Projet Data Science

Raphaël Kanapitsas

Mars 2021

1 Introduction

La rapport porte sur l'étude d'une base de donnée recensant 299 patients (lieu: Pakistan). Pour chaque patient, on dispose de 12 attributs (soit numérique, soit binaire) d'ordre médical, et on sait s'il a souffert d'un arrêt cardiaque ou non, pendant la periode de suivi de l'étude. L'objet sera de prédire, en utilisant les attributs, l'issue du patient.

Dans la première partie, je ferai une présentation rapide de chaque attribut, ainsi qu'une première analyse et visualisation. Je présenterai ensuite la méthodologie utilisée pour créer et tester les modèles. Enfin, seront montrés les résultats obtenus par les différents modèles.

2 La base

La base de donnée est déjà sous une forme utilisable. Il n'y a aucune donnée manquante ou aberrante. En revanche, elle n'est pas tout à fait équilibrée : 68% des patients suivis ont survécu. Un classifieur trivial qui prédirait tout le temps 0 pour le label aura une *accuracy* de 68%.

2.1 Les attributs

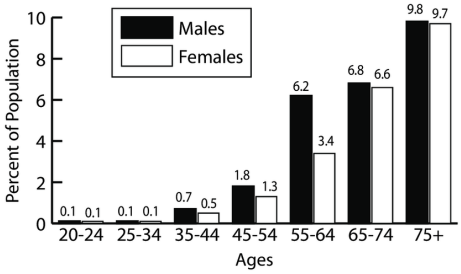
Parmi les attributs, 7 sont à valeurs numériques, et les 5 autres sont à valeurs binaires. La plupart sont des termes médicaux, que j'expliquerai brièvement. En première interprétation, on montrera :

- Pour les variables numériques : un graphique montrant l'issue en fonction de la variable, ainsi qu'une régression logistique sur cette variable.
- Pour les variables binaires : l'espérance conditionnelle $\mathbb{E}(y|x)$, où y est l'issue du patient, et x l'attribut.

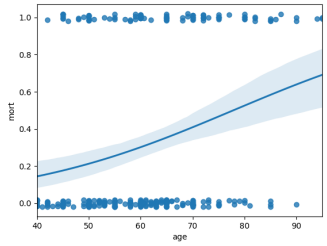
Bien-sûr, ces analyses simples ne prennent pas en compte les dépendances entre les variables, mais elle fournissent une idée de leurs effets et de leurs importances.

2.1.1 Âge

Les patients sont âgés entre 40 et 95 ans, en moyenne ∼60 ans. Il est bien établi que la prévalence des attaques cardiaque augmente significativement avec l'âge (comme la plupart des problèmes de santé), comme le montre la figure ci-dessous. Il semble donc que l'âge ait une influence sur l'issue.



(a) Prévalence en fonction de l'âge



(b) Valeurs de la base

2.1.2 Anémie

L'anémie est un manque de globule rouge dans le sang, ce qui rend plus difficile le transport de l'oxygène pour le corps. Certaines études on montré qu'elle pouvait avoir un effet néfaste sur les problèmes cardiaques. Ici, c'est une variable binaire.

anémie	$\mathbb{E}(y\|\text{anémie})$
0	0,29
1	0,36

L'effet n'est pas très fort, ce ne sera problablement pas une variable très importante pour la suite.

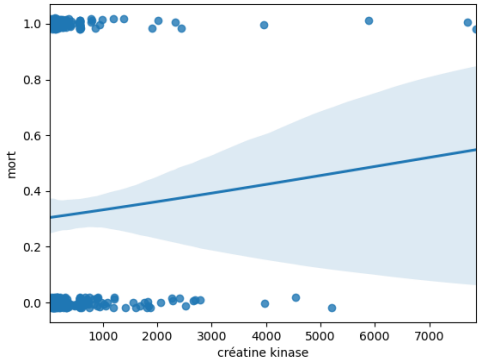
2.1.3 Hypertension

L'hypertension est une pression dans les artères trop élevée. Cela peut être un facteur aggravant d'arrêt cardiaque. Cette variable est binaire, et ici aussi, l'effet est faible.

hypertension	$\mathbb{E}(y\|\text{hypertension})$
0	0,29
1	0,37

2.1.4 Créatine Kinase

La créatine Kinase est une enzyme qui intervient dans les mécanisme impliquant l'ATP, la molécule qui transporte l'énergie au sein des cellule. La créatine kinase est une source d'énergie importante dans le cœur. On voit sur le graphique que la relation n'est pas claire et assez incertaine.



2.1.5 Diabète

Le diabète est une maladie où la régulation de la glycémie (le sucre) se fait mal. Elle conduit a une hyperglycémie. Il est bien établi que le diabète est un facteur aggravant pour les problèmes cardiaques.

diabète	$\mathbb{E}(y\|\text{diabète})$
0	0,32
1	0,32

Mais ce facteur n'a visiblement aucun effet dans cette base, ce qui est assez surprenant. Il y a peut-être des facteurs de confusion.

2.1.6 Fraction d'éjection

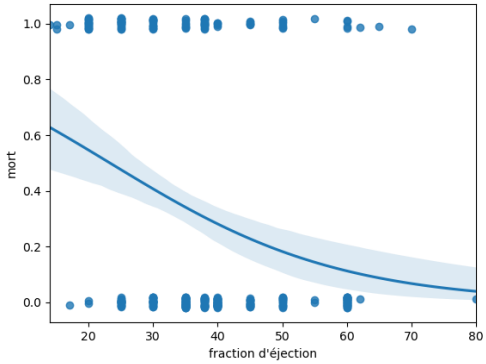
La fraction d'éjection est la proportion de sang qui sort du coœur à chaque contraction. Naturellement, une valeur élevée est signe d'un cœur en bonne santé.

- Une valeur entre 50 et 70% est bonne.
- Entre 40 et 50% passable.
- Moins de 40% problématique.

On s'attend donc à un effet négatif : plus la fraction d'éjection est grande, moins le patient serait susceptible d'avoir un arrêt cardiaque. Comme on le voit sur le graphique, l'effet est significatif. Cet attribut sera important pour la suite.

2.1.7 Sexe

Dans la section sur l'âge, on peut remarquer que la prévalence de problème cardiaques est légèrement plus élevée chez les hommes que chez les femmes.

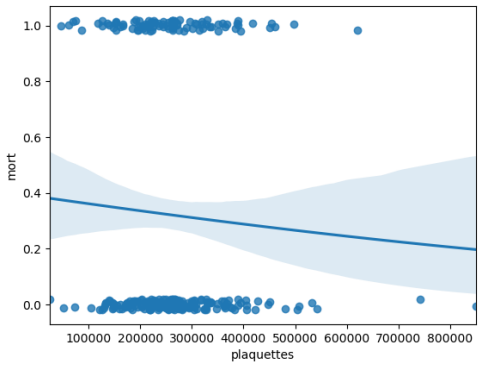


sexe	$\mathbb{E}(y\|\text{sexe})$
femme	0,32
homme	0,32

Dans la base, l'effet est a priori négligeable, ce que je trouve un peu surprenant. À noter que 65% des patients sont ici des hommes.

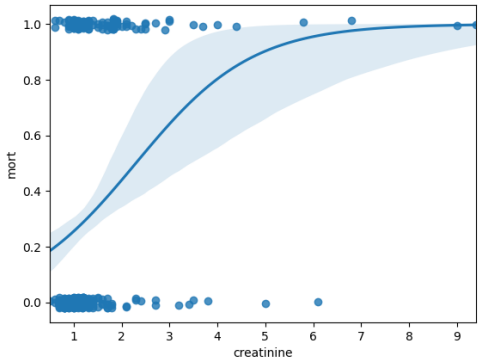
2.1.8 Plaquettes

Les plaquettes sont des fragments de cellules, qui circulent dans le sang. Leur rôle est de permettre la coagulation du sang en cas de rupture des vaisseaux sanguins, afin de combler le trou et éviter une hémorragie. D'un côté, elle peuvent aider, mais en cas de dysfonctionnement, la formation de caillots sanguin peut s'avérer néfaste. La tendance ne semble pas significative.



2.1.9 Créatinine

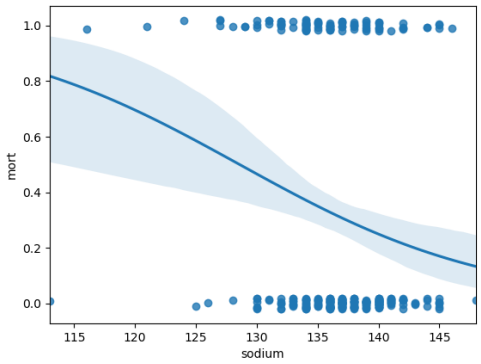
La créatinine est un déchet métabolique issu de la dégradation de la créatine par les reins. Certains patients souffrant de problèmes cardiaques vivent une montée du niveau de créatinine.



Cet attribut semble être l'un ayant le plus d'influence jusqu'à maintenant.

2.1.10 Sodium

Une concentration de sodium trop élevée dans le sang (apportée par l'alimentation par le sel) peut être mauvaise pour le cœur (une concentration trop faible aussi, mais c'est plus rare). L'effet est négatif comme attendu, et semble relativement important.



2.1.11 Tabac

La consommation de tabac n'aide bien-sûr pas les problèmes cardiaques. En effet, elle augmente les chance de formation de plaque dans les artères, et de caillots sanguins, faisant porter un effort excessif sur le coœur. Les fumeurs aurait deux fois plus de chance d'arrêt cardiaque. On s'attend à un effet négatif.

tabac	$\mathbb{E}(y\|\text{tabac})$
0	0,32
1	0,31

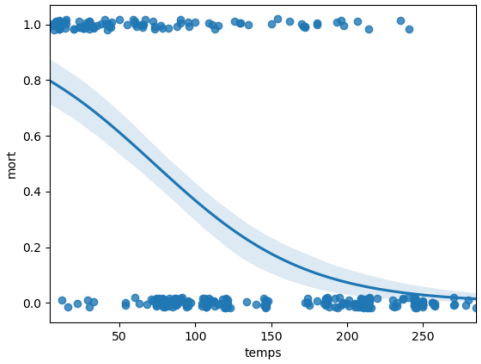
L'effet est négligeable (et va même contre l'intuition).

2.1.12 Temps de suivi

Le dernier attribut est le temps de suivi. C'est le temps du début de l'étude jusqu'au moment où :

- soit le patient décède
- soit on arrête de suivre le patient

Donc si le temps est court, il est probable que le patient soit décédé. Seulement, cette durée n'est évidemment pas connue en cours d'étude, et certainement pas lorsqu'on cherche à prédire l'issue dans le monde réel. Par conséquent, on utilisera pas cette variable pour nos modèle, ce qui reviendrait à connaître le futur.



2.2 Importances des attributs

Pour comparer l'importance des attributs de manière quantitative, j'utiliserai les facteurs de corrélation, ainsi qu'un modèle de forêt aléatoire.

Les deux approches donnent des résultats très similaires. Comme attendu, la variable de temps de suivi est la plus révélatrice de l'issue du patient, mais on s'interdira de l'utiliser dans la suite.

Il reste alors quatre facteurs qui se démarquent dans les deux graphiques : la créatinine, la fraction d'éjection, l'âge et la concentration de sodium. Ces résultats sont en cohérences avec relations que l'on trouvait dans la section précédente.

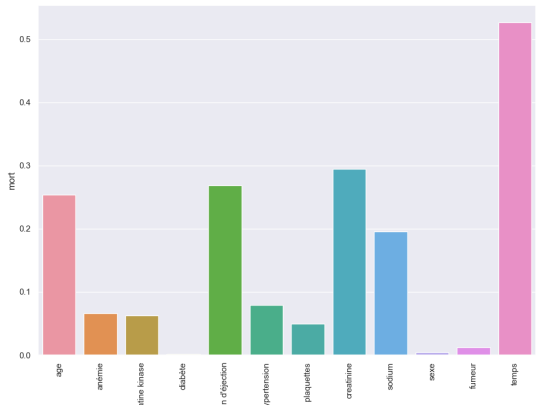


Figure 2: Facteurs de corrélation (en valeur absolue) entre les attributs, et l'issue

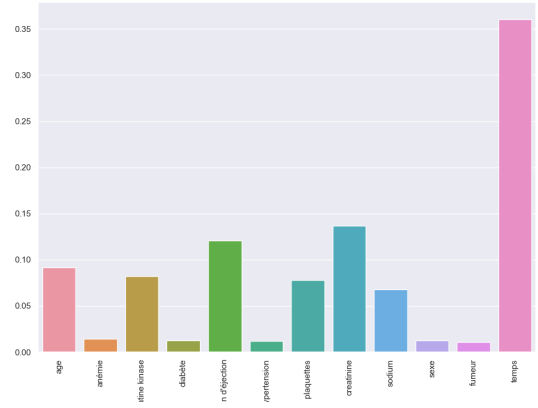
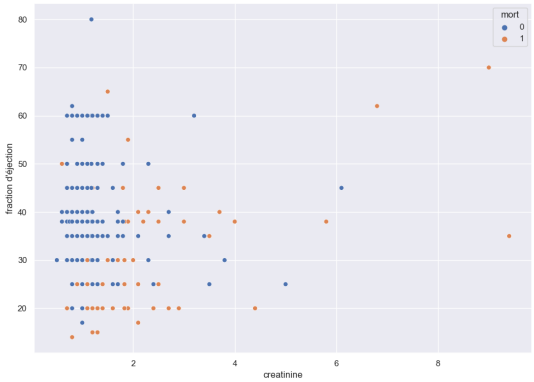


Figure 3: Facteurs d'importance des attributs dans un modèle de forêt aléatoire

2.3 Fraction d'éjection et créatinine

Peut-on inférer l'issue du patient en utilisant seulement les deux attributs les plus important ?



On voit qu'il est facile de séparer les points en deux parties qui sépare **grossièrement** les deux groupes.

3 Méthodologie

Dans cette section, je décrirai la démarche utilisée pour faire les modèles et les tester.

3.1 Validation croisée

Vue la faible taille de la base de donnée (n=299), je trouve préférable d'utiliser la validation croisée pour chercher les bons hyper-paramètres. Il y aura ainsi plus de données sur lesquelles apprendre. Dans les résultat qui vont suivre, j'utiliserai un échantillon test de taille 80 (environ 25% de l'échantillon, seed=42), et une validation croisée 10-fold.

3.2 Sélection des attributs

Pour chaque modèle, on testera la performance en utilisant tous les attributs (sauf le temps de suivi), ou seulement des quatre mentionnés ci-dessus.

4 Résultats

4.1 Modèle simple (benchmark)

Pour commencer, le modèle acceptable le plus simple : une régression logistique sur les 4 attributs importants (avec pénalisation l^2). Vue la différence d'échelle et la présence de variables binaires, il est nécessaire de normaliser les valeurs entre dans l'intervalle $[0, 1]$. Le facteur de complexité C est déterminé à $C = 100$ par validation croisée 10-fold.

Avec ce modèle, j'obtiens une *accuracy* de 76% sur la validation, et 70% sur le test, ce qui est un peu mieux que le classifieur trivial mentionné plus haut.

y	$\mathbb{P}(\hat{y} = y\|y)$
0	0,77
1	0,59

En regardant le taux de bonne réponse pour chaque classe, on voit que le classifieur obtenu est relativement équilibré : on fait significativement mieux que le hasard pour les deux classes.

4.2 Attributs polynomiaux

Les quatre facteurs important pourraient se combiner de manière non-linéaire. En effet, il semble plausible que cumuler deux facteurs aggravant puisse être plus néfaste que la somme de leurs effets isolés. Pour tester cet idée, j'utilise un noyau polynomial avant d'appliquer la régression logistique. Pour plus de cohérence, j'inverse la fraction d'éjection (0 correspond au meilleur, et 1 au pire), pour que tous les facteurs aient un effet négatif lorsque grands. Ainsi, plus les polynômes seront grand, plus le patient cumulera de facteurs négatifs (du moins, c'est l'idée).

En pratique, on obtient 78% sur la validation, mais 66% sur le test : il y a peut-être un peu de sur-apprentissage. Le modèle généralise moins bien.

4.3 Comparaison de modèles plus complexes

J'effectue une validation croisée sur cinq autres modèles, plus complexes :

- Un MLP avec une à trois couches cachées
- SVM avec noyau RBF et polynomial, ainsi que plusieurs niveaux de complexité
- K-plus proches voisins avec des valeurs de K entre 3 et 30
- AdaBoost avec arbres, avec un nombre d'estimateurs entre 1 et 400
- Forêt Aléatoire avec 5 à 800 arbres, et une profondeur maximale entre 1 et le maximum

Voici les résultats pour le meilleur de chaque technique:

Modèle	Score de validation	Meilleurs paramètres
MLP	0,77	1 couche, 50 neurones
SVM	0,79	C=10, noyau polynomial
K-NN	0,77	K=10
AdaBoost	0,80	4 estimateurs
Forêt Aléatoire	0,81	100 arbres, profondeur 2

Tous les modèles ont été entraînés uniquement sur les quatre attributs les plus importants. Lorsqu'on les utilise tous, les scores de validation sont uniformément moins bons (de l'ordre de 0,03 en moins).

Le meilleur modèle au niveau du score semble être la forêt aléatoire. Mais ce modèle se généralise-t-il mieux ? L'*accuracy* test est de 66%. Par classe :

y	$\mathbb{P}(\hat{y} = y\|y)$
0	0,67
1	0,65

La forêt aléatoire a l'avantage d'aussi bien prédire les deux classes. Mais elle reste moins performante que la régression logistique simple.

4.4 Meilleur modèle

Le résultat est quelque-peu décevant : le meilleur modèle est le plus simple, une régression logistique basée sur les quatre attributs que j'ai jugé les plus importants. Je vois deux explications possibles :

- Les données ne suffisent simplement pas pour prédire l'issue avec une meilleure précision.
- La base de donnée est trop petite pour correctement apprendre comment faire mieux : on tombe trop vite dans le sur-apprentissage.

S'il fallait utiliser le modèle en vrai, il faudrait ré-entraîner une régression logistique sur toute les données, avec $C = 100$. Vue la performance, le résultat d'un tel modèle ne serait qu'un indicateur parmi d'autres, et ne devrait pas être pris trop au sérieux. Il serait sans doute possible de faire mieux avec plus de données (au moins un ordre de grandeur en plus).

5 Conclusion

Cette rapide analyse montre trois choses :

Premièrement, des modèles plus complexes ne sont pas forcément meilleurs. Dans le cadre présent, avec très peu de données disponible, le sur-apprentissage n'est jamais loin.

Deuxièmement, il faut un minimum comprendre les données dont on dispose. Ici par exemple, utiliser le temps de suivi aurait mené à de bien meilleures performances. En revanche, le modèle n'aurait pas été utilisable, puisque ce n'est pas une donnée dont on dispose *en pratique*.

Finalement, ne pas utiliser toute les données peut, dans certains cas, mener à un meilleur classifieur. Ici, j'obtenais de meilleures performance en utilisant seulement les quatre attributs les plus importants (ce qui est aussi lié à la connaissance des données). Cela est contre-intuitif. Il est possible que les autres variables ajoute du bruit plutôt que de l'information pertinente, bruit sur lequel il est facile de sur-apprendre.