

Pràctica 2 – Tipologia i cicle de vida de les dades

Assignatura: M2.951 / Semestre: 2022/23-2 / Data: Juny de 2023

Autors:

- Francisco J. Bastida López – fbastidal@uoc.edu
- Ivan Benaiges Trenchs – ibenaiges@uoc.edu

Resolució dels apartats

1. Descripció del dataset

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 la predicció de 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, 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 permet trobar un model per determinar la probabilitat de patir un atac de cor segons els valors de les diferents variables 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
- **cp**: tipus de dolor toràcic que experimenta el pacient
- **trtbps**: pressió arterial en repòs (mm Hg)
- **chol**: nivell de colesterol (md/dl)
- **fb**: nivell de sucre en sang en dejú
- **restecg**: resultat de l'electrocardiograma en repòs
- **thalachh**: freqüència cardíaca màxima registrada pel pacient durant les proves realitzades
- **exng**: angina provocada per l'exercici
- **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
- **caa**: nombre de vasos sanguinis coronaris que mostren obstrucció o estenosi significativa
- **thall**: relacionat amb una malaltia hereditària de la sang anomenada talassèmia

- **output:** probabilitat de patir un atac de cor

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 moltes 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. Llavors s'ha descartat aquest fitxer.

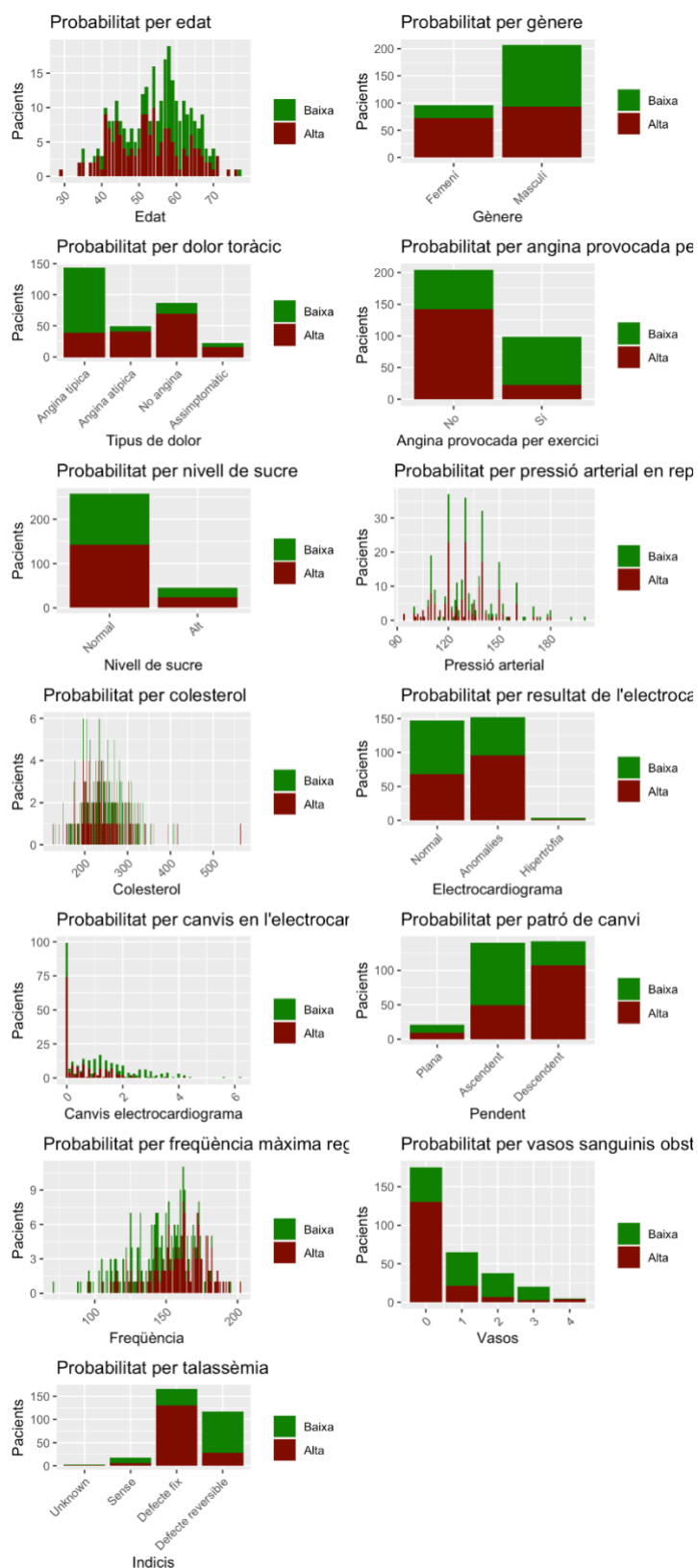
2. Integració i selecció de les dades d'interès

Les dades estan integrades en un únic fitxer CSV que hem carregat. El conjunt de dades té unes dimensions de 303 observacions i 14 variables. És, per tant, un fitxer petit i no s'ha considerat reduir la dimensionalitat ni s'ha aplicat cap filtre a les dades originals al no considerar-ho necessari.

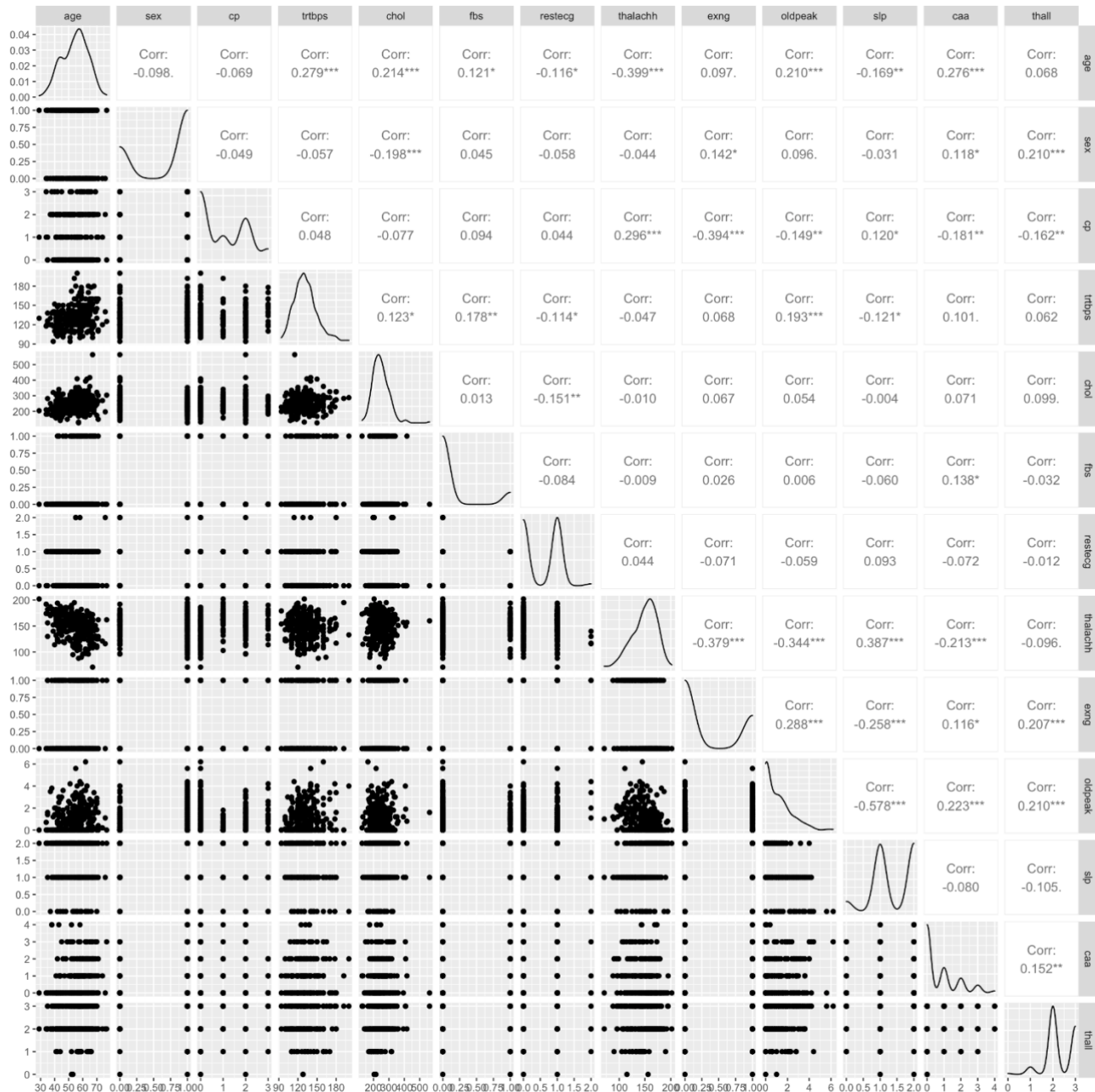
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	54.37	9.08	29	47.5	55.0	61.0	77.0	
sex	0	1	0.68	0.47	0	0.0	1.0	1.0	1.0	
cp	0	1	0.97	1.03	0	0.0	1.0	2.0	3.0	
trtbps	0	1	131.62	17.54	94	120.0	130.0	140.0	200.0	
chol	0	1	246.26	51.83	126	211.0	240.0	274.5	564.0	
fbs	0	1	0.15	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	

3. Neteja de les dades

En el fitxer resultat de l'anàlisi (pra2.html) es pot consulta tot una **anàlisi exploratòria** preliminar de les dades per conèixer la distribució de les seves variables. Especialment diferenciant els pacients amb alta i baixa probabilitat d'un atac de cor segons les diferents variables: sexe, edat, dolor toràcic, angina, nivell de sucre, pressió arterial en repòs, colesterol, resultat de l'electrocardiograma, vasos sanguinis obstruïts i talassèmia.

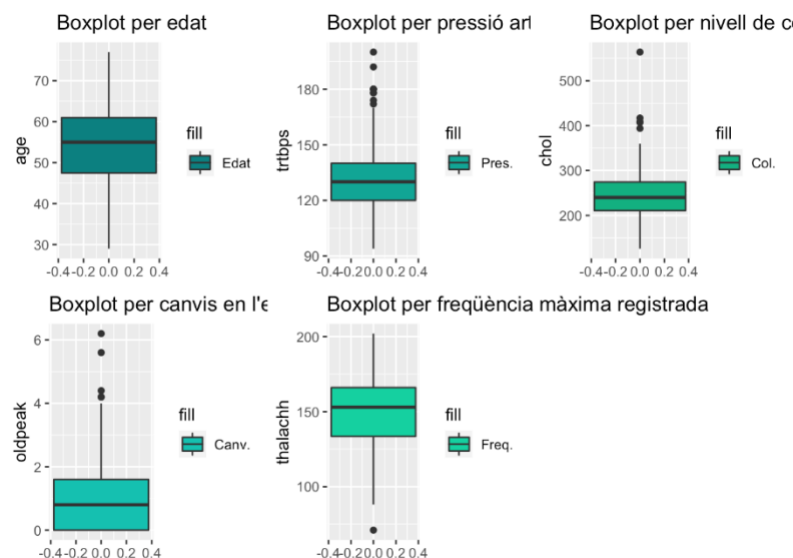


També s'ha dibuixat un núvol de punts de totes les parelles de variables per identificar correlacions visuals, juntament amb les corbes de densitat de cada variable per tenir una primera imatge de les distribucions que tenen les diferents variables.



A continuació s'ha procedit a fer una neteja de dades amb els següents resultats:

- No s'han trobat valors absents (NA) ni nuls.
- S'han identificat alguns valors atípics en algunes variables; però tots dins rangs possibles tot i que no massa habituals i, per tant, no s'han descartat.



- S'ha procedit a fer una normalització de dades numèriques escalant i centrant per obtenir mitjanes zero i desviacions 1. Aquesta tasca ajudarà a l'anàlisi posterior.
- S'han discretitzat les variables numèriques per rangs amb un algorisme de clusterització per poder aplicar posteriorment algorismes de classificació.

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

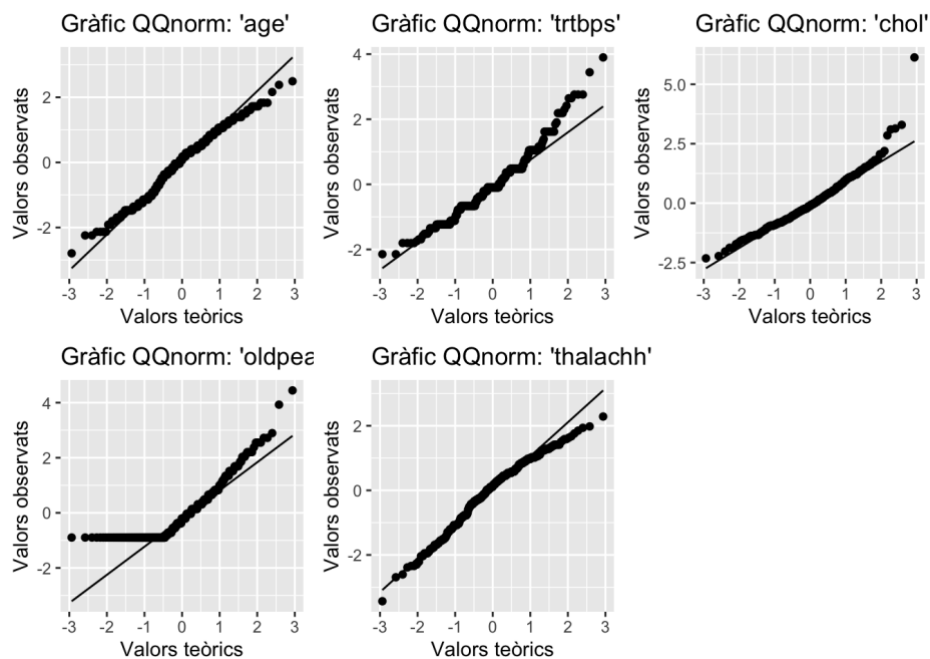
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	0.00	1.0	-2.79	-0.76	0.07	0.73	2.49	
trtbps	0	1	0.00	1.0	-2.15	-0.66	-0.09	0.48	3.90	
chol	0	1	0.00	1.0	-2.32	-0.68	-0.12	0.54	6.13	
thalachh	0	1	0.00	1.0	-3.43	-0.70	0.15	0.71	2.29	
oldpeak	0	1	0.00	1.0	-0.90	-0.90	-0.21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

- S'ha fet una anàlisi de components principals per considerar la possibilitat de reduir-ne la dimensionalitat; però finalment s'ha descartat ja que no s'ha determinat necessari.

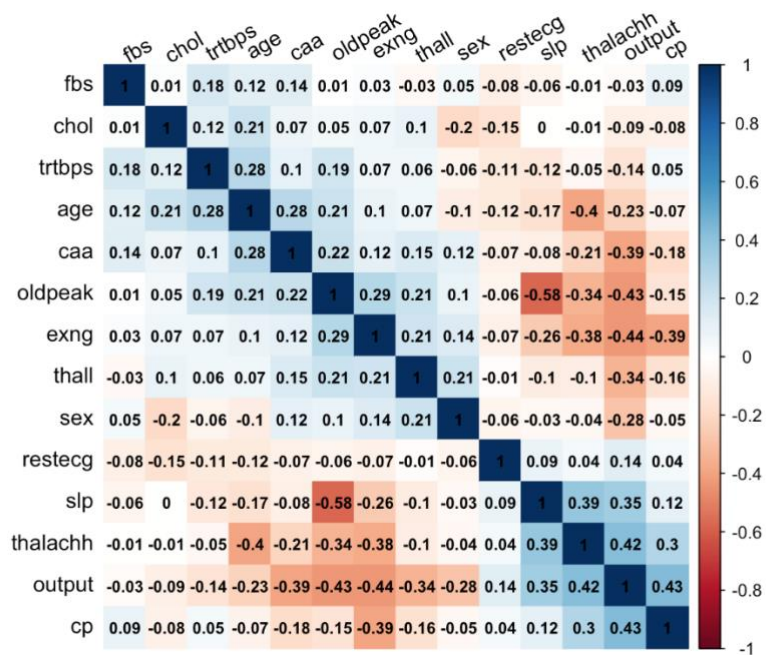
4. Anàlisi

El procés d'anàlisi ha consistit en les següents tasques:

- **Test de normalitat** sobre les diferents variables numèriques, per poder aplicar els contrast d'hipòtesi. Han sortit negatius, no obstant es donen les condicions per poder aplicar el teorema del límit central.



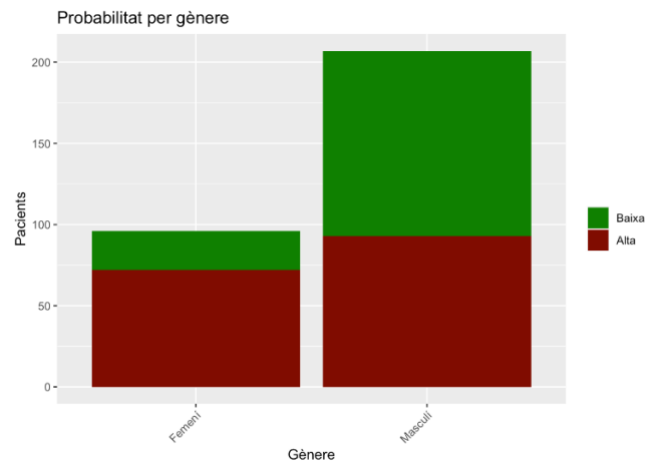
- **Estudi de correlacions** entre les diferents variables, on han aparegut algunes correlacions moderades i poc significatives.



- **Regressió logística**, per poder estimar la probabilitat de tenir un atac de cor a partir de la resta de variables; hem aconseguit un model que explica el 55% de la variabilitat de l'objectiu.


```
## 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 toràcic         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
```

- Un **test d'hipòtesi** que ens ha obligat a acceptar que les dones tenen més probabilitat de tenir un atac de cor que els homes amb un nivell de confiança del 95%. Per fer aquest test ha estat necessari fer un altre test previ per confirmar la homoscedasticitat dels dos grups: homes i dones.



```
print(paste("Valor observat:", round(tobs,2)))
```

```
## [1] "Valor observat: 5.08"
```

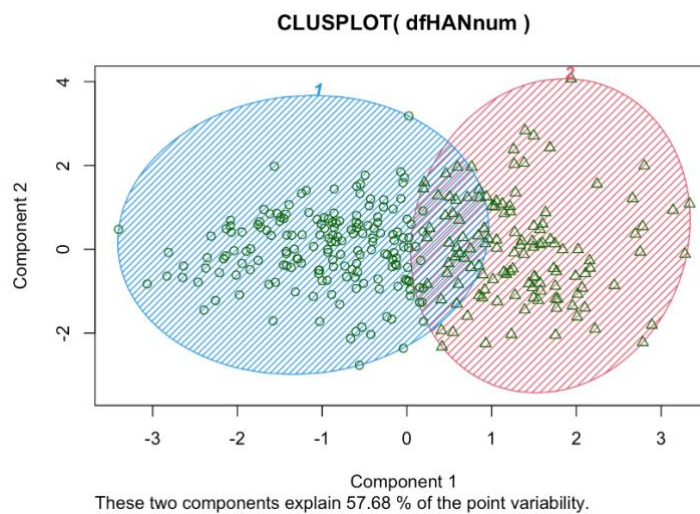
```
print(paste("Interval d'acceptació: [",round(tcrit.L, 2),",", round(tcrit.U, 2), "]"))
```

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

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

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

- A continuació hem aplicat un **algorisme de clusterització no supervisat**, concretament l'algorisme *k_means* amb dos grups, que amb la interpretació adequada ens permet predir la probabilitat d'un atac de cor amb un 71,3% d'encerts.



```
##      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"
```

- Per últim, hem aplicat un **algorisme de classificació supervisat** *randomForest*; per al qual ha estat necessari separar el conjunt de dades en dos estratificats per la variable objectiu: un per entrenar el model i l'altre per validar-ho. Aquest model té una taxa d'encert del 80%.

```
##
## 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
##
## [1] "La precisión del árbol es: 80.2 %"
```

5. Representació dels resultats

L'informe detallat extret des de R-Studio presenta durant tots els punts diferents taules i gràfiques, així com informació més detallada de l'estudi que s'ha portat a terme per aquesta pràctica, i que ajuden a entendre i interpretar correctament els resultats obtinguts:

- En l'anàlisi exploratori podem trobar diagrames de barres senzill i apilats, núvols de punts, corbes de distribució i boxplots.
- Durant la normalització s'han dibuixat gràfics QQnorm per visualitzar la normalitat.
- En la correlació de variables s'ha fet una matriu de colors.

- En el model no supervisat s'han fet diagrames de línies per ajudar a estimar el valor de k i també un núvol de punts per colors.
- En el model supervisat s'ha mostrat la taula de confusió.

6. Resolució del problema

La principal conclusió de l'anàlisi és que el joc de dades utilitzat ens permet predir amb un grau d'encert d'un 80% la probabilitat de tenir un atac de cor amb les dades disponibles. En aquest cas el més important és reduir els falsos negatius (predir baixa probabilitat d'un atac de cor quan és fals) més que els falsos positius (predir un atac de cor quan es fals) ja que els segons són una falsa alarma; però els primers no donen alarma quan en realitat hi ha risc. Segons la taula de confusió, el model té un 10% de casos de falsos negatius.

7. Codi

Tot el codi s'ha fer en R i concretament amb R Studio mitjançant R Markdown. El codi es pot descarregar des de GitHub a <https://github.com/fbastidal/Practica-2>

7. Vídeo

Vídeo explicatiu de la pràctica:

<https://drive.google.com/file/d/17M4xRM3fHLgivtAlr40ySyCwUZot0hxL/view>

Contribucions	Signatura
Investigació prèvia	Francisco Bastida López, Ivan Benaiges Trenchs
Redacció de les respostes	Francisco Bastida López, Ivan Benaiges Trenchs
Desenvolupament del codi	Francisco Bastida López, Ivan Benaiges Trenchs
Participació en el vídeo	Francisco Bastida López, Ivan Benaiges Trenchs