Análisis del conjunto de datos de cáncer de mama de Wisconsin

Maitena Tellaetxe Abete

8 de enero de 2016

1. Índice

2. Introducción

El cáncer de mama es el tumor más frecuente en las mujeres occidentales. En España, en concreto, se diagnostican alrededor de 25.000 nuevos casos al año y se prevé que 1 de cada 8 mujeres lo padecerá a lo largo de su vida. Este tipo de cáncer se puede presentar también en hombres, aunque es mucho menos habitual (1).

El cáncer de pecho es un tumor maligno que se origina de las células del seno. Pese a que se conocen factores de riesgo como la edad, factores genéticos, antecedentes familiares o la obesidad, todavía se desconoce el mecanismo exacto mediante el cual surge y se desarrolla la enfermedad. En cuanto a su diagnóstico, la mamografía y la punción por aspiración con aguja fina (PAAF) son las técnicas más habituales para detectar estos tumores. Desafortunadamente, el diagnóstico realizado tras analizar una mamografía varía considerablemente entre especialistas y la tasa media de identificaciones correctas del PAAF no supera el 90 %. Por tanto, es necesario desarrollar técnicas capaces de realizar un diagnóstico más preciso de cáncer de mama (2).

En este estudio analizaremos el **set de datos diagnóstico de cáncer de mama de Wisconsin**, comúnmente utilizado en la comunidad *machine learning* y disponible en numerosos portales web (3). Dicho set contiene observaciones realizadas a 699 pacientes de los hospitales Universidad de Wisconsin-Madison, en las que se detallan 9 atributos de un test citológico realizado a muestras de tumores obtenidas por PAAF, tal y como muestra el Cuadro 1. Cada uno de estos atributos toma valores entre 1 y 10 que denotan niveles de anomalía, de manera que el 1 se corresponde con un estado normal y el 10 con el más anómalo.

Concretamente, las células de enfermedades mamarias benignas tienen tamaños y membranas nucleares uniformes, nucleolos sin particularidades, contienen cromatina blanda y compacta y tienden a agruparse en monocapas. Además, dichas células suelen visualizarse habitualmente como un núcleo solitario y desnudo, sin citoplasma. Las células malignas, en cambio, suelen ser irregulares en cuanto a tamaño, forma y cromatina, presentan nucleolos más prominentes y en mayor abundancia de lo habitual y se agrupan en multicapas. En ellas destaca un desequilibrio del sistema de autorregulación que

Cuadro 1: Características de la observación citológica de cada tumor

X_1	Grosor de la masa
X_2	Uniformidad de tamaño celular
X_3	Uniformidad de forma celular
X_4	Adhesión marginal
X_5	Tamaño de célula epitelial
X_6	Núcleos desnudos
X_7	Cromatina blanda
X_8	Nucleolos normales
X_9	Actividad mitótica

controla y limita la división celular, por lo que estas células comienzan a dividirse más rápidamente. En consecuencia, el número de mitosis que se dan en tejidos malignos es más elevado que el que se da en los benignos. Se caracterizan además por la pérdida de la capacidad de adhesión, de manera que las células se pueden desprender más fácilmente de los tejidos originales y, si entran en el torrente sanguíneo, pueden llegar a producir tumores en otros tejidos, lo que se conoce como metástasis. (4)

Para cada una de estas observaciones también se conoce si el tumor es maligno o benigno. El objetivo final será analizar si la observación de dichos atributos (o algún subgrupo de ellos) permite diagnosticar la malignidad de los tumores.

3. Análisis descriptivo

Cuando observamos por primera vez un conjunto de datos, es aconsejable realizar un estudio descriptivo para hacernos una idea de los datos con los que vamos a tratar.

```
# Cargamos las librerías necesarias
library(corrplot)
library(corrgram)
library(ggplot2)
library(reshape2)
library(ggbiplot)
```

```
## Loading required package: plyr
##
## Attaching package: 'plyr'
##
## The following object is masked from 'package:corrgram':
##
##     baseball
##
## Loading required package: scales
## Loading required package: grid
```

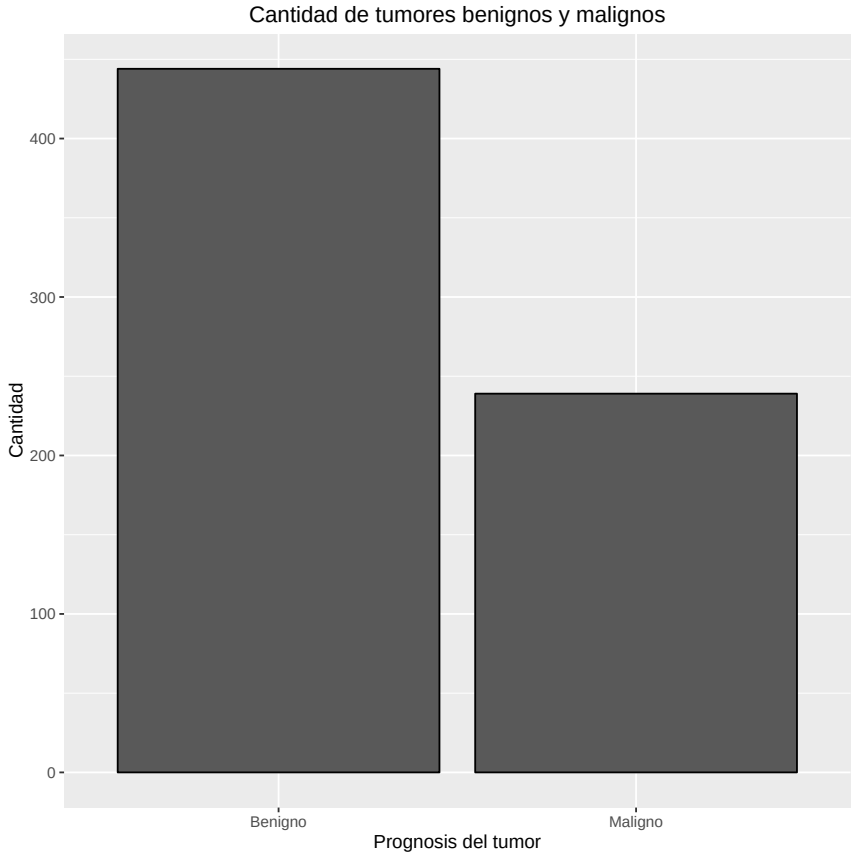
```
library(plyr)
library(caret)
```

```
## Loading required package: lattice
```

```
library(cluster)
library(MASS)
library(mvnormtest)
library(pROC)
```

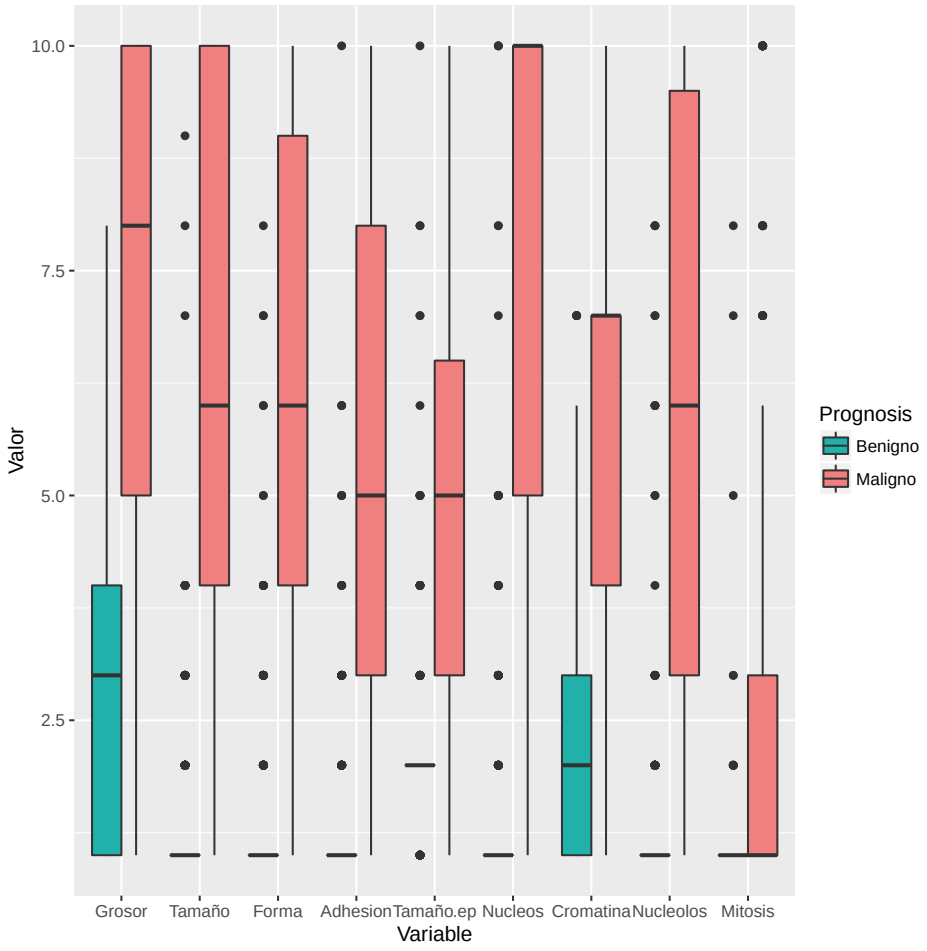
```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
# Cargamos los datos
setwd("~/KISA//EAD/Breast/")
data <- read.csv("breat.data", header = FALSE, na.strings = "?")
str(data)

## 'data.frame': 699 obs. of  11 variables:
##  $ V1 : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 1033078 ..
##  $ V2 : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ V3 : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ V4 : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ V5 : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ V6 : int  2 7 2 3 2 7 2 2 2 2 ...
##  $ V7 : int  1 10 2 4 1 10 10 1 1 1 ...
##  $ V8 : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ V9 : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ V10: int  1 1 1 1 1 1 1 1 5 1 ...
##  $ V11: int  2 2 2 2 2 4 2 2 2 2 ...

# Modificamos ligeramente las anotaciones
colnames(data) <- c("ID", "Grosor", "Tamaño", "Forma", "Adhesion", "Tamaño.ep",
                    "Nucleos", "Cromatina", "Nucleolos", "Mitosis", "Prognosis")
data <- na.omit(data)
data$Prognosis[data$Prognosis == 4] <- "Maligno"
data$Prognosis[data$Prognosis == 2] <- "Benigno"
data$Prognosis <- as.factor(data$Prognosis)

# Analizamos los datos
# Lo primero de todo, vamos a ver que cantidad de benignos y malignos tenemos
table(data$Prognosis)

##
## Benigno Maligno
##     444     239

ggplot(data = data, aes(x = Prognosis)) + geom_bar(color = "black") +
  xlab("Prognosis del tumor") + ylab("Cantidad") +
  ggtitle("Cantidad de tumores benignos y malignos")
```

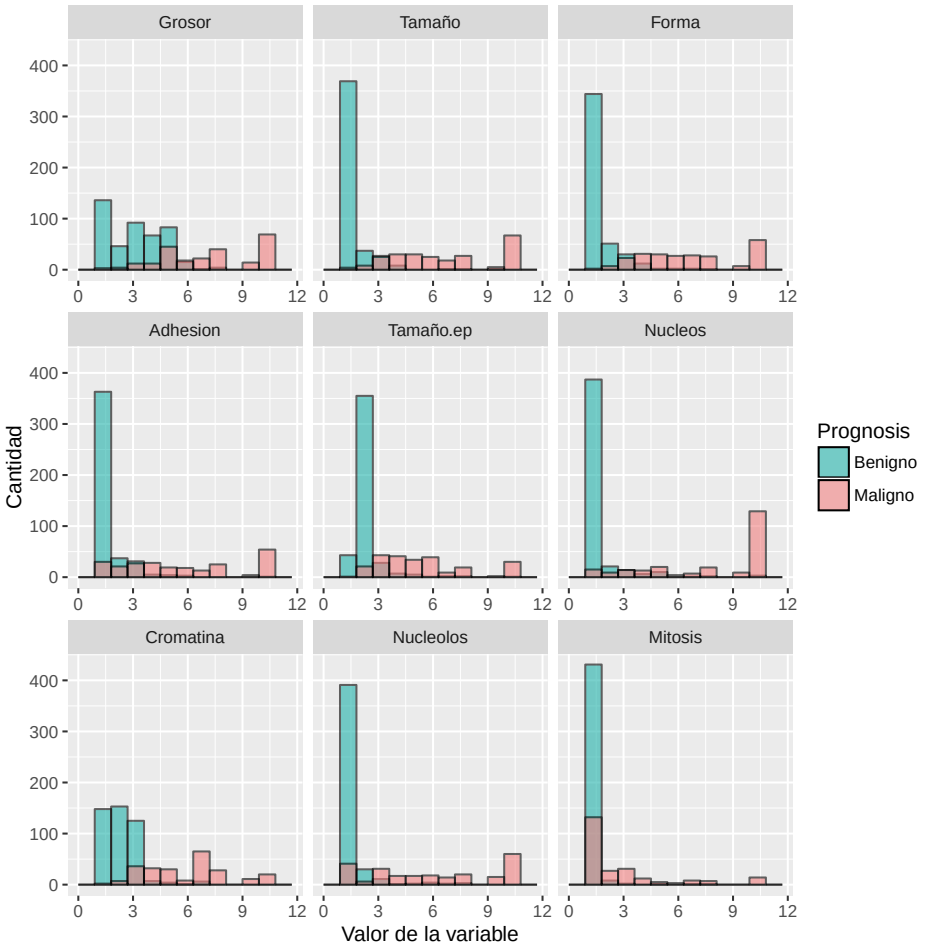


De las 699 observaciones iniciales, solo 683 están completas y será con las que procederemos. De ellas, **444** pertenecen a tumores benignos y **239** a malignos. Puede parecer lógico analizar en primer lugar si los atributos de las pruebas citológicas toman valores notablemente diferentes para los dos grupos.

```
data_melt <- melt(data[-1], id.var = "Prognosis")
ggplot(data = data_melt, aes(x = variable, y = value)) +
  geom_boxplot(aes(fill = Prognosis)) + xlab("Variable") + ylab("Valor") +
  scale_fill_manual(values = c("lightseagreen", "lightcoral"))
```

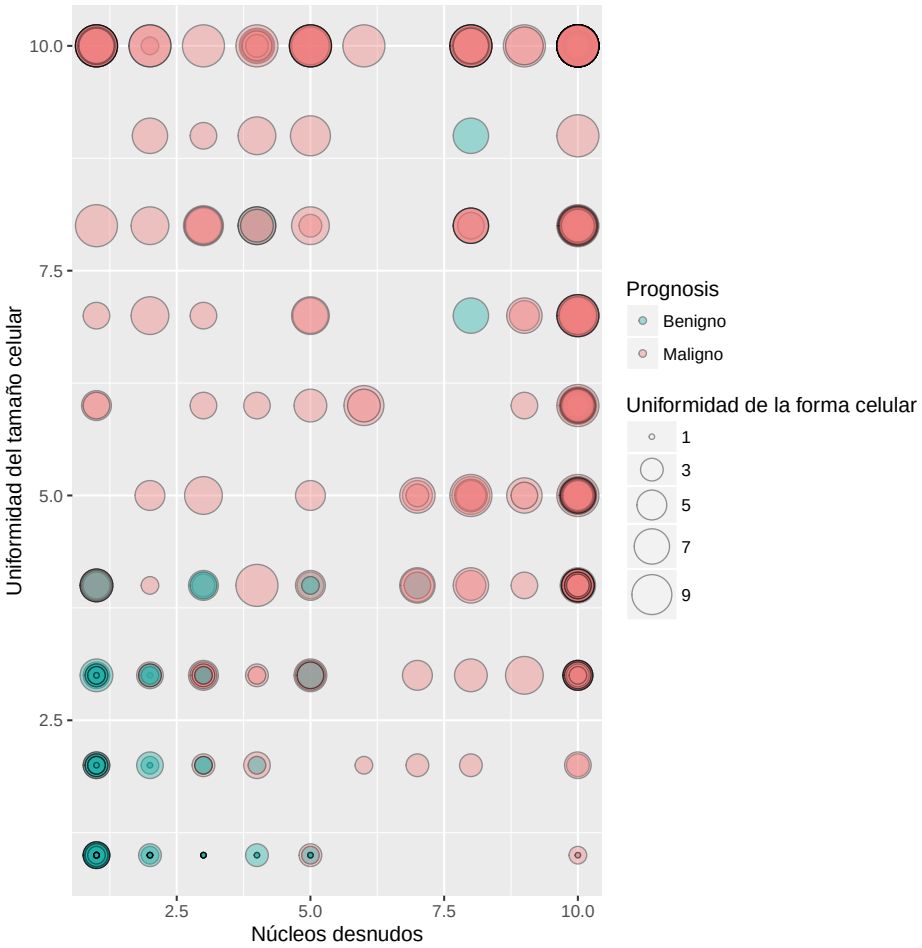



```
# Histograma para cada variable
ggplot(data_melt, aes(x = value, fill = Prognosis)) +
  facet_wrap(~variable, scales = "free_x") +
  geom_histogram(color = "black", alpha = 0.6, position = "identity", bins = 10) +
  ylab("Cantidad") + xlab("Valor de la variable") +
  scale_fill_manual(values = c("lightseagreen", "lightcoral"))
```



Como era de esperar por la definición de las características que estamos analizando, los tumores benignos toman valores visiblemente más bajos que los malignos para todas las variables, de manera que la mediana de casi todas ellas es 1 (solo en el caso del grosor y la cromatina blanda se acerca a 3 y 2 respectivamente) para los tumores benignos, mientras que la mediana para los malignos alcanza el 5 para todas las características excepto para la mitosis, llegando a ser incluso 10 en el caso de los nucleos desnudos. Concretamente, parece que las mayores diferencias se observan en los **nucleos desnudos**, la **uniformidad del tamaño** y la **uniformidad de la forma**. Vamos a ver, por tanto, si con dichos atributos podemos separar los dos grupos.

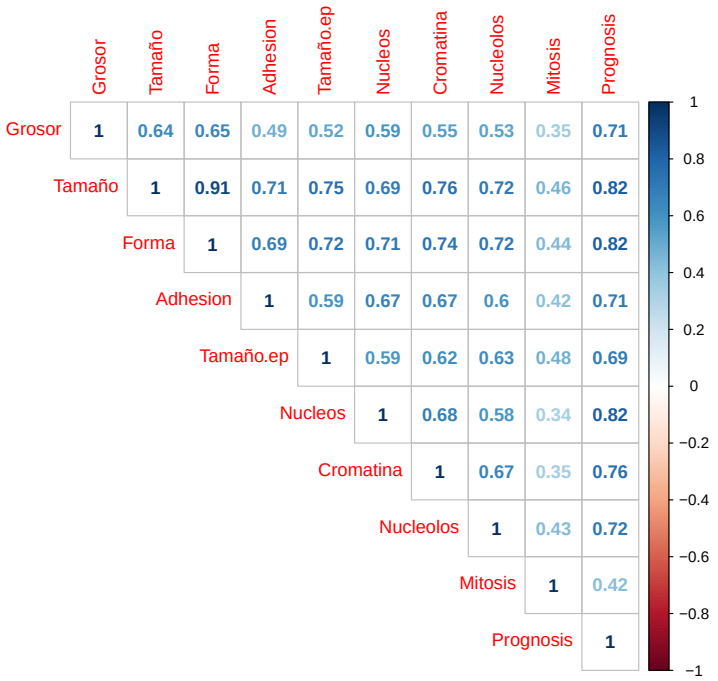
```
ggplot(data, aes(x = Nucleos, y = Tamaño, size = Forma, fill = Prognosis)) +
  geom_point(shape=21, colour='black', alpha = 0.4) +
  scale_size(range = c(min(data$Forma), max(data$Forma)),
             breaks = seq(min(data$Forma), max(data$Forma), by = 2),
             name = "Uniformidad de la forma celular") + xlab("Núcleos desnudos") +
  ylab("Uniformidad del tamaño celular") +
  scale_fill_manual(values = c("lightseagreen", "lightcoral"))
```



El gráfico anterior confirma que la malignidad de un tumor está asociada a valores altos de estos atributos. En general, muestras con valores inferiores a 5 en núcleos densnudos, a 4 en uniformidad del tamaño y a 3 en uniformidad de la forma celular, tienen una prognosis benigna.

Por último estudiaremos la matriz de correlación de las variables para averiguar si hay atributos linealmente relacionados y que, por tanto, nos pueden estar aportando información redundante.

```
corrgram_data <- data[,-1]
corrgram_data$Prognosis <- as.numeric(revalue(corrgram_data$Prognosis, c("Benigno" = 0,
                                                                         "Maligno" = 1)))
corrplot(cor(corrgram_data, use = "complete.obs", method = "pearson"),
         method = "number", type = "upper")
```



La correlación es positiva, como era de esperar, para todas las variables. Mientras que la relación de la variable mitosis con las demás variables es moderada, el resto están fuertemente relacionadas linealmente, siendo especialmente destacable la relacion entre la uniformidad de la forma y la uniformidad del tamaño celular. Además, la prognosis está también estrechamente relacionada con la uniformidad del tamaño celular, la uniformidad de la forma celular y los nucleos desnudos, como habíamos intuido en gráficos anteriores.

4. Reducción de dimensionalidad

Una vez conocidos los datos, una primera acción puede ser intentar caracterizar los tumores que tenemos de una manera más sencilla, identificando cuáles de los atributos contienen la mayor parte de la variabilidad de los datos y son, por tanto, relevantes, y cuáles nos están dando prácticamente la misma información que otros, pudiendo ser, por tanto, descartables. Por ejemplo, sería clínicamente interesante buscar un subconjunto de las 10 variables que facilite el diagnóstico de la prognosis de un tumor. Todas las técnicas que tienen este objetivo se engloban dentro de las metodologías de reducción de dimensionalidad.

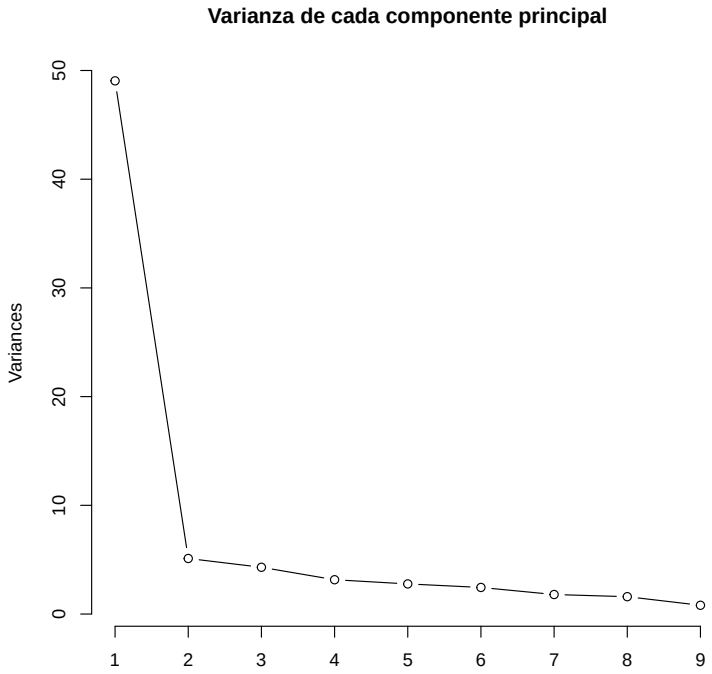
Concretamente, en este estudio se va a proceder con el método de análisis de componentes principales (**PCA**, por sus siglas en inglés), técnica exploratoria en la que se crean nuevas variables (no correladas linealmente) mediante combinación lineal de las originales (posiblemente correladas), de manera que todas las variables resultantes son ortogonales entre sí. Estos nuevos atributos, llamados **componentes principales**, cumplen la propiedad de que cada uno de ellos recoge una mayor proporción de la variabilidad original que el siguiente, es decir, la varianza de mayor tamaño del conjunto de datos es capturada en la primera componente, la segunda varianza más grande en la segunda componente, y así sucesivamente. Una de las técnicas más empleadas para obtener dichas componentes es la descomposición en valores propios de la matriz de covarianzas, cuando los datos están en la misma escala, o correlaciones, cuando no lo están.

La transformacion de los datos al nuevo sistema de coordenadas basado en las componentes principales permite además la visualizacion de clusters (grupos de muestras que tienen un patrón de atributos similar) existentes en los datos, ya que, en muchos casos, las direcciones que contienen la máxima variabilidad son las más relevantes para el clustering. En concreto, hay estudios que han mostrado que la PCA es una relajación del método de clustering K-medias y que las componentes principales son las soluciones continuas al problema de clustering (5).

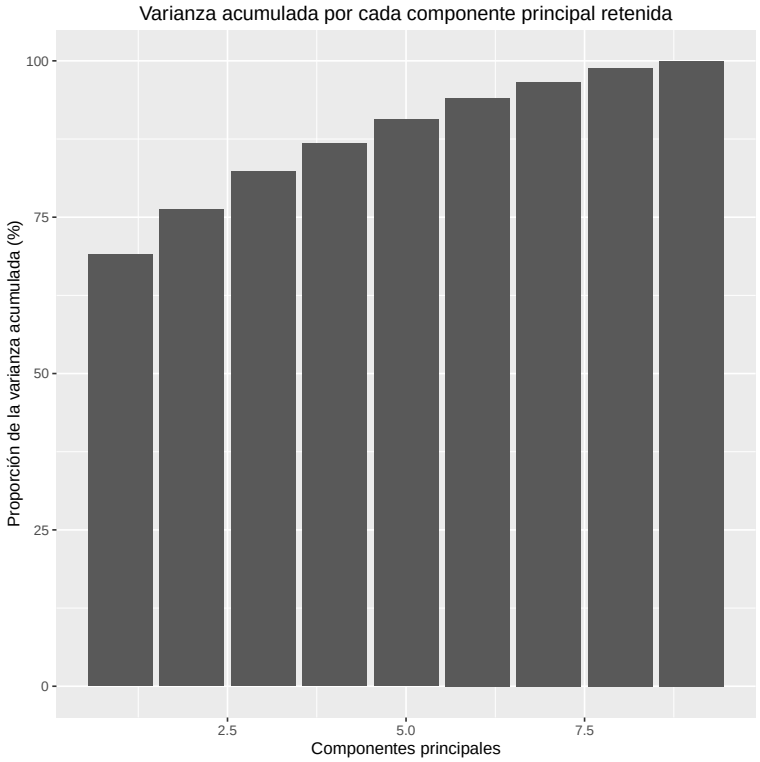
El análisis de PCA para el set que estamos analizando se muestra a continuación. Dado que los datos

están todos en la misma escala, partiremos de la matriz de covarianzas.

```
data0 <- data[,-c(1,11)]
pca_out <-prcomp(data0, scale. = FALSE)
summary(pca_out)

## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     7.0034 2.26069 2.07402 1.77629 1.66450 1.56404
## Proportion of Variance 0.6905 0.07195 0.06056 0.04442 0.03901 0.03444
## Cumulative Proportion  0.6905 0.76246 0.82302 0.86744 0.90644 0.94088
##                           PC7     PC8     PC9
## Standard deviation     1.34041 1.26322 0.89822
## Proportion of Variance 0.02529 0.02247 0.01136
## Cumulative Proportion  0.96618 0.98864 1.00000

plot(pca_out, type = "l", main = "Varianza de cada componente principal")
```




```
var_total <- sum(diag(cov(data0)))
cum_prop <- cumsum(pca_out$sdev^2/var_total)*100
ggplot(data.frame(cum_prop), aes(seq_along(cum_prop), cum_prop)) +
  geom_bar(stat="identity")+ xlab("Componentes principales") +
  ylab("Proporción de la varianza acumulada (%)") +
  ggtitle("Varianza acumulada por cada componente principal retenida")
```



La primera componente principal tiene una desviación estándar de 7 y explica el 70 % de la varianza en los datos; cada una de las siguientes componentes explica menos del 10 % de la variabilidad. Así, para explicar el 100 % de la variabilidad, seguimos necesitando 9 variables, solo una menos que empleando los datos originales. Dado que los tumores son enfermedades muy hetereogéneas, es esperable que unos pocos componentes no sean suficientes para explicar gran parte de la variabilidad, como suele verse en ejemplos más sencillos. Para determinar la cantidad de componentes a retener, empleamos el *scree test*, de acuerdo al cual se plotea la varianza para cada una de las componentes y se descartan aquellas componentes posteriores a la que dibuja un codo. En el caso de nuestros datos, nos quedamos, por tanto, con 2 componentes, que explican el 76 % de la varianza de los datos y que intentaremos interpretar a continuación.

Cada nueva componente viene dada por una combinación lineal de los atributos iniciales, con los siguientes pesos:

```
pca_out$rotation[,1:2]
```

```
##                  PC1          PC2
## Grosor    -0.2967358 -0.073506644
## Tamaño    -0.4039707  0.229928848
## Forma     -0.3927586  0.164700982
## Adhesion  -0.3312021 -0.098197542
## Tamaño.ep -0.2497398  0.200215050
## Nucleos   -0.4426135 -0.780569633
## Cromatina -0.2920783  0.008479735
## Nucleolos -0.3545360  0.469194517
## Mitosis   -0.1245763  0.188068892
```

Podemos observar que todos los pesos de la primera componente son negativos. Teniendo en cuenta que las variables iniciales estan codificadas del 1 al 10, tal que el 1 representa un estado normal y el 10 uno anómalo, se podría concluir que esta componente representa la normalidad y, por tanto, benignidad de los tumores. En lo que se refiere a la segunda componente, contribuyen de manera similar la uniformidad del tamaño, forma, tamaño epitelial y el número de mitosis, y de manera más significativa los nucleos desnudos y los nucleolos. Vamos a computar la correlación de las nuevas variables con las originales en caso de que esto pueda arrojar más luz sobre su interpretación.

```
cor(pca_out$x, data0)[1:2,]
```

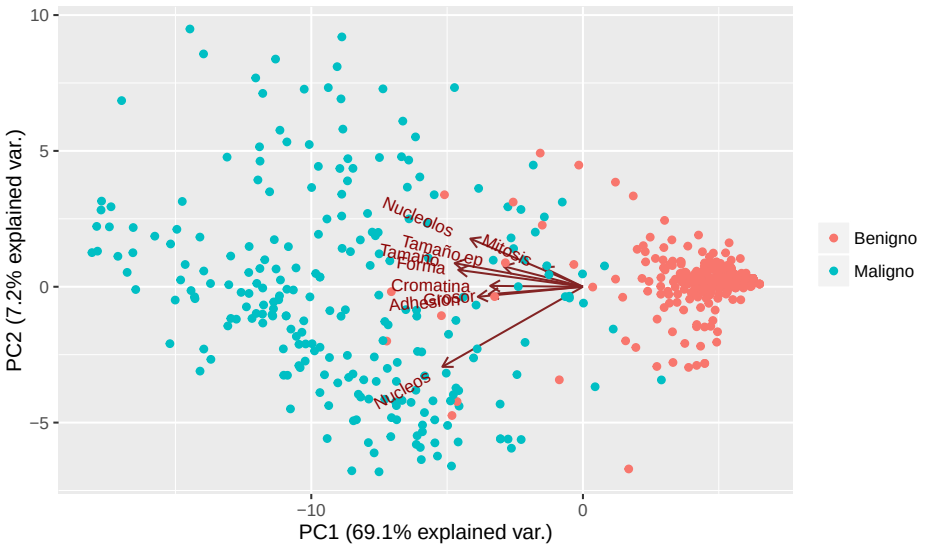
```
##          Grosor     Tamaño      Forma    Adhesion  Tamaño.ep    Nucleos
## PC1 -0.73673517 -0.9230106 -0.9203829 -0.80973466 -0.7867549 -0.8506896
## PC2 -0.05891166  0.1695835  0.1245869 -0.07749673  0.2036018 -0.4842742
##        Cromatina  Nucleolos    Mitosis
## PC1 -0.835016152 -0.8133712 -0.5035313
## PC2  0.007825481  0.3474678  0.2453811
```

La primera componente tiene una alta correlación negativa con todos los atributos, resultado esperable dado que todas las variables están muy correladas entre ellas, destacando en concreto la uniformidad del tamaño y de la forma. Esto confirma la idea de que la primera componente está relacionada con la prognosis, ya que según el correlograma obtenido en el apartado del análisis descriptivo, la prognosis de los tumores se relaciona de manera más significativa con la uniformidad del tamaño, la forma y los núcleos desnudos. Podemos también concluir que todos los atributos parecen ir en bloque, de manera que aquellas muestras que tienen un valor alto para el tamaño, también lo tienen en el resto de atributos y viceversa. Esto concuerda con el hecho de que los tumores malignos/benignos tienden a obtener valores elevados/bajos en todos los atributos.

Por lo que vemos en la segunda componente principal, otra parte de la variabilidad, pero mucho menos significativa que la anterior, está relacionada con los núcleos desnudos (de manera inversa), los nucleolos y el número de mitosis (estos dos últimos de manera directa). Todos estos atributos están asociados al núcleo celular, lo que sugiere que pueden existir tipos de tumores con comportamientos nucleares dispares.

Vamos por último a representar los datos en el nuevo sistema de coordenadas:

```
ggbiplot(pca_out, obs.scale = 1, var.scale = 1, groups = data$Prognosis)  +
  theme(legend.title = element_blank())
```



En efecto, parece que la primera componente está relacionada con la prognosis, de manera que tumores que puntúan alto en ella están asociados a benignidad. Si fijamos el 0 como valor umbral para separar tumores benignos y malignos, tenemos que:

```
predicted <- rep("Maligno", nrow(data))
predicted[pca_out$x[,1] > 0] <- "Benigno"
confusionMatrix(predicted, data$Prognosis)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction Benigno Maligno
##    Benigno     430       5
##    Maligno      14     234
##
##                Accuracy : 0.9722
##                  95% CI : (0.9569, 0.9832)
##     No Information Rate : 0.6501
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9394
##  Mcnemar's Test P-Value : 0.06646
##
##             Sensitivity : 0.9685
##             Specificity : 0.9791
##          Pos Pred Value : 0.9885
##          Neg Pred Value : 0.9435
##              Prevalence : 0.6501
##          Detection Rate : 0.6296
##    Detection Prevalence : 0.6369
##       Balanced Accuracy : 0.9738
##
##        'Positive' Class : Benigno
##
```

Los valores de sensibilidad de 0.97 y especificidad de 0.98, junto con el resto de estadísticos, sugieren que estamos en lo correcto al pensar que la primera componente está muy relacionada con la prognosis de los tumores.

En cuanto a la segunda componente, los tumores benignos en general se agrupan alrededor del 0. Entre los malignos podemos observar que aquellos que tienen valores más altos en la primera componente, que hemos identificado como benignidad, y están más cerca de los tumores benignos, tienen valores negativos, mientras que aquellos con valores más pequeños en la primera componente tienden a puntuar más alto en la segunda. Esta componente está relacionada con elementos nucleares, en concreto, negativamente con los núcleos desnudos y positivamente con los nucleolos normales y la actividad mitótica. Esto sugiere que quizás existan tumores malignos con distintas características en cuanto al núcleo, que a su vez podrían estar relacionadas con factores como la metástasis o la resistencia a medicamentos, por lo que resultaría interesante poder estudiar esto con mayor profundidad.

5. Análisis de clustering

En vista de la estructura de grupos que hemos observado en el apartado anterior, vamos a realizar un análisis de clustering para confirmar los resultados obtenidos. En general, las técnicas de clustering se clasifican en dos grandes grupos: por una parte tenemos el **clustering particional**, que busca agrupar las muestras en grupos o clusters no vacíos, de tal manera que cada elemento pertenece a un único grupo y, por otra, el **clustering jerárquico**, en el que en vez de una única partición, se construye una jerarquía de clusters, como si de un árbol se tratase. En ambos casos los objetos que pertenecen a un grupo son muy homogeneos entre sí, y la heterogeneidad entre los grupos muy elevada.

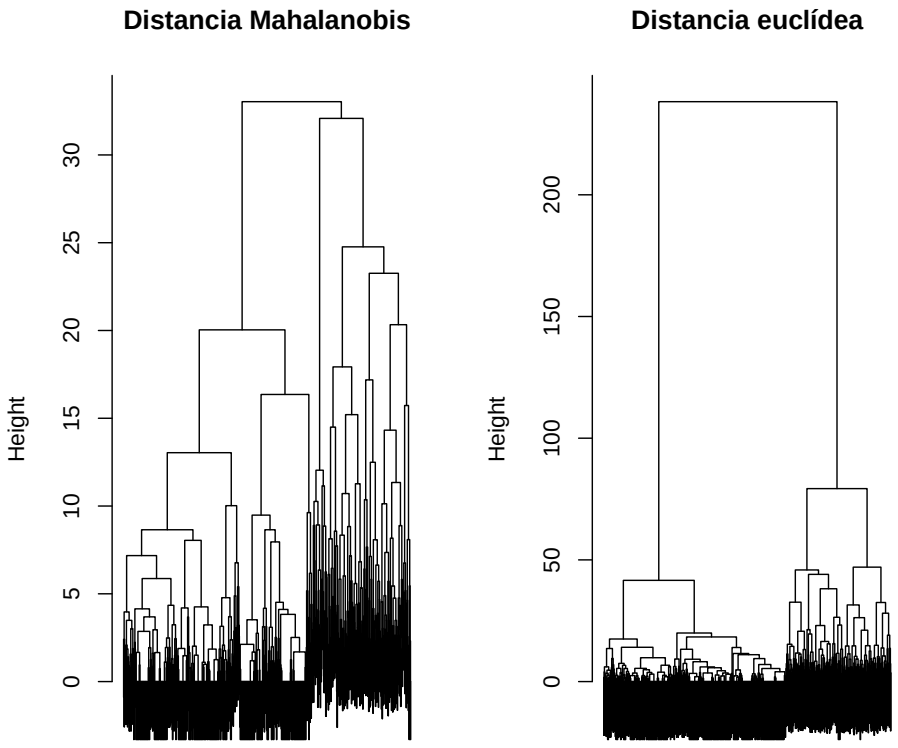
En la práctica, una diferencia significativa entre ambos es que en el clustering particional es necesario predefinir el número de clusters que se desean obtener, mientras que para el jerárquico no. En consecuencia, el clustering particional se recomienda cuando se conoce el número de clusters que deben verse representados. De acuerdo a los resultados que hemos obtenido en el análisis de PCA, esperamos encontrar al menos dos clusters que se corresponden con la prognosis de los tumores, aunque también existe la sospecha de que quizás puede haber dos subtipos de tumores malignos. Por tanto, comenzaremos realizando un clustering jerárquico que no nos limite en este aspecto y tras observar la salida, elegiremos cómo proceder con el particional.

5.1. Clustering jerárquico

Realizaremos un clustering jerárquico **aglomerativo**, de manera que al principio se tendrá un cluster para cada observación, y estos se irán uniendo de dos en dos, creando así nuevos clusters más poblados. En contraposición tenemos el clustering jerárquico divisivo, en el que en cada paso los nuevos clusters se crean diviendo clusters existentes.

El punto de partida de esta técnica es una matriz de distancias, por lo que un primer parámetro a definir es el tipo de distancia a emplear. La distancia de Mahalanobis es la única medida que tiene en cuenta la posible correlación entre las variables explicativas; dado que este es nuestro caso, será la distancia con la que procederemos. Otro factor a tener en cuenta es el criterio que se empleará para fusionar los clusters. Existen varias opciones pero en nuestro caso nos decantamos por el método de varianza mínima de Ward, donde el criterio para la elección de los clusters a fusionar en cada paso se basa en la minimización de la varianza dentro de los grupos. El método de Ward está diseñado para trabajar con una matriz de distancias euclídeas; al ser la distancia de Mahalanobis una generalización multidimensional de la euclídea, también se considera correcto su uso en este caso. Compararemos sin embargo los resultados con ambas distancias para ver posibles diferencias. El último detalle a destacar es que no escalaremos los datos, ya que como hemos ido repitiendo a lo largo del estudio, están todos en la misma escala.

```
mahalanobis_dist_matrix <- function(x) {
  dec <- chol(cov(x))
  tmp <- forwardsolve(t(dec), t(x))
  dist(t(tmp))
}
dist_mat_mah <- as.dist(mahalanobis_dist_matrix(data0))
dist_mat_euc <- dist(data0, method = "euclidean")
clust_data_mah <- hclust(dist_mat_mah, method="ward.D2")
clust_data_euc <- hclust(dist_mat_euc, method="ward.D2")
par(mfrow = c(1,2))
plot(clust_data_mah, main = "Distancia Mahalanobis", xlab = "", sub = "",
     labels = FALSE)
plot(clust_data_euc, main = "Distancia euclídea", xlab = "", sub = "", labels = FALSE)
```



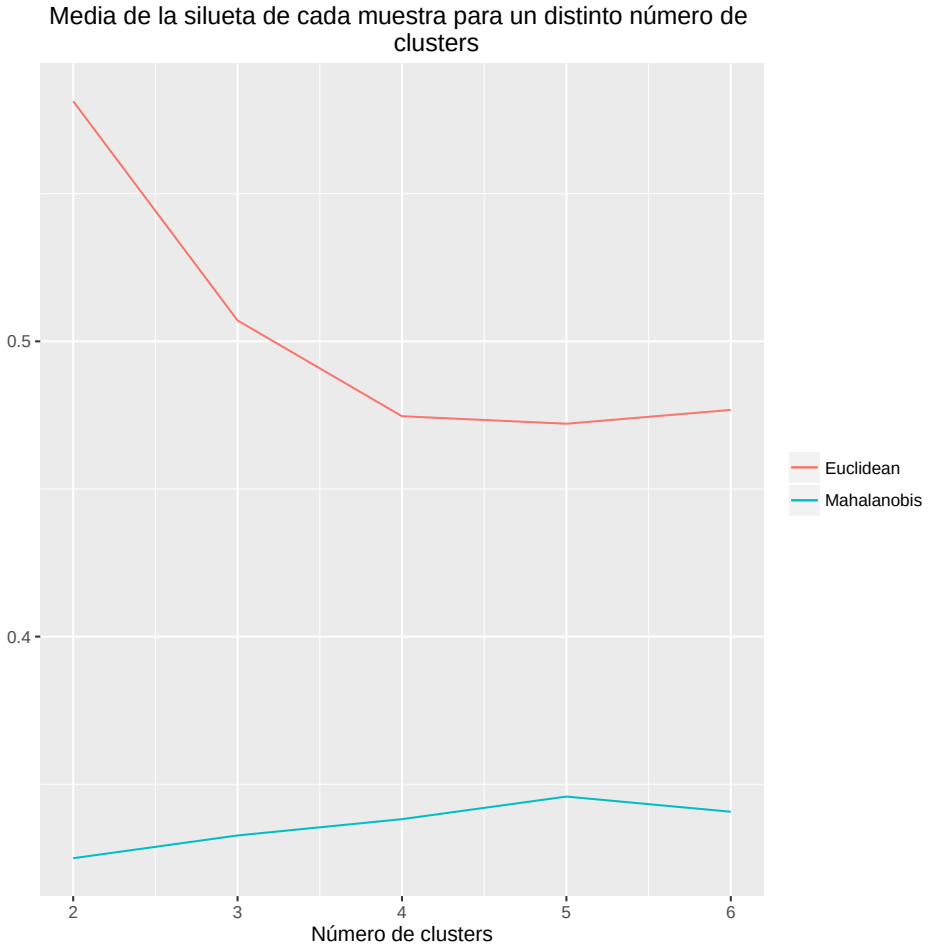
Salta a la vista que los dendrogramas obtenidos mediante los dos métodos son totalmente distintos. Para evaluar la calidad de cada jerarquía y elegir una partición, emplearemos una técnica cuantitativa conocida como el **método de las siluetas**.

Como ya hemos visto, el objetivo general perseguido por las técnicas de clustering consiste en identificar grupos con una similitud intra-cluster alta y una similitud inter-cluster baja. Así, el coeficiente silueta mide cuan buena es la asignación de una muestra a su cluster, comparando las distancias de este elemento respecto a todos los demás elementos del cluster al que pertenece, contra las distancias respecto a los clusters vecinos. Si dicho valor es cercano a -1, la muestra estará mal agrupada; si es aproximadamente 1 lo estará correctamente y si está alrededor de 0, el dato está entre dos clusters. El promedio de las siluetas de los elementos de un cluster da una idea de la calidad de ese cluster, mientras que la media de las siluetas de todo el set de datos nos dice cuan buena es la partición.

Vamos a considerar particiones con un número de clusters k de entre 2 a 6, y nos quedaremos con aquella que maximice el valor medio de la silueta.

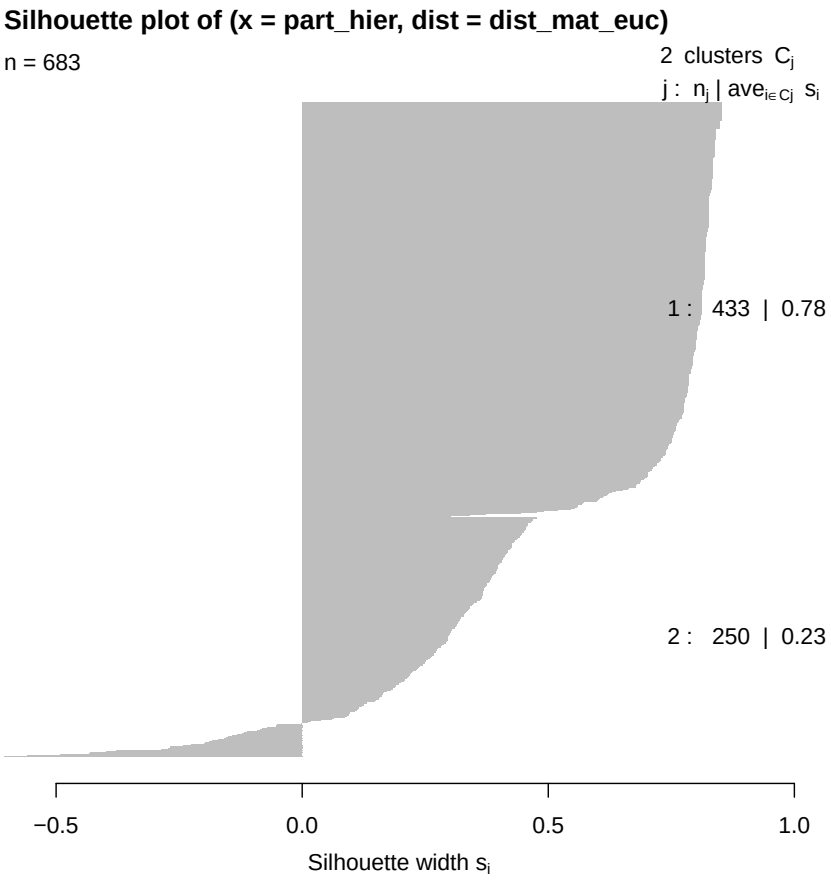
```
# Vamos a ver que k nos da el valor medio de la silueta mayor:
mean_sil <- data.frame(nrow = 5, ncol = 2)
colnames(mean_sil) <- c("Mahalanobis", "Euclidean")
for (k in 2:6)
  {
  clust_groups_mah <- cutree(clust_data_mah, k)
  clust_groups_euc <- cutree(clust_data_euc, k)
  sil_mah <- silhouette(clust_groups_mah, dist_mat_mah)
  sil_euc <- silhouette(clust_groups_euc, dist_mat_euc)
  mean_sil[k-1, 1] <- mean(sil_mah[,3])
  mean_sil[k-1, 2] <- mean(sil_euc[,3])
}

ggplot(mean_sil, aes(x = 2:6)) +
```

```
  geom_line(aes(y = Mahalanobis, colour = "Mahalanobis")) +
  geom_line(aes(y = Euclidean, colour = "Euclidean")) +
  ggtitle("Media de la silueta de cada muestra para un distinto número de
          clusters") + ylab(NULL) + xlab("Número de clusters") +
  theme(legend.title=element_blank())
```



El clustering basado en la matriz de distancias euclídeas es el de mayor calidad, ya que el valor medio de la silueta es mayor para todos los números de k. Concretamente, la mejor partición es la que tiene dos clusters.

```
part_hier <- cutree(clust_data_euc, 2)
sil <- silhouette(part_hier, dist_mat_euc)
summary(sil)

## Silhouette of 683 units in 2 clusters from silhouette.default(x = part_hier, dist = dist_mat_eu
##  Cluster sizes and average silhouette widths:
##       433       250
## 0.7838099 0.2306319
## Individual silhouette widths:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6056  0.3665  0.7588  0.5813  0.8184  0.8517

plot(sil)
```



Pese a que esta era la partición más adecuada, el segundo de los clusters parece no ser muy bueno ya que tiene un valor de silueta promedio bastante bajo y algunos de los tumores tienen coeficientes de silueta negativos. El otro cluster, en cambio, parece representar de manera adecuada la estructura de los datos.

Vamos a probar ahora a realizar un clustering particional empleando el algoritmo de k-medias, fijando el valor de k a 2. Este sencillo método consiste en seleccionar k centroides al azar y asignar a continuación cada muestra al centroide más cercano. Una vez realizado esto, se recalculan los centroides como la media de todas las muestras contenidas en ese cluster y se vuelven a reasignar las muestras siguiendo el mismo criterio. El proceso se repite hasta alcanzar convergencia o al llegar a un límite de iteraciones preestablecido. Cabe destacar que, dado que los centroides iniciales se eligen al azar, no se trata de un método determinista y el resultado depende del paso inicial.

5.2. Clustering particional

```
set.seed(1)
# Nota: los resultados de este método dependen de la elección inicial de los centroides.
# El parámetro nstart de la función kmeans() permite realizar pruebas con el número
# que le indiquemos de configuraciones iniciales, empleando para el análisis final la
# que mejores resultados da. En nuestro caso indicaremos que se realicen pruebas con
# 10 configuraciones.
kmeans_out <- kmeans(data0, 2, nstart = 10)
part_kmeans <- kmeans_out$cluster
```

Vamos por último a comparar los resultados obtenidos mediante los dos métodos de clustering.

5.3. Comparación de los resultados de clustering e interpretación

```
table(part_kmeans, part_hier)

##            part_hier
## part_kmeans   1   2
##           1 433  20
##           2   0 230
```

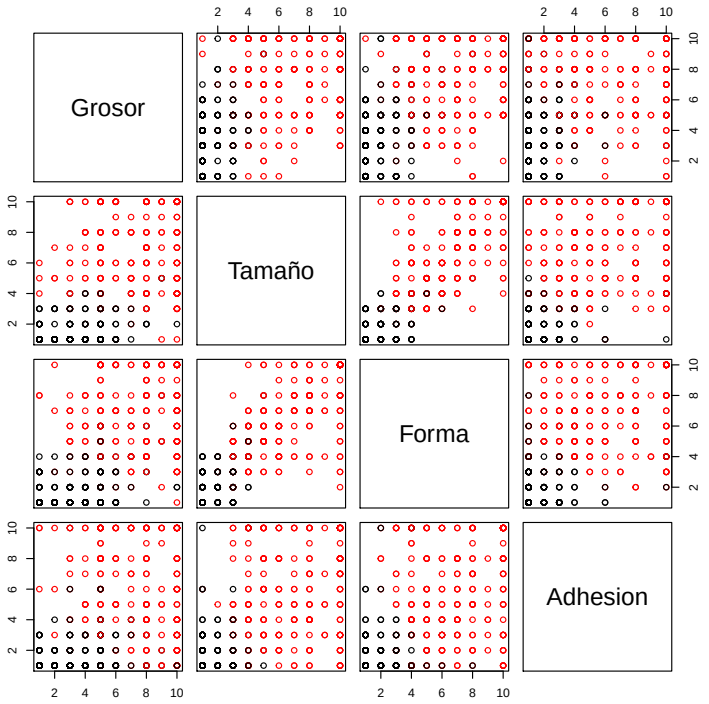
Los resultados obtenidos mediante los dos métodos son muy similares. La diferencia es de únicamente 20 muestras (3 %), que en el método particional son asignadas al cluster más grande y en el jerárquico al más pequeño. Sin embargo, debe recordarse que en el método jerárquico, el cluster más pequeño tenía asignadas muestras con un valor de silueta negativo. Es por tanto razonable estudiar si las diferencias que acabamos de ver entre los dos métodos coinciden con aquellas observaciones.

```
negative_sil <- which(sil[,3] < 0) # Todos estos están clasificados en el cluster 2
differences <- which(part_kmeans/part_hier != 1)
differences

##  13  16  26  51  52  58  60  64  66 102 106 149 233 235 249 349 416 495
##  13  16  25  49  50  56  58  62  64 100 104 145 227 229 242 335 401 480
## 556 658
## 541 642

intersect(negative_sil, differences)

##  [1]  13  16  25  49  50  56  58  62  64 100 104 145 227 229 242 335 401
## [18] 480 541 642

length(intersect(negative_sil, differences))

## [1] 20
```

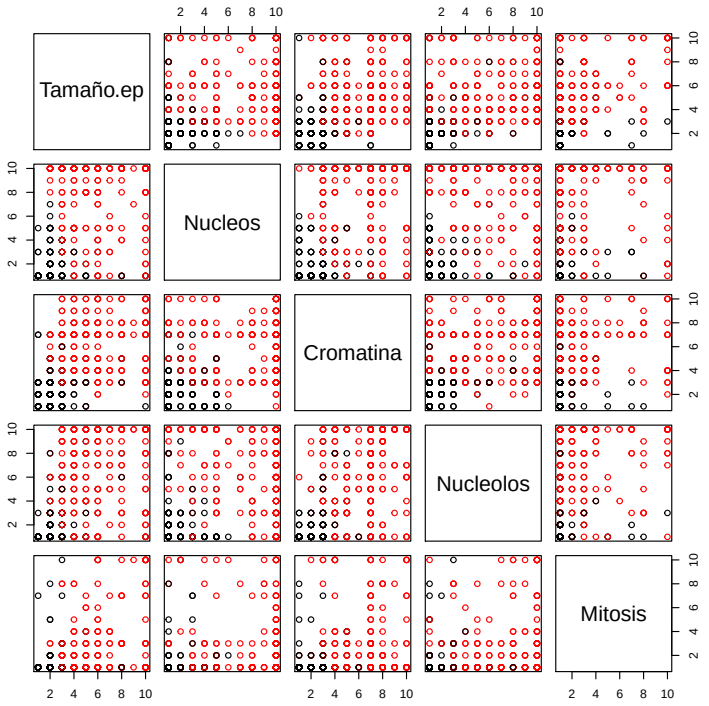
En efecto, los 20 tumores que estamos clasificando de manera distinta entre los dos métodos corresponden a muestras con silueta negativa en el clustering jerárquico.

Los resultados hasta ahora obtenidos sugieren firmemente que los clusters se van a corresponder con tumores de prognosis opuesta. Lo comprobaremos y, para ello, nos quedaremos con los grupos sugeridos por el algoritmo de k-medias, ya que hemos visto que las 20 observaciones que lo diferencian del método jerárquico tenían silueta negativa en este último método.

```
final_clusters <- part_kmeans
final_clusters[final_clusters == 1] <- "Benigno"
final_clusters[final_clusters == 2] <- "Maligno"
data_temp <- cbind(data, final_clusters)
par(mfrow = c(1,2))
plot(data0[,1:4], col = kmeans_out$clust)
```




```
plot(data0[,5:9], col = kmeans_out$clust)
```



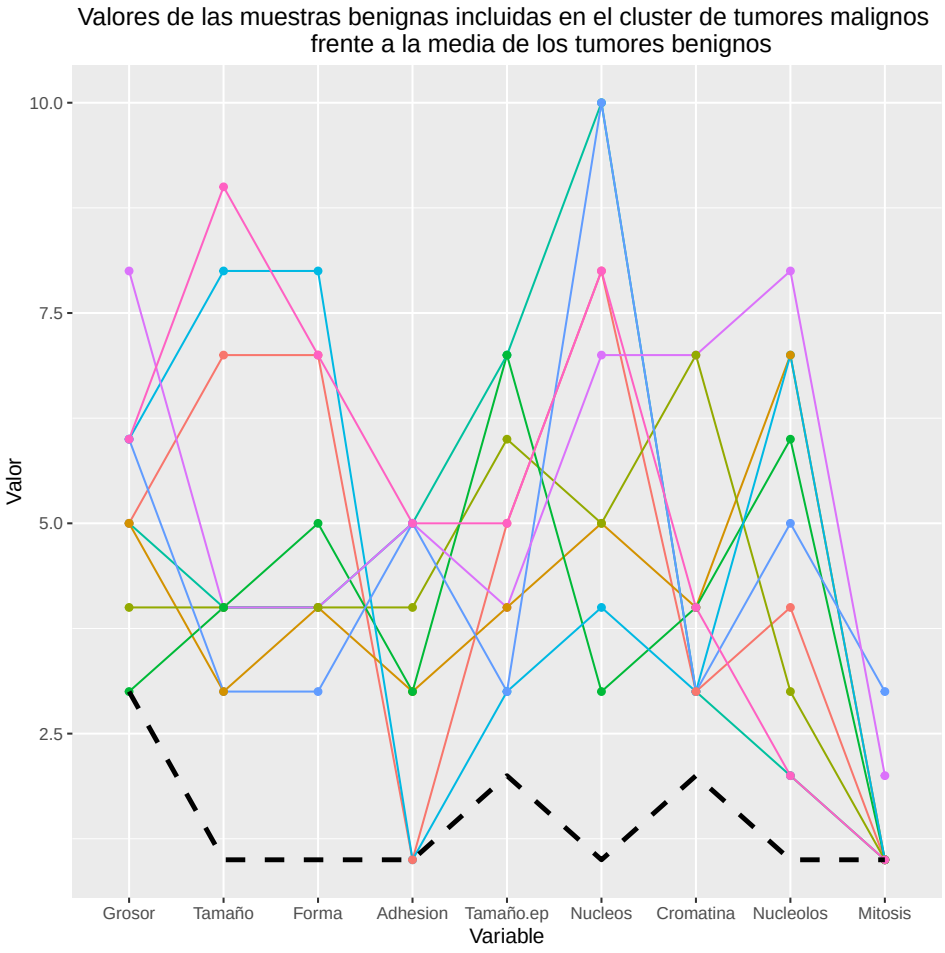

```
confusionMatrix(final_clusters, data$Prognosis)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction Benigno Maligno
##    Benigno     435      18
##    Maligno       9     221
```

```
##
##                Accuracy : 0.9605
##                  95% CI : (0.943, 0.9738)
##     No Information Rate : 0.6501
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9123
##  Mcnemar's Test P-Value : 0.1237
##
##             Sensitivity : 0.9797
##             Specificity : 0.9247
##          Pos Pred Value : 0.9603
##          Neg Pred Value : 0.9609
##              Prevalence : 0.6501
##          Detection Rate : 0.6369
##    Detection Prevalence : 0.6633
##       Balanced Accuracy : 0.9522
##
##        'Positive' Class : Benigno
##
```

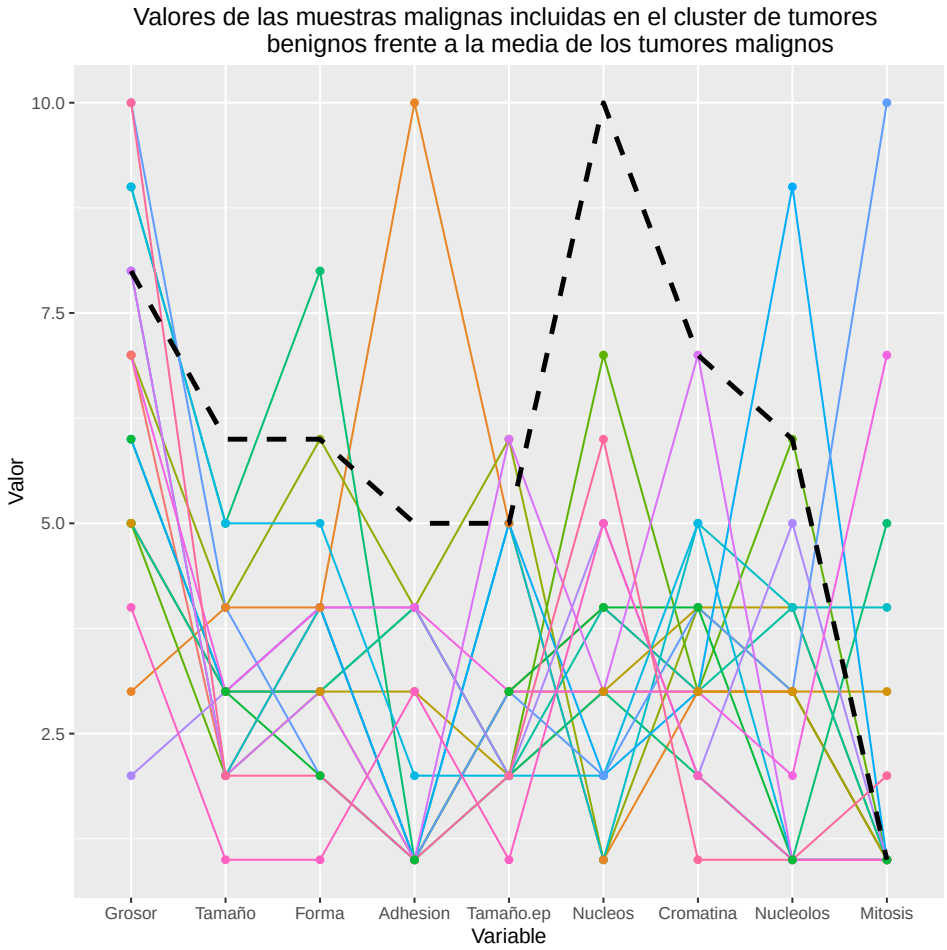
El clustering está agrupando las muestras de tal manera que aquellas que toman bajos valores en todas o casi todas las variables están en un cluster, y las que toman valores altos en otro. Estos gráficos son similares a los que hemos obtenido en el apartado de estadística descriptiva, cuando intentábamos relacionar las cualidades de los tumores con el diagnóstico. Así, cuando calculamos la tabla de contingencia entre los clusters y la prognosis, vemos que, como sospechábamos, concuerdan de manera significativa.

Intentaremos, por último, averiguar la razón por la que los clusters no se corresponden completamente con la prognosis.

```
true_benigns <- which((as.numeric(as.factor(final_clusters)))/(as.numeric(as.factor(
  data$Prognosis))) > 1)
true_maligns <- which((as.numeric(as.factor(final_clusters)))/(as.numeric(as.factor(
  data$Prognosis))) < 1)
true_benign_data <- data[true_benigns,-11]
benign_median <- apply(data[data$Prognosis == "Benigno",-c(1,11)], 2, median)
benign_median <- data.frame(x = 1:9, y = benign_median, ID = NA)
true_malign_data <- data[true_maligns,-11]
malign_median <- apply(data[data$Prognosis == "Maligno",-c(1,11)], 2, median)
malign_median <- data.frame(x = 1:9, y = malign_median, ID = NA)
# Benignos
benign_melt <- melt(true_benign_data, id.var = "ID")
benign_melt$ID <- as.factor(benign_melt$ID)
ggplot(data = benign_melt, aes(x = variable, y = value, group = ID)) +
     geom_line(aes(color = ID)) +
     geom_point(aes(color = ID)) + geom_line(data = benign_median, aes(x = x, y = y),
                                             linetype = "dashed", size = 1.2) +
  theme(legend.position = 'none') + xlab("Variable") + ylab("Valor") +
  ggtitle("Valores de las muestras benignas incluidas en el cluster de tumores malignos
          frente a la media de los tumores benignos")
```




```
# Malignos
malign_melt <- melt(true_malign_data, id.var = "ID")
malign_melt$ID <- as.factor(malign_melt$ID)
ggplot(data = malign_melt, aes(x = variable, y = value, group = ID)) +
     geom_line(aes(color = ID)) +
     geom_point(aes(color = ID)) + geom_line(data = malign_median, aes(x = x, y = y),
                                             linetype = "dashed", size = 1.2) +
    theme(legend.position='none') + xlab("Variable") + ylab("Valor") +
    ggtitle("Valores de las muestras malignas incluidas en el cluster de tumores
            benignos frente a la media de los tumores malignos")
```



En lo que a los tumores benignos que han sido clasificados como malignos se refiere, vemos que todos ellos tienen, para casi todos los atributos, valores más altos que lo esperable para un tumor benigno. El caso de los malignos clasificados como benignos no es tan claro ya que los valores para algunos de los atributos son en varios casos cercanos o incluso superiores a la media.

6. Análisis discriminante

Las técnicas que hemos aplicado hasta ahora eran exploratorias, es decir, inferían patrones basándose únicamente en los datos y sin tener conocimiento alguno acerca de las posibles clases. Ahora vamos a ir un paso más allá y vamos a intentar crear un modelo que nos ayude a separar dos clases concretas (tumores malignos y benignos), creado tras la observacion de muestras junto con sus etiquetas para estas clases.

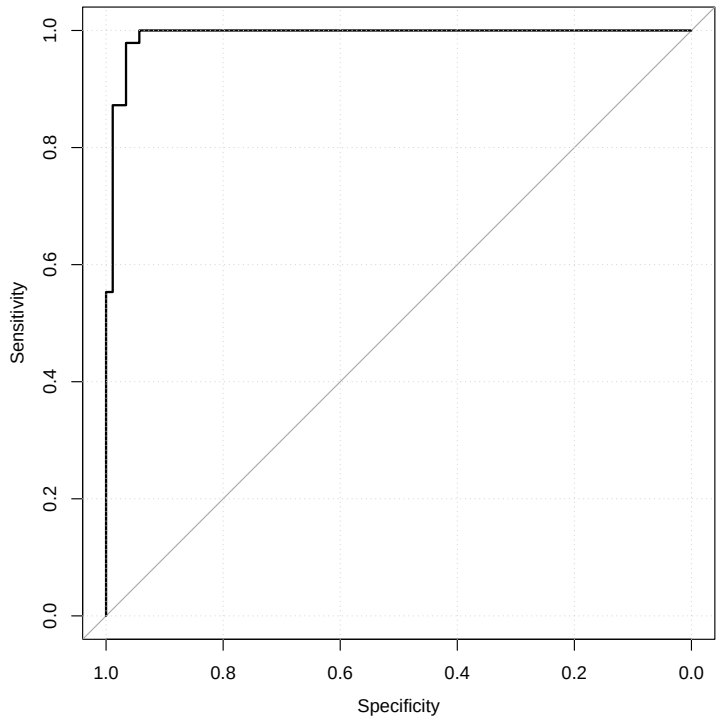
Emplearemos para ello el discriminante lineal. En este método se busca la combinación lineal de las variables descriptivas que maximiza el ratio entre la *matriz de covarianzas entre grupos* y la *matriz de covarianzas dentro de los grupos*, de donde se intuye la misma idea que tiene el clustering o el análisis de componentes principales. Sin embargo, aquí tenemos el añadido de que explota la información que le estamos dando acerca de la clase, ya que que intenta modelar explícitamente la diferencia entre las dos clases.

Antes que ello fijaremos los parámetros de entrenamiento. En este caso vamos a realizar 10 repeticiones de 10-fold cross-validation estratificada, manteniendo una proporción de 80 %-20 % entre los sets de entrenamiento y testeo. El criterio que vamos a emplear para evaluar el clasificador en cada fold va a ser el coeficiente Kappa.

```
set.seed(6)
inTrain <- createDataPartition(y = data$Prognosis, p = .8, list = FALSE)
train <- data[inTrain,]
dim(train)
```

```
## [1] 548  11
```

```r
test <- data[-inTrain,]
dim(test)

## [1] 135  11

fitControl <- trainControl(
  method = "repeatedcv",
  p = 0.8,
  number = 10,
  repeats = 10)
lda_fit <- train(Prognosis ~ ., data = train[,-1],
                 method = "lda",
                 trControl = fitControl)
lda_fit$finalModel

## Call:
## lda(x, grouping = y)
##
## Prior probabilities of groups:
##  Benigno  Maligno
## 0.649635 0.350365
##
## Group means:
##           Grosor   Tamaño    Forma Adhesion Tamaño.ep  Nucleos Cromatina
## Benigno 2.935393 1.289326 1.396067 1.376404  2.129213 1.317416  2.087079
## Maligno 6.979167 6.651042 6.697917 5.755208  5.265625 7.562500  6.031250
##         Nucleolos  Mitosis
## Benigno  1.272472 1.022472
## Maligno  5.916667 2.635417
##
## Coefficients of linear discriminants:
##                    LD1
## Grosor     0.18575785
## Tamaño     0.16298801
## Forma      0.10667130
## Adhesion   0.04025007
## Tamaño.ep  0.05354128
## Nucleos    0.24925644
## Cromatina  0.10045430
## Nucleolos  0.09815562
## Mitosis   -0.01787236

predicted_probs <- predict(lda_fit, test[,-c(1,11)], type = "prob")
roc(test$Prognosis, predicted_probs[,1], plot = TRUE)

##
## Call:
## roc.default(response = test$Prognosis, predictor = predicted_probs[,     1], plot = TRUE)
##
## Data: predicted_probs[, 1] in 88 controls (test$Prognosis Benigno) > 47 cases (test$Prognosis M
## Area under the curve: 0.9915

grid()
```




```
confusionMatrix(test$Prognosis, predict(lda_fit, test[,-c(1,11)]))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Benigno Maligno
##    Benigno      85       3
##    Maligno       2      45
##
##                Accuracy : 0.963
##                  95% CI : (0.9157, 0.9879)
##     No Information Rate : 0.6444
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9188
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9770
##             Specificity : 0.9375
##          Pos Pred Value : 0.9659
##          Neg Pred Value : 0.9574
##              Prevalence : 0.6444
##          Detection Rate : 0.6296
##    Detection Prevalence : 0.6519
##       Balanced Accuracy : 0.9573
##
##        'Positive' Class : Benigno
##
```

El análisis discriminante muestra unos resultados realmente buenos: el area bajo la curva es de 0.99 y la precisión de 0.96, fallando solo en 5 casos de un total de 135 (3 %).

Por último, cabe mencionar que existe una ligera diferencia entre el discriminante lineal y el discriminante lineal de Fisher: mientras que el primero de ellos asume que (1) las variables independientes siguen una distribución normal para cada una de las clases y (2) las matrices de covarianzas dentro de

cada clase deben ser aproximadamente iguales, el segundo no hace asunción alguna. En el caso de que nuestros datos cumplan dichas condiciones habremos realizado un análisis del primer tipo y de no ser así, habremos llevado a cabo el segundo.

```
# Hipotesis de normalidad: shapiro-test multivariante. La hipotesis nula que sigue es
# que los datos provienen de distribuciones normales
benign_index <- which(data$Prognosis == "Benigno")
malign_index <-  which(data$Prognosis != "Benigno")
mshapiro.test(as.matrix(t(data0[benign_index,])))

##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.19275, p-value < 2.2e-16

mshapiro.test(as.matrix(t(data0[malign_index,])))

##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.94609, p-value = 9.898e-08
```

Los datos no cumplen la asunción de normalidad dentro de las clases; por tanto, nuestro ultimo análisis corresponde a un discriminante lineal de Fisher.

7. Discusión

Mediante las tres técnicas que hemos empleado (PCA, clustering y análisis discriminante) hemos obtenido resultados similares, intuibles tras el análisis descriptivo inicial de los datos. En los datos se pueden separar de manera lineal dos grandes grupos, de tal manera que uno se corresponde con tumores de prognosis benigna y el otro con tumores de prognosis maligna.

En cuanto al análisis de componentes principales, la mayor parte de la variabilidad (70 %) viene explicada por la primera componente, que a su vez hemos intuido que se relaciona con la benignidad de un tumor. El hecho de que esta componente tenga una alta correlación con todas las variables es por tanto esperable si tenemos en cuenta que todas ellas definen un estado anómalo de una muestra citológica. De entre las variables más significativas destacan la uniformidad de la forma y el tamaño celular; en efecto varios estudios han demostrado la utilidad de estas características para determinar el diagnóstico de un tumor (6). Este análisis había sugerido la posibilidad de que existiesen subtipos de tumores con un ligero comportamiento nuclear distinto; sin embargo, el posterior estudio de clustering no ha confirmado esa suposición. Si acudimos a la bibliografía vemos que hay distintos tipos de tumores de mama, entre ellos el carcinoma ductal no-invasivo (papilar, micropapilar y cibriforme), el carcinoma ductal invasivo, el carcinoma papilar intraductal y el carcinoma papilar invasivo, que deben ser diferenciados dado su distinto nivel de agresividad. Así, una de las características de los carcinomas ductales invasivos es que muestran hipercelularidad, un acúmulo de láminas de celulas epiteliales con anomalías nucleares y ausencia de núcleos desnudos. El carcinoma micropapilar, por su parte, carece de tejido fibrovascular y muestra agregados celulares con un patrón tubuloalveolar y de papilares angulares (7). Este tipo de características son más específicas que las que estamos estudiando aquí y probablemente nuestras variables no sean capaces de capturar esas cualidades.

Este primer análisis también nos ha mostrado la efectividad de las componentes principales para capturar la estructura de clusters de los datos, que se han confirmado con las técnicas empleadas a continuación.

El hecho de que las variables originales estuviesen fuertemenete correladas sugería el uso de la distancia de Mahalanobis en el clustering. Sin embargo, ha resultado que la distancia euclídea ha sido más efectiva a la hora de mostrar la estructura en grupos de los datos. Los resultados obtenidos mediante el clustering jerárquico y el particional han sido prácticamente iguales, lo que da confianza a la metodología

de trabajo escogida.

Por último, el análisis discriminante ha resultado ser la técnica más eficaz a la hora de encontrar los clusters de los datos. Este es un resultado que a priori no resulta sorprendente dado que estamos diciendo al modelo a qué tipo de clase corresponde cada patrón que observa en los datos. Es destacable que esto no es siempre de utilidad; véase, por ejemplo, el caso de que quisiésemos encontrar subtipos de la enfermedad.

8. Conclusiones

Todas las variables de las que disponemos información en este estudio resultan muy útiles a la hora de discriminar los tumores en benignos y malignos; concretamente, las más signifivativas son la uniformidad del tamaño y forma celular y la que menos el número de mitosis. Sin embargo se debe tener en cuenta que muchas ellas están muy correladas y que no se trata de un caso en el que cada una de ellas está explicando un aspecto distinto de los tumores. Estos datos permiten separar las muestras benignas de las malignas de manera lineal, sin necesidad de un algoritmo complejo.

Sin embargo, estas variables no son suficientes para encontrar subtipos de cáncer de mama o estadíos de la enfermedad en los datos, aspectos cruciales a la hora de determinar el tratamiento a llevar a cabo. Una rápida revisión bibliográfica sugiere que es necesario el conocimiento de otros factores como el perfil genético, receptores celulares, características fibrovasculares o si los ganglios linfáticos están afectados.

References

[1] Murat Karabatak and M. Cevdet Ince. An expert system for detection of breast cancer based on association rules and neural network. *Expert Syst. Appl.*, 36(2):3465–3469, March 2009.

[2] Elif Derya íbeyli. Implementing automated diagnostic systems for breast cancer detection. *Expert Syst. Appl.*, 33(4):1054–1062, November 2007.

[3] Uci machine learning repository. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+ Accessed: 2015-12-27.

[4] Gouda I. Salama, M. B. Abdelhalim, and Magdy Abd elghany Zeid. Breast cancer diagnosis on three different datasets using multi-classifiers", int. Technical report, J. of Comput. and Inform. Technology, 2012.

[5] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 29–, New York, NY, USA, 2004. ACM.

[6] S.-C. Chen, Y.-C. Cheung, C.-H. Su, M.-F. Chen, T.-L. Hwang, and S. Hsueh. Analysis of sonographic features for the differentiation of benign and malignant breast tumors of different sizes. *Ultrasound in Obstetrics and Gynecology*, 23(2):188–193, 2004.

[7] Deepti Aggarwal, Navmeet Soin, Dipti Kalita, Leela Pant, Madhur Kudesia, and Sompal Singh. Cytodiagnosis of papillary carcinoma of the breast: Report of a case with histological correlation. *Journal of cytology / Indian Academy of Cytologists*, 31(2):119–121, April 2014.