Tarea 4

Estudiante

John Daniel hoyos Arias
Ivan Santiago Rojas Martinez
Genaro Alfonso Aristizabal Echeverri

Docente

Cesar Augusto Gomez Velez

Asignatura

Analitica de datos

Sede Medellín
29 de Noviembre del 2022

Índice

1. Ejercicio 1

1. Este ejercicio utiliza el conjunto de datos OJ el cual es parte de la librería ISLR

1.1. a) Cree un conjunto de entrenamiento con una muestra aleatoria de 800 observaciones y un conjunto de prueba que conste del resto de observaciones.

Se procede a cargar las librerías necesarias del **R** y a crear un conjunto de entrenamiento de **800** datos para prueba y **270** datos para entrenamiento fijando una semilla = **1** la cual permitirá la replicabilidad de nuestro informe.

```
require(ISLR)
require(tidyverse)
require(ggthemes)
require(caret)
require(e1071)
require(kableExtra)

set.seed(1)

data('OJ')

inTrain <- sample(nrow(OJ), 800, replace = FALSE)

training <- OJ[inTrain,]
testing <- OJ[-inTrain,]
```

1.2. b) Ajuste un clasificador de soporte vectorial utilizando cost = 0.1, con Purchase como la variable respuesta y las demás como predictores.

1.2.1. Utilice la función summary() para obtener un resumen de estadísticas y describa los resultados obtenidos.

Tomando como variable respuesta **Purchase**. Se ajusta un clasificador de soporte vectorial lineal **(SVM Linear)** con un parámetro de **cost = 0.1**

```
##
## Call:
## svm(formula = Purchase ~ ., data = training, kernel = "linear", cost = 0.1)
##
```

```
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.1
##
## Number of Support Vectors:  342
##
##  ( 171 171 )
##
##
## Number of Classes:  2
##
## Levels:
##  CH MM
```

Este clasificador **SVM** de kernel **lineal** ha sido utilizado con **cost=0.1**, y se obtienen **342** vectores de soporte, **171** en una clase y **171** en la otra.

1.3. c) Que tasas de error de entrenamiento y de prueba obtiene?.

```
##  Accuracy     Kappa
## 0.8350000 0.6532361
```

Se obtiene una tasa de precisión del **83.5 %** para el conjunto de entrenamiento.

```
## Accuracy    Kappa
## 0.837037 0.640914
```

Se obtiene una tasa de precisión del **83.7 %** para el conjunto de prueba.

Estos nos indica que este clasificador de soporte vectorial de kernel lineal tiene una alta capacidad predictiva.

1.4. d)Utilice la función tune() para obtener un valor óptimo del parámetro cost. Considere valores en el rango de 0.01 a 10.

```
## Support Vector Machines with Linear Kernel
##
## 800 samples
##  17 predictor
##   2 classes: 'CH', 'MM'
##
```

```
## Pre-processing: centered (17), scaled (17)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 720, 721, 721, 719, 719, 721, ...
## Resampling results across tuning parameters:
##
##   cost        Accuracy   Kappa
##    0.0100000  0.8249211  0.6321557
##    0.5357895  0.8299523  0.6429017
##    1.0615789  0.8287332  0.6403847
##    1.5873684  0.8287490  0.6405113
##    2.1131579  0.8324836  0.6485113
##    2.6389474  0.8287490  0.6403954
##    3.1647368  0.8300148  0.6428238
##    3.6905263  0.8287648  0.6404514
##    4.2163158  0.8262490  0.6352704
##    4.7421053  0.8275148  0.6375998
##    5.2678947  0.8275148  0.6375998
##    5.7936842  0.8262490  0.6346733
##    6.3194737  0.8275148  0.6371017
##    6.8452632  0.8287336  0.6395174
##    7.3710526  0.8262490  0.6338052
##    7.8968421  0.8274990  0.6366070
##    8.4226316  0.8274990  0.6366070
##    8.9484211  0.8262490  0.6342346
##    9.4742105  0.8262648  0.6343594
##   10.0000000  0.8262648  0.6343594
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cost = 2.113158.
```

1.5. e) Calcule nuevamente las tasas de error de entrenamiento y de prueba usando el valor óptimo obtenido de cost.

```
## Accuracy    Kappa
## 0.833750 0.650414
```

```
##  Accuracy     Kappa
## 0.8481481 0.6660633
```

1.5.1. Repita items de (b) hasta (e) ajustando esta vez una máquina de soporte vectorial (svm) con un nucle radial. Utilizando el valor de default paray

```
##
```

```
## Call:
## svm(formula = Purchase ~ ., data = training, method = "radial", cost = 0.1)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  0.1
##
## Number of Support Vectors:  541
##
##  ( 272 269 )
##
##
## Number of Classes:  2
##
## Levels:
##  CH MM
```

```
##  Accuracy     Kappa
## 0.8262500 0.6288385
```

```
##  Accuracy     Kappa
## 0.7962963 0.5502453
```

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 800 samples
##  17 predictor
##   2 classes: 'CH', 'MM'
##
## Pre-processing: centered (17), scaled (17)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 720, 720, 721, 720, 719, 720, ...
## Resampling results across tuning parameters:
##
##   C          Accuracy   Kappa
##    0.0100000  0.6062566  0.0000000
##    0.5357895  0.8200524  0.6151608
##    1.0615789  0.8238340  0.6233726
##    1.5873684  0.8300377  0.6368084
##    2.1131579  0.8213340  0.6186170
##    2.6389474  0.8213340  0.6172953
##    3.1647368  0.8150682  0.6041695
```

```
##    3.6905263  0.8163182  0.6070327
##    4.2163158  0.8175682  0.6100368
##    4.7421053  0.8163182  0.6071736
##    5.2678947  0.8163024  0.6069460
##    5.7936842  0.8163182  0.6063928
##    6.3194737  0.8175682  0.6087877
##    6.8452632  0.8163336  0.6058321
##    7.3710526  0.8163336  0.6058321
##    7.8968421  0.8175836  0.6081607
##    8.4226316  0.8163178  0.6055942
##    8.9484211  0.8150832  0.6037406
##    9.4742105  0.8138174  0.6012044
##   10.0000000  0.8125674  0.5982272
##
## Tuning parameter 'sigma' was held constant at a value of 0.05
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were sigma = 0.05 and C = 1.587368.
```

```
##  Accuracy     Kappa
## 0.8537500 0.6889952
```

```
##  Accuracy     Kappa
## 0.8222222 0.6082699
```

1.5.2. Repita items (b) hasta (e) utilizando nuevamente una máquina de soporte vectorial pero esta vez con un nucleo polinomial, usando degree = 2.

```
##
## Call:
## svm(formula = Purchase ~ ., data = training, method = "polynomial",
##     degree = 2, cost = 0.01)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  0.01
##
## Number of Support Vectors:  634
##
##  ( 319 315 )
##
##
## Number of Classes:  2
```

```
##
## Levels:
##  CH MM

## Accuracy    Kappa
##  0.60625  0.00000

##  Accuracy     Kappa
## 0.6222222 0.0000000

## Support Vector Machines with Polynomial Kernel
##
## 800 samples
##  17 predictor
##   2 classes: 'CH', 'MM'
##
## Pre-processing: centered (17), scaled (17)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 720, 720, 719, 719, 720, 720, ...
## Resampling results across tuning parameters:
##
##   C           Accuracy   Kappa
##    0.0100000  0.8148953  0.6030074
##    0.5357895  0.8111912  0.5975658
##    1.0615789  0.8074254  0.5906020
##    1.5873684  0.8061596  0.5881041
##    2.1131579  0.8036596  0.5820392
##    2.6389474  0.8036596  0.5819997
##    3.1647368  0.8036754  0.5818901
##    3.6905263  0.8061599  0.5867144
##    4.2163158  0.8049254  0.5843108
##    4.7421053  0.8049254  0.5843108
##    5.2678947  0.8061445  0.5867235
##    5.7936842  0.8073945  0.5891632
##    6.3194737  0.8061287  0.5862157
##    6.8452632  0.8061445  0.5862129
##    7.3710526  0.8074103  0.5891634
##    7.8968421  0.8049099  0.5840027
##    8.4226316  0.8061599  0.5867553
##    8.9484211  0.8061599  0.5867553
##    9.4742105  0.8061599  0.5867553
##   10.0000000  0.8074416  0.5901102
##
## Tuning parameter 'degree' was held constant at a value of 2
```

```
## Tuning
##  parameter 'scale' was held constant at a value of TRUE
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were degree = 2, scale = TRUE and C = 0.01.
```

```
## Accuracy    Kappa
## 0.850000 0.678295
```

```
##  Accuracy     Kappa
## 0.8148148 0.5886654
```

1.6. h) En general cúal método parece proporcionar los mejores resultados en estos datos?.

En general, los modelos son muy similares, pero los núcleo lineal y radial funcionan mejor por un pequeño margen.

2. Ejercicio 2

Se considera el conjunto de datos **USArrests**. En este ejercicio se agruparán los estados en **USArrests** con agrupamiento jerarquico. Este conjunto de datos contiene estadísticas, en arrestos por cada 100,000 residentes por agresión, asesinato y violación en cada uno de los 50 estados de EE. UU. en 1973. También se proporciona el porcentaje de la población que vive en áreas urbanas.

Primeramente, se procede a cargar la base de datos **USArrests** y examinar sus características:

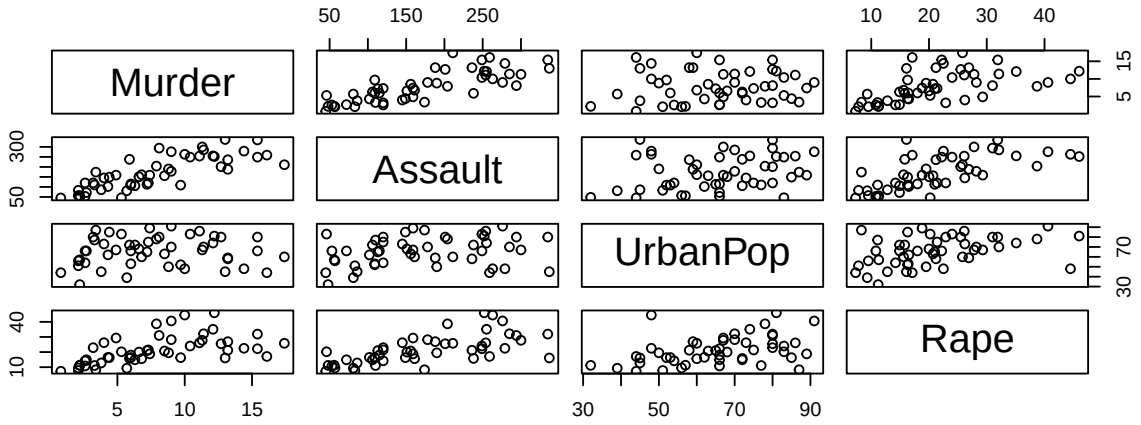
```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

Se observa como la base cuenta con 50 observaciones y 4 variables las cuales todas son numericas y su descripción se presenta seguidamente:

- **Murder:** Arrestos por asesinato (por 100.000).
- **Assault:** Arrestos por asalto (por 100.000).
- **UrbanPop:** Porcentaje de población urbana.
- **Rape:** Arrestos por violaciones (por 100.000).

Adicionalmente, se presentan un análisis descriptivos de estas variables:

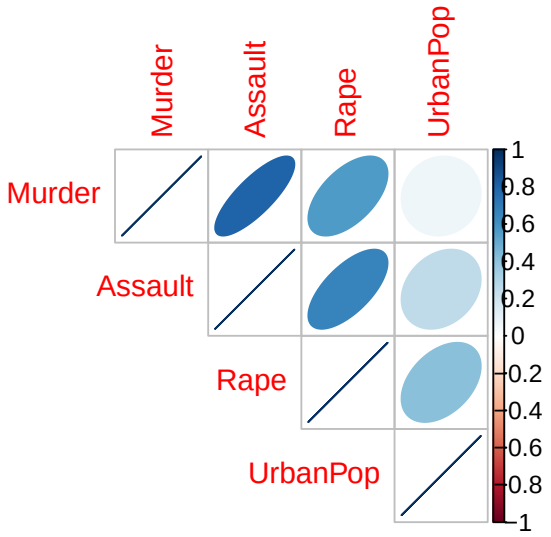


Del anterior gráfico de dispersión entre las variables, se observa como entre cada par de combinación de variables, existe una relación creciente. Lo cual en primera instancia podría ser un indicativo de que posiblemente en los estados donde se presente mayor porcentaje de población urbana también se puede presentar mayores casos de arrestos por asalto, asesinato o violación.

Por otro lado, se presenta una matriz de correlación entre las cuatro variables:

```
##              Murder   Assault   UrbanPop      Rape
## Murder   1.00000000 0.8018733 0.06957262 0.5635788
## Assault  0.80187331 1.0000000 0.25887170 0.6652412
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
## Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

También, se presenta un gráfico de correlaciones de estas variables:



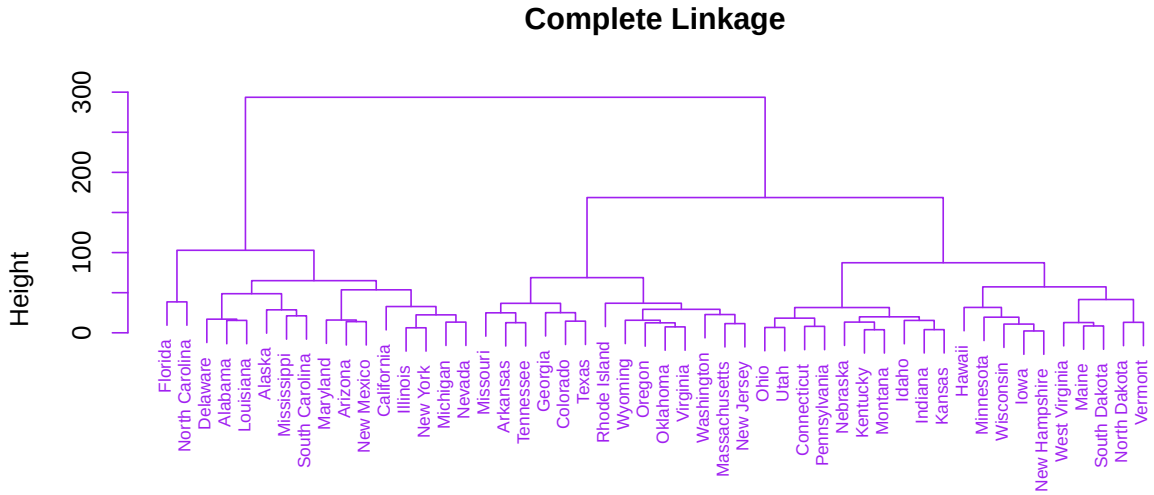
De los resultados obtenidso anteriormente, se observa como:

- Existe una alta correlación positiva entre los arrestos por asesinato y los arrestos por asaltos, la cual es de un 0.8018733, Esto puede indicar que, así como pueden aumentar los arrestos por asalto en un estado de USA, también puede aumentar los arrestos por asesinato en ese mismo estado.
- Existe una alta correlación positiva entre los arrestos por asalto y los arrestos por violaciones, la cual es de un 0.6652412, Esto puede indicar que, así como pueden aumentar los arrestos por asalto en un estado de USA, también puede aumentar los arrestos por violaciones en ese mismo estado.
- Se observa en la matriz de correlaciones como, no existe una aparente correlación significativa entre los arrestos por asesinatos y el porcentaje de población urbana.

2.1. a)

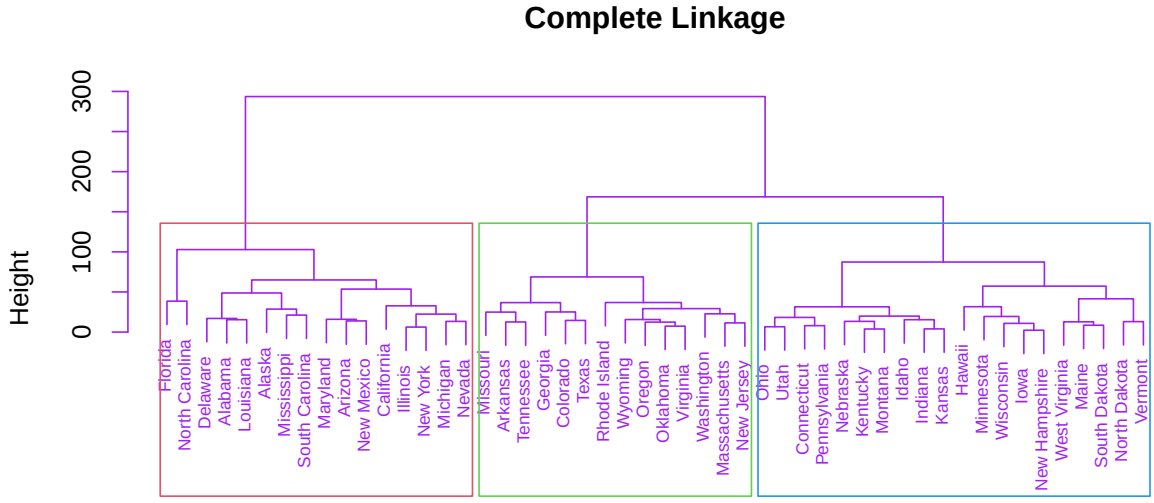
Se utiliza agrupación jerárquica con enlace completo y distancia euclidiana, para agrupar los estados, de la siguiente forma:

Luego se presenta el **dendrograma** de dicho enlace completo:



2.2. b)

Se procede a separar en el **dendograma** a una altura que dé como resultado 3 *clusters*.



Luego, usando la función **cutree()** se puede observar las etiquetas a las que pertenece cada estado según el cluster al que se le asigno.

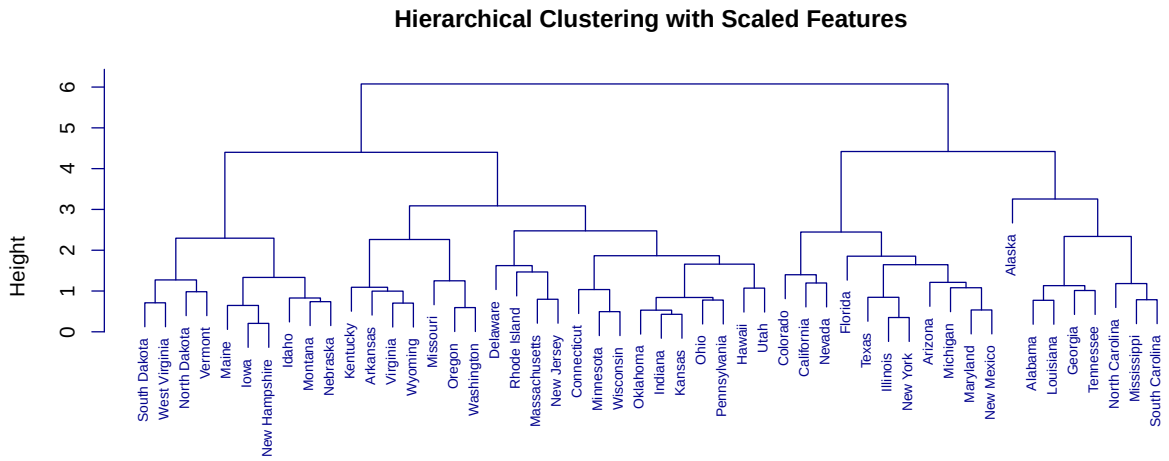
```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

2.3. c)

Ahora se procede a escalar las variables a fin de tener una desviación estándar de uno y luego se realiza la respectiva agrupación jerárquica usando un enlace completo y la distancia euclidiana.

La función **scale()** permite escalar las variables, así como el proceso de agrupación jerárquica se muestra a continuación:

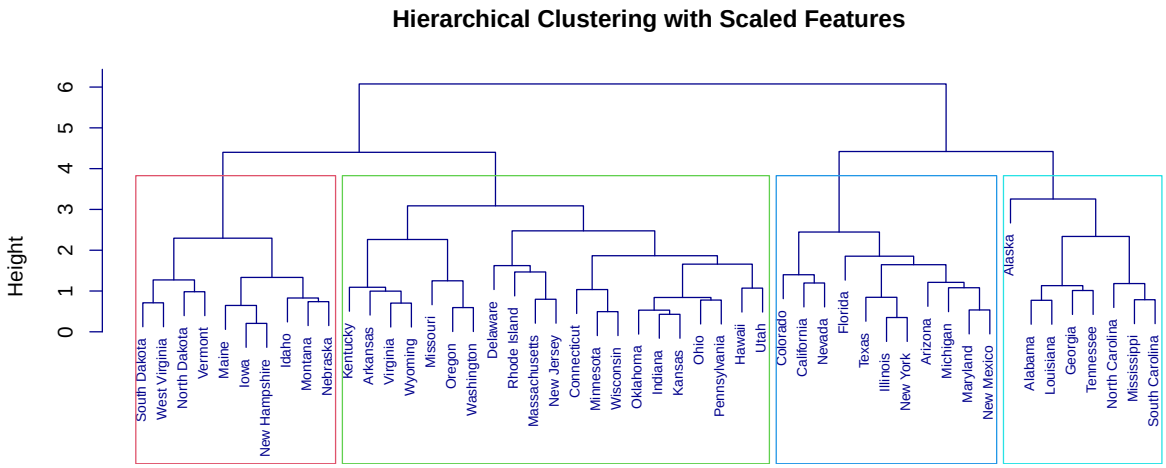
Luego se presenta el **dendrograma** de dicho enlace completo:



2.4. d)

Se observa como, en el **dendrograma** correspondiente a las agrupaciones jerárquicas con las variables escaladas, es notablemente distinto a la agrupación jerárquica generada sin las variables escaladas. Dado que, si bien estamos tratando con los mismos datos, la escalación de variables hace que en cada sub rama existan agrupaciones más uniformes. Inicialmente con las variables sin escalar se observaba claramente una distinción entre tres grupos o *clusters* distintos. En cambio, realizando el escalado de las variables se puede observar una posible distinción entre 4 *clusters*.

A continuación se presenta una posible agrupación entre 4 *clusters*:



Aunque también podría ser una agrupación de tres *clusters*.

En definitiva, se considera que las variables deben ser escaladas previamente, ya que proporciona una mejor estabilidad a la hora de hacer agrupaciones jerárquicas. Porque es bien sabido que la distancia euclidiana no tiene en cuenta el tipo de escala en la cual se encuentran las variables. Lo cual hace que, en ocasiones, usar esta medida no sea del todo preciso y como se pudo observar en los literales anteriores, se obtuvieron *clusters* y **dendrogramas** distintos.