Práctica 2

Juan Pasaret y David Reyes

2023-06-01

1 Descripción del dataset

El dataset seleccionado para la práctica se denomina **Heart Attack Analysis & Prediction Dataset** (https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset)**.**

Es un conjunto de datos de análisis y predición de ataques cardiados. Está diseñado para la práctica de análisis estadístico, ya que el número de muestras es pequeño.

El objetivo de la práctica es predecir las posibilidades de sufrir un ataque al corazón y estudiar qué factores son los más influyentes en esta dolencia.

El conjunto está formado por dos ficheros que contienen las siguientes variables (La descripción de las variables ha sido realizada por la autora en la documentación de kaggle):

- **age**: Edad del paciente
- **sex** : Sexo del paciente
- **cp** : Tipo de dolor torácico
  - Valor 1: typical angina (angina típica)
  - Valor 2: atypical angina (angina atípica)
  - Valor 3: no-anginal pain (dolor no anginoso)
  - Valor 4: asymptomatic (asintomático)
- **trtbps**: presión arterial en reposo (en mm Hg)

- **chol**: colesterol en mg/dl obtenido a través del sensor BMI
- **fbs**: (azúcar en sangre en ayunas > 120 mg/dl) (1 = verdadero; 0 = falso)
- **restecg** : resultados electrocardiográficos en reposo
    - Valor 0: normal
    - Valor 1: tener anomalías en la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST > 0,05 mV)
    - Valor 2: mostrar hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes
- **thalachh**: frecuencia cardíaca máxima alcanzada
- **exng**: angina inducida por el ejercicio (1 = sí; 0 = no)
- **oldpeak**: Pico anterior
- **slp**: Pendiente del segmento ST en el pico del ejercicio
- **caa**: número de vasos principales (0-3)
- **thall**: frecuencia cardíaca máxima alcanzada
- **output**:
    - 0= enfermedad cardíaca diagnosticada
    - 1= enfermedad cardíaca diagnosticada

1.1 Carga del fichero

```
# carga de datos
heart <- read.csv(file = 'heart.csv')
```

1.2 Tipos de variables

Verificamos la estructura del juego de datos principal. Vemos el número de columnas que tenemos y ejemplos de los contenidos de las filas.

```
structure = str(heart)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Vemos que tenemos **14** variables y **303** registros.

A continuación vemos una muestra:

```
head(heart, 5)
```

```
##   age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output
## 1  63   1  3    145  233   1       0      150    0     2.3   0   0     1      1
## 2  37   1  2    130  250   0       1      187    0     3.5   0   0     2      1
## 3  41   0  1    130  204   0       0      172    0     1.4   2   0     2      1
## 4  56   1  1    120  236   0       1      178    0     0.8   2   0     2      1
## 5  57   0  0    120  354   0       1      163    1     0.6   2   0     2      1
```

Vamos a factorizar las variables categóricas. Las variables numéricas no categóricas se transforman en variables numéricas. Se factoriza igualmente la variable output, que se utilizará para evaluar la precisión de nuestro modelo.

Comprobamos los transformaciones realizadas:

Variable	Clase original	Clase modificada
age	integer	integer
sex	integer	factor
cp	integer	factor
trtbps	integer	numeric
chol	integer	numeric
fbs	integer	factor
restecg	integer	factor
thalachh	integer	numeric
exng	integer	factor
oldpeak	numeric	numeric
slp	integer	factor
caa	integer	factor
thall	integer	factor
output	integer	factor

Vemos un sumario de los datos del conjunto después de las transformaciones iniciales:

```
summary(heart)
```

```
##       age            sex                      cp          trtbps
##  Min.   :29.00   Female: 96   Tipical angina  :143   Min.   : 94.0
##  1st Qu.:47.50   Male  :207   Atypical angina : 50   1st Qu.:120.0
##  Median :55.00                Non-anginal pain: 87   Median :130.0
##  Mean   :54.37                Asymptomatic    : 23   Mean   :131.6
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##       chol          fbs                               restecg       thalachh
##  Min.   :126.0   False:258   Normal                       :147   Min.   : 71.0
##  1st Qu.:211.0   True : 45   ST-T wave abnormality        :152   1st Qu.:133.5
##  Median :240.0               Left ventricular hypertrophy:  4   Median :153.0
##  Mean   :246.3                                                  Mean   :149.6
##  3rd Qu.:274.5                                                  3rd Qu.:166.0
##  Max.   :564.0                                                  Max.   :202.0
##   exng        oldpeak              slp      caa                   thall
##  No :204   Min.   :0.00   unsloping  : 21   0:175   null              :  2
##  Yes: 99   1st Qu.:0.00   flat       :140   1: 65   fixed defect      : 18
##            Median :0.80   downsloping:142   2: 38   normal            :166
##            Mean   :1.04                     3: 20   reversable defect:117
##            3rd Qu.:1.60                     4:  5
##            Max.   :6.20
##         output
##  no disease:138
##  disease   :165
##
##
##
##
```

2 Limpieza de datos

2.1 Valores vacíos

Se comprueba que no hay ninguna columna vacía

```
print(sum(is.na(heart)))
```

```
## [1] 0
```

2.2 Duplicados

Se observa que hay 1 duplicado, que se decide quitar del dataset.

```
sum(duplicated(heart))
```

```
## [1] 1
```

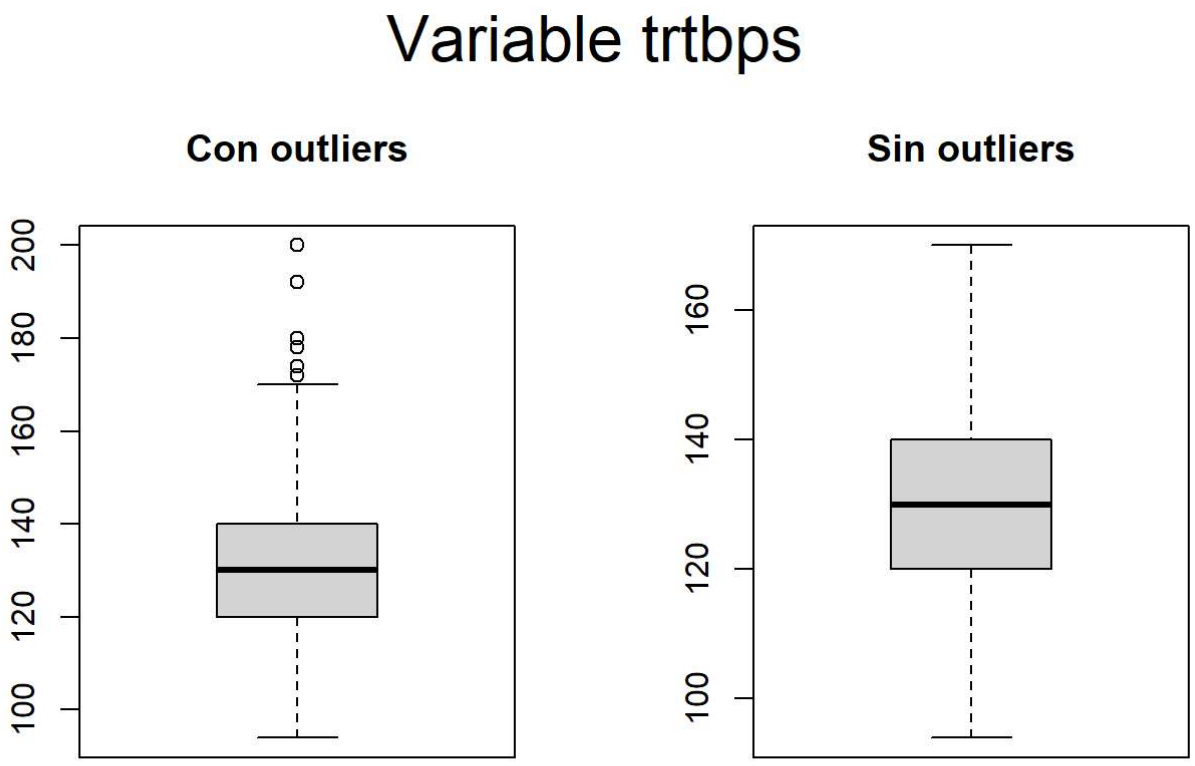
```
heart<-heart[!duplicated(heart),]
```

2.3 Identificación y gestión de valores extremos

Vamos a identificar los valores extremos. Para ello vamos a tomar el criterio de considerar extremos a aquellos valores alejados 1,5 veces la diferencia entre el primer y el tercer cuartil:

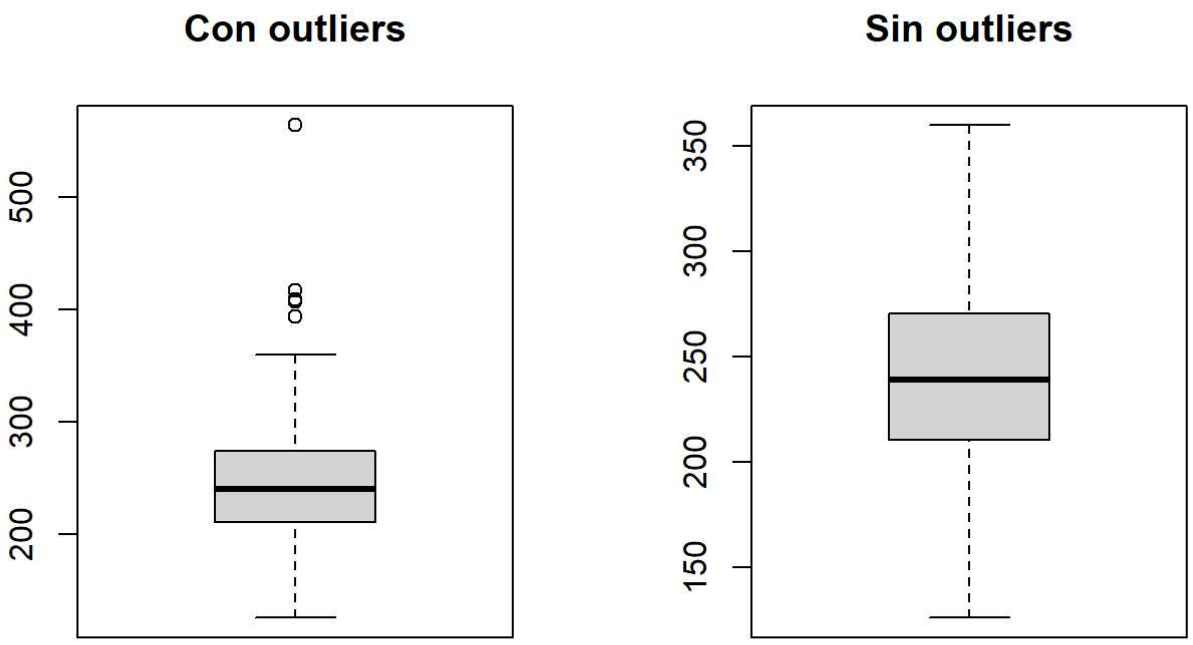
Se identifican valores extremos en las siguientes variables numéricas:
- **trtbps** : filas 9, 102, 111, 204, 224, 242, 249, 261, 267
- **chol** : filas 29, 86, 97, 221, 247
- **oldpeak** : filas 102, 205, 222, 251, 292

Se decide quitar las observaciones extremas del dataset. Ejemplo con la variable **trtbps.** No tenemos conocimiento del motivo de que dicha medidas tengan esos valores tan extremos, pero son claramente anómalos, ya que unas pulsasiones de 140 ya se consideran extremadamente peligrosas, y hablamos de eliminar pulsaciones mayores de 172.:



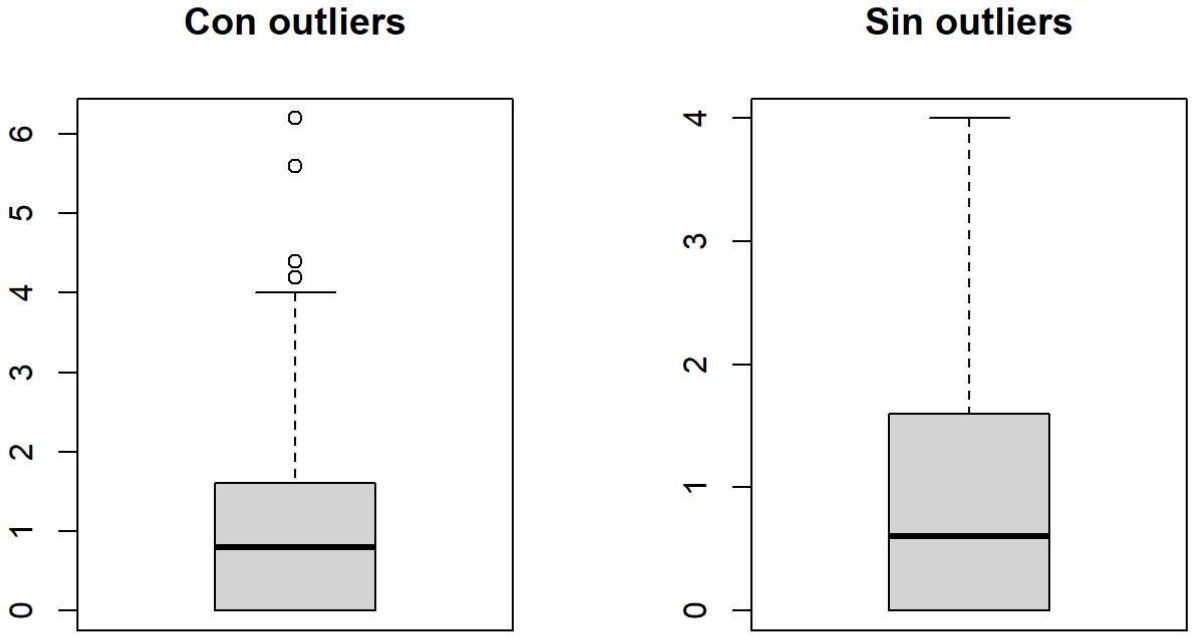
De igual manera actuamos con la variable chol (colesterol).

Variable chol



Y con la variable oldpeak.

Variable oldpeak



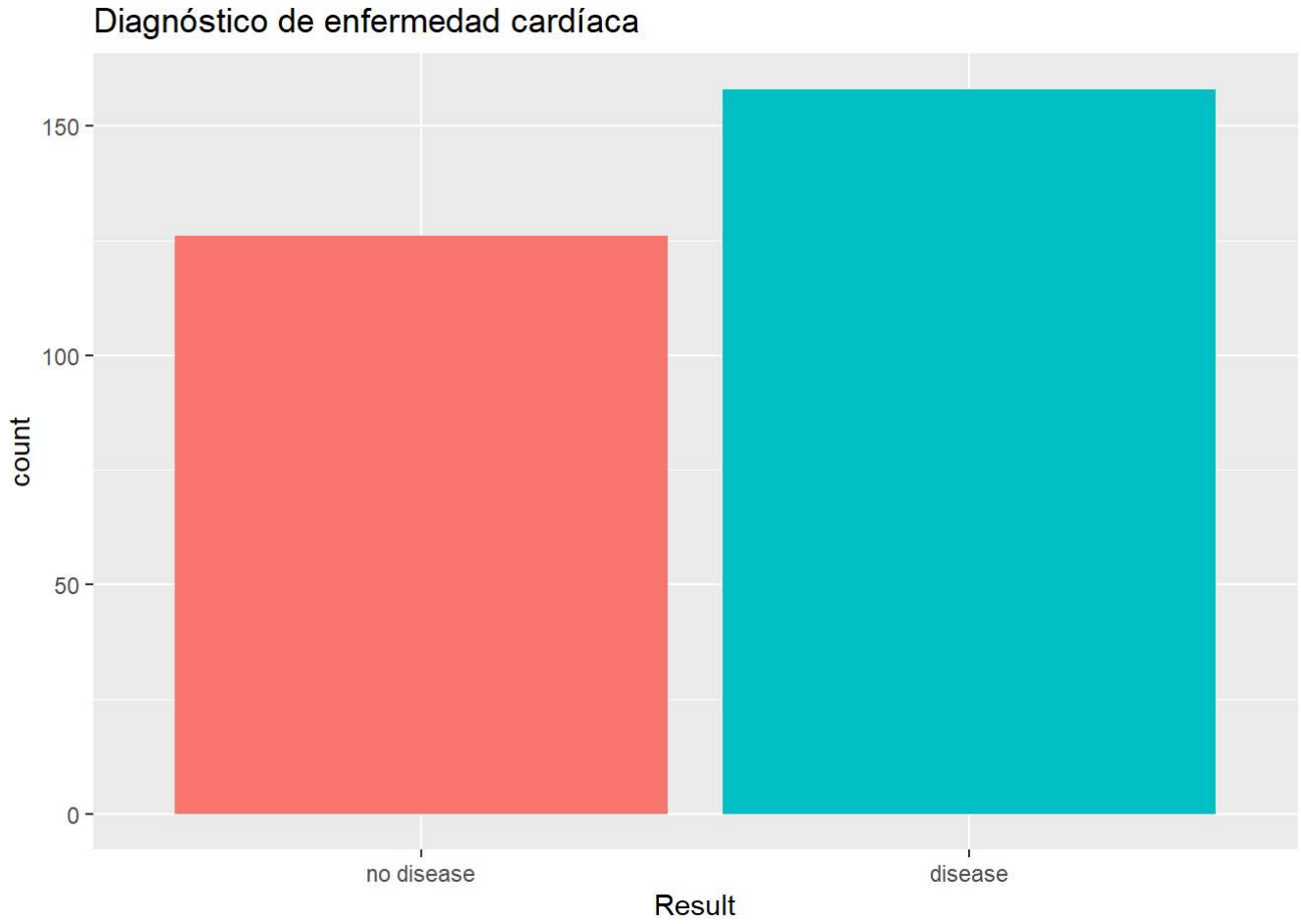
Para fundamentar esta actuaciones, deberíamos tener información sobre a la forma en la que se han llevado a cabo las medidas para poder comprobar que realmente se trata de medidas anómalas.

3 Análisis de los datos

3.1 Análisis estadístico descriptivo

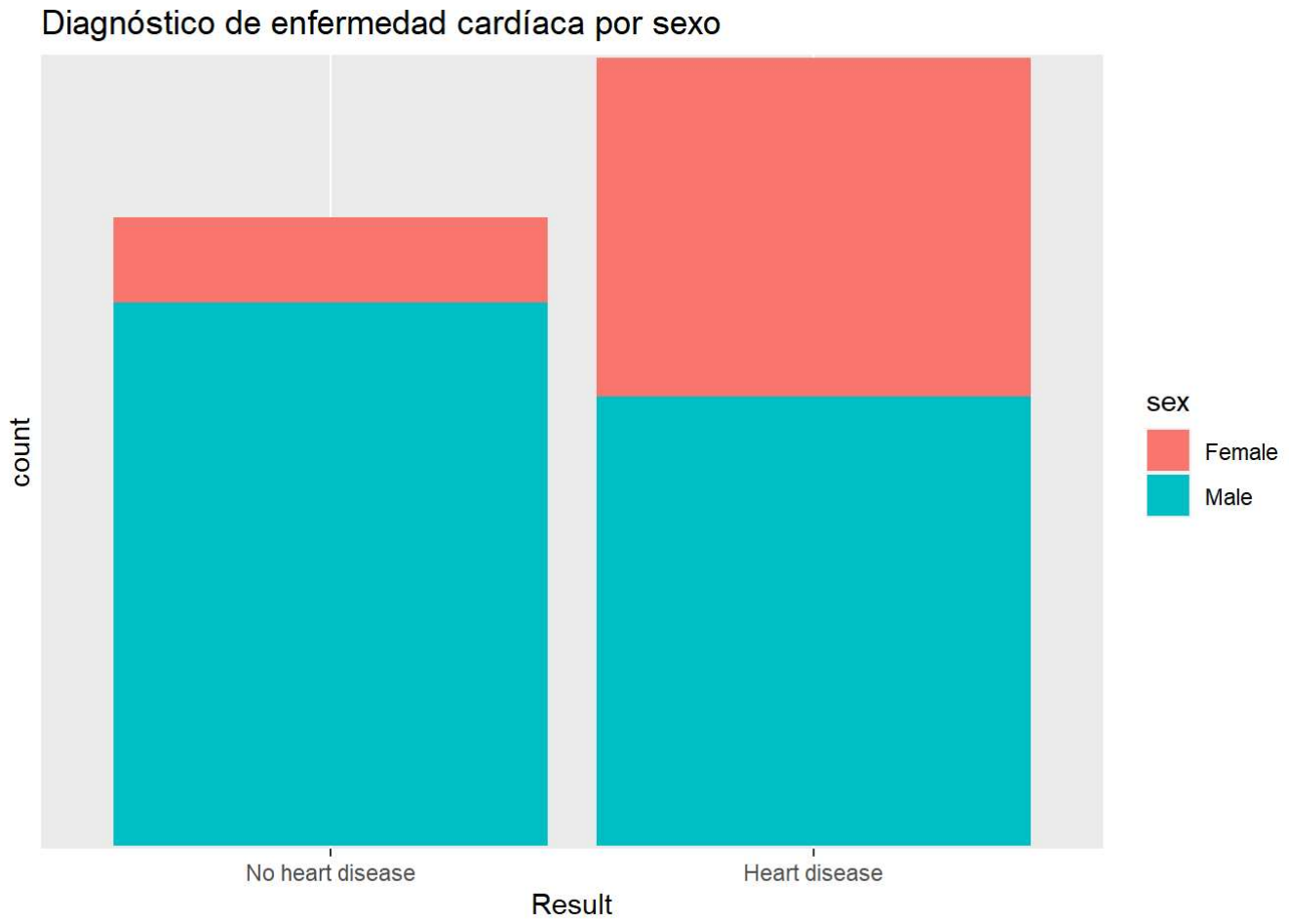
Este análisis empieza por la exploración del conjunto de observaciones en base al diagnóstico obtenido de cada paciente. Se observa un total de 126 pacientes sin enfermedad cardíaca contra 158 pacientes diagnosticados con enfermedad cardíaca:

```
##
## no disease    disease
##        126        158
```



Sin embargo, si se analizan este parámetro en función del sexo (nuestro objecto de estudio), se observa una desproporción en la proporción de mujeres diagnosticadas con dolencia cardíaca frente a los hombres:

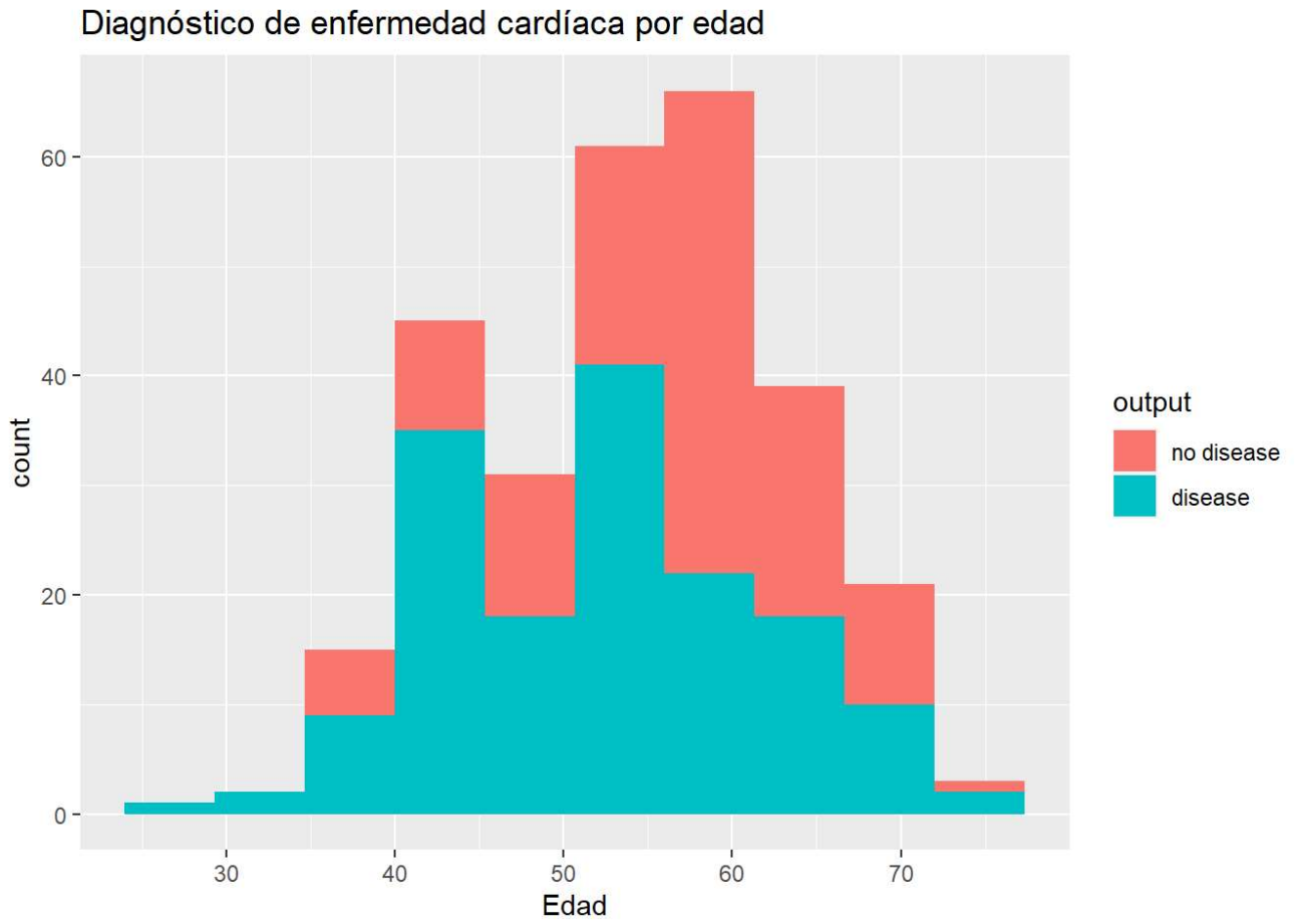
```
##
##              Female Male
##   no disease     17  109
##   disease        68   90
```



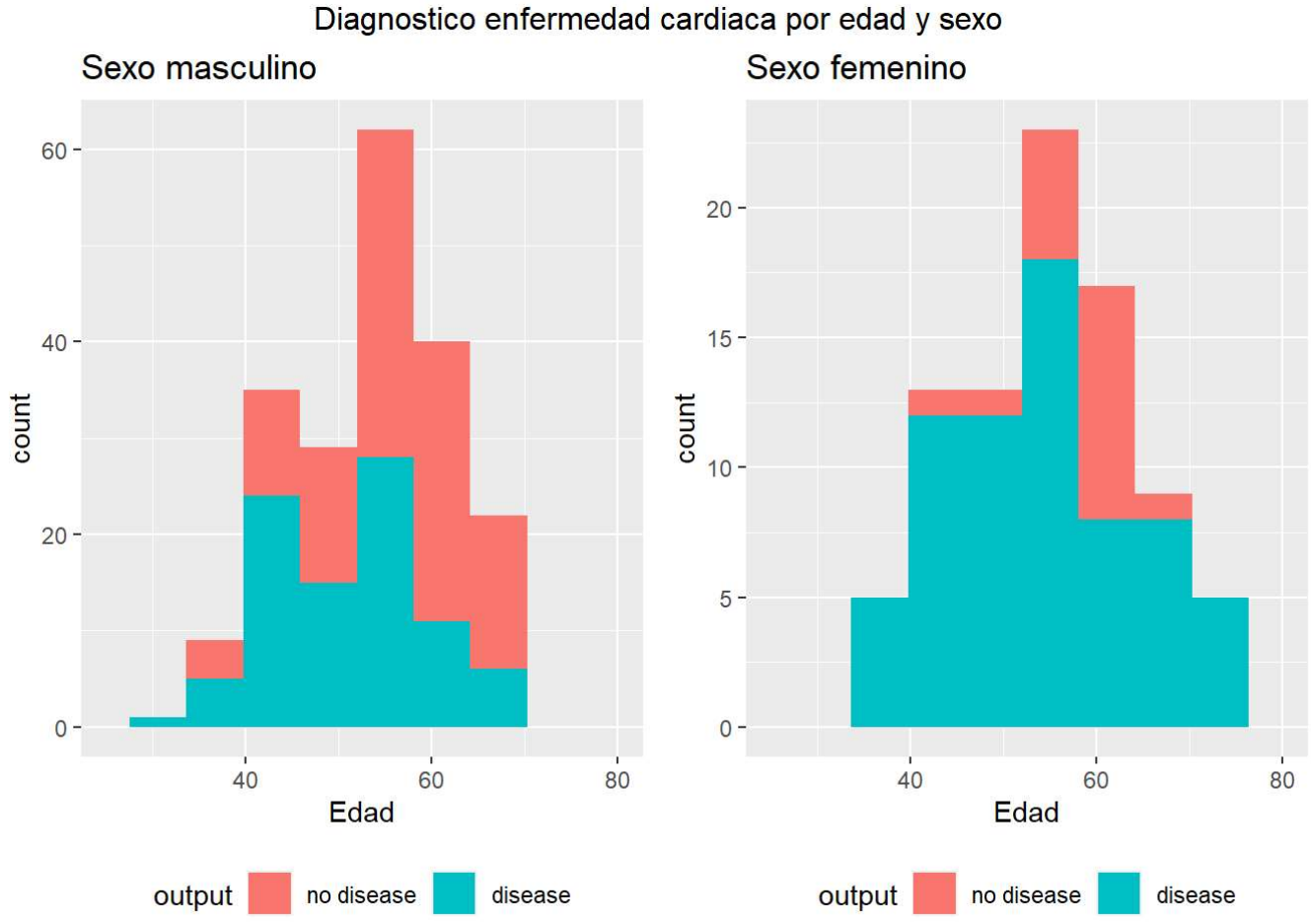
El análisis de componentes principales nos indica que la edad explica el 35% de la varianza en el diagnóstico:

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5
## Standard deviation     1.3338 1.0469 0.9384 0.8680 0.70067
## Proportion of Variance 0.3558 0.2192 0.1761 0.1507 0.09819
## Cumulative Proportion  0.3558 0.5750 0.7511 0.9018 1.00000
```

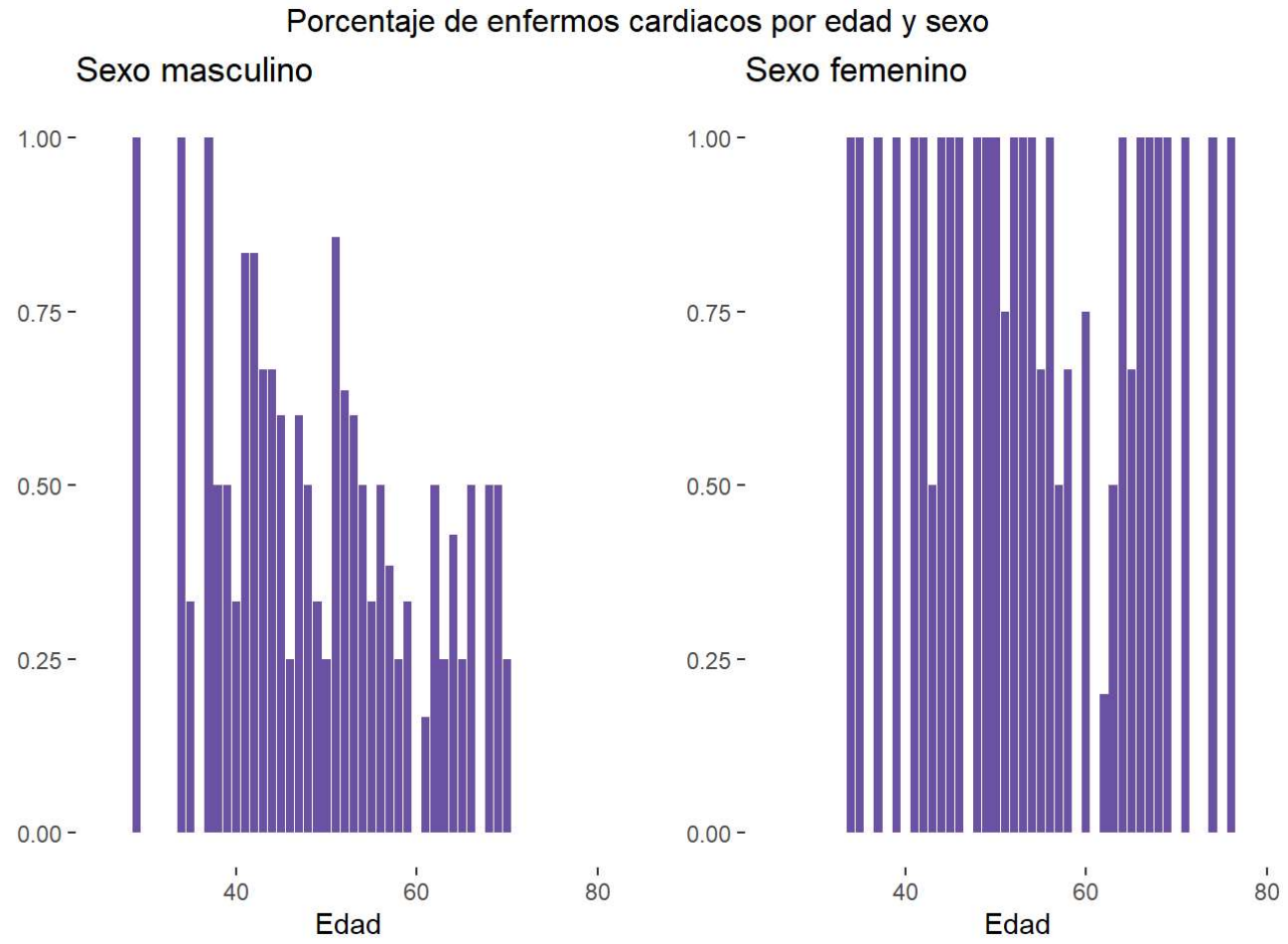
Por ello se visualiza el diagnóstico de enfermedad cardíaca en función de la edad. Se observa que la posibilidad de un diagnóstico de enfermedad cardíaca es mayor entre los 40 y 60 años:



Se estudia si el diagnóstico de enfermedad cardíaca aparece de forma distinta en hombres y mujeres en función de la edad:



Por el gráfico anterior parece que en la muestra de mujeres el diagnóstico de enfermedad cardíaca es más estable, lo cual se confirma con el siguiente gráfico:

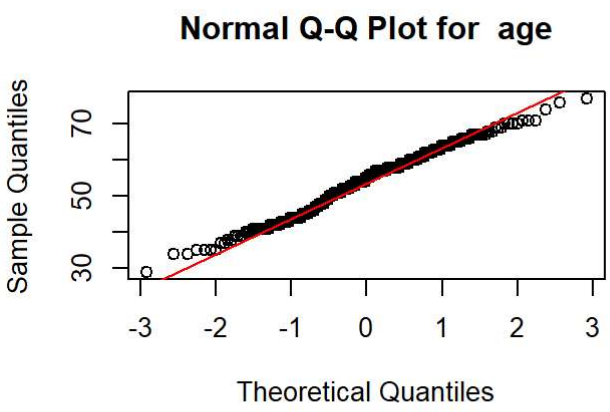


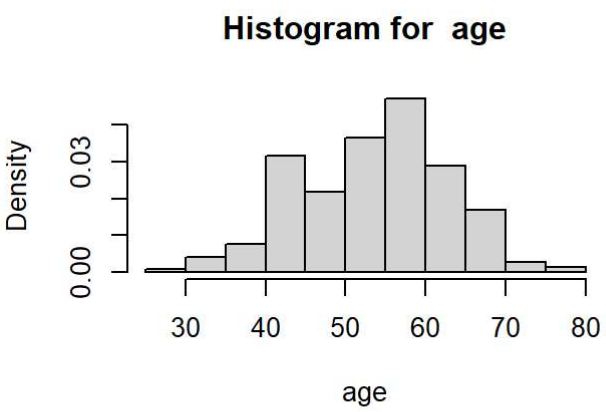
Conclusiones del análisis exploratorio:
- La posibilidad de un diagnóstico de enfermedad cardíaca es mayor entre los 40 y 60 años.
- Parece haber una mayor proporción de mujeres diagnosticadas con enfermedad cardíaca que en hombres.
- Parece que la proporción de mujeres diagnosticadas con enfermedad cardíaca se mantiene estable en el tiempo, mientras que en los hombre decrece.

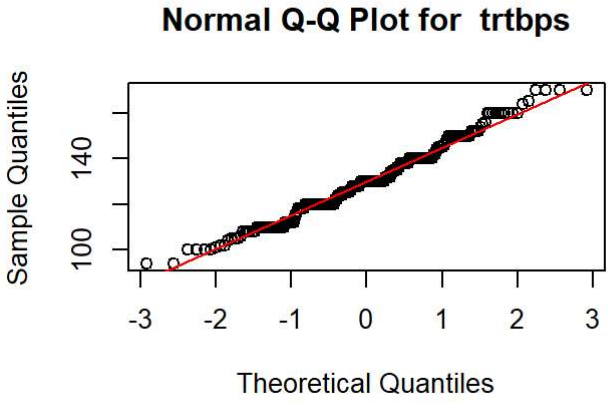
3.2 Comprobación de normalidad

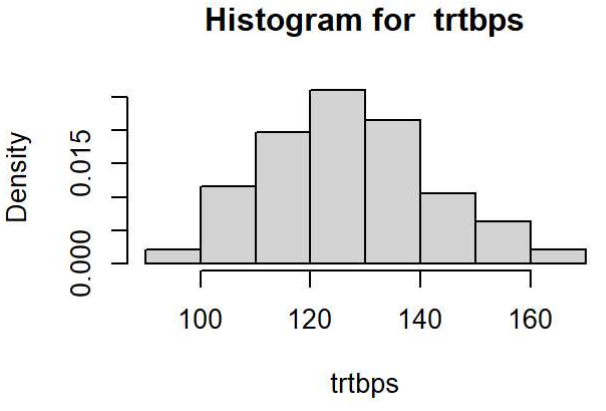
Para revisar si las variables pueden ser candidatas a la normalización miramos las graficas de quantile-quantile plot y el histograma.

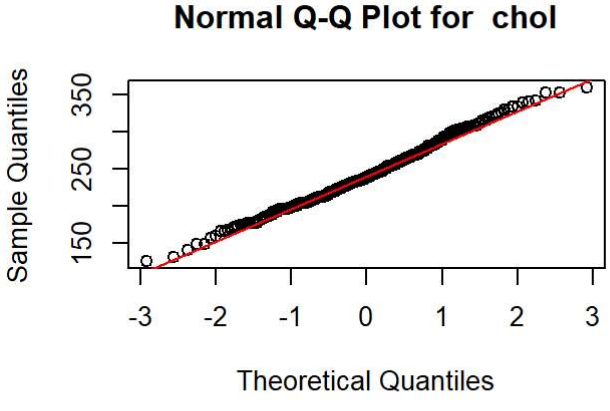
```
par(mfrow=c(2,2))
for(i in 1:ncol(heart)) {
  if (is.numeric(heart[,i])){
    qqnorm(heart[,i],main = paste("Normal Q-Q Plot for ",colnames(heart)[i]))
    qqline(heart[,i],col="red")
    hist(heart[,i],
      main=paste("Histogram for ", colnames(heart)[i]),
      xlab=colnames(heart)[i], freq = FALSE)
  }
}
```

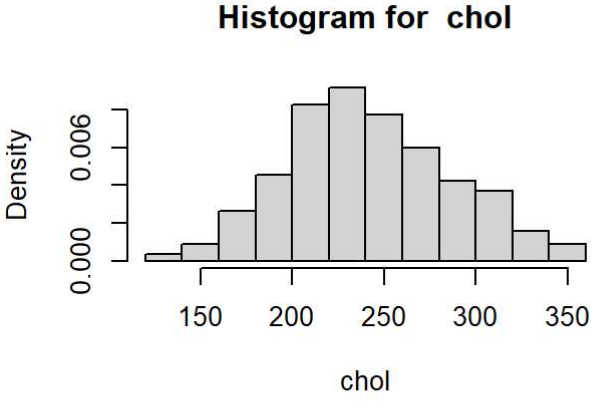

Normal Q-Q Plot for age
Sample Quantiles
30
50
70
-3
-2
-1
0
1
2
3
Theoretical Quantiles

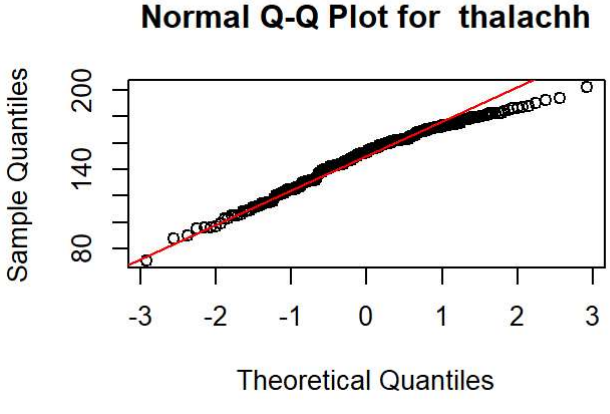

Histogram for age
Density
0.00
0.03
30
40
50
60
70
80
age

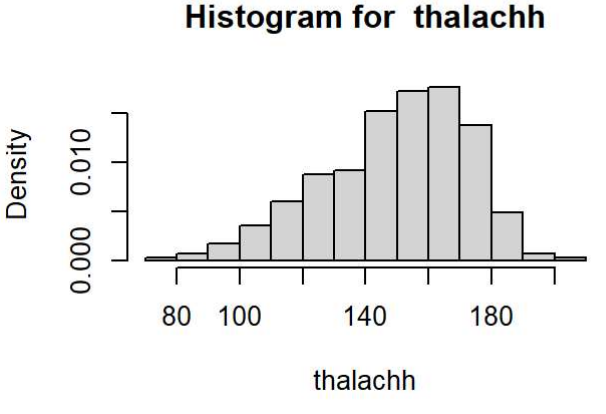

Normal Q-Q Plot for trtbps
Sample Quantiles
100
140
-3
-2
-1
0
1
2
3
Theoretical Quantiles

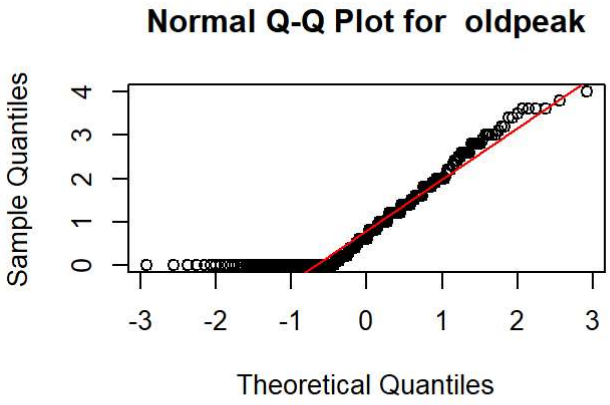

Histogram for trtbps
Density
0.000
0.015
100
120
140
160
trtbps

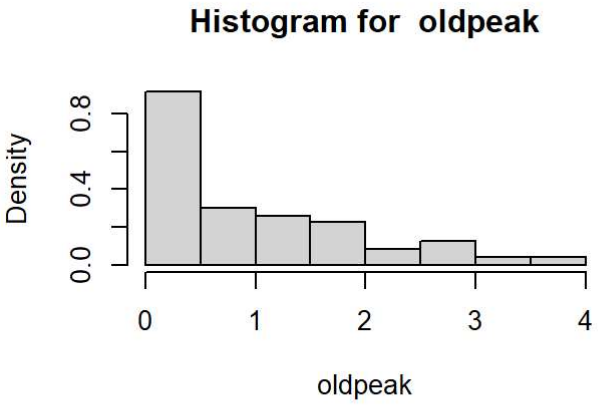

Normal Q-Q Plot for chol
Sample Quantiles
150
250
350
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Histogram for chol
Density
0.000
0.006
150
200
250
300
350
chol


Normal Q-Q Plot for thalachh
Sample Quantiles
80
140
200
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Histogram for thalachh
Density
0.000
0.010
80
100
140
180
thalachh





Los resultados del quantile-quantile plot nos indica que las variables pueden ser candidatas a la normalización si es necesario.

Para revisar si las variables estan normalizadas se aplica el test de Shapiro Wilk en cada variables numérica.

```
## P-valor Shapiro test para variable age :  0.01551123
## P-valor Shapiro test para variable trtbps :  0.002691308
## P-valor Shapiro test para variable chol :  0.1976049
## P-valor Shapiro test para variable thalachh :  4.050401e-05
## P-valor Shapiro test para variable oldpeak :  1.530384e-15
```

El test de normalidad de Shapiro-Wilk trabaja con la hipótesis nula de normalidad de los datos. Se observa un p-valor menor que el nivel de significancia (0.05) para las siguientes variables:
- Age
- Trtbps
- Thalachh
- Oldpeak

Por lo tanto se asume la distribución no normal en estas variables.

Se observa un p-valor mayor que el nivel de significancia para la variable **chol** y por lo tanto se asume que se encuentra distribuida normalmente.

Se realizan transformaciones Box-cox para las variables detectadas no-normales:

```
heart.norm<-heart
heart.norm$age<-BoxCox(heart$age,lambda = BoxCoxLambda(heart$age))
heart.norm$trtbps<-BoxCox(heart$trtbps,lambda = BoxCoxLambda(heart$trtbps))
heart.norm$thalachh<-BoxCox(heart$thalachh,lambda = BoxCoxLambda(heart$thalachh))
heart.norm$oldpeak<-BoxCox(heart$oldpeak,lambda = BoxCoxLambda(heart$oldpeak))
```

Se revisan los p-valor de los test de Shapiro nuevamente para estas variables:

```
## P-valor Shapiro test para variable age :  0.02802765
## P-valor Shapiro test para variable trtbps :  0.0174481
## P-valor Shapiro test para variable chol :  0.1976049
## P-valor Shapiro test para variable thalachh :  0.03779969
## P-valor Shapiro test para variable oldpeak :  3.145655e-18
```

Se observa que se ha mejorado el p-valor para las variables age, trtbps y thalachh. Sin embargo la variable oldpeak ha arrojado un p-valor mucho más pequeño, por lo que esta transformación no se guardará en nuestro conjunto de observaciones:

```
heart$age<-heart.norm$age
heart$trtbps<-heart.norm$trtbps
heart$thalachh<-heart.norm$thalachh
```

3.3 Análisis estadístico inferencial

3.3.1 Contraste de hipótesis de independencia entre las variables sexo-output

Se desea realizar un primer contraste en el que se evalue si las variables categóricas sexo y output se encuentra relacionadas.

Se dispone de 2 muestras independientes (Hombres y mujeres):

```
table(heart$sex,heart$output)
```

```
##
##          no disease disease
##   Female         17      68
##   Male          109      90
```

Para comparar diferencias significativas entre en una variable categórica (output) entre dos grupos definidos se utiliza la función chisq.test():

```
chisq.test(table(heart$sex,heart$output))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(heart$sex, heart$output)
## X-squared = 27.787, df = 1, p-value = 1.354e-07
```

Dado que se recoge un p-valor muy pequeño, se rechaza la hipótesis nula y se confirma que se observan diferencias significativas entre los hombres y mujeres en sus diagnósticos.

3.3.2 Contraste de hipótesis sobre varianza en la edad

Se desea comparar la varianza en la edad de los dos grupos (Hombre, Mujer). Se considera que se trata de dos muestras independientes con varianza desconocida.

Dado que se dispone de una cantidad de muestras significativa (199 y 85) se asume normalidad de en los datos por el Teorema del Límite Central.

```
leveneTest(y = heart$age, group = heart$sex, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   1  0.9979 0.3187
##       282
```

Se obtiene un p-valor mayor que el nivel de significancia, por lo tanto se asume la hipótesis normal: no hay diferencias significativas en la varianza de en la edad en los dos grupos.

3.3.3 Contraste de hipótesis sobre varianza en la edad en muestra diagnosticada

Se desea comparar la proporción dos grupos (Hombre, Mujer) pero esta vez se realiza solamente en el conjunto de muestra diagnosticada. Se considera que se trata de dos muestras independientes con varianza desconocida.

Dado que se dispone de una cantidad de muestras significativa (68 Female y 90 Male) se asume normalidad de en los datos.

```
heart.disease<-heart[which(heart$output=="disease"),]
t.test(heart.disease$age[heart.disease$sex=="Male"],heart.disease$age[heart.disease$sex=="Fem
ale"],alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  heart.disease$age[heart.disease$sex == "Male"] and heart.disease$age[heart.disease
$sex == "Female"]
## t = -2.0005, df = 128.66, p-value = 0.04756
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.1810092  -0.1109999
## sample estimates:
## mean of x mean of y
##  127.1816  137.3276
```

Se obtiene un valor menor que el nivel de significancia, por lo tanto se rechaza la hipótesis normal: hay diferencias significativas en la varianza de en la edad en los dos grupos cuando se trata de población enferma.

4 Construcción de un modelo predictivo

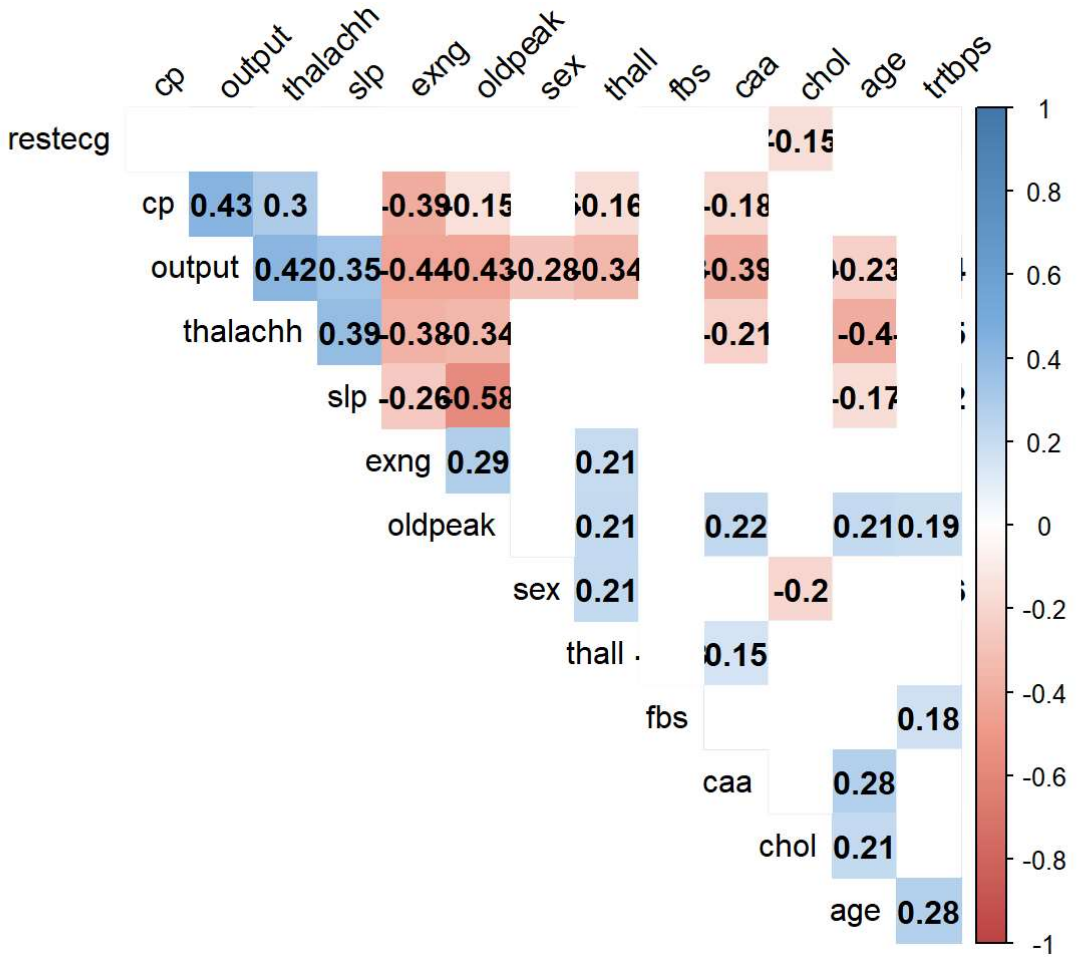
4.1 Creción de conjunto training y test

Se decide utilizar separar el conjunto original en un 80% para training y un 20% para test.

```
set.seed(9)
h<-holdout(heart$output,ratio=0.8)
heartTraining<-heart[h$tr,]
heartTest<-heart[h$ts,]
```

4.2 Identificación de correlación

Se observa que las variables **chol, trtbps y fbs** no están significativamente relacionadas con output.



Dado que se dispone de variables categóricas, se utiliza un modelo de regresión logística en el que se cuenta con las variables de edad y sexo, junto con las variables más correlacionadas con output detectas en el gráfico anterior:

```
Modelo=glm(
  formula=output~sex+age+cp+thalachh+exng+oldpeak+slp+caa+thall,
           data=heartTraining,
           family=binomial)
```

4.3 Bondad del ajuste

Se evalúa la bondad del ajuste, mediante la devianza. Para que el modelo sea bueno, la devianza residual debe ser menor que la devianza nula.

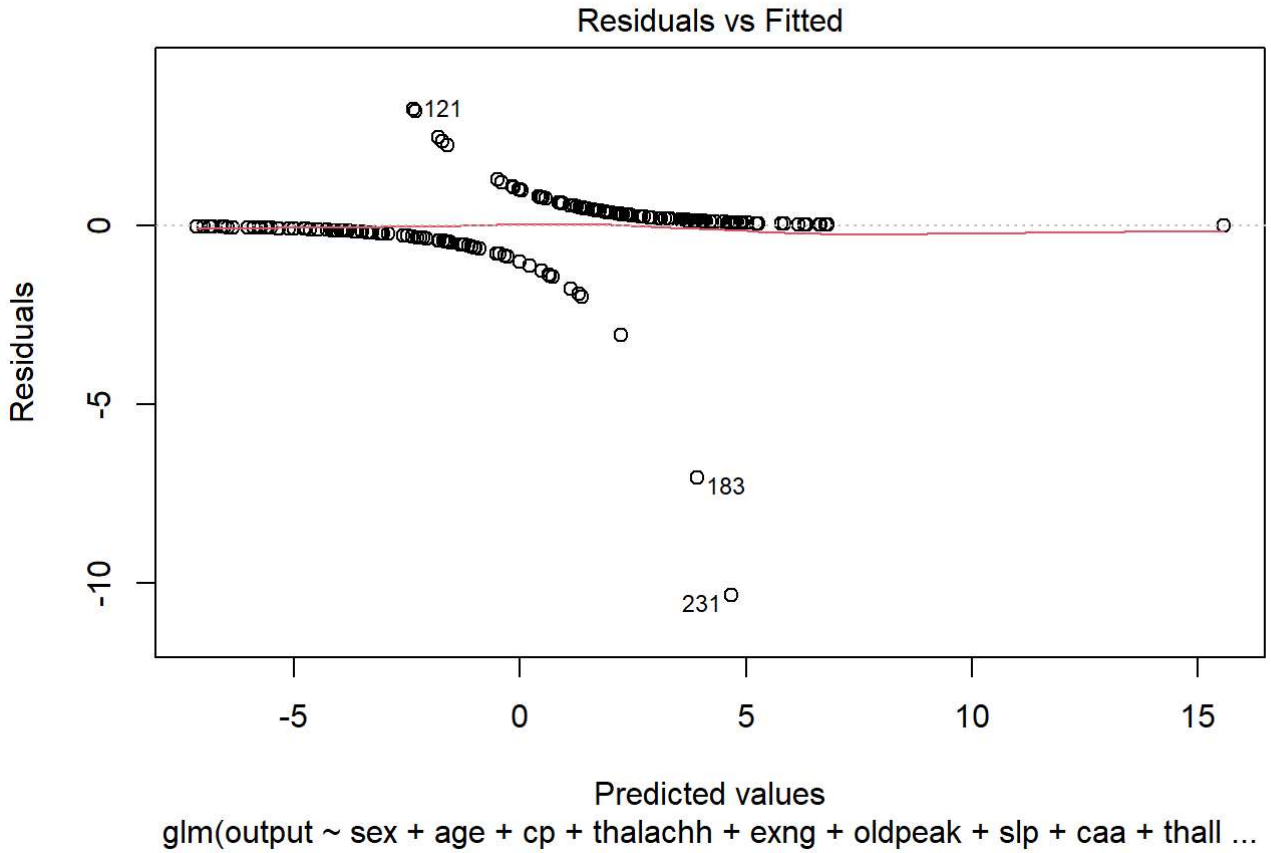
```
dev <- Modelo$deviance
nullDev <- Modelo$null.deviance
Chi_Obs <- nullDev - dev
Chi_Obs
```

```
## [1] 192.8172
```

Dado que el resultado es positivo, podemos deducir que el modelo predice la variable dependiente con mayor precisión.

4.4 Evaluación de residuos

```
plot(Modelo,which=1)
```

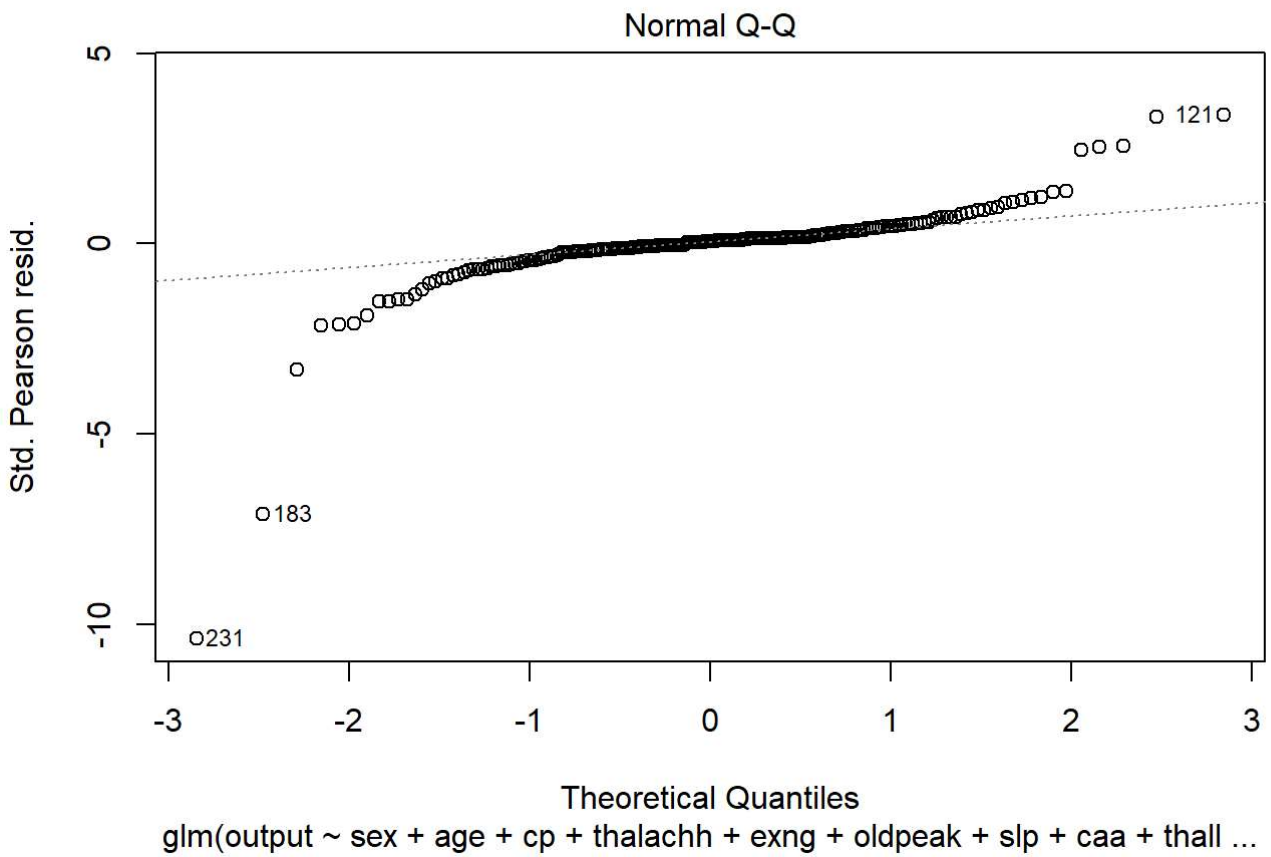


A través de este gráfico se aprecia que la varianza de los residuos en nuestro modelo no es constante, lo cual puede indicar que el modelo eligido no es óptimo.

4.5 Normalidad del modelo

```
plot(Modelo,which=2)
```

```
## Warning: not plotting observations with leverage one:
##   172
```



Por el gráfico qqnorm del modelo, se observa que los valores más bajos son menores que una distribución normal y los más altos son mayores que los de una distribución normal. Junto con el análisis anterior de los residuos, no se puede asumir normalidad en nuestro modelo.

4.6 Matriz de confusión

Se obtiene la matriz de confusión:

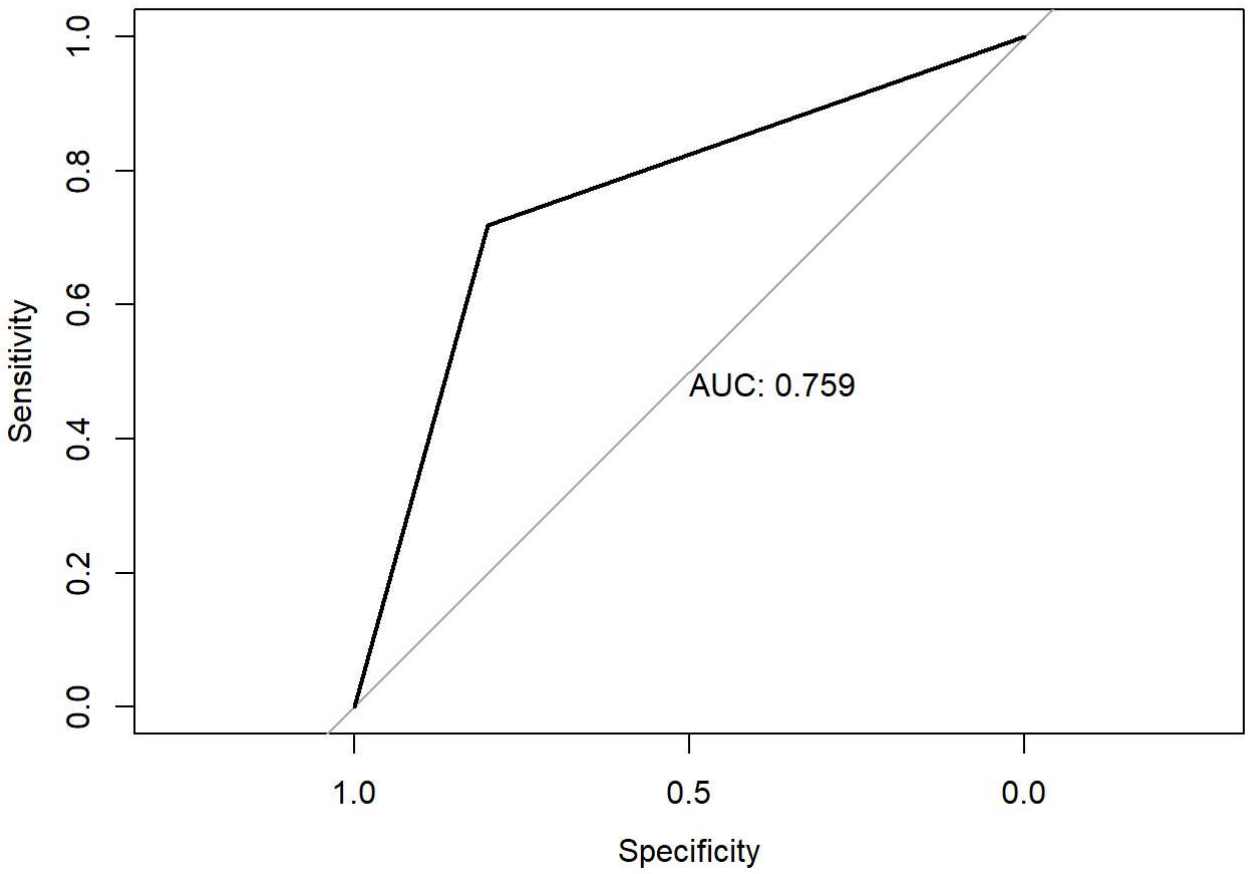
```
##
## output_predicted no disease disease
##                0         20       9
##                1          5      23
```

Por la matriz de confusión se calcula una exactitud de 0.68 (casos correctamente clasificados). Este valor se encuentra dentro un modelo aceptable en su límite inferior.

4.7 Evaluación de la potencia

```
## Setting levels: control = no disease, case = disease
```

```
## Setting direction: controls < cases
```




```
## Area under the curve: 0.7594
```

El valor de área debajo de la curva es de 0.76, lo cual indica un modelo aceptable pero sin contar con resultados óptimos.

```
#Se guarda el dataset trabajado
write.csv(heart, "heartFinal.csv", row.names=FALSE)
```

5 Resolución del problema, conclusiones

- La variable edad es una variable influyente en la predicción de un diagnóstico de enfermedad cardíaca.
- No se infieren diferencias significativas en la población sujeta a estudio medical: es decir, no se infiere que los hombres o las mujeres van más al médico en busca de un diagnóstico.
- Sí se infiere una diferencia signfiicativa poblacional en el diagnóstico de enfermedad cardíaca entre hombres y mujeres: parece haber una mayor proporción de mujeres afectada por esta dolencia.
- El modelo predictivo elegido (regresión logística) es un modelo aceptable pero probablemente pueda ser mejorado por otros (objeto para un potencial posterior estudio).

Contribuciones	Firma
Investigación previa	 DR JP
Redacción de las respuestas	 DR JP
Desarrollo del código	 DR JP
Participación en el video	 DR JP