Tipologia i cicle de vida de les dades

Pràctica 2 (35% nota final)

Autors: Antoni Espadas Navarro et Jordi Samaniego Vidal

Índex

1. Descripció del dataset.

Perquè és important i quina pregunta/problema pretén respondre?

Un atac de cor anàleg a l'infart agut de miocardi (IAM) és una de les malalties més greus en el segment de les malalties cardiovasculars, de fet, la Marató de TV3 d'enguany serà recordada per haver contribuït a la conscienciació al voltant de la salut cardiovascular per haver recaptat 8.034.807 d'euros per a la investigació d'aquesta malaltia.

Es produeix a causa de la interrupció de la circulació sanguínia al múscul del cor que danya al mateix. El diagnòstic de malalties del cor també és una tasca crucial. Els símptomes, l'examen físic i la comprensió dels diferents signes d'aquesta malaltia són necessaris per diagnosticar la malaltia cardíaca. Diferents factors, com el colesterol, les malalties genètiques del cor, la pressió arterial alta, la baixa activitat física, l'obesitat i el tabaquisme poden ser motius per a l'aparició de malalties del cor.

El motiu principal dels atacs cardíacs és l'aturada de sang a les artèries coronàries, per tant, la pregunta que volem respondre és: donats uns factors de risc d'un pacient de gènere masculí o femení, podem predir si aquest patirà un atac de cor? Hi ha diferències entre la possibilitat de patir un atac de cor entre homes i dones?

2. Integració i selecció de les dades d'interès a analitzar.

Primerament, farem una anàlisis exploratori de les dades, les netejarem, farem modificacions oportunes d'acord amb el darrer punt i, finalment, aplicarem dos models predictius de classificació supervisat (regressió logística i random forest) per la variable dependent (objectiu) que es detallen a continuació:

- Edat: Edat del pacient
- Sexe: Sexe del pacient
- exang: angina induïda per l'exercici (1 = sí; 0 = no)
- ca: nombre de vaixells principals (0-3)
- cp : tipus de dolor de pit tipus de dolor de pit
    - Valor 1: angina típica
    - Valor 2: angina atípica
    - Valor 3: dolor no anginós
    - Valor 4: asimptomàtic
- trtbps: pressió arterial en repòs (en mm Hg)
- chol : colestoral en mg/dl obtingut mitjançant el sensor IMC
- fbs : (sucre en sang en dejú > 120 mg/dl) (1 = cert; 0 = fals)
- rest_ecg : resultats electrocardiogràfics en repòs
    - Valor 0: normal
    - Valor 1: amb anormalitat de l'ona ST-T (inversions de l'ona T i/o elevació o depressió ST > 0,05 mV)
    - Valor 2: mostra una hipertròfia ventricular esquerra probable o definitiva segons els criteris d'Estes
- thalach: freqüència cardíaca màxima aconseguida
- **objectiu**: (0) malaltia cardíaca no present, (1) malaltia cardíaca present

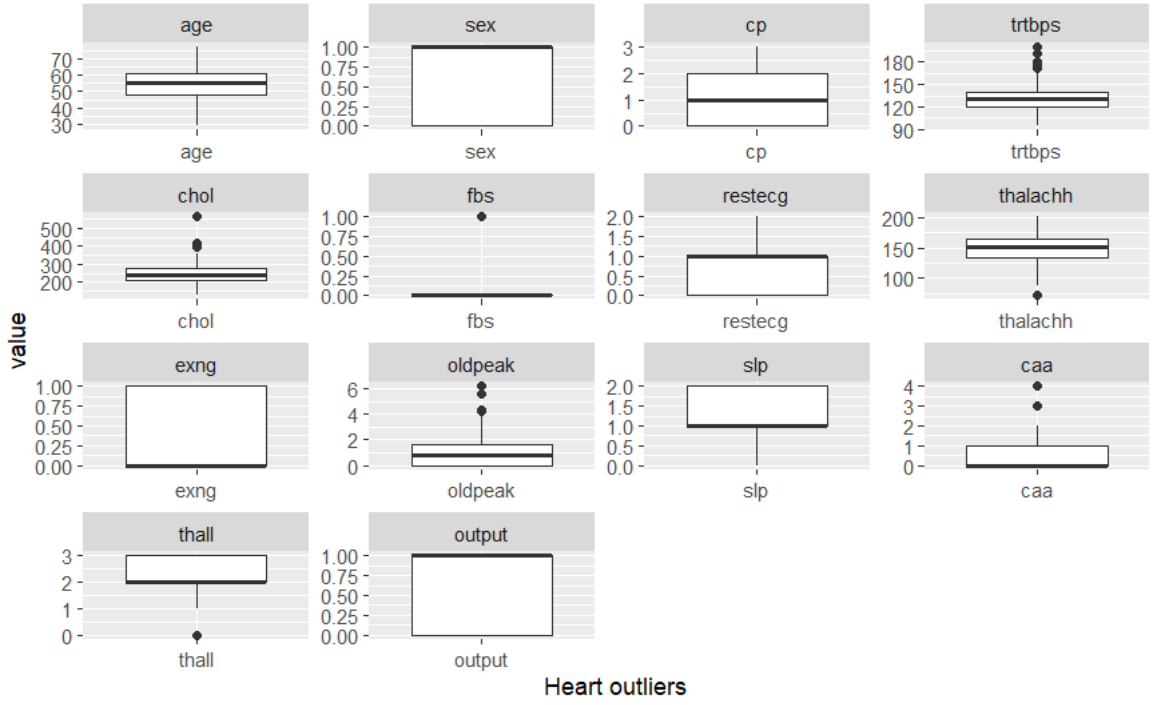
3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits?

En primer lloc, mostrem que no hi ha cap valor NA's ni valors buits (“”), en conseqüència, no hem de realitzar el procés d'eliminació de valors perduts. No obstant, trobem una fila duplicada a que eleminem. Seguidament, procedirem a analitzar si hi ha valors extrems en les variables numèriques en les quals els possibles valors no estiguin acotats (com per exemple output, on només pot valdre 0 o 1).

3.2. Identifica i gestiona els valors extrems.

Per a detectar outliers, mostrarem un gràfic de dispersió, el qual marcarà els possibles valors extrems i, a més, ens permetrà conèixer quina és la dispersió entre els diferents valors. Mitjançant el següent gràfic mostrarem els valors extrems de cada columna:



Age: Observem que no hi ha outliers, la mitjana es troba als 55 anys i on trobem més registres és entre 47 i 62 anys.

- trtbps: Recordem que aquesta variable indica la pressió arterial en repòs. En aquest cas, es detecten alguns possibles valors extrems per sobre de 170. Tot i així, no ho considerarem com a outlier, doncs és possible que una persona tingui una pressió arterial de 200, tot i que pot ser un indicador de problemes mèdics de la persona.

- chol: En aquest cas, observem un registre que podriem considerar com a outlier, doncs un valor de 600 de colesterol és molt atípic. La resta de valors considerats com a extrems, no els considerarem outliers, doncs tot i que és un risc, un valor de 400 de colesterol pot ser correcte. **Per tant, eliminem els registres del dataset que tinguin un nivell de colesterol superior a 500**
- thalachh: Per a la freqüència cardíaca màxima no detectem outliers, doncs és normal que aquesta es trobi entre 90 i 200. Veiem un registre que té un valor proper a 75, tot i que ho considerarem correcte i no l'exclourem.

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar.

(p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?)

Per una banda, realitzarem un contrast de mitjanes per tal de saber si hi ha diferències entre la possibilitat d'un atac de cor entre homes i dones. També, analitzarem les correlacions entre els diferents atributs del dataset "heart" del que disposem. Per últim, crearem dos models supervisats, un de regressió logística i l'altra de Random Forest, per tal de predir les possibilitats d'atac de cor d'un pacient.

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

En primer lloc, abans d'aplicar cap model, és necessari conèixer si el conjunt de dades segueix una distribució normal i si presenta homoscedasticitat, doncs en base a això, podrem aplicar uns models o uns altres.

Així doncs, començarem amb l'anàlisi de normalitat. Cal tenir present, però, que la variable independent que volem predir (output) es tracta d'un atribut binari, és a dir, només pot prendre dos possibles valors. Per tant, la distribució d'una variable dicotòmica mai podrà presentar una distribució normal. Tot i així, apliquem el test de Kolmogorov-Smirnov sota la hipòtesi nul·la de que les dades són normals. També aplicarem el test de Shapiro-Wilk:

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  heart$output
D = 0.36237, p-value < 2.2e-16
alternative hypothesis: two-sided


        Shapiro-Wilk normality test

data:  heart$output
W = 0.63397, p-value < 2.2e-16
```

Observem que obtenim un **p-valor de quasi zero pels dos tests,** fet que ens permet **rebutjar la hipòtesi nul·la**. Per tant, podem dir que **les dades no presenten normalitat**.

Tot i així, en tenir un conjunt significativament gran (més de 30 registres) **pel Teorema del Límit central** podem suposar que **la mitjana de les dades sí que presenta normalitat.**

Per altra banda, procedim a analitzar si el conjunt presenta homoscedasticitat, és a dir, igualtat de variància. Per a fer-ho, utilitzarem la funció var.test(). Aquest test l'aplicarem per saber si la variable output presenta igualtat de variàncies o no entre homes i dones. Per aconseguir-ho utilitzarem la funció var.test(), sota la hipòtesi nul·la d'igualtat de variàncies:

```
        F test to compare two variances

data:  heart$output[heart$sex_cat == "F"] and heart$output[heart$sex_cat == "M"]
F = 0.76833, num df = 94, denom df = 205, p-value = 0.1482
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5492551 1.0993475
sample estimates:
ratio of variances
         0.7683278
```

Obtenim **un p-valor de 0.15**, fet que **no ens permet rebutjar la hipòtesi de igualtat de variàncies**. Per tant, podem dir que **la probabilitat d'un atac de cor entre homes i dones presenta igualtat de variàncies**.

4.3. Aplicació de proves estadístiques per comparar els grups de dades.

En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Ara, aplicarem un contrast de mitjanes per saber si hi ha diferències entre les possibilitats d'un atac de cor entre homes i dones. En aquest cas, utilitzarem les següents hipòtesis nul·la i alternativa:

Hipòtesi nul·la: l'output entre homes i dones és igual.

H0:Output$homes=Output$dones

Hipòtesi alternativa: l'output entre homes i dones és diferent.

H1:Output$homes≠Output$dones

Per aplicar el contrast, cal recordar les dues conclusions extretes prèviament: es tracta d'una distribució normal (TLC) amb variàncies desconegudes i iguals. Així doncs, podrem aplicar una prova de contrast d'hipòtesi de tipus paramètric, com la t d'student.

```
        Two Sample t-test

data:  heart$output[heart$sex_cat == "F"] and heart$output[heart$sex_cat == "M"]
t = 5.0539, df = 299, p-value = 7.553e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1836516 0.4178814
sample estimates:
mean of x mean of y
0.7473684 0.4466019
```

Obtenim un p-valor de pràcticament zero, fet que ens permet rebutjar la hipòtesi nul·la. **Per tant, amb un nivell de confiança superior al 99%, podem dir que hi ha diferències entre la possibilitat de patir un atac de cor entre homes i dones.**

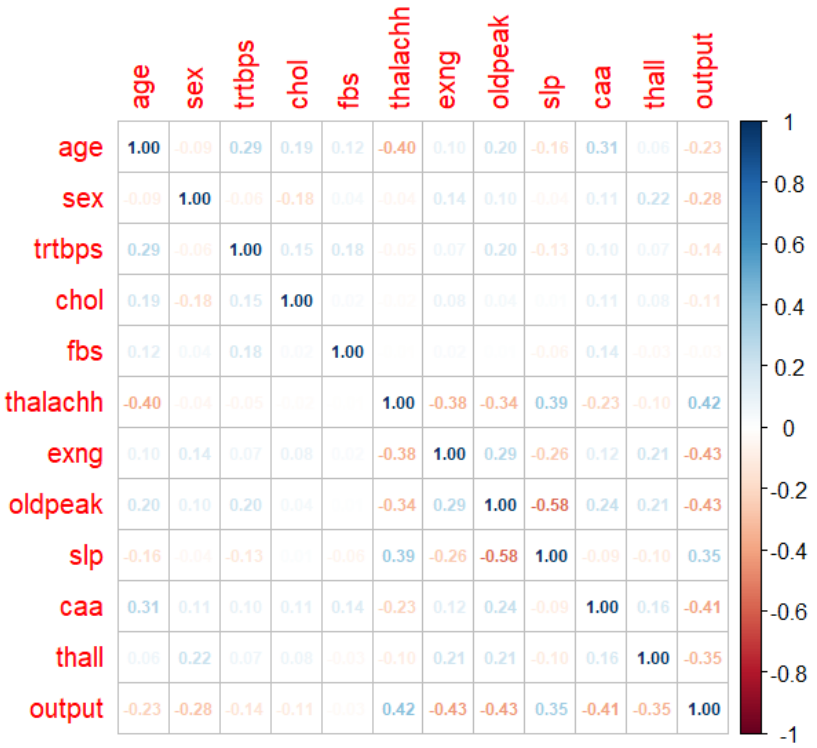
Finalment, per predir si un pacient patirà un atac de cor, hem comentat que utilitzarem dos models: regressió logística i random forest. En conseqüència, hem fet ús de **tècniques de subdivisió del conjunt (partició de 80/20 o estratificada de 2/3)** de dades així com la **validació creuada** de la mateixa, d'altres tècniques per **mesurar la qualitat del model (auc ROC)** i tècniques de correlacions entre variables; així doncs, mostrarem els resultats obtinguts d'ambdòs models acompanyat de taules i gràfiques en el següent punt

5. Representació dels resultats a partir de taules i gràfiques.

Aquest apartat es pot respondre al llarg de la pràctica, sense la necessitat de concentrar totes les representacions en aquest punt de la pràctica.

- Matriu de correlacions

En aquesta secció, procedirem a analitzar quines correlacions presenten els diferents atributs de tipus numèric:



Del gràfic obtingut podem concloure que:

- La variable que més correlació positiva presenta és thalach, es a dir, com més freqüència cardíaca màxima presenta el pacient, més probabilitat d'atac de cor.

- La variable exng presenta una correlació negativa. És a dir, hi ha correlació entre no tenir angina en fer esport i patir un atac de cor.

- La variable caa presenta correlació negativa. Hi ha correlació entre un nombre baix de vasos cardíacs principals i patir un atac de cor.

- La variable oldpeak presenta correlació negativa. Hi ha correlació entre una baixa depressió del ST induïda per l'exercici en relació amb el repòs i patir un atac de cor.

Dit això, abans de realitzar model de regressió, **generarem un set de dades d'entrenament, el qual contindrà el 80% dels registres, i un set de test, per a validar el model construït a partir de les dades d'entrenament.**

- Model de Regressió logística (GLM)

Considerem que és important, però, que la variable a predir, es trobi en la mateixa proporció entre el dataset d'entrenament i de test, per tal de no obtenir un model esbiaixat. Per a fer-ho, utilitzarem la funció createDataPartition (de la llibreria caret), la qual permet assegurar que, en fer les particions, tinguem la mateixa proporció de la variable objectiu en ambdós datasets:

Dit això i donat que la variable a predir (output) conté una informació binària (1 = propens a patir un atac de cor; 0 = no propens a patir un atac de cor), el primer model que crearem serà una regressió logística, la qual permet predir una variable dicotòmica on inclourem totes les variables independents possibles, per tal de poder analitzar si totes elles són significatives o no:

Call:
glm(formula = output ~ age + sex_cat + cp_cat + trtbps + chol +

fbs + restecg_cat + thalachh + exng + oldpeak + slp + caa +
thall, family = binomial, data = train)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6108  -0.3097   0.1651   0.5210   2.6692

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 3.770862   2.936101   1.284 0.199034
age                        -0.003891   0.027543  -0.141 0.887663
sex_catM                   -1.790184   0.538075  -3.327 0.000878 ***
cp_catAnginaAtipica         0.717248   0.648487   1.106 0.268712
cp_catSenseAngina           1.572284   0.535950   2.934 0.003350 **
cp_catAsimptomatic          1.580715   0.696481   2.270 0.023233 *
trtbps                     -0.025854   0.012577  -2.056 0.039809 *
chol                       -0.004151   0.004821  -0.861 0.389304
fbs                         0.696883   0.607743   1.147 0.251516
restecg_catWaveAbnormality  0.674691   0.426458   1.582 0.113631
restecg_catHypertrophy      0.233133   2.215518   0.105 0.916195
thalachh                    0.029285   0.012064   2.427 0.015204 *
exng                       -0.988470   0.478457  -2.066 0.038833 *
oldpeak                    -0.564770   0.256731  -2.200 0.027817 *
slp                         0.345321   0.414594   0.833 0.404893
caa                        -0.776810   0.238856  -3.252 0.001145 **
thall                      -1.005681   0.388177  -2.591 0.009576 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Del resultat obtingut, ens centrem en el p-valor de cada variable. Si aquest supera 0.05, podrem dir amb un 95% de confiança que la variable no és significativa tals com age, trtbps, chol, fbs, restecg_cat, thalachh, exng i slp (excloses del model)

Seguidament, procedim a analitzar les Odds-Ratio de cada variable, les quals poden prendre els següent valors:

```
      (Intercept)                       age                  sex_catM       cp_catAnginaAtipica
       43.4174922                 0.9961168                 0.1669295                 2.0487880
 cp_catSenseAngina       cp_catAsimptomatic                    trtbps                      chol
        4.8176386                 4.8584293                 0.9744773                 0.9958580
              fbs restecg_catWaveAbnormality     restecg_catHypertrophy                  thalachh
        2.0074852                 1.9634265                 1.2625494                 1.0297176
             exng                   oldpeak                       slp                       caa
        0.3721456                 0.5684908                 1.4124439                 0.4598705
            thall
        0.3657954
```

Odd-Ratio proper a 1: Indica que no hi ha relació entre la covariable i la variable independent (output).

Odd-Ratio superior a 1: Es tracta d'un factor de risc, doncs si aquesta covariable està present, el succés (és adir, que el pacient pateixi un atac de cor) és més probable.

Odd-Ratio inferior a 1: Es tracta d'un factor de protecció, doncs si aquesta covariable està present, el succés serà menys probable.

Punts a concloure:

- Les variables age, trtbps, chol, fbs, restecg_cat, thalachh, exng i slp tenen uns valors propers a 1, fet que indicaria que no tenen afectació sobre el succés a predir. Aquest resultat l'hem obtingut amb la anàlisi prèvia, on hem vist que no eren significants.

- Les variables sex, caa i thall són factors de protecció. Un valor baix en aquestes variables afecta en que sigui més probable l'atac de cor. Això vol dir que si el pacient és dona, té pocs vasos cardíacs principals i té una taxa de talessèmia baixa, tindrà més probabilitats de patir una malaltia cardíaca.

- La variable cp és un factor de risc. Un valor elevat en aquesta variable fa que sigui més probable l'atac de cor. Aquesta variable, però, recordem que tot i que sigui de tipus enter té un significat categòric, doncs cada valor indica un tipus diferent de de dolor al pit.

Finalment, generarem un nou model de regressió logística utilitzant només les variables significatives, és a dir: sex_cat, cp_cat, caa i thall:

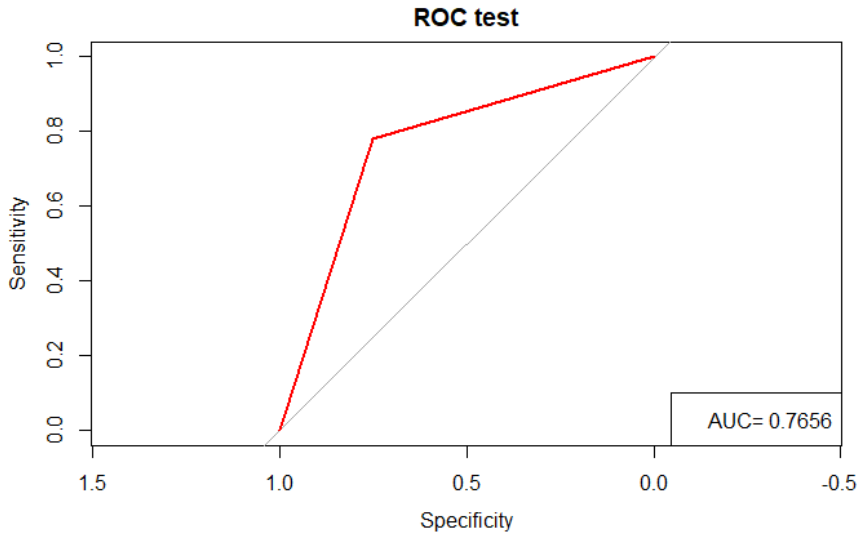
Call:

```
glm(formula = output ~ sex_cat + cp_cat + caa + thall, family = binomial,
    data = train)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2144  -0.6624   0.2876   0.5321   2.7947

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.5040     0.7888   4.442 8.90e-06 ***
sex_catM           -1.2810     0.4083  -3.138 0.001704 **
cp_catAnginaAtipica  2.1029     0.5339   3.939 8.18e-05 ***
cp_catSenseAngina    2.0793     0.4341   4.789 1.67e-06 ***
cp_catAsimptomatic   1.3366     0.5808   2.301 0.021380 *
caa                 -0.8265     0.1958  -4.222 2.43e-05 ***
thall               -1.2094     0.3181  -3.802 0.000144 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Un cop tenim un model amb les variables més òptimes, procedim a executar-lo amb el subconjunt de dades de test i obtenim la matriu de confusió següent:

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  60


             | Prediction
     Reality |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |        21 |         7 |        28 |
             |     0.350 |     0.117 |           |
-------------|-----------|-----------|-----------|
           1 |         7 |        25 |        32 |
             |     0.117 |     0.417 |           |
-------------|-----------|-----------|-----------|
Column Total |        28 |        32 |        60 |
-------------|-----------|-----------|-----------|
```



Resultats a destacar:

- Observem que el model generat ha predit correctament 46 dels 60 registres (77,7%).
- El model té una especificitat de 0.75
- El model té una sensibilitat de 0.78125
- ROC AUC: 0.77

A continuació, per comparar amb el model anterior en concepte de rendiment, mostrarem un exemple **d'aplicació d'un model random forest al conjunt de dades sobre el dataset original (subdividit en train i test de tipus statisfied en una ratio de 2/3), mitjançant una validació creuada amb 4 folds** on es divideixen les dades en entrenament i test per només aplicar l'entrenament mitjançant validació creuada al primer subconjunt.

Posteriorment, mitjançant la funció predict() es prediu el resultat de les dades del subconjunt de test i es representen les diferents mesures de bondat del model, mitjançant la funció confusionMatrix(), especificant com a positius els casos d'infarts.

- Random Forest

Primer de tot, esbrinarem quines són les variables més contributives amb l'ús de la mitjana Decrease Gini:

	MeanDecreaseGini
age	8.160091
sex	3.739620
cp	11.036192
trtbps	7.617683
chol	7.097344
fbs	1.060219
restecg	2.479304
thalachh	12.349974
exng	4.913949
oldpeak	13.652746
slp	4.596728
caa	11.865729
thall	9.531293

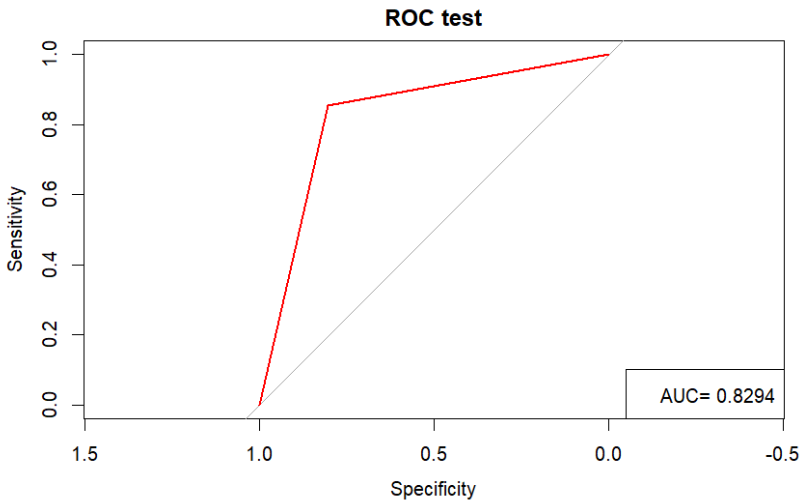
D'acord l'anàlisis EDA, podem descartar fbs (sumat que no havíem trobat conclusions notòries) i la columna restecg. A continuació, entrenem el model i fem les prediccions obtenint el resultat següent:

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  101

             | Prediction
     Reality |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |        35 |        11 |        46 |
             |     0.347 |     0.109 |           |
-------------|-----------|-----------|-----------|
           1 |         9 |        46 |        55 |
             |     0.089 |     0.455 |           |
-------------|-----------|-----------|-----------|
Column Total |        44 |        57 |       101 |
-------------|-----------|-----------|-----------|
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 35  9
         1 11 46

               Accuracy : 0.802
                 95% CI : (0.7109, 0.8746)
    No Information Rate : 0.5446
    P-Value [Acc > NIR] : 5.736e-08

                  Kappa : 0.5994

 Mcnemar's Test P-Value : 0.8231

            Sensitivity : 0.8364
            Specificity : 0.7609
         Pos Pred Value : 0.8070
         Neg Pred Value : 0.7955
             Prevalence : 0.5446
         Detection Rate : 0.4554
   Detection Prevalence : 0.5644
      Balanced Accuracy : 0.7986

       'Positive' Class : 1
```

El model generat ha predit correctament 84 dels 101 registres (~83,2%). Per una banda, tenim una sensibilitat del 85.5% i, per l'altra banda, una especificitat de 80%,

en conseqüència, en tenim una precisió equilibrada de 83% on s'ajusta a un bon rendiment de model. A més, fem un gràfic per assegurar-nos de la qualitat d'aquest:



El test ROC obté una performance de 0.8294 que ens informa d'un valor diagnòstic positiu.

6. Resolució del problema.

A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

- Afirmem que, amb un nivell de confiança superior al 99%, hi ha diferències entre la possiblitat de patir un atac de cor entre homes i dones.
- Les dones són més propenses a patir un infart que els homes
- Les variables age, trtbps, chol, fbs, restecg, thalachh, exng i slp no tenen afectació sobre el succés a predir
- Les variables sex, caa i thall són factors de protecció (un valor baix en aquestes variables afecta en que sigui més probable l'atac de cor)
- La variable cp és un factor de risc (Un valor elevat en aquesta variable fa que sigui més probable l'atac de cor)
- Tot i així, ens fixem que en valor absolut, tenim un nivell bastant baix de correlació entre les variables i el succés output.
- Model GLM: 77,7% accuracy, especificitat de 0.75, sensibilitat de 0.78 i ROC de 0.77
- Model RF: 83% accuracy, especificitat de 80%, sensibilitat de 0.83 i ROC de 0.83

7. Codi.

Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.
enllaç: https://github.com/anthonyswords/HeartAttackEDA

8. Video.

Realitzar un breu vídeo explicatiu de la pràctica (màxim 10 minuts) on tots els integrants de l'equip expliquin amb les seves pròpies paraules el desenvolupament de la pràctica, basant-se en les preguntes de l'enunciat per a justificar i explicar el codi desenvolupat. Aquest vídeo s'haurà de lliurar a través d'un enllaç al Google Drive de la UOC (https://drive.google.com/…), juntament amb l'enllaç al repositori Git lliurat.
enllaç:
https://drive.google.com/drive/folders/1L-T6CwUL_Z0f8nbMN-tnQ3laSrQ0SfHE?usp=share_link

Contribucions	Signatura
Investigació prèvia	A. E. N. *et* J. S. V.
Redacció de les respostes	A. E. N. *et* J. S. V.
Desenvolupament del codi	A. E. N. *et* J. S. V.
Participació al vídeo	A. E. N. *et* J. S. V.