Université Paris Dauphine

Expliquer et prédire les maladies cardiaques

Victoire de Salaberry

2023 – 2024

Projet de visualisation des données - M2 ISF

Table des matières

1 Introduction

Les maladies cardiaques représentent l'une des principales causes de décès dans le monde. Selon l'Organisation Mondiale de la Santé (OMS), elles sont responsables d'un nombre alarmant de décès chaque année : "on estime à 17,7 millions le nombre de décès imputables aux maladies cardio-vasculaires, soit 31% de la mortalité mondiale totale" [1]. Elles englobent un large éventail de conditions affectant le cœur et les vaisseaux sanguins. Ces affections peuvent inclure des maladies telles que l'insuffisance cardiaque, les maladies coronariennes, les arythmies cardiaques entre autres. Les symptômes varient mais peuvent comprendre des douleurs thoraciques, des essoufflements, des palpitations et des vertiges. Ces conditions peuvent être influencées par divers facteurs tels que le mode de vie, les antécédents familiaux, le régime alimentaire, le niveau d'activité physique et d'autres facteurs environnementaux.

Dans ce projet, nous allons nous intéresser à ces symptômes et voir s'ils ont effectivement un lien avec la présence de maladie cardiaque chez un individu. À noter que les facteurs extérieurs ne seront pas pris en compte. Les résultats devront donc être nuancés.

Pour cela, nous allons étudier un certain nombre de variables d'un jeu de données [2] pour voir lesquelles ont le plus d'impact sur la santé cardiaque. Dans un premier temps, nous analyserons le jeu de données. Grâce à des modèles et méthodes statistiques appropriés, nous allons ensuite identifier les variables qui ont un impact significatif sur la variable cible, à savoir la présence ou l'absence de maladies cardiaques chez les patients. Enfin, nous essaierons de prédire au mieux cette variable à partir de modèles statistiques dont nous étudierons les performances.

2 Présentation des données et premières analyses

2.1 Source des données

Le jeu de données sur lequel a été réalisé ce projet provient du site Kaggle [2] qui est une source de données publiques.

2.2 Dimensions du jeu de données

L'étude est réalisée à partir de 303 individus ce qui est correct pour une étude dans le domaine de la santé. En effet, dans ce domaine, il est souvent difficile d'avoir beaucoup d'observations. Pour chaque individu, on a des informations sur 14 variables qui sont décrites dans le paragraphe ci-dessous.

2.3 Description des variables

Les variables sur lesquelles notre étude est basée sont :

- ***target*** : cette variable est la variable cible, celle que l'on cherche à expliquer et prédire. Comme vu en introduction, elle représente l'état du patient. Elle vaut 0 si l'individu est en bonne santé et vaut 1 si l'individu souffre d'une maladie cardiaque.
- ***age*** : cette variable représente l'âge de l'individu. Dans notre jeu de données, les individus ont entre 29 et 77 ans avec une moyenne d'âge de 54 ans.
- ***sex*** : cette variable représente le genre de l'individu. Elle vaut 0 si l'individu est une femme et 1 si l'individu est un homme. Dans notre jeu de données, il y a deux fois plus d'hommes que de femmes.
- ***cp*** : cette variable représente si un patient a des douleurs thoraciques ($1, 2$ ou 3) ou non (0). 1 signifie que l'individu présente des douleurs faibles, 2 des douleurs modérées et 3 de fortes douleurs.

La moitié des individus ne souffre pas de douleur thoracique.

— ***trestbps*** : cette variable représente la pression artérielle systolique de l'individu au repos. Elle est exprimée en mmHg (millimètre de mercure). Une pression artérielle est normale si elle est inférieure à 120 mmHg [3]. La plupart des individus ont une pression élevée voir à risque car supérieure à ce seuil.

— ***chol*** : cette variable représente le niveau de cholestérol sérique de l'individu. Elle est exprimée en mg/dL. En l'absence de facteurs de risque cardiovasculaire, un niveau optimal est un niveau inférieur à 200 mg/dL [4]. La plupart des individus ont un niveau qui dépasse ce seuil.

— ***fbs*** : cette variable représente le taux de glycémie à jeun. Si celui-ci est inférieur à 120 mg/dL, c'est-à-dire si l'individu à une taux de glycémie normal [5], la variable vaut 0, sinon elle vaut 1. Une grande majorité des individus ont un taux de glycémie normal.

— ***restecg*** : cette variable représente les résultats électrocardiographiques au repos (le niveau de l'ECG). Elle vaut 0 si les résultats sont normaux, 1 s'il y a des changements mineurs ou des anomalies mineure dans l'ECG et 2 s'il y a des anomalies plus importantes. Dans notre jeu de données, il y a presque autant d'individus dont les résultats électrocardiographiques sont normaux (0) que de personnes ayant des résultats avec des changements mineurs (1).

— ***thalach*** : cette variable représente la fréquence cardiaque maximale atteinte, exprimée en bpm (battements par minute). Elle est normalement comprise entre 200 et 140 bpm avec une nette diminution avec l'âge [6], ce que nous vérifierons dans ce projet.

— ***exang*** : cette variable représente si l'individu a une angine de poitrine (douleur thoracique) induite par l'exercice (1) ou non (0). Il y a deux fois plus d'individus qui n'ont pas d'angine de poitrine que d'individu qui en ont.

— ***oldpeak*** : cette variable représente la dépression du segment ST induite par l'exercice par rapport au repos, c'est-à-dire les changements dans l'électrocardiogramme (ECG) qui se produisent en réponse à l'exercice physique. Cela fait référence à des pas de dépression du segment ST observée par rapport à un niveau de référence. Elle est exprimée en mm et est normalement très faible. Les changements dans le segment ST sont importants en cardiologie car ils peuvent être indicatifs de problèmes cardiaques [7], ce que nous vérifierons dans ce projet.

— ***slope*** : cette variable représente le niveau de pente de l'individu (le degré de dépression du segment ST). Ces niveaux sont généralement utilisés pour quantifier la gravité des changements dans le segment ST, en particulier, la dépression du segment ST. La plupart des individus ont une dépression du segment ST de niveau 1 ou 2 (environ 85%).

— ***ca*** : cette variable représente le nombre de vaisseaux principaux (0 à 4) colorés par la flourosopie. Plus de la moitié des individus n'ont pas de vaisseaux principaux colorés par la flourosopie.

— ***thal*** : cette variable représente le score de thallium. C'est une mesure qui évalue la gravité de la perfusion myocardique (circulation du sang vers le muscle cardiaque) anormale en fonction de la distribution du thallium radioactif dans le cœur. 0 indique une perfusion myocardique normale. 1 est généralement associé à des défauts mineurs de perfusion myocardique. 2 est généralement associé à des défauts modérés et 3 à des défauts graves de perfusion myocardique. La plupart des individus ont des défauts de perfusion myocardique.

2.4 Principaux retraitements

Le jeu de données ne présente pas de valeur manquante ni de valeur aberrante. Certaines valeurs présentes dans ce jeu de données sont atypiques mais nous avons décidé de les garder car ce sont des valeurs extrêmes mais pas aberrantes, ce ne sont donc probablement pas des erreurs de saisie.

De plus, la variable cible est à peu près équilibré : on a environ 50% (54.5%) d'individus malades et 50% (45.5%) d'individus sains dans le jeu de données. Nous n'avons pas de rééquilibrage à effectuer. La

distribution des varaibles explicatibves a été mentionné lors de la description des variable. Une étude plus précise s a été menée dans l'application. Le jeu de données est donc étudié tel quel.

2.5 Corrélations entre les variables

Pour avoir la meilleure analyse possible du jeu de données, il faut étudier les corrélations entre les variables. Si deux variables sont très corrélées, il faut faire une étude de leur relation et potentiellement retirer une des deux variables. En effet, garder ces deux variables pourraient créer un biais et rendre les coefficients du modèle instables.

La matrice des corrélations sous forme de graphique *corrplot* (fig. 1) permet de représenter les coefficients de corrélation entre les variables par des couleurs.

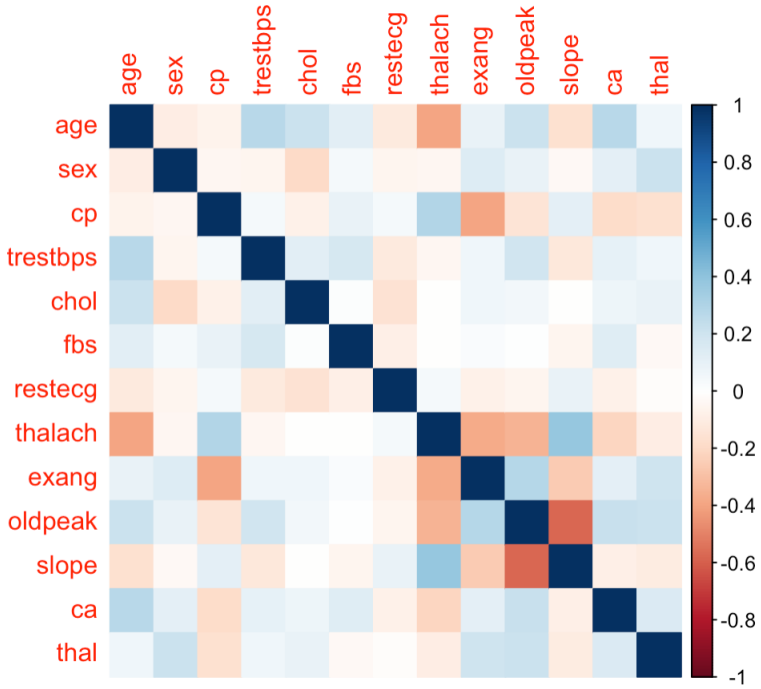

FIGURE 1 – Matrice de corrélations entre les variables explicatives

Les valeurs de corrélation varient de -1 à 1, et indiquent la force et la direction de la relation linéaire entre les variables. Les couleurs varient en fonction de cette valeur : plus la couleur est foncée, plus le coefficient est proche de 1 en valeur absolue et donc plus les variables sont corrélées. Les coefficients positifs sont affichés en rouge, les coefficients négatifs en bleu, et les coefficients proches de zéro en blanc. On peut alors constater que 5 couples de variables sont plus corrélés que les autres : les couples composés des variables *oldpeak* et *slope ; age* et *thalach ; exang* et *cp ; slope* et *thalach* et enfin *thalach* et *exang*.

On peut étudier graphiquement la corrélation des variables de ces couples pour mieux visualiser leur lien. Par exemple, l'affichage du graphique de la fréquence cardiaque maximale atteinte (*thalach*) en fonction de l'âge (*age*) (fig. 2) nous permet de confirmer que ces variables ont un lien qui est décroissant et linéaire. La fréquence cardiaque maximale diminue avec l'âge, ce qui vérifie ce qui a été dit dans la description de la variable *thalach*. Les visualisations graphiques des autres couples de variables ont été réalisées dans l'application.

Finalement, pour décider quelles variables vont être retirées parmi les variables corrélées, nous allons réaliser des tests plus poussés dans la suite.

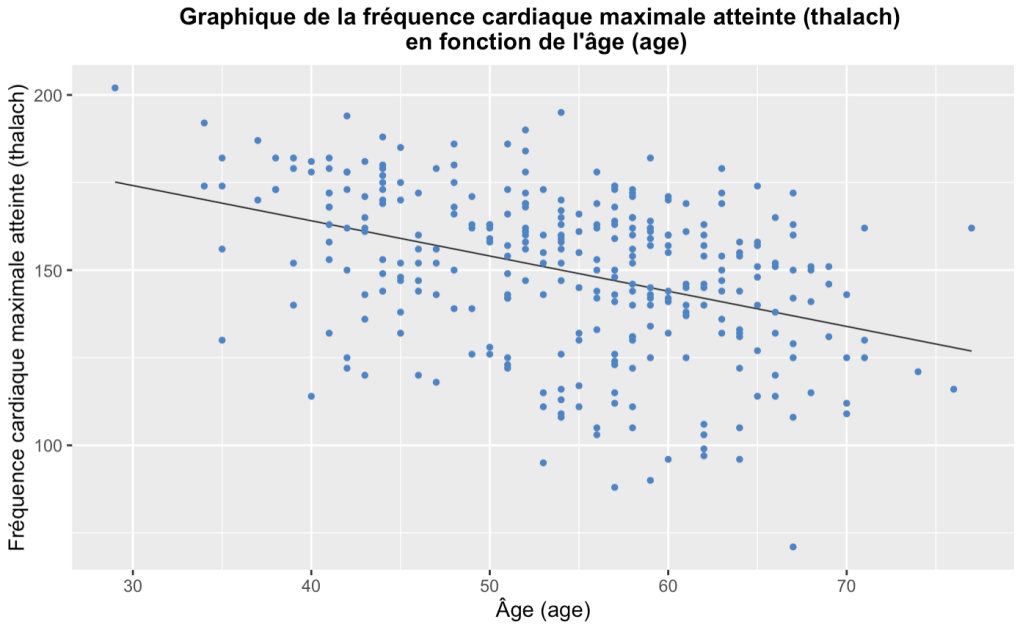

FIGURE 2 – Graphique qui illsutre la relation entre 2 variables explicatives *thalach* et *age*

On regarde ensuite le coefficient de corrélation entre les variables explicatives et la variable cible (fig. 3) afin de répondre à notre objectif d'explication de la variable cible. Les variables qui semblent avoir le plus d'importance sont : *ca, cp, exang, oldpeak, sex, slope, thal* et *thalach*. Comme pour les corrélations entre les variables explicatives, nous allons faire une étude plus poussée pour confirmer ou nuancer cette observation.

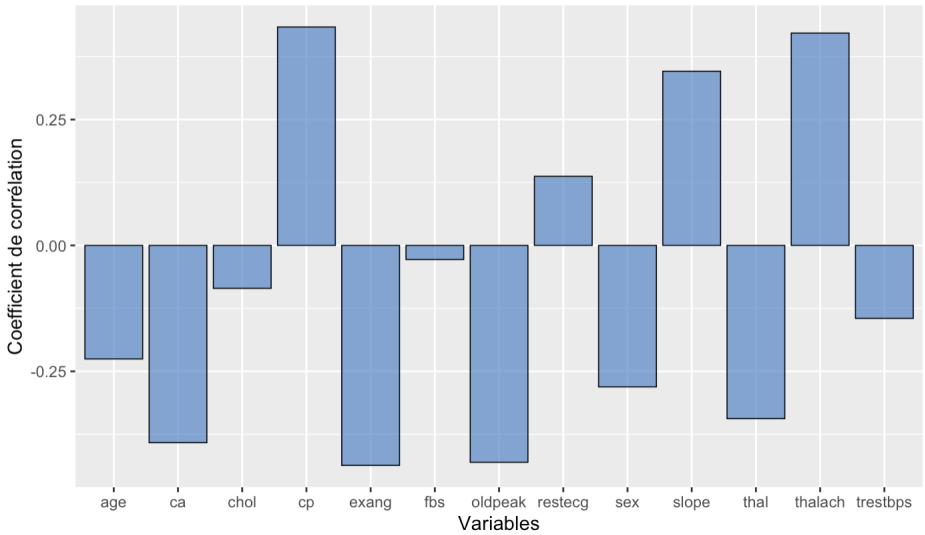

FIGURE 3 – Graphique qui affiche les coefficients de corrélation entre la varible cible et les autres variables

Cependant, on peut, dans un premier temps, observer graphiquement ces corrélations pour les préciser. Par exemple d'après le graphique de la variable *oldpeak* en fonction de la variable *target* (fig. 4), les individus n'ayant pas de changement dans l'electrocardiogramme ont plus tendance à avoir une maladie cardiaque et inversement. Cette observation semble contre-intuitive. Des précisions et explications seront apportées dans la conclusion. Les visualisations graphiques des autres variables corrélées avec la variable cible sont consultables dans l'application.

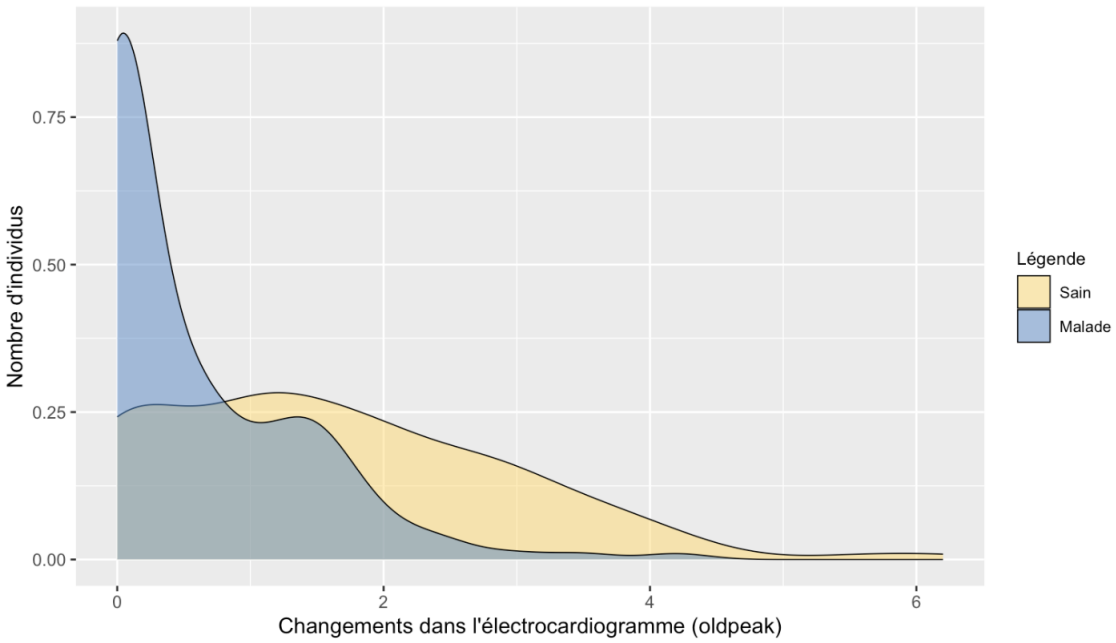

FIGURE 4 – Graphique de la distribution de la la dépression ST induite par l'exercice (*oldpeak*) en fonction de la variable cible (*target*)

2.6 Sampling : division du jeu de données

Afin de finir la préparation de nos données, nous divisons notre jeu de données en 2 ensembles : un ensemble d'entrainement et un ensemble de test.

L'ensemble d'entraînement est utilisé pour construire et entraîner le modèle. C'est sur cet ensemble que le modèle apprend les relations entre les variables explicatives et la variable cible.

Une fois que le modèle est entraîné, l'ensemble de test est utilisé pour évaluer sa performance. Cela permet de mesurer à quel point le modèle est capable de généraliser et de prédire sur de nouvelles données qu'il n'a pas encore vues.

Comme notre jeu de données est plutôt petit, on laisse une grande proportion (80%) de celui-ci pour l'ensemble d'entrainement. On garde 20% des données pour l'ensemble de test. On a alors 242 individus dans l'ensemble d'entrainement et 61 dans l'ensemble de test.

3 Fitting : création et entraînement du modèle

3.1 Création du modèle de régression logistique

Dans un premier temps, on déclare un modèle de régression logistique avec toutes les variables. La régression logistiques est une classe particulière de modèle linéaire généralisé (glm) la plus couramment utilisée pour la modélisation de variables binaires. En effet, c'est un modèle spécialement conçu pour modéliser des variables binaires et qui fournit des résultats facilement interprétables. Comme la variable à prédire est un facteur à deux choix (0 ou 1), on utilise la famille binomiale avec une fonction de lien *logit*.

On analyse ensuite le résumé statistique de ce modèle. Les coefficients de régression dans le modèle logistique mesurent l'impact de chaque variable explicative sur la probabilité de la présence de la maladie cardiaque. Les variables qui semblent les plus significatives sont : *sex, cp, exang, oldpeak, ca* et *thal*. On

remarque que les variables *thalach* et *slope* ne semblent pas significatives d'après ce modèle, or elles avaient un coefficient de corrélation plutôt élevé avec la variable cible. On effectue alors plusieurs tests et méthodes pour décider quelles sont les variables qui sont effectivement significatives.

3.2 Sélection de modèles grâce aux méthodes pas à pas

Dans un premier temps, on fait un test Anova. Ce test permet d'évaluer la significativité des variables explicatives du modèle de régression logistique, en utilisant un test de rapport de vraisemblance spécifique pour les modèles logistiques. Cela permet de déterminer l'importance de chaque variable dans le modèle.

Puis on réalise les méthodes pas à pas *backward, forward* et *both*. Ces méthodes de sélection de variables sont basées sur le critère AIC (*Akaike's Information Criterion*).

L'AIC mesure la qualité de l'ajustement d'un modèle aux données, tout en pénalisant la complexité du modèle pour éviter le surajustement. Il prend en considération à la fois la capacité du modèle à expliquer les données et le nombre de paramètres utilisés. Un modèle avec un AIC plus faible est considéré comme préférable, indiquant un meilleur équilibre entre l'ajustement et la simplicité.

La méthode *backward* commence avec un modèle contenant toutes les variables et supprime itérativement les variables qui ont le moins d'impact sur le modèle, c'est-à-dire les variables dont le retrait du modèle entraîne la diminution du critère AIC.

La méthode *forward* a un principe similaire à celui de la méthode *backward* mais dans le sens inverse : c'est un méthode ascendante et non pas descendante. Elle commence avec un modèle vide et ajoute itérativement les variables qui améliorent le modèle, c'est-à-dire celle qui font diminuer l'AIC du modèle.

Enfin, la méthode *both* est une combinaison de *forward* et *backward* : elle ajoute et élimine alternativement les variables.

3.3 Construction du modèle final

Après avoir fait toutes ces étapes, on peut enfin construire notre modèle final. On garde le modèle construit à partir de la méthode *backward* car c'est celui qui a le plus petit AIC. Notre modèle final est donc composé des variables suivantes : *sex, exang, oldpeak, ca, thal, restecg, cp* et *thalach*.

Remarque : La corrélation entre la variable explicative *slope* et la variable cible est élevée, et pourtant, la méthode *backward* ne garde pas cette variable. Cela s'explique par le fait qu'elle est corrélée avec les variables *oldpeak* et *thalach* qui, elles, sont conservées.

4 Validation du modèle

4.1 Déviance

Nous validons ce modèle car sa déviance (173.07) est bien inférieure à celle du modèle réduit à l'intercept (333.48).

4.2 Résidus

De plus, le graphique des résidus studentisés (les résidus divisés par leur écart-type estimé) nous permet de valider notre modèle. En effet, les résidus suivent un schéma aléatoire donc l'homoscédasticité est vérifiée.

Il y a quelques points à l'extérieur de l'intervalle -2 à 2 qui indiquent des valeurs atypiques, que nous avions déjà traitées précédemment (section 2.4) et que nous avions décidé de garder.

4.3 Outliers

Enfin, en étudiant la présence d'*outliers*, nous obtenons qu'aucun point n'a une distance de Cook supérieure à 1. Il n'y a donc pas d'*outliers* et nous pouvons conserver le modèle tel quel.

5 Optimisation du seuil de décision

On optimise ensuite le seuil à partir duquel on considère qu'un individu à une maladie cardiaque, appelé le seuil de décision. Pour calculer le seuil optimal, il faut qu'on détermine quelle métrique on veut optimiser.

En fonction du problème, on peut préférer un "taux de faux positifs" (FPR, *False Negative Rate*) très faible à un "taux de vrais positifs" (TPR, *True Positive Rate*) élevé. Dans le cas de la santé, on préfère naturellement avoir un petit nombre de faux négatifs (FN) c'est-à-dire un petit nombre de personnes qui n'ont pas été diagnostiquées comme positives alors qu'elles ont une maladie cardiaque.

Pour minimiser le taux de faux négatifs dans un modèle de classification, l'indice le plus approprié à considérer est la sensibilité, également appelée le taux de vrais positifs. La sensibilité mesure la capacité du modèle à détecter correctement les cas positifs (véritables positifs) parmi tous les cas réellement positifs.

Pour maximiser la sensibilité (et donc minimiser les faux négatifs), il faut choisir un seuil de décision qui classe un maximum d'observations positives comme positives, même au détriment de la spécificité. La spécificité ou taux de vrais négatifs (TNR, *True Negative Rate*) est la proportion de vrais négatifs parmi tous les cas réels négatifs. Cela signifie donc que dans notre cas, on tolère davantage de faux positifs (FP) pour éviter les FN.

L'indice de Youden [8] est un critère couramment utilisé pour trouver ce seuil optimal qui maximise la sensibilité. L'indice de Youden est calculé comme suit :

$$IndiceYouden = Sensibilite + Spcificite - 1$$

En maximisant cet indice, on trouve un seuil optimal de 0.556. On peut visualiser ce seuil sur la fig. 5.

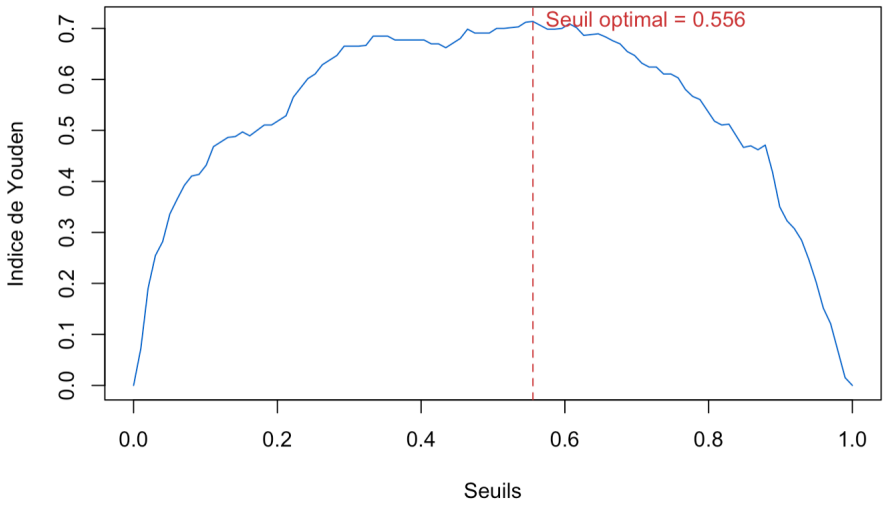

FIGURE 5 – Graphique de l'évolution de l'ndice de Youden en fonction des seuils

6 Prédiction et performances du modèle

6.1 Prédiction

On passe maintenant à la prédiction. On prédit de manière que si la probabilité prédite est supérieure au seuil optimal, l'individu est considéré comme enclin à avoir une maladie cardiaque.

6.2 Matrice de confusion

On évalue ensuite les performances de prédictions de notre modèle. Pour évaluer celles-ci, on peut afficher dans un premier temps la matrice de confusion (fig. 6) [9], qui permet d'observer le nombre de vrais positifs, de faux positifs, de vrais négatifs et de faux négatifs.

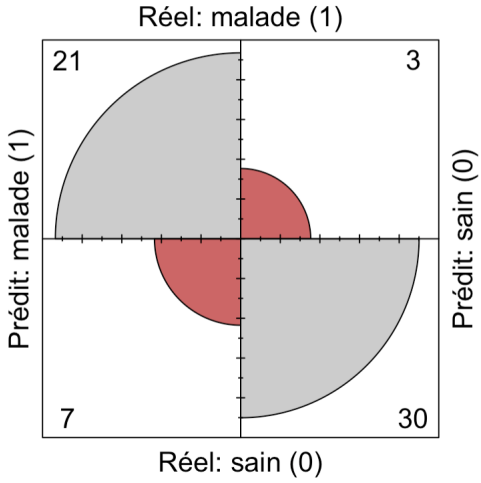

FIGURE 6 – Matrice de confusion du modèle de régression logistique

Le nombre de vrais positifs (respectivement négatifs) représente le nombre d'individu qui ont été prédit malade (resp. sain), et qui le sont vraiment. Inversement, le nombre de faux positifs (resp. négatifs) représente le nombre d'individu qui ont été prédit malade (resp. sain), alors qu'il ne présente aucune maladie cardiaque (resp. ils ont une maladie cardiaque).

On obtient qu'on a prédit 3 faux négatifs, 7 faux positifs, 21 vrais positifs et 30 vrais négatifs. Ce résultat est tout à fait satisfaisait car le nombre de faux négatifs est assez faible, ce qui était notre objectif.

6.3 Indicateurs et métriques de performance

La sensibilité, le F1 Score et la valeur prédictive négative [9] sont des indicateurs importants pour évaluer la capacité du modèle à minimiser les faux négatifs, chacun apportant une compréhension particulière sur cette problématique. Comme on s'est intéressé particulièrement au nombre de faux négatifs, on peut alors s'intéresser à ces métriques pour évaluer les performances du modèle.

— La sensibilité, définie plus haut, met en évidence la capacité du modèle à identifier correctement les vrais positifs. On a, en notant TP le nombre de vrais positifs (*True Positive*), $Sensibilite = \frac{TP}{(TP+FN)}$. Ainsi, plus la sensibilité est élevée, plus il y a des vrais positifs et donc moins il y a de faux négatifs. Dans notre modèle, la sensibilité est de 0.909, ce qui est une bonne sensibilité.

— Le F1 Score est la moyenne harmonique de la précision (définie plus bas) et de la sensibilité. C'est une métrique utile pour évaluer l'équilibre entre la précision et la sensibilité. On a, $F1score = 2 \times \frac{(Precision \times Sensibilite)}{(Precision + Sensibilite)}$. Comme il intègre la sensibilité, une valeur élevée de F1 Score, ce qui est notre cas ici puisque $F1score = 0.857$, indique un faible taux de faux négatifs.

— La valeur prédictive négative (*Neg Pred Value*) est la proportion de vrais négatifs parmi toutes les observations prédites comme négatives par le modèle. On a, en notant TN le nombre de vrais négatifs (*True Negative*), $NegPredValue = \frac{TN}{(TN+FN)}$. Une valeur plus élevée de cette métrique indique une réduction du taux de faux négatifs. Dans notre cas, la valeur prédictive négative est de 0.875, ce qui est plutôt élevée.

De plus, la précision globale (*accuracy*), c'est-à-dire la proportion totale de prédictions correctes effectuées par le modèle, est de 83.6%, ce qui est correct.

Enfin, on mesure l'AUC qui correspond à l'aire sous la courbe ROC. La courbe ROC (*Receiver Operating Characteristic*) est la courbe générée en traçant le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR) à différents seuils. En règle générale, un modèle ayant une bonne capacité de prédiction doit avoir une aire sous la courbe plus proche de 1 (1 étant la valeur idéale) que de 0.5. Dans notre cas, l'AUC est d'environ 0.92 qui est plus proche 1 que de 0.5. Notre modèle est donc adapté aux données.

7 Comparaison avec d'autres modèles

Pour terminer le projet, on compare le modèle de régression logistique à 2 autres modèles : un modèle de forêt aléatoire (*Radom Forest*) et un modèle d'arbre de décision [9].

Chaque modèle a ses propres forces et faiblesses. La régression logistique est un modèle relativement simple et interprétable, tandis que les modèles d'arbres (arbre de décision et forêt aléatoire) sont plus complexes.

La forêt aléatoire est populaire en raison de sa capacité à produire des prédictions précises, à gérer des ensembles de données complexes et bruités, et à éviter le surajustement.

Les arbres de décision, quant à eux, peuvent capturer des relations non linéaires entre les variables explicatives et la variable cible et des interactions complexes. Cependant, ils peuvent être sensibles aux variations des données d'entraînement et avoir tendance à surajuster si l'arbre devient trop complexe.

Comparer ces modèles en évaluant leur performance respective pour prédire la variable cible, nous permet d'avoir une idée de la robustesse des modèles et de leur capacité à généraliser à de nouvelles données. On compare donc les métriques qui nous intéressent, (fig. 7) c'est-à-dire, comme vu dans la partie précédente, la sensibilité, le F1 score, la valeur prédictive négative, la précision, le nombre de faux négatifs et l'AUC. On obtient des modèles qui ont des performances assez similaires.

	Sensibilité	F1 Score	Valeur prédictive négative	Accuracy	Nombre de faux négatifs	AUC
Regression logistique	0.909	0.875	0.811	0.836	3	0.92
Random Forest	0.909	0.870	0.789	0.820	3	0.93
Arbre de décision	0.939	0.909	0.795	0.836	2	0.89

FIGURE 7 – Tableau qui compare les performances des différents modèles

8 Conclusion

Dans le cadre de ce projet, axé sur l'explication et la prédiction des maladies cardiaques, nous avons développé et évalué un modèle de régression logistique, qui est un modèle facilement interprétable. Les performances de ce modèle ont été prometteuses, démontrant une capacité considérable à expliquer et à prédire la présence de maladies cardiaques.

En parallèle, nous avons comparé ce modèle avec des approches alternatives, à savoir un modèle de forêt aléatoire et un modèle d'arbre de décision, qui ont montré des performances globalement similaires à la régression logistique. Cependant, d'autres modèles auraient pu être explorés pour cette tâche, tels que SVM (*Support Vector Machine*), classification naive bayésienne, KNN (*K-Nearest Neighbors*, K plus proches voisins), ou encore le Gradient Boosting.

Par ailleurs, pour des perspectives futures, plusieurs points peuvent être considérés. La comparaison de modèles avec différents découpages *train/test* pourrait offrir une meilleure évaluation de la généralisation des modèles. Introduire des interactions entre les variables pourrait également améliorer la capacité prédictive des modèles.

Dans l'application, on a pu observer des résultats étonnats comme par exemple : D'après les histogrammes des variables explicatives en fonction de la variable cible, les individus qui semblent plus enclin à avoir une maladie cardiaque sont

* les individus ayant des anomalies mineures dans l'ECG, ce qui semble contre-intuitif,
* les individus n'ayant pas d'angine de poitrine induite par l'exercice, ce qui, là-encore, semble contre-intuitif,
* les individus n'ayant pas de vaisseaux principaux colorés par la flourosopie et ceux qui en ont beaucoup (4), ce qui est contradictoire,
* les individus ayant des défauts modérés de perfusion myocardique (niveau 2) mais pas des défauts très élevés (3), ce qui, là aussi, est contradictoire

Ces observations peuvent sembler paradoxales ou contre-intuitives à première vue. Cependant, ces résultats apparemment incohérents peuvent être expliqués par plusieurs facteurs :

— Les maladies cardiaques sont complexes et peuvent être influencées par des interactions subtiles entre plusieurs facteurs. Il se peut que ces variables qui semblent inattendues aient des interactions complexes avec d'autres caractéristiques qui, prises isolément, peuvent sembler contre-intuitives. Par exemple, une anomalie mineure dans l'ECG pourrait être corrélée à d'autres caractéristiques non encore identifiées qui favorisent ou atténuent le risque de maladie cardiaque.

— La taille de notre échantillon étant limitée, certains groupes peuvent être sous-représentés, introduisant ainsi des biais statistiques. À titre d'exemple, on peut citer les 2 variables ci-dessous (on peut se référer à l'application pour le détail de la distribution de toutes les variables) qui ont une catégorie prédominante par rapport aux autres, qui sont sous-représentées :

  - *trestbps* : la pression artérielle au repos de la plupart des individus est supérieure à 120 mmHg, ce qui est un niveau élevé voir à risque.
  - *chol* : la plupart des individus ont un niveau de cholestérol sérique entre 200 et 300 mg/dL, ce qui est très élevé.

  La taille de l'ensemble des données limite donc la capacité des modèles à capturer toute la complexité des relations entre les variables explicatives et la maladie cardiaque.

— Des variables importantes n'ont pas été prise en compte dans le projet, comme précisé en introduction. Des facteurs environnementaux, comportementaux (mauvaise alimentation, sédentarité, tabagisme, consommation d'alcool...), génétiques ou d'autres conditions médicales spécifiques pourraient

influencer les relations observées. Pour de futures études, prendre en compte ces facteurs externes pourrait affiner la prédiction et l'explication des maladies cardiaques.

Références

[1] OMS. Maladies cardiovasculaires. 17 mai 2017. https ://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Kaggle. Heart Disease Dataset. https ://www.kaggle.com/datasets/yasserh/heart-disease-dataset.

[3] OMS. Les différents niveaux de la tension artérielle. https ://www.emro.who.int/fr/media/world-health-day/public-health-problem-factsheet-2013.html : :text=Chez

[4] Ameli. Cholestérol et/ou triglycérides élevés : diagnostic et surveillance. 19 janvier 2021. https ://www.ameli.fr/paris/assure/sante/themes/trop-cholesterol-triglycerides-dans-sang-dyslipidemie/diagnostic-surveillance.

[5] Doctissimo. Glycémie et diabète : comment savoir si mon taux est normal, élevé ou bas ? 6 novembre 2022. https ://www.doctissimo.fr/html/dossiers/diabete/articles/901-diabete-chiffres-faits.html.

[6] Wikipédia. Fréquence cardiaque maximale. 17 octobre 2023. https ://fr.wikipedia.org/wiki/Fr

[7] Dr. Abdelouaheb Farhi. Segment st. 24 janvier 2019. https ://ecgformation.com/blog/segment-st.

[8] Loic Desquilbet. Tutoriel sur les courbes roc et leur création grâce au site internet easyroc. 22 mars 2023. https ://hal.science/hal-02870055/document.

[9] Jorge OCHOA. Les algorithmes supervisés. 2023-2024. Cours de Machine learning - Master 2 ISF.