Técnicas de análisis Multivariado

Alumnos:
Huertas Quispe, Anthony Enrique ***Cod:*** *20173728*
Torres Salinas, Karina Hesi ***Cod:*** *20164111*
Córdova Proleón, Christian Therius ***Cod:*** *20173970*

Semestre: *2017-II*

Tema: Lista 2

Prof. Enver Tarazona

Pontificia Universidad Católica del Perú
Escuela de Posgrado
Maestría en Estadística

Pregunta 1 (4 puntos)

El conjunto de datos del archivo wisc_bc_data.csv contiene información para 569 instancias relacionadas a biopsias por sospecha de cáncer de mama, cada una de ellas conteniendo 32 atributos. Uno de los atributos es un número de identificación, otro es el diagnóstico de cáncer y los 30 atributos restantes están relacionados a mediciones numéricas realizadas en laboratorio. El diagnóstico fue codificado como M que indica que es maligno o B que indica que es benigno. El objetivo del análisis es predecir si un paciente tiene cáncer basado en los resultados del laboratorio. Compare varios modelos de clasificación (logístico, LDA, QDA y RDA) y encuentre aquel que tenga mejor desempeño para el conjunto de datos de entrenamiento. Utilice diversos criterios de comparación la exactitud total, índice Kappa de Cohen, la sensibilidad, especificidad y el AUC (área bajo la curva ROC). Presente una gráfica comparativa que muestre las curvas ROC para los clasificadores usados.

Desarrollo. Se estableció en Anexo 1, como categoría de referencia al diagnóstico de cáncer "maligno". La base de datos "sf wisc_bc_data.csv" que cuenta con 569 registros se dividió en 2 base de datos. La primera parte tendría el 80 % de la base de datos original, esto nos servirá como entrenamiento para la construcción del modelo; mientras que el otro 20 % de la base de datos original la usaremos para la validación del modelo.

Indicadores	Logístico	LDA	QDA	RDA
Exactitud total	0.930	0.947	0.965	0.956
Kappa	0.854	0.889	0.927	0.908
Sensibilidad	0.913	0.976	0.957	0.977
Especificidad	0.941	0.931	0.971	0.944
AUC	0.927	0.938	0.964	0.949

Del cuadro resumen, se tiene que el modelo con el análisis discriminante cuadrático es el que tiene una mejor precisión o tasa de aciertos (96.5 %), mientras que con el modelo logístico se consigue la menor precisión (93.0 %). En cuanto al índice Kappa de Cohen, se tiene una mejor concordancia (casi perfecta) utilizando el modelo discriminante cuadrático en comparación a los otros tres modelos, pues se tiene un índice de 0.927.

La sensibilidad con mayor valor se consigue usando el modelo discriminante lineal (0.976) y el modelo discriminante regularizado (0.977). La especificidad más grande se obtendría usando el modelo discriminante cuadrático (0.971), lo que significa que el 97.1 % son correctamente diagnosticados como cáncer "benigno". La curva ROC que presenta una mayor área es también utilizando el modelo discriminante cuadrático (0.964).

En conclusión, el mejor modelo que presenta un mejor desempeño con la base de datos para la evaluación del modelo, sería empleando el discriminante cuadrático.

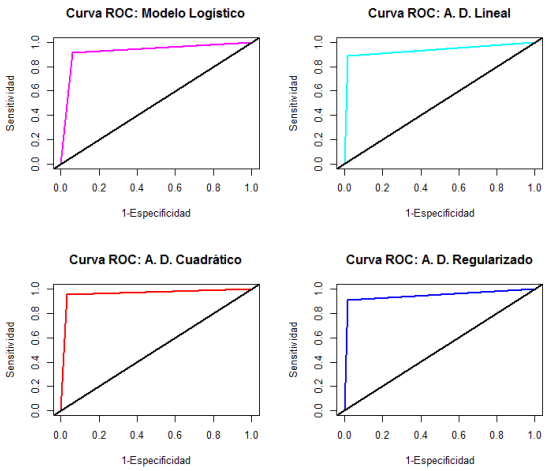


Figura 1: Curvas ROC

Pregunta 2 (4 puntos)

Wolford y Hollingsworth (1974) estaban interesados en las confusiones que se producen cuando una persona intenta identificar las letras del abecedario visualizadas solo durante algunos milisegundos. Se construyó una matriz de confusión que muestra la frecuencia con la que cada letra de estímulo se llamó erróneamente por otra. Una sección de esta matriz se muestra en la tabla a continuación.

Letter	C	D	G	H	M	N	Q	W
C	-							
D	5	-						
G	12	2	-					
H	2	4	3	-				
M	2	3	2	19	-			
N	2	4	1	18	16	-		
Q	9	20	9	1	2	8	-	
W	1	5	2	5	18	13	4	-

Realice un análisis de Escalamiento Multidimensional e interprete los resultados. ¿Qué se puede concluir en relación a la confusión entre las letras?

Desarrollo. Para ver que podemos concluir en relación a la confusión entre las letras, hay que encontrar una matriz de similaridad entre los elementos a partir de la matriz de confusión, para luego construir una matriz de distancias de disimilaridades d_{ij}.

Entonces de las similaridades s_{ij}, escogemos una similaridad máxima $c \geq \text{máx}\{s_{ij}\}$, tal que la matriz de distancias de similaridades tendría los elementos:

$$d_{ij} = \begin{cases} c - s_{ij} & \text{si } i \neq j, \\ 0 & \text{si } i = j \end{cases}$$

Dado que las disimilaridades dependen del valor de c elegido el escalamiento multidimensional no métrico sería el más apropiado. Sea la matriz de confusión:

Listing 1: Matriz de confusión.

```
letras=as.dist(as.matrix(letras))
letras
```

```
   C  D  G  H  M  N  Q
D  5
G 12  2
H  2  4  3
M  2  3  2 19
N  2  4  1 18 16
Q  9 20  9  1  2  8
W  1  5  2  5 18 13  4
```

Figura 2: Matriz de confusión

Escogemos $c = 21$ que es el máximo de los $s_{ij}+1$, entonces tenemos nuestra matriz de distancias de disimilaridades:

Listing 2: Matriz de distancias de disimilaridades.

```
letras=21-letras
letras
```

```
   C  D  G  H  M  N  Q
D 16
G  9 19
H 19 17 18
M 19 18 19  2
N 19 17 20  3  5
Q 12  1 12 20 19 13
W 20 16 19 16  3  8 17
```

Figura 3: Matriz de distancias de disimilaridades

Usamos la función smacofsym porque estamos usando una matriz simétrica, el número de dimensiones en el cual queremos representar la distancia entre las letras es 2, entonces de los resultados podemos observar que hemos necesitado 30 iteraciones para generar la estimación final. La medida de estrés es de 0.06, en general podríamos decir que tenemos una representación buena y eso es adecuado.

Listing 3: Matriz de distancias de disimilaridades.

```
res=smacofSym(letras,ndim=2, type = "ordinal")
res
summary(res)
```

```
Call:
smacofSym(delta = letras, ndim = 2, type = "ordinal")

Model: Symmetric SMACOF
Number of objects: 8
Stress-1 value: 0.06
Number of iterations: 30
```

```
Configurations:
       D1      D2
C -0.7234 -0.2231
D -0.3098  0.5664
G -0.6208 -0.5749
H  0.5748 -0.3876
M  0.6656 -0.0829
N  0.4167  0.0494
Q -0.4811  0.2942
W  0.4781  0.3584
```

```
Stress per point (in %):
    C     D     G     H     M     N     Q     W
 2.50  5.76 15.36 15.38 12.94 23.24  7.55 17.27
```

Figura 4: Cuadro 1.

Del cuadro anterior observamos que los puntos de estrés son altos, siendo la letra N la que está peor representada y la que tiene menor estrés es la letra C aunque no es cercano a cero.

Evaluando los resultados de manera más general hemos visto que la medida de estrés es regularmente baja con el cuál podríamos tener resultados con el que podríamos estar satisfechos.

Del siguiente mapa perceptual observamos que hay cercanía entre las letras D y Q, así como las letras C y G, otro grupo que podría presentar ciertas características similares son las letras W, N ,M y H. La dimensión 1 si bien no tiene una explicación ayuda a diferenciar un poco más claro los grupos de letras (W, N, M y H) , (D , Q) y (C, G).

Para sustentar que los resultados sean coherentes observamos la siguiente gráfica de Shepard, donde la idea es observar una función que sea no decreciente, observamos la gráfica vemos que

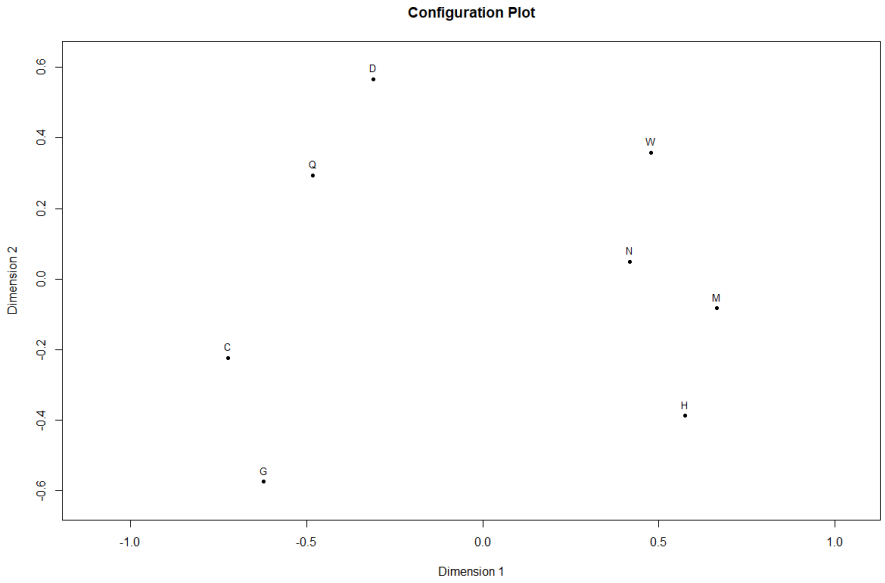


Figura 5: Configuración.

tenemos una función no decreciente y eso nos ayuda a poder evaluar que estamos teniendo resultados adecuados para representar a las letras en 2 dimensiones.

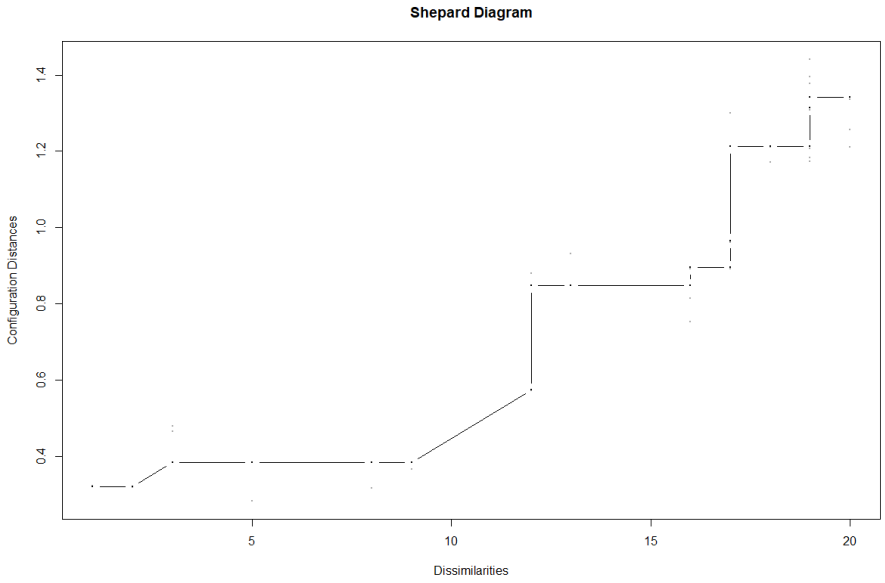


Figura 6: Diagrama Shepard.

En el siguiente gráfico de residuos, esperamos que haya cierta relación lineal, del gráfico se observa que hay una tendencia lineal entre las disimilaridades y las distancias que tenemos, asimismo no se observan outliers.

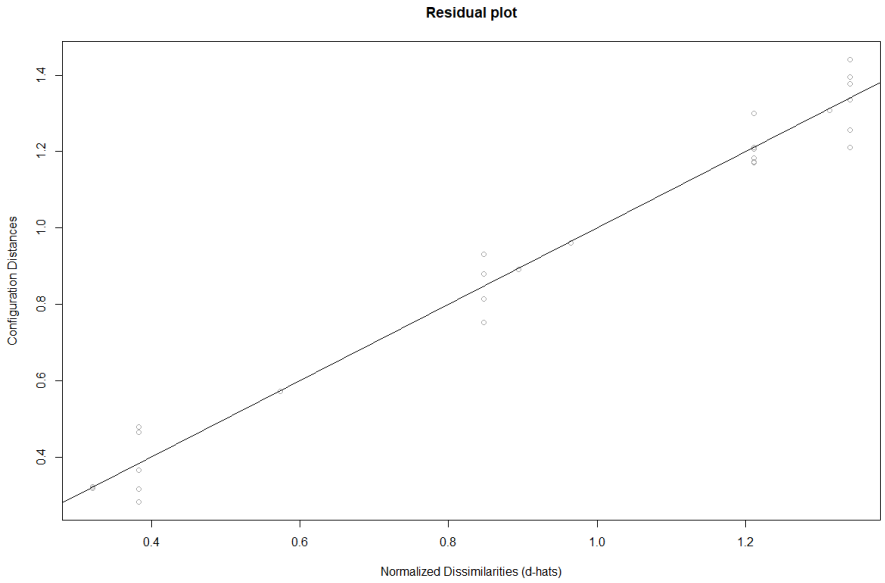


Figura 7: Gráfica de residuales.

Para evaluar que tan asociadas están esas 2 medidas, se sacó la correlación al cuadrado saliendo 0.9758 , un valor muy cercano a 1.

En el siguiente gráfico podemos observar el estrés individual para cada uno de las letras, ya habíamos observado antes que el que se encuentra mejor representado es la letra C y la letra N es el que tiene mayor medida de estrés en comparación a todos los demás.

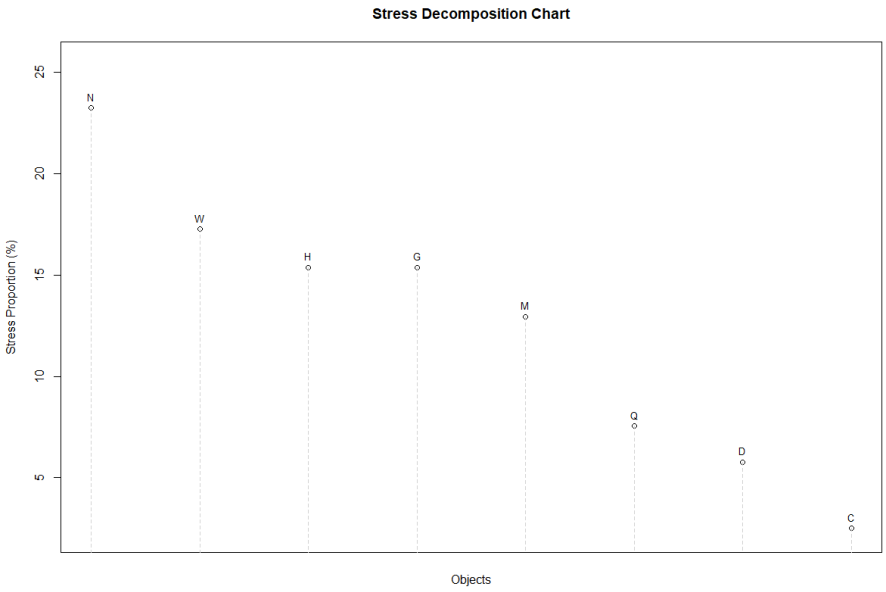


Figura 8: Gráfico de Estrés.

Con el gráfico de burbuja tenemos la representación del mapa perceptual y en simultáneo cuál es la medida de estrés individual que tenemos para cada una de las letras:

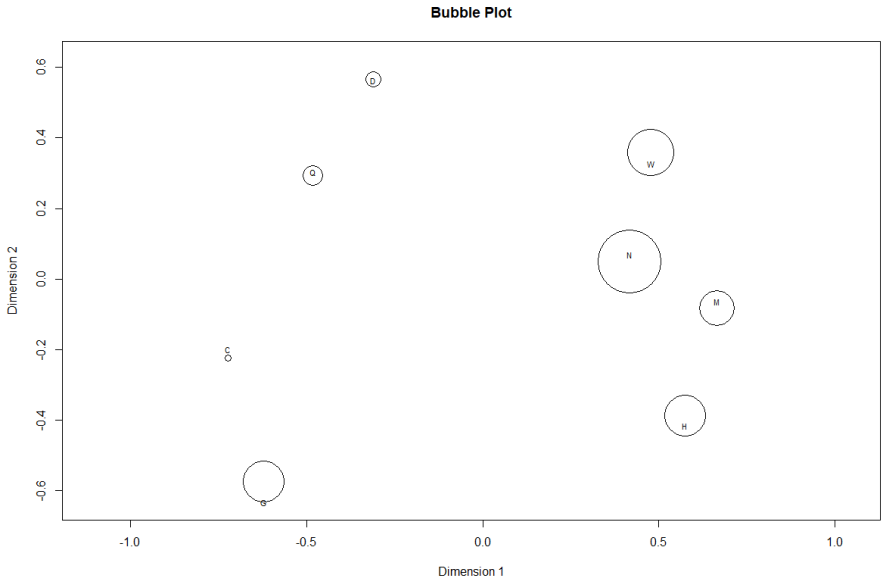


Figura 9: Gráfico de burbuja.

Usando un agrupamiento jerárquico aglomerativo con enlace promedio, se forman los siguientes cluster: (W, H, M, N) , (D, Q) y (C, G).

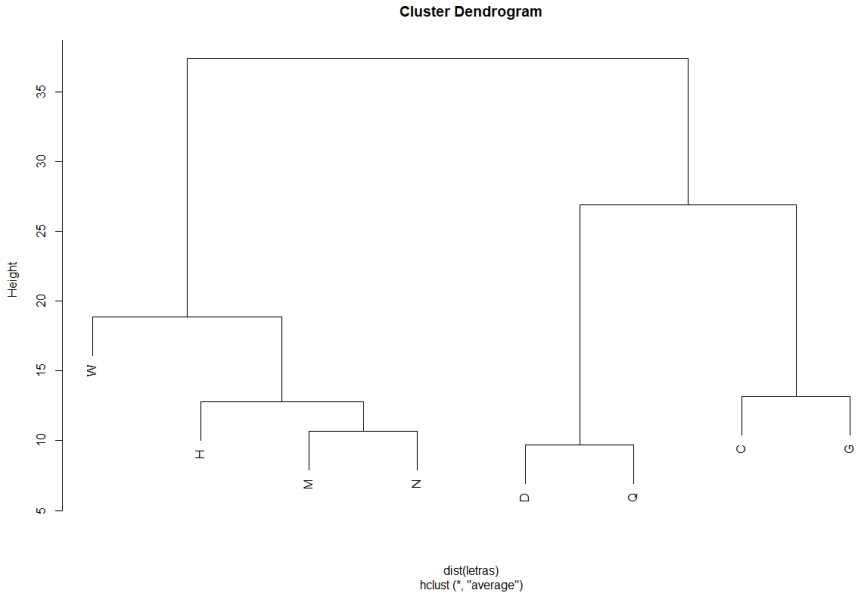


Figura 10: Dendograma.

Lo que se concluye es que dado que las personas tienen poco tiempo para visualizar las letras del abecedario, tienden a confundirlas de acuerdo a ciertas características y formas de éstas, por ejemplo (D y Q) son letras cerradas y parecieran redondas, las letras (C y G) son semi cerradas y el último grupo (W, H, M y N) tienen trazos marcados tanto verticales como diagonales unidas por ángulos.

Pregunta 3 (8 puntos)

Aplique al menos 2 de las metodologías presentadas en el curso (o una distinta) a un conjunto de datos reales.

Desarrollo. Se hará uso de una base de datos original desarrollada por *School of Information and Computer Science, University of California*. Esta base de datos cuenta con seis variables biomecánicas (explicativas) medidas en pacientes ortópedicos con el fin de categorizar si un paciente presenta alguna malformación (anormal) o no la presente (normal); :

- incid.pelvica: Incidencia pélvica.
- incl.pelvica: Inclinación pélvica.
- angulo.lumbar: Ángulo de lordosis lumbar.
- pend.sacral: Ángulo lumbar.
- rad.pelvico: Radio pélvico.
- grado.spond: Grado de spondylolisthesis.
- clase: Anormal y Normal.
- subclase: Anormal (Hernia, Spondylolisthesis) y Normal.

Clustering

El objetivo será agrupar los datos mediante un lenguaje no supervisado, de tal forma que lo obtenido infiera los grupos que originalmente son presentados es decir Normal y Anormal.

Primero diseñaremos una primera base haciendo uso de todos las variables, adicionando otras más, para luegos tomar el respectivo subconjunto de variables con las cuales trabajaremos en esta metodología.

Listing 4: Base de datos.

```
head(datos)
```

	incid.pelvica	incl.pelvica	angulo.lumbar	pend.sacral	rad.pelvico	grado.spond	clase
1	63.02782	22.552586	39.60912	40.47523	98.67292	-0.254400	Abnormal
2	39.05695	10.060991	25.01538	28.99596	114.40543	4.564259	Abnormal
3	68.83202	22.218482	50.09219	46.61354	105.98514	-3.530317	Abnormal
4	69.29701	24.652878	44.31124	44.64413	101.86850	11.211523	Abnormal
5	49.71286	9.652075	28.31741	40.06078	108.16872	7.918501	Abnormal
6	40.25020	13.921907	25.12495	26.32829	130.32787	2.230652	Abnormal

	subclase	clase.grupo	subclase.grupo
1	Hernia	1	1
2	Hernia	1	1
3	Hernia	1	1
4	Hernia	1	1
5	Hernia	1	1
6	Hernia	1	1

Figura 11: Datos Generales; Clase.grupo: 2 = Normal, 1 = Anormal; Clase.grupo: 2 = Normal, 1 = Anormal Hernia, 3 = Anormal Spondylolisthesis

Se desarrollará un resumen descriptivo de las variables explicativas, y un análisis de caja usando los grupos de clase indicados en la base.

```
incid.pelvica     incl.pelvica     angulo.lumbar     pend.sacral      rad.pelvico
Min.   : 26.15   Min.   :-6.555   Min.   : 14.00   Min.   : 13.37   Min.   : 70.08
1st Qu.: 46.43   1st Qu.:10.667   1st Qu.: 37.00   1st Qu.: 33.35   1st Qu.:110.71
Median : 58.69   Median :16.358   Median : 49.56   Median : 42.40   Median :118.27
Mean   : 60.50   Mean   :17.543   Mean   : 51.93   Mean   : 42.95   Mean   :117.92
3rd Qu.: 72.88   3rd Qu.:22.120   3rd Qu.: 63.00   3rd Qu.: 52.70   3rd Qu.:125.47
Max.   :129.83   Max.   :49.432   Max.   :125.74   Max.   :121.43   Max.   :163.07
 grado.spond          clase
Min.   :-11.058   Abnormal:210
1st Qu.:  1.604   Normal  :100
Median : 11.768
Mean   : 26.297
3rd Qu.: 41.287
Max.   :418.543
```

Figura 12: Resumen descriptivo

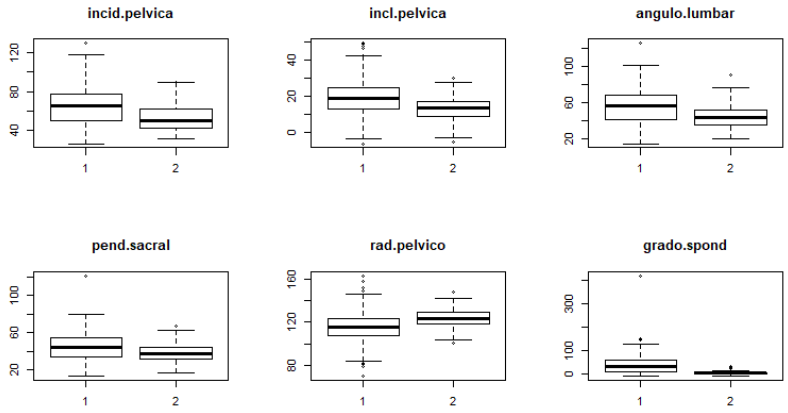


Figura 13: Diagrama de Caja - Grupos de clase originales

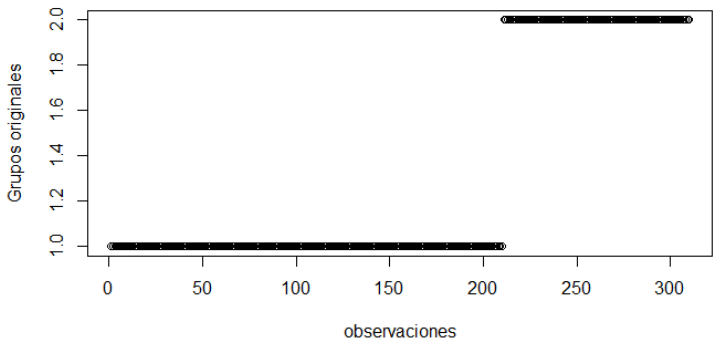


Figura 14: Observaciones vs Grupo de Clase

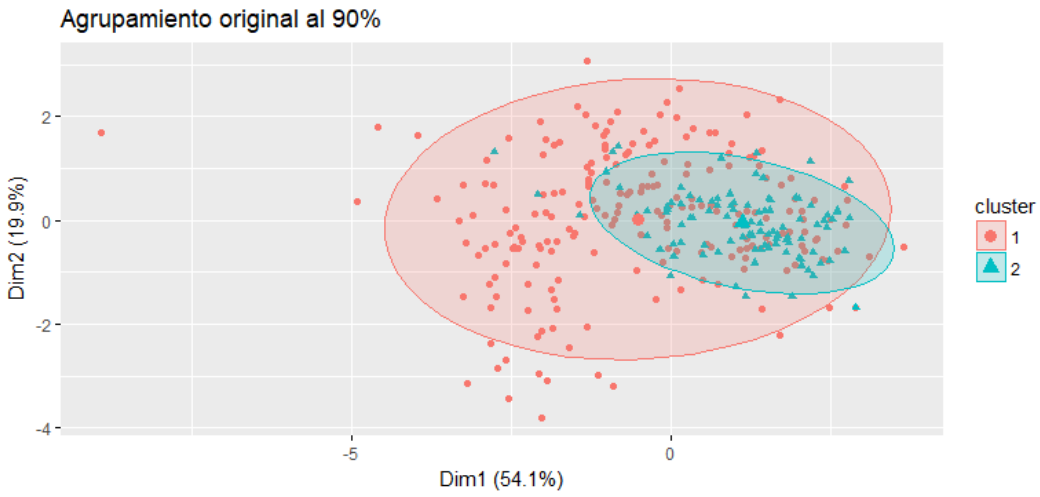


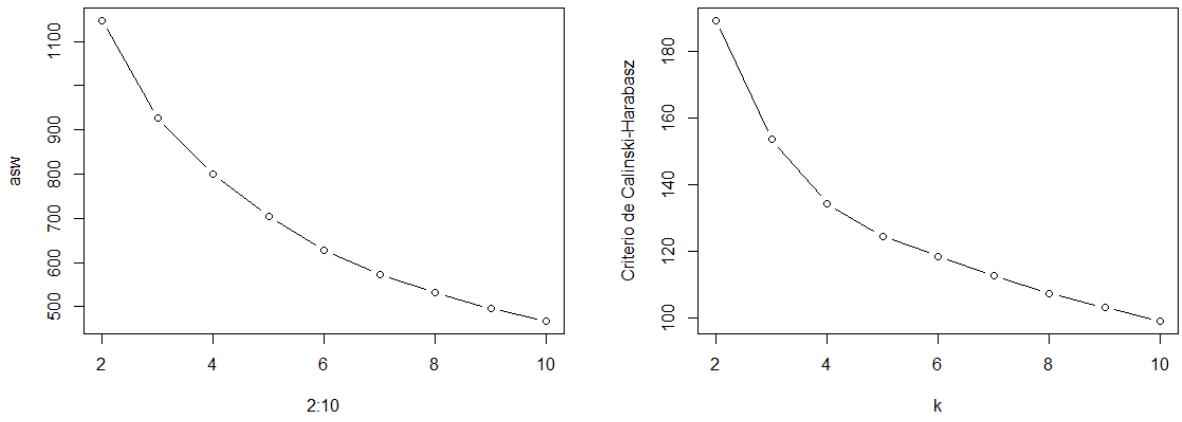
Figura 15: Agrupamiento por clases originales en un 90 %

Los gráficos 13, 14 y 15 nos representan los datos respecto a sus grupos originales, donde 1 = Anormal y 2 = Normal. El fin del uso de esta metodología es modelar los datos para que se agrupen mediante la información procedente de sus variables explicativas sin variables respuesta; y observar que efectivamente el ajuste por clustering nos lleva a un aproximado a la realidad. Esto con el objetivo de poder usar un modelo para agrupar datos nuevos.

Como hemos observado solo estamos analizando sobre la clase y no la subclase, es decir con 2 agrupamientos. Analicemos si en efecto, tiene sentido el uso de dos conglomerados mediante distintos métodos.

Listing 5: kmeans.

```
SC.cluster(datos1,1)
silueta(datos1,1)
kmeansruns(scale(datos1),criterion="asw")
Cal.Har(datos1,1)
kmeansruns(scale(datos1),criterion="ch")
```



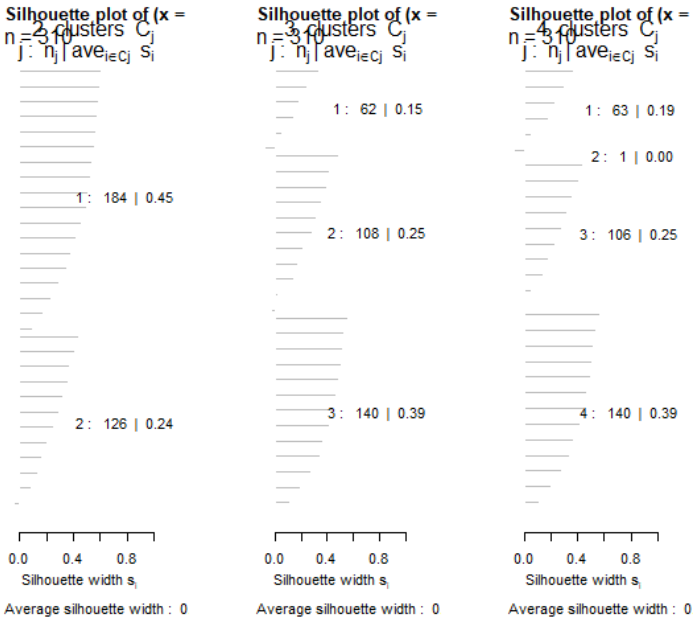
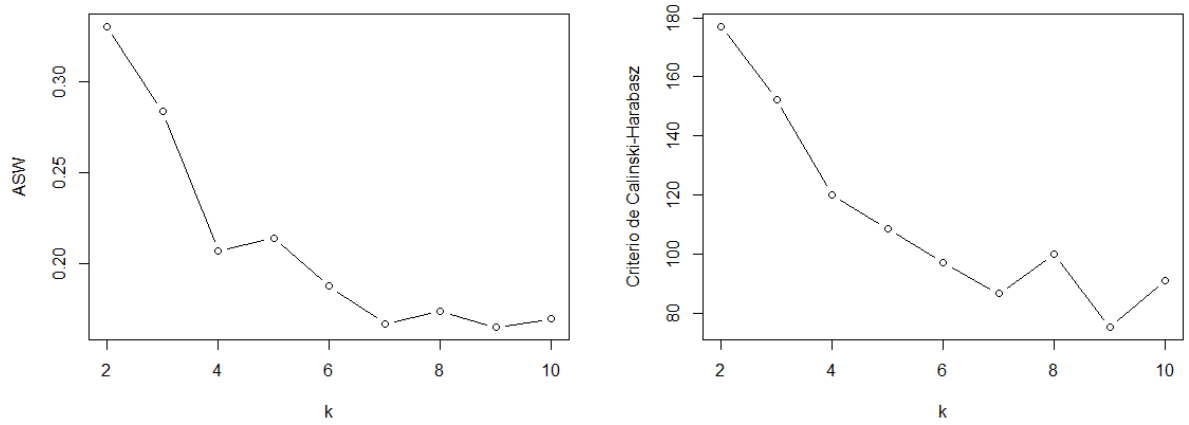


Figura 16: kmeans - Silueta

Listing 6: PAM.

```
SC.cluster(datos1,2)
silueta(datos1,2)
pamk(scale(datos1),criterion="asw")
Cal.Har(datos1,2)
pamk(scale(datos1),criterion="ch")
```



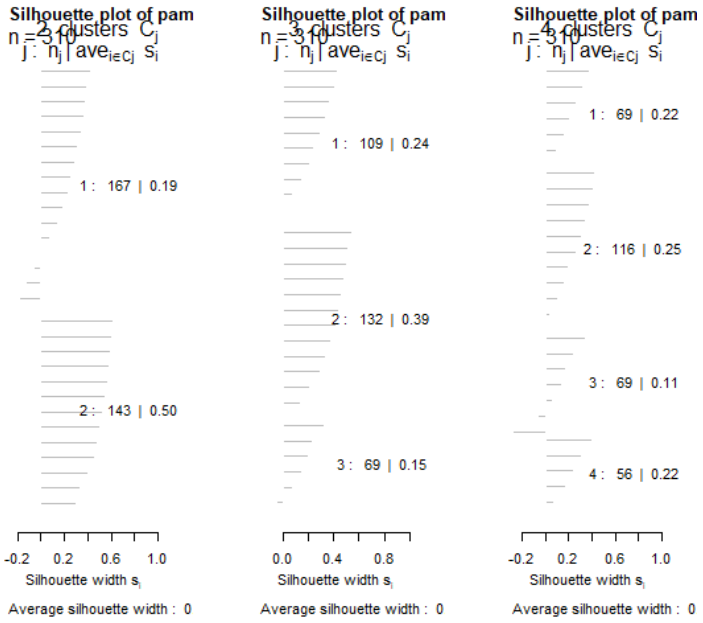
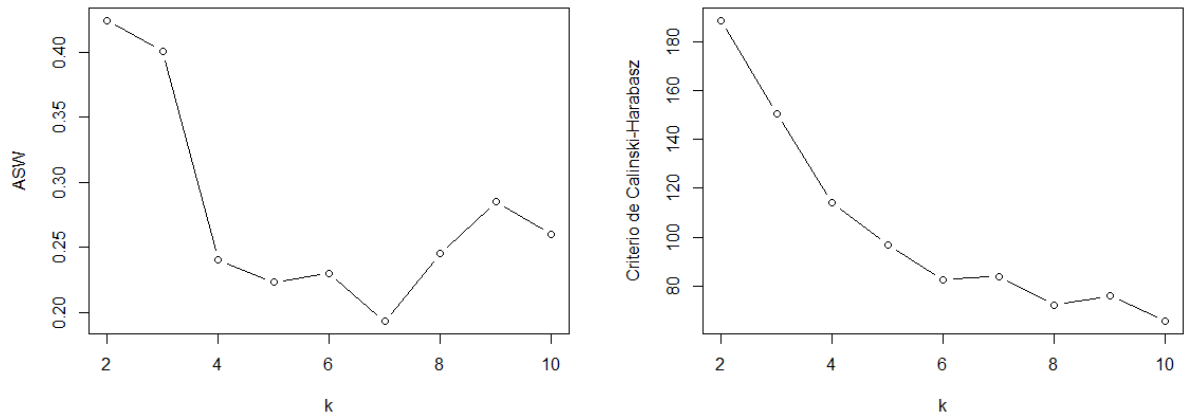


Figura 17: PAM - Silueta

Listing 7: Clara.

```
SC.cluster(datos1,3)
silueta(datos1,3)
pamk(scale(datos1),criterion="asw",usepam=FALSE)
Cal.Har(datos1,3)
pamk(scale(datos1),criterion="ch",usepam=FALSE)
```



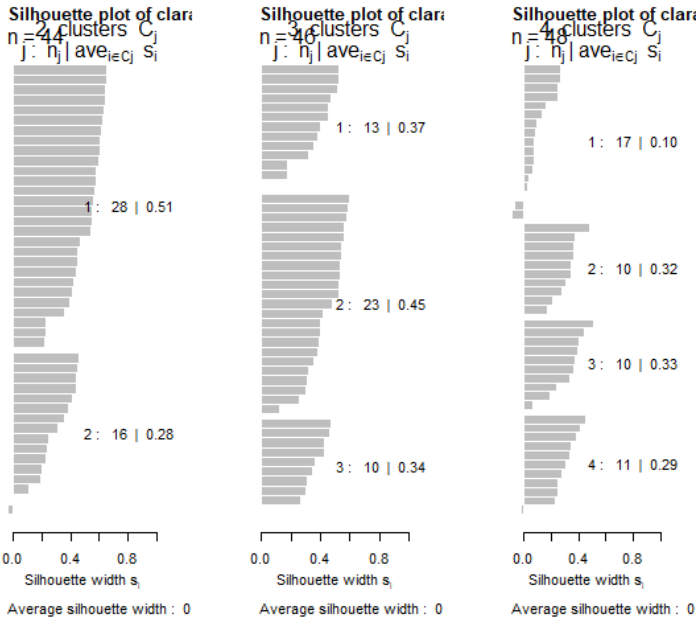
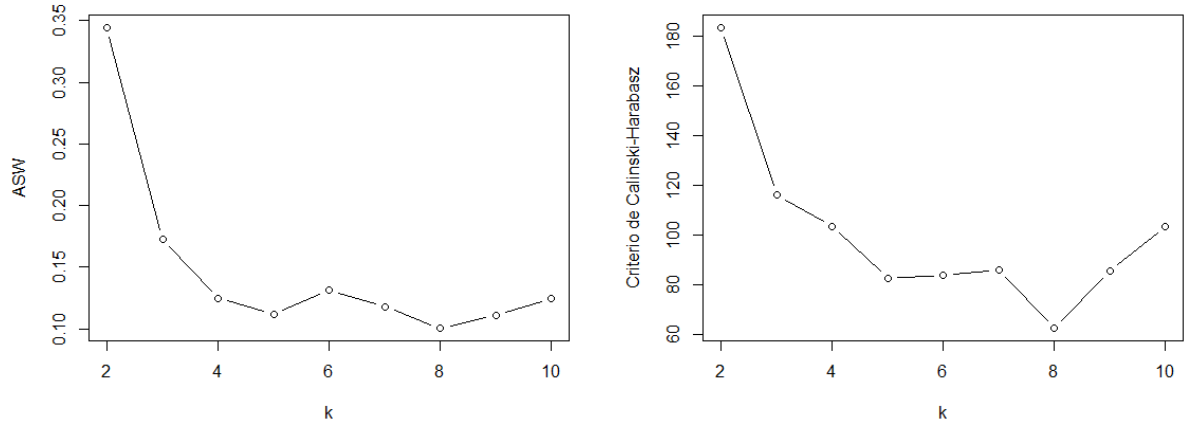


Figura 18: PAM - Silueta

Listing 8: Fanny.

```
SC.cluster(datos1,4)
silueta(datos1,4)
Cal.Har(datos1,4)
```



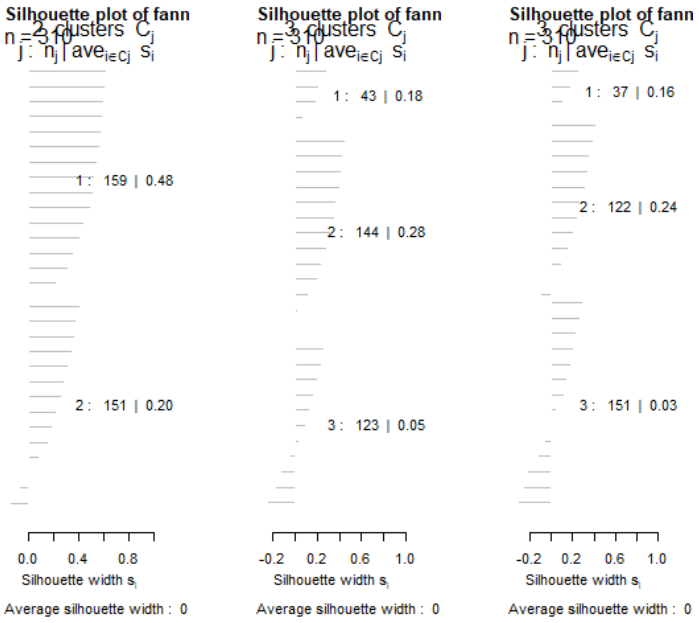


Figura 19: PAM - Silueta

Después de haber obtenido todos los gráficos, en efecto parece ser eficiente el uso de dos conglomerados, tanto por el índice de mayor valor generado en la silueta de cluster 2 como por los criterios de suma de cuadrados y Criterio de Calinski-Harabasz.

A continuación, Se observarán como se han agrupado los datos mediante los métodos.

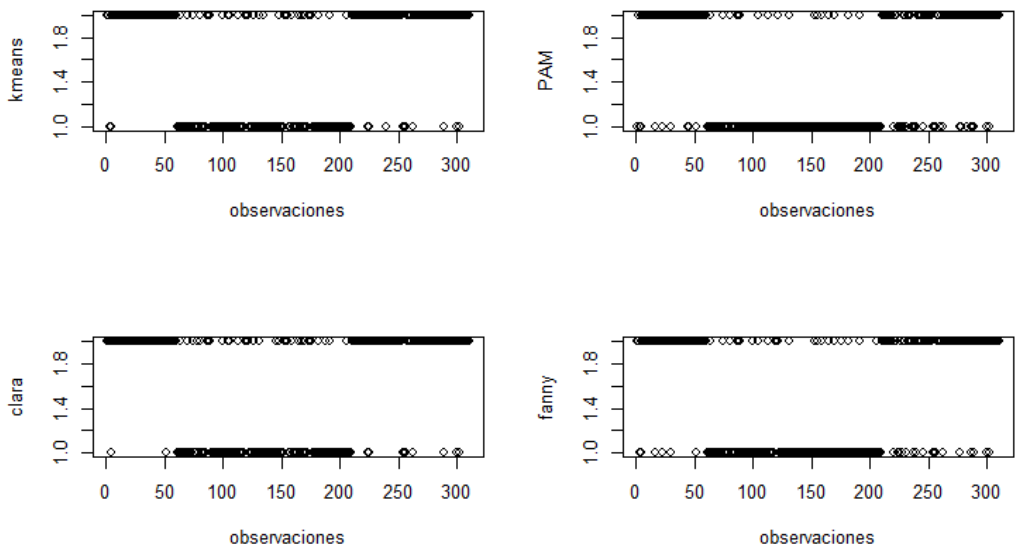


Figura 20: Agrupamiento

Como se observa, en relación al agrupamiento real (figura 14), la tendencia parece darse. Usaremos a continuación métodos de cluster jerárquicos diseñados por arboles de decisión.

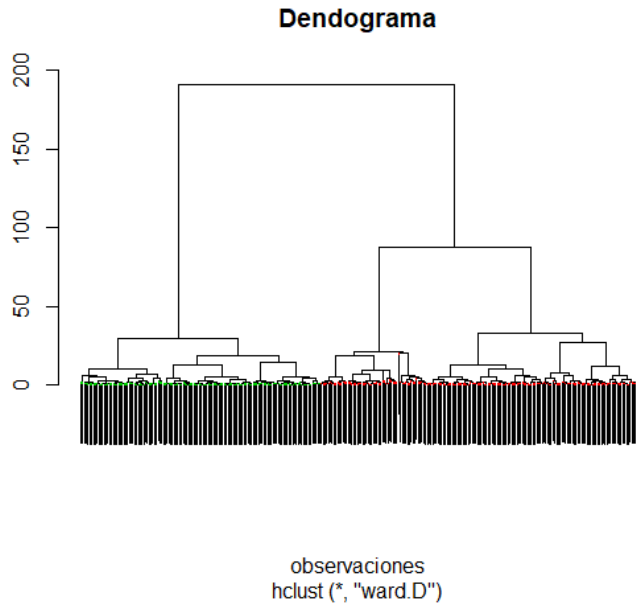


Figura 21: Dendograma HIERARCHICAL

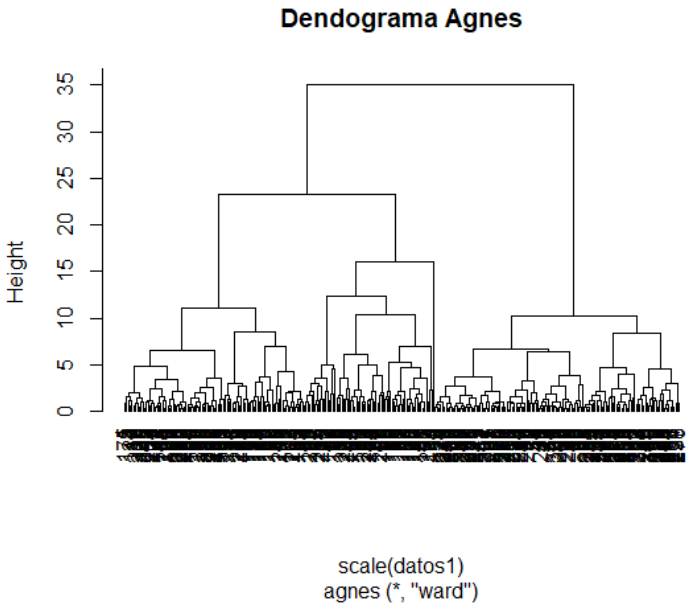


Figura 22: Dendograma Agnes

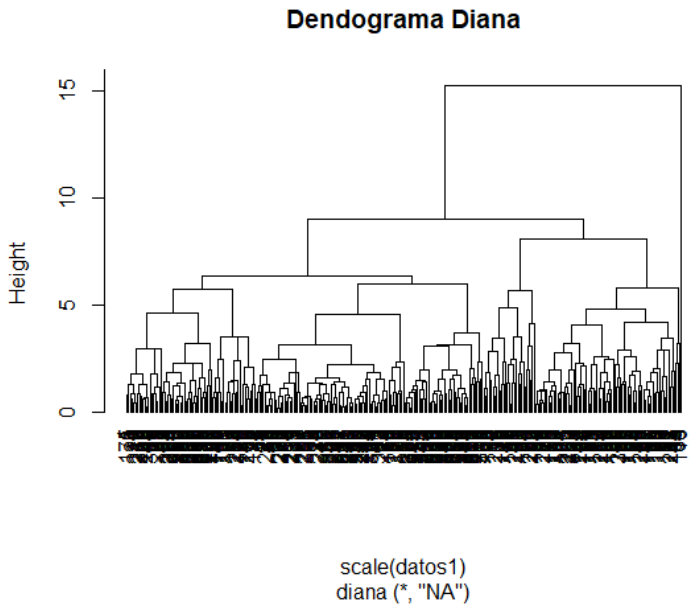


Figura 23: Dendograma Diana

Analicemos gráficamente, el agrupamiento de los datos para observar mejor la tendencia en la agrupación.

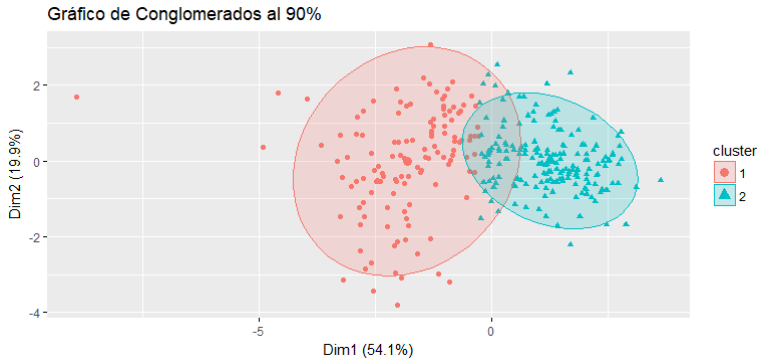


Figura 24: kmeans - Agrupamiento por clases en un 90 %

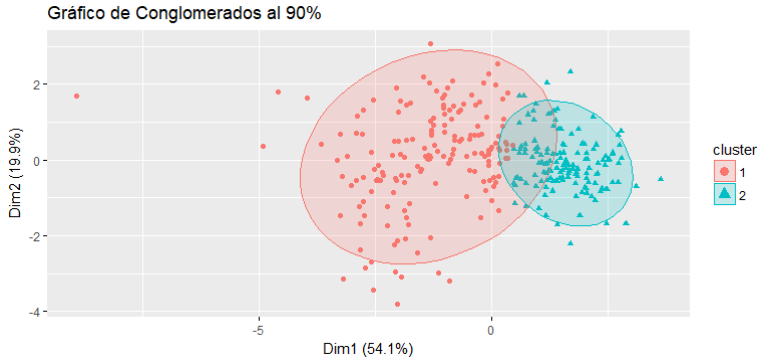


Figura 25: PAM - Agrupamiento por clases en un 90 %

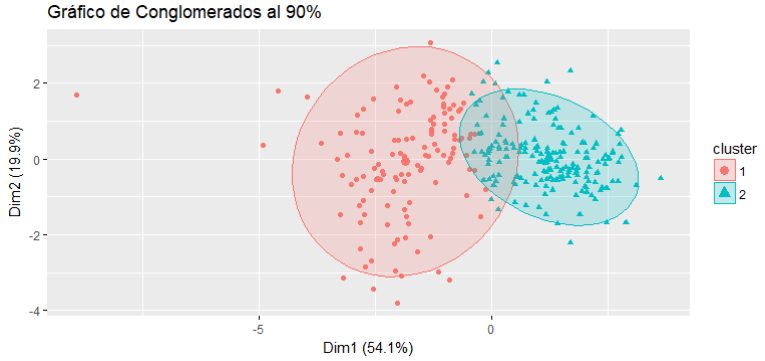


Figura 26: Clara - Agrupamiento por clases en un 90 %

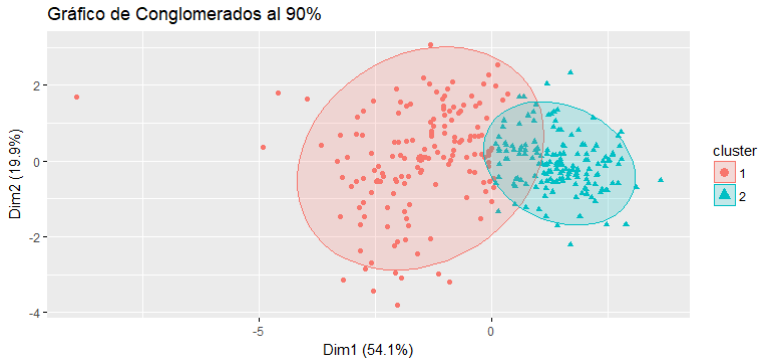


Figura 27: Fanny - Agrupamiento por clases en un 90 %

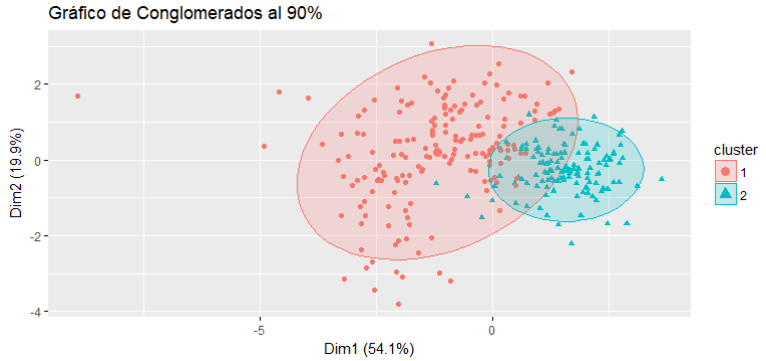


Figura 28: HIERARCHICAL - Agrupamiento por clases en un 90 %

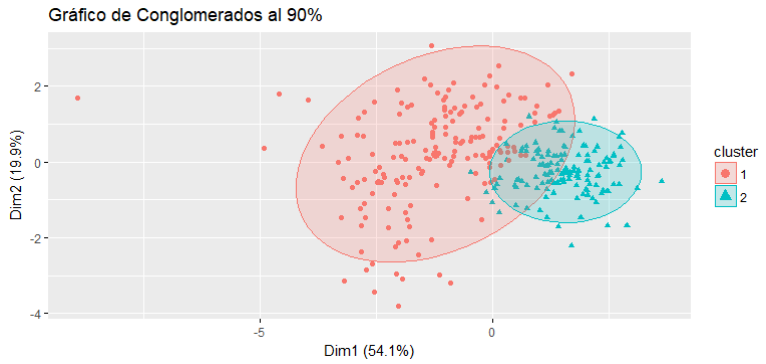


Figura 29: Agnes - Agrupamiento por clases en un 90 %

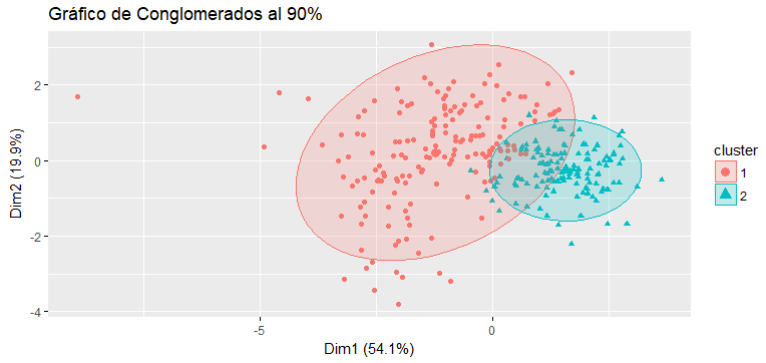


Figura 30: Diana - Agrupamiento por clases en un 90 %

En efecto, se sigue un agrupamiento acorde a lo establecido por la base de datos original, sin embargo es necesario realizar un análisis, con mejor resumen por diagrama de caja, de estos nuevos agrupamiento para corroborar que el modelo trabaja de forma adecuada.

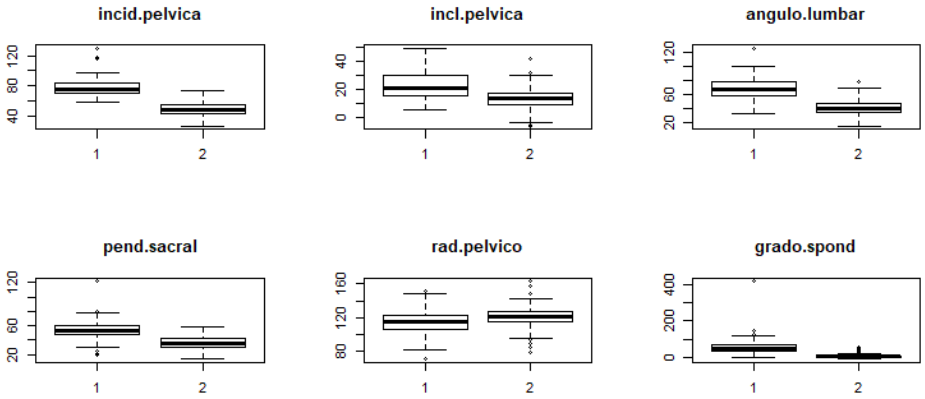


Figura 31: kmeans

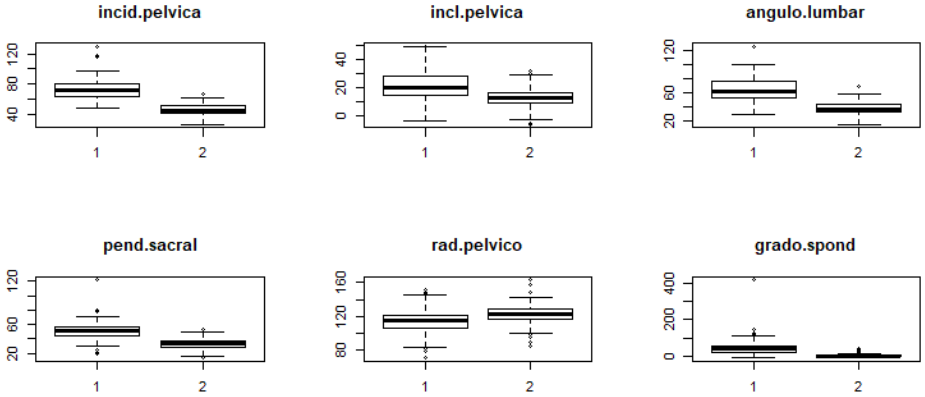


Figura 32: PAM

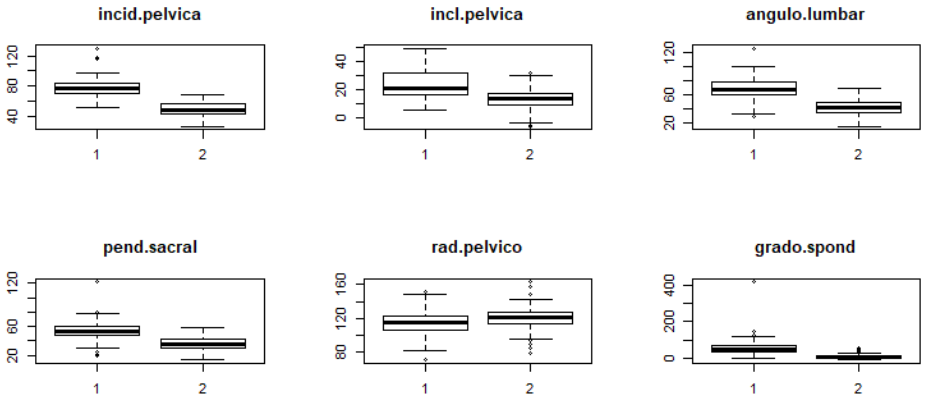


Figura 33: Clara

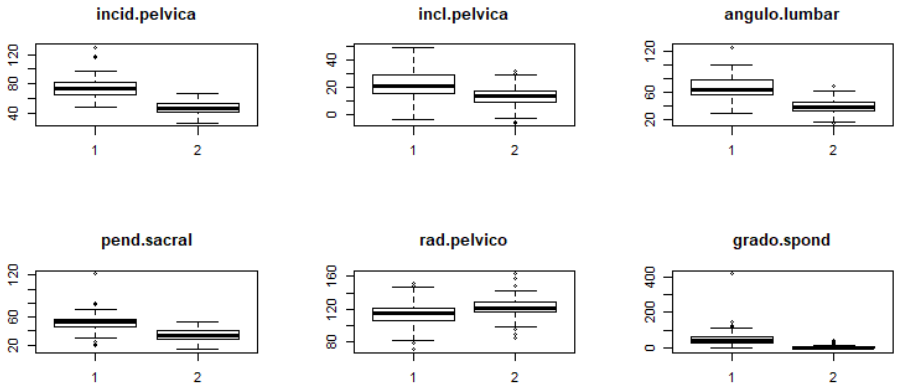


Figura 34: Fanny

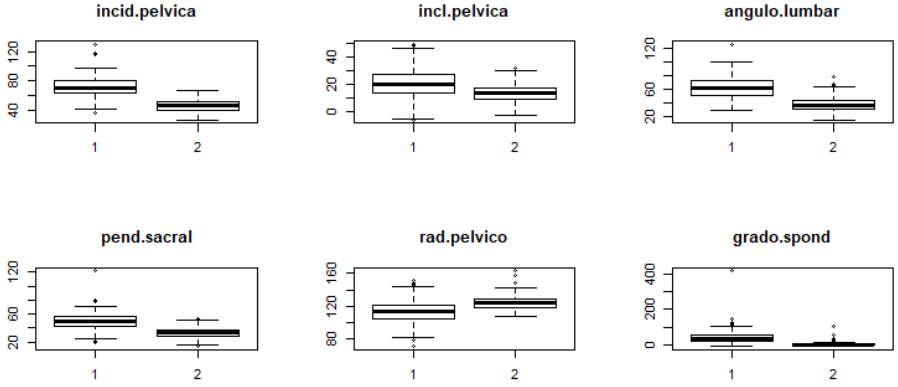


Figura 35: Hierarchical

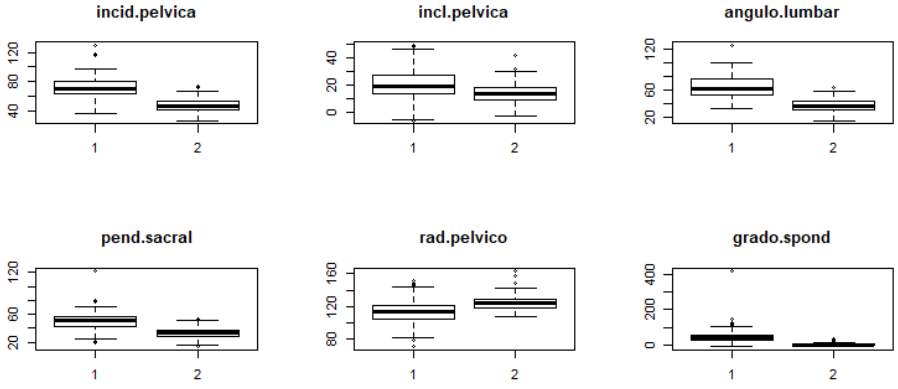


Figura 36: Agnes

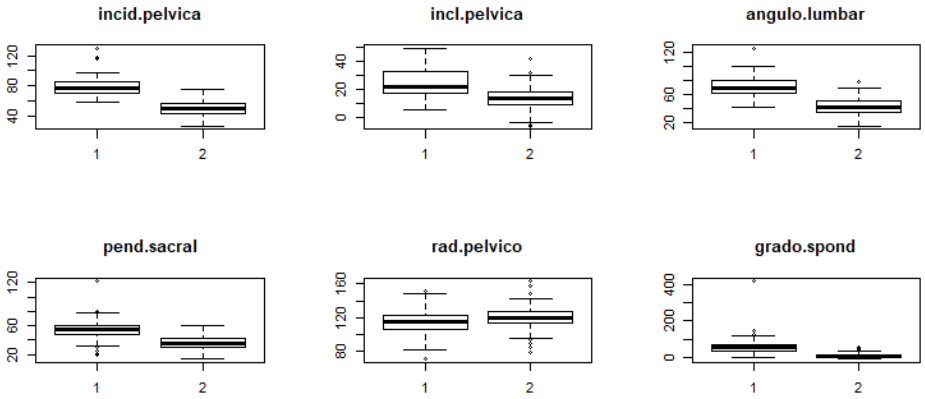


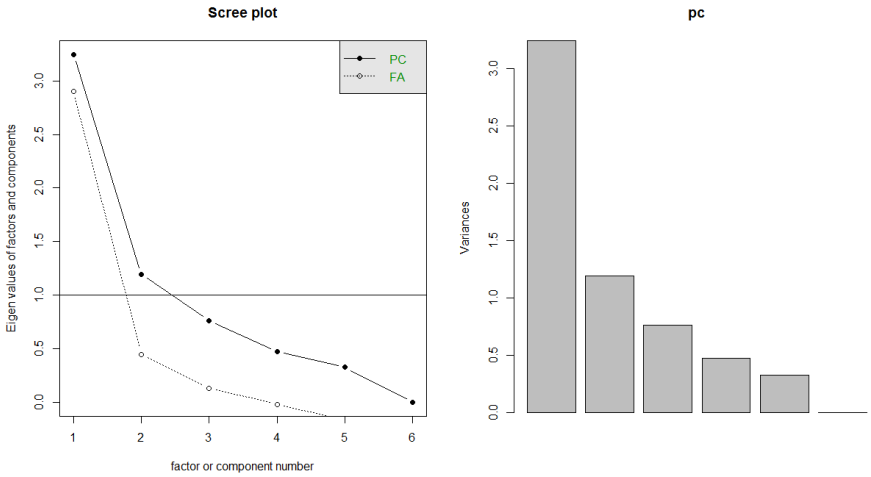
Figura 37: Diana

Evaluando el diagrama de cada de la Figura 13. En conclusión se ha efectuado un correcto agrupamiento por parte de los modelos.

Análisis de Componentes Principales (PCA)

Esta metodología buscará relacionar las distintas variables de la base de datos, con el objetivo de analizar correspondencias entre ellas.

El gráfico de Sedimentación nos ayudará a elegir el número de componente que retengan la mayor información posible. En el siguiente gráfico (línea "PC"), se observa que a partir del tercer componente, la pendiente formada por los autovalores se estabiliza, por lo que se decide usar solo dos componentes, de otra manera, si tomamos en cuenta la regla de Kaiser que nos sugiere tomar tantos componentes como autovalores mayores a la unidad, optaríamos también usar dos componentes. Usar las dos primeras componentes, implica también que ambas expliquen el 86.8 % de la varianza total. La tercera componente y cuarta componente solo aportarían el 8.9 % y 4.3 % de la varianza total.

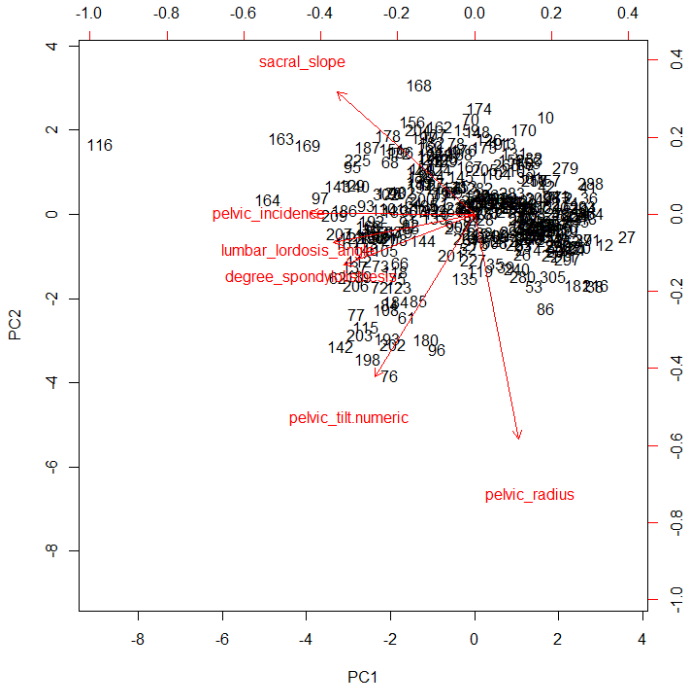


```
Importance of components:
                          PC1   PC2   PC3    PC4    PC5      PC6
Standard deviation     1.802 1.093 0.872 0.6874 0.5710 1.94e-10
Proportion of Variance 0.541 0.199 0.127 0.0788 0.0543 0.00e+00
Cumulative Proportion  0.541 0.740 0.867 0.9457 1.0000 1.00e+00
```

Ahora bien, si calculamos las cargas de las dos primeras componentes, se observa que las variables pelvic_incidence, lumbar_lordosis_angle, degree_spondylolisthesis y sacral_slope pueden ser explicados en la primera componente, mientras que las variables pelvic_tilt.numeric y pelvic_radius pueden ser explicadas en la segunda componente, pues éstas tienen un mayor peso (en valor absoluto) respecto solo a las componentes mencionadas.

```
                            PC1         PC2         PC3         PC4       PC5          PC6
pelvic_incidence      -0.53514   0.0021937  -0.096069   0.1027990 -0.42346 -7.1729e-01
pelvic_tilt.numeric   -0.32358  -0.5275454  -0.648701   0.0064412 -0.15056  4.1649e-01
lumbar_lordosis_angle -0.45797  -0.0928751   0.152338   0.5480997  0.67677  1.4568e-11
sacral_slope          -0.44591   0.3961573   0.360313   0.1272009 -0.43150  5.5860e-01
pelvic_radius          0.14350  -0.7277556   0.585991   0.1742130 -0.27576  3.4466e-12
degree_spondylolisthesis -0.42398 -0.1627769  0.271184  -0.8015281  0.27885 -8.3036e-12
```

En el gráfico biplot, se aprecia que un buen grupo de observaciones que se contrapone o es explicado de manera inversa con las variables de pelvic_incidence, lumbar_lordosis_angle y degree_spondylolisthesis, en tanto, las variables sacral_slope y pelvic_radius se relacionan de manera inversa.



Regresión Logística:

Se hará una evaluación por modelo logístico. Asumiendo clases donde 0 = anormal y 1=normal.

```
Call:
  glm(formula = clase.grupo ~ pelvic_incidence + pelvic_tilt.numeric +
        lumbar_lordosis_angle + sacral_slope + pelvic_radius + degree_spondylolisthesis
      family = binomial, data = datos)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-2.2678  -0.3639  -0.0289   0.4081   2.7317

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.530e+01  3.315e+00  -4.615 3.93e-06 ***
  pelvic_incidence          2.517e+07  4.017e+07   0.627    0.531
pelvic_tilt.numeric      -2.517e+07  4.017e+07  -0.627    0.531
lumbar_lordosis_angle     1.794e-02  2.290e-02   0.784    0.433
sacral_slope             -2.517e+07  4.017e+07  -0.627    0.531
pelvic_radius             1.077e-01  2.318e-02   4.645 3.39e-06 ***
  degree_spondylolisthesis -1.693e-01  2.335e-02  -7.248 4.23e-13 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 389.86  on 309  degrees of freedom
Residual deviance: 177.87  on 303  degrees of freedom
AIC: 191.87

Number of Fisher Scoring iterations: 8
```

Figura 38: Estimación del modelo

```
       ytrue
ypred   0   1
0 196  28
1  14  72
> testerr <- mean(ypred!=ytrue)
> testerr
[1] 0.1354839
> (S <- mc[2,2]/sum(mc[,2]))
[1] 0.72
> (E <- mc[1,1]/sum(mc[,1]))
[1] 0.9333333
```

Figura 39: Matriz de confusión (error de clasificación)

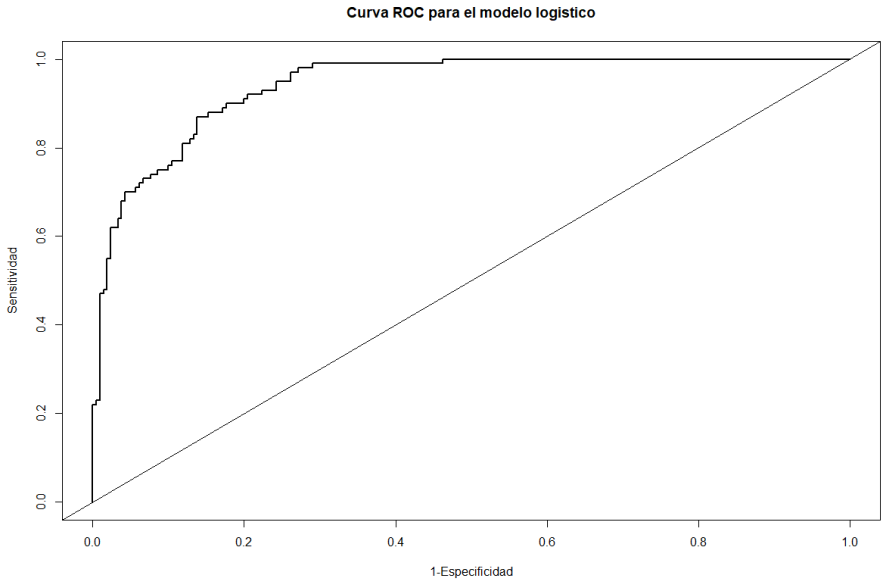


Figura 40: CURVA ROC

	incid.pelvica	incl.pelvica	angulo.lumbar	pend.sacral	rad.pelvico	grado.spond	clase
1	63.02782	22.552586	39.60912	40.47523	98.67292	-0.254400	Abnormal
2	39.05695	10.060991	25.01538	28.99596	114.40543	4.564259	Abnormal
3	68.83202	22.218482	50.09219	46.61354	105.98514	-3.530317	Abnormal
4	69.29701	24.652878	44.31124	44.64413	101.86850	11.211523	Abnormal
5	49.71286	9.652075	28.31741	40.06078	108.16872	7.918501	Abnormal
6	40.25020	13.921907	25.12495	26.32829	130.32787	2.230652	Abnormal

	subclase	clase.grupo	subclase.grupo	clas.grupo	ypred	yprob
1	Hernia	0	1	0	1	0.16443417
2	Hernia	0	1	0	1	0.22865463
3	Hernia	0	1	0	1	0.62861031
4	Hernia	0	1	0	1	0.04161865
5	Hernia	0	1	0	1	0.22036960
6	Hernia	0	1	0	1	0.52556338

Figura 41: Base de datos con categorías con las probabilidades y categorías predichas

Anexo 1

Listing 9: Código R. Pregunta 1

```r
rm(list=ls(all=TRUE))

### Instalar y cargar paquetes

library(MASS)
library(rpart)
library(caret)
library(klaR)
library(pROC)

### Cargar base de datos: wisc_bc_data.csv

# datos = read.csv(https://raw.githubusercontent.com/joshishwetha/dsx-spark/master/data.csv)[-1]
# datos = read.csv("http://www3.nd.edu/~steve/computing_with_data/Data/wisc_bc_data.csv")[-1]
datos = read.csv("wisc_bc_data.csv")[-1]
datos$diagnosis = ifelse(datos$diagnosis=='M',1,0)
datos$diagnosis = as.factor(datos$diagnosis)


### Base de datos de entrenamiento y de test
N = nrow(datos)
p = 0.8
n = round(N*p,0)

set.seed(1234)
indic = sample(N,n,replace=FALSE)
datos.train = datos[indic,]
datos.test = datos[-indic,]


#--------------------#
# MODELO LOGISTICO #
#--------------------#

## Seleccion de Variables
model.glm = glm(diagnosis ~ ., family=binomial, data=datos.train)
model.step = stepAIC(model.glm)
model.step
# X1 = datos.test[c(2,6,7,9,11,12,14,17,19,23,24,25,26,28)]   #Variables elegidas segun stepAIC
X1 = datos.test[attr(model.step$terms,"term.labels")]
head(X1)

## Matriz de confusion
pi.glm = predict(model.step, type="response", newdata=X1)
cut = 0.5
y.glm = as.numeric(pi.glm >= cut)
y.true = datos.test$diagnosis
mc.glm = confusionMatrix(y.true, y.glm, positive="1")
roc.glm = roc(response=y.true, predictor=y.glm)
area.glm = roc.glm$auc


```

```
#--------------------------------------#
# ANALISIS DISCRIMINANTE LINEAL (LDA) #
#--------------------------------------#

## Seleccion de Variables
wilks.lda = greedy.wilks(diagnosis ~ ., data=datos.train)
model.lda = lda(wilks.lda$formula, data=datos.train)
model.lda
X2 = datos.test[attr(model.lda$means,"dimnames")[[2]]]    #Variables elegidas segun
    greedy.wilks
head(X2)

## Matriz de confusion
y.lda = predict(model.lda, newdata=X2)$class
y.true = datos.test$diagnosis
mc.lda = confusionMatrix(y.true, y.lda, positive="1")
roc.lda = roc(response=y.true, predictor=as.numeric(y.lda))
area.lda = roc.lda$auc


#-----------------------------------------#
# ANALISIS DISCRIMINANTE CUADRATICO (QDA) #
#-----------------------------------------#

## Seleccion de Variables
wilks.qda = greedy.wilks(diagnosis ~ ., data=datos.train)
model.qda = qda(wilks.qda$formula, data=datos.train)
model.qda
X3 = datos.test[attr(model.qda$means,"dimnames")[[2]]]    #Variables elegidas segun
    greedy.wilks
head(X3)

## Matriz de confusion
y.qda = predict(model.qda, newdata=X3)$class
y.true = datos.test$diagnosis
mc.qda = confusionMatrix(y.true, y.qda, positive="1")
roc.qda = roc(response=y.true, predictor=as.numeric(y.qda))
area.qda = roc.qda$auc


#-------------------------------------------#
# ANALISIS DISCRIMINANTE REGULARIZADO (RDA) #
#-------------------------------------------#

## Seleccion de Variables
wilks.rda = greedy.wilks(diagnosis ~ ., data=datos.train)
model.rda = rda(wilks.rda$formula, data=datos.train)
model.rda
X4 = datos.test[attr(model.rda$means,"dimnames")[[1]]]    #Variables elegidas segun
    greedy.wilks
head(X4)

## Matriz de confusion
y.rda = predict(model.rda, newdata=X4)$class
y.true = datos.test$diagnosis
mc.rda = confusionMatrix(y.true, y.rda, positive="1")
roc.rda = roc(response=y.true, predictor=as.numeric(y.rda))
area.rda = roc.rda$auc
```

```r

#————————#
# RESUMEN #
#————————#

GLM = c(mc.glm$overall[1:2], mc.glm$byClass[1:2], AUC=area.glm)
LDA = c(mc.lda$overall[1:2], mc.lda$byClass[1:2], AUC=area.lda)
QDA = c(mc.qda$overall[1:2], mc.qda$byClass[1:2], AUC=area.qda)
RDA = c(mc.rda$overall[1:2], mc.rda$byClass[1:2], AUC=area.rda)
tabla = round(cbind(GLM, LDA, QDA, RDA),3)
tabla


#————————————————————#
# Grafica de Curvas ROC #
#————————————————————#

par(mfrow=c(2,2))
plot(1—roc.glm$specificities, roc.glm$sensitivities, type="l", lwd=2, col="magenta",
     ylab="Sensitividad", xlab="1—Especificidad",main="Curva ROC: Modelo Logistico")
     abline(a=0, b=1, lwd=2)
plot(1—roc.lda$specificities, roc.lda$sensitivities, type="l", lwd=2, col="cyan",
ylab="Sensitividad", xlab="1—Especificidad",main="Curva ROC: A. D. Lineal")
abline(a=0, b=1, lwd=2)
plot(1—roc.qda$specificities, roc.qda$sensitivities, type="l", lwd=2, col="red",
ylab="Sensitividad", xlab="1—Especificidad",main="Curva ROC: A. D. Cuadratico")
abline(a=0, b=1, lwd=2)
plot(1—roc.rda$specificities, roc.rda$sensitivities, type="l", lwd=2, col="blue",
ylab="Sensitividad", xlab="1—Especificidad",main="Curva ROC: A. D. Regularizado")
abline(a=0, b=1, lwd=2)
```

Anexo 2

Listing 10: Código R. Pregunta 2

```r
## Pregunta 2

rm(list=ls())

library(MASS)
library(smacof)
library(cluster)

letras=read.csv(file.choose(),sep=";")
filas=letras[,1]
letras=letras[,-1]
rownames(letras)=filas
letras=as.dist(as.matrix(letras))
letras=21-letras


# Escalamiento no metrico


res=smacofSym(letras,ndim=2, type = "ordinal")
res
summary(res)

plot(res,plot.type="confplot")
plot(res,plot.type="Shepard")
plot(res,plot.type="resplot")
(cor(res$confdist, res$dhat))^2
plot(res,plot.type="stressplot")
plot(res,plot.type="bubbleplot")


## Cluster jerarquico

a=hclust(dist(letras),method="average")
plot(a)
```

Anexo 3

Listing 11: Codigo R. Pregunta 3.

```r
##### ------------------------------------------------------------------------
##### TRABAJO GRUPAL — PREG. 3 — ANALISIS DE DATOS
#####
##### BASE DE DATOS: CARACTERISTICAS BIOMECANICAS DE PACIENTES ORTOPEDICOS
##### ------------------------------------------------------------------------

library(cluster)
library(fpc)
library(NbClust)
library(clValid)
library(dendextend)
library(circlize)
library(sparcl)

## LECTURA DE DATOS
datos <- read.csv("C:.../ortopedico.csv")
head(datos)
sapply(datos,class)

attach(datos)

## ADICION DE GRUPOS ORIGINALES
dim.class=c(summary(class))
datos$clase.grupo=c(rep(1,dim.class[1]),rep(2,dim.class[2]))

dim.sub=c(summary(sub_class))
datos$subclase.grupo=c(rep(1,dim.sub[1]),rep(3,dim.sub[3]),rep(2,dim.sub[2]))

colnames(datos) = c("incid.pelvica","incl.pelvica","angulo.lumbar","pend.sacral",
"rad.pelvico","grado.spond","clase","subclase",
colnames(datos[,9:10]))

head(datos)

## ANALISIS EXPLOTARIO
dim(datos)
summary(datos[,1:7])

par(mfrow=c(2,3))
for (i in 1:6) {
boxplot(datos[,i]~datos$clase.grupo, main = names(datos[i]), type="l")
}

par(mfrow=c(1,1))
plot(datos$clase.grupo,ylab= "Grupos originales",xlab = "observaciones")

## DATOS + GRUPO CLASS
datos1 <- datos[,1:6]

par(mfrow=c(1,1))
fviz_cluster(list(data = datos1,cluster=datos[,9]),
ellipse.type="norm",ellipse.level=0.9,
geom = "point",
main = "Agrupamiento original al 90%")

```

```
##### ------------------------------------------------------------
##### PRIMERA METODOLOGIA: CLUSTERING
##### ------------------------------------------------------------

###-------------- PASO 1. DEFINICION DE FUNCIONES (ALTERNATIVO)

# 1: kmeans, 2: PAM, 3: clara, 4: fanny
metodo <- function(datos1,h,nm){
set.seed(2010)
if(nm == 2){ pam(scale(datos1),h)
}else{if(nm == 3){ clara(scale(datos1),h)
}else{if(nm == 4){ fanny(scale(datos1),h,maxit=5000)
}else{ kmeans(scale(datos1),h,nstart = 100)}}}
}

# SUMA DE CUADRADOS DENTRO DE CADA CLUSTER
SC.cluster <- function(datos1, nm){
asw<-numeric()
if(nm ==1){for(h in 2:10){asw[h-1]=metodo(datos1,h,nm)$tot.withinss}
plot(2:10,asw,type="b")
}else{for(h in 2:10){asw[h-1]=metodo(datos1,h,nm)$silinfo$avg.width}
plot(2:10,asw,type="b",xlab="k",ylab="ASW")
}
}

# SILUETA
silueta <- function(datos1, nm){
diss.datos=daisy(scale(datos1))
par(mfrow=c(1,3))
if(nm ==1){for(h in 2:4){
plot(silhouette(metodo(datos1,h,nm)$cluster,diss.datos))}
}else{for(h in 2:4){plot(metodo(datos1,h,nm),which.plots=2)}}
}

# CRITERIO DE CALINSKI - HARABASZ
Cal.Har <- function(datos1, nm){
ch<-numeric()
par(mfrow=c(1,1))
if(nm ==1){for(h in 2:10){
ch[h-1] = calinhara(scale(datos1),metodo(datos1,h,nm)$cluster)}
}else{for(h in 2:10){
ch[h-1] = calinhara(scale(datos1),metodo(datos1,h,nm)$clustering)}}
plot(2:10,ch,type="b",xlab="k", ylab="Criterio de Calinski-Harabasz")
}

# GRAFICAS CLUSTER
plot.cluster <- function(datos1,clus){
par(mfrow=c(1,1))
fviz_cluster(list(data = datos1,cluster=clus),
ellipse.type="norm",ellipse.level=0.9,
geom = "point",
main = "Grafico de Conglomerados al 90%")
}

###-------------- PASO 2. DETERMINANDO NUMERO DE CONGLOMERADOS
res.cluster=list()

## KMEANS
```

```
SC.cluster(datos1,1)

silueta(datos1,1)
kmeansruns(scale(datos1),criterion="asw")

Cal.Har(datos1,1)
kmeansruns(scale(datos1),criterion="ch")

set.seed(2000)
res.cluster[[1]]=kmeans(scale(datos1),2)$cluster

par(mfrow=c(1,1))
plot.cluster(datos1,res.cluster[[1]])

## PAM
SC.cluster(datos1,2)

silueta(datos1,2)
pamk(scale(datos1),criterion="asw")

Cal.Har(datos1,2)
pamk(scale(datos1),criterion="ch")

set.seed(2000)
res.cluster[[2]]=pam(scale(datos1),2)$clustering

par(mfrow=c(1,1))
plot.cluster(datos1,res.cluster[[2]])

## CLARA
SC.cluster(datos1,3)

silueta(datos1,3)
pamk(scale(datos1),criterion="asw",usepam=FALSE)

Cal.Har(datos1,3)
pamk(scale(datos1),criterion="ch",usepam=FALSE)

set.seed(2000)
res.cluster[[3]]=clara(scale(datos1),2)$clustering
for (i in 1:dim(datos)[1]) {
if(res.cluster[[3]][i] == 1){
res.cluster[[3]][i]=2
}else{res.cluster[[3]][i] = 1
}
}

par(mfrow=c(1,1))
plot.cluster(datos1,res.cluster[[3]])

## FANNY
SC.cluster(datos1,4)
silueta(datos1,4)
Cal.Har(datos1,4)

set.seed(2000)
res.cluster[[4]]=fanny(scale(datos1),2)$clustering
for (i in 1:dim(datos)[1]) {
if(res.cluster[[4]][i] == 1){
```

```
res.cluster[[4]][i]=2
}else{res.cluster[[4]][i] = 1
}
}

par(mfrow=c(1,1))
plot.cluster(datos1,res.cluster[[4]])

#table(datos[,9],res$cluster)

par(mfrow=c(2,2))
plot(res.cluster[[1]],xlab = "observaciones",ylab = "kmeans")
plot(res.cluster[[2]],xlab = "observaciones",ylab = "PAM")
plot(res.cluster[[3]],xlab = "observaciones",ylab = "clara")
plot(res.cluster[[4]],xlab = "observaciones",ylab = "fanny")

## CLUSTER JERARQUICO: HIERARCHICAL
set.seed(2000)
hc <- hclust(dist(scale(datos1)),method="ward.D")
res.cluster[[5]] = cutree(hc , 2)

dend <- as.dendrogram(hc) %>%
color_branches(k=2) %>%
color_labels

par(mar = rep(0,4))
circlize_dendrogram(dend, labels_track_height = 0.3,
dend_track_height = .6,
main = "Dendongrama")

par(mfrow=c(1,1))
ColorDendrogram(hc, y = res.cluster[[5]],
labels = names(res.cluster[[5]]),
main = "Dendograma",
xlab= "observaciones",
branchlength = 0.4)

par(mfrow=c(1,1))
plot.cluster(datos1,res.cluster[[5]])

## CLUSTER JERARQUICO: AGNES
agnes.single=agnes(scale(datos1),method="single")
agnes.single$ac
par(mfrow=c(1,1))
pltree(agnes.single,cex=1,hang=-1, main = "Dendograma Agnes")

agnes.ward=agnes(scale(datos1),method="ward")
agnes.ward$ac
par(mfrow=c(1,1))
pltree(agnes.ward,cex=1,hang=-1, main = "Dendograma Agnes")

diss.datos=daisy(scale(datos1))

agnes.ward=as.hclust(agnes.ward)
par(mfrow=c(1,3))
for(h in 2:4){
res=cutree(agnes.ward,k=h)
plot(silhouette(res,diss.datos))
}
```

```
232
233 set.seed(2000)
234 res.cluster[[6]]=cutree(agnes.ward,k=2)
235
236 par(mfrow=c(1,1))
237 plot.cluster(datos1,res.cluster[[6]])
238
239 ## DIANA
240 diana.metodo=diana(scale(datos1))
241
242 par(mfrow=c(1,1))
243 pltree(diana.metodo,cex=1,hang=-1, main = "Dendograma Diana")
244
245 diana.metodo=as.hclust(diana.metodo)
246 par(mfrow=c(1,3))
247 for(h in 2:4){
248 res=cutree(diana.metodo,k=h+1)
249 plot(silhouette(res,diss.datos))
250 }
251
252 set.seed(2000)
253 res.cluster[[7]]=cutree(diana.metodo,k=3)
254 for (i in 1:dim(datos)[1]) {
255 if(res.cluster[[7]][i] == 1){
256 res.cluster[[7]][i]=2
257 }else{res.cluster[[7]][i] = 1
258 }
259 }
260
261 par(mfrow=c(1,1))
262 plot.cluster(datos1,res.cluster[[7]])
263
264 par(mfrow=c(1,2))
265 plot(res.cluster[[5]],xlab = "observaciones",ylab = "Hierarchical")
266 plot(res.cluster[[6]],xlab = "observaciones",ylab = "Agnes: Wald")
267
268 par(mfrow=c(1,1))
269 plot(res.cluster[[7]],xlab = "observaciones",ylab = "Diana")
270
271 ###---------- PASO 3. PERFILADO Y CARACTERIZACION DE CLUSTERS
272
273 ## ADICION DE CLUSTER A LA BASE
274 datos.new <- datos[,c(1:6,9)]
275 for (i in 1:7) {
276 datos.new <- cbind(datos.new,res.cluster[[i]])
277 }
278
279 colnames(datos.new)<-c(colnames(datos.new[,1:7]),"cluster.km","cluster.PAM",
280 "cluster.clara","cluster.fanny","cluster.h",
281 "cluster.agnes","cluster.diana")
282 head(datos.new)
283
284 # TABLA DE MEDIAS
285 med=list()
286 for (i in 1:7) {
287 med[[i]] = aggregate(x = datos.new[,1:6],
288 by = list(datos.new[,i+7]),
289 FUN = mean)
290 }
```

```
med[[1]]

# ANALISIS EXPLORATIVO POR CLUSTER
for (j in 1:7) {
par(mfrow=c(2,3))
for (i in 1:6) {
boxplot(datos.new[,i]~datos.new[,7+j], main=names(datos.new[i]), type="l")
}
}

## OTRAS TECNICAS
clmethods <- c("hierarchical","kmeans","pam","agnes","diana")

# Medidas de validacion interna
intern <- clValid(scale(datos1), nClust = 2:10,
clMethods = clmethods, validation = "internal")
summary(intern)

par(mfrow=c(1,1))
plot(intern)

# Medidas de estabilidad
stab <- clValid(scale(datos1), nClust = 2:10, clMethods = clmethods,
validation = "stability")
summary(stab)

# Mostrar solo scores optimos
optimalScores(intern)
optimalScores(stab)


#### SEGUNDO METODO PCA

## Cargar base de datos: ortopedico.csv
datos = read.csv("C:../ortopedico.csv")[,-7]
head(datos)

## Grafico de Sedimentacion
library(psych)
par(mfrow=c(1,2))
scree(datos)
pc = prcomp(x=datos, scale=TRUE, center=TRUE)
plot(pc)
par(mfrow=c(1,1))

## Proporcion de varianza explicada
summary(pc)

## Carga de las componentes
pc$rotation

## Grafico biplot
biplot(pc,scale=0)
```