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

## TAREA 5

***Lea el archivo de datos “deudas”. Aplique el método KNN para identificar a los clientes que tendrán impagos a partir de partir de las variables cuantitativas conocidas (variables en las columnas 5 a 8 del conjunto de datos). Utilice una muestra aleatoria de 80 casos para el entrenamiento y los 20 restantes para validación.***

***{utilice  entrenamiento <- deudas[muestra, 5:8]     y     prueba <- deudas[-muestra, 5:8]}***

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")

Ahora creamos el subset que necesitaremos para entrenar el método con 80 muestras aleatorias y 20 muestras para validar el modelo:

muestra <- sample(1:nrow(deudas), 80)

entrenamiento <- deudas[muestra,5:8]

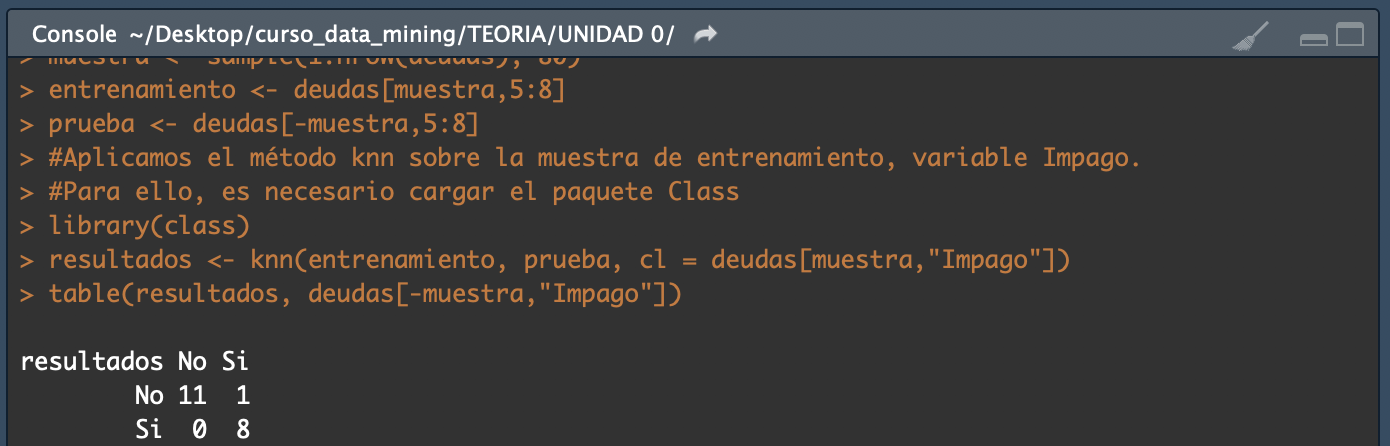
prueba <- deudas[-muestra,5:8]

Aplicamos el método knn sobre la muestra de entrenamiento, variable Impago. Para ello, es necesario cargar el paquete Class

library(class)

resultados <- knn(entrenamiento, prueba, cl = deudas[muestra,"Impago"])

#Tabla de frecuencias cruzada entre los resultados y la clasificación "natural" de las muestras:

table(resultados, deudas[-muestra,"Impago"])

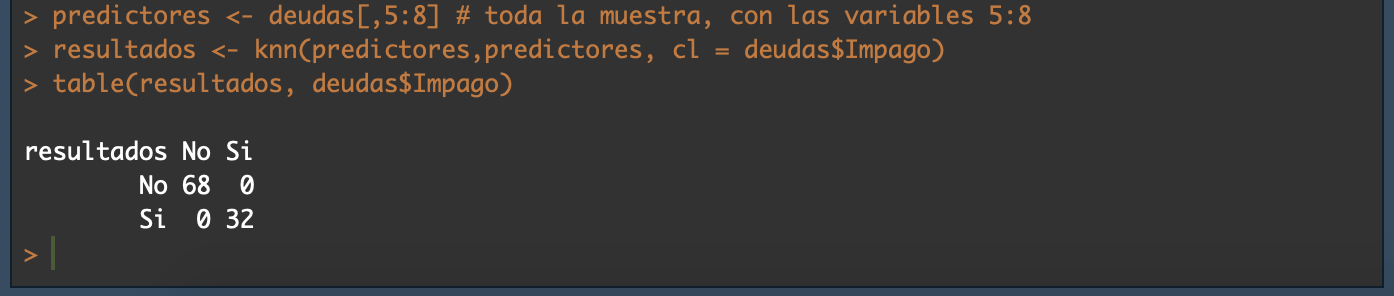
Podemos ver que han sido clasificados correctamente todos los clientes excepto 1.

Aplicamos el método a toda la muestra "deudas":

predictores <- deudas[,5:8] # toda la muestra, con las variables 5:8

resultados <- knn(predictores,predictores, cl = deudas$Impago)

table(resultados, deudas$Impago)



Vemos que en este caso el modelo mejora; todos los clientes han sido clasificados correctamente.

***Lea el conjunto de datos "vinos" (40 casos para entrenamiento, el resto para validación). Aplique el método SVM para identificar la variedad de uva utilizada en la elaboración del vino a partir de las concentraciones de ácidos orgánicos.***

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")

Ahora creamos el subset que necesitaremos para entrenar el método con 40 muestras aleatorias y el resto de las muestras para validar el modelo:

muestra <- sample(1:nrow(vinos), 40)

entrenamiento <- vinos[muestra,]

prueba <- vinos[-muestra,]

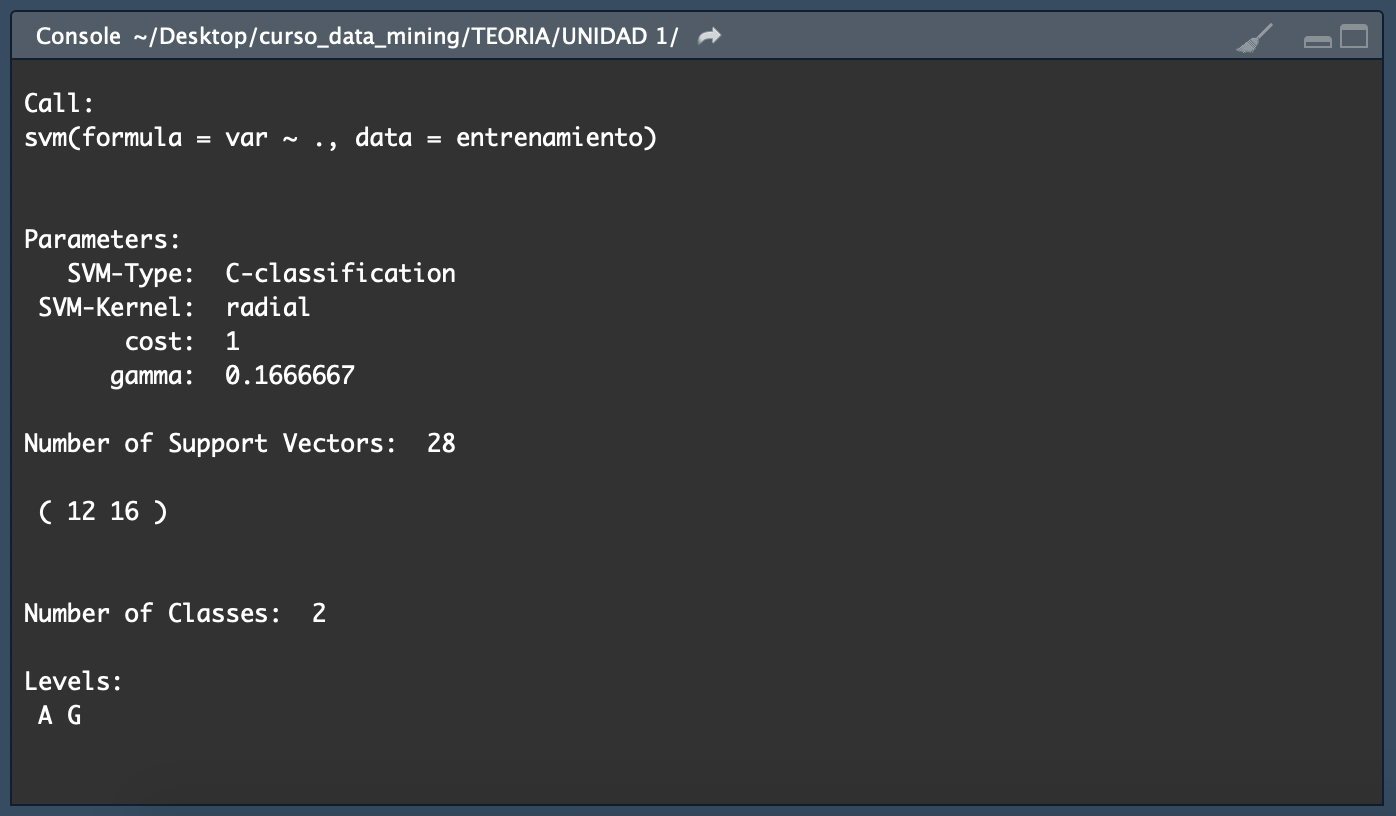
Ahora aplicamos el método SVM para construir el modelo, con la variable "var" dependiente de todas las restantes:

library(e1071)

modelo <- svm(var ~ ., data = entrenamiento)

Exploramos el modelo:

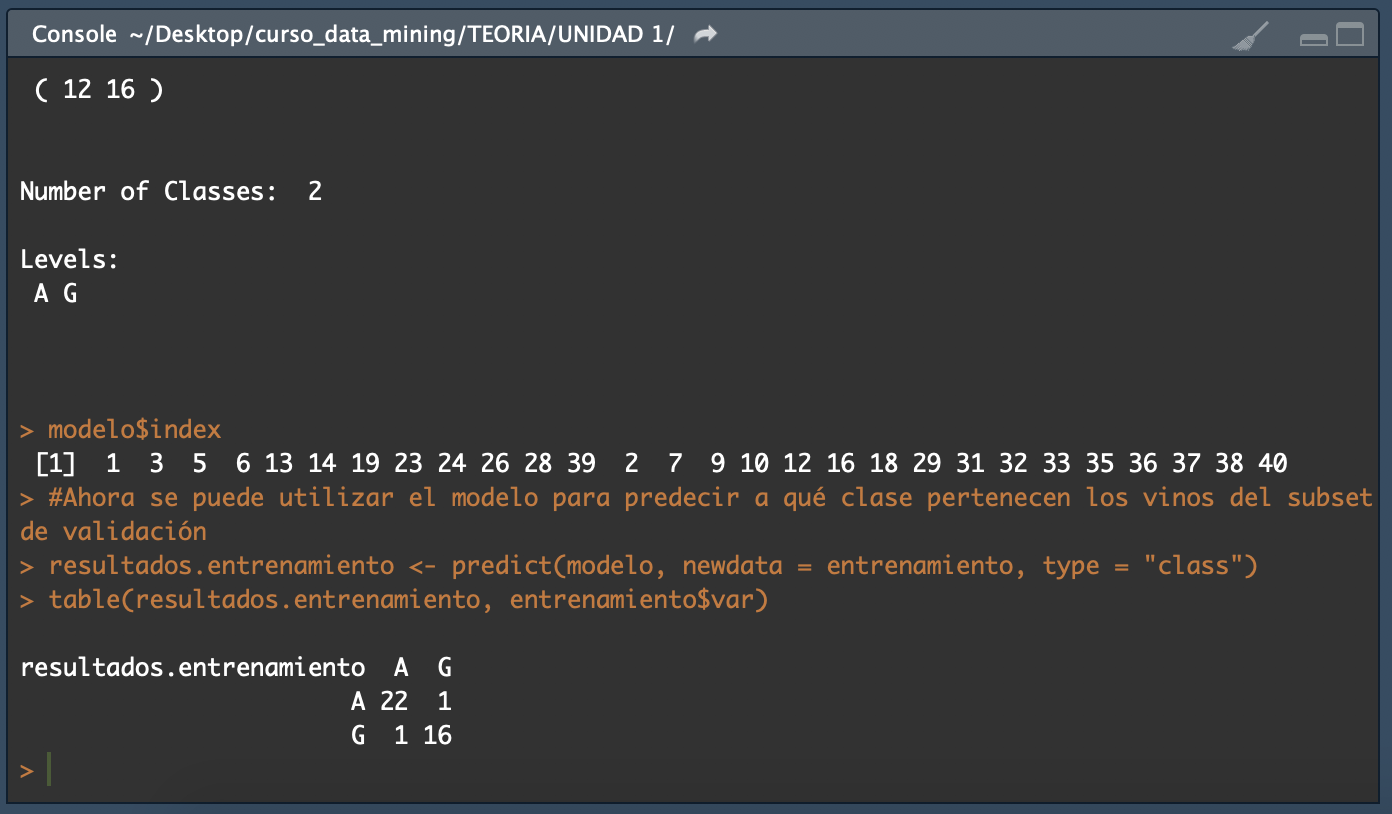
summary(modelo)

Vemos que la transormación es radial (kernel = radial), opción por defecto. En este caso, el valor de la constante gamma es 0.16, que es el inverso del número de variables explicativas (6).

Hay 28 vectores soporte: 12 a un lado del hiperplano y 16 al otro lado.

Ahora se puede utilizar el modelo para predecir a qué clase pertenecen los vinos del subset de entrenamiento, los mismos que se usaron para la construcción del modelo.

resultados.entrenamiento <- predict(modelo, newdata = entrenamiento, type = "class")

table(resultados.entrenamiento, entrenamiento$var)

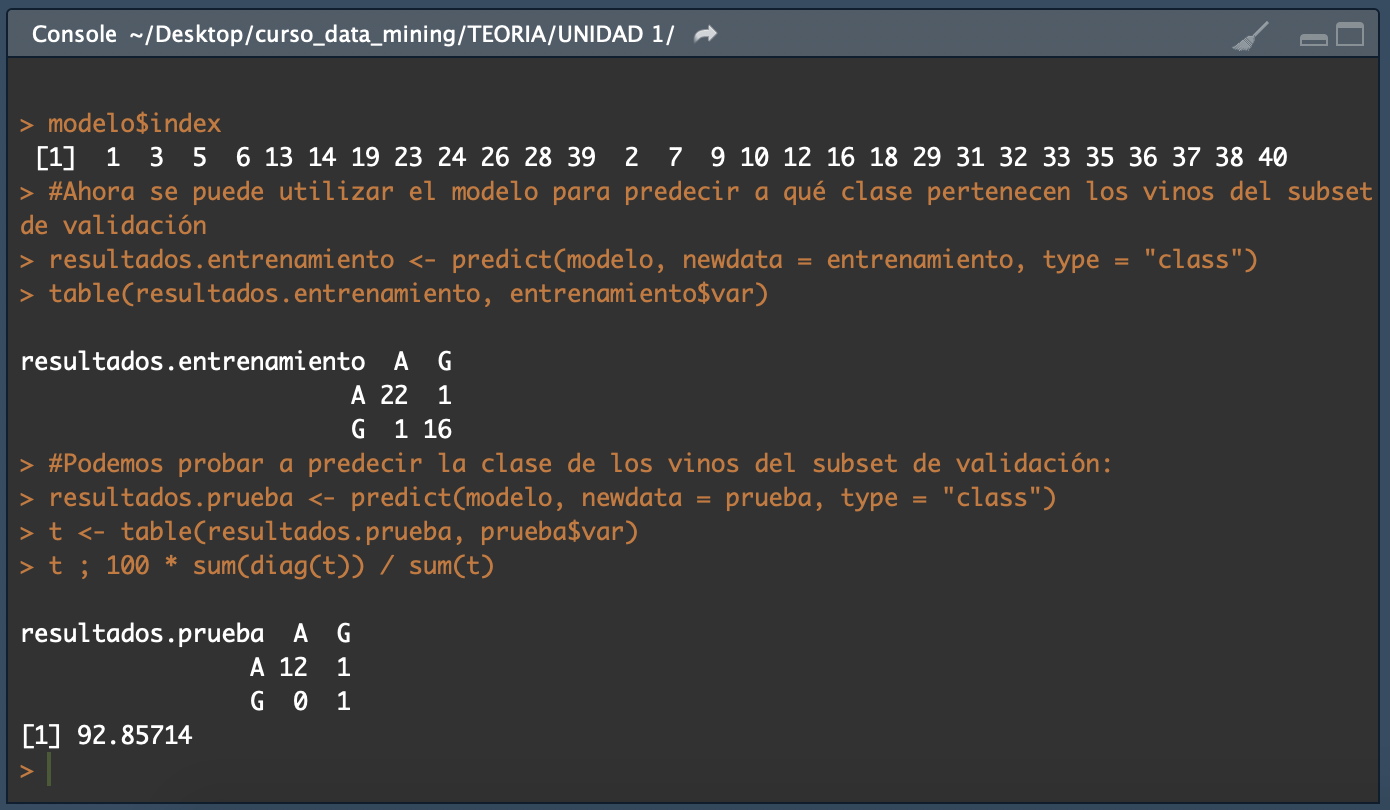
Vemos que sólamente 2 vinos son clasificados incorrectamente.

Podemos probar a predecir la clase de los vinos del subset de validación:

resultados.prueba <- predict(modelo, newdata = prueba, type = "class")

t <- table(resultados.prueba, prueba$var)

t ; 100 \* sum(diag(t)) / sum(t)



En este caso, sólamente 1 vino fue clasificado incorrectamente como Godello. El 92.85% de las muestras fueron clasificadas correctamente.