**Ejercicio Propuesto 2 (Resuelto)**

**Una compañía de telecomunicaciones realiza un estudio con el fin de reducir el abandono de sus clientes. Para ello dispone de un archivo de datos, donde cada caso corresponde a un cliente distinto del que registra diversa información demográfica y del uso del servicio. El objetivo es segmentar su base de clientes por patrones de uso del servicio. Si los clientes se pueden clasificar por el uso, la empresa puede ofrecer paquetes más atractivos para sus clientes. Las variables que indican el uso y no uso de los servicios están contenidas en el archivo**[**Telecomunicaciones1.s**](http://wpd.ugr.es/~bioestad/wp-content/uploads/Telecomunicaciones1.sav)**av o**[**Teleconunicaciones1.dat**](https://wpd.ugr.es/~bioestad/wp-content/uploads/Telecomunicaciones1.dat)

**El archivo de datos telecomunicaciones1.dat contiene 1000 datos y está formado por las siguientes variables: región, permanencia, edad, estado\_civil, dirección, ingresos\_familiares, nivel\_educativo, empleo, género, n-pers\_hogar, llamadas\_gratuitas, alquiler\_equipo, tarjeta\_llamada, inalámbrico, larga\_distancia\_mes, llamadas\_gratuitas\_mes, equipo\_mes, tarjeta\_mes, inalámbrico\_mes, líneas\_múltiples, mensaje\_voz, servicio\_busca, internet, identificador\_llamada, desvío\_llamadas, llamada\_a\_tres, facturación\_electrónica.**

**Se pide**:

**a) Utilizar el procedimiento Análisis de conglomerados jerárquico para estudiar las relaciones entre los distintos servicios.**

**b) Realizar de nuevo el estudio con la medida de distanciade Jaccard y comparar los resultados**.

**Solución**

Primero leemos el fichero de datos:

> setwd("C:/Users/Usuario/Desktop/Datos")

> datos <- read.table("Telecomunicaciones1.dat", header = TRUE)

A continuación, filtramos las variables que queremos elegir:

**Variables**: Servicio de llamadas gratuitas, Alquiler de equipo, Servicio de tarjeta de llamada, Servicio inalámbrico, Líneas múltiples, mensajes de voz, servicio de busca, internet, Identificador de llamadas, llamadas en espera, Desvío de llamadas, llamadas a tres, Facturación electrónica

> datos <- datos[, c("llamada\_gratuitas", "alquiler\_equipo","tarjeta\_llamada", "Inhalámbrico", "lineas\_múltiples", "mensaje\_voz", "servicio\_busca", "internet", "identificador\_llamada", "llamada\_espera", "desvio\_llamadas", "llamada\_a\_tres", "facturación\_electrónica")]

Ya tenemos el data frame con las variables que nos interesan.  A continuación calculamos la matriz de distancias.

**a) Utilizar el procedimiento Análisis de conglomerados jerárquico para estudiar las relaciones entre los distintos servicios.**

Elegimos como medida **la binaria**, y como método **Vinculación inter-grupos,**es decir **“average”**. Los datos los metemos como su tabla traspuesta ya que se van a aglomerar las variables en vez de los casos

> dmx <- dist(t(datos), method = "binary")  
> cl <- hclust(dmx, method = "average")

> cl

Call:

hclust(d = dmx, method = "average")

Cluster method : average

Distance : binary

Number of objects: 13

Ya podemos dibujar nuestro dendrograma.

> cl$height

[1] 0.3116197 0.3264605 0.3321387 0.3440881 0.3991507 0.3994253 0.4368252 0.4458629

[9] 0.4632616 0.5989666 0.5996411 0.6844400

> cl$merge

[,1] [,2]

[1,] -1 -10

[2,] -9 -11

[3,] 1 2

[4,] -12 3

[5,] -2 -8

[6,] -4 -7

[7,] -13 5

[8,] -3 4

[9,] -6 6

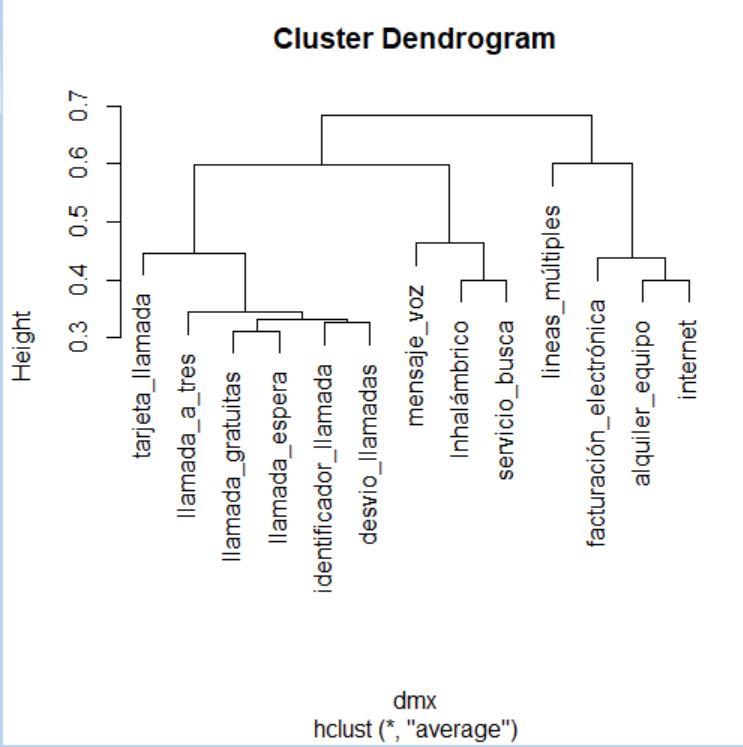
[10,] 8 9

[11,] -5 7

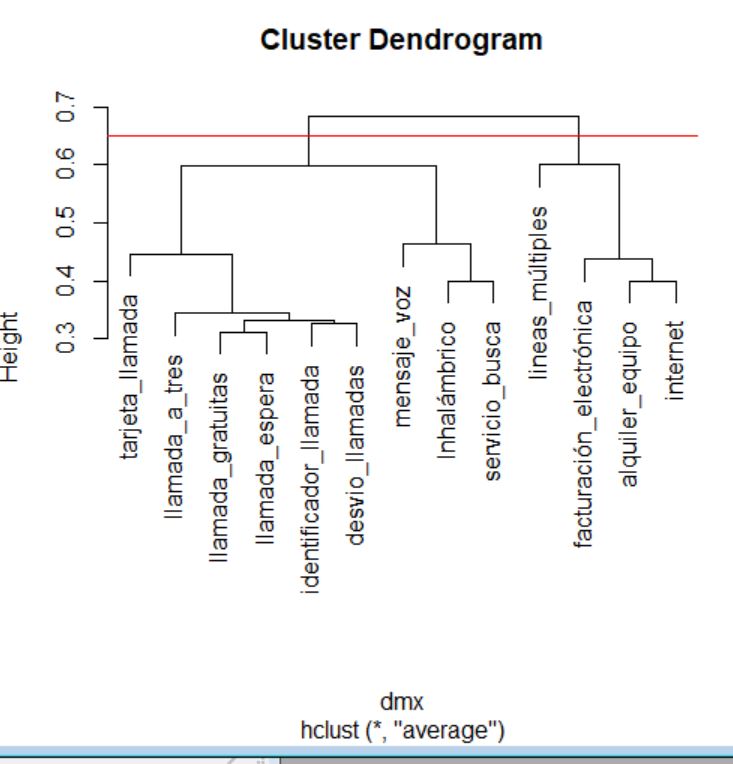
[12,] 10 11

En las medidas binarias, la columna de los coeficientes informa de las medidas de similitud, por lo tanto, los valores de este coeficiente van disminuyendo en cada etapa del análisis. Es difícil interpretar los resultados, por lo que recurrimos al Dendrograma.

> plot(cl)



> abline(h=.65, col="red")



El dendrograma muestra que los patrones de uso de **Líneas múltiples** y **Servicio de tarjeta de llamadas** son distintos de los otros servicios. Estos otros se agrupan en tres grupos. Un grupo incluye **inalámbrico**, **servicio\_busca** y **mensaje\_voz**. Otro incluye **alquiler\_equipo**, **internet**, y **facturación\_electrónica**.

El último grupo contiene las variables **llamadas\_gratuitas,** **llamadas\_espera**, **identificador\_llamada**, **desvío\_llamadas** y **llamada\_a\_tres**. El grupo **servicio inalámbrico** está más cerca del grupo de **Internet** que el grupo **LlamEsp**.

**b) Realizar de nuevo el estudio con la medida de distanciade Jaccard y comparar los resultados**.

Para ejecutar un análisis de conglomerados con la medida de **distanciade Jaccard**, utilizamos la función **"vegdist"**. Para ello,  instalamos y cargamos el paquete **“vegan”**.

> install.packages("vegan")

> library("vegan")

En la función **vegdist**, ponemos el primer argumento la tabla traspuesta de los datos, elegimos el método y seleccionamos True en la opción binaria. Volvemos a elegir el método **Vinculación inter-grupos.**

> dmx2 <- vegdist(t(datos), method = "jaccard", binary=T)

> cl2 <- hclust(dmx2, method = "average")

> cl2

Call:

hclust(d = dmx2, method = "average")

Cluster method : average

Distance : binary jaccard

Number of objects: 13

> cl2$height

[1] 0.3116197 0.3264605 0.3321387 0.3440881 0.3991507 0.3994253 0.4368252 0.4458629

[9] 0.4632616 0.5989666 0.5996411 0.6844400

> cl2$merge

[,1] [,2]

[1,] -1 -10

[2,] -9 -11

[3,] 1 2

[4,] -12 3

[5,] -2 -8

[6,] -4 -7

[7,] -13 5

[8,] -3 4

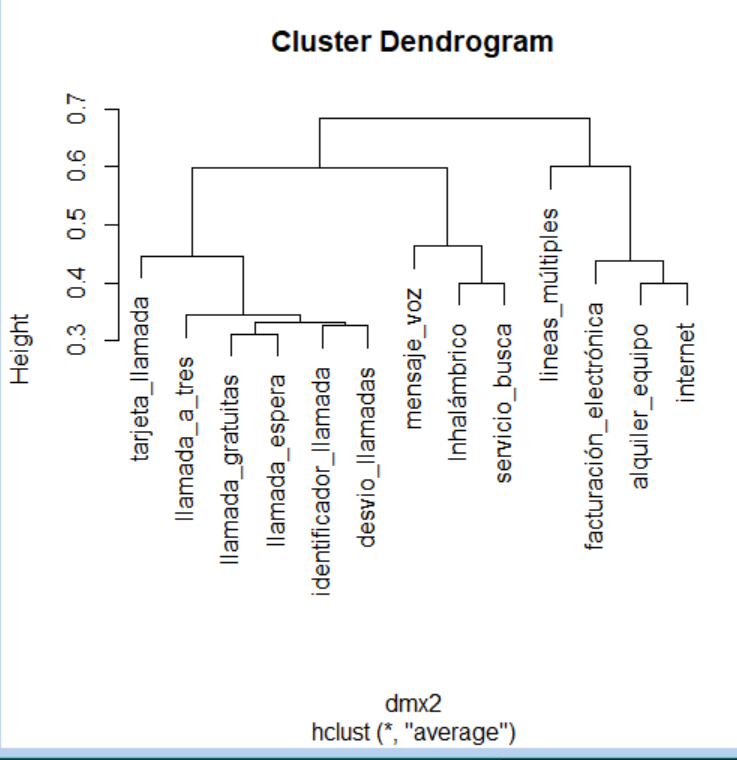
[9,] -6 6

[10,] 8 9

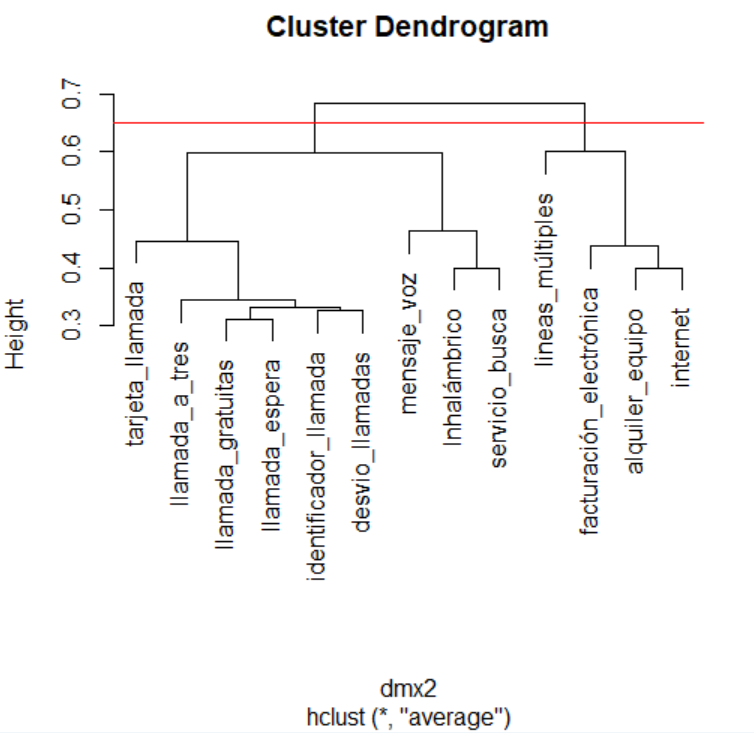
[11,] -5 7

[12,] 10 11

> plot(cl2)



abline(h=.65, col="red")



Utilizando **la medida de Jaccard,** los tres grupos básicos son los mismos, pero el grupo **servicio inalámbrico** está más cerca del grupo **LlamEsp** que el grupo **Interne**t.

La diferencia entre la simple adaptación y las medidas de Jaccard es que la medida Jaccard no considera dos servicios similares si una persona no está suscrita. Es decir, casación simple considera que los servicios inalámbricos y de Internet son similares cuando un cliente esté en ambos o en ninguno, mientras que Jaccard considera que son similares sólo cuando un cliente tiene dos servicios. Esto provoca una diferencia en las soluciones de clúster porque hay muchos clientes que no tienen servicios inalámbricos o Internet. Por lo tanto, estos grupos son más similares en la solución de casación simple que la solución de Jaccard. La medida que se utiliza depende de la definición de "similares" que se aplica a la situación.