

## **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

<b>1. Ejercicio 1</b>	<b>4</b>
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. . . .	4
1.2. b) Ajuste un clasificador de soporte vectorial utilizando $\text{cost} = 0.1$ , con Purchase como la variable respuesta y las demás como predictores. . . . .	4
1.3. c) Que tasas de error de entrenamiento y de prueba obtiene?. . . . .	4
1.4. d) Utilice la función <code>tune()</code> para obtener un valor óptimo del parámetro <code>cost</code> . Considere valores en el rango de 0.01 a 10. . . . .	5
1.5. e) Calcule nuevamente las tasas de error de entrenamiento y de prueba usando el valor óptimo obtenido de <code>cost</code> . . . . .	6
1.5.1. Repita items de (b) hasta (e) ajustando esta vez una máquina de soporte vectorial (svm) con un núcleo radial. Utilizando el valor de default <code>paray</code> . . . . .	6
1.5.2. Repita items (b) hasta (e) utilizando nuevamente una máquina de soporte vectorial pero esta vez con un núcleo polinomial, usando <code>degree = 2</code> . . . . .	7
1.6. h) En general cuál método parece proporcionar los mejores resultados en estos datos?. . . . .	9
<b>2. Ejercicio 2</b>	<b>9</b>
2.1. a) . . . . .	12
2.2. b) . . . . .	12
2.3. c) . . . . .	14
2.4. d) . . . . .	14

# 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.

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

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

```
##
## 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
```

Esto informa, que el kernel lineal ha sido utilizado con  $\text{cost}=0.1$ , y que se obtuvieron 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
```

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

**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: 721, 720, 720, 720, 721, 719, ...
## Resampling results across tuning parameters:
##
##   cost          Accuracy      Kappa
##   0.0100000    0.8199215    0.6202565
##   0.5357895    0.8273760    0.6360834
##   1.0615789    0.8236101    0.6284665
##   1.5873684    0.8261105    0.6333280
##   2.1131579    0.8261105    0.6333280
##   2.6389474    0.8273605    0.6362121
##   3.1647368    0.8261105    0.6338114
##   3.6905263    0.8248605    0.6309732
##   4.2163158    0.8248605    0.6309732
##   4.7421053    0.8261105    0.6338114
##   5.2678947    0.8273605    0.6361662
##   5.7936842    0.8273605    0.6361662
##   6.3194737    0.8260947    0.6331693
##   6.8452632    0.8260947    0.6331693
##   7.3710526    0.8260947    0.6331693
##   7.8968421    0.8273605    0.6361662
##   8.4226316    0.8273605    0.6361662
##   8.9484211    0.8273605    0.6361662
##   9.4742105    0.8248447    0.6308145
##   10.0000000   0.8248447    0.6308145
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cost = 0.5357895.
```

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

```
## Accuracy      Kappa
## 0.8350000 0.6524601
```

```
## Accuracy      Kappa
## 0.8444444 0.6585983
```

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, 719, 719, 721, 719, 720, ...
## Resampling results across tuning parameters:
##
##      C          Accuracy    Kappa
##  0.0100000  0.6062600  0.0000000
##  0.5357895  0.8249523  0.6261917
##  1.0615789  0.8249527  0.6267051
##  1.5873684  0.8237178  0.6242083
##  2.1131579  0.8199828  0.6162839
##  2.6389474  0.8174982  0.6101331
##  3.1647368  0.8162482  0.6068971
##  3.6905263  0.8149982  0.6045585
##  4.2163158  0.8137324  0.6020878
##  4.7421053  0.8124824  0.5992752
##  5.2678947  0.8124670  0.5992911
##  5.7936842  0.8112170  0.5963850
##  6.3194737  0.8099824  0.5935266
##  6.8452632  0.8099824  0.5935266
##  7.3710526  0.8124828  0.5988491
##  7.8968421  0.8124828  0.5988491
##  8.4226316  0.8124828  0.5988491
##  8.9484211  0.8137486  0.6019866
##  9.4742105  0.8137486  0.6019866
## 10.0000000  0.8124828  0.5995232
##
## 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 = 1.061579.

##      Accuracy      Kappa
## 0.8475000 0.6751539

##      Accuracy      Kappa
## 0.8148148 0.5935694
```

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

```
##
## 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
## 0.622222 0.000000

## 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: 721, 720, 719, 719, 720, 721, ...
## Resampling results across tuning parameters:
##
##      C          Accuracy      Kappa
## 0.010000 0.8249986 0.6233962
## 0.5357895 0.8224662 0.6237344
## 1.0615789 0.8224975 0.6238801
## 1.5873684 0.8162783 0.6107559
## 2.1131579 0.8199816 0.6194664
## 2.6389474 0.8187316 0.6172318
## 3.1647368 0.8187475 0.6166477
## 3.6905263 0.8137625 0.6061815

```

```

##      4.2163158  0.8137467  0.6062677
##      4.7421053  0.8162158  0.6109565
##      5.2678947  0.8174816  0.6133548
##      5.7936842  0.8162158  0.6109565
##      6.3194737  0.8149812  0.6080981
##      6.8452632  0.8149812  0.6080981
##      7.3710526  0.8174816  0.6138489
##      7.8968421  0.8187316  0.6162900
##      8.4226316  0.8174816  0.6139101
##      8.9484211  0.8174816  0.6139101
##      9.4742105  0.8174816  0.6139101
##     10.0000000  0.8174816  0.6139101
##
## 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 cuál 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 jerárquico. 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:



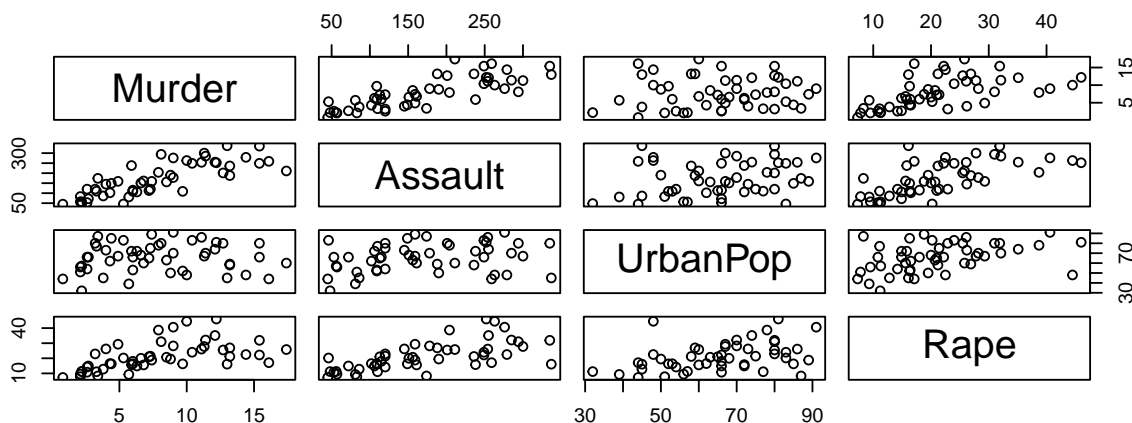
```
##           Murder Assault UrbanPop Rape
## Alabama      13.2      236       58 21.2
## Alaska       10.0      263       48 44.5
## Arizona       8.1      294       80 31.0
## Arkansas      8.8      190       50 19.5
## California    9.0      276       91 40.6
## Colorado     7.9      204      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



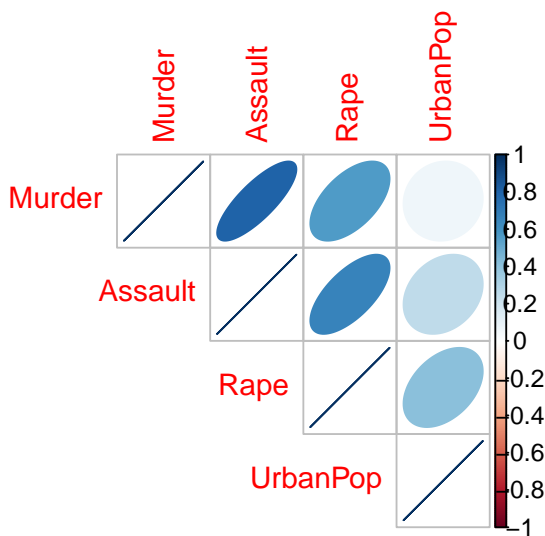
Del anterior gráfico de dispersión entre las variables, se observa como entre cada par de combinación de variables, existe una relación creciente. Lo cual en primera instancia podría ser un indicativo de que posiblemente en los estados donde se presente mayor porcentaje de

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 1.00000000 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:



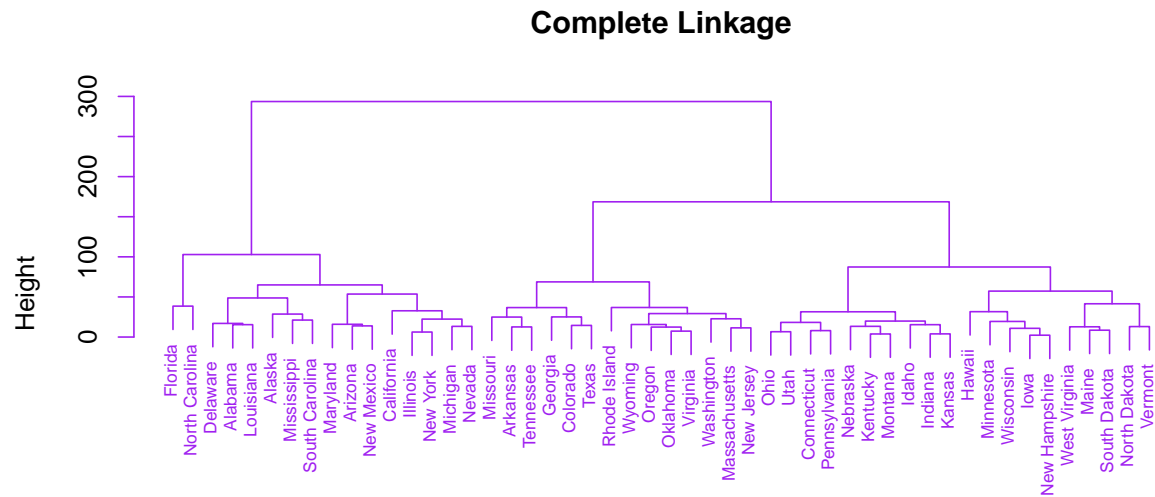
De los resultados obtenidos anteriormente, se observa como:

- 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)

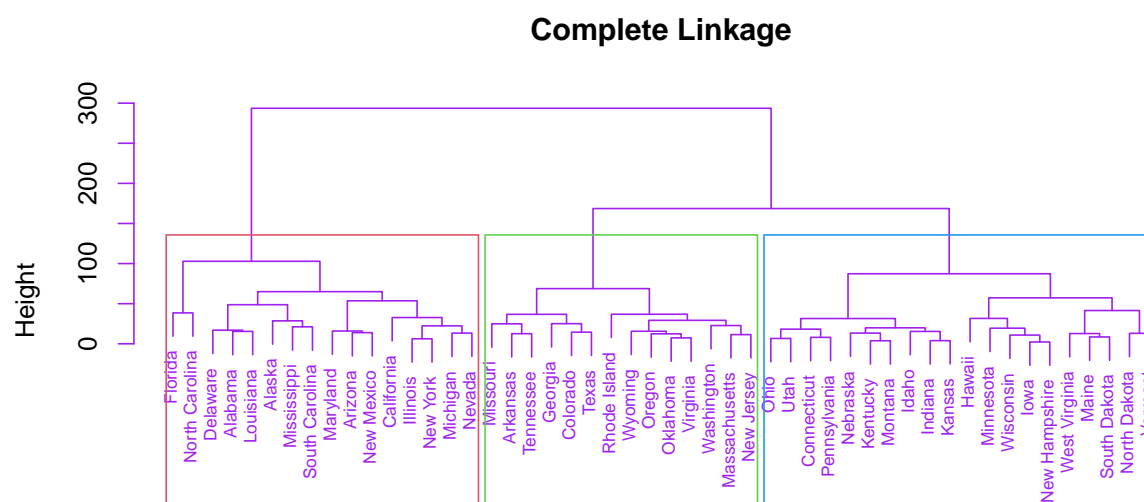
Se utiliza agrupación jerárquica con enlace completo y distancia euclidiana, para agrupar los estados, de la siguiente forma:

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



## 2.2. b)

Se procede a separar en el **dendrograma** 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.

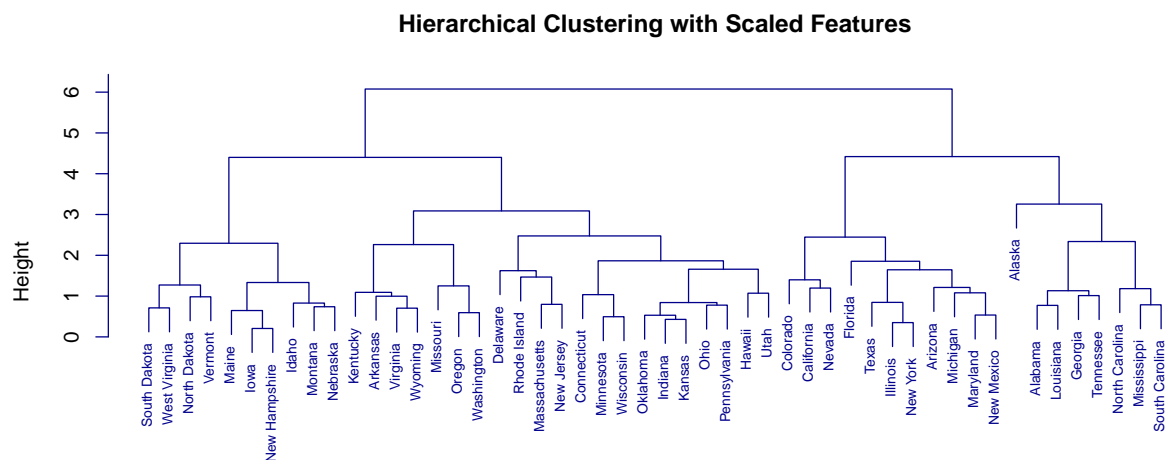
##	Alabama	Alaska	Arizona	Arkansas	California
##	1	1	1	2	1
##	Colorado	Connecticut	Delaware	Florida	Georgia
##	2	3	1	1	2
##	Hawaii	Idaho	Illinois	Indiana	Iowa
##	3	3	1	3	3
##	Kansas	Kentucky	Louisiana	Maine	Maryland
##	3	3	1	3	1
##	Massachusetts	Michigan	Minnesota	Mississippi	Missouri
##	2	1	3	1	2
##	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:

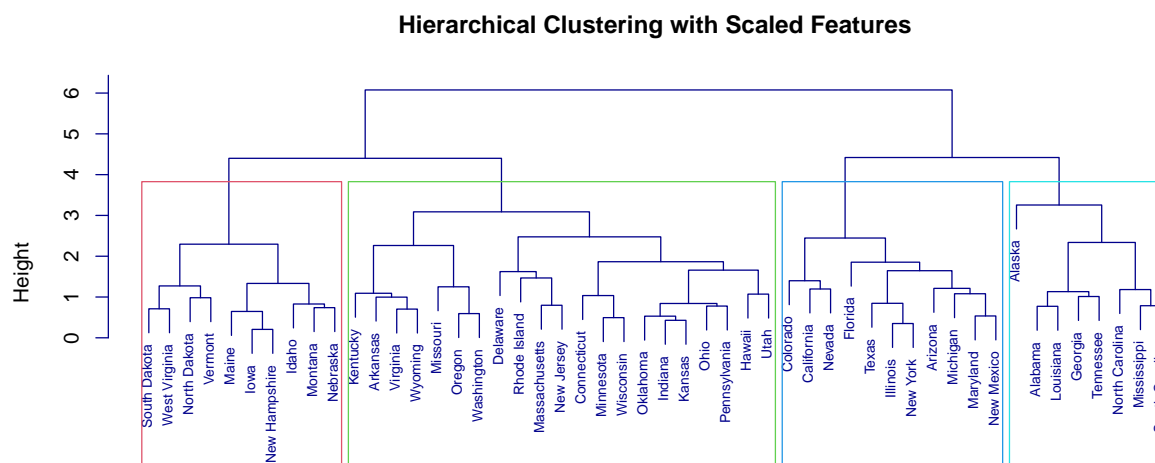
Luego se presenta el **dendrograma** de dicho enlace completo:



### 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.