Heart Disease Group Project

Enzo Ramirez Castellanos / Aymeric Maillard

5/10/2020

Contents

Introduction

La crise sanitaire du Covid-19 que traverse la population mondiale est à l'origine de plusieurs centaines de milliers de morts. L'utilisation de données, de statistiques préfigure comme une approche rapide d'endiguement du virus, différant de la nécessaire lenteur de l'expérimentation médicale. Les données et leur analyse ont toujours été au coeur de la médecine, l'apprentissage statistique pourrait en devenir un véritable allié dans le traitement de situation d'urgence mondiale ou nationale. Des algorithmes d'intelligence artificielle de l'entreprise Bluedot, entreprise canadienne spécialisée dans la prédiction d'épidémie, avait , en effet, anticipé, fin décembre 2019, la propagation du virus. L'appréhension de phénomènes sanitaires par les données est donc possible et peut s'appliquer à d'autres enjeux médicaux comme les maladies cardiaques. Elles constituent, ainsi, la première cause de décès au monde avec près de 17,5 millions de décès chaque année. Leur détection et leur traitement sont donc primordiaux et l'utilisation des techniques de machine learning permettrait alors d'éviter des accidents cardiovasculaires potentiellement mortels. On peut alors se demander si l'intelligence artificielle constitue une réponse cohérente au dépistage des maladies cardiovasculaires.

Les maladies cardiovasculaires sont des pathologies qui touchent le coeur et les vaisseaux sanguins. D'ici 2030, 23,6 millions de personnes mourront chaque année de maladies cardiovasculaires selon l'Organisation Mondiale de la Santé (OMS). Le surpoids constitue un facteur déterminant pour ce type de maladie et l'obésité concernera un quart de la population mondiale en 2050. En effet, la mondialisation des échanges a conduit à la détérioration de l'alimentation et l'accroissement de l'obésité à travers le globe. Dans les pays en développement, les coûts des soins liés aux cardiopathies sont élevés et minent le développement économique des familles les moins favorisés. Une détection davantage précoce aurait comme intérêt de limiter ces frais et d'améliorer l'espérance de vie. Le décèlement des cardiopathies représentera un enjeu certain pour la médecine moderne et l'emploi de nouveaux outils comme l'intelligence artificielle pourrait se montrer efficace. Afin de dépister ces pathologie, il demeure important d'identifier les déterminants et de recueillir ainsi des données porteuses d'intérêt. En 2008, un score a été développé suite à l'étude de Framingham afin de déceler les personnes à risque. Ce score s'appuie principalement sur l'âge, le sexe, le tabagisme, le taux de cholestérol et la pression artérielle. Ces caractéristiques sont présentes dans 4 bases de données de l'Université de Californie Irvine (UCI) publiée en 1988. De nombreux travaux ont été réalisés depuis, sur la prédiction des maladies cardiovasculaires, à l'aide de ces données ou non. Ont été appliquées à la question de multiples algorithmes de machine learning (Arbres de décision, Classification naïve bayésienne, Réseau de neurones, Bagging, ou encore Support vector machine) et les résultats présentent différents niveaux de précision selon les paramètres et les données utilisés. On retrouve, par exemple, une étude réalisée par Detrano R. et. al. Les résultats obtenus en termes de précision avec un seuil de 0.5 sont approximativement de 77%. Ils sont obtenus avec une "logistic-regression-derived discriminant function". D'autres résultats sont obtenus par David W. Aha & Dennis Kibler avec la base de données de Cleveland. Ils obtiennent une précision de 77% avec la méthode NT growth et une précision de 74.8% avec le méthode C4. On retrouve aussi les résultats obtenues par Gennari, J.H. et. al. avec le "CLASSIT conceptual clustering system", ils obtiennent une

précision de 78.9% sur la base de données de Cleveland.

Concernant les données de l'UCI, la plupart du temps, seule la base de données de Cleveland et 14 variables sur 76 ont été utilisés pour les travaux. Dans ce papier, sera utilisé et soumis à un tri l'ensemble des données et des variables. Cette approche se veut plus complète et plus robuste à l'application des différents algorithmes. A l'instar des autres études, seront utilisé des algorithmes d'apprentissage automatique comme Random Forest ou encore K-Nearest-Neighbors. Ces modèles répondent à des problèmes de classification et sont alors à même de détecter ou non la présence d'une maladie cardiovasculaire. Afin de prolonger la réflexion, des méthodes de boosting seront également implémentées comme Adaboost, LogitBoost ou encore Extreme Gradient Boosting. Avec un nombre de dimension important, cette approche permettra de synthétiser l'information présente dans la base de données. Enfin, une méthode de stacking sera introduite afin d'aboutir à un modèle plus précis. On pourra alors répondre à la question suivante : L'intelligence artificielle constitue-t-elle une réponse cohérente au dépistage des maladies cardiovasculaires? Nous présenterons dans un premier temps les données afin de définir nos variables, puis nous procéderons à une analyse statistique en exposant une approche descriptive et des modélisations.

Données

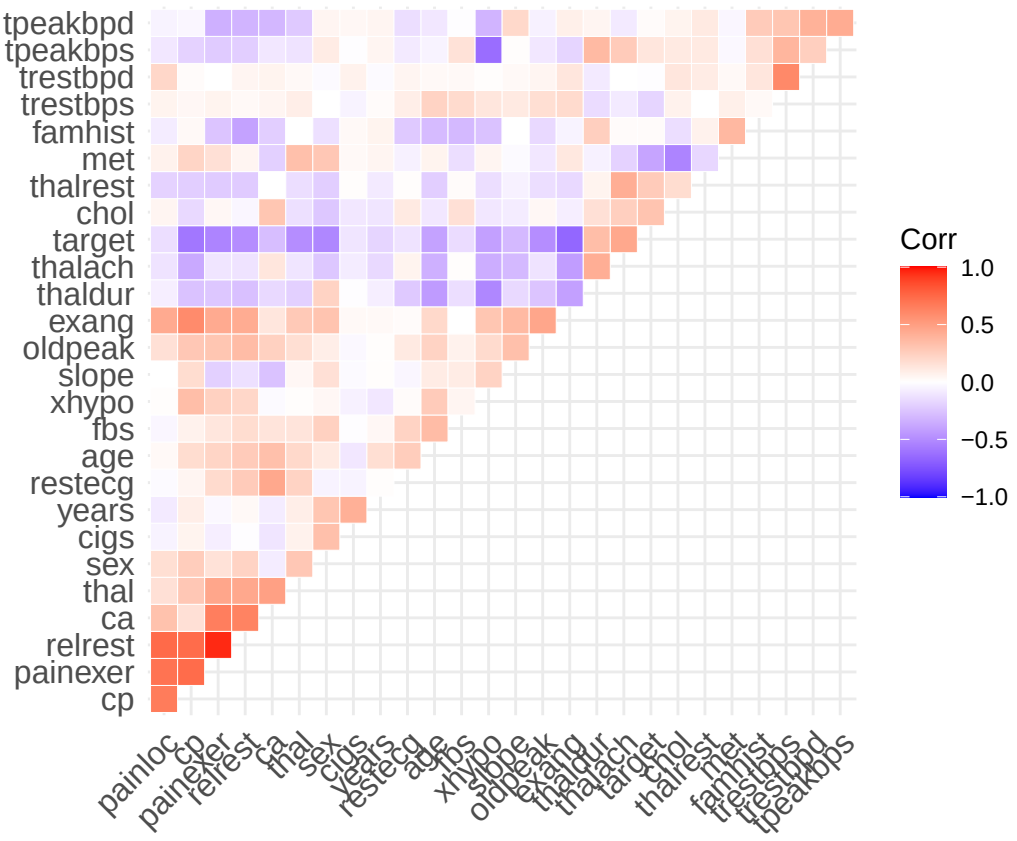
Les données proviennent d'informations rassemblées par deux centres hospitaliers universitaires suisses (Bâle et Zurich), un institut de cardiologie hongrois (Budapest) et deux centres médicaux américains (Californie et Ohio). Elles concernent 76 informations médicales diverses sur 899 patients au total. La variable expliquée décrit les résultats d'une angiographie qui est une technique d'imagerie médicale portant sur les vaisseaux sanguins. Si la valeur de la variable est égal à zéro alors le patient est considéré comme sain et non atteint par une maladie cardiovasculaire (rétrécissement du diamètre des vaisseaux sanguins $< 50\%$). Si la valeur est supérieure à zéro (1, 2, 3 ou 4), le patient est considéré comme malade (rétrécissement du diamètre des vaisseaux sanguins $> 50\%$). Nous n'avons pas trouvé d'informations sur les différences entre les valeurs positives. Un électrocardiogramme et un exercice physique test a été effectué pour chaque patient. Les autres variables concernent, dans un premier angle, des informations personnelles aux individus comme l'âge, le sexe, le fait qu'il soit fumeur et depuis combien de temps. Dans un second angle, les informations font écho à des données purement médicales comme les différents résultats de l'électrocardiogramme, le taux de cholestérol, les résultats de l'exercice physique, le type de douleur thoracique... Notre analyse se focalise donc sur l'utilisation de nouveaux algorithmes pour améliorer la précision des prédictions des modèles actuels. Aussi, l'échantillon de données utilisé dans la plupart des modèles existants n'utilise qu'une partie de la donnée disponible. Nous utiliserons donc l'intégralité des données afin d'établir un modèle qui génélarisera mieux sur les données non observées. Le premier constat sur les données utilisés dans les autres études c'est qu'elles se focalisent sur un échantillon restreint (Cleveland) et uniquement sur 14 variables des 76 disponibles. Nous avons donc récupéré l'intégralité des données et toutes les variables pour en faire un seul échantillon. Parmis les 76 variables, plusieurs ne sont pas exploitables. On ne peut pas déterminer à quoi elles font référence, d'autres comptent trop de données

manquantes, et certaines ne présentent aucun intérêt dans l'analyse (ex : ID). Une fois ces variables retirées, on distingue les variables continues des variables catégoriques. On traitera les données manquantes pour les variables continues en calculant la moyenne de la série et son écart type pour deux catégories de classe d'âges, et en distribuant aléatoirement une valeur comprise entre [moyenne - ecart type ; moyenne + écart type]. Pour les variables catégoriques nous remplacerons la valeur manquante par le mode de la série.

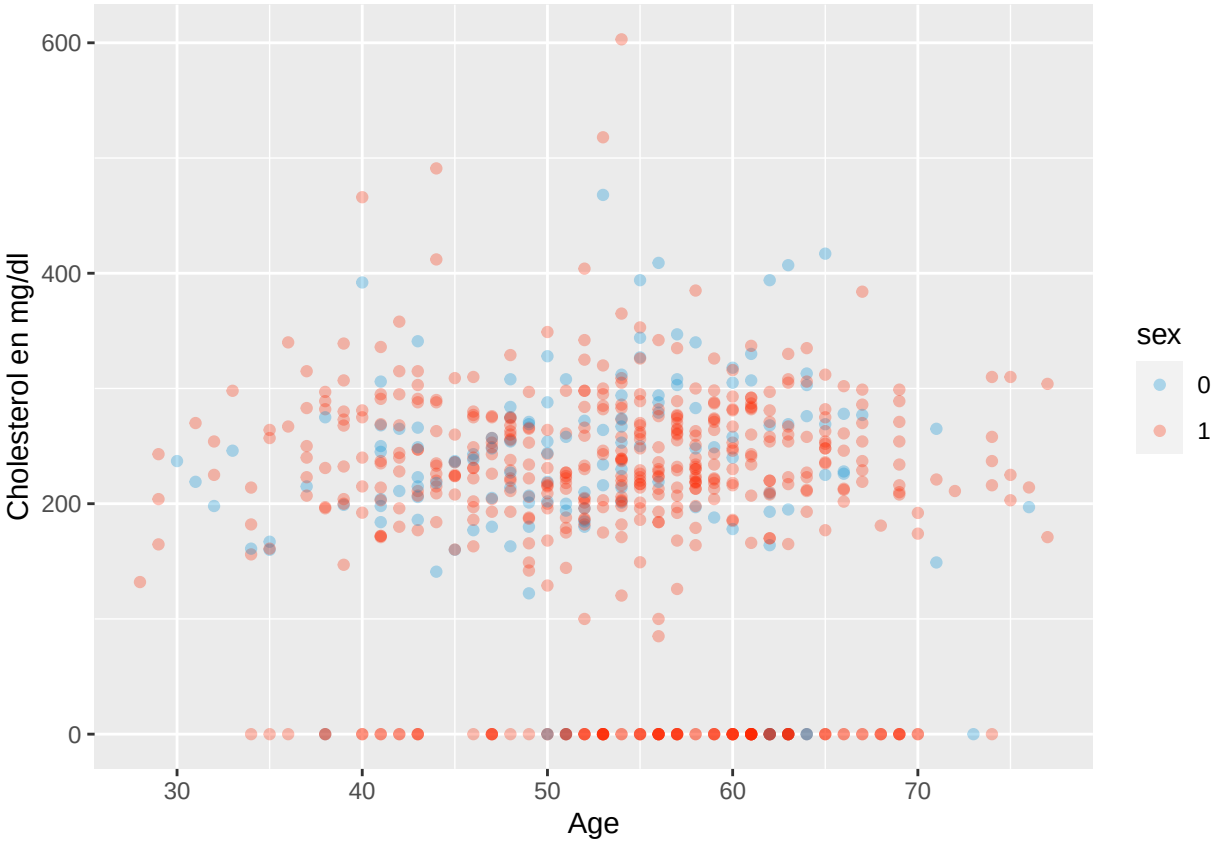
Analyse descriptive

```
##       age           trestbps          chol            cigs
##  Min.   :28.00   Min.   : 80.0   Min.   :  0.0   Min.   : 0.00
##  1st Qu.:47.00   1st Qu.:120.0   1st Qu.:175.0   1st Qu.: 4.60
##  Median :54.00   Median :130.0   Median :223.0   Median :20.00
##  Mean   :53.49   Mean   :132.1   Mean   :199.3   Mean   :18.67
##  3rd Qu.:60.00   3rd Qu.:140.4   3rd Qu.:268.0   3rd Qu.:28.50
##  Max.   :77.00   Max.   :200.0   Max.   :603.0   Max.   :99.00
##      years          thaldur            met            thalach
##  Min.   : 0.00   Min.   : 1.500   Min.   :-12.40   Min.   : 70.0
##  1st Qu.: 5.20   1st Qu.: 6.000   1st Qu.:  5.00   1st Qu.:120.0
##  Median :18.60   Median : 8.000   Median :  7.00   Median :139.9
##  Mean   :17.72   Mean   : 8.686   Mean   : 16.38   Mean   :137.6
##  3rd Qu.:26.40   3rd Qu.:10.800   3rd Qu.: 11.00   3rd Qu.:155.0
##  Max.   :60.00   Max.   :24.000   Max.   :200.00   Max.   :202.0
##     thalrest         tpeakbps        tpeakbpd         trestbpd
##  Min.   : 37.00   Min.   : 84.0   Min.   : 26.00   Min.   : 58.00
##  1st Qu.: 65.00   1st Qu.:156.0   1st Qu.: 80.00   1st Qu.: 80.00
##  Median : 74.00   Median :170.0   Median : 90.00   Median : 80.00
##  Mean   : 75.69   Mean   :172.2   Mean   : 87.67   Mean   : 83.64
##  3rd Qu.: 84.00   3rd Qu.:190.0   3rd Qu.:100.00   3rd Qu.: 90.00
##  Max.   :139.00   Max.   :240.0   Max.   :134.00   Max.   :120.00
##     oldpeak        target    sex     painloc painexer relrest cp      fbs
##  Min.   :-2.6000   Yes:362   0:134   0: 37   0:186    0:153   1: 37   0:563
##  1st Qu.: 0.0000   No :295   1:523   1:620   1:471    1:504   2:124   1: 94
##  Median : 0.5000                                               3:138
##  Mean   : 0.8725                                               4:358
##  3rd Qu.: 1.5000
##  Max.   : 6.2000
##  famhist restecg exang   xhypo   slope   ca      thal
##  0:153   0:395   0:424   0:639   0:135   0:568   0:498
##  1:504   1:128   1:233   1: 18   1:473   1: 45   1: 34
##          2:134                   2: 49   2: 28   2:125
##                                          3: 16
##
```

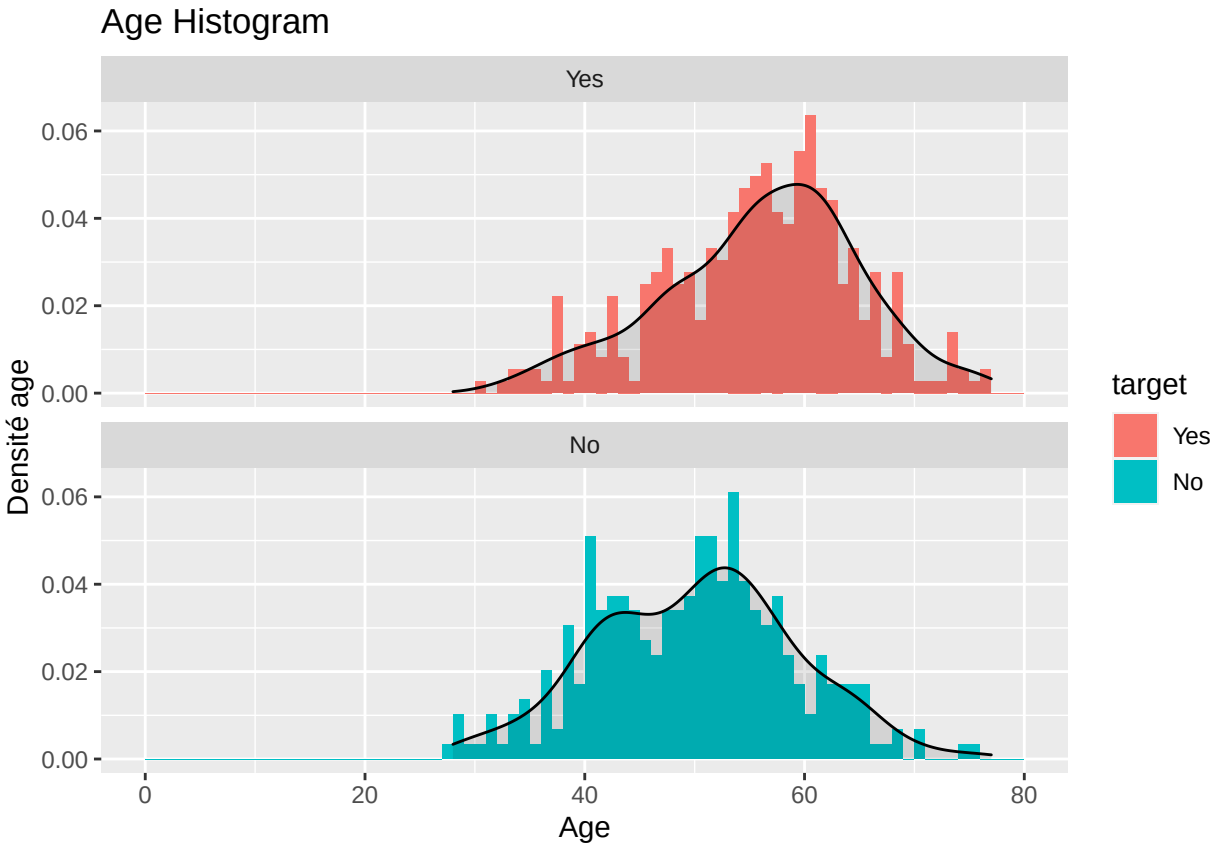
##



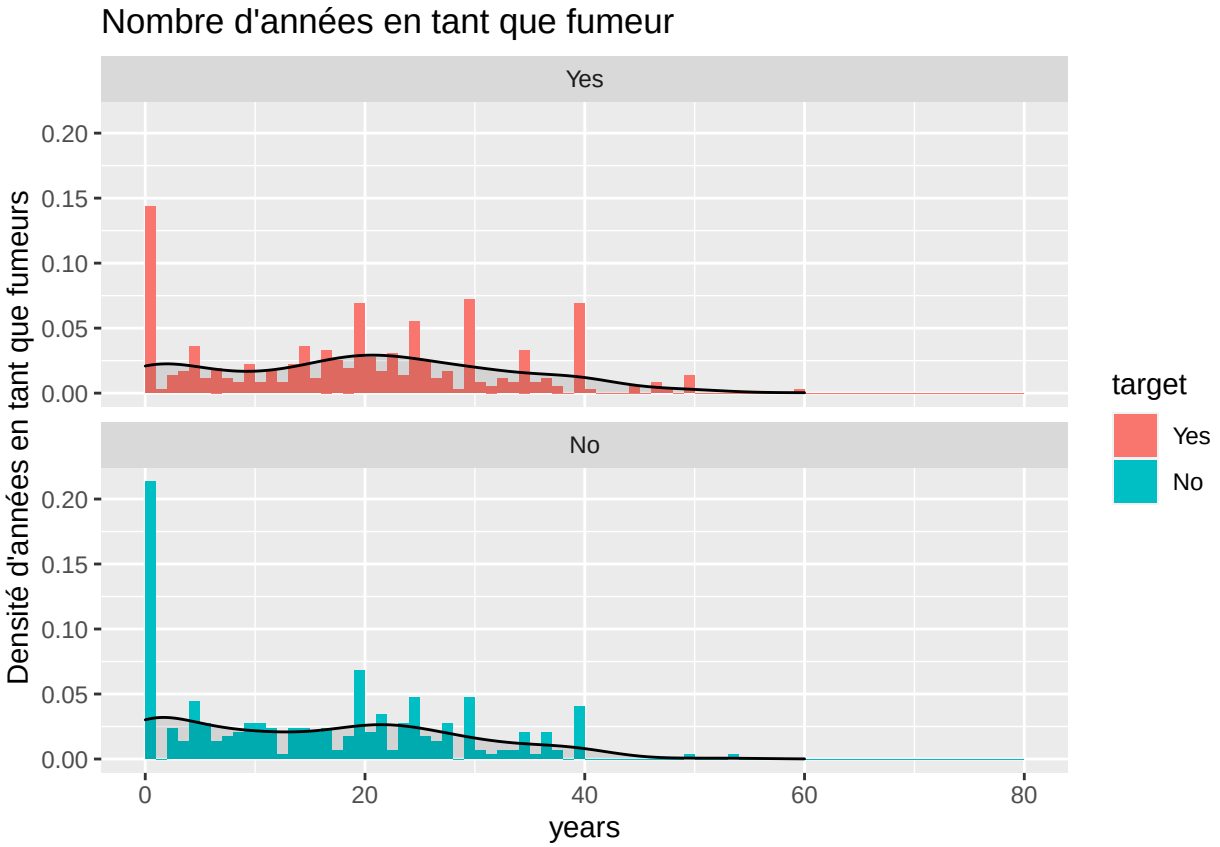
Comme on peut l'observer sur la représentation des corrélations entre nos variables, elles sont rarement corrélées à plus de 50%. Si on se plonge plus en détail sur certaines de nos variables, on observe une distribution du taux de cholestérol similaire entre hommes et femmes à travers les âges.



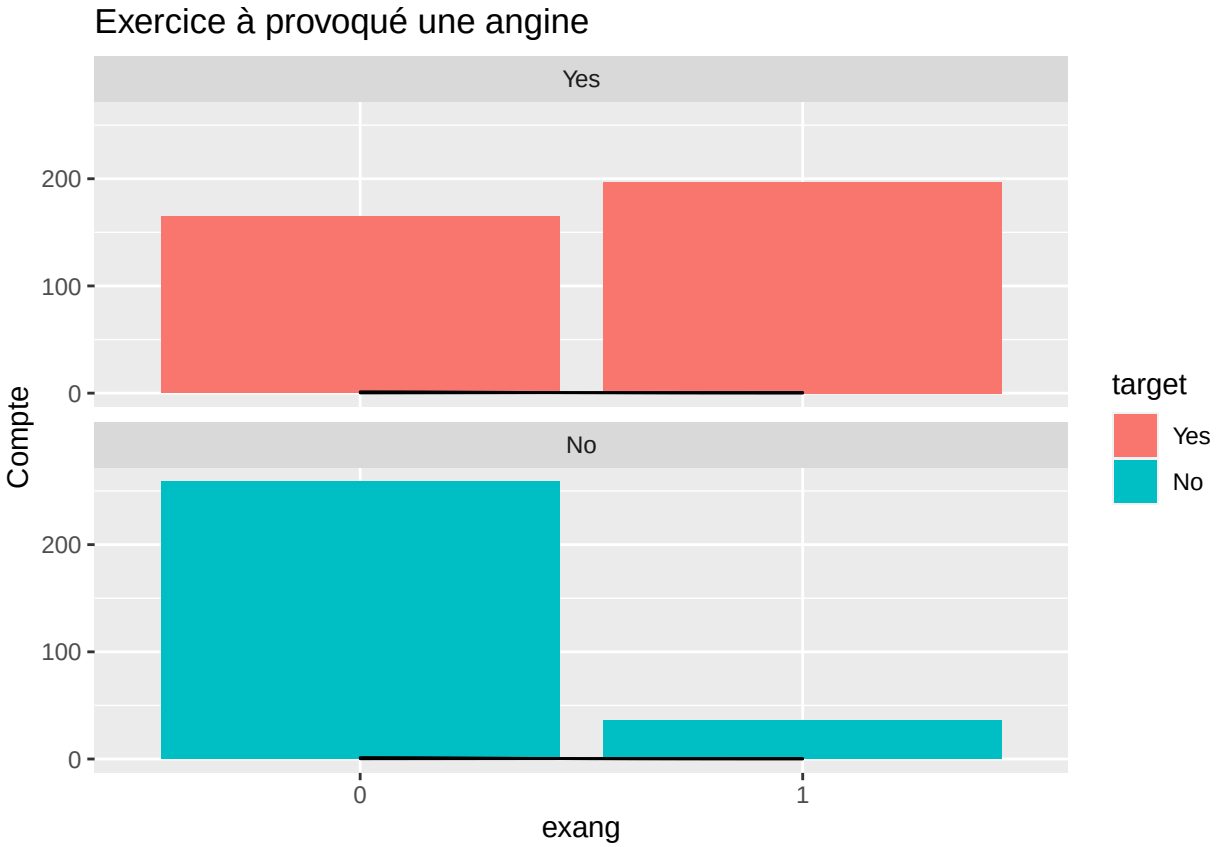
Si on se plonge plus en détail sur certaines de nos variables, on observe une distribution du taux de cholestérol similaire entre hommes et femmes à travers les âges, voir ci-dessus.



Dans cette représentation on peut voir la répartition de l'âge entre populations saines et atteintes de maladies cardiovasculaires. On remarque que la majorité des personnes atteintes se trouve vers la soixantaine, alors que pour la population saine on retrouve une concentration vers la cinquantaine voire quarantaine.



Dans le graphique ici, on représente la densité du nombres d'années en tant que fumeur entre personnes saines et malades. On remarque une plus forte concentration des personnes non fumeurs chez les personnes saines.

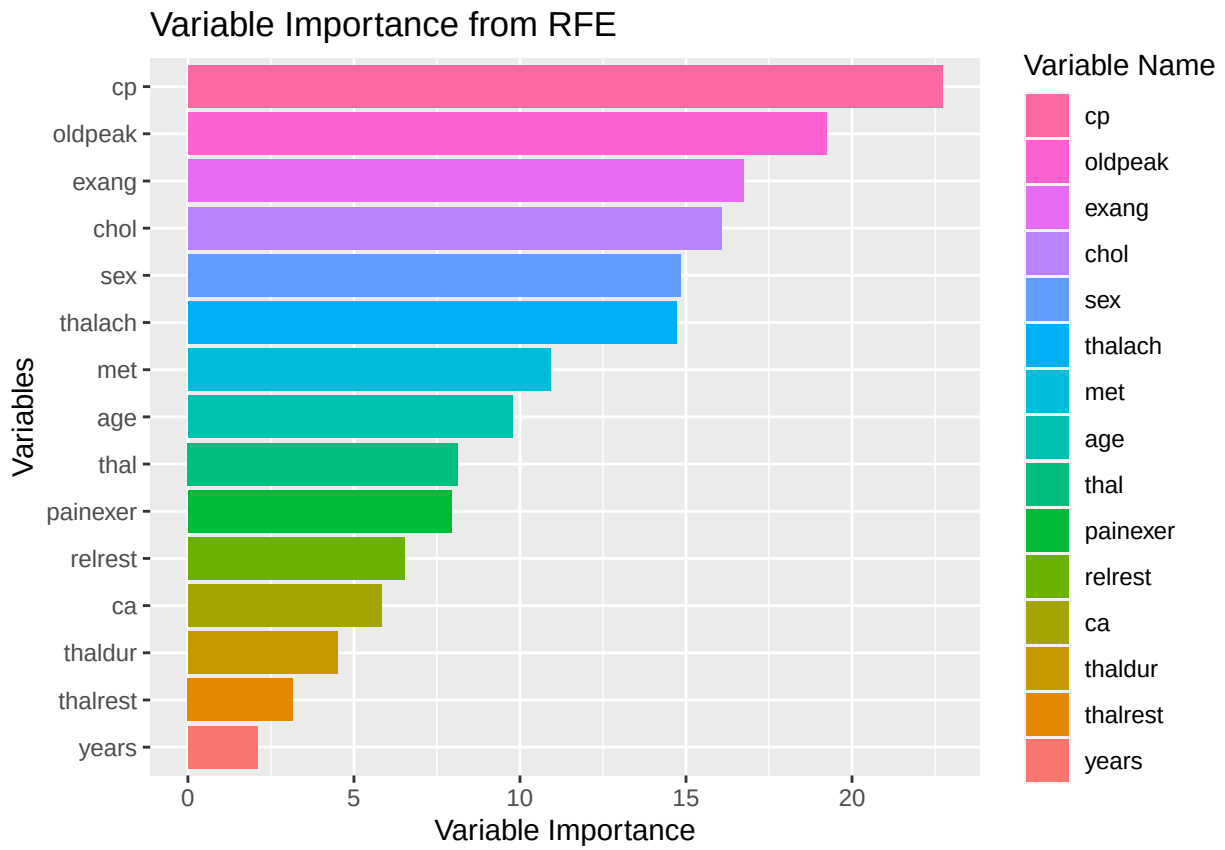


Cette dernière représentation nous montre le nombre de personnes dont un exercice a provoqué une angine entre les personnes saines et malades. On voit une très nette différence entre les deux groupes, les personnes saines sont peu nombreuses à avoir développé une angine à l'inverse, c'était très souvent le cas pour les personnes malades.

Sélection de variables

Afin d'opérer un tri statistique des variables de notre échantillon, il nous semble pertinent de déterminer quelles sont celles qui apportent le plus à notre modèle mais aussi celles qui n'apportent rien ou viennent créer du bruit. Pour procéder à la sélection des variables nous avons utilisé deux méthodes. Un modèle avec une pénalisation (L1) qui mettra en évidence les variables inutiles dans notre modèle en mettant leur coefficient à 0, nous utiliserons la méthode Lasso pour cela. Puis une méthode RFE (recursive feature elimination) qui est une sélection de variables dite backward, qui consiste à calculer après chaque itération du processus l'importance des variables et de retirer celles qui sont les moins importantes. En combinant les résultats obtenus nous avons éliminé plusieurs variables de notre échantillon : le nombre de cigarettes par jour (cigs), la pression artérielle au pic de l'exercice physique test (tpeakbpd), l'historique familiale des maladies cardiovasculaires (famhist), la pression artérielle au repos (trestbps) et une variable sans description (xhypo). Le nombre de cigarettes par jour et l'historique familial auraient pu être des variables intéressante cependant 46% de ces données sont manquantes. Cela rend ces deux variables peu consistantes.

Grâce au modèle RFE on peut obtenir une représentation des variables les plus importantes. Comme elle l'indique, les 5 variables les plus importantes sont le type de douleur dans la poitrine (cp), le sous décalage du segment ST induite par l'exercice par rapport au repos ou non (oldpeak), le fait que l'exercice ait induit une angine ou non (exang), le niveau de cholestérol en mg/dl (chol) et le sexe (sex). Il semble logique que des variables comme l'exercice à induit une angine de poitrine soient pertinentes. En effet c'est un indicateur et un signal envoyé par le coeur pour dire qu'il est en manque d'oxygène. Ceci est très sensiblement lié à la variable sur le niveau de cholestérol, puisqu'un niveau de cholestérol élevé témoignent certainement d'artères obstruées et donc d'un apport en oxygène difficile. Dans les deux cas, il semble certain que ces variables renvoient des informations pertinentes pour le modèle. On retrouve également le type de douleur dans la poitrine qui peut être normal, anormal, ou bien non situé dans la poitrine. Ces informations peuvent renvoyer des informations cruciales sur l'évaluation des symptômes pour le modèle. Enfin, on retrouve le sexe qui joue un rôle important. Cela semble suggérer qu'il y a une dichotomie entre les genres qui permet une meilleure classification. On pourrait imaginer que la constitution cellulaire ou bien que les modes de vies et/ou de consommation des deux genres sont des facteurs sous jacents de ce résultat.



Modélisation

Dans notre étude nous utilisons d'abord plusieurs algorithmes d'apprentissage, qui sont exécutés de manière individuelle les uns des autres. Nous utilisons, ensuite, une méthode d' « ensemble » pour combiner les modèles avec les prédictions les moins colinéaires. En effet, c'est en combinant des modèles qui ne prédisent pas les mêmes résultats qu'on pourra obtenir un modèle qui tirera parti des performances de chaque modèle sur une partie ou l'ensemble de l'espace. La méthode d'ensemble « stacking » est utilisée pour les combinés. On utilisera un algorithme de classification binaire dans un cas et une forêt aléatoire dans l'autre. Finalement on conservera la combinaison avec la forêt aléatoire dans notre cas, car elle présente le taux de faux-négatif le moins élevée. Les faux-négatif représentent les patients malades qui ne sont pas détectés par le modèle. L'enjeu réside donc dans leur minimisation.

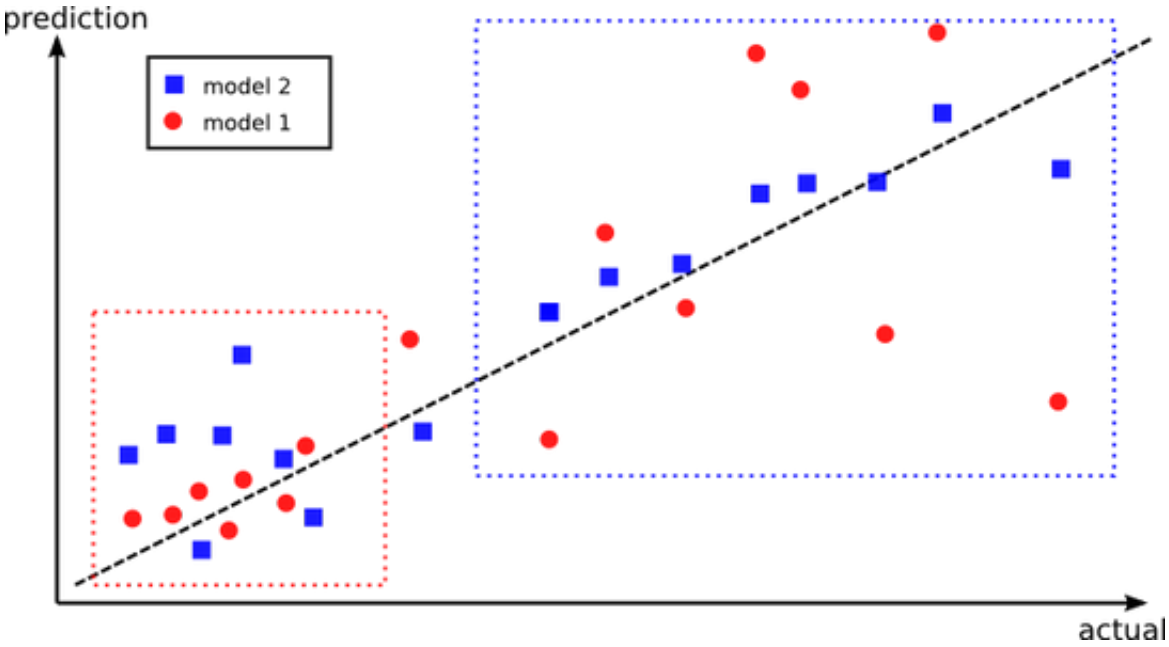


Figure 1: Caption for the picture.

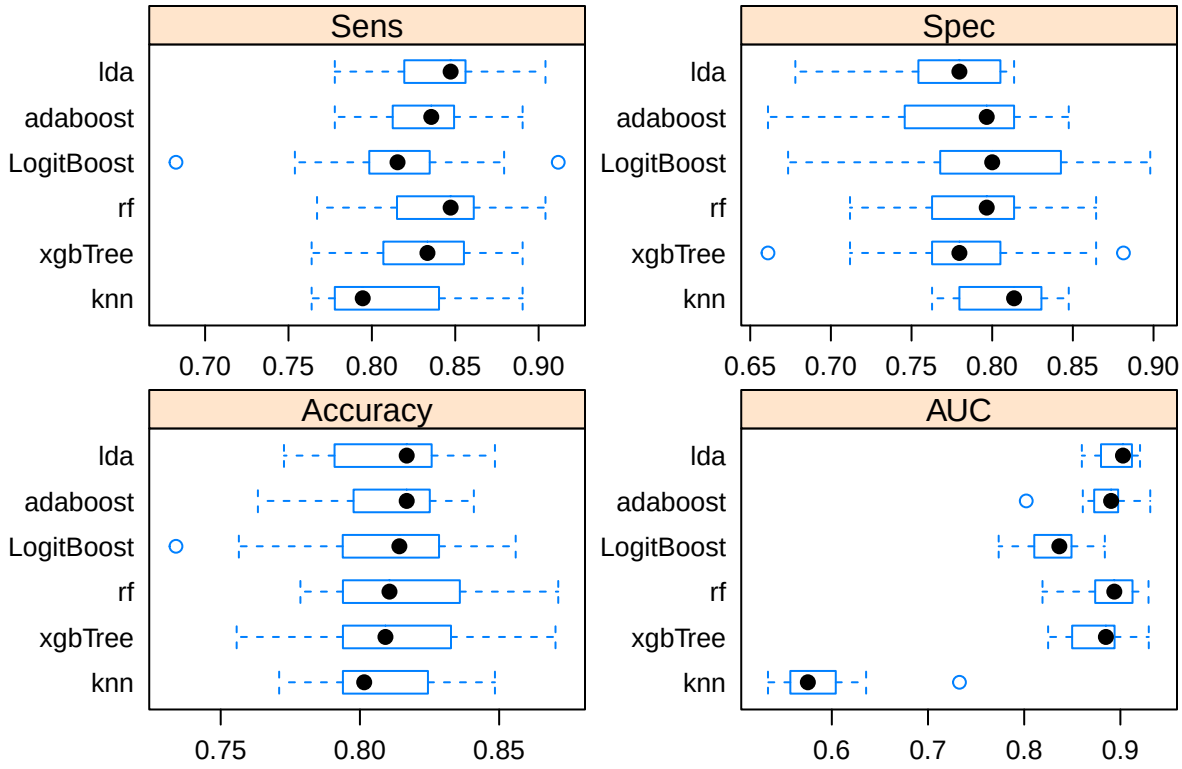
Sur cette image on peut voir comment le modèle rouge semble être meilleur en général. Cependant on remarque que le modèle bleu est plus performant sur une partie des données en combinant ces deux résultats on peut obtenir un meilleur modèle. C'est sur cette idée que le stacking repose.

La méthode pour combiner nos modèles s'effectue en plusieurs étapes. La première consiste à entraîner chaque modèle individuellement sur nos données d'entraînement. Celle ci nous permet d'obtenir pour chaque modèle des prédictions sur notre variable dépendante. En rassemblant chaque prédiction, on obtient un nouvel ensemble de données d'entraînement pour notre seconde étape. Chacune des prédictions est considérée comme une variable et chaque modèle comme une observation. Ici, plusieurs possibilités s'offrent à nous. Par exemple, on pourrait utiliser la règle de vote majoritaire et considérer que la prédiction retenue est celle avec le plus d'occurrence à chaque fois. On pourrait également calculer la

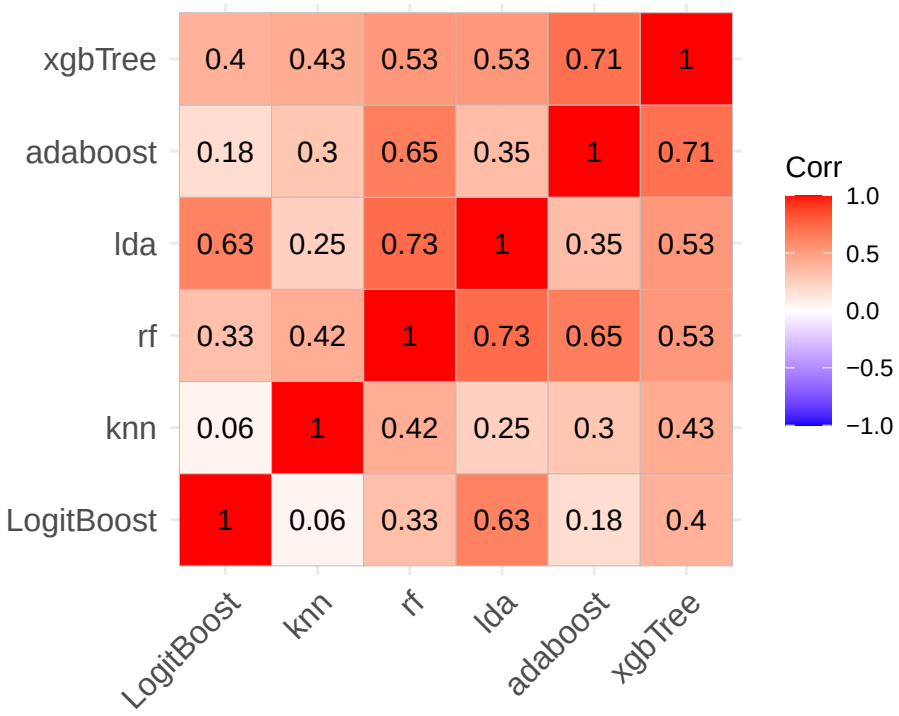
moyenne des probabilités de prédictions pour chaque classe et baser notre règle de décision là-dessus. Enfin la méthode dite de « stacking » que nous utiliserons, diffère sur deux points. La première différence est sur la façon dont on constitue notre nouvelle ensemble de données. La second différence est sur la manière dont les modèles sont combinés. Cette méthode permet d'extraire le meilleur de chaque modèle afin d'en constituer un plus performant et stable.

Le stacking va séparer notre échantillon en N sous échantillons. On va entrainer notre modèle sur N-1 sous-echantillons et prédire le N-ème sous-échantillon pour les N sous-échantillons avec chacun des modèles. On va ensuite combiner les probabilités obtenues à l'aide d'un algorithme (e.g. Regression Logistic). Nos prédictions sont ainsi assemblées à travers un nouvel algorithme. Celui-ci nous permet d'obtenir un modèle plus stable, plus robuste et plus précis. Nos modèles de bases sont : - eXtreme Gradient Boosting, Random Forest, Boosted Logistic Regression, Linear-Discriminant-Analysis, k-Nearest Neighbors, AdaBoost Classification Trees. Nous avons décidé d'utiliser des algorithmes avec des méthodes d'apprentissages différentes pour diversifier l'apprentissage de nos données et corriger les erreurs des modèles entre eux. Parmis tous nos modèles, on retrouve les forêts aléatoires qui permettent d'entraîner chaque arbres indépendamment, en utilisant un sous échantillon de données aléatoires sur un sous ensemble de l'espace en ne sélectionnant qu'une partie des variables. Cela permet au modèle d'éviter le sur-apprentissage et donne des résultats généralement plus robustes. On retrouve une méthode de noyau, K-nearest-neighbors, qui classifie les observations par un calcul de distance et sélectionne les K voisins les plus proches. On retrouve des modèles de boosting, qui est une méthode consistant à faire apprendre au modèle de ces erreurs à chaque étapes. En boosting, on combine de multiples "apprentisseurs" faibles pour former un "apprentisseur" fort. Ici, nous avons utilisé adaboost qui signifie "adaptive boosting". Celui-ci fait attention aux classes faussement prédites et adapte les poids de son algorithme pour corriger les erreurs. Chaque variables se voit attribuer un coefficient qui mesure sa contribution au modèle. Nous avons également utilisé une méthode de "gradient boosting", eXtreme Gradient Boosting Trees qui permet un paramétrage plus profond qu'un modèle classique. Ce modèle de boosting fonctionne de la même manière que les autres modèles mais à la différence d'adaboost celui ne modifie pas les poids à chaque étape d'apprentissage. Celui-ci va apprendre des erreurs de son prédécesseur dans un nouvel arbre de décisions.

Chacun de nos modèles a été paramétré spécifiquement. La validation croisée était un paramètre essentiel pour évaluer notre modèle, et s'assurer de sa stabilité à travers les répétitions. Etant donnée la taille de notre échantillon d'entraînement, le séparer en 5 sous échantillons semble être plus judicieux. On conserve 80 % des données pour l'apprentissage et 20 % pour l'évaluation. Pour confirmer la robustesse et la fiabilité de notre modèle nous avons pris soin de répéter 3 fois chaque itération de la validation croisée avec des échantillons non similaires aléatoirement sélectionnés. Globalement, nos modèles de type arbres de décisions étaient constitués de 150 ou 200 arbres avec pour chaque arbre un sous échantillons de 60% des variables et l'utilisation de 50% ou 75% des données uniquement. Le niveau d'apprentissage a été paramétré à 0.3 ou 0.4 pour éviter un apprentissage trop rapide ou trop lent. Combiné aux 150/200 arbres ainsi que la proportion des données et variables utilisées, nous avons décidé de paramétrer la profondeur de chaque arbre à 2,3 ou 5 maximum. Aucune régularisation n'a été introduite dans nos modèles.



Lors de la première étape dans l'élaboration de notre modèle nous avons comparé la performance de chacun individuellement.



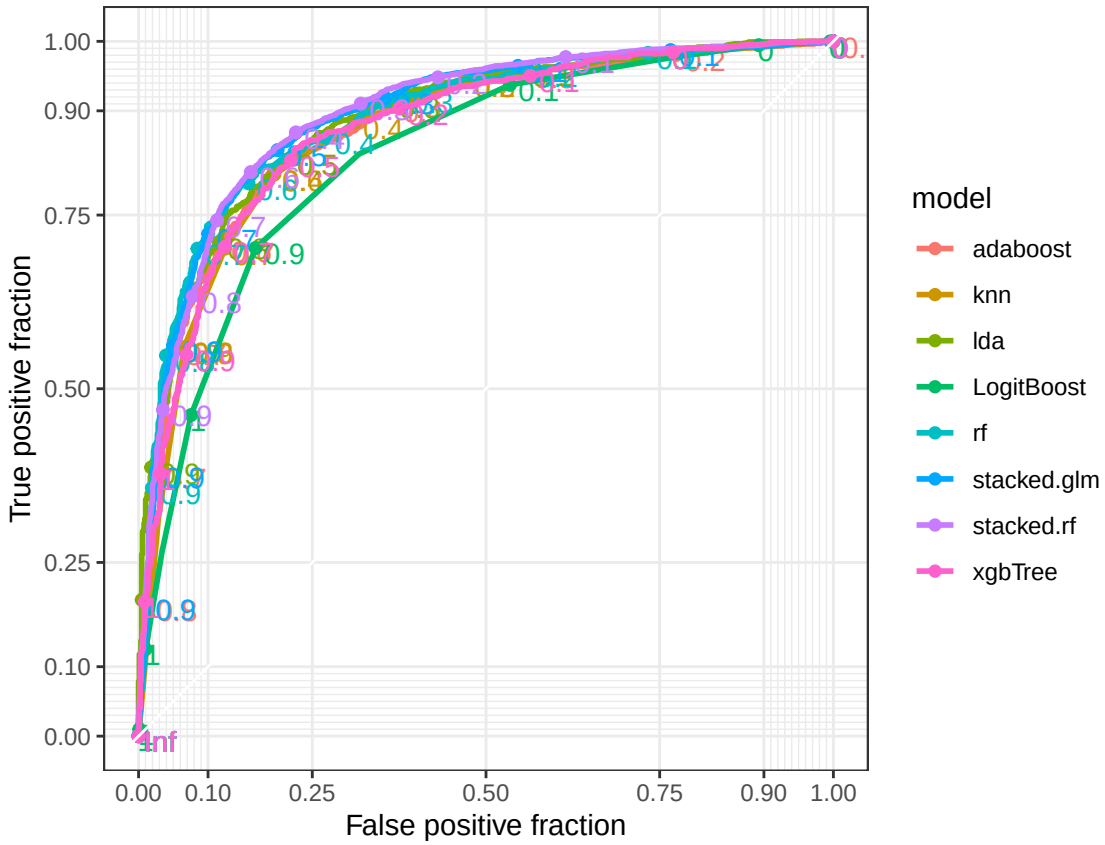

```
##        knn LogitBoost   adaboost        lda    xgbTree         rf
##  0.4104881  0.4328429  0.5313412  0.5803921  0.6013577  0.6084632
```

Finalement notre modèle final est un ensemble des modèles les plus performants avec les colinéarités de prédiction les plus faibles sur les données de test. On obtient ainsi un modèle plus stable et plus performant qui combine les résultats de plusieurs modèles.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Yes  No
##        Yes 118  25
##        No   15  84
##
##                Accuracy : 0.8347
##                  95% CI : (0.7818, 0.8792)
##     No Information Rate : 0.5496
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6634
##
```

```
##  Mcnemar's Test P-Value : 0.1547
##
##             Sensitivity : 0.8872
##             Specificity : 0.7706
##          Pos Pred Value : 0.8252
##          Neg Pred Value : 0.8485
##              Prevalence : 0.5496
##          Detection Rate : 0.4876
##    Detection Prevalence : 0.5909
##       Balanced Accuracy : 0.8289
##
##        'Positive' Class : Yes
##


## Confusion Matrix and Statistics
##
##           Reference
## Prediction Yes  No
##        Yes 121  28
##        No   12  81
##
##                Accuracy : 0.8347
##                  95% CI : (0.7818, 0.8792)
##     No Information Rate : 0.5496
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.6617
##
##  Mcnemar's Test P-Value : 0.01771
##
##             Sensitivity : 0.9098
##             Specificity : 0.7431
##          Pos Pred Value : 0.8121
##          Neg Pred Value : 0.8710
##              Prevalence : 0.5496
##          Detection Rate : 0.5000
##    Detection Prevalence : 0.6157
##       Balanced Accuracy : 0.8264
##
##        'Positive' Class : Yes
##
```



Nous avons, tout d'abord, comparé les courbes ROC de chaque modèle avec celles du modèle d'ensemble. On remarque bien la disparité de chaque modèle. L'amélioration de l'aire sous la courbe est un des premiers facteurs qui montre que le modèle choisi semble meilleur, on remarque aussi une trajectoire plus lisse et donc des prédictions plus stables et plus robustes. Par ailleurs, il est important de rappeler que l'objectif est d'obtenir un meilleur niveau de précision mais surtout de réduire le nombre de faux-négatif, autrement dit que l'on prédit comme sain alors qu'ils ne le sont pas. On remarque que notre modèle présente des mesures de précision performantes, mais pas encore parfaites.

Conclusion

Nous nous étions demandés si l'apprentissage automatique pourrait aider le dépistage des maladies cardiovasculaires. Les différentes modélisation nous donnent des taux de précision aux alentours de 81%. De plus, la méthode du stacking nous permet d'obtenir une précision de 83%. Nos méthodes sont plus précises que les différents travaux cités en introduction. On peut conclure qu'avec ce score de précision, notre algorithme est capable d'aider le personnel soignant et de se substituer potentiellement à l'avis médical. Ainsi un patient pourrait se faire déceler une maladie cardiovasculaire avec les seules informations de notre modèle. Le problème majeur demeure les faux négatifs qui serait diagnostiqué sain alors qu'ils sont atteints d'une maladie cardiovasculaire. L'intelligence montre, ici, ses limites puisqu'elle mettrait en danger certaines personnes. En regardant la matrice de confusion de la méthode de stacking, sur 242 personnes, 12 seraient mal détectées et encoureraient un risque d'attaque cardiaque sans le savoir. Les résultats sont, tout de même, significatif et on peut dire que l'intelligence artificielle pourrait apporter un soutien à la médecine. Le traitement de situation d'urgence par le personnel soignant pourrait être soulager par le machine learning en décelant les cas susceptibles d'être atteints. On peut certainement améliorer le modèle en traitant les données de manière plus approfondis. De plus, le matériel à notre disposition ne nous permettait pas d'exploiter le plein potentiel de toutes nos méthodes du fait d'une puissance requise trop élevé. Il serait alors intéressant d'élargir les paramètres utilisés pour étudier de meilleures combinaisons. Compte tenu des limites matérielles et temporelles, il est important de souligner les résultats obtenus qui témoignent de la puissance de notre modèle mais aussi de son potentiel à devenir meilleur.

References

https://www.researchgate.net/publication/222467943_Stacked_Generalization

https://www.who.int/cardiovascular_diseases/about_cvd/fr/

https://www.sciencesetavenir.fr/sante/canada-l-intelligence-artificielle-pour-traquer-le-coronavirus_141664

https://www.futura-sciences.com/sante/definitions/coeur-maladie-cardiovasculaire-15398/

http://www.francesoir.fr/societe-science-tech/insuffisants-cardiaques-du-machine-learning-pour-prevenir-les-deces-par-avc

Jahangiry, Leila et al. "Framingham risk score for estimation of 10-years of cardiovascular diseases risk in patients with metabolic syndrome." Journal of health, population, and nutrition vol. 36,1 36. 13 Nov. 2017, doi:10.1186/s41043-017-0114-0

Gennari, J.H. et. al., "Models of incremental concept formation", Artificial Intelligence, 40, 11–61

Detrano,R. et. al., "International application of a new probability algorithm for the diagnosis of coronary artery disease", American Journal of Cardiology, 304-310

Annexes

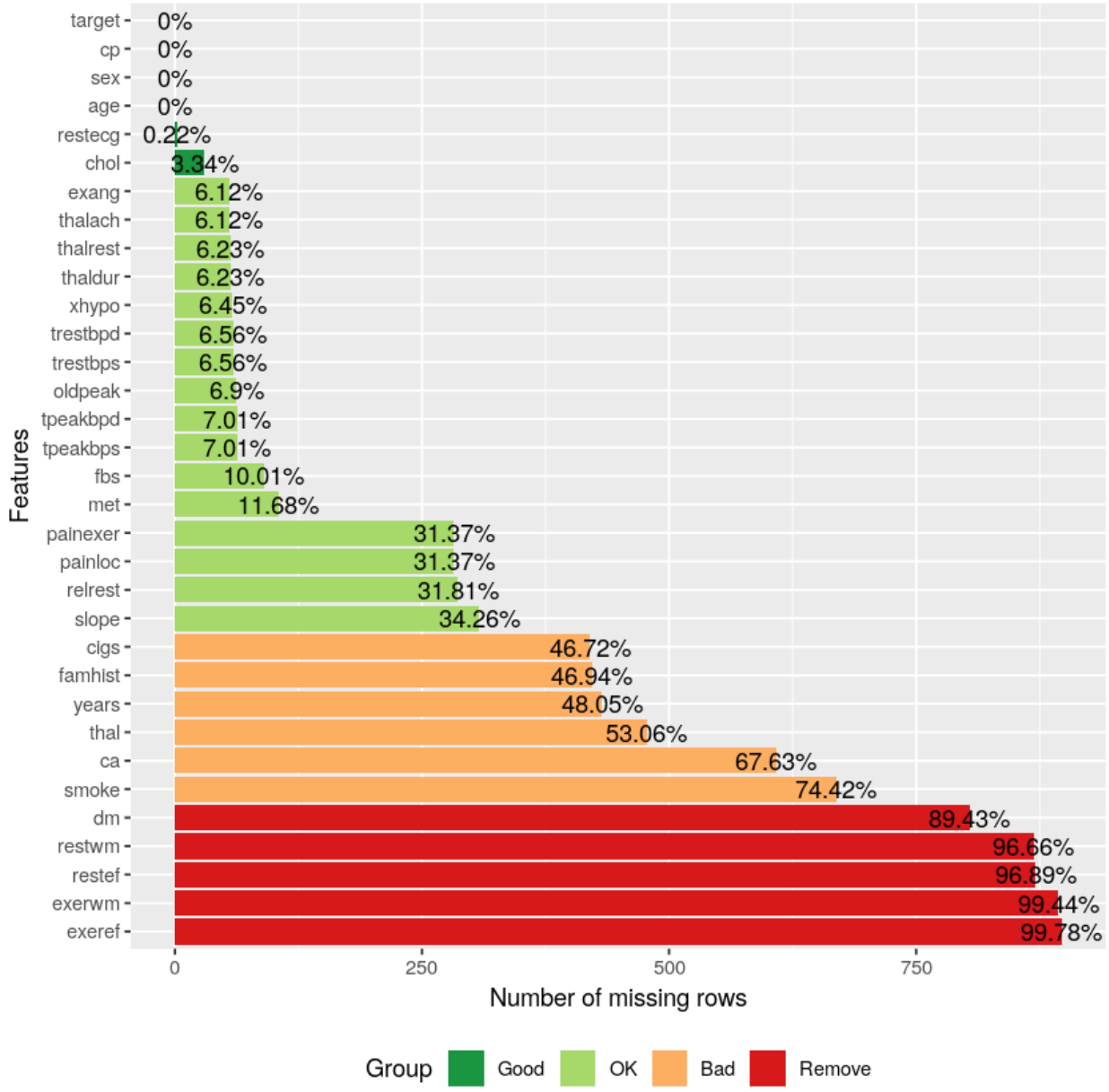


Figure 2: Caption for the picture.

Informations complémentaires (obtenir les données)

- Le fichier R (dans le drive) permet d'obtenir la base de données utilisée dans le fichier Rmarkdown. Pour exécuter le fichier R, aucun besoin d'enregistrer les bases sur son ordinateur au préalable. Le script du code se charge de récupérer les données directement sur internet puis les traitent et les enregistrent sous le nom indiqué à la fin du code.
- La base de données obtenue est tout de même placée dans un drive en ligne à cette adresse : https://drive.google.com/open?id=1vN78wF4pIlUrzw6ZaW7Ailz1kbEHeFME
- Les bases de données brutes utilisées dans le script R, pour constituer la base finale ce trouve à cette adresse : https://archive.ics.uci.edu/ml/datasets/Heart+Disease