Apports des méthodes de Machine Learning et de Deep Learning dans la prédiction des durées de séjours hospitalières et des ré-hospitalisations

Franck Jaotombo

Apports des méthodes de Machine Learning et de Deep Learning dans la prédiction des durées de séjours hospitalières et des ré-hospitalisations

*****

Contributions of Machine Learning and Deep Learning methods in predicting hospital Length of Stay and Readmissions

NNT/NL : 0000AIXM0000/000ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 08 Décembre 2022 par

Franck JAOTOMBO

Apports des méthodes de Machine Learning et de Deep Learning dans la prédiction des durées de séjours hospitalières et des ré-hospitalisations

Discipline
Sciences de la Vie et de la Santé

Spécialité
Recherche Clinique et Santé Publique

École doctorale
ED 62

Laboratoire/Partenaires de recherche
Centre d'Etude et de Recherche sur les Services de Santé et la Qualité de Vie (CEReSS)

Composition du jury

Antoine DUCLOS CHU – Hospices Civils de Lyon	Rapporteur
Marianne CLAUSEL Institut Elie Cartan	Rapporteuse
Patrice FRANCOIS CHU La Tronche Grenoble	Examinateur
Laurent BOYER Aix-Marseille Université	Directeur de thèse
Badih GHATTAS Aix-Marseille Université	Co-directeur de thèse
Cyrille COLIN Pôle IMER - Hospices Civils de Lyon	Président du Jury

Affidavit

Je soussigné, Franck JAOTOMBO, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Laurent BOYER et Badih GHATTAS, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille, le 1er octobre 2022

Affidavit

I, undersigned, Franck JAOTOMBO, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Laurent BOYER et Badih GHATTAS, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Place Marseille, date 1er octobre 2022

Liste de publications et participation aux conférences

1) Liste des publications[1] réalisées dans le cadre du projet de thèse :

1. Jaotombo, F., Pauly, V., Auquier, P., Orleans, V., Boucekine, M., Fond, G., ... & Boyer, L. (2020). *Machine-learning prediction of unplanned 30-day rehospitalization using the French hospital medico-administrative database*. Medicine, 99(49)
2. Jaotombo, F., Pauly, V., Auquier, P., Orleans, V., Boucekine, M., Fond, G., ... & Boyer, L. *Machine-learning prediction for hospital length of stay using a French medico-administrative database*. Journal of Market Access & Health Policy, (Accepté le 16/11/2022)
3. Jaotombo, F.¶, Adorni L.¶, Ghattas, B., Boyer, L. *Predicting hospital readmission with machine learning using the MIMIC III discharge notes: the impact of vectorization* (Révision pour resoumission)
4. Jaotombo, F.¶, Adorni L.¶, Ghattas, B., Boyer, L. *Predicting prolonged Hospital Length of Stay by finding the best trade-off between performance and interpretability using structured and unstructured data* (Article en cours de relecture pour soumission)

…

2) Participation aux conférences[2] et écoles d'été au cours de la période de thèse :

1. Prédiction à partir de données textuelles en utilisant du Machine Learning
   Journée de Recherche : "Recherche quantitative inductive en GRH : enjeux et méthodes " – 8 Juillet 2022 IAE Paris

[1] Cette liste comprend les articles publiés, les articles soumis à publication et les articles en préparation ainsi que les livres, chapitres de livre et/ou toutes formes de valorisation des résultats des travaux propres à la discipline du projet de thèse. La référence aux publications doit suivre les règles standards de bibliographie et doit être conforme à la charte des publications d'AMU.

[2] Le terme « conférence » est générique. Il désigne à la fois « conférence », « congrès », « workshop », « colloques », « rencontres nationales et/ou internationales » … etc.
Indiquer si vous avez fait une présentation orale ou sous forme de poster.

Remerciements

Avant toute chose, je remercie sincèrement les Professeurs Marianne CLAUSEL, Antoine DUCLOS, Patrice FRANCOIS et Cyrille COLIN d'avoir accepté d'évaluer et de siéger dans ce jury de thèse.

Toute ma gratitude va tout naturellement à Badih GHATTAS, mon co-directeur de thèse qui m'a fait confiance dans une période délicate de transition personnelle et professionnelle et m'a présenté à Laurent BOYER qui a été pour moi un véritable « benefactor », non seulement en m'accueillant au sein de l'APHM mais aussi en acceptant de diriger cette thèse au sein du labo de Santé Publique.

Je remercie vivement Pascal AUQUIER, notre directeur d'unité et toute l'équipe de mon laboratoire d'accueil, ainsi que tous les ingénieurs et personnel du Département d'Information Médicale pour leurs inestimables contributions à ce travail, par leurs retours constructifs, la collecte des données et la mise à disposition d'un environnement de travail positif et bienveillant. Je salue tout particulièrement Vanessa PAULY, Veronica ORLEANS, Gilbert PIRONTI, Clément FRABOULET et François ANTONINI qui m'ont chacun d'une façon ou une autre initié à l'informatique médicale et aux données de santé.

Un travail de thèse a bien souvent des victimes collatérales en termes de temps indisponibles à la famille. Je ne saurais jamais remercier mes proches pour leur soutien inconditionnel lorsque je m'enferme dans mon bureau pour traiter de tous ces éléments abstraits, conceptuels ou techniques qui leur prive de ma présence.

Une mention spéciale pour mes amis et anciens étudiants Joël SOLLARI et Luca ADORNI qui ont été des soutiens sans faille pour discuter, réfléchir et coder les différents algorithmes utilisés dans cette thèse.

Je suis aussi largement débiteur de mes amis proches qui se reconnaîtront ici par leurs soutiens sans faille dans les moments joyeux tout autant que difficiles. Merci mes amis… !

Enfin je remercie mes collègues de l'emlyon pour leur soutien, leur bienveillance et leur compréhension – ils se reconnaîtront !

REMERCIEMENTS 6

RESUME 12

Mots clés 13

ABSTRACT 14

Key words 16

PARTIE 1 – CADRE THEORIQUE 17

QUALITÉ DES SOINS : UNE INTRODUCTION 18

Dimensions de la Qualité des Soins (Indicateurs de la Qualité des Soins) 19

1- Efficacité 19

Définitions : 19

Quelques indicateurs : 19

2- Sécurité 20

Définitions : 20

Quelques indicateurs : 20

3- Accessibilité 20

Définitions : 20

Indicateurs : 20

4- Réactivité 21

Définitions : 21

Indicateurs : 21

5- Efficience 21

Définitions : 21

Indicateurs : 21

Indicateurs retenus dans cette thèse 21

Durée de Séjour 22

Définition de la durée de séjour prolongée (de façon générale) 23

Les facteurs déterminants 23

Définition du séjour prolongé en réanimation 24

Particularités du LOS en réanimation 24

En conclusion 25

Réhospitalisation (Réadmission) 25

1- Les Réhospitalisations dans leur contexte international 26

I – Les USA 26

II – Autres pays 26

III – En France 27

2- Prévenir la réhospitalisation sous 30 jours 27

Les interventions pré décharge (c-à-d en préparation de la sortie après le séjour hospitalier). 27

Les interventions post décharge incluent : 27

Pour faire le lien (la transition) entre ces deux périodes : 27

3- Une brève synthèse de la littérature sur les modèles de prédiction de la RH30 27

Les prédicteurs de la RH30 27

Le pouvoir discriminant des modèles (performances des classifieurs mesurées par le C-statistic ou ROC AUC) . 28

MACHINE LEARNING : UN RÉSUMÉ CONCEPTUEL 29

Le Machine Learning en Bref 29

1- Apprentissage supervisé 29

Prédiction 29

Inférence & Interprétabilité 29

2- Apprentissage non supervisé 30

Quelques spécificités des sciences de la santé. 30

Modèles de Machine Learning classiques (supervisés) 30

1- KNN ou K Nearest Neighbors 30

Principe: 30

Interprétabilité 31

2- Régression Logistique Pénalisée 31

Principe : 31

Interprétabilité : 32
3- Arbres de décision de type CART (Classification and Regression Trees) 32
Principe : 32
Elagage par complexité (Complexity pruning) 34
Interprétabilité : 34
4- Random Forest 35
Principe : 35
Interprétabilité : 35
Famille de Boosting *35*
Principe du Boosting : 35
5- Gradient Boosting 36
6- XGBOOST (eXtreme Gradient Boosting) 36
Interprétabilité : 37
7- Light GBM 37
Interprétabilité : 38
8- Catboost 38
Traitement des variables catégorielles. 38
Traitement des biais de gradient. 39
Inteprétabilité : 39
9- Réseaux de Neurones Feed Forward (Perceptrons Multicouches) 39
Principe : 39
Interprétabilité : 40
Limites des modèles classiques *41*
LES TRANSFORMERS (SELON L'APPROCHE HUGGING FACE) 42
Partie Encodeur *42*
1- Tokenization 43
2- Encodage 43
3- La couche « Multihead-Self-Attention » 44
4- Scaled dot-product Attention 44
5- Multi-head Attention 44
6- Couches Réseau de Neurones + Normalization 45
7- Couche embedding de position 46
8- Interprétabilité 46
LE PROBLÈME DE L'INTERPRÉTABILITÉ EN MACHINE LEARNING 47
Une classification des méthodes d'interprétabilité *47*
1- Intrinsèque ou post-hoc ? 47
Les méthodes intrinsèques 47
2- Spécifique au modèle ou indépendant du modèle ? 47
3- Local ou global ? 48
Exemples de modèles intrinsèquement interprétables *48*
1- Les modèles linéaires (généralisés) 48
2- Les arbres de décision 49
Exemples de méthodes globales indépendantes des modèles *49*
1- Importance de variables par permutation (Permutation feature importance : PFI) 49
2- Modèle global de substitution (Global surrogate model) 50
Exemple de méthodes locales indépendantes des modèles *50*
1- Modèle de substitution locale (LIME) 50
Remarque : 51
2- Shapley Values 51
Remarques : 52
3- SHAP (Shapley Addition exPlanations) 52
KernelSHAP (LIME Linéaire + Shapley Values) 53
Autres variantes de SHAP 53
4- SHAP pour l'explicabilité globale 54
Importance des Variables (SHAP Feature Importance) 54
Shap Summary Plot 54
Remarques: 55
PARTIE 2 – METHODOLOGIE 56

PRÉSENTATION DES DONNÉES 57
Jeu de données de l'APHM *57*
Les variables à expliquer dans le jeu de données de l'APHM 58
La base de données MIMIC III *61*
Les Variables MIMIC III retenues 61
Enjeux autour des données déséquilibrées *65*
1- L'enjeu de la distribution des classes 65
2- Le problème de l'évaluation de la performance (métriques) 66
3- Les métriques utilisées en classification 67
Les métriques dépendantes du seuil de classification 67
Les métriques indépendantes du seuil de classifications 67
ARTICLE 1 69
MACHINE-LEARNING PREDICTION OF UNPLANNED 30-DAY REHOSPITALIZATION USING THE FRENCH HOSPITAL MEDICO-ADMINISTRATIVE DATABASE 69
Contexte *69*
Données déséquilibrées *69*
Résumé *71*
Machine-learning prediction of unplanned 30-day rehospitalization using the French hospital medico-administrative database *72*
Abstract *74*
1- Introduction 75
2- Methods 75
Study design 75
Study setting and inclusion criteria 76
Study outcome 76
Collected data 76
Statistical models 76
Statistical analyses 77
3- Results 78
Rates of unplanned 30-day all-cause rehospitalization 78
Predictive model performance 78
Variable importance 78
4- Discussion 78
Conclusion 80
References *81*
SYNTHÈSE DES CONCLUSIONS – ARTICLE 1 94
ARTICLE 2 96
MACHINE-LEARNING PREDICTION FOR HOSPITAL LENGTH OF STAY USING A FRENCH MEDICO-ADMINISTRATIVE DATABASE 96
Contexte *96*
1- Algorithmes et modèles : 97
2- Entraînements des modèles : 97
Résumé *98*
Machine-learning prediction for hospital length of stay using a French medico-administrative database . *99*
Abstract *100*
1- Introduction 101
2- Methods 101
Study design 101
Study setting and inclusion criteria 102
Study outcomes 102
Collected data 102
Statistical models 102
Statistical analyses 103
3- Results 104
Characteristics of the population 104
Factors associated with LOS 104
Predictive model performance 105
Variable importance 105

4- Discussion 105
Conclusion 107
Synthèse des Conclusions – Article 2 122
ARTICLE 3 125
PREDICTING HOSPITAL READMISSION WITH MACHINE LEARNING USING THE MIMIC III DISCHARGE NOTES: THE IMPACT OF VECTORIZATION 125
Contexte 125
1- La vectorisation des données textuelles 125
2- Bag of Words (BOW) 126
3- LSA : Latent Semantic Analysis 127
4- LDA : Latent Dirichlet Allocation 128
5- Modèles de Machine Learning 130
6- Critère d'inclusion 130
7- En résumé : 130
Résumé 131
Predicting hospital readmission with machine learning using the MIMIC III discharge notes: the impact of vectorization 132
Abstract 133
1- Introduction 134
A general international & historical background on the 30-day readmission indicator (RH30) 134
Predicting 30-days readmission from clinical notes 135
2- Materials and methods 136
Data 136
Inclusion criteria 136
Handling duplicates 136
Outcome: 30-days readmissions 136
Predictors 136
Text cleaning 137
Text vectorization 137
Statistical models 138
Statistical analyses 139
3- Results 139
Before tuning: on the default parameters 139
Tuned models: 141
Comparing variable importance through permutation importance method 142
Topic modeling 142
4- Discussion 142
Synthèse des Conclusions – Article 3 153
1- La performance des données textuelles tabulaires vectorisées 153
2- Un gain significatif en interprétabilité 153
3- Complementarité des informations avec les données structurées 154
PARTIE 3 – FUSION DE DONNEES STRUCTUREES ET NON STRUCTUREES 155
PRÉDIRE (ET EXPLIQUER) LA DURÉE DE SÉJOUR À PARTIR DE DONNÉES STRUCTURÉES, NON STRUCTURÉES ET MIXTES DU MIMIC III 156
Contexte 156
Les données 157
1- Critères d'inclusion 157
2- La variable à expliquer 157
3- La variable d'intérêt binarisée (los_cat) 158
4- Les prédicteurs 158
Préparation et Fusion des Données 159
1- Package AutoGluon pour les données structurées 159
Analyse des données structurées tabulaires 161
Permutation importance 162
Package AutoGluon pour les données non-structurées textuelles 165
1- Phase 1 : Aucun Pré-traitement des Données 165

2- Phase 2 : Pré-traitement « léger » des Données ... 169
3- Phase 3 : Pré-traitement « plus lourd » des données ... 170
Package AutoGluon pour les données mixtes : structurées + textuelles ... 171
1- Mécanisme de Fusion de Multimodal AutoGluon (MAG) ... 171
2- Processus d'agrégation : ... 172
3- Processus Ensembliste ... 173
AutoGluon Multimodal via TabularPredictor ... 174
1- Phase 1 : Aucun prétraitement des données ... 174
2- Phase 2 : Prétraitement « léger » des données textuelles ... 176
3- Phase 3 : Prétraitement « plus lourd » des données textuelles ... 178
Conclusion de la partie : ... 180
AutoGluon TabularPredictor avec fusion LDA ... 181
1- Vectorisation TF ... 181
Racinisation ... 181
Lemmatisation ... 182
2- Vectorisation binaire ... 184
Racinisation ... 184
Lemmatisation ... 184
3- Vectorisation TFIDF ... 186
Racinisation ... 187
Lemmatisation ... 187
Mise en Perspective des Résultats ... 189

PARTIE 4 – DISCUSSION ET CONCLUSION ... 192

SYNTHÈSE ET PERSPECTIVES GLOBALES ... 193
Concernant les aspects techniques ... 193
1- Pertinence du Machine Learning (ML) en santé publique ... 193
2- Boîtes noires ? ... 193
3- Le compromis performance-explicabilité ... 193
4- L'enjeu des données déséquilibrées ... 194
5- Partitionnements et rééchantillonnages ... 194
6- Performance selon les données et fusion des données ... 194
Données tabulaires ... 194
Données textuelles ... 195
Données mixtes fusionnées ... 195
7- Les principales conclusions que l'on tire : ... 195
Concernant l'aspect santé publique ... 196
1- La réadmission à 30 jours (RH30) ... 196
2- La durée de séjour (LOS) ... 197
Limites et Perspectives de Recherche ... 199
RÉFÉRENCES ... 200

Résumé

Cette thèse traite de la prédiction des durées de séjours hospitalières et de réhospitalisations à partir de méthodes de Machine Learning et de Deep Learning appliquées à l'ensemble des données hospitalières exploitables (structurées et non structurées), largement sous-utilisées à l'heure actuelle. La prédiction des durées de séjour hospitalières est un enjeu organisationnel important pour améliorer l'accès, la qualité et l'efficience des soins. La prévention des réhospitalisations constitue un enjeu important pour la qualité et la sécurité des prises en charge du patient hospitalisé ; les réhospitalisations ont un impact négatif sur la qualité de vie des patients et de leurs proches en plus des risques iatrogènes inhérents à toute hospitalisation, et alourdissent le coût de la prise en charge.

La démarche suivie au cours de cette thèse a consisté à utiliser des méthodes de Machine Learning et de Deep Learning pour rechercher le meilleur compromis possible entre performance et interprétabilité. Nous démontrons que les données structurées bien choisies permettent d'obtenir une très bonne performance (ROC AUC variant de 0.789 à 0.972 sur nos données), avec une interprétabilité satisfaisante mais peu spécifique. Les données textuelles seules ont une performance plus que satisfaisante (ROC AUC variant de 0.723 à 0.848), mais avec une interprétabilité beaucoup plus spécifique et détaillée sur les séjours à risque. Le meilleur compromis entre performance et interprétabilité est donné par les données mixtes, avec d'un côté une très bonne performance (ROC AUC variant entre 0.938 et 0.966) et simultanément des descriptions très détaillées des séjours à risques.

Ce document de thèse est constitué de 4 parties réparties comme suit.

1- La première partie couvre le cadre théorique de la thèse contenant un rappel du contexte et de la revue de littérature sur la qualité des soins dont la réadmission à 30 jours et la durée de séjour (prolongée) sont des indicateurs importants. Elle enchaîne ensuite sur un rappel conceptuel des différents algorithmes de machine learning utilisés ainsi que l'enjeu autour du compromis entre la performance des modèles et l'interprétabilité des résultats.
2- La deuxième partie traite de la méthodologie explicitant les données et les variables utilisées, ainsi que le choix des métriques. S'en suivent les trois articles qui ont été publiés ou soumis dans le cadre de la thèse, chacun précédé d'une contextualisation avec parfois un rappel des enjeux méthodologiques ou techniques, puis suivi d'une synthèse des conclusions.
    a. Le premier article prédit la réhospitalisation à 30 jours en utilisant les données structurées provenant de l'Assistance Publique – Hôpitaux de Marseille (APHM) basées sur les données issues du Programme de Médicalisation des Systèmes d'Information (PMSI), et des méthodes classiques de Machine Learning.
    b. Le deuxième article prédit les séjours prolongés, toujours avec les données de l'APHM, en utilisant également des méthodes classiques de Machine Learning.
    c. Le troisième article prédit la réadmission à 30 jours en soins intensifs, en utilisant les données publiques MIMIC III du Beth Israel Deaconess Medical Center de Boston. Cette étude a la particularité de réaliser ses prédictions exclusivement avec des données textuelles.
3- La troisième partie cherche à dépasser les limites de ces premières études en prédisant le séjour prolongé en soins intensifs, en fusionnant les données structurées du MIMIC III avec des données textuelles non structurées. Pour ce faire, différentes méthodes de fusion sont expérimentées :

a. Utiliser les données tabulaires comme principal support et inclure une représentation tabulaire des données, via une vectorisation (« embedding ») des documents textuels par Transformers
b. Utiliser les données tabulaires comme principal support et inclure une représentation tabulaire des données par une vectorisation Bag of Words + LDA (Latent Dirichlet Allocation) ;

4- La quatrième partie discute des résultats dans leur globalité, puis conclut la thèse avec des considérations à la fois techniques, et de santé publique.

Mots clés

Machine Learning, Transformers, Fusion de données, Latent Dirichlet Allocation, Réhospitalisation, Durée de Séjour, Qualité des soins

Abstract

This thesis is centered on predicting hospital length of stay and readmissions using Machine Learning and Deep Learning methods applied to all usable hospital data (structured and unstructured), still largely underused. Predicting length of stay is an important organizational issue for improving access, quality, and efficiency of care. Preventing readmission is an important step in enhancing the quality and safety of the hospitalized patient's care. In addition to the iatrogenic risks inherent to any hospitalization, readmissions increase the cost of care, and engender negative impacts on the patients' quality of life as well as on their relatives.

The approach adopted in this thesis is based on using Machine Learning and Deep Learning methods to find the best possible trade-off between performance and interpretability. We demonstrate that with a well-chosen structured data, one can obtain a very good performance (ROC AUC varying from 0.789 to 0.972 on our data), with a satisfactory but not very specific interpretability. Textual data alone provide a somewhat satisfactory performance (ROC AUC varying from 0.723 to 0.848), but with a much more specific and detailed interpretability on risky stays. The best compromise between performance and interpretability is given by mixed data, with on the one hand a very good performance (ROC AUC varying between 0.938 and 0.966) along with a very detailed description of risky stays.

This thesis contains 4 main parts distributed as following:

1- The first part covers the theoretical framework of the thesis with a reminder of the context and a literature review on the quality of care, of which 30-days readmission and (prolonged) length of stay are important indicators. It then continues with a conceptual reminder of the different machine learning algorithms used as well as the issue around the trade-off between the performance of the models and the interpretability of the results.

2- The second part deals with the methodology explaining the data and the variables used, as well as the choice of metrics. This is followed by the three articles that were published or submitted as part of the thesis, each preceded by a contextualization with – sometimes – a reminder of the methodological or technical issues, then a summary of the conclusions.

a. The first article predicts 30-days readmissions using structured data from the "Assistance Publique – Hôpitaux de Marseille" (APHM) based on the data from the "Programme de Médicalisation des Systèmes d'Information" (PMSI), using classical Machine Learning methods.
b. The second article predicts prolonged LOS, again with APHM data, also using classical Machine Learning methods.
c. The third article predicts 30-day readmission in intensive care units, using the public MIMIC III data from the Boston Beth Israel Deaconess Medical Center. This study is specifically making its predictions from textual data alone.

3- The third part seeks to go beyond the limits of these first studies by predicting the prolonged stays in intensive care, by merging structured with unstructured textual data of the MIMIC III. To do this, different fusion methods are tested:

a. Use tabular data as the main support and include a tabular representation of the data, via an embedding of textual documents by Transformers;

b. Use tabular data as the main support and include a tabular representation of the data by vectorization Bag of Words + LDA (Latent Dirichlet Allocation).

4- The fourth part discusses the results in their entirety, then concludes the thesis with both technical and public health considerations.

Key words

Machine Learning, Transformers, Data Fusion, Latent Dirichlet Allocation, Readmission, Length of Stay, Quality of Care

Partie 1 – Cadre Théorique

Qualité des Soins : une introduction

Donabedian (2005) définit la qualité comme étant des soins qui « maximisent le bien-être des patients après avoir pris en compte le rapport bénéfice/risque à chaque étape du processus de soins ».

Pour l'OMS (Roemer et al., 1988, p. 82)) il s'agit de la capacité de « garantir à chaque patient l'assortiment d'actes thérapeutiques… lui assurant le meilleur résultat en termes de santé, conformément à l'état actuel de la science, au meilleur coût pour le même résultat, au moindre risque iatrogénique, pour sa plus grande satisfaction en termes de procédures, résultats, contacts humains… ».

Quant à Institut de Médecine des Etats Unis (IOM) (2001), c'est « la capacité des services de santé destinés aux individus et aux populations d'augmenter la probabilité d'atteindre les résultats de santé souhaités, en conformité avec les connaissances professionnelles du moment ». Il y a plusieurs implications à cette définition de l'IOM (Or & Com-Ruelle, 2008) :

- Un large éventail de services de santé (incluant les maladies mentales) ;
- L'importance d'envisager plusieurs perspectives (individuelles, collectives, etc.) ;
- Elle s'applique à différents services qui touchent à la santé (médicaux et paramédicaux, différents types d'établissements) ;
- L'amélioration de soins fournis par les professionnels de santé ET le système de soin en entier, ce qui entraîne la probabilité d'atteindre les résultats souhaités – mais ce n'est pas garanti ; et contribue en tous les cas à plus de bien que de mal.
- Les connaissances continuent cependant à évoluer et les professionnels sont supposés se maintenir à jour.

Cette définition est reprise par l'OMS[3] avec une légère variation, puis un développement : « La qualité des soins est la mesure dans laquelle les services de santé destinés aux individus et aux populations augmentent la probabilité de parvenir à l'état de santé souhaité. Elle dépend de connaissances professionnelles fondées sur des bases factuelles et est essentielle pour instaurer la couverture sanitaire universelle. »

« Les soins de santé de qualité peuvent être définis de bien des manières mais il est de plus en plus admis dans le monde qu'ils devraient être :

- **efficaces**, c'est-à-dire reposer sur des bases factuelles et être dispensés à ceux qui en ont besoin ;
- **sûrs**, c'est-à-dire éviter de porter préjudice à ceux qui en bénéficient ;
- **centrés sur la personne**, c'est-à-dire correspondre aux préférences, aux besoins et aux valeurs de chaque individu.

Afin de tirer profit des bénéfices de soins de qualité, ceux-ci doivent être :

- **dispensés en temps utile**, c'est-à-dire que les délais d'attente et parfois les retards préjudiciables tant pour les bénéficiaires que pour les prestataires doivent être réduits ;
- **équitables**, c'est-à-dire que la qualité des soins ne doit pas varier en fonction du genre, de l'origine ethnique, de la situation géographique et du statut socioéconomique ;

[3] https://www.who.int/fr/health-topics/quality-of-care#tab=tab_1 (au 08/06/2022)

- **intégrés**, c'est-à-dire que l'ensemble des services de santé doivent être disponibles à toutes les étapes de la vie ;
- **efficients**, c'est-à-dire que les ressources disponibles doivent être exploitées au maximum en évitant le gaspillage. »

Afin de rendre ces différentes définitions utilisables sur le terrain, il est indispensable de traduire ces différentes définitions en dimensions opérationnelles, notamment à travers des indicateurs. Pour ce faire, certains auteurs, se sont entre autres appuyés sur des travaux internationaux.

Dimensions de la Qualité des Soins (Indicateurs de la Qualité des Soins)

Or et Com-Ruelle (2008) suggèrent de conserver 5 dimensions basées notamment sur les travaux de l'OCDE (Kelley & Hurst, 2006). Pour chacune, nous tenterons de proposer une définition et fournir d'éventuels indicateurs.

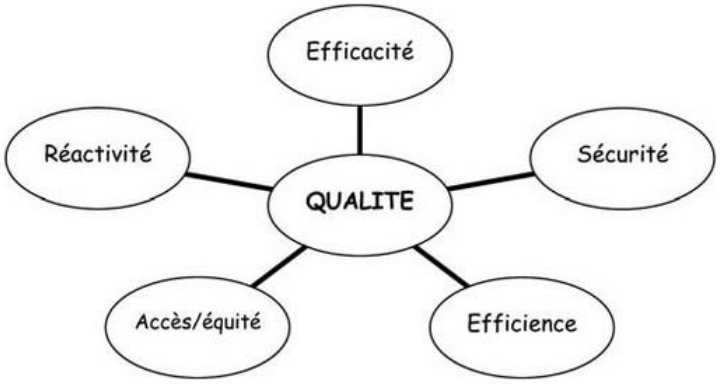


Figure 1-1 – Dimensions de la Qualité - D'après Or et Com-Ruelle (2008)

1- Efficacité

Définitions :

- la capacité de réaliser des résultats (des soins) souhaitables, à condition qu'ils soient bien dispensés à ceux qui en ont besoin et pas aux autres (Arah et al., 2003)
- l'aptitude à atteindre ou à réaliser toute amélioration possible en termes de résultats sanitaires (Donabedian, 2003)

Cette dimension est associée à la **pertinence** c-à-d que les soins fournis correspondent aux besoins cliniques et sont basés sur de solides recommandations médicales.

Quelques indicateurs :

- mortalité ou morbidité par cause, par groupe cible, etc. ;
- taux d'incidence (nombre de nouveaux cas rapporté à la population) de certaines maladies infectieuses comme le Sida, ou d'infections évitables par vaccination comme les hépatites A et B, le tétanos, ou encore d'affections graves telles que le cancer ;
- qualité technique des soins, par exemple :
    - le taux de patients fumeurs ayant reçu un conseil d'arrêt du tabac lors d'une consultation médicale (soins efficaces) ;
    - le taux d'admissions inappropriées pour l'asthme (sur des critères cliniques) ;
    - le taux de prescription d'antibiotiques pour infection virale (soins inefficaces) ;

- les taux d'hospitalisations évitables, les taux de réadmission et de mortalité post hospitalière ;
- les taux standardisés de mortalité par cancer du sein et du côlon ;
- les taux de survie à cinq ans après le diagnostic d'un cancer du sein et d'un cancer de la prostate ;
- les taux de dépistage du cancer du col d'utérus et de vaccination contre la grippe de la population de plus de 65 ans.

2- Sécurité

Définitions :

- La capacité d'empêcher ou d'éviter les résultats indésirables ou les dommages qui proviennent des processus de soins eux-mêmes (Donabedian, 2003).

Cette définition met l'accent sur la prévention et sur la réduction des défauts de qualité des soins.

Quelques indicateurs :

- Infections nosocomiales (des plaies, celles liées aux soins médicaux, aux escarres, etc.) ;
- Événements sentinelles (accidents liés à la transfusion, erreurs de groupage sanguin, oublis de corps étrangers dans le champ opératoire) ;
- Complications opératoires et postopératoires (embolies pulmonaires ou accidents ; d'anesthésie) ;
- Autres événements indésirables.

3- Accessibilité

Définitions :

- La facilité avec laquelle on accède aux bons services de santé au bon moment.
  Cela inclut les perspectives géographique, financier, socio-psychologique, et implique :
    - La disponibilité a priori des services de santé ;
    - L'équité (la capacité d'un système de santé à traiter de manière juste toutes les personnes concernées, indépendamment de leur âge, de leur sexe, de leur race et de leurs ressources financières) : la distribution des soins entre différents groupes de populations quelles que soient leurs situations géographique, économique et sociale ;
    - La ponctualité :
        - degré avec lequel les patients obtiennent les soins nécessaires rapidement ;
        - l'accès aux soins dans des délais opportuns (obtenir les soins lorsqu'on en a besoin) ;
        - la coordination de soins (trajectoire de soins).

Indicateurs :

- Délais d'attente pour la chirurgie programmée ;
- Délais d'attente aux urgences ;
- Sorties retardées ;
- Temps d'accès aux médecins généralistes et spécialistes ;
- Problèmes d'accès liés au coût des soins.

4- Réactivité

Définitions :

- La façon dont le système prend en charge les patients pour répondre à leurs attentes légitimes non liées à la santé (Kelley & Hurst, 2006)

Cela implique :

  - de mettre le patient au centre des soins en intégrant différents éléments comme l'écoute, l'empathie, la confidentialité, mais aussi l'information dont le patient dispose sur sa maladie et la possibilité d'un choix éclairé de sa part ;
  - la continuité de soins : la coordination des soins de santé dans le temps et à travers différents professionnels et établissements pour un même utilisateur.

Indicateurs :

- La majorité des mesures concernent la coordination entre les soins hospitaliers et ambulatoires, et notamment la réhospitalisation sous 30 jours.
- Autres indicateurs :
  - Mesures de la continuité des soins pour des conditions cliniques ;
  - Qualité de soins perçue par le patient :
    - Patient-Reported Outcome Measures (PROMs);
    - Patient-Reported Experience Measures (PREMs).

5- Efficience

Définitions :

- L'utilisation optimale des ressources disponibles pour obtenir les bénéfices ou les résultats les meilleurs (Kelley & Hurst, 2006).
- La capacité d'un système de santé à fonctionner à moindres frais sans diminuer les résultats possibles et souhaitables (Donabedian, 2003).

L'IOM soutient par exemple que les soucis d'efficience (contraintes de ressources) ne devraient pas être introduits dans la définition de la qualité, celle-ci ne devant pas varier en fonction des moyens disponibles.

Indicateurs :

- Les niveaux de ressources consacrés au système de santé
- La durée de séjour.

Indicateurs retenus dans cette thèse

Dans nos travaux, nous avons retenu deux indicateurs majeurs en santé publique : le taux de réadmission sous 30 jours, qui est à la fois un indicateur d'efficacité et de réactivité (continuité des soins), et la durée de séjour qui est principalement un indicateur d'efficience (management des flux, programmation des lits). Nous présentons ces deux indicateurs de façon plus détaillé dans les parties suivantes.

Durée de Séjour

« La durée moyenne de séjour à l'hôpital est souvent considérée comme un indicateur **d'efficience**.

- Un séjour plus court contribue à la réduction des coûts.
- **Les séjours plus longs peuvent être un indicateur de soins de mauvaise qualité** pour les raisons suivantes :
    - les traitement des patients peuvent subir des retards à cause de différents processus hospitaliers – ce qui indique un problème d'efficience ;
    - des soins de mauvaise qualité, peuvent nécessiter le recours à des traitements complémentaires ou engendre un temps de récupération plus long ;
    - une mauvaise coordination entre les différentes parties du système de santé, peut engendrer des attentes entre les différentes prises en charges, prolongeant ainsi la durée séjour.
- Dans le même temps, certaines personnes peuvent sortir trop tôt, alors qu'un séjour hospitalier plus long aurait pu améliorer leurs résultats ou réduire le risque de réadmission » (OECD, 2017).

En 2015, la durée moyenne de séjour à l'hôpital, toutes causes confondues, dans les pays de l'OCDE était d'environ **huit** jours. En 2019 elle était de 7.6 jours[4].

Au-delà des différences en besoins cliniques, plusieurs facteurs peuvent expliquer les variations entre les pays. En particulier, deux facteurs conjoints peuvent inciter les hôpitaux à garder les patients plus longtemps :

- Une offre abondante de lits ;
- La structure de paiement des hôpitaux.

Différentes stratégies ont été mises en œuvre pour réduire les durées de séjour (LOS) :

- L'adoption de dispositifs qui permettent de fixer les coûts des prestations et les rémunérations en amont (par exemple : en Allemagne, en France et en Pologne) basées sur les GHM (groupes homogènes de malades).
- Ces modes de paiement incitent les établissements à réduire le coût de chaque séjour.
- Des réductions stratégiques du nombre de lits d'hôpital, accompagnées d'un développement des services de soins de proximité.
- Promouvoir l'adoption d'actes chirurgicaux moins invasifs, et développer les programmes de sorties précoces qui permettent aux patients de retourner chez eux et d'y recevoir des soins de suivi, ou aider les hôpitaux à améliorer la coordination des soins.
- Réduire les sorties retardées à savoir le nombre de jours qu'une personne passe à l'hôpital après qu'un médecin l'a déclarée apte à sortir ou à être transférée.

Lorsque l'on s'intéresse à la variable LOS, on constate qu'elle a une distribution particulièrement asymétrique contenant un bon nombre d'observation hors de la boîte à moustache (outliers). Une des façons de gérer cette asymétrie caractéristique est de dichotomiser la variable en séjours dits « normaux » ou « courts », et en séjours « prolongés » ou « longs ». Les différents éléments présentés ci-avant suggèrent que ce sont justement ces séjours prolongés que nous devons chercher à comprendre, expliquer ou prédire.

[4] https://www.oecd-ilibrary.org/sites/265429e7-fr/index.html?itemId=/content/component/265429e7-fr#figure-d1e27704

Nous présentons ci-après deux façons différentes et non exhaustives de définir la durée de séjour prolongée : l'une en hospitalisation générale, et l'autre plus spécifique aux soins intensifs.

Définition de la durée de séjour prolongée (de façon générale)

Nous prendrons comme référence une étude basée sur 18 années d'observations réalisées au Mexican National Institutes of Health (MNIH) à Mexico (Marfil-Garza et al., 2018b)

La durée de séjour à l'hôpital (LOS) fait référence au nombre total de jours-lits occupés par un patient pendant son hospitalisation, et elle a été utilisée comme substitut traditionnel pour évaluer (Marshall et al., 2005) :

- l'efficience des soins de santé,
- l'efficacité des stratégies préventives et thérapeutiques,
- les méthodes de diagnostic,
- l'utilisation, l'allocation et l'administration des ressources hospitalières.

L'indicateur opérationnel de la durée de séjour à l'hôpital est la durée moyenne du séjour, et par cette mesure, les patients peuvent être classés en deux groupes : ceux ayant une durée de séjour normale (NLOS) et ceux ayant une durée de séjour prolongée (PLOS).

Pour Marfil-Garza (2018b), PLOS représente les séjours à partir du la 95ème centile du LOS (Baek et al., 2018), ce qui représente 34 jours.

De façon générale, il ne semble pas y avoir de consensus sur le seuil utilisé pour définir ce qu'est un séjour prolongé (Williams et al., 2010).

Les facteurs déterminants

Dans le cas particulier qui nous intéresse on peut les résumer comme suit (Marfil-Garza et al., 2018b) :

- Le diagnostic le plus courant : les néoplasmes hématologiques ;
- Le type de chirurgie le plus courant : la chirurgie de l'intestin grêle ;
- L'âge : plus jeune en moyenne ;
- Le sexe : masculin ;
- Le ratio médecin-patient : en moyenne plus faible ;
- Les admissions aux urgences et le week-end ;
- L'admission pour chirurgie ;
- Le nombre de comorbidités ;
- La résidence en dehors de Mexico ;
- Un statut socio-économique inférieur.

On notera en particulier les plus grands risques de PLOS :

- Greffe de moelle osseuse ;
- Les maladies infectieuses complexes telles que les mycoses systémiques et les parasitoses ;
- Les maladies abdominales complexes telles que la fistule intestinale.

De plus, le risque de mortalité chez les patients atteints de PLOS est trois fois plus élevé (3,7 % contre 13,3 %, $p < 0{,}001$).

Définition du séjour prolongé en réanimation

La durée de séjour en réanimation est un outil important dans l'évaluation de la qualité des soins et de l'activité d'un service de réanimation (Carpentier et al., 2015). Pour le praticien, assurer au quotidien une prise en charge optimale du patient avec une durée de séjour la plus courte possible est un enjeu essentiel compte-tenu de l'association étroite entre allongement de la durée de séjour et augmentation de la morbimortalité en réanimation (Combes et al., 2003; Montuclard et al., 2000; Rimachi et al., 2007).

Selon les données de la littérature, il est difficile d'établir une définition précise de la durée à partir de laquelle le séjour en réanimation doit être considéré comme prolongé. Il existe notamment une certaine hétérogénéité suivant le type de réanimation considéré.

- Pour un service de réanimation chirurgicale cardiaque, un séjour pourra être considéré comme prolongé dès une durée supérieure à trois à sept jours (Hassan et al., 2012; Mahesh et al., 2012) ;
- Concernant les services de réanimation médicale, chirurgicale et polyvalente, une durée de 14 jours apparaît relativement consensuelle et fréquemment retenue (Becker et al., 1984; Laupland et al., 2006; Zampieri et al., 2014) ;
- La ventilation mécanique (VM) est souvent retenue comme prolongée à partir de 21 jours (MacIntyre et al., 2005; White, 2012).

L'ensemble de ces patients pour lesquels la durée de séjour s'avère prolongée (≥14 jours) représente ainsi entre 4 à 11 % des admissions en réanimation (ARABI et al., 2002; Heyland et al., 1998; Silberman et al., 2013), soit un effectif relativement faible, mais correspondant, a contrario, à une occupation allant jusqu'à 40 voire 50 % des jours-lits de réanimation (ARABI et al., 2002; Zampieri et al., 2014). Cette situation n'est donc pas sans conséquence sur l'activité et l'utilisation des ressources du service concerné.

Particularités du LOS en réanimation

« Quatre à 11 % des patients admis en réanimation vont nécessiter un séjour prolongé en service de réanimation. Il s'agit essentiellement de patients admis pour détresse respiratoire aiguë, état de choc ou polytraumatisme et pour lesquels des durées prolongées de séjour et de VM apparaissent étroitement liées. (…) La prise en charge de ces patients, une fois stabilisés, et dont le séjour se prolonge en réanimation en raison souvent d'une dépendance ventilatoire, doit avoir pour objectif d'initier le plus rapidement possible leur autonomisation et leur réhabilitation. Le recours à la trachéotomie s'inscrit souvent dans cette prise en charge lorsque le sevrage ventilatoire s'avère long et difficile. Une attention particulière doit être portée à l'état neuropsychique de ces patients, chez lesquels une souffrance psychique intense est fréquemment retrouvée, pouvant ralentir, voire entraver le processus de récupération et de guérison et dont la prévention repose sur des mesures simples visant à favoriser le bien-être du patient. La prise en charge de la famille ne doit pas non plus être négligée, car elle est exposée elle aussi à cette grande souffrance psychique, d'autant plus que le séjour se prolonge. Les unités de sevrage ou de post réanimation semblent constituer ainsi des alternatives intéressantes pour poursuivre une prise en charge spécifique dans un environnement maintenu sécurisé, mais plus adapté aux besoins du patient, tout en permettant de préserver les capacités d'accueil des unités de réanimation et de diminuer l'impact économique engendré par un séjour prolongé. » (Carpentier et al., 2015, pp. 386–387).

En conclusion

Il ressort de ces deux grandes perspectives que la définition du séjour dit prolongé dépend des circonstances et du type d'admission tout autant que des services et des institutions de soins concernés. C'est une des raisons qui justifient le choix de critères statistiques pour définir le seuil de classement en séjour « normal » vs « prolongé ». En effet, le choix de ce seuil n'est pas toujours clair ni explicite (Williams et al., 2010, p. 462), et aucun ne fait l'unanimité. Certains utilisent des critères ad-hoc pour obtenir une distribution relativement équilibrée du LOS (Chrusciel et al., 2021), d'autres utilisent des centiles (75, 90 ou 95) (Blumenfeld et al., 2015; Collins et al., 1999).

Réhospitalisation (Réadmission)

La réhospitalisation fait partie des indicateurs nationaux (France) de qualité et de sécurité des soins dans les établissements de santé – IQSS.

« Afin de promouvoir l'amélioration de la qualité et de la sécurité des soins et de répondre à la demande des patients et des usagers, le ministère chargé de la santé conduit une politique de transparence sur les résultats des actions menées en la matière au sein des établissements de santé. Ces actions sont suivies à l'aide d'indicateurs de mesure dédiés : les indicateurs de qualité et de sécurité des soins (IQSS) en établissements de santé. »[5].

La réhospitalisation est ainsi explicitement mentionnée comme étant un indicateur[6] de qualité (sécurité) des soins.

Nous avons cependant noté dans les précédents paragraphes que la réhospitalisation fait également partie des indicateurs d'efficacité et de réactivité, lesquels sont deux dimensions de la qualité selon la perspective de l'OCDE. On voit ainsi que la réhospitalisation constitue effectivement un indicateur récurrent de la qualité. Plus précisément, pour le ministère de la santé et des solidarités la réhospitalisation constitue aussi, sous ses différentes formes, un indicateur de coordination (DGOS & ATIH, 2022) :

a. *Le taux de réhospitalisation dans un délai de 1 à 7 jours en médecine-chirurgie-obstétrique (MCO) : RH7.*
   Indicateur de vigilance et d'alerte qui concerne en premier lieu les établissements de santé MCO. Il vise à renforcer la réflexion des équipes sur leurs pratiques. Calculé pour chaque établissement de santé, ajusté à l'activité et croisé avec d'autres informations (la durée moyenne de séjour par exemple), il a vocation de faciliter l'identification d'actions d'amélioration par les équipes hospitalières.
b. *Le taux de réhospitalisation dans un délai de 1 à 30 jours en médecine et chirurgie (RH30) ainsi que le taux d'hospitalisation potentiellement évitable (HPE).*
   Sont des indicateurs de vigilance et d'alerte concernant la problématique de la coordination ville / hôpital (Premier recours/établissements de santé). Ces deux indicateurs territoriaux sont calculés par zone géographique et non par établissement de santé.

Il peut être utile de situer cette conception Française dans un cadre plus général. Nous reprenons ci-après l'essentiel d'un document préparé par la direction générale de l'offre des soins (DGOS) et

[5] https://solidarites-sante.gouv.fr/soins-et-maladies/qualite-des-soins-et-pratiques/qualite/les-indicateurs/article/les-indicateurs-de-qualite-et-de-securite-des-soins-dans-les-etablissements-de

[6] https://solidarites-sante.gouv.fr/soins-et-maladies/qualite-des-soins-et-pratiques/qualite/les-indicateurs/article/les-indicateurs-de-rehospitalisation-et-de-coordination

l'agence technique de l'information sur l'hospitalisation (ATIH), positionnant la perspective des autorités françaises sur la question de la réhospitalisation (DGOS & ATIH, 2022)

1- Les Réhospitalisations dans leur contexte international

I – Les USA

Les premiers travaux aux USA sur les réhospitalisations remontent aux années 80 et concernent notamment les personnes âgées de plus de 65 ans, couvertes par l'assurance maladie publique Medicare. Il s'agissait à l'époque d'identifier des prises en charge en ambulatoire ou à domicile qui auraient dû permettre de limiter les réadmissions pour des personnes souffrant de certaines maladies chroniques comme le diabète ou la bronchopneumopathie chronique obstructive (BPCO)(Burgess & Hockenberry, 2014). À la même période, d'autres travaux ont montré la part prépondérante des dépenses des patients atteints de pathologies chroniques, nécessitant des hospitalisations multiples (Schroeder et al., 1979; Zook et al., 1980; Zook & Moore, 1980). Une autre étude démontre que 22 % des hospitalisations de 1974 à 1977 étaient suivies d'une réhospitalisation dans les 60 jours suivant la sortie du patient alors que près de 60 % des dépenses hospitalières de Medicare concernaient 12,5 % des bénéficiaires hospitalisés au moins trois fois sur la période d'étude (Anderson & Steinberg, 1984).

Sous la recommandation du Medicare Payment Advisory Commission (MedPAC), le congrès US a décidé de sélectionner le délai de 30 jours comme indicateur pour mesurer le taux de réadmission, permettant de développer une politique d'incitation à la réduction des taux de réhospitalisation (McIlvennan et al., 2015). Certaines réhospitalisations peuvent être le fruit d'une gestion inadéquate de la continuité des soins par les établissements de santé, notamment en sortie d'hospitalisation : programmation de la sortie, délivrance des informations et éducation du patient et suivi du patient à son domicile en sont des exemples. L'indicateur de réhospitalisation sous 30 jours (RH30) serait donc sensible à la coopération et à la coordination des secteurs ambulatoires et hospitaliers (Raleigh, 2014).

Parmi les bénéficiaires de Medicare aux États-Unis, les patients avec les taux de recours aux soins primaires les plus élevés avaient les taux de réadmission à 30 jours les plus faibles suite à une prise en charge chirurgicale. Ce recours, notamment en sortie d'hospitalisation, semble être un facteur clé pour réduire les réadmissions à 30 jours. Ainsi, les patients qui bénéficiaient d'une prise en charge par les acteurs de soins primaires après leur hospitalisation pour chirurgie étaient moins réhospitalisés dans les 30 jours que les autres, notamment pour les prises en charge avec complications (Brooke et al., 2014). Ces résultats suggérant une incidence de l'offre de soins primaires sont concordants avec des études en cours en France sur le risque de réhospitalisation des patients après infarctus du myocarde et des patients atteints d'insuffisance cardiaque. La prise en compte de la transition hôpital-ville et de ses facteurs explicatifs ainsi que l'accessibilité à la médecine générale pourraient constituer autant de pistes pour réduire les réhospitalisations à 30 jours.

II – Autres pays

En Allemagne, l'indicateur de réadmission à 30 jours a été mis en place dans le cadre d'une politique visant à mesurer les effets inattendus de la tarification à l'activité, notamment l'augmentation du nombre d'hospitalisations.

L'Angleterre a utilisé l'indicateur RH30 dans le cadre d'un programme d'incitation financière à la diminution des réadmissions des patients par les services d'urgence. Les établissements de santé sont ainsi responsables de leurs patients après leur hospitalisation.

Le Danemark rend public les résultats de réhospitalisations des établissements afin d'appuyer les patients dans leurs choix (Kristensen et al., 2015).

III – En France

RH3 et RH7 : correspondent à des taux de réadmission dans les 3 jours après une chirurgie ambulatoire (HAS et ATIH) et des taux de réhospitalisation entre 1 et 7 jours.

Il convient de bien distinguer les réhospitalisations entre 1 et 7 jours et sur une période plus étendue de 1 à 30 jours. En effet, alors que les réhospitalisations les plus précoces semblent liées à la qualité des soins au cours du séjour d'hospitalisation (infection nosocomiale, mauvaise prise en charge, etc.), les réhospitalisations survenant au-delà de la première semaine suivant la sortie d'hospitalisation seraient le reflet de l'organisation du système de soins (coordination des acteurs, accès aux soins, etc.) (Chin et al., 2016; Graham et al., 2015; Joynt & Jha, 2012).

Pour les deux indicateurs de réhospitalisation précoce, le résultat de l'indicateur est spécifique à un établissement. Pour les réhospitalisations dans un délai plus long, le résultat de l'indicateur fait sens au niveau d'un territoire.

2- Prévenir la réhospitalisation sous 30 jours

Etant entendu qu'il est souhaitable de réduire la réadmission sous 30 jours afin d'améliorer la qualité des soins, et de réduire les coûts, il convient de faire le point sur les interventions agissant dans ce sens (O. Hansen et al., 2011).

Les interventions pré décharge (c-à-d en préparation de la sortie après le séjour hospitalier).

Elles incluent:

- L'éducation du patient ;
- La planification de la décharge (sortie) ;
- La réconciliation médicamenteuse ;
- La programmation de rendez-vous de suivis.

Les interventions post décharge incluent :

- Des suivis réguliers et opportuns téléphoniques ;
- La disponibilité d'une hotline ;
- Une communication régulière et opportune avec des fournisseurs de services ambulatoires ;
- Des visites de suivis à domicile.

Pour faire le lien (la transition) entre ces deux périodes :

- L'intervention de coachs de transition ;
- La continuité des soins entre le séjour hospitalier et post hospitalier ;
- Des instructions de décharge adaptées, centrées sur le patient.

3- Une brève synthèse de la littérature sur les modèles de prédiction de la RH30

Les prédicteurs de la RH30

Des revues de littérature systématiques suggèrent que les variables qui prédisent le plus souvent la RH30 sont en priorité les comorbidités, les variables sociodémographiques, la durée de séjour(LOS),

les précédentes admissions suivies des tests de laboratoire, des médicaments, puis du type d'admission, parmi les plus importantes (Zhou et al., 2016).

Malgré l'inclusion plus récemment de variables plus complexes et des éléments issues du Big Data (signes vitaux, résultats de laboratoires, complexité des procédés chirurgicaux, etc.) ces données restent incomplètes et ne tiennent pas compte de variables susceptibles d'être particulièrement pertinentes pour la réhospitalisation (la fragilité au moment de la décharge, indicateurs de risques de multimorbidité, disponibilité de soins ou d'accompagnement post-décharges, instabilité du domicile, etc.). Il reste également à exploiter plus pleinement les données issues des notes cliniques des infirmiers et des médecins (Mahmoudi et al., 2020).

Le pouvoir discriminant des modèles (performances des classifieurs mesurées par le C-statistic ou ROC AUC)

En termes de performances, les modèles prédictifs étudiés par Zhou et al. (2016) dans une revue de littérature systématique indiquent une large variation : de 0.21[7] à 0.88. Les résultats semblent cependant indiquer qu'un modèle « acceptable » prédisant le RH30 devrait avoir un ROC AUC d'au moins 0.70, et que les meilleurs modèles démontrent un ROC AUC approximativement à 0.88.

Dans le cas d'une autre revue de littérature systématique de Mahmoudi et al. (2020), la valeur du ROC AUC pour les modèles traditionnels (Régression Logistique) vaut en moyenne 0.74, et pour les modèles de Machine Learning 0.76. L'utilisation de données textuelles dans des modèles a fait progresser le ROC AUC moyen de 0.75 à 0.78.

Après avoir fourni un contexte général sur la qualité des soins en santé publique, en France comme à l'international et notamment dans les pays de l'OCDE, nous allons maintenant procéder à une présentation synthétique et essentiellement conceptuelle du Machine Learning et des approches que nous utilisons dans cette thèse.

[7] Cette valeur implique que le modèle prédit moins que le hasard pur. Il y a probablement une erreur d'étiquetage de l'outcome, ce qui donnerait une valeur de l'AUC = 0.79

Machine Learning : un résumé conceptuel

L'utilisation de techniques de Machine Learning (et de Data Mining) est aujourd'hui largement adoptée et établie en santé publique (Asadullah et al., 2021; dos Santos et al., 2019) et plus largement en médecine ou en informatique médicale (Ravì et al., 2017). Mais il convient sans doute au préalable de se faire une idée plus précise de ce qu'est le machine learning.

Le Machine Learning en Bref

Soient p variables $X = [X_1, \dots, X_p]$ dites prédicteurs ou variables explicatives (*features*) dans un jeu de données à n observations (*instances*).

1- Apprentissage supervisé

Supposons qu'il existe une variable à expliquer (*outcome*) Y telle qu'il existe une relation entre Y et X sous la forme générale :

$$Y = f(X) + \varepsilon$$

Où f représente une fonction fixe mais inconnue des X et ε un terme d'erreur aléatoire indépendant de X et de moyenne nulle.

Soit $\hat{Y} = \hat{f}(X)$ une estimation de f. De façon générale, le Machine Learning revient à résoudre de façon numérique et algorithmique un problème d'optimisation. Typiquement, on minimise une fonction objective ou $Loss(y, \hat{y})$ où $y\ et\ \hat{y}$ représentent une réalisation de Y et $\hat{Y}$ sur un échantillon donné. Le modèle le plus ajusté est celui disposant des (hyper)paramètres pour lesquels cette fonction est optimale.

En essence donc, le machine learning – en tant qu'apprentissage supervisé – fait référence à un ensemble d'approches pour estimer f (James et al., 2013).

L'apprentissage supervisée peut être divisée en deux catégories : dans le cas où Y est quantitative, on parle de **régression** et dans le cas ou Y est catégorielle, on parle de **classification**.

Il existe deux motivations principales à l'estimation de f : la prédiction ou l'inférence (James et al., 2013).

Prédiction

Dans la prédiction on peut traiter $\hat{f}$ comme une *boîte noire* dans le sens où on ne se soucie pas de sa forme exacte, du moment qu'elle prédit correctement Y. L'enjeu ici est donc la *performance*.

Dans l'apprentissage supervisé, on s'appuie sur des couples d'instances (y_i, x_i) et d'un échantillon dit d'apprentissage (*training set*) pour « entraîner » f et qu'elle puisse « apprendre » à partir de ces exemples. La performance est ensuite évaluée sur un échantillon dit de test (*test set*) qui n'a pas été utilisé pour entraîner le modèle f. Entre ces deux phases, on peut se servir d'un échantillon dit de validation (*validation set*) pour ajuster (*tuning*) le modèle.

Inférence & Interprétabilité

Ici on ne peut plus traiter f comme une *boîte noire*. Il s'agit d'expliciter la relation entre X et Y et plus précisément comment les changements dans X peuvent affecter Y. Trois questions peuvent notamment se poser :

- Quelles sont les prédicteurs les plus associés avec la variable cible, et à quel point ; c'est l'étude de l'importance des variables (*features importance*).
- Quelle est la nature de la relation entre chaque prédicteur et la variable à expliquer (positive, négative…).
- La relation entre Y et chaque prédicteur de X peut-elle prendre une forme explicite ou analytique ?

Ce paradigme de l'inférence est étroitement lié à une idée essentielle qui est l'interprétabilité des modèles définie comme *la capacité d'expliquer ou de présenter un modèle d'une façon compréhensible pour l'humain* (Carvalho et al., 2019; Doshi-Velez & Kim, 2017).

En effet, plus les modèles sont complexes plus il est difficile d'en obtenir une expression analytique ou explicite, mais la question de la relation entre la cible est les prédicteurs reste pertinente et dans certaines disciplines comme la santé publique, essentielle.

2- Apprentissage non supervisé

Lorsque la variable explicative (ou cible) est absente, alors l'enjeu consiste à trouver des schémas de relations entre les prédicteurs (corrélations) ou entre les instances (similarités). Cela permet de regrouper les variables les plus « liées », par exemple en extrayant un maximum d'information dans les nuages de points (Analyse en Composante Principale) ou bien de regrouper les observations par similarité (Clustering).

La présente thèse se limite au cadre de l'apprentissage supervisé, et plus spécifiquement, à la classification.

3- Quelques spécificités des sciences de la santé.

De façon schématique, les données utilisées en recherche médicale sont de deux types :

- Les données tabulaires structurées classiques qui réunissent des variables socio-démographiques, cliniques ou hospitalières ;
- Les données plus complexes structurées et non structurées comme celles issues de l'imagerie ou des notes cliniques (comptes rendus d'hospitalisation, etc.).

Les données tabulaires classiques sont généralement traitées avec des modèles classiques de Machine Learning (Asadullah et al., 2021; dos Santos et al., 2019), alors que les données plus complexes ou non structurées sont plutôt traitées avec des modèles de Deep Learning (Ravì et al., 2017).

Nous allons présenter ci-après de façon plus formelle les différents modèles de Machine Learning mobilisés dans cette thèse. Une première approche d'interprétabilité sera abordée pour chaque classifieur, mais la question de l'interprétabilité sera développée plus longuement ultérieurement.

Modèles de Machine Learning classiques (supervisés)

1- KNN ou K Nearest Neighbors

Principe:

Soient Y une variable aléatoire qualitative (catégorielle) à expliquer de modalités $j \in [1, \dots, k]$ et des variables aléatoires explicatives $X = [X_1, \dots, X_p]$ prenant respectivement des valeurs $y = (y_i)$ et $x = (x_i)$ $o\grave{u}$ $i \in [1, \dots, n]$ identifie les instances.

Le but du Machine Learning – en tant qu'apprentissage supervisé – est de trouver une fonction (classifieur) qui minimise l'erreur (ou maximise la performance) estimée sur un échantillon test x_0. Dans un problème de classification cela revient à assigner x_0 à une classe j qui maximise la probabilité conditionnelle du classifieur de Bayes donné par (James et al., 2013) :

$$\mathcal{P}(Y = j|X = x_0)$$

Dans la pratique, on ne connaît pas la distribution conditionnelle de Y, sachant X mais on peut l'estimer entre autres par KNN :

Étant donné un entier positif K et une observation x_0 de l'échantillon test, le classifieur KNN identifie d'abord les points K de l'échantillon d'apprentissage qui sont les plus proches de x_0, représentés par $\mathcal{N}_0$. Il estime ensuite la probabilité conditionnelle de la classe j comme la fraction de points dans $\mathcal{N}_0$ dont les valeurs de réponse sont égales à j :

$$\hat{\mathcal{P}}(Y = j|X = x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} I(y_i = j)$$

KNN applique la règle de Bayes pour classer l'observation x_0 dans la classe de plus grande probabilité. Dans la pratique, cela revient à classer x_0 dans la classe majoritaire parmi les K plus proches voisins. La valeur optimale de K peut être déterminée par validation croisée.

Interprétabilité :

Il n'existe pas d'interprétabilité globale (au niveau de chaque variable) pour KNN. Cependant, on peut envisager une forme locale d'interprétabilité en étudiant les voisins dont la classe a été assignée à une nouvelle observation x_0 (Molnar, 2022). Pour obtenir une interprétation globale, on peut toujours faire appel à une méthode agnostique (indépendante du modèle).

2- Régression Logistique Pénalisée

Principe :

Intéressons-nous au cas d'une classification binaire qui constitue l'essentiel de ce qui est traité dans cette thèse. La variable à expliquer Y ne peut avoir que deux valeurs $1\ ou\ 0$. Nous choisissons d'associer $Y = 1$ à notre modalité d'intérêt.

Soient $X = [X_1, \dots, X_p]$, $\boldsymbol{X} = [1, X]$ $\beta^T = [\beta_0, \beta_1, \dots., \beta_p]$ et x un échantillon de réalisation de X.

$$\pi(x) = \mathcal{P}(Y = 1\,|X = x) = \sigma(\boldsymbol{X}\beta) = \frac{1}{1 + e^{-\boldsymbol{X}\beta}}$$

En régression logistique, on cherche à minimiser la fonction $-log(vraisemblance)$ donnée par :

$$L(\beta) = \sum_{i=1}^{n}\{y_i ln\sigma(X_i\beta) + (1 - y_i)\ln\,(1 - \sigma(X_i\beta)\,\}$$

Dans la régression logistique pénalisée, on cherche à minimiser une fonction proportionnelle à $(-\ log\ vraisemblance\ +\ un\ terme\ de\ pénalité$ donnée par :

$$J(\beta) = \lambda\left[\frac{(1-\alpha)}{2}\|\beta\|^2 + \alpha|\beta|\right] - L(\beta)$$

Où ‖ ‖ représente la norme Euclidienne L2 et | | la norme L1

$\alpha = 0$ correspond à la régression « Ridge » et $\alpha = 1$ correspond à la régression « Lasso » – sinon on est dans un modèle « Elasticnet » où α représente le degré de mélange (entre Ridge et Lasso), et λ un hyperparamètre de pénalité. En effet, pour que le terme de pénalité reste petit, λ contraint les coefficients β de la régression à tendre vers zéro quand il augmente. Les coefficients β ne sont donc pas uniques, mais varient avec λ (Hastie et al., 2009; James et al., 2013).

Interprétabilité :

Pour des prédicteurs quantitatifs, chaque variation unitaire du prédicteur x_j entraîne une variation de la cote $\frac{\mathcal{P}(y=1|X)}{\mathcal{P}(y=0|X)}$ d'une quantité égale à e^{β_j} (ou de façon équivalente une variation du $logit$ de β_j).

Pour des prédicteurs catégoriels, il est plus simple de réaliser une disjonction complète (one hot encoding) et de prendre une catégorie de référence. Chaque variation unitaire du prédicteur x_j entraîne une variation de la cote $\frac{\mathcal{P}(y=1|X)}{\mathcal{P}(y=0|X)}$ d'une quantité égale à e^{β_j} par rapport à la catégorie de référence.

4- Arbres de décision de type CART (Classification and Regression Trees)

Principe :

Les arbres de décision consistent à subdiviser l'espace des prédicteurs en sous-espaces (un ensemble de « rectangles ») et de prédire la valeur de la variable à expliquer par une constante. Sur la figure 1-1, l'arbre de décision est représenté par la partie dans le quadrant inférieur gauche. Les points de coupures avec les différents seuils t_1 à t_4 sont les **nœuds** (internes) et les régions R_1 à R_5 représentent les nœuds terminaux où **feuilles** de l'arbre. Les segments qui relient les nœuds sont les **branches**.

On considère ici le cas où l'algorithme suit une subdivision binaire récursive :

- On sélectionne le prédicteur X_j et le seuil s de telle sorte que la subdivision de l'espace du prédicteur en région $R_L(j,s) = \{X|\ X_j < s\}$ et $R_R(j,s) = \{X|\ X_j > s\}$ mène à une réduction aussi grande que possible de la fonction d'impureté du nœud parent, vers les nœuds enfants.
- La fonction d'impureté $\Phi(p_1, \dots, p_K)$ mesure la pureté des données dans une catégorie k. Pour une classification à K classes, elle est définie dans $[0,1]^K$ par (Breiman et al., 1984) :
    1. $\sum_j^K p_j = 1,\ p_j \geq 0$
    2. Φ est maximum quand tous les p_j sont égaux
    3. Φ est minimum quand un des $p_j = 1$ et tous les autres 0
    4. Φ est symétrique par permutation des p_j
- Etant donnée une fonction d'impureté Φ, on définit la mesure de l'impureté du nœud t par :

$$i(t) = \phi(p(1|t), p(2|t), \dots, p(K|t))$$

Où $p(j|t)$ est la probabilité a posteriori de la classe j sachant qu'un point se trouve dans le nœud t

Ainsi, si $i(t)$ désigne la mesure d'impureté pour le nœud t, parent des nœuds $t_R\ et\ t_L$

$$\Delta i(sp, t) = i(t) - p_R i(t_R) - p_L i(t_L)$$

Représente la qualité de la partition (split) sp c-à-d la différence de mesure d'impureté entre le nœud parent t et la somme pondérée des mesures d'impuretés entre les nœuds enfants de droite (t_R) et de gauche (t_L) pour la partition sp - où $p_R\ et\ p_L$ sont les proportions respectives des échantillons dans les nœuds correspondants.

L'algorithme de classification des arbres de décision recherche parmi toutes les partitions sp celle qui maximise $\Delta i(sp, t)$.

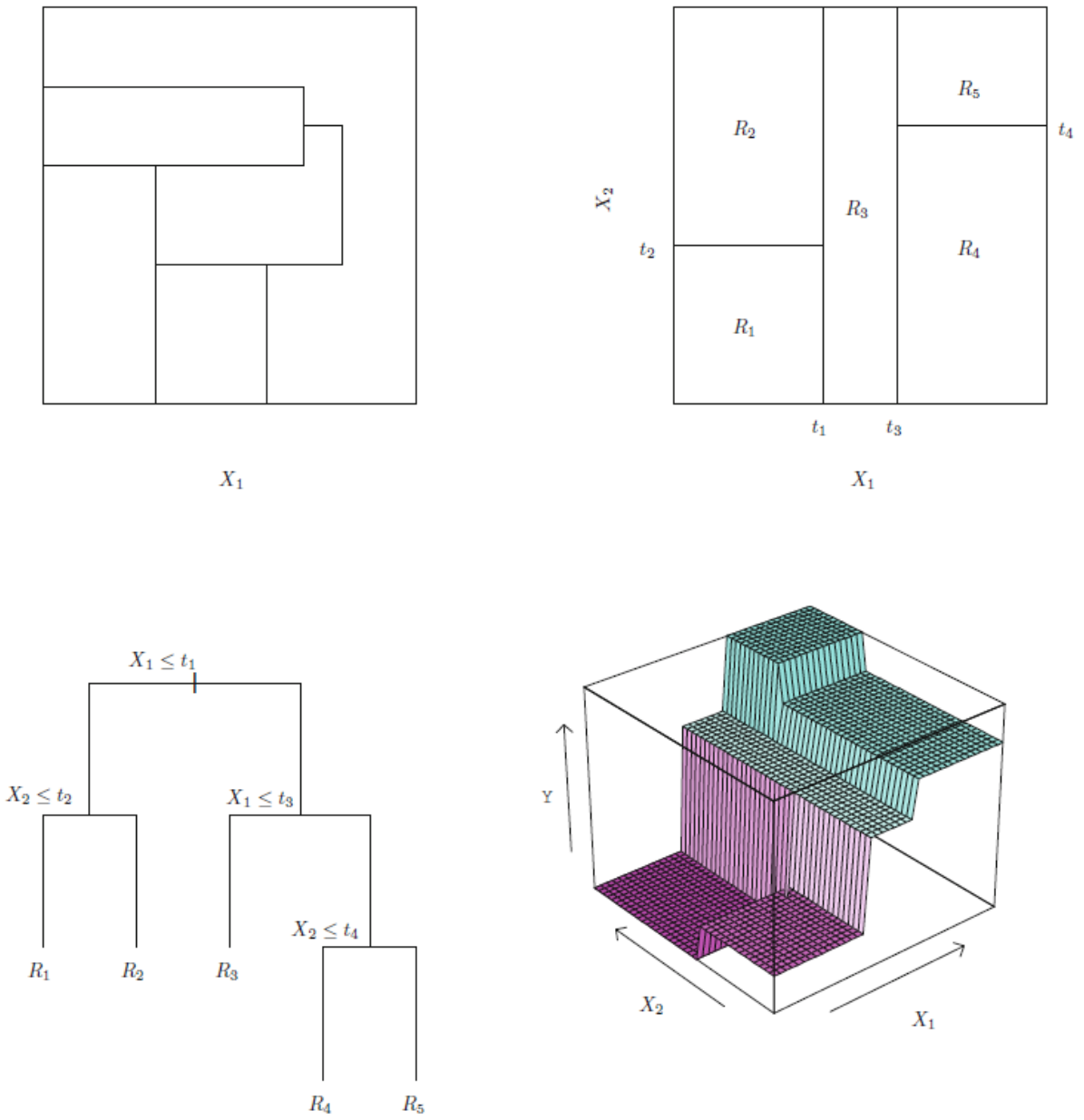


Figure 1-2. Subdivision récursive binaire d'un espace à deux prédicteurs

D'après James et al. (2013, p. 308)

Dans la pratique, les fonctions d'impuretés les plus utilisées sont :

- L'indice de Gini :

$$G(\hat{p}_{mk}) = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- La fonction Entropie :

$$D(\hat{p}_{mk}) = -\sum_{k=1}^{K} \hat{p}_{mk}(log\, \hat{p}_{mk})$$

Avec $\hat{p}_{mk}$ = proportion d'observations de l'échantillon d'apprentissage dans la région R_m, provenant de la classe k :

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

- On assigne chaque observation à la classe majoritaire de chaque nœud.
- On répète la même opération au sein de chaque nouvelle région et on poursuit ceci jusqu'à qu'à atteindre un critère d'arrêt prédéfini (exemple : le nombre minimum d'observations par région est atteint).

Elagage par complexité (Complexity pruning)

Il convient ensuite de tailler les branches de l'arbre selon un critère dit de « coût de complexité (complexity cost) » minimal.

- Soit $R(T)$ le taux d'erreur de classification d'un arbre T à partir de l'échantillon d'apprentissage.
- Soit $|T|$ le nombre de nœuds terminaux du sous-arbre T de l'arbre maximal T_{max} . On l'appelle complexité.
- Soit $\alpha \geq 0$ le paramètre de complexité.

La fonction qui mesure le « coût de complexité » (complexity cost) est donnée par :

$$R_\alpha(T) = R(T) + \alpha|T|$$

Elle doit être minimisée en α pour tailler l'arbre afin de trouver le sous-arbre optimal

Déroulé :

- On commence par rechercher l'arbre maximal T_{max}
- Pour tout nœud t, parent des nœuds t_R et t_L tel que $R(t) = R(t_R) + R(t_L)$ on taille t_R et t_L
- On répète ceci récursivement jusqu'à ne plus satisfaire la condition d'égalité
- Le sous-arbre obtenu est dit T_1 et le sous-arbre optimal y est inclus

Interprétabilité :

La **représentation graphique** d'un arbre de décision correspond à une interprétabilité maximale explicite. Chaque nœud est une condition à satisfaire pour passer au nœud suivant, jusqu'à arriver à la décision finale (la classe d'attribution).

En partant du nœud racine, on suit les nœuds successifs. Les arêtes indiquent dans quels sous-ensembles on est situé. Le dernier nœud dans la progression (la feuille) indique la classe prédite. Chaque arête représente une opération d'intersection logique « ET ».

De plus, on peut calculer **l'importance de chaque variable** en mesurant la variation de la fonction d'impureté entre les nœuds enfants et le nœud parent, chaque fois que la variable est utilisée dans l'arbre.

5- Random Forest

Principe :

Le Random Forest (RF) est un cas particulier de Bagging (Bootstrap Aggregation). Bien que nous nous intéressions ici au Bagging appliqué aux arbres de décision, le principe peut être étendu à d'autres algorithmes d'apprentissages (classifieurs ou régresseurs)

Pour un échantillon de taille n, le Bagging consiste à tirer des échantillons avec remise de même taille n, et de construire un arbre de décision avec chaque échantillon dit de bootstrap. Pour prédire une nouvelle observation, on choisit la classe majoritaire parmi l'ensemble des prédictions faites par chaque arbre. Les arbres ne sont pas taillés.

Alors que dans le Bagging, on utilise tous les prédicteurs pour construire chaque arbre, dans Random Forest, on le fait à partir d'un échantillon de prédicteurs, tirés sans remise. La taille de l'échantillon des prédicteurs est un paramètre qu'on peut déterminer par validation croisée.

Interprétabilité :

Bien qu'il n'y ait pas d'interprétabilité intrinsèque pour Random Forest, on peut prolonger le calcul de l'importance des variables à chaque arbre de décision du modèle puis en faire la moyenne sur la totalité des arbres. Afin de bien mettre en évidence l'importance relative des variables entre elles, on utilise la variable la plus importante comme étalon (100%) et on mesure l'importance des autres variables relativement à celle-ci. Dans le processus de bootstrap, le tiers environ de l'échantillon n'est pas utilisée pour la modélisation : c'est un échantillon dit « out of bag » (OOB). On peut utiliser cet échantillon OOB pout estimer la performance du RF et l'importance des variables.

Famille de Boosting

Comme le Bagging, le Boosting est une approche générale qu'on peut utiliser pour différents algorithmes d'apprentissage, dans la régression comme dans la classification. On se limitera ici à la classification et aux arbres de décision.

Principe du Boosting :

On appelle weak learner un classifieur f − avec un nombre nœuds (splits) d et un paramètre d'atténuation λ − qui prédit juste un peu mieux que le hasard (ou un modèle aléatoire). Un strong learner est au contraire un classifieur qui performe significativement mieux qu'un modèle aléatoire. Le boosting consiste à transformer un ensemble de weak-learners en un strong learner.

La plupart des algorithmes de boosting consistent à entraîner des weak learners, puis de les ajouter de façon séquentielle et pondérée en fonction de leur performance. Après l'addition de chaque weak learner, on procède à une repondération des données de sorte que les observations mal classées gagnent en poids et celles qui sont correctement classées perdent en poids. Ainsi la nouvelle série de weak learners pourra se concentrer davantage sur ce que les précédents ont manqué.

James et al. (2013) décrit le boosting d'arbres de décision de la façon suivante (pp 321-322) :

« A partir du modèle courant, nous ajustons un arbre de décision aux résidus du modèle. Autrement dit, nous ajustons un arbre en utilisant les résidus courants, plutôt que l'outcome y, comme réponse. Nous ajoutons ensuite ce nouvel arbre de décision dans la fonction ajustée afin d'actualiser les résidus. Chacun de ces arbres peut être assez petit, avec seulement quelques nœuds terminaux, déterminés par le paramètre d dans l'algorithme. En ajustant de petits arbres aux résidus, nous améliorons lentement le modèle $\hat{f}$ dans les zones où il n'est pas bien performant. Le paramètre d'atténuation λ ralentit encore plus le processus, permettant à des arbres de formes plus nombreuses et différentes d'attaquer les résidus. En général, les approches d'apprentissage statistique qui apprennent lentement ont tendance à bien fonctionner. A noter qu'en boosting, contrairement au bagging, la construction de chaque arbre dépend fortement des arbres précédents (déjà existants). »

Il existe plusieurs déclinaisons du boosting. L'essentiel de leur différence réside dans les méthodes de pondération des données d'apprentissage et de la variable à expliquer (residus).

6- Gradient Boosting

Dans le cas simple d'une classification binaire, l'algorithme du gradient boosting est donné par (M. Kuhn & Johnson, 2013) :

1. On initialise toutes les prédictions par $\widehat{f_i}^{(0)} = log\frac{\hat{p}}{1-\hat{p}}$ (Bernouilli) où $\hat{p}$ est la proportion d'une des classes de l'échantillon d'apprentissage (la classe positive)
2. Pour j variant de 1 à M (nombre d'arbres) répéter :
3. $\hat{p}_i = \frac{1}{1+\exp[-f(x)]}$
    a. Calculer le résidu (gradient) $z_i = y_i - \hat{p}_i$
    b. Échantillonner de façon aléatoire les données d'apprentissage
    c. Apprentissage sur un arbre sur le sous échantillon, utilisant les résidus comme variable à expliquer
    d. Estimer les résidus de Pearson des feuilles
    e. $r_i = \frac{1/n\sum_i^n(y_i-\hat{p}_i)}{1/n\sum_i^n(1-\hat{p}_i)\hat{p}_i}$
    f. Actualiser le modèle courant par $\hat{f}_i = \hat{f}_i + \lambda\,\hat{f}_i^{(j)}$
4. Fin

Conceptuellement, le but du gradient boosting est de minimiser la fonction Loss (Objective) en ajoutant des weak learners, et en utilisant une optimisation par descente du gradient.

7- XGBOOST (eXtreme Gradient Boosting)

Comme son nom l'indique, il s'agit d'une amélioration significative du gradient boosting à travers plusieurs contributions supplémentaires (T. Chen & Guestrin, 2016) :

1. En ajoutant un terme de régularisation à la fonction objective (Chen & Guestrin, 2016, p. 786, equation 2) qui pénalise la complexité des arbres additifs du modèle. Cette fonction objective pénalisée tend à favoriser les modèles les plus simples (moins complexes, i.e. avec moins de feuilles), puis en optimisant chaque nouvel arbre ajouté séquentiellement (Chen & Guestrin, 2016, p. 786, equations 5-7). Xgboost construit également les « weak tree learners » d'une façon différente des autres approches de Boosting.
2. En ajoutant un terme d'atténuation (shrinkage) à la contribution de chaque fonction additive, à chaque itération.

3. En utilisant un sous-échantillonnage des variables (comme dans Random Forest) et/ou un sous-échantillonnage des observations.
4. Utilisation d'une nouvelle distribution des données dite « weighted quantile sketch » pour déterminer les seuils de partitions optimaux. Autrement dit, xgboost ne teste pas chaque seuil de chaque prédicteur, mais des quantiles pondérés par la Hessienne (au niveau de chaque instance) en utilisant des algorithmes appelés « sketches » qui répartissent préalablement les données pour un traitement en parallèle. Ceci est fait de telle sorte que sont regroupés dans les mêmes quantiles les observations dont les poids cumulés sont similaires.
5. Une prise en compte spécifique de la sparsity (rareté des données) qui multiple par un facteur 50 la vitesse de l'algorithme, et qui est en mesure d'entraîner des modèles avec des valeurs manquantes.
6. Utilisation de l'apprentissage par bloc parallèles pour réduire l'importante durée de tri et de recherche de seuils de partitions.
7. Prise en compte de la taille optimale des blocs pour la gestion optimale du cache et de la parallélisation. Etant donnée que la mémoire cache est plus rapide que les mémoires RAM et ROM, xgboost calcule les Gradients et Hessiens à partir de cette mémoire – ce qui représente un gain de temps de calcul significatif.
8. Utilisation du *block compression* et *block sharding* pour faciliter les calculs à la volée. Lorsque les données sont de trop grande taille, au moins une partie doit être sauvegardée au niveau du disque dur (ROM). Xgboost essaie de minimiser le temps de traitement en compressant les données. Le temps CPU pour réaliser cette opération est plus courte que le temps pour lire les données du disque dur. Et quand il y a plus d'un disque dur, xgboost utilise une méthode de gestion des bases de données dite « sharding » pour accélérer le temps d'accès.

Interprétabilité :

L'interprétabilité dans le gradient boosting comme dans xgboost peut également être évaluée ici de la même façon que dans les ensembles d'arbres comme random forest, c'est-à-dire la contribution de chaque variable à la réduction de la fonction objective ou à l'augmentation du score dans tous les arbres du modèle.

8- Light GBM

LightGBM est une amelioration apportée à gradient boosting à travers l'utilisation de deux nouvelles techniques : Gradient-based One-Side Sampling (GOSS) et Exclusive Feature Bundling (EFB) (Ke et al., 2017).

En effet, malgré les différentes innovations déjà apportées à gradient boosting par d'autres algorithmes tels que xgboost, l'efficience et la mise à l'échelle (scaling) reste limitée lorsqu'il s'agit de données à dimensions élevées et de large taille. Ce problème vient notamment de la nécessité de parcourir chaque instance de chaque variable pour sélectionner le point de partition optimal (de réduction maximale de la fonction objective entre le nœud parent et les nœuds enfants).

GOSS contribue à solutionner le problème et excluant une quantité significative d'instances à faible gradient, et en n'utilisant que le reste pour calculer le gain d'information (la réduction de l'impureté). GOSS peut ainsi faire une estimation précise et fiable du gain d'information à partir d'une taille d'échantillon significativement plus petite.

L'EFB regroupe des variables mutuellement exclusives (i.e. qui prennent rarement des valeurs non-nulles simultanément) pour réduire le nombre de variables. On arrive à trouver le regroupement optimal de variables mutuellement exclusives et réduire ainsi le nombre de variables, sans

significativement affecter la détermination des points de partition, et en utilisant un algorithme glouton (Ke et al., 2017).

Interprétabilité :

L'interprétabilité pour LightGBM passe aussi par l'importance de chaque variable, qui peut intrinsèquement être évaluée de deux façons :

- Par la variation totale de la fonction d'impureté dans les arbres du modèle
- Par le nombre de fois où la variable a été sollicitée dans les différents nœuds des arbres du modèle

9- Catboost

Catboost est un autre algorithme qui cherche à corriger une faille courante à savoir la *prediction shift* : une forme spéciale de fuite d'information dans les données (target leakage). La prediction shift est engendrée par deux phénomènes (Dorogush et al., 2018; Prokhorenkova et al., 2018) :

Traitement des variables catégorielles.

Les algorithmes de machine learning ne peuvent traiter que des données numériques. Les données catégorielles doivent donc être transformées en données numériques (« numéricalisées ») d'une façon ou d'un autre, soit pendant la phase de pré-traitement, soit – ce qui est souvent préférable – pendant la phase d'apprentissage.

Pour des variables catégorielles à faible cardinalité, la méthode la plus couramment utilisée est la disjonction complète (dummy coding ou one-hot encoding).

Pour des variables à forte cardinalité on peut procéder à ce qui est appelé target encoding (Micci-Barreca, 2001). Pour cela, on calcule une statistique à partir des valeurs de la variable cible (target) des données d'apprentissage. Par exemple, on peut remplacer chaque catégorie par la valeur moyenne de la variable cible sur l'ensemble des données d'apprentissage correspondant aux modalités de cette catégorie (mean target encoding) (Dorogush et al., 2018).

$Soit\ le\ jeu\ de\ données\ D = \{(X_i, Y_i)\}\ \ i = 1, \dots, n$

$avec\ X_i = \left(x_i^1, \dots, x_i^m\right)\ un\ vecteur\ de\ m\ variables\ et\ Y_i\ la\ cible$

$pour\ une\ variable\ catégorielle\ X^k\ on\ substitue\ la\ modalité\ x_i^k\ par$

$$\frac{\sum_{j=1}^{n}\left[x_j^k = x_i^k\right].Y_j}{\sum_{j=1}^{n}\left[x_j^k = x_i^k\right]}$$

$$o\grave{u}\ \left[x_j^k = x_i^k\right] = 1\ si\ \ x_j^k = x_i^k\ et\ 0\ sinon$$

Cette méthode utilise donc des catégories qui sont basées sur la variable cible, ce qui engendre automatiquement un surapprentissage (overfitting).

Pour corriger cette faille, catboost utilise une permutation aléatoire $\sigma = (\sigma_1, \dots, \sigma_n)$ et

$la\ modalité\ x_{\sigma_p}^k\ est\ substituée\ par$

$$\frac{\sum_{j=1}^{n}\left[x_{\sigma_j}^k = x_{\sigma_i}^k\right].Y_{\sigma_j} + a.P}{\sum_{j=1}^{n}\left[x_{\sigma_j}^k = x_{\sigma_i}^k\right] + a}$$

où $P\ repr\acute{e}sente\ une\ valeur\ apriori\ (probabilit\acute{e}\ a\ priori\ d'\hat{e}tre\ dans\ la\ classe\ positive)$

$et\ a\ >0\ son\ coefficient(un\ param\grave{e}tre\ )$

Traitement des biais de gradient.

Comme tous les autres types de gradient boosting, catboost utilise chaque nouvel arbre construit pour approximer les gradients du model courant. Mais les gradients utilisés à chaque étape sont estimés avec les mêmes données qui ont servi à construire le modèle. Ceci engendre un problème de biais et d'overfitting dus à un écart dans la distribution des gradients estimés par rapport à la distribution effective (Dorogush et al., 2018).

Pour résoudre les biais de gradient, catboost procède en deux étapes :

- Il choisit la meilleure structure d'arbres en énumérant les différentes partitions, construit des arbres avec ces partitions, définit des valeurs pour chaque feuille obtenue, évalue la performance de chaque arbre et choisit la meilleure partition.
- Les valeurs des feuilles dans chaque étape sont calculées comme approximations des gradients.

Soit F^i le modèle construit après la construction des i premiers arbres.

Soit $g^i(X_k, Y_k)$ la valeur du gradient sur le $k-i\grave{e}me$ observation de l'échantillon d'apprentissage à la phase i.

Pour que le gradient $g^i(X_k, Y_k)$ ne soit pas biaisé au regard de F^i, on doit entraîner F^i sans X_k

C'est ce que réalise catboost. Pour les détails, cf. (Dorogush et al., 2018; Prokhorenkova et al., 2018)

Enfin, catboost réalise plusieurs permutations de l'échantillon d'apprentissage à chaque séquence de construction du modèle.

Inteprétabilité :

L'interprétabilité pour catboost passe aussi par l'importance de chaque variable, qui peut intrinsèquement être évaluée de plusieurs façons :

- Par la variation totale de la fonction d'impureté dans les arbres du modèle
- Par la variation de la performance dans l'ensemble des arbres du modèle
- Par la valeur de l'interaction pour chaque pair de variable
- Par les SHAP values qui permettent d'évaluer le rôle et l'importance d'une variable de façon agnostique (indépendante du modèle utilisé) – thème que nous détaillerons davantage ultérieurement

10- Réseaux de Neurones Feed Forward (Perceptrons Multicouches)

Principe :

Ces réseaux de neurones dits « Artificial Neural Networks » (ANN) sont essentiellement des modèles statistiques non-linéaires constitués des éléments suivants :

- Une première couche d'entrée (input), les variables explicatives ;
- Une ou plusieurs couches « cachées » (hidden layers) constituées chacune de plusieurs neurones (units) ;
- Une couche de sortie (output), la variable à expliquer.

Chaque neurone des couches intermédiaires et de la couche de sortie est constitué de la composition de deux fonctions : une fonction non linéaire dite « d'activation » appliquée à une fonction linéaire, elle-même appliquée aux neurones de la couche précédente.

$$Soit\ la\ fonction\ d'activation\ \sigma_c\ de\ la\ couche\ c$$

$$Soit\ Z_c\ le\ vecteur\ des\ neurones\ de\ la\ c-ième\ couche\ du\ réseau$$

$$Soit\ [W_c]\ la\ matrice\ poids\ associée\ aux\ neurones\ de\ la\ couche\ c$$

$$Soit\ B_c\ le\ vecteur\ biais\ de\ la\ couche\ c$$

$$Alors:\ Z_c = \sigma_c([W_c]Z_{c-1} + B_c)$$

Les dimensions de la matrice poids W_c sont donc égales à :

$$(nombres\ de\ neurones\ de\ la\ couche\ [c]) \times (nombre\ de\ neurones\ de\ la\ couche\ [c-1])$$

Le nombre de neurones de la couche de sortie et la fonction d'activation correspondante dépendent du type de problème traité :

- Pour une classification à K classes, on envisage K neurones avec une fonction d'activation softmax donnant la probabilité de chaque classe x_k :

$$(softmax)_k(X) = \frac{\exp(x_k)}{\sum_{i=1}^{K} \exp(x_i)}$$

- Pour une classification binaire, $K = 1$ et la fonction d'activation est la sigmoïde :

$$\sigma(X) = \frac{1}{1 + \exp(-X)}$$

- Pour une régression, $K = 1$ et la fonction d'activation est l'Identité :

$$\sigma_o(X) = X$$

Ainsi donc la fonction d'apprentissage $f\ telle\ que\ Y = f(X)$ est une succession de compositions de fonctions. On détermine cette fonction par optimisation (minimisation) d'une fonction objective, typiquement : l'erreur quadratique pour un problème de régression :

$$L(W) = \left\| Y - \hat{f}_W(X) \right\|^2$$

Et la déviance (crossentropy) pour la classification :

$$L(W) = -\sum_{i,k} y_{ik} ln \hat{f}_{W,k}(x_i)$$

- L'algorithme d'optimisation utilisée est la descente du gradient :

$$W \leftarrow W - \alpha \frac{\partial}{\partial W} L(W)$$

$$o\grave{u}\ \alpha\ représente\ le\ taux\ d'apprentissage(learning\ rate)$$

Interprétabilité :

Pour les perceptrons multicouches, nous utilisons des méthodes d'interprétabilité indépendantes des modèles (Model-Agnostic Methods). Celles-ci peuvent apporter une explication globale comme l'importance des variables par permutation (permutation feature importance). Elles peuvent

également fournir une explication locale comme LIME ou SHAP (Molnar, 2022) que nous développerons ultérieurement.

Limites des modèles classiques

Les différents modèles présentés jusque-là sont généralement plus adaptés pour des données tabulaires et/ou structurées. On peut également les utiliser pour traiter des données séquentielles comme des textes, des images ou de la vidéo mais ils s'avèrent souvent beaucoup moins performants que les modèles dits de « deep learning » (Tunstall et al., 2022, Ravì et al., 2017, Asadullah et al., 2021; dos Santos et al., 2019).

Dans cette thèse, nous nous intéressons également à la prédiction des différents indicateurs de qualité de soins à partir des données cliniques en texte libre, essentiellement des données non structurées. La particularité de ces données (textuelles) est qu'elles exigent de tenir compte du contexte d'apparition de chaque terme (ou token), c'est-à-dire que le sens d'un mot dépend des mots qui l'entourent (i.e. qui le précédent et le suivent) dans une fenêtre donnée. La séquence des mots n'est donc pas arbitraire et tout traitement de données qui efface cette séquence engendre une importante perte d'information. Les effets de cette perte d'information sont ainsi plus considérables dans des tâches comme la traduction ou la génération de textes que dans des tâches comme la classification de textes ou le topic modeling. En effet, dans ces derniers cas, le contexte peut être partiellement retrouvé à travers la cooccurrence des mots, même si l'ordre des séquences stricto sensu a été perdu.

Dans le cas qui nous intéresse, les modèles de deep learning permettent de prédire différents indicateurs de qualité de soins en tenant compte du contexte de chaque terme (token) des compte-rendus cliniques, ce qui peut potentiellement améliorer significativement la capacité de données textuelles à prédire ces indicateurs, et à éventuellement apporter une interprétabilité plus détaillée des facteurs de risques (les prédicteurs). En théorie, plusieurs types de modèles peuvent être utilisés tels que les Convolutional Neural Networks (CNN) ou les Recurrent Neural Networks (RNN), plus spécifiquement les modèles dits Bi-LSTM (Bidirectional Long Short Term Memory). Cependant, pour des raisons à la fois théoriques et pratiques, nous nous contenterons de ne présenter que le plus performant et le plus complet des modèles en Natural Language Processing (NLP) : les Transformers.

Les Transformers (selon l'approche Hugging Face)

L'architecture des Transformers est constituée de deux parties :

- **Encodeur** : convertit une séquence de textes tokenisés (ou simplement de tokens) en une séquence de vecteurs embedding encore appelés *hidden state* ou *context* dans la mesure où ces vecteurs tiennent compte non seulement des représentations des tokens eux-mêmes, mais aussi du contexte dans lequel ils sont utilisés.
- **Décodeur** : utilise les vecteurs contextes pour générer itérativement une séquence de tokens, un token à la fois

Cependant, si les Transformers ont été originalement conçus pour des tâches d'encodage-décodage (sequence to sequence ou many to many), le bloc d'encodage et le bloc de décodage ont vite été adaptés comme des modèles autonomes (à part entière).

- Le modèle d'encodage seul transforme une séquence de texte en une représentation numérique riche (embedding) particulièrement bien adapté à des tâches comme la classification. C'est cet aspect que nous mobilisons dans cette thèse, notamment avec le modèle BERT (Bidirectional Encoder Representations from Transformers). Une particularité de ce modèle est d'être bidirectionnel dans le sens où l'encodage tient à la fois compte des tokens avant et après le token à encoder ;
- Le modèle de décodage quant à lui cherche à prédire le token suivant, sachant la séquence de tokens précédents, et ce de façon itérative. L'encodage est donc considéré comme unidirectionel, de la gauche vers la droite.

Dans cette thèse nous n'aurons recours qu'à la partie encodeur, et c'est donc celle-ci que nous allons développer.

Partie Encodeur

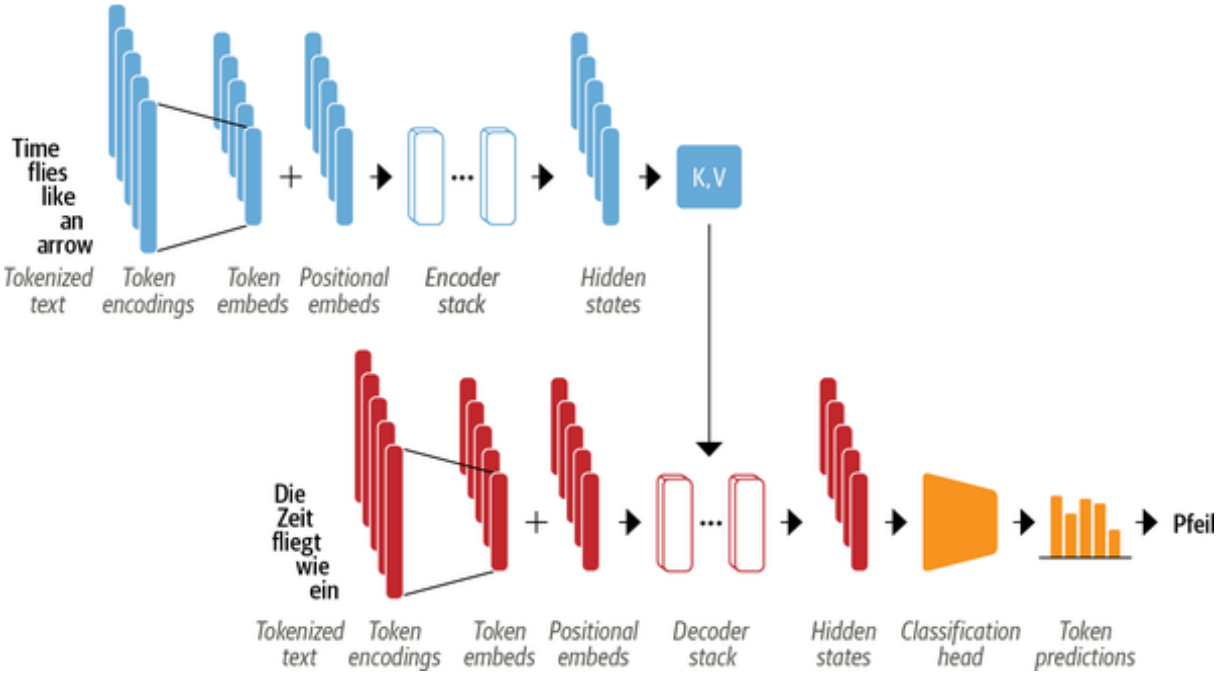


Figure 1-3 – Architecture des Transformers – d'après (Tunstall et al., 2022, p. 58)

Les algorithmes de Machine Learning et de Deep Learning ne peuvent pas directement traiter les textes. Il est donc indispensable de transformer ces derniers en nombres. Ce processus se déroule en deux phases complétées par l'architecture de l'encodeur. Il en résulte la succession de couches suivantes :

1- Tokenization

Cela consiste à découper une séquence de textes en unités atomiques utilisées par le modèle (tokens). Cette décomposition peut se passer au niveau de chaque caractère, ou de chaque mot. Les Transformers utilisent une tokenisation spéciale à mi-chemin entre ces deux options : *subword tokenization*.

Exemple (Tunstall et al., 2022):

$$text = "Tokenizing\ text\ is\ a\ core\ task\ of\ NLP."$$

$$Tokens = [\text{'}[CLS]\text{'}, \text{'}token\text{'}, \text{'}\#\#izing\text{'}, \text{'}text\text{'}, \text{'}is\text{'}, \text{'}a\text{'}, \text{'}core\text{'}, \text{'}task\text{'}, \text{'}of\text{'}, \text{'}nl\text{'}, \text{'}\#\#p\text{'}, \text{'}.\text{'}, \text{'}[SEP]\text{'}]$$

On observe deux choses dans cet exemple :

- La tokenisation engendre deux caractères spéciaux [CLS] et [SEP] qui correspondent respectivement au début et à la fin de la séquence.
- Certains mots comme tokenizing ont été décomposés en deux tokens, le préfixe ## indiquant que les caractères qui les précèdent ne sont pas des espaces blancs et que lorsque l'on reconvertit les tokens en séquences, les tokens avec ces préfixes doivent être fusionnés avec le token qui le précède.

2- Encodage

Les tokens sont ensuite associés chacun à un identifiant numérique (un entier naturel). Dans notre exemple cela donne (Tunstall et al., 2022) :

$$Encoded_{text} =$$

$$\{\text{'}input\_ids\text{'}: [101, 19204, 6026, 3793, 2003, 1037, 4563, 4708, 1997, 17953,$$

$$2361, 1012, 102], \ldots \}$$

La tokenization crée un dictionnaire de tokens identifiés chacun par un nombre, et la taille du dictionnaire varie facilement de 20k à 200k. Si on utilise la taille du vocabulaire comme espace vectoriel, on aurait des dimensions beaucoup trop importantes et peu performantes au regard des ressources que les modèles associés peuvent mobiliser. Il est donc de mise de procéder à une réduction de dimension. On retiendra la méthode du *word embedding* qui procède à la vectorisation des tokens par un processus d'apprentissage, tenant notamment compte du contexte (Mikolov et al., 2013).

On procède donc à deux étapes d'embedding, la première utilisant directement un modèle de Transformer (exemple : BERT), et la deuxième contenant l'information relative des positions des tokens. L'encodeur du transformer est constitué en réalité de blocs d'encodeurs de même que la partie décodeur est constituée de blocs de décodeurs.

Ces blocs d'encodeurs sont alors envoyés dans un mécanisme dit « Multihead-Self-Attention ». L'idée majeure derrière le « Self-Attention » est qu'au lieu d'utiliser une représentation vectorielle

(embedding) unique pour chaque token, nous pouvons utiliser toute la séquence pour calculer une moyenne pondérée de chaque embedding. Puisque cette vectorization tient compte de tout le contexte séquentiel, on l'appelle « contextual embedding »

3- La couche « Multihead-Self-Attention »

En pratique, cette couche applique trois transformations linéaires indépendantes à chaque embedding pour générer les vecteurs Q (Query), K (Key) et V (Value). Ces transformations projettent les embeddings, et chaque projection apporte son propre ensemble de paramètres qui sont appris et permettent à cette couche d'auto-attention de se concentrer sur différents aspects sémantiques de la séquence – pour le détails, cf. (Vaswani et al., 2017).

4- Scaled dot-product Attention

Une fonction d'attention peut être décrite comme une application de trois vecteurs Q (Query), K (Key) et V (Value) vers un vecteur attention pondéré par une fonction de compatibilité. En particulier le « scaled dot-product attention » est donné par :

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Autrement dit, plus les vecteurs Q et K sont proches (ou compatibles), plus grand sera le coefficient de V et donc plus V aura de l'influence dans le mécanisme d'attention.

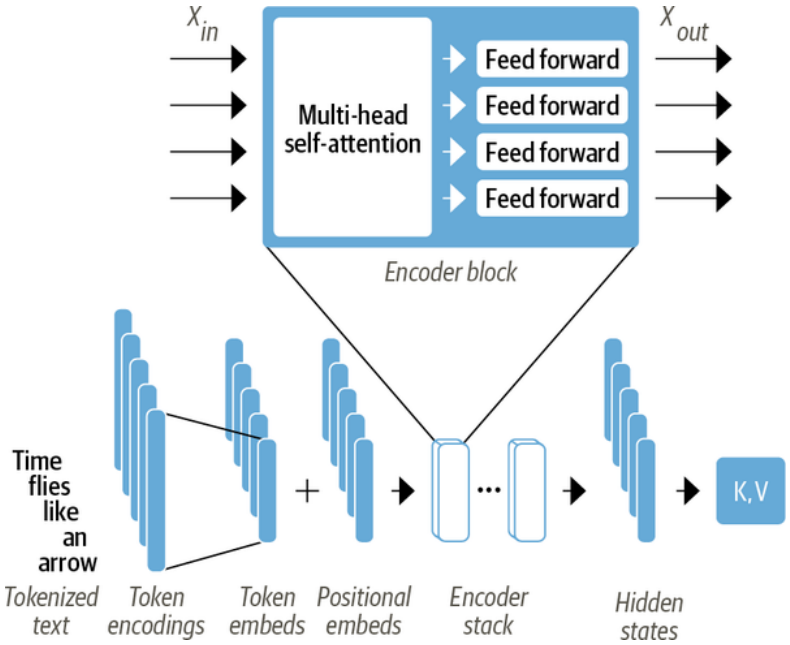


Figure 1-3 – La couche multi-headed self-attention (Tunstall et al., 2022, p. 60)

5- Multi-head Attention

Au lieu d'appliquer la fonction d'attention ci-avant une seule fois dans l'espace vectoriel de dimension du model d_{model} on réalise h projections linéaires sur trois différents sous-espaces contenant les vecteurs Q (Query), K (Key) et V (Value) de dimensions respectives d_K, d_K, d_V et on applique en parallèle la fonction d'attention

$$Multihead(Q, K, V) = Concat(head_1, \dots, head_h)W^o$$

$$o\grave{u}\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

$$et\ W_i^Q, W_i^K, W_i^V\ sont\ les\ matrices\ de\ projection$$

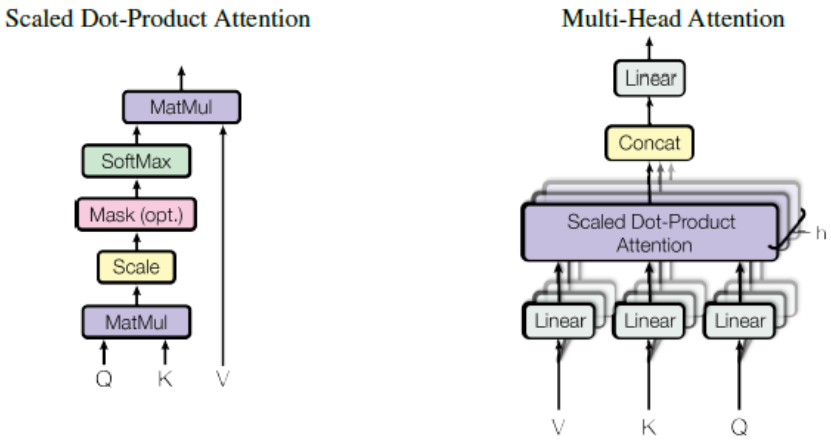


Figure 1-4 – Fonction d'Attention d'après (Vaswani et al., 2017, p. 4)

6- Couches Réseau de Neurones + Normalization

L'encodeur des Transformers contient également deux couches denses de réseaux de neurones qui au lieu de traiter la séquence d'embeddings comme un seul vecteur, traite chaque embedding de façon indépendante. Pour cette raison cette couche est souvent appelée « couche feed-forward de position ». On peut voir cela comme des neurones de convolution de fenêtre 1. Typiquement, dans la littérature, la taille de la première couche cachée de ces deux couches denses est de 4 fois la taille des embeddings, avec une fonction d'activation $GELU$ (Hendrycks & Gimpel, 2020).

La fonction d'activation GELU (Gaussian Error Linear Unit) est définie comme suit :

$$Soit\ \Phi(x)\ la\ fonction\ de\ répartition\ de\ la\ distribution\ gaussienne\ standardisée$$

$$GELU(x) = x\Phi(x) = xP(X \leq x)\ avec\ X \sim \mathcal{N}(0,1)$$

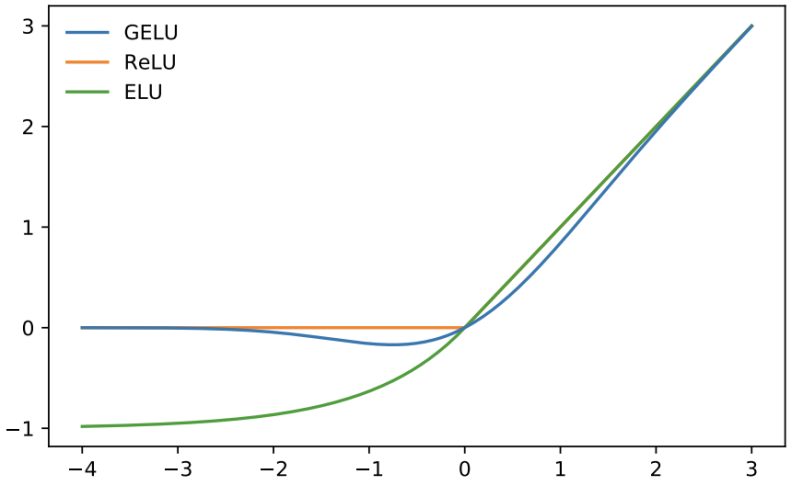


Figure 1-5 – Fonction d'activation GELU

Les Transformers utilisent également des couches de normalisation et de skip-connections. La normalisation consiste à transformer une variable en un vecteur de moyenne = 0 et d'écart-type 1. Un skip-connection transfère un tenseur à la couche suivante sans le traiter pour être ensuite ajouté aux tenseurs qui ont été traités. Les couches de normalisation peuvent être placées en position ante ou post la couche de réseaux de neurones feed-forward.

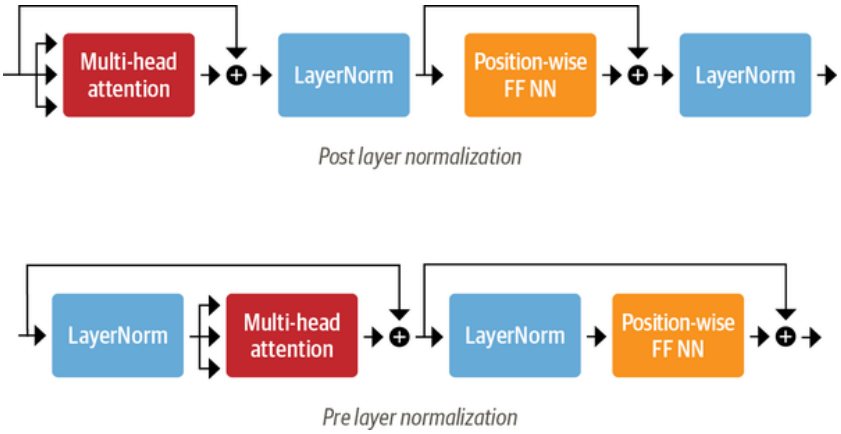


Figure 1-6 – Couches de reseaux de neurones feed-forward + normalisation.

D'après (Tunstall et al., 2022, p. 72)

7- Couche embedding de position

Le principe revient à faire une extension de l'embedding des tokens avec un embedding de position qui tient compte de la position de chaque token de la séquence. On procède à l'embedding de position de la même façon qu'on réalise l'embedding des tokens mais en utilisant un index de position (position_ids) comme entrée plutôt qu'un index d'identification (input_ids).

L'embedding final retenu est la somme des embeddings de token avec l'embedding de position.

8- Interprétabilité

Pour mieux expliquer la classification de textes avec les Transformers on peut utiliser des méthodes locales qui peuvent être généralisées au niveau global, comme LIME et SHAP.

Le problème de l'interprétabilité en Machine Learning

En santé publique comme dans de nombreuses autres disciplines, il ne suffit pas de savoir qu'un modèle est un bon, très bon ou excellent prédicteur d'un phénomène ou d'une variable d'intérêt, il est aussi important de savoir pourquoi, ou au minimum comment le modèle en question procède pour arriver à ces prédictions et les éléments déterminants de ces prédictions (Doshi-Velez & Kim, 2017), ne serait-ce que parce que la loi peut l'exiger[8].

La question de l'interprétabilité est devenue de plus en plus centrale en Machine Learning (Miller, 2019), notamment pour répondre aux nombreuses critiques sur la dimension « boîte noire » de ces différents modèles (Guidotti et al., 2018). C'est également devenu aujourd'hui pratiquement un champ à part entière que nous ne sommes pas en mesure de couvrir de façon satisfaisante et encore moins exhaustive dans cette thèse (Linardatos et al., 2021). Nous nous contenterons donc de présenter les méthodes d'explicabilité les plus courantes, et en particulier celles que nous avons mobilisées.

Une classification des méthodes d'interprétabilité

Celle-ci peut être réalisée en fonction de plusieurs critères (Molnar, 2022), présentées ici comme des questions.

1- Intrinsèque ou post-hoc ?

Intrinsèque fait référence au fait que l'explicabilité est inhérente à la structure même des modèles – d'une certaine façon à leur simplicité. C'est le cas par exemple pour les modèles linéaires et les arbres de décision.

Les méthodes *post-hoc* sont quant à elles mises en œuvre après l'apprentissage. Elles peuvent aussi être utilisées sur les modèles intrinsèquement interprétables. « Permutation feature importance » est un exemple type de méthode post-hoc.

Les méthodes intrinsèques

Peuvent être essentiellement différenciées en fonction des résultats présentés (Molnar, 2022) :

- ***Des statistiques synthétiques*** (descriptives) : par exemple une valeur par prédicteur ou bien des pairs d'intensité d'interactions entre des prédicteurs.
- ***Des résumés visuels*** : exemple les « Partial Dependency Plots » (PDP).
- ***Les paramètres internes*** : exemple les coefficients de régressions linéaires ou la visualisation des « features detectors » dans les réseaux de neurones de convolution (CNN).
- ***Des points de données*** : les observations elles-mêmes sont interprétables, par exemple en identifiant des prototypes de classes prédites ou par des explications contrefactuelles.
- ***Approximation par un modèle intrinsèquement interprétable***.

2- Spécifique au modèle ou indépendant du modèle ?

Sont spécifiques aux modèles des méthodes qui ne peuvent être appliquées qu'à ce type de modèle. Par exemple, l'interprétation des paramètres d'une régression linéaire est spécifique au modèle, dans la mesure où on ne peut pas l'appliquer à d'autres modèles.

[8] https://gdpr.eu/

Les méthodes indépendantes des modèles peuvent être utilisées à n'importe quel modèle de Machine Learning après entraînement (Post Hoc). Elles analysent généralement la relation entre chaque variable d'entrée et la variable de sortie par pairs et n'ont pas accès à des informations internes aux modèles, comme les poids (paramètres).

3- Local ou global ?

La méthode d'interprétation explique-t-elle une prédiction individuelle ou bien le comportement de la totalité du modèle ? Ou bien se situe-t-elle quelque part entre ces deux pôles ? Cela couvre la question de la portée du modèle :

- ***La Transparence ou à quel point le modèle s'éloigne de la boîte noire***. Il s'agit de déterminer comment l'algorithme crée le modèle et notamment d'expliciter la relation entre les variables explicatives et la variable à expliquer. Par exemple, les modèles linéaires sont plutôt explicites sur le rôle des paramètres dans la réalisation des prédictions, ce qui n'est pas le cas des réseaux de neurones profonds.

Au-delà de la transparence l'interprétabilité peut être considérée à différents niveaux

- ***L'interprétabilité globale et holistique du modèle***. On peut être en mesure de déterminer comment le modèle entraîné fait ses prédictions de façon globale en tenant compte de chacune de ses composantes. On peut constater que même pour un modèle linéaire, au-delà de quelques variables explicatives il est difficile de visualiser les choses et donc de satisfaire ce niveau d'interprétabilité.
- ***L'interprétabilité globale du modèle à un niveau modulaire***. On peut cette fois chercher à déterminer de quelle façon différentes parties du modèle influencent les prédictions. Typiquement dans un modèle linéaire multivariée, on est en mesure d'expliquer l'impact de la variation d'un prédicteur en contrôlant tous les autres – même si on n'est pas en mesure de déterminer comment ces prédicteurs agissent simultanément sur la variable à expliquer.
- ***L'interprétabilité locale d'une prédiction***. On veut cette fois-ci déterminer pourquoi le modèle émet une certaine prédiction pour une instance (observation) particulière. Un modèle peut-être très complexe par exemple, mais lorsque l'on s'intéresse très localement à son comportement, il peut correspondre à un modèle linéaire.
- ***L'interprétabilité locale d'un groupe de prédictions***. On cherche alors déterminer pourquoi le modèle réalise ces prédictions particulières pour un groupe d'observations. On peut alors soit appliquer une méthode globale aux instances de ce groupe, soit en chercher une explication individuelle à chaque instance pour ensuite agréger sur tout le groupe.

Exemples de modèles intrinsèquement interprétables

1- Les modèles linéaires (généralisés)

$Soient\ Y\ la\ variable$ à $expliquer\ suivant\ une\ loi\ \ de\ la\ famille\ exponentielle,$

$\boldsymbol{X} = [1, X^1, \dots, X^p]\ la\ matrice\ des\ prédicteurs, g\ une\ fonction\ inversible\ dite\ de\ lien,$

$et\ \boldsymbol{\beta} = [\beta_0, \dots, \beta_p]\ les\ paramètres\ du\ modèles$ (à $estimer$)

$Un\ modèle\ lineaire\ généralisé\ vérifie : g(E(y|x)) = \boldsymbol{X\beta}$

L'interprétabilité du modèle est ici globale et au minimum modulaire, étant donné qu'une variation unitaire du prédicteur X^j, toutes choses étant égales par ailleurs, correspond à une variation de $g(E(y|x))$ égale à β_j

Exemples :

- Pour la régression linéaire, g est la fonction Identité, donc une variation d'une unité de X^j entraîne une variation de y égale à β_j
- Pour la régression logistique, g est la fonction $logit$, donc une variation d'une unité de X^j entraîne une variation du $logit(P(y = 1|x)$ égale à β_j

De plus, **l'importance des variables** peut être évaluée en utilisant la statistique de Wald, celle du score ou celle de la Déviance (équivalent au test de rapport de vraisemblance). Plus ces statistiques sont élevées en valeurs absolue, plus importante est le prédicteur.

Enfin, au niveau local, on peut explicitement prédire ET expliquer les résultats de chaque nouvelle instance.

2- Les arbres de décision

Dans les arbres de décision, chaque nouvelle observation peut être associée à une région de l'espace des prédicteurs correspondant à une valeur agrégée de la variable d'intérêt (une moyenne ou une proportion de valeurs de y dans la région considérée). On se situe donc ici dans une interprétabilité de type globale et modulaire.

$$f(\boldsymbol{X}) = \sum_{m=1}^{M} C_m \,.\, \mathbb{I}_{(\boldsymbol{X} \in R_m)}$$

$$o\grave{u}\ les\ R_m\ repr\acute{e}sentent\ les\ diff\acute{e}rentes\ partitions\ de\ l'espace\ des\ variables$$

Comme nous l'avons déjà signalé, on peut aussi déterminer l'importance intrinsèque de chaque variable par sa contribution globale dans la variation de la fonction objective dans la totalité de l'arbre.

Au niveau local, toute nouvelle observation, peut être explicitement située dans une région des prédicteurs permettant d'en déduire la prédiction.

Exemples de méthodes globales indépendantes des modèles

1- Importance de variables par permutation (Permutation feature importance : PFI)

La méthode PFI est de nature globale. Le principe est simple : pour chaque prédicteur on effectue une permutation aléatoire des instances et on calcule la variation de la performance (ou de l'erreur) résultante. Cette variation mesure l'importance du prédicteur. Plus cette variation est grande, plus le prédicteur est important. Idée initialement proposée pour le Random Forest (Breiman, 2001), elle a été récemment étendue à tous les modèles par Fisher et al. (2019), résumée par l'algorithme suivant :

Soit $\hat{f}$ la fonction estimée par l'algorithme d'apprentissage, X la matrice de prédicteurs, y la variable cible et $L(y, \hat{f})$ la mesure de l'erreur agrégée correspondante.

1. On estime la valeur de L avant permutation par $e_{orig} = L(y, \hat{f}(X))$
2. Pour $j \in \{1,2, \dots, p\}$, répéter :

a. Générer une nouvelle matrice de prédicteurs X_{perm} où les instances de X_j ont été aléatoirement permutées
b. Estimer $e_{perm} = L(y, \hat{f}(X_{perm}))$
c. L'importance du prédicteur j est donnée par $FI_j = \frac{e_{perm}}{e_{orig}}$ ou par $FI_j = e_{perm} - e_{orig}$

3. Trier les variables par valeurs descendantes de FI (puis éventuellement normaliser en divisant par le max des FI et en multipliant par 100)

Cette méthode présente des inconvénients, lorsque les prédicteurs sont corrélés ou multicolinéaires :

- On voit bien que si deux variables sont corrélées et qu'on permute l'un d'entre elles, on se retrouve avec des observations qui n'ont aucun fondement dans la réalité, par exemple un individu qui fait 2m et 50kg. Alors quelle crédibilité accorder à la variation du score ou de la fonction objective dans un tel cas ?
- La présence d'un prédicteur corrélée à un autre peut diminuer l'importance d'une variable dans la mesure où les informations qu'elle contient sont disponibles dans l'autre variable, de telle sorte que lorsque l'on efface cette information via la permutation, elle reste disponible dans l'autre, au moins en partie.

Des alternatives à ces différents inconvénients ont été envisagées et ne seront pas traitées dans cette thèse et ceci d'autant plus que chaque approche d'interprétabilité apporte ses propres lots d'inconvénients (Fisher et al., 2019; Wei et al., 2015).

2- Modèle global de substitution (Global surrogate model)

Ici aussi, le principe est simple : il s'agit de réaliser une approximation de la fonction « boîte noire » f aussi fidèlement que possible avec une fonction intrinsèquement interprétable g telle qu'un modèle linéaire généralisé ou un arbre de décision. Ci-après les étapes pour obtenir un modèle de substitution :

1. Choisir un jeu de données de prédicteurs $\boldsymbol{X}$ – idéalement le même que celui utilisé pour entraîner le modèle « boîte noire » f , sinon issu de la même distribution.
2. Obtenir les prédictions du modèle f pour les données $\boldsymbol{X}$
3. Sélectionner un modèle intrinsèquement interprétable g
4. Entraîner le modèle g sur les données $\boldsymbol{X}$ et les prédictions du modèle f
5. Mesurer à quel point le modèle g est en mesure répliquer le modèle f
6. Interpréter le modèle de substitution g

La performance évaluée au point 5 doit idéalement utiliser une mesure de performance qui culmine à 100% (R^2, Variance Expliquée ou R^2 ajusté, Accuracy, F1, ROC AUC, PRC AUC)

Exemple de méthodes locales indépendantes des modèles

1- Modèle de substitution locale (LIME)

LIME (Local Interpretable Model-agnostic Explanations) est un modèle de substitution entraîné pour expliquer des prédictions individuelles (Ribeiro et al., 2016).

Le principe : étant donné un modèle boîte noire f, LIME teste ce qui arrive aux prédictions lorsqu'on procède à une variation des données dans f. LIME engendre de nouvelles données issues de perturbations dans l'échantillon et des prédictions correspondantes de f. Sur ces nouvelles données, LIME entraîne un modèle intrinsèquement interprétable g, pondéré par la proximité entre les

observations de l'échantillon et les instances considérées. Le modèle entraîné g devrait être une bonne approximation de f localement, mais pas nécessairement globalement (fidélité locale) (Molnar, 2022) :

$Soit\ L\ la\ fonction\ objective\ du\ modèle\ f, g\ le\ modèle\ inteprétable\ de\ substitution$

$de\ complexité\ \Omega, et\ \pi_x\ une\ mesure\ de\ la\ proximité\ du\ voisinage\ de\ l'instance\ x.$

$Le\ modèle\ d'explication\ est\ donné\ par$:

$$explication(x) = \underset{gEG}{\operatorname{argmin}}\{L(f, g, \pi_x) + \Omega(g)\}$$

Autrement dit, le modèle d'explication pour l'instance x est le modèle g qui minimise la fonction objective (Loss) L, laquelle mesure la proximité de la prédiction de f avec g tout en minimisant la complexité $\Omega(g)$ du modèle. G représente la famille de toutes les explications possibles (exemple, l'ensemble des modèles linéaires). La mesure de proximité π_x détermine l'étendue du voisinage de x que l'on considère dans l'explication.

Les étapes explicites pour entraîner un modèle de substitution locale sont fournies ci-après :

1. Choisir l'instance d'intérêt dont on cherche à expliquer la prédiction via la boîte noire f;
2. Procéder à une perturbation des données et récupérer les prédictions de ces données perturbées ;
3. Pondérer les nouvelles données en fonction de leur proximité de l'instance d'intérêt ;
4. Entraîner un modèle pondéré intrinsèquement interprétable sur les données perturbées ;
5. Expliquer la prédiction en interprétant le modèle local.

Un bon choix de g est le Lasso qu'on entraîne en jouant avec les valeurs du paramètre de régularisation λ.

Remarque :

Il existe au moins deux problèmes sérieux avec LIME :

1. Le problème concernant la définition du voisinage n'est pas résolu pour les données tabulaires (Garreau & Luxburg, 2020)
2. Les explications sont encore instables et peuvent notamment passer à côté de sévères biais (Alvarez-Melis & Jaakkola, 2018; Slack et al., 2020)

2- Shapley Values

On peut considérer qu'une prédiction est constituée d'une part de joueurs (chaque prédicteur du modèle), et de l'autre, d'un gain (ou récompense), qui est la prédiction. Les Shapley Values nous indiquent comment répartir équitablement les gains parmi les prédicteurs. La Shapley value (H. W. Kuhn, 2020) est une méthode pour attribuer les gains en fonction de leur contribution au gain total.

Se référant donc à la théorie classique des jeux, ici le jeu est représenté par la tâche de prédire pour une instance des données. Le gain est la différence entre la prédiction effective de cette instance et la prédiction moyenne pour toutes les instances. Les joueurs sont les valeurs des prédicteurs qui « collaborent » pour obtenir le gain (prédiction spécifique) associé à l'instance considérée.

La Shapley value pour un prédicteur est la contribution marginale moyenne du prédicteur à travers toutes les possibles coalitions. Autrement dit :

- On considère tous les sous-ensembles possibles des prédicteurs sans le prédicteur d'intérêt ;

- Pour chacun de ces sous-ensembles, on prédit l'outcome avec et sans le prédicteur d'intérêt ;
- On calcule la différence de prédiction pour obtenir la contribution marginale du prédicteur ;
- La Shapley value est la moyenne des contributions marginales.

Explicitement, la Shapley Value du prédicteur j est donnée par :

$$\phi_j(val) = \sum_{S \subseteq \{1,\dots,p\} \backslash \{j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\,(val_x(S \cup \{j\}) - val_x(S))$$

$$val_x(S) = \int \hat{f}\,(x_1, \dots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

De plus la Shapley Value satisfait des propriétés dites :

Efficiency : la somme des contributions de chaque prédicteur est égale à la différence entre la prédiction de l'instance x et la moyenne des prédictions de toutes les instances.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

Symmetry : la contribution de deux prédicteurs j et k devraient être la même si elles contribuent de façon égale à toutes les coalitions.

$$Si\ \forall\ S \subseteq \{1, \dots, p\} \backslash \{j, k\}\ \ val_x(S \cup \{j\}) = val_x(S \cup \{k\})\ \ alors\ \phi_j = \phi_k$$

Dummy : un prédicteur j qui ne change pas la valeur prédite – quelle que soit la coalition à laquelle elle est ajoutée – devrait avoir une Shapley value de 0.

$$Si\ \forall\ S \subseteq \{1, \dots, p\}\ \ val_x(S \cup \{j\}) = val_x(S)\ \ alors\ \phi_j = 0$$

Additivity : pour un jeu à gains combinés $val + val^+$ le Shapley values sont données par :

$$\phi_j + \phi_j^+$$

Remarques :

- Le temps de calcul de la Shapley value est généralement prohibitif, ce qui rend nécessaire de passer par une approximation, notamment par une méthode Monte-Carlo (Štrumbelj & Kononenko, 2014).
- Les Shapley values ne renvoient pas à un modèle de prédiction.
- L'existence d'une corrélation entre les prédicteurs peut être problématique comme chaque fois qu'on a recours à des permutations pour déterminer l'importance des variables.
- Les Shapley values ne s'interprètent pas comme une mesure de la contribution unique d'un prédicteur lorsqu'on l'injecte dans le modèle ou qu'on en efface l'information. On ne peut l'expliquer indépendamment des autres prédicteurs du modèle.

3- SHAP (Shapley Addition exPlanations)

Basée sur les Shapley values, SHAP (Lundberg & Lee, 2017), est une méthode pour expliquer les prédictions individuelles. Contrairement aux Shapley values qui sont exclusivement locales, SHAP propose plusieurs méthodes d'interprétation globale basées sur l'agrégation des Shapley values. SHAP cherche à faire le lien entre LIME et les Shapley values.

Le but de SHAP est d'expliquer la prédiction d'une instance x en calculant la contribution de chaque prédicteur à la prédiction. Par rapport à la Shapley value classique, pour SHAP l'explication de la

Shapley value est représentée comme un modèle linéaire – c'est ce qui la rapproche du LIME linéaire :

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

Où M est le nombre maximal de coalition, g le modèle de substitution, $z' \in \{0,1\}^M$ le vecteur de coalition (simplified inputs), ϕ_j la Shapley value pour le prédicteur j

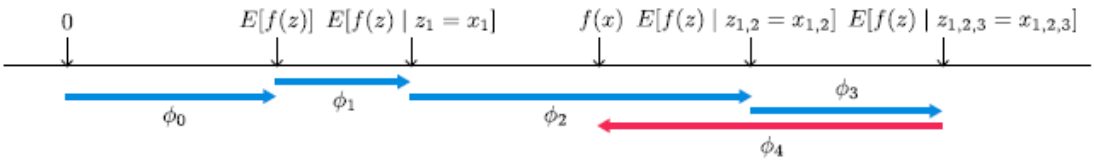


Figure 1: SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value $E[f(z)]$ that would be predicted if we did not know any features to the current output $f(x)$. This diagram shows a single ordering. When the model is non-linear or the input features are not independent, however, the order in which features are added to the expectation matters, and the SHAP values arise from averaging the ϕ_i values across all possible orderings.

Figure 1-7 - D'après (Lundberg & Lee, 2017, p. 5)

KernelSHAP (LIME Linéaire + Shapley Values)

Estime pour une instance x, les contributions de la valeur de chaque prédicteur à la prédiction de x.

Les solutions de l'équation

$$explication(x) = \underset{gEG}{\operatorname{argmin}}\{L(f, g, \pi_x) + \Omega(g)\}$$

Sont associées aux résultats suivants :

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}\,|z'|(M-|z'|)}$$

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z')$$

$|z'|$: le nombre d'éléments présents (non-nuls) dans l'instance z'

Les coefficients estimés par le modèle, les ϕ_j sont naturellement les Shapley values.

Autres variantes de SHAP

Il existe d'autres variantes de SHAP optimisées pour des familles de modèles de machine learning :

- Pour les familles d'arbres (Lundberg et al., 2020)
- Pour le Deep Learning (H. Chen et al., 2021)
- Pour les Transformers

4- SHAP pour l'explicabilité globale

Les Shapley values peuvent être combinées en explications globales. Si on lance SHAP pour chaque instance, on obtient une matrice de Shapley values : une ligne par instance et une colonne par prédicteur.

Importance des Variables (SHAP Feature Importance)

Il suffit juste de faire la moyenne par colonne de la matrice de Shapley values (Figure 1-7) :

$$I_j = \frac{1}{n}\sum_{i=1}^{n}\left|\phi_j^{(i)}\right|$$

Ci-après deux illustrations données sur le site web de Lundberg.

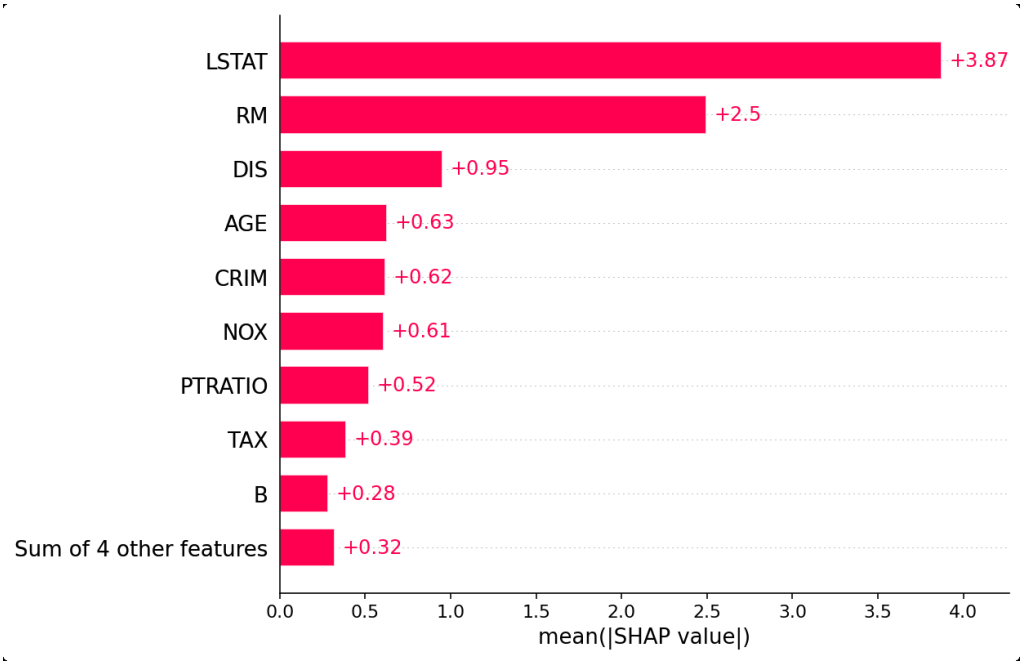


Figure 1-8 – from https://github.com/slundberg/shap

Shap Summary Plot

C'est un graphique qui représente simultanément l'importance des variables et leurs effets.

Chaque point du graphique est une Shapley value pour une instance et un prédicteur. L'abscisse représente les Shapley values et l'ordonnée l'importance de la variable.

Plus une variable est haute sur l'axe des y plus elle importante. Si elle est positive, c'est qu'elle pousse à augmenter la valeur de la prédiction ; si elle est négative elle pousse vers sa réduction.

Si elle est rouge cela signifie qu'elle a une valeur élevée pour cette ligne et si elle est bleue c'est qu'elle a une valeur faible sur cette ligne.

Par exemple (Figure 1-8), LSTAT est une variable importante pour le modèle et elle varie dans le sens opposé des prédictions : plus elle est élevée plus elle tend à diminuer la valeur de la prédiction et plus elle est faible, plus elle tend à l'augmenter

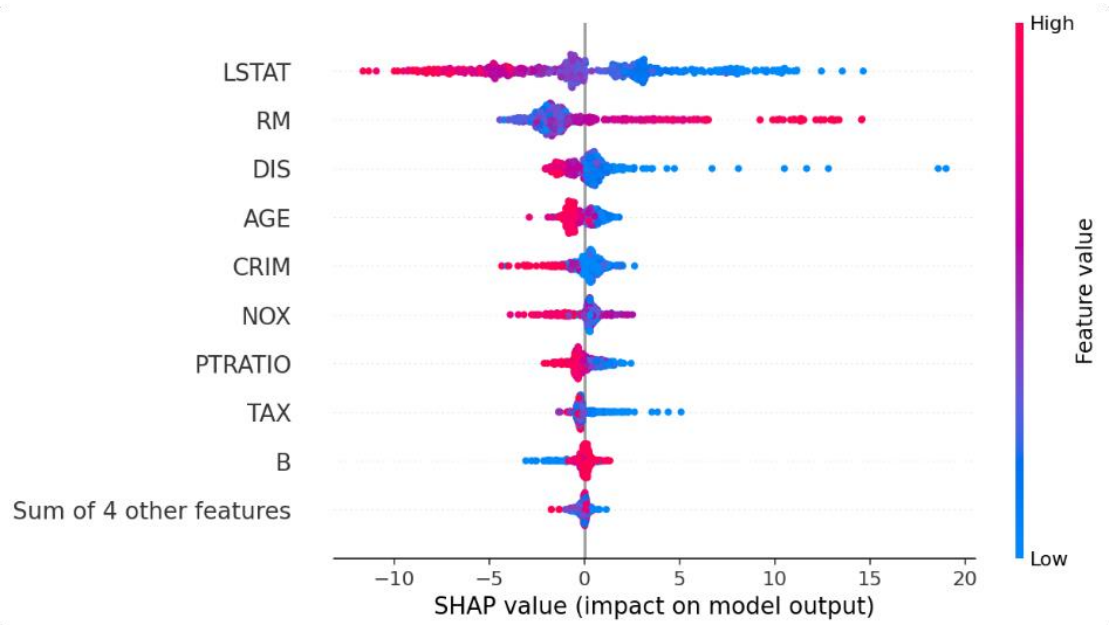


Figure 1-9 – from https://github.com/slundberg/shap

Remarques:

SHAP présente l'avantage d'avoir de solides fondements théoriques dans la théorie des jeux, de proposer une importance de variable qui reflète bien la contribution de chaque prédicteur et de proposer une explication basée sur le contraste par rapport à la prédiction moyenne.

SHAP permet également d'unifier différentes approches d'explicabilité en Machine Learning dont LIME et les Shapley values, DeepLIFT, etc. (Lundberg & Lee, 2017). On peut même se permettre de combiner interprétabilité locale et globale.

Malheureusement, SHAP a pour inconvénient majeur d'avoir un temp de calcul long (davantage encore sur les données textuelles) ce qui peut encore passer si on ne s'intéresse qu'à quelques instances isolées. Mais cela rend pratiquement impossible d'obtenir des évaluations globales de l'importance des variables dès que les données sont un peu conséquentes.

De plus, le calcul de l'importance de nouvelles données exige d'avoir accès à toutes les données et l'interprétation demande une réelle prudence (Slack et al., 2020).

Partie 2 – Méthodologie[9]

[9] Github de codes de la thèse : https://github.com/jaotombo/these_ML_qualite_soins

Présentation des données

Dans cette thèse, nous avons utilisé deux jeux de données : la première est issue de l'APHM (Assistance Publique Hôpitaux de Marseille), et la seconde provient du Boston Beth Israel Deaconess Medical Center (Boston BIDMC).

Les données de l'APHM sont, comme toutes les données médicales, sensibles et protégées par des lois très strictes – notamment par le RGPD (Règlement Général sur la Protection des Données).

L'accès à ces données est donc soumis à des droits très spécifiques qui ne sont pas nécessairement ouverts aux doctorants. En l'occurrence, les données utilisées ici ont été extraites par les ingénieurs du Département d'Informatique Médicale (DIM), dans un cadre limité pour la recherche et disponible uniquement sur site. Malgré la très grande richesse des informations disponibles dans les bases de données administratives et du PMSI (Programme de Médicalisation des Systèmes d'information), les données telles que les informations sur la biologie ou sur les notes cliniques n'ont été pas facilement disponibles, et de toutes façons quand les notes cliniques ont été disponibles, les conditions de leur exploitation n'ont pas pu être réunies.

En conséquence, seule une partie des données structurées de l'APHM ont été mobilisées dans cette thèse.

Fort heureusement, les bases de données MIMIC III (Medical Information Mart for Intensive Care III) du BIDMC nous ont permis de poursuivre la recherche sur des données également très riches, contenant des informations sur la biologie des patients et des comptes-rendus cliniques très détaillés.

Jeu de données de l'APHM

On considère ici tous les séjours en soins-aigus[10] (hospitalisation complète) du 1er Janvier au 31 Décembre 2015 en excluant les soins ambulatoires (chirurgie ambulatoire, radiothérapie, dialyse, chimiothérapie, transfusions) et la mortalité hospitalière.

Les données comportent 118650 lignes et 19 colonnes (voir Tableau 2-1 ci-après) :

- ***Des variables sociodémographiques :***
    - Age : l'âge des patients mesuré en années.
    - Age_cat : on remplace généralement l'âge par la catégorie d'âge pour mieux identifier les profils de patients (moins de 18 ans, 18-44 ans, 45-64 ans, 65-84 ans, 85 ans ou plus).
    - Sexe : sexe biologique.
    - AME (State Funded Medical Assistance) : l'aide médical de l'état : « est un dispositif permettant aux étrangers en situation irrégulière de bénéficier d'un accès aux soins »[11].

[10] Les soins aigus sont une branche des soins de santé secondaires où un patient reçoit un traitement actif mais à court terme pour une blessure grave ou un épisode de maladie, une condition médicale urgente ou pendant la convalescence après une intervention chirurgicale

[11] https://www.service-public.fr/particuliers/vosdroits/F3079

- CMU (Free Universal Health Care) : couverture maladie universelle : « La complémentaire santé solidaire vous aide pour vos dépenses de santé. Elle remplace la couverture maladie universelle complémentaire (CMU-C) »[12].

- ***Des variables cliniques :***
  - Sévérité : la sévérité associée au séjour hospitalier définie en 3 catégories (aucune ou faible, modérée, haute).
  - Catégorie de maladies fournie par le diagnostic associé (DA) : « un diagnostic ou une pathologie majorant l'effort de soins et les moyens utilisés, par rapport à la morbidité principale. Les DA influencent le groupage des séjours dans un Groupe Homogène de Malades GHM[13] »[14]. Une variable à 24 modalités.
  - La comorbidité mesurée par les 17 indicatrices de Charlson (Quan et al., 2005).

- ***Des variables hospitalières :***
  - Durée de séjour (LOS) : le temps écoulé entre les coordonnées temporelles d'admission et de décharge.
  - Type de séjour (MCO) : Médical, Obstétrique ou Chirurgical.
  - Origine du patient : Domicile ou bien autres institutions hospitalières.
  - Destination du patient : Domicile ou bien autres institutions hospitalières.
  - Le patient est-il admis via le service d'urgence : Oui / Non.
  - Le patient a-t-il été admis dans un service dans les 6 précédents mois : Oui / Non.
  - Le patient a-t-il été réadmis de façon non planifiée via les urgences sous 30 jours : Oui / Non.

Les variables à expliquer dans le jeu de données de l'APHM

On cherchera à expliquer alternativement la réadmission sous 30 jours via urgence ou la durée de séjour, lesquels sont deux indicateurs de qualité de soins encore pertinents à ce jour.

Tableau 2-1 : Variables du Jeu de Données APHM*

Sociodemographic Characteristics	Modality
Gender	1-Male
State Funded Medical Assistance	1-Yes
Free Universal Health Care	1-Yes
Age	Quantitative

*(en anglais pour maintenir la continuité avec les articles tous rédigés en anglais)

[12] https://www.service-public.fr/particuliers/vosdroits/F10027

[13] Un groupe homogène de malades regroupe les prises en charge de même nature médicale et économique et constitue la catégorie élémentaire de classification en MCO (Médecine, Chirurgie, Obstétrique).

[14] https://solidarites-sante.gouv.fr/professionnels/gerer-un-etablissement-de-sante-medico-social/financement/financement-des-etablissements-de-sante-10795/financement-des-etablissements-de-ante-glossaire/article/diagnostic-associe-da

Hospitalization Characteristics	Modality
Length of Stay	Quantitative
Type of Hospital Stay	1-Medical
	2- Obstetrics
	3-Surgical
Origine of Patient	1-Home
Hospitalization via Emergency Departments	1-Yes
Destination on discharge	1-Home
At least one previous hospitalization via emergency departments 6 months before	1-No hospitalization
	2-At least one non-emergency
	3-At least one with emergency

Tableau 2-1 : Variables du Jeu de Données APHM

Clinical Characteristics	Modality
Category Of Disease	01-Digestive
	02-Orthopedic – Trauma
	03-Multiple or complex trauma
	04-Rheumatology
	05-Nervous system
	06-Vascular catheterization
	07-Cardiovascular
	08-Pulmonary
	09-Ear Nose and Throat - Stomatology
	10-Ophthalmology
	11-Gynecology-Breast
	13-Uronephrology and reproductive organs
	14-Hematology
	15-Chemotherapy - radiotherapy
	16-Infectious diseases
	17-Endocrinology
	18-Cutaneous and subcutaneous
	19-Burns
	20-Psychiatry
	21-Toxicology - Intoxication - Alcohol
	22-Chronic pain palliative care
	23-Organ Transplant
	24-Interdisciplinary activities and follow-up of patients
Severity	1-None or Low
	2-Moderate - High
	3-Not Determined (short stay)

Tableau 2-1 : Variables du Jeu de Données APHM

Charlson Comorbidities	
Renal Disease (CH)	1-Yes
Rheumatologic Disease (CH)	1-Yes
Peripheral Vascular Disease (CH)	1-Yes
Peptic Ulcer Disease (CH)	1-Yes
Hemiplegia or Paraplegia (CH)	1-Yes
Moderate or Severe Liver Disease (CH)	1-Yes
Mild Liver Disease (CH)	1-Yes
Metastatic Solid Tumor (CH)	1-Yes
Any Malignancy including Leukemia and Lymphomia (CH)	1-Yes
AIDS/HIV (CH)	1-Yes
Diabetes with Chronic Complications (CH)	1-Yes
Diabetes without Chronic Complications (CH)	1-Yes
Dementia (CH)	1-Yes
Cerebrovascular Disease (CH)	1-Yes
Chronic Pulmonary Disease (CH)	1-Yes
Congestive Heart Failure (CH)	1-Yes
Myocardial Infarction (CH)	1-Yes

La base de données MIMIC III

Le MIMIC III est publiquement disponible pour la recherche sous condition d'une formation de quelques heures et d'un engagement écrit concernant les contraintes de son utilisation.

Il s'agit d'une grande base de données comprenant des données anonymisées liées à la santé, associées à plus de quarante mille patients qui ont séjourné dans les unités de soins intensifs du Beth Israel Deaconess Medical Center (BIDMC) entre 2001 et 2012.

La base de données comprend des informations telles que les données démographiques, les mesures des signes vitaux effectuées au chevet du patient ($\sim 1\ point\ de\ données\ par\ heure$), les résultats des tests de laboratoire, les procédures, les médicaments, les notes des soignants, les rapports d'imagerie et la mortalité (à l'intérieur et à l'extérieur de l'hôpital).

MIMIC est adapté à un large éventail d'études analytiques en épidémiologie, à l'amélioration des règles de décision cliniques et au développement d'outils électroniques. Il est caractérisé par trois dimensions essentielles :

- disponible gratuitement pour les chercheurs du monde entier ;
- englobe une population diversifiée et très importante de patients en soins intensifs ;
- contient des données à haute résolution temporelle, y compris les résultats de laboratoire, la documentation électronique, les tendances et les courbes des moniteurs de chevet.

Le MIMIC III est structuré comme une base de données relationnelle dont les clés primaires sont typiquement ROW_ID, SUBJECT_ID et HADM_ID selon ce l'on cherche à réaliser. Les variables retenues dans l'analyse sont issues de ces différentes tables obtenues par différentes requêtes (cf. github de la thèse à l'adresse suivante : https://github.com/jaotombo/these_ML_qualite_soins)

Les Variables MIMIC III retenues

Ces variables vont être sélectionnées pour l'essentiel dans les tables de données suivantes :

- ADMISSONS et PATIENTS pour les données sociodémographiques telles que ethnicity, admission_type, insurance, religion, marital_status, readmit_dt, Gender, DOB (Date of Birth)
- LABEVENTS, CHARTEVENTS, CPTEVENTS, OUTPUTEVENTS pour les données d'analyses biologiques (urea, platelets, magnesium, albumin, calcium, resprate, hr, sysbp, diaspb, temp)
- TRANSFERS pour les parcours du patient : origin_patient, dest_discharge
- D_ICD_DIAGNOSES pour déterminer le code ICD9 (ou CIM9) par chapitre : International Classification of Disease Chapters[15].
- NOTEEVENTS contient les comptes rendus et autres notes cliniques dont les notes de décharges, discharge

D'autres variables ont été calculées selon des algorithmes directement fournies sur le site internet du MIMIC[16], en l'occurrence nous avons choisi d'inclure deux scores de sévérité :

- le SAPS (Simplified Acute Physiology Score) II (Le Gall et al., 1993)
- le SOFA (Sepsis-related Organ Failure Assessment) (Singer et al., 2016)

L'ensemble des variables, leur modalité et les tables de la base de données relationnelles où elles ont été extraites sont fournies ci-après (Tableau 2-2) :

[15] https://icd.codes/icd9cm

[16] https://github.com/MIT-LCP/mimic-code/tree/main/mimic-iii/concepts/severityscores

Tableau 2-2 : Variables du Jeu de Données MIMIC III

Sociodemographic Characteristics	Modality	MIMIC III Tables
Ethnicity	asian	admissions
	black	
	hispanic	
	white	
	unknown	
	other	
Insurance	medicaid	admissions
	medicare	
	government	
	private	
	self pay	
Religion	catholic	admissions
	jewish	
	protestant quaker	
	undefined	
	other	
Marital Status	couple	admissions
	separated	
	single	
	unknown	
	widowed	
Gender / Sex	1-Male	patients
	2-Female	
Age	< 18 years	admissions + patients (computed)
	18-44 years	
	45-64 years	
	65-84 years	
	85+ years	

Tableau 2-2 : Variables du Jeu de Données MIMIC III

Hospitalization Characteristics	Modality	MIMIC III Tables
Admission Type	emergency	admissions
	elective	
	urgent	
Admission Location	home	admissions
	other	
ICU LOS	quantitative	icustays
Readmission Time Delta	quantitative	admissions (computed)
Type Of Stay	1-medical	services (computed)
	2-obstetrics	
	3-surgical	
At least one previous hospitalization via emergency departments 6 months before	1-no hospitalization	admissions (computed)
	2-at least one non-emergency	
	3-at least one with emergency	
Hospitalization via Emergency Departments	yes	admissions
	no	
Origin of Patient	other	admissions (computed)
	home	
Destination on discharge	home	admissions
	other	
Discharge Notes	text	noteevents
Length of Hospital Stay (LOS)	quantitative	admissions (computed)

Clinical Characteristics	Modality	MIMIC III Tables
ICD Chapter	Digestive System	diagnoses_icd + procedures_icd (computed)
	Respiratory System	
	Circulatory System	
	Neoplasms	
	Injury Poisoning	
	Genitourinary System	
	Symptoms Signs Ill-Defined Conditions	
	Musculoskeletal System Connective Tissue	
	Endocrine Nutritional Metabolic Immunity Disorders	
	Mental Disorders	
	Nervous System & Sense Organs	
	Complications Pregnancy Childbirth Puerperium	
	Skin Subcutaneous Tissue	
	Infectious Parasitic	
	Congenital Anomalies	

	Blood & Blood-Forming Organs	
	Supp Factors Health Status	

Tableau 2-2 : Variables du Jeu de Données MIMIC III

Clinical Characteristics	Modality	MIMIC III Tables
Simplified Acute Physiology Score (SAPS II)	quantitative	multiple tables (computed)
Sepsis-related Organ Failure Assessment (SOFA)	quantitative	multiple tables (computed)
Urea nitrogen min	quantitative	labevents (computed)
Urea nitrogen max	quantitative	
Urea nitrogen mean	quantitative	
Platelets min	quantitative	
Platelets max	quantitative	
Platelets mean	quantitative	
Magnesium max	quantitative	
Albumin min	quantitative	
Calcium min	quantitative	
Respiratory rate min	quantitative	chartevents
Respiratory rate max	quantitative	
Respiratory rate mean	quantitative	
Glucose min	quantitative	
Glucose max	quantitative	
Glucose mean	quantitative	
Heart Rate min	quantitative	
Heart Rate max	quantitative	
Heart Rate mean	quantitative	
Systolic Blood Pressure min	quantitative	
Systolic Blood Pressure max	quantitative	
Systolic Blood Pressure mean	quantitative	
Diastolic Blood Pressure min	quantitative	
Diastolic Blood Pressure max	quantitative	
Diastolic Blood Pressure mean	quantitative	
Temperature min	quantitative	
Temperature max	quantitative	
Temperature mean	quantitative	
Urine Ouput min	quantitative	outputevents
Urine Ouput max	quantitative	
Urine Ouput mean	quantitative	
Patient Weight	quantitative	chartevents

Enjeux autour des données déséquilibrées

Les situations de données déséquilibrées en classification sont nombreuses en Machine Learning (KaurHarsurinder et al., 2019) et tout particulièrement dans les Sciences de la Santé. Les deux jeux de données utilisées dans cette thèse font appel à des variables d'intérêts très largement déséquilibrées.

On parle de données déséquilibrées lorsque les classes de la variable à expliquer ont des proportions suffisamment différentes pour que les performances des modèles utilisés soient biaisées en faveur de l'une d'entre elles (KaurHarsurinder et al., 2019; Wang & Sun, 2021).

Pour bien situer l'enjeu des données déséquilibrées, prenons l'exemple d'un jeu de données où un modèle de Machine Learning (classifieur) réalise une classification binaire avec une performance (en termes d'accuracy[17]) de 95%. Autrement dit le taux d'erreur de classement est de 5%. A première vue, cette performance est bonne, voire très bonne. Mais en réalité cette opinion est implicitement basée sur l'hypothèse d'un jeu de données relativement équilibrée – par exemple dans le cas d'une classification binaire, sur l'hypothèse que chaque classe contient approximativement 50% des observations, ou de façon plus étendue, sur une distribution uniforme de proportions approximativement égales à $1/k$ pour une classification à k classes.

Supposons cependant que nos données soit fortement déséquilibrées soit de 97% vs 3%. Un classifieur tout à fait trivial prédisant la classe majoritaire pour toutes les observations obtiendrait un accuracy de 97%, et réaliserait donc une meilleure performance que notre modèle de Machine Learning.

1- L'enjeu de la distribution des classes

Les données déséquilibrées posent un problème évident en termes d'apprentissage : le modèle voit beaucoup plus d'instances de la classe majoritaire que de la classe minoritaire et est dont susceptible de mieux apprendre de la première.

Une des façons de solutionner ce problème est de sous-échantillonner la classe majoritaire, mais il en résulte une évidente perte d'information donc un apprentissage moins complet qui se traduira par une perte de performance auprès d'un échantillon test (que le modèle n'a jamais « vu ») (KaurHarsurinder et al., 2019).

Une autre façon de solutionner le problème est de sur-échantillonner la classe minoritaire, mais il en résulte une tendance marquée au surapprentissage (overfitting) qui se traduira également par une perte de performance auprès d'un échantillon test (que le modèle n'a jamais « vu ») (KaurHarsurinder et al., 2019).

Une autre alternative, SMOTE (Chawla et al., 2002) combine un sur-échantillonnage de la classe minoritaire avec un sous-échantillonnage de la classe majoritaire. Le processus est décrit comme suit (Chawla et al., 2002, p. 328) :

« Nous générons des exemples synthétiques d'une manière moins spécifique à l'application, en opérant dans « l'espace des variables » plutôt que dans « l'espace des instances ». La classe minoritaire est suréchantillonnée en prenant chaque échantillon de classe minoritaire et en introduisant des exemples synthétiques le long des segments de ligne joignant n'importe lesquels des (tous les) k voisins les plus proches de la classe minoritaire. En fonction de la quantité de

[17] Le terme *Accuracy* est volontaire maintenu en anglais pour éviter la confusion avec *Precision* qui se traduit de la même façon en anglais

suréchantillonnage requise, les k plus proches voisins sont choisis au hasard. Notre implémentation utilise actuellement cinq plus proches voisins. Par exemple, si la quantité de suréchantillonnage nécessaire est de 200 %, seuls deux voisins parmi les cinq voisins les plus proches sont choisis et un échantillon est généré dans la direction de chacun. Les échantillons synthétiques sont générés de la manière suivante : Prenez la différence entre le vecteur de variables (échantillon) considéré et son voisin le plus proche. Multipliez cette différence par un nombre aléatoire entre 0 et 1, et ajoutez-le au vecteur de variables considéré. Cela provoque la sélection d'un point aléatoire le long du segment de ligne entre deux entités spécifiques. Cette approche force effectivement la région de décision de la classe minoritaire à devenir plus générale. (…)

La classe majoritaire est sous-échantillonnée en retirant au hasard des échantillons de la population de la classe majoritaire jusqu'à ce que la classe minoritaire devienne un pourcentage spécifié de la classe majoritaire. »

En résumé :

- On tire un échantillon aléatoire de la classe minoritaire.
- Pour les observations de cet échantillon, on identifie les k voisins les plus proches.
- On prend ensuite l'un de ces voisins et on identifie le vecteur entre le point de données actuel et le voisin sélectionné.
- On multiplie le vecteur par un nombre aléatoire entre 0 et 1.
- Pour obtenir le point de données synthétique, on l'ajoute au point de données actuel.

Cette opération revient en fait à déplacer légèrement le point de données dans la direction de son voisin. De cette façon, on s'assure que notre point de données synthétique n'est pas une copie exacte d'un point de données existant tout en vérifiant qu'il n'est pas non plus trop différent des observations connues dans la classe minoritaire.

2- Le problème de l'évaluation de la performance (métriques)

Un autre problème engendré par les données déséquilibrées se situe au niveau de la mesure de la performance. En effet, plusieurs des métriques couramment utilisées dans la classification (exemple : l'accuracy) présupposent implicitement l'équilibre des données. Par exemple, implicitement les algorithmes de classification binaire classent les observations en positif si la probabilité empirique estimée est égale 0.5 ou plus alors qu'il serait plus pertinent pour des données déséquilibrées de choisir un seuil plus élevé ou plus bas, selon que la classe majoritaire soit respectivement la classe positive ou la classe négative.

Une des façons de tenir compte de ce déséquilibre dans les métriques est de pondérer les estimations par le poids relatif de chaque classe (Guyon et al., 2015; Kelleher et al., 2015). D'autres améliorations ont été apportées récemment dans le cas de déséquilibres de classes indépendantes de la distribution des classes (leurs proportions)(Gupta et al., 2020).

Une autre façon de tenir compte de ces déséquilibres est de choisir un seuil de classification différent, c'est à dire de déterminer le seuil optimal s pour classer une prédiction en positif tel que :

$$\mathcal{P}(Y = 1|X) \geq s \Longrightarrow Y = 1$$

Par défaut (hypothèse équilibrée) $s = 0.5$

Le seuil s peut être choisi de façon à optimiser la métrique de notre choix : l'Accuracy, le F1 score, le ROC AUC, ou le PRC AUC (ci-après).

3- Les métriques utilisées en classification

Nous allons considérer principalement deux types de métriques :

Les métriques dépendantes du seuil de classification

Il s'agit de toutes les métriques qui sont basées sur la matrice de confusion résumée par le graphique suivant (cas de la classification binaire) (Tableau 2-3) :

Tableau 2-3 : Matrice de Confusion d'une réponse binaire

Matrice de Confusion	Négative Prédite	Positive Prédite	Totaux Observés
Négative Observée	VN (Vraie Négative)	FP (Fausse Positive)	Total Négatives
Positive Observée	FN (Fausse Négative)	VP (Varie Positive)	Total Positives
Totaux Prédits	Total Négatives Prédites	Total Positives Prédites	Taille échantillon

En découlent les métriques suivantes :

$$Accuracy = \frac{VN + VP}{Taille\ échantillon}$$

$$Taux\ d'Erreur = 1 - Accuracy$$

$$Sensibilité = \frac{VP}{Total\ Positives} = Recall$$

$$Spécificité = \frac{VN}{Total\ Négatives}$$

$$Précision = \frac{VP}{Total\ Positives\ Prédites}$$

$$F_1 = \frac{2.Precision.Recall}{Precision + Recall}$$

Pour les données déséquilibrées, ce sont principalement ces métriques qui doivent être l'objet de pondération, comme évoqué ci-avant (en 2-) afin de tenir compte de la distribution de chaque classe (Gupta et al., 2020).

Les métriques indépendantes du seuil de classifications

Le ROC AUC et le PRC AUC sont deux métriques tout à fait pertinentes qui ne dépendent pas des seuils de classification car elles tiennent compte de tous les seuils de classification possibles.

a. **Le ROC AUC ou AUROC**: Area Under the Receiver Operator Characteristics. Comme son nom l'indique, il s'agit de la surface contenue sous la courbe ROC. Cette dernière est obtenue en reliant tous les points d'un couple (tfp, tvp), lorsque l'on fait varier le seuil du maximum du score prédit (par exemple : la probabilité de la classe positive prédite par le classifieur) au minimum du score prédit, et où :

$$tfp = \frac{FP}{Total\ Négatives} = 1 - Spécificité$$

$$tvp = \frac{VP}{Total\ Positives} = Sensibilité = Recall$$

b. **Le PRC AUC ou AUPRC**: Area Under the Precision-Recall Curve. Comme son nom l'indique, il s'agit de la surface contenue sous la courbe Precision-Recall. Cette dernière est obtenue en reliant tous les points d'un couple $(Recall, Precision)$, lorsque l'on fait varier le seuil du maximum du score prédit (par exemple : la probabilité de la classe positive prédite par le classifieur) au minimum du score prédit.

Le choix du ROC AUC est justifié (Fawcett, 2006) non seulement parce qu'il est indépendant des seuils de classification mais également par la prise en compte des coûts d'erreurs d'allocation qui peuvent engendrer de sérieux biais pour les métriques dépendants des seuils de classification.

Dans la littérature médicale on utilise très largement le ROC AUC, mais rarement le PRC AUC. Pourtant, certains auteurs semblent indiquer que le ROC AUC pourrait être trop inflationniste lorsque la proportion de classe positive est très faible, auquel cas ces auteurs recommandent d'avoir plutôt recours aux PRC AUC (Cook & Ramadas, 2020; Ozenne et al., 2015).

c. **Le H-measure**. Hand (2009) formule également une autre critique du ROC AUC avec une solide démonstration mathématique à l'appui. Cet article est d'ailleurs largement cité à tel point que l'on se demande pourquoi la solution alternative qu'il propose ne semble toujours pas être adoptée.
L'argument de Hand (2009) est centré autour de l'idée centrale que lorsqu'on utilise le ROC AUC, les coûts des erreurs de classifications (ou d'allocations) dépendent du classifieur, ce qui ne devrait pas être le cas si on veut utiliser une métrique dépendante uniquement de la prédiction du classifieur mais pas de la distribution des données (condition vérifiée par le ROC AUC), ni des scores de classification qui dépendent du classifieur (condition non vérifiée par le ROC AUC).
C'est donc ce dernier point qui est problématique et soulevé par Hand : « AUC is equivalent to averaging the misclassification loss over a cost ratio distribution, which depends on the score distributions » (Hand, 2009, p. 109) qui propose d'utiliser un score indépendant du classifieur, à savoir le H-Measure. Cependant, cette première version du H-measure ne semble pas être optimale dans la mesure où elle fait une hypothèse sur la distribution statistique à utiliser qui ne tient pas vraiment compte des données déséquilibrées, ce qui a entraîné un ajustement plus récent (Hand & Anagnostopoulos, 2014).

Dans la pratique, le H-measure utilise une distribution $Beta$ qui ne peut pas être considérée comme objective puisqu'elle ne peut s'appliquer à tous types de problèmes. Il est sans doute plus avisé de comparer les performances de classifieurs en tenant compte de plusieurs métriques. Nous avons trouvé que même sans utiliser le H-measure, la combinaison des deux mesures de l'AUC (ROC et PRC) fournissent de très bonnes indications sur la performance relative des modèles entre-eux.

Article 1

Machine-learning prediction of unplanned 30-day rehospitalization using the French hospital medico-administrative database

Contexte

Cette publication explore s'il est effectivement pertinent d'utiliser des modèles de Machine Learning en lieu et place de la Régression Logistique classique utilisée dans les problèmes de classification en Sciences de la Santé. Elle constitue un prolongement d'un précédent article utilisant le même jeu de données de l'APHM (présenté ci-avant), et cherchant à créer un score de risque de réhospitalisation (Pauly et al., 2019).

A notre connaissance, c'est la première tentative d'appliquer des algorithmes de Machine Learning sur ces données et d'étudier s'il est pertinent à termes de déployer ceux-ci dans le système d'information d'un service hospitalier (notamment de l'APHM) afin d'améliorer la qualité des soins, la satisfaction des patients, la transition entre les services hospitaliers et le domicile, ainsi que la coordination entre les différents types de services (hospitalisation complète, de jour ou à domicile).

Utilisant la Régression Logistique (LR) classique comme référence, les autres algorithmes étudiées sont :

- La Régression Logistique Pénalisée (ELNET)
- Les arbres de décision (CART)
- Les Support Vector Machines (SVM)
- Le Random Forest (RF)
- Le Gradient Boosting (GBM)
- Les Réseaux de Neurones à une couche cachée (NNET)

Cependant pour la publication, ELNET n'a pas été retenue étant donné que les résultats sont pratiquement identiques à LR. SVM n'a pas été retenu pour des raisons de convergence – de fait toutes les expérimentations que nous avons réalisées au fil des années démontrent que pour des tailles d'échantillon au-delà de 30k lignes, les packages courants de SVM (généralement utilisant libsvm) tendent à ne converger que très difficilement, quel que soit le noyau utilisé. Or, nos données contiennent 118650 lignes.

Cette partie de l'étude a été réalisée sous R avec le package CARET (M. Kuhn, 2008).

Données déséquilibrées

Dans cette publication, la variable d'intérêt (cible, outcome, etc.) est binaire. Il s'agit de la réadmission non planifiée sous 30 jours via les urgences ($y = RHurg30$). Le codage de y est réalisé comme suit:

$$si\ réadmis\ sous\ 30\ jours, y = 1\ (positif)\ sinon\ y = 0\ (négatif)$$

La distribution de y est de 3.5% de positifs (4127 séjours) et 96.5% de négatifs, ce qui est extrêmement déséquilibré. Nous avons choisi l'option d'évaluer la performance des modèles avec deux indicateurs indépendants des seuils de classification : le ROC AUC et le H-measure.

Observational Study

Medicine®

OPEN

Machine-learning prediction of unplanned 30-day rehospitalization using the French hospital medico-administrative database

Franck Jaotombo, PhD[a,b], Vanessa Pauly, PhD[a,c], Pascal Auquier, MD, PhD[a], Veronica Orleans, PhD[c], Mohamed Boucekine, MD, PhD[a], Guillaume Fond, MD, PhD[a], Badih Ghattas, PhD[b], Laurent Boyer, MD, PhD[a,c,*]

Abstract

Predicting unplanned rehospitalizations has traditionally employed logistic regression models. Machine learning (ML) methods have been introduced in health service research and may improve the prediction of health outcomes. The objective of this work was to develop a ML model to predict 30-day all-cause rehospitalizations based on the French hospital medico-administrative database.

This was a retrospective cohort study of all discharges in the year 2015 from acute-care inpatient hospitalizations in a tertiary-care university center comprising 4 French hospitals. The study endpoint was unplanned 30-day all-cause rehospitalization. Logistic regression (LR), classification and regression trees (CART), random forest (RF), gradient boosting (GB), and neural networks (NN) were applied to the collected data. The predictive performance of the models was evaluated using the H-measure and the area under the ROC curve (AUC).

Our analysis included 118,650 hospitalizations, of which 4127 (3.5%) led to rehospitalizations via emergency departments. The RF model was the most performant model according to the H-measure (0.29) and the AUC (0.79). The performances of the RF, GB and NN models (H-measures ranged from 0.18 to 0. 29, AUC ranged from 0.74 to 0.79) were better than those of the LR model (H-measure = 0.18, AUC = 0.74); all P values <.001. In contrast, LR was superior to CART (H-measure = 0.16, AUC = 0.70), P < .0001.

The use of ML may be an alternative to regression models to predict health outcomes. The integration of ML, particularly the RF algorithm, in the prediction of unplanned rehospitalization may help health service providers target patients at high risk of rehospitalizations and propose effective interventions at the hospital level.

Abbreviations: AME = Aide Médicale d'Etat, APHM = Assistance Publique – Hôpitaux de Marseille, AUC = area under the curve, CART = classification and regression trees, CMU = Couverture Maladie Universelle, DT = decision tree, GB = gradient boosting, LOS = length of stay, LR = logistic regression, MDG = mean decrease in Gini, ML = machine learning, NN = neural networks, PMSI = Programme de Médicalisation des Systèmes d'Information, RF = random forest, ROC = receiving operating characteristic, VI = variable importance.

Keywords: health service research, machine learning, patient rehospitalization, prediction

1. Introduction

Reducing 30-day rehospitalizations is a priority of health care policies in Western countries.[1,2] Unplanned rehospitalizations are common[3,4] and costly,[4,5] reflecting poor quality inpatient care,[6–8] and poorly coordinated transitions between hospitals and homes.[9] Despite the growing literature on this issue, unplanned rehospitalizations are still poorly understood and controlled.[3] We need to better identify patients at high risk of rehospitalization to improve the quality of care and reduce rehospitalizations and associated health care costs.[10]

Editor: Phil Phan.

The authors have no funding and conflicts of interests to disclose.

Supplemental Digital Content is available for this article.

The data that support the findings of this study are available from a third party, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are available from the authors upon reasonable request and with permission of the third party.

[a] Aix-Marseille University, EA 3279 - Public Health, Chronic Diseases and Quality of Life - Research Unit, La Timone Medical University, 27, boulevard Jean-Moulin, [b] Mathematics Institute of Marseille, Aix-Marseille University, Marseille, France, [c] Service d'Information Médicale, Public Health Department, La Conception Hospital, Assistance Publique - Hôpitaux de Marseille, 147 Boulevard Baille, Marseille, France.

[*] Correspondence: Laurent Boyer, Aix-Marseille University, Faculté de Médecine - Secteur Timone, EA 3279: CEReSS -Centre d'Etude et de Recherche sur les Services de Santé et la Qualité de vie, 27 Boulevard Jean Moulin, Marseille 13005, France (e-mail: laurent.boyer@ap-hm.fr).

Résumé

Introduction. La prévision des réhospitalisations non planifiées a traditionnellement utilisé des modèles de régression logistique. Les méthodes de Machine Learning (ML) ont été introduites dans la recherche sur les services de santé et peuvent améliorer la prédiction des résultats pour la santé. L'objectif de ce travail était de développer un modèle de ML pour prédire les réhospitalisations toutes causes à 30 jours à partir de la base de données médico-administrative des hôpitaux français.

Méthodes. Il s'agit d'une étude rétrospective de cohorte de l'ensemble des sorties de l'année 2015 d'hospitalisations en soins aigus dans un centre universitaire de soins tertiaires regroupant quatre hôpitaux français. Le critère d'évaluation de l'étude était une réhospitalisation non planifiée sous 30 jours, toutes causes. Ont été appliqués aux données collectées : la régression logistique (LR), les arbres de classification et de régression (CART), le Random Forest (RF), le gradient boosting (GB) et les réseaux de neurones (NN). La performance prédictive des modèles a été évaluée à l'aide du H-measure et de l'aire sous la courbe ROC (AUC).

Résultats. Notre analyse a porté sur 118 650 hospitalisations dont 4 127 (3,5 %) ont conduit à des réhospitalisations via les urgences. Le modèle RF était le modèle le plus performant selon le H-measure (0,29) et l'AUC (0,79). Les performances des modèles RF, GB et NN (H-mesures variant de 0,18 à 0,29, AUC variant de 0,74 à 0,79) étaient meilleures que celles du modèle LR (H-mesure=0,18, AUC=0,74) ; toutes les valeurs p<0,001. En revanche, LR était supérieur à CART (mesure H = 0, 16, AUC = 0, 70), p <0, 0001.

Discussion. L'utilisation du ML peut être une alternative aux modèles de régression pour prédire les résultats de santé. L'intégration du ML, en particulier l'algorithme RF, dans la prédiction des réhospitalisations non planifiées peut aider les prestataires de services de santé à cibler les patients à haut risque de réhospitalisation et à proposer des interventions efficaces au niveau hospitalier.

Mots-clés : Apprentissage automatique ; Réhospitalisation des patients ; Prédiction; Recherche sur les services de santé.

Machine-learning prediction of unplanned 30-day rehospitalization using the French hospital medico-administrative database

Short title: Machine learning and rehospitalization

Jaotombo Franck (PhD)[1,2], Pauly Vanessa (PhD)[1,3], Pascal Auquier (MD, PhD)[1], Orleans Veronica (PhD)[3], Boucekine Mohamed (MD, PhD)[1], Fond Guillaume (MD, PhD)[1], Ghattas Badih (PhD)[2], Boyer Laurent (MD, PhD)[1,3]

1. Aix-Marseille University, EA 3279 - Public Health, Chronic Diseases and Quality of Life - Research Unit, La Timone Medical University, 27, boulevard Jean-Moulin, F-13005 Marseille.
2. Mathematics Institute of Marseille, Aix-Marseille University, Marseille, France.
3. Service d'Information Médicale, Public Health Department, La Conception Hospital, Assistance Publique - Hôpitaux de Marseille, 147 Boulevard Baille, F-13005 Marseille.

Category: Observational Study

Word count: 3,064

2 tables, 2 figures

52 references

* Correspondence should be sent to:

Prof. Laurent Boyer

Aix-Marseille Univ, Faculté de Médecine - Secteur Timone, EA 3279: CEReSS -Centre d'Etude et de Recherche sur les Services de Santé et la Qualité de vie, 27 Boulevard Jean Moulin, 13005 Marseille, France.

Tel: (33 686 93 62 76), e-mail: laurent.boyer@ap-hm.fr

List of abbreviations

AME: Aide Médicale d'Etat

APHM: Assistance Publique – Hôpitaux de Marseille

AUC: Area under the curve

CART: Classification and regression trees

CMU: Couverture Maladie Universelle

DT: Decision tree

GB: Gradient boosting

LOS: Length of stay

LR: Logistic regression

MDG: mean decrease in Gini

ML: Machine learning

RF: Random forest

ROC: Receiving Operating Characteristic

NN: Neural networks

PMSI: Programme de Médicalisation des Systèmes d'Information

VI: Variable importance

Abstract

Introduction. Predicting unplanned rehospitalizations has traditionally employed logistic regression models. Machine learning (ML) methods have been introduced in health service research and may improve the prediction of health outcomes. The objective of this work was to develop a ML model to predict 30-day all-cause rehospitalizations based on the French hospital medico-administrative database.

Methods. This was a retrospective cohort study of all discharges in the year 2015 from acute-care inpatient hospitalizations in a tertiary-care university center comprising four French hospitals. The study endpoint was unplanned 30-day all-cause rehospitalization. Logistic regression (LR), classification and regression trees (CART), random forest (RF), gradient boosting (GB) and neural networks (NN) were applied to the collected data. The predictive performance of the models was evaluated using the H-measure and the area under the ROC curve (AUC).

Results. Our analysis included 118,650 hospitalizations, of which 4,127 (3.5%) led to rehospitalizations via emergency departments. The RF model was the most performant model according to the H-measure (0.29) and the AUC (0.79). The performances of the RF, GB and NN models (H-measures ranged from 0.18 to 0. 29, AUC ranged from 0.74 to 0.79) were better than those of the LR model (H-measure=0.18, AUC=0.74); all p-values<0.001. In contrast, LR was superior to CART (H-measure=0.16, AUC=0.70), p<0.0001.

Discussion. The use of ML may be an alternative to regression models to predict health outcomes. The integration of ML, particularly the RF algorithm, in the prediction of unplanned rehospitalization may help health service providers target patients at high risk of rehospitalizations and propose effective interventions at the hospital level.

Keywords: Machine learning; Patient rehospitalization; Prediction; Health Service Research.

1- Introduction

Reducing 30-day rehospitalizations is a priority of health care policies in Western countries [1,2]. Unplanned rehospitalizations are common [3,4] and costly [4,5], reflecting poor quality inpatient care [6-8] and poorly coordinated transitions between hospitals and homes [9]. Despite the growing literature on this issue, unplanned rehospitalizations are still poorly understood and controlled [3]. We need to better identify patients at high risk of rehospitalization to improve the quality of care and reduce rehospitalizations and associated health care costs [10].

In a recent work [11], we developed an easy-to-use predictive rehospitalization risk score of unplanned 30-day all-cause rehospitalization using a logistic regression (LR) model based on 13 variables from the French hospital medico-administrative database (Programme de Médicalisation des Systèmes d'Information - PMSI). This predictive rehospitalization risk score yielded better discriminatory properties than the LACE index score [12] (c-statistic = 0.74 vs. 0.66, respectively). The LACE index score is one of the most widely used predictive tools in the world and the current instrument recommended by the French Health Authority. Despite this improvement, this new score presented moderate discriminative ability and needs to be more accurate. The other prediction models in the literature present similar properties, with c-statistics of approximately 0.70 (e.g., hospital score = 0.72 [13]). The common point among the prior work is to use traditional statistical methods such as logistic regression (LR) models. Recently, machine learning (ML) methods have been introduced in health service research and have shown a better level of prediction than traditional statistical approaches in several domains [14-21]. ML methods offer key benefits over traditional statistical approaches because they account for nonlinear relationships between the outcome and the predictors and yield more stable predictions [22]. ML methods account for interactions between predictors which relaxes the homogeneity assumption that there are no interactions among predictors. To our knowledge, ML methods have rarely been applied to improve the prediction of all-cause rehospitalization. A recent study [23] developed models using an ML approach to predict 30-day all-cause rehospitalization in patients hospitalized for heart failure but without prediction improvement when compared to LR models (c-statistics<0.61). Another recent study [24] reported that automated ML better predicted readmissions than commonly used readmission scores in 3 US hospitals (n =16 649).

Thus, the objective of this work was to compare the predictive performance of traditional logistic and ML models to predict 30-day all-cause rehospitalizations on a large population-based study from the French hospital medico-administrative database, based on the following two criteria: the area under the receiving operating characteristic curve and the H-measure. For this purpose, we selected the best ML methods: random forest (RF), neural networks (NN) and gradient boosting (GB)[25], which we compared with 2 reference methods: LR and classification and regression trees (CART) methods.

2- Methods

Study design

This was a retrospective cohort study of all acute-care inpatient hospitalization cases discharged from January 1 to December 31, 2015, from the largest university health center in south France (Assistance Publique – Hôpitaux de Marseille, APHM). All data were collected from the French Hospital database (PMSI - Programme de Médicalisation des Systèmes d'Information) [26]. The PMSI is the French medico-administrative database for all hospitalizations based on diagnosis-related groups that we could group into significant diagnostic categories. Research on such retrospective data are excluded from the framework of the French Law Number 2012-300 of 5 March 2012 relating to the

research involving human participants, as modified by the Order Number 2016-800 of 16 June 2016. Neither the French competent authority (Agence Nationale de Sécurité du Médicament et des Produits de Santé, ANSM) approval nor the French ethics committee (Comités de Protection des Personnes, CPP) approval is required in this context.

Study setting and inclusion criteria

The APHM is a public tertiary-care center comprising four hospitals (La Timone, La Conception, Sainte-Marguerite, and North) with 3,400 beds and 2,000 physicians. Approximately 300,000 hospitalizations are recorded every year at the APHM, involving approximately 210,000 patients. All acute-care hospitalizations were included in this study. We excluded hospitalizations in the ambulatory care unit (i.e., ambulatory surgery, radiotherapy, dialysis, chemotherapy, and transfusions) as well as in-hospital mortalities.

Study outcome

The study outcome was unplanned 30-day all-cause rehospitalization (a binary variable where positive rehospitalization is coded y=1), defined as any cause of readmission via emergency departments in any acute care wards within 30 days of discharge. To calculate this outcome, a unique and individual PMSI identifying variable was used to track rehospitalizations, 30 days following discharge. No more than one rehospitalization for each discharge was taken into account. Readmission via the emergency department was employed to identify unplanned rehospitalizations [27].

Collected data

The dataset collected from the PMSI used 29 predictor variables based on a previous work [11]:

- sociodemographic characteristics: age, gender, state-funded medical assistance (Aide Médicale d'Etat, AME) (i.e., health coverage for undocumented migrants), and free universal health care (Couverture Maladie Universelle, CMU) (i.e., universal health coverage for those not covered by employment/business-based schemes);

- clinical characteristics: category of disease based on the 10th revision of the International Statistical Classification of Diseases, disease severity (no or low severity, moderate – high severity or not determined for short hospitalizations) based on an algorithm issued from the PMSI and 17 comorbidities from the Charlson comorbidity index [28] (supplementary file 4);

- hospitalization characteristics: patient origin (home or other hospital institution), hospitalization via emergency departments, LOS, destination after hospital discharge (home or transfer to other hospital institution), and hospitalization via emergency departments in the previous 6 months.

Statistical models

Five distinct types of predictive models were fitted to the data: LR considered as the reference, CART, RF, GB, and one hidden-layer NN. These models have been explained elsewhere in detail [29]; a brief summary is presented here.

LR is a linear model of the exponential family such that $\ln\left(\frac{\pi}{1-\pi}\right) = w^T x$, where π = P(y=1|x) and w is the weight vector to be estimated from the data.

CART [30] is a binary decision tree (DT) method that involves segmenting the predictor space into a number of simple regions. CART can be applied to both regression and classification problems, as in our study. A DT is constructed through an iterative process by applying a binary splitting rule. For

each variable x_j in the data, a rule of the form $x_j < a$ (a ∈ R is a threshold) is used to split the initial set of observations (denoted t_0, the root of the tree) into two subsets t_l and t_r (the sibling nodes). Each observation falling in those regions is then predicted by the highest frequency class. The best split is defined as the one minimizing a loss function (i.e., the Gini index). Once the best split has been defined, the same process is applied to the two nodes t_l and t_r and repeated until a predefined minimum number of observations is reached. Then, a pruning algorithm can be used to search for an optimal tree, given a penalty criterion (complexity parameter) applied to the objective function. A DT can be represented graphically and thus can be directly interpretable, given its simple structure.

RF [31] is an ensemble learning method based on aggregating ntree trees similar to the ones constructed with CART, each one grown from a bootstrap sample of the original data set. Each tree in the forest uses only a random subset of mtry predictors at each node. The trees are not pruned. Each value predicted by RF is the average of the values predicted by the ntree trees.

GB [32] is also an ensemble learning method based on DT but does not involve bootstrap sampling. Given a loss function (i.e., squared error for regression) and a weak learner (i.e., regression trees), the GB algorithm seeks to find an additive model that minimizes the loss function. It is initialized with the best guess of the response (i.e., the mean of the response in regression), then the gradient (i.e., residual) is calculated, and a model is then fit to the residuals to minimize the loss function. The current model thus obtained is added to the previous model, adjusted by a shrinkage parameter, and the procedure continues for a user-specified number of iterations, leading to a n.trees total number of trees, a tree depth equal to interaction.depth and a given minimum number of observations in the trees terminal nodes, n.minobsinnode.

NN [33] are nonlinear statistical models for regression or classification. They are structured in layers of "neurons" where the input layer is made of the predictor variables, the output layer contains as many neurons as there are classes (two in our study), and one to many intermediate (size) layers of "weights" called hidden layers. Each neuron is a linear combination of the neurons of the previous layer, to which is applied an activation function, typically the sigmoid function: $g(x) = \frac{1}{1+\exp(-x)}$. The weights are the parameters of the model, and they are estimated through a back-propagation algorithm called gradient descent. The loss function used is the cross-entropy to which a decay penalty is applied.

Statistical analyses

The statistical unit of the data was hospitalization. Descriptive analyses for the sociodemographic, clinical, and hospitalization data were expressed as frequencies and percentages. Chi-squared tests were employed to compare sociodemographic, clinical, and hospitalization data between unplanned 30-day all-cause rehospitalized (y=1) and nonrehospitalized patients (y=0).

To train and evaluate the different models (i.e., LR, CART, RF, NN, and GB), the dataset was split into a 70% training sample and a 30% test sample, stratified on the outcome variable. On the training set, we performed a 5-fold cross validation repeated 5 times to tune the hyperparameters. We kept the optimal hyperparameter values for which the loss was minimum. The tuning process and the values of the optimal hyperparameters are presented in supplementary file 2. On the test set, we assessed the performance of each model using the optimal hyperparameters. We randomly split the test set in two parts: 70% of the sample as a training set and 30% of the sample as a test set. This procedure was repeated 100 times and we computed the average of H-measure and AUC for each model. Since we evaluate different classification rules and the outcome distribution is unbalanced, we used the H-measure, which has the advantage of being classifier-independent and is relevant for heavily unbalanced datasets [34]. The area under the Receiving Operating Characteristic (ROC) curve (AUC) was

also used because it is threshold independent and is a widely used measure. The H-measure and the AUC of each prediction model were compared using a paired t-test.

Finally, we presented variable importance (VI) (i.e., the most important discriminators between classes) for LR and the optimal prediction model (i.e., RF). VI for the LR is given by the reduction in the deviance each variable brings to the null model. For the RF algorithm, VI is calculated by the mean decrease in Gini (MDG) over all the mtry trees for each variable. We applied a corrected feature importance measure to consider categorical variables with a large number of categories which can bias RF models[35]. The changes in Gini are aggregated for each variable and normalized [31]. A high value of the aggregate of the changes indicates great variable importance. All analyses were implemented with R (version 3.5.0) using the caret R (version 6.0.80), hmeasure (version 1.0) and pROC packages (version 1.12.1).

3- Results

Rates of unplanned 30-day all-cause rehospitalization

A total of 289,358 hospitalizations (112,662 patients) were recorded in the year 2015 at this French University Hospital. After excluding mortalities and hospitalizations for ambulatory surgery, radiotherapy, and dialysis, 118,650 hospitalizations (82,862 patients) were included. The most common diseases were digestive disease, nervous system conditions, and cardiovascular and pulmonary diseases. In total, 4,127 (3,294 patients) (3.5%) hospitalizations resulted in rehospitalizations via emergency departments 30 days after discharge. Rehospitalization rates according to sociodemographic, clinical, and hospitalization characteristics are presented in supplementary file 1.

Predictive model performance

The predictive performance of each model is presented in Table 1, and the comparison of each model's H-measure and AUC is presented in Table 2. The RF model was the most performant model with the highest H-measure (0.290) and AUC (0.794), superior to all the other models (all p-values<0.0001). The performance of the RF, GB and NN models (H-measures ranged from 0.184 to 0.290, AUC ranged from 0.741 to 0.794) was superior to that of the LR model (H-measure=0.184, AUC=0.740); all p-values<0.0001. In contrast, LR was superior to CART (H-measure=0.162, AUC=0.707), p<0.0001.

From the optimal cut-point estimated for RF model, the specificity was high (0.99) and the sensitivity was low (0.18).

Variable importance

The variable importance is presented for the RF and LR models in Figures 1 and 2. The seven most important variables are identical (with slightly difference in ranking) and their contributions to reducing the deviance are comparable: "at least one previous hospitalization via emergency departments 6 months before", "category of disease", "hospitalization via emergency departments", "length of stay", "age", "severity" and "type of hospital stay".

The variable importance of the other models is presented in supplementary file 3.

4- Discussion

In this large sample of acute care inpatients (82,862 patients and 118,650 hospitalizations), ML methods (i.e., RF, GB and NN), except for CART, are superior to LR for predicting 30-day all-cause rehospitalizations. To date, the majority of studies have focused on particular conditions, e.g.,

patients with specific diagnoses [36]. This finding confirms the importance of ML models in predicting rehospitalization, despite previous contradictory results on this subject [23]. RF achieves the best performance among all models according the H-measure and the AUC. This result is consistent with recent studies reporting that RF is a relevant and accurate method for predicting health outcomes [37,38,39,40,] although some studies report no improvement in ML models compared to LR [23].

RF is an easy-to-understand method providing an original variable's importance index that helps identify the top-ranked variables associated with 30-day all-cause rehospitalizations [31]. This property of RF should be highlighted regarding the traditional trade-off between accuracy and interpretability in statistical modeling [41]. Contrary to LR, ML models (e.g., RF, GB, NN) are considered to be black boxes because there is not always a clear interpretable connection between outcomes and predictors. However, there has been a tremendous amount of work in developing ways to explain black box models. Variable importance is one of them. In our study, two important findings should be highlighted.

First, the seven most important variables are identical (with slightly difference in ranking) and their contributions to reducing the deviance are comparable between RF and LR. This homogeneity of findings between the two methods is reassuring for the interpretation of results by health care providers. Hospitalization via Emergency Departments and previous hospitalization via emergency departments 6 months before are generally associated with higher readmission in previous works [42]. Older adults are also described as at higher risk of readmission in previous studies [4,5]. Concerning the category of disease, medical-psychiatric comorbidity was highly related to rehospitalizations, confirming previous studies on this complex population [43,44]. This finding justifies the identification of hospitalized patients with psychiatric conditions to better address their behavioral needs. The length of hospital stay was inconstantly associated with higher readmission in previous works [45-47]. French hospitals are under pressure to save on costs, and reducing LOS is strongly advocated. Future studies should thus explore the consequences of this health policy in the French context, particularly its impact on rehospitalization and, more generally, on quality of care.

Second, there are more variables of importance above a threshold of 10% in RF (7 variables: state funded medical assistance, gender, destination on discharge, congestive heart failure, chronic pulmonary disease, dementia, free universal health care and malignancy) than in LR (only 1 variable: dementia). This suggests that RF is better able to identify discriminating variables than LR, including clinical and socio-economic variables. For example, socioeconomic status (i.e., state funded medical assistance and free universal health care in our study) was associated with rehospitalization in our study, confirming recent findings on social risk (poverty, disability, housing instability, residence in a disadvantaged neighborhood) and rehospitalization [48]. Interestingly, previous studies also reported gender inequalities [49] and risk associated with congestive heart failure [4], chronic pulmonary disease [50], dementia [51].

Despite our findings in favor of RF and ML methods, two issues must be considered in future work: a moderate improvement, especially for the AUC, between ML and LR, and the use of data at discharge.

Our study included a relatively small set of variables (29 variables), relevant for classical statistical methods based on standard parametric models but suboptimal for ML methods in some respects. Several additional pieces of information could be relevant to predict rehospitalization, including structured (e.g., socioeconomic status, drugs, and self-reported functional status) and unstructured (e.g., clinical notes from physicians, nurses, and other professionals) data available in electronic medical records. These data could improve prediction by offering richer medical information than those found in the only medico-administrative databases. Previous studies reported that the

performance of ML methods could be improved by taking into account a larger number of variables [52]. Future studies should include all data available in electronic medical records.

As for the majority of predictive risk scores, our study was based on data at discharge, while predictive risk scores should ideally give information early enough during hospitalization to trigger care intervention [53]. To date, instruments based on discharge data have been proven to lead to models with better performance [53,54] than models based solely on admission data. An important perspective would be to implement real-time predictive rehospitalization risk scores during hospitalization, updated for all new available data, and then propose early alerts for high risk of rehospitalization. A recent study reported that ML methods can be used in real-time predictions using routinely collected clinical data exclusively, without the need for any manual processing [55]. Another recent study trained and tested a neural network model to predict the risk of patients' rehospitalization within 30 days of their discharge based on real-time data from EHR, and thus applicable at the time discharge from hospital [56].

Our findings must be interpreted in the context of our study's limitations. Despite the large overall sample size of this multihospital study, our findings may not be applicable to all French hospitals, particularly general hospitals where patients have potentially different characteristics from those of university hospitals. In addition, the four university hospitals included in our study were located in only one geographical area, and social and healthcare geographical characteristics (e.g., poverty, density of physicians, number of beds, and private hospitals) are known to influence the risk of rehospitalization [53,57]. Future studies should thus be conducted in different categories of hospitals and in several geographical areas to confirm the properties and importance of our predictive risk score. Our model does not factor in deaths outside the hospital because we do not account for this information in our database. Other studies with available data on outpatient events are needed to investigate to what extent this could impact our predictive risk score using a competing risk model as an example. We excluded ambulatory surgery from the analyses. This specific topic should be studied in the French context, strongly marked by pressures for reducing length of stay. Lastly, the caret R package offers the possibility of using other statistical models that could be studied in future work (e.g., Multi-Layer Perceptron Neural Network, Support Vector Machine, Bayesian Network).

Conclusion

The use of ML may be an alternative to regression models to predict health outcomes. The integration of ML, particularly the RF algorithm, in the prediction of unplanned rehospitalization, may help health service providers target patients at high risk of rehospitalizations and propose effective interventions at the hospital level.

Conflict of interest: None

Funding: None

Acknowledgments: None

References

1. Boutwell AE, Johnson MB, Rutherford P, et al. An early look at a four-state initiative to reduce avoidable hospital readmissions. Health Aff (Millwood). 2011;30(7):1272-1280.

2. HAS. Haute Autorité de Santé. Note méthodologique et de synthèse documentaire «Sortie d'hospitalisation supérieure à 24 heures–Établissement d'une check-list». In Available from: http://wwwhas-santefr/portail/upload/docs/application/pdf/2015-05/note_documentaire_check-list_sortie_hospitalisation_webpdf. 2015.

3. Gusmano M, Rodwin V, Weisz D, Cottenet J, Quantin C. Comparison of rehospitalization rates in France and the United States. J Health Serv Res Policy. 2015;20(1):18-25.

4. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med. 2009;360(14):1418-1428.

5. Friedman B, Basu J. The rate and cost of hospital readmissions for preventable conditions. Med Care Res Rev. 2004;61(2):225-240.

6. Ashton CM, Kuykendall DH, Johnson ML, Wray NP, Wu L. The association between the quality of inpatient care and early readmission. Ann Intern Med. 1995;122(6):415-421.

7. Balla U, Malnick S, Schattner A. Early readmissions to the department of medicine as a screening tool for monitoring quality of care problems. Medicine (Baltimore). 2008;87(5):294-300.

8. Francois P, Bertrand D, Beden C, Fauconnier J, Olive F. [Early readmission as an indicator of hospital quality of care]. Rev Epidemiol Sante Publique. 2001;49(2):183-192.

9. Coleman EA, Parry C, Chalmers S, Min SJ. The care transitions intervention: results of a randomized controlled trial. Arch Intern Med. 2006;166(17):1822-1828.

10. Leppin AL, Gionfriddo MR, Kessler M, et al. Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. JAMA Intern Med. 2014;174(7):1095-1107.

11. Pauly V, Mendizabal H, Gentile S, Auquier P, Boyer L. Predictive risk score for unplanned 30-day rehospitalizations in the French universal health care system based on a medico-administrative database. Plos one. 2018, In press.

12. van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ. 2010;182(6):551-557.

13. Donze JD, Williams MV, Robinson EJ, et al. International Validity of the HOSPITAL Score to Predict 30-Day Potentially Avoidable Hospital Readmissions. JAMA Intern Med. 2016;176(4):496-502.

14. Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. PLoS One. 2017;12(4):e0175383.

15. Ahn JM, Kim S, Ahn KS, Cho SH, Lee KB, Kim US. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. PLoS One. 2018;13(11):e0207982.

16. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3(3):243-250.

17. Gholipour C, Rahim F, Fakhree A, Ziapour B. Using an Artificial Neural Networks (ANNs) Model for Prediction of Intensive Care Unit (ICU) Outcome and Length of Stay at Hospital in Traumatic Patients. J Clin Diagn Res. 2015;9(4):OC19-23.

18. Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. PLoS One. 2017;12(5):e0177726.

19. Kuo PJ, Wu SC, Chien PC, et al. Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: a cross-sectional retrospective study in southern Taiwan. BMJ Open. 2018;8(1):e018252.

20. LaFaro RJ, Pothula S, Kubal KP, et al. Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre-Incision Variables. PLoS One. 2015;10(12):e0145395.

21. Stylianou N, Akbarov A, Kontopantelis E, Buchan I, Dunn KW. Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. Burns. 2015;41(5):925-934.

22. Kuhn M, Johnson K. Applied predictive modeling. Springer. 2013;26.

23. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. JAMA Cardiol. 2017;2(2):204-209.

24. Morgan DJ, Bame B, Zimand P, et al. Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. JAMA Netw Open. 2019;2(3):e190348.

25. Fernandez-Delgado M, Cernadas E, Barro S. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? . Journal of Machine Learning Research. 2014;15 3133-3181

26. Boudemaghe T, Belhadj I. Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI). Int J Epidemiol. 2017;46(2):392-392d.

27. Bottle A, Aylin P, Majeed A. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. J R Soc Med. 2006;99(8):406-414.

28. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005;43(11):1130-1139.

29. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning Springer Series in Statistics. New York: Springer. 2017.

30. Breiman L. Classification and Regression Trees. 1st ed Wadsworth; International Group. 1984.

31. Breiman L. Random forests. Mach Learn. 2001(45):5-32.

32. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001:1189–1232.

33. Arbib MA. The handbook of brain theory and neural networks. MIT press. 2003.

34. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009(21):1263-1284.

35. Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340-1347.

36. Garcia-Arce A, Rico F, Zayas-Castro JL. Comparison of Machine Learning Algorithms for the Prediction of Preventable Hospital Readmissions. J Healthc Qual. 2018;40(3):129-138.

37. Hsieh MH, Hsieh MJ, Chen CM, Hsieh CC, Chao CM, Lai CC. Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. Sci Rep. 2018;8(1):17116.

38. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med. 2016;23(3):269-278.

39. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res. 2017;121(9):1092-1101.

40. Artetxe A, Beristain A, Grana M. Predictive models for hospital readmission risk: A systematic review of methods. Comput Methods Programs Biomed. 2018;164:49-64.

41. Nanayakkara S, Fogarty S, Tremeer M, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. PLoS Med. 2018;15(11):e1002709.

42. Brennan JJ, Chan TC, Killeen JP, Castillo EM. Inpatient Readmissions and Emergency Department Visits within 30 Days of a Hospital Admission. West J Emerg Med. 2015;16(7):1025-1029.

43. Jansen L, van Schijndel M, van Waarde J, van Busschbach J. Health-economic outcomes in hospital patients with medical-psychiatric comorbidity: A systematic review and meta-analysis. PLoS One. 2018;13(3):e0194029.

44. Boaz TL, Becker MA, Andel R, McCutchan N. Rehospitalization risk factors for psychiatric treatment among elderly Medicaid beneficiaries following hospitalization for a physical health condition. Aging Ment Health. 2017;21(3):297-303.

45. Bueno H, Ross JS, Wang Y, et al. Trends in length of stay and short-term outcomes among Medicare patients hospitalized for heart failure, 1993-2006. JAMA. 2010;303(21):2141-2147.

46. Kaboli PJ, Go JT, Hockenberry J, et al. Associations between reduced hospital length of stay and 30-day readmission rate and mortality: 14-year experience in 129 Veterans Affairs hospitals. Ann Intern Med. 2012;157(12):837-845.

47. Sud M, Yu B, Wijeysundera HC, et al. Associations Between Short or Long Length of Stay and 30-Day Readmission and Mortality in Hospitalized Patients With Heart Failure. JACC Heart Fail. 2017;5(8):578-588.

48. Joynt Maddox KE, Reidhead M, Hu J, et al. Adjusting for social risk factors impacts performance and penalties in the hospital readmissions reduction program. Health Serv Res. 2019;54(2):327-336.

49. Gonzalez JR, Fernandez E, Moreno V, et al. Sex differences in hospital readmission among colorectal cancer patients. J Epidemiol Community Health. 2005;59(6):506-511.

50. Buhr RG, Jackson NJ, Dubinett SM, Kominski GF, Mangione CM, Ong MK. Factors Associated with Differential Readmission Diagnoses Following Acute Exacerbations of Chronic Obstructive Pulmonary Disease. J Hosp Med. 2020;15(2):e1-e9.

51. Pickens S, Naik AD, Catic A, Kunik ME. Dementia and Hospital Readmission Rates: A Systematic Review. Dement Geriatr Cogn Dis Extra. 2017;7(3):346-353.

52. Couronne R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018;19(1):270.

53. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306(15):1688-1698.

54. Nguyen OK, Makam AN, Clark C, et al. Predicting all-cause readmissions using electronic health record data from the entire hospitalization: Model development and comparison. J Hosp Med. 2016;11(7):473-480.

55. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med. 2018.

56. Jamei M, Nisnevich A, Wetchler E, Sudat S, Liu E. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PLoS One. 2017;12(7):e0181173.

57. Hernandez AF, Greiner MA, Fonarow GC, et al. Relationship between early physician follow-up and 30-day readmission among Medicare beneficiaries hospitalized for heart failure. JAMA. 2010;303(17):1716-1722.

Figure 1. Importance of variables in LR model.

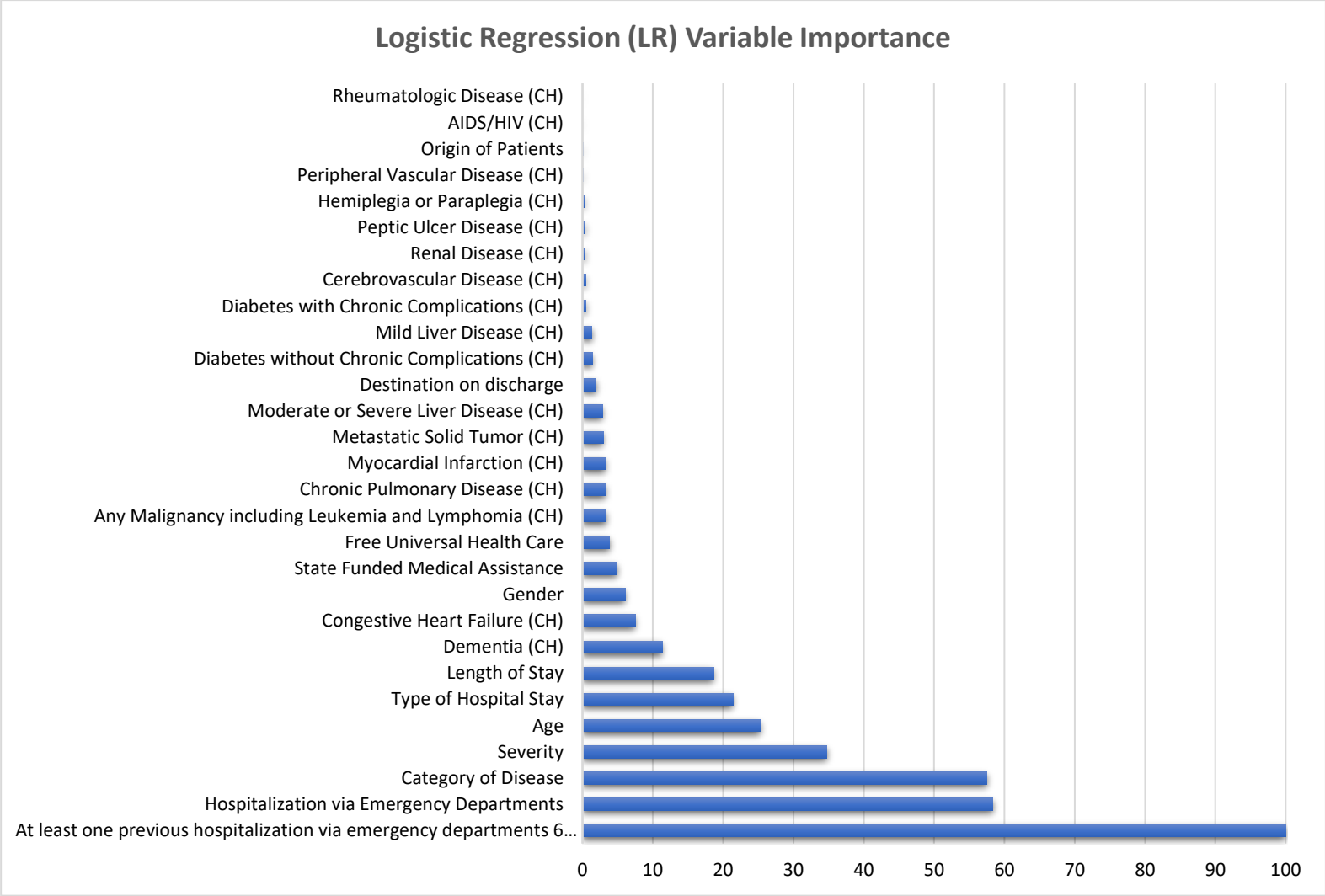


Figure 2. Importance of variables in RF model.

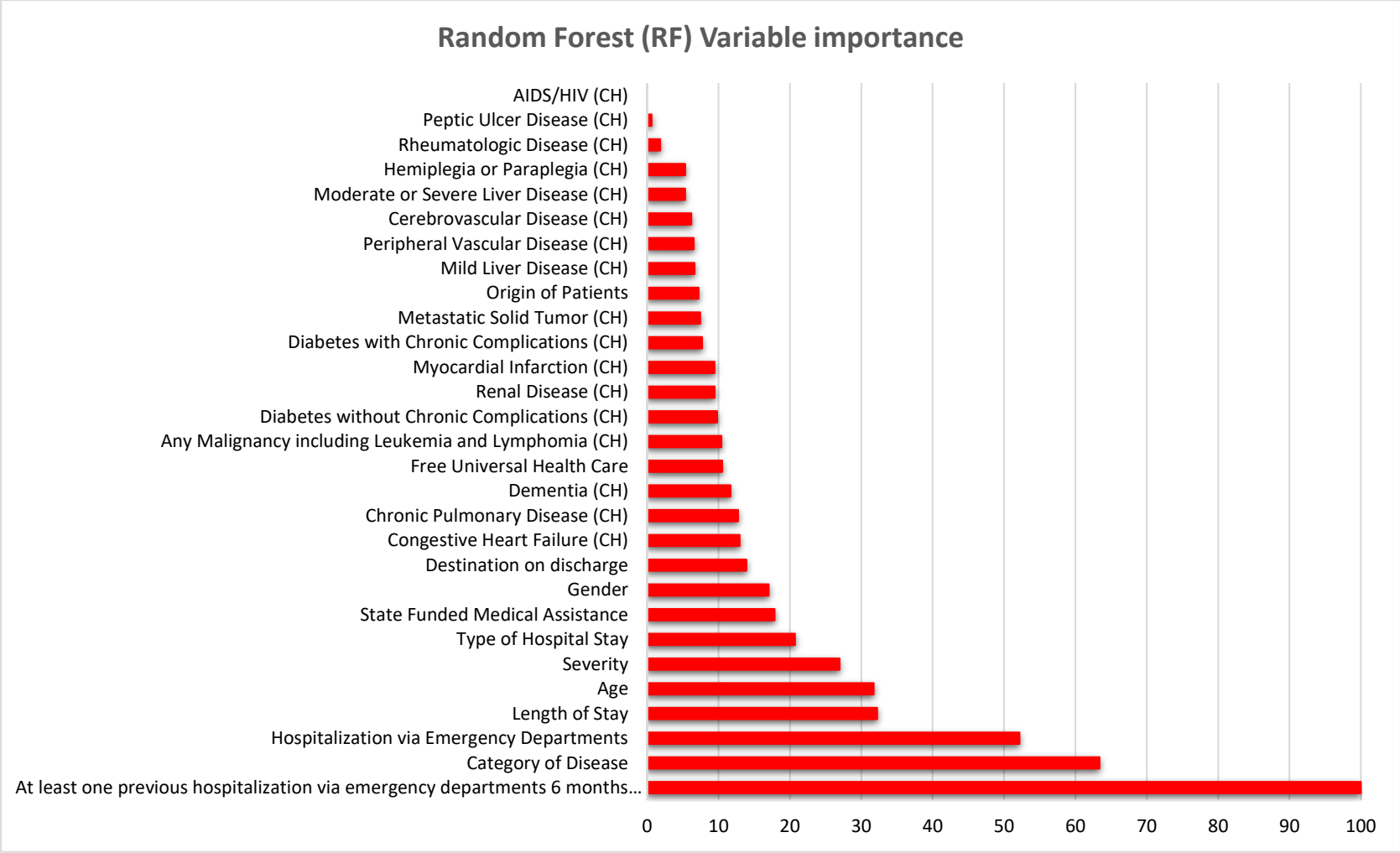


Table 1. Predictive performance for LR, CART, RF, GB and NN models

	H (95%CI)	AUC (95%CI)
LR	0.1833 (0.1815;0.1851)	0.7399 (0.7387;0.7410)
CART	0.1625 (0.1609;0.1641)	0.7067 (0.7055;0.7078)
RF	0.2900 (0.2879;0.2921)	0.7941 (0.7929;0.7954)
GB	0.2179 (0.2160;0.2198)	0.7620 (0.7608;0.7631)
NN	0.1843 (0.1825;0.1861)	0.7417 (0.7405;0.7429)

LR: logistic regression

CART: classification and regression trees

RF: random forest

GB: gradient boosting

NN: neural networks

H: H-measure

AUC: area under the ROC curve

95%CI: 95% confidence interval

Table 2. Comparison of AUC and H-measures of ML models to LR and RF (paired t-tests).

Ref. Model: LR	Index	Statistic	p-value
H t-tests	H-GB	54.44	< 0.0001
	H-NN	75.28	< 0.0001
	H-CART	96.27	< 0.0001
	H-RF	76.09	< 0.0001
AUC t-tests	AUC-GBM	38.85	< 0.0001
	AUC-NN	61.39	< 0.0001
	AUC-CART	106.39	< 0.0001
	AUC-RF	62.35	< 0.0001

Ref. Model: RF	Index	Statistic	p-value
H t-tests	H-GB	-27.10	< 0.0001
	H-NN	-0.75	0.4558
	H-CART	17.31	< 0.0001
	H-LR	-76.09	< 0.0001
AUC t-tests	AUC-GB	-27.18	< 0.0001
	AUC-NN	-2.16	0.0332
	AUC-CART	41.89	< 0.0001
	AUC-LR	-62.35	< 0.0001

LR: logistic regression

CART: classification and regression trees

RF: random forest

GB: gradient boosting

NN: neural networks

Supplementary file 1. Sample characteristics and rates of unplanned 30-day all-cause rehospitalizations

	30-day rehospitalization rates		p-value	All (n=118 650)	
Sociodemographic characteristics					
Age	N (4 127)	%	<0.0001	N	% of all
>= 75 years old	1 201	5.6		21 651	18.3
>=18 and <75 years old	2 156	2.9		75 063	63.3
>= 1 and <=17 years old	363	2.9		12 596	10.6
<1 year	407	4.4		9 340	7.9
Gender			<0.0001		
Male	2 377	4.0		59 874	50.5
Female	1 750	3.0		58 776	49.5
State-funded medical assistance			<0.0001		
Yes	142	7.4		1 921	1.6
No	3 985	3.4		116 729	98.4
Free universal health care			<0.0001		
Yes	600	4.7		12 894	10.9
No	3 527	3.3		105 756	89.1
Clinical characteristics					
Category of disease			<0.0001		
Digestive	600	4.9		12 159	10.3
Orthopedic – Trauma	218	2.7		8 000	6.7
Multiple or complex trauma	11	3.8		288	0.2
Rheumatology	81	2.2		3 656	3.1
Nervous system	428	3.6		12 030	10.1
Vascular catheterization	160	2.4		6 598	5.6
Cardiovascular	340	3.7		9 167	7.7
Pulmonary	604	6.6		9 108	7.7
Ear Nose and Throat - Stomatology	80	1.6		4 903	4.1
Ophthalmology	13	0.8		1 579	1.3
Gynecology-Breast	45	2.0		2 231	1.9
Obstetrical	98	1.3		7 360	6.2
Newborns and perinatal diseases	224	3.8		5 827	4.9
Uronephrology and reproductive organs	274	4.1		6 754	5.7
Hematology	148	4.9		3 014	2.5
Chemotherapy - radiotherapy	155	2.0		7 852	6.6
Infectious diseases	76	5.3		1 444	1.2
Endocrinology	151	3.1		4 856	4.1
Cutaneous and subcutaneous	67	2.1		3 146	2.7
Psychiatry	83	6.6		1 264	1.1
Toxicology - Intoxication - Alcohol	96	6.0		1 611	1.4

	30-day rehospitalization rates		p-value	All (n=118 650)	
Chronic pain palliative care	20	5.7		351	0.3
Organ Transplant	3	1.3		240	0.2
Interdisciplinary activities and follow-up of patients	149	3.1		4 825	4.1
Burns	3	0.8		387	0.3
Severity			<0.0001		
No or low severe	2 121	2.7		78 014	65.8
Moderate - high severe	1 180	6.2		19 191	16.2
Not determined (short stay)	826	3.9		21 445	18.1
Charlson Comorbidity Index			<0.0001		
0	2 513	3.0		84 364	71.1
1	582	4.6		12 786	10.8
2	416	4.0		10 356	8.7
3	191	5.0		3 802	3.2
4 and higher	425	5.8		7 342	6.2
Type of hospital stay			<0.0001		
Surgical	922	2.6		35 480	29.9
Medical	3 107	4.1		75 810	63.9
Obstetrical	98	1.3		7 360	6.2
Hospital stay characteristics					
Origin of patient			0.25		
From home	3 878	3.5		111 973	94.4
Other (other hospital-institution)	249	3.7		6 677	5.6
Hospitalization via emergency departments			<0.0001		
No	1 06	2.2		71 846	60.5
Yes	2 521	5.4		46 804	39.5
Length of hospital stay			<0.0001		
One day	265	4.4		5 979	5.0
From 2 to 3 days	1 201	2.9		42 083	35.5
From 4 to 8 days	1 417	3.1		46 251	39.0
9 and higher	1 244	5.1		24 337	20.5
Destination on discharge			<0.0001		
Return to home	3 425	3.4		101 859	85.9
Other (other hospital-institution)	702	4.2		16 791	14.1
Hospitalization via emergency departments 6 months before			<0.0001		
No hospitalization	1 935	2.6		74 452	62.8
At least one previous hospitalization but not via emergency departments	719	2.5		28 591	24.1
At least one previous hospitalization via emergency departments 6 months before	1 473	9.4		15 607	13.2

Supplementary file 2. Tuning process and values of hyperparameters.

We randomly split the sample in two parts: 70% of the sample as a training set and 30% of the sample as a test set.

On the training set, we performed a 5-fold cross validation repeated 5 times to tune the hyperparameters for each classifier. For each value of hyperparameter within a fixed range defined in the caret Package, we performed the following procedure:
We randomly split the training set into 5 approximately equal subsets or folds (k=5 folds)
We took 1 subset (first fold) as test set and the remaining subsets (k-1=4 folds) as training set. We fitted the model on the training set and evaluated it on the test set using the loss function.
This process was repeated 5 times for testing each combination of the subsets and we computed the average loss over the 5 estimations.
We repeated this whole cycle (1,2,3) 5 times, we computed the average loss and saved the loss score for the hyperparameter value.
Then we selected the optimal hyperparameter value for which the average loss was the minimum.

On the test set, we assessed the performance of each model using the optimal hyperparameter obtained previously. We randomly split the test set in two parts: 70% of the sample as a training set and 30% of the sample as a test set. This procedure was repeated 100 times and we computed the average of H-measure and AUC for each model.

R package	Model	Hyperparameter	Description	Optimal Value
rpart	CART	Cp	cost complexity pruning - penalty applied to the loss function	0.00011538
gbm	GB	n.trees	number of trees	200
		interaction.depth	tree depth	5
		Shrinkage	shrinkage parameter	0.1
		n.minobsinnode	minimum number of observations in the terminal nodes	10
Random Forest	RF	Ntrees	number of trees	500
		mtry	number of randomly selected predictors	7
nnet	NN	size	number of units in hidden layers	1
		decay	penalty applied to the loss function	0.1

Supplementary file 3. Variable Importance for GB, CART and NN.

CART	VI	GB	VI	NN	VI
Category of Disease	100,00	Category of Disease	100,00	Hospitalization via Emergency Departments	100,00
At least one previous hospitalization via emergency departments 6 months before	67,17	At least one previous hospitalization via emergency departments 6 months before	73,05	At least one previous hospitalization via emergency departments 6 months before	98,65
Hospitalization via Emergency Departments	63,31	Hospitalization via Emergency Departments	37,16	Severity	56,18
Age	61,27	Age	15,32	Length of Stay	41,11
Severity	46,67	State Funded Medical Assistance	13,70	Age	39,90
State Funded Medical Assistance	41,53	Severity	13,57	Gender	32,92
Gender	38,79	Length of Stay	11,97	Category of Disease	19,27
Type of Hospital Stay	26,07	Destination on discharge	7,64	Type of Hospital Stay	18,84
Length of Stay	24,89	Gender	5,32	Free Universal Health Care	18,06
Chronic Pulmonary Disease	14,84	Free Universal Health Care	4,54	Congestive Heart Failure	15,68
Destination on discharge	10,19	Dementia	3,76	Dementia	13,64
Dementia	5,39	Congestive Heart Failure	3,72	Destination on discharge	13,20
Peripheral Vascular Disease	5,13	Diabetes without Chronic Complications	3,26	Myocardial Infarction	10,85
Free Universal Health Care	3,31	Chronic Pulmonary Disease	3,11	Any Malignancy including Leukemia and Lymphomia	10,46
Renal Disease	2,91	Renal Disease	2,86	State Funded Medical Assistance	8,46
Mild Liver Disease	2,28	Metastatic Solid Tumor	2,76	Metastatic Solid Tumor	7,67
Diabetes without Chronic Complications	2,14	Myocardial Infarction	2,62	Diabetes without Chronic Complications	7,51
Any Malignancy including Leukemia and Lymphomia	1,80	Type of Hospital Stay	2,37	Chronic Pulmonary Disease	7,47
Hemiplegia or Paraplegia	1,66	Moderate or Severe Liver Disease	2,24	Mild Liver Disease	4,47
Metastatic Solid Tumor	1,22	Hemiplegia or Paraplegia	2,15	Diabetes with Chronic Complications	3,85
Congestive Heart Failure	1,03	Any Malignancy including Leukemia and Lymphomia	1,79	Moderate or Severe Liver Disease	3,70

Diabetes with Chronic Complications	0,76	Origin of Patients	1,69	Renal Disease	3,60
AIDS/HIV	0,42	Peripheral Vascular Disease	1,42	Cerebrovascular Disease	2,90
Origin of Patients	0,28	Cerebrovascular Disease	1,27	Origin of Patients	2,54
Cerebrovascular Disease	0,21	Mild Liver Disease	1,24	Peripheral Vascular Disease	1,50
Myocardial Infarction	0,00	Diabetes with Chronic Complications	0,69	Hemiplegia or Paraplegia	1,04
Moderate or Severe Liver Disease	0,00	Rheumatologic Disease	0,31	Rheumatologic Disease	0,01
Peptic Ulcer Disease	0,00	Peptic Ulcer Disease	0,10	Peptic Ulcer Disease	0,01
Rheumatologic Disease	0,00	AIDS/HIV	0,00	AIDS/HIV	0,00

GB: gradient boosting

CART: classification and regression trees

NN: neural networks

VI: variable importance

Synthèse des Conclusions – Article 1

A la question : est-il pertinent d'utiliser des modèles de Machine Learning en lieu et place de la Régression Logistique classique utilisée dans les problèmes de classification en Sciences de la Santé, nos résultats suggèrent que la différence de performance est suffisamment significative pour conclure que c'est le cas.

Ces résultats sont d'autant plus honorables que nos données n'incluent pas des variables considérées comme particulièrement importantes dans la littérature, comme les tests de laboratoires, les signes vitaux ou les médicaments (Zhou et al., 2016) ou plus récemment celles qui sont contenues dans les données de santé électroniques comme la complexité des procédés chirurgicaux ou les notes cliniques des infirmiers et des médecins (Mahmoudi et al., 2020).

De plus, la performance de nos modèles se trouve plutôt dans le médian supérieur des valeurs du ROC AUC mentionnées dans les différentes revues de littérature systématique (Y. Huang et al., 2021; Mahmoudi et al., 2020; Zhou et al., 2016)

Constatant que les variables et les observations utilisées proviennent des données administratives du PMSI, ces résultats confirment que ces dernières peuvent utilement contribuer à la prévention ou à la réduction des réhospitalisations via les urgences à 30 jours et confirment ainsi de récents travaux allant dans ce sens (Nicolet et al., 2022). Les variables les plus importantes de nos modèles correspondent d'ailleurs à celles qui sont identifiées comme telles dans les revues de littérature systématiques les plus récentes (Y. Huang et al., 2021; Zhou et al., 2016).

Admettant que la réhospitalisation à 30 jours (RH30) est un indicateur de qualité et de coordination des soins entre autres choses, en France comme dans les autres nations de l'OCDE, il est tout à fait naturel de mettre en œuvre des interventions visant à accompagner cette coordination, telles que des interventions pré-décharges, des interventions post décharges et des interventions de transition (O. Hansen et al., 2011). Clairement les variables les plus importantes de nos modèles permettent de cibler plus finement ces séjours / patients à risques, c'est-à-dire ceux qui ont eu de multiples réhospitalisation lors des derniers mois, et encore davantage s'ils sont passés par les urgences, certaines catégories de maladies (qu'il reste à identifier), la sévérité de la condition d'admission, la durée de séjour que nous nous doutons plutôt prolongée et l'âge que nous devinons plus avancé.

Nous voyons déjà poindre ici les premières limites de cette recherche. Nous avons choisi l'option de traiter les variables catégorielles sans passer par une disjonction complète (one hot encoding) des modalités. Ceci a été possible en utilisant la déviance comme mesure de l'importance pour la Régression Logistique (LR), les qualités intrinsèques des arbres pour les arbres de décision (CART), Gradient Boosting (GB) et Random Forest (RF), puis le score individuel ROC AUC des variables pour les Réseaux de Neurones (NN). Si, cette option peut avoir contribué à l'amélioration de la performance des modèles basés sur les arbres, elle apporte avec elle son lot d'opacité dans le mesure où ne savons pas en fin de compte quelles modalités contribuent le plus à la capacité discriminative de notre modèle, notamment pour des variables comme les catégories de maladie (24 modalités), ou le type d'hospitalisation.

Acceptant de prendre le risque de perdre un peu en performance pour gagner en interprétabilité, nos prochains travaux examineront la contribution des modalités une à une. Ce compromis est au cœur de l'enjeu de l'explicabilité ou de l'interprétabilité en Machine Learning (ElShawi et al., 2021; Stiglic et al., 2020).

Il convient aussi de souligner que les méthodes de calcul de l'importance des variables dans cet article entrent dans la catégorie des modèles intrinsèquement interprétables. Or le calcul de l'importance étant inhérent au modèle, on ne peut pas vraiment comparer les valeurs entre elles, aussi convient-t-il d'utiliser une méthode du calcul de l'importance des variables qui soit la même pour tous les classifieurs, c'est-à-dire une méthode « agnostique » (ou indépendant du modèle). Nos prochains travaux utiliseront une telle méthode – l'importance des variables par permutation (permutation feature importance ou PFI).

D'autre part, l'indisponibilité d'algorithmes gérant les GPU pour les réseaux de neurones dans le package CARET (et sur R de façon générale) lors des expériences sur les différents modèles constitue une autre limite, nous contraignant à des Réseaux de Neurones à une seule couche (cachée). Cela explique sans doute l'absence de différence significative entre les performances de LR et de NN.

L'ensemble de ces éléments nous incitent à poursuivre la suite des travaux sur Python, avec les modules Scikit Learn, Keras, puis Hugging Face ou Pytorch ultérieurement.

Pour conclure, il nous faut mentionner une particularité de ces travaux qui n'est pas courante en Sciences de la Santé : il s'agit du recours au rééchantillonnage. En effet si le recours à des échantillons distincts pour l'entraînement, la validation et/ou le test est maintenant une évidence auprès des chercheurs et des praticiens (Bacchi, Tan, et al., 2020) il n'en reste pas moins que le partitionnement des données en sous-échantillons, même quand ils sont stratifiés sur la variable catégorielle à expliquer, peut engendrer des modèles excessivement optimistes ou tout à fait le contraire. Afin d'éviter ce potentiel biais, il est recommandé de procéder à des rééchantillonnages de l'échantillon d'apprentissage et de validation pour lisser les aléas, mais aussi pour être en mesure de produire des statistiques de test. On sait ainsi que la différence de performance entre les classifieurs n'est pas due aux aléas des partitionnements des échantillons.

Article 2[18]

Machine-learning prediction for hospital length of stay using a French medico-administrative database

Contexte

Nous nous intéressons à présent à un autre indicateur de qualité des soins : la durée de séjour. Comme nous l'avons mentionné auparavant, la durée de séjour (LOS) indique plus précisément l'efficacité des pratiques des soins, mais aussi l'efficience dans l'organisation des soins et dans l'utilisation des ressources hospitalières.

Ici nous mesurons le LOS comme la différence entre la date d'admission et la date de décharge correspondant à un séjour donné. Etant donné que notre variable à expliquer (LOS) est quantitative, nous pouvons traiter le problème d'apprentissage supervisé comme une régression. Cependant, à la fois pour des raisons statistiques : la distribution du LOS est très fortement asymétrique à droite, mais aussi pour des raisons théoriques : ce qui nous intéresse c'est de prévenir les séjours prolongés, nous avons choisi de dichotomiser le LOS en séjour ordinaire (ou court) vs séjour prolongé (ou long).

Les données utilisées sont celles de l'APHM, issues du PMSI, déjà mobilisées dans le précédent article.

Etant donné qu'il n'existe pas de seuil consensuel pour départager les séjours courts des séjours longs, même si la durée de 14 jours semble être celle qui est la plus couramment utilisée, nous avons opté pour un critère statistique simple. Nous considérons comme séjour prolongé (non ordinaire) toutes les durées qui excèdent la clôture supérieure de Tukey (1977a) :

$$UF = Q_3 + 1.5 \times IQR$$

$$o\grave{u}\ Q_3\ repr\acute{e}sente\ la\ troisi\grave{e}me\ quartile\ et\ IQR\ l'intervalle\ interquartile$$

Par un heureux hasard, pour les données de l'APHM, ce seuil représente bien 14 jours et la 90ème centile.

Autrement dit, la proportion de positifs (séjours prolongés) et de négatifs (séjours ordinaires) est respectivement de 10% vs 90%, ce qui correspond à une distribution également très déséquilibrée.

Critères d'Inclusion

En choisissant le LOS comme variable d'intérêt, certaines variables écrasent les autres et ne permettent pas de saisir toute la nuance des facteurs de risques. Par exemple la variable sévérité écrase toutes les autres pour prédire le séjour prolongé et neutralise toute autre information. On l'exclut donc du modèle.

De plus, les séjours obstétriques et les décès ont été exclus de l'analyse, ainsi que les patients de moins de 18 ans, et les séjours de moins de 24 heures.

[18] Actuellement en révision mineure pour la revue « Journal of Market Access & Health Policy » ID 225605906

1- Algorithmes et modèles :

Les analyses mobiliseront principalement les packages Scikit Learn et Keras pour 5 modèles de Machine Learning :

- Régression Logistique (LR)
- Arbres de décision (CART)
- Random Forest (RF)
- Gradient Boosting (GB)
- Support Vector Machine (SVM)
- Artificial Neural Network (NN)

Pour des raisons de convergence (très longue), associées à des performances peu convaincantes, SVM a été éliminé pour l'article.

2- Entraînements des modèles :

Les données ont été partitionnées en échantillons test, apprentissage et validation.

L'échantillon test est mis de côté, en acceptant l'éventualité qu'il puisse être (de façon aléatoire) particulièrement favorable ou défavorable aux modèles appris – comme c'est souvent le cas dans la réalité.

L'échantillon d'apprentissage est utilisé pour entraîner le modèle avec un processus de validations croisées pour optimiser les hyperparamètres. Le meilleur modèle obtenu par validations croisées est ensuite évalué sur l'échantillon de validation.

Pour éviter les aléas des tirages ce processus est répété dix fois, puis on sauvegarde le meilleur modèle sur dix rééchantillonnages de partitionnements apprentissage + validation.

Avec ce meilleur modèle on partitionne encore 100 fois l'échantillon apprentissage + validation, de façon à obtenir 100 valeurs de la performance, permettant de faire un test de comparaisons de moyennes avec effet de taille. La performance est mesurée uniquement avec le ROC AUC.

Résumé

Introduction. La prédiction de la durée du séjour hospitalier (LOS) a traditionnellement utilisé des modèles de régression logistique ou linéaire. Des méthodes Machine Learning (ML) ont été introduites dans la recherche en Sciences de la santé et peuvent améliorer la prédiction de la durée de séjour. L'objectif de ce travail est de développer un modèle ML pour prédire le LOS à partir d'une base de données médico-administrative hospitalière française.

Méthodes. Il s'agit d'une étude de cohorte rétrospective de l'ensemble des sorties de l'année 2015 d'hospitalisations en soins aigus dans un centre universitaire de soins tertiaires regroupant quatre hôpitaux français. La variable d'intérêt est la durée de séjour transformée en une variable binaire (longue ou courte durée de séjour) selon le 90e centile (14 jours). Ont été appliqués aux données collectées : la régression logistique (LR), les arbres de classification et de régression (CART), la forêt aléatoire (RF), le gradient boosting (GB) et les réseaux de neurones (NN). La performance prédictive des modèles a été évaluée à l'aide de l'aire sous la courbe ROC (AUC).

Résultats. Notre analyse a inclus 73 182 hospitalisations, dont 7 341 (10,0 %) ont subi une durée de séjour prolongée. Le classifieur GB est le modèle le plus performant avec l'AUC la plus élevée (0,810), comparée à tous les autres modèles (toutes les valeurs de p <0,0001). La performance des modèles RF, GB et NN (AUC variant de 0,808 à 0,810) est supérieure à celle du modèle LR (AUC=0,795) ; toutes les valeurs de p<0,0001. En revanche, LR est supérieure à CART (AUC=0,786), p<0,0001. La variable la plus prédictive du LOS prolongé est la destination du patient après hospitalisation vers d'autres établissements. Le profil clinique type de ces patients (17,5 % de l'échantillon) était le patient âgé, admis en urgence, pour un traumatisme, une pathologie neurologique ou cardiovasculaire, le plus souvent institutionnalisé, avec plus de comorbidités notamment des problèmes de santé mentale, de démence et d'hémiplégie.

Discussion. L'intégration du ML, en particulier de l'algorithme GB, peut être utile aux professionnels de la santé et aux gestionnaires de lits pour mieux identifier les patients à risque de séjour prolongé. Ces résultats soulignent la nécessité de renforcer les hôpitaux grâce à une allocation ciblée pour répondre aux besoins d'une population vieillissante.

Mots-clés : Apprentissage automatique ; Réseau de neurones; Prédiction; Recherche sur les services de santé ; Santé publique.

Machine-learning prediction for hospital length of stay using a French medico-administrative database

Short title: Machine Learning and Length of Stay

Jaotombo Franck (PhD)[1,2], Pauly Vanessa (PhD)[1,3], Fond Guillaume (MD, PhD)[1], Orleans Veronica (PhD)[3], Pascal Auquier (MD, PhD)[1], Ghattas Badih (PhD)[2], Boyer Laurent (MD, PhD)[1,3]

1. Aix-Marseille University, EA 3279 - Public Health, Chronic Diseases and Quality of Life - Research Unit, La Timone Medical University, 27, boulevard Jean-Moulin, F-13005 Marseille.
2. Aix-Marseille University, I2M, CNRS, UMR 7373, Marseille, France.
3. Service d'Information Médicale, Public Health Department, La Conception Hospital, Assistance Publique - Hôpitaux de Marseille, 147 Boulevard Baille, F-13005 Marseille.

Category: Observational Study

Word count: 5451 words

3 tables, 2 figures

* Correspondence should be sent to:

Prof. Laurent Boyer

Aix-Marseille Univ, Faculté de Médecine - Secteur Timone, EA 3279: CEReSS -Centre d'Etude et de Recherche sur les Services de Santé et la Qualité de vie, 27 Boulevard Jean Moulin, 13005 Marseille, France.

Tel: (33 686 93 62 76), e-mail: laurent.boyer@ap-hm.fr

Abstract

Introduction. Predicting hospital length of stay (LOS) has traditionally employed logistic or linear regression models. Machine learning (ML) methods have been introduced in health service research and may improve the prediction of LOS. The objective of this work was to develop a ML model to predict hospital LOS based on a French hospital medico-administrative database.

Methods. This was a retrospective cohort study of all discharges in the year 2015 from acute-care inpatient hospitalizations in a tertiary-care university center comprising four French hospitals. The study outcomes were LOS transformed into a binary variable (long vs. short LOS) according the 90th percentile (14 days). Logistic regression (LR), classification and regression trees (CART), random forest (RF), gradient boosting (GB) and neural networks (NN) were applied to the collected data. The predictive performance of the models was evaluated using the area under the ROC curve (AUC).

Results. Our analysis included 73,182 hospitalizations, of which 7,341 (10.0%) led to long LOS. The GB classifier was the most performant model with the highest AUC (0.810), superior to all the other models (all p-values<0.0001). The performance of the RF, GB and NN models (AUC ranged from 0.808 to 0.810) was superior to that of the LR model (AUC=0.795); all p-values<0.0001. In contrast, LR was superior to CART (AUC=0.786), p<0.0001. The variable most predictive of the long LOS was the destination of the patient after hospitalization to other institutions. The typical clinical profile of these patients (17.5% of the sample) was the elderly patient, admitted in emergency, for a trauma, a neurological or a cardiovascular pathology, more often institutionalized, with more comorbidities notably mental health problems, dementia and hemiplegia.

Discussion. The integration of ML, particularly the GB algorithm, may be useful for health care professionals and bed managers to better identify patients at risk of prolonged LOS. These findings underscore the need to strengthen hospitals through targeted allocation to meet the needs of an aging population.

Keywords: Machine learning; Neural network; Prediction; Health Services Research; Public Health.

1- Introduction

In 2019, healthcare expenditure (consumption of care and medical goods, CSBM) amounted to €208 billion in France, of which €97 billion was for hospital care (46.7%) (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques, 2020). In addition to being the largest contributor to health care spending, hospital expenditure accelerated in 2019 (+2.4%) to the point of increasing faster than the CSBM (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques, 2020). In France as in other Western countries, strategies to control health expenditure are similar and are notably based on the reduction in length of stay (LOS) (Baumann & Wyss, 2021). Numerous studies show that some of the beds occupied in hospitals in France are inadequately occupied, with approximately 10% of medical and surgical beds being inadequately occupied on a given day (5% in surgery, 17.5% in medicine) (2011). LOS, defined as the interval time between admission and discharge (i.e., total bed-days occupied by a patient), is thus considered as an important indicator to evaluate quality of care and hospital performance. Prolonged LOS is associated with more consumption of hospital resources and costs, more complications (e.g., hospital-acquired infection, falls), increased mortality and deteriorated patient experience (Marfil-Garza et al., 2018a; Rojas-García et al., 2018). In addition, prolonged LOS may impact negatively on admission of critically ill patients and denies timely access to treatment (Tefera et al., 2020). For all these reasons, we need to better identify patients at high risk of prolonged LOS to improve the quality of care and reduce associated health care costs.

Recently, machine learning (ML) methods have been introduced in health service research and have shown a better level of prediction than traditional statistical approaches such as logistic regression (LR) in several domains (Acion et al., 2017; Ahn et al., 2018; Chekroud et al., 2016; Gholipour et al., 2015). ML methods offer key benefits over traditional statistical approaches because they account for nonlinear relationships between the outcome and the predictors and yield more stable predictions (M. Kuhn & Johnson, 2013). ML methods account for interactions between predictors which relaxes the homogeneity assumption that there are no interactions among predictors. To our knowledge, ML methods have rarely been applied to improve the prediction of LOS. A recent study developed models using a ML approach (i.e., elastic-net, gradient boosted trees, random forest, support vector machines, logistic regression, a Eureqa classifier, generalized additive models, a Vowpal Wabbit classifier, K-nearest neighbors classifiers, residual neural network, a Rulefit classifier) to predict LOS in patients hospitalized for COVID-19 (N=966 patients) (Ebinger et al., 2021) Another recent study explored two ML methods, the Random Forest (RF) and the Gradient Boosting model (GB), using an open source available dataset (Mekhaldi et al., 2020). Last, Bacchi et al. applied neural network model to 313 patients admitted in general medical stay (Bacchi, Gluck, et al., 2020) Altogether, these findings suggest that ML approach may help hospital systems prepare for bed capacity needs.

Thus, the objective of this work was to predict LOS using ML methods on a large population-based study from a French hospital medico-administrative database, based on the area under the receiving operating characteristic curve. For this purpose, we selected the following ML methods (Fernández-Delgado et al., 2014): random forest (RF), neural networks (NN), gradient boosting (GB), decision trees (CART), in addition to LR.

2- Methods

Study design

This was a retrospective cohort study of all acute-care inpatient hospitalization cases discharged from January 1 to December 31, 2015, from the largest university health center in south France

(Assistance Publique – Hôpitaux de Marseille, APHM). All data were collected from the French Hospital database (PMSI - Programme de Médicalisation des Systèmes d'Information) (Boudemaghe & Belhadj, 2017). The PMSI is the French medico-administrative database for all hospitalizations based on diagnosis-related groups that we could group into significant diagnostic categories. Research on such retrospective data are excluded from the framework of the French Law Number 2012-300 of 5 March 2012 relating to the research involving human participants, as modified by the Order Number 2016-800 of 16 June 2016. Neither the French competent authority (Agence Nationale de Sécurité du Médicament et des Produits de Santé, ANSM) approval nor the French ethics committee (Comités de Protection des Personnes, CPP) approval is required in this context.

Study setting and inclusion criteria

The APHM is a public tertiary-care center comprising four hospitals (La Timone, La Conception, Sainte-Marguerite, and North) with 3,400 beds and 2,000 physicians. Approximately 300,000 hospitalizations are recorded every year at the APHM, involving approximately 210,000 patients. The inclusion criteria were all acute-care hospitalizations for patients older than 18 years old and with a length of stay (LOS) > 24 hours (to exclude ambulatory care such as ambulatory surgery, radiotherapy, dialysis, chemotherapy, and transfusions that we did not want to predict). Were also excluded in-hospital mortalities and obstetrical stays.

Study outcomes

The study outcome was LOS transformed into a binary variable (long vs. short LOS). We considered a long stay any potential outlier of the quantitative LOS variable, according to Tukey's criterion (Everitt & Skrondal, 2010; Tukey, 1977b) that is, any stay greater than the boxplot upper fence = Q3 + 1.5*IQR where Q3 indicates the third quartile and IQR the interquartile interval. In our case, the cut point coincides with the 90th percentile (14 days).

Collected data

The dataset collected from the PMSI used 27 predictor variables:

- sociodemographic characteristics: age, gender, state-funded medical assistance (Aide Médicale d'Etat, AME) (i.e., health coverage for undocumented migrants), and free universal health care (Couverture Maladie Universelle, CMU) (i.e., universal health coverage for those not covered by employment/business-based schemes);

- clinical characteristics: category of disease based on the 10th revision of the International Statistical Classification of Diseases and 17 comorbidities from the Charlson comorbidity index (Quan et al., 2005);

- hospitalization characteristics: patient origin (home or other hospital institution), hospitalization via emergency departments, destination after hospital discharge (home or transfer to other hospital institution), and hospitalization via emergency departments in the previous 6 months.

Statistical models

Five distinct types of predictive models were fitted to the data: LR, CART, RF, GB, and three-hidden layers NN. These models have been explained elsewhere in detail (Hastie et al., 2009); a brief summary is presented here.

LR is a general linear model of the exponential family such that $\ln\left(\frac{\pi}{1-\pi}\right) = w^T x$, where $\pi = P(y=1|x)$, where y is a binary outcome, x the predictors and w is the weight vector to be estimated from the data.

CART (Breiman et al., 1984) is a binary decision tree (DT) method that involves segmenting the predictor space into a number of simple regions. CART can be applied to both regression and classification problems, as in our study. A DT is constructed through an iterative process by applying a binary splitting rule. For each explanatory variable x_j in the data, a rule of the form $x_j < a$ (a ∈ R is a threshold) is used to split the initial set of observations (denoted t_0, the root of the tree) into two subsets t_l and t_r (the sibling nodes). Each observation falling in those regions is then predicted by the highest frequency class. The best split is defined as the one minimizing a loss function (e.g., the Gini index, or the Entropy). Once the best split has been defined, the same process is applied to the two nodes t_l and t_r and repeated until a predefined minimum number of observations is reached. Then, a pruning algorithm can be used to search for an optimal subtree, given a penalty criterion (complexity parameter) applied to the objective function. A DT can be represented graphically and thus can be directly interpretable, given its simple structure.

RF (Breiman, 2001) is an ensemble learning method based on aggregating n_estimators trees similar to the ones constructed with CART, each one grown using a bootstrap sample of the original data set. Each tree in the forest uses only a random subset of max_features predictors to determine the best split at each node. The trees are not pruned. The prediction by RF is the majority vote over the predictions made by the n_estimators trees. Other hyperparameters such as the minimum number of samples required to split an internal node (min_samples_split) or the maximum depth of a tree (max_depth) may be used to tune further the RF model.

GB (Friedman, 2001) is also an ensemble learning method based on DT but does not involve bootstrap sampling. It is built sequentially using a weak learner (e.g., shallow classification trees). The GB is initialized with the best guess of the response (e.g., the majority vote), then the gradient is calculated, and a model is then fit to the residuals to minimize the loss function. The current model thus obtained is added to the previous model, adjusted by a learning_rate parameter. The user may specify the number of trees (n_estimators), a tree depth equal to max_depth and a given minimum number of observations in the trees terminal nodes, min_samples_leaf.

NN (Arbib, 1995) are nonlinear statistical models for regression or classification. They are structured in layers of “neurons” where the input layer is made of the predictor variables, followed by intermediate layers called hidden layers, and the output layer. Each neuron is a linear combination of the neurons of the previous layer, to which is applied a non-linear activation function, typically the relu function: $relu(x) = \max(0, x)$. Usually, the activation function used in the output layer is the softmax for multiclass classification and the sigmoid for binary classification. Thus, the output layer contains as many neurons as there are classes, but only one for binary classification. The weights of the linear combinations are the parameters of the model, and they are estimated through an optimization algorithm called (stochastic) gradient descent. The loss function optimized in binary classification is the cross-entropy to which a decay penalty is applied.

Statistical analyses

Descriptive analyses for the sociodemographic, clinical, and hospitalization data were expressed as frequencies and percentages. For each predictor (sociodemographic, clinical, and hospitalization data), the two categories of LOS (long vs. short) were compared by estimating their difference in proportions through a statistical test of proportions. The effect size of this difference is then estimated with Cohen’s d standardized difference (SD). SD use effect size methods to identify meaningful differences between groups that, unlike p-values, are not influenced by sample size. Values greater than 0.20 are clinically significant (Goulet-Pelletier & Cousineau, 2018).

In the following, model performance is estimated through the area under the receiver operating characteristic curve (ROC, AUC). Indeed, given that our outcome class proportions are quite unbalanced (90% short vs 10% long LOS), threshold-dependent measures of performance such as the accuracy or the F1 are less reliable (J. Huang & Ling, 2005; Wardhani et al., 2019).

To train and evaluate the different models (i.e., LR, CART, RF, NN, and GB), the dataset was split into 80% full training sample and 20% hold out test sample, stratified on the outcome variable. The first step was to tune each of the different model (i.e., CART, RF, NN, and GB - LR, as the reference model has no hyperparameter to be tuned). The 80% full training sample is again split into 80% training set and 20% validation set. We performed a 10-fold cross validation to tune the hyperparameters with the training set, then assessed model performance with the validation set for that specific resampling split, and the optimal hyperparameters for that resampling split is saved. This process is repeated 10 times over 10 different resampling splits. The hyperparameters corresponding to the highest performance over these 10 resampling splits are now used to compare each of the 5 models 100 times over 100 different resampling splits. The performance of each model is saved for each split and the mean performances of the different models over 100 splits are compare using paired t-test (post hoc tests with Bonferroni correction). Given the large sample size, the p-value of the test statistic is completed with the Cohen's size effect, to appreciate the amplitude of the difference in performance. In addition, we computed the performance of each model (classifier) on the hold out test sample which the model has never "seen"– this is not only a supplementary indication on the classifier's performance but also provides the means to check for overfitting.

Lastly, we computed variable importance (VI), averaged over the 100 resampling splits. VI provides a simple way to inspect each model and gain insights on which variables are most influential in predicting the outcome, and to what extent. Here, permutation feature importance is used to estimate variable importance. Permutation feature importance is defined as the decrease in a model score when a single feature value is randomly shuffled (Altmann et al., 2010a; Breiman, 2001). The larger the decrease in score, the more important the variable.

All analyses were implemented in Python 3.7 (Van Rossum & Drake, 2009) with Sci Kit Learn 0.24.1 (Pedregosa et al., 2011) and Keras 2.4.0 (Chollet, 2015)

3- Results

Characteristics of the population

A total of 289,358 hospitalizations were recorded in 2015. After exclusion of non-adult stays with death and hospitalizations for ambulatory and obstetrical care, 73,182 hospitalizations were retained. The most common diseases were digestive disease and nervous system conditions. In total, 7341 (10.03%) hospitalizations resulted in long LOS. The characteristics of the sample are presented in Table 1.

Factors associated with LOS

Based on the Cohen's d standardized difference in proportions, the destination of discharge to other institutions shows a significant and sizeable higher proportion of long LOS than to home (d=0.727 p-value < 0.0001). Next comes those who are admitted for Chemotherapy and Radiotherapy who display a sizeable and significant lower level of long LOS (d=-0.390, p-value < 0.0001), followed by the origin of patient where other institutions are associated to higher proportion of long LOS (d=0.294, p-value<0.0001). Table 1 displays all the significant difference in proportion of LOS for which the size effect is at least equal to 0.2 (small effect).

Predictive model performance

The predictive performance of each model is presented in Table 2, and the comparison of each model's AUC is presented in Table 3. The GB classifier was the most performant model with the highest AUC (0.810), superior to all the other models (all p-values<0.0001). The performance of the RF, GB and NN models (AUC ranged from 0.808 to 0.810) was superior to that of the LR model (AUC=0.795); all p-values<0.0001. In contrast, LR was superior to CART (AUC=0.786), p<0.0001. As the values are close, the size effects are also provided by the Cohen's d, which confirms small effects between GB and RF or NN but large effects between all others. Thus, the seemingly small difference in value between the AUC of LR and the other classifiers, when accounting for their standard errors are in fact very large ones. However, the performance of NN and RF are identical. The ROC curve for the best model (i.e., GB) is presented in Figure 1.

Variable importance

The variable importance of the best model (i.e., GB) is presented in Figure 2. In the GB classifier as well as in all the others, the variable most predictive of the categorical LOS was the destination of the patient after hospitalization. Destination to other institutions but not home was associated to long LOS. The typical clinical profile of these patients (17.5% of the sample) was the elderly patient, admitted in emergency, for a trauma, a neurological or a cardiovascular pathology, more often institutionalized, with more comorbidities, notably dementia and hemiplegia (supplementary file #1). This is coherent with the bivariate analysis. Two of the other most important variables were also identified in the bivariate analysis: the origin of the patient from other institutions was predictive of long LOS, whereas the admission for chemotherapy or radiotherapy was associated with short LOS. The model also included admission for orthopedic trauma and surgical type of hospital stay to be predictive of long LOS.

The variable importance of the other models is presented in supplementary file #2.

4- Discussion

One of the strategies to address the sustainability of health care systems is to reduce the length of inpatient hospital stay. Reducing LOS is expected to release bed capacity as well as staff time and to reduce costs associated with inappropriate patient days in hospital. In addition, prolonged LOS is associated with more medical complications and longer discharge delays. Therefore, improving LOS prediction with the best artificial intelligence method remains a key challenge, especially to enable better bed planning, care delivery and cost optimization. Linear and logistic regression methods have been supplanted by ML and deep learning (DL) models, yet it remains challenging to identify, benchmark and select optimal prediction methods given the discrepancy in data sources, inclusion criteria, choice of input variables, and metrics used (Bacchi, Tan, et al., 2020; Lequertier et al., 2021).

In our study, GB displays the best performance level for predicting LOS. In a recent study (Rachda Naila et al., 2021), LOS prediction was modeled with multiple linear regression, support vector machine, RF and GB. GB outperformed all the other models using a basic training-test split with a 70%-30% ratio. In another study, RF slightly outperformed GB (Mekhaldi et al., 2020). NN as a multiple layer perceptron (MLP) is often used as a benchmark to other ML models but GB consistently outperforms NN on tabular datasets (Bacchi, Gluck, et al., 2020; Fernández-Delgado et al., 2014). This is verified again here for the three-hidden layers NN (5 layers MLP).

Scientific efforts to provide accurate prediction of LOS have been steady for half of a century (Lequertier et al., 2021). While the use of ML in health-related research has become more and more popular, its application on LOS remains scattered. A recent systematic review conducted by Bacchi et

al. (Bacchi, Tan, et al., 2020) identified only 21 articles predicting LOS including regression and classification as well as different medical specialties group patients. Several shortcomings have been highlighted by the authors and considered in our work.

The failure to provide the criteria of inclusion as well as the lack of demographic and clinical information such as disease prevalence details: this issue has been carefully considered in our work with detailed clinical and organizational information.

The lack of information regarding the distribution of the LOS outcome and the handling of the outliers: in our study, we considered a long stay any potential outlier of the quantitative LOS variable, according to a valid and reproducible criterion: Tukey's criterion (Everitt & Skrondal, 2010; Tukey, 1977b). The distribution of long and short LOS is provided for the whole dataset and for each variable.

The absence of separate datasets for training and assessment leading to overfitting (i.e., inflation of the model performance) (Bacchi, Tan, et al., 2020): model assessment must be implemented on a dataset never seen by the trained model. Selecting randomly a test-training split of the data set might lead to an overly optimistic or pessimistic outcome (Hastie et al., 2009; Lequertier et al., 2021). Hence, cross validation is recognized as an alternative. However, k-fold cross validation may also lead to overfitting unless separate validation sets are used (Bacchi, Tan, et al., 2020; Cawley & Talbot, 2010). Thus, some authors suggest that rigorous performance evaluation requires multiple randomized partitioning of the available data, with model selection performed separately in each trial (Cawley & Talbot, 2010, p. 2103). In this study, we have used separate validation sets for model selection and hyperparameter tuning and another different holdout test set to check for overfitting.

Beyond these noted limitations, the authors reported other areas on which improvements are needed.

The systematic review reported the absence of feature importance. One reason why this is not implemented is that most of the learners use their inbuilt feature importance computation, while others do not. Permutation importance may be called for estimating feature importance in a way that is equivalent for all ML models. Thus, in our case all the learners concur that the feature most predictive (by far) of long LOS is the Destination of Patient on Discharge to other but home.

The review also alerts on the use of resampling-based statistical tests to compare performance. To account for any randomness involved in training-validation splits we may supplement any performance comparison with, say 100 resampling of the training and validation set. From this perspective, each learner becomes comparable to an experimental condition and each resampling to a statistical unit. It now becomes possible to apply a means comparison between the learners over 100 samples, using for example post-hoc methods and Bonferroni correction. And the observed difference cannot only be estimated in terms of statistical significance but also in terms of effect size (Bland & Altman, 2015). Under this perspective, the use of the holdout test sample becomes at best a way of verifying the absence of overfitting.

Finally, our findings identify important levers for action for health care professionals, planners and health policy. Destination to other institutions, especially for elderly patient, admitted in emergency, for a trauma, a neurological or a cardiovascular pathology, more often institutionalized and with more comorbidities were associated with substantial long LOS. Previous studies have shown that discharge destination have significant impact on LOS. In a sample of 313 144 medical records of all patients older than 18, discharge destination was one of the main LOS predictors (Brasel et al., 2007). In addition, another study confirmed that older patients had prolonged LOS (>17days) which was

associated with discharge to places other than usual residence (Lisk et al., 2019). Indeed, hospitalizations are frequently associated in older people with an increased risk of functional decline both during hospitalization and following discharge (Koskas et al., 2019) These findings provide a rationale for increased staffing for elderly patients requiring intensive care in hospitals, particularly for those with cognitive impairment and multiple comorbidities. Needing more caring time than usual was reported for 20% of older patients in general and for 57% of the patients with dementia (Hendlmeier et al., 2019). Considering the demographic change, this situation will worsen and there is thus an urgent need to strengthen hospitals with targeted allocation to meet the needs of an aging population.

Perspectives and limitations. Our findings must be interpreted in the context of our study's limitations. Despite the large overall sample size of this multi-center study, our findings may not be applicable to all French hospitals, particularly concerning general hospitals whose patients offer potentially different characteristics from those of university hospitals. In addition, the four university hospitals included in our study were located in only one geographical area, even though social and healthcare geographical characteristics (e.g., poverty, density of physicians, number of beds, and private hospitals) are known to influence LOS. Future studies should thus be conducted in different categories of hospitals and several geographical areas to confirm the properties and interest of our method. An external validation in addition to the internal validation performed in this study will guarantee the generalizability of this method.

Conclusion

The integration of ML, particularly the GB algorithm, may be useful for health care professionals and planners to better identify patients at risk of prolonged LOS. These findings underscore the need to strengthen hospitals through targeted allocation to meet the needs of an aging population.

Acknowledgments: None

Conflict of interest: None

Funding: None

Figure 1 – Best Model: Gradient Boosting Mean ROC Curve

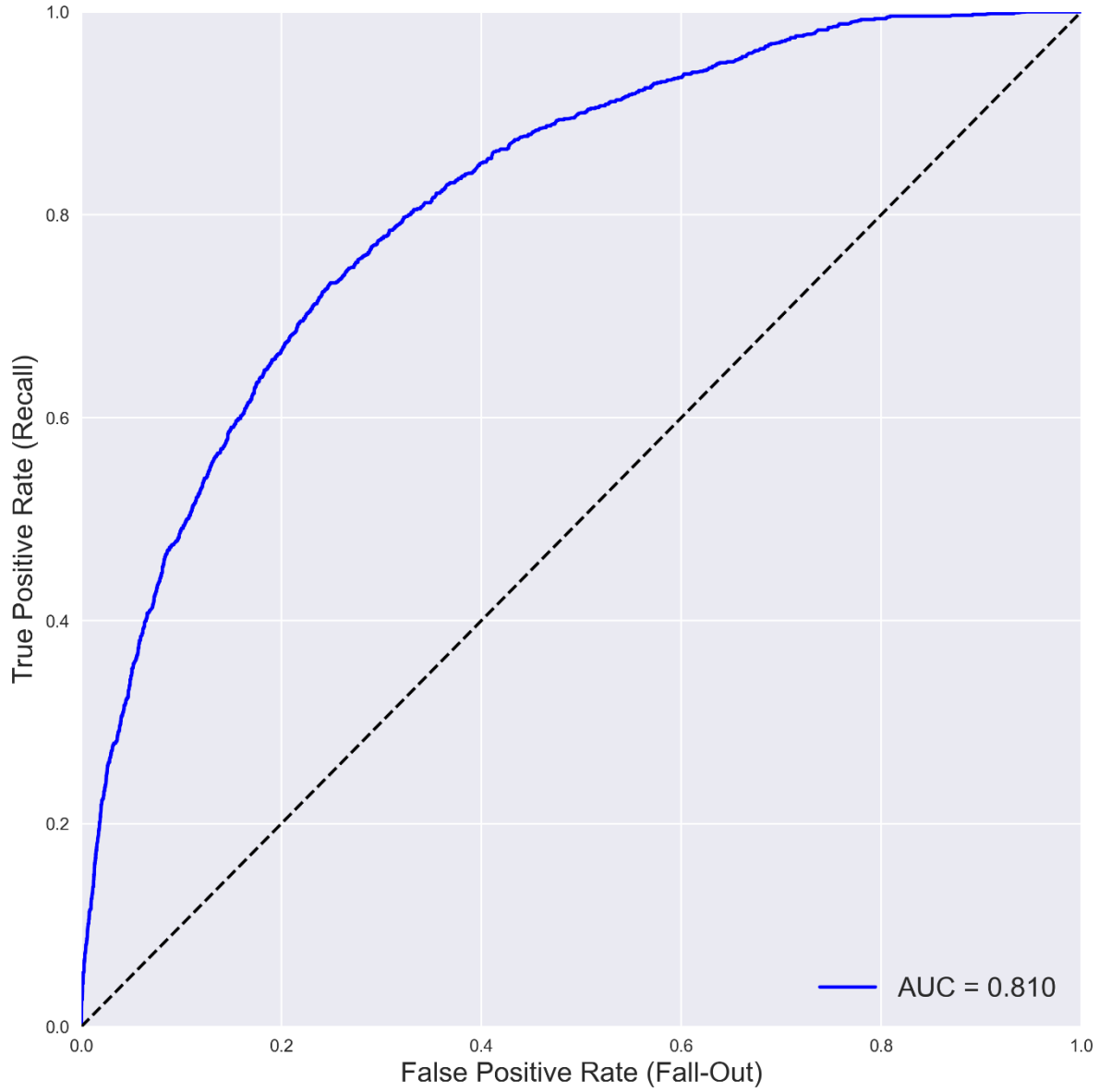
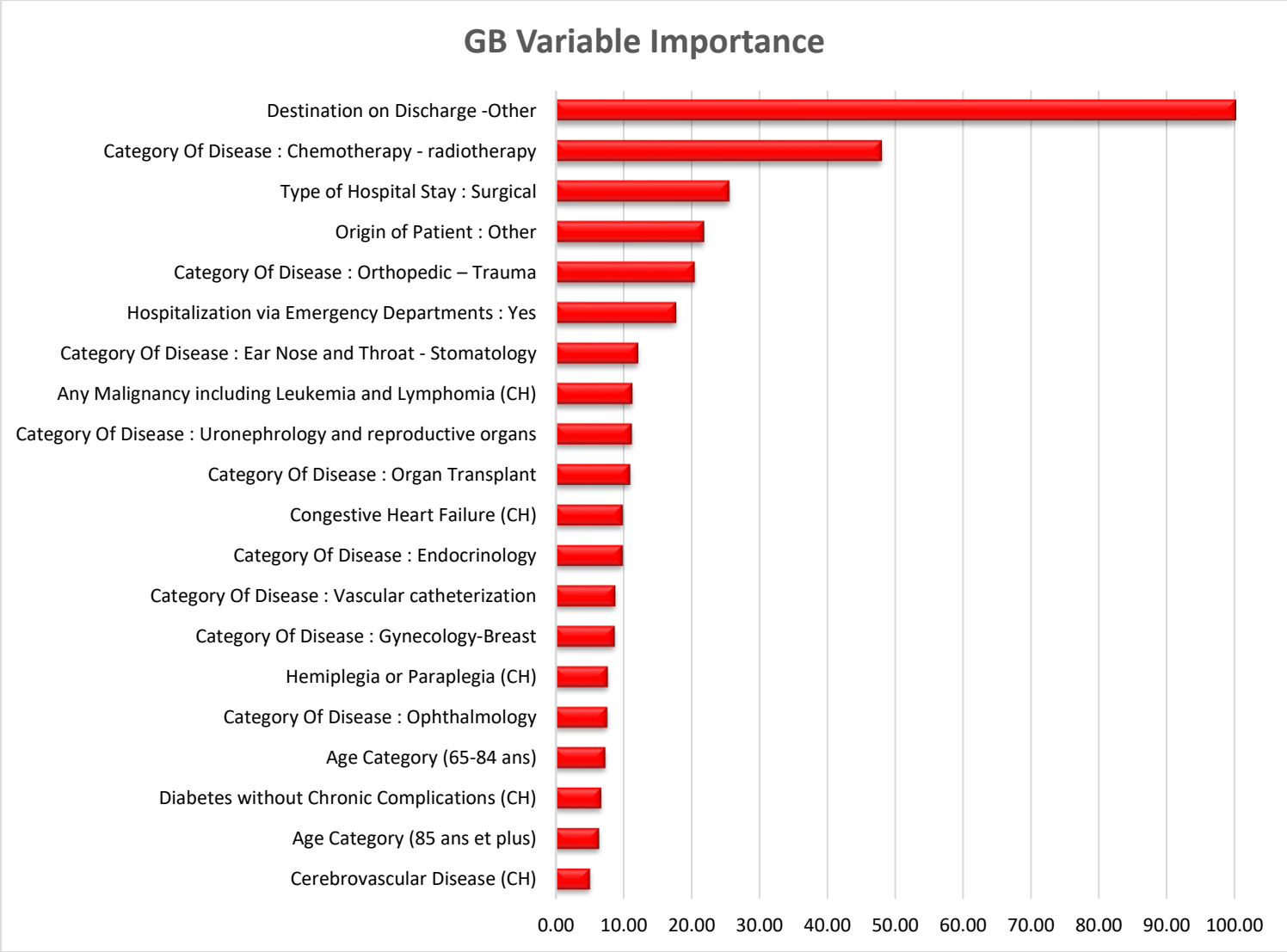


Figure 2 – Gradient Boosting Variable importance (20 highest)



[109]

Table 1 – Sample characteristics (significant effect size are highlighted in yellow)

Variable	Modality	N	(%)	LOS-	(%-)	LOS+	(%+)	Comparing modality LOS proportions	
	Total	73182	100	65841	89.97	7341	10.03	modality p-value	Cohen's d (long - short)
Gender	1-Male	39065	53,38	34805	52,86	4260	58,03	0,000	0,104
	2-Female	34117	46,62	31036	47,14	3081	41,97	0,000	-0,104
State Funded Medical Assistance	No	72383	98,91	65134	98,93	7249	98,75	0,161	-0,017
	Yes	799	1,09	707	1,07	92	1,25	0,161	0,017
Free Universal Health Care	No	66991	91,54	60210	91,45	6781	92,37	0,007	0,034
	Yes	6191	8,46	5631	8,55	560	7,63	0,007	-0,034
Type of Hospital Stay	1-Medical	44396	60,67	40555	61,60	3841	52,32	0,000	-0,188
	3-Surgical	28786	39,33	25286	38,40	3500	47,68	0,000	0,188
Origin of Patient	1-Home	68024	92,95	61792	93,85	6232	84,89	0,000	-0,294
	2-Other	5158	7,05	4049	6,15	1109	15,11	0,000	0,294
Hospitalization via Emergency Departments	No	52412	71,62	47841	72,66	4571	62,27	0,000	-0,223
	Yes	20770	28,38	18000	27,34	2770	37,73	0,000	0,223
Destination on discharge	1-Home	60401	82,54	56409	85,67	3992	54,38	0,000	-0,727
	2-Other	12781	17,46	9432	14,33	3349	45,62	0,000	0,727
At least one previous hospitalization via emergency departments 6 months before	1-No hospitalization	43198	59,03	39123	59,42	4075	55,51	0,000	-0,079
	2-At least one non emergency	20133	27,51	18263	27,74	1870	25,47	0,000	-0,051
	3-At least one with emergency	9851	13,46	8455	12,84	1396	19,02	0,000	0,169
Age Category	2(18-44 ans)	16028	21,90	15057	22,87	971	13,23	0,000	-0,253
	3(45-64 ans)	24501	33,48	22307	33,88	2194	29,89	0,000	-0,086
	4(65-84 ans)	26290	35,92	23084	35,06	3206	43,67	0,000	0,177
	5(85 ans et plus)	6363	8,69	5393	8,19	970	13,21	0,000	0,163

Variable	Modality	N	(%)	LOS-	(%-)	LOS+	(%+)	Comparing modality LOS proportions	
	Total	73182	100	65841	89.97	7341	10.03	modality p-value	Cohen's d (long - short)
Category Of Disease	01-Digestive	8427	11,52	7412	11,26	1015	13,83	0,000	0,078
	02-Orthopedic – Trauma	5955	8,14	5590	8,49	365	4,97	0,000	-0,141
	03-Multiple or complex trauma	250	0,34	121	0,18	129	1,76	0,000	0,161
	04-Rheumatology	2430	3,32	2175	3,30	255	3,47	0,440	0,009
	05-Nervous system	8665	11,84	7314	11,11	1351	18,40	0,000	0,207
	06-Vascular catheterization	5929	8,10	5547	8,42	382	5,20	0,000	-0,128
	07-Cardiovascular	6313	8,63	5192	7,89	1121	15,27	0,000	0,232
	08-Pulmonary	6038	8,25	5338	8,11	700	9,54	0,000	0,050
	09-Ear Nose and Throat - Stomatology	2974	4,06	2862	4,35	112	1,53	0,000	-0,168
	10-Ophthalmology	1330	1,82	1296	1,97	34	0,46	0,000	-0,138
	11-Gynecology-Breast	1863	2,55	1829	2,78	34	0,46	0,000	-0,184
	13-Uronephrology and reproductive organs	4842	6,62	4507	6,85	335	4,56	0,000	-0,099
	14-Hematology	2031	2,78	1757	2,67	274	3,73	0,000	0,060
	15-Chemotherapy - radiotherapy	5395	7,37	5365	8,15	30	0,41	0,000	-0,390
	16-Infectious diseases	788	1,08	668	1,01	120	1,63	0,000	0,054
	17-Endocrinology	3655	4,99	3480	5,29	175	2,38	0,000	-0,152
	18-Cutaneous and subcutaneous	2393	3,27	2161	3,28	232	3,16	0,578	-0,007
	19-Burns	171	0,23	117	0,18	54	0,74	0,000	0,083
	20-Psychiatry	613	0,84	480	0,73	133	1,81	0,000	0,097
	21-Toxicology - Intoxication - Alcohol	656	0,90	609	0,92	47	0,64	0,014	-0,032
	22-Chronic pain palliative care	86	0,12	82	0,12	4	0,05	0,097	-0,023
	23-Organ Transplant	224	0,31	21	0,03	203	2,77	0,000	0,234
	24-Interdisciplinary activities and follow-up of patients	2154	2,94	1918	2,91	236	3,21	0,147	0,018

Variable	Modality	N	(%)	LOS-	(%-)	LOS+	(%+)	Comparing modality LOS proportions	
	Total	73182	100	65841	89.97	7341	10.03	modality p-value	Cohen's d (long - short)
Renal Disease (CH)	No	69327	94,73	62725	95,27	6602	89,93	0,000	-0,205
	Yes	3855	5,27	3116	4,73	739	10,07	0,000	0,205
Rheumatologic Disease (CH)	No	72567	99,16	65314	99,20	7253	98,80	0,000	-0,040
	Yes	615	0,84	527	0,80	88	1,20	0,000	0,040
Peripheral Vascular Disease (CH)	No	71163	97,24	64232	97,56	6931	94,41	0,000	-0,161
	Yes	2019	2,76	1609	2,44	410	5,59	0,000	0,161
Peptic Ulcer Disease (CH)	No	72942	99,67	65662	99,73	7280	99,17	0,000	-0,076
	Yes	240	0,33	179	0,27	61	0,83	0,000	0,076
Hemiplegia or Paraplegia (CH)	No	70744	96,67	64076	97,32	6668	90,83	0,000	-0,277
	Yes	2438	3,33	1765	2,68	673	9,17	0,000	0,277
Moderate or Severe Liver Disease (CH)	No	72733	99,39	65489	99,47	7244	98,68	0,000	-0,082
	Yes	449	0,61	352	0,53	97	1,32	0,000	0,082
Mild Liver Disease (CH)	No	71201	97,29	64166	97,46	7035	95,83	0,000	-0,090
	Yes	1981	2,71	1675	2,54	306	4,17	0,000	0,090
Metastatic Solid Tumor (CH)	No	70756	96,68	63815	96,92	6941	94,55	0,000	-0,118
	Yes	2426	3,32	2026	3,08	400	5,45	0,000	0,118
Any Malignancy including Leukemia and Lymphomia (CH)	No	68124	93,09	61720	93,74	6404	87,24	0,000	-0,223
	Yes	5058	6,91	4121	6,26	937	12,76	0,000	0,223
AIDS/HIV (CH)	No	72824	99,51	65525	99,52	7299	99,43	0,283	-0,013
	Yes	358	0,49	316	0,48	42	0,57	0,283	0,013
Diabetes with Chronic Complications (CH)	No	69977	95,62	63144	95,90	6833	93,08	0,000	-0,124
	Yes	3205	4,38	2697	4,10	508	6,92	0,000	0,124
Diabetes without Chronic Complications (CH)	No	67245	91,89	60956	92,58	6289	85,67	0,000	-0,223
	Yes	5937	8,11	4885	7,42	1052	14,33	0,000	0,223
Dementia (CH)	No	71196	97,29	64236	97,56	6960	94,81	0,000	-0,144
	Yes	1986	2,71	1605	2,44	381	5,19	0,000	0,144

Cerebrovascular Disease (CH)	No	70966	96,97	64165	97,45	6801	92,64	0,000	-0,223
	Yes	2216	3,03	1676	2,55	540	7,36	0,000	0,223
Chronic Pulmonary Disease (CH)	No	70512	96,35	63566	96,54	6946	94,62	0,000	-0,094
	Yes	2670	3,65	2275	3,46	395	5,38	0,000	0,094
Congestive Heart Failure (CH)	No	68976	94,25	62510	94,94	6466	88,08	0,000	-0,248
	Yes	4206	5,75	3331	5,06	875	11,92	0,000	0,248
Myocardial Infarction (CH)	No	68635	93,79	61959	94,10	6676	90,94	0,000	-0,120
	Yes	4547	6,21	3882	5,90	665	9,06	0,000	0,120

Table 2 – Performance of the tuned classifiers over 100 (re)sampling experiments

	100 sampling experiments **Mean AUC**
Logistic Regression (LR)	**0,7947**
Classification Trees (CT)	**0,7858**
Random Forest (RF)	**0,8086**
Gradient Boosting (GB)	**0,8101**
Neural Networks (NNET)	**0,8085**

Table 3 – AUC paired T-tests of classifiers' performance over 100 experiments (*Bonferonni corrected, with effect size*)

Classifier A	Classifier B	T Statistic	dof	p-uncorrected	p-corrected	Cohen's d
CT	GB	-57,78	99	0,0000	0,0000	-3,75
CT	LR	-19,66	99	0,0000	0,0000	-1,34
CT	NNET	-52,98	99	0,0000	0,0000	-3,53
CT	RF	-60,60	99	0,0000	0,0000	-3,51
EN	GB	-71,35	99	0,0000	0,0000	-2,38
GB	LR	72,95	99	0,0000	0,0000	2,43
GB	NNET	6,90	99	0,0000	0,0000	0,26
GB	RF	7,98	99	0,0000	0,0000	0,26
LR	NNET	-49,40	99	0,0000	0,0000	-2,19
LR	RF	-48,87	99	0,0000	0,0000	-2,18
NNET	RF	-0,04	99	0,9702	1,0000	0,00

Supplementary file #1: Profiles of Patients Discharge (*significant effect size are highlighted in yellow*)

Variable	Modality	N	HOME	(% home)	OTHER	(% other)	Cohen's d
Gender	Male	39065	32498	53.80%	6567	51.38%	-0.049
	Female	34117	27903	46.20%	6214	48.62%	0.049
State Funded Medical Assistance	No	72383	59719	98.87%	12664	99.08%	0.021
	Yes	799	682	1.13%	117	0.92%	-0.021
Free Universal Health Care	No	66991	55021	91.09%	11970	93.65%	0.097
	Yes	6191	5380	8.91%	811	6.35%	-0.097
Type of Hospital Stay	Medical	44396	37644	62.32%	6752	52.83%	-0.193
	Surgical	28786	22757	37.68%	6029	47.17%	0.193
Origin of Patient	Home	68024	57255	94.79%	10769	84.26%	-0.349
	Other	5158	3146	5.21%	2012	15.74%	0.349
Hospitalization via Emergency Departments	No	52412	45019	74.53%	7393	57.84%	-0.358
	Yes	20770	15382	25.47%	5388	42.16%	0.358
At least one previous hospitalization via emergency departments 6 months before	No hospitalization	43198	35784	59.24%	7414	58.01%	-0.025
	At least one non-emergency	20133	17188	28.46%	2945	23.04%	-0.124
	At least one with emergency	9851	7429	12.30%	2422	18.95%	0.184
Age Category	(18-44 years)	16028	14713	24.36%	1315	10.29%	-0.378
	(45-64 years)	24501	21187	35.08%	3314	25.93%	-0.200
	(65-84 years)	26290	20601	34.11%	5689	44.51%	0.214
	(85 years and more)	6363	3900	6.46%	2463	19.27%	0.390
LOS Categorical	0 (short)	65841	56409	93.39%	9432	73.80%	-0.549
	1 (long)	7341	3992	6.61%	3349	26.20%	0.549

Variable	Modality	N	HOME	(% home)	OTHER	(% other)	Cohen's d
Category Of Disease	01-Digestive	8427	7604	12.59%	823	6.44%	-0.211
	02-Orthopedic – Trauma	5955	3968	6.57%	1987	15.55%	0.289
	03-Multiple or complex trauma	250	115	0.19%	135	1.06%	0.110
	04-Rheumatology	2430	1953	3.23%	477	3.73%	0.027
	05-Nervous system	8665	6170	10.22%	2495	19.52%	0.264
	06-Vascular catheterization	5929	5112	8.46%	817	6.39%	-0.079
	07-Cardiovascular	6313	4471	7.40%	1842	14.41%	0.226
	08-Pulmonary	6038	4871	8.06%	1167	9.13%	0.038
	09-Ear Nose and Throat - Stomatology	2974	2802	4.64%	172	1.35%	-0.194
	10-Ophthalmology	1330	1276	2.11%	54	0.42%	-0.152
	11-Gynecology-Breast	1863	1821	3.01%	42	0.33%	-0.211
	13-Uronephrology and reproductive organs	4842	4197	6.95%	645	5.05%	-0.080
	14-Hematology	2031	1718	2.84%	313	2.45%	-0.025
	15-Chemotherapy - radiotherapy	5395	5161	8.54%	234	1.83%	-0.306
	16-Infectious diseases	788	609	1.01%	179	1.40%	0.036
	17-Endocrinology	3655	3328	5.51%	327	2.56%	-0.150
	18-Cutaneous and subcutaneous	2393	2075	3.44%	318	2.49%	-0.056
	19-Burns	171	120	0.20%	51	0.40%	0.037
	20-Psychiatry	613	388	0.64%	225	1.76%	0.103
	21-Toxicology - Intoxication - Alcohol	656	508	0.84%	148	1.16%	0.032
	22-Chronic pain palliative care	86	80	0.13%	6	0.05%	-0.029
	23-Organ Transplant	224	177	0.29%	47	0.37%	0.013
	24-Interdisciplinary activities and follow-up of patients	2154	1877	3.11%	277	2.17%	-0.059

Variable	Modality	N	HOME	(% home)	OTHER	(% other)	Cohen's d
Renal Disease (CH)	No	69327	57394	95.02%	11933	93.37%	-0.071
	Yes	3855	3007	4.98%	848	6.63%	0.071
Rheumatologic Disease (CH)	No	72567	59878	99.13%	12689	99.28%	0.016
	Yes	615	523	0.87%	92	0.72%	-0.016
Peripheral Vascular Disease (CH)	No	71163	58819	97.38%	12344	96.58%	-0.047
	Yes	2019	1582	2.62%	437	3.42%	0.047
Peptic Ulcer Disease (CH)	No	72942	60207	99.68%	12735	99.64%	-0.007
	Yes	240	194	0.32%	46	0.36%	0.007
Hemiplegia or Paraplegia (CH)	No	70744	59013	97.70%	11731	91.78%	-0.268
	Yes	2438	1388	2.30%	1050	8.22%	0.268
Moderate or Severe Liver Disease (CH)	No	72733	60030	99.39%	12703	99.39%	0.001
	Yes	449	371	0.61%	78	0.61%	-0.001
Mild Liver Disease (CH)	No	71201	58708	97.20%	12493	97.75%	0.035
	Yes	1981	1693	2.80%	288	2.25%	-0.035
Metastatic Solid Tumor (CH)	No	70756	58442	96.76%	12314	96.35%	-0.022
	Yes	2426	1959	3.24%	467	3.65%	0.022
Any Malignancy including Leukemia and Lymphomia (CH)	No	68124	56409	93.39%	11715	91.66%	-0.066
	Yes	5058	3992	6.61%	1066	8.34%	0.066
AIDS/HIV (CH)	No	72824	60104	99.51%	12720	99.52%	0.002
	Yes	358	297	0.49%	61	0.48%	-0.002
Diabetes with Chronic Complications (CH)	No	69977	57859	95.79%	12118	94.81%	-0.046
	Yes	3205	2542	4.21%	663	5.19%	0.046
Diabetes without Chronic Complications (CH)	No	67245	55856	92.48%	11389	89.11%	-0.117
	Yes	5937	4545	7.52%	1392	10.89%	0.117
Dementia (CH)	No	71196	59178	97.98%	12018	94.03%	-0.202
	Yes	1986	1223	2.02%	763	5.97%	0.202
Cerebrovascular Disease (CH)	No	70966	58889	97.50%	12077	94.49%	-0.154
	Yes	2216	1512	2.50%	704	5.51%	0.154
Chronic Pulmonary Disease (CH)	No	70512	58288	96.50%	12224	95.64%	-0.044
	Yes	2670	2113	3.50%	557	4.36%	0.044

Congestive Heart Failure (CH)	No	68976	57411	95.05%	11565	90.49%	-0.177
	Yes	4206	2990	4.95%	1216	9.51%	0.177
Myocardial Infarction (CH)	No	68635	56866	94.15%	11769	92.08%	-0.082
	Yes	4547	3535	5.85%	1012	7.92%	0.082

Supplementary File #2: 20 most important variables for each other classifier

	Logistic Regression Modalities	Importance
1	Destination on Discharge - Other	100,00
2	Category Of Disease : Chemotherapy - Radiotherapy	65,09
3	Category Of Disease : Orthopedic – Trauma	38,28
4	Origin of Patient : Other	21,23
5	Category Of Disease : Gynecology-Breast	17,72
6	Category Of Disease : Uronephrology and reproductive organs	15,93
7	Type Of Hospital Stay : Surgical	15,22
8	Age Category (65-84 years old)	14,43
9	Category Of Disease : Ear Nose and Throat - Stomatology	14,31
10	Hospitalization via Emergency Departments : Yes	11,87
11	Category Of Disease : Endocrinology	10,72
12	Category Of Disease : Vascular catheterization	9,66
13	Category Of Disease : Ophthalmology	9,54
14	Category Of Disease : Organ Transplant	9,05
15	Hemiplegia or Paraplegia (CH)	8,67
16	Age Category (85 years and more)	8,12
17	Congestive Heart Failure (CH)	7,30
18	Any Malignancy including Leukemia and Lymphomia (CH)	5,88
19	Dementia (CH)	5,36
20	Age Category (45-64 years old)	5,11

Classification Trees

	Modalities	Importance
1	Destination on Discharge - Other	100,00
2	Type Of Hospital Stay : Surgical	31,00
3	Hospitalization via Emergency Departments : Yes	29,45
4	Origin of Patient : Other	21,01
5	Category Of Disease : Chemotherapy - Radiotherapy	17,86
6	Any Malignancy including Leukemia and Lymphomia (CH)	17,09
7	Category Of Disease : Orthopedic – Trauma	15,34
8	Category Of Disease : Organ Transplant	10,97
9	Congestive Heart Failure (CH)	10,64
10	Dementia (CH)	8,71
11	Age Category (65-84 years old)	7,10
12	Age Category (85 years and more)	6,90
13	At least one previous hospitalization via emergency departments 6 months before	6,89
14	Hemiplegia or Paraplegia (CH)	6,44
15	Category Of Disease : Vascular catheterization	6,43
16	Category Of Disease : Ear Nose and Throat - Stomatology	6,25
17	Category Of Disease : Nervous system	5,65
18	CategoryOfDisease_07-Cardiovascular	4,88
19	Age Category (45-64 years old)	4,53
20	Cerebrovascular Disease (CH)	4,31

Random Forest

	Modalities	Importance
1	Destination on Discharge - Other	100,00
2	Origin of Patient : Other	18,52
3	Type Of Hospital Stay : Surgical	17,92
4	Hospitalization via Emergency Departments : Yes	17,30
5	Any Malignancy including Leukemia and Lymphomia (CH)	15,24
6	Category Of Disease : Chemotherapy - Radiotherapy	12,17
7	Congestive Heart Failure (CH)	9,68
8	Category Of Disease : Organ Transplant	9,29
9	Category Of Disease : Orthopedic – Trauma	8,79
10	Dementia (CH)	7,50
11	Hemiplegia or Paraplegia (CH)	7,35
12	Category Of Disease : Cardiovascular	7,08
13	Cerebrovascular Disease (CH)	5,49
14	Category Of Disease : Nervous system	5,21
15	Renal Disease (CH)	4,81
16	Age Category (85 years and more)	4,65
17	Category Of Disease : Multiple or complex trauma	4,31
18	At least one previous hospitalization via emergency departments 6 months before	4,12
19	Category Of Disease : Vascular catheterization	3,89
20	Age Category (65-84 years old)	3,82

	Neural Networks Modalities	Importance
1	Destination on Discharge - Other	100,00
2	Category Of Disease : Chemotherapy - Radiotherapy	48,74
3	Type Of Hospital Stay : Surgical	32,62
4	Category Of Disease : Orthopedic – Trauma	23,51
5	Origin of Patient : Other	20,23
6	Hospitalization via Emergency Departments : Yes	19,30
7	Category Of Disease : Uronephrology and reproductive organs	15,67
8	Category Of Disease : Vascular catheterization	15,13
9	Category Of Disease : Ear Nose and Throat - Stomatology	14,09
10	Category Of Disease : Endocrinology	11,33
11	Any Malignancy including Leukemia and Lymphomia (CH)	10,62
12	Congestive Heart Failure (CH)	10,22
13	Age Category (65-84 years old)	8,82
14	Hemiplegia or Paraplegia (CH)	8,62
15	Category Of Disease : Gynecology-Breast	8,56
16	Category Of Disease : Organ Transplant	8,12
17	At least one previous hospitalization via emergency departments 6 months before	8,10
18	Category Of Disease : Ophthalmology	7,07
19	Dementia (CH)	6,82
20	Age Category (85 years and more)	6,49

Synthèse des Conclusions – Article 2

Une récente revue de littérature systématique sur le LOS (Lequertier et al., 2021) nous fournit une excellente façon de discuter nos résultats.

Données : les auteurs ont regroupé les jeux de données utilisés dans la Recherche sur le LOS en 12 catégories :

1. Données administratives
2. Données démographiques et anthropométriques
3. Diagnostics et historique médical
4. Soins et procédures
5. Examens biologiques et paramètres physiologiques
6. Médicaments
7. Evènements adverses
8. Scores de risques
9. Temporalité et fréquence d'admission
10. Caractéristiques hospitalières
11. Caractéristiques du personnel soignant
12. Notes cliniques

Les données de la cohorte 2015 de l'APHM que nous avons utilisées excluent les points 4, 5, 6, 7, 11 et 12. Ce constat ouvre déjà une piste d'amélioration sur les jeux de données que nous pourrions mobiliser dans de futures recherches.

D'autre part, il est souligné (Lequertier et al., 2021) que la différence entre les jeux de données, les critères d'inclusion et les choix des variables rendent les benchmarks difficiles dans l'étude du LOS. Si dans notre cas, les critères d'inclusion sont explicitement annoncés, le choix du jeu de données n'échappe effectivement pas à cette critique. En effet, même en supposant la cohorte de 2015 comme étant typique d'une période d'hospitalisation annuelle, ces données restent très spécifiques à l'environnement socio-économique et géographiques des 4 hôpitaux de l'APHM.

Méthodes de prédiction : Lequertier et al (2021) ont identifié trois catégories :

1. Des modèles « traditionnelles » de régression qui visent à prédire le LOS en tant que variable quantitative ou binaire (régression linéaire, ridge, lasso, logistique), ou multiclasse (analyse discriminante). Cette catégorie inclut les modèles multiniveaux et les modèles temporels de survie de Cox.
2. Des modèles classiques de Machine Learning (arbres de décision de type CART, support vector machines), les modèles ensemblistes à base d'arbres de décision (Random Forest, Boosted Trees) ou d'autres modèles ensemblistes, réseaux Bayesiens, Modèles de Markov.
3. Des modèles basés sur les réseaux de neurones (Multiperceptrons), les réseaux de neurones profonds ou deep learning (réseaux de neurones récurrents (RNN), réseaux de neurones utilisant des mécanismes d'attention (Transformers) et modèles multimodaux).

Les modèles que nous avons mobilisés dans cette recherche puisent dans chacune de ces trois catégories, sans les inclure toutes. Dans la catégorie 1, nous avons choisi la Régression Linéaire ; dans la catégorie 2, nous avons sélectionné les arbres de décision (CART), Random Forest et Gradient Boosting, et dans la catégorie 3, nous avons mobilisé des réseaux de neurones dense (multiple perceptrons) à trois couches cachées.

Choix de la métrique : Lequertier et al (2021) constatent que la plupart des articles sélectionnés pour la revue de littérature ne justifient pas le choix de la métrique. Nous avons déjà souligné auparavant que dans le cas de classifications (binaires) pour des données déséquilibrées, une des meilleures options est d'utiliser la mesure de l'AUC (ROC et/ou PRC) puisqu'elles ne dépendent pas des seuils de classification.

Traitements des outliers : nous sommes bien conscients des biais que peuvent engendrer l'exclusion des outliers (Lequertier et al., 2021) – surtout quand justement ce sont ces observations atypiques qui sont les plus largement associées aux problèmes de non-qualité comme les séjours prolongés ou la réhospitalisation. Notre choix de dichotomiser la variable d'intérêt au seuil de Tukey tient explicitement compte de ce problème et représente donc une des solutions possibles.

Partitionnement des échantillons test, apprentissage et validation : (Lequertier et al., 2021) recommandent d'utiliser la k-validation croisée pour estimer la performance des modèles afin d'éviter notamment les aléas excessivement pessimistes ou optimistes du partitionnement en échantillon test vs échantillon d'apprentissage. Nous partageons effectivement leur prudence sur cet aléa de partitionnement et que le rééchantillonnage réalisé par k validation croisée permet de lisser le phénomène.

Malgré tout, si cette méthode permet de comparer les modèles entre eux, elle reste problématique pour obtenir une estimation fiable de la performance des modèles. En effet tant que l'on utilise des données que le modèle « connait » déjà, on ne peut pas être certain d'éviter le phénomène d'overfitting. Or c'est bien ce qui se passe dans la k validation croisée (Powell et al., 2020) sans utilisation d'un échantillon test brut, mis de côté avant toute analyse. Aussi est-il recommandé d'utiliser un échantillon de test [à n'utiliser qu'à la toute fin de l'analyse pour évaluer la performance sur des données inconnues du modèle], un échantillon d'apprentissage [sur lequel on applique une k-crossvalidation pour sélectionner les meilleurs paramètres], et un échantillon de validation [pour tester la performance du meilleur modèle sur des données qui n'ont pas été utilisées pour l'entraîner].

Malgré tout, on ne peut exclure là aussi que la performance obtenue avec l'échantillon de validation ne soit pas dépendante de cette partition spécifique (échantillon d'apprentissage, échantillon de validation). Il est donc recommandé de pousser le rééchantillonage plus loin, en rééchantillonnant de multiples fois la partition (échantillon d'apprentissage, échantillon de validation). Les performances obtenues après ces deux niveaux de rééchantillonnage (k cross validation et échantillon apprentissage + validation) donnent des résultats bien plus fiables. De plus, le principe de conserver un échantillon test dont le modèle n'a jamais eu connaissance reste indispensable, même si celui-ci contient toutes sortes d'aléas – ce qui serait de toutes façons le cas avec des données réelles.

Utilisation des données publiquement disponibles pour pouvoir mieux comparer les modèles et leurs performances (Lequertier et al., 2021) . Cette exigence ne peut hélas pas être satisfaite avec les données de l'APHM sans passer préalablement par un rigoureux processus d'anonymisation des données et l'autorisation des autorités compétentes. Elle constitue néanmoins une piste d'amélioration que nous prendrons en considération dans la suite des travaux.

Implications en Santé Publique : Plus les modèles de prédictions du LOS sont performants, plus ils seront fiables pour en déterminer les facteurs de risques ou les déterminants permettant de mieux gérer les ressources des services hospitaliers, d'améliorer leur efficacité à travers la gestion des lits, la programmation des décharges, et de façon générale la qualité des soins.

Les prédicteurs les plus important des séjours prolongés (LOS long) sont [coefficients ou sens de la relation entre crochets] :

- la destination du patient post-décharge vers d'autres établissements hospitaliers (plutôt que le domicile) [+++++],
- des séjours de Chimio et de Radiothérapie [---],
- une hospitalisation de type Chirurgie [++],
- une hospitalisation provenant d'autres établissements hospitaliers (plutôt que le domicile)[+],
- un séjour de nature orthopédique ou traumatique [-]

Autrement dit, une transition vers ou en provenance d'un lieu autre que le domicile est associé à une plus forte probabilité d'un séjour prolongé, de même qu'une hospitalisation en chirurgie, alors qu'un séjour pour radio ou chimiothérapie, en service orthopédique ou traumatique est associé à une probabilité plus faible d'être prolongé.

Ce sont là des informations utiles dans la programmation des lits comme dans l'allocation des ressources hospitalières en général. Concernant les destinations post hospitalisation on peut porter une attention particulière aux patients :

- âgés de plus de 65 ans et davantage encore les 85+ ans
- admis via les services d'urgence
- provenant d'établissements autres que le domicile
- admis pour des traumatismes complexes ou multiples
- admis pour des pathologies du système nerveux ou cardiovasculaire

Déploiement du modèle dans le système d'information de l'hôpital : bien entendu, un des objectifs à termes est de déployer un modèle comme le GB dans le système d'information de l'hôpital afin qu'il puisse par exemple signaler les séjours à risques dès l'admission du patient, mais aussi – pourquoi pas – à chaque fois que ses conditions changent, ou quand il est transféré vers un autre service, y compris à la sortie de l'hospitalisation.

Article 3[19]

Predicting hospital readmission with machine learning using the MIMIC III discharge notes: the impact of vectorization

Contexte

Dans cet article, nous revisitions à nouveau la réadmission à 30 jours mais en y apportant quelques changements majeurs :

- Les données sont cette fois issues du MIMIC III que nous avons présenté dans un chapitre précédent.
- Les variables se limitent aux notes cliniques contenues dans la table NOTEEVENTS.

L'objectif de l'article est d'explorer dans quelle mesure on peut utiliser les données textuelles pour prédire la réadmission à partir de modèles de machine learning classiques, et le rôle que peuvent jouer différentes façons de vectoriser les données textuelles. Sont volontairement exclus les modèles de deep learning utilisant des données séquentielles, notamment les réseaux de neurones de convolution (CNN) ou les réseaux de neurones récurrents (RNN), voire les réseaux de neurones avec attention comme les Transformers. L'idée est de rester dans une configuration de données tabulaires pour chercher à obtenir des informations beaucoup plus spécifiques à partir des données textuelles, quitte à sacrifier un peu de performance – en prolongement du compromis entre performance et explicabilité (développé ultérieurement). De fait, nous verrons que le niveau d'explicabilité apporté par ces modèles peut difficilement être atteint par des modèles séquentiels bien plus performants.

1- La vectorisation des données textuelles

Dans le jeu de données tabulaires dont chaque ligne représente un séjour, nous avons d'un côté la colonne RH30, variable binaire telle que :

$$y = 1 \; si \; le \; patient \; a \; été \; réadmis \; sous \; 30 \; jours, sinon \; y = 0$$

De l'autre, une colonne contenant toutes les notes cliniques du séjour correspondant. Pour des raisons pratiques et pour rester dans la continuité de précédentes recherches, nous ne conserverons parmi les notes cliniques que les notes de sortie ou de décharge (discharge notes).

La vectorisation consiste principalement à transformer chaque ligne de données textuelles en un vecteur ligne dont chaque élément représente un terme (token) du vocabulaire de l'ensemble du corpus (tous les lignes de textes). Il en résulte que la taille de notre tableau est :

- Nombre de lignes = nombre de séjours dans la période considérée (nombre de lignes du jeu de données MIMIC III)
- Nombre de colonnes = taille du vocabulaire + 1 (la colonne pour la variable d'intérêt RH30)

Or, typiquement, la taille d'un vocabulaire varie de 20k à 200k mots, donc pour ne pas se retrouver devant un tableau inutilement grand, exigeant des ressources considérables, il existe différentes

[19] Article soumis à la revue « PLOS ONE », le 18 Mai 2022 - PONE-D-22-14479

façons de prétraiter les données, et de réduire le nombre de colonnes (la dimension) du tableau dont voici quelques exemples non exhaustives :

- Eliminer les ponctuations.
- Eliminer les symboles inutiles tels que : *, /, \, #, [,] , (,), { , }.
- Eventuellement éliminer les chiffres.
- Passer tous les caractères en minuscules.
- Eliminer les mots vides ou les mots outils ($stopwords$) : ce sont typiquement les mots qui sont ignorés par les moteurs de recherche (prépositions, articles, pronoms, etc.).
- Lemmatiser le texte : il s'agit de ramener chaque terme à sa forme canonique. Par exemple on ramène les verbes à leur forme infinitive, les noms à leur forme au singulier, et le genre au masculin.
- Raciniser le texte : processus qui consiste à réduire les mots à leur radical (appelés également stemmes, bases ou racines), la partie du mot restante une fois que l'on a supprimé son préfixe et suffixe.

Dans les vectorisations conservant la forme tabulaire des données, ce pré-traitement des données est indispensable. Après cette phase on peut procéder à différentes formes de vectorisation dont les suivantes :

2- Bag of Words (BOW)

Il s'agit essentiellement de compter les occurrences de chaque terme (token) du vocabulaire dans chaque ligne de texte. Ainsi obtenons-nous un tableau dite DTM (document terms matrix) dont les lignes sont les documents prétraités de chaque texte clinique (discharge notes) et les colonnes sont les tokens du vocabulaire prétraité, et chaque cellule contient le nombre d'occurrence de chaque token par texte. Un tel tableau est généralement rempli de beaucoup de 0 avec ici et là quelques nombres non-nuls. On parle de parcimonie (sparsity).

Bien évidemment ce mode de vectorisation efface l'ordre des mots comme si tous les mots étaient mis dans un sac, et seul compte leur poids (Bag of Words : Sac de Mots), mais il en efface également le contexte.

Une façon de conserver le contexte est d'étendre le vocabulaire à des combinaisons de mots adjacents. Ainsi la phrase « j'aime le chocolat » après pré-traitement (lemmatisation) donne : {aimer, chocolat} peut être étendu et devient : {aimer, chocolat, aimer chocolat}. Ces éléments étendus sont appelés n-grams. Dans notre exemple, nous avons deux unigrams {aimer, chocolat} et un bigram {aimer chocolat}.

Enfin, le nombre d'occurrence du tableau DTM peut être pondéré de différentes façons dont (Gefen et al., 2017) :

- **Pondération binaire** : 0 quand l'occurrence est nulle, et 1 si elle est non-nulle
- **Pondération identité** : on conserve la fréquence d'occurrence comme telle - également appelé TF (Terms Frequency)
- **Pondération TfIdf** : on multiplie la fréquence par une quantité dite inverse de la fréquence du document (texte) (Terms Frequency, Inverse Document Frequency)

Les BOW, malgré tous ces pré-traitements des données donnent lieu à des vecteurs à grandes dimensions et avec peu d'informations. On peut donc envisager de leur appliquer des

transformations qui réduisent les vecteurs d'une dimension typique d'environ 10-20k de tokens à environ 100-300 tokens.

3- LSA : Latent Semantic Analysis

L'idée sous-jacente à l'analyse sémantique latente (LSA) est que les cooccurrences de tokens dans différents textes suggèrent que ces termes sont liés par un sens ou un concept latent commun. Cette idée reste valable même si les tokens ne coexistent pas dans le même document tant qu'ils coexistent avec d'autres termes communs.

À titre d'exemple, considérons les mots {chat, chien, souris, vert, bleu, rouge}. Si l'on devait comparer les cooccurrences de ces mots dans l'usage typique de la langue tels qu'ils apparaissent dans les articles de journaux, il est probable que chat, chien et souris coexisteraient fortement les uns avec les autres, et moins avec vert, bleu et rouge, tandis que vert, bleu et rouge coexisteraient fortement les uns avec les autres, mais moins avec chat, chien, et souris. (…) On pourrait alors projeter sa propre connaissance du monde pour supposer que "chat", "chien" et "souris" sont des animaux tandis que "vert", "bleu" et "rouge" pourraient être des couleurs ou des partis politiques. De plus, il est probable que les documents dans lesquels ces 6 mots apparaissent pourraient également être classés en deux groupes basés sur ceux où "chat", "chien" et "souris" coexistent principalement et ceux dans lesquels "vert", " bleu » et « rouge » coexistent pour la plupart. On pourrait supposer que le premier groupe de documents traite peut-être d'animaux, et le second groupe de documents de couleurs ou de politique.

Alternativement, considérons les termes "rouge", "rose", "rubis", "vin", "bourgogne" et "bordeaux". Ces six termes peuvent être considérés comme reflétant un concept latent partagé de "teintes de rouge", ce qui est le cas. Cependant, notons que « vin », « Bourgogne » et « Bordeaux » peuvent en outre avoir leur propre concept latent partagé (c'est-à-dire « vins ») et que « Bourgogne » et « Bordeaux » peuvent également coexister en tant que lieux en France. Comme ces exemples l'impliquent, l'analyse des cooccurrences de mots peut fournir des informations pertinentes, même si cela doit passer par un travail d'interprétation.

Examinons à présent les mots « rouge » et « merlot ». Ces mots apparaîtraient probablement ensemble fréquemment dans un corpus construit à partir d'une série de blogs sur le vin, et leur coexistence commune dans les documents pourrait alors être utilisée pour identifier qu'ils sont liés. Cependant, la simple recherche de mots qui apparaissent ensemble directement n'identifierait jamais "Cheval Blanc" et "Franzia" comme étant liés puisqu'ils n'apparaîtraient pas fréquemment ensemble dans le même document. La LSA, pourrait pourtant identifier que ces deux termes sont liés par leur fréquente cooccurrence partagée avec d'autres termes tels que « producteur » ou « merlot ». Ainsi, malgré le fait que Château Cheval Blanc produise le merlot le plus cher jamais vendu et que les frères Franzia emballent le vin dans des cartons, la LSA peut identifier qu'ils sont apparentés… (d'après Gefen et al., 2017, p. 452).

En pratique, la LSA est une technique de réduction de dimensions (de la taille du vocabulaire prétraité à un espace vectoriel à 100-300 dimensions). On procède en appliquant une transformation dite SVD (Singular Value Decomposition) sur la matrice DTM (ou plus couramment sa transposée, la matrice TDM). En sortie, après la SVD, on obtient trois matrices : (1) une matrice des termes [tokens], (2) une matrice des documents [textes] et (3) une matrice qui, multipliée par les deux autres, reconstruira la matrice TDM d'origine.

En termes plus familiers aux sciences sociales, la SVD est conceptuellement équivalente au PCA (principal component analysis), permettant d'identifier les facteurs sous-jacents à une matrice. SVD et PCA sont mathématiquement étroitement liés, sauf que là où la CPA engendre une matrice transformée, la SVD en crée deux. Dans la SVD comme dans la PCA, les facteurs sous-jacents sont supposés avoir une signification abstraite de niveau supérieur qui est commune à tous les éléments qui composent ce facteur (Landauer et al., 1998) .

Soit X la matrice TDM. Avec :

$$X = \begin{bmatrix} x_1^1 & \cdots x_i^j \cdots & x_1^p \\ \vdots & x_i^j & \vdots \\ x_n^1 & \cdots x_n^j \cdots & x_n^p \end{bmatrix}$$

Soit T_i le vecteur termes de la ligne i de X tel que :

$$T_i^t = \begin{bmatrix} x_i^1 & \cdots & x_i^p \end{bmatrix}$$

Soit D_j le vecteur documents de la colonne j de X tel que :

$$D_j = \begin{bmatrix} x_1^j \\ \vdots \\ x_n^j \end{bmatrix}$$

Alors :

$$T_i^t T_p = [XX^t]_i^p \; est\ la\ corrélation\ entre\ les\ termes\ de\ la\ ligne\ i\ et\ p$$

$$D_j^t D_q = [X^t X]_j^q \; est\ la\ corrélation\ entre\ les\ documents\ de\ la\ colonne\ j\ et\ q$$

$$Il\ existe\ deux\ matrices\ orthogonales\ U\ et\ V\ et\ une\ matrice\ diagonale\ \Sigma\ telles\ que$$

$$X = U\Sigma \mathrm{V}^t$$

$$Elles\ sont\ la\ décomposition\ en\ valeurs\ singulières\ (SVD)\ de\ X$$

La réduction de dimension se produit lorsque l'on ne retient que les k plus grandes valeurs de la matrice Σ et qu'on réduise les autres à zéro. On obtient ainsi la SVD de rang k de X, soit :

$$X_k = U_k \Sigma_k \mathrm{V}_k^t$$

Les vecteurs $U_k\ et\ V_k$ représentent l'espace sémantique du corpus initial capturé par X

4- LDA : Latent Dirichlet Allocation

LDA constitue une extension et amélioration de la LSA (Blei et al., 2003).

L'idée centrale derrière la LDA, est un processus génératif imaginaire qui suppose que les auteurs composent d documents en choisissant une distribution discrète de t sujets (topics) dans lesquels puiser, et en tirant w mots à partir d'une distribution discrète de mots typiques pour chaque sujet (voir Figure 2-1). En d'autres termes, une distribution de probabilité sur un ensemble fixe de sujets définit chaque document, et, à son tour, une distribution de probabilité sur un vocabulaire limité de mots définit chaque sujet. Alors que la LDA suppose que tous les documents sont générés à partir du même ensemble fixe de sujets, chaque document présente ces sujets dans des proportions différentes pouvant aller de 0 % (si un document ne traite pas du tout d'un sujet) à 100 % (si un

document traite exclusivement d'un sujet). L'algorithme LDA estime les distributions de sujets et de mots cachés en fonction des occurrences de mots observées par document. La LDA peut effectuer cette estimation via des approches d'échantillonnage (par exemple, l'échantillonnage de Gibbs) ou des approches d'optimisation (par exemple, Variational Bayes) (Debortoli et al., 2016).

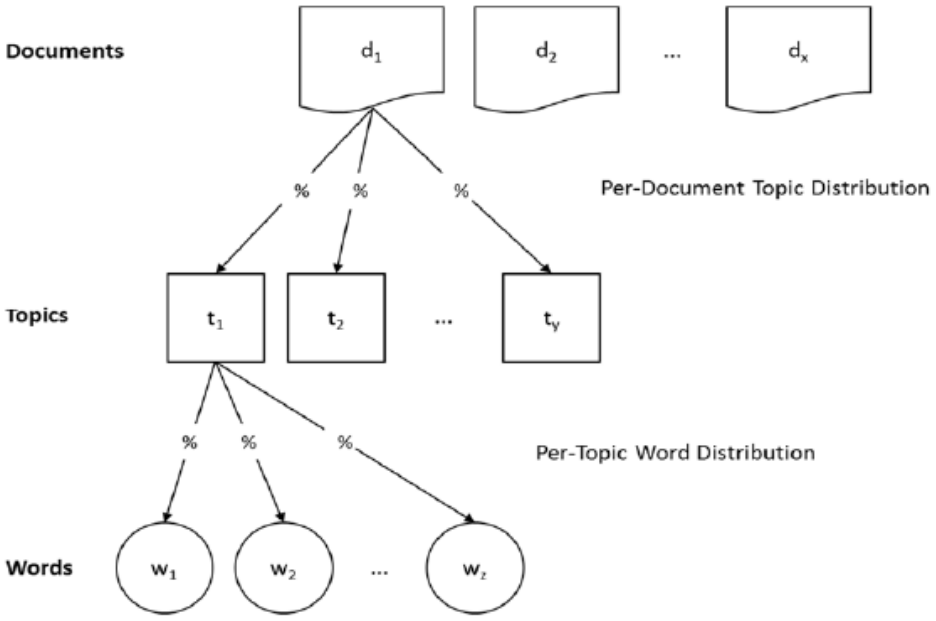


Figure 2-1 – Processus du LDA - d'après (Debortoli et al., 2016, p. 114)

La LDA fait partie d'un champ plus étendu dit modélisation probabiliste. Dans la modélisation probabiliste générative, on traite les données comme résultant d'un processus génératif qui inclut des variables cachées. Ce processus génératif définit une distribution de probabilité conjointe sur les variables aléatoires observées et les variables aléatoires cachées. On analyse les données en utilisant cette distribution conjointe pour calculer la distribution conditionnelle des variables cachées sachant les variables observées. Cette distribution conditionnelle est aussi appelée distribution a posteriori.

La LDA s'inscrit précisément dans ce cadre. Les variables observées sont les mots des documents ; les variables cachées sont la structure du sujet ; et le processus de génération est tel que décrit ci-dessus. Le problème consistant à déduire la structure de sujet caché (hidden topic) à partir des documents revient à celui de déterminer la distribution a posteriori, c-à-d la distribution conditionnelle des variables cachées sachant les documents (Blei, 2012).

De façon plus formelle, soient :

- $\beta_{1:K}$ les sujets, où chaque β_k est une distribution sur le vocabulaire.
- θ_d les proportions de sujets pour le $d$ième document, où $\theta_{d,k}$ est la proportion du sujet k dans le document d.
- z_d les affectations des sujets pour le $d$ième document, où $z_{d,n}$ est l'affectation à un sujet pour le $n$ième mot du document d.
- w_d les mots observés pour le document d, où $w_{d,n}$ est le $n$ième mot du document d, qui est un élément du vocabulaire fixe.

Avec cette notation, le processus génératif pour LDA correspond à la distribution conjointe suivante des variables cachées et observées :

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D},) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})$$

En pratique, la LDA peut également être utilisée comme méthode de réduction de dimensions, chaque dimension devenant potentiellement un sujet ou une thématique (topic).

5- Modèles de Machine Learning

Contrairement aux articles précédents, nous avons recours à une liste un peu différente de modèles de Machine Learning, notamment :

- Régression Logistique avec une pénalisation L1 (Lasso)
- Arbre de classification
- Random Forest
- Gradient Boosting
- LightGBM
- Catboost

Autrement dit, nous avons éliminé le réseau de neurones dense de la liste et inclus deux variations supplémentaires de Boosting, réputées plus performantes.

6- Critère d'inclusion

Sont incluses dans l'analyse toutes les réadmissions à 30 jours pour les adultes d'au moins 18 ans, excluant les décès.

7- En résumé :

Dans cet article, on cherche à prédire et à expliquer la réadmission sous 30 jours en comparant 5 types de vectorisation :

- 3 vectorisations en utilisant BOW avec les trois types de pondération (binaires, identité ou TF, TFIDF)
- Vectorisation avec réduction de dimension LSA
- Vectorisation avec réduction de dimension LDA

Résumé

Introduction. La réadmission à 30 jours est un indicateur utilisé partout dans le monde pour mesurer la qualité des soins et considéré dans certains pays comme un meilleur indicateur de la performance du système de santé que celle des établissements de santé (hôpitaux). Le but de cette étude est de prédire la réadmission à 30 jours à partir de données textuelles cliniques et ainsi de recueillir des informations plus précises et détaillées sur les patients à risque.

Méthodologie. Cette recherche s'appuie sur les notes de sortie clinique de la base de données publique MIMIC-III (Medical Information Mart for Intensive Care III). La cible de l'étude est la variable binaire réadmission à 30 jours parmi les admissions aux soins intensifs. Les modèles de Machine Learning utilisés sont la régression logistique pénalisée, les arbres de décision, la forêt aléatoire, le Gradient Boosting, LightGBM et Catboost, et sont comparés sur la base de cinq types de vectorisation. Les trois premiers sont Bag of Words pondérées respectivement par la fréquence des termes, par des coefficients binaires (0 ou 1), ou par TfIdf (Terms Frequency Inverse Document Frequency). Les deux autres sont des réductions de dimension sur base de LSA (Latent Semantic Analysis) et de LDA (Latent Dirichlet Allocation). La performance est évaluée avec l'aire sous la courbe ROC (ROC AUC).

Résultats. Notre analyse inclut 42825 admissions avec 2444 réadmissions à 30 jours (5,71%). Catboost est le modèle le plus performant (AUC = 0,723). L'importance des variables donnée par les modèles performants identifie des tokens (termes) et des thématiques liés à l'insuffisance (rénale, cardiaque, respiratoire), la trachéotomie, et les types et la fréquence des traitements comme principaux prédicteurs des réadmissions à 30 jours. Le profil des réadmis sont majoritairement ceux assurés par Medicaid et Medicare, célibataires et âgés.

Discussion. L'application de données textuelles aux modèles d'apprentissage automatique (ML) apporte des détails complémentaires cliniques plus spécifiques aux prédicteurs de la réadmission avec un niveau de performance acceptable. L'interprétabilité des modèles de ML est significativement améliorée par l'importance des variables et par la modélisation des thématiques (LDA). Le choix de la vectorisation est aussi important que le choix des modèles de ML dans ce processus.

Mots clés

Apprentissage automatique, réadmission à l'hôpital, sac de mots, modélisation de sujets, notes cliniques

Predicting hospital readmission with machine learning using the MIMIC III discharge notes: the impact of vectorization

Short title: Machine learning, hospital readmission and discharge notes

Franck Jaotombo[1,2,3]*¶, Luca Adorni[4]¶, Badih Ghattas[2&], Laurent Boyer[1,5&]

1. Aix-Marseille University, EA 3279 - Public Health, Chronic Diseases and Quality of Life Research Unit, La Timone Medical University, 27, boulevard Jean-Moulin, 13005 Marseille, France
2. Aix-Marseille University, I2M, CNRS, UMR 7373, Marseille, France.
3. EMLYON Business School, 23 avenue Guy de Collongue, 69130 Ecully, France
4. Bocconi University, Via Roberto Sarfatti, 25, 20100 Milano MI, Italy
5. Service d'Information Médicale, Public Health Department, La Conception Hospital, Assistance Publique - Hôpitaux de Marseille, 147 Boulevard Baille, 13005 Marseille, France

*Corresponding author

E-mail : jaotombo@em-lyon.com (FJ)

Abstract

Introduction. 30-day readmission is an indicator used all over the world to measure quality of care and considered in some countries to be a better indicator to the performance of the healthcare system than to the health institutions (hospitals). The goal of this study is to predict 30-day readmission from clinical text data and thus to gather more specific and detailed information on the patients at risk.

Methods. This research draws on the freely available database MIMIC-III (Medical Information Mart for Intensive Care III) clinical discharge notes. The study outcome is the binary 30-day readmission amongst ICU admissions. Penalized Logistic Regression, Decision Trees Classifier, Random Forest, Gradient Boosting, LightGBM and Catboost are used and compared based on five types of vectorizations. The three first are Bag of Words weighed respectively with Terms Frequency, Binary, and Terms Frequency Inverse Document Frequency. The two others are Latent Semantic Analysis and Latent Dirichlet Allocation. Performance is assessed with the Receiving Operator Characteristic Area Under the Curve (ROC AUC).

Results. Our analysis includes 42825 admissions with 2444 30-day readmissions (5.71%). Catboost is the highest performing model (AUC = 0.723). The feature importance of all the well performing models identify tokens and topics related to failure (renal, heart, respiratory), tracheostomy, treatment types and frequency as the main predictors of 30-day readmissions. The profile of the readmitted are mostly those insured under Medicaid and Medicare, single and old aged.

Discussion. Applying text data to Machine Learning (ML) models brings more specific clinical complementary details to the predictors of readmission with an acceptable level of performance. The interpretability of the ML models is enhanced by the features' importance and topic modelling. The choice of the vectorization is as important as the choice of the ML in this process.

Keywords

Machine Learning, Hospital Readmission, Bag of Words, Topic Modelling, Clinical Notes

1- Introduction

Given the economic weight of readmissions, the 30-day rehospitalization indicator (RH30) is widely used around the world to assess the quality of hospital discharges. It can be calculated for specific treatments (e.g. for heart failure) or be used globally for all-cause rehospitalizations.

A general international & historical background on the 30-day readmission indicator (RH30)

As early as the 1980s, in the USA, one can find studies on the rehospitalizations of people over the age of 65, covered by the public health insurance Medicare. The goal was to limit readmissions by identifying patients eligible for ambulatory or home care such as those suffering from certain chronic diseases like diabetes or chronic obstructive pulmonary disease (COPD) (Burgess & Hockenberry, 2014). RH30 was subsequently used to measure the potential impact of the shift to case-based reimbursement policy on the quality of care. However, the hypothesis according to which quality of care will decrease with the average length of stay has not been demonstrated in the United States (Burgess & Hockenberry, 2014; Epstein et al., 2011; Gilman et al., 2014) nor in France (Yilmaz & Vuagnat, 2015).

During the same period, more economics-oriented approach sought to quantify the volume of activity and health insurance expenses associated with rehospitalizations. These demonstrated how the expenses of patients with chronic pathologies - requiring multiple hospitalizations - weighed well above the others (Schroeder et al., 1979; Zook et al., 1980; Zook & Moore, 1980). In 1984, a large-scale retrospective study on Medicare beneficiaries (Anderson & Steinberg, 1984) showed that 22% of hospitalizations from 1974 to 1977 were followed by rehospitalization within 60 days of the patients' discharge, and nearly 60% of Medicare hospital expenditures were imputed to 12.5% of beneficiaries hospitalized at least three times within the study period.

Other studies followed in the USA, using 30 days after patients' discharge as a period to observe readmissions rate. It was estimated to roughly 20% for people aged 65+ years in 2004 (Jencks et al., 2009). By way of comparison, it was less than 15% in France based on 2010 data (Gusmano et al., 2015). Given their magnitude and the avoidable nature of some readmissions, the Medicare Payment Advisory Commission (MedPAC) recommended that the US Congress take measures (Medicare Payment Advisory Commission (U.S.), 2007).

Consequently, as part of the Patient Protection and Affordable Care Act (Obamacare), a more proactive approach to reducing rehospitalization rates has been implemented, combining support for the institutions, provision of appropriates tools and, where applicable, financial penalties (McIlvennan et al., 2015). Financial sanctions have since been triggered for hospitals with high rates of rehospitalization for number of identified pathologies. This list has since been expanded.

The implementation of programs to reduce rehospitalizations is part of a broader policy to improve coordination of care between the city and the hospital. Indeed, other measures enacted within this framework aim to provide incentives to the stakeholders. The goal is to promote coordination between players and thus achieve a triple objective: improving patient experience, enhancing the health of the population, and controlling expenses. The 30-day all-cause rehospitalization rate is one of the indicators to which these organizations commit.

Because the healthcare institutions are at the heart of the issue on rehospitalizations, they are the best to judge and decide the appropriateness of a patients' return home and their care by community doctors, accounting for their behavior and socio-economic profile (Burgess &

Hockenberry, 2014). This indicator would therefore be sensitive to the cooperation and coordination of the outpatient and inpatient sectors (Raleigh, 2014). It will help assess how some rehospitalizations may be the result of inadequate management of the continuity of care by health institutions, particularly on discharge from hospital, for example: scheduling discharge, providing information and educating the patient, and following-up the patient at home (Horwitz et al., 2011).

Several countries, such as the United States, England and Denmark, use the RH30 to measure the quality of care. Germany uses it to monitor the consequences of introducing an activity-based pricing (Kristensen et al., 2015). The Commonwealth Fund presents this indicator together with the rate of potentially avoidable hospitalizations as measures of the performance of the health systems of each American state (Radley et al., 2015). At an international level, in the interim report on the facts-based integrated health and patient-centered review, rehospitalizations is one of the two indirect indicators selected by the World Health Organization to assess the health care system, along with potentially avoidable hospitalizations (World Health Organization, 2015).

Given the current aging of the population and the increasing prevalence of chronic diseases and polypathologies in many countries, there is a need to develop indicators to support a more integrated health care provision (e.g. to report on their organization and operation in the territories). Some indicators of quality and performance are assessed at the level of an individual player (health establishment or city doctor), other indicators are assessed at the level of a territory or a group of actors. For example, there are indicators measuring rehospitalization readmission rates within 3 days after surgery and rehospitalization rates between 1 and 7 days. A clear distinction should be made between rehospitalizations between 1 and 7 days and over a longer period up to 30 days (Chin et al., 2016; Graham et al., 2015; Joynt & Jha, 2012). Indeed, while the rehospitalizations closest to the initial hospitalization seem to be linked to the quality of care during the stay, the more distant readmissions would be linked to the management of the primary health care and to the coordination between the city and hospital players at the level of a territory (Kangovi & Grande, 2011). Early readmission indicators, are facility-specific whereas rehospitalizations within a longer period (RH30), makes sense at the level of a territory

Thus, 30-day rehospitalization is a better indicator to the performance of the healthcare system than to the health institutions (hospitals) (Adeyemo & Radley, 2007).

Predicting 30-days readmission from clinical notes

There has been an increasing interest in predicting readmission using clinical notes over the last years (Craig et al., 2017; K. Huang et al., 2020; McCoy et al., 2015; Navathe et al., 2018; Rumshisky et al., 2016). This interest is motivated by gathering more specific and detailed information on the patients at risk. Unstructured text data provide richer insights on the patients since they describe symptoms, diagnosis, history, and other relevant clinical information, among others.

Hospital readmission is bad for the health of patients, lowers patients' quality of life, and is costly to the healthcare system. Cases of unavoidable readmission exist, but the variation of readmission rates across hospitals suggests that some cases are predictable and avoidable. A challenge in reducing readmission risk is the difficulty in identifying individuals at greatest risk, who might benefit from personalized interventions. Thus, text data may provide the missing information unavailable in traditional structured datasets.

Clinical text data are not usually readily available for confidentiality reasons; therefore the clinical notes of the Medical Information Mart for Intensive Care III (MIMIC III) has been largely used to explore different text-based machine learning models among which the prediction of readmissions

(K. Huang et al., 2020; Liu et al., 2019). Some of these models use classical Machine Learning models which require some prior feature engineering, whereas others use deep learning models with less feature engineering (Teo et al., 2020). Yet, while several studies have compared different machine learning models (Mahmoudi et al., 2020) in predicting hospital readmission, there is hardly – if any – study investigating the impact of text preprocessing and vectorization on the discriminating power of these models (ROC AUC based performance).

Text data must be transformed into a quantitative (numerical) vector before it can be processed by a Machine Learning model. This process, called vectorization can be accomplished through different means such as a bag of words (BOW) or an embedding. In this study we compare BOW weighing such as binary (BIN), terms frequency (TF), terms frequency inverse document frequency (TFIDF) and two latent methods for dimensional reduction - Latent Semantic Analysis (LSA) and Latent Dirichlet Analysis (LDA). These are used in different machine learning models: Penalized Logistic Regression (Lasso), Classification Trees (CT), Random Forest (RF), Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM) and Catboost (CB).

2- Materials and methods

Data

MIMIC-III is a large, freely available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (Boston, USA) between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside (approx. 1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital) (Johnson et al., 2016).

Inclusion criteria

The admissions table contains 58976 hospital admissions for 46520 patients. The clinical notes table contains over 2 million notes of 15 different types, but as per previous studies, we will keep only the discharge summaries (J. Huang et al., 2019). Only patients older than 18 years old with at least one stay in Intensive Care Unit (ICU) and a discharge summary are selected for this study. Patients who died during their first admission or later readmissions are also excluded.

Handling duplicates

The selected dataset contains duplicates of discharge summaries, as well as (subject id, admission id) duplicates. The text duplicates have been removed altogether. However, different texts associated to (subject id, admission id) duplicates have been merged, and duplicates removed.

Outcome: 30-days readmissions

Readmissions within 30 days are computed as the difference between the time of admission and the immediate time of discharge. We are considering only ICU admissions and account for all readmissions, including several subsequent readmissions for the same subject. All duplicates however are removed. The final count of readmissions is 2444 (5.7%) vs 40391 (94.3%) for a total of 42825 admissions and 33348 patients.

Predictors

The only predictor is a $Text$ variable which includes the discharge note for a given patient on a given admission stay. However, after preprocessing, this single column is transformed into a Document to

Matrix (DTM) table of up to 10 000 columns where the rows of the table represent each admission and the columns each term (or token). Very low frequency terms (5 of less) and very high frequency terms (in 80% or more of the admissions) are removed as they are considered irrelevant to the analysis: the former because they are too rare and the later because they are too common (not unlike stopwords – see hereafter).

Text cleaning

Text data from the discharge notes were cleaned according to the following steps. The recurring patterns containing the deidentified information for the patients (such as [** patient infos **]) were removed, using regular expressions, followed by other symbols or abbreviations such as Dr, M.D., firstname, lastname, *, ?, @, different types of brackets and other punctuations. Then regular and clinical stopwords were removed after lowercasing and lemmatizing. A stopword is defined as "any of a number of very commonly used words, as 'a', and, 'in', and 'to', that are normally excluded by computer search engines or when compiling a concordance" (Dictionary.com, 2022).

Each of the remaining terms is a unigram token. These terms are then extended to all contiguous sequences of 2 tokens, each of which becomes a bigram token. Our vocabulary will thus include unigrams and bigrams.

Text vectorization

Vectorization is the process through which the prepared free text is transformed into meaningful numbers. One way to do this is through Bag of Words (BOW) Vectorization. From the corpus of the text in the dataset, after cleaning and preprocessing, we may build a vocabulary of all the words (or tokens). A BOW may be pictured as a table (DTM = Document Terms Matrix) in which each test of a corpus is associated to a dictionary table containing each token of the corpus vocabulary, and their occurrence in the given text. Each occurrence frequency is then weighed accordingly. Here we have used three of the most used weighing:

- **Terms Frequency** (TF: $weight_{i,j} = frequency_{i,j}\ \ i.e.\ \ frequency\ of\ term\ i\ in\ document\ j$),
- **Terms Frequency Inverse Document Frequency** ($TFIDF$: $weight_{i,j} = frequency_{i,j} \times log_2 \frac{Document\ size}{frequency_i}$)
- **Binary Frequency** (BIN: $weight_{i,j} = 1\ if\ term\ i\ is\ in\ document\ j, 0\ otherwise$)

Another way to accomplish our goal is through word embeddings. It is a word representation where each token is encoded in a much smaller vector of real numbers (typically 50 to 300) compared to the size of the vocabulary (typically in tens of thousands), i.e. that of a full BOW representation. The mapping may draw on different approaches including neural-networks-based language modeling such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), on dimensional reduction of the word co-occurrence matrices (Lebret & Collobert, 2017), on probabilistic models (Globerson et al., 2007), or on explicit representations in terms of the context in which the word appears (Levy & Goldberg, 2014). Here we use two BOW-based embeddings through dimensional reduction: LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation).

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the meaning of words in their contextual usage (Landauer et al., 1998). The underlying idea behind latent semantic analysis (LSA) is that co-occurrences of terms (e.g., words) across many documents (e.g., book chapters or paragraphs) suggest that those terms are somehow related in that they are either

synonymous or reflect a shared latent concept. It applies a mathematical technique called Singular Value Decomposition (SVD) on the DTM or the TDM (Terms Document Matrix) to reduce the size of the vocabulary (Gefen et al., 2017).

Latent Dirichlet Allocation (LDA) is an extension of LSA. The core idea behind LDA (Blei et al., 2003), is an imaginary generative process that assumes that authors compose d documents by choosing a discrete distribution of t topics to write about and drawing w words from a discrete distribution of words that are typical for each topic. In other words, a probability distribution over a fixed set of topics defines each document, and, in turn, a probability distribution over a confined vocabulary of words defines each topic. The LDA algorithm computationally estimates the hidden topic and word distributions given the observed per-document word occurrences (Debortoli et al., 2016).

Statistical models

Five distinct types of predictive models were fitted to the data: L1 penalized logistic regression or Lasso (LR), Decision Tree Classifier (DTC), Random Forest (RF), and three variations of boosting model: Gradient Boosting (GB), LightGBM (LGBM), and Catboost (CB).

LR is a generalized linear model of the exponential family such that $\ln\left(\frac{\pi}{1-\pi}\right) = w^T x$, where $\pi = P(y = 1|x)$, and y is a binary outcome, x the predictors, and w the weight vector to be estimated from the data. However, in the case of the Lasso, the loss function includes a L1 penalty term $\lambda \sum|w_j|$ (Hastie et al., 2009) where λ is the regularization parameter that may be tuned through cross validation.

DTC (Breiman et al., 1984) is a binary decision tree method that involves segmenting the predictor space into a number of simple regions. DTC is constructed through an iterative process by applying a binary splitting rule. For each explanatory variable x_j in the data, a rule of the form $x_j < a$ (a ∈ R is a threshold) is used to split the initial set of observations (denoted t_0, the root of the tree) into two subsets t_l and t_r (the sibling nodes). Each observation falling in those regions is then predicted by the highest frequency class. The best split is defined as the one minimizing a loss function (e.g., the Gini index, or the Entropy). Once the best split has been defined, the same process is applied to the two nodes t_l and t_r and repeated until a predefined minimum number of observations is reached. Then, a pruning algorithm can be used to search for an optimal subtree, given a penalty criterion ($complexity\ parameter$) applied to the objective function. DTC can be represented graphically and thus can be directly interpretable, given its simple structure.

RF (Breiman, 2001) is an ensemble learning method based on aggregating $n\_estimators$ trees similar to the ones constructed with DTC, each one grown using a bootstrap sample of the original data set. Each tree in the forest uses only a random subset of $max\_features$ predictors to determine the best split at each node. The trees are not pruned. The prediction by RF is the majority vote over the predictions made by the $n\_estimators\ trees$. Other hyperparameters such as the minimum number of samples required to split an internal node ($min\_samples\_split$) or the maximum depth of a tree (max_depth) may be used to tune further the RF model.

GB (Friedman, 2001) is also an ensemble learning method based on a decision tree but does not involve bootstrap sampling. It is built sequentially using a weak learner (e.g., shallow classification trees). The GB is initialized with the best guess of the response (e.g., the majority vote), then the gradient is calculated, and a model is then fit to the residuals to minimize the loss function. The current model thus obtained is added to the previous model, adjusted by a $learning\_rate$ parameter. The user may specify the number of trees ($n\_estimators$), a tree depth equal to

max_depth and a given minimum number of observations in the trees' terminal nodes, $min\_samples\_leaf$.

LGBM is a speeding of GB up to 20 times by using two novel techniques, called Gradient-based One-Slide Sampling (GOSS) and exclusive Feature Bundling (EFB). For more details, refer to (Ke et al., 2017).

CB introduces two critical algorithmic advances to GB: ordered boosting, a permutation-driven alternative to the classic algorithm, and an innovative algorithm for processing categorical features. Both techniques were created to fight a prediction shift caused by a special kind of target leakage present in all currently existing implementations of gradient boosting algorithms. For more details, see (Prokhorenkova et al., 2018).

Statistical analyses

In the following, model performance is estimated through the area under the receiver operating characteristic curve (ROC AUC). Indeed, given that our outcome class proportions are quite unbalanced (94.29% non-readmitted vs 5.71% readmitted), threshold-dependent measures of performance such as the accuracy or the F1 are less reliable (J. Huang & Ling, 2005; Wardhani et al., 2019). ROC AUC values closer to 1 are best, and a value equal to 0.5 indicates a random classifier, i.e., which predictions are not better than chance. In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based on the test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding (Hosmer et al., 2013).

To train and evaluate the different models (i.e., LR, DTC, RF, GB, LGBM, CB), the dataset was split into 80% full training sample and 20% hold out test sample, stratified on the outcome variable. The first step was to train the models on the training sample with their default settings, then estimate performance on the holdout sample. The second step was to tune each of the different models through 5-fold cross validation on the training set, then select the model with optimal parameter to estimate performance on the holdout sample.

Lastly, we computed variable or feature importance (VI). VI provides a simple way to inspect each model and gain insights on which variables are most influential in predicting the outcome, and to what extent. VI is computed in two ways. First, we used the inbuilt feature importance for each classifier which is different from one classifier to another. This however forbids a fair comparison between classifiers, so we used permutation feature importance to estimate variable importance. Permutation feature importance is defined as the decrease in a model score (here ROC AUC) when a single feature value is randomly shuffled (Altmann et al., 2010b). The larger the decrease in score, the more important the variable.

All analyses were implemented in Python 3.8.8 (Van Rossum & Drake, 2009) with Sci Kit Learn 0.24.1 (Pedregosa et al., 2011).

3- Results

Before tuning: on the default parameters

Table 1 shows that CB is the best model overall and TFIDF is the corresponding best vectorization. Overall, the boosting models are performing better than the other models before tuning and on almost all vectorizations.

Table 2 shows the variable (inbuilt) importance for the best result: CB on TFIDF. The lemmatized most predictive token for Rehospitalization is <time> which is here an indicator of timing as well as of frequency (when browsing through the discharge notes). This token is used as a reference unit of importance. We will address only those features within 20% of importance as compared to it.

Table 1. ROC AUC Performance on the test sample, on default parameters

	Vectorization				
Models	TF	BIN	TFIDF	LSA	LDA
LR	0,652	0,595	0,716	0,650	0,716
DTC	0,538	0,533	0,531	0,529	0,520
RF	0,664	0,682	0,658	0,644	0,685
GB	0,706	0,713	0,707	0,684	0,699
LGBM	0,701	0,686	0,702	0,673	0,701
CB	0,708	0,714	0,723	0,682	0,716

Next most important feature is <trach> which is mostly related to track collars and trach masks which are medical devices installed after a tracheostomy. Indeed, the token <tracheostomy> is one of the top 5 most important features predicting readmission. The next token is the bigram <tablet po> which indicates medication by mouth intake (see Medical Abbreviation, in Appendix 1). This bigram is related to another important bigram feature <po tid> related to medication frequency, and the token <tablet> indicative of medication. Then comes the token <failure> which in the original text is related to heart failure, renal failure, and respiratory failure. <hospice> is another token, indicator of the hospital facility. <recently> is another important feature mostly related to a severe medical condition as is the token <multiple>.

The other less important features do, however, provide further details into the type of medical information conducive to rehospitalizations, such as <chronic> indicating persistency of some conditions, <dialysis>, <transplant>, <dnr> (do not resuscitate), <palliative>, or <esdr> (end-stage-renal disease). They suggest either serious conditions or old-aged patients as predictive of readmissions.

Table 2. Variable (inbuilt) importance for Catboost, TFIDF Vectorization (30 most important features)

	Features	Importance		Features	Importance
1	time	100,00	16	dialysis	16,75
2	trach	42,79	17	qid	15,59
3	tablet po	35,26	18	transplant	15,24
4	failure	30,02	19	dnr	14,85
5	tracheostomy	28,59	20	barbitrt neg	14,64
6	po tid	27,32	21	palliative	14,63
7	hospice	27,10	22	weight lb	14,54
8	am	25,54	23	disp tablet	14,27
9	tablet	24,78	24	esrd	14,22
10	tid	23,85	25	cpap	12,98
11	recently	22,62	26	picc	12,46
12	multiple	21,92	27	transfer hospital	12,13
13	chronic	19,29	28	motor	12,05
14	sig	18,62	29	subdural	11,81
15	po daily	17,10	30	process	11,71

Tuned models:

Table 3 displays results after tuning. We can see a slight improvement of the binary vectorization for the Logistic Regression and an overall improvement for Random Forest on all vectorizations. None of our tuning attempts have been able to bring about any improvement in the performance of the other classifiers. This shows the extent to which the boosting models are already pre-tuned quite well. As we can see here, Catboost on binary vectorization remains the best model by a very small margin.

Table 3. ROC AUC Performance on the test sample, on tuned parameters

	Vectorization				
Models	TF	BIN	TFIDF	LSA	LDA
LR	0,652	0,717	0,716	0,625	0,717
DTC	0,419	0,534	0,598	0,534	0,603
RF	0,673	0,698	0,700	0,672	0,705
GB	0,691	0,445	0,638	0,632	0,699
LGBM	0,679	0,645	0,684	0,693	0,663
CB	0,710	0,719	0,682	0.623	0.698

Table 4 shows the results of the inbuilt feature importance of CB in a binary vectorization format. Even after tuning, the most important variables are basically the same but in different orders. For example, the token <time> is now closely matched by <tid>. This reinforces the idea that frequency of medication intake is an important predictor of readmission. <hospice> and <chronic> have also moved to higher ranks which strengthens the idea that hospital facility, persistence of a medical condition and old age are another important predictor of readmission. Severe medical conditions as related to <multiple> remains among the highest 20% most important features compared to <time> as are <trach> and <tracheostomy>. There are also some remarkable new tokens: <crash> associated

to car accidents, <sulfate> referring to discharge medication as is <load dilantin>. Another fairly important token is <am> which in context is very often associated to blood tests.

Table 4. Variable (inbuilt) importance for Catboost, BIN Vectorization (30 most important features)

	Features	Importance		Features	Importance
1	time	100,00	16	dnr	16,95
2	tid	82,95	17	recently	15,91
3	hospice	50,74	18	suicide	15,25
4	chronic	45,66	19	po tid	14,86
5	trach	38,35	20	gram stain	13,98
6	tracheostomy	34,97	21	transfer hospital	13,58
7	palliative	31,34	22	qid	12,91
8	po daily	28,77	23	picc	12,81
9	crash	24,20	24	myalgias past	12,78
10	am	23,74	25	am blood	12,73
11	sulfate	22,27	26	tachypnea	12,40
12	load dilantin	21,37	27	respiratory failure	12,25
13	multiple	20,56	28	transplant	12,16
14	cpap	19,49	29	sputum	11,76
15	tablet po	19,45	30	esrd	11,04

For the other tokens please see our medical abbreviation list (Appendix 1)

Comparing variable importance through permutation importance method

As mentioned earlier, permutation importance allows a direct comparison between classifiers. Here we selected only the two best models in each vectorization for comparison, and we confine our analysis to the features within 20% of the most important feature. For these classifiers and in all types of vectorizations, the common most important tokens are <time>, <trach>, <tracheostomy>, then comes <tid>, <failure> and <am>. This gives consistency and confirmation to the previous observations (see Appendix 2).

Topic modeling

Table 3 shows that LR fares the best with LDA with a respectable ROC AUC = 0.717. Topic modeling has thus reduced the size of our data table from 10000 to 300 columns without a major loss in performance. But what about its interpretability?

We selected the 10 most important topics of LDA, based on LR coefficients and examined the 20 topics with the highest values in the topics-terms matrix. Each of these 10 topics are indeed clearly expressing some medical conditions or some treatments, most of them related to the most important tokens from previous paragraphs. For example, Topic 0 is mostly related to renal disease, Topic 1 to tracheostomy, Topic 3 to respiratory failure, Topic 4: heart failure, Topic 5 : treatments of infections, Topic 6 : liver disease, Topic 7 : gastric treatments, Topic 8 : frequency of treatments and Topic 9 : blood treatments (see Appendix 0).

4- Discussion

Hospital readmission is bad for the health of patients, lowers patients' quality of life, and is costly to the healthcare system. Cases of unavoidable readmission exist, but the variation of readmission rates

across hospitals suggests that some cases are predictable and avoidable. A challenge in reducing readmission risk is the difficulty in identifying individuals at greatest risk, who might benefit from personalized interventions. Thus, text data may provide the missing information unavailable in traditional structured datasets.

A recent work (Orangi-Fard et al., 2022) also uses the MIMIC III discharge notes to predict ICU readmission from discharge summaries. The researchers used SVB-RBF, AdaBoost, QDA, LASSO and RIDGE Regression to explore readmission, also using Bag of Words Vectorization (BOW) with TF weights. Using all the features, their best performance equals ROC AUC = 0.71, and 0.74 after a special feature selection. Unfortunately, their results remain a black box as they do not provide any information on feature importance. Another recent work (K. Huang et al., 2020) uses the MIMIC III discharge notes on the most powerful deep learning-based Natural Language Processing (NLP), trained on clinical data (Clinical Bert). Their performance does not exceed a ROC AUC of 0.680 on their best model. However, this model offers some means to interpret the results. Another study using different clinical (doctor) notes (Craig et al., 2017) on convolutional neural networks (CNN) does not exceed a ROC AUC Performance of 0.70 but does provide some insights into the relationship between the features and the outcome.

In this sample of 42825 admissions, we are extensively exploring different BOW vectorization. Three of them use a vocabulary of 10000 unigram and bigram tokens weighed with TF, BIN and TFIDF and two others use reduced dimensions of 300 topics through LSA and LDA topic modeling. These 5 types of BOW vectorizations are then used to compare 6 models of ML (LR, DTC, FR, GB, LGBM and CB). Our best performance is reached by CB with the default parameters on TFIDF weight for a ROC AUC = 0.723. After tuning, the best performance is again on the tuned CB but on BIN weight for a ROC AUC = 0.719. Tuned LR also fares well both on BIN weight and on LDA.

One interesting contribution of this paper is the demonstration that dimension reductions such as topic modelling (from 10000 unigrams and bigrams to 300 topics in our case), not only may preserve (or even enhance) performance, but it also brings a rich way to interpret the conditions of the patients that is complementary to structured datasets as well as to regular BOW vectorizations.

It is thus clear to us that optimizing a ML model using text data - all preprocessing being equal - will require exploring not only different models but also several vectorizations. Furthermore, this vectorization paves the way to open the black box and provides interpretation to the results, through feature importance.

In our case, using our best Machine Learning model (CB) to predict 30-days readmission from discharge notes, we are able to provide further information on the individuals at risk: they tend to require specific timing and frequent attention such as for medication, they are likely to be admitted in a hospice facility and palliative care, and they are mostly likely going to be the sick elderly. Chronic conditions, tracheostomy and different types of failures (heart, respiratory and renal) are also predictive of readmission as do crash accidents, organ transplants, suicide and certain medications such as dilantin loading. LDA's emerging topics not only confirm but also enhance the information provided by the most important tokens. Indeed, the main topics predicting rehospitalization include those already identified by the BOW tokens but extends to others e.g topics on liver disease, and different treatments such as those for infectious diseases, blood clotting, and digestive.

These findings suggest that ML (especially CB) maybe a good means to improve quality of care by identifying those patients at risks, through their medical conditions, their types of treatments and medications, their type of stay, their hospital facility or their age based on their clinical discharge

notes, in way that is not easily identifiable through the information in the structured dataset. As RH30 is linked to the management of the primary health care and to the coordination between the city and hospital players at the level of a territory (Kangovi & Grande, 2011), it may be relevant to pay a special attention to the management of the continuity of care of the above-referred patients and their conditions, especially on discharge from the hospital.

In this study, the readmissions rate is significantly higher for the patients who are single (marital status), and who are insured under Medicaid or Medicare (Appendix 3). Therefore, the above-described types of conditions are likely to relate to them. It must be noted that the age category of the readmitted are not significantly different from the non-readmitted patients, thus structured data alone is not able to provide these types of information that we gathered on the conditions of the (old) patients through text mining.

This study is not without limitations. The discharge notes represent only 1 among 15 types of clinical notes in the MIMIC III database. We would expect much better performance and more detailed clinical information were we to include more types. The reliability of our results would also be enhanced if we used resampling and validations samples in the tuning and scoring of our models, instead of a single split test and training set with cross validation. Such an approach however would be very time consuming. Future study will explore deep learning models and a comparison between structured, unstructured and mixed datasets.

Appendix 0 – Latent Dirichlet Analysis Topics (20 highest coefficient tokens for the 10 most important features)

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4	
Renal disease		Tracheostomy		Respiratory Failure		Tracheostomy & Respiratory Failure		Heart Failure	
hd	10081,98	tube	16922,64	bipap	2473,87	trach	2212,28	daily	3538,72
dialysis	6194,96	feed	6090,23	sleep	1661,05	tablet	1926,92	heart	2178,97
esrd	3847,34	tube feed	5156,87	respiratory	1396,87	peg	1667,25	tablet	2103,55
hemodialysis	3247,73	peg	2032,10	pco	1335,20	sig	1504,95	heart failure	1935,69
tablet	2601,31	placement	2005,30	cpap	1160,15	tracheostomy	1398,04	failure	1923,26
renal	2215,95	place	1908,21	abg	1107,58	tube	1370,15	chf	1871,71
tid	1812,79	feeding	1702,97	ph	1090,73	respiratory	1289,63	lasix	1702,41
catheter	1770,05	tube place	1683,11	hypercarbic	1025,71	rehab	1138,17	weight	1665,43
daily	1733,05	gastrostomy	1618,60	failure	893,07	daily	1081,26	carvedilol	1608,13
sig	1679,56	percutaneous	1529,86	osa	889,30	place	1048,86	cardiomyopathy	1580,59
esrd hd	1672,30	continue	1430,09	respiratory failure	874,18	vent	1037,72	icd	1276,66
capsule	1474,12	wean	1394,30	hypercarbic respiratory	820,52	prn	981,98	chronic	1135,20
meal	1245,30	tube placement	1385,37	apnea	765,05	tablet po	960,91	ef	1098,73
acid	1241,02	tracheostomy	1237,41	base	742,68	wean	938,49	digoxin	1056,32
folic	1237,07	cc	1172,45	sleep apnea	636,99	tid	932,99	increase	1044,07
folic acid	1235,61	ventilator	1067,36	hypercarbia	589,12	failure	906,67	po daily	1035,26
complex	1221,48	gastrostomy tube	1044,00	po pco	582,37	placement	766,11	continue	1025,52
po tid	1214,59	peg tube	1024,50	pco ph	551,92	respiratory failure	743,06	spironolactone	1017,01
continue	1156,52	status	1008,90	sleep study	497,83	bid	742,56	chronic systolic	953,43
sevelamer	1153,61	tolerate	1008,32	hypoventilation	495,91	ventilator	732,44	home	942,06

Topic 5		Topic 6		Topic 7		Topic 8		Topic 9	
Treatment (Infections)		Liver disease		Treatment (Gastric)		Treatment (Frequency)		Treatment (Blood)	
sig	6069,25	ascite	4282,70	tablet	11474,28	tablet	9539,45	flush	4990,76
tablet	3161,72	cirrhosis	3992,64	tablet po	3869,75	sig	7221,05	prn	3706,78
vancomycin	2192,55	liver	3575,53	dissolve	3587,39	sig tablet	4330,01	heparin	3264,81
diff	2150,16	lactulose	2578,18	rapid dissolve	3471,27	tablet po	4028,83	iv	3041,76
appl	2098,49	encephalopathy	1820,21	tablet rapid	3450,77	tablet sig	3708,22	picc	2007,05
topical	2019,08	hepatic	1791,84	rapid	3355,07	capsule	3032,04	iv prn	1768,38
appl topical	2002,08	paracentesis	1778,51	daily	2706,02	release	2848,28	saline	1533,47
sig appl	1866,76	fluid	1514,71	tablet hospital	2631,04	prn	2261,55	prn flush	1302,22
daily	1689,47	portal	1322,99	hospital tablet	2532,29	qhs	2060,32	flush saline	1279,97
sig tablet	1515,19	daily	1233,32	tablet tablet	2412,05	delay release	1773,13	flush iv	1259,73
tablet sig	1451,37	sbp	1141,98	po daily	1737,59	delay	1718,15	daily	1098,19
continue	1415,83	albumin	1135,68	tablet day	1524,18	hr	1711,60	daily prn	1032,09
rehab	1342,14	abdominal	1123,71	day tablet	1425,01	tid	1699,10	luman	932,93
soln	1308,21	hepatic encephalopathy	993,16	po bid	1330,21	daily	1631,78	heparin flush	901,77
recon soln	1307,21	liver disease	991,23	bid	1311,44	sig capsule	1393,78	dependent	876,52
flagyl	1292,33	spironolactone	945,06	lansoprazole	1197,54	qam	1342,95	heparin dependent	855,15
recon	1277,72	continue	935,28	dissolve tablet	1101,82	po daily	1340,18	sig	756,82
tablet po	1260,02	rifaximin	902,97	lansoprazole tablet	1031,00	continue	1324,44	prn luman	742,00
colitis	1241,92	negative	828,20	dissolve po	796,27	release sig	1302,80	sodium chloride	736,72
miconazole	1122,02	start	798,58	dissolve daily	791,00	iv	1260,03	flush picc	732,74

Appendix 1 – Medical abbreviations

aicd	automatic implantable cardiac defibrillator
alk	alkaline
am	am blood tests
ama	American Medical Association; antimitochondrial antibodies; against medical advice
asd	atrial septal defect; autism spectrum disorders
ast	aspartate aminotransferase (formerly SGOT); aspartate transaminase, astigmatism
bid	twice a day (bis in die)
bipap	bilevel positive airway pressure
bp	base pair; blood pressure; bullous pemphigoid
ccu	critical care unit; coronary/cardiac care unit
cn	tomorrow night (cras nocte), child nutrition; cognitively normal; cranial nerve
cpap	continuous positive airway pressure
ct	chest tube; clinical trial; cognitive therapy; computed/computerized tomography
disp	dispense
dka	diabetic ketoacidosis
dm	dermatomyositis; dextromethorphan; diabetes mellitus
dnr	do not resiscitate
dvt	deep venous thrombosis
dz	disease
ed	emergency department; erectile dysfunction; effective dose
esrd	end-stage renal disease
gi	gastrointestinal; glycemic index
hct	hematocrit
hd	health department; hearing distance; hemodialysis; herniated disk; Hodgkin's disease; Huntington's disease
hr	hour, heart rate
icd	International Classification of Diseases; implantable cardioverter defibrillator; intrauterine contraceptive device; ischemic cardiac disease
inr	international normalized ratio
iv	intravenous
lb	live birth / lower body / pound
lb	pound
le	left eye; lower extremity; lupus erythematosus
mcg	microgram
min	minute
mmhg	millimeter of mercury
mr	medical record / mental retardation / mitral regurgitation / magnetic resonance
neg	negative
nh	nursing home
pcp	primary care physician; pneumocystis carinii pneumonia; phencyclidine (anesthetic/hallucinogenic); primary care provider
peg	percutaneous endoscopic gastrostomy; pneumoencephalogram; polyethylene glycol
ph	public health; parathyroid hormone; past history; poor health, hydrogen ion concentration (measure of acidity/alkalinity)
picc	peripherally inserted central catheter
po	by mouth / per os

ppi proton pump inhibitor: patient package insert
ptt partial thromboplastin time
qd every day (quaque die), once daily
qid 4 times a day
rdw red cell distribution width
rle right lower extremity
rr recovery room; relative risk; respiratory rate
sig surface immunoglobulin / write on label
staph staphylococcal infections
stent Subclass of:Prostheses and Implants
th thyroid hormone
tid 3 times a day / ter in die
trach tracheotomy; trachea(l); tracheostomy

Appendix 2 – Permutation importance of the two best classifiers per vectorization (30 top features)

	Terms Frequency (TF) Vectorization			
Models	Default GB - AUC = 0,706		Tuned CB - AUC = 0,710	
Rank	features	importance	features	importance
1	time	100,00	time	100,00
2	trach	49,65	trach	47,12
3	tracheostomy	24,01	tracheostomy	28,95
4	failure	23,88	esrd	8,48
5	tid	21,09	dialysis	6,97
6	chronic	17,01	am	5,94
7	transplant	13,93	failure	4,91
8	multiple	11,40	multiple	4,29
9	dialysis	9,27	chronic	3,99
10	peg	5,35	transplant	3,41
11	epoetin alfa	5,10	esrd hd	3,00
12	dka	4,96	gastroparesis	2,88
13	nh	4,83	end stage	2,85
14	hospice	4,80	slide	2,75
15	esrd	4,37	hemodialysis	2,67
16	sulfate	4,27	feed	2,53
17	varix	4,11	albuterol	2,47
18	undergo	4,08	end	2,29
19	gastroparesis	4,03	intermittently	2,27
20	end	3,79	cpap	2,27
21	physiology	3,65	prednisone tablet	2,16
22	esrd hd	3,65	peg	2,09
23	cpap	3,36	overload	2,02
24	weight lb	3,12	daily	1,66
25	tachypnea	3,07	nephrocap	1,59
26	feed	2,99	chf	1,41
27	recently	2,92	spironolactone	1,37
28	intermittently	2,87	gap	1,37
29	codeine	2,82	strength	1,33
30	albuterol	2,74	negative staph	1,26

	Binary (BIN) Vectorization			
Models	Tuned LR - AUC = 0,717		Tuned CB - AUC = 0,719	
Rank	features	importance	features	importance
1	time	100,00	time	100,00
2	trach	52,90	trach	87,99
3	multiple	38,00	tracheostomy	35,55
4	tracheostomy	37,08	tid	32,36
5	failure	21,23	chronic	24,23
6	am	18,69	subdural	19,06
7	chronic	18,23	am	18,21
8	loss	18,02	recently	15,73
9	weigh	17,58	palliative	15,43
10	building	17,51	building	15,12
11	end	17,49	hospice	14,18
12	po daily	16,79	multiple	13,33
13	tid	15,33	subdural hematoma	13,01
14	subdural	15,07	tachypnea	12,99
15	hospitalization	14,88	suicide	12,51
16	albuterol	14,15	heart failure	12,17
17	am blood	13,67	hemodialysis	12,09
18	dialysis	13,35	esrd	10,63
19	amiodarone	13,29	sputum	9,51
20	prednisone	13,28	hospital aged	8,96
21	dnr	13,06	dialysis	8,13
22	transplant	11,64	show growth	8,02
23	slightly	10,84	prednisone tablet	7,98
24	pain control	10,62	amiodarone	7,89
25	feed	10,37	hypercarbic respiratory	7,85
26	mr	9,73	transplant	7,84
27	ptt inr	8,90	size left	7,44
28	leave	8,84	feed	7,21
29	hospice	8,68	cpap	7,14
30	cc	8,35	gastroparesis	6,98

	Terms Frequency Inverse Document Frequency (TFIDF) Vectorization			
Models	Tuned LR - AUC = 0,716		Default CB - AUC = 0,723	
Rank	features	importance	features	importance
1	subdural	100,00	time	100,00
2	hospice	85,11	trach	80,38
3	tracheostomy	68,41	tracheostomy	32,19
4	time	67,63	am	21,40
5	trach	65,53	failure	18,00
6	tamponade	47,01	issue continue	17,13
7	am blood	44,93	multiple	16,81
8	dka	42,06	chronic	14,08
9	peg	40,10	hospice	13,17
10	tid	32,05	tamponade physiology	12,40
11	postoperative	31,08	subdural	11,64
12	am	28,76	dialysis	11,55
13	hemodialysis	26,04	loss	10,10
14	amiodarone	25,43	transplant	9,58
15	clonidine	25,32	saline	9,25
16	lactulose	22,14	palliative	8,16
17	cancer	20,69	drug	8,14
18	loss	19,46	end	7,71
19	overdose	19,32	displacement	7,42
20	hypercarbic respiratory	18,41	esrd	7,07
21	hd	18,35	albuterol	6,03
22	patch	17,73	hospitalization	5,99
23	withdrawal	17,36	spironolactone	5,57
24	esrd	16,93	aspirin	5,47
25	neg	16,60	colonoscopy	5,29
26	sleep	16,23	similar	5,18
27	icd	16,22	lung disease	5,13
28	multiple	15,82	floor	5,06
29	similar	15,61	amiodarone	5,02
30	bipap	15,56	slightly	4,90

Appendix 3 - Readmission profiles

Variable	Modality	N	(%)	NoReadm-	(%-)	Readm+	(%+)	modality p-value	Cohen's d
agecat	2(18-44 years old)	6387	14,91%	6035	14,95%	352	14,40%	0,465	-0,015
	3(45-64 years old)	15074	35,20%	14237	35,26%	837	34,25%	0,310	-0,021
	4(65-84 years old)	17283	40,36%	16255	40,25%	1028	42,06%	0,077	0,037
	5(85+ years old)	4081	9,53%	3854	9,54%	227	9,29%	0,676	-0,009
gender	F	18655	43,56%	17597	43,58%	1058	43,29%	0,781	-0,006
	M	24170	56,44%	22784	56,42%	1386	56,71%	0,781	0,006
marital_status	DIVORCED	2753	6,43%	2579	6,65%	174	7,20%	0,295	0,022
	LIFE PARTNER	15	0,04%	15	0,04%	0	0,00%	0,333	-0,028
	MARRIED	20594	48,09%	19468	50,21%	1126	46,59%	0,001	-0,073
	SEPARATED	487	1,14%	446	1,15%	41	1,70%	0,016	0,046
	SINGLE	11141	26,02%	10407	26,84%	734	30,37%	0,000	0,078
	UNKNOWN (DEFAULT)	268	0,63%	256	0,66%	12	0,50%	0,331	-0,022
	WIDOWED	5934	13,86%	5604	14,45%	330	13,65%	0,278	-0,023
insurance	Government	1231	2,87%	1189	2,94%	42	1,72%	0,000	-0,081
	Medicaid	3904	9,12%	3610	8,94%	294	12,03%	0,000	0,101
	Medicare	23322	54,46%	21808	54,01%	1514	61,95%	0,000	0,161
	Private	13951	32,58%	13366	33,10%	585	23,94%	0,000	-0,204
	Self Pay	417	0,97%	408	1,01%	9	0,37%	0,002	-0,078
	TOTAL	42825	100,00%	40381	94,29%	2444	5,71%		

Synthèse des Conclusions – Article 3

Une récente revue de littérature systématique sur la réhospitalisation à 30 jours concluait de la façon suivante (Mahmoudi et al., 2020, p. 7) :

> Malgré le nombre croissant de la littérature en faveur des méthodes de Machine Learning comme alternative, en supplément des avantages potentiels de leur utilisation pour prédire les réhospitalisations, il reste à aborder trois critères importants qui ont été mis en évidence dans cette revue systématique. Premièrement, la sélection des variables reste un critère important qui repose sur la disponibilité d'un ensemble exhaustif et diversifié d'éléments de données, tels que le statut socio-économique et fonctionnel. Des études ultérieures devraient envisager de mettre en œuvre des éléments de données suffisamment granulaires via le text mining, en les fusionnant avec des unités géographiques d'analyse plus petites (secteur de recensement ou niveau du quartier), ou en encourageant les systèmes de santé à collecter ces attributs spécifiques. Deuxièmement, les méthodes de Machine Learning peinent à atteindre la parcimonie en raison du recours à plusieurs centaines à des milliers de variables pour prédire un résultat. L'utilisation de méthodes de Machine Learning, bien qu'à la mode et offrant un exercice académique potentiel, ne répond pas à des questions cliniques importantes sur la mise en œuvre et l'interprétabilité des résultats. Troisièmement, les méthodes de Machine Learning varient considérablement dans leur interprétabilité, créant des barrières et des obstacles à l'adhésion clinique et à leur mise en œuvre dans les systèmes de santé. Bien que les méthodes d'apprentissage automatique interprétables aient été absentes de cette revue systématique, l'évolution du domaine nécessite le développement et la mise en œuvre de méthodes de Machine Learning interprétables pour établir l'utilité clinique et inspirer des changements potentiels dans les modèles de pratique.

Le présent article traite très largement des points et des préoccupations mentionnés ci-avant.

1- La performance des données textuelles tabulaires vectorisées

Les résultats obtenus ici ne pâlissent pas face à d'autres travaux prédisant la rehospitalisation sous 30 jours (Zhou et al., 2016), et certainement pas aux travaux n'utilisant que les textes cliniques (K. Huang et al., 2020; Orangi-Fard et al., 2022). Avec un ROC AUC de 0.723, la capacité de discrimination de notre meilleur modèle est considérée de façon générale comme satisfaisante (Hosmer et al., 2013), mais se situant également dans la fourchette haute pour les réhospitations sous 30 jours.

Autrement dit : il est pertinent d'utiliser des notes cliniques pour prédire et mieux expliquer la RH30

2- Un gain significatif en interprétabilité

Les différentes formes de vectorisation, qu'elles soient à haute dimension (BOW) ou à dimension réduite (LDA), non seulement convergent dans leurs prédictions mais fournissent en plus des informations très spécifiques sur les patients à risques, telles que :

- le type de soin (attention et médication fréquente),
- les pathologies ou type de prise en charge (trachéotomie, insuffisance rénale, respiratoire, cardiaque),
- le type d'hospitalisation et l'âge (hospice, crash automobile),
- la persistance des conditions (dialyse, transplantation, chronicité)
- des conditions spécifiques de fin de vie (ne pas réanimer, soins palliatifs, phase finale d'une pathologie rénale),
- des médicaments (dilantin, vancomycine, carvedilol, digoxyn, miconazole, rifaximin, etc.),

- des types de traitements (renal, respiratoire, cardiaque, infectueux, gastriques, sanguins, etc.)

Ce premier article, ayant recours exclusivement à des prédicteurs textuels apporte ainsi une solution concrète à la critique sur l'interprétabilité des modèles de Machine Learning.

3- Complementarité des informations avec les données structurées

Même si nous n'avons pas utilisé les informations cliniques classiques pour prédire le RH30, des analyses descriptives (bivariées) permettent d'examiner les profils des patients réadmis. Ainsi constate-t-on assez logiquement que les personnes les plus à risques sont celles qui sont sous assurance Medicare (information indisponible via les tokens des données textuelles).

A contrario, les données textuelles fournissent de multiples informations sur le risque des personnes âgées (hospice) sans qu'il n'apparaisse une distinction significative dans les classes d'âge via analyse des données structurées.

La suite de nos travaux cherche à combiner les informations fournies par les données structurées et non structurées afin d'améliorer simultanément la performance et l'interprétabilité.

Partie 3 – Fusion de Données Structurées et Non Structurées

Prédire (et expliquer) la durée de séjour à partir de données structurées, non structurées et mixtes du MIMIC III

Contexte

Cette partie de la thèse cherche à tenir compte autant que possible des limites des précédents articles :

- Le jeu de données de l'APHM gagnerait à être enrichi par des variables structurées supplémentaires telles les tests de laboratoire, les traitements médicamenteux, les signes vitaux, etc. (Zhou et al., 2016)

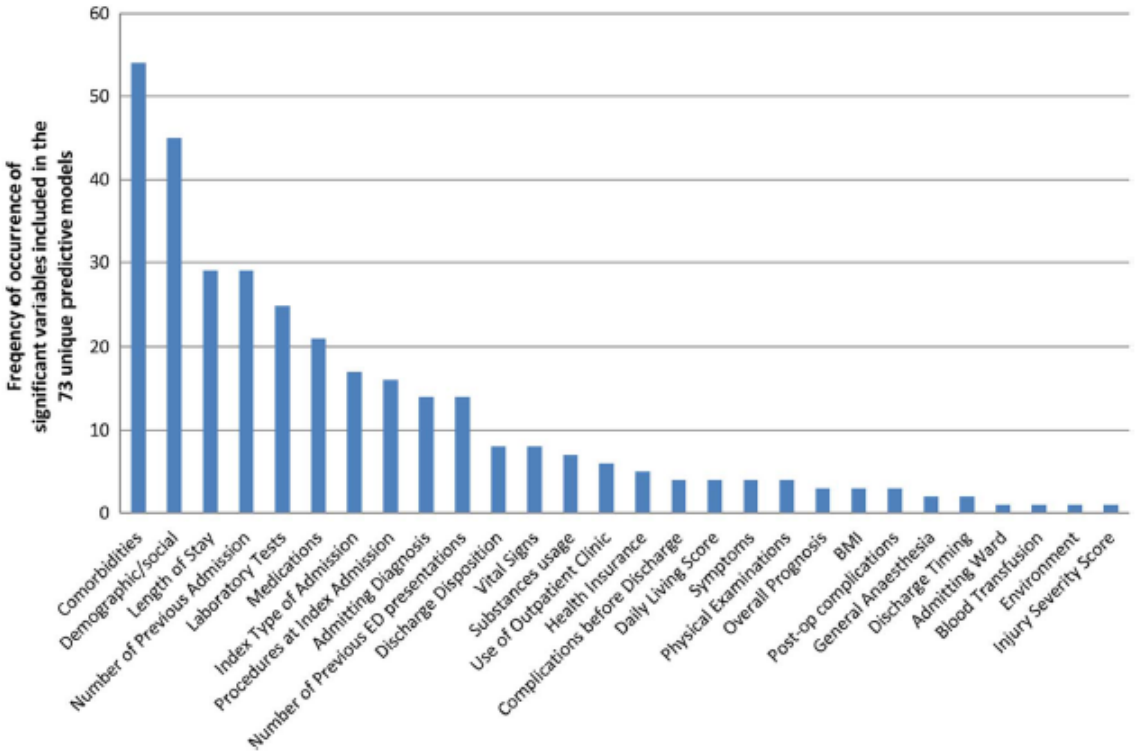


Figure 3-1. Importance des Variables prédisant le LOS - D'après Zhou et al. (2016, p. 21)

- Si l'interprétabilité est améliorée par la disjonction complète des variables catégorielles (one-hot encoding) et par la prise en compte des données textuelles, on peut chercher à augmenter la qualité du compromis performance-interprétabilité. Pour cela, il reste à explorer la différence de performance et d'interprétabilité entre les (meilleurs) modèles utilisant les données structurées seules, ceux utilisant les données textuelles seules, et ceux utilisant une combinaison de données structurées (tabulaires) et données textuelles (tabulaires ou séquentielles)

Pour tenter de dépasser ces limites nous allons procéder en plusieurs étapes :

- La base de données utilisée est le MIMIC III
- On sélectionnera des variables biologiques qui permettront d'étendre significativement la richesse et la diversité de nos données structurées (Kareliusson et al., 2015; Nguyen et al.,

2016) au-delà de celles typiquement utilisées dans les données de l'APHM (analyses de laboratoire, constante physiologiques etc.).

- On inclura les notes cliniques des patients au moment de la sortie (discharge notes)
- On procédera à trois groupes d'analyses (données structurées seules, données textuelles seules, et données mixtes). Pour chacune d'entre elles on cherchera à évaluer la performance et à étudier l'interprétabilité en utilisant des méthodes globales (permutation importance) et/ou des méthodes locales (SHAP, LIME)

Les données

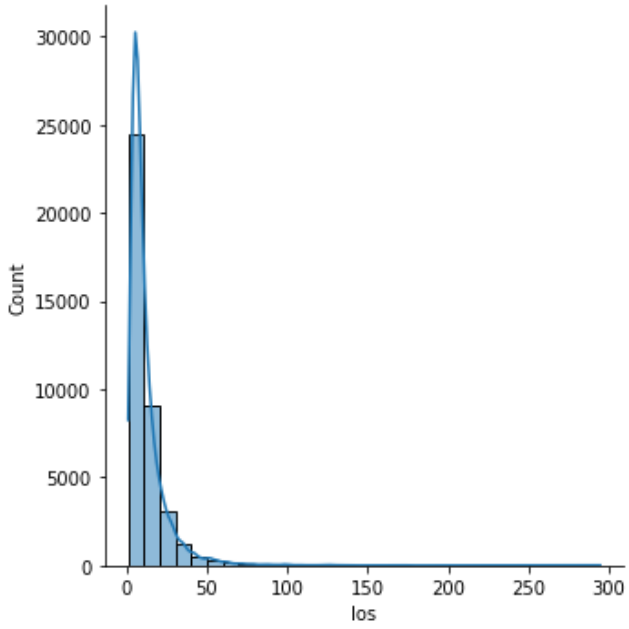
1- Critères d'inclusion

Après avoir exclu les décès, les patients de moins de 18 ans et les durées de séjour inférieures à 24 heures, le jeu de données comporte 39105 séjours

2- La variable à expliquer

La variable d'intérêt est la durée de séjour, calculée comme la différence entre la date précise de décharge et la date précise d'admission.

Rappelons que les séjours du MIMIC III sont principalement associés à des soins intensifs. La distribution de la durée de séjour est donc significativement différente de celle des données APHM :




```
count    39105.000000
mean        12.171897
std         12.777174
min          1.000000
25%          5.141667
50%          8.306944
75%         14.506250
max        294.660417
Interquartile Interval = 9.365
Lower Fence = -8.905, Upper Fence = 28.553
```

Figure 3-2. distribution du LOS

On constate que la clôture supérieure est de 28 jours, c'est-à-dire le double de la durée pour les données de l'APHM. Cela représente 92.56% de séjours ordinaires et 7.44% de séjours prolongés pour les soins intensifs. On se situe donc au-delà de la 92ème centile, ce qui reste dans la fourchette de la 90ème centile et de la 95ème centile (Marfil-Garza et al., 2018b).

3- La variable d'intérêt binarisée (los_cat)

Est définie comme suit:

$$y = 1 \; si \; durée \; de \; séjour > 28 \; sinon \; y = 0$$

4- Les prédicteurs

Les variables des données structurées sont les suivantes.

Tableau 3-1 Données Structurées pour prédire le LOS

Clinical Characteristics
Simplified Acute Physiology Score (SAPS II)
Sepsis-related Organ Failure Assessment (SOFA)
Urea nitrogen min
Urea nitrogen max
Urea nitrogen mean
Platelets min
Platelets max
Platelets mean
Magnesium max
Calcium min
Respiratory rate min
Respiratory rate max
Respiratory rate mean
Glucose min
Glucose max
Glucose mean
Heart Rate min
Heart Rate max
Heart Rate mean
Systolic Blood Pressure min
Systolic Blood Pressure max
Systolic Blood Pressure mean
Diastolic Blood Pressure min
Diastolic Blood Pressure max
Diastolic Blood Pressure mean
Temperature min
Temperature max
Temperature mean
Urine Ouput min
Urine Ouput max
Urine Ouput mean
ICD Chapter

Tableau 3-1 Données Structurées pour prédire le LOS

Sociodemographic Characteristics
Ethnicity
Insurance
Religion
Marital Status
Gender / Sex
Age
Hospitalization Characteristics
Admission Type
Admission Location
Readmission Time Delta
At least one previous hospitalization via emergency departments 6 months before
Hospitalization via Emergency Departments
Origin of Patient
Destination on discharge
Discharge Notes
Length of Hospital Stay (LOS)

Les données non structurées sont constituées des notes cliniques de sortie (discharge notes)

Préparation et Fusion des Données

1- Package AutoGluon pour les données structurées

Pour cette partie de la thèse, on utilise le package « AutoGluon Tabular » (Erickson et al., 2020).

AutoGluon-Tabular est une bibliothèque AutoML en Python, simple d'utilisation, optimisée pour des données tabulaires. Il procède à un traitement avancé des données, avec un apprentissage en profondeur et ensembliste (Ensemble Learning) de modèles par couches successives.

Il reconnaît automatiquement le type de données dans chaque colonne, et effectue un prétraitement robuste des données, incluant une gestion spéciale des champs de texte. AutoGluon s'adapte à divers modèles allant des arbres boostés prêts à l'emploi aux modèles de réseaux de neurones customisés.

Il utilise une nouvelle approche d'apprentissage ensembliste[20] des modèles : ces derniers sont empilés en plusieurs couches et entraînés couche par couche de façon à garantir que les données brutes puissent être traduites en prédictions de haute qualité dans une contrainte de temps donnée. Le surapprentissage (overfitting) est atténué tout au long de ce processus en divisant les données de différentes manières avec un suivi minutieux des exemples « Out of Fold » (hors partitions d'apprentissage).

En appelant la méthode $fit()$, AutoGluon réalise automatiquement les tâches suivantes : prétraite les données brutes, identifie le type de prédiction (binaire, classification multi-classes ou régression),

[20] On utilisera alternativement le terme « ensembliste » pour désigner l'expression « Ensemble Learning »

découpe les données en différents partitions d'entraînement et de validation, entraîne individuellement chaque modèle, et enfin génère un modèle ensembliste optimisé qui surpasse facilement tous les modèles entraînés individuellement.

AutoGluon utilise un ensemble de modèles sur mesure dans un ordre prédéfini. On garantit ainsi que des modèles performants et fiables, tels que les forêts aléatoires, soient entraînés avant des modèles plus consommateurs de ressources et moins fiables, tels que les k plus proches voisins. Cette stratégie est essentielle lorsqu'on veut imposer des délais à $fit()$. En particulier, parmi la liste on trouve des réseaux de neurones, des arbres boostés $LightGBM$ (Ke et al., 2017), des arbres boostés $CatBoost$ (Dorogush et al., 2018), des forêts aléatoires ($Random\ Forest$), des arbres extrêmement aléatoires ($Extra\ Trees$) et des k-plus proches voisins (kNN).

Les données tabulaires ne bénéficient pas des structures des images et des textes qui peuvent être exploitées via des convolutions (CNN) ou des récurrences (RNN). Au lieu de cela, les ensembles de données tabulaires sont composés de divers types de valeurs, et donc les réseaux de neurones denses (multiperceptrons) sont généralement l'architecture de choix. Cependant, les caractéristiques brutes d'une table de données sont mieux adaptées aux modèles d'arbres, qu'à une couche dense de réseaux de neurones qui réalise une combinaison linéaire de toutes les variables avant d'appliquer une fonction d'activation dans chaque neurone d'une couche cachée. Néanmoins, Mendoza et al. (2016) ont démontré que des réseaux de neurones correctement entraînés peuvent apporter des gains de performance significatifs lorsqu'ils sont inclus dans un ensemble existant d'autres types de modèles (approche ensembliste d'apprentissage ou Ensemble Learning).

L'architecture réseau utilisée par AutoGluon est illustrée à la Figure 3-3.

Le réseau applique une couche d'intégration distincte à chaque variable catégorielle, où la dimension d'embedding est sélectionnée proportionnellement au nombre de niveaux uniques observés (modalités) pour cette variable (Guo & Berkhahn, 2016). Pour les données multivariées, les couches d'intégration individuelles permettent au réseau d'apprendre séparément chaque variable catégorielle avant que sa représentation ne soit intégrée avec d'autres variables. Les embeddings de variables catégorielles sont concaténées avec les variables numériques dans un grand vecteur qui est à la fois transmis dans un réseau dense à 3 couches et directement connecté aux prédictions de sortie via une connexion linéaire.

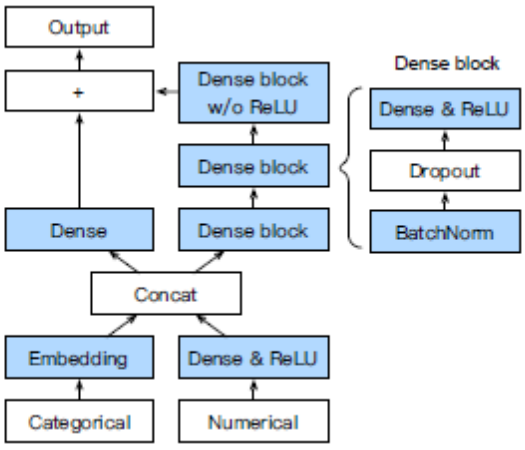


Figure 3-3. Architecture des Réseaux de Neurones pour les données tabulaires pour AutoGluon. Les couches en bleu correspondent à des hyperparamètres appris – D'après Erickson (2020)

Les ensembles qui combinent les prédictions de plusieurs modèles sont connus depuis longtemps pour surpasser les modèles individuels, réduisant souvent considérablement la variance des prédictions finales (Dietterich, 2000).

AutoGluon introduit une nouvelle forme de procédé ensembliste multicouches, illustrée à la Figure 3.4. Ici, la première couche comporte plusieurs modèles de base, dont les sorties sont concaténées puis envoyées dans la couche suivante, qui se compose elle-même de plusieurs modèles empilés par couches. Ces couches empilées servent alors de modèles de base à une couche supplémentaire.

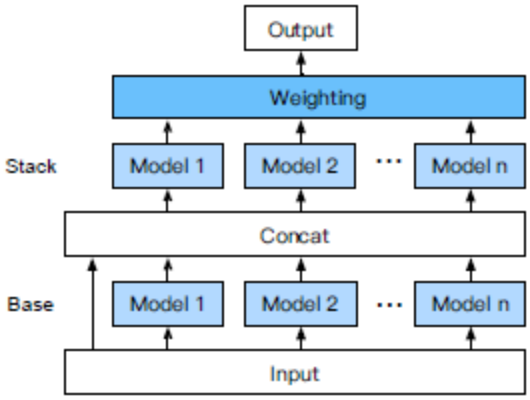


Figure 3-4. Stratégie d'empilement de couches utilisée par AutoGluon. L'exemple comporte n modèles de base et 2 couches d'empilement – D'après Erickson (2020)

Analyse des données structurées tabulaires

Puisque nos données sont tabulaires, on mobilise l'objet <TabularPredictor> avec les paramètres suivants :

- $label$ = « los_cat » (le nom donné à la variable LOS binarisée)
- $eval\_metric$ = « roc_auc » (la métrique pour évaluer la performance des modèles)
- pour les différents classifieurs, nous avons conservé les paramètres par défaut[21]

Par défaut la méthode . $fit()$ d'apprentissage du <TabularPredictor> entraîne 14 modèles dont le dernier est un modèle $Ensembliste\ Pondéré$ des 13 autres (voir Figure 3-4).

Les résultats sont fournis par le Tableau 3-2 ci-après :

[21] Donnés ici : https://auto.gluon.ai/stable/api/autogluon.predictor.html#module-0

Tableau 3-2 : Performance (ROC AUC) des données structurées tabulaires

model	score_test	score_val	fit_time	stack_level	fit_order
WeightedEnsemble_L2	0.964	0.972	154.759	2	14
CatBoost	0.963	0.968	64.047	1	7
LightGBM	0.963	0.966	3.528	1	4
LightGBMXT	0.962	0.969	8.161	1	3
XGBoost	0.962	0.966	2.887	1	11
LightGBMLarge	0.962	0.967	4.490	1	13
RandomForestEntr	0.952	0.961	4.972	1	6
ExtraTreesEntr	0.950	0.960	1.657	1	9
ExtraTreesGini	0.948	0.957	1.889	1	8
NeuralNetTorch	0.947	0.950	38.172	1	12
RandomForestGini	0.945	0.957	4.879	1	5
NeuralNetFastAI	0.945	0.951	33.068	1	10
KNeighborsDist	0.854	0.833	0.036	1	2
KNeighborsUnif	0.837	0.812	0.034	1	1

Bien entendu, le modèle retenu est le meilleur modèle, à savoir le modèle Ensembliste.

Le tableau des performances dans différentes métriques est fourni ci-après.

Tableau 3-3 : Métriques des Performances

metrics	performances
Cohen's Kappa	0.670
PRC AUC	0.789
ROC AUC	0.964
Accuracy	0.962
Balanced Accuracy	0.783
mcc (*)	0.686
f1	0.689
precision	0.863
recall	0.574

(*) Matthews Correlation Coefficient (Matthews, 1975) est l'équivalent du Phi de Pearson mesurant le degré d'association entre deux variables catégorielles via un tableau de contingence (Yule, 1912)

Les valeurs indiquent de très bonnes performances de classification. Il s'agit maintenant de déterminer quelles sont les variables qui sont les plus prédictives (les 20 ou 30 premières) des séjours prolongés, compte tenu de ces performances.

On procèdera de deux façons : en utilisant l'importance par permutation et en utilisant SHAP.

Permutation importance

Les résultats de l'importance des variables par permutation sont donnés par la Figure 3.5

On observe que la durée de séjour prolongée est largement prédite par les taux de nitrogène uréique et de plaquettes. Ces résultats sont conformes à la littérature qui associe un taux élevé de nitrogène uréique à un risque de séjour prolongé en soins intensifs pour cause de pancréatites nécrosantes

(Faisst et al., 2010) ou encore à une mortalité plus élevée pour cause d'embolisme pulmonaires (Tatlisu et al., 2017) ou encore pour les patients plus âgés (Dundar et al., 2021).

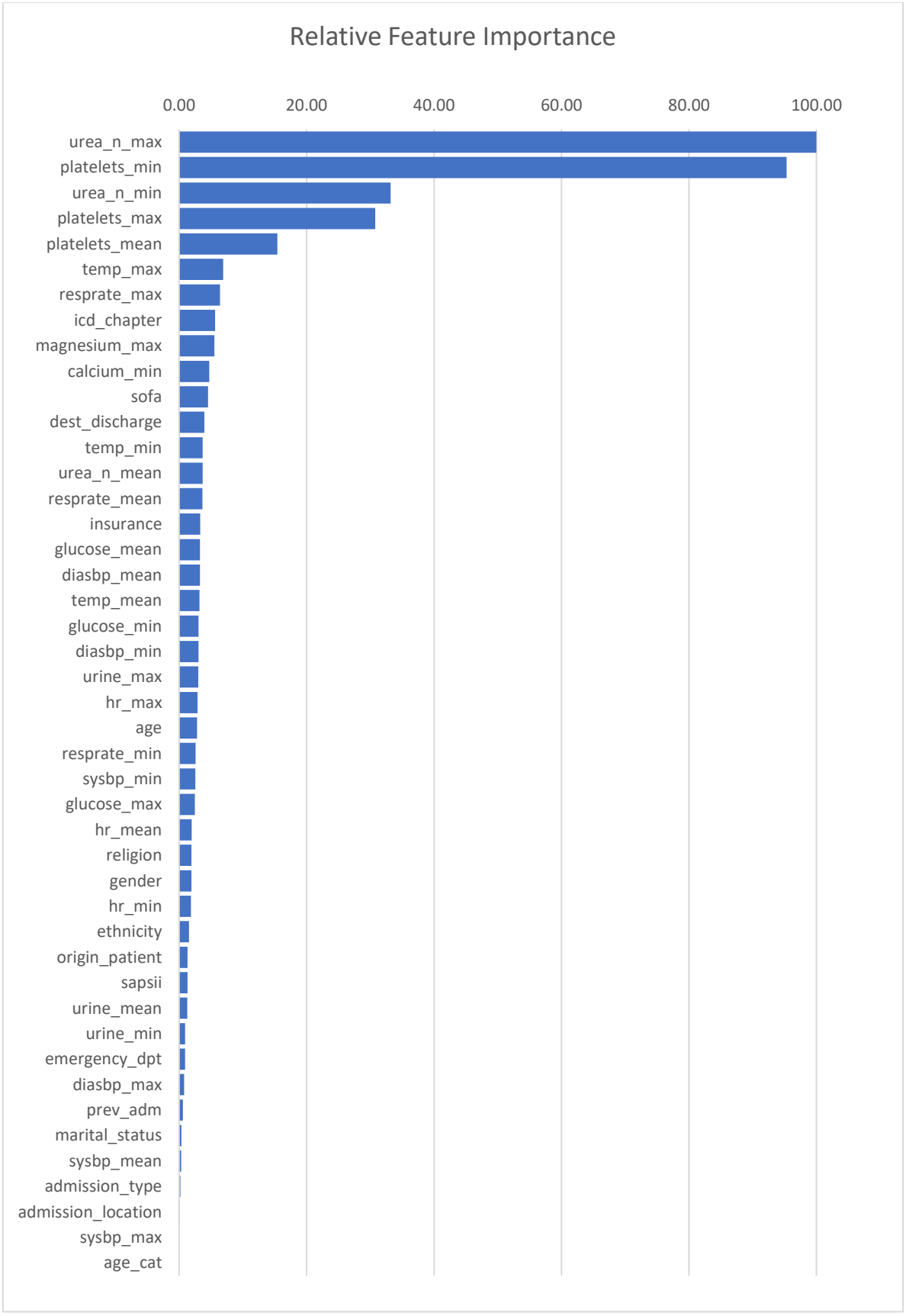


Figure 3.5. Permutation Importance

De la même façon un taux faible de plaquettes est associé à toutes sortes de complications et de prolongement du séjour hospitalier, par exemple par augmentation du risque d'infection (Qu et al., 2018) ou suite à des complications post opératoires (Abanoz & Engin, 2021; Amygdalos et al., 2020)

On notera cependant que l'importance des variables par permutation reste ambigüe sans une analyse bivariée complémentaire, pour indiquer notamment si la relation entre le prédicteur et la variable d'intérêt est dans un sens croissant ou décroissant. Par exemple en, examinant la Figure 3-5, et en examinant la littérature, on se doute que lorsque le taux de nitrogène uréique augmente alors la probabilité de prolonger le séjour également (relation positive ou croissante) et vice versa, lorsque le taux de plaquette diminue, alors la probabilité de prolonger le séjour augmente (relation négative ou décroissante). Mais cette information n'est pas explicitement disponible dans le graphique.

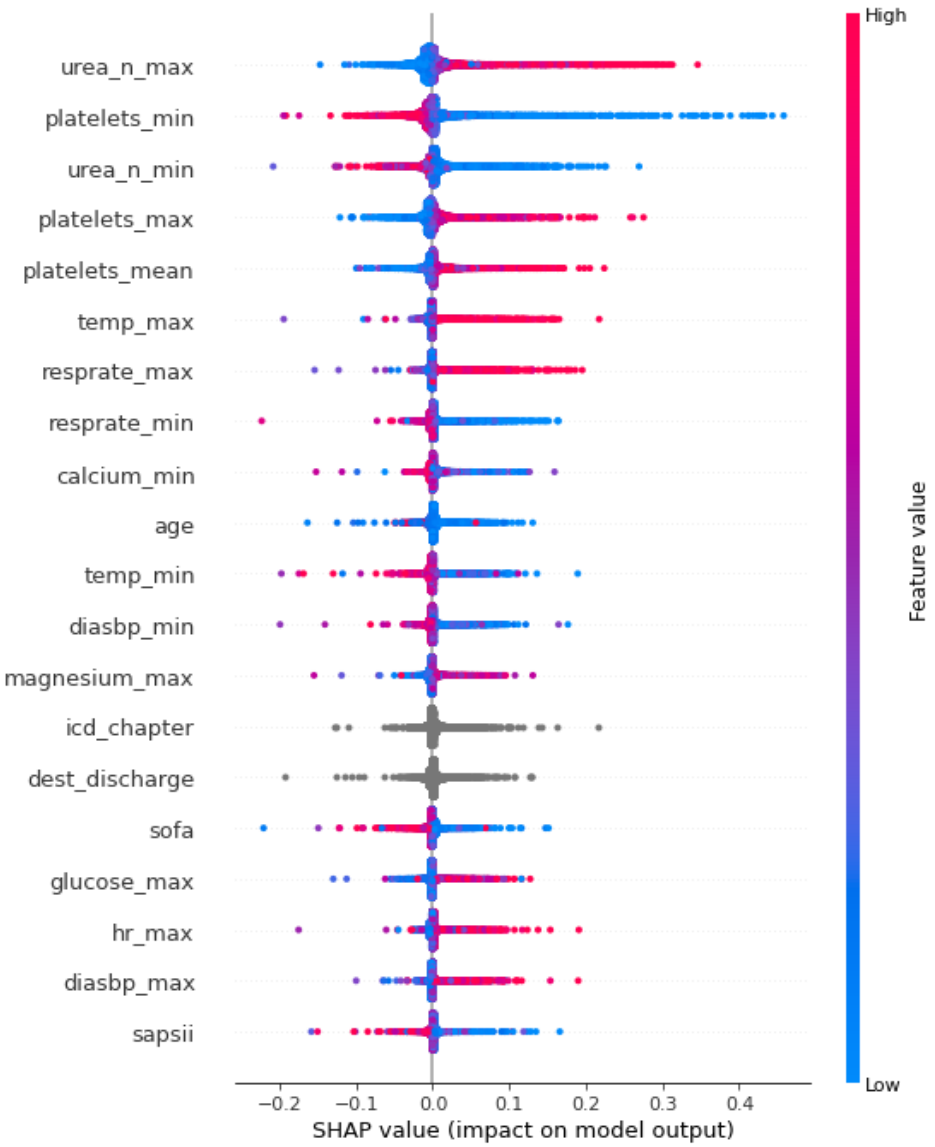


Figure 3-6. SHAP Feature Importance

Pour mieux appréhender ces nuances, on peut s'appuyer sur la représentation graphique donnée par la méthode $shap.plots.beeswarm()$ de SHAP (Figure 3-6). Les résultats entre les deux graphiques convergent, même si on observe des différences pour les variables moins importantes. Les valeurs

positives de SHAP vont dans le sens du séjour prolongé et les valeurs négatives indiquent les séjours ordinaires. La couleur rouge indique que là où elle situe les valeurs augmentent tandis que pour la couleur bleue, les valeurs diminuent. On lit ainsi que lorsque le taux maximum de nitrogène uréique augmente (rouge), le séjour tend à se prolonger (SHAP positif) et à contrario lorsque le taux minimum de plaquettes diminue (bleue) alors le séjour tend à se prolonger.

Gardant à l'esprit que chaque point de couleur représente un séjour, on voit bien que même quand la tendance est rouge, il existe des points bleus, et vice versa, indiquant que la tendance qui se vérifie au niveau global connaît des exceptions au niveau local.

Malgré tout, dans ce graphique il manque l'information sur le poids global relatif de chaque variable d'un côté et de l'autre la difficulté à vraiment interpréter les variables catégorielles (grisées). Il faut pour cela mobiliser une autre fonctionnalité de SHAP via la méthode $shap.plots.bars()$.

Package AutoGluon pour les données non-structurées textuelles

Nous nous intéressons à présent à la prédiction du séjour prolongé à partir de textes exclusivement, mais au lieu d'utiliser des modèles classiques de Machine Learning sur des données tabulaires, nécessitant une vectorisation du texte, on s'intéresse à des modèles de Deep Learning en utilisant des données séquentielles.

Le module <TextPredictor> d'AutoGluon permet de réaliser très simplement cette tâche en partant d'une dataframe de données textuelles et en utilisant ce qui se fait de mieux dans la classification de textes, à savoir les Transformers. En particulier <TextPredictor> est compatible avec les Transformers de Hugging Face.

De plus, étant donné que nos données sont très spécifiques à une discipline (médicale, clinique biologique) nous allons choisir un modèle pré-entraîné sur des données cliniques très complètes : le Bio Clinical BERT[22] (Alsentzer et al., 2019).

1- Phase 1 : Aucun Pré-traitement des Données

Un des avantages reconnus du Deep Learning est la possibilité de réduire au strict minimum la préparation des données (feature engineering) avant de lancer l'entraînement des modèles. De fait, dans les travaux avec les modèles de langues, comme la traduction ou la génération de texte, il faut éviter d'altérer les données pendant l'apprentissage afin de ne pas perturber la prise en compte du contexte. Dans la classification d'images, le prétraitement des données est réalisé directement par l'architecture du modèle et on peut supposer qu'il pourrait en être de même dans la classification de textes.

C'est donc avec cet a priori positif que nous avons lancé l'analyse :

- La variable d'intérêt : « los_cat » (le LOS binarisé)
- Le prédicteur : « discharge » (notes de décharges)
- L'hyperparamètre du modèle d'entraînement est : "emilyalsentzer/Bio_ClinicalBERT"

Les résultats de l'analyse sont donnés par le Tableau 3.4 suivant :

[22] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

Tableau 3.4 : Performance du modèle avec le prédicteur « discharge » seul

métriques	performances
Cohen's Kappa	0.873
PRC AUC	0.935
roc_auc	0.988
f1	0.882
acc	0.984

Ces résultats sont extrêmement bons ! Bien meilleurs que ceux des données tabulaires… !

Cependant pour évaluer la valeur clinique de ces résultats, il nous faut en creuser l'interprétabilité. Pour les modèles de deep learning, l'importance par permutation ne fonctionne pas, et SHAP prend un temps infini (l'estimation pour une seule observation prend plusieurs dizaines de minutes – or nous avons près de 8000 observations). Nous avons donc conservé LIME pour l'estimation de l'importance en prenant la moyenne de la valeur absolue des poids (coefficients) du modèle linéaire (Lasso) de substitution sur l'ensemble de l'échantillon de test. Le résultat est donné par la Figure 3-7 ci-après.

Ce résultat est troublant. En examinant plus attentivement le texte, on note que dans le processus d'anonymisation, les champs de dates (certes modifiés) ont été maintenus dans le texte :

```
Admission Date:  [**2150-4-17**]          Discharge Date:   [**2150-4-21**]

Date of Birth:  [**2090-5-19**]             Sex:   M

Service: MEDICINE

Allergies:
Patient recorded as having No Known Allergies to Drugs

Attending:[**First Name3 (LF) 12174**]
Chief Complaint:
coffee ground emesis

Major Surgical or Invasive Procedure:
EGD
Right IJ CVL


History of Present Illness:
Mr. [**Known lastname 52368**] is a 59M w HepC cirrhosis c/b grade I/II
esophageal
varices and portal gastropathy (last EGD [**3-/2150**]), who p/w
coffee-ground emesis and melena x2 days.
(…)
```

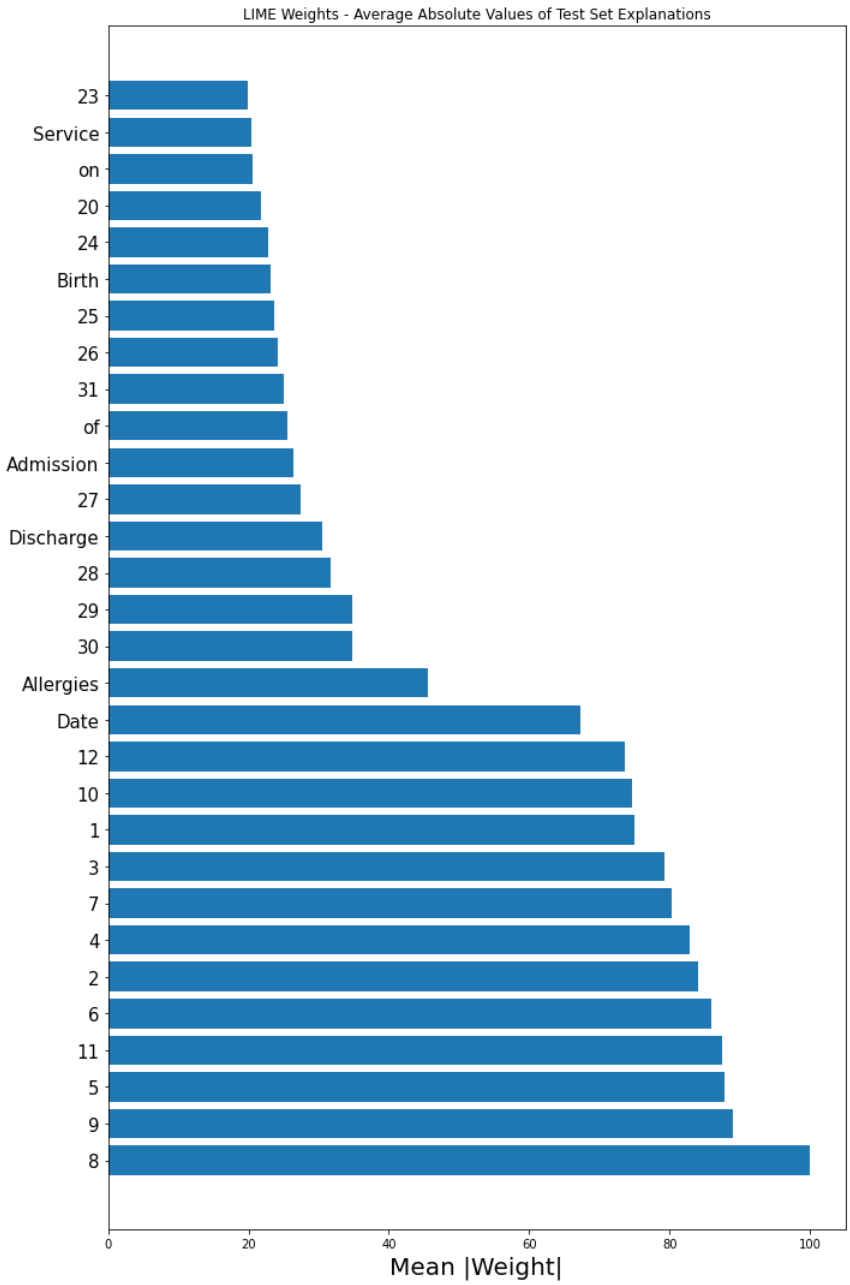


Figure 3-7. LIME Global Feature Importance for the text only “discharge” feature

Pour mieux comprendre le phénomène, on décide de recoder tout ce qui se trouve dans le champ $admission\ date$ et $discharge\ date$ en $date\_placeholder$ et on relance l’analyse. Le résultat obtenu est donné ci-après (Figure 3-8).

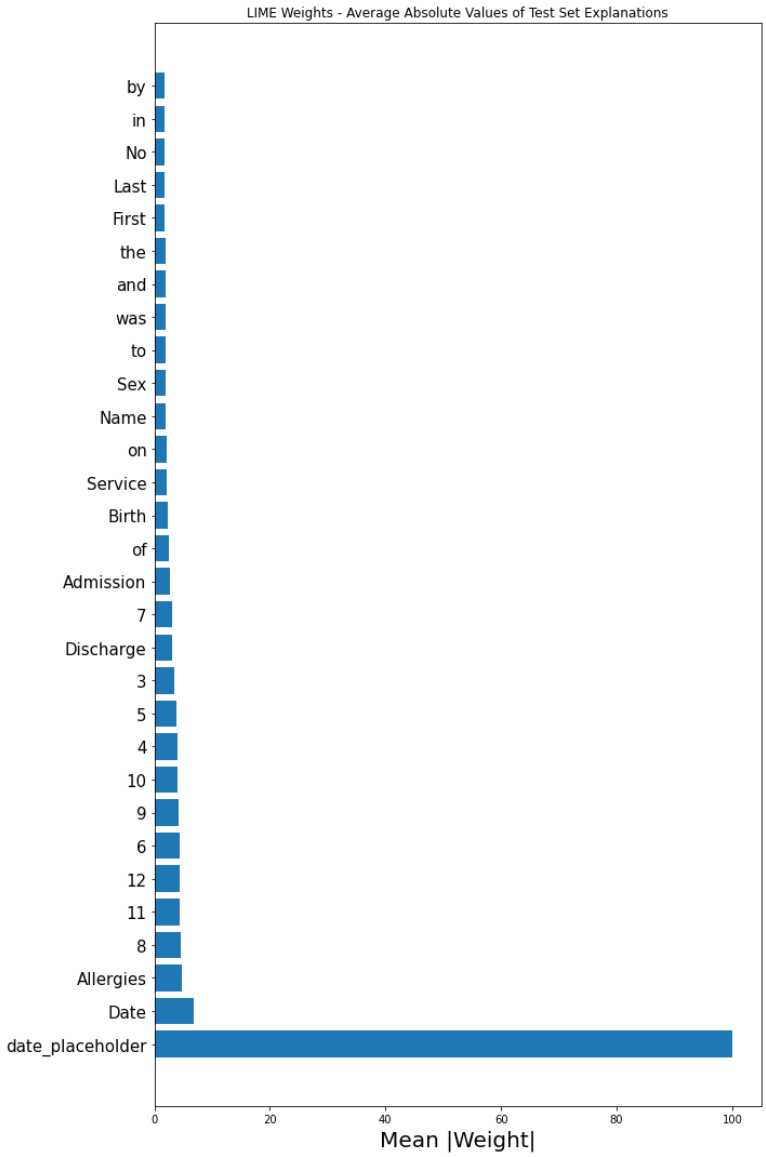


Figure 3-8. LIME Global Feature Importance with date placeholder

On se rend alors à l'évidence que les tokens renseignant les dates d'admission et la date de décharge représentent le principal prédicteur, et il est donc normal qu'ils prédisent aussi bien le LOS. On se trouve dans un cas affirmé de « data leakage » : visiblement le modèle a réussi à reconnaître les dates d'admission et de prédiction à partir du texte et à s'en servir pour prédire le LOS.

Il en résulte en fin de compte qu'un prétraitement des données peut s'avérer nécessaire dans le cas qui nous occupe.

2- Phase 2 : Pré-traitement « léger » des Données

On décide donc d'éliminer tous les textes ou groupes de textes qui contiennent des dates, et on relance l'analyse. Les résultats sont donnés par le Tableau 3-5 ci-après :

Tableau 3-5 : Performance du modèle avec pré traitement « léger » du texte « discharge »

métriques	performances
Cohen's Kappa	0.200
PRC AUC	0.413
roc_auc	0.859
f1	0.219
acc	0.931

On notera une nette baisse de la performance, même si celle-ci reste tout à fait honorable (on se fiera principalement au ROC AUC et au PRC AUC, les autres indicateurs étant dépendants d'un seuil de classification peu adapté à des données très déséquilibrées comme les nôtres.

Quant à l'interprétabilité, on obtient à présent un graphique nettement plus lisible (Figure 3-9) :

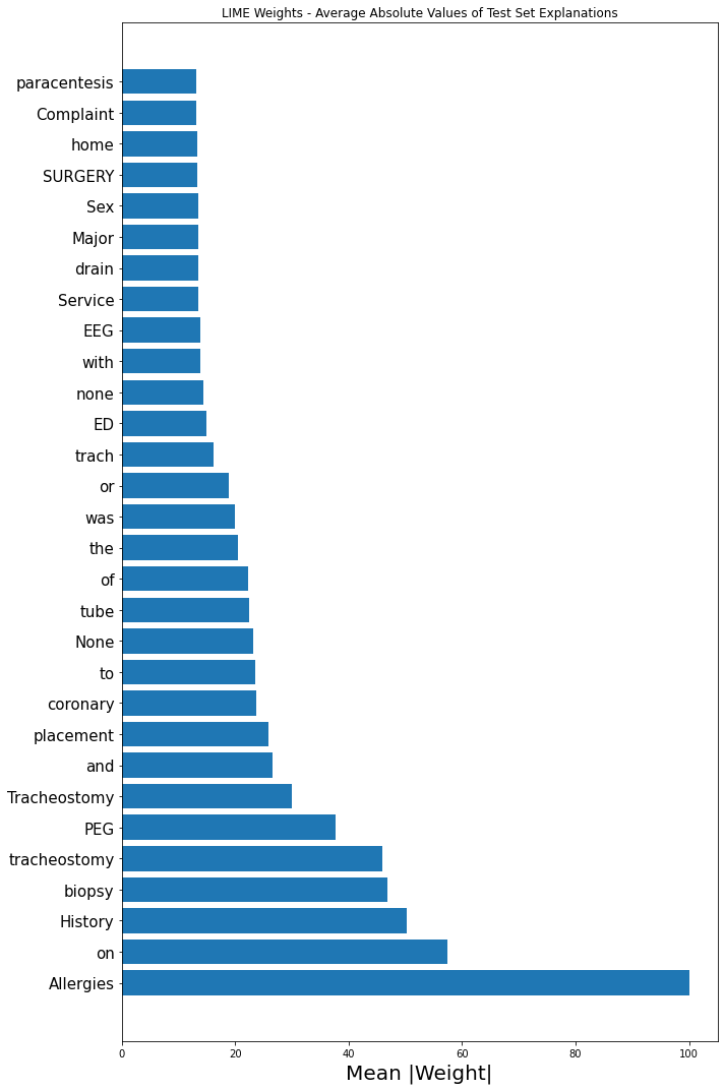


Figure 3-9 - LIME Global Feature Importance with light Preprocessing

3- Phase 3 : Pré-traitement « plus lourd » des données

Les résultats de la Figure 3-9 incluent des stopwords. On procède donc à un prétraitement qui inclut l'élimination des ponctuations, une mise en minuscule du texte, une élimination des mots vides (stopwords) et alternativement une racinisation (stemming) ou une lemmatisation (lemmatization). La performance est donnée par le Tableau 3-6 ci-après :

Tableau 3-6 : Performance du modèle avec pré traitement « plus lourd » du texte « discharge »

performances	**stemming**	**lemmatization**
Cohen's Kappa	0.248	0.329
PRC AUC	0.382	0.401
roc_auc	0.848	0.859
f1	0.275	0.364
acc	0.929	0.927

La performance s'est encore très légèrement dégradée pour la racinisation mais reste toujours bonne. La lemmatisation constitue distinctement un meilleur prétraitement que la racinisation.

L'interprétabilité quant à elle a fait un bond et permet d'enrichir nettement l'interprétation pour les risques de réhospitalisations sous 30 jours comme le montre la Figure 3-10 :

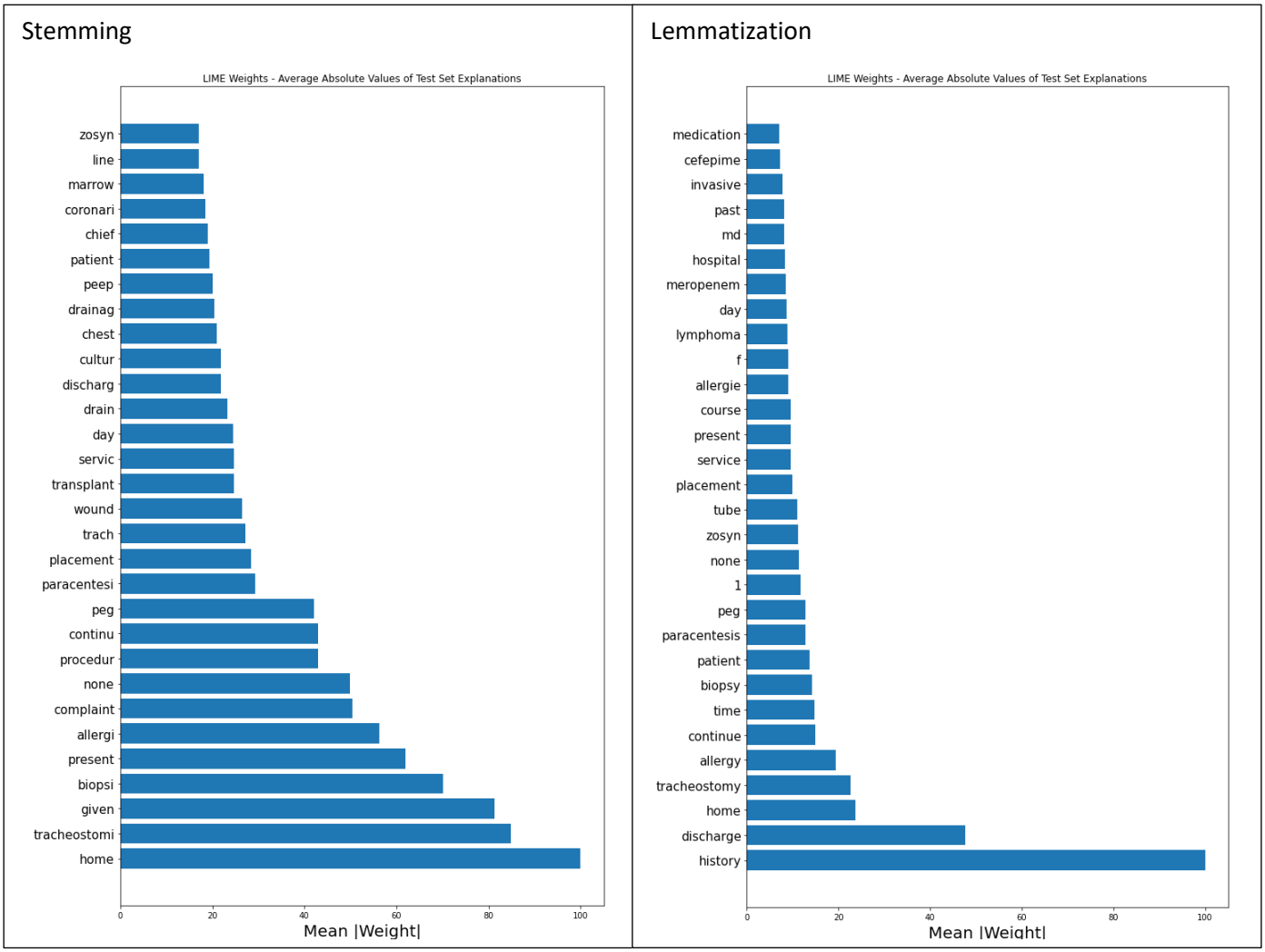


Figure 3-10 - LIME Global Feature Importance

Il existe une différence notable dans les deux représentations qui peuvent tant provenir du pré-traitement que de l'instabilité de LIME. Il est sans doute plus prudent de ne conserver que ce qu'ils ont en commun, comme <tracheostomy> qui ressort déjà dans la littérature comme un important facteur de risque. Ainsi on retiendra tout particulièrement : home, discharge, paracentesis, allergies, biopsy, history, zosyn, etc. On peut également simplement favoriser davantage les résultats de la lemmatisation qui conduisent à de meilleures performances, ou considérer les informations non-confirmées comme des indicateurs supplémentaires à considérer.

Package AutoGluon pour les données mixtes : structurées + textuelles

Combiner les données structurées et non-structurées est une pierre angulaire en Machine Learning notamment dans l'image captioning (Hossain et al., 2019). Dans les sciences de la santé, un des enjeux majeurs est de combiner les données structurées cliniques, hospitaliers et socio-démographiques avec de l'image (imagerie médicale) ou du texte (notes cliniques).

De récents travaux ont fusionné des données structurées et non structurées issues de la base MIMIC, en utilisant des embeddings de données de textes séquentielles, et des données temporelles également séquentielles avec des données structurées statiques, dans une architecture CNN et LSTM (Zhang et al., 2020). Comme dans notre étude, les variables d'intérêts sont alternativement le séjour prolongé (au-delà de 7 jours), et la réadmission non planifiée sous 30 jours, ainsi que la mortalité hospitalière.

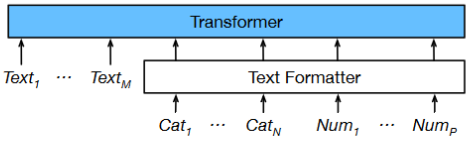
La fonction multimodale d'AutoGluon permet théoriquement d'accomplir la même tâche mais en utilisant les Transformers.

1- Mécanisme de Fusion de Multimodal AutoGluon (MAG)

Les principe de MultiModal AutoGluon (MAG) consiste à traiter des jeux de données qui contiennent à la fois du texte libre, des données quantitatives (numériques), et des données catégorielles – le tout dans un format tabulaire (Shi et al., 2021).

D'un côté, il est admis que les réseaux de neurones profonds (comme les Transformers) sont performants pour les textes, mais que les ensembles d'arbres sont plus performants pour les données structurées. Dans tous les cas, pour pouvoir tirer profits des informations du texte dans des données tabulaires, les textes doivent être vectorisés. La difficulté est de faire en sorte que les informations nécessaires à la prédiction ne soient pas exclusivement dominées ni par le texte, ni par les données structurées. Plusieurs stratégies peuvent ainsi être adoptées (Shi et al., 2021, pp. 2–3) :

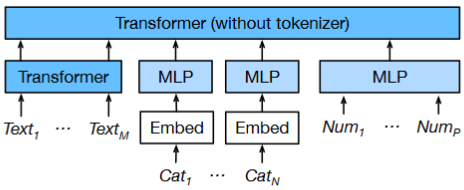
i. **All text**. On convertit les valeurs numériques en valeurs catégorielles (en les traitant comme des chaînes de caractères). On peut alors utiliser un Transformer pré-entraîné pour tout encoder.



(a) *All-Text.* Convert numeric and categorical values into additional text tokens.

Figure 3-11. D'après Shi et al. (2020, p.3)

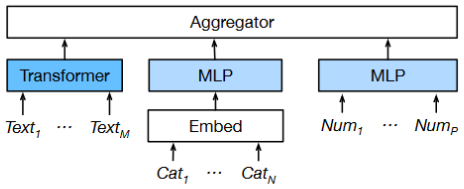
ii. **Fuse Early**. On apprend au modèle des embeddings pour représenter les valeurs numériques et catégorielles comme des tokens spécifiques. Autrement dit, chaque variable catégorielle est vectorisée séparément en embedding avant d'être envoyée dans un réseau de neurones dense (multiperceptrons = MLP) alors que toutes les variables quantitatives sont envoyées directement dans un MLP. Ces trois types d'embedding sont ensuite transférés à un Transformer (sans tokenizer) à six couches avec un système d'attentions qui tiennent compte de l'interaction entre vecteurs textes, catégoriels et numériques.



(b) *Fuse-Early.* Transformer operates on learned embeddings for each feature.

Figure 3-12. D'après Shi et al. (2020, p.3)

iii. **Fuse Late**. Des traitements neuronaux (embedding) sont d'abord réalisés en amont avec chaque type de données, et l'agrégation ne se fait qu'à la sortie. Chaque type (modality) de données structurées est d'abord envoyé à des MLP, alors que le texte est envoyé à un Transformer pré-entraîné, puis les ultimes représentations vectorielles sont fusionnées (mean pooling, max pooling ou concaténation), puis l'ensemble est envoyé dans deux couches denses, pour la prédiction.



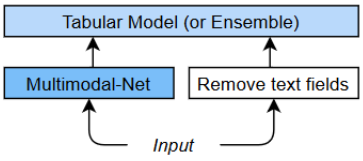
(c) *Fuse-Late.* Separate branches encode each modality, aggregate via mean/max/concat.

Figure 3-13. D'après Shi et al. (2020, p.3)

C'est ce troisième mécanisme de fusion qui sera ultérieurement utilisé dans AutoGluon Multimodal <TabularPredictor>

2- Processus d'agrégation :

Lorsque l'on encode le texte avec un Transformer, ce dernier est remplacé dans la table de données par autant de colonnes que la dimension de l'embedding. On peut envisager trois façons de vectoriser un texte via Transformer.



(a) *Embedding-as-Feature*

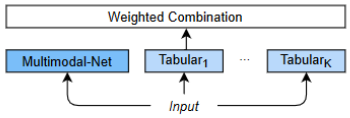
Figure 3-14. D'après Shi et al. (2020, p.4)

i. **Pré-Embedding**. On utilise un Transformer pré-entraîné et on l'entraîne sur nos jeux de données labélisés.
ii. **Text-Embedding**. Dans le pré-embedding, le Transformer est trop général et ne connaît pas la spécificité de nos données textuelles. Text-embedding commence par une fine-tuning du Transformer pré-entrainé sur le domaine spécifique au texte.
iii. **Multimodal-Embedding**. Dans l'embedding des données textuelles réalisées via Transformer, on applique l'auto-attention (Self Attention) en tenant compte des valeurs numériques et catégorielles comme information supplémentaire au contexte. Puis en utilisant les différents types de fusion multimodale obtenus dans la phase précédente on entraîne (fine tuning) le vecteur représentant l'ensemble des colonnes - quel que soit leur type - avec les données labélisées.
On retiendra cette architecture* pour *AutoGluon Tabular Multimodal

3- Processus Ensembliste

Le processus ensembliste est bien adapté aux données multimodales où les différents modèles ont été entraînés avec différentes modalités (types de données). Un possible inconvénient est que l'ensemble résultant peut être incapable d'exploiter les interactions prédictives (non-linéaires) entre les différentes modalités.

Weighted Ensemble - On pourrait juste choisir de prendre la moyenne pondérée des prédictions du Transformer et des autres modèles tabulaires, comme cela a été le cas avec AutoGluon Tabular. Les modèles sont entraînés de façon indépendante via un processus de partitionnement en apprentissage + validation. Ensuite on applique une « Forward Selection » sur chaque modèle pondéré à partir de l'échantillon test.

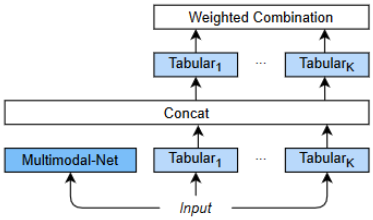


(b) *Weighted-Ensemble*

Figure 3-15. D'après Shi et al. (2020, p.4)

Stacked Ensemble – Il existe aussi la possibilité de procéder à des empilements au lieu d'utiliser des combinaisons linéaires. Cela permet à un autre modèle de ML d'apprendre la meilleure stratégie d'agrégation. Le modèle « empileur » utilise les prédictions de tous les autres modèles (y compris le Transformer), concaténé avec les variables tabulaires originales des données. On essaie ensuite chacun des modèles d'AutoGluon Tabular comme empileur. Puis on se sert d'un ensemble pondéré via sélection d'ensemble, appliquée aux empileurs tabulaires. Le Transformer n'est pas considéré comme empileur. Pour éviter l'overfitting, les modèles empileurs ne sont entraînés que sur les données exclus de l'entraînement des modèles individuels.

C'est cette dernière architecture qui est appliquée dans AutoGluon Tabular Multimodal



(c) *Stack-Ensemble*

Figure 3-15. D'après Shi et al. (2020, p.4)

AutoGluon Multimodal via TabularPredictor

1- Phase 1 : Aucun prétraitement des données

Sans appliquer un quelconque pré-traitement aux données cette architecture multimodale appliquée à nos données donne les résultats suivants (Tableau 3-7) :

Tableau 3-7 : Performance de l'architecture AutoGluon multimodal

	performances
Cohen's Kappa	0.593
PRC AUC	0.786
roc_auc	0.972
accuracy	0.956
balanced_accuracy	0.734
mcc	0.624
f1	0.614
precision	0.876
recall	0.473

Ici encore, le modèle ensembliste est celui qui a été sélectionné pour avoir démontré la meilleure performance sur la métrique du ROC AUC (Tableau 3-8) :

Tableau 3-8 : Leaderboard de l'architecture AutoGluon multimodal

model	score_test	score_val	fit_time	stack_level	fit_order
WeightedEnsemble_L2	0.972	0.978	5554.066	2	8
LightGBM	0.966	0.972	49.324	1	1
LightGBMXT	0.966	0.975	69.338	1	2
LightGBMLarge	0.965	0.973	71.797	1	6
CatBoost	0.962	0.973	1629.474	1	3
XGBoost	0.962	0.970	58.407	1	4
TextPredictor	0.959	0.963	3750.012	1	7
NeuralNetTorch	0.941	0.950	32.373	1	5

L'importance des variables du modèle multimodal sans prétraitement des données textuelles est donnée ci-après en utilisant LIME (Figure 3-16).

On notera tout particulièrement que dans ce mode de fusion, les données textes n'apparaissent pas du tout.

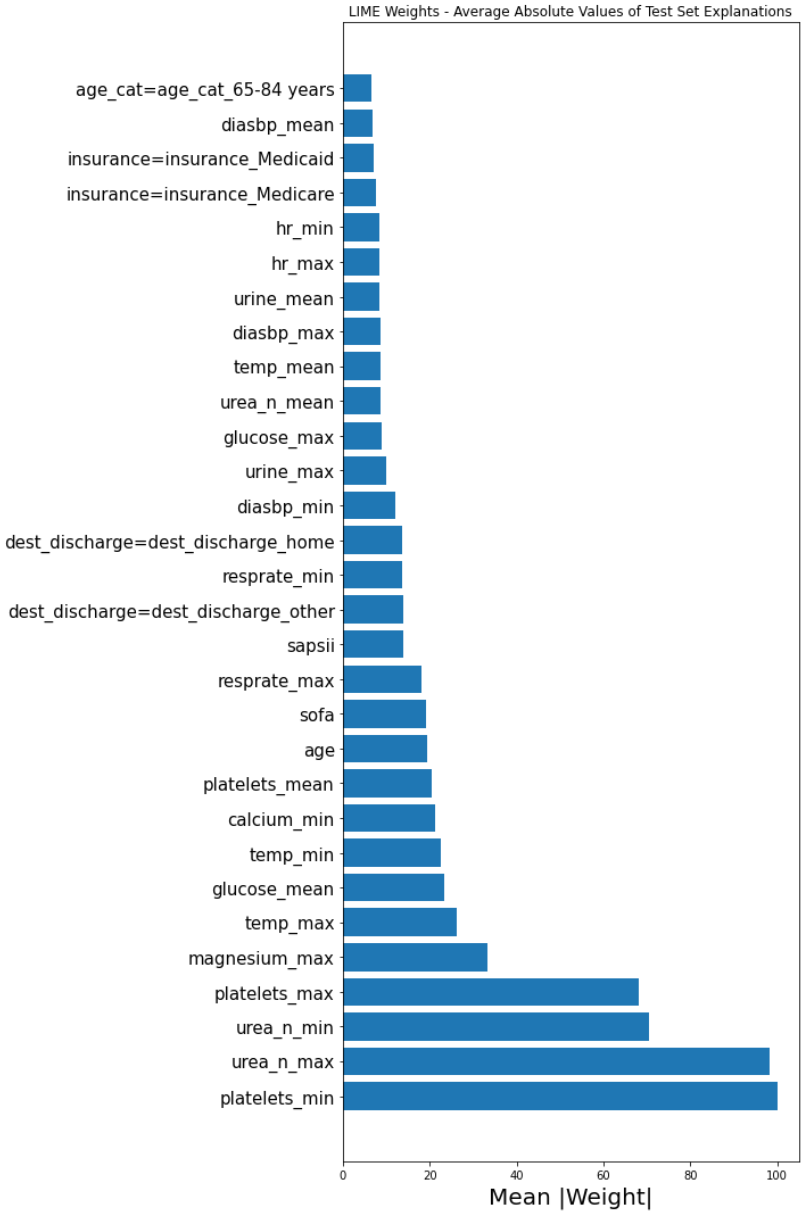


Figure 3-16 : Importance des Variables en utilisant LIME – Modèle Multimodal sans prétraitement

Or, lorsque l'on examine l'importance par permutation (Figure 3-17 ci-après), on note que la colonne texte joue un rôle très important dans la prédiction des séjours prolongés.

Dans ce cas précis cependant, on constate que l'apport du texte n'améliore par notablement la performance des données purement tabulaires (Tableau 3.3), même si le texte dans sa globalité semble avoir un poids significativement au-dessus des données tabulaires.

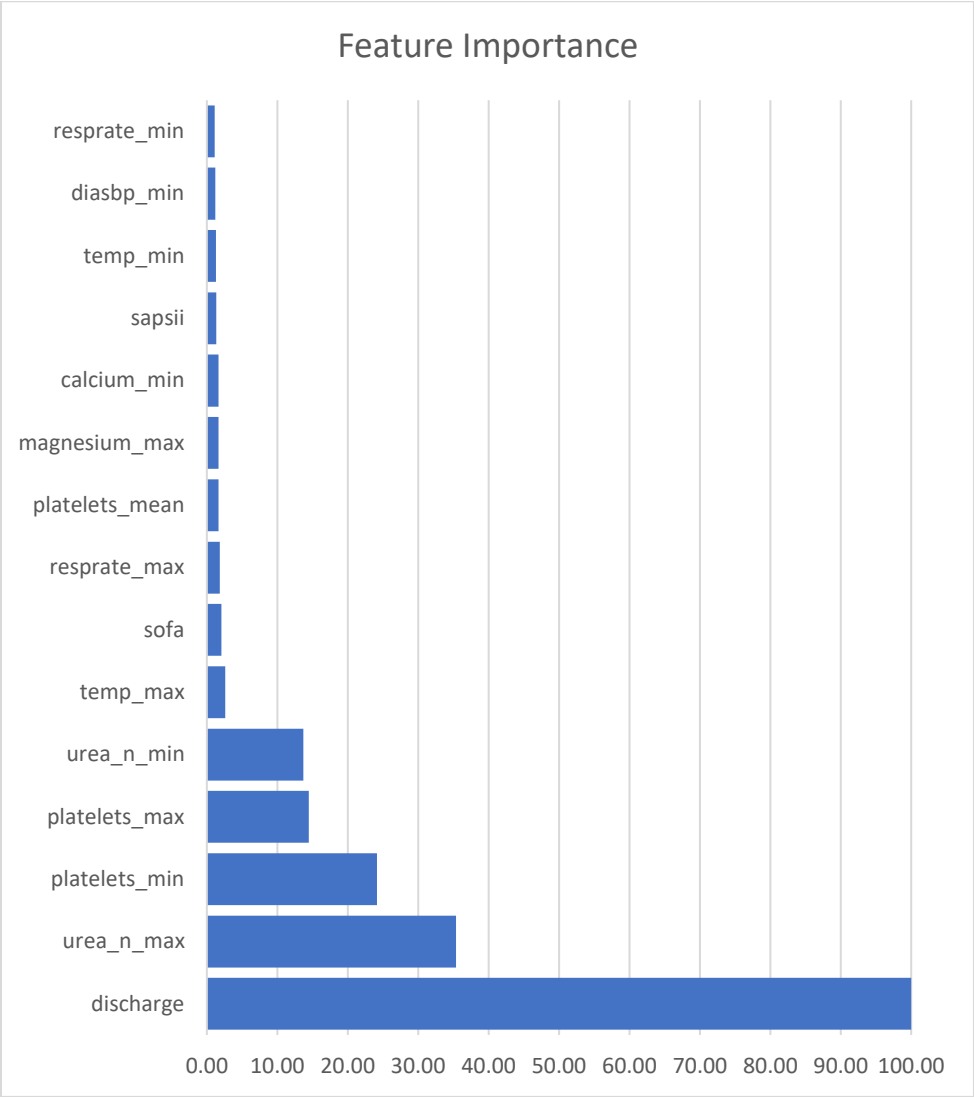


Figure 3-17 : Permutation Importance – Modèle Multimodal sans prétraitement

2- Phase 2 : Prétraitement « léger » des données textuelles

Le prétraitement léger ne change pas grand-chose à la performance…

Tableau 3-9 : Performance de l'architecture AutoGluon multimodal avec pré-traitement « léger »

	performances
Cohen's Kappa	0.601
PRC AUC	0.780
roc_auc	0.968
accuracy	0.956
balanced_accuracy	0.742
mcc	0.628
f1	0.623
precision	0.856
recall	0.490

Mais change le meilleur modèle qui cette fois-ci est LightGBM

Tableau 3-10 : Leaderboard de l'architecture AutoGluon multimodal avec prétraitement « léger »

model	score_test	score_val	fit_time	stack_level	fit_order
LightGBM	0.969	0.973	93.467	1	1
WeightedEnsemble_L2	0.968	0.975	332.009	2	8
LightGBMXT	0.966	0.974	92.777	1	2
LightGBMLarge	0.965	0.971	103.487	1	6
XGBoost	0.961	0.968	73.457	1	4
CatBoost	0.960	0.970	1514.348	1	3
NeuralNetTorch	0.941	0.948	41.038	1	5
TextPredictor	0.941	0.944	4234.040	1	7

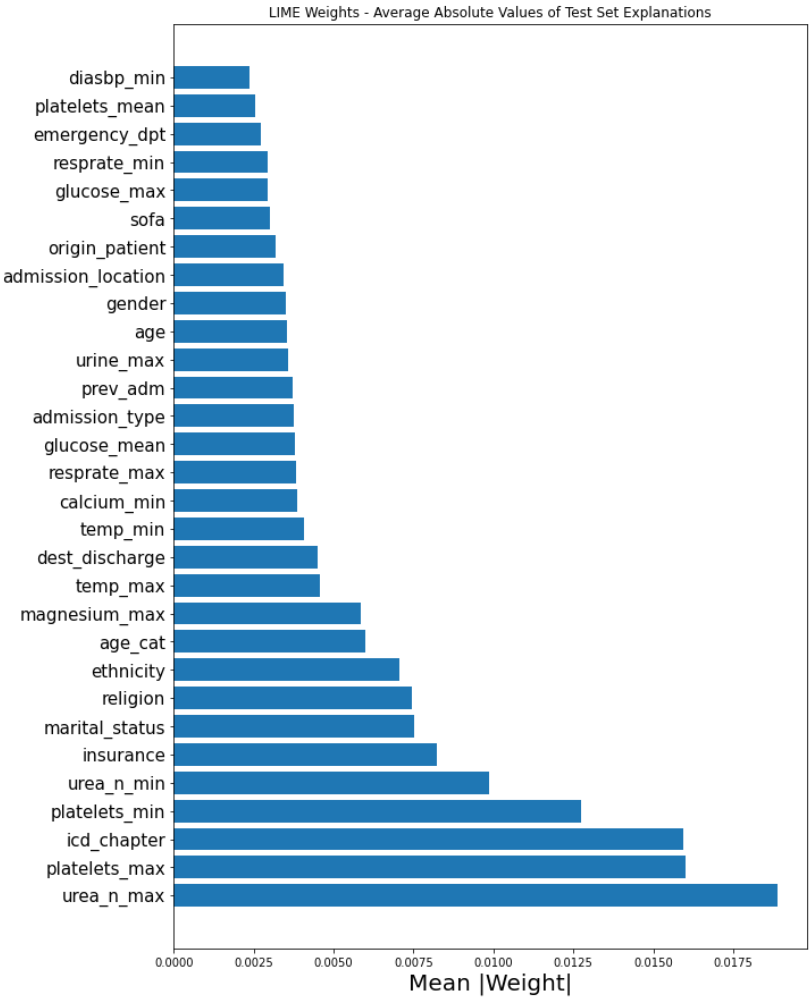


Figure 3-18 : Importance des Variables en utilisant LIME – Modèle Multimodal avec prétraitement « léger »

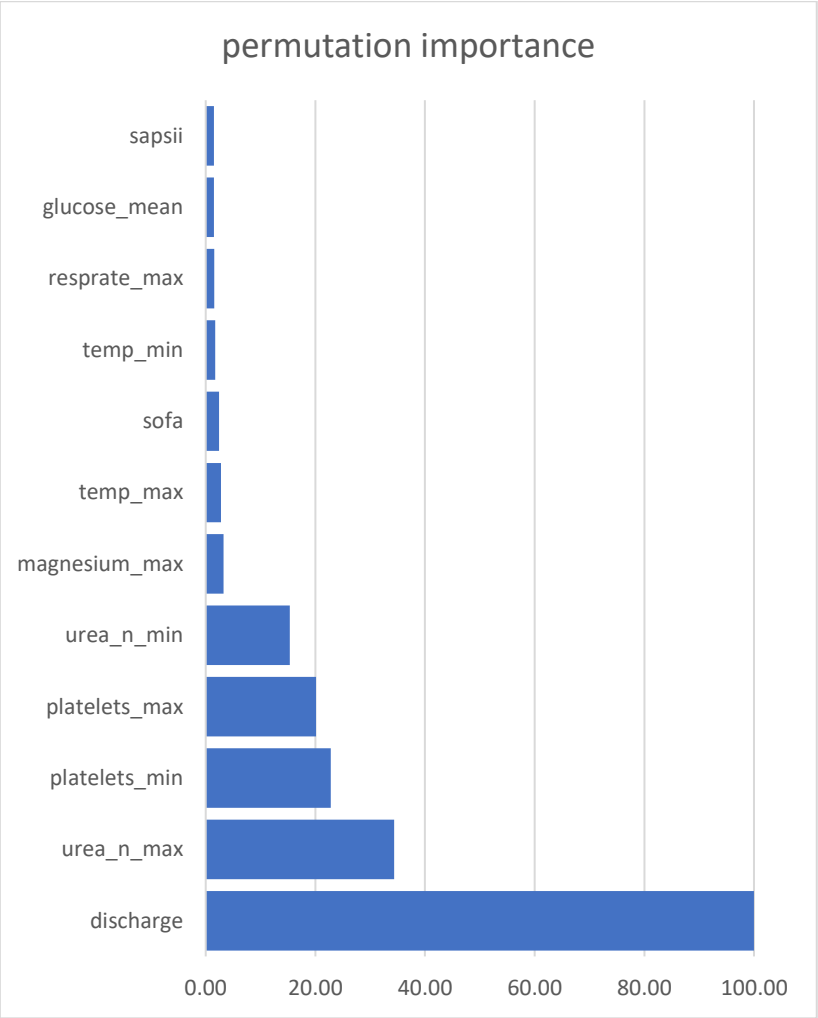


Fig 3-19 : Permutation Importance – Modèle Multimodal avec prétraitement « léger »

3- Phase 3 : Prétraitement « plus lourd » des données textuelles

Le prétraitement plus lourd ne change pas grand-chose à la performance…

Tableau 3-11 : Performance de l'architecture AutoGluon multimodal avec pré-traitement « plus lourd »

performance	stemming	lemmatization
Cohen's Kappa	0.636	0.616
PRC AUC	0.784	0.779
roc_auc	0.969	0.968
accuracy	0.959	0.957
balanced_accuracy	0.761	0.749
mcc	0.658	0.642
f1	0.657	0.637
precision	0.865	0.867
recall	0.529	0.503

Tableau 3-12 : Leaderboard de l'architecture AutoGluon multimodal avec prétraitement « plus lourd »

model	score_test	score_val	fit_time	stack_level	fit_order
WeightedEnsemble_L2	0.969	0.977	296.215	2	8
LightGBM	0.968	0.976	81.661	1	1
LightGBMLarge	0.967	0.973	88.035	1	6
LightGBMXT	0.966	0.975	84.739	1	2
XGBoost	0.964	0.968	66.540	1	4
CatBoost	0.962	0.970	2365.028	1	3
NeuralNetTorch	0.952	0.960	40.527	1	5
TextPredictor	0.943	0.943	4262.961	1	7

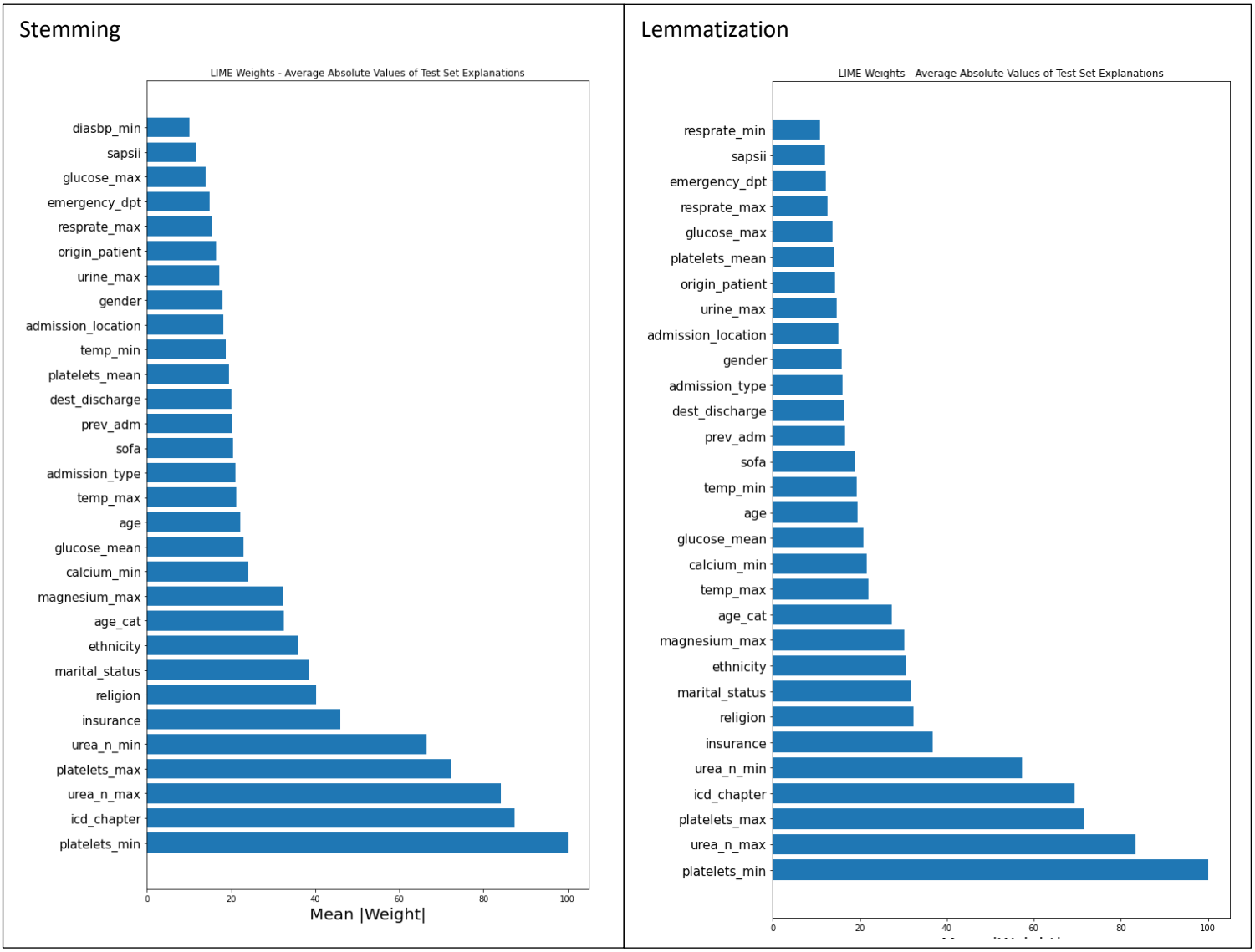


Figure 3-20 : Importance des Variables en utilisant LIME – Modèle Multimodal avec prétraitement « plus lourd »

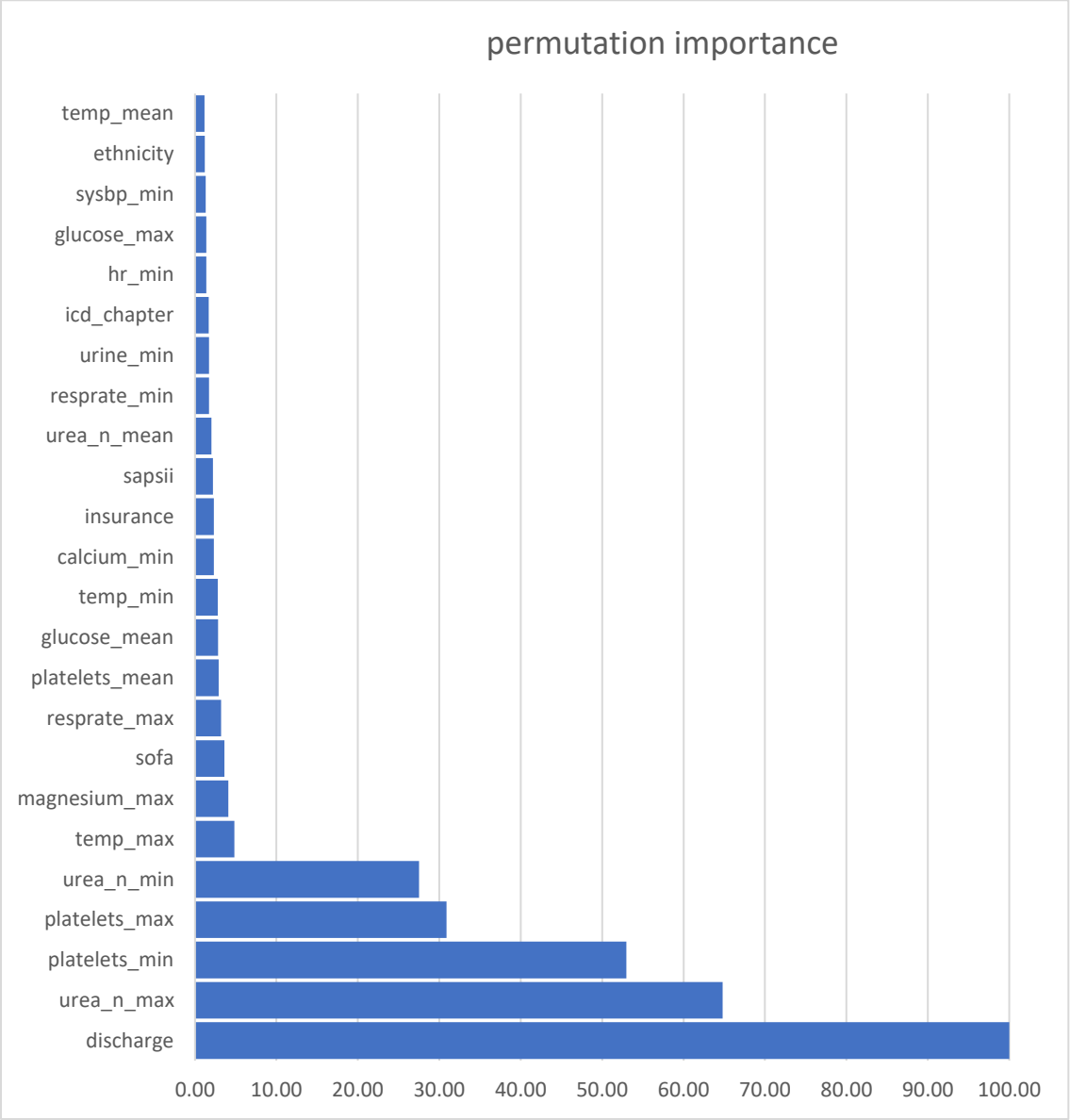


Figure 3.21 : Permutation Importance – Modèle Multimodal avec prétraitement « plus lourd (lémmatisation) »

Conclusion de la partie :

La fonction multimodale via <TabulaPredictor> d'AutoGluon, en traitant le texte comme une colonne unique (selon un mécanisme de fusion « Fuse Late ») met bien évidence que le texte est la variable la plus importante (selon la permutation feature importance). Autrement dit, dans leur globalité, les données textuelles constituent le meilleur prédicteur des séjours prolongés. Mais quand il s'agit de considérer chaque token des données textuelles comme variables individuelles (via LIME), alors les données structurées l'emportent et on ne voit apparaître aucun token des données textuelles.

AutoGluon TabularPredictor avec fusion LDA

Le principe consiste à vectoriser les données textuelles via BOW-LDA puis à les concaténer directement aux données structurées. Nous conservons le même processus de vectorisation que ce qui a été utilisé dans l'article #3, en examinant à la fois le cas où on utilise la racinisation d'un côté et la lemmatisation de l'autre.

1- Vectorisation TF

Les performances du modèle dont données ci-après (Tableau 3-13) :

Tableau 3-13 : Performance Fusion Données Structurées + LDA BOW TF

Stemming

	performances
Cohen's Kappa	0.605
PRC AUC	0.773
roc_auc	0.965
accuracy	0.955
balanced_accuracy	0.748
mcc	0.626
f1	0.627
precision	0.830
recall	0.503

Lemmatization

	performances
Cohen's Kappa	0.593
PRC AUC	0.770
roc_auc	0.966
accuracy	0.956
balanced_accuracy	0.732
mcc	0.628
f1	0.614
precision	0.889
recall	0.469

Comparés aux résultats des données tabulaires pures (PRC AUC = 0.789 – ROC AUC=0.964), il y a globalement une légère détérioration des résultats qui restent néanmoins excellents.

L'importance des variables apporte cependant des informations plus fines et plus précises comme l'indiquent les tableaux suivants (Tableau 3-14)

On constate une fois encore la domination des variables biologiques. Néanmoins il existe quelques thèmes dignes d'intérêts qui sont plus pertinents que d'autres variables biologiques ou même socio-démographiques (Tableau 3-15).

Racinisation

F16 : La thèmatique évoquée concerne les pathologies du rein (hémodyalise, dialyse, esdr : end stage renal disease, catheter, renal, sevelamer : un médicament utilisé pour traiter l'insuffisance rénale, fréquence de prise de médication comme tid ou daily, au moment du repas ou avec le repas)

F217 : La thématique renvoie d'un côté à la continuité des soins (vna : visiting nurse association, facil vna, vns service) et à la prise en charge à domicile (home, home service, discharge home) mais aussi au thème du service cardiothoracique (cardiothorac, service, allergies)

F41 : On trouve des termes associés à des auscultations de différentes partie du corps

F71 : Indique principalement une pathologie des intestins (crohn, colon, hernia) combiné à des examens cardiaques (qt, qtc, ekg)

Tableau 3-14 : Feature Importance AutoGluon Tabular & LDA BOW-TF (20 highest)

Stemming

	percent
urea_n_max	100.00
platelets_min	97.43
platelets_max	68.69
urea_n_min	38.25
platelets_mean	13.20
urine_max	11.87
F16	10.27
glucose_max	7.52
F217	6.85
sofa	6.45
urea_n_mean	6.00
F41	5.98
F71	5.73
magnesium_max	5.27
age	5.03
resprate_max	4.69
glucose_mean	4.53
urine_mean	4.27
urine_min	4.22
temp_max	4.20

Lemmatization

	percent
urea_n_max	100.00
platelets_max	91.45
platelets_min	76.69
urea_n_min	35.69
urine_max	24.61
platelets_mean	16.06
glucose_max	15.02
urine_mean	9.66
F180	8.77
age	7.25
sapsii	6.94
glucose_mean	6.75
urea_n_mean	6.60
resprate_max	6.58
sofa	6.23
diasbp_max	6.07
hr_max	5.59
hr_mean	5.57
urine_min	5.48
F162	5.35

Lemmatisation

F180 : Pathologie du colon potentiellement avec chirurgie et complications, et ou évacuation dans une poche externe

F162 : Pathologie du sang (leucémie) et/ou de la moëlle osseuse

Ainsi même s'il n'y a pas nécessairement de réel gain en performance, le recours aux données textuelles apporte beaucoup d'informations en termes d'interprétabilité. On notera également que les informations apportées par la lemmatisation et par la racinisation sont complémentaires, ce qui suggère de ne pas exclure un mode de prétraitement au profit d'un autre.

Tableau 3-15 : Thématique LDA BOW TF

Thème F16	**Thème F217**	**Thème F180**
hd	vna	collection
dialysi	home	fluid
hemodialysi	home servic	fluid collection
esrd	servic facil	drain
line	facil vna	drainage
cathet	disposit home	abdominal
tunnel	facil	ostomy
renal	dr	abscess
tid	discharg home	output
daili	vna servic	ileostomy
tablet	complet	ct
meal	md	fistula
wmeal	md complet	place
tid wmeal	cardiothorac	small
esrd hd	mr	leak
day meal	servic cardiothorac	catheter
sevelam	week	surgery
po tid	cardiothorac allergi	abdomen
wmeal time	appoint	site
cap	dr week	bowel

Thème F41	**Thème 71**	**Thème F162**
deni	crohn	bone
clear	prolong	marrow
tender	qt	bone marrow
bowel	crohn diseas	biopsy
gallop	qtc	cell
rub gallop	interv	platelet
wheez	block	multiple
well perfus	diseas	count
rub	av block	show
warm well	colon	myeloma
perfus	av	marrow biopsy
bowel sound	inguin hernia	transfusion
murmur rub	week	leukemia
sound present	ekg	neutropenia
oropharynx	inguin	multiple myeloma
normal s1	gi	fever
sclera anicter	hernia	thrombocytopenia
heent sclera	take	start
sclera	appoint	neutropenic
club	conduct	disease

2- Vectorisation binaire

La performance du meilleur modèle sous vectorisation binaire est donnée ci-après (Tableau 3-16)

Tableau 3-16 : Performance Fusion Données Structurées + LDA BOW BIN

Stemming

	performances
Cohen's Kappa	0.644
PRC AUC	0.800
roc_auc	0.970
accuracy	0.959
balanced_accuracy	0.766
mcc	0.665
f1	0.665
precision	0.865
recall	0.540

Lemmatization

	performances
Cohen's Kappa	0.593
PRC AUC	0.770
roc_auc	0.966
accuracy	0.956
balanced_accuracy	0.732
mcc	0.628
f1	0.614
precision	0.889
recall	0.469

La fusion des données apporte une légère amélioration de la performance pour le stemming et une perte de performance pour la lemmatisation, mais les résultats restent excellents.

L'importance des variables permettent d'examiner l'interprétabilité des résultats (Tableaux 3-17 et 3-18)

Racinisation

F217 ne semble rien indiquer de particulièrement différencié. On y trouve pêle-mêle des tokens de dosage médicamenteux, puis deux conditions : embolie pulmonaire et paralysie. En examinant plus loin le vecteur (au-delà des 20 premières variables les plus importantes), on y trouve des termes associés à la vieillesse (voire au décès) et des traitements tels que lorazepam, clonazepam, dilantin. On reconnaît ici certains des risques associés à un taux de nitrogène uréique trop élevé ou un taux de plaquettes trop bas, ou possiblement des embolies pulmonaires conséquentes aux anesthésies.

Lemmatisation

F188 renvoie à la thématique des maladies infectieuses et les traitements y afférant.

F143 renvoie à des conditions respiratoires qui peuvent être sévères mais aussi à des procédures : intubation, extubation, sédation.

Tableau 3-17 : Feature Importance AutoGluon Tabular & LDA BOW-BIN

Stemming

	percent
urea_n_max	100.00
platelets_min	85.65
platelets_max	44.61
urea_n_min	37.55
F217	18.33
platelets_mean	6.76
resprate_max	5.65
sofa	5.46
magnesium_max	5.20
temp_max	5.15
temp_min	3.39
age	3.13
sapsii	2.98
glucose_mean	2.91
calcium_min	2.68
hr_max	2.68
resprate_min	2.67
urine_min	2.53
temp_mean	2.51
glucose_max	2.44

Lemmatization

	percent
urea_n_max	100.00
platelets_min	98.46
platelets_max	96.40
urea_n_min	32.94
urine_max	22.79
platelets_mean	15.63
glucose_max	13.01
F188	12.68
urine_mean	9.98
F143	9.01
age	8.04
urea_n_mean	6.84
glucose_mean	6.63
magnesium_max	6.49
sapsii	5.96
urine_min	5.95
diasbp_min	5.64
diasbp_max	5.59
hr_mean	5.57
sofa	5.42

Tableau 3-18 : Thématique LDA BOW BIN

Thème F217	Thème F188	Thème F143
10mg po	culture	intubate
100mg po	vancomycin	sedate
100mg	infectious	intubation
10mg	infectious disease	extubate
25mg po	feed	fio2
20mg po	tube feed	peep
embolus	antibiotic	intubate sedate
pulmonari embolus	nutrition	patient intubate
25mg	patient continue	endotracheal
patient made	grow	tube
made	disease	intubated
400mg	blood culture	extubation
palsi	negative	sedation
1000mg	discontinue	ac
secur	fluid	vent
23 day	improve	propofol
formal	place	respiratory failure
refus	urine	airway
spoke	infection	endotracheal tube
effort	hematocrit	respiratory

3- Vectorisation TFIDF

La performance du meilleur modèle utilisant la vectorisation TFIDF est donnée ci-après (Tableau 3-19)

Tableau 3-19 : Performance Fusion Données Structurées + LDA BOW TFIDF

Stemming

	performances
Cohen's Kappa	0.657
PRC AUC	0.783
roc_auc	0.965
accuracy	0.960
balanced_accuracy	0.776
mcc	0.674
f1	0.677
precision	0.856
recall	0.560

Lemmatization

	performances
Cohen's Kappa	0.629
PRC AUC	0.769
roc_auc	0.965
Accuracy	0.958
balanced_accuracy	0.759
Mcc	0.649
f1	0.650
Precision	0.850
Recall	0.526

Comparées aux données tabulaires pures, la fusion des données avec vectorisation TFIDF est légèrement inférieure en termes de performance, mais les résultats restent excellents. Examinons à présent l'éventuel gain en interprétabilité via l'importance des variables (Tableau 3-20).

Tableau 3-20 : Feature Importance AutoGluon Tabular & LDA BOW-TFIDF

Stemming

	percent
urea_n_max	100.00
platelets_min	82.82
platelets_max	37.53
urea_n_min	35.50
platelets_mean	8.32
icd_chapter	7.75
magnesium_max	5.83
temp_max	4.81
sofa	4.79
resprate_max	4.60
calcium_min	3.73
temp_min	3.52
age	2.97
dest_discharge	2.85
resprate_min	2.73
diasbp_min	2.63
resprate_mean	2.59
glucose_max	2.53
glucose_min	2.47
glucose_mean	2.37

Lemmatization

	percent
platelets_min	100.00
urea_n_max	95.44
platelets_max	56.05
urea_n_min	34.01
platelets_mean	8.64
magnesium_max	6.42
F266	6.41
urine_max	5.24
age	4.98
temp_max	4.91
resprate_max	4.80
glucose_max	4.32
temp_min	4.22
F38	4.22
sofa	4.08
F168	3.87
glucose_mean	3.75
calcium_min	3.17
diasbp_min	3.15
resprate_min	3.00

Racinisation

Avec le stemming, on observe qu'aucune des thématiques LDA ne ressortent dans les 20 variables les plus importantes.

Lemmatisation

En ce qui concerne la lemmatisation, trois thématiques ressortent très clairement :

F266 renvoie à la chirurgie cardiaque et notamment au pontage coronarien.

F38 renvoie visiblement également à des pathologies cardiovasculaires et aux traitements associés.

F168 est une thématique essentiellement de procédures de soins. Si on cherche plus loin dans le vecteur, des thématiques relatives à l'insuffisance respiratoire ou rénale émergent.

Tableau 3-21 : Thématique LDA BOW-TFIDF

Thème F266	**Thème F38**	**Thème F168**
tablet	tablet	tablet
daily	cardiac	sig
sig	ventricular	daily
aortic	cath	sig one
valve	daily	pt
mg tablet	rca	mg tablet
sig one	artery	tablet po
tablet sig	stenosis	tablet sig
coronary	coronary	right
incision	stent	one tablet
artery	catheterization	give
tablet po	lad	need
coronary artery	plavix	continue
one tablet	aortic	start
please	left ventricular	po daily
refills0	chest pain	mg po
mitral	systolic	ct
release	valve	note
bypass	mitral	urine
leave	regurgitation	sp

Mise en Perspective des Résultats

Dans cette partie de la thèse, nous cherchons à prédire et à expliquer le LOS en utilisant successivement des données structurées, des données textuelles et une fusion des deux.

Pour l'aspect technique, cette étude peut être mise en relief en la comparant à une des rares autres étude sur le MIMIC fusionnant données textuelles (notes cliniques), données temporelles (biologie et signes vitaux) avec des données statiques (socio-démographiques et hospitalières) (Zhang et al., 2020).

Les notes cliniques choisies sont différentes des nôtres, et d'autres part dans notre cas les données temporelles ont été simplement résumées en min, max et moyennes. Ceci mis à part, les structures des données sont comparables.

La différence majeure entre les deux approches se situe plutôt dans les méthodes de fusion de données et dans le choix des classifieurs. En ce qui concerne les méthodes classiques, Zhang et al. (2020) utilisent la régression logistique et Random Forest alors que notre étude couvre une large gamme de modèles d'arbres et de réseaux de neurones. S'agissant des modèles pour traiter les textes de façon séquentielle, Zhang et al. (2020) utilisent des réseaux de neurones de convolution (CNN) et des réseaux de neurones récurrents (LSTM) alors que nous utilisons des Transformers.

En ce qui concerne les variables d'intérêts, les auteurs examinent la Mortalité Hospitalière (MH), la réadmission sous 30 jours non planifiée (RH30NP) et la durée prolongée au-delà de 7 jours (LOS7). Dans notre cas, les variables d'intérêts sont tous les réadmissions sous 30 jours (RH30) et la durée de séjour prolongée au-delà de 28 jours (LOS28). Dans les deux études les distributions des variables comparables sont respectivement données par (pourcentage de positifs vs pourcentage de négatifs) :

Zhang et al. (2020) : RH30NP (5.7% vs 94.3%) – notre étude : RH30 (5.7% vs 94.3%)

Zhang et al. (2020) : LOS7 (49.9% vs 50.1%) – notre étude : LOS28 (7.4% vs 92.6%)

En termes de performances, pour le LOS, le meilleur modèle de Zhang et al (2020) atteint ROC AUC=0.787 et PRC AUC=0.666. Nos meilleurs modèles atteignent ROC AUC=0.964 et PRC AUC=0.789 pour le tabulaire pur, et pour la fusion multimodale ROC AUC=0.968 et PRC AUC=0.780. Ces résultats sont particulièrement bons si on tient compte du déséquilibre du LOS28 comparé au LOS7.

Pour la réadmission à 30 jours, le meilleur modèle de Zhang et al (2020) atteint ROC AUC=0.676 et PRC AUC = 0.084. Notre meilleur modèle sur le texte seul atteint ROC AUC = 0.723 (Article #3)

En conséquence, en termes de fusion tout autant qu'en termes de prédiction, nos modèles sont performants.

Il existe d'autres travaux qui approchent le LOS en utilisant des données non structurées textuelles seules sans chercher à faire une fusion. L'une d'entre elle (Chrusciel et al., 2021) utilise une approche originale consistant d'abord à extraire les concepts médicaux (Unified Medical Language System – UMLS) à partir du texte libre et à utiliser ensuite les 969 concepts ainsi extraits comme les catégories d'une variable catégorielle. De cette façon en passant pas une disjonction complète (one hot encoding) on obtient un jeu de données structurées classique à large dimension traité avec Random Forest. L'importance des variables est également fournie par les auteurs en comparant d'un côté un modèle avec les données socio-démographiques et autres données structurées avec les données démographiques et données non structurées. Ces données ne se comparent pas vraiment avec nos travaux dans la mesure où les auteurs (Chrusciel et al., 2021) s'intéressent au LOS7, ce qui donne des données plutôt équilibrées et justifie l'utilisation du F1 comme principal indicateur de performance.

Le F1 pour des données déséquilibrées est moins fiable que le ROC AUC ou le ROC PRC comme nous l'avons déjà évoqué en amont.

D'autre part, il y a un avantage évident à rechercher des informations textuelles plus précises sur les déterminants des risques, c'est-à-dire à viser plus d'interprétabilité. C'est probablement la principale limite (que les auteurs reconnaissent d'ailleurs) de cette approche et que nous avons tenté de résoudre de différentes manières dans nos travaux, à la fois sur les données non structurées seules et les données mixtes fusionnées.

Une autre étude, essentiellement exploratoire utilise des textes cliniques de médecins pour prédire LOS > 2 jours et la destination du patient (Bacchi, Gluck, et al., 2020). Cette étude utilise les données textuelles de deux façons différentes : la première par une vectorisation BOW sur les fréquences absolues (pondération TF), puis compare Régression Logistique (LR), Random Forest (RF), et Réseaux de Neurones Denses (ANN) ; la seconde par une vectorisation séquentielle qui mobilise ensuite des Réseaux de Neurones de Convolution (CNN). Le meilleur résultat est obtenu par l'ANN à 6 couches cachées avec un AUC = 0.75, mais il faut rappeler que la taille d'échantillon est de 313 patients. On rappellera que notre performance pour les textes seuls atteint un ROC AUC de 0.859 avec un PRC AUC de 0.413.

Comparé à la plupart des autres études sur le LOS avec des données structurées Bacchi, Gluck, et al. (2020) proposent une interprétabilité basée sur les coefficients des tokens pour la régression logistique qui est en quelque sorte utilisée comme modèle du substitution – mais on ne sait pas à quel point ces tokens se retrouvent (restent valides) effectivement dans les autres modèles et notamment dans le meilleur modèle.

Comme dans la plupart des études qui intègrent les données textuelles dans les prédictions en sciences de la santé, Weissman et al (2018) soulignent que parmi les facteurs qui limitent la possibilité de réduire les risques de mortalité hospitalière ou de séjours prolongés se trouve la non prise en compte des données textuelles, l'utilisation quasi exclusive de la régression logistique et la non reproductibilité des études, notamment par la non disponibilité de données communes.

En conséquence, leur étude s'intéresse à la mortalité hospitalière et aux séjours prolongés > 7 jours (mais également une variable composite mortalité hospitalière ou LOS > 21 jours). Les données mobilisées utilisent uniquement des données structurées puis une fusion des données structurées et non structurées. Les notes cliniques choisies sont également issues du MIMIC III et incluent toutes les notes et non exclusivement les notes de décharges (comme dans notre cas). Les données textuelles sont ensuite vectorisées en BOW-TF après une présélection ad-hoc avec d'un côté des termes a priori et de l'autre les 500 termes les plus prédictifs par sélection via Lasso. La fréquence totale de chaque ligne de données textuelles est également incluse comme variable. Les modèles utilisés sont la Régression Logistique (LR), la Régression Logistique Pénalisée Elasticnet (EN), Random Forest (RF) et Gradient Boosting (GB)

En termes de performance, leurs meilleurs résultats sont obtenus par GB (ROC AUC=0.89), ce qui est meilleur que nos résultats des modèles purement textuels mais bien inférieurs à nos modèles tabulaires ou mixtes. Cette recherche (Weissman et al., 2018) décline aussi l'importance des variables dont le mode de calcul peut-être objet de débats, mais qui a le mérite d'indiquer les 10 prédicteurs considérés comme les plus important. Ainsi, pour les données mixtes, on voit une plus grande explicabilité émerger avec des tokens tels que « settings », « vent changes », « extubated », « poor prognosis » etc. comme prédicteurs de risques.

Pour conclure, nous pouvons estimer que la performance de nos modèles, comme leur explicabilité semblent aller bien au-delà de ces différents travaux recensés.

Partie 4 – Discussion et Conclusion

Synthèse et perspectives globales

Concernant les aspects techniques

Dans le fond, ce parcours de recherche s'est intéressé à quelques points saillants qu'il convient maintenant de souligner :

1- Pertinence du Machine Learning (ML) en santé publique

Certaines études affirment qu'il n'y aurait pas de gain de performance des modèles de ML par rapport aux modèles de régression logistique (Christodoulou et al., 2019) alors que d'autres affirment le contraire (Austin et al., 2021, 2022). L'ensemble des études que nous avons menées vont dans le sens de différences notables voire significatives entre les modèles classiques (régression logistique) et les modèles d'arbres boostés ainsi que le Random Forest. La différence entre les modèles de Régression Logistique (LR) et ceux qui sont pénalisés (Lasso, Ridge, Elasticnet) n'a jamais été ni notable ni significative. Et les arbres de décision ont toujours été les moins performants.

Un des avantages des modèles de ML vis-à-vis des modèles linéaires est leur capacité à tenir compte de relations non linéaires avec la variable explicative, et d'interactions entre les prédicteurs, sans que celles-ci n'aient à être explicitées. Autrement dit, les modèles de ML tiennent plus facilement compte de la complexité de l'information dans les données.

Enfin, certains auteurs insistent sur l'importance d'avoir assez de données pour les modèles de ML (Riley et al., 2020) qui parce qu'ils ont souvent plus de paramètres à estimer devraient contenir plus d'observations (une des règles informelles appliquées est celle des 10 observations par paramètre).

En conséquence, nous répondons par l'affirmative. De façon générale, les modèles de Machine Learning, et plus précisément les ensembles d'arbres sont recommandés en sciences de la santé, idéalement avec des tailles d'échantillons suffisantes – ce qui est généralement le cas en santé publique, du moins dans les cas que nous traitons dans cette thèse.

2- Boîtes noires ?

La question du « Black Box Problem » a été traitée de façon extensive par certains auteurs (Guidotti et al., 2018) et est à l'origine de tout un champ de recherche dit « explainable AI » (Linardatos et al., 2021). Certes, les modèles de ML et de Deep Learning (DL) ne peuvent pas être expliqués d'une façon aussi « mécanique » que les modèles linéaires mais les méthodes permettant d'examiner des prédictions au niveau local et global, non seulement existent mais continuent à se développer (ElShawi et al., 2021; Lundberg et al., 2020). On doit pourtant se rendre à l'évidence que dans nos travaux sur la qualité des soins, que ce soit dans la réadmission ou la durée de séjour, les recherches mobilisant les modèles de ML et de DL ne saisissent pas systématiquement l'opportunité de clarifier les déterminants majeurs de risques, par exemple et au minimum, en mobilisant l'importance des variables. Même les revues de littérature systématique sur ces sujets (Bacchi, Tan, et al., 2020; Y. Huang et al., 2021; Lequertier et al., 2021; Mahmoudi et al., 2020) n'abordent pas vraiment la question de l'interprétabilité qui est pourtant un élément central de la modélisation statistique classique. Comme si la performance prenait maintenant complètement le pas sur l'explication.

3- Le compromis performance-explicabilité

Une question centrale se pose, notamment dans le cas de la qualité des soins, à savoir s'il faut absolument améliorer la performance des modèles pour que ceux-ci puissent par exemple – une fois déployés dans le système d'information de l'hôpital – alerter en amont ou en temps réel les

soignants quant à d'éventuels séjours ou patients à risques. Ou faut-il plutôt accepter de perdre en performance pour avoir plus d'explicabilité.

Il ne semble pas vraiment y avoir de raccourci à cette question et il se peut ce que ces deux facettes soient complémentaires. Par exemple dans cette thèse, nous avons vu que le choix de privilégier les données textuelles ont permis d'avoir beaucoup plus de richesse en termes d'explicabilité, notamment dans la vectorisation en BOW et en LDA (Article #3), mais la performance tout en restant satisfaisante est plus modeste. De la même façon dans la partie 4 de cette thèse, nous avons bien noté que les modèles avec textes seuls sont significativement moins performants que les modèles mixtes ou tabulaires, bien qu'ils restent encore très bons. ***Cependant, lorsque les modèles tabulaires sont performants, alors au minimum, l'inclusion de vecteurs de textes thématiques comme la LDA peut significativement en améliorer l'interprétabilité sans renoncer à la performance.***

4- L'enjeu des données déséquilibrées

Concernant les données déséquilibrées, voire très déséquilibrées comme dans nos études, certains auteurs et praticiens recommandent différentes méthodes de rééchantillonnage de l'échantillon d'apprentissage (KaurHarsurinder et al., 2019). Notre propre expérimentation sur le sujet ne nous a pourtant pas convaincu, de même que les différents projets donnés aux étudiants.

Nous défendons ici l'idée que pour des données déséquilibrées – comme c'est le cas dans toute cette thèse – les métriques faisant une hypothèse de distribution équilibrée a priori ne sont pas fiables. En conséquence ce manque de fiabilité est transféré dans la sensibilité, la spécificité, la précision et le recall. Nous ne sommes pas seuls à défendre cette idée (Bradley, 1997) et plusieurs discussions très intéressantes développent certains des arguments[23,24] . En conséquence, nous privilégions le ROC AUC et le PRC AUC. Mais ces derniers ont également été l'objet de récentes critiques complétées par des propositions d'alternatives (Carrington et al., 2022), aussi la question semble rester ouverte.

5- Partitionnements et rééchantillonnages

Si la partition en échantillon test vs échantillon d'apprentissage est maintenant couramment adoptée dans l'utilisation de ML en sciences de la santé, il est encore extrêmement rare de trouver des comparaisons de performance basées sur des tests statistiques, l'exception étant Zhang et al. (2020). Or le fait de réaliser plusieurs tirages aléatoires successifs de l'échantillon test vs échantillon de validation permet de réaliser ensuite un test statistique de comparaison de moyennes. L'échantillon test qui a été mis de côté peut ensuite être utilisé comme un jeu de données entièrement inconnu des modèles. Nous noterons cependant que cette approche n'a été utilisée que dans nos deux premiers articles, mais pas dans la suite, essentiellement pour une question de temps et de ressources.

6- Performance selon les données et fusion des données

Données tabulaires

On notera la très grande différence de performance entre les résultats obtenus avec les données de l'APHM et celles du MIMIC III. Certes les variables d'intérêts sont légèrement différentes dans leur définition et plus encore, contrairement aux données APHM, le MIMIC III correspond principalement

[23] https://www.fharrell.com/post/class-damage/

[24] https://stats.stackexchange.com/questions/312780/why-is-accuracy-not-the-best-measure-for-assessing-classification-models

à des séjours en soins intensifs. Malgré tout, on peut bien voir que ce qui a surtout fait la différence c'est l'inclusion de variables biologiques (analyses et signes vitaux). Ce sont d'ailleurs principalement les taux de nitrogène uréiques et le taux de plaquettes qui prédisent le plus fortement la durée de séjour prolongée.

Données textuelles

Elles ont été traitées de deux façons.

La première utilise 5 formes de vectorisation tabulaires : 3 Bag Of Words (BOW) respectivement avec pondération Binaire (BIN), Terms Frequency (TF) et Terms Frequency Inverse Document Frequency (TFIDF), puis 2 Topic Modeling Latent Semantic Analysis (LSA) et Latent Dirichlet Allocation (LDA).

La deuxième utilise une vectorisation séquentielle avec un modèle préentrainé Bio+Clinical BERT, complété par une fine-tuning sur les données MIMIC.

Données mixtes fusionnées

Elles ont été traitées de deux façons.

- En utilisant l'architecture multimodale TabularPredictor d'AutoGluon. Ici les données tabulaires sont primaires et la vectorisation (embedding) des textes via Transformer est concaténée dans une couche d'empilement supérieure, puis le tout est agrégé par processus ensembliste pondéré.
- En fusionnant les données tabulaires avec une vectorisation BOW-LDA comme dans l'article #3. Cela revient à une pure concaténation des données tabulaires avec la représentation à 300 dimensions (colonnes) en LDA du corpus de texte.

7- Les principales conclusions que l'on tire :

- Les données tabulaires sont plus performantes que les données textuelles seules.
- Les données textuelles seules sont quand même performantes tout en apportant une plus grande richesse en termes d'interprétabilité.
- Les données mixtes sont aussi performantes que les données textuelles ou tabulaires seules mais apportent une plus grande richesse à l'interprétabilité.
- La fusion des données tabulaires avec la vectorisation LDA est celle qui apporte le meilleur compromis performance-interprétabilité.

On notera également que les modèles de textes purs ou mixtes traitant les textes avec des Transformers sont significativement plus performants que tous les autres modèles équivalents de textes purs ou mixtes consultés dans la littérature.

Concernant l'aspect santé publique

1- La réadmission à 30 jours (RH30)

Est un indicateur de la qualité et de la continuité des soins entre la période d'hospitalisation et la période qui s'en suit, mais aussi un indicateur de la coordination entre la ville et l'hôpital (ou les soins aigus et les soins ambulatoires)(DGOS & ATIH, 2022, p. 30). Pour les séjours de l'Assistance Publique, Hôpitaux de Marseille, les variables qui prédisent le mieux la RH30 (via les urgences) sont :

- L'existence d'au moins une hospitalisation via les urgences lors des 6 derniers mois ;
- La catégorie de maladie (diagnostic associé au groupe homogène de malades : GHM) ;
- Une hospitalisation via le service des urgences ;
- La durée de séjour (LOS) ;
- L'âge ;
- Le type de séjour (Médical, Chirurgical ou Obstétrique).

Pour les séjours du Beth Israel Deaconess Medical Center (BIDMC), les tokens les plus prédictifs de RH30 (en soins intensifs) sont :

- le type de soin (attention et médication fréquente),
- les pathologies (trachéotomie, insuffisance rénale, respiratoire, cardiaque),
- le type d'hospitalisation et l'âge (hospice, soins palliatifs, crash automobile),
- la persistance des conditions (dialyse, transplantation d'organes, chronicité),
- les conditions spécifiques de fin de vie (ne pas réanimer, soins palliatifs, phase finale d'une pathologie rénale),
- les médicaments (dilantin, vancomycine, carvedilol, digoxyn, miconazole, rifaximin, etc.),
- types de traitements (rénal, respiratoire, cardiaque, infectieux, gastriques, sanguins, etc.)

Une des façons d'utiliser les modèles de Machine Learning est de déployer les modèles dans le système d'information hospitalier. Ainsi pour l'Assistance Publique, 5 des 6 variables peuvent être déterminées à l'admission et permettent déjà de prédire la probabilité de RH30 ; quant à la dernière variable, elle pourrait être connue à la sortie (moment de la décharge). Il suffirait donc que le modèle (ici de Random Forest) puisse tourner à chaque admission et à chaque décharge.

Pour la BIDMC, comme pour les séjours en soins intensifs en général, bien que nous ne l'ayons pas fait ici, on aurait parfaitement pu obtenir des variables structurées en déroulant la même routine que celle de l'APHM, mais de plus, on peut aussi traiter les notes cliniques obtenues auprès des soignants, et en fonction des tokens contenus dans ces notes (de décharge par exemple), le risque de réhospitalisation est évalué.

Dès que ces patients sont identifiés à l'admission, les différentes interventions peuvent être mises en œuvre, puis poursuivies (O. Hansen et al., 2011). Pendant le séjour et en préparation de la période post-hospitalière, le patient est sensibilisé et éduqué, la sortie est soigneusement préparée et planifiée, on s'assure que le patient n'est pas en conflit avec ses traitements médicamenteux et qu'il en assume la responsabilité ou s'organise pour être en mesure de les suivre, et surtout on programme des rendez-vous de suivis réguliers.

Au moment de la décharge, les interventions consistent à mettre en œuvre des instructions adaptées centrées sur le patient, donc ajustées à sa nature, à son environnement et à son style de vie ; on peut avoir recours à des coachs de transition internes à l'hôpital ou bien indépendants travaillant avec l'hôpital et/ou le patient. Cette personne pourrait d'ailleurs être une infirmière libérale qui pourrait ensuite assurer la continuité de soins.

Après la décharge devraient s'en suivre une visite régulière du médecin traitant à domicile, un appel régulier du secrétariat médical de l'hôpital, une relation régulière avec les services hospitaliers de jour, et à l'ère du numérique pourquoi pas une hotline audio, visio ou au strict minimum par mail qui évalue la pertinence de faire intervenir un tiers adapté ?

2- La durée de séjour (LOS)

Est aussi un indicateur de la qualité des soins mesurant plus précisément l'efficacité et/ou l'efficience. Plus précisément les séjours prolongés sont un indicateur de non-qualité. Un séjour peut être prolongé à cause d'une mauvaise qualité de prise en charge du moment du diagnostic jusqu'au moment de la décharge (diagnostics inappropriés ou hâtifs, impossibilité de délivrer les traitements pour des raisons humaines ou logistiques, nécessité d'avoir recours à des traitements complémentaires ou supplémentaires, disponibilité des lits et du personnel).

Les séjours trop courts peuvent aussi être un indicateur de non-qualité, notamment parce qu'une décharge prématurée est cause de réhospitalisation.

Dans cette thèse, nous nous intéressons surtout aux séjours prolongés donc nous avons opté pour une binarisation de la variable LOS. C'est un choix qui est tout à fait courant dans la littérature, cependant le seuil de binarisation peut être un objet de débats. Certains auteurs ont par exemple fait le choix de binariser la variable LOS de façon à obtenir une distribution quasiment équilibrée, ce qui revient à couper au niveau de la médiane. Ce sera généralement le cas de ceux qui choisissent un seuil de 7 jours (Marfil-Garza et al., 2018b). D'un point de vue purement statistiques, ce choix justifie l'utilisation de métriques telles que l'accuracy, la sensibilité, la spécificité. Cependant d'un point de vue qualitatif, si on reste dans la logique de prédire des séjours prolongés, cela suppose que la moitié des séjours seraient des séjours prolongés ou en exagérant un peu, la moitié des séjours seraient à risques, autrement dit la qualité des soins en termes d'efficacité et/ou d'efficience est mauvaise. On voit bien que le choix de ce seuil de binarisation n'est pas du tout innocent ni sans conséquences car selon la définition que l'on retient, les facteurs (prédicteurs) de risques ne seront peut-être pas du tout les mêmes. Nous proposons l'idée que le choix du seuil de discrétisation reflète implicitement les croyances que l'on peut avoir sur le système de soins : plus il est bas, plus nous estimons qu'il existe beaucoup trop des séjours inefficaces ou inefficients dans le système des soins et qu'il convient donc d'être plus exigeants et plus efficaces. Les facteurs de risques obtenus avec ces seuils peuvent alors être des leviers pour mettre en œuvre les mesures adaptés, nécessaires, ou souhaités.

Dans cette thèse, nous prenons le parti de considérer les séjours prolongés comme très minoritaires, voire rares. D'un point de vue statistique on peut considérer comme rare un outlier et une façon simple et pratique de déterminer un outlier est d'appliquer le critère de Tukey qui détermine les clôtures inférieures et supérieures d'un boxplot. Il faut donc admettre implicitement notre vision optimiste du système des soins et de leur qualité. Avec ce choix, le seuil de binarisation pour les données de l'APHM – qui est un centre hospitalier universitaire constitué par 4 hôpitaux – est de 14 jours ($90^{ème}$ centile). Pour les données MIMIC III du BIDMC qui représentent essentiellement des séjours en soins intensifs, le seuil est de 28 jours ($92^{ème}$ centile).

Les facteurs de risques (les plus importants) de séjours prolongés pour les données de l'APHM sont :

- La destination à la sortie d'hospitalisation. Si elle est autre que le domicile alors le risque de séjour prolongé est significativement plus important.
- La catégorie de maladie (diagnostic associé au GHM). L'admission pour chimiothérapie et radiothérapie ont des risques significativement plus faibles d'être un séjour prolongé.
- Le type d'hospitalisation : un séjour en chirurgie présente un risque plus élevé d'être prolongé.

- L'origine du patient : s'il est d'ailleurs que le domicile alors le risque de séjour prolongé est plus élevé.
- La catégorie de maladie. L'admission pour orthopédie et traumatologie est associée un risque plus faible de séjour prolongé.
- Une hospitalisation via le service des urgences est associée à un risque plus important d'être un séjour prolongé.

Pour les séjours en soins intensifs du BIDMC, au niveau des données structurées, les facteurs les plus prédictifs de séjour prolongé sont les taux de nitrogène uréiques et le taux de plaquettes :

- Plus le taux maximum de nitrogène uréique est élevé et plus la probabilité de séjour prolongé augmente, et il en est de même pour le taux maximum de plaquettes ;
- Plus le taux minimum de nitrogène uréique est bas plus la probabilité de séjour prolongé augmente, et il en est de même pour le taux minimum de plaquettes.

L'inclusion des thématiques textuelles vectorisées via topic modeling (LDA) apporte des éclairages supplémentaires sur les facteurs de risques donc certains convergent avec la littérature. On notera tout particulièrement que le type de prétraitement du texte, par exemple racinisation (stemming) vs lemmatisation fait ressortir différents types de facteurs de risques ou prédicteurs de séjour prolongés.

La fusion des données utilisant la LDA avec le stemming comme la lemmatisation permettent d'une façon concrète et pratique de résoudre le compromis performance-interprétabilité. En effet les modèles obtenus restent à un niveau très élevé de discrimination (mesuré par le ROC AUC et le PRC AUC) et en même temps apportent des éclairages très détaillés à travers les éléments de textes et les thématiques associées.

Ainsi, ressortent tout particulièrement comme facteurs de risques ou de mesures correctives, les suivantes :

- Les thématiques associées aux pathologies des reins comme la dialyse et les traitements tels que le sevelamer ;
- Les pathologies des intestins comme les hernies ou la maladie de Crohn, mais également les chirurgies et procédures qui les accompagnent (fistules, stomies) ;
- Les pathologies cardiaques et les mesures qui les accompagnent (service cardiothoracique, chirurgie cardiaque, pontage coronarien, stent) ;
- Différentes pathologies et soins associées au sang (leucémie) et à la moëlle osseuse ;
- Les conditions de fin de vie et les traitements qui les accompagnent (lorazepam, clonazepam, dilantin, paralysie, embolie) ;
- Des thématiques associées à l'insuffisance respiratoire (intubation, extubation, sédation, propofol) ;
- Des thématiques associées aux maladies infectieuses et à leurs traitements ;
- Des thématiques uniquement consacrées aux soins, à la médication aux traitements.

Une des thématiques qui émergent très clairement est celle de la continuité des soins : la prise en charge à domicile, la visite des infirmières et des médecins et tout le dispositif post-hospitalisation. La littérature regrette souvent que cette dimension ne soit pratiquement jamais vraiment étudiée dans la littérature (Mahmoudi et al., 2020; O. Hansen et al., 2011), mais ici elle émerge spontanément comme un des facteurs des risques.

Limites et Perspectives de Recherche

Dans cette thèse, nous avons cherché à tenir compte de toutes les limites que nous avons identifiées dans la littérature et dans la circonférence des variables que nous avons cherchées à expliquer. En cela, plusieurs des méthodes que nous avons mobilisées et les résultats que nous avons obtenus restent rares voire uniques en sciences de la santé. Par exemple, le recours au LDA pour faciliter l'interprétabilité des données textuelles est au mieux de notre connaissance unique, ainsi que sa fusion avec des données tabulaires. De la même façon l'approche explicite de rechercher ce meilleur compromis entre interprétabilité et performance, n'a pas non plus à notre connaissance, d'équivalent.

Malgré ces quelques contributions uniques, nous avons identifié plusieurs limites à nos travaux qu'il convient de mentionner.

Une limite majeure réside dans les données APHM dans la mesure où ni les variables de biologie, ni les données textuelles cliniques n'ont pu être exploitées, pour des raisons principalement logistiques. Il aurait été tout à fait intéressant de comparer les données équivalentes de l'APHM et du BIDMC, permettant ainsi de contraster non seulement les Pays mais également les types d'institutions ou les services. Le rôle joué par les variables de biologie dans les données MIMIC III milite fortement pour la pertinence d'inclure ce type de variables dans les données de l'APHM.

L'ensemble des analyses réalisées dans ces travaux utilisent des données rétrospectives dont un certain nombre ne sont disponibles qu'à la décharge (comme les notes cliniques de décharge ou la destination patient après décharge). Or, il est sans doute préférable de n'utiliser que des données disponibles au moment de l'admission ou même en temps réel pour maîtriser les facteurs de risques de la qualité des soins, et davantage encore pour les séjours prolongés. Il serait intéressant d'appliquer les différents algorithmes utilisés ici sur des données patients disponibles uniquement à l'admission. Une autre alternative serait de déployer les meilleurs modèles comme tels dans le système hospitalier en imputant au départ les valeurs manquantes par des valeurs types (comme la médiane ou le mode) puis d'évaluer les performances à mesure que les données deviennent disponibles en temps réel.

Concernant les données textuelles du MIMIC III, nos analyses sont certes limitées par le choix de ne travailler que sur une petite partie des textes disponibles (les discharge notes) alors qu'il existe jusqu'à 14 autres différents types de textes cliniques. Une des pistes majeures d'amélioration serait de tenir compte de la totalité des données textuelles.

En termes de fusion de données, il existe une autre façon de fusionner les données textuelles sous Transformers avec les données structurées – en utilisant AutoGluon <TextPredictor> - que nous n'avons pas exploité, essentiellement pour des raisons techniques. Ce sera l'objet d'études ultérieures. Quelques explorations préliminaires suggèrent que l'on pourrait gagner en performance sans perdre en explicabilité par rapport aux textes encodés sous Transformers seuls.

Enfin, l'existence de package particulièrement performants tels qu'AutoGluon ouvre la possibilité d'explorer la fusion de données de différentes manières sans nécessairement avoir recours à des codes trop complexes.

Références

Abanoz, M., & Engin, M. (2021). The effect of the relationship between post-cardiotomy neutrophil/lymphocyte ratio and platelet counts on early major adverse events after isolated coronary artery bypass grafting. *Turkish Journal of Thoracic and Cardiovascular Surgery*, *29*(1), 36–44. https://doi.org/10.5606/tgkdc.dergisi.2021.20873

Acion, L., Kelmansky, D., Laan, M. van der, Sahker, E., Jones, D., & Arndt, S. (2017). Use of a machine learning framework to predict substance use disorder treatment success. *PLOS ONE*, *12*(4), e0175383. https://doi.org/10.1371/journal.pone.0175383

Adeyemo, D., & Radley, S. (2007). Unplanned general surgical re-admissions—How many, which patients and why? *Annals of the Royal College of Surgeons of England*, *89*(4), 363–367. https://doi.org/10.1308/003588407X183409

Ahn, J. M., Kim, S., Ahn, K.-S., Cho, S.-H., Lee, K. B., & Kim, U. S. (2018). A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLOS ONE*, *13*(11), e0207982. https://doi.org/10.1371/journal.pone.0207982

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). *Publicly Available Clinical BERT Embeddings* (arXiv:1904.03323). arXiv. https://doi.org/10.48550/arXiv.1904.03323

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010a). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010b). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Alvarez-Melis, D., & Jaakkola, T. S. (2018). *On the Robustness of Interpretability Methods* (arXiv:1806.08049). arXiv. https://doi.org/10.48550/arXiv.1806.08049

Amygdalos, I., Czigany, Z., Bednarsch, J., Boecker, J., Santana, D. A. M., Meister, F. A., von der Massen, J., Liu, W.-J., Strnad, P., Neumann, U. P., & Lurje, G. (2020). Low Postoperative Platelet Counts Are Associated with Major Morbidity and Inferior Survival in Adult Recipients of Orthotopic Liver Transplantation. *Journal of Gastrointestinal Surgery*, *24*(9), 1996–2007. https://doi.org/10.1007/s11605-019-04337-3

Anderson, G. F., & Steinberg, E. P. (1984). Hospital readmissions in the Medicare population. *The New England Journal of Medicine*, *311*(21), 1349–1353. https://doi.org/10.1056/NEJM198411223112105

ARABI, Y., VENKATESH, S., HADDAD, S., SHIMEMERI, A. A., & MALIK, S. A. (2002). A prospective study of prolonged stay in the intensive care unit: Predictors and impact on resource utilization. *International Journal for Quality in Health Care*, *14*(5), 403–410. https://doi.org/10.1093/intqhc/14.5.403

Arah, O. A., Klazinga, N. S., Delnoij, D. M. J., ten Asbroek, A. H. A., & Custers, T. (2003). Conceptual frameworks for health systems performance: A quest for effectiveness, quality, and improvement. *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care*, *15*(5), 377–398. https://doi.org/10.1093/intqhc/mzg049

Arbib, M. A. (Ed.). (1995). *The Handbook of Brain Theory and Neural Networks*. A Bradford Book.

Asadullah, M., Rashid, M., Bosu, P., Ahmed, E., & Tamanna, S. (2021). Machine Learning in Public Health: A Review. *Global Journal of Research In Engineering*. https://engineeringresearch.org/index.php/GJRE/article/view/2150

Austin, P. C., Harrell, F. E., Lee, D. S., & Steyerberg, E. W. (2022). Empirical analyses and simulations showed that different machine and statistical learning methods had differing performance for predicting blood pressure. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-13015-5

Austin, P. C., Harrell, F. E., & Steyerberg, E. W. (2021). Predictive performance of machine and statistical learning methods: Impact of data-generating processes on external validity in the

"large N, small p" setting. *Statistical Methods in Medical Research*, *30*(6), 1465–1483. https://doi.org/10.1177/09622802211002867

Bacchi, S., Gluck, S., Tan, Y., Chim, I., Cheng, J., Gilbert, T., Menon, D. K., Jannes, J., Kleinig, T., & Koblar, S. (2020). Prediction of general medical admission length of stay with natural language processing and deep learning: A pilot study. *Internal and Emergency Medicine*, *15*(6), 989–995. https://doi.org/10.1007/s11739-019-02265-3

Bacchi, S., Tan, Y., Oakden-Rayner, L., Jannes, J., Kleinig, T., & Koblar, S. (2020). Machine Learning in the Prediction of Medical Inpatient Length of Stay. *Internal Medicine Journal*. https://doi.org/10.1111/imj.14962

Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., & Yoo, S. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLoS ONE*, *13*(4), e0195901. https://doi.org/10.1371/journal.pone.0195901

Baumann, A., & Wyss, K. (2021). The shift from inpatient care to outpatient care in Switzerland since 2017: Policy processes and the role of evidence. *Health Policy*, *125*(4), 512–519. https://doi.org/10.1016/j.healthpol.2021.01.012

Becker, G. J., Strauch, G. O., & Saranchak, H. J. (1984). Outcome and Cost of Prolonged Stay in the Surgical Intensive Care Unit. *Archives of Surgery*, *119*(11), 1338–1342. https://doi.org/10.1001/archsurg.1984.01390230104026

Bland, J. M., & Altman, D. G. (2015). Statistics Notes: Bootstrap resampling methods. *BMJ*, *350*, h2622. https://doi.org/10.1136/bmj.h2622

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Blumenfeld, Y. J., El-Sayed, Y. Y., Lyell, D. J., Nelson, L. M., & Butwick, A. J. (2015). Risk Factors for Prolonged Postpartum Length of Stay Following Cesarean Delivery. *American Journal of Perinatology*, *32*(9), 825–832. https://doi.org/10.1055/s-0034-1543953

Boudemaghe, T., & Belhadj, I. (2017). Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI). *International Journal of Epidemiology*, *46*(2), 392–392d. https://doi.org/10.1093/ije/dyw359

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Brasel, K. J., Lim, H. J., Nirula, R., & Weigelt, J. A. (2007). Length of Stay: An Appropriate Quality Measure? *Archives of Surgery*, *142*(5), 461–466. https://doi.org/10.1001/archsurg.142.5.461

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brooke, B. S., Stone, D. H., Cronenwett, J. L., Nolan, B., DeMartino, R. R., MacKenzie, T. A., Goodman, D. C., & Goodney, P. P. (2014). Early Primary Care Provider Follow-up and Readmission After High-Risk Surgery. *JAMA Surgery*, *149*(8), 821–828. https://doi.org/10.1001/jamasurg.2014.157

Burgess, J. F., & Hockenberry, J. M. (2014). Can all cause readmission policy improve quality or lower expenditures? A historical perspective on current initiatives. *Health Economics, Policy, and Law*, *9*(2), 193–213. https://doi.org/10.1017/S1744133113000340

Carpentier, D., Beduneau, G., & Girault, C. (2015). Séjour prolongé en réanimation. *Réanimation*, *24*(4), Article 4. https://doi.org/10.1007/s13546-015-1089-8

Carrington, A. M., Manuel, D. G., Fieguth, P., Ramsay, T. O., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., Sheikh, Y., Mcinnes, M., & Holzinger, A. (2022). Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/TPAMI.2022.3145392

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), Article 8. https://doi.org/10.3390/electronics8080832

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, *11*(70), 2079–2107.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet. Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

Chen, H., Lundberg, S. M., & Lee, S.-I. (2021). *Explaining a Series of Models by Propagating Shapley Values* (arXiv:2105.00108). arXiv. https://doi.org/10.48550/arXiv.2105.00108

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chin, D. L., Bang, H., Manickam, R. N., & Romano, P. S. (2016). Rethinking Thirty-Day Hospital Readmissions: Shorter Intervals Might Be Better Indicators Of Quality Of Care. *Health Affairs (Project Hope)*, *35*(10), 1867–1875. https://doi.org/10.1377/hlthaff.2016.0205

Chollet, F. (2015). Keras. *GitHub Repository*. https://github.com/fchollet/keras%7D%7D

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Chrusciel, J., Girardon, F., Roquette, L., Laplanche, D., Duclos, A., & Sanchez, S. (2021). The prediction of hospital length of stay using unstructured data. *BMC Medical Informatics and Decision Making*, *21*. https://doi.org/10.1186/s12911-021-01722-4

Collins, T. C., Daley, J., Henderson, W. H., & Khuri, S. F. (1999). Risk Factors for Prolonged Length of Stay After Major Elective Surgery. *Annals of Surgery*, *230*(2), 251.

Combes, A., Costa, M.-A., Trouillet, J.-L., Baudot, J., Mokhtari, M., Gibert, C., & Chastre, J. (2003). Morbidity, mortality, and quality-of-life outcomes of patients requiring ≥14 days of mechanical ventilation. *Critical Care Medicine*, *31*(5), 1373–1381. https://doi.org/10.1097/01.CCM.0000065188.87029.C3

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal*, *20*(1), 131–148. https://doi.org/10.1177/1536867X20909693

Craig, E., Arias, C., & Gillman, D. (2017). Predicting readmission risk from doctors' notes. *ArXiv:1711.10663 [Stat]*. http://arxiv.org/abs/1711.10663

Debortoli, S., Müller, O., Junglas, I., & Brocke, J. vom. (2016). Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, *39*(1). https://doi.org/10.17705/1CAIS.03907

DGOS, & ATIH. (2022). *Les réhospitalisations à 30 jours (RH30)*. Ministère de la Santé et de la Prévention.

Dictionary.com. (2022). *Stopword* (Dictionary.com, LLC). https://www.dictionary.com/browse/stopword

Direction de la Recherche, des Études, de l'Évaluation et des Statistiques. (2020). *Les dépenses de santé en 2019—Résultats des comptes de la santé—Édition 2020*. Vie publique.fr. https://www.vie-publique.fr/rapport/276352-les-depenses-de-sante-en-2019-resultats-des-comptes-de-la-sante

Donabedian, A. (2003). *An Introduction to Quality Assurance in Health Care*. Oxford University Press.

Donabedian, A. (2005). Evaluating the Quality of Medical Care. *The Milbank Quarterly*, *83*(4), 691–729. https://doi.org/10.1111/j.1468-0009.2005.00397.x

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: Gradient boosting with categorical features support* (arXiv:1810.11363). arXiv. https://doi.org/10.48550/arXiv.1810.11363

dos Santos, B. S., Steiner, M. T. A., Fenerich, A. T., & Lima, R. H. P. (2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers & Industrial Engineering*, *138*, 106120. https://doi.org/10.1016/j.cie.2019.106120

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Dundar, Z. D., Kucukceran, K., & Ayranci, M. K. (2021). Blood urea nitrogen to albumin ratio is a predictor of in-hospital mortality in older emergency department patients. *The American Journal of Emergency Medicine*, *46*, 349–354. https://doi.org/10.1016/j.ajem.2020.10.008

Ebinger, J., Wells, M., Ouyang, D., Davis, T., Kaufman, N., Cheng, S., & Chugh, S. (2021). A Machine Learning Algorithm Predicts Duration of hospitalization in COVID-19 patients. *Intelligence-Based Medicine*, *5*, 100035. https://doi.org/10.1016/j.ibmed.2021.100035

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650. https://doi.org/10.1111/coin.12410

Epstein, A. M., Jha, A. K., & Orav, E. J. (2011). The relationship between hospital admission rates and rehospitalizations. *The New England Journal of Medicine*, *365*(24), 2287–2295. https://doi.org/10.1056/NEJMsa1101942

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data* (arXiv:2003.06505). arXiv. https://doi.org/10.48550/arXiv.2003.06505

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics* (4th edition). Cambridge University Press.

Exertier, Minodier, Roland, Huault, Collin, Richard, Cash, Saab, Groheux, Raimbault, Paufiques, & Eyssartier. (2011). *Les inadéquations hospitalières en France: Fréquence, causes et impact économique* (pp. 33–45).

Faisst, M., Wellner, U. F., Utzolino, S., Hopt, U. T., & Keck, T. (2010). Elevated blood urea nitrogen is an independent risk factor of prolonged intensive care unit stay due to acute necrotizing pancreatitis. *Journal of Critical Care*, *25*(1), 105–111. https://doi.org/10.1016/j.jcrc.2009.02.002

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Fisher, A. J., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.*

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Garreau, D., & Luxburg, U. (2020). Explaining the Explainer: A First Theoretical Analysis of LIME. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 1287–1296. https://proceedings.mlr.press/v108/garreau20a.html

Gefen, D., Endicott, J., Fresneda, J., Miller, J., & Larsen, K. (2017). A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community. *Communications of the Association for Information Systems*, *41*(1). https://doi.org/10.17705/1CAIS.04121

Gholipour, C., Rahim, F., Fakhree, A., & Ziapour, B. (2015). Using an Artificial Neural Networks (ANNs) Model for Prediction of Intensive Care Unit (ICU) Outcome and Length of Stay at Hospital in Traumatic Patients. *Journal of Clinical and Diagnostic Research : JCDR*, *9*(4), OC19–OC23. https://doi.org/10.7860/JCDR/2015/9467.5828

Gilman, M., Adams, E. K., Hockenberry, J. M., Wilson, I. B., Milstein, A. S., & Becker, E. R. (2014). California Safety-Net Hospitals Likely To Be Penalized By ACA Value, Readmission, And Meaningful-Use Programs. *Health Affairs*, *33*(8), 1314–1322. https://doi.org/10.1377/hlthaff.2014.0138

Globerson, A., Chechik, G., Pereira, F., & Tishby, N. (2007). Euclidean Embedding of Co-occurrence Data. *Journal of Machine Learning Research*, *8*(76), 2265–2295.

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

Graham, K. L., Wilker, E. H., Howell, M. D., Davis, R. B., & Marcantonio, E. R. (2015). Differences between early and late readmissions among patients: A cohort study. *Annals of Internal Medicine*, *162*(11), 741–749. https://doi.org/10.7326/M14-2159

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

Gupta, A., Tatbul, N., Marcus, R., Zhou, S., Lee, I., & Gottschlich, J. (2020). *Class-Weighted Evaluation Metrics for Imbalanced Data Classification* (arXiv:2010.05995). arXiv. https://doi.org/10.48550/arXiv.2010.05995

Gusmano, M., Rodwin, V., Weisz, D., Cottenet, J., & Quantin, C. (2015). Comparison of rehospitalization rates in france and the United States. *Journal of Health Services Research and Policy*, *20*(1), 18–25. https://doi.org/10.1177/1355819614551849

Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Ho, T. K., Macià, N., Ray, B., Saeed, M., Statnikov, A., & Viegas, E. (2015). Design of the 2015 ChaLearn AutoML challenge. *2015 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2015.7280767

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

Hand, D. J., & Anagnostopoulos, C. (2014). A better Beta for the H measure of classification performance. *Pattern Recognition Letters*, *40*, 41–46. https://doi.org/10.1016/j.patrec.2013.12.011

Hassan, A., Anderson, C., Kypson, A., Kindell, L., Ferguson, T. B., Chitwood, W. R., & Rodriguez, E. (2012). Clinical Outcomes in Patients With Prolonged Intensive Care Unit Length of Stay After Cardiac Surgical Procedures. *The Annals of Thoracic Surgery*, *93*(2), 565–569. https://doi.org/10.1016/j.athoracsur.2011.10.024

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2nd ed.). Springer-Verlag. //www.springer.com/us/book/9780387848570

Hendlmeier, I., Bickel, H., Heßler-Kaufmann, J. B., & Schäufele, M. (2019). Care challenges in older general hospital patients. *Zeitschrift Für Gerontologie Und Geriatrie*, *52*(4), 212–221. https://doi.org/10.1007/s00391-019-01628-x

Hendrycks, D., & Gimpel, K. (2020). *Gaussian Error Linear Units (GELUs)* (arXiv:1606.08415). arXiv. https://doi.org/10.48550/arXiv.1606.08415

Heyland, D. K., Konopad, E., Noseworthy, T. W., Johnston, R., & Gafni, A. (1998). Is It 'Worthwhile' To Continue Treating Patients With a Prolonged Stay (>14 Days) in the ICU?: An Economic Evaluation. *Chest*, *114*(1), 192–198. https://doi.org/10.1378/chest.114.1.192

Horwitz, L., Partovian, C., Lin, Z., Herrin, J., Grady, J., Conover, M., Montague, J., Dillaway, C., Bartczak, K., Ross, J., Bernheim, S., Drye, E., & Krumholz, H. M. (2011). *Hospital-Wide (All-Condition) 30-Day Risk-Standardized Readmission Measure*.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression, 3rd Edition* (3rd edition). Wiley.

Hossain, MD. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys*, *51*(6), 118:1-118:36. https://doi.org/10.1145/3295748

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

Huang, J., Osorio, C., & Sy, L. W. (2019). An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Computer Methods and Programs in Biomedicine*, *177*, 141–153. https://doi.org/10.1016/j.cmpb.2019.05.024

Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *ArXiv:1904.05342 [Cs]*. http://arxiv.org/abs/1904.05342

Huang, Y., Talwar, A., Chatterjee, S., & Aparasu, R. R. (2021). Application of machine learning in predicting hospital readmissions: A scoping review of the literature. *BMC Medical Research Methodology*, *21*(1), 96. https://doi.org/10.1186/s12874-021-01284-z

Institute of Medicine (US) Committee on Quality of Health Care in America. (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*. National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK222274/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer-Verlag. //www.springer.com/us/book/9781461471370

Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *The New England Journal of Medicine*, *360*(14), 1418–1428. https://doi.org/10.1056/NEJMsa0803563

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.35

Joynt, K. E., & Jha, A. K. (2012). Thirty-day readmissions—Truth and consequences. *The New England Journal of Medicine*, *366*(15), 1366–1369. https://doi.org/10.1056/NEJMp1201598

Kangovi, S., & Grande, D. (2011). Hospital readmissions—Not just a measure of quality. *JAMA*, *306*(16), 1796–1797. https://doi.org/10.1001/jama.2011.1562

Kareliusson, F., De Geer, L., & Tibblin, A. O. (2015). Risk prediction of ICU readmission in a mixed surgical and medical population. *Journal of Intensive Care*, *3*(1), 30. https://doi.org/10.1186/s40560-015-0096-1

KaurHarsurinder, Singh, P., & Kaur, M. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning. *ACM Computing Surveys (CSUR)*. https://doi.org/10.1145/3343440

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.

Kelley, E., & Hurst, J. (2006). *Health Care Quality Indicators Project: Conceptual Framework Paper*. OCDE. https://doi.org/10.1787/440134737301

Koskas, P., Pons-Peyneau, C., Romdhani, M., Houenou-Quenum, N., Galleron, S., & Drunat, O. (2019). Hospital Discharge Decisions Concerning Older Patients: Understanding the Underlying Process. *Canadian Journal on Aging / La Revue Canadienne Du Vieillissement*, *38*(1), 90–99. https://doi.org/10.1017/S0714980818000442

Kristensen, S. R., Bech, M., & Quentin, W. (2015). A roadmap for comparing readmission policies with application to Denmark, England, Germany and the United States. *Health Policy (Amsterdam, Netherlands)*, *119*(3), 264–273. https://doi.org/10.1016/j.healthpol.2014.12.009

Kuhn, H. W. (2020). *Classics in Game Theory*. Princeton University Press.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer-Verlag. //www.springer.com/us/book/9781461468486

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Laupland, K. B., Kirkpatrick, A. W., Kortbeek, J. B., & Zuege, D. J. (2006). Long-term Mortality Outcome Associated With Prolonged Admission to the ICU. *Chest*, *129*(4), 954–959. https://doi.org/10.1378/chest.129.4.954

Le Gall, J. R., Lemeshow, S., & Saulnier, F. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA*, *270*(24), 2957–2963. https://doi.org/10.1001/jama.270.24.2957

Lebret, R., & Collobert, R. (2017). Word Embeddings through Hellinger PCA. *ArXiv:1312.5542 [Cs]*. http://arxiv.org/abs/1312.5542

Lequertier, V., Wang, T., Fondrevelle, J., Augusto, V., & Duclos, A. (2021). Hospital Length of Stay Prediction Methods: A Systematic Review. *Medical Care*, *59*(10), 929–938. https://doi.org/10.1097/MLR.0000000000001596

Levy, O., & Goldberg, Y. (2014). Linguistic Regularities in Sparse and Explicit Word Representations. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 171–180. https://doi.org/10.3115/v1/W14-1618

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), Article 1. https://doi.org/10.3390/e23010018

Lisk, R., Uddin, M., Parbhoo, A., Yeong, K., Fluck, D., Sharma, P., Lean, M. E. J., & Han, T. S. (2019). Predictive model of length of stay in hospital among older patients. *Aging Clinical and Experimental Research*, *31*(7), 993–999. https://doi.org/10.1007/s40520-018-1033-7

Liu, X., Chen, Y., Bae, J., Li, H., Johnston, J., & Sanger, T. (2019). Predicting Heart Failure Readmission from Clinical Notes Using Deep Learning. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2642–2648. https://doi.org/10.1109/BIBM47256.2019.8983095

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), Article 1. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

MacIntyre, N. R., Epstein, S. K., Carson, S., Scheinhorn, D., Christopher, K., & Muldoon, S. (2005). Management of Patients Requiring Prolonged Mechanical Ventilation: Report of a NAMDRC Consensus Conference. *Chest*, *128*(6), 3937–3954. https://doi.org/10.1378/chest.128.6.3937

Mahesh, B., Choong, C. K., Goldsmith, K., Gerrard, C., Nashef, S. A. M., & Vuylsteke, A. (2012). Prolonged Stay in Intensive Care Unit Is a Powerful Predictor of Adverse Outcomes After Cardiac Operations. *The Annals of Thoracic Surgery*, *94*(1), 109–116. https://doi.org/10.1016/j.athoracsur.2012.02.010

Mahmoudi, E., Kamdar, N., Kim, N., Gonzales, G., Singh, K., & Waljee, A. K. (2020). Use of electronic medical records in development and validation of risk prediction models of hospital readmission: Systematic review. *BMJ*, *369*, m958. https://doi.org/10.1136/bmj.m958

Marfil-Garza, B. A., Belaunzarán-Zamudio, P. F., Gulias-Herrero, A., Zuñiga, A. C., Caro-Vega, Y., Kershenobich-Stalnikowitz, D., & Sifuentes-Osornio, J. (2018a). Risk factors associated with prolonged hospital length-of-stay: 18-year retrospective study of hospitalizations in a tertiary healthcare center in Mexico. *PloS One*, *13*(11), e0207203. https://doi.org/10.1371/journal.pone.0207203

Marfil-Garza, B. A., Belaunzarán-Zamudio, P. F., Gulias-Herrero, A., Zuñiga, A. C., Caro-Vega, Y., Kershenobich-Stalnikowitz, D., & Sifuentes-Osornio, J. (2018b). Risk factors associated with prolonged hospital length-of-stay: 18-year retrospective study of hospitalizations in a tertiary

healthcare center in Mexico. *PLOS ONE*, *13*(11), e0207203. https://doi.org/10.1371/journal.pone.0207203

Marshall, A., Vasilakis, C., & El-Darzi, E. (2005). Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions. *Health Care Management Science*, *8*(3), 213–220. https://doi.org/10.1007/s10729-005-2012-z

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*. https://doi.org/10.1016/0005-2795(75)90109-9

McCoy, T. H., Castro, V. M., Cagan, A., Roberson, A. M., Kohane, I. S., & Perlis, R. H. (2015). Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study. *PLOS ONE*, *10*(8), e0136341. https://doi.org/10.1371/journal.pone.0136341

McIlvennan, C. K., Eapen, Z. J., & Allen, L. A. (2015). Hospital readmissions reduction program. *Circulation*, *131*(20), 1796–1803. https://doi.org/10.1161/CIRCULATIONAHA.114.010270

Medicare Payment Advisory Commission (U.S.) (Ed.). (2007). *Report to the Congress: Promoting greater efficiency in Medicare*. Medicare Payment Advisory Commission.

Mekhaldi, R. N., Caulier, P., Chaabane, S., Chraibi, A., & Piechowiak, S. (2020). Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting. In Á. Rocha, H. Adeli, L. P. Reis, S. Costanzo, I. Orovic, & F. Moreira (Eds.), *Trends and Innovations in Information Systems and Technologies* (pp. 202–211). Springer International Publishing. https://doi.org/10.1007/978-3-030-45688-7_21

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, *3*(1), 27–32. https://doi.org/10.1145/507533.507538

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2022). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/

Montuclard, L., Garrouste-Orgeas, M., Timsit, J.-F., Misset, B., De Jonghe, B., & Carlet, J. (2000). Outcome, functional autonomy, and quality of life of elderly patients with a long-term intensive care unit stay. *Critical Care Medicine*, *28*(10), 3389–3395.

Navathe, A. S., Zhong, F., Lei, V. J., Chang, F. Y., Sordo, M., Topaz, M., Navathe, S. B., Rocha, R. A., & Zhou, L. (2018). Hospital Readmission and Social Risk Factors Identified from Physician Notes. *Health Services Research*, *53*(2), 1110–1136. https://doi.org/10.1111/1475-6773.12670

Nguyen, O. K., Makam, A. N., Clark, C., Zhang, S., Xie, B., Velasco, F., Amarasingham, R., & Halm, E. A. (2016). Predicting all-cause readmissions using electronic health record data from the entire hospitalization: Model development and comparison. *Journal of Hospital Medicine*, *11*(7), 473–480. https://doi.org/10.1002/jhm.2568

Nicolet, G., Quantin, C., Duclos, A., Chollet, F., Cottenet, J., & Mercier, G. (2022). Développement d'un modèle prédictif du risque de réhospitalisation non programmée à partir des données PMSI nationales. *Revue d'Épidémiologie et de Santé Publique*, *70*, S24. https://doi.org/10.1016/j.respe.2022.01.104

O. Hansen, L., S. Young, R., Hinami, K., Leung, A., & V. Williams, M. (2011). Interventions to Reduce 30-Day Rehospitalization: A Systematic Review. *Annals of Internal Medicine*. https://www.acpjournals.org/doi/10.7326/0003-4819-155-8-201110180-00008

OECD. (2017). *Health at a Glance 2017: OECD Indicators*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2017_health_glance-2017-en

Or, Z., & Com-Ruelle, L. (2008). La qualité de soins en France: Comment la mesurer pour l'améliorer ? *Journal d'économie médicale*, *26*(6–7), 371–385. https://doi.org/10.3917/jgem.086.0371

Orangi-Fard, N., Akhbardeh, A., & Sagreiya, H. (2022). Predictive Model for ICU Readmission Based on Discharge Summaries Using Machine Learning and Natural Language Processing. *Informatics*, *9*(1), Article 1. https://doi.org/10.3390/informatics9010010

Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, *68*(8), 855–859. https://doi.org/10.1016/j.jclinepi.2015.02.010

Pauly, V., Mendizabal, H., Gentile, S., Auquier, P., & Boyer, L. (2019). Predictive risk score for unplanned 30-day rehospitalizations in the French universal health care system based on a medico-administrative database. *PLOS ONE*, *14*(3), e0210714. https://doi.org/10.1371/journal.pone.0210714

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014). *GloVe: Global Vectors for Word Representation*. Empirical Methods In Natural Language Processing.

Powell, M., Hosseini, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). *I Tried a Bunch of Things: The Dangers of Unexpected Overfitting in Classification* (p. 078816). bioRxiv. https://doi.org/10.1101/078816

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

Qu, M., Liu, Q., Zhao, H.-G., Peng, J., Ni, H., Hou, M., & Jansen, A. J. G. (2018). Low platelet count as risk factor for infections in patients with primary immune thrombocytopenia: A retrospective evaluation. *Annals of Hematology*, *97*(9), 1701–1706. https://doi.org/10.1007/s00277-018-3367-9

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., Beck, C. A., Feasby, T. E., & Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, *43*(11), 1130–1139. https://doi.org/10.1097/01.mlr.0000182534.19832.83

Rachda Naila, M., Caulier, P., Chaabane, S., Chraibi, A., & Piechowiak, S. (2021). A Comparative Study of Machine Learning Models for Predicting Length of Stay in Hospitals. *Journal of Information Science and Engineering*, *37*, 1025–1038.

Radley, D., Mccarthy, D., & Hayes, S. (2015). *Aiming Higher: Results from the Commonwealth Fund Scorecard on State Health System Performance*. https://doi.org/10.15868/socialsector.26933

Raleigh, V. B. M. (2014). Integrated care and support Pioneers: Indicators for measuring the quality of integrated care. *Policy Innovation Research Unit (PIRU)*. https://researchonline.lshtm.ac.uk/id/eprint/4398580/

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G.-Z. (2017). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4–21. https://doi.org/10.1109/JBHI.2016.2636665

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & Smeden, M. van. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

Rimachi, R., Vincent, J. L., & Brimioulle, S. (2007). Survival and quality of life after prolonged intensive care unit stay. *Anaesthesia and Intensive Care*, *35*(1), 62–67. https://doi.org/10.1177/0310057X0703500108

Roemer, M. I., Montoya-Aguilar, C., & Organization, W. H. (1988). *Quality assessment and assurance in primary health care*. World Health Organization. https://apps.who.int/iris/handle/10665/40663

Rojas-García, A., Turner, S., Pizzo, E., Hudson, E., Thomas, J., & Raine, R. (2018). Impact and experiences of delayed discharge: A mixed-studies systematic review. *Health Expectations : An International Journal of Public Participation in Health Care and Health Policy*, *21*(1), 41–56. https://doi.org/10.1111/hex.12619

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, *6*(10), Article 10. https://doi.org/10.1038/tp.2015.182

Schroeder, S. A., Showstack, J. A., & Roberts, H. E. (1979). Frequency and clinical description of high-cost patients in 17 acute-care hospitals. *The New England Journal of Medicine*, *300*(23), 1306–1309. https://doi.org/10.1056/NEJM197906073002304

Shi, X., Mueller, J., Erickson, N., Li, M., & Smola, A. (2021, July 14). *Multimodal AutoML on Structured Tables with Text Fields*. 8th ICML Workshop on Automated Machine Learning (AutoML). https://openreview.net/forum?id=OHAIVOOl7Vl

Silberman, S., Bitran, D., Fink, D., Tauber, R., & Merin, O. (2013). Very Prolonged Stay in the Intensive Care Unit After Cardiac Operations: Early Results and Late Survival. *The Annals of Thoracic Surgery*, *96*(1), 15–22. https://doi.org/10.1016/j.athoracsur.2013.01.103

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., & Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 801–810. https://doi.org/10.1001/jama.2016.0287

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180–186). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375830

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, *10*(5), e1379. https://doi.org/10.1002/widm.1379

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Tatlisu, M. A., Kaya, A., Keskin, M., Avsar, S., Bozbay, M., Tatlisu, K., & Eren, M. (2017). The association of blood urea nitrogen levels with mortality in acute pulmonary embolism. *Journal of Critical Care*, *39*, 248–253. https://doi.org/10.1016/j.jcrc.2016.12.019

Tefera, G. M., Feyisa, B. B., Umeta, G. T., & Kebede, T. M. (2020). Predictors of prolonged length of hospital stay and in-hospital mortality among adult patients admitted at the surgical ward of Jimma University medical center, Ethiopia: Prospective observational study. *Journal of Pharmaceutical Policy and Practice*, *13*, 24. https://doi.org/10.1186/s40545-020-00230-6

Teo, K., Yong, C. W., Chuah, J. H., Murphy, B. P., & Lai, K. W. (2020). Discovering the Predictive Value of Clinical Notes: Machine Learning Analysis with Text Representation. *Journal of Medical Imaging and Health Informatics*, *10*(12), 2869–2875. https://doi.org/10.1166/jmihi.2020.3291

Tukey, J. (1977a). *Exploratory Data Analysis*. Addison-Wesley.

Tukey, J. (1977b). *Exploratory Data Analysis* (1st edition). Pearson.

Tunstall, L., Werra, L. von, & Wolf, T. (2022). *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. *CreateSpace*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, W., & Sun, D. (2021). The improved AdaBoost algorithms for imbalanced data classification. *Information Sciences*, *563*, 358–374. https://doi.org/10.1016/j.ins.2021.03.042

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., & Lestantyo, P. (2019). Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data. *2019 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, 14–18. https://doi.org/10.1109/IC3INA48034.2019.8949568

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, *142*, 399–432. https://doi.org/10.1016/j.ress.2015.05.018

Weissman, G. E., Hubbard, R. A., Ungar, L. H., Harhay, M. O., Greene, C. S., Himes, B. E., & Halpern, S. D. (2018). Inclusion of Unstructured Clinical Text Improves Early Prediction of Death or Prolonged ICU Stay. *Critical Care Medicine*, *46*(7), 1125–1132. https://doi.org/10.1097/CCM.0000000000003148

White, A. C. (2012). Long-Term Mechanical Ventilation: Management Strategies. *Respiratory Care*, *57*(6), 889–899. https://doi.org/10.4187/respcare.01850

Williams, T. A., Ho, K. M., Dobb, G. J., Finn, J. C., Knuiman, M., & Webb, S. A. R. (2010). Effect of length of stay in intensive care unit on hospital and long-term mortality of critically ill adult patients. *British Journal of Anaesthesia*, *104*(4), 459–464. https://doi.org/10.1093/bja/aeq025

World Health Organization. (2015). *People-centred and integrated health services: An overview of the evidence: interim report* (WHO/HIS/SDS/2015.7). World Health Organization. https://apps.who.int/iris/handle/10665/155004

Yilmaz, E., & Vuagnat, A. (2015). *Tarification à l'activité: Quel impact sur les réadmissions à l'hôpital ?* DRESS.

Yule, G. U. (1912). On the Methods of Measuring Association between Two Attributes. *Journal of the Royal Statistical Society*, *75*(6), 579–642. https://doi.org/10.1111/j.2397-2335.1912.tb00463.x

Zampieri, F. G., Ladeira, J. P., Park, M., Haib, D., Pastore, C. L., Santoro, C. M., & Colombari, F. (2014). Admission factors associated with prolonged (>14 days) intensive care unit stay. *Journal of Critical Care*, *29*(1), 60–65. https://doi.org/10.1016/j.jcrc.2013.09.030

Zhang, D., Yin, C., Zeng, J., Yuan, X., & Zhang, P. (2020). Combining structured and unstructured data for predictive models: A deep learning approach. *BMC Medical Informatics and Decision Making*, *20*(1), 280. https://doi.org/10.1186/s12911-020-01297-6

Zhou, H., Della, P. R., Roberts, P., Goh, L., & Dhaliwal, S. S. (2016). Utility of models to predict 28-day or 30-day unplanned hospital readmissions: An updated systematic review. *BMJ Open*, *6*(6), e011060. https://doi.org/10.1136/bmjopen-2016-011060

Zook, C. J., & Moore, F. D. (1980). High-cost users of medical care. *The New England Journal of Medicine*, *302*(18), 996–1002. https://doi.org/10.1056/NEJM198005013021804

Zook, C. J., Savickis, S. F., & Moore, F. D. (1980). Repeated hospitalization for the same disease: A multiplier of national health costs. *The Milbank Memorial Fund Quarterly. Health and Society*, *58*(3), 454–471.