**CURSO DE FORMACIÓN EN DATA MINING CON "R"**

## TAREA 1

***Lea el archivo de datos “vinos.RData”. Aplique el método k-medias para obtener dos clases utilizando como variables explicativas las concentraciones de los distintos ácidos orgánicos. Relacione la clasificación obtenida con la variedad de uva, con el fin de averiguar hasta que punto esa variedad es un criterio principal de la clasificación “natural” de los vinos o es necesario buscar criterios adicionales (como la zona, el año de la vendimia, u otros).***

En primer lugar, seleccionamos nuestro directorio de trabajo. En mi caso:

setwd("/Users/jlsovaz/Desktop/curso\_data\_mining/TEORIA/UNIDAD 1/")

Establecemos una semilla para lograr reproducibilidad en los resultados obtenidos:

set.seed(123)

Importamos el dataset del primer problema en formato .RData con la función load:

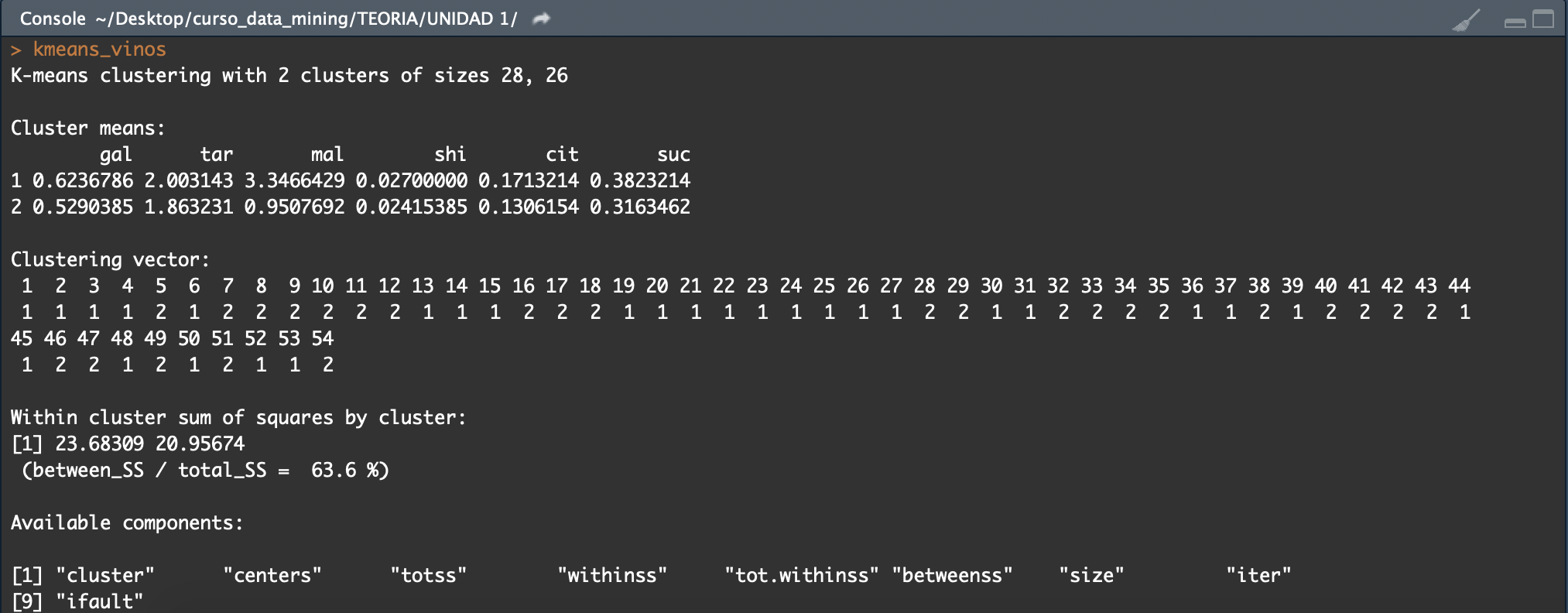
load("vinos.RData")

Realizar método k-means en el dataframe "vinos" utilizando dos clases para clasificar cada muestra en función de la información que nos dan sobre las concentraciones de los distintos ácidos orgánicos. Elegimos dos clases porque es la clasificación "natural" de nuestras muestras en este caso: Albariño y Godello. Así veremos si las variables cuantitativas son un buen criterio para clasificar los vinos en las dos clases mencionadas.

kmeans\_vinos <- kmeans(vinos[,2:ncol(vinos)], centers = 2)

Exploramos el modelo calculado:

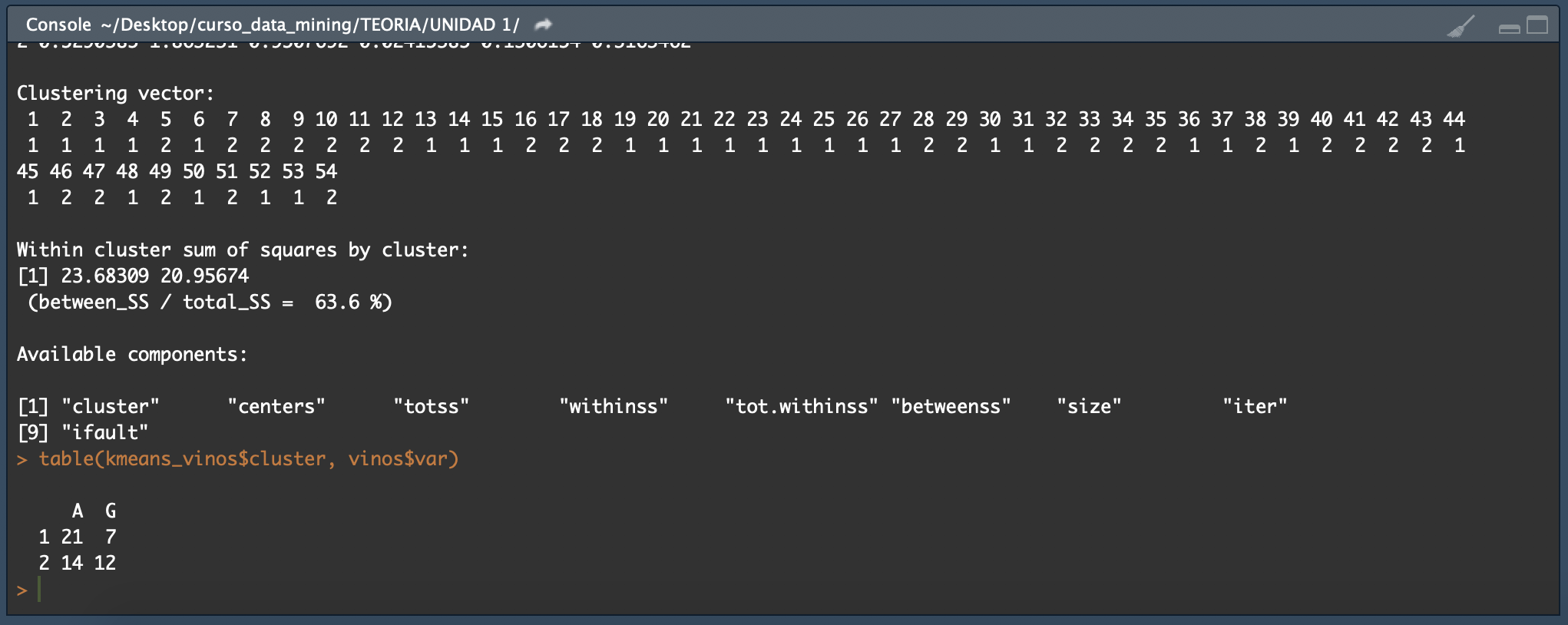
kmeans\_vinos

Aparece la siguiente salida:

Vemos que este método clasifica 28 vinos en la clase 1 y 26 vinos en la clase 2. Tenemos las medias (centros) para cada ácido orgánico en cada una de las clases. A simple vista podemos ver que algunos ácidos orgánicos (*gal, shi, cit, suc*) presentan medias parecidas para las dos clases; mientras que *tar* y *mal* tienen medias muy distintas entre las dos clases.

En este caso, el % de variabilidad explicada por esta clasificación es del 63.6%, lo que podría indicar que no es una buena clasificación.

De hecho, si escribimos:

table(kmeans\_vinos$cluster, vinos$var)

Podemos ver que 21 Albariños son clasificados en una clase y los otros 14 son clasificados en la clase 2. En relación al Godello, 7 son designados a la clase 1; 12 vinos a la clase 2. Parece ser que la concentración de ácidos orgánicos no es un buen criterio para clasificar los vinos en Albariño y Godello. Podrían ser necesarias otras variables cuantitativas para poder discernir mejor entre la clasificación natural de los vinos.

***Lea el conjunto de datos "deudas.RData". Se trata de una muestra de 100 clientes de un banco, algunos de los cuales han presentado impagos, de los que se dispone de información relativa a su nivel de ingresos, relación entre deudas e ingresos, importe de las deudas por tarjeta de crédito, e importe de otras deudas, entre otras variables. Utilice esas 4 variables (columnas 5 a 8) para obtener con el método EM una clasificación con dos grupos o clases.  Averigue si la clasificación obtenida está relacionada con la variable “Impago”.***

En primer lugar, seleccionamos nuestro directorio de trabajo. En mi caso:

setwd("/Users/jlsovaz/Desktop/curso\_data\_mining/TEORIA/UNIDAD 0/")

Establecemos una semilla para lograr reproducibilidad en los resultados obtenidos:

set.seed(123)

Importamos el dataset del primer problema en formato .RData con la función load:

load("deudas.RData")

Realizamos el método EM en el dataframe "deudas" utilizando dos clases para clasificar cada muestra en función de la información que nos dan sobre ingresos, deudas\_ingresos, deudas\_tarjeta y otras deudas. Así veremos si las variables cuantitativas son un buen criterio de clasificación de los clientes en dos categorías: presentan o no presentan impagos.

Cargamos el paquete necesario para realizar el método EM:

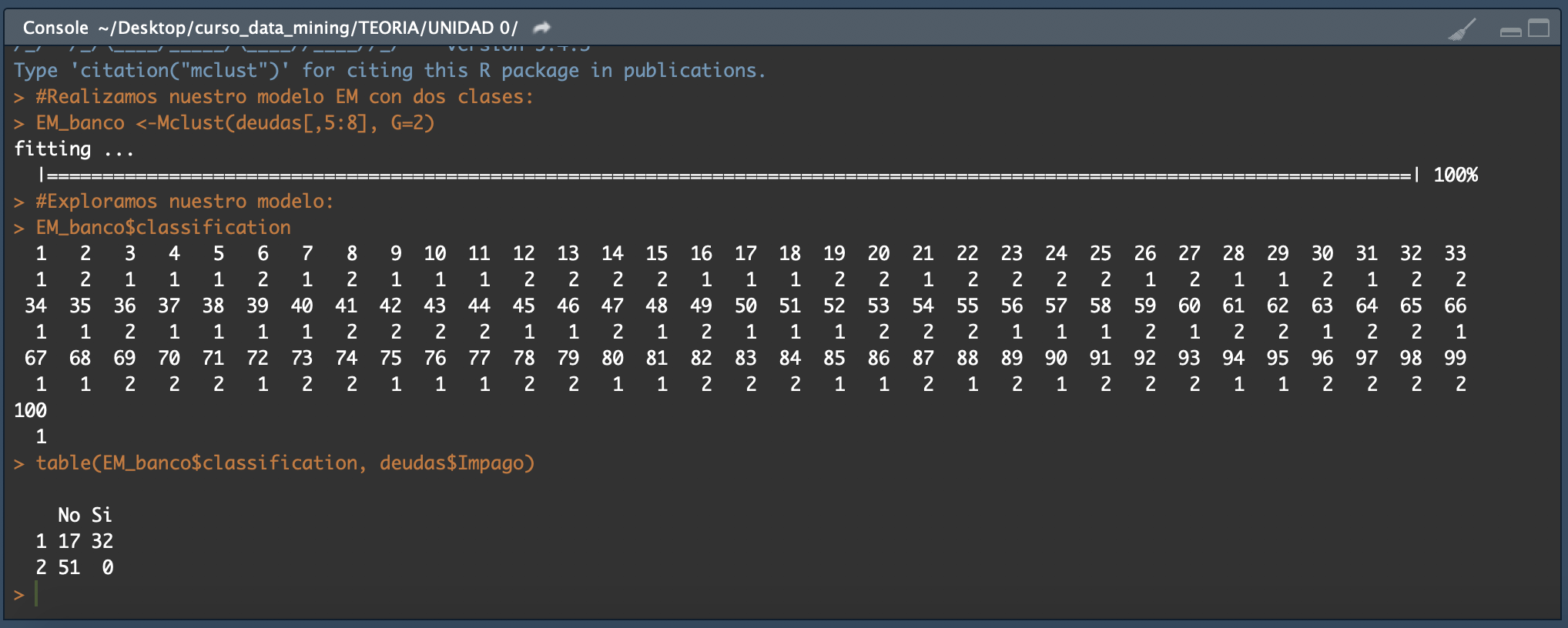
library(mclust)

Realizamos nuestro modelo EM con dos clases:

EM\_banco <-Mclust(deudas[,5:8], G=2)

Exploramos nuestro modelo y vemos como clasificación se relaciona con la variable Impago:

EM\_banco$classification

table(EM\_banco$classification, deudas$Impago)

En este caso, la clasificación es buena, 17 de las personas que no presentan impago fueron a la clase 1, mientras que 51 fueron a la clase 2. Por otro lado, la totalidad de las personas que presentan impago fueron clasificadas en una sóla clase. Se puede concluir que las variables cuantitativas usadas clasifican (al menos) perfectamente a las personas que presentan impago.