Trabajo de Clasificación en Introducción a Ciencia de Datos. Dataset pima.

Danel Arias

2023-12-11

Descripción del trabajo

El trabajo consiste en la realización de un análisis exploratorio de datos (EDA) sobre un dataset y realizar una clasificación con diferentes métodos para el dataset. El dataset indicado es el dataset pima, que contiene datos sobre la diabetes de mujeres de la tribu pima.

Descripción del dataset

Según podemos leer en KEEL se trata de un dataset sobre la diabetes de los indios Pima extraido del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. Donde se impusieron varias restricciones a la selección de estos casos de una base de datos más amplia. En particular, todos los pacientes son mujeres de al menos 21 años de ascendencia india Pima. La etiqueta de clase representa si la persona tiene o no diabetes.

Las diferentes variables son:

- Preg: Número de veces que ha estado embarazada. [0-17]
- Plas: Concentración plasmática de glucosa a las 2 horas en una prueba oral de tolerancia a la glucosa [0-199]
- Pres: Presión arterial diastólica (mm Hg) [0-122]
- Skin: Espesor del pliegue cutáneo del tríceps (mm) [0-99]
- Insu: Insulina sérica de 2 horas (mu U/ml) [0-846]
- Mass: Índice de masa corporal (peso en kg/(altura en m)^2) [0-67.1]
- Pedi: Función de pedigrí de la diabetes [0.078-2.42]
- Age: Edad (años) [21-81]
- Class: Variable de clase ('tested_negative' o 'tested_positive')

En total el dataset tiene 8 variables y 768 observaciones. Además, según la descripción en KEEL este dataset no contiene valores perdidos (NA's).

Preparación inicial del entorno

Carga de librerías

```r
library(ggplot2)
library(moments)
library(corrplot)
library(dplyr)
library(caret)
library(class)
library(MASS)
library(tidyverse)
library(ggthemes)
library(gridExtra)
library(klaR)
```

Carga de datos

Cargamos los datos de pima/pima.dat. Al ser un fichero .dat debemos obviar las filas que no son datos, estas empiezan en '@'

```r
pima <- read.table("pima/pima.dat", header = FALSE,
                   sep = ",", comment.char = "@")
```

Añadimos los nombres de las columnas. Los nombres de las columnas están en pima.dat. Y se encuentran en la línea tras '@inputs' así que lo leemos y lo añadimos.

```r
nombres <- grep("^@inputs", readLines("pima/pima.dat"), value = TRUE) %>%
  # Eliminamos el @inputs
  str_remove("@inputs ") %>%
  #Eliminamos espacions
  str_remove_all(" ") %>%
  # Separamos por comas
  str_split(",") %>%
  # Convertimos a vector
  unlist()

# El nombre de la última variable se encuentra tras @outputs
nombre_output <- grep("^@outputs", readLines("pima/pima.dat"), value = TRUE) %>%
  # Eliminamos el @outputs
  str_remove("@outputs ")

# Añadimos el nombre de la última variable
nombres <- c(nombres, nombre_output)

# Añadimos los nombres al dataset
colnames(pima) <- nombres
```

Análisis exploratorio de datos

Primero echamos un vistazo a los datos son un `summary()`. Podemos ver que cuadra con la descripción proporcionada.

```
##       Preg             Plas            Pres             Skin
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##       Insu            Mass            Pedi             Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##     Class
##  Length:768
##  Class :character
##  Mode  :character
##
##
##
```

Revisamos si tenemos NA's por si acaso y confirmamos lo que se nos indicaba en KEEL, no hay NA's.

```
colSums(is.na(pima))
```

```
##  Preg  Plas  Pres  Skin  Insu  Mass  Pedi   Age Class
##     0     0     0     0     0     0     0     0     0
```

Análisis variable a variable

En este apartado analizaremos cada variable por separado.

Variable 'Class'

Comezamos con la variable 'Class' que es una variable categórica binaria. La convertimos a factor para poder trabajar con ella y visualizamos los valores que toma (Figura 1) donde vemos que hay más mujeres que no tienen diabetes que mujeres que sí la tienen.

Variable 'Preg'

Para comenzar con las variables numéricas estudiamos la variable 'Preg'. Sabemos que pertenece al rango [0-17]. Para más detalle, vemos un histograma en la Figura (2). Vemos que hay un gran número de mujeres que no han estado embarazadas, y que el número de embarazos va disminuyendo conforme aumenta el número de embarazos.

Estudiamos también los cuantiles con un boxplot (Figura 3). Donde vemos que el 25% de las mujeres no han estado embarazadas y y que solo un 25% ha estado embarazada más de 6 veces. Observamos que hay valores atípicos principalmente en el extremo superior, estos parecen ser valores correctos, por lo que no los eliminaremos.

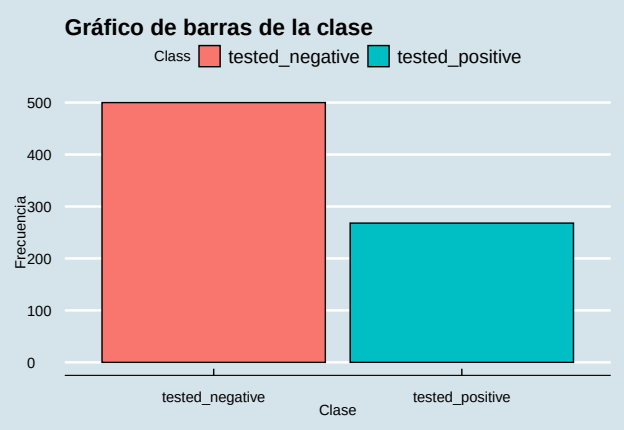

Figure 1: Gráfico de barras de la clase

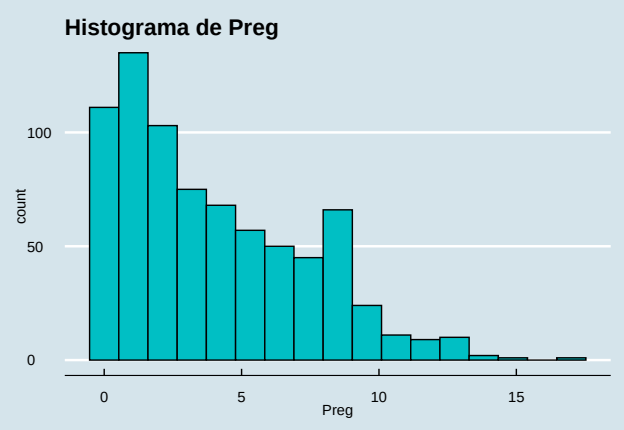

Figure 2: Histograma de Preg

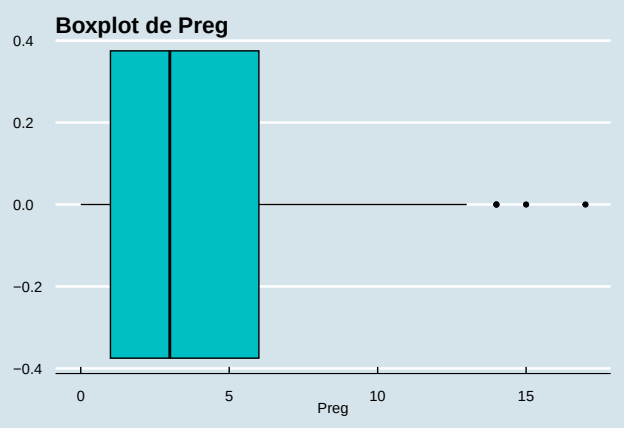

Figure 3: Boxplot de Preg

```
##   0%  25%  50%  75% 100%
##    0    1    3    6   17
```

Estudiamos la forma de la variable Preg. Para ello miramos la media, sd, skewness y kurtosis.

```
## [1] "Media: 3.8451"
```

```
## [1] "Deviación estándar: 3.3696"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 0.8999"
```

```
##
##  D'Agostino skewness test
##
## data:  pima$Preg
## skew = 0.89991, z = 8.90464, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Dado que la skewness > 0 la distribución está sesgada a la derecha. Ademas el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada a la derecha.

```
## [1] "Kurtosis: 3.1504"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pima$Preg
## kurt = 3.15038, z = 0.93339, p-value = 0.3506
## alternative hypothesis: kurtosis is not equal to 3
```

Dados que la kurtosis > 0 la distribución tiene colas pesadas. En cambio el valor de p-value del test de Anscombe es mayor que 0.05, por lo que es posible que no se pueda afirmar que la distribución no es normal. Podemos verlo mejor con gráfico de densidad (ver Figura 4). Realizamos el test de Shapiro-Wilk para confirmar la normalidad de la variable. Donde vemos que el valor de p-value es menor que 0.05, por lo que la variable Preg no sigue una distribución normal. Lo vemos también el QQ-plot (Figura 5).

```
##
##  Shapiro-Wilk normality test
##
## data:  pima$Preg
## W = 0.90428, p-value < 2.2e-16
```

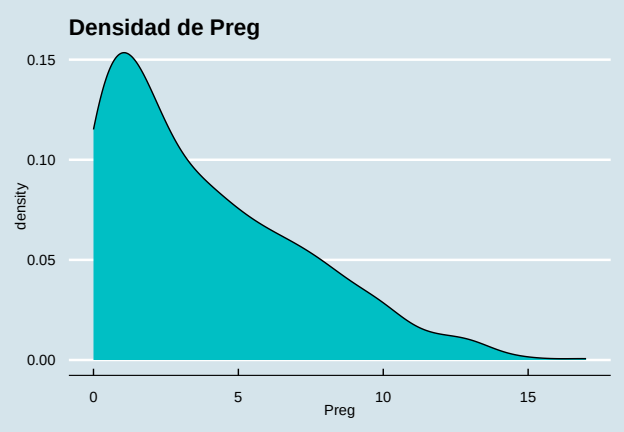

Figure 4: Densidad de Preg

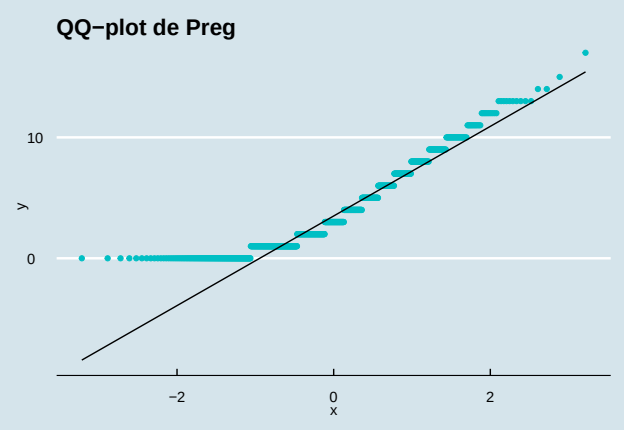

Figure 5: QQ-plot de Preg

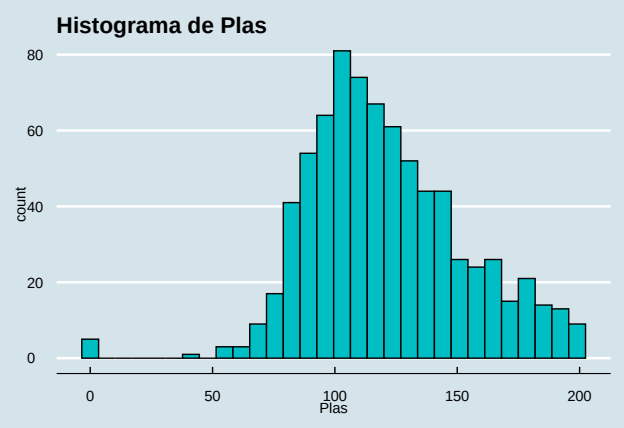

Figure 6: Histograma de Plas

Variable 'Plas'

Estudiamos la variable 'Plas' perteneciente al rango [0-199]. Para más detalle, vemos un histograma en la Figura (6). Vemos que hay un gran número de mujeres con una concentración de glucosa en plasma entre 80 y 130.

Llama la atención que haya un pequeño pico en 0, esto puede indicar que hay valores perdidos. Dada la naturaleza de la variable, la concentración plasmática de glucosa no puede ser 0. Asignamos estos valores como NA's y toca valorar como imputarlos. Dado que son pocos valores y que la distribución es bastante simétrica, imputaremos con la media.

```
## [1] "Resumen antes de imputar:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    44.0    99.0   117.0   121.7   141.0   199.0       5
```

```
## [1] "Resumen después de imputar:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.00   99.75  117.00  121.69  140.25  199.00
```

Estudiamos también los cuantiles con un boxplot (Figura 7). Donde se observa que el 25% de las mujeres tienen una concentración de glucosa en plasma menor de 100, y que el 25% de las mujeres tienen una concentración de glucosa en plasma mayor de 140.

```
##     0%    25%    50%    75%   100%
##  44.00  99.75 117.00 140.25 199.00
```

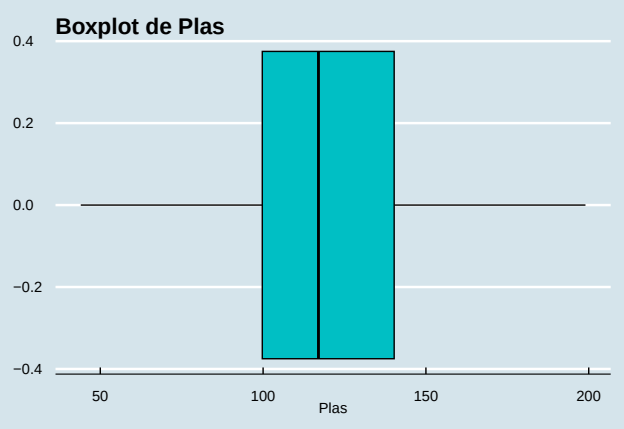

Figure 7: Boxplot de Plas

Estudiamos la forma de la variable Plas, Para ello miramos la media, sd, skewness y kurtosis.

```
## [1] "Media: 121.6868"
```

```
## [1] "Deviación estándar: 30.4359"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 0.5317"
```

```
##
##  D'Agostino skewness test
##
## data:  pima$Plas
## skew = 0.53168, z = 5.71818, p-value = 1.077e-08
## alternative hypothesis: data have a skewness
```

Dado que skewness > 0 la distribución está sesgada a la derecha. Además el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada a la derecha.

```
## [1] "Kurtosis: 2.7347"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pima$Plas
## kurt = 2.7347, z = -1.6211, p-value = 0.105
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. En cambio el valor de p-value del test de Anscombe es mayor que 0.05, por lo que no se puede afirmar que la distribución no es normal. Visualizamos esto mejor con un gráfico de densidad (ver Figura 8) donde no se ve claramente que la distribución rechaze la normalidad. Por lo tanto realizamos el test de Shapiro-Wilk para ver si la variable Plas sigue una distribución normal y para visualizarlo mejor usamos un QQ-plot (ver Figura 9). El valor de p-value es menor que 0.05, por lo que podemos afirmar que la variable Plas no sigue una distribución normal.

```
##
##  Shapiro-Wilk normality test
##
## data:  pima$Plas
## W = 0.9699, p-value = 1.777e-11
```

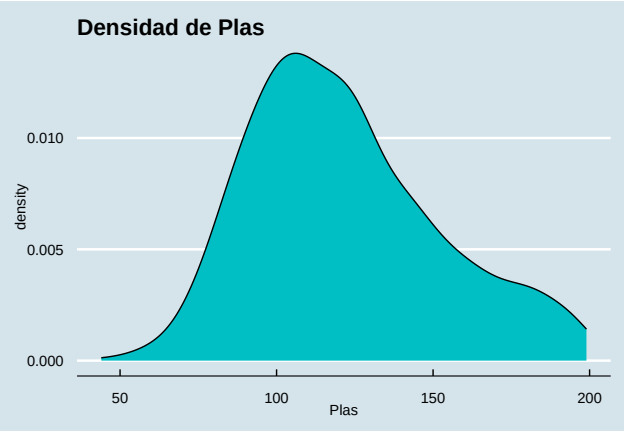

Figure 8: Densidad de Plas

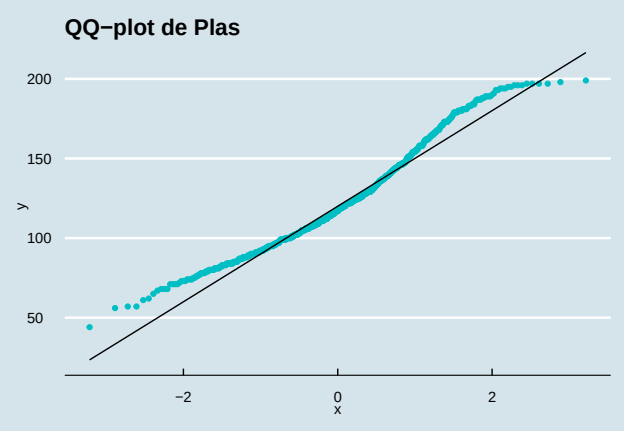

Figure 9: QQ-plot de Plas

Variable 'Pres'

Estudiamos a continuación la variable 'Pres' que se refiere a la presión arterial diastólica la cual pertenece al rango [0, 122]. Visualizamos la distribución de la variable con un histograma (Figura 10). Donde vemos que la mayoría de los valores se encuentran entre 40 y 100, y que hay un pico en 0.

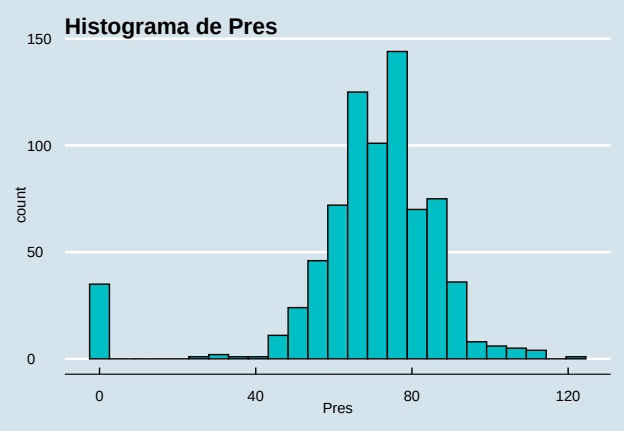

Figure 10: Histograma de Pres

El pico en el valor 0 es interesante ya que puede indicar que hay valores perdidos. Dada la naturaleza de la variable, la presión arterial diastólica no puede ser 0. Asignamos estos valores como NA's y toca valorar como imputarlos. Dado que son pocos valores y que la distribución es bastante simétrica, imputaremos con la media.

```
## [1] "Resumen antes de imputar:"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   24.00   64.00   72.00   72.41   80.00  122.00      35

## [1] "Resumen después de imputar:"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.00   64.00   72.20   72.41   80.00  122.00
```

Estudiamos también los cuantiles con un boxplot (Figura 11) y vemos que el 25% de las mujeres tienen una presión arterial diastólica menor de 64, y que el 25% de las mujeres tienen una presión arterial diastólica mayor de 80.

```
##        0%       25%       50%       75%      100%
##  24.00000  64.00000  72.20259  80.00000 122.00000
```

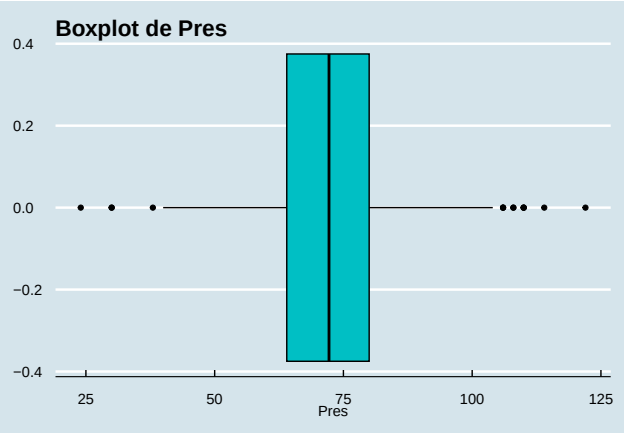

Figure 11: Boxplot de Pres

Estudiamos la forma de la variable Pres. Para ello miramos la media, sd, skewness y kurtosis.

```
## [1] "Media: 72.4052"
```

```
## [1] "Desviación estándar: 12.0963"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 0.137"
```

```
##
##  D'Agostino skewness test
##
## data:  pima$Pres
## skew = 0.13704, z = 1.55814, p-value = 0.1192
## alternative hypothesis: data have a skewness
```

Dados que skewness > 0 la distribución está sesgada a la derecha. En cambio el valor de p-value del test de D'agostino es mayor que 0.05, por lo que no podemos afirmar con seguridad que la distribución este sesgada.

```
## [1] "Kurtosis: 4.0828"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pima$Pres
## kurt = 4.0828, z = 4.2911, p-value = 1.778e-05
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución no es normal. Podemos verlo mejor con un gráfico de densidad (Figura 12). Estudiamos de todas formas la normalidad de la variable Pres con el test de Shapiro-Wilk y lo visualizamos con un QQ-plot (Figura 13).

```
##
##  Shapiro-Wilk normality test
##
## data:  pima$Pres
## W = 0.98804, p-value = 6.463e-06
```

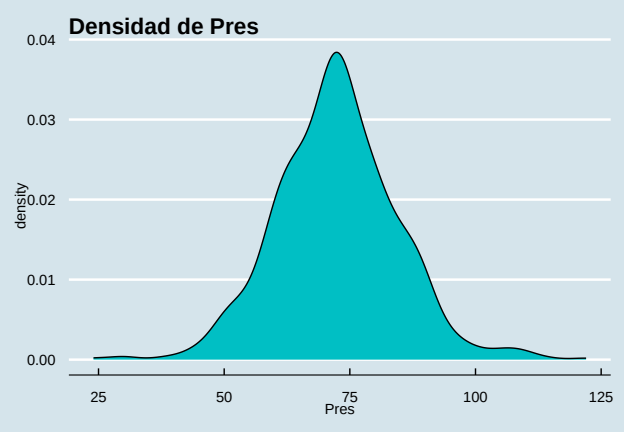

Figure 12: Densidad de Pres

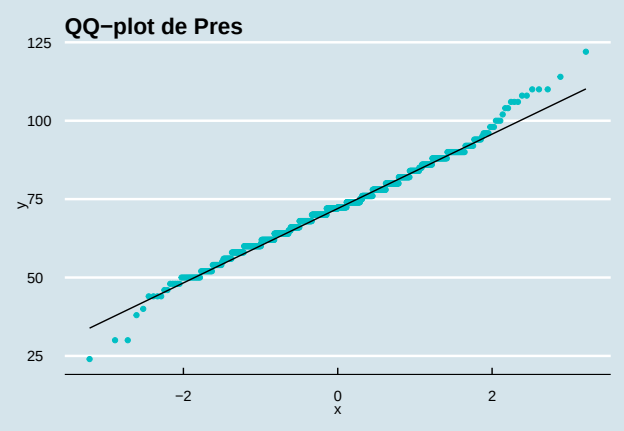

Figure 13: QQ-plot de Pres

Variable 'Skin'

Ahora la variable 'Skin' que se refiere al grosor del pliegue cutáneo del tríceps. Vemos que los valores pertenecen al rango [0-99], para más detalle vamos a ver el histograma (Figura 14). Vemos que el grueso de los valores se encuentran entre 5 y 60, con un gran pico en 0 y con un pequeño pico en 100. Esto puede indicar que hay valores perdidos,

Estudiando esos posibles valores perdidos vemos que hay 227 valores en 0, dado que Skin se refiere al grosor del pliegue cutáneo del tríceps no tiene sentido que haya valores en 0, por lo que se tratan de valores perdidos.

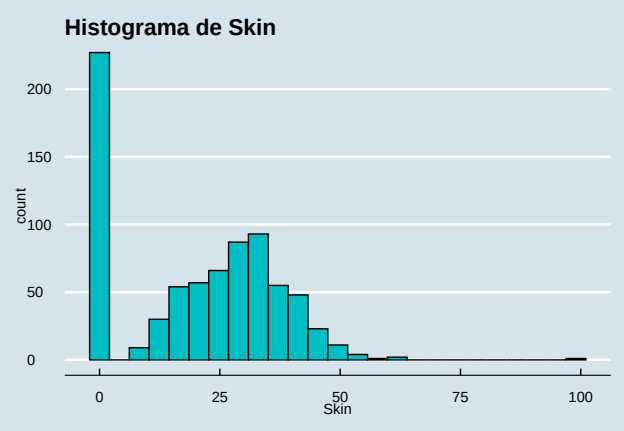

Figure 14: Histograma de Skin

Podemos ver que representan un 30% de los valores, por lo que eliminarlos nos dejaría con sólo el 70% de los datos, por lo que no es una opción. Por otro lado, en el caso de imputarlos de forma generalizada (ya sea media o mediana), al ser tantos valores perdidos, la imputación podría sesgar los datos. Lo más correcto sería usar algún tipo de técnica que nos permita extraer información estimada de los valores perdidos usando la información disponible del resto de variables del dataset. Esto se podría conseguir con un KNN o una regresión lineal, pero este no entra dentro del objetivo principal de este trabajo. Por estas razones y dado que es una clasificación, se decide imputar siguiendo la media de cada clase. Se elige la media dado que la distribución es relativamente simétrica.

```
## [1] "Resumen antes de imputar:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    7.00   22.00   29.00   29.15   36.00   99.00     227
```

```
## [1] "Media clase 'tested_negative': 27.2355"
```

```
## [1] "Media clase 'tested_positive': 33"
```

```
## [1] "Resumen después de imputar:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   25.00   28.00   29.25   33.00   99.00
```

Estudiamos los cuantiles de la variable Skin y lo visualizamos con un boxplot (Figura 15). Donde vemos que el 25% de las mujeres tienen un grosor del pliegue cutáneo del tríceps menor de 25 y que solo un 25% tiene un grosor del pliegue cutáneo del tríceps mayor de 33. En el boxplot podemos apreciar como la mayoría de valores están alrededor de la media, pero con algunos outliers en ambos extremos.

```
##   0%  25%  50%  75% 100%
##    7   25   28   33   99
```

Estudiamos la forma de la variable Skin. Para ello miramos la media, sd, skewness y kurtosis.

```
## [1] "Media: 29.247"
```

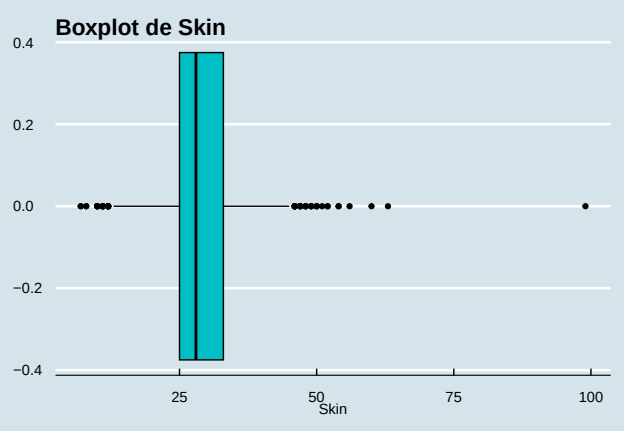

Figure 15: Boxplot de Skin

```
## [1] "Desviación típica: 8.9239"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 0.7603"
```

```
##
##  D'Agostino skewness test
##
## data:  pima$Skin
## skew = 0.76033, z = 7.77815, p-value = 7.36e-15
## alternative hypothesis: data have a skewness
```

Dados que skewness > 0 la distribución está sesgada a la derecha. Además el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada.

```
## [1] "Kurtosis: 7.8557"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pima$Skin
## kurt = 7.8557, z = 9.4860, p-value < 2.2e-16
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor de 0.05, por lo que la ditribución tiene colas pesadas, y por lo tanto no es normal. Podemos verlo mejor con un gráfico de densidad (Figura 16) donde observamos que la distribución presenta 2 picos, esto es consecuencia de la imputación de los valores perdidos. Esto se puede visualizar mejor en el gráfico de densidad separado por clase (Figura 17). Por otro lado, realizamos el test de Shapiro-Wilk para confirmar que la distribución no es normal y lo visualizamos con un QQ-plot (Figura 18).

```
##
##  Shapiro-Wilk normality test
##
## data:  pima$Skin
## W = 0.95267, p-value = 5.366e-15
```

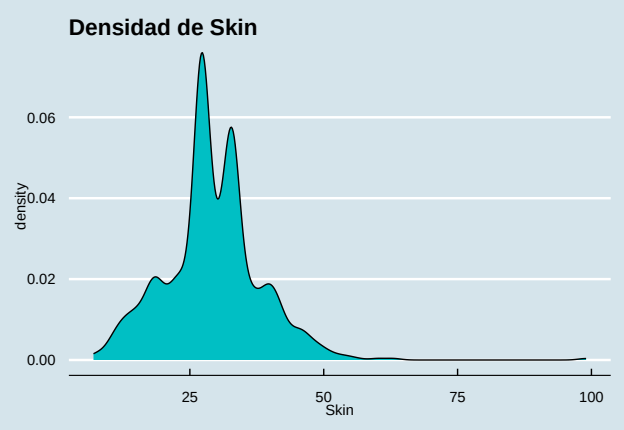

Figure 16: Densidad de Skin

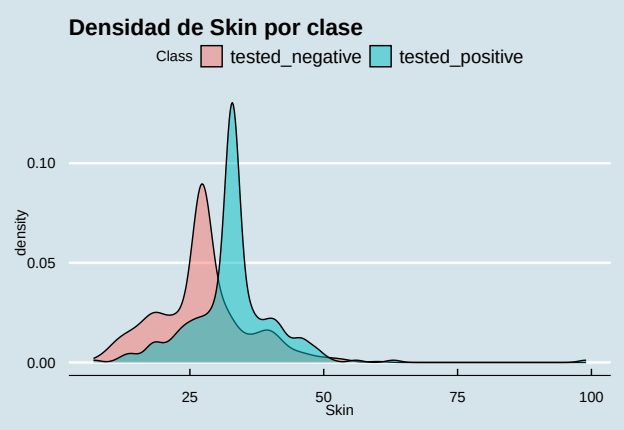

Figure 17: Densidad de Skin por clase

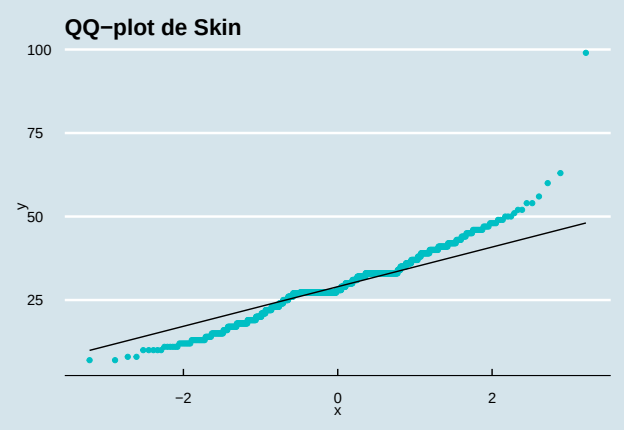

Figure 18: QQ-plot de Skin

Variable 'Insu'

Continuamos estudiando ahora la variable Insu. Esta pertenece al rango [0-846] y se refiere a la insulina sérica de 2 horas (mu U/ml). Para más detalle vemos un histograma (Figura 19). Se puede observar que la mayoría de los valores se encuentran entre 0 y 200, con un gran pico en 0 y con algunos valores a partir de 500 hasta 846. Esto puede indicar que hay valores perdidos. Dado mi poco conocimiento sobre el dataset y tras una búsqueda en internet, he encontrado que la insulina sérica de 2 horas podría ser 0 en casos de ayuno por lo que no podemos afirmar que los valores en 0 sean valores perdidos. Lo mismo con los valores del rango superior, a partir de 800 se considera hiperinsulinemia, por lo que tampoco podemos afirmar que sean valores perdidos.

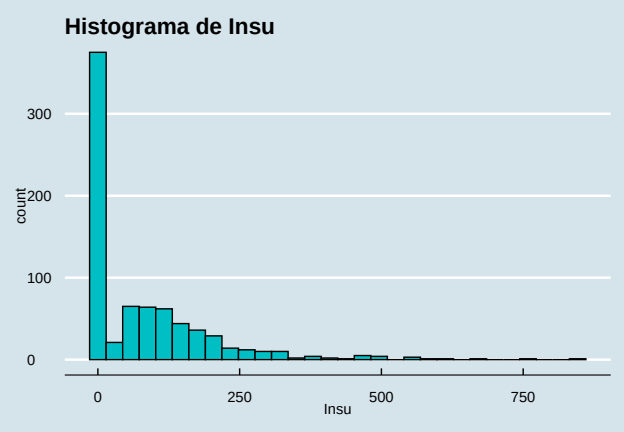

Figure 19: Histograma de Insu

Siguiendo con el análisis, vemos los cuantiles y lo visualizamos con un boxplot (Figura 20). Observamos que el 25% de las mujeres tienen un nivel de insulina de 0, y que solo un 25% tiene un nivel de insulina mayor de 127. Podemos ver que parece ser una distribución con valores atípicos en la parte superior, dejando una larga cola.

```
##     0%    25%    50%    75%   100%
##   0.00   0.00  30.50 127.25 846.00
```

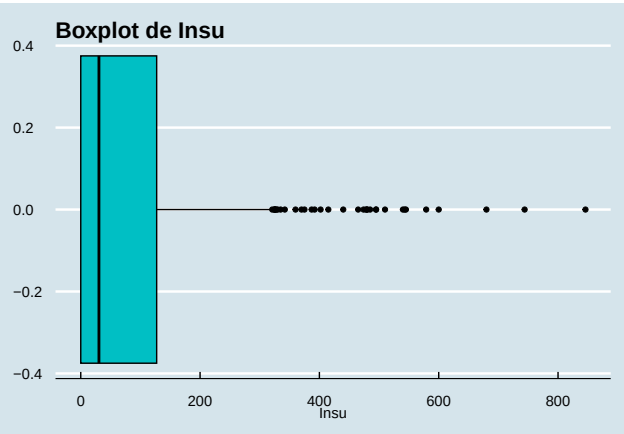

Figure 20: Boxplot de Insu

Estudiamos de seguido la forma de la variable Insu. Miramos la media, sd, skewness y kurtosis.

```
## [1] "Media: 30"
```

```
## [1] "Desviación típica: 115.244"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 2.2678"
```

```
##
##  D'Agostino skewness test
##
## data:  pima$Insu
## skew = 2.2678, z = 16.3881, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Dados que skewness > 0 la distribución está sesgada a la derecha. Además el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada a la derecha.

```
## [1] "Kurtosis: 10.1596"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pima$Insu
## kurt = 10.160, z = 10.909, p-value < 2.2e-16
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor que 0.05, por lo que podemos afirmar que la ditribución no es normal. Visualizamos la forma con un gráfico de densidad (Figura 21). Podemos ver que la distribución es bastante asimétrica con una cola en la parte superior muy pronunciada. Confirmamos la no normalidad con el test de Shapiro-Wilk y visualizamos un QQ-plot (Figura 22).

```
##
##  Shapiro-Wilk normality test
##
## data:  pima$Insu
## W = 0.72202, p-value < 2.2e-16
```

Variable 'Mass'

Ahora estudiamos la variable Mass. Esta pertece al rango [0-67.1] y se refiere al índice de masa corporal (peso en kg/(altura en m)^2). Para más detalle vemos un histograma (Figura 23). Se puede observar que la mayoría de los valores se encuentran entre 20 y 45, con un pequeño pico en 0. Esto puede indicar que hay valores perdidos.

Dada la naturaleza de la variable, no tiene sentido que haya valores en 0 dado que implicaría un peso de 0kg. Sustituimos estos valores por NA's e imputamos los valores perdidos con la media ya que la variable es bastante simétrica.

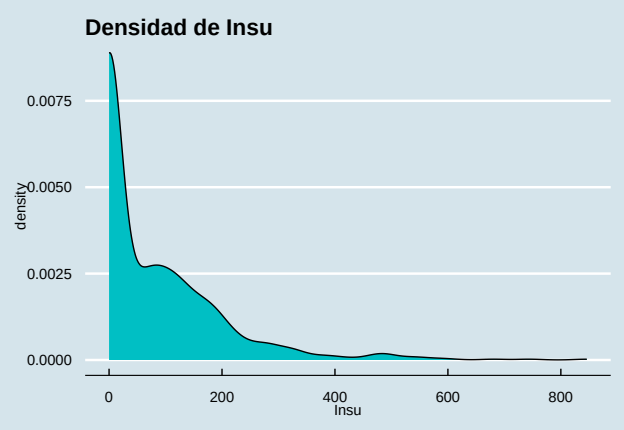

Figure 21: Densidad de Insu

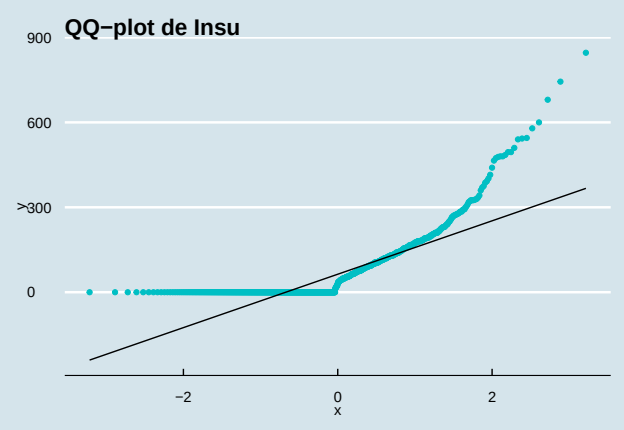

Figure 22: QQ-plot de Insu

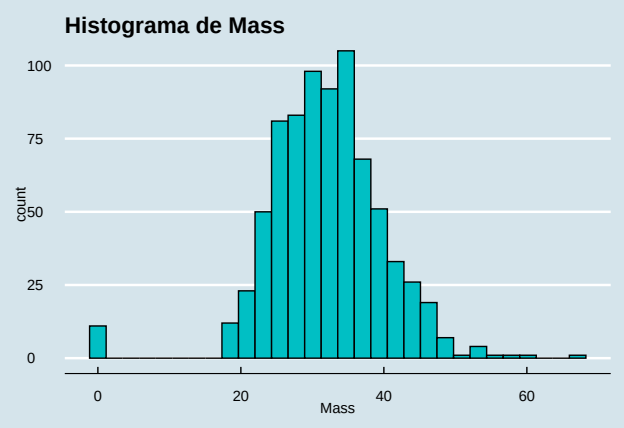

Figure 23: Histograma de Mass

```
## [1] "Resumen antes de imputar:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.20   27.50   32.30   32.46   36.60   67.10      11
```

```
## [1] "Resumen después de imputar:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.20   27.50   32.40   32.46   36.60   67.10
```

Estudiamos los cuantiles y lo visualizamos con un boxplot (Figura 24). Observamos que el 25% de las mujeres tienen un índice de masa corporal menor de 27.5 y que solo un 25% tiene un índice de masa corporal mayor de 36.6, teniendo una cola superior pronunciada.

```
##   0%  25%  50%  75% 100%
## 18.2 27.5 32.4 36.6 67.1
```

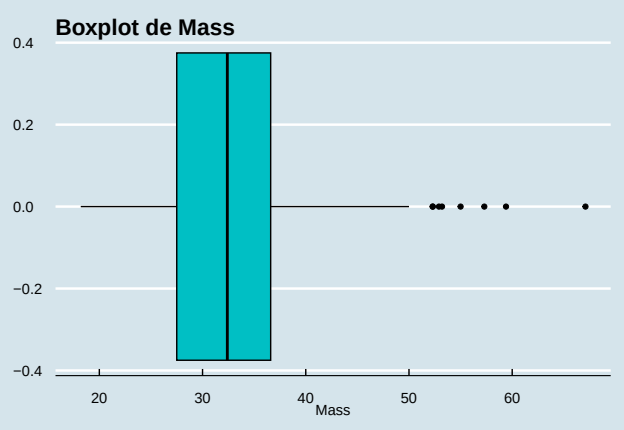

Figure 24: Boxplot de Mass

A continuación estudiamos la forma de la variable Mass. Miramos la media, sd, skewness y kurtosis.

```
## [1] "Media: 32"
```

```
## [1] "Desviación típica: 6.8752"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 0.5971"
```

```
##
##  D'Agostino skewness test
##
## data:  pima$Mass
## skew = 0.59708, z = 6.33556, p-value = 2.365e-10
## alternative hypothesis: data have a skewness
```

Dados que skewness > 0 la distribución está sesgada a la derecha. Además el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada a la derecha.

```
## [1] "Kurtosis: 3.9057"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pima$Mass
## kurt = 3.9057, z = 3.7992, p-value = 0.0001452
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor que 0.05, por lo que podemos afirmar que la ditribución no es normal. Visualizamos la forma de la variable con un gráfico de densidad (Figura 25). Podemos ver que la distribución es algo simétrica con una cola en la parte superior. Estudiamos la normalidad con el test de Shapiro-Wilk y lo visualizamos con un QQ-plot (Figura 26).

```
##
##  Shapiro-Wilk normality test
##
## data:  pima$Mass
## W = 0.97946, p-value = 6.526e-09
```

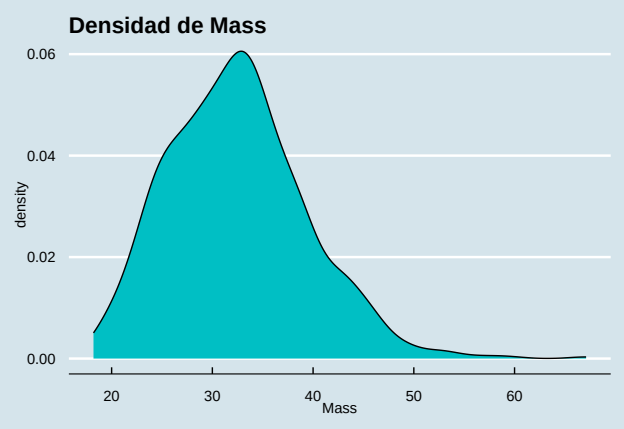

Figure 25: Densidad de Mass

Variable 'Pedi'

Seguimos el análisis con la variable Pedi. Esta pertece al rango [0.078-2.42] y se refiere a la función de pedigree de la diabetes. Para más detalle vemos un histograma (Figura 27). Se puede observar que la mayoría de los valores se encuentran entre 0 y 1.5, con una cola en la parte superior que alcanza hasta 2.5.

Estudiamos a continuación los cuantiles y lo visualizamos con un boxplot (Figura 28). Vemos que el 25% de las mujeres tienen un valor de Pedi menor de 0.24 y que solo un 25% tiene un valor de Pedi mayor de 0.63. Parece ser una distribución con valores atípicos en la parte superior.

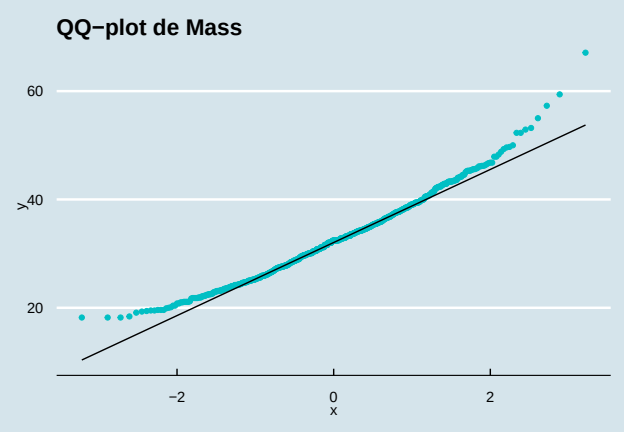

Figure 26: QQ-plot de Mass

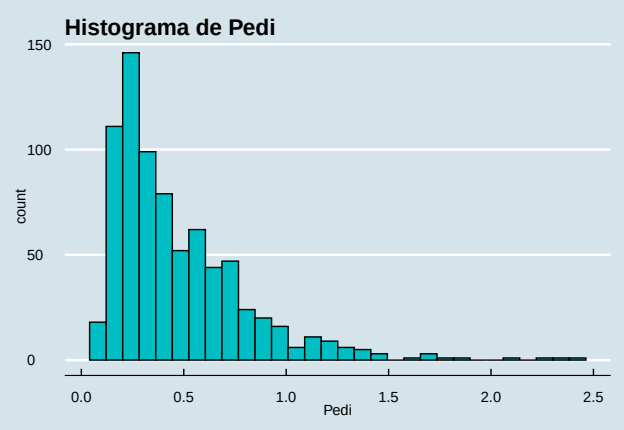

Figure 27: Histograma de Pedi

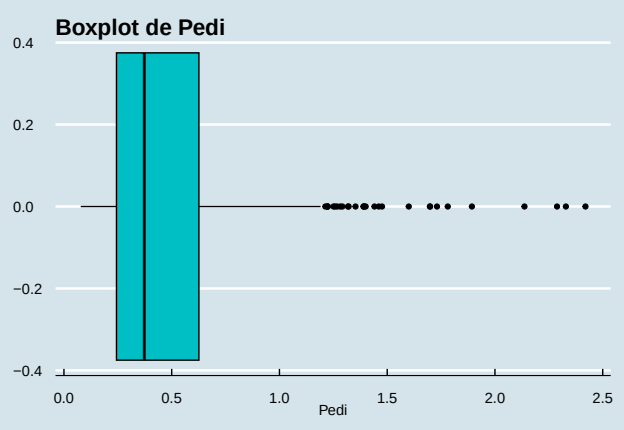

Figure 28: Boxplot de Pedi

```
##      0%     25%     50%     75%    100%
## 0.07800 0.24375 0.37250 0.62625 2.42000
```

Estudiamos a continuación la forma de la variable Pedi mirando la media, sd, skewness y kurtosis.

```
## [1] "Media: 0"
```

```
## [1] "Desviación típica: 0.3313"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 1.9162"
```

```
##
##  D'Agostino skewness test
##
## data:  pima$Pedi
## skew = 1.9162, z = 14.9069, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Dados que skewness > 0 la distribución está sesgada a la derecha. Además el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada a la derecha.

```
## [1] "Kurtosis: 8.5508"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pima$Pedi
## kurt = 8.5508, z = 9.9811, p-value < 2.2e-16
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor que 0.05, por lo que podemos afirmar que la ditribución no es normal. Visualizamos la forma de la variable con un gráfico de densidad (Figura 29). Podemos ver como la distribución es bastante asimétrica con una cola en la parte superior. Confirmamos la no normalidad con el test de Shapiro-Wilk y lo visualizamos con un QQ-plot (Figura 30).

```
##
##  Shapiro-Wilk normality test
##
## data:  pima$Pedi
## W = 0.83652, p-value < 2.2e-16
```

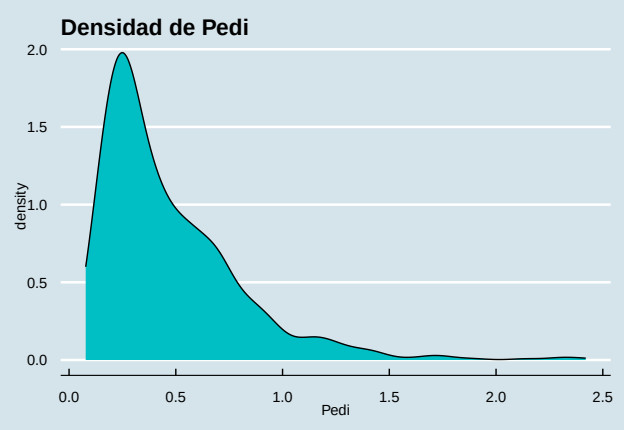

Figure 29: Densidad de Pedi

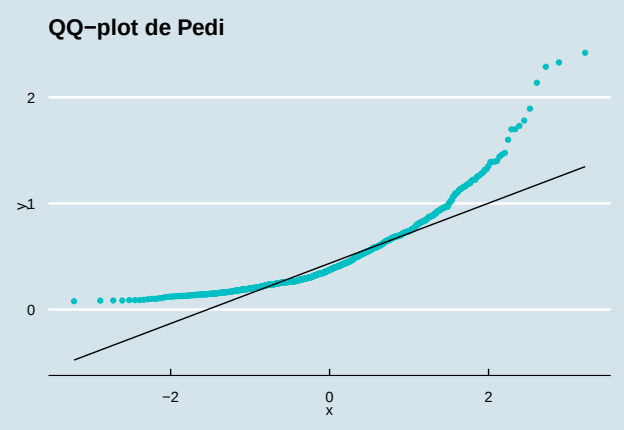

Figure 30: QQ-plot de Pedi

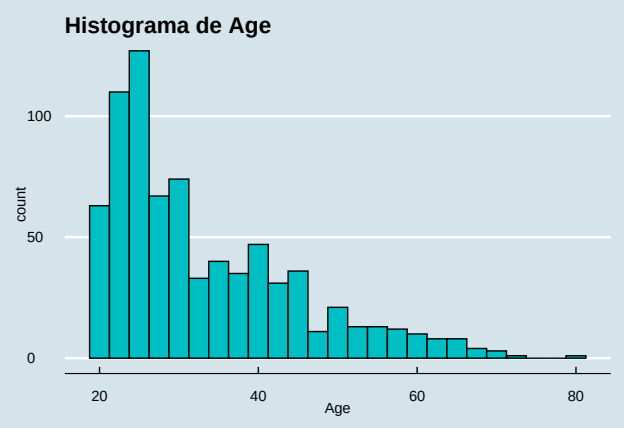

Figure 31: Histograma de Age

Variable 'Age'

Estudiamos, por último, la variable Age. Esta pertece al rango [21-81] y se refiere a la edad de la persona. Para más detalle vemos un histograma (Figura 31). Se puede observar que la mayoría de los valores se encuentran entre 20 y 40, con una cola en la parte superior que alcanza hasta 80.

Estudiamos a continuación los cuantiles y lo visualizamos con un boxplot (Figura 32). Vemos que el 25% de las mujeres tienen una edad menor de 24 y que solo un 25% tiene una edad mayor de 41. Parece ser una distribución con valores atípicos en la parte superior.

```
##   0%  25%  50%  75% 100%
##   21   24   29   41   81
```

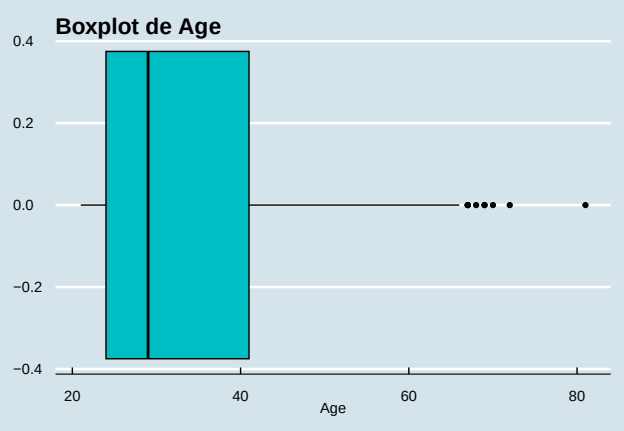

Figure 32: Boxplot de Age

Estudiamos la forma de la variable Age mirando la media, sd, skewness y kurtosis.

```
## [1] "Media: 29"
```

```
## [1] "Desviación típica: 11.7602"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 1.1274"
```

```
##
##  D'Agostino skewness test
##
## data:  pima$Age
## skew = 1.1274, z = 10.5504, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Dados que skewness > 0 la distribución está sesgada a la derecha. Además el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada a la derecha.

```
## [1] "Kurtosis: 3.6312"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pima$Age
## kurt = 3.6312, z = 2.9267, p-value = 0.003425
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor que 0.05, por lo que podemos afirmar que la ditribución no es normal. Visualizamos la forma de la variable con un gráfico de densidad (Figura 33). Podemos ver como la distribución es bastante asimétrica con una cola en la parte superior. Confirmamos la no normalidad con el test de Shapiro-Wilk y lo visualizamos con un QQ-plot (Figura 34).

```
##
##  Shapiro-Wilk normality test
##
## data:  pima$Age
## W = 0.87477, p-value < 2.2e-16
```

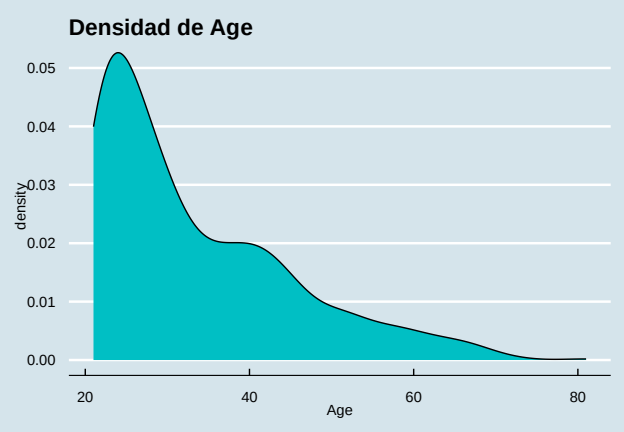

Figure 33: Densidad de Age

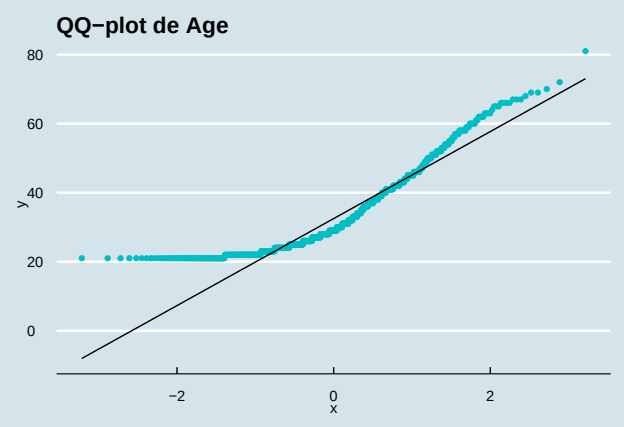

Figure 34: QQ-plot de Age

Análisis bivariable

Tras el análisis de cada variable por separado, haremos un análisis bivariable, donde veremos la relación entre las variables.

Correlación

Para ver la correlación entre las variables numéricas usaremos el método de Spearman pues las variables no siguen una distribución normal (ver Figura 35).

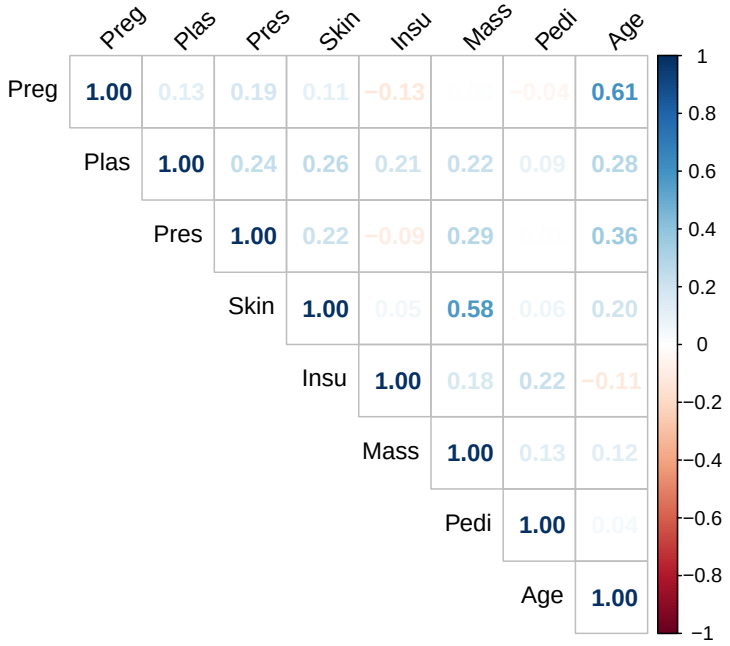

Figure 35: Correlación entre variables numéricas

Vemos que las variables que más correlación tienen, aunque no muy significativa, son Skin con Mass y Age con Preg. Lo cual tiene sentido pues la edad de la persona está relacionada con el número de embarazos ya que a mayor número de embarazos, necesariamente, mayor edad. Además, la variable Skin está relacionada con Mass pues a mayor masa corporal, mayor grosor de la piel.

Scatterplots

Para ver mejor las relaciones de las variables con mayor correlación vamos a hacer un scatterplot de cada una de estas parejas (ver Figura 36). Vemos que para la pareja Skin y Mass hay una relación lineal positiva entre ambas variables y los puntos se aconglomeran pero con una clara tendencia. Por otro lado, para Age y Preg vemos que hay una relación lineal positiva entre ambas variables pero los puntos se dispersan significativamente de la recta.

Relación con la variable Class

Ahora vamos a ver la relación entre las variables numéricas y la variable Class. Para ello hacemos un boxplot de cada variable numérica con respecto a la variable Class (Figura 37). Vemos que todas las variables, en general, indican que las mujeres con diabetes tienen valores más altos que las mujeres sin diabetes. Principalmente son interesantes las variables Plas y la variable Skin cuyos percentiles 25 en la clase positiva, son mayores que los percentiles 75 de la clase negativa.

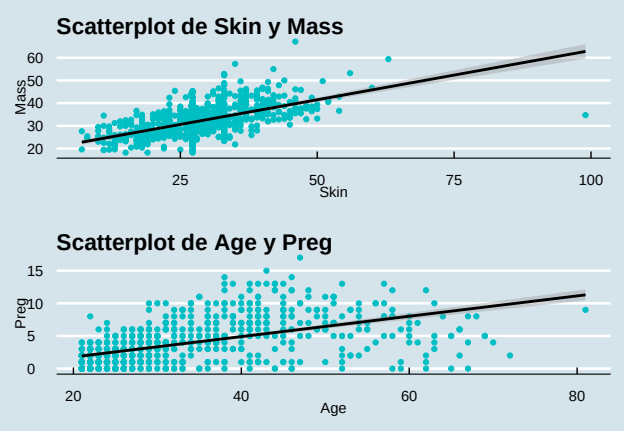

Figure 36: Scatterplots de las variables con mayor correlación

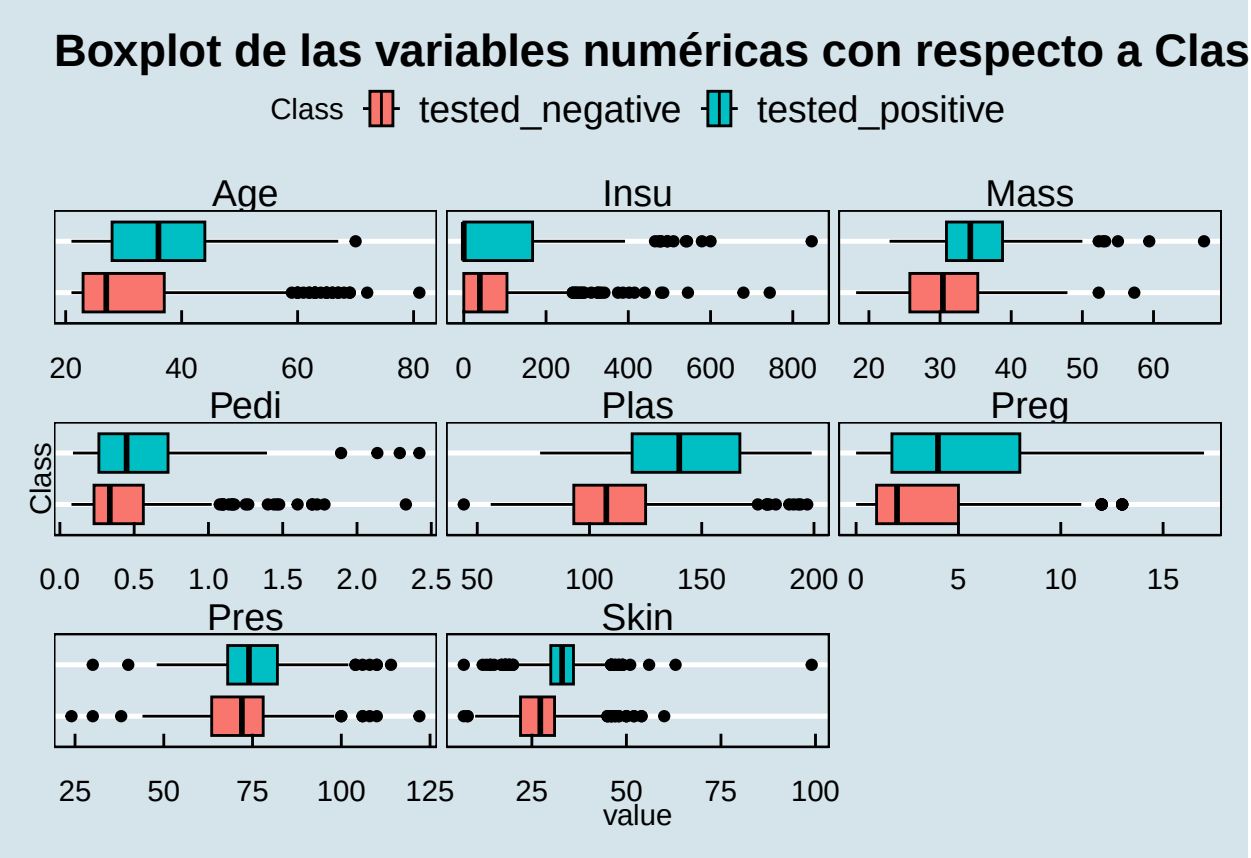

Figure 37: Boxplot de las variables numéricas con respecto a Class

Modelos de clasificación

Metología de trabajo

Para la realización de los modelos de clasificación, se pide usar las técnicas de K-NN, LDA y QDA. Para ello, primeramente se preparán los datos y se separa el dataset en train y test. Posteriormente, se entrena el modelo con los datos de train y se evalúa con los datos de test. Finalmente, se evalúa el modelo con los datos de test y se calcula la matriz de confusión y el accuracy.

Preparación de los datos

Para poder utilizar los modelos de clasificación, debemos preparar los datos. Para ello, primero separamos la variable Class del resto de variables, normalizamos los datos y separmos el conjunto de datos en train y test (80% y 20% respectivamente).

```
set.seed(123)

pima_data <- pima %>% dplyr::select(-Class)

pObj <- preProcess(pima_data, method=c('range'))
pima_scaled <- predict(pObj, pima_data)

test_size <- 0.2
train_size <- 1 - test_size

train_indices <- createDataPartition(pima$Class, p=train_size, list=FALSE)
train <- pima_scaled[train_indices, ]
test <- pima_scaled[-train_indices, ]
train_lbls <- pima[train_indices, ]$Class
test_lbls <- pima[-train_indices, ]$Class
```

K-NN (K-Nearest Neighbors)

K-NN se basa en la idea de que los puntos de datos con etiquetas similares se encuentran cerca unos de otros en el espacio. Por lo tanto elegir un correcto valor de K es crucial. Para ello, se usa la función 'train' de la librería 'caret' para encontrar el valor de K óptimo. Usaremos la función `trainControl()` para realizar validación cruzada repetida 10 veces y con 5 repeticiones.

```
## [1] "K óptimo: 27"
```

Vemos que el valor óptimo es 27 (Figura 38). Ahora, evaluamos el modelo con el conjunto de test y calculamos la matriz de confusión y el accuracy. Podemos ver que el accuracy es cercano al 80% pero nos interesa más el recall, que es la capacidad del modelo de encontrar todos los casos positivos. En este caso, el recall es de 0.58 relativamente menor que el accuracy general.

```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        tested_negative tested_positive
##   tested_negative              90              22
##   tested_positive              10              31
```

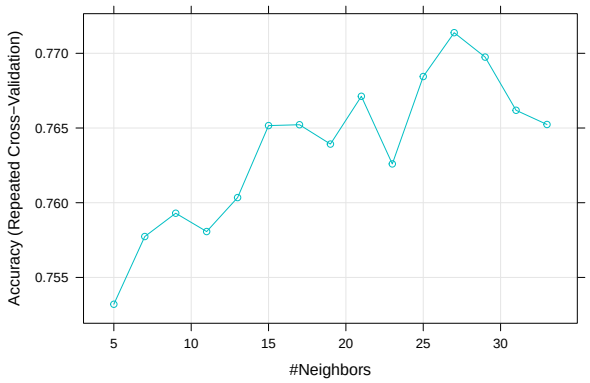

Figure 38: Accuracy en función de K

```
##
##                Accuracy : 0.7908
##                  95% CI : (0.7178, 0.8523)
##     No Information Rate : 0.6536
##     P-Value [Acc > NIR] : 0.0001499
##
##                   Kappa : 0.5122
##
##  Mcnemar's Test P-Value : 0.0518299
##
##               Precision : 0.7561
##                  Recall : 0.5849
##                      F1 : 0.6596
##              Prevalence : 0.3464
##          Detection Rate : 0.2026
##    Detection Prevalence : 0.2680
##       Balanced Accuracy : 0.7425
##
##        'Positive' Class : tested_positive
##
```

10 - fold cross validation

Ahora, vamos a realizar validación cruzada de 10 folds para ver si el accuracy es similar al obtenido anteriormente. El dataset ya viene con las particiones preparadas así que las cargamos en memoria y realizamos el entrenamiento y evaluación del modelo. Los resultados se pueden ver en la Figura 39 y estos son similares a los obtenidos anteriormente.

Discriminant Analysis

En este apartado usaremos LDA (Linear Discriminant Analysis) y su versión cuadrática QDA (Quadratic Discriminant Analysis). Para ello, usaremos la librería 'MASS' y su funciones 'lda()' y 'qda(). Ambas que las variables siguen una distribución normal, LDA que las matrices de covarianza son iguales para todas las clases y QDA permite que cada variable tenga su matriz de covarianza. Por lo tanto, debemos asegurarnos de que las variables estén normalizadas para cada clase. Visualizamos la distribución de cada variable para

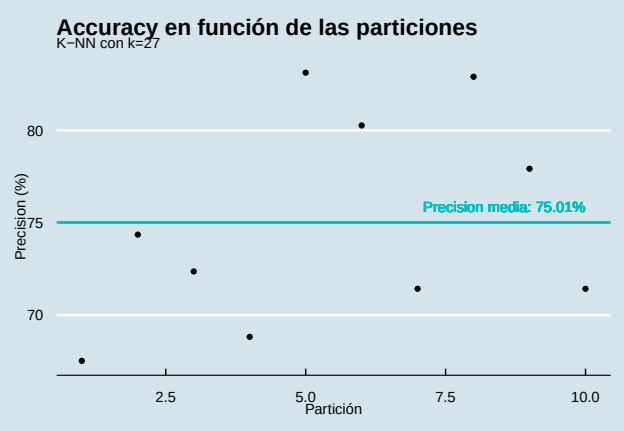

Figure 39: Accuracy en función de las particiones para KNN

cada clase (Figura 40). Vemos que hay que normalizar todas las variables por su media y desviación típica. Los resultados podemos verlos en la Figura 41.

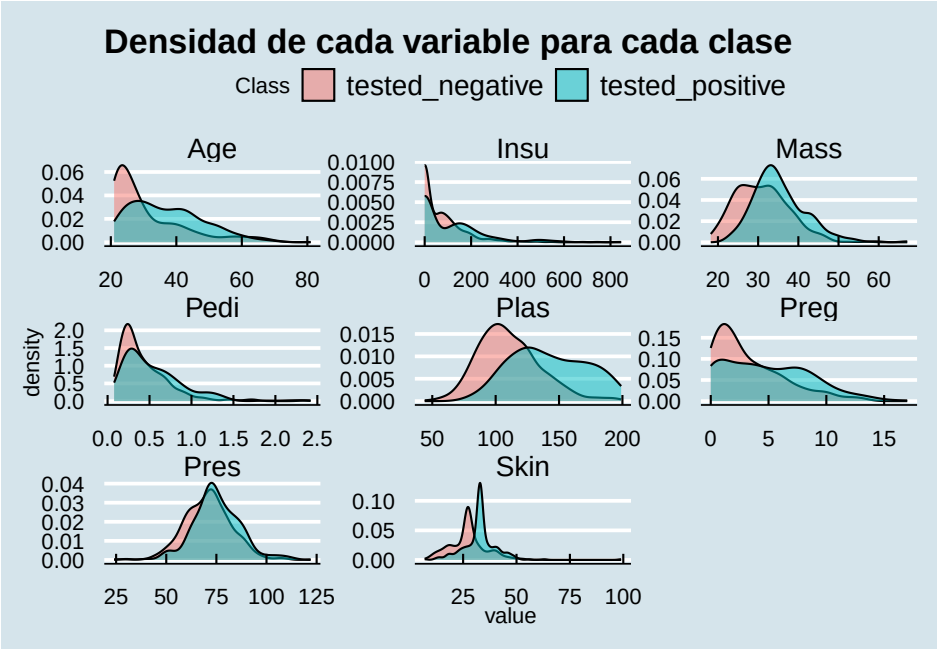

Figure 40: Distribución de cada variable para cada clase

```
## [1] "Datos normalizados:"
```

```
##       Preg              Plas              Pres              Skin        
##  Min.   :-1.3006   Min.   :-2.6989   Min.   :-3.9339   Min.   :-3.0747  
##  1st Qu.:-0.7616   1st Qu.:-0.7165   1st Qu.:-0.5983   1st Qu.:-0.4973  
##  Median :-0.2314   Median :-0.1096   Median :-0.0784   Median : 0.0000  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.5705   3rd Qu.: 0.6591   3rd Qu.: 0.5921   3rd Qu.: 0.3548  
##  Max.   : 3.2434   Max.   : 3.4911   Max.   : 4.2800   Max.   : 7.8050  
##       Insu              Mass              Pedi              Age        
```

```
##  Min.   :-0.7235   Min.   :-1.9506   Min.   :-1.2421   Min.   :-1.4649
##  1st Qu.:-0.6958   1st Qu.:-0.7515   1st Qu.:-0.7238   1st Qu.:-0.7355
##  Median :-0.4531   Median :-0.1212   Median :-0.3000   Median :-0.3591
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4052   3rd Qu.: 0.6108   3rd Qu.: 0.4635   3rd Qu.: 0.5409
##  Max.   : 6.8296   Max.   : 4.8089   Max.   : 6.3502   Max.   : 4.2691
##             Class
##  tested_negative:500
##  tested_positive:268
##
##
##
##
```

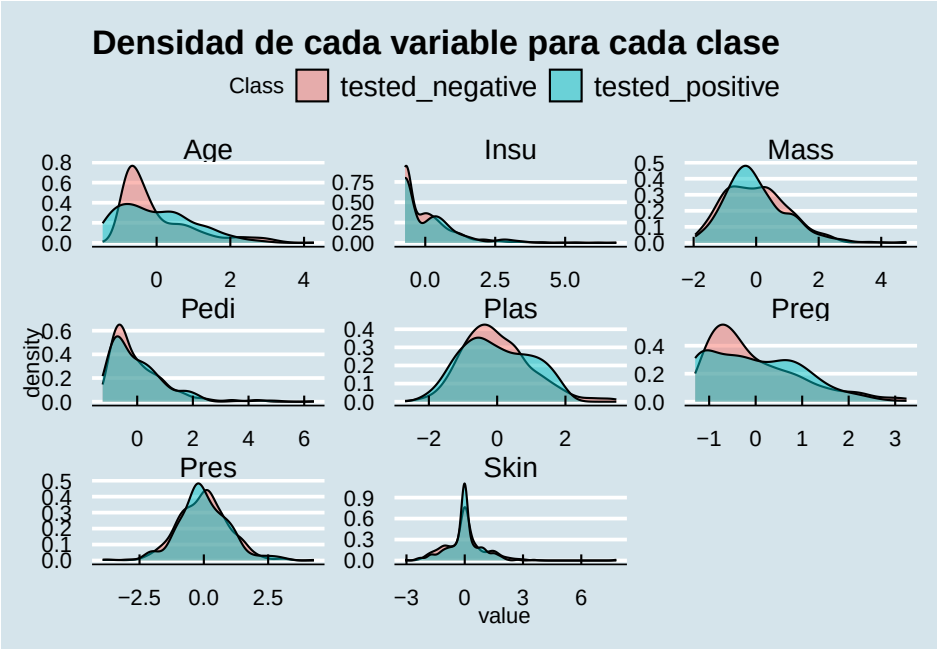

Figure 41: Distribución de cada variable para cada clase

Dadas las asunciones de LDA, comprobamos que se cumplen. Para ello, realizamos un test de Shapiro-Wilk para cada variable separada por clase (Figura 42) y un test de Barlett para la homocedasticidad (Figura 43). Vemos que ninguna variable sigue una distribución normal, aún así, dado que es un requisito y tanto LDA como QDA son suficientemente robustos sobre algunas desviaciones de la normalidad, los usaremos. Por otro lado, vemos que la homocedasticidad se incumple para las variables 'Preg', 'Plas', 'Pedi' e 'Insu'.

Por último nos queda separar el dataset normalizado en train y test (80% y 20% respectivamente) siguiendo la misma metodología que en la sección de preparación de datos.

Es necesario comprobar el balance de las clases ya que LDA y QDA sufren de un problema de sesgo hacia la clase mayoritaria. En este caso, vemos que el dataset está desbalanceado, por lo que haremos un submuestreo de la clase mayoritaria para equilibrar el dataset del entrenamiento.

```
## [1] "Balance de clases (previo):"
```

```
##
## tested_negative tested_positive
##             400             215
```

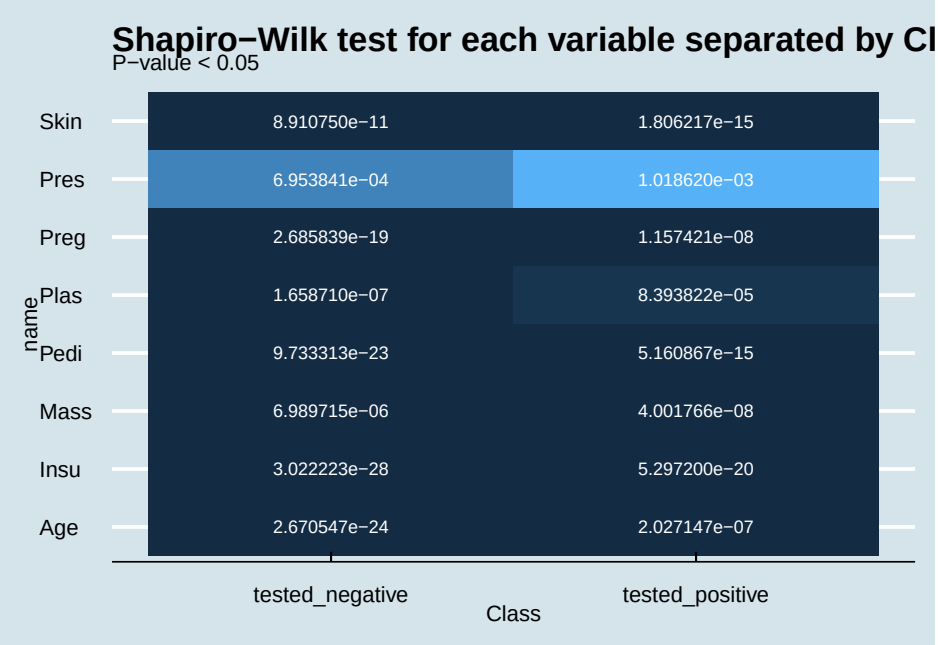

Figure 42: Test de Shapiro-Wilk para cada variable separada por clase

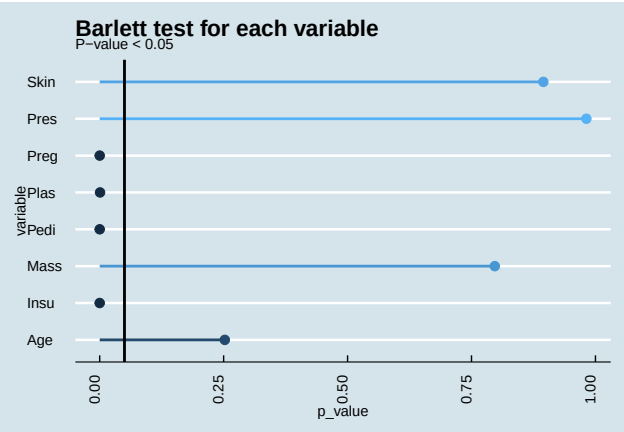

Figure 43: Test de Barlett para cada variable separada por clase

```
## [1] "Balance de clases (posterior):"

##
## tested_negative tested_positive
##             215             215
```

LDA

Como se ha mostrado, solo las variables Skin, Pres, Mass y Age no rechazan la homocedasticidad. Por ello, entrenaremos el modelo solo con estas variables. Para visualizar la clasificación usaremos la librería 'klaR'. Entrenaremos 2 modelos, uno con el dataset desbalanceado y otro con el dataset balanceado.

```
## [1] "Modelo con dataset desbalanceado:"

## Call:
## lda(train_lbls ~ Skin + Pres + Mass + Age, data = train)
##
## Prior probabilities of groups:
## tested_negative tested_positive
##       0.6504065       0.3495935
##
## Group means:
##                          Skin         Pres         Mass         Age
## tested_negative  0.010847726  0.009582014 -0.006401664 0.033640008
## tested_positive -0.009350619 -0.011060642  0.003943845 0.004477913
##
## Coefficients of linear discriminants:
##             LD1
## Skin -0.7137423
## Pres -0.4211953
## Mass  0.6711865
## Age  -0.4373904

## [1] "Modelo con dataset balanceado:"

## Call:
## lda(train_bal_lbls ~ Skin + Pres + Mass + Age, data = train_bal)
##
## Prior probabilities of groups:
## tested_negative tested_positive
##             0.5             0.5
##
## Group means:
##                          Skin        Pres         Mass         Age
## tested_negative  0.050489335  0.01119844 -0.075128018 0.096928668
## tested_positive -0.009350619 -0.01106064  0.003943845 0.004477913
##
## Coefficients of linear discriminants:
##             LD1
## Skin -0.7511871
## Pres -0.1577628
## Mass  0.8714400
## Age  -0.3555460
```

Evaluamos ambos modelos con el conjunto de test. No solo observamos la precisión sino el recall, ya que nos interesa que el modelo sea capaz de detectar la clase positiva. Podemos observar como en precisión el modelo con el dataset desbalanceado tiene mejor rendimiento (65% frente a 43%), pero esto se debe a que con el modelo desbalanceado se clasifican todos los casos como negativos (Recall = 0). En cambio, el modelo con el dataset balanceado tiene un recall de 0.434, por lo que es capaz de detectar la clase positiva. Visualizamos la clasificación por cada conjunto de 2 variables en la Figura 44.

```
## [1] "Modelo con dataset desbalanceado:"
```

```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        tested_negative tested_positive
##   tested_negative             100              53
##   tested_positive               0               0
##
##                Accuracy : 0.6536
##                  95% CI : (0.5725, 0.7286)
##     No Information Rate : 0.6536
##     P-Value [Acc > NIR] : 0.5373
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 9.148e-13
##
##               Precision :     NA
##                  Recall : 0.0000
##                      F1 :     NA
##              Prevalence : 0.3464
##          Detection Rate : 0.0000
##    Detection Prevalence : 0.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : tested_positive
##
```

```
## [1] "Modelo con dataset balanceado:"
```

QDA

En este apartado usaremos QDA para clasificar los datos. Al igual que en el apartado anterior, entrenaremos 2 modelos, uno con el dataset desbalanceado y otro con el dataset balanceado. Siendo que QDA no requiere que las variables cumplan la homocedasticidad, usaremos todas las variables para entrenar el modelo.

```
## [1] "Modelo con dataset desbalanceado:"
```

```
## Call:
## qda(train_lbls ~ ., data = train)
##
## Prior probabilities of groups:
## tested_negative tested_positive
```

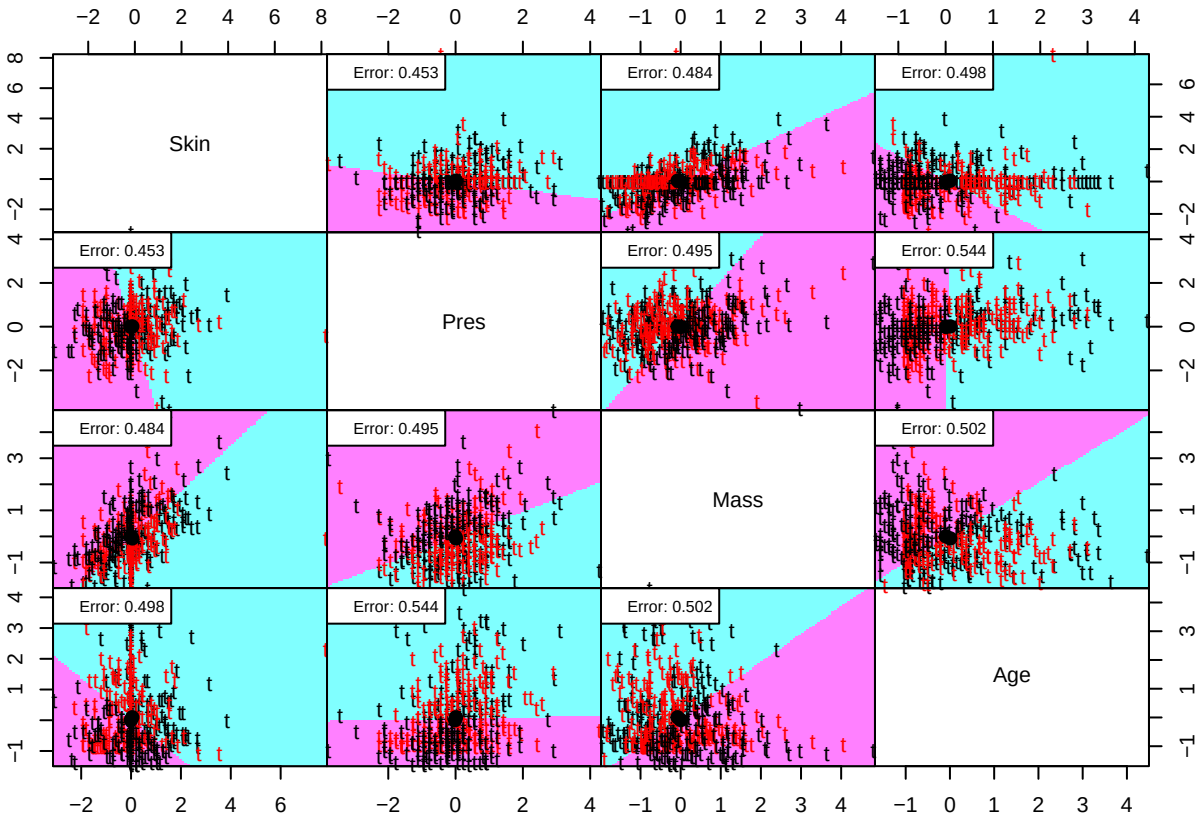

Figure 44: Visualización de la clasificación del modelo LDA

```
##       0.6504065       0.3495935
##
## Group means:
##                         Preg         Plas         Pres         Skin         Insu
## tested_negative  0.01723461 0.0002035090  0.009582014  0.010847726  0.002710759
## tested_positive -0.05982274 0.0004369632 -0.011060642 -0.009350619 -0.012851278
##                         Mass       Pedi         Age
## tested_negative -0.006401664 0.01076950 0.033640008
## tested_positive  0.003943845 0.04428763 0.004477913
```

```
## [1] "Modelo con dataset balanceado:"
```

```
## Call:
## qda(train_bal_lbls ~ ., data = train_bal)
##
## Prior probabilities of groups:
## tested_negative tested_positive
##             0.5             0.5
##
## Group means:
##                         Preg         Plas        Pres         Skin        Insu
## tested_negative  0.002975206 0.0120453171  0.01119844  0.050489335  0.05484183
## tested_positive -0.059822742 0.0004369632 -0.01106064 -0.009350619 -0.01285128
##                         Mass       Pedi         Age
## tested_negative -0.075128018 0.07003042 0.096928668
## tested_positive  0.003943845 0.04428763 0.004477913
```

Evaluamos ambos modelos con el conjunto de test. Al igual que en el apartado anterior, no solo observamos la precisión sino el recall, ya que nos interesa que el modelo sea capaz de detectar la clase positiva. Podemos observar como en precisión el modelo con el dataset desbalanceado tiene mejor rendimiento (64% frente a 50%), pero esto se debe a que con el modelo desbalanceado se clasifican casi todos los casos como negativos (Recall = 0.09). En cambio, el modelo con el dataset balanceado tiene un recall de 0.4, por lo que es capaz de detectar la clase positiva. Visualizamos la clasificación por cada conjunto de 2 variables en la Figura 45.

```
## [1] "Modelo con dataset desbalanceado:"
```

```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        tested_negative tested_positive
##   tested_negative              93              48
##   tested_positive               7               5
##
##                Accuracy : 0.6405
##                  95% CI : (0.5591, 0.7164)
##     No Information Rate : 0.6536
##     P-Value [Acc > NIR] : 0.667
##
##                   Kappa : 0.0297
##
##  Mcnemar's Test P-Value : 6.906e-08
##
##               Precision : 0.41667
```

```
##                  Recall : 0.09434
##                      F1 : 0.15385
##              Prevalence : 0.34641
##          Detection Rate : 0.03268
##    Detection Prevalence : 0.07843
##       Balanced Accuracy : 0.51217
##
##        'Positive' Class : tested_positive
##
```

```
## [1] "Modelo con dataset balanceado:"
```

```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        tested_negative tested_positive
##   tested_negative              55              32
##   tested_positive              45              21
##
##                Accuracy : 0.4967
##                  95% CI : (0.415, 0.5786)
##     No Information Rate : 0.6536
##     P-Value [Acc > NIR] : 1.0000
##
##                   Kappa : -0.0508
##
##  Mcnemar's Test P-Value : 0.1715
##
##               Precision : 0.3182
##                  Recall : 0.3962
##                      F1 : 0.3529
##              Prevalence : 0.3464
##          Detection Rate : 0.1373
##    Detection Prevalence : 0.4314
##       Balanced Accuracy : 0.4731
##
##        'Positive' Class : tested_positive
##
```

10 - fold Cross Validation

En este apartado usaremos 10-fold cross validation para evaluar el rendimiento de los modelos LDA y QDA. Para ello, usaremos las particiones provistas en el dataset. Se visualizan los resultados por partición de LDA y QDA en las Figuras 46 y 47 respectivamente.

Anexo : Código de la práctica

Dado que esto se ha realizado en un archivo de RMarkdown, se puede ver el código de la práctica en el archivo `pima-cladificacion.Rmd` que se encuentra en el repositorio de GitHub. Click aquí

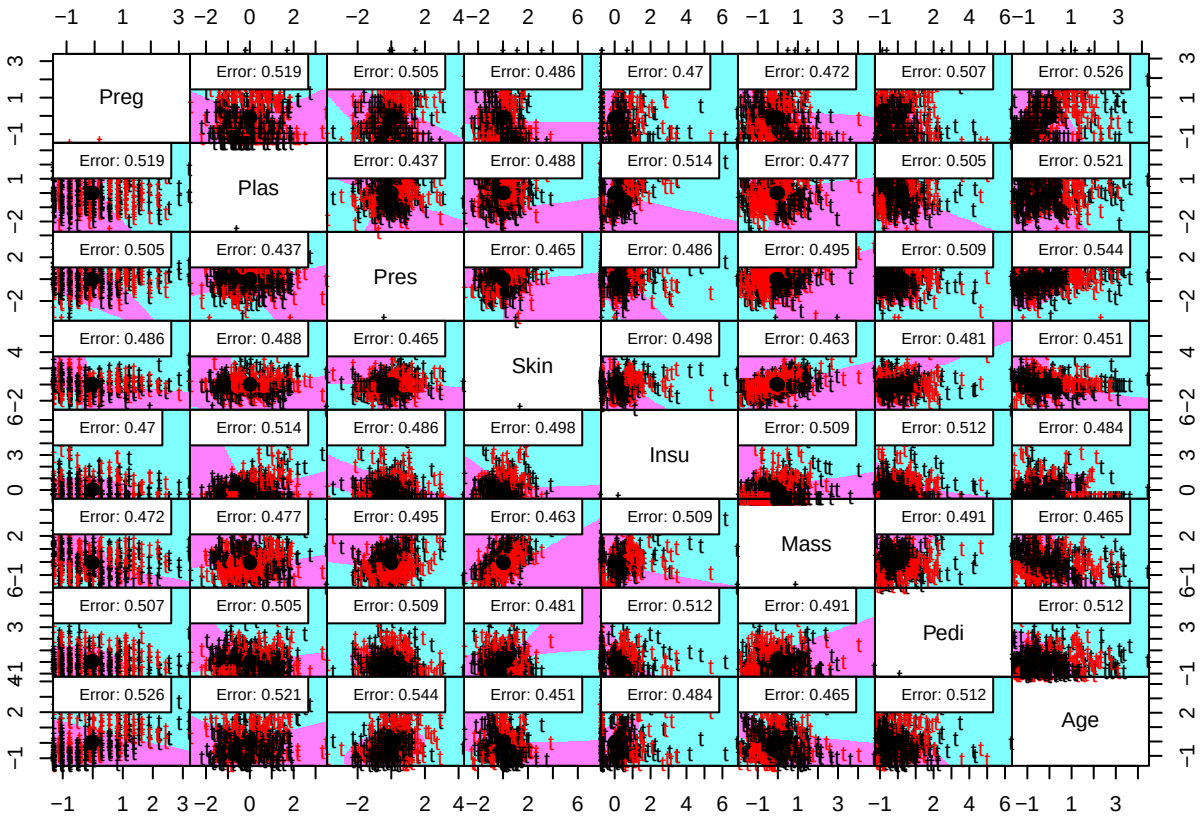

Figure 45: Visualización de la clasificación del modelo QDA

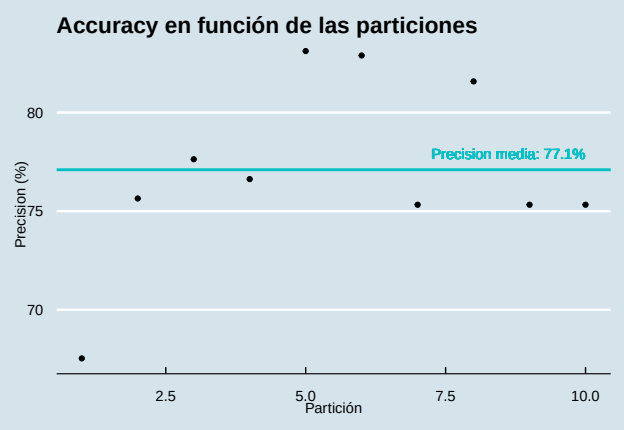

Figure 46: Accuracy en función de las particiones para LDA

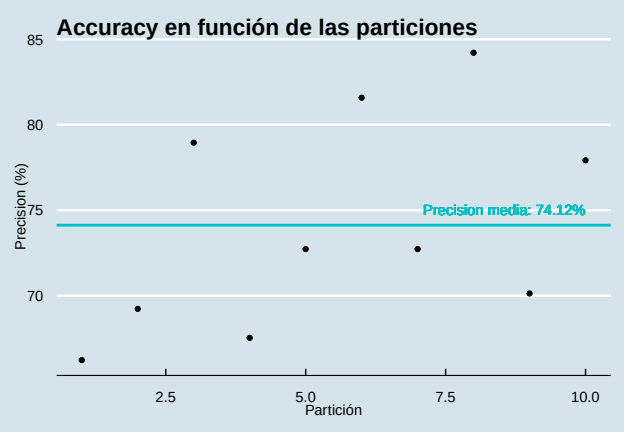

Figure 47: Accuracy en función de las particiones para QDA