Práctica 2

Mario García Puebla & Antonio García-Bustamante

4 de enero, 2023

Índice

1 Descripción del dataset

1.1 Variables

Para esta práctica, se va a analizar el dataset **"Heart Attack Analysis & Prediction Dataset"** obtenido en Kagle. Este conjunto de datos comprende datos de 303 personas distribuidos en 14 columnas. Estas 14 columnas son las siguientes:

- **age**: edad del paciente
- **sex**: género del paciente.
  - Valor 0: Mujer
  - Valor 1: Hombre
- **cp**: tipo de dolor en el pecho
  - Valor 1: angina típica
  - Valor 2: angina atípica
  - Valor 3: dolor no anginal
  - Valor 4: asintomático
- **trtbps**: presión arterial en reposo en mmHg
- **chol**: colesterol en mg/dl obtenido a través del sensor de IMC
- **fbs**: glucosa en ayunas > 120 mg/dl (1 = verdadero; 0 = falso)
- **rest_ecg**: resultados electrocardiográficos en reposo
  - Valor 0: normal
  - Valor 1: con anormalidades en la onda ST-T (inversiones de onda T y/o elevación o depresión de ST de > 0,05 mV)
  - Valor 2: muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes
- **thalach**: frecuencia cardíaca máxima alcanzada
- **exng**: presencia o no de angina inducida por el ejercicio (1 = sí; 0 = no)
- **oldpeak**: Depresión del ST inducida por el ejercicio en relación con el reposo ("ST" se refiere a las posiciones en el gráfico del ECG)
- **slope**: la pendiente del segmento ST de ejercicio máximo — Valor 0: pendiente descendente
  - Valor 1: plano
  - Valor 2: pendiente ascendente
- **caa**: cantidad de vasos mayores (0-3)
- **thal**: Un trastorno sanguíneo llamado talasemia
  - Valor 0: NULL
  - Valor 1: defecto fijo (no hay flujo sanguíneo en alguna parte del corazón)
  - Valor 2: flujo sanguíneo normal
  - Valor 3: defecto reversible (se observa un flujo sanguíneo pero no es normal)
- **output**: 0 = menor probabilidad de ataque al corazón, 1 = mayor probabilidad de ataque al corazón

1.2 Importancia del dataset

Este dataset acerca del análisis y la predicción de ataques cardiacos es fundamental ya que estos ataques son una de las principales causas de muertes en todo el mundo. Si se puede acceder a los valores de las variables necesarias para predecir estos ataques, a través de análisis de sangre por ejemplo, se puede determinar el riesgo de un ataque cardiaco, y tomar las medidas oportunas para prevenirlo.

Con este estudio se pretende poder solucionar el problema de ataques cardiacos no detectados o inesperados, pudiendo tomar medidas en casos complejos para poder reaccionar a tiempo, y así reducir el número de muertes por esta enfermedad.

2. Intregación y selección de los datos

En este apartado, se va a cargar el archivo a analizar y se va a hacer un pequeño resumen de las variables que se encuentran.

```
#Lectura del fichero
datos <- read.csv("heart.csv")
head(datos)
```

```
##   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
## 6  57   1  0    140  192   0       1      148    0     0.4   1   0     1      1
```

Para el desarrollo de la práctica se empleará el dataset "heart.csv". La finalidad de la limpieza y el análisis es predecir la probabilidad de un paciente de sufir un infarto mediante la variable Output. Para comprobar la integridad del conjunto, primero se comprueban los tipos de los datos:

```
#Resumen de los datos
summary(datos)
```

```
##       age             sex               cp            trtbps     
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
##       chol            fbs             restecg          thalachh     
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
##       exng           oldpeak          slp             caa        
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
```

```
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##      thall           output
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

Como se puede comprobar, todos los datos son numéricos. A priori no se puede descartar ningún atributo por no tener relevancia en el conjunto.

3 Limpieza de los datos

En el tercer apartado se va a llevar a cabo la limpieza del conjunto de datos, estudiando los valores nulos y los valores extremos

3.1 Valores nulos o vacíos

Para empezar, se va comprobar los elementos vacíos de cada variable:

```
# Valores vacíos
sapply(datos, function(x) sum(is.na(x)))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

Como se puede ver en la ejecución obtenida, ninguna variable posee valores vacíos. Ahora se va a comprobar si alguna variable que no le corresponde posee valores nulos, solo se hará la comprobación sobre las siguientes 5 variables, ya que el resto de ellas pueden tomar valores nulos.

```
#Valores nulos
any(datos$age == 0)
```

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

```
any(datos$trtbps == 0)
```

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

```
any(datos$chol == 0)
```

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

```r
any(datos$thalachh == 0)
```

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

```r
any(datos$thall == 0)
```

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

Se puede ver que ninguna variable tiene valores nulos a excepción de **thall**, por lo que habrá que eliminar estos valores nulos.

```r
#Eliminación de valores nulos
dim(datos)
```

```
## [1] 303  14
```

```r
datos <- datos[!datos$thall==0,]
dim(datos)
```

```
## [1] 301  14
```

Quedan eliminados los dos valores que tenían valores a 0 en la variales **thall**.

3.2 Valores extremos

A continuación se muestran los valores extremos de los datos del conjunto:

- Valores extremos del atributo **age**:

```r
stats <- boxplot.stats(datos$age)
extremos <- c(stats$out)
print(extremos)
```

```
## integer(0)
```

- Valores extremos del atributo **trtbps**:

```r
stats <- boxplot.stats(datos$trtbps)
extremos <- c(stats$out)
print(extremos)
```

```
## [1] 172 178 180 180 200 174 192 178 180
```

- Valores extremos del atributo **chol**:

```
stats <- boxplot.stats(datos$chol)
extremos <- c(stats$out)
print(extremos)
```

```
## [1] 417 564 394 407 409
```

- Valores extremos del atributo **thalachh**:

```
stats <- boxplot.stats(datos$thalachh)
extremos <- c(stats$out)
print(extremos)
```

```
## [1] 71
```

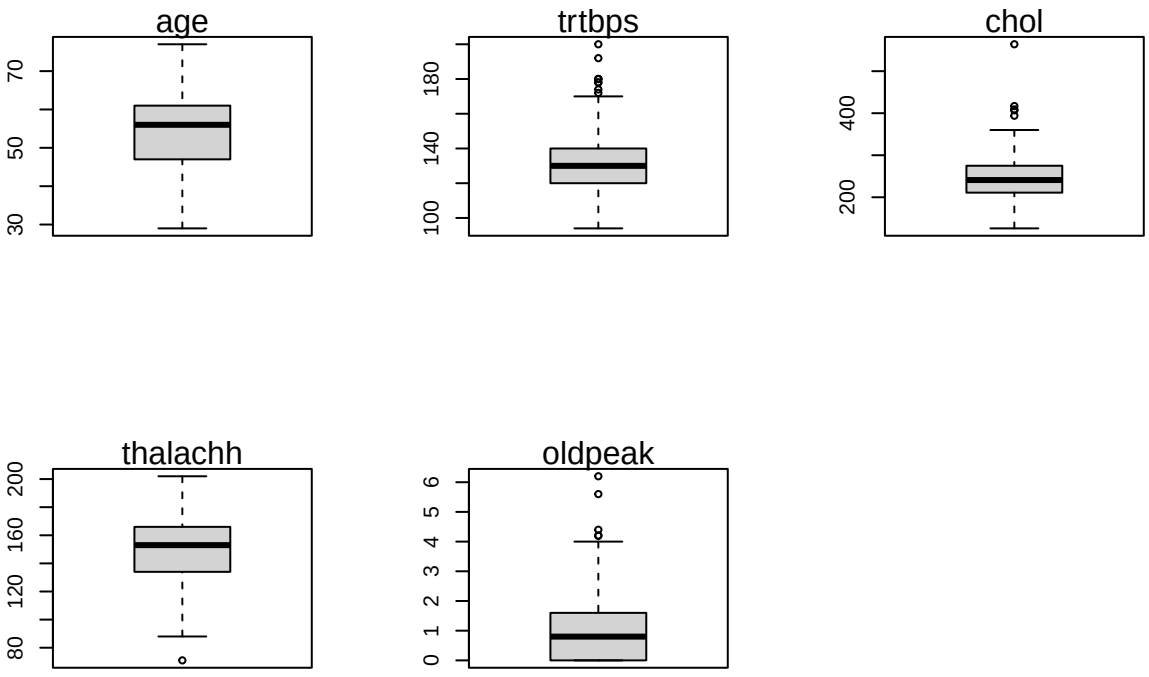
- Valores extremos del atributo **oldpeak**:

```
stats <- boxplot.stats(datos$oldpeak)
extremos <- c(stats$out)
print(extremos)
```

```
## [1] 4.2 6.2 5.6 4.2 4.4
```

A continuación se muestran los valores extremos representados mediante diagramas de cajas y bigotes

```
par(mfrow=c(2,3))

boxplot(datos$age)
mtext("age", side=3)

boxplot(datos$trtbps)
mtext("trtbps", side=3)

boxplot(datos$chol)
mtext("chol", side=3)

boxplot(datos$thalachh)
mtext("thalachh", side=3)

boxplot(datos$oldpeak)
mtext("oldpeak", side=3)
```



Como se puede comprobar tanto en los gráficos como en los valores extraidos, hay varios datos que se encuentran en los extremos de la distribución de datos.

Tras analizar cada caso individualmente, se comprueba que no es necesario modificar ni eliminar los registros que los contienen ya que, aunque sen valores que se alejan del punto medio de la distribución, son valores lógicos.

Se convierte el conjunto definitivo a formato csv.

```
write.csv(datos, "heart_clean.csv")
```

4 Análisis de los datos

4.1. Selección de los grupos de datos a analizar

La proporción de ataques al corazón entre hombres y mujeres no es necesariamente la misma. A menudo, los hombres tienen un mayor riesgo de sufrir un ataque al corazón que las mujeres. Sin embargo, esto puede variar dependiendo de factores como la edad, la raza y el estilo de vida. Con el paso del tiempo, el riesgo de sufrir un ataque al corazón en las mujeres se acerca al de los hombres.

Los datos de la Organización Mundial de la Salud muestran que, en general, la tasa de mortalidad por enfermedad coronaria (que incluye los ataques al corazón) es mayor en los hombres que en las mujeres en todas las edades. Pero, a medida que las mujeres envejecen, su riesgo de sufrir un ataque al corazón se aproxima al de los hombres.

Por lo tanto el conjunto se dividirá en uno de hombres y otro de mujeres.

```r
hombres <- subset(datos, sex == 1)
mujeres <- subset(datos, sex == 0)

print(head(hombres,4))
```

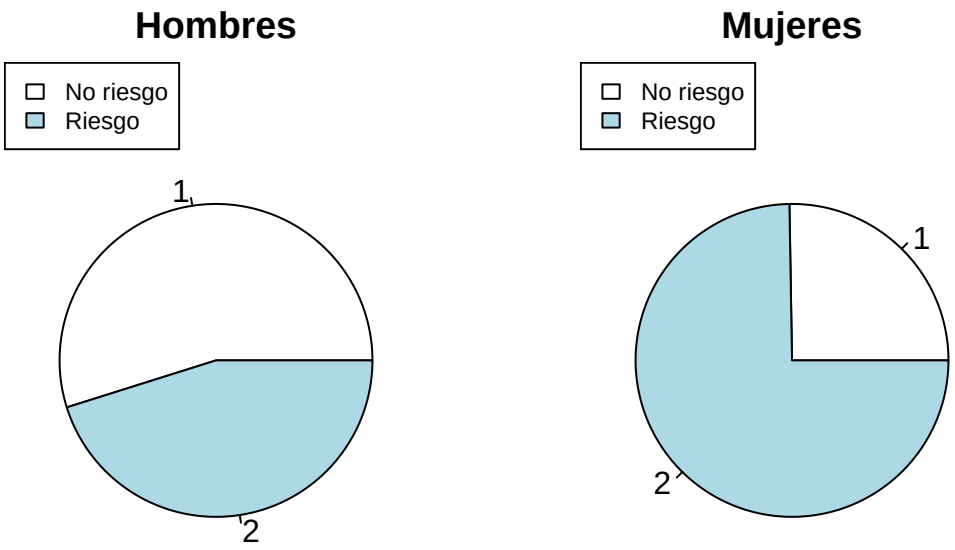
```
##   age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output
## 1  63   1  3    145  233   1       0      150    0     2.3   0   0     1      1
## 2  37   1  2    130  250   0       1      187    0     3.5   0   0     2      1
## 4  56   1  1    120  236   0       1      178    0     0.8   2   0     2      1
## 6  57   1  0    140  192   0       1      148    0     0.4   1   0     1      1
```

```r
print(head(mujeres,4))
```

```
##    age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
## 3   41   0  1    130  204   0       0      172    0     1.4   2   0     2
## 5   57   0  0    120  354   0       1      163    1     0.6   2   0     2
## 7   56   0  1    140  294   0       0      153    0     1.3   1   0     2
## 12  48   0  2    130  275   0       1      139    0     0.2   2   0     2
##    output
## 3       1
## 5       1
## 7       1
## 12      1
```

En el conjunto, las proporciones de ataques al corazón son las siguientes:

```r
par(mar=c(0,2,2,2))
par(mfrow=c(1,2))
# Calculamos la frecuencia de cada valor de "output" en cada conjunto de datos
hombres_output_0 <- sum(hombres$output == 0)
hombres_output_1 <- sum(hombres$output == 1)
mujeres_output_0 <- sum(mujeres$output == 0)
mujeres_output_1 <- sum(mujeres$output == 1)

# Creamos el primer gráfico de sectores con los datos de hombres
pie(c(hombres_output_0, hombres_output_1), main = "Hombres")
legend("topleft", legend = c("No riesgo","Riesgo"),
       fill =  c("white", "lightblue"),cex= 0.75)
# Creamos el segundo gráfico de sectores con los datos de mujeres
pie(c(mujeres_output_0, mujeres_output_1), main = "Mujeres")
legend("topleft", legend = c("No riesgo","Riesgo"),
       fill =  c("white", "lightblue"),cex= 0.75)
```



Aparentemente, a raíz de observar los gráficos, se puede decir que las mujeres tienen mayor probabilidad de tener ataques al corazón que los hombres.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

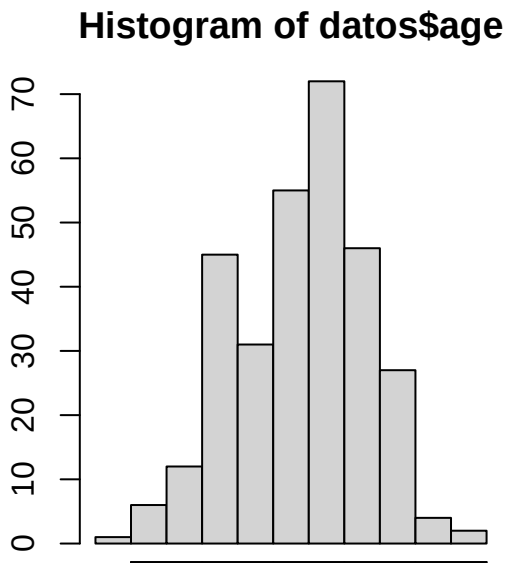
Para comprobar la normalidad de las distribuciones de las variables del dataset, se puede recurrir a la prueba de normalidad de Shapiro-Wilk. Se supondrá un nivel de significación de 0,05. Si el valor p obtenido del test es menor que el nivel de significación, entonces se puede rechazar la hipótesis nula y concluir que los datos no siguen una distribución normal.

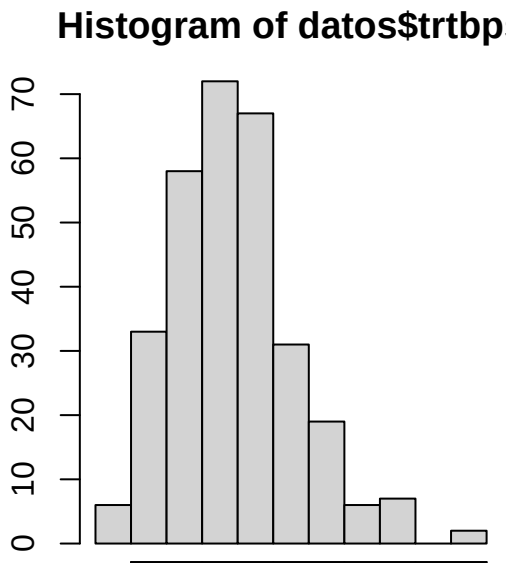
```
sapply(datos, function(x) shapiro.test(x)$p.value)
```

```
##          age          sex           cp       trtbps         chol          fbs
## 4.470113e-03 3.163166e-26 2.241062e-19 1.781220e-06 6.971614e-09 5.646991e-30
##      restecg     thalachh         exng      oldpeak          slp          caa
## 1.657018e-23 6.860354e-05 4.440286e-26 1.015340e-16 3.104756e-21 8.777887e-22
##        thall       output
## 2.137537e-21 6.742351e-25
```

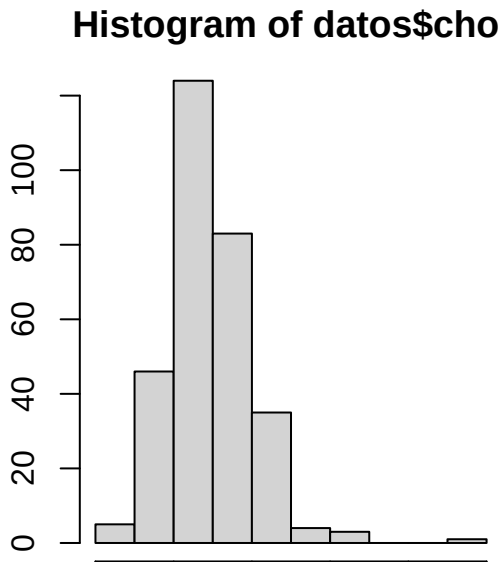
Se observa que el test de Shapiro-Wilk indica que las variables no siguen una distribución normal. Se va representar algunas de estas variables para comprobarlo.

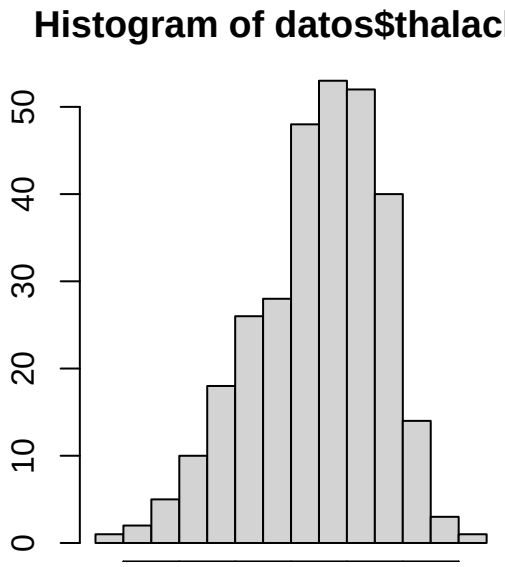
```
par(mar=c(0,2,2,2))
par(mfrow=c(1,2))
hist(datos$age)
hist(datos$trtbps)
```



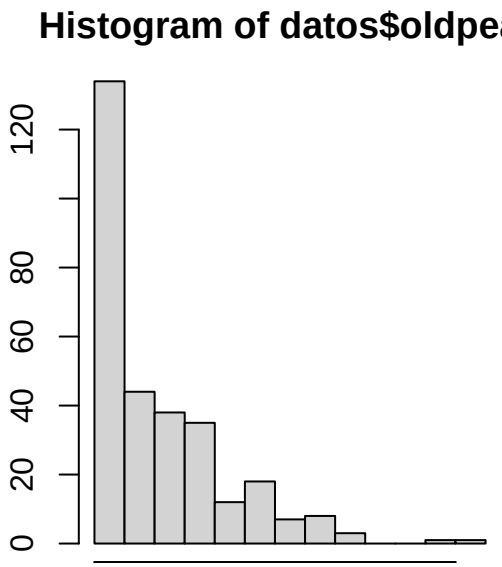


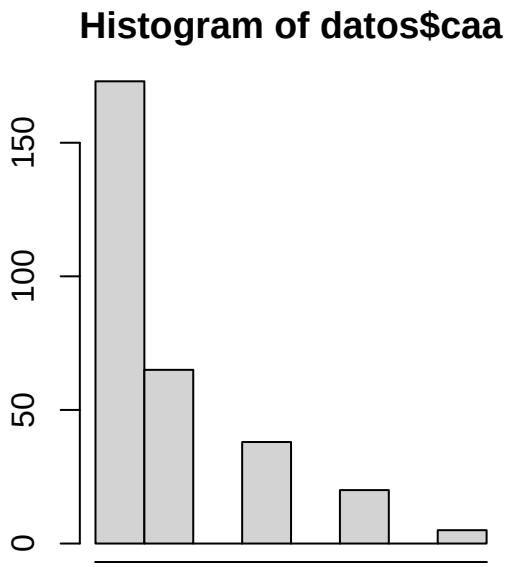

```
par(mfrow=c(1,2))
hist(datos$chol)
hist(datos$thalachh)
```






```
par(mfrow=c(1,2))
hist(datos$oldpeak)
hist(datos$caa)
```





Aunque parezca que alguna distribución de las variables **age, trtbps, chol y thalachh** sigan distribuciones normales, con el test de Shapiro-Wilk se confirma que no lo son.

Para comprobar la homogeneidad de la varianza de dos grupos (hombres y mujeres) primero hay que tener en cuenta que las variables no están distribuidas normalmente. Una opción es recurrir al test de Fligner-Killeen. Si el valor p resultante de la prueba es mayor que el nivel de significación escogido, no es posible rechazar la hipótesis nula de que la varianza de los dos grupos de datos es igual.

```
resultado <- fligner.test(hombres, mujeres)
print(resultado)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  hombres
## Fligner-Killeen:med chi-squared = 2018.5, df = 13, p-value < 2.2e-16
```

El p-value tiene un valor de 2,2e-16, mucho menor que el nivel de significación escogido (0,05) por lo tanto se puede rechazar la hipótesis nula de que la varianza de los dos grupos de datos es igual.

4.3 Pruebas estadísticas

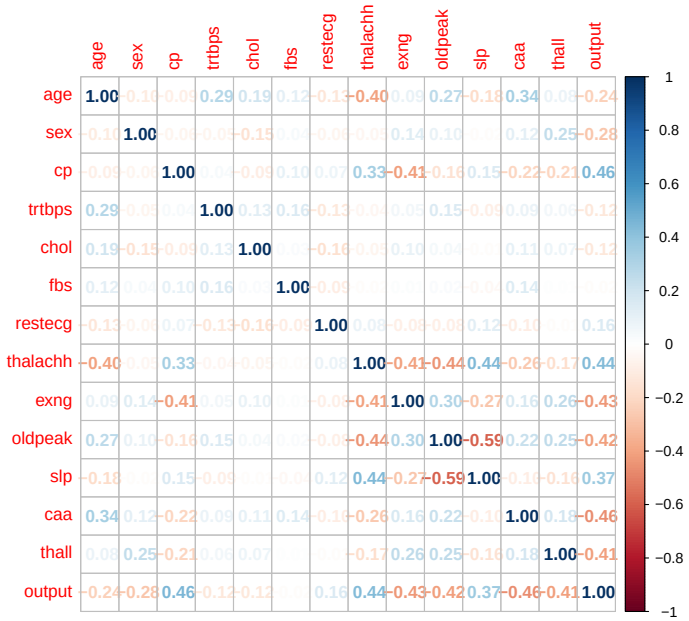
4.3.1 Variables más influyentes

Para comenzar el apartado de las pruebas estadísticas, se va a llevar a cabo un estudio de las variables más determinantes e influyentes para poder predecir un ataque al corazón. Para ello, se va a realizar un análisis de correlación de las variables.

En los apartados anteriores se ha observado que ninguna variable sigue una distribución normal, por lo que habrá que hacer uso del **coeficiente de correlación de Spearman** para llevar a cabo este análisis.

Primero se va a hacer una pequeña visualización de la correlación entre todas las variables.

```
#Análisis de correlación
correlacion <- cor(datos,method="spearman")
corrplot(correlacion, method = "number")
```



Una vez visto el gráfico, se van a observar las variables con una mayor correlación respecto a la variable **output**.

```
#Análisis de correlación
correlacion1 <- cor(datos,datos$output,method="spearman")
print(correlacion1)
```

```
##                [,1]
## age      -0.23932550
## sex      -0.27616965
## cp        0.45740334
## trtbps   -0.12287920
## chol     -0.12331272
## fbs      -0.01838199
## restecg   0.15610077
## thalachh  0.43610178
## exng     -0.43274331
## oldpeak  -0.41958201
## slp       0.36777169
## caa      -0.46049843
## thall    -0.41105362
## output    1.00000000
```

Con estos datos, se observa que no hay ninguna variable que destaque por encima del resto, en donde se encuentra que las variables más influyentes sobre la variable **output** son **cp**, **thalachh**, **exgn**, **oldpeak**, **caa** y **thall**.

4.3.2 ¿Es mayor la probabilidad de ataque al corazón siendo mujer que siendo hombre?

Siguiendo con la división de datos realizada en los apartados anteriores, se quiere aplicar un contraste de hipótesis sobre dos muestras, hombres y mujeres, para poder observar si la probabilidad de ataque al corazón es superior siendo mujer que siendo hombre

Primero hay que hacer la división en los dos conjuntos de datos.

```
#División de los datos
hombres <- datos[datos$sex ==1,]$output
mujeres <-datos[datos$sex ==0,]$output
print(hombres)
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
print(mujeres)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0
## [77] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

En segundo lugar, las muestras de datos que se van a utilizar deben tener un tamaño mayor a 30. Se comprueba si esto se cumple.

```
#Comprobación de tamaño
print(length(hombres))
```

```
## [1] 206
```

```
print(length(mujeres))
```

```
## [1] 95
```

Con el tamaño mayor de 30, se puede hacer uso de un test no paramétrico.

Con las comprobaciones pertinentes realizadas, se tiene un contraste de hipótesis no paramétrico de dos muestras independientes sobre la diferencia de medias.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Se lleva a cabo el test no paramétrico de Wilcoxon, en donde se toma $\alpha = 0,05$.

```
#Test paramétrico
wilcox.test(mujeres,hombres)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  mujeres and hombres
## W = 12680, p-value = 1.731e-06
## alternative hypothesis: true location shift is not equal to 0
```

En los resultados se puede observar que el p-value obtenido es menor que el valor de significancia fijado, por lo que se rechaza la hipótesis nula.

De esto se concluye que la **probalidad de ataque al corazón es mayor siendo mujer que siendo hombre**.

4.3.3 Modelo de regresión lineal

Como última prueba estadística de este análisis, se va a llevar a cabo el cálculo de un modelo de regresión lineal, con el objetivo de poder predecir si una persona tiene probabilidad de sufrir un ataque al corazón.

Para ello se va a construir un modelo con todas las variables posibles, y en función de los resultados se irán revisando para construir el mejor modelo posible, con el mayor coeficiente de determinación.

```
#Modelo de regresión lineal
modelo1 <- lm(output  ~ age + sex   + cp + trtbps + chol + fbs + restecg
              + thalachh + exng +   oldpeak +   slp +   caa +   thall ,data = datos)
summary(modelo1)
```

```
##
## Call:
## lm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
##     restecg + thalachh + exng + oldpeak + slp + caa + thall,
##     data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95363 -0.20886  0.05418  0.25234  0.93893
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8754880  0.2958259   2.959 0.003339 **
## age         -0.0007656  0.0026886  -0.285 0.776046
## sex         -0.1902520  0.0471352  -4.036 6.97e-05 ***
## cp           0.1097030  0.0223928   4.899 1.61e-06 ***
## trtbps      -0.0020255  0.0012522  -1.618 0.106861
## chol        -0.0003688  0.0004200  -0.878 0.380661
## fbs          0.0338335  0.0600407   0.564 0.573528
## restecg      0.0541234  0.0399866   1.354 0.176949
## thalachh     0.0030743  0.0011412   2.694 0.007476 **
## exng        -0.1339049  0.0514066  -2.605 0.009671 **
## oldpeak     -0.0581810  0.0228305  -2.548 0.011343 *
## slp          0.0776777  0.0423667   1.833 0.067770 .
## caa         -0.1028272  0.0217880  -4.719 3.70e-06 ***
## thall       -0.1417136  0.0373186  -3.797 0.000178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3527 on 287 degrees of freedom
## Multiple R-squared:  0.5218, Adjusted R-squared:  0.5001
## F-statistic: 24.09 on 13 and 287 DF,  p-value: < 2.2e-16
```

Para el primer modelo se obtiene un $R^2 = 0{,}5$, pero se observa que hay varias variables que no aportan valor al modelo, por lo que se va a hacer un estudio de las variables más interesantes para este modelo.

```
#Estudio de las mejores variables para el modelo
step(object = modelo1, direction = "both", trace = 1)
```

```
## Start:  AIC=-613.75
## output ~ age + sex + cp + trtbps + chol + fbs + restecg + thalachh +
##     exng + oldpeak + slp + caa + thall
##
##            Df Sum of Sq    RSS     AIC
## - age       1   0.01008 35.706 -615.67
## - fbs       1   0.03950 35.736 -615.42
## - chol      1   0.09589 35.792 -614.94
## - restecg   1   0.22787 35.924 -613.84
## <none>                  35.696 -613.75
## - trtbps    1   0.32542 36.022 -613.02
## - slp       1   0.41810 36.114 -612.25
## - oldpeak   1   0.80774 36.504 -609.02
## - exng      1   0.84391 36.540 -608.72
## - thalachh  1   0.90268 36.599 -608.24
## - thall     1   1.79355 37.490 -601.00
## - sex       1   2.02632 37.722 -599.13
## - caa       1   2.77026 38.466 -593.25
## - cp        1   2.98510 38.681 -591.58
##
## Step:  AIC=-615.67
## output ~ sex + cp + trtbps + chol + fbs + restecg + thalachh +
##     exng + oldpeak + slp + caa + thall
##
##            Df Sum of Sq    RSS     AIC
## - fbs       1   0.03732 35.744 -617.35
## - chol      1   0.10982 35.816 -616.74
## - restecg   1   0.23475 35.941 -615.69
## <none>                  35.706 -615.67
## - trtbps    1   0.37055 36.077 -614.56
## - slp       1   0.41503 36.121 -614.19
## + age       1   0.01008 35.696 -613.75
## - oldpeak   1   0.81432 36.521 -610.88
## - exng      1   0.83462 36.541 -610.71
## - thalachh  1   1.13249 36.839 -608.27
## - thall     1   1.79752 37.504 -602.88
## - sex       1   2.01895 37.725 -601.11
## - caa       1   2.93577 38.642 -593.88
## - cp        1   2.97506 38.681 -593.58
##
## Step:  AIC=-617.35
## output ~ sex + cp + trtbps + chol + restecg + thalachh + exng +
```

```
##     oldpeak + slp + caa + thall
##
##            Df Sum of Sq    RSS     AIC
## - chol      1   0.10966 35.853 -618.43
## - restecg   1   0.22380 35.967 -617.47
## <none>                  35.744 -617.35
## - trtbps    1   0.34273 36.086 -616.48
## - slp       1   0.40037 36.144 -616.00
## + fbs       1   0.03732 35.706 -615.67
## + age       1   0.00791 35.736 -615.42
## - exng      1   0.82286 36.566 -612.50
## - oldpeak   1   0.84893 36.593 -612.29
## - thalachh  1   1.13317 36.877 -609.96
## - thall     1   1.80932 37.553 -604.49
## - sex       1   2.00171 37.745 -602.95
## - caa       1   2.90345 38.647 -595.84
## - cp        1   3.10872 38.852 -594.25
##
## Step:  AIC=-618.43
## output ~ sex + cp + trtbps + restecg + thalachh + exng + oldpeak +
##     slp + caa + thall
##
##            Df Sum of Sq    RSS     AIC
## <none>                  35.853 -618.43
## - restecg   1    0.2817 36.135 -618.07
## + chol      1    0.1097 35.744 -617.35
## - trtbps    1    0.3771 36.230 -617.28
## - slp       1    0.3812 36.234 -617.25
## + fbs       1    0.0372 35.816 -616.74
## + age       1    0.0206 35.833 -616.60
## - exng      1    0.8542 36.707 -613.34
## - oldpeak   1    0.8700 36.723 -613.21
## - thalachh  1    1.1128 36.966 -611.23
## - sex       1    1.8921 37.745 -604.95
## - thall     1    1.9224 37.776 -604.71
## - caa       1    2.9676 38.821 -596.49
## - cp        1    3.1675 39.021 -594.95

##
## Call:
## lm(formula = output ~ sex + cp + trtbps + restecg + thalachh +
##     exng + oldpeak + slp + caa + thall, data = datos)
##
## Coefficients:
## (Intercept)          sex           cp       trtbps      restecg     thalachh
##    0.746452    -0.177338     0.111827    -0.002090     0.059217     0.003166
##        exng      oldpeak          slp          caa        thall
##   -0.134055    -0.060116     0.073896    -0.103180    -0.145893
```

Tras realizar un estudio de las mejores variables, se obtiene el modelo final.

```
#Modelo de regresión lineal final
modelo_final <- lm(output  ~ sex    + cp + trtbps + restecg
```

```
              + thalachh + exng +   oldpeak +   slp +   caa +   thall ,data = datos)
summary(modelo_final)
```

```
##
## Call:
## lm(formula = output ~ sex + cp + trtbps + restecg + thalachh +
##     exng + oldpeak + slp + caa + thall, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96348 -0.20353  0.06157  0.25082  0.91984
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.746452   0.238174   3.134 0.001901 **
## sex         -0.177338   0.045331  -3.912 0.000114 ***
## cp           0.111827   0.022093   5.062  7.4e-07 ***
## trtbps      -0.002090   0.001197  -1.746 0.081792 .
## restecg      0.059217   0.039229   1.510 0.132249
## thalachh     0.003166   0.001055   3.000 0.002932 **
## exng        -0.134055   0.051000  -2.629 0.009032 **
## oldpeak     -0.060116   0.022661  -2.653 0.008422 **
## slp          0.073896   0.042082   1.756 0.080143 .
## caa         -0.103180   0.021060  -4.899  1.6e-06 ***
## thall       -0.145893   0.036998  -3.943 0.000101 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3516 on 290 degrees of freedom
## Multiple R-squared:  0.5197, Adjusted R-squared:  0.5031
## F-statistic: 31.38 on 10 and 290 DF,  p-value: < 2.2e-16
```

Con este nuevo modelo, el coeficiente de determinación, R^2, mejora muy poco, $R^2 = 0{,}5031$.

Con el modelo final determinado, ahora se puede predecir la probabilidad de una persona de tener probabilidades de sufrir un ataque al corazón.

```
#Predicción
prediccion <- data.frame(
sex = 1,
cp = 0,
trtbps = 127,
restecg = 1,
thalachh = 170,
exng = 0,
oldpeak = 0,
slp = 1,
caa = 0,
thall = 2
)

prediccion1 <- data.frame(
sex = 0,
```

```
cp = 0,
trtbps = 127,
restecg = 1,
thalachh = 170,
exng = 0,
oldpeak = 0,
slp = 1,
caa = 0,
thall = 2
)

prediccion2 <- data.frame(
sex = 1,
cp = 0,
trtbps = 127,
restecg = 1,
thalachh = 170,
exng = 0,
oldpeak = 3,
slp = 1,
caa = 0,
thall = 2
)

prediccion3 <- data.frame(
sex = 0,
cp = 0,
trtbps = 127,
restecg = 1,
thalachh = 170,
exng = 0,
oldpeak = 3,
slp = 1,
caa = 0,
thall = 2
)

predict(modelo_final, prediccion)
```

```
##         1
## 0.6832027
```

```
predict(modelo_final, prediccion1)
```

```
##         1
## 0.8605407
```

```
predict(modelo_final, prediccion2)
```

```
##         1
## 0.5028539
```

```
predict(modelo_final, prediccion3)
```

```
##         1
## 0.6801919
```

Tras realizar 4 predicciones con muestras de test diferentes, se puede concluir que el riesgo de ataque al corazón aumenta cuando el paciente es mujer y cuando el valor de oldpeak es menor.

5 Resolución del problema

Mediante el análisis se ha propuesto resolver tres cuestiones:

- ¿Qué factores aumentan el riesgo de ataques al corazón?
- ¿Son hombres y mujeres igualmente propensos a padecer ataques al corazón?
- ¿Qué probabilidad tiene un paciente de sufrir un ataque al corazón?

Para responder a la primera pregunta, primero se ha realizado un estudio sobre la normalidad de las variables del conjunto, sometiéndolas a un test de Shapiro-Wilk y representando sus distribuciones mediante histogramas. Al analizar las distribuciones, aparentemente las variables presentan distribuciones normales, pero el test confirma que no lo son. A continuación, se analiza que variables son las más influyentes sobre la variable de la salida mediante un análisis de correlación, estas son: **cp, thalachh, exgn, oldpeak, caa y thall**.

Para la segunda pregunta, se ha divido el conjunto original en dos subconjuntos, uno con todos los hombres y otro con todas las mujeres. Primero se ha observado la proporción de pacientes con riesgo de ataque al corazón es superior dentro de las mujeres que dentro de los hombres. A continuación, se someten ambos conjuntos a un test de Fligner-Killeen con la finalidad de comprobar la homogeneidad de sus varianzas. Tras analizar los resultados del test, se confirma que no son iguales. Posteriormente, teniendo en cuenta las condiciones de no normalidad y no homocedasticidad, se aplica a los conjuntos el test no paramétrico de Wilcoxon, obtenidndo que la probalidad de ataque al corazón es mayor siendo mujer que siendo hombre.

Tratando la tercera cuestión, se ha creado un modelo de regresión lineal. Tras el primer modelado, se han encontrado ciertas variables que no aportan información al modelo, manteniendo en el modelo definitivo la variable a predecir (output) y las variables **sex, cp, trtbps, restecg, thalachh, exng, oldpeak, slp, caa y thall**. Con este modelo se obtiene un coeficiente de determinación de 0,5031. Tras realizar 4 predicciones con muestras de test diferentes, se puede concluir que el riesgo de ataque al corazón aumenta cuando el paciente es mujer y cuando el valor de oldpeak es menor.

6 Contribuciones al trabajo

Contribuciones	Firma
Investigación Previa	Mario García Puebla, Antonio García-Bustamante Usano
Redacción de las respuestas	Mario García Puebla, Antonio García-Bustamante Usano
Desarrollo del código	Mario García Puebla, Antonio García-Bustamante Usano
Participación en el video	Mario García Puebla, Antonio García-Bustamante Usano