Pràctica 2

Tipologia i cicle de vida de les dades

Juny 2023
Pau Bernabé
Sergi Crespi

1. Descripció del dataset

El dataset escollit es pot trobar a l'adreça: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
S'ha escollit aquest dataset, ja que és el set de dades proposat i és un bon exemple de dades mèdiques, en les quals es poden veure característiques dels pacients de la mostra, des de l'edat i el sexe fins a nivells de colesterol, ritme cardíac, tipus de dolor al pit que tenien... que poden ser bones característiques de cara a classificar, o fins i tot arribar a predir, segons aquestes característiques, quins pacients podrien arribar a tenir un atac de cor. Tanmateix, amb aquesta mostra de la població del dataset, es podrien arribar a extreure algunes conclusions tals com si el gènere i l'edat, així com els nivells de colesterol, per exemple, poden ser factors que contribueixin al fet de tenir un atac de cor.
Un exemple de les dades seria el següent:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

2. Integració i selecció

Les dades que tenim al dataset, i els seus tipus, així com una breu descripció:

Columna	Tipus	Descripció
age	Integer	Edat del pacient de la mostra
sex	Integer	Sexe del pacient de la mostra
cp	Integer	Tipus de dolor al pit, 1- Angina típica, 2- Angina atípica, 3- Dolor no d'angina, 4- Asimptomàtic.
trtbps	Integer	Pressió de la sang en estat de descans (en mm Hg)
chol	Integer	Nivel de colesterol captat mitjançant un sensor BMI, en mg/dl

fbs	Integer	Fasting blood sugar, si és major de 120mg/dl té el valor d'1. Si no ho és, de 0
restecg	Integer	Resultats de l'electrocardiogràfic en estat de repòs. 0- Valor normal, 1- Té l'abnormalitat d'ona ST-T, 2- Mostra hipertròfia ventricular al ventricle esquerra segons el criteri d'Estes
thalachh	Integer	Màxim ritme cardíac
exng	Integer	Angina de pit induida per exercici, 1- Sí, 2-No
oldpeak	Float	Valor de la depressió en el segment ST d'un electrocardiograma
slp	Integer	La pendent del pic al segment ST de l'electrocardiograma, 0- pendent cap avall, 1- plana, 2- pendent cap amunt
caa	Integer	Nombre de vasos sanguinis majors colorejats per la fluoroscòpia (de 0 a 3)
thall	Integer	Thalassemia, desordre a la sang, 1- presente el defecte, 2- normal, 3- defecte reversible
output	Integer	Possibilitat d'un atac de cor, 0- Menor possibilitat, 1- Major possibilitat

Es pot dividir el set de dades en les variables que són **numèriques** i les que són **categòriques**. Les variables numèriques són: `age`, `trtbps`, `chol`, `thalachh` i `oldpeak`. Les variables categòriques son: `sex`, `cp`, `fbs`, `restecg`, `exng`, `slp`, `caa`, `thall` i `output`.

A priori s'han fet servir totes les dades disponibles al dataset. En aquest pas, també hi ha la possibilitat d'emprar dades d'altres datasets o bé combinar les variables ja presents per tal de formar-ne de noves. En aquest cas, no s'han dut a terme aquestes iniciatives ja que, les dades actuals ja són rellevants. S'ha revisat també quin era el nombre de valors únics al set de dades, i queda la següent fotografia:

	Valors únics
age	41
sex	2
cp	4
trtbps	47
chol	145
fbs	2
restecg	3
thalachh	89
exng	2
oldpeak	38
slp	3
caa	5
thall	4
output	2

3. Neteja de les dades

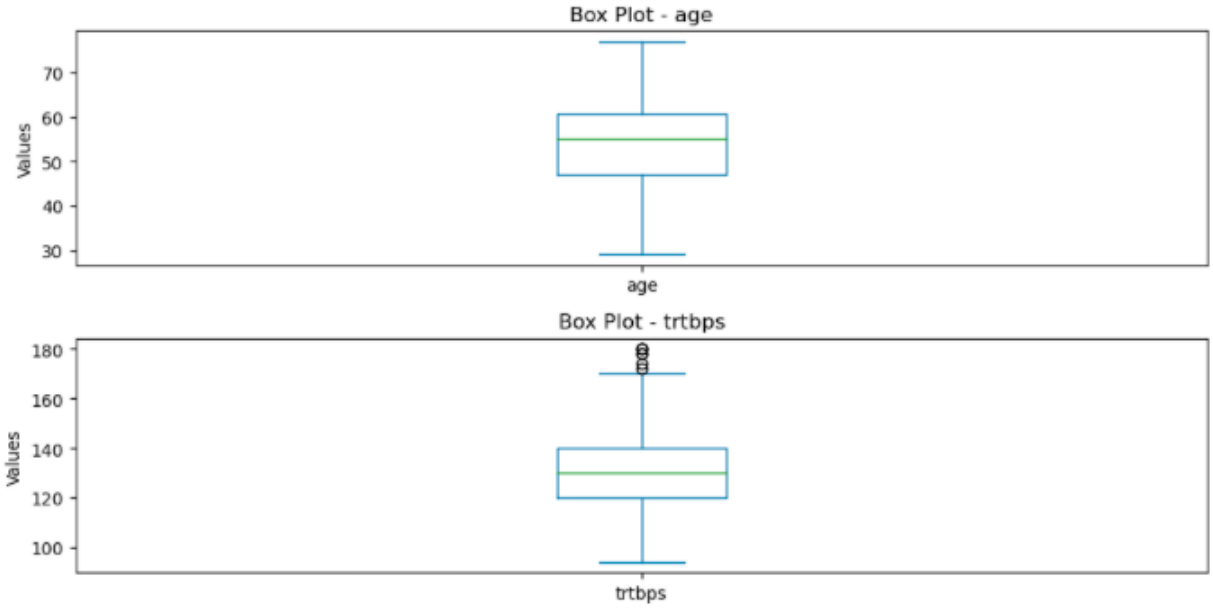
En el procés de neteja de dades, s'ha revisat primerament que no hi hagués valors nuls, i no n'hi ha:

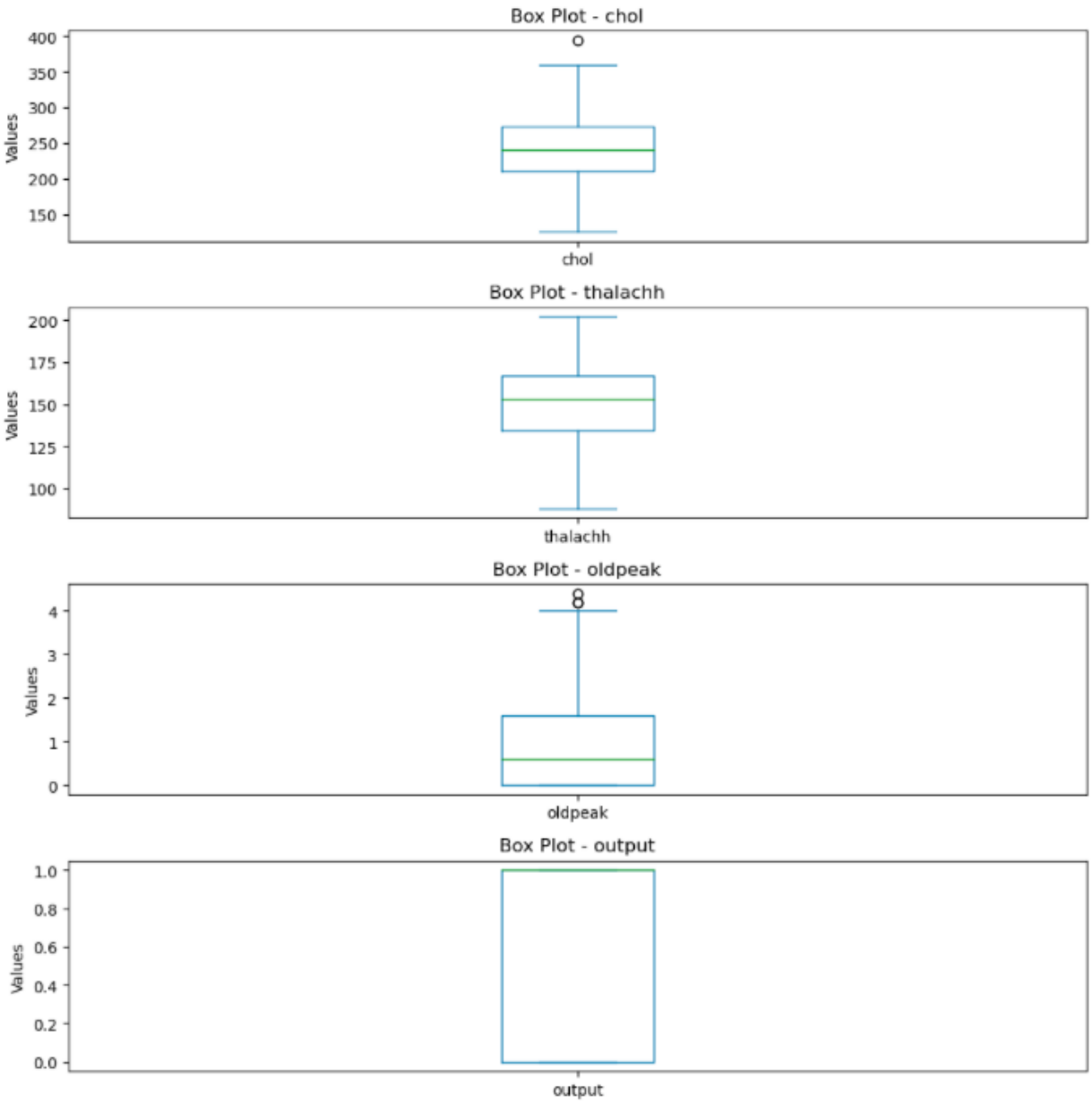
	Valors nuls
age	0
sex	0
cp	0
trtbps	0
chol	0
fbs	0
restecg	0
thalachh	0
exng	0
oldpeak	0
slp	0
caa	0
thall	0
output	0

També s'ha revisat que no hi haguessin NAs, i tampoc n'hi ha:

	Valors absents(NA)
age	0
sex	0
cp	0
trtbps	0
chol	0
fbs	0
restecg	0
thalachh	0
exng	0
oldpeak	0
slp	0
caa	0
thall	0
output	0

Es revisa si hi ha valors extrems, i veiem que si hi ha outliers. Tot i així, no hi ha gaires mostres en el conjunt de dades, així que s'ha de tenir en compte en el tractament d'aquests valors extrems. Primerament s'havia emprat una metodologia diferent a la que hi ha al codi finalment, que era tractar els outliers per sota del primer quartil i per sobre del tercer quartil, assignant el valor d'aquests, però desvirtuava la mostra. Finalment s'ha optat per a treue els outliers els valors dels quals superin 3 cops la desviació estàndard de la seva mitjana. S'han eliminat 9 registres del dataset. Si veiem els boxplots una vegada tractats els outliers:


Box Plot - age
Values
70
60
50
40
30
age
Box Plot - trtbps
Values
180
160
140
120
100
trtbps


Box Plot - chol
Values
chol
Box Plot - thalachh
Values
thalachh
Box Plot - oldpeak
Values
oldpeak
Box Plot - output
Values
output

4. Anàlisi de les dades

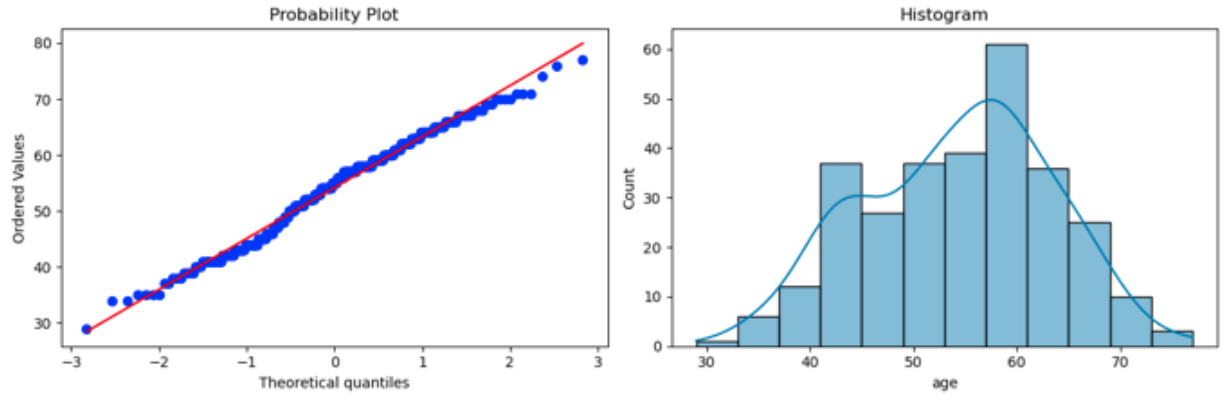
4.1 Selecció dels grups de dades que es volen analitzar/comparar

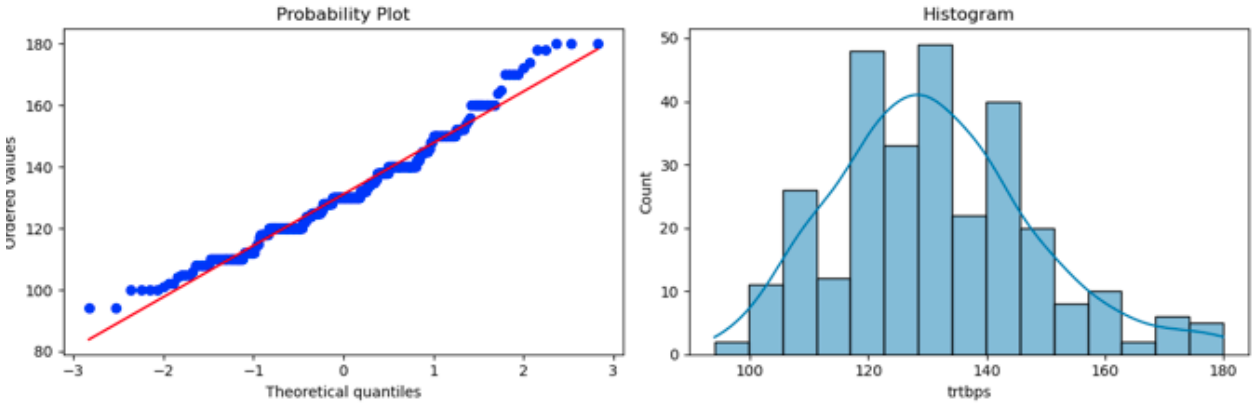
Es tractaran amb totes les dades disponibles del dataset, després de les correccions aplicades (eliminació d'outliers).

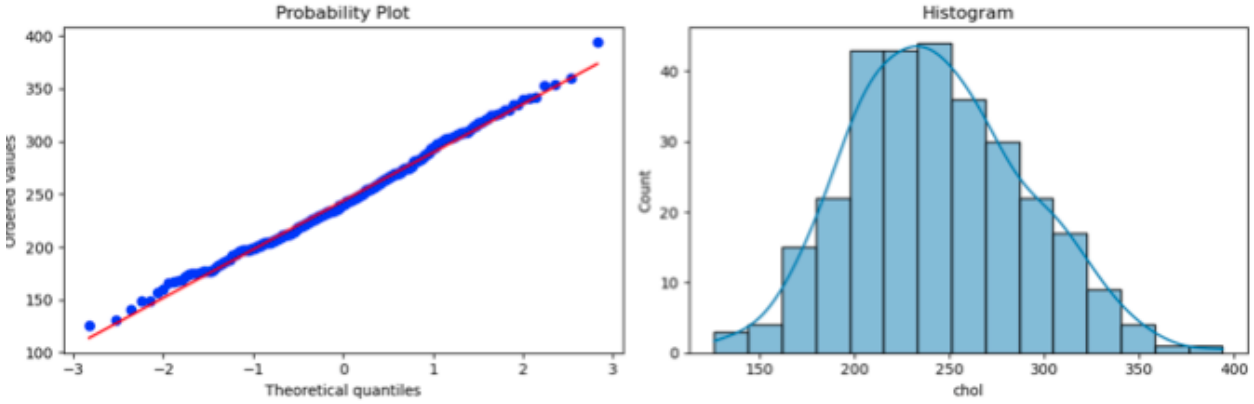
4.2 Anàlisi de la normalitat i homogeneïtat de la variància

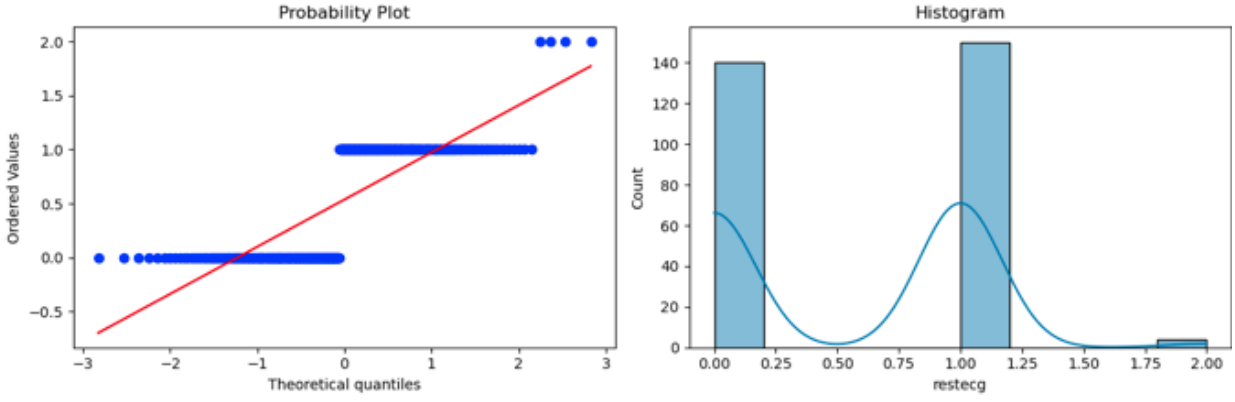
Normalitat

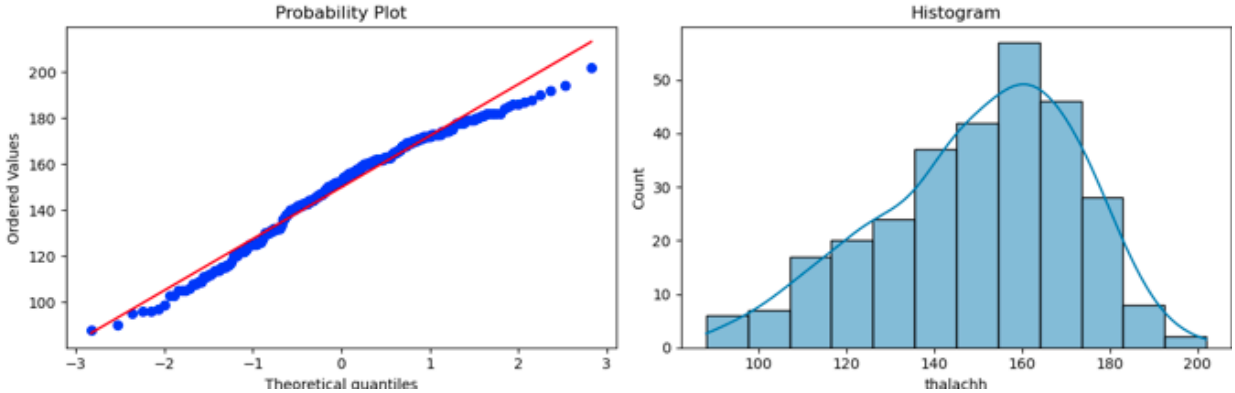
Cal revisar el q-q plot i el seu histograma per a veure si els atributs poden ser candidats a la normalització:




Probability Plot
Ordered Values
Theoretical quantiles
Histogram
Count
trtbps


Probability Plot
Ordered Values
Theoretical quantiles
Histogram
Count
chol


Probability Plot
Ordered Values
Theoretical quantiles
Histogram
Count
restecg


Probability Plot
Ordered Values
Theoretical quantiles
Histogram
Count
thalachh

On veiem que alguns atributs podrien ser candidats a una distribució normal. Fem el test de Shapiro-Wilk de cara a comprovar-ho, i els resultats:

	Columna	Test Statistic	p-value	dist.normal
0	age	0.987135	1.010232e-02	NO
1	sex	0.579333	4.296983e-26	NO
2	cp	0.793169	4.976597e-19	NO
3	trtbps	0.973450	2.891042e-05	NO
4	chol	0.992778	1.663795e-01	SI
5	fbs	0.420666	1.145049e-29	NO
6	restecg	0.680289	3.198513e-23	NO
7	thalachh	0.975179	5.586504e-05	NO
8	exng	0.591151	8.702644e-26	NO
9	oldpeak	0.853299	4.870273e-16	NO
10	slp	0.738087	3.019214e-21	NO
11	caa	0.721607	7.682854e-22	NO
12	thall	0.751531	9.685561e-21	NO
13	output	0.632162	1.147621e-24	NO

Es pot veure als resultats que els atributs no tenen una distribució normal, ja que el p-value és inferior al 0.05, per tant, es rebutja la hipòtesi nul·la i s'afirma que no és normal. Tot i així, pel teorema central del límit, si hi ha més de 30 elements, condició que es compleix, es pot afirmar que es pot aproximar a una distribució normal de mitjana 0 i desviació estàndard 1.

Anàlisi de la homogeneïtat de la variància

```
stats.levene(df['output'], df['chol'], center='median')
```

```
LeveneResult(statistic=510.1096393976572, pvalue=9.996251613386838e-82)
```

Emprant el mètode de Levene per comprovar la homogeneïtat de la variància en la variable **output** i a la variable **chol**, el p-valor és menor a 0.05 i es pot afirmar que no hi ha diferències entre les variàncies d'aquests dos grups.

Es fan més anàlisis de la homogeneïtat de la variància en el següent apartat.

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

Test d'hipòtesi

Com que únicament la variable `chol` compleix el requisit de normalitat i homogeneïtat de la variància, s'hi podran aplicar proves paramètriques.

Pels altres grups, s'hauran d'aplicar proves no paramètriques. No obstant això, per aplicar aquest mètode cal comprovar la igualtat de variàncies entre els grups. En aquest cas s'empra el mètode de fligner.

Test d'hipòtesi per les variables output i chol

```
stats.ttest_ind(df['output'], df['chol'], equal_var=True)
```

```
Ttest_indResult(statistic=-90.89521560278598, pvalue=0.0)
```

Amb aquest test es pot confirmar que els dos grups, output i chol, tenen diferències estadísticament significatives. Això es pot afirmar si es té en compte el p-valor<0.05.

Test d'hipòtesi per les variables output i sex

El primer pas serà comprovar la homogeneïtat de la variància:

```
stats.fligner(df['output'], df['sex'])
```

```
FlignerResult(statistic=12.165990605003254, pvalue=0.00048668810271415596)
```

Emprant el mètode de fligner per comprovar la homogeneïtat de la variància en la variable output i a la variable sex, el p-valor és menor a 0.05 i es pot afirmar que no hi ha diferències entre les variàncies d'aquests dos grups. Ara es procedeix a aplicar el test de Wilcoxon.

```
stats.wilcoxon(df['output'], df['sex'])
```

```
WilcoxonResult(statistic=6370.0, pvalue=0.0023074891975701935)
```

Amb aquest test es pot confirmar que els dos grups, output i sex, no tenen diferències estadísticament significatives. Això es pot afirmar si es té en compte el p-valor<0.05.

Test d'hipòtesi per les variables output i age

Com en el cas anterior, el primer pas serà comprovar la homogeneïtat de la variància:

```
stats.fligner(df['output'], df['age'])
```

```
FlignerResult(statistic=333.66752504627107, pvalue=1.5275965360150919e-74)
```

Emprant el mètode de fligner per comprovar la homogeneïtat de la variància en la variable output i a la variable age, el p-valor és menor a 0.05 i es pot afirmar que no hi ha diferències entre les variàncies d'aquests dos grups.

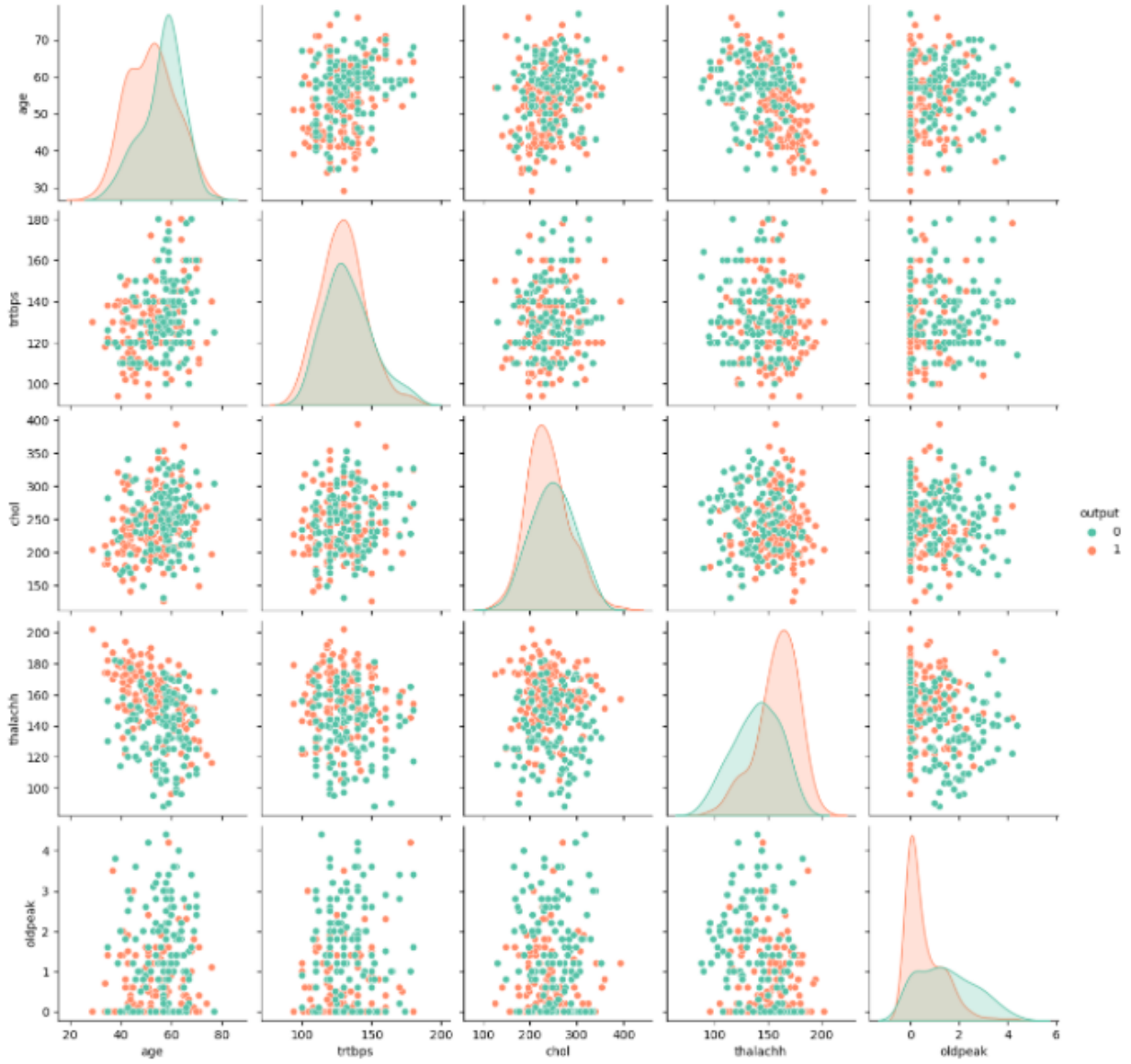
```
stats.wilcoxon(df['output'], df['age'])
```

```
WilcoxonResult(statistic=0.0, pvalue=5.636539390181292e-50)
```

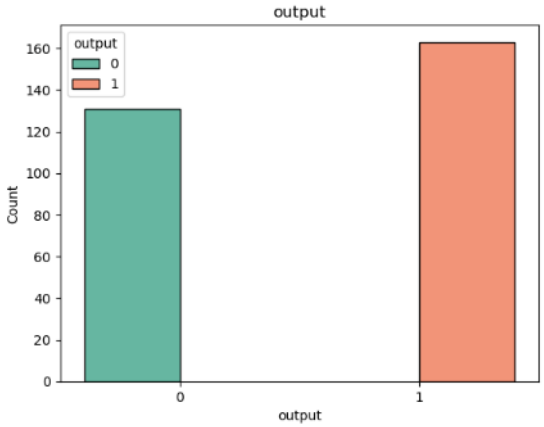
Amb aquest test es pot confirmar que els dos grups, output i age, no tenen diferències estadísticament significatives. Això es pot afirmar si es té en compte el p-valor<0.05.

Anàlisi de les variables

Anem a veure les dades desde diferents prismes de cara a conèixer-les millor. Primerament comencem amb un kde, Kernel Density Estimation, per a veure la distribució de les variables. Al kde el que es fa és sumar els kernels de cadascun dels valors del dataset i sumar-los, així surt una gràfica de la qual podrem interpretar on hi ha més densitat de data points. Obtenim la següent visualització:

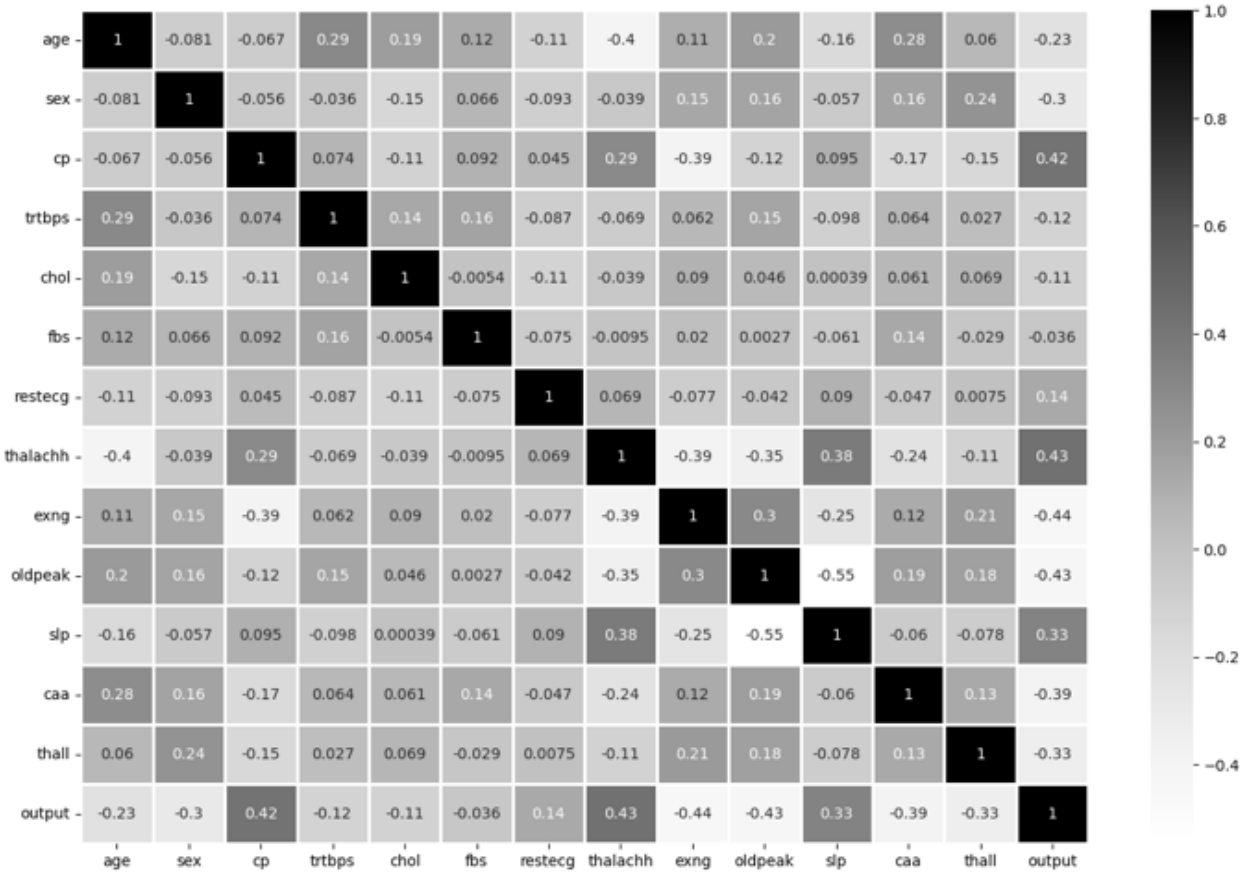


Respecte a les variables categòriques, se'n pot veure la distribució amb uns diagrames de barres:

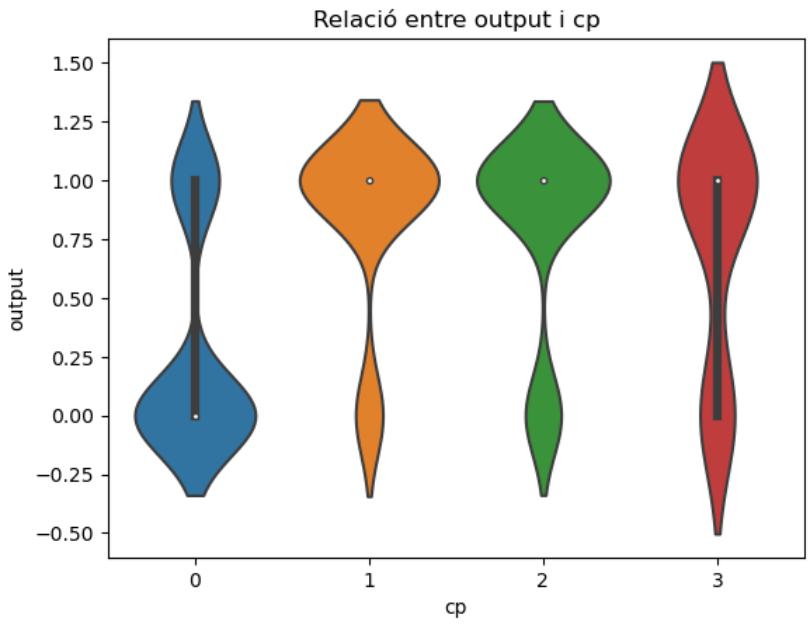


Veiem que hi ha lleugerament més casos de major risc d'atac de cor que de menor risc d'atac de cor en el conjunt de dades. S'han destacat dues d'aquestes visualitzacions però la resta estan presents al codi.

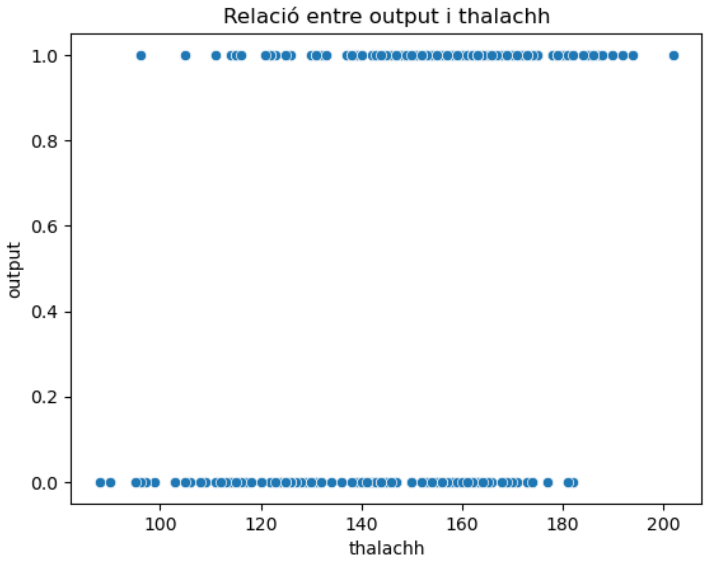
També es pot revisar la correlació de les variables. En surt la següent matriu:



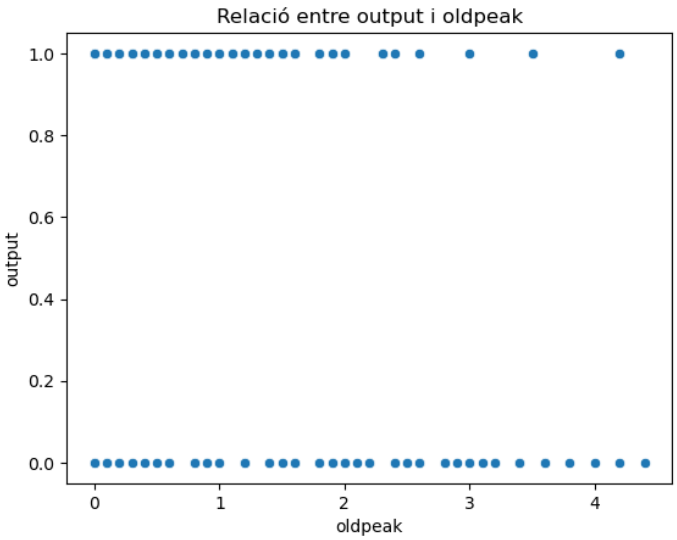
On es destacarien les correlacions de thalachh amb output (0.43) i de cp amb output (0.42).



La variable cp està força correlacionada amb la variable objectiu. Analitzant el diagrama de violí, la variable output tendeix a indicar problemes cardíacs a mesura que el tipus de dolor de pit (cp) es major. Ja que, els quatre tipus estan ordenats de menor a major gravetat.

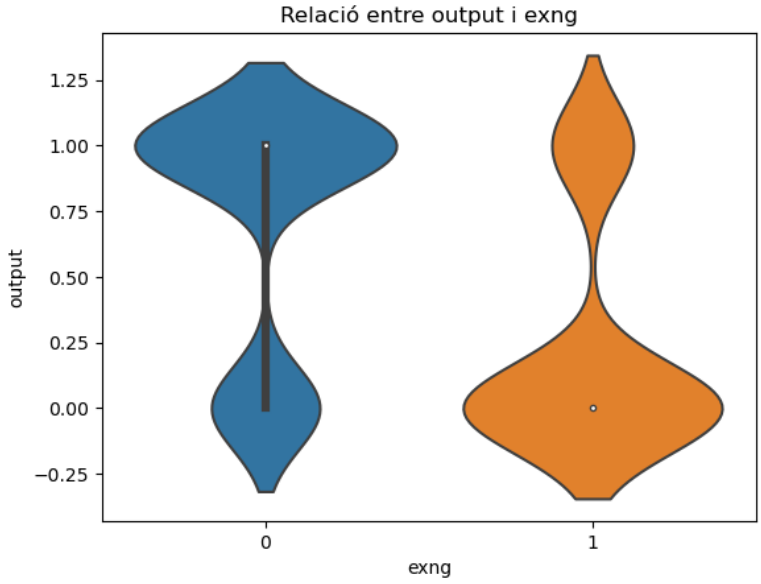


La variable thalachh té un nivell de correlació també elevat. Es pot veure que la variable objectiu indica problemes cardíacs a mesura que el ritme cardíac (thalachh) es major.

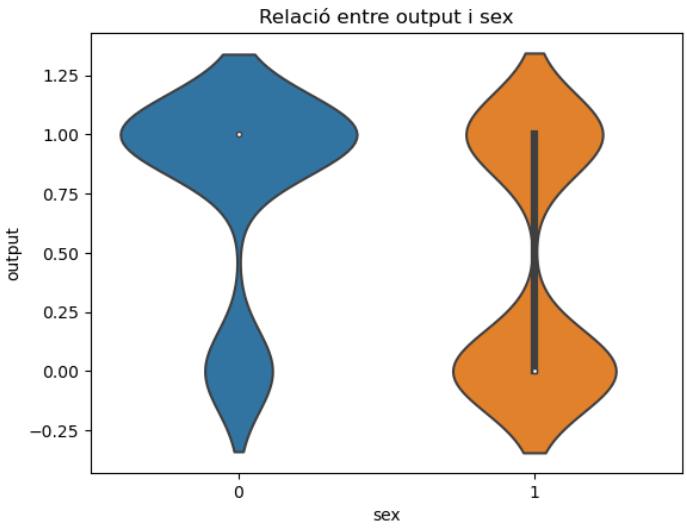


La variable oldpeak té un nivell de correlació també elevat. Es pot veure que la variable objectiu indica problemes cardíacs a mesura que el oldpeak es menor.

La variable exng té un nivell de correlació elevat. Es pot veure que la variable objectiu indica problemes cardíacs a mesura que exng és 1, és a dir, quan l'angina de pit no és induïda per exercici.



La variable sex no destaca per la seva correlació (0.23). No obstant això, es interessant com afecta a cadascun dels sexes. Els homes (0) pateixen molts més problemes cardíacs que les dones (1) tal com indica la figura.



Predicció

Un dels problemes que es volia resoldre era poder predir, segons els valors de les variables del joc de dades, si hi hauria major o menor possibilitat de tenir un atac de cor. S'han fet prediccions de la variable objectiu emprant diferents mètodes: regressió lineal i regressió logística.

Un cop fet la separació de train-test s'obtenen els següents resultats:

Regressió lineal

Amb una accuracy del 84% i una matriu de confusió amb uns resultats acceptables.

	0	1
0	17	8
1	1	33

Regressió logística

S'obtenen els mateixos resultats que amb el mètode anterior. No obstant això, per predir classes és més adequat la regressió logística.

	0	1
0	17	8
1	1	33

5. Representació dels resultats

S'han anat adjuntant les representacions i taules al llarg de la resta de la memòria.

6. Resolució del problema

Es treballa sobre dades mèdiques i es vol saber informació que faci possible millorar els tractaments dels pacients. Amb aquestes dades i tècniques s'ha aconseguit:

- S'ha aconseguit desenvolupar un model que permet saber amb una precisió del 85% i amb els paràmetres emprats si un pacient tindrà o no un atac de cor
- S'han descobert quins factors tenen una correlació més forta al fet de tenir un atac de cor, com son: exng, cp i d'altres.

7. Codi

El codi es pot trobar al repositori de github: https://github.com/sercres/pra2_tipologia

8. Vídeo

El vídeo es pot trobar:
https://drive.google.com/file/d/1u1I2zaEoNzyp8Aox5PGaeCu1hZUy-ZJZ/view?usp=sharing

Contribucions	Signatura
Investigació prèvia	P. B., S. C.
Redacció de les respostes	P. B., S. C.
Desenvolupament del codi	P. B., S. C.
Participació al vídeo	P. B., S. C.