

## M2.951 - Tipologia i cicle de vida de les dades: Pràctica 2

Autors: Francisco J. Bastida López ([fbastidal@uoc.edu](mailto:fbastidal@uoc.edu)) / Ivan Benaiges Trenchs  
([ibenaiges@uoc.edu](mailto:ibenaiges@uoc.edu))

Semestre 2022/23-2 / Juny 2023

---

### Descripció del dataset

Inicialment s'havia pensat aprofitar les dades recopilades durant la primera pràctica, relacionades amb dades de navegació de vaixells. No obstant, després de valorar-ho, no trobavem com poder utilitzar-lo de forma bastant gràfica i que ens permetés realitzar un estudi estadístic com el que es demana a l'enunciat, pel que finalment hem decidit utilitzar el conjunt de dades que es posa com a exemple a l'enunciat, ja que considerem que resulta molt interessant i pot donar més joc al estar relacionat amb temes de salut.

El conjunt de dades seleccionat "[Heart Attack Analysis & Prediction dataset](#)" proporciona informació sobre diferents factors que podrien estar relacionats amb malalties cardiovasculars, més concretament amb predir la probabilitat de patir un atac de cor.

Es tracta d'un conjunt de dades que ens permetrà realitzar un estudi en major profunditat de possibles problemes del cor, molt més centrat en els atacs de cor, al tenir una variable que està directament relacionada amb el resultat que es va obtenir durant l'estudi dels pacients i que ens permeti trobar un model per determinar la probabilitat de patir un atac de cor segons els valors dels diferents factors d'estudi.

Per poder facilitar aquesta anàlisi, el dataset conté el fitxer *heart.csv* amb les següents variables/atributs:

- **age**: edat del pacient
- **sex**: gènere del pacient (0 = femení, 1 = masculí)
- **cp**: tipus de dolor toràcic que experimenta el pacient (0 = angina típica, 1 = angina atípica, 2 = dolor no relacionat amb angina, 3 = asimptomàtic (sense dolor toràcic))
- **trtbps**: pressió arterial en repòs (mm Hg)
- **chol**: nivell de colesterol (md/dl)
- **fbs**: nivell de sucre en sang en dejú (0 = normal, 1 = alt)

- **restecg**: resultat de l'electrocardiograma en repòs (0 = normal, 1 = anomalies en l'ona ST-T, 2 = hipertrofia ventricular esquerra probable o definitiva segons els criteris d'Estes)
- **thalachh**: freqüència cardíaca màxima registrada pel pacient durant les proves realitzades
- **exng**: angina provocada per l'exercici (0 = no, 1 = sí)
- **oldpeak**: canvis en el segment ST de l'electrocardiograma després de l'exercici físic
- **slp**: patró de canvi en el segment ST de l'electrocardiograma durant una prova d'esforç o situacions d'estrès (0 = pendent plana, 1 = pendent ascendent (canvi en el patró elèctric del cor), 2 = pendent descendent (canvi en el patró elèctric del cor))
- **caa**: nombre de vasos sanguinis coronaris que mostren obstrucció o estenosi significativa (0 = sense obstrucció detectada, 1 = obstrucció en un dels vasos sanguinis, 2 = obstrucció en dos dels vasos sanguinis, 3 = obstrucció en tres dels vasos sanguinis, 4 = obstrucció en els quatre vasos sanguinis)
- **thall**: relacionat amb una malaltia hereditària de la sang anomenada talassèmia (1 = no s'ha detectat cap indicatiu, 2 = presència d'un defecte fix, 3 = presència d'un defecte reversible)
- **output**: probabilitat de patir un atac de cor (0 = sense o poca probabilitat, 1 = major probabilitat)

Per altra banda, hi ha disponible un segon fitxer, *o2Saturation.csv*, que conté múltiples observacions relacionades amb el nivell de saturació d'oxigen (una única variable). El fitxer conté moltíssimes més observacions (3586 en total) sense cap tipus d'identificador que ens permeti poder realitzar una integració de les dades d'ambdós fitxers per tenir un conjunt més complet.

## Integració i selecció de les dades d'interès a analitzar

### Càrrega de dades

Abans de començar, és necessari carregar el fitxer en un data frame que ens permeti treballar de forma còmoda amb les dades. Una vegada carregades totes les dades, procedim a un primer anàlisi ràpid a partir del resum estadístic del data frame.

*Data summary*

Name	dfHeartAttack
Number of rows	303
Number of columns	14















Column type frequency:

numeric	14
---------	----

Group variables

None

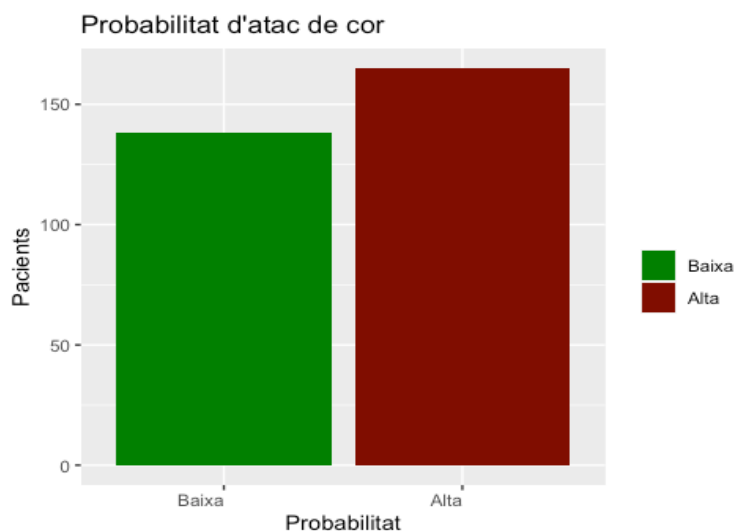
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	54.37	9.08	29	47.5	55.0	61.0	77.0	
sex	0	1	0.68	0.47	0	0.0	1.0	1.0	1.0	
cp	0	1	0.97	1.03	0	0.0	1.0	2.0	3.0	
trtbps	0	1	131.62	17.54	94	120.0	130.0	140.0	200.0	
chol	0	1	246.26	51.83	126	211.0	240.0	274.5	564.0	
fbs	0	1	0.15	0.36	0	0.0	0.0	0.0	1.0	
restecg	0	1	0.53	0.53	0	0.0	1.0	1.0	2.0	
thalachh	0	1	149.65	22.91	71	133.5	153.0	166.0	202.0	
exng	0	1	0.33	0.47	0	0.0	0.0	1.0	1.0	
oldpeak	0	1	1.04	1.16	0	0.0	0.8	1.6	6.2	
slp	0	1	1.40	0.62	0	1.0	1.0	2.0	2.0	
caa	0	1	0.73	1.02	0	0.0	0.0	1.0	4.0	
thall	0	1	2.31	0.61	0	2.0	2.0	3.0	3.0	
output	0	1	0.54	0.50	0	0.0	1.0	1.0	1.0	

Veiem que tenim un joc de dades compost per un total de 303 observacions i 14 variables.

## Anàlisi exploratori

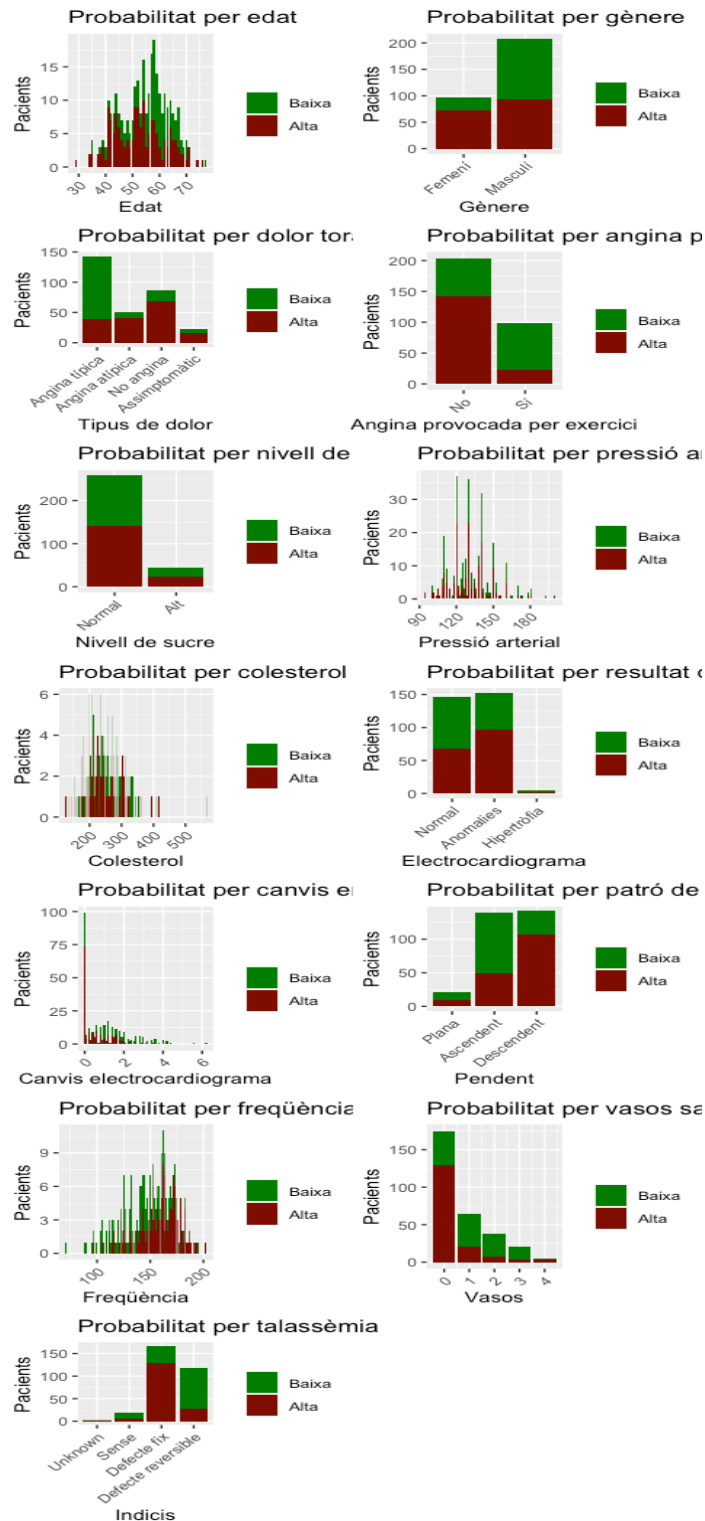
Realitzem un primer anàlisi exploratori que ens permeti entendre millor el conjunt de dades.



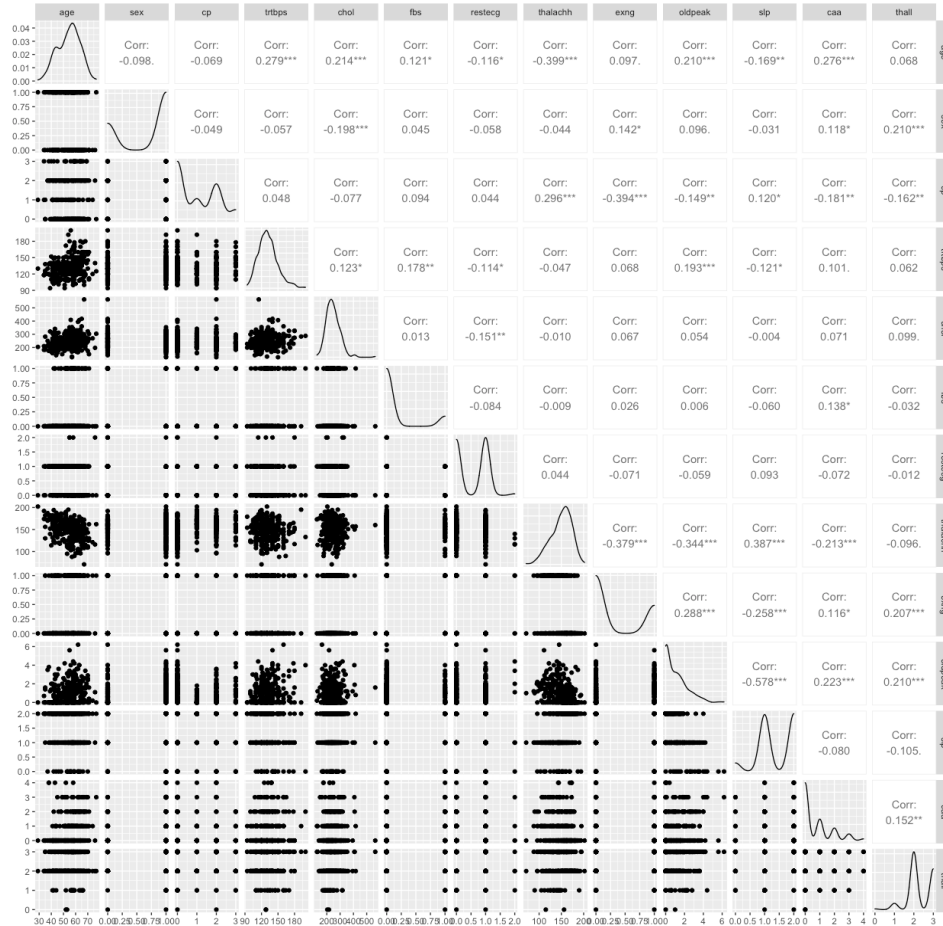
Veiem que les dades estan molt repartides entre els pacients que tenen una

probabilitat major i menor de patir un atac de cor segons els factors registrats.

Vegem la probabilitat de patir un atac de cor en base a les diferents variables:



Finalment, creem un gràfic on es mostrin les possibles relacions existents entre les variables per veure, de forma ràpida, si podem veure algunes correlacions de forma directa:



Observant els diversos factors que es troben dins el conjunt de dades i tots els gràfics creats, sembla que la variància entre els diferents valors, així com la, a priori, poca correlació entre les variables i la seva relació amb la probabilitat de patir un atac de cor podria ser interessant mantenir tots els atributs que tenim per poder utilitzar-los durant la resta de l'anàlisi.

## Integració, selecció i reducció de dades

La **integració** consisteix en la combinació de dades de diferents fonts, per tal de crear una estructura de dades coherent. En el cas d'estudi, aquesta integració ja està realitzada i tenim, de cada observació, totes les variables en columnes.

La **selecció** consisteix en filtrar o seleccionar les dades d'interès. A partir de l'anàlisi exploratori que hem realitzat consideram vàlides totes les dades i no cal filtrar ni reduir la dimensionalitat del dataset.

## Neteja de les dades

### Valors nul·ls

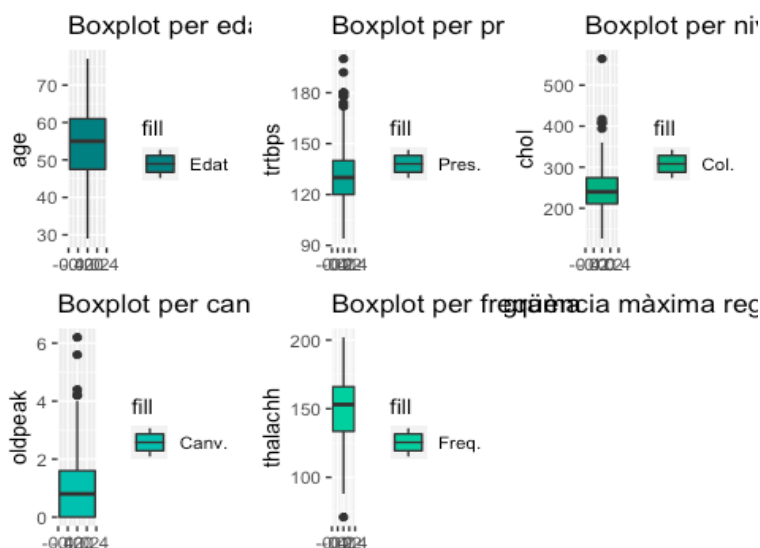
Comprobem si existeixen valors nul·ls en les dades, tot i que, aparentment, al revisar els resultats del resum estadístic anterior sembla que no existeixi cap valor d'aquest tipus.

Es confirma que no existeixen valors d'aquest tipus en el conjunt de dades, pel que no serà necessari realitzar cap acció al respecte. En cas de tenir valors d'aquest tipus hauriem de pensar si treure'ls o bé realitzar una aproximació del possible valor a partir de la resta de valors de la variable en qüestió.

### Valors atípics

Respecte a les variables categòriques, podem comprovar en el resum estadístic que no hi ha cap valor fora del rang vàlid; per tant no tenen valors atípics.

Respecte a les variables numèriques, utilitzarem els diagrames de caixa per poder veure ràpidament si existeix algun valor atípic o *outlier* en el joc de dades:



S'observen valors extrems a partir dels gràfics per algunes de les variables. No obstant, veiem que són valors que estan dintre dels paràmetres que es poden considerar com a vàlids:

- Pressió arterial per sobre de 170: tot i que estigui per sobre del normal, és un valor possible i, per tant, els mantindrem dintre del joc de dades.
- Colesterol per sobre de 400: novament ens trobem davant de valors extrems, però que es troben dins d'un rang possible, pel que mantindrem aquestes observacions dins el joc de dades a analitzar.

- Canvis en l'electrocardiograma per sobre de 4: tot i que està fora dels valors més comuns, no està tant distanciat com per considerar treure'ls de l'anàlisi.
- Freqüència màxima registrada per sota de 100: tot i que existeixi un valor més baix, aquest és un valor possible i, per tant, el mantindrem dins el joc de dades.

En resum, no s'han trobat valors anòmals que puguin considerar-se fora dels valors possibles, tot i que sí hi ha alguns valors extrems degut a les condicions físiques i/o de salut dels diferents pacients. Possiblement aquests valors poden ser importants a l'hora de fer estimacions. No procedeix eliminar cap valor atípic.

## Normalització i discretització

La normalització de les dades ens permet obtenir valors en escales que permetin comparar la magnitud de forma similar entre els diferents rangs de valors que tenen les variables.

Per altra banda, la discretització ens permet agrupar observacions numèriques per tenir noves categories que puguin resultar útils durant l'anàlisi (per exemple en algorismes de classificació d'arbre):

### Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
sex	0	1	FALSE	2	Mas: 207, Fem: 96
cp	0	1	FALSE	4	Ang: 143, Dol: 87, Ang: 50, Sen: 23
fbs	0	1	FALSE	2	Nor: 258, Alt: 45
restecg	0	1	FALSE	3	Ano: 152, Nor: 147, Hip: 4
exng	0	1	FALSE	2	No : 204, Si : 99
slp	0	1	FALSE	3	Pen: 142, Pen: 140, Pen: 21
caa	0	1	FALSE	5	Sen: 175, Obs: 65, Obs: 38, Obs: 20
thall	0	1	FALSE	4	Pre: 166, Pre: 117, No : 18, ? : 2
age.d	0	1	FALSE	5	[53: 104, [61: 71, [46: 57, [40: 52
trtbps.d	0	1	FALSE	3	[11: 153, [13: 97, [94: 53
chol.d	0	1	FALSE	5	[23: 101, [19: 100, [28: 65, [12: 32
thalachh.d	0	1	FALSE	3	[13: 141, [16: 85, [71: 77
resultat	0	1	FALSE	2	Ata: 165, Ata: 138

### Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	0.00	1.0	-2.79	-0.76	0.07	0.73	2.49	—■■—
trtbps	0	1	0.00	1.0	-2.15	-0.66	-0.09	0.48	3.90	■■—
chol	0	1	0.00	1.0	-2.32	-0.68	-0.12	0.54	6.13	■—
thalachh	0	1	0.00	1.0	-3.43	-0.70	0.15	0.71	2.29	—■■■
oldpeak	0	1	0.00	1.0	-0.90	-0.90	-0.21	0.48	4.44	■—
output	0	1	0.54	0.5	0.00	0.00	1.00	1.00	1.00	■—■

## Anàlisi de components principals

Una possible forma de reduir-ne la dimensionalitat és considerar els components principals de les variables numèriques:

```
pca.acc <- prcomp(dfHANnum, scale. = TRUE)

summary( pca.acc )

## Importance of components:
##               PC1      PC2      PC3      PC4      PC5
## Standard deviation   1.3441 1.0380 0.9399 0.8713 0.68799
## Proportion of Variance 0.3613 0.2155 0.1767 0.1518 0.09467
## Cumulative Proportion 0.3613 0.5768 0.7535 0.9053 1.00000
```

Podem veure que partim de cinc variables i necessitam quatre per arribar a descriure el 90% de la variabilitat total, per la qual cosa no suposa una gran reducció de dimensionalitat.

## Anàlisi de les dades

### Test de normalitat

Per facilitar els càlculs futurs, es procedirà a un anàlisi de la normalitat dels valors numèrics del conjunt de dades, el que ens permetrà saber si és possible aplicar certs tests més endavant.

Comencem amb un test de Shapiro-Wilk de normalitat:

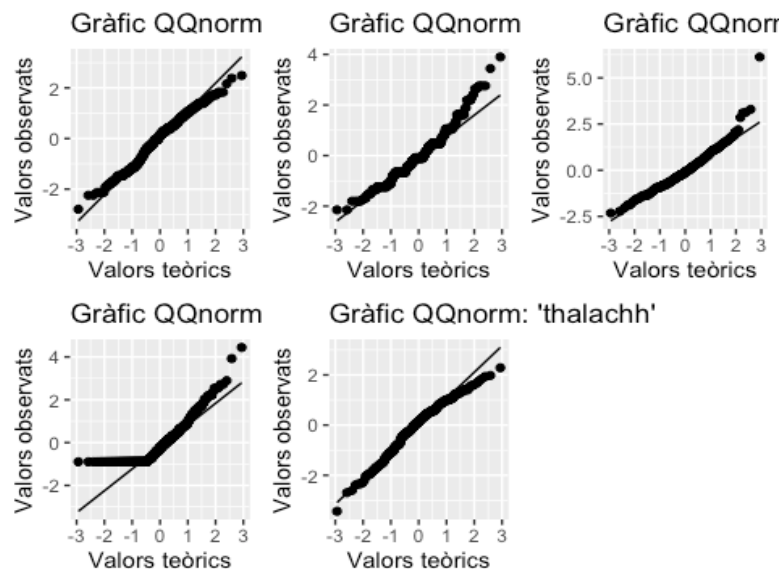
```
apply( dfHANnum, 2, shapiro.test )

## $age
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.98637, p-value = 0.005798
##
##
## $trtbps
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.96592, p-value = 1.458e-06
##
##
## $chol
##
```



```
## Shapiro-Wilk normality test
##
## data: newX[, i]
## W = 0.94688, p-value = 5.365e-09
##
##
## $thalachh
##
## Shapiro-Wilk normality test
##
## data: newX[, i]
## W = 0.97632, p-value = 6.621e-05
##
##
## $oldpeak
##
## Shapiro-Wilk normality test
##
## data: newX[, i]
## W = 0.84418, p-value < 2.2e-16
```

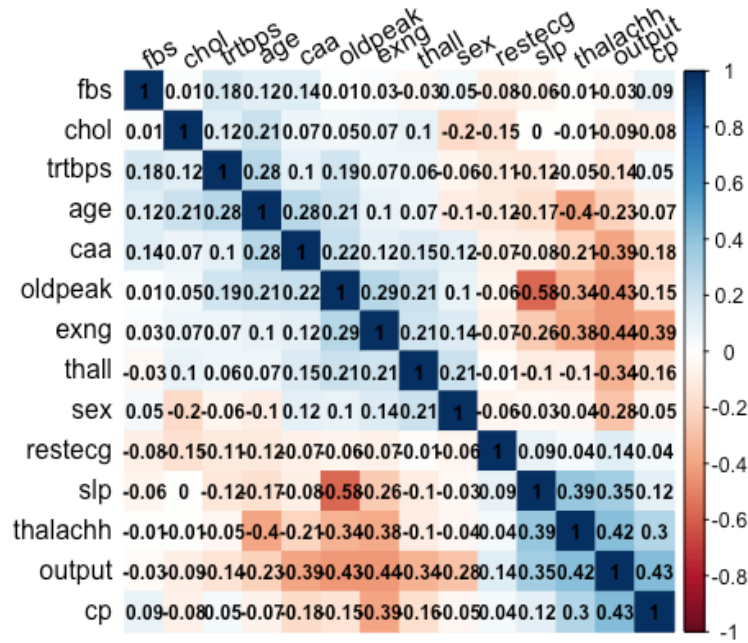
Tots els test mostren que les variables no segueixen una distribució totalment normal. No obstant, es compleixen els requisits per poder aplicar el *Teorema del Límit Central*. Mostrem a continuació els gràfics QQ per mostrar gràficament la normalitat de les dades:



Amb els resultats obtinguts, podem observar que les variables numèriques, tot i no tenir una distribució normal, sí semblen tendir a aquesta. L'únic cas que podríem considerar que no és tant així seria per la variable *oldpeak*, ja que és la única que no té una distribució tant propera a la línia que podríem considerar que defineix la normalitat.

## Correlació de les variables

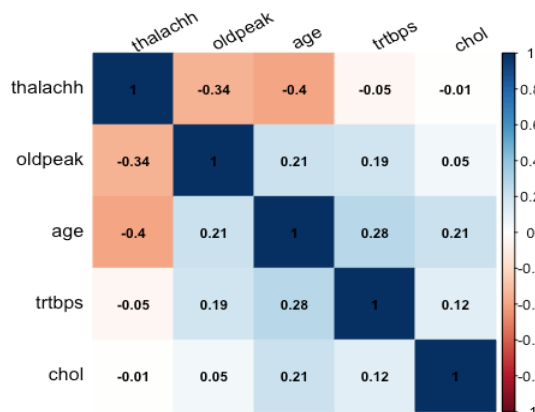
Tot i que ja s'ha vist anteriorment que semblava que no hi havia correlacions molt fortes entre les diferents variables, procedim a analitzar-ho més a fons:



La més forta d'aquestes correspon amb una correlació negativa entre el patró de canvi de l'electrocardiograma en una situació d'estrés (variable *slp*) i el canvi en el segment ST de l'electrocardiograma (variable *oldpeak*). Mentre que la més forta amb valor positiu correspon a una correlació entre el tipus de dolor toràctic (variable *cp*) i la probabilitat d'un atac de cor (variable *output*).

Cal tenir en compte que aquestes correlacions no són molt fortes i el signe ens indica que quan una augmenta o disminueix, la variable correlacionada actua de la mateixa manera en certa proporció segons la potència d'aquesta correlació.

Podem fer la correlació únicament sobre les variables numèriques:



Amb els resultats obtinguts, no hi ha cap evidència que, per les correlacions existents, es pugui disminuir el nombre de variables.

## Regressió logística

La regressió logística és un model que intentar predir el resultat (output) a partir de la resta de variables mitjançant un model de regressió estadística.

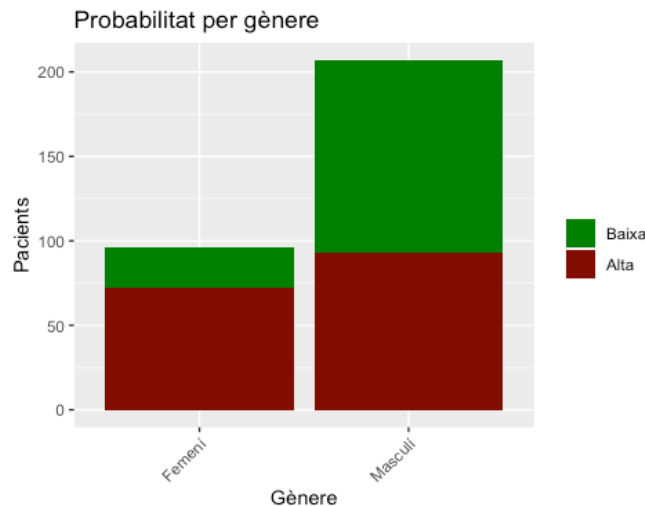
```
regressio <- lm(output ~ age + sex + cp + trtbps + chol + fbs + restecg +
thalachh + exng + oldpeak + slp + caa + thall, data = dfHAN)
summary(regressio)
```

```
##
## Call:
## lm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
##     restecg + thalachh + exng + oldpeak + slp + caa + thall,
##     data = dfHAN)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01468 -0.18519  0.03069  0.22575  1.02624
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.30215    0.25796   1.171 0.242466
## age                           0.02464    0.02406   1.024 0.306638
## sexMasculí                    -0.16407    0.04823  -3.402 0.000766 ***
## cpAngina típica               -0.16886    0.06397  -2.640 0.008767 **
## cpDolor no relacionat amb angina  0.05637    0.06102   0.924 0.356389
## cpSense dolor toràctic         0.09685    0.08875   1.091 0.276084
## trtbps                       -0.04104    0.02137  -1.920 0.055844 .
## chol                         -0.01653    0.02084  -0.793 0.428416
## fbsNormal                    -0.03335    0.05749  -0.580 0.562362
## restecgHipertrofia ventricular -0.13405    0.17685  -0.758 0.449098
## restecgNormal                -0.04699    0.04083  -1.151 0.250823
## thalachh                     0.04285    0.02563   1.672 0.095634 .
## exngSi exercici              -0.09352    0.04999  -1.871 0.062437 .
## oldpeak                     -0.04828    0.02674  -1.806 0.072023 .
## slpPendent descendent         0.13900    0.04953   2.806 0.005364 **
## slpPendent plana              0.06540    0.08593   0.761 0.447281
## caaObstrucció de quatre vasos  0.38693    0.16945   2.284 0.023149 *
## caaObstrucció de tres vasos   0.04660    0.09558   0.488 0.626212
## caaObstrucció en un vas       0.07225    0.07324   0.987 0.324711
## caaSense obstrucció           0.34261    0.06825   5.020 9.2e-07 ***
## thallNo hi ha antecedents      0.22837    0.25437   0.898 0.370068
## thallPresència de defecte reversible 0.07849    0.24369   0.322 0.747623
## thallPresència defecte fix     0.28800    0.24183   1.191 0.234694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.335 on 280 degrees of freedom
## Multiple R-squared:  0.5818, Adjusted R-squared:  0.549
## F-statistic: 17.71 on 22 and 280 DF, p-value: < 2.2e-16
```

Podem veure que amb la regressió múltiple  $R^2 = 0.5489541$ , que ens indica que, amb aquest model, tot el conjunt de variables disponibles expliquen el 54.9% de la variable objetivo output.

## Test d'hipòtesi

Ens demanem si les dones tenen major probabilitat de tenir un atac de cor que els homes. Podem veure la gràfica comparativa per gènere:



En la mostra hi ha un percentatge major de dones amb alta probabilitat de tenir un atac de cor que d'homes; però ens demanem si aquesta diferència és significativa ja que podria ser fruit de l'atzar de la mostra.

## Hipòtesi

La hipòtesi nul·la, o  $H_0$  és: **La probabilitat de tenir un atac de cor de les dones és igual a la probabilitat de tenir un atac de cor dels homes**

La hipòtesi alternativa, o  $H_1$  és: **La probabilitat de tenir un atac de cor de les dones és superior a la probabilitat de tenir un atac de cor dels homes**

## Contrast

Es tracta d'un test de contrast unilateral sobre dues mostres (les probabilitats de tenir atac de cor de les dones i les dels homes) en relació a les seves probabilitats de tenir un atac de cor.

Per determinar el test a aplicar, és pertinent aplicar el teorema del límit central que estableix que el contrast d'hipòtesi sobre una mitjana d'una mostra s'aproxima a una distribució normal encara que la població original no segueixi una distribució normal, sempre que la mida de la mostra sigui suficientment gran (superior a 30). Això ja ho hem vist en l'apartat del Test de normalitat.

Perquè es puguin donar les condicions para aplicar el teorema del límit central, el tamany de les dues mostres ha de ser superior a 30; vegem-ho:

```
mostra_dones <- dfHAN$output[dfHAN$sex=="Femení"]
mostra_homes <- dfHAN$output[dfHAN$sex=="Masculí"]
print( paste("La mostra de les dones és de", length(mostra_dones), "observacions") )

## [1] "La mostra de les dones és de 96 observacions"

print( paste("La mostra dels homes és de", length(mostra_homes), "observacions") )

## [1] "La mostra dels homes és de 207 observacions"
```

Podem veure que les dues mostres compleixen les hipòtesis del teorema del límit central i, per tant podem assumir que segueixen lleis normals.

Les variàncies de les poblacions (la població de les dones i dels homes) no les coneixem. Només tenim, òbviament, les respectives variàncies mostrals. A continuació, necessitam saber si les variàncies de les dues poblacions són iguals o no, Per això cal fer un **test d'homoscedasticitat**:

Assumim (per l'argument anterior) que les dues mostres correponen a dues poblacions normals independents  $N(\mu_1, \sigma_1)$  i  $N(\mu_2, \sigma_2)$ . Aleshores la variable aleatòria següent segueix una distribució F d'Snedecor amb  $n_1 - 1$  i  $n_2 - 1$  graus de llibertat:

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

On  $s_1$  i  $s_2$  són les desviacions estàndards mostrals i  $\sigma_1$  i  $\sigma_2$  les desviacions poblacionals.

Sota la hipòtesi nul·la  $H_0: \sigma_1^2 = \sigma_2^2$ , el test estadístic és:

$$f_{obs} = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$

On F és una distribució d'Snedecor amb  $n_1 - 1$  i  $n_2 - 1$  graus de llibertat.

La hipòtesi alternativa  $H_1: \sigma_1^2 \neq \sigma_2^2$

Facem el càlcul en R:

```
var.test( mostra_dones, mostra_homes)

##
## F test to compare two variances
##
## data:  mostra_dones and mostra_homes
## F = 0.76208, num df = 95, denom df = 206, p-value = 0.1343
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5455293 1.0885394
```

```
## sample estimates:
## ratio of variances
##      0.7620767
```

Podem veure que el valor de l'estadístic cau en la zona d'acceptació de l'hipòtesi nul·la. De fet, obtenim un valor de  $p$  superior a  $\alpha$  i, per tant, no podem rebutjar la hipòtesi nul·la. Per tant, les variances poblacionals són essencialment iguals.

En resum, el test a aplicar és el corresponent a dues mostres independents corresponents a poblacions que se poden aproximar a una llei normal, sobre la mitjana amb variances desconegudes però iguals.

### Càlculs

El test estadístic de la diferència de les dues mitjanes segueix una distribució t d'Student amb  $n_1 + n_2 - 2$  graus de llibertat:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

El valor  $S$  és la desviació típica comuna que es calcula com:

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

On  $s_1^2$  i  $s_2^2$  són les variances mostrals estimades.

Sota la hipòtesi nul·la  $H_0: P(t_{\alpha/2} \leq t_{obs} \leq t_{1-\alpha/2}) = 1 - \alpha$

Per tant, la zona d'acceptació és  $[t_{\alpha/2}, t_{1-\alpha/2}]$ .

Facem el càlcul en R:

```
alfa      <- 0.05
# Tamany de les mostres
nd <- length(mostra_dones)
nh <- length(mostra_homes)
# Mitjanes de les mostres
md <- mean(mostra_dones)
mh <- mean(mostra_homes)
# Desviació estàndard de les mostres
sd <- sd(mostra_dones)
sh <- sd(mostra_homes)

# Valor de l'estadístic
S      <- sqrt( ( (nd-1)*sd^2 + (nh-1)*sh^2 ) / (nd+nh-2) )
tobs   <- (md-mh)/(S*sqrt( 1/nd + 1/nh ))
# Valors crítics
tcrit.L <- qt( alfa/2, df=nd+nh-2)
tcrit.U <- qt( alfa/2, df=nd+nh-2, lower.tail=FALSE)
# Valor de p
```

```
valor_p <- pt( abs(tobs), df=nd+nh-2, lower.tail=FALSE)*2
print(paste("Valor observat:", round(tobs,2)))

## [1] "Valor observat: 5.08"

print(paste("Interval d'acceptació: [",round(tcrit.L, 2),",", round(tcrit.U, 2), "]" ))

## [1] "Interval d'acceptació: [ -1.97 , 1.97 ]"

print(paste("Valor de p:", round(valor_p, 2)))

## [1] "Valor de p: 0"
```

### Interpretació

Podem veure que l'observació de l'estadístic considerat és 5.08, valor que cau fora la zona d'acceptació: [-1.97, 1.97]. Per tant podem rebutjar la hipòtesi nul·la amb un nivell de confiança del 95%.

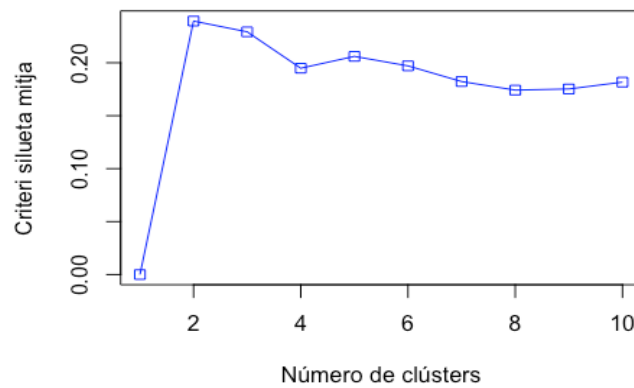
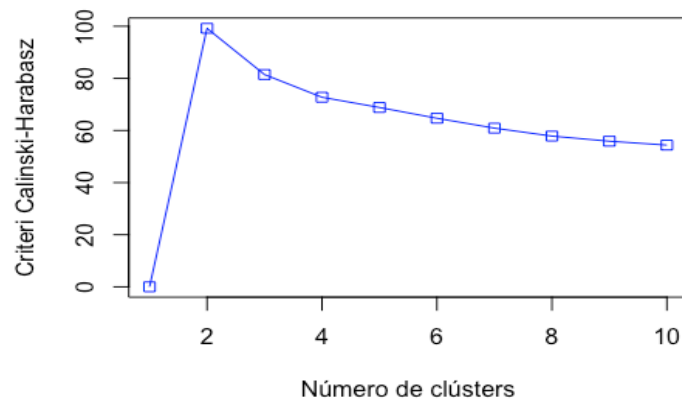
El valor de p (probabilitat de l'error que s'estaria cometent si es rebutja la hipòtesi nul·la essent aquesta certa) és de 0. Un valor inferior a  $\alpha$ .

Com a conseqüència, hem de rebutjar la hipòtesi nul·la a favor de l'alternativa: **La probabilitat de tenir un atac de cor de les dones és superior a la probabilitat de tenir un atac de cor dels homes**

### Model no supervisat

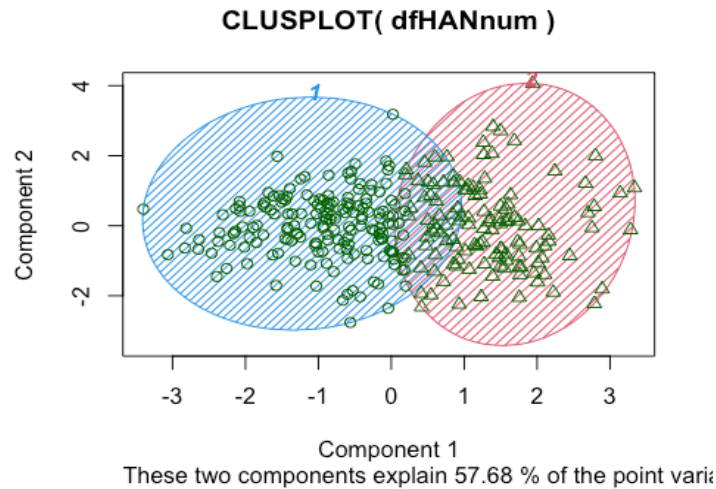
Anem a generar un model no supervisat basat en l'algorisme *k\_means*. El paràmetre fundamental és el valor de k, que coincideix amb el nombre de grups que volem trobar. En realitat en el nostre problema volem classificar les observacions en dos grups: els que tenen risc d'atac de cor i els que no; per tant, el valor que ens interessa és k=2; però abans analitzarem quin és el valor de k amb el que podem obtenir uns clústers més diferenciats:

### Estimació del millor valor de k pel mètode Calinski-Harabasz



Podem veure que el millor valor de  $k$  és 2, per tant, fem una clusterització en 2 grups i analitzem la qualitat. Per poder visualitzar el resultat, ho projectarem sobre les dues primeres components principals:





```
sk2 <- silhouette(y_cluster2, d)
qualitat2 <- mean(sk2[,3])
```

```
qualitat2
```

```
## [1] 0.2435173
```

La qualitat de la clusterització segons la silueta és 0.2435173; un valor molt baix. Visualment no s'aprecia una clara separació dels dos grups.

Comparem els dos grups trobats amb la variable output:

```
resultat1 <- matrix( rep(0,4), nrow=2, ncol = 2,
  dimnames=list( c("Grup 1", "Grup 2"),
    c("Atac probable", "Atac poc probable")))

for (i in 1:nrow(dfHANnum)) {
  resultat1[fit2$cluster[i],dfHAN$output[i]+1] <-
  resultat1[fit2$cluster[i],dfHAN$output[i]+1] + 1
}

total <- sum(resultat1)

opc1 <- round(100 * (resultat1["Grup 1", "Atac probable"] + resultat1["Grup 2", "Atac poc
probable"] ) / total, 1)
opc2 <- round( 100 * (resultat1["Grup 1", "Atac poc probable"] + resultat1["Grup 2",
"Atac probable"] ) / total, 1)
millor_opc <- max( opc1, opc2)
resultat1

##      Atac probable Atac poc probable
## Grup 1           50           128
## Grup 2           88           37

print(paste("Si Grup 1 = Atac probable llavors, percentatge encert:", opc1 ))

## [1] "Si Grup 1 = Atac probable llavors, percentatge encert: 28.7"

print(paste("Si Grup 1 = Atac poc probable llavors, percentatge encert:", opc2))
```

```
## [1] "Si Grup 1 = Atac poc probable llavors, percentatge encert: 71.3"
```

Podem veure que la millor opció d'interpretació dels resultats dels dos grups ens dona un percentatge d'encert de 71.3%.

### Conclusió:

Els dos grups trobats considerats com a clústers no estan gens separats en distància. De fet si calculam la mesura de la silueta sobre la classificació coneguda, obtenim una qualitat molt baixa de 0.2393498. L'algorisme de clusterització només encerta en un 71.3% dels casos amb la classificació real.

### Model supervisat

Ara aplicarem un model supervisat basat en la generació de regles de classificació a partir d'un arbre de decisions. Farem servir les dades categòriques amb la discretització de les variables numèriques; però abans hem de separar el data set en dos: un per entrenar l'algorisme i l'altre per validar-ho. Farem servir la proporció habitual de 2/3 per entrenament i 1/3 per test.

Una vegada feta la separació aleatòria de les mostres, convé realitzar una mínima anàlisi de dades per a assegurar-nos de no obtenir classificadors esbiaixats pels valors que conté cada mostra. En aquest cas, verificarem que la proporció del prèstecs bons i dolents és més o menys constant en els dos conjunts:

Es pot veure que els percentatges de probabilitat d'atacs de cor del conjunt d'entrenament i del conjunt de test són semblants (45.5% i 45.5% respectivament). Per tant, consideram correctes els jocs d'entrenament i el de test.

Anem a generar el model de regles de decisió per determinar si la probabilitat d'atac de cor és o no alta. Ho farem amb l'algorisme *random Forest*, tècnica en què genera diversos classificadors amb els seus arbres de decisió per, finalment, registrar-ne tots els resultats i determinar-ne la classe final.

```
rf <- randomForest(resultat ~ ., data=train)
print(rf)

##
## Call:
## randomForest(formula = resultat ~ ., data = train)
##           Type of random forest: classification
##           Number of trees: 500
## No. of variables tried at each split: 3
##
##           OOB estimate of  error rate: 13.86%
## Confusion matrix:
##           Atac poc probable Atac probable class.error
## Atac poc probable           50           5 0.09090909
## Atac probable              9           37 0.19565217
```

Validem el model amb el joc de test:

```
pred = predict(rf, newdata=test[1:12])
cm = table(test[,13], pred)

preciso_model <- round( 100 * sum(diag(cm)) / sum(cm), digits = 2 )
print(sprintf("La precisi3n del 6rbol es: %.1f %%", preciso_model))

## [1] "La precisi3n del 6rbol es: 80.2 %"
```

Podem veure que obtenim un model de decisi3 d'arbre amb una taxa de predicci3 del (80.2%).

Vegem la taula de confusi3:

```
CrossTable(test[,13], pred, prop.chisq = FALSE, prop.c = FALSE, prop.r =TRUE,dnn =
c('Reality', 'Prediction'))

##
##
## Cell Contents
## |-----|
## |                      N |
## |      N / Row Total |
## |      N / Table Total |
## |-----|
##
##
## Total Observations in Table:  202
##
##
##      Reality | Prediction
##      Reality | Atac poc probable | Atac probable | Row Total |
## -----|-----|-----|-----|
## Atac poc probable |          92 |          18 |          110 |
##                  |      0.836 |      0.164 |      0.545 |
##                  |      0.455 |      0.089 |              |
## -----|-----|-----|-----|
## Atac probable |          22 |          70 |          92 |
##                  |      0.239 |      0.761 |      0.455 |
##                  |      0.109 |      0.347 |              |
## -----|-----|-----|-----|
## Column Total |          114 |          88 |          202 |
## -----|-----|-----|-----|
##
##
```

## Resoluci3 del problema

A partir del fitxer de dades, hem fet unes tasques de preprocessament:

- An6lisi exploratori
- Neteja de dades, on hem vist que les dades estaven molt netes i 6nicament hem fet una normalitzaci3 i una discretitzaci3 de valors.

Amb les dades preprocessades, hem fet una s6rie d'an6lisi:

- Una anàlisi de components principals, on hem vist que amb 4 components principals es pot explicar el 90% de la variabilitat total.
- Un test de normalitat, que han sortit negatius; no obstant és possible aplicar el teorema del límit central perquè tenim moltes dades.
- Un estudi de correlacions entre les variables, que ha conclos que aquestes son molt febles.
- Una regressió logística, que ens ha permet disposar d'un model de predicció de la probabilitat d'atac de cor amb una fiabilitat del 55%.
- Un test d'hipòtesi, que ens ha permet confirmar una diferència que s'intuïa en les gràfiques: que les dones tenen més probabilitat de tenir un atac de cor que els homes.
- Hem aplicat un model no supervisat per clusteritzar les dades en dos grups (una vegada normalitzades les dades), i hem onbtingut un model predictiu amb un percentatge d'encerts del 71%.
- Per últim hem obtingut un model supervisat basat en un algorisme de classificació *randomForest* amb una taxa de predicció del 80%.

En definitiva, amb aquesta pràctica hem seguit les fases del cicle de vida de les dades: captura, emmagazematge, preprocesat, anàlisi, visualització i publicació.