TD2 Groupe E – Printemps 2024

ESCLEINE Justine - GAJAN Antoine - HAN Yushi

16/06/2024

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.

Face à l'augmentation des arrêts cardiaques chez les jeunes [1], nous avons choisi de travailler sur le jeu de données "Heart Attack" [2]. Dans ce rapport, nous présenterons et détaillerons les études menées sur celui-ci.

1 Présentation du dataset

1.1 Les individus

Le jeu de données Heart Attack contient des informations médicales provenant d'une étude menée sur un échantillon de 303 individus. Chaque individu est décrit par 14 variables descriptives, dont une variable cible qui indique la présence ou l'absence de maladie cardiaque.

1.2 Les variables

Les variables explicatives incluent des facteurs démographiques comme l'âge ou le sexe du patient et des facteurs cliniques comme la fréquence cardiaque.

Variable	Description
Age	Âge du patient
Sex	Genre du patient
ChestPain	Type de douleur thoracique
Restbps	Pression artérielle au repos
Chol	Cholestérol en mg/dl
Fbs	Taux de sucre dans le sang > 120 mg/dl
RestECG	Résultats de l'électrocardiogramme au repos
Thalach	Fréquence cardiaque maximale atteinte
Exang	Angine induite par l'exercice
Oldpeak	Dépression de l'onde ST
Slope	Pente du segment ST à l'exercice maximal
Ca	Nombre de vaisseaux principaux colorés par la fluoroscopie (0-3)
Thal	Thallium scintigraphie
Target	Maladie cardiaque (1 = malade, 0 = sain)

TABLE 1 – Description des variables

Les 14 variables peuvent ainsi être classées en 2 catégories :

- Quantitative : Age, Restbps, Chol, Oldpeak, Thalach
- Qualitative : Sex, ChestPain, Fbs, RestECG, Exang, Slope, Ca, Thal, Target

Le jeu de données a été préalablement nettoyé et pourra être manipulé sans autre transformation.

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 Caractéristiques démographiques

Sur les 303 individus, plus de la moitié (165) ont été touchés par une crise cardiaque. Pour comprendre les caractéristiques des personnes ayant été touchés, nous avons étudié la proportion de malades en fonction de l'âge et du sexe.

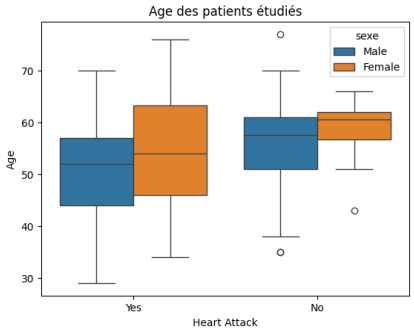

FIGURE 1 – Age selon l'incidence de la pathologie

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.

Nous avons procédé de manière analogue pour étudier l'influence du genre. Les résultats suivants ont été observés :

Genre	Pas de maladie	Maladie cardiaque
Femme	24	72
Homme	114	93

TABLE 2 – Incidence de la maladie cardiaque en fonction du genre

La population étudiée est principalement masculine (207 individus, soit 68% de l'échantillon). Sur l'échantillon observé, la proportion de femmes touchée est plus élevée que pour les hommes.

2.1.2 Visualisation univariée

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

Ainsi, pour représenter les variables qualitatives, nous avons conçu des diagrammes en barres. Ils nous permettent de nous rendre compte des valeurs observées au sein de la population pour les différentes variables.

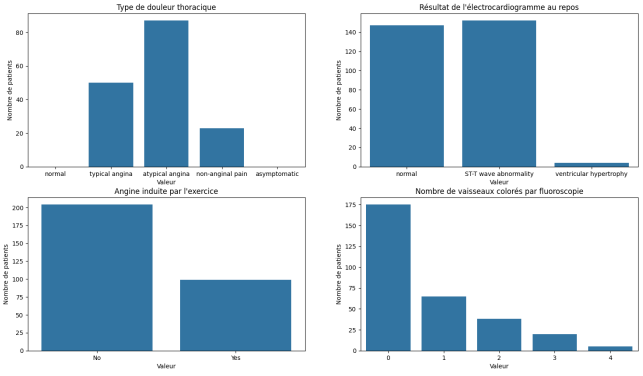

FIGURE 2 – Représentation univariée des variables qualitatives

Nous avons fait de même avec les variables quantitatives en proposant des histogrammes. Ils nous permettront de détecter les tendances générales de distribution des données.

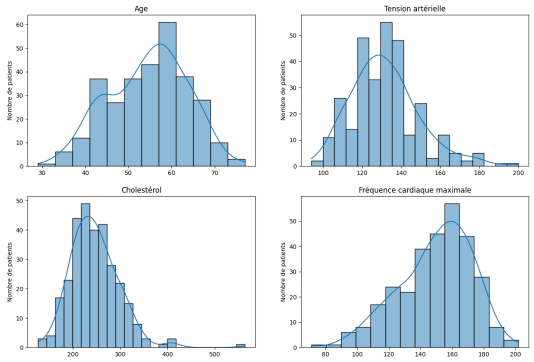

FIGURE 3 – Représentation univariée des variables quantitatives

On remarque un effet d'échelle entre les différentes variables. Leur ordre de grandeur diffère. Par exemple, la variable associée au taux de cholestérol a des valeurs de l'ordre de la centaine, tandis que l'âge se compte en dizaines. Cet effet devra être pris en considération par la suite.

2.1.3 Visualisation multivariée

Pour comprendre les éventuels liens entre certaines variables quantitatives, nous avons conçu des nuages de points pour chaque paire de variables. Ces nuages de points ne nous ont pas permis de détecter de liens apparents entre deux variables quantitatives. Cela n'exclut toutefois pas l'existence d'un lien entre un ensemble de variables, d'où la nécessité de s'intéresser aux corrélations entre les variables.

2.1.4 Etude des corrélations entre les variables

Dans cette section, nous allons étudier les corrélations entre les différentes variables quantitatives, selon l'incidence de la maladie cardiaque. On a ainsi obtenu les résultats ci-dessous :

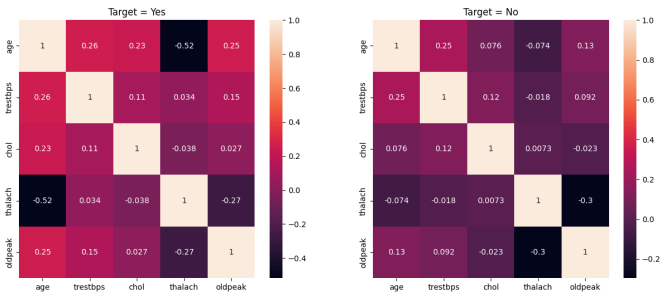

FIGURE 4 – Matrice de corrélation

Sur cette matrice, les corrélations sont relativement faibles en valeur absolue (inférieures à 30%). L'âge et la fréquence cardiaque chez les patients touchés présentent une corrélation plus importante (50%). Une analyse approfondie de cette relation pourrait aider à mieux comprendre les facteurs de risque aux crises cardiaques chez les jeunes patients, contribuant ainsi à l'amélioration des stratégies de prévention et de traitement.

2.2 Analyse en composantes principales

Bien qu'aucun lien apparent n'ait été détecté entre 2 variables, il nous a semblé pertinent d'effectuer une ACP pour disposer d'une visualisation conservant le maximum d'inertie possible.

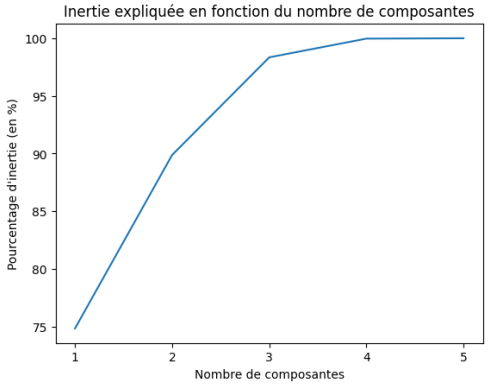

FIGURE 5 – Méthode du coude

La méthode du coude indique que le nombre optimal de composantes à conserver est de 3. Elles correspondent respectivement aux variables chol, thalach et trestbps, qui ont une variance plus élevée en raison de l'effet d'échelle. Les autres variables quantitatives contribuent à moins de 10% de l'inertie totale.

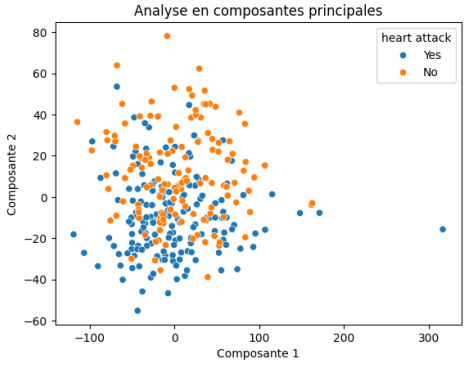

FIGURE 6 – Analyse en composantes principales

Le faible nombre de variables quantitatives et l'absence de corrélation entre elles ne permettent pas d'obtenir des résultats concluants. En effet, l'ACP ne nous permet pas d'identifier clairement une séparation en deux groupes des individus dans le premier plan factoriel, basée sur le fait qu'ils aient ou non subi une crise cardiaque.

3 Méthodes de clustering

Même s'il n'est pas possible de discerner une frontière de décision à l'œil nu, nous avons appliqué diverses méthodes d'apprentissage non supervisé pour regrouper nos données.

3.1 Classification ascendante hiérarchique

Nous avons effectué une CAH avec distance de Ward pour maximiser l'inertie.

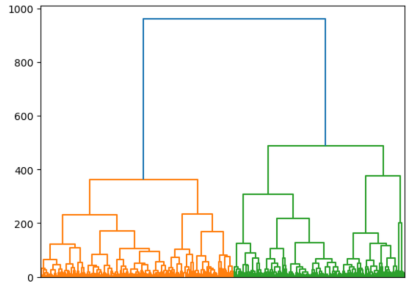

FIGURE 7 – Dendrogramme issu de la CAH

Le modèle n'a cependant pas effectué l'apprentissage escompté. En réalité, le modèle a séparé les données en raison de l'effet d'échelle selon le taux de cholestérol. Pour le groupe en orange, il suit une distribution centrée à 209 mg/dl avec un écart type de 25 mg/dl, tandis que pour le groupe vert, la distribution est centrée à 288 mg/dl avec un écart type de 42 mg/dl.

3.2 K-Means

L'algorithme des K-Means n'est pas parvenu non plus à effectuer le regroupement escompté, avec un score de rand de 0.51. Comme la CAH, on remarque que la séparation s'est faite selon la première composante, c'est-à-dire, selon la variable chol.

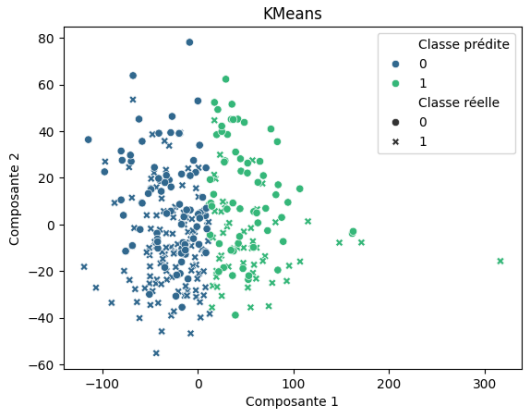

FIGURE 8 – Résultats issus du K-Means

4 Apprentissage supervisé

Nous nous intéressons à présent à l'apprentissage supervisé qui utilise des données étiquetées pour apprendre les relations entre les caractéristiques.

4.1 Régression logistique

Nous avons entrainé un modèle de régression logistique avec 70% de données d'entrainement et 30% de test. La régression logistique basée sur nos 13 variables descriptives nous a permis d'obtenir une accuracy de 87%, avec la matrice de confusion suivante sur les données de test :

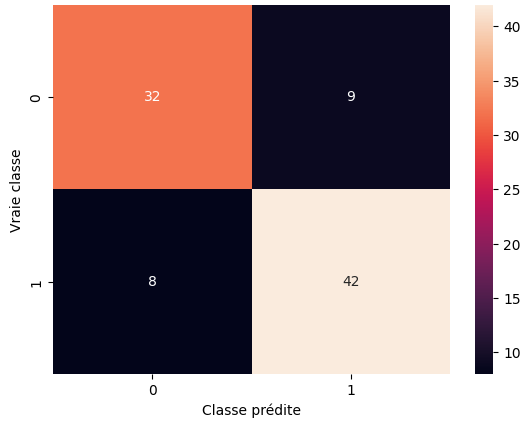

FIGURE 9 – Matrice de confusion issue de la régression logistique

Ce modèle nous a ainsi permis d'obtenir des résultats pertinents sur le jeu de test. En effet, bien que le dataset présente une répartition similaire des cas positifs et négatifs, le modèle n'a pas surévalué les faux positifs.

L'entrainement d'un modèle de régression logistique quadratique plus général n'a pas conduit à une amélioration des résultats. Ceci s'explique par le peu de données qui peut ajouter de l'overfitting et le peu de lien entre les différentes variables.

Métrique	**Reg. log.**	**Reg. log. polynomiale**
Accuracy	0.87	0.79
ROC-AUC	0.88	0.86

TABLE 3 – Comparaison des performances des modèles

4.2 K plus proches voisins (KNN)

Nous nous sommes ensuite intéressés au modèle KNN. En effet, nous voulions savoir s'il était possible de prédire l'incidence de la maladie cardiaque d'un nouveau patient en fonction de cas proches. Nous avons donc sélectionné l'hyperparamètre (nombre de voisins) à l'aide d'une validation croisée.

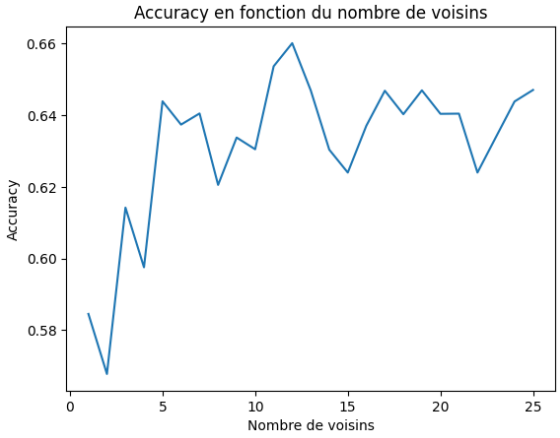

FIGURE 10 – Choix du nombre de voisins

Sur ce graphique, on remarque que le nombre de voisins permettant d'obtenir la meilleure accuracy est 12. L'apprentissage d'un KNN avec cet hyperparamètre a conduit à une accuracy de 0.66. Ces moins bons résultats sont liés à la distribution des données. En effet, KNN fonctionne mieux avec des distributions de données où les points similaires sont proches les uns des autres. Au vu des résultats de l'ACP, KNN a eu des difficultés pour capturer les tendances sous-jacentes.

4.3 Classification bayésienne

Nous avons entraîné divers modèles s'appuyant sur la théorie bayésienne : una analyse discriminante quadratique (QDA), une analyse discriminante linéaire (LDA) et un classifieur bayésien naïf (NB). Le QDA, malgré sa robustesse théorique grâce à la prise en compte des corrélations entre les variables, n'a obtenu qu'une précision de 0,79. Cette performance limitée s'explique par sa complexité, la quantité limitée de données et notre hypothèse de normalité des distributions, souvent inexacte en pratique. En revanche, le classifieur bayésien naïf, avec son hypothèse simpliste d'indépendance conditionnelle entre les caractéristiques, a obtenu une accuracy de 0,83. Sa simplicité, son efficacité, et l'indépendance relative des variables observée dans notre étude, expliquent ses meilleures performances.

Métrique	QDA	LDA	NB
Accuracy	0.79	0.80	0.83
ROC-AUC	0.86	0.90	0.89

TABLE 4 – Comparaison des performances des modèles

4.4 Arbres de décision

Nous avons opté pour l'apprentissage d'un arbre de décision afin de comprendre comment classifier et prédire de nouvelles données. Pour améliorer la lisibilité du modèle et l'interprétabilité des résultats, nous avons limité sa profondeur maximale à 3. Cet arbre, bien que très simpliste, permet d'obtenir une accuracy de 78%.

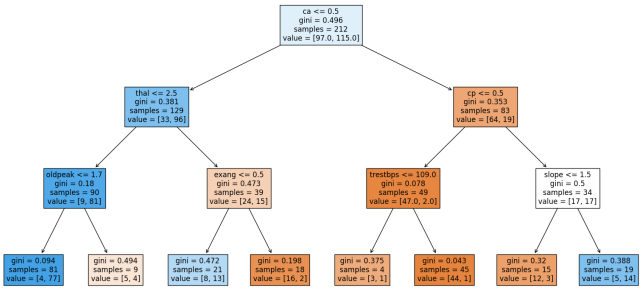

FIGURE 11 – Arbre de décision

Cet arbre facilite la prédiction de la variable cible pour un nouveau patient.

La mise en place d'une forêt aléatoire nous a permis d'améliorer les résultats obtenus, avec une accuracy de 0.82. Nous avons alors décidé de visualiser l'importance des différentes variables dans la forêt.

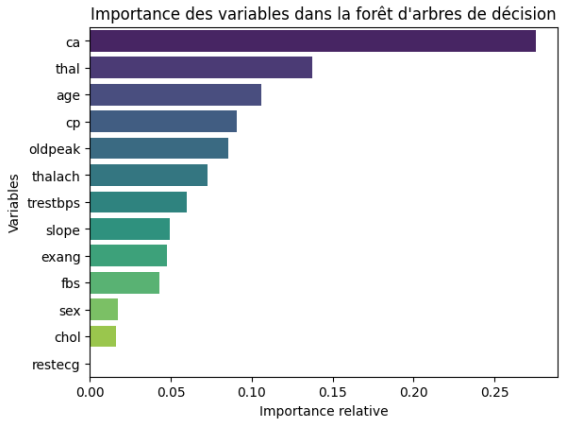

FIGURE 12 – Importance des variables dans la forêt d'arbres de décision

D'après ce graphique, le nombre de vaisseaux colorés par fluoroscopie (variable "ca") semble être le principal critère de décision. On remarque que la variable "chol" ne figure pas parmi les critères essentiels pour prédire l'incidence de la maladie. Cette observation justifie les moins bons résultats obtenus lors de l'apprentissage non supervisé, où l'accent était mis sur cette variable.

Conclusion

Ce projet a été une opportunité pour mettre en pratique les concepts théoriques que nous avons étudiés tout au long du semestre. À travers l'analyse exploratoire des données et la mise en œuvre de méthodes d'apprentissage, nous avons approfondi notre compréhension des données et développé des modèles capables de prédire la présence d'une maladie cardiaque.

Les méthodes d'apprentissage supervisé se sont avérées efficaces pour prédire notre variable cible en fonction des 13 variables descriptives, avec une accuracy de 80% pour la régression logistique. Augmenter la taille du jeu de données ou se rapprocher d'experts pour comprendre les données pourraient permettre d'améliorer les résultats.

Références

[1] N. Geographic. Santé : médecine, cardiologie : de plus en plus de jeunes adultes font des crises cardiaques, 2024.

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