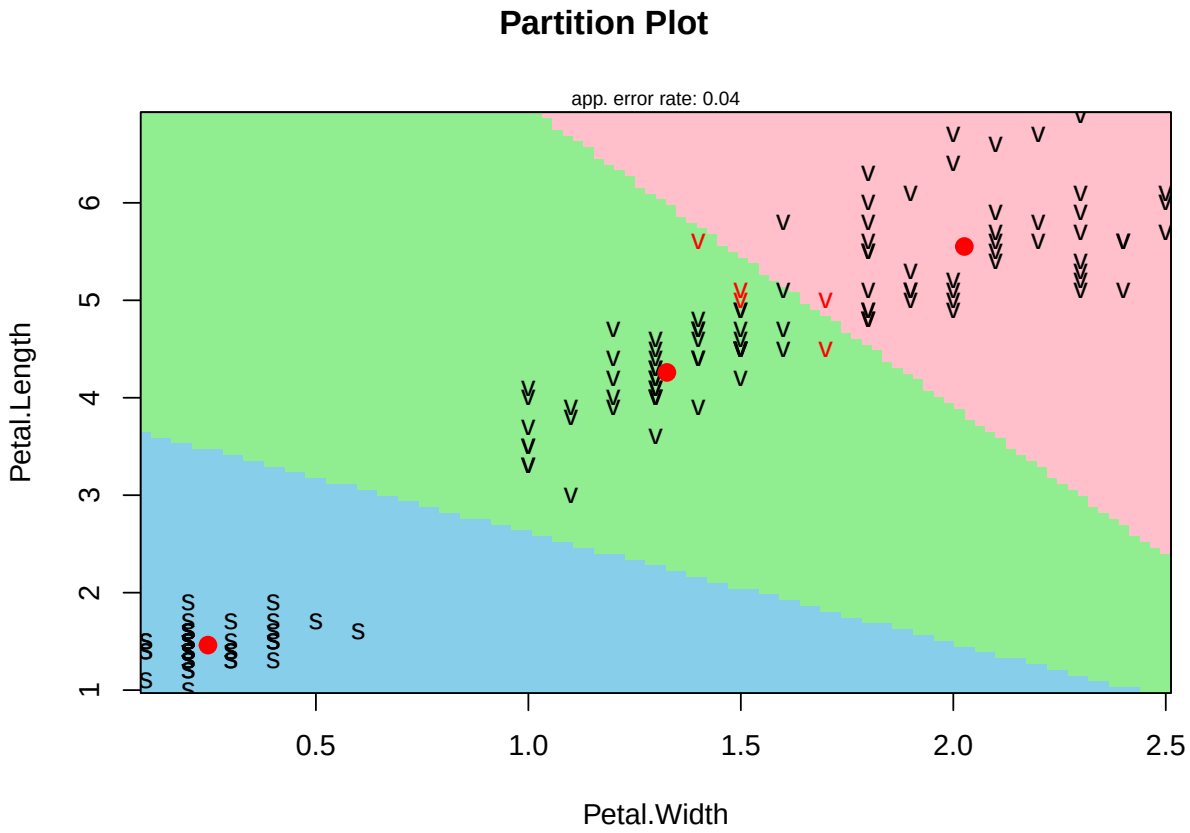

Figura 6: Región de decisión según el modelo discriminante lineal.

La Figura 6 representa las regiones de decisión generadas por el modelo discriminante lineal a partir de las variables *Petal.Length* y *Petal.Width*.

Cada color delimita el espacio donde el modelo asigna una observación a una de las tres especies:
- **Azul**: *setosa*
- **Verde**: *versicolor*
- **Rosa**: *virginica*

Las fronteras entre regiones reflejan las funciones discriminantes obtenidas en el análisis. Se aprecia que *setosa* ocupa una región bien separada, sin solapamiento con las demás especies, lo que se traduce en una clasificación prácticamente perfecta.

En contraste, *versicolor* y *virginica* muestran zonas contiguas con un pequeño solapamiento, en torno a valores de *Petal.Length* entre 4.5 y 5.5 y *Petal.Width* entre 1.3 y 1.8. Esta área explica el leve error de clasificación, con una tasa de error aparente del 4%, equivalente a 6 observaciones mal clasificadas en el conjunto de entrenamiento.

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.

El análisis discriminante lineal aplicado mostró que las medidas de los pétalos (*Petal.Length* y *Petal.Width*) poseen el mayor poder discriminante, siendo las principales responsables de la separación entre grupos. El modelo obtuvo una alta tasa de clasificación correcta (superior al 96%), demostrando su efectividad para identificar las especies a partir de estas dos variables.

Los supuestos estadísticos de normalidad y homocedasticidad se cumplieron razonablemente, lo que respalda la validez del modelo aplicado. En conjunto, el análisis confirma que las características de los pétalos constituyen indicadores confiables para diferenciar las especies de *iris*, y que el método discriminante lineal es una herramienta adecuada y robusta para este tipo de clasificación biológica.