Apprentissage automatique dans la prédiction des durées de séjour hospitalier.

MEKHALDI Rachda Naila[1], CAULIER Patrice[1], CHAABANE Sondes[1], PIECHOWIAK Sylvain[1], TAILLARD Julien[2], HANSSKE Arnaud[3]

[1] Univ. Polytechnique Hauts-de-France, CNRS UMR 8201 - LAMIH - Laboratoire d'Automatique, de Mécanique et d'Informatique Industrielles et Humaines, F-59313 Valenciennes, France, (+33) (0) 3 27 51 13 55, {rachdanaila.mekhaldi, patrice.caulier,sondes.chaabane,sylvain.piechowiak}@uphf.fr

[2] Alicante, 50 rue Philippe de Girard, 59113 SECLIN, (+33) 3 28 55 92 50, julien.taillard@alicante.fr

[3] Groupement des Hôpitaux de l'Institut Catholique de Lille (GHICL), rue du grand but BP 249, 59462 Lomme, (+33) 3 20 22 50 50, Hansske.Arnaud@ghicl.net.

Résumé. Au cours des dernières années, l'estimation de la durée de séjour hospitalier (DDS) au moment de l'admission du patient fait l'objet de plusieurs études. La DDS est un indicateur d'évaluation du rendement des établissements de soins et d'efficacité de la performance des services hospitaliers. La prédiction des DDS contribue à l'optimisation des ressources des hôpitaux, à l'amélioration de l'organisation des soins et à une meilleure planification des activités. Dans cette communication, nous exposons la démarche suivie pour implémenter un modèle de prédiction de durées de séjour hospitalier en exploitant des algorithmes d'apprentissage automatique. Pour chaque étape du processus de prédiction, les difficultés rencontrées sont soulevées et discutées. L'implication de l'expertise et son importance dans dans le projet sont également mises en évidence. Finalement, un exemple illustratif utilisant l'ensemble de données « Microsoft for length of stay prediction » est présenté.

Mots clés : Durée de séjour hospitalier, exploration de données, apprentissage automatique, modèle de prédiction.

1. Introduction

Au cours de ces dernières années, dans le domaine hospitalier, un intérêt croissant est donné à la collaboration entre les chercheurs universitaires, les professionnels de santé et les développeurs d'applications dans le domaine hospitalier. Cette collaboration, cherche à rassembler les efforts de chaque partie pour améliorer les performances d'efficacité des services des établissements de soins. La Durée de Séjour Hospitalier (DDS) constitue un des indicateurs de base d'évaluation. Cette durée représente l'intervalle de temps entre l'admission du patient dans un service (ou hôpital) et sa sortie du service (ou hôpital). Elle peut aussi être étayée par le parcours suivi lors du séjour. Une des problématiques autour de cet indicateur est la prédiction de la durée de séjour. La prédiction de DDS présente plusieurs avantages. Pour l'organisation de l'hôpital, elle aide à optimiser les ressources matérielles et humaines de l'hôpital, à planifier au plus juste les activités de soins et, enfin, à améliorer la qualité des services. En 2017, selon (ATIH, 2018), la région Hauts-de-France a connu près de 1 711 900 hospitalisations. Du côté des professionnels de santé, la prédiction des DDS contribue à diminuer leur surcharge du travail et à améliorer la planification de leurs activités. En 2018, selon les données PMSI (Programme de Médicalisation des Systèmes d'Information), cette surcharge du travail dans les services MCO (Médecine, Chirurgie, Obstétrique) a atteint plus de 12,3 millions de patients (ATIH, 2019).

Finalement, pour les patients, la prédiction de DDS permet de réduire les délais d'attente et ainsi améliorer le service rendu.

La durée de séjour à l'hôpital est une variable complexe qui dépend de plusieurs facteurs hétérogènes relatifs aux patients et à l'organisation des établissements de soins. Des méthodes d'intelligence artificielle (IA), dont l'apprentissage automatique et le datamining sont utilisées pour l'analyse de ces facteurs. Dans (Shea *et al.*, 1995), les auteurs ont montré que la durée de séjour est définie par type de service. En effet, la modélisation du séjour hospitalier dans un service programmé diffère de celle d'un service d'urgence (RIGAL, 2009). Dans (Lafaro *et al.*, 2015) et (Hachesu *et al.*, 2013), les auteurs se sont intéressés aux facteurs liés au service de cardiologie. D'autres travaux ont analysé les facteurs qui influencent les DDS dans le service de soins intensifs (Gentimis *et al.*, 2017) et (Maharlou *et al.*, 2018). D'autres se sont focalisés sur le service de chirurgie (Chuang *et al.*, 2016) et (Khosravizadeh *et al.*, 2016).

Quoique toutes ces recherches, basées sur des techniques d'IA, ont tenté d'étudier les facteurs impactant les DDS dans différents services, l'implication des experts dans le domaine médical est aussi fondamentale et, donc, inévitable. Le rôle déterminant de l'expertise se situe, principalement, lors des phases d'analyse des facteurs et de validation des résultats. La construction d'un modèle de prédiction est influencée par la conception de l'étude. L'objectif de ce papier est double. D'une part, présenter les algorithmes issus de la littérature pour la prédiction des DDS et le processus classique à suivre en s'appuyant sur les connaissances dans le domaine hospitalier. D'autre part, exposer la complexité du traitement des données médicales à partir d'un exemple illustratif fondé sur un ensemble de données open source de Microsoft pour la prédiction des durées de séjour hospitalier (Microsoft, 2017).

L'article est organisé comme suit. La partie suivante dresse une panoplie de travaux dans le domaine de l'apprentissage automatique pour la prédiction des durées de séjour hospitalier. Ensuite, nous exposons le processus classique de classification automatique. Finalement, un cas d'étude illustre les difficultés relatives au prétraitement des données médicales dans un processus de classification automatique est donné.

2. Prédiction des DDS basée sur les modèles d'apprentissage automatique

Les algorithmes d'intelligence artificielle, dont l'apprentissage automatique pour les problèmes de prédiction, ont fait un retour fracassant ces dernières années. Dans le volet de la prédiction des durées de séjour hospitalier, diverses méthodes ont été comparées. Dans (Hachesu *et al.*, 2013), les modèles d'arbres de décisions, de machines à vecteur support (SVM pour Support Vector Machines) et les réseaux de neurones artificiels (RN) sont implémentés pour la prédiction des DDS chez les patients atteints d'une coronaropathie. L'étude a montré que l'étape de sélection de variables est très importante pour l'obtention de meilleures performances dans la phase d'apprentissage automatique. Les auteurs (Chuang, Hu and Lo, 2018) ont obtenu une meilleure prédiction des longues DDS après une opération chirurgicale avec les forêts décisionnelles (RF pour Random Forest) comparant aux SVM, la régression logistique, les algorithmes CART. En plus des réseaux de neurones, la régressions linéaire est utilisée pour prédire les DDS dans le service cardiovasculaire (Tsai *et al.*, 2016). Pour le même type de service, plusieurs modèles de RN sont employés dans (Lafaro *et al.*, 2015). Dans le même contexte des RN, la méthode de Back Propagation, est appliquée dans (Li *et al.*, 2013) pour la prédiction des DDS dans le service de chirurgie. Dans l'étude menée par (Chuang *et al.*, 2016), les arbres de décisions, les SVM et le RF sont implémentés pour prédire la DDS avant une opération chirurgicale. Le meilleur résultat est obtenu par la méthode RF. Dans un autre travail, les systèmes d'inférence flous, les réseaux de neurones en employant la fonction sigmoid sont utilisés (Maharlou *et al.*, 2018). Plusieurs méthodes sont implémentées pour prédire la DDS dans le service pédiatrique d'urgence. Nous citons le naïf bays, les SVM et

les arbres de décision (Benbelkacem *et al.*, 2019). Sur la base de données MIMIC3 (Medical Information Mart for Intensive Care) (Johnson *et al.*, 2016), les réseaux de neurones et les forêts décisionnelles sont utilisés (Gentimis *et al.*, 2017).

Les différentes recherches ont révélé, que les algorithmes d'apprentissage automatique, particulièrement, ceux d'apprentissage supervisé sont largement utilisés pour la prédiction des DDS. Il s'agit d'un même processus de classification et de différents types d'algorithmes pour anticiper la DDS. Nous avons constaté que les algorithmes d'apprentissage du type supervisé comme les réseaux de neurones, les arbres de décision, les SVM et les Random Forest sont les plus utilisés dans la prédiction des DDS. Dans le cadre de notre étude, nous devons dans un premier temps définir l'algorithme à adopter parmi cet ensemble. Ceci dépendra de l'ensemble de données d'entrée ainsi que de l'objectif de l'étude.
Dans la section suivante, nous allons détailler le processus de classification automatique en positionnant le rôle des différents intervenants dans chaque phase.

3. Processus de classification automatique

Au vu de la revue de la littérature présentée plus haut, l'estimation de la durée de séjour du patient au moment de son admission peut faire appel aux méthodes d'Intelligence Artificielle dont l'apprentissage automatique et le data mining. Le processus d'application de ces méthodes part de l'utilisation de grandes quantités de données au dépoilement d'un modèle de prédiction. Les étapes de ce processus s'articulent autour des points suivants :

- Collecte des données : Une fois l'objectif de l'étude défini, à partir de différentes sources, nous procédons à la collecte de données à partir de différentes sources. Plusieurs formats coexistent. Pour un modèle d'apprentissage automatique, le format matriciel est utilisé. Les données peuvent aussi être incomplètes et, parfois, aberrantes ou imprécises. Les lignes représentent les observations de l'ensemble de données et les colonnes représentent les variables utilisées.
- Prétraitement des données : Cette étape consiste à préparer les données pour la phase de classification. Elle englobe la description des données, l'élimination des données manquantes, traitement des données aberrantes et une éventuelle normalisation de données.
- Création du modèle d'apprentissage : lors de cette phase, nous choisissons l'algorithme d'apprentissage automatique puis nous entamons à l'apprentissage du modèle.
- Evaluation et dépoilement du modèle : la phase finale consiste à évaluer le modèle et à optimiser les paramètres des algorithmes. Le recours à l'expertise est ici crucial. Enfin, il s'agit de visualiser les résultats obtenus et d'utiliser le modèle pour prédire de nouvelles situations.

Ce processus est classique et ne met pas en évidence les compétences requises dans le domaine médical. Ces compétences sont nécessaires pour valider nos résultats dans chaque étape décrite ci-dessus. Dans le cadre de notre étude, nous traitons des données spécifiques au domaine médical et qui sont très particulières. Afin de comprendre et savoir comment exploiter ces données, nous impliquons l'aide des experts. L'implication des experts est donc primordiale et son rôle sera en grande partie dans la phase d'analyse des facteurs qui influencent les DDS et de validation des résultats. Nous proposons donc le positionnement de l'expert dans le processus de classification automatique présenté dans la figure suivante.

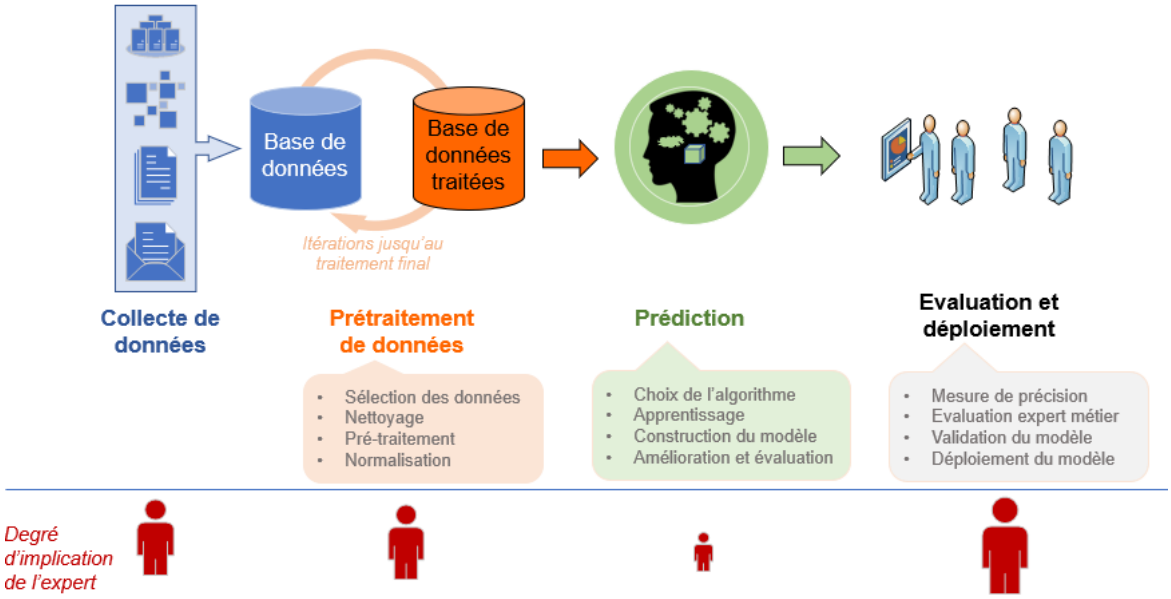

Fig1 : Processus de classification automatique.

La figure précédente montre le degré d'implication de l'expert médical dans chaque étape du processus de classification automatique. La phase principale de ce processus est la phase de prétraitement de données. En effet, le résultat de cette étape est le point de départ des algorithmes de prédiction employés par la suite. Cette étape dépend fortement de l'ensemble de données, dans la section qui suit, nous présentons les différents problèmes rencontrés lors du prétraitement des données médicales.

4. Données médicales dans le processus de classification automatique

4.1 Description de la base de données

Pour la prédiction des durées de séjours des patients au moment de l'admission, nous avons pris l'ensemble de données de « *Microsoft Length of stay prediction* ». Cet ensemble de données est disponible en open source. Il contient 28 variables dont « *Length of stay* » qui représente le nombre de jours d'hospitalisation du patient. Le nombre d'observations de l'ensemble de données est 100000.

La base de données contient des attributs de type catégoriel comme le genre du patient (Homme ou Femme) et d'autres de type continue discret comme les mesures des analyses du laboratoire (Hématocrite). Le traitement de chaque type est différent. Dans ce qui suit, nous allons mettre en évidence les difficultés rencontrées lors du traitement de cet ensemble de données en particulier, et des données médicales en général pour chaque étape du processus de classification automatique.

4.2 Cas d'étude : Ensemble de données Microsoft.

A. Collecte de données.

Pour construire l'ensemble de données, la première difficulté rencontrée est celle de ***l'accès aux données***. En effet, plusieurs lois qui s'imposent liées à la protection des données de la vie privé des patients. Plus

particulièrement, en France, l'utilisation des données de Programme de Médicalisation des Systèmes d'Informations (PMSI) est soumise à des règles relatives à la confidentialité des informations médicales conservées sur support informatique ou transmises par voie électronique (Le service public de la diffusion du droit, 2007). Il existe également d'autres article judiciaire selon (Système National des Données de Santé, 2019) pour définir les permissions requises afin d'accéder aux données.

La deuxième difficulté dans cette étape concerne ***la disponibilité des données*** au moment de l'admission. En effet, plusieurs modèles d'apprentissage automatique se basent sur des ensembles de données contenant des rapports de sortie, de diagnostic médical, des complications suite à l'intervention, etc (Maharlou *et al.*, 2018). Ces variables ne peuvent être disponibles qu'à la fin du séjour du patient. L'estimation de la DDS nécessite donc toutes ces variables au départ. Dans cette étape, il est important de faire appel aux experts dans le domaine. L'ATIH (Agence Technique de l'Information sur l'Hospitalisation) a mis à disposition des groupes d'experts d'information médicale pour accompagner les chercheurs dans leurs travaux (Agence technique de l'information sur l'hospitalisation, 2019a).

Les méthodes que nous allons décrire par la suite sont appliquées sur l'ensemble de données de Microsoft décrit auparavant.

B. Prétraitement des données.

Cette phase est la plus critique. C'est la base de données résultante qui sera l'entrée de notre algorithme de classification. Les principales tâches à effectuer se résument dans ce qui suit :

- *Elimination des données manquantes* : Ce sont les données qui ne sont pas disponible lors du prétraitement. L'ensemble de données utilisé de « *Microsoft Length of stay prediction* » ne comporte pas de variables avec des données manquantes par conséquent, aucun traitement n'a été nécessaire.
- *Données non équilibrées* : les classes des variables sont déséquilibrées. Il existe des variables avec une classe majoritaire et d'autre minoritaire. Pour remédier à ce problème, une méthode qui consiste à équilibrer artificiellement les données est appliquée. Ceci mène à mettre autant d'observation pour chaque classe de l'ensemble de données. Le tableau suivant illustre ce cas pour la variable « *pneum* ».

Classe : 0	Classe : 1
96055 observations	3945 observations

Table 1 : Répartition des catégories de la variable « pneum ».

- *Distribution des données* : L'étude de la distribution des données est primordiale pour des opérations ultérieures. Le but est d'avoir une distribution gaussienne, afin de pouvoir appliquer des méthodes de normalisation des données qui prennent comme hypothèse de départ la distribution normale des données. Certaines variables dans notre base de données ne suivent pas cette distribution. Ce qui rend leur traitement plus complexe. Nous employons donc une transformation à ces variables pour rendre leur distribution sous une forme plus proche d'une gaussienne. La méthode souvent utilisée pour la transformation des données est la fonction logarithmique. Nous retrouvons également d'autres méthodes de transformation de données comme la racine carrée et l'exponentielle. Pour voir l'allure de la distribution de nos variables, nous avons tracé les histogrammes de chacune. L'exemple suivant montre une distribution gaussienne de la variable « *hematocrit* » (à droite) et une distribution non gaussienne de la variable « *neutrophils* » (à gauche).

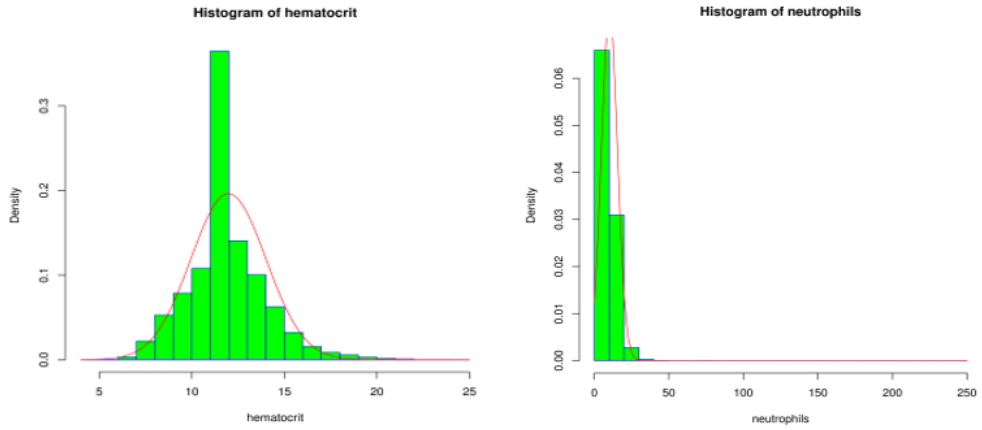

Fig 2 : Histogramme de la distribution des données : « hematocrit » et « neutrophils ».

L'histogramme de la variable « neutrophils » après transformation en utilisant la fonction logarithmique est le suivant :

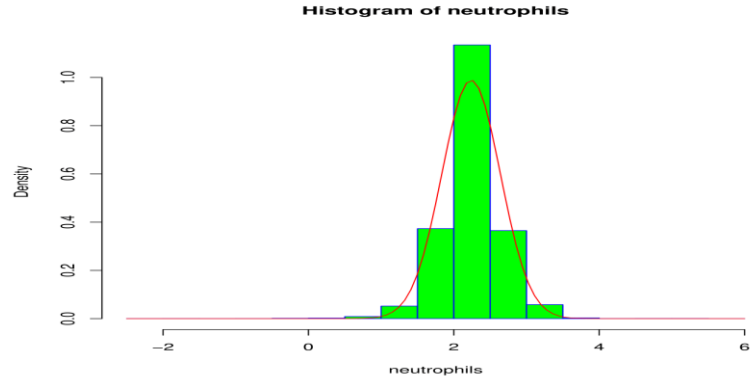

Fig 3 : Distribussion de la variable « neutrophils » après transformation.

- *Normalisation des données* : afin de remettre les valeurs des variables sur la même grandeur d'échelle, une normalisation des données s'applique. Plusieurs méthodes existent comme la Z-score et la Min Max normalisation. Nous avons utilisé la normalisation Z-score nommée aussi centrer-réduire.

- *Traitement des données aberrantes* : La présence de ces données influence leur qualité. Dans le domaine médical, les mesures des analyses du laboratoire présentent beaucoup de données aberrantes. Afin de décider si nous devons les éliminer ou de les transformer, nous devons nous référer à un expert de l'information médicale. Le diagramme suivant illustre cette situation pour la variable « *hematocrit* ».

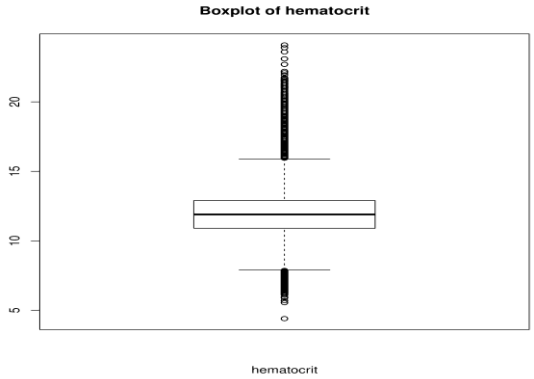

Fig 4 : Boxplot pour la variable « hematocrit ».

- *Codification des variables* : Le choix de l'algorithme d'apprentissage dépend du type des variables, le type de la classe à prédire et d'autres paramètres. La codification des variables est donc importante dans la construction du modèle. Pour la base de données « *Microsoft Length of stay prediction* », nous avons codifié nos variables catégorielles en valeur numérique. Pour le genre par exemple, F <- 0 et M <- 1. Pour la variable « rcount » qui contient plus de deux catégories, la codification est comme suit :

« 0 »	« 1 »	« 2 »	« 3 »	« 4 »	« 5+ »
0	1	2	3	4	5

5. Conclusion

L'intérêt de la prédiction de la durée de séjour dans un domaine hospitalier prend plus d'ampleur ces dernières années. Dans cet article, nous avons présenté l'utilité de la prédiction des DDS. Nous avons cité plusieurs travaux dans ce contexte et avons exposé la démarche méthodologique pour un processus de classification automatique classique. Pour mettre en œuvre notre démarche, nous avons utilisé la base de données de Microsoft qui est en open source. En ce qui concerne cette démarche, nous nous sommes focalisés sur la mise en évidence des différentes difficultés rencontrées lors de la manipulation des données médicales.
Cet article ouvre plusieurs perspectives. Les principaux axes de recherche sont le prétraitement des bases de données pour mieux les représenter. Aussi, concevoir et mettre en place une solution permettant l'interaction entre différents acteurs (chercheurs universitaires, professionnels de santé, économistes) pour résoudre les problèmes dues au conflit d'intérêts.

6. Remerciements

Ce travail de recherche est co-financé par la région Hauts-de-France (fonds FEDER) et la société Alicante (https://www.alicante.fr/). Nous les remercions vivement pour leur soutien et leurs contributions. Nos remerciements s'adressent également au GHICL pour son implication dans la rédaction de ce papier.

7. Références bibliographiques

Agence technique de l'information sur l'hospitalisation (2019a) 'Experts information médicale | Publication ATIH'.

Agence technique de l'information sur l'hospitalisation (2019b) *Médecine, chirurgie, obstétrique : Chiffres clés*.

Agence Technique de l'Information sur l'Hospitalisation, A. (2018) *Hosipitalisation chiffres: Hauts-de-France*.

Benbelkacem, S. *et al.* (2019) 'Machine Learning for Emergency Departement Management', *International Journal of Information Systems in the Service Sector*, 11(3).

Chuang, M., Hu, Y. and Lo, C. (2018) 'Predicting the prolonged length of stay of general surgery patients : a supervised learning approach', *INTERNATIONAL TRANSACTIONS INOPERATIONAL RESEARCH*, 25(1), pp. 75–90. doi: 10.1111/itor.12298.

Chuang, M. Te *et al.* (2016) 'The Identification of Prolonged Length of Stay for Surgery Patients', in *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, pp. 3000–3003. doi: 10.1109/SMC.2015.522.

Gentimis, T. *et al.* (2017) 'Predicting Hospital Length of Stay using Neural Networks on MIMIC III Data', in *IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pp. 1194–1201. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191.

Hachesu, P. R. *et al.* (2013) 'Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients', *Healthcare Informatics Research*, 19(2), pp. 121–129.

Johnson, A. E. W. *et al.* (2016) 'MIMIC-III, a freely accessible critical care database', *Scientific Data*, 3, pp. 1–9. doi: 10.1038/sdata.2016.35.

Khosravizadeh, O. *et al.* (2016) 'Factors affecting length of stay in teaching hospitals of a middle-income country', *Electronic physician*, 8(10), pp. 3042–3047. doi: 10.19082/3042.

Lafaro, R. J. *et al.* (2015) 'Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre- Incision Variables', *plos one*, 10(12), pp. 1–19. doi: 10.1371/journal.pone.0145395.

Li, J. S. *et al.* (2013) 'Applying a BP neural network model to predict the length of hospital stay', in *Springer-Verlag Berlin Heidelberg 2013*, pp. 18–29. doi: 10.1007/978-3-642-37899-7_2.

Maharlou, H. *et al.* (2018) 'Predicting length of stay in intensive care units after cardiac surgery: Comparison of artificial neural networks and adaptive neuro-fuzzy system', *Healthcare Informatics Research*, 24(2), pp. 109–117. doi: 10.4258/hir.2018.24.2.109.

Microsoft (2017) *Predicting Hospital Length of Stay*, *file:///C:/Users/rmekhald/Downloads/For%20the%20Data%20Scientist.html*.

RIGAL, M. (2009) *Management des lits et durée moyenne de séjour : Exemple de recherche d ' optimisation au Centre Hospitalier d ' Avignon*.

Le service public de la diffusion du droit (2007) 'Décret n°2007-960 : confidentialité des informations médicales.'

Shea, S. *et al.* (1995) 'Computer-generated informational messages directed to physicians: Effect on length of hospital stay', *Journal of the American Medical Informatics Association*, 2(1), pp. 58–64. doi: 10.1136/jamia.1995.95202549.

Système National des Données de Santé (2019) 'Impact du nouveau cadre juridique sur l'utilisation des données du PMSI', pp. 10–11.

Tsai, P. F. J. *et al.* (2016) 'Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network', *Journal of Healthcare Engineering*, 2016, p. 11 pages. doi: 10.1155/2016/7035463.