Activitat grupal.

Pràctica 2: Tipologia i cicle de vida de les dades

1. **Descripció del data set.** Perquè és important i quina pregunta/problema pretén respondre?

*El dataset ofereix informació de 303 pacients a través de 14 variables.*

*L’objectiu es veure si es pot construir un model predictiu per tal d’explicar les causes d’un atac de cor.*

*Hem seleccionat la base de dades de Kaggle proposades. Heart attack analysis prediction dataset* [*link*](https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?resource=download)

*L’anàlisi l’hem realitzat amb R.*

|  |  |  |
| --- | --- | --- |
| *Variable* | *Descripció* | *tipus* |
| *Age* | *Edat del pacient* | *contínua* |
| *Sex* | *Sexe del pacient* | *Dicotòmica*  *(0= Home; 1= Dona)* |
| *cp (chest pain type)* | *dolor al pit* | *1. Typical angina*  *2. Atypical angina*  *3. Non-angina pain*  *4. Asymptomatic.* |
| *Tstbps.- resting blood pressure (mm Hg)* | *Pressió sanguínia en repós* | *Contínua* |
| *Chol colesterol mg/dl via BMI sensor* | *Colesterol via sensor IBM* | *Contínua* |
| *fbs : (fasting blood sugar > 120 mg/dl)* | *Sucre en sang* | *Dicotòmica*  *(1 = true; 0 = false)* |
| *rest\_ecg : resting electrocardiographic results* | *Resultats electrocardiogràfics en repos* | *Value 0: normal*  *Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)*  *Value 2: showing probable or definite left ventricular hypertrophy by Estes’ criteria* |
| *thalach : maximum heart rate achieved* |  | *Continua* |
| *exang: exercise induced angina* |  | *Dicotòmica*  *(1 = yes; 0 = no)* |
| *oldpeak: previous peak* |  |  |
| *slp: slope* |  |  |
| *caa: number of major vessels (0-3)* |  | *0-3* |
| *thall: thal rate* |  |  |
| *Output* |  | *dicotòmica*  *(0= less chance of heart attack*  *1= more chance of heart attack)* |

1. **Integració i selecció de les dades d’interès a analitzar.**

*Pot ser el resultat d’addicionar diferents datasets o una subselecció útil de les dades originals, en base a l’objectiu que es vulgui aconseguir.*

1. **Neteja de les dades.**
   1. Les dades contenen zeros o elements buits? Gestiona cadascun d’aquests casos.

*En general no hem trobat valors NA.*

*La variable oldpeak té molts valors zero, tenim el dubte de si són valors perduts, però segurament és que en el període anterior no tenien dades i per això decidim no fer res al respecte*

* 1. Identifica i gestiona els valors extrems.

*Busquem valors extrems (outliers) en les columnes trtbps, thalachh, chol i oldpeak dibuixant uns gràfics boxplot*

|  |  |
| --- | --- |
|  | *La suma de files amb outliers a la columna trtbps és de 13* |
|  | *La suma de les files amb outliers a la columna thalachh és de 3* |
|  | *La suma de les files amb outliers a la columna chol és de 8* |
|  | *La suma de les files amb outliers a la columna oldpeak és de 8* |

*En trobar outlier es pot procedir de diferents maneres, en funció de la causa de que hi hagi un valor que es distancia de la resta, es pot decidir d’imputar un valor constant o si s’escau es podria també d’assignar el valor de la mitjana per el grup al que pertanyi l’observació, per exemple es podria imputar el valor de la mitjana d’homes o dones.*

*Hi ha el métode kNN (k-Nearest Neighbours) però es pot veure afectat per la tria de la k.*

*També es pot suprimir el registre perquè els resultats no es vegin alterats per aquest valors erronis.*

*Hem decidit d’eliminar els registres amb valors perduts. Per fer-ho hem fet servir la comanda subset per descartar els casos extrems.*

*El dataset queda amb el nombre de files de 279*

1. **Anàlisi de les dades.**
   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?)

*Hem vist que al dataset hi ha variables contínues i variables categòriques: dicotòmiques i de 3 o 4 categories. I en aquest sentit hem orientat l’anàlisi.*

*A partir de la variable output que predir els que tenen risc d’atac de cor. En aquest sentit hem construït un model de regressió lineal.*

*Per gènere comparativa risc de patir un atac de cor homes o dones. Observant si hi ha independència entre les diferents variables en funció del gènere. Aquesta comparativa l’hem portat a terme amb tests de chi quadrat (o el test exacte de Fisher en cas de no ser apropiat aplicar el test de Chi Quadrat per tenir valors inferiors a 5 en alguna de les caselles de la taula de contingència del creuament de les variables 2 a 2. S’ha portat a terme amb les diferents variables categòriques o dicotòmiques en funció del gènere.*

* 1. Comprovació de la normalitat i homogeneïtat de la variància.

*S’ha comprovat la normalitat de les variables amb el test de Normalitat de Shapiro-Wilk i Tenint que la H0 és que la variable segueix una distribució normal, si el p-valor és més petit que el nivell de significació α rebutgem la H0 i per tant no té una distribució normal, en el cas de tenir un p-valor més gran (com en la majoria de les variables que hem realitzat el test) no podem rebutjar la hipòtesi nul·la i per tant hem de concloure que la variable Chol segueix una distribució normal.*

*Però donat el Teorema central del límit en el cas de tenir una n prou gran es pot considerar que segueix una distribució normal i es poden aplicar tècniques d’estadística paramètrica però els resultats seran poc robustos.*

*Per veure l’homogeneïtat de la variància es poden fer servir els tests: test de Levene, que s’aplica quan les dades segueixen una distribució normal, així com el test de Fligner-Killeen, que es tracta de l’alternativa no paramètrica, utilitzada quan les dades no compleixen amb la condició de normalitat.*

*En el nostre cas hem procedit a fer un test d’hipòtesis de comparació de mitjanes entre homes i dones F-test de raó de variances de la variable que té distribució Normal chol en funció del gènere i no podem rebutjar la hipòtesi nul·la de que el quocient de les variances dels dos grups és diferent a zero, per tant podem dir que la variança és homogènia.*

* 1. 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.

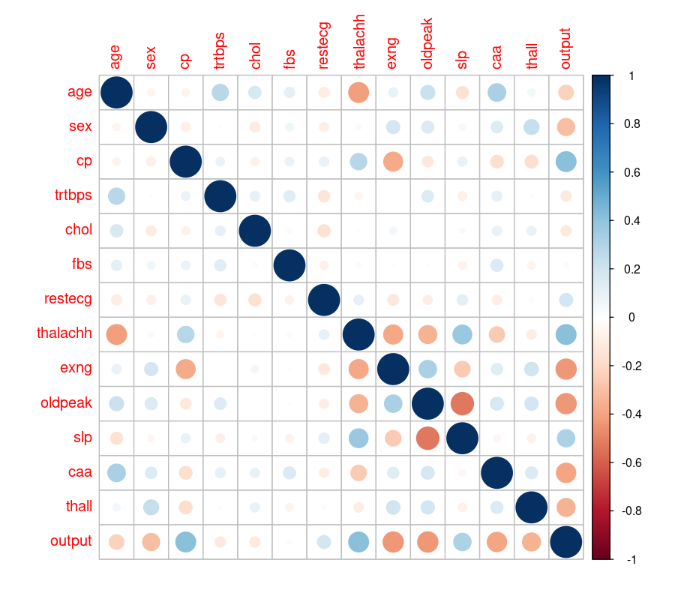
1. *Contrastos d’hipòtesis.*

|  |  |
| --- | --- |
| *Tests de Normalitat: Només en el cas de la variable chol no es rebutja la Ho* | | |age |trtbps |chol |thalachh |oldpeak |  |:------|:------|:--------|:------|:---------|:--------|  |pvalue |0.0203 |0.003591 |0.1485 |7.648e-05 |1.42e-15 | |
| *Homocedasticitat* |  |

1. *Correlació*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *age* | *sex* | *cp* | *trtbps* | *chol* | *fbs* | *restecg* | *thalachh* | *exng* | *oldpeak* | *slp* | *caa* | *thall* | *output* |
| *age* | *1.00* | *-0.06* | *-0.06* | *0.28* | *0.16* | *0.11* | *-0.11* | *-0.42* | *0.09* | *0.21* | *-0.15* | *0.33* | *0.06* | *-0.23* |
| *sex* | *-0.06* | *1.00* | *-0.09* | *0.01* | *-0.11* | *0.06* | *-0.09* | *-0.03* | *0.18* | *0.16* | *-0.05* | *0.14* | *0.24* | *-0.31* |
| *cp* | *-0.06* | *-0.09* | *1.00* | *0.08* | *-0.07* | *0.08* | *0.10* | *0.28* | *-0.38* | *-0.12* | *0.09* | *-0.17* | *-0.17* | *0.41* |
| *trtbps* | *0.28* | *0.01* | *0.08* | *1.00* | *0.10* | *0.13* | *-0.14* | *-0.06* | *0.00* | *0.15* | *-0.08* | *0.11* | *-0.02* | *-0.12* |
| *chol* | *0.16* | *-0.11* | *-0.07* | *0.10* | *1.00* | *0.03* | *-0.16* | *-0.01* | *0.06* | *-0.01* | *0.03* | *0.09* | *0.09* | *-0.11* |
| *fbs* | *0.11* | *0.06* | *0.08* | *0.13* | *0.03* | *1.00* | *-0.08* | *-0.03* | *0.01* | *0.02* | *-0.07* | *0.16* | *-0.06* | *-0.03* |
| *restecg* | *-0.11* | *-0.09* | *0.10* | *-0.14* | *-0.16* | *-0.08* | *1.00* | *0.10* | *-0.12* | *-0.09* | *0.12* | *-0.09* | *0.03* | *0.18* |
| *thalachh* | *-0.42* | *-0.03* | *0.28* | *-0.06* | *-0.01* | *-0.03* | *0.10* | *1.00* | *-0.38* | *-0.34* | *0.37* | *-0.25* | *-0.10* | *0.42* |
| *exng* | *0.09* | *0.18* | *-0.38* | *0.00* | *0.06* | *0.01* | *-0.12* | *-0.38* | *1.00* | *0.32* | *-0.26* | *0.13* | *0.20* | *-0.43* |
| *oldpeak* | *0.21* | *0.16* | *-0.12* | *0.15* | *-0.01* | *0.02* | *-0.09* | *-0.34* | *0.32* | *1.00* | *-0.53* | *0.18* | *0.19* | *-0.43* |
| *slp* | *-0.15* | *-0.05* | *0.09* | *-0.08* | *0.03* | *-0.07* | *0.12* | *0.37* | *-0.26* | *-0.53* | *1.00* | *-0.05* | *-0.08* | *0.32* |
| *caa* | *0.33* | *0.14* | *-0.17* | *0.11* | *0.09* | *0.16* | *-0.09* | *-0.25* | *0.13* | *0.18* | *-0.05* | *1.00* | *0.15* | *-0.39* |
| *thall* | *0.06* | *0.24* | *-0.17* | *-0.02* | *0.09* | *-0.06* | *0.03* | *-0.10* | *0.20* | *0.19* | *-0.08* | *0.15* | *1.00* | *-0.34* |
| *output* | *-0.23* | *-0.31* | *0.41* | *-0.12* | *-0.11* | *-0.03* | *0.18* | *0.42* | *-0.43* | *-0.43* | *0.32* | *-0.39* | *-0.34* | *1.00* |

*Podem observar que la major correlació negativa amb output és oldpeak (-0.43 )i la major correlació positiva cp i thalachh (0.28) que estan allunyats de -1 o de 1*



1. *Regressió.*

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | *R2 Ajustada* |
| lm(formula = output ~ age + sex + cp + trtbps + chol + fbs +  restecg + thalachh + exng + oldpeak + slp + caa + thall,  data = train) |  |  |  |
|  |  |  |  |

1. *Tests Chi Quadrat de variables cualitatives.*
2. **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.
3. **Resolució del problema.** A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?
4. **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
5. **Vídeo.** 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.

|  |  |
| --- | --- |
| **Contribucions** | **Signatura** |
| Investigació prèvia | SGM, LTA |
| Redacció de respostes | SGM, LTA |
| Desenvolupament del codi | SGM, LTA |
| Participació al vídeo | SGM, LTA |

# Bibliografía