# Pràctica2 - M2.951_20221

Gerard Ramos Gambús: gambusrg@uoc.edu
Oriol Caravaca Müller: ocaravacam@uoc.edu
03/01/2023

# Contingut

Descripció del *dataset*....3
Integració i selecció....4
Neteja de les dades....4
Anàlisi de les dades....6
Anàlisi visual....6
Normalitat i Homoscedasticitat....8
Proves estadístiques....10
Test d'hipòtesis....11
Regressions....11
Representació dels resultats....15
Resolució del problema....16
Codi....17
Video....18

# Descripció del *dataset*

El *dataset* triat és "Heart Attack Analysis & Prediction dataset", publicat per Rashik Rahman a Kaggle amb llicencia CC0 1.0 Universal (CC0 1.0) *Public Domain Dedication.* Aquest *dataset* conte 13 variables corresponents a dades demogràfiques i resultats de proves mèdiques cardiovasculars de 303 pacients. Així mateix, ens ofereix una variable objectiu que ens indica quin es el risc de patir un infart de cada un d'aquest pacients.

Aquest *dataset* és d'interès perquè ens permet detectar factors de risc i crear models predictius que poden ajudar a la diagnosis i detecció de problemes cardiovasculars.

A continuació es detalla el tipus i descripció de les variables del *dataset*:

| Variable | Tipus | Descripció |
|---|---|---|
| **age** | Categòrica | Edat de la persona |
| **sex** | Categòrica | Gènere de la persona |
| **cp** | Categòrica | Tipus de dolor toràcic<br>1= angina típica<br>2 = angina atípica<br>3 = dolor no anginós<br>4 = asimptomàtic |
| **trtbps** | Numèrica | Pressió arterial en repòs (en mm Hg) |
| **chol** | Numèrica | Colestorol en mg/dl obtingut mitjançant el sensor IMC |
| **fbs** | Categòrica | Sucre en sang en dejú > 120 mg/dl<br>(1 = cert; 0 = fals) |
| **restecg** | Categòrica | Resultats electrocardiogràfics en repòs<br>0 = normal<br>1 = tenir una anomalia de l'ona ST-T<br>2 = mostra una hipertròfia ventricular esquerre probable o definitiva |
| **thalachh** | Numèrica | Freqüència cardíaca màxima aconseguida |
| **exng** | Categòrica | Angina induïda per l'exercici<br>(1 = sí; 0 = no) |
| **oldpeak** | Numèrica | Depressió del ST induïda per l'exercici en relació amb el repòs |
| **slp** | Categòrica | El pendent del segment ST de l'exercici màxim<br>0 = sense pendent<br>1 = pla<br>2 = pendent avall |
| **caa** | Numèrica | Nombre de vasos principals (0-3) |
| **thall** | Categòrica | Talassèmia<br>0 = nul<br>1 = defecte fixat<br>2 = normal<br>3 = defecte reversible |
| **output** | Categòrica | 0 = menys probabilitat d'atac cardíac<br>1 = més probabilitat d'atac cardíac |

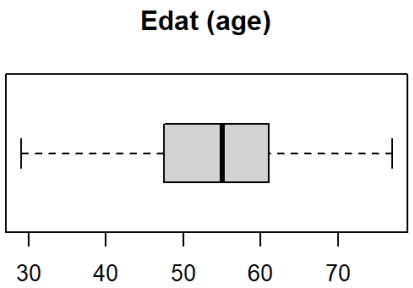
# Integració i selecció

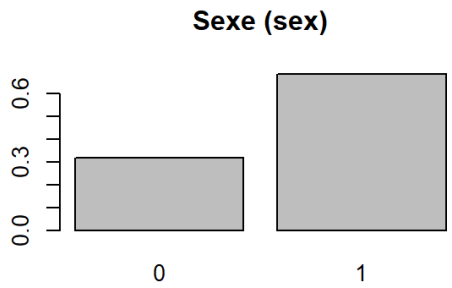
Per a la realització d'aquesta practica es contempla la utilització del *dataset* complet amb l'objectiu de fer-ne un anàlisis i implementar un model de regressió logística.

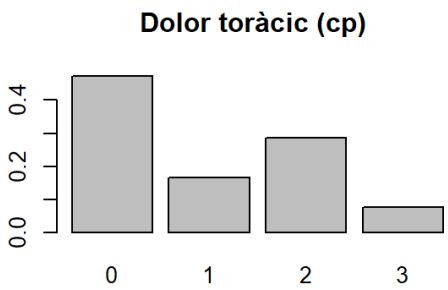
Posteriorment de l'estudi del model de regressió logística se'n descarten les variables age, trtbps, fbs, restecg i oldpeakper considerar-se insignificants amb un nivell de significació per sota del 90%.

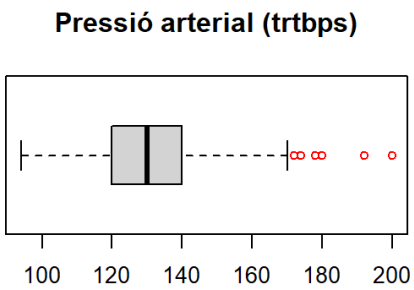
# Neteja de les dades

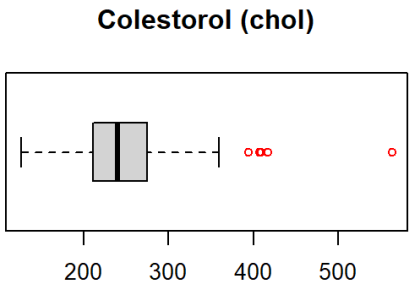
Per a la neteja de les dades s'han fer comprovacions sobre tot el conjunt de dades, tenint en compte si hi havia espais en blanc o valors Nulls. Addicionalment s'ha passat les variables categòriques a factors i se n'ha visualitzat la distribució per detectar-ne valors atípics.

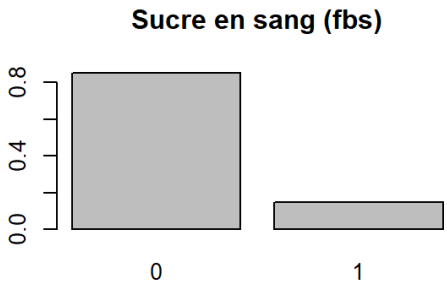


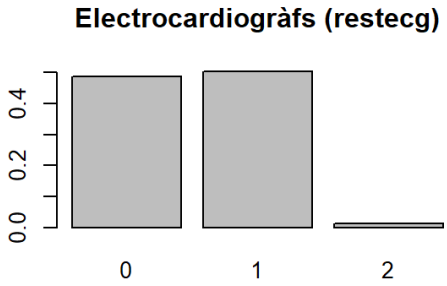



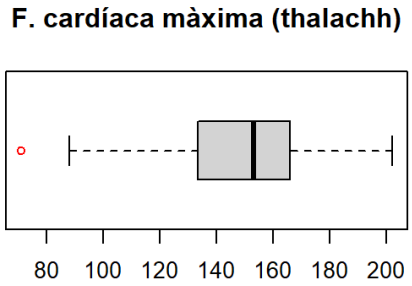



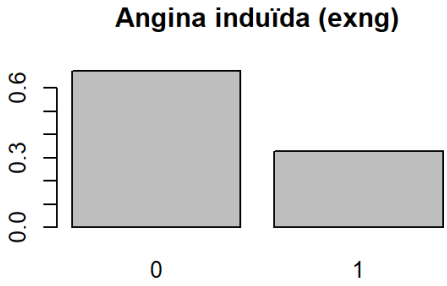



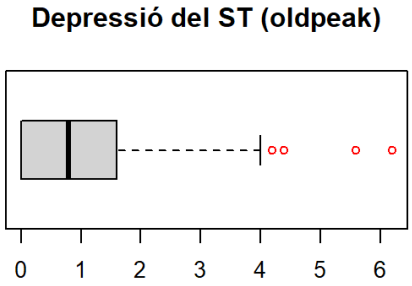

Colestorol (chol)
200
300
400
500

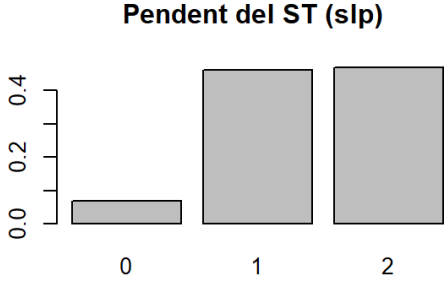

Sucre en sang (fbs)
0.0
0.4
0.8
0
1

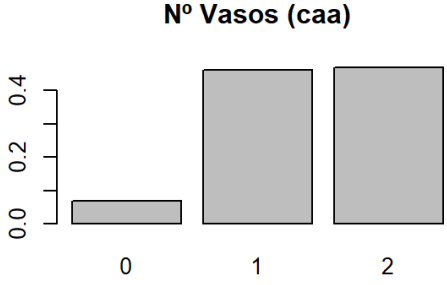

Electrocardiogràfs (restecg)
0.0
0.2
0.4
0
1
2

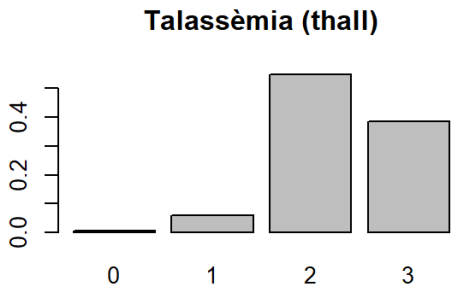

F. cardíaca màxima (thalachh)
80
100
120
140
160
180
200

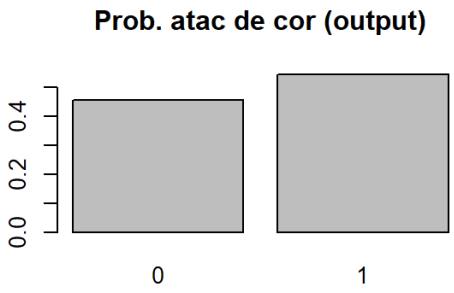

Angina induïda (exng)
0.0
0.3
0.6
0
1


Depressió del ST (oldpeak)
0
1
2
3
4
5
6


Pendent del ST (slp)
0.0
0.2
0.4
0
1
2


Nº Vasos (caa)
0.0
0.2
0.4
0
1
2



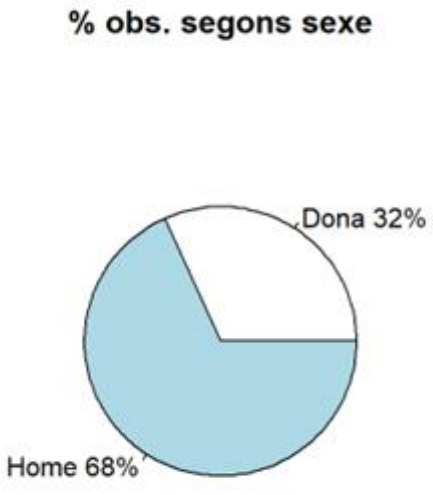

Seguidament, per cada variable continua amb valors atípics se'n eliminen els valors i s'imputen utilitzant l'algoritme kNN amb una k de 5. Per les variables categòriques thall i caa s'imputen els valors atípics manualment.

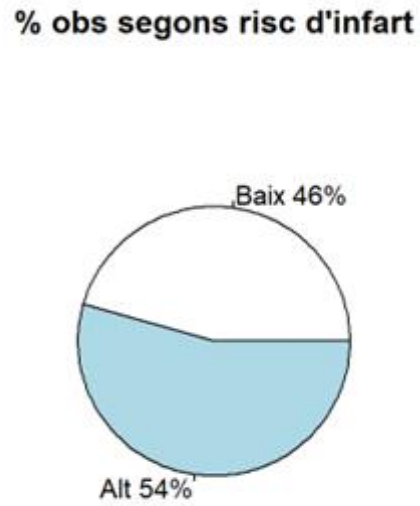
# Anàlisi de les dades

## Anàlisi visual

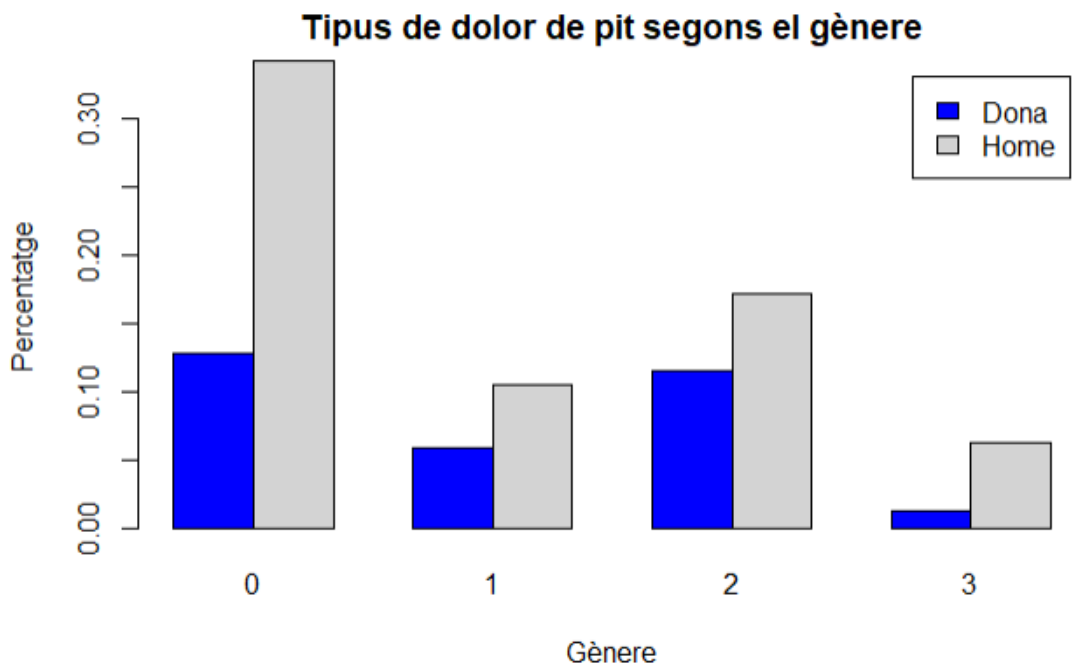
En primer lloc el que s'ha fet és fer un anàlisi visual per comprendre com son i com estan relacionades les dades entre si. Per això s'han creat un conjunt de diagrames que faciliten aquesta comprensió.

En primer lloc s'han creat diagrames de sectors de la distribució de les observacions segons sexe i risc de infart. Les dades estan compostes de un 68% homes i un 32% dones, addicionalment un 54% del pacients tenen probabilitats altes probabilitats de patir un infart i un 46% baixes.

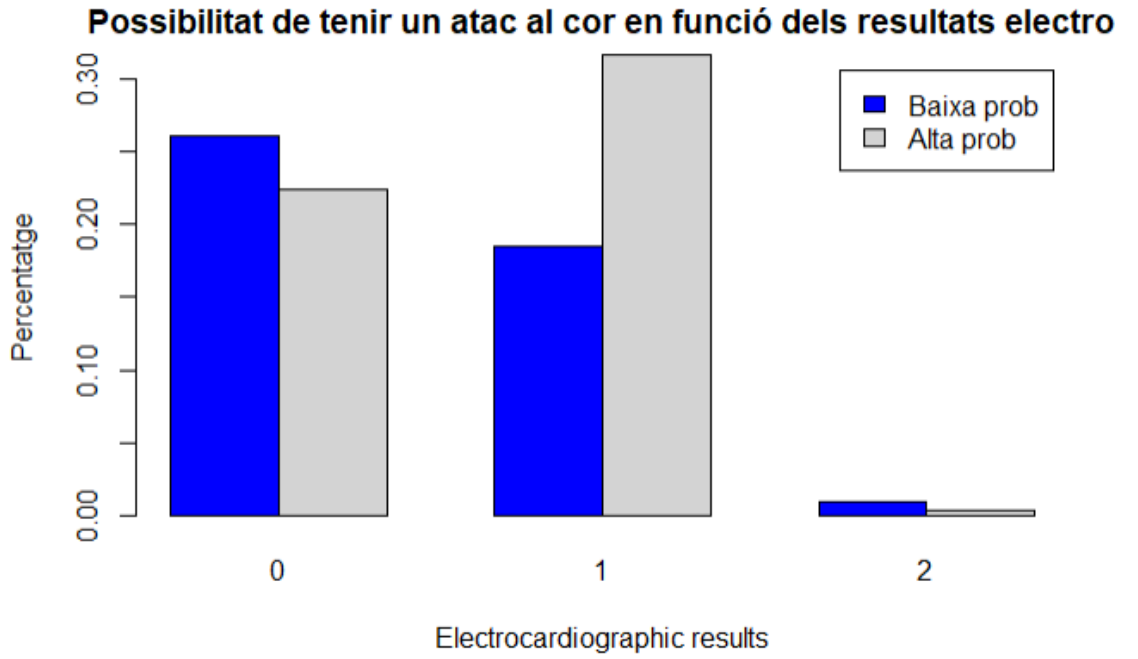



En segon lloc, s'han creat diagrames de barres per veure com es distribueixen les dades.



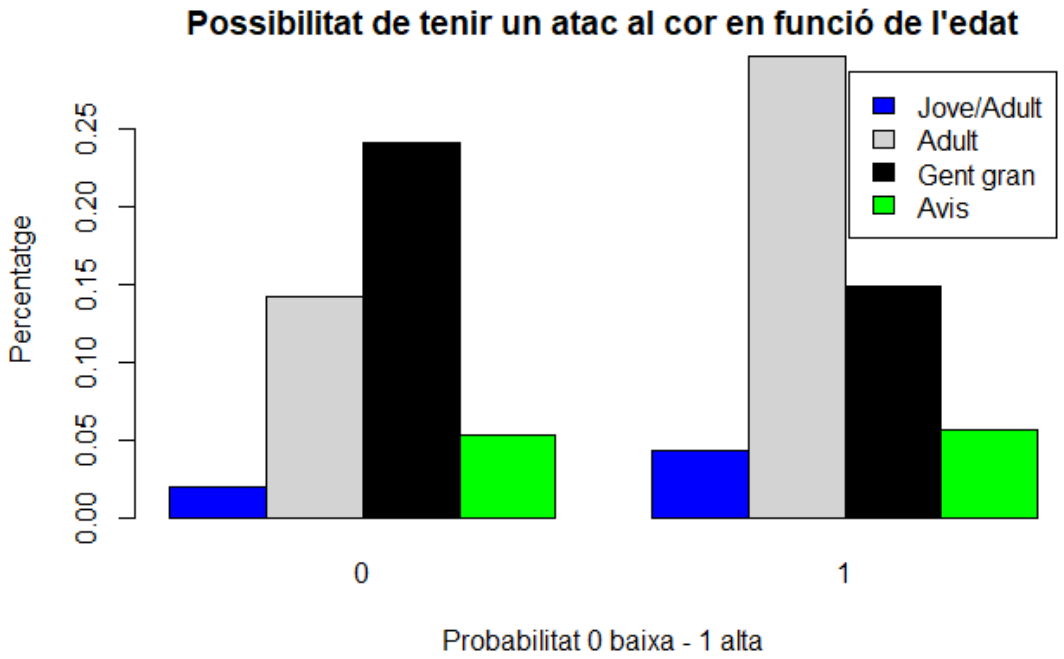

Atenent al gràfic, els homes tenen moltes més angines típiques que les dones, però de moment no podem fer cap suposició sobre que representi si les angines típiques tenen més per a l'hora de patir un infart.




---

Segons els resultats electrocardiogràfics, hi ha una alta probabilitat de tenir un atac al cor si hi ha anomalies de l'ona ST-T.

Seguidament s'ha fet una discretització per intervals (no iguals) de l'edat per determinar quin rang d'edats son els més propensos a patir infarts.




Les edats s'han separat en Jove/Adult (27-40 anys), Adult(41-55 anys), Gent gran(56-65 anys) i Avis(66-78 anys).

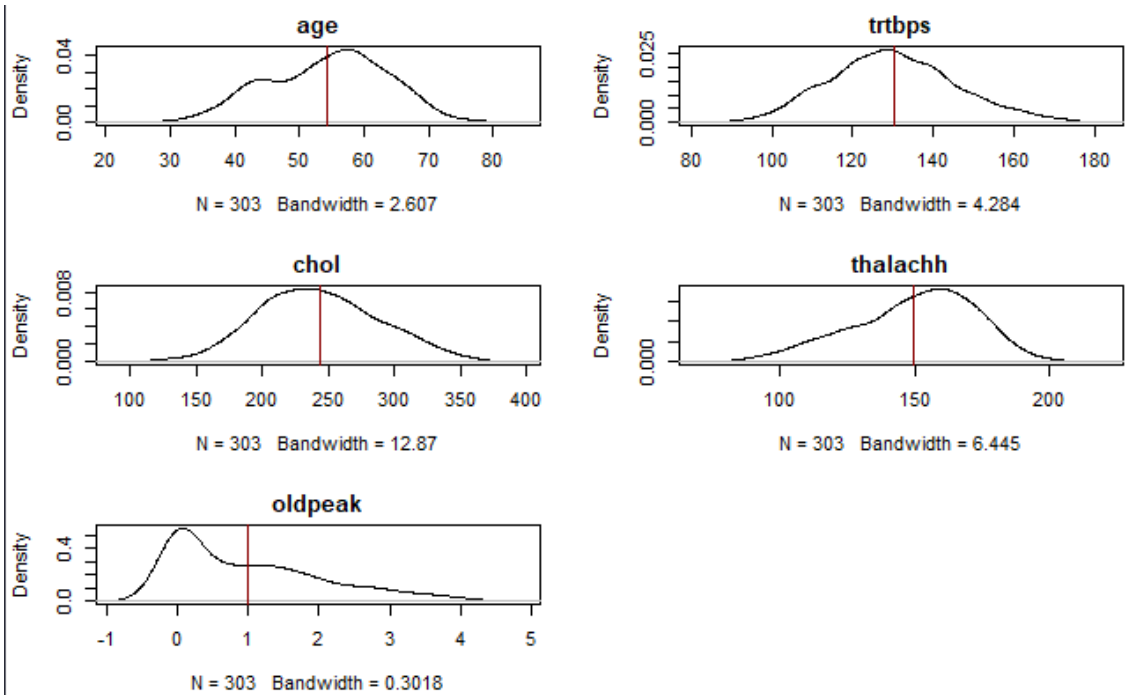
El que es pot extreure del gràfic és que les persones més propenses a tenir un atac de cor son les els adults. D'altra banda, dels que tenen baixa probabilitat de patir un infart, la gent gran és la que més possibilitats té de patir-lo.

## Normalitat i Homoscedasticitat

Per a cada variable continua s'aplica un test Shapiro-Wilk i se'n mostra la distribució amb un gràfic de densitat per comprovar-ne la normalitat. Només la variable 'chol' passa el test, afortunadament sabem que gracies el teorema del límit central si el nombre d'observacions es major que 30 les variables es poden tractar coma variables amb distribució normal.

Els resultats obtinguts han estat els següents:

```
[1] "Shapiro-Wilk per a la variable age"

        Shapiro-Wilk normality test

data:  x
W = 0.98637, p-value = 0.005798

[1] "age No segeuix una distribució normal"
[1] "Per a la variable age la distribució dels pacients sans No segueix una distribució normal."
[1] "Per a la variable age la distribució dels pacients malalts No segueix una distribució normal."
[1] "Per a la variable age la distribució de les dones segueix una distribució normal."
[1] "Per a la variable age la distribució dels homes No segueix una distribució normal."
[1] "Shapiro-Wilk per a la variable trtbps"

        Shapiro-Wilk normality test

data:  x
W = 0.98499, p-value = 0.002977

[1] "trtbps No segeuix una distribució normal"
[1] "Per a la variable trtbps la distribució dels pacients sans No segueix una distribució normal."
[1] "Per a la variable trtbps la distribució dels pacients malalts No segueix una distribució normal."
[1] "Per a la variable trtbps la distribució de les dones segueix una distribució normal."
[1] "Per a la variable trtbps la distribució dels homes No segueix una distribució normal."
[1] "Shapiro-Wilk per a la variable chol"

        Shapiro-Wilk normality test

data:  x
W = 0.99371, p-value = 0.2397

[1] "chol segeuix una distribució normal"
[1] "Per a la variable chol la distribució dels pacients sans segueix una distribució normal."
[1] "Per a la variable chol la distribució dels pacients malats segueix una distribució normal."
[1] "Per a la variable chol la distribució de les dones segueix una distribució normal."
[1] "Per a la variable chol la distribució dels homes segueix una distribució normal."
[1] "Shapiro-Wilk per a la variable thalachh"

        Shapiro-Wilk normality test

data:  x
W = 0.9774, p-value = 0.000103

[1] "thalachh No segeuix una distribució normal"
[1] "Per a la variable thalachh la distribució dels pacients sans segueix una distribució normal."
[1] "Per a la variable thalachh la distribució dels pacients malats segueix una distribució normal."
[1] "Per a la variable thalachh la distribució dels dones No segueix una distribució normal."
[1] "Per a la variable thalachh la distribució dels homes No segueix una distribució normal."
[1] "Shapiro-Wilk per a la variable oldpeak"

        Shapiro-Wilk normality test

data:  x
W = 0.86178, p-value = 8.284e-16

[1] "oldpeak No segeuix una distribució normal"
[1] "Per a la variable oldpeak la distribució dels pacients sans No segueix una distribució normal."
[1] "Per a la variable oldpeak la distribució dels pacients malalts No segueix una distribució normal."
[1] "Per a la variable oldpeak la distribució dels dones No segueix una distribució normal."
[1] "Per a la variable oldpeak la distribució dels homes No segueix una distribució normal."
```




A simple vista no totes les variables segueixen una distribució normal, però s'assumeix que si degut als gràfics i al TLC, el qual determina que si la mida de la mostra és superior a 30 elements i les mitjanes de les mostres s'aproximen a la mitjana de la població, s'aproxima a una distribució normal.

///HOMOSCEDASTICITAT

Per cada variable continua es comprova la Homoscedasticitat entre la seva distribució en relació a la probabilitat de patir un infart i al sexe.

## Proves estadístiques

Per realitzar les proves estadístiques s'han aplicat:

- Test d'hipòtesis
- Regressió lineal
- Correlacions

Per aplicar el test d'hipòtesis s'han desenvolupat tres possibles hipòtesis.

## Test d'hipòtesis

**1-El sexe influeix en el risc de infart?.**

**Hipòtesi nul·la i l'alternativa**

- H0 : No existeix correlació entre les variables.
- H1 : Existeixen correlació entre les variables.

Aplicant el Chi-square test, s'ha obtingut **p-value = 1.007e-06**. Com que p < 0.05, rebutgem la hipòtesis nul·la.

El sexe és un component que afecta al risc d'infart.

**2-L'edat afecta al risc de infart?.**

**Hipòtesi nul·la i l'alternativa**

- H0 : No existeix correlació entre les variables.
- H1 : Existeixen correlació entre les variables.

Aplicant el Chi-square test, s'ha obtingut **p-value = 0.1309**. Com que p > 0.05, acceptem la hipòtesis nul·la.

L'edat no és un component que afecta al risc d'infart.

**3-El colesterol en sang afecta al risc de infart?.**

**Hipòtesi nul·la i l'alternativa**

- H0 : No existeix correlacio entre les variable
- H1 : Existeixen correlació entre les variables.

Aplicant el Chi-square test, s'ha obtingut **p-value = 0.07738**. Com que p > 0.05, acceptem la hipòtesis nul·la.

El colesterol en sang no és un component que afecta al risc d'infart.

## Regressions

Per fer l'anàlisi de regressió, s'ha creat un primer model.

s'ha dividit el conjunt d'entrenament en train i test, amb un 80/20 del total de les dades respectivament. Posteriorment s'ha creat un model **glm**(Generalized Linear Model) el qual hem

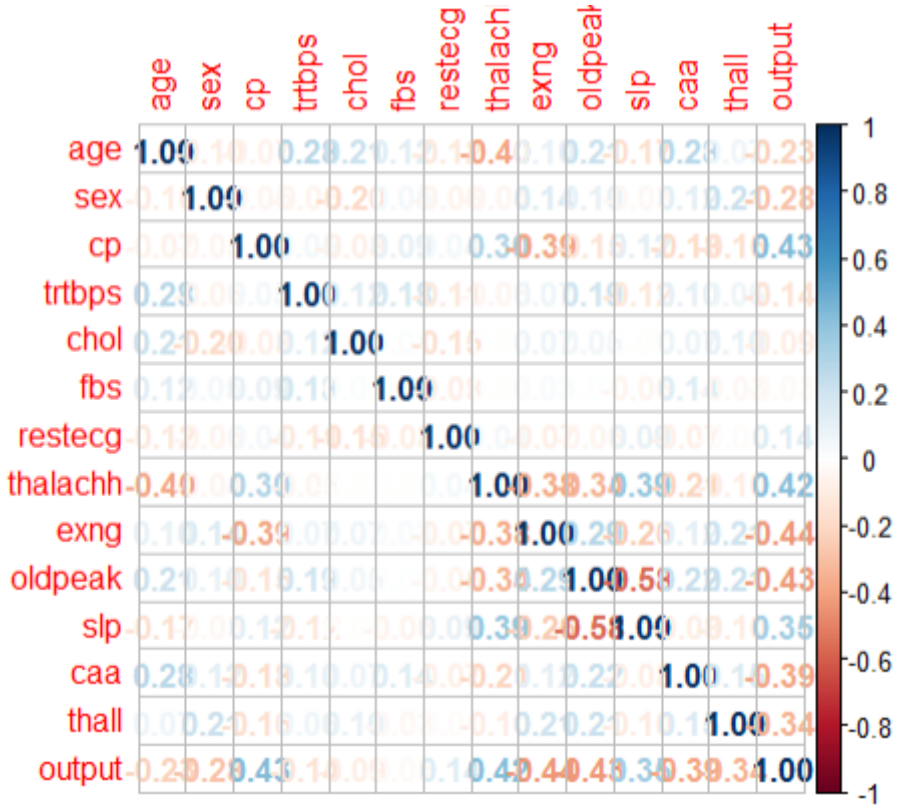
determinat que la variable output seria la dependent, i age, sex, cp, trtbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall les variables independents del model.

Al fer un summary del model hem obtingut els següents resultats.

```
Call:
glm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
    restecg + thalachh + exng + oldpeak + slp + caa + thall,
    family = binomial(link = logit), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8887  -0.3495   0.1213   0.4514   2.4790

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.938055   3.508733   1.407 0.159320
age          0.000914   0.026889   0.034 0.972883
sex1        -1.640559   0.605532  -2.709 0.006743 **
cp1          1.426524   0.608042   2.346 0.018971 *
cp2          2.306104   0.596296   3.867 0.000110 ***
cp3          2.320979   0.749677   3.096 0.001962 **
trtbps      -0.016383   0.014886  -1.101 0.271073
chol        -0.011358   0.005579  -2.036 0.041776 *
fbs1         0.612866   0.620442   0.988 0.323256
restecg1     0.223445   0.427856   0.522 0.601501
restecg2    -1.160596   2.569262  -0.452 0.651468
thalachh     0.013502   0.013821   0.977 0.328611
exng1       -1.066680   0.509039  -2.095 0.036129 *
oldpeak     -0.213412   0.272149  -0.784 0.432938
slp1        -0.179770   0.836466  -0.215 0.829833
slp2         1.190436   0.908663   1.310 0.190163
caa1        -2.422607   0.592489  -4.089 4.33e-05 ***
caa2        -3.287934   0.868511  -3.786 0.000153 ***
caa3        -1.587840   0.813587  -1.952 0.050979 .
thall2      -0.426763   0.917414  -0.465 0.641802
thall3      -2.023916   0.889231  -2.276 0.022844 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 344.97  on 249  degrees of freedom
Residual deviance: 154.62  on 229  degrees of freedom
AIC: 196.62

Number of Fisher Scoring iterations: 6
```

Veient els valor de p, i els significance codes, podem determinar que les variables explicatives o independents estadísticament més significants son cp2, caa1 i caa2 com les més significatives, ja que els seus valors estan en el rang [0, 0.001].

Després s’ha realitzat una anàlisi de col·linealitat del model amb una matriu de correlacions i amb la funció VIF de R.




| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| age | 1.408091 | 1 | 1.186630 |
| sex | 1.624867 | 1 | 1.274703 |
| cp | 2.029756 | 3 | 1.125228 |
| trtbps | 1.305989 | 1 | 1.142799 |
| chol | 1.161899 | 1 | 1.077914 |
| fbs | 1.145959 | 1 | 1.070495 |
| restecg | 1.162327 | 2 | 1.038322 |
| thalachh | 1.598011 | 1 | 1.264125 |
| exng | 1.266385 | 1 | 1.125338 |
| oldpeak | 1.492541 | 1 | 1.221696 |
| slp | 1.782699 | 2 | 1.155499 |
| caa | 2.068038 | 3 | 1.128738 |
| thall | 1.719640 | 2 | 1.145142 |

Com es pot comprovar, les dues formes demostren que no hi ha una col·linealitat significativa entre les variables.

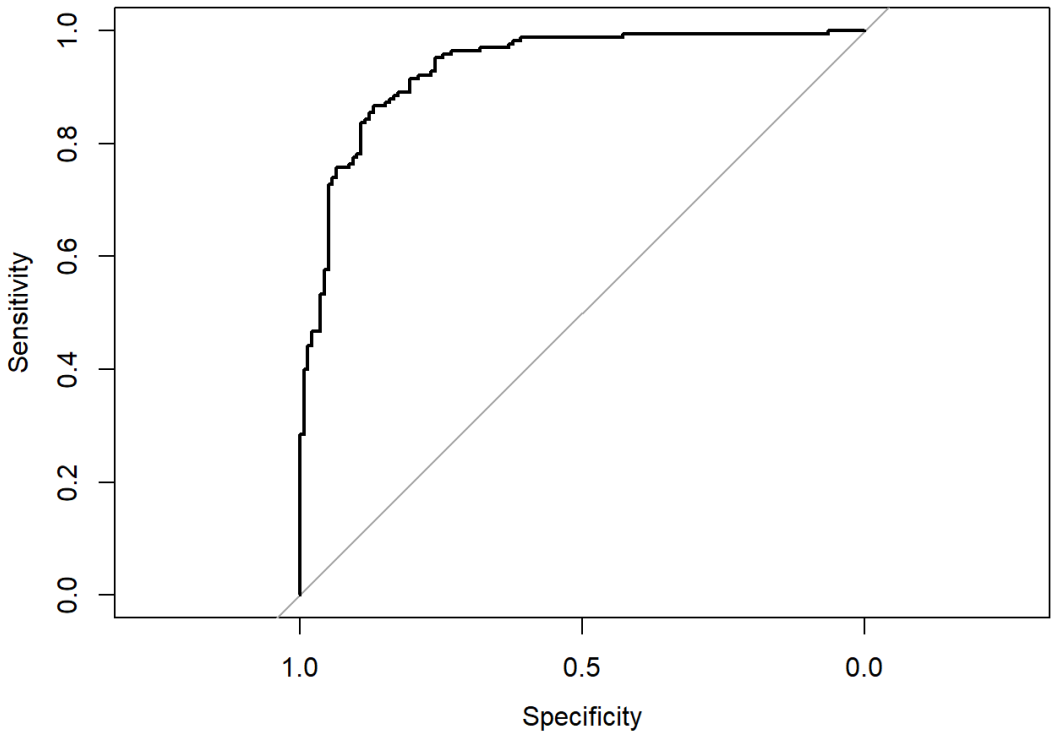
Seguidament s'ha creat el model final excloent les variables que no eren significatives.

```
Call:
glm(formula = output ~ sex + cp + chol + thalachh + exng + slp +
    caa + thall, family = binomial(link = logit), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7373  -0.4114   0.1253   0.4343   2.4399

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.4493589  2.4197780   1.012  0.31143
sex1        -1.5068069  0.5754569  -2.618  0.00883 **
cp1          1.5524497  0.6000346   2.587  0.00967 **
cp2          2.3233009  0.5856190   3.967 7.27e-05 ***
cp3          2.0364152  0.6902511   2.950  0.00318 **
chol        -0.0118800  0.0052449  -2.265  0.02351 *
thalachh     0.0142680  0.0126365   1.129  0.25885
exng1       -1.0561865  0.4937803  -2.139  0.03244 *
slp1        -0.0004576  0.7610421  -0.001  0.99952
slp2         1.4471072  0.7926501   1.826  0.06790 .
caa1        -2.3422115  0.5736851  -4.083 4.45e-05 ***
caa2        -3.2417954  0.7816311  -4.147 3.36e-05 ***
caa3        -1.6774652  0.7770182  -2.159  0.03086 *
thall2      -0.2724351  0.9054472  -0.301  0.76350
thall3      -1.9217124  0.8746657  -2.197  0.02801 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 344.97  on 249  degrees of freedom
Residual deviance: 158.40  on 235  degrees of freedom
AIC: 188.4

Number of Fisher Scoring iterations: 6
```

Per avaluar la bonat de l’ajust del model es s’utilitza el test Chi-quadrat obtenint un valor de p de 0. Per tant es considera que el model es bo.

Seguidament Utilitzant el conjunt de test per avaluar el model, Obtenint:

Accuracy: 0.8490566,
Sensitivity: 0.8695652
Specificity:0.8333333
AUC: 0.9339
Amb una curva roc que pren la següent forma:




Un cop verificat que el model es suficientment bo i la seva capacitat de predicció, es procedeix a fer un anàlisis de les OR(odds ratios) i sen visualitza les probabilitats de risc de infart tot i diferenciat per sexe.

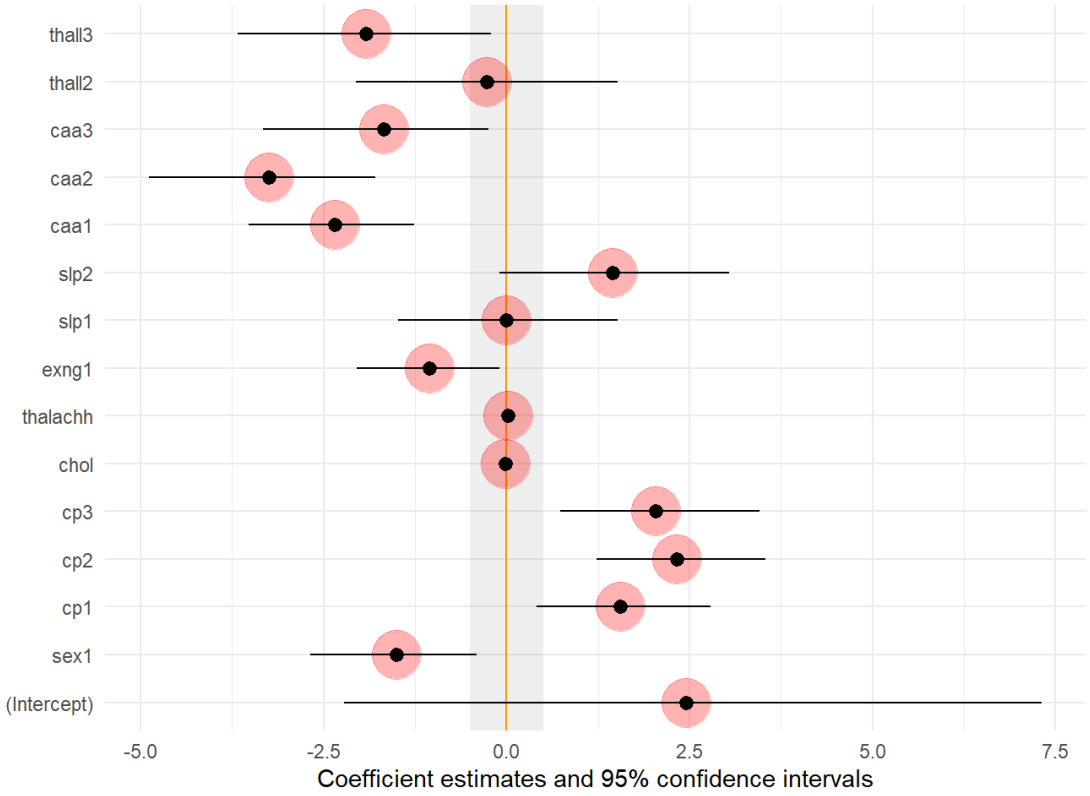
De l'estudi de les Odd ratios s'en extreuen les següents conclusions:

> Les odds-ratio de cp s'en dedueix que el dolor toràcic es el principal indicador de un alt risc de infart. Tenint els pacients amb una angina atípica 10 vegades la probabilitat d'ocurrencia d'un risc alt de infart, els pacients amb una angina típica 4.7 vegades i els pacients amb un dolor agngios 7.7 vegades més probabilitats que la resta de pacients.
>
> Aixi mateix, el pendent del segment ST és també un bon indicador de la probabilitat de ocurrencia de un risc alt de infart, sent els pacients amb un pendent negatiu 4.2 vegades més propensos a tenir un risc alt de infart.
>
> Per altra banda es pot veure que com a principals factors protectors contre risc de infart trobem els vasos afectats, i el sexe on, per exemple, l'odds-ratio estimat per a sex=1 és 0.22, de manera que l'ocurrència de un risc alt de infart és 0,22 vegades menor, en relació al sexe=0.

# Representació dels resultats

De cara a la representació dels resultats es mostra la representació dels coeficients i variables del model de regressió logística. On es pot veure com afecten cada una de les variables al risc de infart.

Sent aquelles variables a la esquerra de la gràfica del model variables protectores i aquelles variables a la dreta factors de risc de cara a patir un infart. Així mateix aquelles variables que cauen a la zona gris central no tenen afectació real en la probabilitat de risc de infart.




# Resolució del problema

Com a resultat de l'analisis portat a terme utilitzant el dataset triat s'extreuen les següents conclusions:

-Ni la edat ni el colesterol tenen una afectació real a l'hora de determinar el risc d'infart.

- Ni la pressió arterial ni els resultats electrocardiogràfics són prou significants per determinar el risc d'infart.

- Els principals factors de risc de cara a tenir un risc elevat de infart són els dolors toràcics.

-Els principals factors protectors de cara a tenir un risc elevat de infart són un nombre de vasos, tenir una talassèmia reversible i el sexe.

- Es pot obtenir un model predictor amb una altra probabilitat de predir correctament el risc de infart.

Tanmateix, donades la naturalesa de algunes de les conclusions extretes, es considera que l'estudi no és concloent, es requereix doncs, més investigació per poder verificar-les i demostrar que no són fruit de un data set poc representatiu de la població general.

# Codi

Per portar a terme aquesta practica es treballa amb R i s'utilitzen les següents llibreries auxiliars: Corrplot, VIM,car, caret,ggplot2,modelsummary i pROC.

Algunes de les parts importants del codi són:

Carrega de dades

```
heart <- read.csv("heart.csv")
```

Imputació de valors atípics

```
columnesImputar<-colnames(heart)[colSums(is.na(heart)) > 0]
heart <-  kNN(heart, variable = columnesImputar, k = 5)
```

Test de correlació entre 2 variables categòriques:

```
chisq.test(heart$sex, heart$output, correct=FALSE)
```

Obtenció del model de regressió

```
model <- glm(output~age+sex+cp+...,data=train,   family=binomial(link=logit))
```

Obtenció matriu de correlació

```
corrplot(cor(heart_numeric), method = "number")
```

Evaluació del model

```
}
pred<-predict(modelF, newdata = test,type ="response")
predict<-lapply(pred,convert)

cf<-confusionMatrix(data=as.factor(as.numeric(predict)), reference = as.factor(test$output))

prob=predict(modelF, heart, type="response")
r=roc(heart$output, prob, data=heart)
plot(r)
auc(r)
```

Obtenció de les odd-ratio

```
exp(coefficients(modelF))
```

Obtenció de la representació del model final

```
modelplot(modelF , background = b)
```

# Video

https://drive.google.com/file/d/1jV898vDMAkJXMZWWAu0KqQoMmYdY5Tex/view?usp=sharing

| Contribucions | Signatura |
|---|---|
| Investigació prèvia | Gerard Ramos Gambús // Oriol Caravaca Müller |
| Redacció de les respostes | Gerard Ramos Gambús // Oriol Caravaca Müller |
| Desenvolupament del codi | Gerard Ramos Gambús // Oriol Caravaca Müller |
| Participació al vídeo | Gerard Ramos Gambús // Oriol Caravaca Müller |