Analisis Multivariado de Datos Antropometricos

Brahian Serna

2022-11-24

Este análisis se hizo empleando el lenguaje de programación R como se darán cuenta próximamente y la idea es poner en practica lo aprendido sobre análisis de clusteres en los datos y la aplicación de algunos algoritmos de clasificación.

A continuacion importamos las librerias que usaremos a lo largo de este notebook.

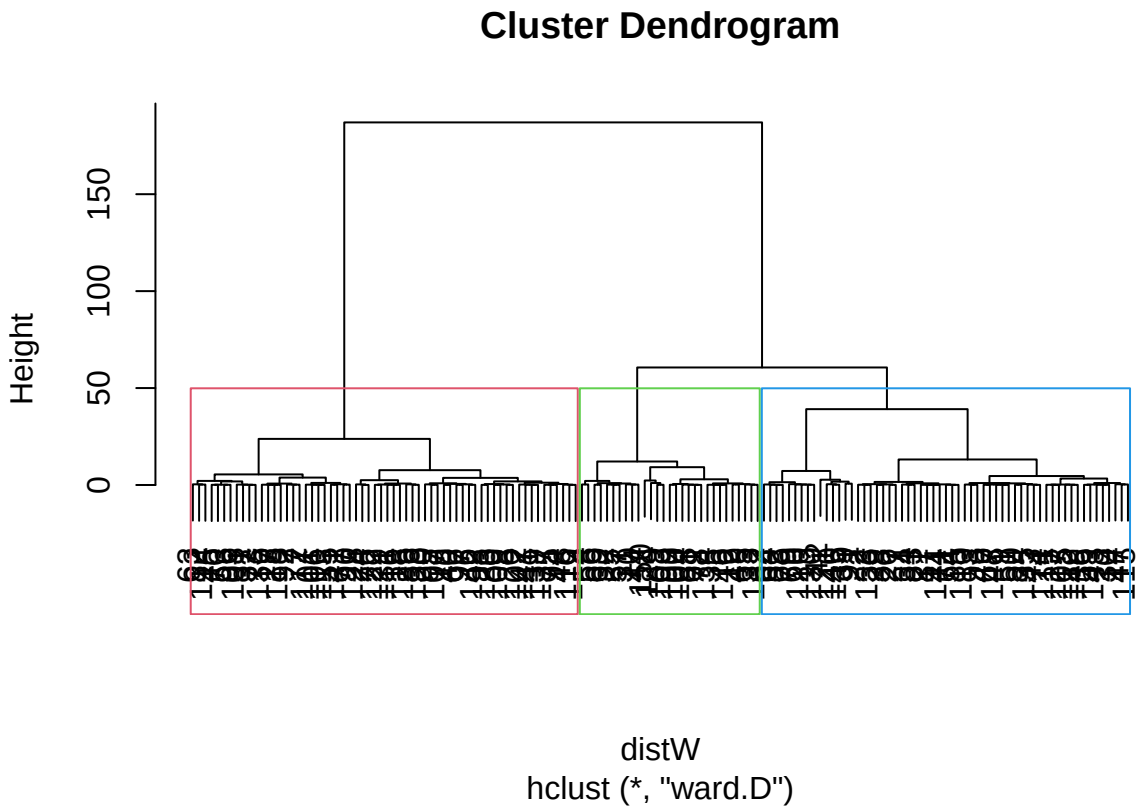
```
library(FactoClass)
library(factoextra)
library(ggplot2)
library(caTools)
library(class)
library(mvShapiroTest)
library(MASS)
```

Importamos el conjunto de datos:

```
uno <- read.table(file.choose(), header=T)

# Copiar el siguiente código en R sin modificar nada

genera <- function(cedula){
set.seed(cedula)
data <- uno[sample(1:2100,150),]
data
}

# Para crear la base de datos con la cual trabajara, debe ejecutar la siguiente línea:

base <- genera(83721)
datos <-data.frame(base$Sexo, base$p1, base$p7, base$p16, base$p22, base$p26,
                        base$p27, base$p29, base$p38)

datos$CAT_IMC <- ifelse(base$imc<18.5 , "Bajo_peso",
                            ifelse(base$imc<25,"Normal",
                                   "sobre peso"))
```

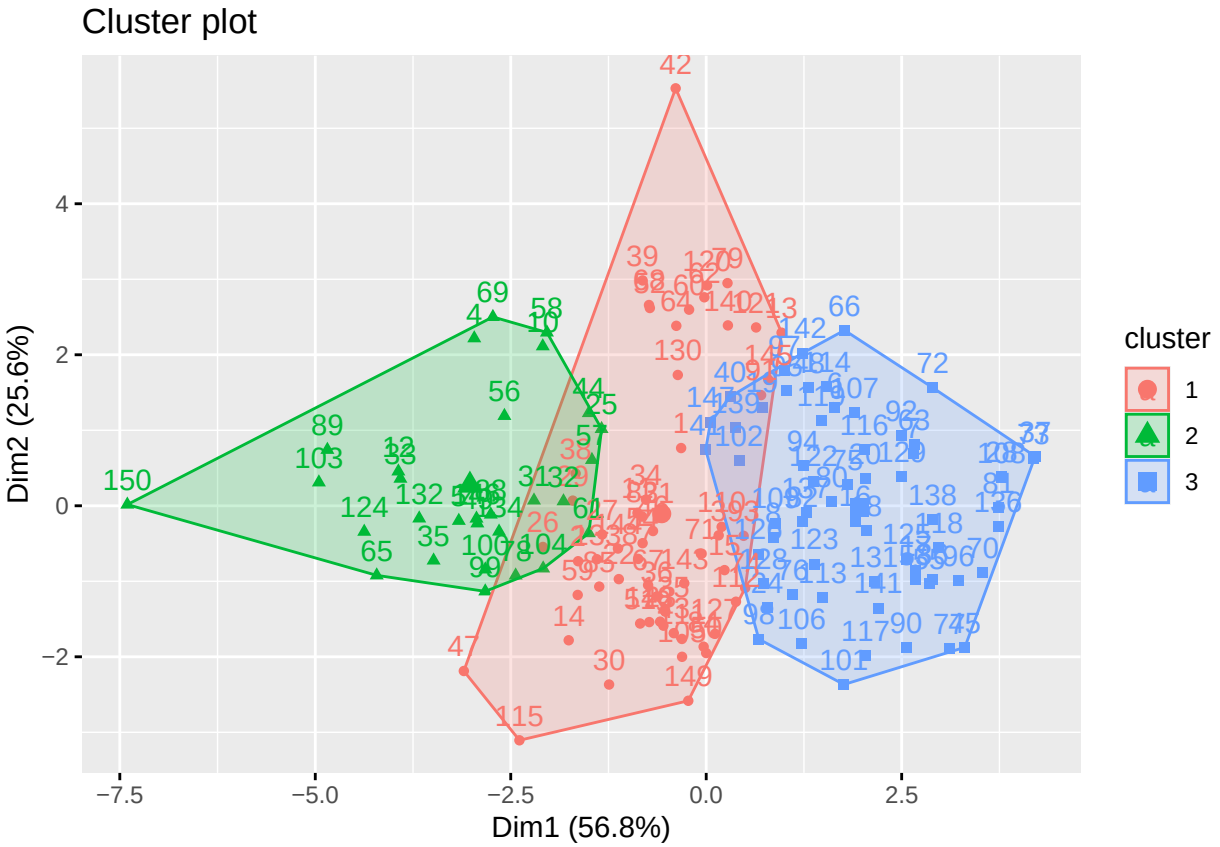
A continuación se muestra el endograma asociado a los datos usando el método de Ward's

```
hw_datos <- ward.cluster(dist(datos), h.clust=1)
plot(hw_datos)
rect.hclust(hw_datos, k=3, border=2:10)
```



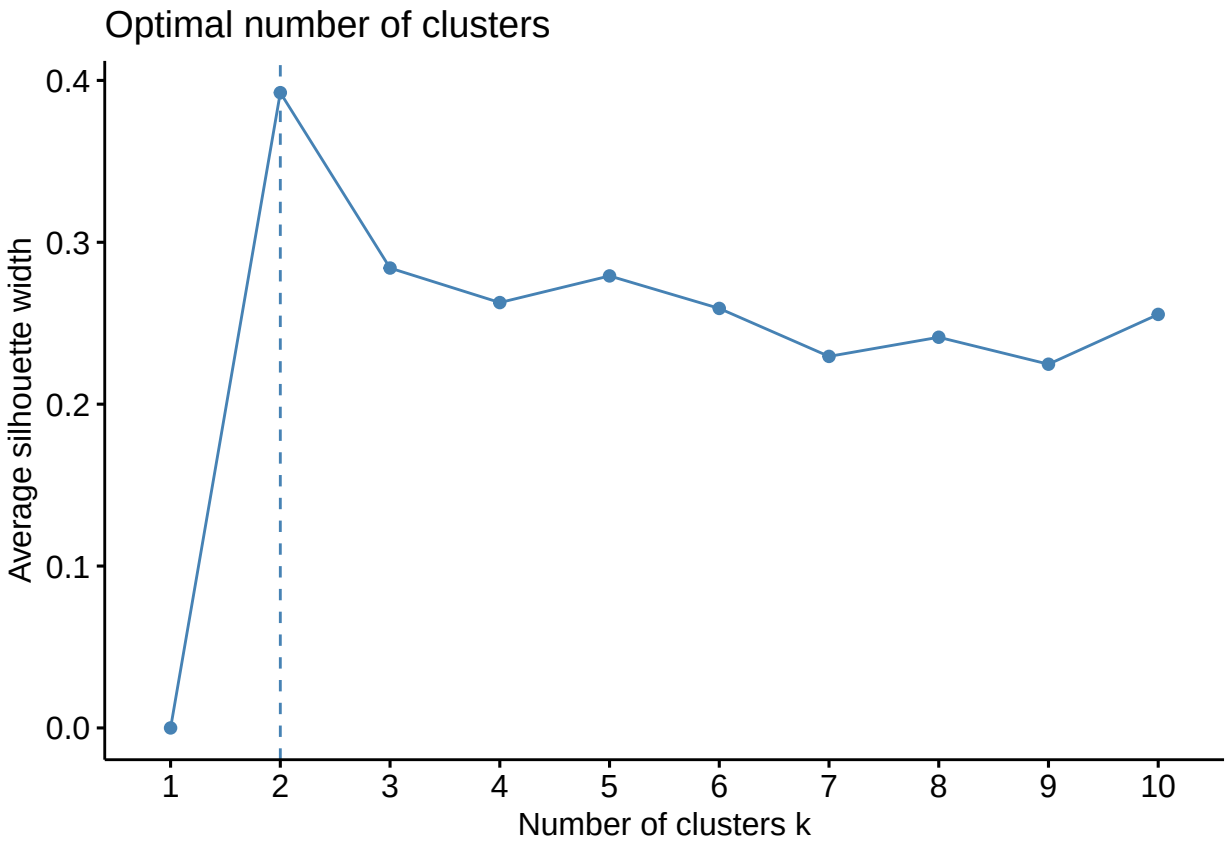
Se observa que se pueden sospechar de 2 o 3 grupos con altas correlaciones o características similares, se puede sospechar de que se trata de hombres y mujeres dado que como ya se ha mencionado en previas entregas estos poseen características o cualidades antropometricas significativamente diferentes, para este caso vamos a quedarnos con 3 grupos

```
clu_datos <- cutree(hw_datos, k = 3)
fviz_cluster(list(data=datos[,2:9], cluster=clu_datos))
```



Observe que las 2 primeras componentes principales logra explicar en 82,4% de la variabilidad de las observaciones, ahora, se aprecia que el traslape de los dos cluster es poca, la mayoría de las observaciones están bien separadas por lo que se confirman que seleccionar 2 cluster fue una aproximación a los grupos distintos de los que se tienen registros a hora bien esto no significa que 2 cluster son los óptimos para clasificar y esto lo comprobaremos a continuación.

```
fviz_nbclust(x=datos[,2:9], FUNcluster = kmeans, method="silhouette")
```



Obseve que cuando se tienen 2 cluster se observa que las observaciones se encuentra bien sobre su grupo esto se sabe gracias al calculo de la silueta promedio de los puntos para los k cluster y con esto se concluye que el numero optimo para clasificar por grupos son 2

Regla para discriminar poblaciones de hombres y mujeres

A continuación se va a obtener una regla que discrimine a las poblaciones.

Vectores de medias de hombres y mujeres

```
datosF <- datos[datos$base.Sexo == "Hom",1:9]
datosM<- datos[datos$base.Sexo =="Muj",1:9]
datosF$base.Sexo <- NULL
datosM$base.Sexo <- NULL
n1 <- nrow(datosF)
n2 <- nrow(datosM)

#vectores de medias
xbF <- round(apply(datosF,2,mean),5)
xbM <- round(apply(datosM,2,mean),5)
```

Matrices de varianzas y covarianzas

```
s1 <- round(cov(datosF),5)
s2 <- round(cov(datosM),5)
```

Observemos si hay normal multivariada, para ello hacemos la prueba de Shapiro ya que se tiene una buena cantidad de datos para esta prueba donde H0: las observaciones distribuyen normal multivariada vs H1: las observaciones no distribuyen normal multivariada

```
maux <- as.matrix(datos[,2:9], ncol=8)
mvShapiro.Test(maux)
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  maux
## MVW = 0.98388, p-value = 0.0004087
```

Observe que no tenemos normalidad multivariada

Observemos si las matrices de varianzas y covarianzas son iguales

```
library(heplots)
boxM(datos[,2:9],datos[,1])
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  datos[, 2:9]
## Chi-Sq (approx.) = 107, df = 36, p-value = 5.71e-09
```

las matrices de varianzas y covarianzas de ambas poblaciones no son iguales de la teoria se sabe que el test Box's M muestra mucha evidencia de que la matriz de covarianza no es igual para ambos sexos, esto hace que sea adecuado hacer uso de Clasificación mediante discriminante Cuadratico de Fischer con Matrices de Covarianza diferentes, se puede usar un discriminante de fisher.

en donde un regla de clasificacion para las poblaciones de hombre y mujeres esta dadon por la siguiente funcion

$$\widehat{W} = \left(\bar{X}_1 - \bar{X}_2\right)' S^{-} 1_p \left[x - \frac{1}{2}\left(\bar{X}_1 + \bar{X}_2\right)\right]$$

En lo que si esta expresión $\left(\bar{x}_1 - \bar{x}_2\right)' S_p^{-1} x$ es mayor o igual que $\frac{1}{2}\left(\widehat{y}_1 + \widehat{y}_2\right)$ entonces se clasifica como Hombre y en caso contrario se clasifica como mujer

Algoritmo de clasificacion:

Análisis discriminante cuadrático

Suponiendo que tenemos datos de solo las medidas antropometricas y queremos predecir de acuerdo a los datos los hombres y mujeres, podemos construir un modelo apartir de algoritmos de clasificacion que nos ayude a clasificar los hombres y mujeres, estos algorimos son importantes ya que en la industria se resuelven problemas como riesgo crediticio en los bancos (paga o no paga), o si compra o no compra el producto en una empresa que quiera vender.

Separemos un 30% de las observaciones y tomemos lo como base de prueba y el restante de entrenamiento de la siguiente forma

```
aux <- sample(1:150, 45)
testing <- datos[aux, ]
aprendizaje <- datos[-aux, ]
```

listo, ahora se montará el modelo que usa el algoritmo que mencionamos en el punto anterior usando la base de aprendizaje para clasificar nuestra información que apartamos en una base de datos llamada testing

```
datos_qda <- qda(base.Sexo ~base.p1+base.p7+base.p16+base.p22+base.p26+
                   base.p27+base.p29+base.p38,data=aprendizaje) #Entrenamiento del modelo

pred_datos <- predict(object=datos_qda, testing[,2:9]) #predicciones con la base testing
table(testing$base.Sexo, pred_datos$class, dnn = c("Real", "Predicha")) #matriz de confusión
```

```
##      Predicha
## Real  Hom Muj
##   Hom  28   0
##   Muj   1  16
```

Nuestro modelo tiene una muy buena capacidad predictiva, observe que de las 45 observaciones que se aparto en una base de datos llamada testing se tiene que el modelo tuvo una tasa de mala clasificacion estimada de 1/45

Regresion logistica

Dado que el problema es de clasificacion de dos poblaciones se puede emplear y montar un modelo de regresion logistica y comparemos su efectividad con relacion al modelo anterior, para ello usamos las misma variables testing y aprendizaje previamente creadas.

```
aprendizaje$base.Sexo <- ifelse(aprendizaje$base.Sexo=="Hom",1,0)#Hombre=1 y Mujeres=0

log_datos <- glm(base.Sexo ~base.p1+base.p7+base.p16+base.p22+base.p26+
                   base.p27+base.p29+base.p38, family=binomial, data=aprendizaje)
```

Hagamos las prediciones con la base testing y comparemos con las observaciones reales

```
pred_datos <- predict(log_datos, type='response', newdata=testing[,2:9])
pred_valid <- ifelse(pred_datos > 0.5, "Hom", "Muj")

# Por defecto la regresion logistica calcula la probabilidad de una
# respuesta = 1 "Exito"
# Es por esto que las categorías se cambian de  1 - 2 y de 0 - 1

# pred_valid_hem <- factor(pred_valid, levels=c(1,2), labels=c("NoPortador", "Portador"))
matrizConfusion <- table(testing$base.Sexo, pred_valid)
matrizConfusion
```

```
##      pred_valid
##       Hom Muj
##   Hom  28   0
##   Muj   0  17
```

Observe la anterior matriz de confusión que la capacidad predictiva con la regresion logistica tambien es muy buena, dandonos una tasa de mala clasificacion de 2/45

En conclucion ambos modelos no difieren mucho en cuanto a los resultados obtenidos, esto puede resultar dado que se tienen una buena cantidad de datos que le da mejor capacidad predictiva a estos algoritmo y modelos

K-vencions más cercanos

A continuacion se construye se generan 45 numero aleatorios del 1 al 150 para separa dos bases de datos, la de prueba y la de entrenamiento, para posteriormente aplicar el modelo asociado a K vecinos más cercanos el cual una nueva observación es clasificado como perteneciente a una clase si la mayoría de sus vecinos pertenece a dicha clase, Se calculan todas las distancias de la nueva observación a los grupos ya existentes en la base de datos, para por ultimo quedarse con la distancia más pequeña y lograr la clasificación de la nueva observación.

```
aux <- sample(1:150, 45)
testing <- datos[aux, ]
aprendizaje <- datos[-aux, ]

pred_knn <- knn(train=aprendizaje[,2:9], test=testing[,2:9], cl=aprendizaje$CAT_IMC, k = 3)
mat_Confu <- table(datos[aux,]$CAT_IMC, pred_knn)
print(mat_Confu)
```

```
##             pred_knn
##              Bajo_peso Normal sobre peso
##   Normal             0     23          0
##   sobre peso         0      7         15
```

Observando la anterior matriz de confusión se tiene como resultado que de la 45 datos que se extrajo para hacer el test se aprecio que el modelo solo esta fallando en 7 es decir que la tasa de mala calificacion es de 7/45