1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

Hem triat el dataset **"Heart Attack Analysis & Prediction Dataset"** de kaggle. És un conjunt de dades que inclou informació dels pacients que han patit atacs cardíacs. Hi trobem atributs demogràfics com l'edat, el genere i el pes del pacient, atributs clínics com la pressió arterial en repos i els nivells de colesterol. També hi trobem proves de laboratori com els nivells de glucèmia i els resultats de un electrocardiograma.

- Age : Edat del pacient
- Sex Genere del pacient
- exng: Engina de pit provocat per l'exercici
- ca: Número de vassos majors.
- cp : Tipus de dolor al pit
- trtbps : Pressió arterial en repos
- chol : Nivell de colesterol
- fbs : Sucre en sang alt
- restecg : Resultats del electrocardiograma
- thalachh : Freqüencia cardiaca máxima
- oldpeak: Depresión del segmento ST del electrocardiograma
- slp: Pendent del segment ST de l'exercici maxim
- thal: Trastorn de la sang anomenat talassèmia
- target Possibilitat d'un atac cardiac

Les preguntes que volem respondre analitzant aquest data set són: Quins factors de risc augmenten la probabilitat de patir un atac cardíac? És possible predir el risc de patir un atac cardíac?

És important perquè els atacs cardíacs són una de les principals causes de mort a nivell mundial. Per tant, conèixer els factors de risc i la probabilitat de patir-e un pot ajudar a prevenir els atacs i salvar vides.

2. 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.

Pel que fa la selecció de les dades d'interès agafarem tots els atributs del dataset per el moment, ja que tots estan relacionats amb els atacs cardíacs. A kaggle, en el dataset trobem un altre dataset format per el atribut de saturació d'oxigen en sang. Semblava que eren dels mateixos pacients que l'altre dataset i podíem fusionar els dos datasets. He mirat les dimensions del dataset i un te 3585 registres i l'altre 303, per tant no sabem la correspondència i no podem fusionar aquests 2 datasets. Per tant, com que no tenim més dades dels mateixos pacients que el dataset original no podem fer cap integració.

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.

Hem observat les dades si tenien valors nulls, blancs, unknown, none... però hem vist que les dades estan 100% completes i que no falta cap dada per completar en el dataset. Hem vist que de zeros n'hi ha molts però no cal gestionar-los perquè no representen elements per definir, sinó atributs binaris o categòrics els quals l'etiqueta zero forma part del dataset i és una classe que pertany l'atribut. Un exemple seria el gènere el qual és 0 o 1 depenent de si és noi o noia. Per tant no cal gestionar aquest tipus de casos. En el cas d'aplicar-ho podríem haver aplicat tècniques com per exemple agafar la mitjana del atribut si és lineal o agafat l'etiqueta més repetida si és categòric, entre d'altres.

3.2. Identifica i gestiona els valors extrems.

Mirarem si hi ha outliers, que són dades que es troben molt allunyades de la distribució normal d'una variable. Veient el dataset i els histogrames de cada atribut veiem a simple vista que no hi ha cap valor extrem molt evident. Aquest dataset conté moltes dades binaries i categòriques en les quals no es troben outliers. Entrant més en detall observem que és difícil saber que és considera outlier però que generalment es considera quan el valor es troba allunyat 3 desviacions estàndard respecte a la mitjana del conjunt. Saben aquesta definició, hi ha l'atribut colesterol, que te una mitjana de 246 i una desviació estàndard de 52, que fent el càlcul 246 + 52 * 3 = 402, és a dir, valors superiors a 402 es considerarien valors extrems. Mirant el histograma de l'atribut veiem que hi ha una dada del colesterol que està entre 520 i 564. Per tant, la podríem substituir per la mitjana 246, però en aquest cas observem que només és un valor. Veient en nivell de completesa de les dades, la idea seria crear primer un model sense gestionar aquest valor extrem i després un altre gestionant el valor ficant un 246. En aquest cas, com que no creem cap model gestionarem el valor i el substituirem per 246.

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?).

Una idea seria analitzar/comparar per diferents grups de dades, com per exemple agrupar per atributs demogràfics , atributs clínics i proves de laboratori. Desde el meu punt de vista penso que no és necessari i millor analitzar/comparar tots els atributs amb tots per no perdre cap coneixement de les dades ni saltar cap correlació important entre atributs.

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

Resolt en el codi

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.

Resolt en el codi

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.

Resolt en el codi

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

Les conclusions són positives, ja que les preguntes que ens havíem fet a l'inici del projecte hem trobat una resposta. Durant el codi hem analitzat les dades i hem vist atributs que són molt importants per saber si una persona patirà un atac cardíac. També hem creat un classificador utilitzant l'algorisme de l'arbre de decisió. Resoldre aquestes dos preguntes són importants ja que ens ajudarà a prevenir futurs atacs de cor, gracies a estudis d'aquest tipus podem reduir el nombre d'atacs de cor, que és la finalitat principal del projecte.