Rapport pour la présentation de Biostatistiques - MVA 2022/2023

Dorian Gailhard dorian.gailhard@telecom-paris.fr

June 27, 2023

1 Introduction

Je faisais partie du même groupe que Samuel Bensoussan et Harith Proietti. Notre article s'intitulait "Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study" [lien] et s'intéressait aux chances de survie de patients admis en soin intensifs pour arrêt cardiaque, afin de déterminer leurs chances de survie dans les 24h. Les auteurs ont pour but de développer un meilleur indicateur - ne se servant que des données pouvant être obtenue dans les 24h après l'admission en soins intensifs - que ceux déjà existants, notamment APACHE III et ANZROD.

2 Résumé de l'article

2.1 Les données

Les auteurs s'appuient sur des données du ANZICS Institutional Data Access/Ethics Committee, qui concernent des patients d'Australie et de Nouvelle-Zélande. Ces données ont été recueillies en hôpital dans les 24h suivant l'admission des patients pour arrêt cardiaque, elles concernent 39 566 individus et contiennent entre autre la mort ou la survie du patient dans les 24h suivant l'admission, son âge, son sexe et les valeurs minimales et maximales de variables physiologiques mesurées dans les 24h (mais pas de données électrocardiographiques et échocardiographiques qui n'étaient pas disponibles).

Les données sont équilibrées entre 21 547 survivants et 18 019 patients décédés. Pour les données manquantes, les patients ont été séparés par catégories en fonction de leur âge, sur des tranches de 10 ans, et leurs données manquantes ont été remplacées par la moyenne de leur groupe pour les données continues ou le mode principal pour les données catégoriques. Les variables continues ont été normalisées.

2.2 L'entraînement des modèles

Les auteurs ont testés six modèles différents : une régression linéaire, un modèle Random Forest, un modèle Support Vector Classifier, un modèle gradient boosting, un modèle ensembliste (combinaison de modèles puis moyenne sur les résultats de ces modèles) et un réseau de neurones artificiels.

90% du jeu de données est consacré à l'entraînement, et 10% sert de jeu de test. L'entraînement a été fait avec une méthode 5-fold cross validation. Pour aider à l'interprétabilité du résultat, ils implémentent une méthode LIME (local interpretable model-agnostic explanation) qui consiste à entraîner un classificateur linéaire sur les prédictions du modèle plus complexe au voisinage d'un point - i.e. à approximer localement un modèle par une régression linéaire. Un exemple figure ci-après :

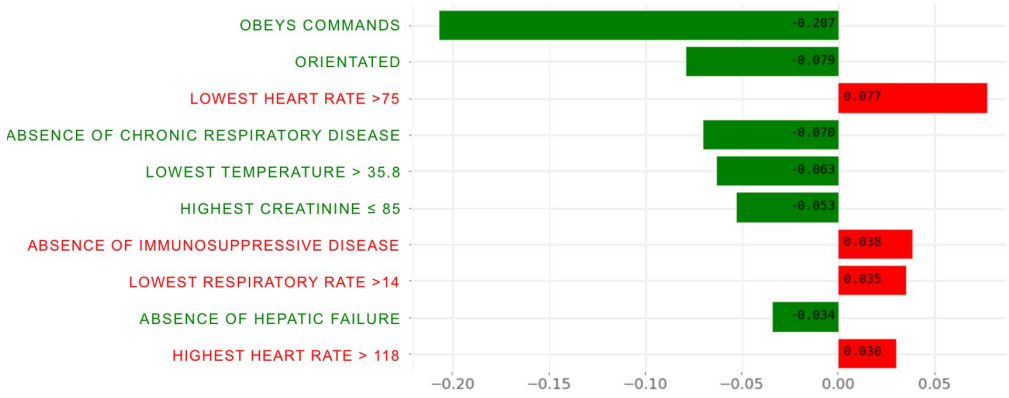

Figure 1: Exemple de linéarisation locale du modèle conçu par les chercheurs.

2.3 Résultats

Les techniques de Machine Learning (en particulier la méthode ensembliste) obtiennent de meilleurs résultats que les indicateurs APACHE III et ANZROD (AUROC de 0.87 pour la méthode ensembliste contre 0.80 et 0.81 pour ces deux indicateurs, les courbes sont inclues ci-après).

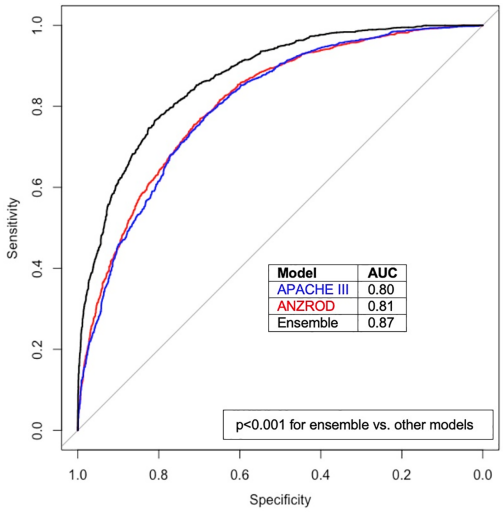

Figure 2: Comparaison des résultats entre la méthode ensembliste et les deux incateurs

Les méthodes de machine learning sont aussi plus équilibrées, là où le modèle APACHE surestimait la mortalité, en particulier chez les patients âgés, et le modèle ANZROD sous-estimait la

mortalité chez les patients jeunes. Elles bénéficient en outre d'une meilleure discrimination entre les patients, comme le montrent les courbes suivantes :

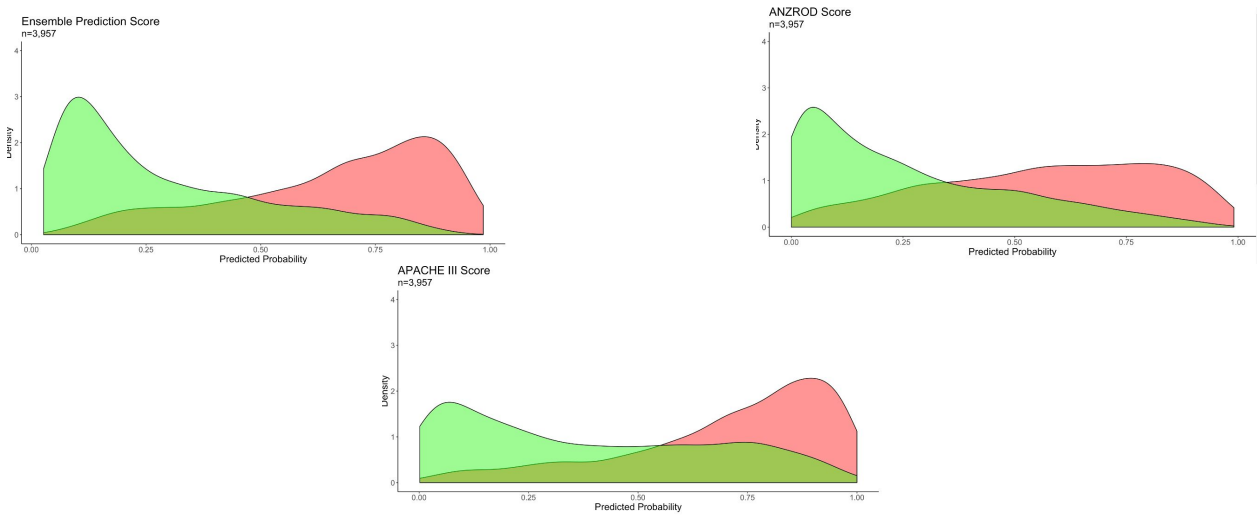

Figure 3: En abscisse figurent les probabilités de survie ou décès (selon l'issue réelle du patient) calculées par le modèle, en ordonnée figure la proportion de patients de chaque classe à la laquelle la probabilité a été attribué. En vert figure les survivants, en rouge les patients décédés. La méthode ensembliste a moins de recouvrement entre les classes que les deux indicateurs, et se rapproche plus du cas idéal qui est d'avoir un dirac en 0 pour les survivants et un dirac en 1 pour les patients décédés.

3 Remarques

3.1 Données

Premièrement, puisque les données concernent toutes des patients australiens et néo-zélandais, la question de la généralisation se pose : les résultats obtenus ici ne concernent-ils que cette zone géographique / ces groupes ethniques là ou le modèle peut-il être appliqué tel quel à d'autres patients. Ensuite vient évidemment la question de la présence de biais dans le jeu de données : les auteurs eux-mêmes remarquent que la taille de l'échantillon - 39 566 patients - reste limitée.

Le choix des variables peut aussi être discuté, les auteurs ayant par exemple choisi de ne conserver que les valeurs minimales et maximales des variables physiologiques disponibles. Un modèle plus complexe pouvant prendre en compte l'évolution de ces caractéristiques durant le passage en soins intensifs gagnerait probablement beaucoup en précision car ne garder que les valeurs extrémales est une grande perte d'information. Il est aussi dit en introduction que les mesures électrocardiographiques et échocardiographiques n'ont pas été incorporées car non-disponibles, il y a probablement là aussi quelque-chose à faire. Une remarque que font aussi les auteurs est que le modèle n'incorpore que les données disponibles "sur le tas" i.e. mesurables directement sur le patient pendant son passage à l'hôpital. Mon avis est que cette décision est bonne car la décision doit être prise rapidement (24h) et qu'un laps de temps si court ne permet probablement pas aux services de trouver des informations sur le patient, à moins de concevoir une base de données facilement accessible, mais je ne sais pas si cela est envisageable légalement. Néanmoins il est

évident qu'incorporer des données exhaustives sur le patient collectées tout au long de sa vie à chaque contact avec le monde hospitalier serait un grand gain d'information qui conduirait à une hausse significative de la performance des modèles.

Il y a enfin les choix liés aux données manquantes (qui concerne toujours des données en entrée, jamais la survie ou la mort d'un patient). Les auteurs les remplacent par la moyenne ou le mode majoritaire du groupe d'âge dans lequel se situe le patient (sorte d'imputation de la moyenne, qui conduit à sous-estimer les variances). Cette technique assez simple pourrait être remplacée par d'autres techniques vues dans le cours, notamment une imputation par régression (qui néanmoins surestime les corrélationss entre les variables) ou comme le soulignent les auteurs, une imputation multiple par équations chaînées (néanmoins très coûteux en temps machine, et il y a ici plusieurs dizaines de milliers de patients).

3.2 Entraînement

Une première remarque concerne le choix des différents modèles testés : une régression linéaire, un modèle Random Forest, un modèle Support Vector Classifier, un modèle gradient boosting, un modèle ensembliste et un réseau de neurones artificiels. Il ne semble pas y avoir de cohérence ou de raisonnement derrière ces choix, un peu comme si les auteurs avaient voulu tester tout ce qui peut se faire en machine learning, sans avoir d'a priori sur la modélisation du problème. Il n'y a d'ailleurs pas de spécifications des différents paramètres utilisés pour l'entraînement ou du choix de l'architecture du réseau de neurone. Cela ressemble à une approche assez "naïve" du machine learning, où l'on mélange un jeu de données avec des architectures en espérant que cela marche. Il y a sûrement beaucoup à gagner en pensant plus soigneusement le choix de l'architecture et des algorithmes utilisés, en particulier pour permettre l'incorporation de plusieurs relevés des variables physiologiques et non seulement les valeurs extrémales.

Quelque-chose qui me vient aussi à l'esprit est que le modèle renvoie un unique résultat provenant de paramètres fixés. On peut probablement beaucoup tirer parti d'un modèle probabiliste comme des réseaux de neurones artificiels bayésiens, qui ne possèdent pas de paramètres fixés mais de distributions de probabilités de ces paramètres, estimées en fonction des données. Le résultat final du modèle est alors l'intégration des prédictions de chaque ensemble de paramètres possible, pondérées par leurs probabilités. Cela donnerait un modèle beaucoup plus robuste qui incorporerait l'aléa des données utilisées pour l'entraînement au lieu de les considérer comme exactes. Une méthode plus simple pour obtenir un score de confiance du modèle serait de discrétiser les probabilités de mortalité (par exemple en 11 classes $[0, 0.1], \cdots, [0.9, 1]$). et de concevoir un modèle qui classifierait les patients dans chacune de ces classes. La probabilité de chaque classe pourrait alors être comprise comme un score de confiance et aiderait sûrement à éviter les faux positifs ou négatifs.

Un dernier point plus anecdotique est le choix d'avoir consacré 90% des données à l'entraînement et seulement 10% au jeu de test, comparé aux traditionnels 75% et 25%, ce qui laisse seulement 4000 échantillons pour le test, ce qui pourrait être considéré comme peu.

3.3 Modèle LIME pour l'interprétabilité

Je trouve personnellement que l'utilité du modèle LIME est exagérée. Certes il s'agit d'un premier pas pour pouvoir expliquer les résultats des modèles de machine learning, mais je ne

pense pas qu'approximer par une régression linéaire de manière locale puisse vraiment aider un mèdecin à prendre une décision. En tant qu'approximation locale, on perd énormément de la dynamique globale, notamment en cas d'existence de plusieurs régimes de fonctionnement. On pourrait imaginer un cas où plusieurs variables serait corrélées positivement aux chances de survie dans un régime particulier, et corrélées négativement dans un autre régime, où le régime serait déterminé par d'autres variables évoluant lentement, et cette dépendance au régime serait alors invisible localement et n'apparaîtrait pas sur le modèle LIME. D'autre part, comme l'approximation réalisée est une régression linéaire, il n'y a pas de relation entre les variables, elles sont traitées comme indépendantes, et encore une fois on perd à ne pas tirer parti d'un contexte et à les traiter individuellement.

Cette méthode d'interprétabilité peut certainement renforcer la confiance d'un mèdecin dans le résultat de l'algorithme mais reste plus un indicateur qu'une véritable explication, et le modèle est si simple que j'ai du mal à croire qu'un médecin apprendrait quelque-chose à lire son "raisonnement". Cela peut à la limite servir de résumé des caractéristiques principales du patient mais je ne pense pas qu'il soit possible d'en tirer autre chose qu'une légère augmentation de la confiance en l'algorithme.