Joan Giné i Gerard Costa

Tipologia i Cicle de Vida de les dades: Pràctica II. Com realitzar la neteja i l'anàlisi de les dades?

1. Descripció del joc de dades escollit

El joc de dades escollit [1] descriu un seguit de variables mèdiques d'un conjunt de pacients. Més concretament, aquest joc de dades se centra en els infarts, els quals són un trastorn vascular que es presenta quan hi ha una massa de sang coagulada que bloqueja de manera total o parcial l'interior d'un vas sanguini, ja sigui una vena o una artèria [3].

Donada l'actual incidència d'aquesta malaltia, és d'especial interès la recerca en la prevenció i el seu tractament. En aquesta línia, aquest joc de dades ens permet, des de l'enfocament de la present pràctica, veure correlacions entre variables o patrons entre les dades que ajudin a explicar o a detectar factors de risc o de protecció enfront aquesta malaltia. Tanmateix, estudis més avançats podrien incloure modelitzacions estadístiques que poguessin predir-la de forma molt acurada, com també l'enriquiment de la mostra considerada. El joc de dades consta de 303 registres.

Així doncs, el present anàlisi busca, en línies generals, respondre si hi ha alguna o algunes variables mèdiques que constitueixin un factor de risc o de protecció significatius de cara al desenvolupament d'un infart, i que aquestes relacions no siguin intuïtives. En aquest sentit, es durà a terme una anàlisi estadística per tal de cercar aquestes relacions entre variables.

A continuació es llisten i s'expliquen cadascuna d'elles.

- **age**: variable quantitativa que indica l'edat del pacient expressada en anys.
- **sex**: sexe del pacient. És una variable qualitativa dicotòmica que pot prendre els valors 1, home, o 0, dona.
- **exng**: el nom de la variable és una abreviació de l'expressió anglesa "exercise induced angina" (angina de pit provocada per l'exercici). Pot prendre els valors 1 i 0, de manera que és una variable qualitativa dicotòmica.
- **caa**: és una variable qualitativa ordinal que pot prendre els valors 0, 1, 2, 3 i 4. És el nombre de vasos sanguinis principals del cor vistos amb la tècnica de la fluoroscòpia.
- **cp**: és una variable qualitativa ordinal que indica el tipus de dolor de pit que experimenta el pacient. Les categories associades a cada valor numèric s'expliquen a continuació:
    - **0**: angina típica.
    - **1**: angina atípica.
    - **2**: dolor que no és d'angina.
    - **3**: asimptomàtic.
- **trtbps**: el nom de la variable és una abreviació de l'expressió anglesa "resting blood pressure", que és la pressió sanguínia en repòs expressada en mil·límetres de mercuri. És, per tant, una variable quantitativa.

- **chol**: variable quantitativa que expressa la concentració de colesterol en sang, expressada en mg/dl.
- **fbs**: el nom de la variable és una abreviació de l'expressió anglesa "fasting blood sugar". És una variable qualitativa dicotòmica que pren el valor 1 en cas que la concentració de sucre en sang en dejú excedeixi els 120 mg/dl i 0 altrament.
- **restecg**: és una variable categòrica ordinal que indica els resultats d'un electrocardiograma. Les categories associades a cada valor numèric es llisten a continuació.
    - **0**: resultat normal.
    - **1**: presenta una anormalitat d'ona ST-T.
    - **2**: presenta hipertròfia ventricular esquerra, ja sigui en una alta probabilitat o bé real.
- **thalachh**: és una variable quantitativa que indica la freqüència cardíaca màxima adquirida.
- **oldpeak**: és una variable quantitativa que indica depressió ST induïda per exercici respecte el repòs. La depressió ST s'observa en un electrocardiograma, i indica normalment isquèmia de miocardi durant l'exercici físic.
- **slp**: pendent del pic durant la prova d'exercici físic. És una variable categòrica ordinal. Les categories associades a cada valor numèric es llisten a continuació.
    - **1**: pendent positiu.
    - **2**: pendent nul.
    - **3**: pendent negatiu.
- **thall**: és una variable categòrica ordinal que indica els resultats del flux sanguini. Les categories associades a cada valor numèric es llisten a continuació.
    - **0**: valor faltant.
    - **1**: defecte fix (existeixen parts del cor on no hi ha flux sanguini).
    - **2**: flux sanguini normal.
    - **3**: defecte reversible (s'observa flux sanguini però no és normal)
- **output**: és la variable resposta. És una variable categòrica dicotòmica que pren el valor 1 en cas d'alta probabilitat d'atac de cor, o 0 altrament.

2. Integració i selecció de les dades d'interès

En aquest apartat es durà a terme una subselecció de les dades presentades a l'apartat anterior. L'objectiu que persegueix aquesta anàlisi estadística és trobar factors de risc o de protecció enfront de la ocurrència d'un atac de cor. No obstant, volem centrar-nos en aquelles variables que siguin fàcilment interpretables. En aquest sentit, en el joc de dades seleccionat n'hi ha algunes que involucren un coneixement que queda fora de l'abast d'aquells qui no tenen formació en medicina. En conseqüència, proposem d'eliminar-ne les següents:

- **restecg**: aquesta variable involucra la interpretació d'un electrocardiograma, fet que dificulta la comprensió de quin efecte pot tenir en la ocurrència d'un atac de cor si no es tenen els coneixements de base necessaris.
- **thall**: aquesta variable involucra la comprensió del funcionament del flux sanguini a algunes zones concretes del cor.
- **slp**: involucra novament la interpretació d'un electrocardiograma.

- **oldpeak**: aquesta variable involucra un concepte de medicina anomenat “ST slope”, que s’observa també en un electrocardiograma.

Per tant, resten 10 variables de les 14 que originalment consten al joc de dades, essent 4 d’elles contínues i 6 de categòriques.

3. Neteja de les dades

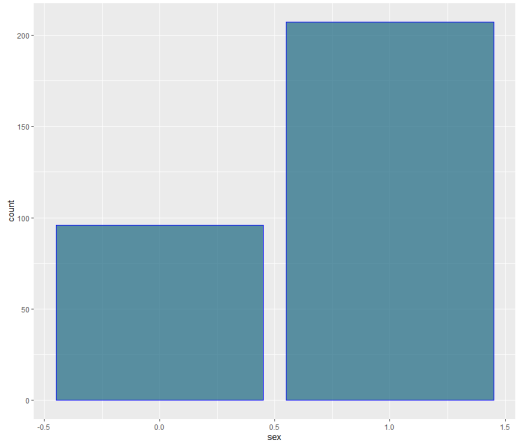
Malgrat que es plantegi descartar 4 variables, l’anàlisi de valors faltants i de valors extrems es durà a terme d’igual manera sobre les 14 variables originals.

Després de carregar el joc de dades a R i d’explorar si hi ha algun valor faltant, s’observa que no n’hi ha. Així doncs, aquests s’hauran de trobar emmascarats com a valors informats a 0. A la figura següent s’observa el que ha retornat la terminal de R.

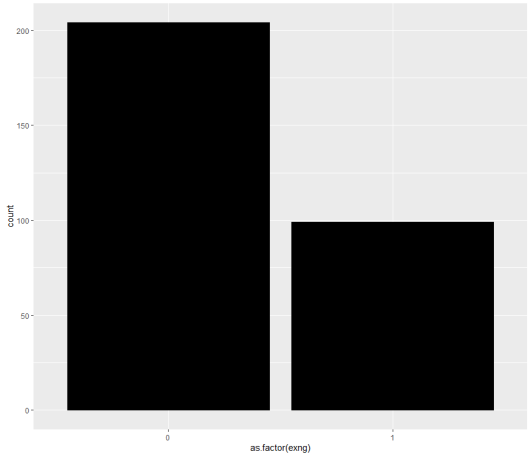
```
> any(is.na(df))
[1] FALSE
```

D’aquesta manera, s’efectuarà una anàlisi per a cada variable amb l’objectiu de detectar valors faltants. Comencem amb les variables categòriques. En aquest cas ens hem de fixar si hi ha valors que no constin a la seva descripció.

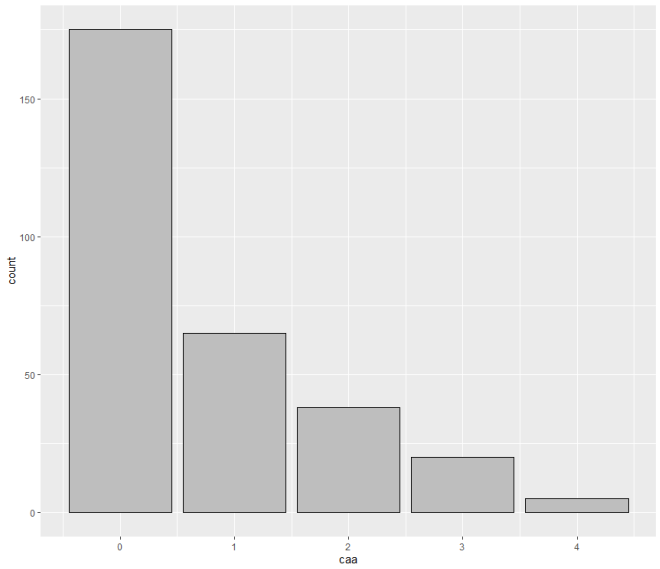
- **Sex**: no s’observen valors faltants. No obstant, s’observen molts més pacients de sexe masculí que de femení.



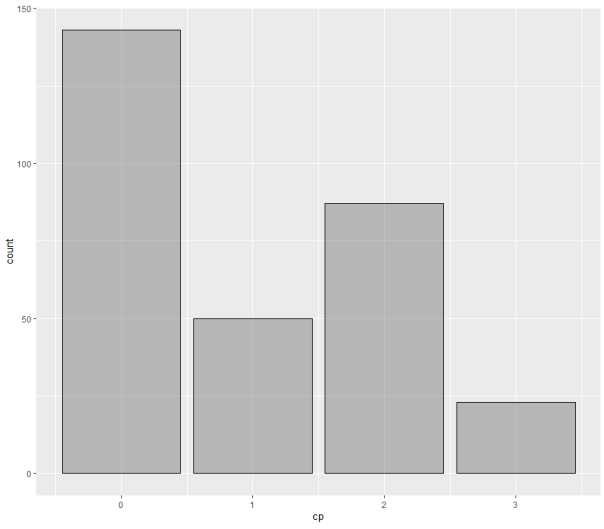
- **Exng**: no s’observen valors faltants.



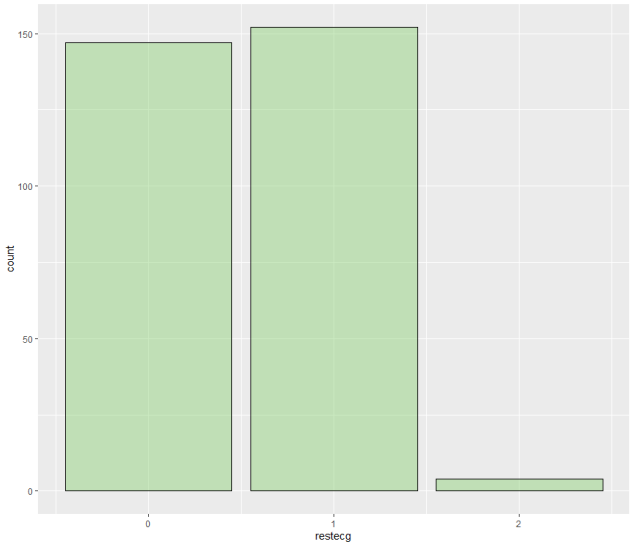
- **Ca**: no s'observen valors faltants.



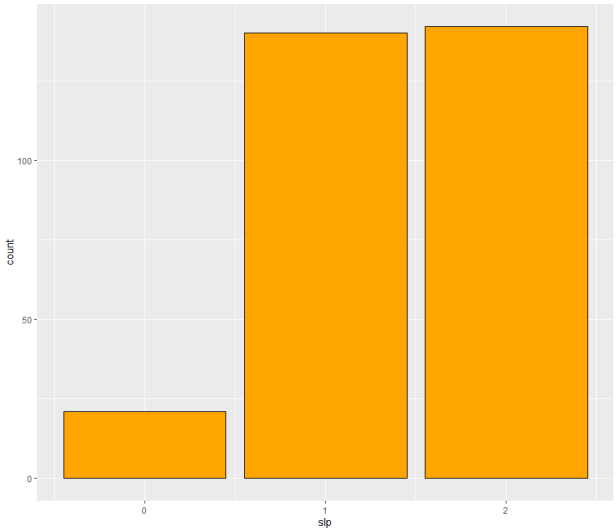
- **Cp**: no s'observen valors faltants.



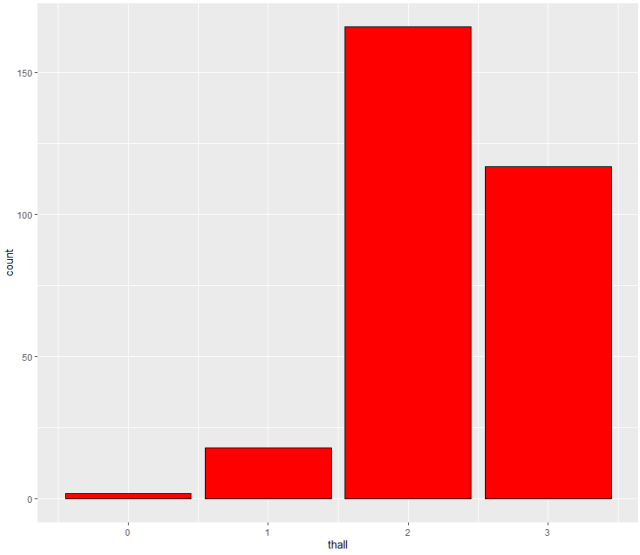
- **Rest_ecg**: no s'observen valors faltants.



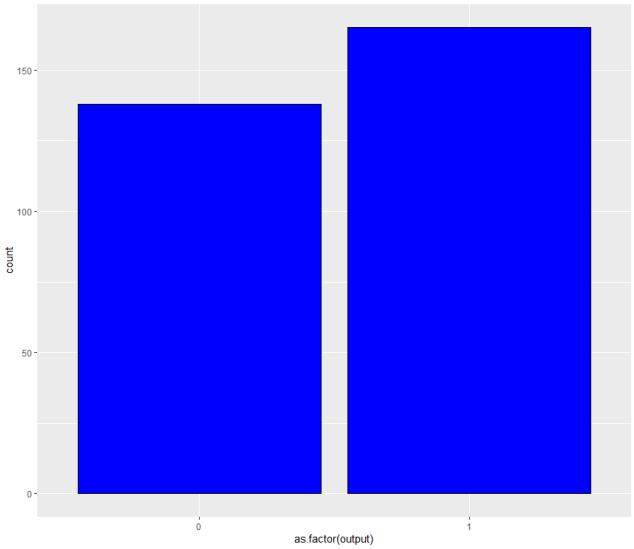
- **Slp:** s'observen 21 valors informats com a '0', categoria no prevista en aquesta variable. S'eliminarà la variable del joc de dades, com s'ha comentat.



- **Thal**: s'observen dos valors faltants informats com a '0', categoria que la variable no té prevista. S'eliminarà la variable del joc de dades, com s'ha comentat.



- **Output**: no s'observen valors faltants.



Cal dir que, en el cas de les variables categòriques, la majoria d'elles tenen el valor 0 com a possible factor. Així doncs, podria haver-hi casos en els quals la variable estigui no informada però que tingui el valor 0 per defecte. No obstant, no podem detectar aquests casos. A continuació, comprovem si hi ha valors incongruents a les variables quantitatives.

- **Age**: no s'observen valors anòmals.

```
> summary(df$age) # Ok, sense valors extrems
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   47.50   55.00   54.37   61.00   77.00
```

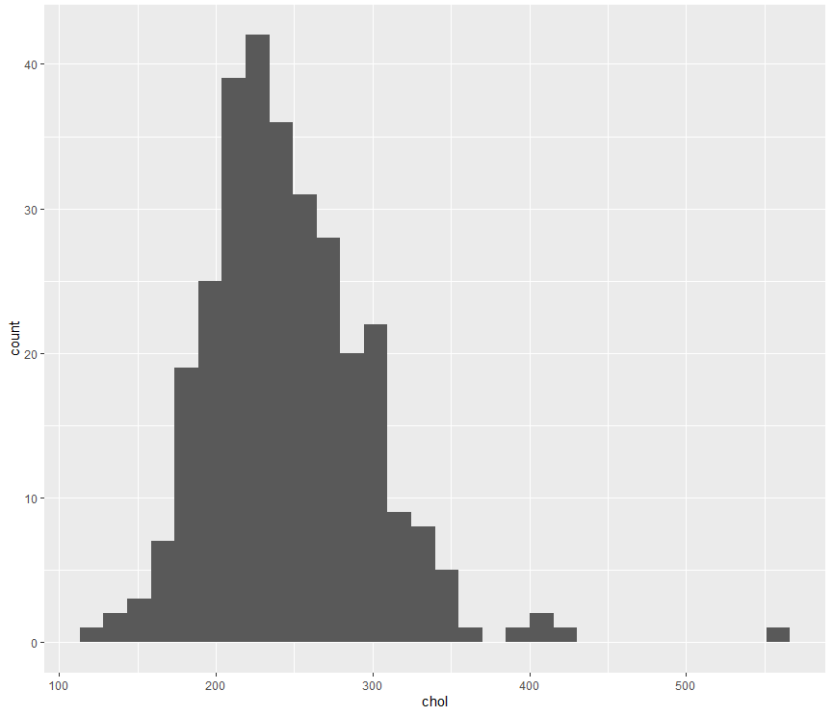
- **Trtbps**: no s'observen valors extrems ni faltants.

```
> summary(df$trtbps) # Ok, sense valors extrems
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   94.0   120.0   130.0   131.6   140.0   200.0
```

- **Chol**: s'observen valors extrems.

```
> summary(df$chol) # "Ok"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  126.0   211.0   240.0   246.3   274.5   564.0
```

S'adjunta a continuació un histograma d'aquesta variable.



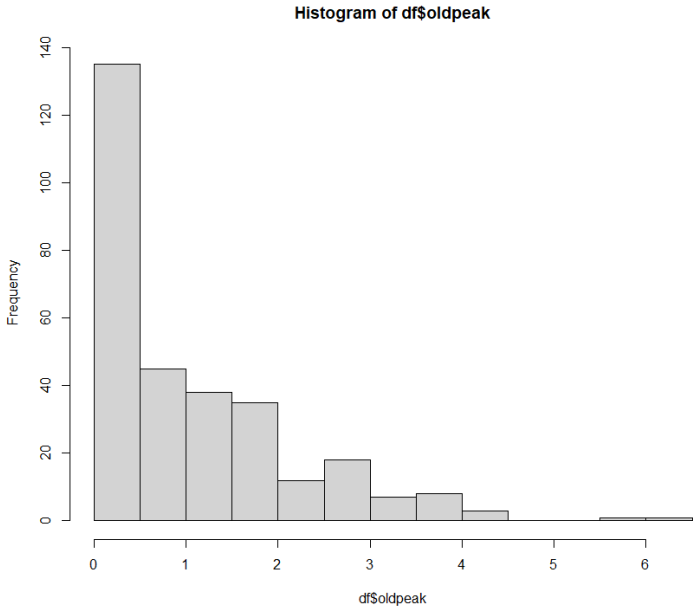
Si bé aquesta distribució té una cua dreta, podem observar que aquesta ocorre al voltant dels 400 mg/dL, que, segons [5], és una concentració perillosament alta però real. No obstant, hi ha un valor, que es correspon amb el màxim de 564 mg/dL vist anteriorment, que es considera poc realista, més encara quan és un sol cas aïllat. Per tant, s'optarà per eliminar aquest registre de la mostra.

- **Thalachh**: no s'observen valors faltants ni valors extrems.

```
> summary(df$thalachh) # Ok
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   71.0   133.5   153.0   149.6   166.0   202.0
```

- **Oldpeak:** s'observa un pic de freqüència a 0, que representa poc menys que la meitat de la mostra**.** Aquesta és una variable difícilment interpretable que es descartarà, però el fet que hi hagi tants valors informats com a 0 podria deure's a que, almenys en part, hi hagi valors sense informar. A continuació s'adjunta un histograma d'aquesta variable.

```
> summary(df$oldpeak) # Ok
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.80    1.04    1.60    6.20
```



Cal destacar que s'observen dos registres de valors extrems al voltant del valor 6.

4. Anàlisi de les dades

A la secció anterior ja s'ha fet una anàlisi univariant de les dades. El següent pas és aprofundir en aquesta anàlisi tot combinant les diferents variables que hem seleccionat.

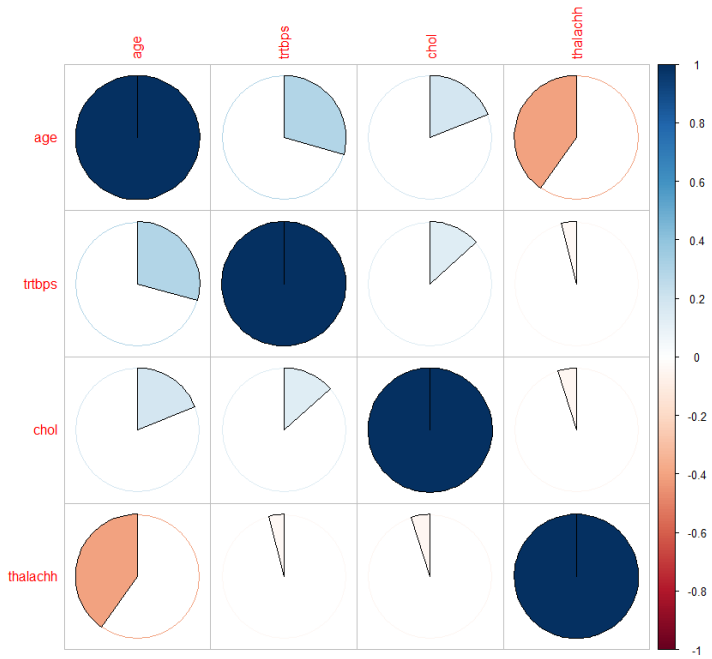
El primer pas que durem a terme serà fer una anàlisi de correlacions. Com que disposem de variables qualitatives i quantitatives, primerament farem dues anàlisis de correlació separades: una per a les variables quantitatives i una altra per a les qualitatives. Comencem per les quantitatives.

El primer que cal fer és comprovar si es compleixen els supòsits per a calcular una matriu de correlacions amb el coeficient de Pearson. En aquest sentit, contrastem primer si les variables quantitatives segueixen una distribució normal. Es presenten a continuació els p-valors resultants d'haver realitzat el contrast de Shapiro-Wilker a les variables consdierades.

Variable	**p-valor Test Shapiro-Wilker**
age	$6.69 \cdot 10^{-3}$
trtbps	$1.60 \cdot 10^{-6}$
chol	$1.10 \cdot 10^{-3}$
thalachh	$7.68 \cdot 10^{-5}$

En tots els casos es rebutja la hipòtesi nul·la. Les dades no estan distribuïdes normalment en cap cas. Sabent això, ja és segur que haurem d'efectuar una prova no paramètrica per tal averiguar

la correlació entre les variables, però igualment efectuem un contrast d'homogeneïtat de variància o homoscedasticitat. S'ha efectuat el contrast de Fligner per cadascuna de les variables, i segmentant-ho segons els factors de les variables 'sex' i 'output'. El resultat en ambdós casos és que el p-valor sempre és inferior a 10^{-16}, de manera que cal descartar la hipòtesi nul·la d'homoscedasticitat. Així doncs, s'ha calculat una matriu de coeficients de correlació de Spearman, que és l'alternativa no paramètrica. S'adjunta a continuació.



Observem que la correlació entre les variables és generalment força baixa; l'únic parell de variables que presenten una correlació moderadament alta són "thalachh" i "age". S'adjunta un scatterplot a continuació.

Inclús en aquest cas, visualment és difícil de veure una correlació entre les dues variables.

A continuació s'ha fet un estudi de correlacions entre les variables discretes. En cada cas, s'ha comprovat si es compleixen les condicions per a aplicar el test de correlació de χ^2, és a dir, que la freqüència esperada de cada element de la matriu de contingència sigui igual o superior a 5; en cas contrari, s'ha dut a terme el test exacte de Fisher. A continuació es recopilen els resultats en dues taules. Partim d'un nivell de significació $\alpha = 0.05$.

variable_1	variable_2	validesa_test	p_valor	Resultat Test
fbs	exng	Test de Fisher	0,73	Variables no correlacionades
output	fbs	Test de Fisher	0,75	Variables no correlacionades
fbs	sex	Test de Fisher	0,49	Variables no correlacionades
cp	sex	Test de Fisher	0,10	Variables no correlacionades
exng	sex	Test de Fisher	0,02	Variables correlacionades
output	sex	Test de Fisher	1,11E-06	Variables correlacionades
output	exng	Test de Fisher	3,44E-14	Variables correlacionades
cp	exng	Test de Fisher	3,05E-15	Variables correlacionades
output	cp	Test de Fisher	1,79E-18	Variables correlacionades
fbs	cp	Test de Chi quadrat	0,26	Variables no correlacionades
fbs	caa	Test de Chi quadrat	0,12	Variables no correlacionades
caa	sex	Test de Chi quadrat	0,11	Variables no correlacionades
caa	exng	Test de Chi quadrat	0,01	Variables correlacionades
cp	caa	Test de Chi quadrat	7,38E-06	Variables correlacionades
output	caa	Test de Chi quadrat	3,65E-15	Variables correlacionades

Cal destacar que hi ha dues variables que estan correlacionades amb el sexe del pacient, concretament "exng" i "output". D'una banda, la referència [10] ens indica que la probabilitat d'infart és la mateixa per a homes com per a dones. Per tant, cal concloure que aquest resultat és una conseqüència d'un possible biaix en la mostra: cal tenir en compte que només comptem amb 302 registres i que la proporció d'homes és predominant. D'altra banda, la referència [7] ens indica que les dones tenen una probabilitat similar o lleugerament superior respecte els homes de patir angina de pit. Per tant, aquesta correlació entre "exng" i "sex" també podria reflectir aquesta característica.

D'altra banda, analitzem les restants correlacions amb la variable "output". A primera vista sembla coherent la correlació amb "cp" i "exng"; no obstant, sembla contradictori que una major concentració de sucre en sang ("fbs") no impliqui una major probabilitat d'infart. Potser es deu al fet que aquesta variable s'ha segmentat només en dues categories, establint una frontera als 120 mg/dl. Per últim, la variable "caa" indica el nombre de vasos sanguinis disponibles al cor. També té sentit que, en funció del nombre de vasos sanguinis disponibles, la probabilitat d'infart variï.

Ja acabant, quant a la variable "fbs" obervem que no està correlacionada amb cap de les variables "exng", "cp" i "caa". Si bé és cert que la concentració de sucre en sang, a no ser que sigui en proporcions molt importants, no té unes conseqüències tan importants com, per exemple, un elevat percentatge de colesterol en sang o sobrepès, creiem novament que la segmentació d'aquesta variable li ha restat poder explicatiu.

Per últim, observem correlació entre “cp” i “exng”, i entre “cp” i “caa”. És congruent que el dolor de pit (“cp”) estigui correlacionat amb l’angina de pit (“exng”) i el nombre de vasos sanguinis disponibles (“caa”). També existeix correlació entre “caa” i “exng”, fet que és coherent.

Fins ara hem analitzat les variables categòriques i quantitatives de forma separada. A continuació, doncs, es durà a terme una anàlisi que involucri variables dels dos tipus. En aquesta línia, es proposa de fer un estudi d’anàlisi de la variància. Com que abans hem vist que les variables quantitatives no seguien una distribució normal, en lloc d’emprar ANOVA farem tests de Kruskal-Wallis per tal de buscar les diferències entre els factors de les diferents variables categòriques en relació a les contínues. Treballem novament amb un nivell de significació $\alpha = 0.05$. Els resultats són els següents:

variable_cat	variable_cont	p_value	Resultat Test
sex	age	0.11	H0 certa
sex	trtbps	0.30	H0 certa
sex	chol	0.01	Diferències significatives entre grups
sex	thalachh	0.51	H0 certa
exng	age	0.10	H0 certa
exng	trtbps	0.38	H0 certa
exng	chol	0.09	H0 certa
exng	thalachh	3.74E-12	Diferències significatives entre grups
caa	age	3.81E-10	Diferències significatives entre grups
caa	trtbps	0.25	H0 certa
caa	chol	0.034	Diferències significatives entre grups
caa	thalachh	4.83E-6	Diferències significatives entre grups
cp	age	0.01	Diferències significatives entre grups
cp	trtbps	0.05	H0 certa
cp	chol	0.39	H0 certa
cp	thalachh	2.20E-10	Diferències significatives entre grups
fbs	age	0.04	Diferències significatives entre grups
fbs	trtbps	0.01	Diferències significatives entre grups
fbs	chol	0.72	H0 certa
fbs	thalachh	0.81	H0 certa
output	age	2.23E-5	Diferències significatives entre grups
output	trtbps	0.04	Diferències significatives entre grups
output	chol	0.03	Diferències significatives entre grups
output	thalachh	1.14E-13	Diferències significatives entre grups

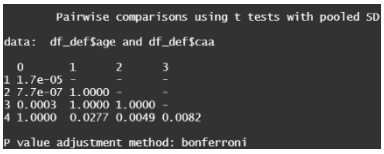
A l'obtenir aquesta taula després de fer el test a totes les combinacions possibles, veiem que hi ha diferències entre les variables:

- **Sex i chol**: homes i dones tenen diferents nivells de concentració de colesterol en sang. Novament, això pot ser conseqüència del desequilibri entre proporcions home-dona a la mostra.
- **Exng i thalachh**: segons si van tenir una angina de pit provocada per l'exercici o no, els pacients tenien diferents nivells de freqüència cardíaca màxima, fet que és congruent.
- **Caa i age**: hi ha diferencies d'edat entre els individus amb diferents nombres de vasos sanguinis del cor.
- **Caa i chol**: existeixen diferències de nivell de colesterol entre els individus amb diferents nombre de vasos sanguinis del cor.
- **Caa i thalachh**: hi ha diferències de freqüència cardíaca màxima segons el nombre de vasos sanguinis del cor.
- **Cp i age**: hi ha diferències d'edat entre els diferents grups d'individus amb el tipus de dolor de pit diferent.
- **Cp i thalachh**: existeixen diferències de freqüència cardíaca màxima entre els grups d'individus amb tipus de dolor de pit diferent.
- **Fbs i age**: per tant, a edats diferents, concentracions de sucre diferents.
- **Fbs i trtbps**: hi ha diferències en la pressió sanguínia en repòs dels pacients de grups amb la concentració de sucre diferent.
- **Output i age**: per tant, hi ha diferències d'edat entre els pacients susceptibles a tenir un atac de cor i els que no.
- **Output i trtbps**: existeixen diferències de pressió sanguínia entre els pacients susceptibles a tenir un atac de cor i els que no.
- **Output i chol**: hi ha diferències de nivell de concentració de colesterol en sang entre els pacients susceptibles a tenir un atac de cor i els que no.
- **Output i thalachh**: per tant, hi ha diferències en els nivells de freqüència cardíaca màxima entre els pacients susceptibles a tenir un atac de cor i els que no.

Ara bé, el test de Kruskal Wallis no ens diu, dins d'una variable categòrica, a quins factors la variable contínua té mitjanes diferents, i com de diferents són.

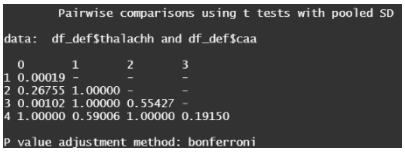
En els casos en els quals en una variable categòrica hi hagi més de dos factors, podem fer una prova post hoc per veure entre quins d'ells hi ha diferències. Per això, farem un t-test per a totes les possibles parelles utilitzant l'ajust de Bonferroni. I els resultats es mostren a continuació.

- Diferències d'edat entre els diferents grups de nombre de vasos sanguinis (age i caa): Sembla que les majors diferències es troben entre els que no se'ls veu cap vas sanguini i els que se'ls hi veu algun, excepte entre 0 i 4. També entre 4 i 1, 4 i 2, i 4 i 3 també hi trobem diferències en la mitjana d'edat.

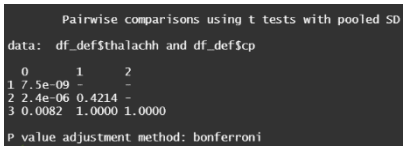


```
        Pairwise comparisons using t tests with pooled SD

data:  df_def$age and df_def$caa

  0       1       2       3
1 1.7e-05 -       -       -
2 7.7e-07 1.0000  -       -
3 0.0003  1.0000  1.0000  -
4 1.0000  0.0277  0.0049  0.0082

P value adjustment method: bonferroni
```

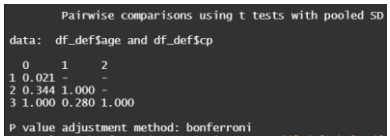
- Diferències de freqüència cardíaca màxima entre els diferents grups de nombre de vasos sanguinis: les úniques diferències són entre 0 i 1 o 0 i 3.



- Diferències de freqüència cardíaca màxima entre els diferents grups de dolor de pit del pacient: entre l'angina típica (0) i els altres grups (1, 2, 3) trobem diferències de BPM.



- Diferències entre l'edat i el tipus de dolor de pit del pacient: Sobretot es veu diferència d'edat entre el tipus 0 i el tipus 1, és a dir, entre l'angína típica i l'atípica.



Seguidament, i ara que ja sabem quines són les variables més relacionades entre elles i, sobretot, més relacionades amb la variable resposta, "output", provarem d'ajustar un parell de models predictius.

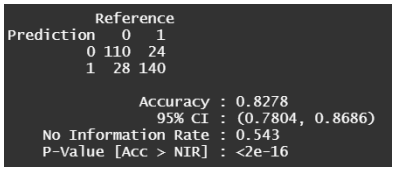
Primer comencem per una **regressió logística**, indicada per predir variables dicotòmiques, com és el nostre cas.

Primerament, com que s'ha vist que la variable fbs no aporta gaire poder explicatiu, es descarta de cara al model. No obstant, s'ha comprovat que el model amb la variable fbs és de pitjor qualitat que el proposat. En resum, les variables seleccionades són age, sex, exng, cp, caa, chol, thalachh i trtbps.

Al model resultant veiem que el regressor corresponent a age no és estadísticament significatiu, de manera que descartem aquesta variable del model i el tornem a entrenar.

El resultat és que aquesta exclusió fa disminuir l'indicador AIC de 245 a 244, i és per això que el segon model de regressió logística sense la variable age és millor.

La primera prova de prediccions la fem amb les mateixes dades que d'entrenament i obtenim una molt bona Accuracy, vegi's la matriu de confusió:



No obstant, podria ser que fos degut a un sobreajust, ja que s'ha entrenat el model amb la totalitat de la mostra i això pot fer que el model resultant no sàpiga generalitzar degudament.

La mètrica més interessant segons el context hauria de ser la taxa de vertaders positius, és a dir, quin percentatge de possibles atacs de cor ha encertat bé el model respecte el total de positius reals, que és la mètrica que s'hauria de maximitzar. En aquest cas és:

$$Precision = \frac{140}{140 + 24} \approx 85.38\%$$

Seguidament es torna a entrenar el model anterior, però dividint el dataset de forma aleatòria segons mostres de train i test amb una proporció de 70% i 30% de la mostra total, respectivament.

Tornem a predir, però ara amb les dades de test. El resultat no s'allunya massa de l'anterior; per tant, és un model prou bo i robust ja que no està sobreentrenat i pot predir correctament un 80% de observacions.

```
          Reference
Prediction  0  1
         0 32  9
         1  9 40

               Accuracy : 0.8
                 95% CI : (0.7025, 0.8769)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 3.697e-07
```

Mirant la matriu, calculem també la taxa de vertaders positius i veiem que la dada continua sent bona.

$$Precision = \frac{40}{40+9} \approx 81.63\%$$

Veiem, no obstant, un lleuger decreixement en la precisió del model.

Seguidament provem un altre **mètode supervisat, el Random Forest**. Per això aprofitarem els dos datasets creats anteriorment de train i test per tal d'entrenar i validar el model.

Els resultats també són molt bons; aconsegueix predir correctament un 80% de les observacions de la mostra de test; de fet, prediu exactament el mateix que la regressió logística.

```
          Reference
Prediction  0  1
         0 32  9
         1  9 40

               Accuracy : 0.8
                 95% CI : (0.7025, 0.8769)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 3.697e-07
```

Noti's que, a l'igual que en les altres prediccions, la precisió obtinguda també és la mateixa que en el cas anterior.

$$Precision = \frac{40}{40+9} \approx 81.63\%$$

6. Resolució del problema- Conclusions

En la present pràctica s'ha dut a terme una anàlisi estadística sobre un joc de dades mèdiques; concretament, sobre infarts. S'han descartat variables que eren difícilment interpretables sense una base de coneixement de medicina, i a continuació s'ha eliminat un registre de la mostra que contenia un valor extrem en la concentració de colesterol.

A efectes de fer una primera anàlisi exploratòria, s'ha explorat la distribució de cadascuna de les variables. S'han trobat dues variables categòriques amb valors faltants que, pel motiu esmentat al paràgraf anterior, s'han eliminat del joc de dades. Cal destacar que una variable força important com és el sexe del pacient conté força més registres de dona que no pas d'home. D'altra banda, s'ha eliminat un registre per causa d'un valor extrem a la variable 'chol', com també s'ha explicat.

A continuació s'ha analitzat la correlació entre variables categòriques i contínues separadament. Quant a les variables categòriques, destaca la correlació d'algunes variables amb 'sex', que, com s'ha dit, segurament es deu al desequilibri entre les subpoblacions dels dos sexes. D'altra banda,

quant a la variable output, que és la variable resposta, es troba que totes les variables considerades hi estan correlacionades menys 'fbs', la qual, de fet, no està correlacionada amb cap altra variable, segurament per haver-la segmentat només en dues categories, fet que li ha llevat el poder explicatiu que altrament podria haver tingut. S'observen correlacions plausibles entre altres variables com ara 'caa' i 'exng'.

Referent a les variables quantitatives, s'ha dut a terme una matriu de correlacions segons el coeficient de Spearman, que és l'alternativa no paramètrica al de Pearson. S'ha trobat que la correlació més significativa ha estat entre 'thalachh' i 'age'; altrament, els valors trobats han estat força baixos.

A continuació s'ha plantejat una altra anàlisi a efectes d'explorar correlacions entre variables quantitatives i qualitatives. Al fer els test de Kruskal-Wallis per veure si hi havia diferències de mitjanes de de les diferents variables contínues entre els factors de les diferents variables categòriques. Hem trobat algunes relacions curioses, com per exemple que sí que hi ha diferències significatives del nivell de colesterol segons el sexe. O que, per exemple, totes les variables contínues examinades excepte una presenten diferències segons el nombre de vasos sanguinis del pacient.

Resulta d'especial interès haver explorat les variacions de les variables quantitatives segons la variable resposta, output.

En primer lloc, s'ha practicat una regressió logística amb totes les variables del dataset menys 'fbs'. S'ha vist, primerament, que la variable quantitativa age no tenia un regressor estadísticament significatiu, de manera que s'ha retirat del model i s'ha reentrenat amb la resta de variables. No obstant, s'ha vist que la mitjana d'edat segons la variable output sí que presenta diferències significatives. En conseqüència, probablement aquesta variable no contribueixi al model per causa de la colinealitat que presenta amb thalachh, que hem vist que tenia un coeficient de correlació d'aproximadament -0,4. La precisió obtinguda pel model ha estat del 80%.

Addicionalment, s'ha construït un model supervisat de tipus Random Forest, entrenat mitjançant cross-validation, que també ha pogut predir exactament les mateixes mostres que el model de regressió logística.

A més, també s'ha vist que la quantitat de mostres on hi ha probabilitat d'atac de cor i que han estat ben predites (Precisió) és d'un 81,63% en els dos models. Entenem que amb el context és l'indicador que cal maximitzar, ja que l'important és predir correctament els possibles atacs de cor. Per tant, si en un futur s'haguessin d'aplicar més modificacions, algoritmes o paràmetres, s'hauria de comprovar si l'algoritme prediu millor aquests possibles atacs de cor mediant la precisió.

Ja acabant, la realització d'aquesta anàlisi ens ha permès d'aprofundir en un problema de salut pública que és endèmic al nostre país, com són les malalties cardiovasculars en general i els infarts en particular. S'ha explorat el joc de dades en profunditat. No s'han trobat relacions entre variables que fossin especialment contraintuïtives, però sí que hem pogut detectar biaixos en la mostra que ens han permès no extreure conclusions errònies (sobretot, quant a les diferències entre sexes). D'altra banda, s'han modelitzat algoritmes supervisats i el rendiment trobat és prou bo, si bé és necessari obtenir rendiments extraordinàriament elevats en aquestes situacions.

Referències

1. Kaggle (2021): *Heart Attack Analysis & Prediction* Dataset. Extret de https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
2. Watanabe, T., Akutsu, Y., Yamanaka, H., Michihata, T., Okazaki, O., Katagiri, T., & Harumi, K. (2000). Exercise-induced ST-segment depression: imbalance between myocardial oxygen demand and myocardial blood flow. *Acta cardiologica*, *55*(1), 25–31. https://doi.org/10.2143/AC.55.1.2005714
3. Gencat: Canal Salut (2023): *Infart agut de miocardi*. Extret de https://canalsalut.gencat.cat/ca/salut-a-z/i/infart-agut-de-miocardi/index.html
4. MedicalNewsToday (16 de gener de 2019): *What casues high cholesterol?*. Extret de https://www.medicalnewstoday.com/articles/9152#_noHeaderPrefixedContent
5. ELO Smart Nutrition (2023): *Total Cholesterol: 400 mg/dL*. Extret de https://www.elo.health/biomarkers/total-cholesterol-overview/400/
6. TempleHealth (6 de juliol de 2020): *Heart Attack Symtpoms: Are They Different For Men And Women?* Extret de https://www.templehealth.org/about/blog/heart-attack-symptoms-men-women-differences
7. AHA Journals (17 març 2008): *Prevalence of Angina in Woman versus Men.* Extret de https://www.ahajournals.org/doi/10.1161/circulationaha.107.720953

Contribucions

Contribucions	**Signatura**
Investigació prèvia	Joan Giné, Gerard Costa
Redacció de les respostes	Joan Giné, Gerard Costa
Desenvolupament del codi	Joan Giné, Gerard Costa
Participació al vídeo	Joan Giné, Gerard Costa