POLYTECHNIQUE MONTRÉAL

IND6212 - Rapport Projet fin de session

Prédire des anomalies cardiaques chez les patients

Fait par :
Lyes Heythem Bettache 1923715
Manal Naji 1963765
Encadré par : M Bruno Agard
25 Avril 2019

1. Introduction :

La cardiologie demeure un domaine de la médecine très intéressant mais compliqué. Ce qui justifie la multitude des études effectuées dedans afin de bien comprendre et maîtriser la prévention, le diagnostic, le traitement et la réadaptation des maladies cardiovasculaires. Dans un soucis de compréhension et d'analyse des facteurs qui causent les anomalies cardiaques chez les patients, la base de données, objet de notre étude, a été créée dont le lien est le suivant: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29?fbclid=IwAR03X_eEnAh-2YjTiFqPHep86gt7_jE8AQ9o1ehU27lhSbQTW0TuSljP01I. L'objectif est de prédire si les patients possèdent des anomalies cardiaques et déterminer les facteurs influents sur leurs états en se basant sur 13 variables potentielles mesurés à partir des différents tests effectués sur l'échantillon. Cette typologie a une double visée : elle permet d'une part une description synthétique de l'échantillon des personnes enquêtées et d'autre part être utilisé comme facteur explicatif dans l'étape de recherche des déterminants de l'état de santé.

Après visualisation des données via l'outil Excel, nous observons que l'anomalie cardiaque est soit présente chez le patient ou absente.

Nous posons : Classe (1) : Absence de maladies cardiaques
Classe (2) : Présence de maladies cardiaques

Notre base de données est caractérisée comme suit :

- Nombre des attributs : 13 caractères
- Nombre d'objectifs : 269 individus

Les attributs sont présentés comme suit :

Attributs	Symbole
age	age
sex	sex
chest pain type (4 valeurs)	cpt
resting blood pressure	Restbps
serum cholestoral en mg/dl	Chol
fasting blood sugar > 120 mg/dl	Fbs
resting electrocardiographic results (valeurs 0,1,2)	Restecg
maximum heart rate achieved	Thalach
exercise induced angina	Exang

oldpeak = ST depression induced by exercise relative to rest	Oldpeak
the slope of the peak exercise ST segment	Slope
number of major vessels (0-3) colored by flourosopy	Ca
thal: 3 = normal; 6 = fixed defect; 7 = reversable defect	thal

Tableau 1: Liste des attributs et leurs symboles

2. Préparation des données

a) Nettoyage de données :

Afin de vérifier la qualité de nos données, nous avons trouvé qu'il est primordial d'examiner la pertinence de toutes les colonnes. La première étape était de chercher les valeurs nulles. Nous remarquons que toutes les données sont bel et bien présentes. La deuxième était de vérifier la consistance des données. Cet examen nous a permis de déduire la nécessité de tous les attributs puisque notre base ne contient aucune valeur manquante ou incorrecte (Age négatif par exemple).

b) Découpage des attributs :

Pour une bonne discrétisation des attributs, nous avons opté pour des découpages en des segments ordonnées des valeurs numériques tirés de notre base de données, des données statistiques (tableau 1) et à l'aide des recherches élaborés par des spécialistes du domaine de la cardiologie afin que nos intervalles soient réalistes. Ensuite, nous avons représenté la classe des intervalles obtenus comme « facteurs ».

	age	sex	paintype	bloodpres	serumCh	bloodsugar	electrocardioR	MaxHeartRaA	ExIndAn	oldpeak	SPESTseg	NbrMajorV	thal
count	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000
mean	54.386617	0.676580	3.171004	131.237918	249.524164	0.148699	1.018587	149.832714	0.327138	1.048327	1.583643	0.661710	4.702602
std	9.093583	0.468653	0.950518	17.808766	51.734796	0.356455	0.997959	23.068318	0.470042	1.147014	0.615011	0.934847	1.941503
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000	3.000000
25%	48.000000	0.000000	3.000000	120.000000	213.000000	0.000000	0.000000	133.000000	0.000000	0.000000	1.000000	0.000000	3.000000
50%	55.000000	1.000000	3.000000	130.000000	245.000000	0.000000	2.000000	154.000000	0.000000	0.800000	2.000000	0.000000	3.000000
75%	61.000000	1.000000	4.000000	140.000000	277.000000	0.000000	2.000000	166.000000	1.000000	1.600000	2.000000	1.000000	7.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	3.000000	7.000000

Tableau 2: les données statistiques

Les intervalles obtenus sont les suivants :

Attributs	Intervalles
Serum cholestoral in mg/dl (chol)	• hautl: si chol>239 • limiteH : si 200≤chol≤ 239 • normal: si chol<200
Resting blood pressure (restbps)	• haut: si restbps>140 • normal : si 120≤restbps≤ 140 • bas: si restbps<120
Maximum heart rate achieved (thalach):	• haut: si thalach>170 • normal : si 85≤thalach≤ 170 • bas: si thalach<85
oldpeak	• Co: si oldpeak>1.6 • Bo : si 0.6≤oldpeak≤ 1.6

	• Ao: si oldpeak<0.6
Age	• A (senior): si l'age>55 • B (adulte) : si 45≤age≤ 55 • C (jaun) : si l'age<45

Tableau 3: Liste des attributs et leurs intervalles

Pour les attributs Cpt et Ca, ils sont déjà catégorisés. Pour cela, nous les avons factorisés directement.

c) Randomisation des données :

A l'aide des fonctionnalités du logiciel R, nous avons ajusté l'ordre des colonnes afin d'éliminer les biais en relation avec l'ordre de saisie des données.

3. Exploration de la base des données préparée

1ère étape : détermination de l'influence de chaque caractéristique sur l'état cardiaque des patients :

Dans le cadre de la visualisation de l'effet des attributs sur la performance et leurs degrés d'importance, nous avons commencé par la forêt d'arbre (Random forest) à l'aide de l'indice de Gini. L'utilité de cet outil se résume seulement en l'explication de l'effet des différents attributs sur l'état cardiaque des patients (présence ou absence d'anomalie). Pourtant, nous allons discuter par la suite l'outil principal adopté pour la classification.
La figure ci-dessous représente le classement des différents attributs utilisés par rapport à la prédiction des anomalies fournie par l'outil Random Forest :

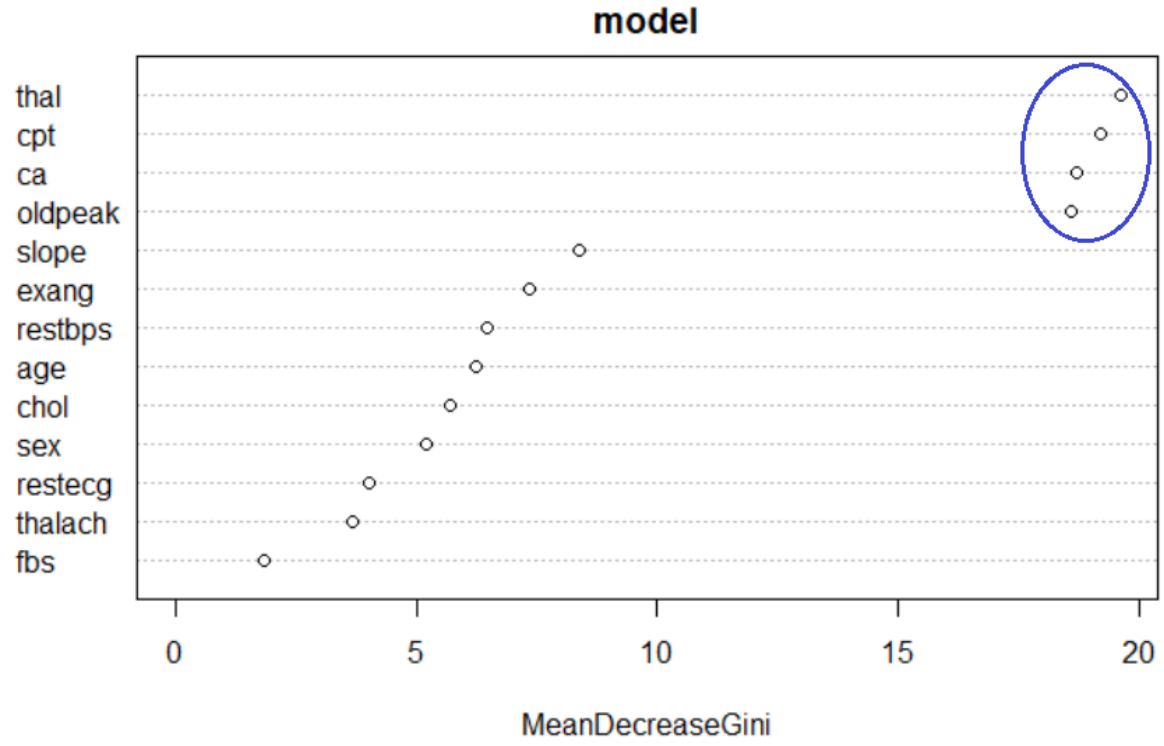

Figure 1: Degré d'importance des attributs

D’après la figure ci-dessus, nous observons que l’indice de Thalassemia (Thal), Cpt, Ca et oldpeak possèdent un degré d’importance plus élevé par rapport aux autres indices sur l’état cardiaque des patients (présence ou absence d'anomalie).
Pourtant, nous ne pouvons pas conclure l’importance des autres indices vu que les pourcentages des attributs majoritaires ne dépassent pas 20%.

2ème étape : Classification

Afin d’évaluer les performances du modèle de classification, nous avons divisé le jeu de données en deux parties : ensemble d’apprentissage 75% et ensemble de test 25%.

- **Phase d’apprentissage :**

Dans cette partie, nous avons entraîné notre modèle avec l’ensemble d'apprentissage.

Choix de l’outil de classification : l’arbre de décision :

Afin de prédire l’état cardiaque, nous avons choisi l’arbre C4.5 pour les raisons suivantes : Premièrement, nous avons besoin d’un arbre n-aire vu que nous possédons des attributs de plus de deux valeurs possibles (exemple **restbps** : élevé, normal et faible).
Deuxièmement, cette méthode permet un élagage automatique.

Présentation des résultats :

La figure ci-dessous représente l’arbre de décision élagué C4.5 obtenu par la fonction C5.0 du logiciel R. Les résultats seront élaborés par la suite :

```
Decision tree:

thal in {3,6}:
:...oldpeak = Co:
:   :...slope = 1: 1 (2)
:   :   slope = 2: 2 (14/3)
:   :   slope = 3:
:   :   :...ca in {0,3}: 1 (3)
:   :       ca in {1,2}: 2 (2)
:   oldpeak in {Ao,Bo}:
:   :...ca = 0: 1 (77/6)
:       ca = 3: 2 (3/1)
:       ca = 1:
:       :...sex = 0: 1 (9)
:       :   sex = 1:
:       :   :...cpt in {1,2}: 1 (2)
:       :       cpt in {3,4}: 2 (5)
:       ca = 2:
:       :...restecg in {0,1}: 1 (2)
:           restecg = 2:
:           :...age = A: 1 (1)
:               age in {B,C}: 2 (3)
thal = 7:
:...cpt = 4: 2 (52/4)
    cpt in {1,2,3}:
    :...fbs = 1: 1 (4)
        fbs = 0:
        :...ca in {2,3}: 2 (3)
            ca = 1:
            :...slope = 1: 1 (1)
            :   slope in {2,3}: 2 (5)
            ca = 0:
            :...restecg in {1,2}: 1 (6)
                restecg = 0:
                :...slope = 1: 1 (4/1)
                    slope in {2,3}: 2 (3)
```

Figure 2: l’arbre C4.5 élagué

Ci-dessous nous voyons le pourcentage d’usage des attributs pour la construction de l’arbre de décision. Ces résultats sont tout à fait conformes avec les degrés d’importance des attributs obtenus au niveau de la forêt d’arbre (Random forest).

```
Attribute usage:

100.00% thal
 64.18% ca
 61.19% oldpeak
 42.29% cpt
 16.92% slope
 12.94% fbs
  9.45% restecg
  7.96% sex
  1.99% age
```

Figure 3: Pourcentage de répartition des attributs

- **Validation avec l'ensemble de test :**

Afin de valider nos résultats, nous avons testé notre modèle avec l'ensemble de test.
La figure suivante contient la matrice de confusion des données de la base d'apprentissage.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1 31  6
         2  7 24
```

Figure 4: la matrice de confusion sur la base de test

Nous pouvons clairement observer que la prédiction a été bonne pour l'ensemble de test contenant 68 objets. Elle ne l'était pas pour 7 objets de la classe 2 et 6 objets de la classe 1. Le taux de la bonne prédiction est donc : 80.88 %.
Le taux d'erreur obtenu ainsi est 19.12%

```
               Accuracy : 0.8088
                 95% CI : (0.6953, 0.8941)
    No Information Rate : 0.5588
    P-Value [Acc > NIR] : 1.323e-05

                  Kappa : 0.6136
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8158
            Specificity : 0.8000
         Pos Pred Value : 0.8378
         Neg Pred Value : 0.7742
             Prevalence : 0.5588
         Detection Rate : 0.4559
   Detection Prevalence : 0.5441
      Balanced Accuracy : 0.8079

       'Positive' Class : 1

Taux d'erreur test est: 0.191176470588235
```

Figure 5: Résultat

5. Discussion des résultats

Les résultats obtenus montrent que les critères les plus discriminants sont les attributs thal, ca, cpt oldpeak Concernant les variables “exang”,”chol”, “resptbps” et “thalach”, ils ont été rejetés par l’arbre de décision vu que leur influence sur l’état cardiaque des patients reste minime voir négligé.
Cependant, nous avons un nombre de faux négatifs égal à 6. En d’autres termes, nous avons des patients affectés par des maladies cardiaques mais qui ont été jugés saints. Ainsi, il y’a un erreur de 16,21% (6/(31+6)=0,1621) par rapport les patients saints. Mathématiquement, nous pouvons dire que notre résultat est bon mais dans le cas réal cela pose un problème, puisque cette erreur n’est pas négligeable en cas de prise en considération l’importance de la vie des patients.
D’un autre côté, nous avons un nombre de faux positifs égale 7. Ceci dit, il existe des patients saints mais qui ont été jugés affectés par des maladies cardiaques. Nous avons ainsi un erreur de 22,58% (7/(24+7)=0,2258) par rapport les patients affectés par des maladies cardiaques. Mathématiquement, le résultat est estimé bon mais dans le cas réel cela pose un problème. En effet, cette erreur n’est pas négligeable si on prend en considération le coût des tests et l’état moral des patients.

6. Conclusion et limites

En guise de conclusion, nous constatons que cette étude demeure utile dans le cadre du pronostic des maladies cardiaques chez les patients avant leurs arrivées et par la suite sauver leurs vies en leur permettant de tenir en compte les préventions possibles. En revanche, nous avouons que la taille de l’échantillon était modeste. Ceci dit, nous ne pouvons pas généraliser les résultats obtenus ni d’en tirer des théories palpables.
Pour cela, nous recommandons de refaire l’étude sur des échantillons plus larges. Ainsi, nous proposons d’ajouter de plus de variables à savoir que la base n’a contenu que des indices physiques internes en négligeant les facteurs externes pouvant avoir influence sur l’état cardiaque des humains (état psychique, entourage, environnement, etc.).

Bibliographie :

1. Méthodes d’exploration des données :
    - Random forest :https://www.r-bloggers.com/random-forests-in-r/
    - Notes de cours
2. Description des données :
    - https://www.medicalnewstoday.com/articles/315900.php
    - https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings?fbclid=IwAR0j4x_eld2TOsh_K5jRCjm_Jj5i7gys7pIQctC6Vo45VgS7MYmalMjWfNI
    - https://www.verywellfit.com/maximum-heart-rate-1231221