Unidad 7: Técnicas de Clasificación

Análisis de Clústers

La herramienta que el equipo de campaña debe aplicar para asegurarse de que el candidato lleve la información precisa a quien lo necesita es el *Análisis de Clústers*. Esta es una técnica de que nos permite generar agrupaciones de casos de nuestra base de datos. Esta técnica se enfoca en encontrar semejanzas entre los casos. Esto nos permitirá realizar una descripción más detallada de nuestros casos y focalizar proyectos o estrategias específicas para cada grupo con el que se trabaja. En este proyecto se presentan dos tipos de clústers: jerárquico y no-jerárquico. Veamos cómo funciona el *Análisis de Clústers* en R.

Abriendo base de datos de indicadores financieros de la Unión Europea. Se utiliza el paquete *foreign* para exportar data de SPSS

Clúster Jerárquico:

```
library(foreign)

data<-read.spss("UE.sav",use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
#abriendo data de la Union Europea
```

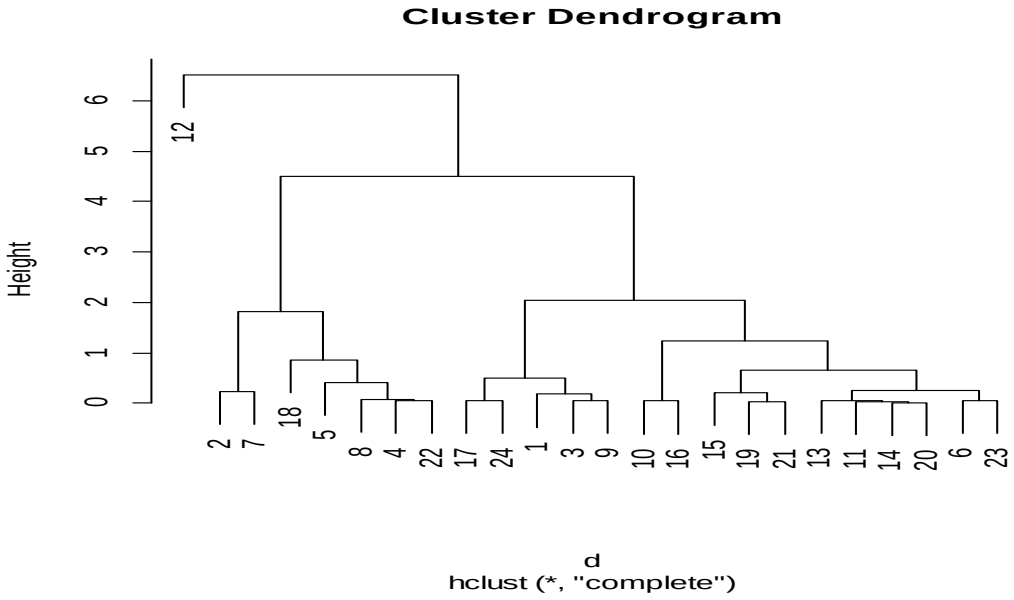
Un requisito para realizar el *Análisis de Clústers* es seleccionar las variables de la base de datos con la que se trabajará. La estructura del clúster está conformada como mínimo por una variable escalar siempre estandarizada, que actuará como dependiente, y una categórica que nos permita incluir los casos. Esto se debe a que la diferenciación de grupos se debe hacer en función a las características que el investigar requiere. En este caso, se seleccionan las variables 1: país (nominal), 19: ZTotalReserves (escalar) y 29: ZGDPDeflator (escalar).

```
data2<-subset(data,select=c(1,19,29), na.value=NULL) #seleccionando las variables a utilizar

str(data2) #mostrando las variables
'data.frame':   30 obs. of  3 variables:
 $ PaÃ.s        : Factor w/ 25 levels "            ",..: 2 4 5 8 9 13 15 16 17 20 ...
 $ ZTotalReserves: num  -0.21 -0.208 -0.209 -0.213 -0.21 ...
 $ ZGDPDeflator  : num  -0.0923 1.9759 -0.1848 0.9583 0.6848 ...
```

```
attach(data2)

pais<- PaÃ.s

mydata <- na.omit(data2)#quitando los casos perdidos
```

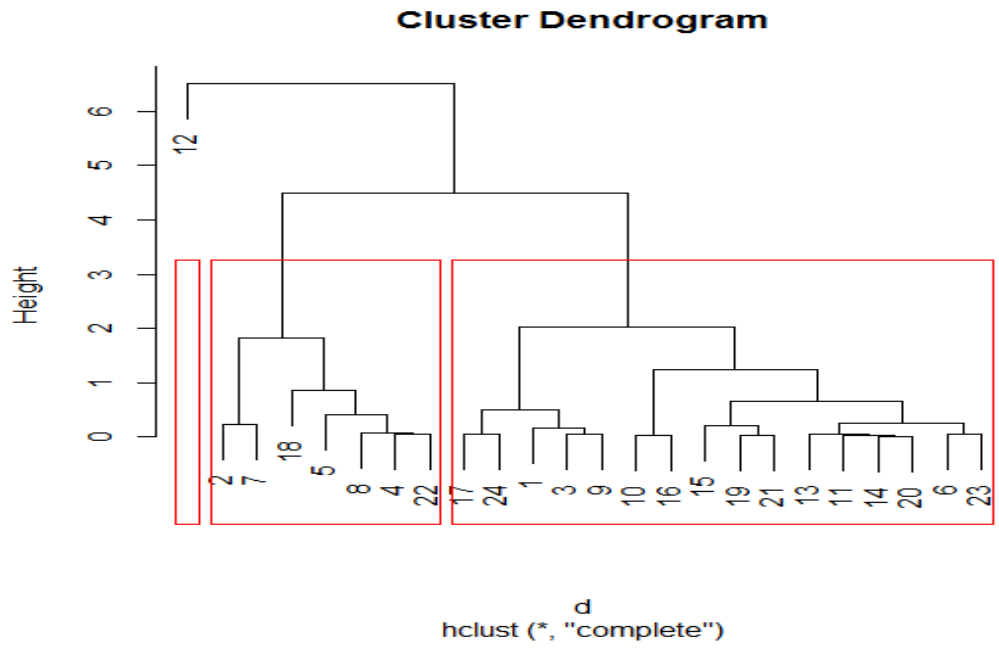
Luego de que se limpiaran los datos, el paso siguiente es la aplicación de la técnica. En primer lugar, se tiene que hallar las distancias entre los casos con el comando *dist.* Después con el comando *hcclust* se aplica la técnica del clúster y se finaliza pidiendo el dendograma con *plot.*

```
d <- dist(as.matrix(mydata))#guardando un nuevo objeto

hc <- hclust(d)#creando cluster

plot(hc)#graficando dendograma
```

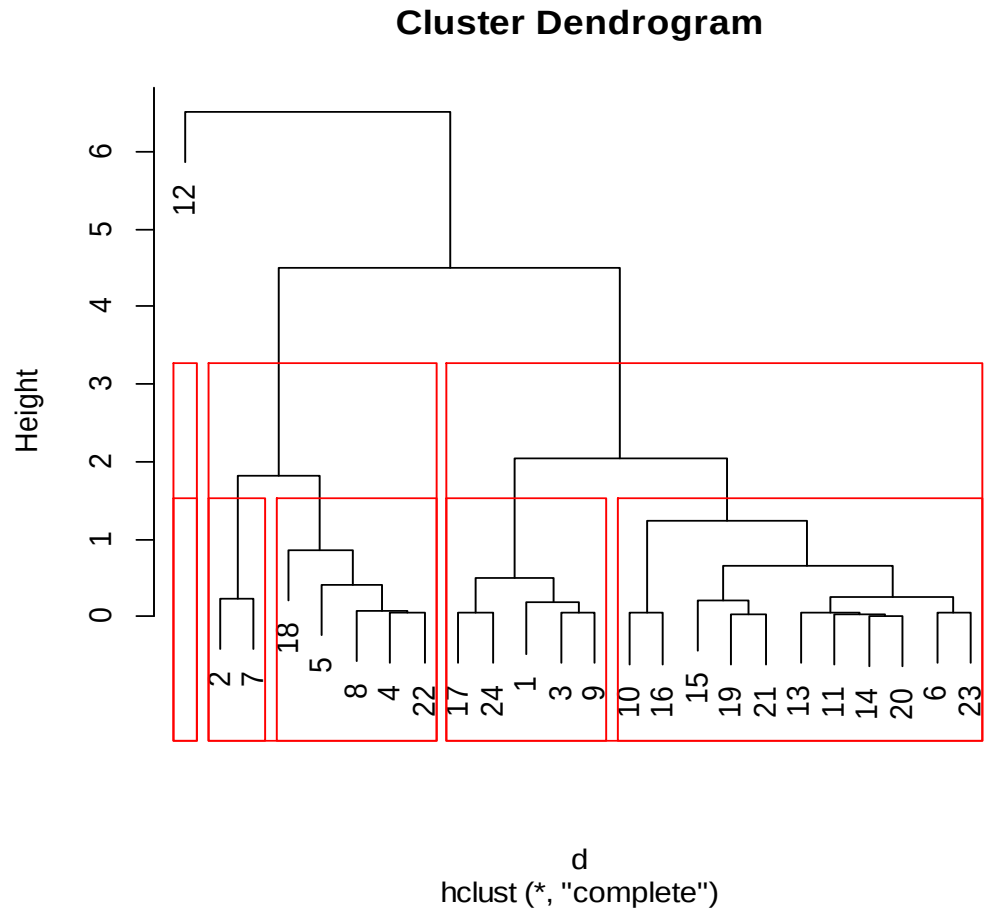


A partir de este dendograma, lo ideal es trazar una línea horizontal imaginaria sobre los niveles en que se entrecruzan los casos para determinar en primera instancia nuestras posibilidades de agrupamiento. Veamos estos dos ejemplos con el comando *rect.hclust* que me permitirá señalar las posibilidades de agrupación.

```
rect.hclust(hc, k=3, border="red")#graficando en tres grupos
```



rect.hclust(hc, k=5, border="red")#graficando en cinco grupos



Las posibilidades de agrupación que se han elegido son un primer grupo de 3 clústers y un segundo grupo de 5 clústers. Un tema a tener en cuenta es si estamos creando grupos distintos. Para esto, los grupos tienen que diferenciarse internamente –entre los mismos casos de un grupo- y externamente –entre los grupos. Este procedimiento se va a verificar utilizando estadísticos que ya hemos visto: varianza y Anova.

El objetivo es comparar las varianzas o las desviaciones estándar de los grupos con respecto a una variable dependiente escalar para comprobar que efectivamente hay diferenciación en los casos del grupo. De igual manera, Anova nos permitirá confirmar la diferencia entre los grupos con respecto a la variable escalar.

```
install.packages("dendextend")#instalando el paquete dendextend para cortar la data

library(dendextend)#ejecutar el paquete dendextend
```

Con el paquete *dendextend* y el comando *cutree* se corta el dendograma y se establece una nueva matriz que incluye las marcas de clase de los clusters 3, 4 y 5. A partir de esta separación, ya es posible analizar el relacionamiento interno y externo de cada variable.

```
matrizdecluster <-cutree(hc,k=3:5) #creamos una tabla con clusters de tres, cuatro y cinco grupos
```

	3	4	5
1	1	1	1
2	2	2	2
3	1	1	1
4	2	2	3
5	2	2	3
6	1	3	4
7	2	2	2
8	2	2	3
9	1	1	1
10	1	3	4
11	1	3	4
12	3	4	5
13	1	3	4

14	1	3	4
15	1	3	4
16	1	3	4
17	1	1	1
18	2	2	3
19	1	3	4
20	1	3	4
21	1	3	4
22	2	2	3
23	1	3	4
24	1	1	1

Es importante que luego de haber creado esta matriz, la misma se junte a la base de datos anterior, con la que se estaba trabajando, para crear una data conjunta. Este procedimiento se realiza usando el comando “data.frame”.

```
nuevosdatos <- data.frame(matrizdecluster,mydata) #creacion de una data en conjunto

str(nuevosdatos) #verificación de la unificación de las dos bases de datos
```

```
'data.frame':   24 obs. of  6 variables:
 $ X3          : int  1 2 1 2 2 1 2 2 1 1 ...
 $ X4          : int  1 2 1 2 2 3 2 2 1 3 ...
 $ X5          : int  1 2 1 3 3 4 2 3 1 4 ...
 $ PaÃ.s       : Factor w/ 25 levels " ",..: 2 4 5 8 9 13 15 16 17 20 ...
 $ ZTotalReserves: num  -0.21 -0.208 -0.209 -0.213 -0.21 ...
 $ ZGDPDeflator  : num  -0.0923 1.9759 -0.1848 0.9583 0.6848 ...
```

```
str(nuevosdatos$X3)
```

```
 int [1:24] 1 2 1 2 2 1 2 2 1 1 ...
```

Vemos que la matriz con los clústers y la data segmentada de nuestra data original ya se encuentran ubicadas bajo el mismo nombre. Ahora, se puede realizar la diferenciación interna y externa de los grupos. Primero, se compararán las desviaciones estándar de cada clúster con respecto a las dos variables dependientes.

```
aggregate(nuevosdatos$ZGDPDeflator, by = list(nuevosdatos$X3), FUN = sd) #pedir stand.desv para la primera variable dependiente ZGDPDeflator
```

	Group.1	x
1	1	**0.48**
2	2	**0.56**
3	3	**NA**

```
aggregate(nuevosdatos$ZGDPDeflator, by = list(nuevosdatos$X4), FUN = sd)
```

	Group.1	x
1	1	**0.18**
2	2	**0.56**
3	3	**0.36**
4	4	**NA**

```
aggregate(nuevosdatos$ZGDPDeflator, by = list(nuevosdatos$X5), FUN = sd)
```

	Group.1	x
1	1	**0.18**
2	2	**0.13**
3	3	**0.25**
4	4	**0.36**
5	5	**NA**

```
aggregate(nuevosdatos$ZTotalReserves, by = list(nuevosdatos$X3), FUN = sd) #pedir stad.desv para la segunda variable dependiente ZTotalReserves
```

	Group.1	x
1	1	**0.007**
2	2	**0.009**
3	3	**NA**

```
aggregate(nuevosdatos$ZTotalReserves, by = list(nuevosdatos$X4), FUN = sd)
```

	Group.1	x
1	1	**0.008**
2	2	**0.009**
3	3	**0.007**
4	4	**NA**

```
aggregate(nuevosdatos$ZTotalReserves, by = list(nuevosdatos$X5), FUN = sd)
```

	Group.1	x
1	1	**0.008**
2	2	**0.001**
3	3	**0.010**
4	4	**0.007**
5	5	**NA**

Como se puede apreciar, las desviaciones estándar de X3 suelen ser las más altas, mientras que las de X5 suelen ser las más bajas. Esto debido a que a una mayor cantidad de grupos en un clúster, menor dispersión entre los casos. De esta forma, X4 se presenta como la opción intermedia y con buenas medidas de dispersión.

Ahora, para probar la diferenciación externa de los grupos del clúster con respecto a las variables dependiente se utiliza Anova. Como ya se ha mencionado, la H0 de Anova prueba la igualdad de medias entre los grupos y la H1 señala la diferencia de medias entre los grupos. Por lo tanto, lo que esperamos de los clústers es que la significancia de Anova sea menor a 0.05 para decir que sus grupos internos son distintos.

```
fit <- aov(nuevosdatos$ZGDPDeflator ~ nuevosdatos$X3) #pedir ANOVA para la primera variable dependiente ZGDPDeflator
```

```
summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
nuevosdatos$X3	1	11.71	11.705	22.8	**9.1e-05 *** **

Residuals 22 11.29 0.513

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
fit1 <- aov(nuevosdatos$ZGDPDeflator ~ nuevosdatos$X4)
```

```
summary(fit1)
```

Df Sum Sq Mean Sq F value Pr(>F)

nuevosdatos$X4 1 4.287 4.287 5.04 **0.0352 ***

Residuals 22 18.713 0.851

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
fit2 <- aov(nuevosdatos$ZGDPDeflator ~ nuevosdatos$X5)
```

```
summary(fit2)
```

Df Sum Sq Mean Sq F value Pr(>F)

nuevosdatos$X5 1 4.257 4.257 4.997 **0.0359 ***

Residuals 22 18.743 0.852

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
fit3 <- aov(nuevosdatos$ZTotalReserves ~ nuevosdatos$X3)#pedir ANOVA para la segunda variable dependiente ZTotalReserves
```

```
summary(fit3)
```

Df Sum Sq Mean Sq F value Pr(>F)

nuevosdatos$X3 1 8.281 8.281 12.38 **0.00194 ****

Residuals 22 14.719 0.669

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
fit4 <- aov(nuevosdatos$ZTotalReserves ~ nuevosdatos$X4)

summary(fit4)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
nuevosdatos$X4  1  3.861   3.861   4.438 0.0468 *
Residuals      22 19.139   0.870
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

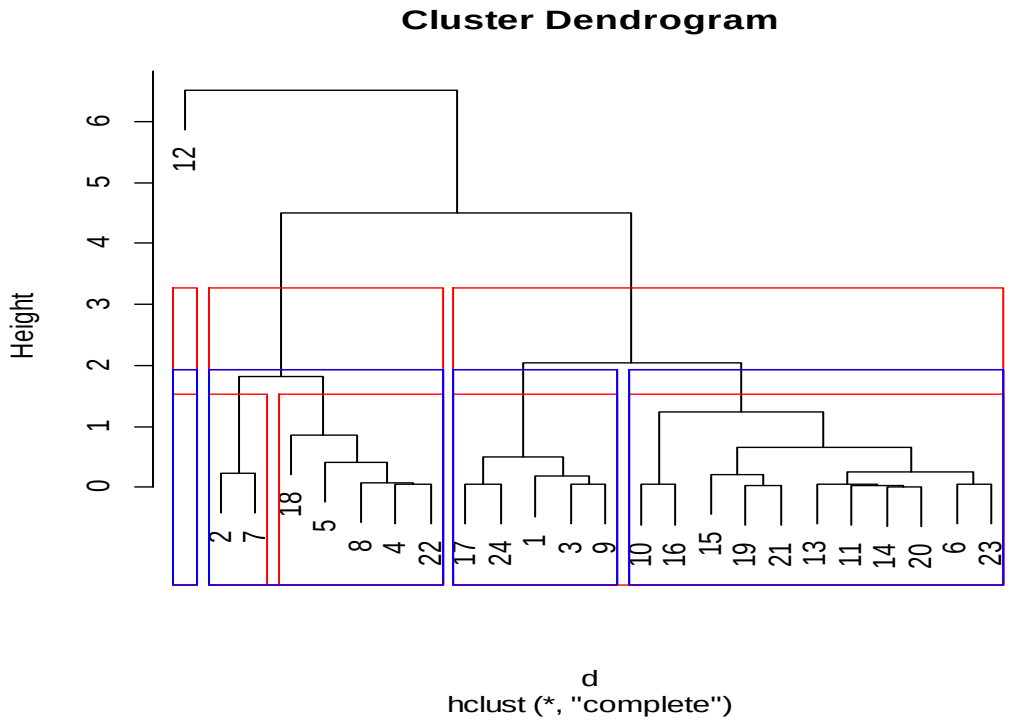
```
fit5 <- aov(nuevosdatos$ZTotalReserves ~ nuevosdatos$X5)

summary(fit5)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
nuevosdatos$X5  1  2.505  2.5046   2.688  0.115
Residuals      22 20.495  0.9316
```

Según el análisis los clústers que pasan la prueba para ambas variables dependientes son *X3* con **9.1e-05** y **0.00194** y *X4* con **0.0352** y **0.0468** en la primera y segunda variable, respectivamente. En otras palabras, ambas agrupaciones son compatibles para ser el clúster definitivo.

Por último, para designar el clúster final se mezclan ambos procedimientos. Primero, para la desviación estándar el grupo a utilizar era *X4*; mientras que para Anova, los grupos a utilizar eran *X3* y *X4*. Por lo tanto, se debe designar la opción *X4* como el clúster definitivo. La agrupación sería de esta manera:

```
rect.hclust(hc, k=4, border="blue") #sombrear la opción de cluster elegida en azul
```



Análisis de Función Discriminante

El Análisis de Función Discriminante tiene como objetivo principal la predicción de un valor respecto de una clasificación. Esto significa que esta herramienta va a crear una nueva variable que almacena los grupos a los que pertenecen los valores; de esta manera, al introducir un valor nuevo en la base de datos, la variable recién creada lo clasificará y en lo incluirá en uno de los grupos señalados.

La Función Discriminante requiere de una variable de agrupación categórica y de un grupo de variables escalares; la primera servirá para generar la clasificación; por eso, es recomendable que la variable categórica no sea la misma que define a los casos de nuestra data. Por ejemplo, en nuestro caso hipotético la variable *país* nos muestra los casos, pero utilizaremos la variable *tipo de gobierno* para generar la clasificación. Las variables escalares van a actuar como independientes y generarán una clasificación solo en torno a esas variables. Entonces ¿qué distingue a la Función Discriminante de una clasificación nominal y un análisis de clusters? La diferencia radica en la capacidad de predicción de nuevos valores de la Función Discriminante. Para ejecutarla en R se debe instalar el paquete *car* y usar el comando *MASS*. Se utilizará como variable categórica: *Zona Europea* y como dependientes numéricas: *Foreign Exchange, GoodsValueofexports* y *GoodsValueofImports*. Veamos primero los niveles de la variable categórica:

```
install.packgaes("car")#instalar el paquete car
```

```
library(MASS)#ejecutar MASS

levels(ZonaEuropea)

 [1] "Balticos"     "Escandinavos" "Centro"       "Este"


install.packages("biotools")#instalar el paquete biotools

library(biotools)#ejecutar biotools


data2<-subset(data,select=c(13,14), na.value=NULL)#seleccionar una subdata


boxM(data2, ZonaEuropea)#pedir M de box

        Box's M-test for Homogeneity of Covariance Matrices

data:  data2

Chi-Sq (approx.) = NaN, df = 9, p-value = NA


nuevo<- lda(ZonaEuropea~ ForeignExchange+ GoodsValueofexports+ GoodsValueofImports,
data=data) #crear el modelo de funcion discriminante


nuevo# observar los estadisticos de la funcion discriminante


Call:

lda(ZonaEuropea ~ ForeignExchange + GoodsValueofexports + GoodsValueofImports,

    data = data)


Prior probabilities of groups:

    Balticos Escandinavos       Centro        Este

       0.375        0.125        0.375        0.125


Group means:

            ForeignExchange GoodsValueofexports GoodsValueofImports

Balticos           6635.158           46462.48           50366.11

Escandinavos      26558.537          126089.94          118465.34

Centro            14328.844          510480.96          517064.03

Este              36886.249          138075.83          145550.51
```

Coefficients of linear discriminants:

	LD1	LD2	LD3
ForeignExchange	-6.211583e-05	-5.773811e-05	1.178213e-06
GoodsValueofexports	-7.344452e-06	8.766925e-08	-1.363641e-05
GoodsValueofImports	1.249548e-05	-1.581681e-06	1.350466e-05

Proportion of trace:

LD1	LD2	LD3
0.7742	0.2243	0.0015

```
new<- predict(nuevo, newdata=data[,c(2,8)])$class # predicir los grupo de los casos
```

```
table(new,data[,1])
```

new	Austria	Belgium	Bulgaria	Croatia	Czech Republic	Denmark	Estonia
Balticos	0	1	0	1	1	0	1
Escandinavos	0	0	0	0	0	0	0
Centro	0	0	1	0	0	0	0
Este	0	0	0	0	0	1	0

new	Finland	France	Germany	Hungary	Ireland	Italy	Lithuania	Luxembourg
Balticos	1	0	0	0	1	0	1	1
Escandinavos	0	0	0	1	0	0	0	0
Centro	0	1	1	0	0	1	0	0
Este	0	0	0	0	0	0	0	0

new	Malta	Netherlands	Poland	Romania	Slovenia	Spain	Sweden
Balticos	1	0	0	1	0	1	0	1
Escandinavos	0	0	0	0	1	0	0	0
Centro	0	1	0	0	0	0	1	0
Este	0	0	1	0	0	0	0	0

new	United Kingdom
Balticos	0
Escandinavos	0
Centro	1
Este	0

La Función Discriminante nos dará una aproximación a la clasificación que nosotros creíamos que teóricamente podría funcionar. Por eso, si bien la variable categórica clasifica los grupos, la función genera una clasificación con algunos resultados distintos, esto se debe a los valores de las variables dependientes. Podemos apreciar, entonces, que luego de generar la función con el comando *lda*, se debe ejecutar la predicción con el comando *predict*. De esta forma, la función nos señala en código binario a qué categoría pertenece cada caso.

Ahora es tu turno

Ejercicios

- Usted se encuentra trabajando como asesor de campaña de un candidato presidencial. El equipo de "análisis de información" del partido ha recogido una muestra significativa en Loreto sobre las preferencias programáticas de esa población. El candidato tiene programado un viaje a cinco zonas distintas de ese departamento donde, por pura intuición, cree que puede encontrar grupos de habitantes con objetivos y necesidades muy distintas. Efectivamente, los resultados de la muestra indican que los temas principales a discutir para los loretanos son distintos entre sí como por ejemplo: *el desarrollo comercial del punto fronterizo, la protección de la selva amazónica y la comunicación entre el Estado y las comunidades no contactadas.* ¿Cómo se asegura el equipo de campaña que el candidato responda las inquietudes que más le preocupan a cada población? Utilice la data "Ejercicios U7" para resolver esta interrogante

- Usted es parte de la Dirección de Índices Financieros de la Unión Europea y ya tiene clasificado a los países según su zona europea: centro, este, báltico y escandinavo*;* y sus ingresos en exportaciones y su Producto Bruto Interno. Sin embargo, con la situación que acontece en España: la posible separación de Cataluña, usted quiere saber en qué grupo se encontraría este posible nuevo miembro de la Unión Europea respecto de sus exportaciones y del Producto Bruto Interno. ¿Qué técnica estadística nos permite generar esta clasificación y la posterior predicción de los nuevos valores? Utiliza la data "UEunidad2" para resolver este ejercicio.