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*

* *Age.- edat del pacient*
* *Sex.- sexe del pacient*
* *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)*
* *Chol colesterol mg/dl via BMI sensor*
* *fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)*
* *rest\_ecg : resting electrocardiographic results 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*
* *exang: exercise induced angina (1 = yes; 0 = no)*
* *oldpeak: previous peak*
* *slp: slope*
* *caa: number of major vessels (0-3)*
* *thall: thal rate*
* *output : 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.

*Sembla una base de dades depurada.*

* 1. Identifica i gestiona els valors extrems.

*D’entrada no n’he vist cap.*

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?)

*Hi ha diferents opcions:*

*Per gènere*

*A partir de la variable output que separa els que tenen risc d’atac de cor dels que no en tenen.*

* 1. Comprovació de la normalitat i homogeneïtat de la variància.
  2. 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. **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.
2. **Resolució del problema.** A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?
3. **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
4. **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