TD2 Groupe E – Printemps 2024

ESCLEINE Justine - GAJAN Antoine - HAN Yushi

21/04/2024

Résumé

Ce rapport rend compte de l'étude exploratoire des données réalisée dans le cadre de l'UV SY09 sur le jeu de données Heart Attack [2]. Ce document a pour objectif de présenter les premières avancées sur le projet et de définir les prochains objectifs à atteindre pour la fin du semestre.

Introduction

Dans le cadre de l'UV de Sciences des Données (SY09) de l'Université de Technologie de Compiègne, les étudiants sont amenés à travailler en groupe sur un projet d'analyse d'un jeu de données pour mettre en pratique les connaissances acquises tout au long du semestre.

Ainsi, nous avons choisi de travailler sur le jeu de données "Heart Attack". Ce choix nous a semblé comme une évidence face à l'augmentation du nombre de patients de plus en plus jeunes touchés par des arrêts cardiaques [1].

Dans ce rapport intermédiaire à mi-parcours, nous présenterons notre jeu de données et détaillerons les premières études menées sur le jeu de données. Enfin, nous aborderons les étapes à venir et les prochains jalons que nous envisageons.

1 Présentation du dataset Heart Attack

1.1 Les données

Le jeu de données Heart Attack présente les données médicales issues d'une analyse menée sur un échantillon de 303 individus, chacun étant représenté par un ensemble de 13 attributs.

Les différents attributs sont les suivants :

- 1. Age : âge du patient (entier positif)
- 2. Sex : genre du patient (1 = M, 0 = F)
- 3. ChestPain : type de douleur thoracique (1 = angine typique, 2 = angine atypique, 3 = douleur non angineuse, 4 = asymptomatique)
- 4. Restbbs : pression artérielle au repos en mmHg (entier positif)
- 5. Chol : cholestérol en mg/dl (entier positif)
- 6. Fbs : taux de sucre dans le sang $>$ 120 mg/dl (1 = vrai, 0 = faux)
- 7. RestECG : résultats de l'électrocardiogramme au repos (0 = normal, 1 = anomalie de l'onde ST-T, 2 = hypertrophie ventriculaire gauche)
- 8. MaxHR : fréquence cardiaque maximale atteinte (entier positif)
- 9. Exang : angine induite par l'exercice (1 = oui, 0 = non)
- 10. Oldpeak : dépression de l'onde ST induite par l'exercice par rapport au repos (réel positif)
- 11. Slope : pente du segment ST à l'exercice maximal (1 = ascendant, 2 = plat, 3 = descendant)
- 12. Ca : nombre de vaisseaux principaux colorés par la fluoroscopie (0-3)
- 13. Thal : thallium scintigraphie (3 = normal, 6 = défaut fixe, 7 = défaut réversible)
- 14. Target : maladie cardiaque (1 = malade, 0 = sain)

Nous avons remarqué que le jeu de données était déjà propre et ne nécessiterait pas de travail préalable à sa manipulation.

1.2 Types des variables

Les 13 variables peuvent ainsi être classées en plusieurs catégories :

- Quantitative : Age, Restbbs, Chol, MaxHR, Oldpeak
- Qualitative : ChestPain, RestECG, Slope, Ca, Thal, Target

Comprendre le type de chaque variable est essentiel pour une analyse de données approfondie. Cette connaissance orientera nos décisions d'analyse, car les approches diffèrent selon que les variables sont quantitatives ou qualitatives.

2 Analyse exploratoire des données

Dans cette partie, nous détaillerons les premières techniques mises en place pour analyser notre jeu de données. Nous décrirons la méthodologie utilisée et exposerons les conclusions qui en découlent.

2.1 Découverte des données

2.1.1 Etude des caractéristiques démographiques des individus

Dans le souhait de mieux appréhender notre jeu de données, il est indispensable de commencer par effectuer une étude des caractéristiques démographiques pour comprendre quels types d'individus ont été étudiés et en quelle proportion.

Nous avons ainsi commencé par analyser la proportion de malades sur la population étudiée. Sur les 303 individus, plus de la moitié (165 sur 303) ont été touchées par une crise cardiaque.

Afin de mieux comprendre les caractéristiques des personnes ayant eu un arrêt cardiaque, nous avons dans un premier temps étudié la proportion de malades en fonction de l'âge et du sexe.

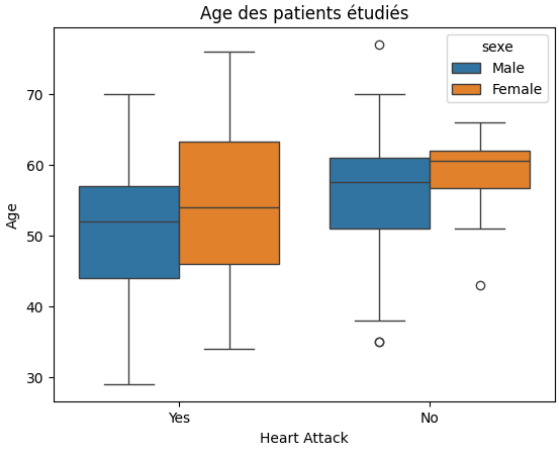

FIGURE 1 – Age des patients en fonction de la pathologie

Agés de 29 à 70 ans avec une moyenne de 55 ans, on remarque que les patients touchés par une crise cardiaque sont plus jeunes, avec un écart-type plus important.

Il serait à présent de connaître le genre des personnes étudiées et touchées par un arrêt cardiaque. Nous avons ainsi conçu le graphique suivant :

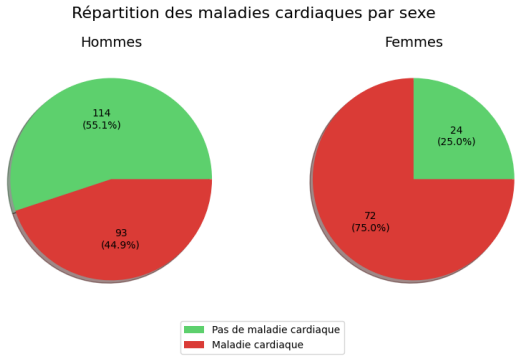

FIGURE 2 – Genre des patients en fonction de la pathologie

On remarque que la population étudiée est principalement masculine, représentant 207 individus, soit 68% de l'échantillon total. Sur l'échantillon observé, les femmes ont tendance à être plus souvent touchées par des arrêts cardiaques que les hommes.

2.1.2 Visualisation univariée

Nous avons réalisé une visualisation univariée pour chacune des variables pour nous permettre de mieux comprendre les caractéristiques observées chez les individus.

Ainsi, pour les variables qualitatives, nous avons conçu des diagrammes en barres et obtenu les résultats suivants :

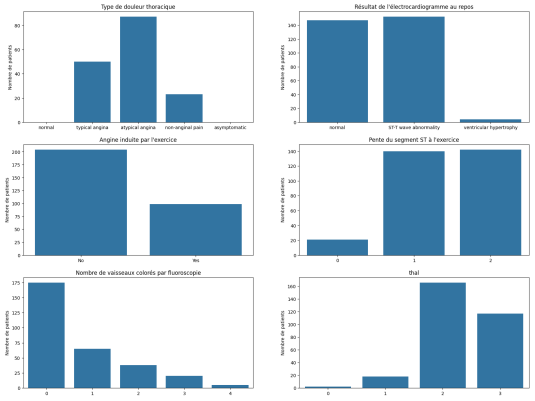

FIGURE 3 – Représentation univariée des variables qualitatives

Nous avons fait de même avec les variables quantitatives et proposé des histogrammes. La réalisation de ces derniers nous permettra de détecter les tendances générales observées dans notre jeu de données.

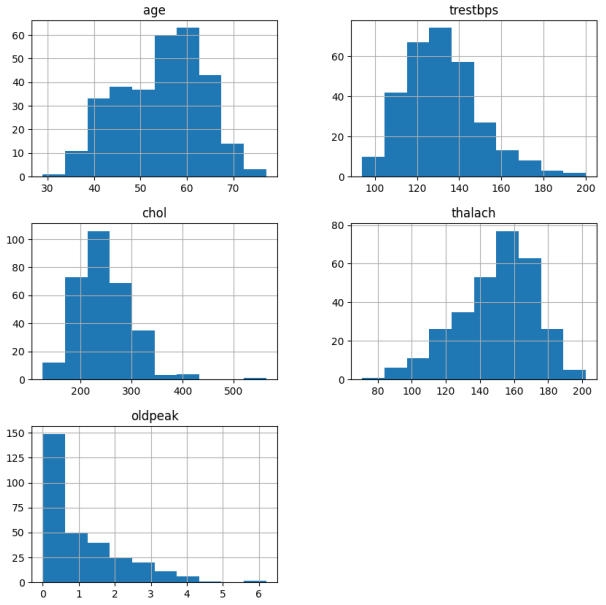

FIGURE 4 – Représentation univariée des variables quantitatives

On remarque un effet d'échelle entre les différentes variables. Leur ordre de grandeur diffère. Par exemple, la variable thalach, qui modélise fréquence cardiaque maximale, a des valeurs de l'ordre de 100, tandis que la variable oldpeak, modélisation la dépression ST, contient des valeurs entre 0 et 10. Cet effet devra probablement être pris en considération par la suite.

2.1.3 Visualisation multivariée

La visualisation univariée nous a permis de comprendre les valeurs observées pour chacune des variables. Il serait à présent pertinent de mieux comprendre les éventuels liens entre certaines variables quantitatives. Ainsi, nous avons choisi de concevoir des nuages de points pour chaque paire de variables. L'obtention de cette visualisation en 2 dimensions nous donnera un premier aperçu des corrélations entre les variables.

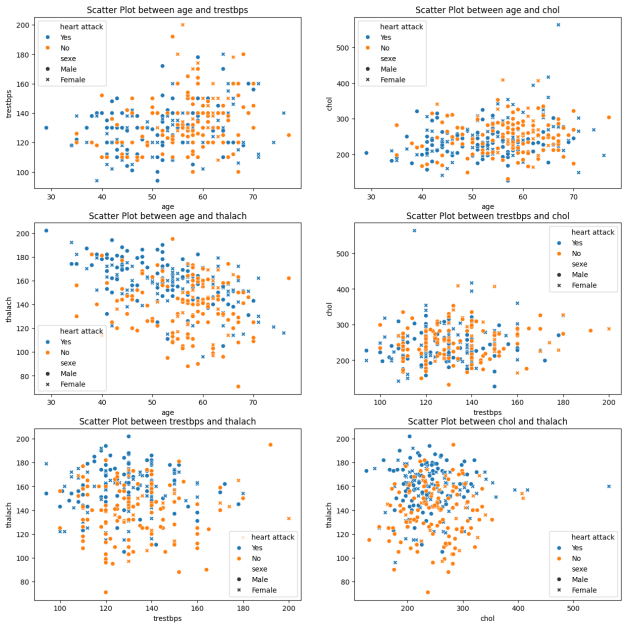

FIGURE 5 – Représentation multivariée des variables quantitatives

Ces nuages de points ne nous ont pas permis de détecter de liens apparents entre deux variables quantitatives. Cela n'exclut toutefois pas qu'un lien entre un ensemble de variables existe. C'est pourquoi il serait pertinent de s'intéresser aux corrélations entre les variables.

2.1.4 Etude des corrélations entre les variables

A présent, il peut être intéressant de comprendre le lien entre les variables. Pour cela, nous allons étudier la corrélation entre les variables quantitatives. On a ainsi obtenu les résultats ci-dessous :

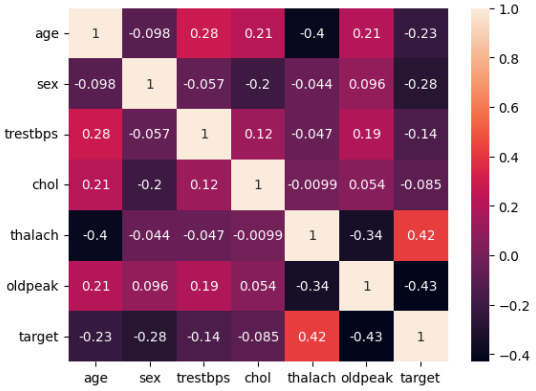

FIGURE 6 – Matrice de corrélation

Sur cette matrice de corrélation, on ne remarque pas de lien apparent entre 2 variables. En effet, les corrélations sont relativement faibles en valeur absolue (inférieures à 10%). Seules certaines variables présentent une corrélation plus importante (de l'ordre de 40%). Toutefois, cette faible corrélation ne permet malheureusement pas de déduire des liens apparents entre les variables. Cela s'explique probablement par le fait que le jeu de données a été préalablement nettoyé et que seules les caractéristiques non corrélées aient été conservées pour mener l'étude.

2.2 Analyse en composantes principales

Bien que peu de lien linéaire aient été observées entre les variables, il nous a semblé pertinent d'effectuer une ACP pour disposer d'une visualisation en 2 dimensions des individus en conservant le maximum d'inertie possible.

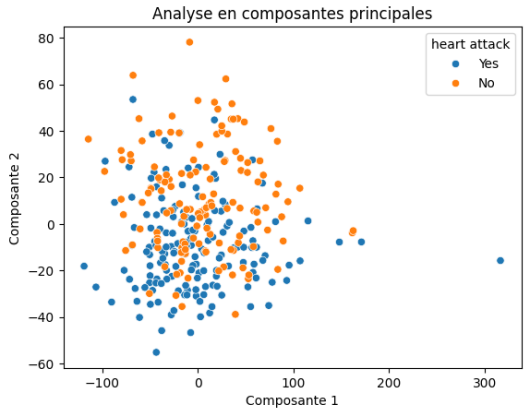

FIGURE 7 – Analyse en composantes principales

Suite à cette ACP en 2 dimensions, on remarque qu'il est difficile de construire une séparation en 2 groupes des individus, basé sur le fait qu'ils aient été atteints d'une crise cardiaque ou non.

Conclusion

Pour conclure, cette première phase exploratoire nous a permis de découvrir notre jeu de données et ces spécificités. Par la suite, nous pourrons appliquer les connaissances théoriques vues lors de la seconde moitié du semestre à notre cas pratique telles que la classification supervisée.

Références

[1] N. Geographic. Santé : médecine, cardiologie : de plus en plus de jeunes adultes font des crises cardiaques, 2024. Article sur le site Web de National Geographic.

[2] P. Sheta. Heart attack analysis & prediction dataset, 2022. Kaggle Dataset.