Thése de doctorat

Pour l'obtention du grade de Docteur de

l'Université Polytechnique Hauts-de-France et de l'INSA Hauts-de-France

Discipline : **Informatique**

Présentée et soutenue par :

RACHDA NAILA MEKHALDI

Le 27 janvier 2022 à Valenciennes

École Doctorale : Polytechnique Hauts-de-France (ED PHF).
Laboratoire : LAMIH UMR CNRS 8201.

CONCEPTION ET DÉVELOPPEMENT DES MÉTHODES DE PRÉDICTION DE LA DURÉE DE SÉJOUR HOSPITALIER CENTRÉES SUR DES TECHNIQUES DE "MACHINE LEARNING".

Président du jury	:	CRISTIAN PREDA	Professeur, Université des Sciences et Technologies de Lille.
Rapporteur	:	PARISA GHODOUS	Professeur, Université Claude Bernard Lyon 1.
Rapporteur	:	VIRGINIE GOEPP-THIEBAUD	Maître de conférences, HDR, INSA de Strasbourg.
Directeur de thèse	:	SYLVAIN PIECHOWIAK	Professeur, UPHF.
Co-encadrant de thèse	:	SONDES CHAABANE	Maître de conférences, INSA HDF, UPHF.
Co-encadrant de thèse	:	PATRICE CAULIER	Maître de conférences, INSA HDF, UPHF.
Invité	:	DAVID DELERUE	Directeur société Alicante.

Janvier 2022

Ce(tte) œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale 4.0 International.

Résumé

Au cours des dernières années, les établissements de soins cherchent sans cesse à optimiser le fonctionnement de leurs services tout en assurant la qualité de ces services. La Durée De Séjour hospitalier (DDS) est un indicateur d'évaluation du rendement des établissements de soins et d'efficacité de la performance des services hospitaliers. De ce fait, l'estimation de la DDS au moment de l'admission du patient et durant son séjour hospitalier fait l'objet de plusieurs études. 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.

D'abord, une étude bibliographique est réalisée afin de recenser les différents modèles de DDS existants dans un environnement hospitalier. Nous avons ensuite déduit un modèle générique caractérisant la DDS dans plusieurs unités médicales en rajoutant de nouvelles informations définies en se basant sur les besoins quotidiens des hôpitaux. La démarche suivie pour la prédiction de DDS s'appuie principalement sur les techniques d'apprentissage automatique et de fouille de données. Deux différents modèles de prédiction sont proposés. Un modèle statique de prédiction de DDS qui concerne la prédiction de DDS au moment d'admission du patient. Un deuxième modèle ajuste la DDS initialement prédite en intégrant de nouvelles données disponibles au cours du séjour hospitalier. Ce modèle est nommé modèle séquentiel de prédiction de DDS. La complexité des données médicales est une des difficultés principales auxquelles il faut faire face. Les données issues du Programme de Médicalisation des Systèmes d'Informations (PMSI) sont utilisées pour la mise en pratique de nos contributions.

Mots clefs : Durée De Séjour hospitalier (DDS), Programme de Médicalisation des Systèmes d'Informations (PMSI), modèles de DDS, apprentissage automatique, fouille de données.

Abstract

These years, there has been a considerable interest in controlling hospital costs to provide better services. In the context of optimizing health care services and with recent advances in technology, diverse relevant medical information concerning both patient and administrative procedures are available to be exploited in many research areas. These databases are a rich source of information and are extremely interesting. Healthcare institutions, academic researchers and industry organizations in various areas are working in coordination to improve the quality of care and the management of healthcare systems. The Length Of Stay (LOS) is considered as one of the basic indicators to evaluate the performance of care services and care quality which explain the growing interest of predicting the LOS in hospitals. Estimating the LOS at the admission time and during the hospital stay provides an approximation of the patient's discharge date involving an appropriate planification of care activities. As a result, expecting the acute value of the LOS is useful to highlight a planning strategy for the hospital's logistics.

A review of contemporary literature revealed factors impacting the LOS and a generic model of LOS is proposed. This model includes patient's demographic information, medical information and administrative information. Moreover, it includes extra-information related to the period and day of admission and availability of human and material resources. Depending on theses factors, we proposed a model for predicting the LOS based on machine learning process. Thus, two models are presented : the first model predict the LOS on the time of admission and the second model integrate new data available during the hospital stay and so adjust the LOS prediction. The developed models are based on real data from the Program of Medicalization of Information System (PMIS) and can be used for real time LOS prediction in any French healthcare institution.

Keywords : Length Of Stay (LOS) prediction, Information Systems Medicalization Program (PMSI), Length Of Stay characterization, machine learning, Data mining.

Dédicaces

À mes fidèles supporteurs,

Mes **chers parents** pour leur amour inconditionnel, leur soutien sans failles et pour tout ce qu'ils ont fait pour moi. Merci pour m'avoir conduit là où je suis aujourd'hui.

Ma soeur **Selma**, pour son affection, pour sa présence à mes côtés dans les pires et les meilleurs moments et pour m'avoir toujours encouragé.

Mon petit frère **Mohammed Khalil**, pour nos moments de folie, de bonheur et de complicité et pour toute la joie de vivre qu'il m'apporte chaque jour.

Aux membres de ma grande famille, à tous ceux que j'aime et qui m'aiment.

Vous êtes ma source de force pour aller toujours plus loin.

Naila

Remerciements

Je souhaiterais tout d'abord remercier mon directeur de thèse, Professeur **Sylvain PIECHOWIAK** pour ses conseils scientifiques, ses relectures soignées, et sa participation aux travaux de recherche.

Je voudrais également remercier mes co-encadrants de thèse, Madame **Sondes CHABAANE** et Monsieur **Patrice Caulier** pour leur encadrement, leurs encouragements et pour la qualité de leurs conseils et directives scientifiques.

J'exprime ma gratitude à Madame **Parisa GHODOUS**, Professeur de l'Université Claude Bernard de Lyon et à Madame **Virginie GOEPP-THIEBAUD**, Maître de conférences, HDR de l'INSA de Strasbourg, d'avoir rapporté mon manuscrit. Je leur suis extrêmement reconnaissante de l'intérêt qu'elles ont porté à ce travail. Je les remercie également pour les discussions enrichissantes que nous avons eues.

Je remercie Monsieur **Cristian PREDA**, Professeur de l'Université des Sciences et Technologies de Lille, d'avoir accepté de présider mon jury de thèse et m'avoir fait l'honneur d'assister à ma soutenance afin d'évaluer mon travail. Je le remercie également pour ses remarques pertinentes.

Je tiens à remercier la **région Hauts de France** et le **programme FEDER** pour avoir financé cette thèse. Mes remerciements vont également à l'entreprise **Alicante** qui ont également participé dans le financement de cette thèse.

Durant mes années de thèse, j'ai eu la chance de connaître des personnes formidables. Des personnes avec qui je garde de beaux souvenirs et qui ont été complices des bons moments. Je tiens à remercier toutes ces personnes et tous mes amis en France pour leurs encouragements, leur soutien tout au long de cette thèse. Je les remercie également de m'avoir permis de me changer les idées au cours de cette période.

Table des matières

Introduction générale

> *Nothing in life is to be feared, it's only to be understood.*
>
> — **Marie Sklodowska-Curie**
> Prix Nobel de physique et de chimie.

Introduction

Avec la numérisation des informations médicales, la quantité de données stockées dans les Systèmes d'Informations Hospitaliers (SIH) ne cesse d'augmenter désormais. Une grande richesse et une importante diversité de ces données est présente. Il est nécessaire de les exploiter pour la recherche médicale. Les données médicales sont utilisées par les professionnels de santé, les chercheurs académiques, les industriels dans le but d'améliorer les systèmes de santé actuels.

En outre, la hausse du nombre des patients que connaît les établissements de soins est coûteuse pour les systèmes de santé en terme économique et clinique. Les montants d'hospitalisation varient en fonction des unités médicales mais restent toujours élevés. En France, le nombre d'hospitalisation a atteint 7,1 millions durant l'année 2019. Ce qui est équivaux à 10,6 millions de séjours hospitaliers [Age20]. Les informations concernant ces séjours hospitaliers sont stockées dans les SIH et utilisées pour améliorer l'organisation des hôpitaux, l'optimisation de leurs ressources et le suivi des parcours des patients.

Contexte et problématique

Les établissements de soins, constituent des systèmes socio-techniques complexes au sein desquels interagissent, d'une part, des services de spécialités médicales (chirurgie, pharmacie, biologie) et, d'autre part, de soutiens transversaux (administratif, financier, logistique). Dans le contexte des établissements de soins actuel et institutionnel, les services hospitaliers ont tenté d'optimiser leurs ressources afin d'améliorer leur rendement. Des indicateurs d'efficacité des systèmes de santé se sont imposés à savoir le taux de mortalité, le taux de réadmission et la Durée De Séjour hospitalier (DDS). La DDS est un des indicateurs de base de l'évaluation de l'efficacité des services hospitaliers [AMB18]. Nous nous y sommes particulièrement intéressés. La DDS représente l'intervalle de temps entre l'admission du patient et sa sortie. Face aux besoins sanitaires croissants de la population, à la surcharge du travail des professionnels de santé et à l'allongement des délais d'attente des patients, l'estimation de la DDS doit être établie au moment de l'admission du patient, suivie et mise à jour tout au long du séjour hospitalier. La prédiction de la DDS contribue à :

- La planification des activités de soins des services médicaux.
- L'amélioration de l'organisation de l'hôpital.
- L'analyse des flux patient et le suivi de leur parcours.

- La gestion des lits hospitaliers.
- L'optimisation des ressources matérielles et humaines de l'hôpital.

Un des avantages les plus importants de la prédiction de la DDS est la maîtrise des contraintes budgétaires à la quelle les hôpitaux sont tenus. En France, le Programme de Médicalisation des Systèmes d'Informations (PMSI) est développé pour calculer les allocations budgétaires attribuées aux établissements de soins d'une manière équitable.

Les données issues du PMSI représentent une source immense d'informations. Ces données regroupent les données médicales des patients, les séjours hospitaliers, l'organisation des services de soins ainsi que des données économiques. Elles proviennent de plusieurs sources et sont particulièrement difficile à gérer. Effectivement, elles sont hétérogènes, volumineuses et contiennent énormément de données incomplètes, erronées, imprécises et atypiques. De plus, elles représentent des données à caractère personnel. Dès lors le respect de leur confidentialité est indispensable. En outre, les données du PMSI ne sont pas disponibles en même temps dans les SIH. Elles se complètent en fonction de l'évolution du séjour hospitalier.

Plus particulièrement, un défi crucial est de mettre en place un modèle de prédiction de Durée De Séjour hospitalier à chaque admission d'un patient. De plus, à chaque arrivée d'une nouvelle information sur le séjour du patient, la prédiction de la DDS doit être affinée en rajoutant les nouvelles informations dans les SIH.

La prédiction de la Durée De Séjour hospitalier est un problème complexe, de part par la nature dynamique du milieu hospitalier et d'autre part, par la richesse et la diversité des données disponibles dans les SIH. Ainsi, pour concevoir un outil de prédiction de DDS au moment de l'admission du patient et tout au long du séjour hospitalier, il est primordial d'analyser les différents paramètres constituant le milieu hospitalier, particulièrement ceux qui constituent un séjour hospitalier. La définition de la DDS est alors liée à la définition des différents éléments composant un séjour hospitalier.

Objectifs

L'objectif principal de la thèse est la conception d'un outil intelligent capable de prédire la durées de séjour hospitalier à partir des données disponibles dans les SIH. Cet outil se base sur des données réelles stockées dans les SIH et issues du PMSI. Il est capable d'assurer la prédiction de la DDS au moment de l'admission du patient et au fil du séjour hospitalier. La DDS est considérée comme une variable complexe

et dépend de plusieurs facteurs. La Durée De Séjour peut être caractérisée par l'état médical du patient, ses informations démographiques, la gestion administrative de l'établissement de soins et sa gestion économique. Le motif d'hospitalisation du patient, son historique médical, les complications médicales sont les informations centrales de la prédiction des DDS. Les informations démographiques comme l'âge du patient, sa situation familiale, son sexe et son adresse sont également prises en compte dans la définition du DDS. Comme la DDS participe à la planification des activités des hôpitaux et à la logistique hospitalière, il est primordial d'intégrer des données administratives dans les modèles de prédiction de la DDS. Ces données administratives peuvent inclure le type de l'unité médicale dans laquelle le patient est admis, ses conditions d'entrée et de sortie et les transferts réalisés entre les unités médicales. La prédiction de DDS vise aussi à maîtriser les contraintes budgétaires des hôpitaux. Des informations liées à la gestion économique de l'hôpital sont considérées. Les données de facturation, les modalités de remboursements et le type d'assurance du patient sont intégrées.

La brève introduction précédente permet de déterminer les objectifs de ce projet de thèse. Une démarche méthodologique est mise en place pour concevoir le modèle de prédiction de DDS. Nous présentons dans la figure suivante cette démarche.

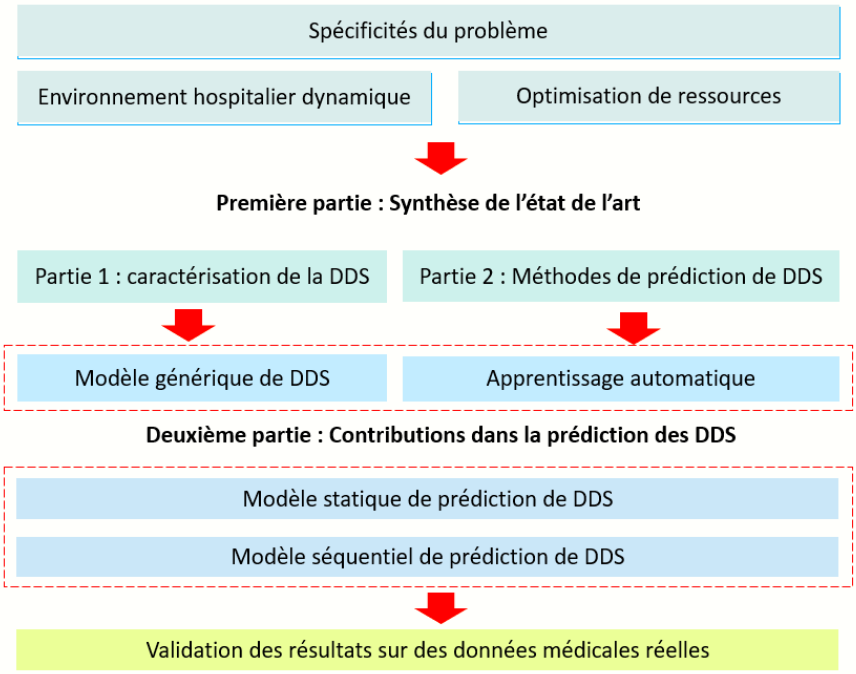


Fig. 0.1: Démarche méthodologique de prédiction de Durée De Séjour hospitalier.

Une nouveau modèle de prédiction de DDS est développée. Ce modèle s'appuie sur des techniques issues du domaine de l'intelligence artificielle à savoir l'apprentissage automatique et la fouille de données. Cette thèse vise à développer un outil de

prédiction de DDS pour contribuer à l'optimisation des ressources hospitalières et à la planification des activités de soins. Ce modèle s'appuie sur des données réelles issues du PMSI. Une première étape consiste à caractériser la DDS par un modèle générique. Ce modèle englobe plusieurs facteurs impactant la DDS définis selon la bibliographie, les besoins des hôpitaux et les évènements possibles de la vie quotidienne. Il constitue l'entrée aux méthodes de prédiction à base d'apprentissage automatique.

La section suivante détaille les contributions liées à ces objectifs.

Contributions

Pour parvenir à répondre aux objectifs tracés pour cette thèse, il faut présenter une solution à divers problèmes liés entre eux. Par définition, la DDS est fortement associée à une unité médicale ou une pathologie. De ce fait, nous avons d'abord déterminé un périmètre d'étude de la DDS. Ensuite, nous avons analysé les travaux précédents pour déterminer les facteurs qui impactent la DDS dans un milieu hospitalier. L'ensemble des facteurs déduits est l'entrée aux modèles de prédiction. La phase de prédiction de DDS emploie des techniques d'apprentissage automatique. Nous avons proposé deux modèles de prédiction. Un premier modèle pour prédire la DDS au moment de l'admission du patient. Un second modèle est proposé afin d'intégrer de nouvelles données disponibles tout au long du séjour. Nous résumons les principales contributions de nos travaux dans ce qui suit :

1. Proposition d'un modèle générique de DDS. Ce modèle considère plusieurs facteurs dans différentes unités médicales en complément à d'autres variables définies à partir des besoins quotidiens des hôpitaux.
2. Étude de la DDS dans plusieurs unités médicales. Les unités médicales choisies se distinguent d'une part, par les profils des patients, et d'autre part, par leur gestion hospitalière.
3. Modèle statique de prédiction de DDS : proposition d'un modèle de prédiction de DDS au moment de l'admission du patient en adaptant les méthodes d'apprentissage automatique pour faire face à la complexité présente dans l'ensemble des données utilisées dans l'étude.
4. Modèle séquentiel de prédiction de DDS : proposition d'un modèle de prédiction de DDS à différents instants du séjour hospitalier. Ce modèle affine la DDS en intégrant de nouvelles données disponibles après l'admission du patient. Une nouvelle manière de structurer et codifier ces données comme une succession d'évènements est exposée.

5. Mise en place des méthodes proposées en utilisant un ensemble de données réelles issues du Programme de Médicalisation des Systèmes d'Informations (PMSI).

Ces apports sont expliqués en détails dans le présent mémoire.

Guide de lecture

Le mémoire de thèse est organisé en cinq chapitres.

Dans le chapitre 1, nous introduisons les Systèmes d'Informations Hospitaliers (SIH) et démontrons la richesse et la complexité des informations stockées. Nous présentons, plus particulièrement, le Programme de Médicalisation des Systèmes d'Informations (PMSI) utilisé. Afin d'évaluer les performances des hôpitaux pour améliorer la qualité et l'efficacité des soins, nous nous sommes intéressé au séjour hospitalier. La Durée De Séjour hospitalier (DDS) est introduite comme un indicateur d'évaluation de base des performances des systèmes de santé. Nous clôturons le chapitre en mettant en évidence l'application des techniques d'intelligence artificielle et particulièrement les méthodes d'apprentissage automatique dans le domaine de la santé en général.

Dans le chapitre 2, nous présentons une synthèse critique des travaux liés aux problématiques de prédiction des DDS. Nous présentons les différents modèles caractérisant la DDS et les méthodes utilisées pour sa prédiction. Afin de bien étudier les travaux existants et pour positionner notre travail parmi eux, nous présentons d'abord une définition de la DDS. Les facteurs qui l'impactent dans un milieu hospitalier sont ensuite déterminés. Un modèle générique de DDS en considérant d'autres facteurs que nous avons jugés importants dans l'étude de la DDS est conçu. Les méthodes de prédiction de DDS sont exposées en mettant l'accent sur celles qui ont montré leur efficacité. Sur la base des limites rencontrées, nous exposons notre proposition.

Le chapitre 3 expose la démarche méthodologique de la prédiction des DDS au moment de l'admission du patient. Il comporte les contributions apportées dans le cadre d'un modèle statique de prédiction de DDS en considérant les données disponibles au moment de l'admission du patient. Nous exposons en détail les formalismes utilisés et les différentes étapes de l'approche.

Le chapitre 4 concerne le modèle séquentiel de prédiction de DDS. Ce modèle en complément du premier proposé, prend en considération de nouvelles variables dis-

ponibles après l'admission du patient et tout au long du séjour hospitalier. Plusieurs algorithmes sont exposés. Pour mettre en place nos algorithmes, nous proposons une technique pour structurer et codifier les données séquentielles tout en préservant la chronologie de leur arrivée. De ce fait, l'aspect temporel des données médicales issues du PMSI est mis en place.

Finalement, le chapitre 5 traite l'application de nos contributions sur la base de données issues du Programme de Médicalisation des Systèmes d'Informations (PMSI). Différentes unités médicales sont sélectionnées dont le service de cardiologie, le service de pédiatrie, le service de médecine polyvalente et le service de néonatologie. Nous présentons et discutons les résultats obtenus.

Nous terminons ce rapport par une conclusion générale en analysant les résultats obtenus et présentant les perspectives pour nos futurs travaux.

1

Les systèmes d'informations hospitaliers et la gestion hospitalière.

> *I was taught that the way of progress was neither swift nor easy.*
>
> — **Marie Sklodowska Curie**
> Prix Nobel de physique et de chimie.

1.1 Introduction

De nos jours dans la plus part des pays, les établissements de soins font face à une forte croissance du nombre des patients admis. Ceci a engendré une augmentation de la consommation des ressources dans les services concernés. Les hôpitaux sont amenés à maîtriser l'utilisation de leurs ressources et à exploiter la masse de données collectées pour assurer l'efficacité et l'efficience de la qualité des services de soins proposés. De ce fait, les banques de données médicales et administratives sont apparues et leur exploitation est devenue importante dans la recherche en santé. De manière générale, les Systèmes d'Informations Hospitaliers (SIH) s'occupent de la gestion de l'ensemble des informations, de leurs règles d'utilisation et de leur circulation. De plus, ils font face au stockage et au traitement des données pour répondre aux besoins quotidiens des établissements de soins [Com20]. Les performances et la qualité des services de soins reposent sur la qualité et la quantité des informations collectées dans les SIH. La Durée De Séjour hospitalier (DDS) constitue un des indicateurs d'évaluation le plus utilisé et sa prédiction basée sur les données disponibles dans les SIH a été au centre d'un grand nombre de travaux de recherche. Le problème de la prédiction des durées de séjours hospitaliers a été abordé sous différents angles dans les différentes recherches précédentes.

L'objectif de ce chapitre est de présenter la terminologie utilisée et les concepts fondamentaux nécessaires au positionnement de nos travaux et à la compréhension du présent mémoire.

Ce chapitre est consacré aux SIH et il est organisé comme suit. La première partie présente les données stockées dans les SIH et leurs sources. La seconde partie présente les principales propriétés des données médicales. La troisième partie détaille le Programme de Médicalisation des Systèmes d'Informations (PMSI). Le PMSI est un sous-système du SIH implémenté dans tous les hôpitaux Français. Ensuite, des exemples d'application des méthodes d'Intelligence Artificielle, et particulièrement l'apprentissage automatique dans le domaine de la santé sont exposés. Enfin, la durée de séjour hospitalier est introduite comme un critère d'évaluation de base de la qualité des systèmes de santé.

1.2 Les systèmes d'informations hospitaliers

Les progrès considérables en traitement de l'information ont permis une augmentation des collectes des données dans toutes les organisations. Le domaine de la santé n'a pas échappé à cette croissance. Ceci a incité de nombreux hôpitaux à

développer et mettre en place un système d'information des données médicales et administratives appelé Système d'Informations Hospitalier (SIH). De plus, le passage du traitement et du stockage sur papier au traitement et au stockage sur ordinateur a montré l'efficacité de la mise en place des SIH dans l'organisation de la structure des soins [Elg05]. Une comparaison entre les deux méthodes de stockage a montré les avantages des systèmes d'informations hospitaliers qui appuient leur besoin et leurs objectifs [Deg13]. D'abord, le dossier papier est un support volumineux et lourd ce qui rend son transport entre les services médicaux lent contrairement au SIH qui dispose d'une méthode de compression du volume de l'information et de ce fait, le temps nécessaire à la collecte d'informations et à son transport diminue. L'avantage des SIH est sa capacité à partager les informations entre plusieurs services, à l'opposé du dossier papier qui n'est pas partageable. Enfin, le système numérisé permet l'acquisition, la sauvegarde et le traitement des informations médicales automatiquement [Elg05].

D'une manière générale, on appelle système d'information l'ensemble des outils matériels, des logiciels et des réseaux de télécommunications utilisés pour recueillir, créer et distribuer des données utiles dans des organisations [VS08]. En particulier, un Système d'Informations Hospitalier (SIH) désigne un système conçu pour gérer l'ensemble des données médicales et administratives d'un hôpital. Il se constitue d'un groupe d'éléments en communication qui rassemblent, traitent et fournissent les informations nécessaires à son activité [Deg13].

Dans le milieu hospitalier, plusieurs types d'informations peuvent être distingués. Les informations disponibles concernent les informations sur l'état de santé du patient et les informations administratives. Elles se présentent sous format électronique connu comme le dossier médical électronique (DME) du patient. Ces informations comprennent les données démographiques sur les patients, les étapes de son suivi, les complications, les médicaments, les signes vitaux, les antécédents médicaux, les immunisations, les données de laboratoire et les rapports de radiologie [HIM]. Les informations administratives concernent la gestion opérationnelle d'un hôpital en matière de soins de santé. Elles englobent les informations de la gestion des patients (parcours, facturation, actes médicaux), la gestion de la finance et de la comptabilité (budget, ressources matérielles, achats) et la gestion des ressources humaines (affectations, planning, paye).

Le Système d'Informations Hospitalier se compose principalement de trois sous-systèmes qui communiquent entre eux afin d'assurer une meilleure structuration et organisation du SIH [Deg13]. La figure 1.1 illustre ces sous-systèmes qui sont : le sous-système de production des soins, le sous-système d'information logistique et le sous-système de pilotage.

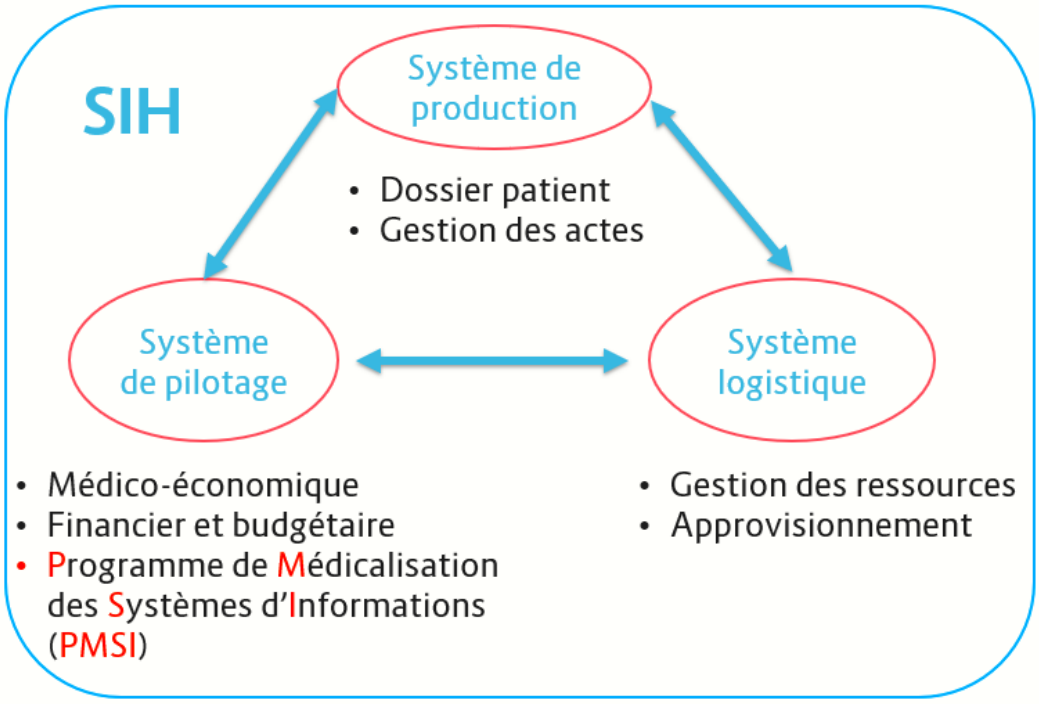


Fig. 1.1: Composantes des Systèmes d'Informations Hospitaliers [Deg13].

- Le sous-système de production : Ce volet s'occupe de l'administration des données patients, les unités de soins, la communication entre ces unités et la gestion de la recherche et de l'enseignement médicaux. Il contient toutes les données liées au patient comme par exemple : le diagnostic médical, les prescriptions et la réalisation des actes médicaux, l'édition des comptes rendus et les résumés de dossier sont présentes au sein de ce sous-système.

- Le sous-système d'information logistique : L'objectif est donc de mieux organiser les activités et les structurer afin d'assurer une meilleure qualité de soins des patients [BBB20]. Le sous-système d'information logistique permet de gérer les différents ressources matérielles, humaines, physiques et financières de l'hôpital. Il englobe la gestion de stocks et des approvisionnements, la gestion des locaux, la gestion des facturations et des commandes, la gestion des lits d'hospitalisation et de soins ainsi que les archives et la documentation des établissements de soins.

- Le sous-système de pilotage : Il veille à la prise en charge de la gestion médico-économique de l'hôpital. Il concerne la qualité des soins et la gestion des risques. De plus, il s'intéresse à l'allocation budgétaire des différentes unités de soins [FAU21]. En France, le Programme de Médicalisation des Systèmes d'Informations (PMSI) est mis en place comme un sous-système de pilotage dans les établissements de soins, qu'ils soient du secteur public ou privé.

Ces sous-systèmes sont souvent en interaction afin d'assurer la continuité des services de soins, améliorer leur qualité et gérer les ressources et les contraintes budgétaires. Compte tenu du grand volume des données des SIH, divers formats de stockage sont apparus. Ces données proviennent de multiples sources et font l'objet de plusieurs études dans le domaine médical. Dans la section suivante, nous tentons de répondre aux questions suivantes : quelles sont donc les sources de ces données ? quel est leur contenu ?

1.3 Sources des données des SIH

Les progrès technologiques et les progrès des processus de traitement des données ont permis une augmentation exponentielle de la quantité des données collectées dans le domaine de la santé. Le volume des données contenues dans les SIH ne cessent de croître. En fonction de leur type, les données sont recueillies à partir de différentes sources. Ces sources de données sont nombreuses et diffèrent selon le type de collecte, le format de représentation et la nature des informations. Les principales sources des données médicales sont : les dossiers médicaux, les enquêtes auprès des patients et les données administratives utilisées pour payer les factures ou gérer les soins. Dans ce qui suit, nous détaillons les sources de données.

1.3.1 Dossier médical du patient

L'avènement des dossiers médicaux électroniques (DME) améliore l'accessibilité aux dossiers des patients. L'idée du dossier médical électronique du patient est apparue au début des années 1970 [SW73]. L'objectif est de responsabiliser les patients et leur permettre de s'engager dans la continuité des soins, le choix du traitement et la réduction des erreurs médicales [JK13 ; SW73]. Le contenu des dossiers médicaux électroniques est très large et varié. Il comporte les données démographiques du patient acquises au moment de son admission : sa date de naissance, son adresse, son statut marital et son sexe. Il contient également les données liées à son état de santé comme les résultats des analyses biologiques et les transcriptions médicales, les résultats d'examens radiologique, le diagnostic médical, les antécédents médicaux et les rapports textuels cliniques [PHS14].

1.3.2 Les données administratives

Dans le cadre de la prestation et du paiement des soins, les établissements de santé génèrent des données administratives concernant les services et les frais liés à ces services. Ces données portent sur les types d'assurances des patients ou d'autres

demandes de paiement et peuvent inclure des diagnostics, des procédures, des prescriptions médicales et le détail des appareils médicaux utilisés dans le traitement. Elles peuvent inclure les données des facturations et des remboursements des séjours hospitaliers des patients. Les données de facturation sont souvent liées aux motifs d'hospitalisation représentés à l'aide de la Codification Internationale des Maladies (CIM) et aux procédures que le patient a subi au cours de son séjour [PHS14]. Les données administratives comportent aussi des informations sur le type de l'unité médical, l'admission du patient, le nombre d'unités dans lesquelles le patient est passé (ou le nombre de jours passés dans chaque unité) [Age18].

1.3.3 Les données issues des enquêtes et de la recherche clinique

Une source importante des données médicales est apparue avec l'explosion de l'utilisation d'internet comme moyen de communication. Les données de santé peuvent provenir des échanges des patient sur les réseaux sociaux et des recherche effectuées sur le web [PHS14]. Elles proviennent également des études cliniques [INS16] réalisées par les professionnels de santé, les scientifique et les industriels.

Des méthodes d'intégration et de stockage des données médicales sont mises en place afin de faciliter leur sauvegarde et leur traitement et manipulations ultérieures. En France, le Programme de Médicalisation des Systèmes d'Informations (PMSI) est mis en place et il est considéré comme une source cruciale des données de santé. Il vise à définir l'activité des établissements de soins et à calculer les allocations budgétaires équitables basée sur l'activité des services hospitaliers. Les données médicales en général, et celles issues du PMSI particulièrement se distinguent d'autres types de données par certaines propriétés. Dans ce qui suit, nous présentons les principales propriétés des données médicales.

1.4 Propriétés des données médicales

Les données médicales sont des sources importantes pour de nombreuses recherches académiques et industrielles. La collecte, l'utilisation et l'exploitation de ces données constituent un enjeux principal dans l'extraction de connaissances, l'analyse, la fouille de données et la conception des systèmes de décision médicaux. De manière plus large, l'information est la clé d'une meilleure organisation et le point de départ de nouveaux développements dans plusieurs systèmes. Les progrès technologiques tels que l'augmentation des espaces de stockage de l'information médicale nous ont aidés à générer de plus en plus de données. L'utilisation des données médicales

participe alors à l'organisation des établissements de soins, l'identification de profils homogènes de patients, le suivi des parcours des patients et la recherche de leur diagnostic médical. Cependant, avant d'utiliser ces données, il est primordial de procéder à leur annotation, de les intégrer et de les pré-renseigner de manière appropriée afin de faciliter leur compréhension. La compréhension et la manipulation des données médicales se heurtent à des défis liés à leur complexité, la richesse des informations aussi qu'à des contraintes de confidentialité. Dans ce qui suit, nous présentons les principales caractéristiques et propriétés des données médicales en vue de leur exploitation par des procédures automatisées.

La figure 1.2 illustre et résume d'une façon générale les propriétés des données médicales et leurs ressources citées auparavant. Elle met en évidence les différentes difficultés auxquelles il faut faire face lors de leur utilisation. Les données du secteur médical sont alors démarquées par leurs propriétés contrairement à d'autres types de données. Particulièrement, les données issues du PMSI que nous avons employées dans notre étude présentent toutes ces propriétés.

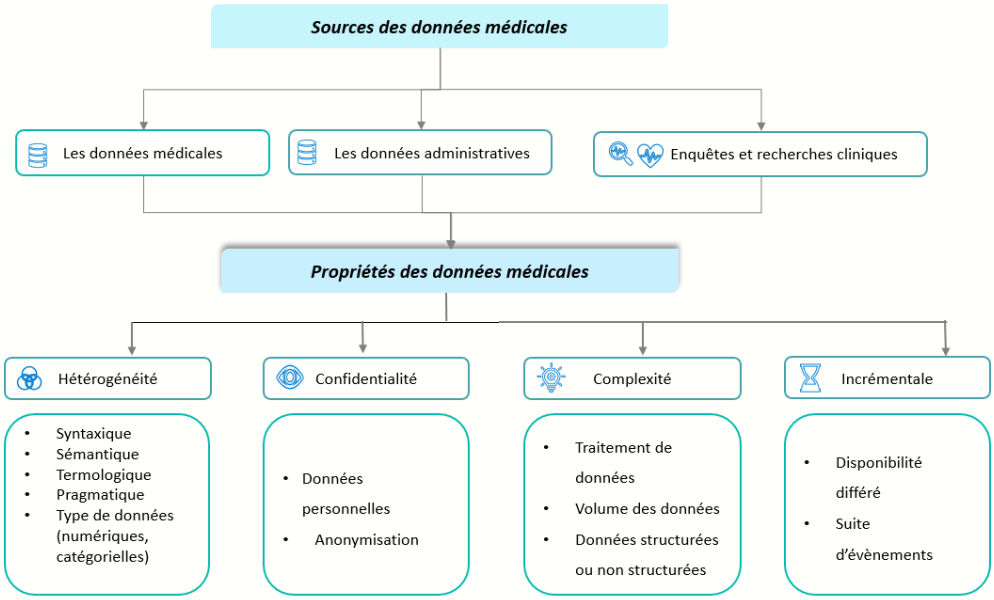


Fig. 1.2: Données médicales : sources et propriétés.

Nous détaillons dans les sections suivantes chaque propriété des données médicales.

1.4.1 Confidentialité

Selon l'article 4 du Règlement Général sur la Protection des Données (RGPD) de l'Union Européenne : « les données relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent

des informations sur l'état de santé de cette personne » sont définies comme données à caractère personnel. Ces donnes doivent donc être protégées et une politique et une démarche de sécurité de ces données doivent être définies pour les protéger. Si la protection des données est un enjeu majeur, d'autres risques liés au matériel et à l'infrastructure informatique sont également des points d'attention récurrents. Pour cette raison, la Commission Nationale de l'Informatique et des Libertés (CNIL) a été créée par la loi Informatique et Libertés du 6 janvier 1978. Son rôle est de veiller à la protection des données personnelles contenues dans les fichiers et traitements informatiques ou papiers, aussi bien publics que privés [CNI19]. Sur le plan juridique, en France, après un examen détaillé de la protection des données et des risques éventuels de ré-identification des personnes du fait du partage des données et de leur utilisation croissante par de nombreux organismes, la loi de modernisation du système de santé de 2016 a été mise en place et donne naissance au Système National des Données de Santé (SNDS). Le SNDS repose sur quatre principes [Tup+17] :

- La préservation des intérêts des personnes physiques.
- Le respect du secret professionnel.
- Le respect de la sécurité des données (Confidentialité, intégrité et confinement) par les responsables du traitement des données.
- La conservation des données pour une durée limitée à 20 ans.

Les données médicales sont exploitées dans plusieurs recherches et études académiques et industrielles. Elles peuvent être utilisées dans la conception des systèmes d'aide à la décision du domaine médical, l'amélioration des prestations de soins de santé, l'optimisation des ressources matérielles et humaines des hôpitaux [HW14]. Un processus d'anonymisation ou de pseudo-anonymisation des données est donc utile avant toute manipulation. L'anonymisation des données médicales est définie comme la suppression de tout caractère identifiant un ensemble de données d'une manière irréversible. Toutes les informations directement ou indirectement identifiables sont supprimées ou modifiées afin d'empêcher toute ré-identification des personnes. Quant à la pseudo-anonymisation, elle permet le retour à l'information originale en cas de besoin particulier. Elle consiste à remplacer les données à caractère personnel par des pseudonymes. Cette technique est réversible et permet donc la ré-identification ou l'étude de corrélations entre les informations codifiées en cas de besoin particulier [DRE15]. De cette manière, la réutilisation des données médicales est possible ce qui suscite un intérêt et une demande croissante.

1.4.2 Données incrémentales

A l'aire du Big data, les données médicales prennent une place importante. L'analyse des données massives est un domaine en pleine croissance qui peut fournir des informations utiles dans le domaine des soins de santé. Ce domaine fait face à plusieurs difficultés dans le secteur de la santé telles que la gestion du volume, la gestion de la variété, la disponibilité des données, leur exactitude, leur intégrité et leur interprétation en continue. Comme les données médicales proviennent de plusieurs sources et sous divers formats, leur collecte est rarement synchrone. Un challenge auquel il faut faire face quant au traitement des données médicales, concerne la gestion des flux et du traitement par les différents dispositifs des établissements et systèmes de santé. L'intégration de l'information médicales ne se limite pas à un instant précis du séjour hospitalier comme le moment de l'admission du patient mais il s'étale tout au long du séjour hospitalier pour s'adapter au quotidien hospitalier. Le séjour hospitalier commence par le moment d'admission du patient et s'achève au moment de sa sortie de l'hôpital. C'est la ou réside la propriété séquentielle des données médicales. L'approche proposée pour la manipulation des données médicales doit prendre en compte cet aspect temporel et assure d'une façon continue l'ajout de nouvelles informations cliniques ou administratives. Dans les systèmes d'aide à la décision ou de prédiction, les éléments collectés sont insérés dès leur disponibilité dans le modèle comme des évènements successifs. Un exemple qui caractérise cette particularité est de modéliser le séjour hospitalier par un processus de trois étapes : moment d'admission du patient, le séjour hospitalier et la sortie du patient [Yan+10]. Au moment d'admission du patient, des informations démographiques comme l'âge, l'adresse, le genre et l'état civil sont acquises ainsi que des informations administratives comme le type d'admission au service concerné, le motif d'hospitalisation et l'unité médicale dans laquelle le patient est admis. Au cours du séjour hospitalier, d'autres informations médicales et administratives s'ajoutent. Par exemple les actes médicaux réalisés pour le patient, les complications médicales et les transferts entre unités médicales. Au final, à la sortie du patient, les rapports médicaux effectués par les médecins ou les infirmiers sont élaborés. Les informations de facturation, la durée de séjour du patient et son mode de sortie sont également établies en fin de séjour.

1.4.3 Hétérogénéité

De nos jours, il est nécessaire d'utiliser conjointement des données provenant de systèmes d'information qui utilisent différentes sources de connaissances comme par exemple, les rapports médicaux textuels et les résultats d'imagerie médicale pour l'enregistrement des données et les utiliser dans la résolution de nombreux

problèmes dans le domaine médical. L'exploration de ces données dites hétérogènes pour extraire des connaissances est un processus fastidieux imposant des contraintes opérationnelles importantes. Les données hétérogènes sont des données dont les types et les formats présentent une grande variabilité [Wan17]. Selon [JO14], il existe principalement 4 types d'hétérogénéité :

- L'hétérogénéité syntaxique : Elle se produit lorsque deux sources de données ne sont pas exprimées dans le même langage.
- L'hétérogénéité sémantique ou conceptuelle : Elle désigne les différences de modélisation d'un même domaine d'intérêt.
- L'hétérogénéité terminologique : Elle désigne les variations de noms lorsqu'on se réfère aux mêmes entités à partir de différentes sources de données.
- L'hétérogénéité pragmatique : Elle correspond à des interprétations différentes des entités.

De plus, nous rajoutons l'hétérogénéité par type de données. Elle réside dans ce cas dans la présence de données quantitatives ou dites numériques et qualitatives ou dites catégorielles. Les données quantitatives sont celles qui peuvent être comptées ou comparées sur une échelle numérique. On distingue alors les données quantitatives continues et discrètes. Pour le type qualitatif, on sépare le qualitatif nominal et le qualitatif ordinal. Par exemple l'âge d'un patient est une donnée numérique discrète, sa taille est une donnée numérique continue, son genre est une donnée catégorielle nominal et son niveau d'étude est une donnée catégorielle ordinale. Nous définissons aussi le type de donnée catégorielle multivaluée comme par exemple les diagnostics médicaux si le patient possède plusieurs diagnostics.

Le format des données médicales peut être structuré ou non structuré. Le format des données structuré est organisé et formaté. Par conséquent, il est facile de saisir, rechercher et manipuler les données structurées. A l'inverse, les données non structurées comme par exemple les rapports médicaux en format textuel ou les images de radiologie médicale, souvent classées comme des données qualitatives, sont plus difficiles à traiter et à analyser [Pic18].

Un processus d'intégration des données hétérogènes est crucial pour permettre aux utilisateurs de définir leurs requêtes sans connaître leurs sources et donner une vue uniforme de l'ensemble de ces sources [Bou+02].

1.4.4 Complexité

La grande quantité d'informations générées par les systèmes d'informations de santé, la variété des sources des données médicales et l'hétérogénéité des données rendent leur traitement et leur analyse plus difficile et plus complexe soulevant ainsi plusieurs défis. Parmi ces défis, nous retrouvons la présence de plusieurs variables ce qui engendre une grande dimension. De plus, ces données sont souvent incomplètes et contiennent des variables fortement corrélées entre elles résultant de la redondance de l'information [CDD01]. Les données médicales présentent également d'autres problèmes comme la présence des données aberrantes ou des erreurs dans les informations enregistrées. Ces problèmes imposent des méthodes de pré-traitement des données avant de les utiliser afin de rendre leur exploitation plus facile et fiable. La complexité des données médicales rend primordial l'implication de l'expertise médicale dans leur exploitation par les utilisateurs afin de valider, interpréter et mieux valoriser leur contenu [NQS17 ; SS13].

Ces propriétés reflète la particularité des données médicales sauvegardées dans les SIH et les sous-systèmes qui le compose. Dans nos travaux de recherche, nous abordons les données du sous-système de pilotage et qui sont issues du PMSI. Par la suite, nous exposons les objectifs du PMSI, sa mise en place et son contexte économique ainsi que son mode de fonctionnement.

1.5 Le Programme de Médicalisation des Systèmes d'Informations (PMSI)

La maîtrise des dépenses hospitalières et les contraintes budgétaires des établissements de soins deviennent de plus en plus pressantes [De +20]. En France, le programme de Médicalisation des Systèmes d'Informations (PMSI) est né en 1985 pour répondre à un double objectif. Il est apparu dans un premier temps dans une phase expérimentale puis dans une phase opérationnelle dès 1989 [Evi89] . Ce programme, considéré comme un outil de description et de mesure médico-économique de l'activité hospitalière, encourage l'échange entre les différents collaborateurs dans les établissements de soins : médecins, soignants et administratifs. Dans ce qui suit, nous présentons les objectifs de la création du PMSI, le contexte de sa mise en place et son fonctionnement.

1.5.1 Objectif du PMSI

Le PMSI est un outil médico-économique qui cible d'un côté l'organisation des soins des établissements de santé et d'un autre côté, leur organisation économique. Le PMSI décrit l'activité hospitalière en rajoutant une dimension médicale dans l'information collectée. Il mesure l'utilisation des ressources et non l'évaluation des soins. Il concerne ce qui est effectivement réalisé et non ce qui devrait être fait [sol02]. Comme la mise en place du PMSI vise à la création d'un système d'allocation budgétaire équitable pour tous les établissements de soins, la relation entre la mesure de l'activité hospitalière et les ressources utilisées est mise en évidence. Le PMSI possède donc deux rôles principaux : donner à l'établissement une visibilité sur son activité et aligner le financement à la nature et au volume de l'activité [Hol15].

D'un point de vue économique, le PMSI permet d'évaluer les coûts des soins sur l'échelle nationale. Cette évaluation est basée sur un critère construit à partir des Groupes Homogènes de Malades (GHM) et est calculée à l'aide de l'Indice Synthétique d'Activité (ISA). La mesure de l'activité de chaque hôpital de cette manière donne la possibilité d'établir l'Echelle Nationale de Coût (ENC) par GHM. Par conséquent, un budget théorique est défini pour chaque établissement, avec un ajustement progressif du budget historique [Elg05]. En se basant sur une meilleure connaissance du volume et de la nature des activités des hôpitaux, des objectifs médicaux du PMSI sont également apparus. Dans le cadre des recherches médicales, le PMSI cible l'amélioration de la qualité des soins et de l'activité médicale en les évaluant. Ceci est réalisé en organisant les dossiers médicaux des patients et le planning des soignants, les stratégies de diagnostic et de soins, en promouvant la recherche épidémiologique et l'aide à la recherche clinique [Gra14].

Avec cette relation entre la gestion médicale et la gestion économique des hôpitaux, chaque institution de soins est dans l'obligation de mieux gérer ses activités afin d'assurer le juste prix des soins. Si les gains sont au-delà de l'objectif, le reste servira à d'autres dépenses comme, par exemple, les investissements ou les recrutements. Le PMSI a révélé aussi les standards des prises en charges des patients et des indicateurs de qualité des soins [Elg05]. Maintenant que les objectifs du PMSI sont présentés, nous expliquons sa mise en place et son contexte économique.

1.5.2 Mise en place et contexte économique

L'objectif principal de la mise en place du PMSI est de calculer le budget qui sera destiné aux unités médicales concernées afin de réduire les inégalités entre les établissements de santé. A l'origine, le PMSI s'inspire d'un modèle américain qui

se base sur une classification de groupes de diagnostics (Diagnosis Related Groups DRG). Ce modèle a été créé par le professeur Robert Fetter et de son équipe de l'Université de Yale [Mor09] aux États Unis. Le but était l'analyse de l'activité hospitalière. Puis, le DRG a été utilisé par l'administration américaine pour effectuer les paiements forfaitaires des séjours hospitaliers des personnes âgées et handicapées. Ensuite, il a été progressivement étendu à plusieurs centres hospitaliers. En France, le PMSI a été introduit comme un dispositif épidémiologique et il a été intégré à la réforme du système de santé dans le milieu des années 80 par Jean de Kervasdoué [FAU21]. L'objectif de d'allocation budgétaire est apparu par la suite [Elg05]. Les données issues du PMSI incluent des données médico-administratives synthétiques recueillies d'une manière normalisée et standardisée. Selon l'activité hospitalière, le PMSI est séparé en 4 champs principaux appelés PMSI-MCO, PMSI-SSR, PMSI-HAD et PMSI-PSY.

1.5.2.1 Médecine, chirurgie, obstétrique et odontologie (PMSI-MCO)

Le PMSI-MCO collecte les informations des séjours hospitaliers des patients admis en médecine, chirurgie, obstétrique ou odontologie [FAU21]. Les séjours peuvent être avec ou sans hébergement ou des soins d'affections graves pendant leur phase aigüe. Le PMSI-MCO comporte des informations administratives, démographiques, médicales et de prise en charge du patient dans les unités médicales dans lesquelles il a été admis. Ces informations sont rapportées lors de la sortie du patient sous format standardisé dans un Résumé Standardisé du Séjour (RSS). Ce RSS est composé d'autant de Résumés d'Unité Médicale (RUM) que d'unités médicales que le patient a fréquenté pendant son séjour. Le RSS est anonymisé par la suite et devient donc Résumé Standardisé Anonymisé (RSA). Une classification à la fois médicale et économique est réalisée en affectant l'ensemble des séjours hospitaliers à un Groupe Homogène de Malades (GHM). Cette classification est une description de l'activité hospitalière et est le fondement de l'optimisation du chiffre d'affaire de l'établissement.

1.5.2.2 Soins de suite et de réadaptation (PMSI-SSR)

Le PMSI-SSR concerne la description médico-économique de l'activité des hôpitaux de l'ensemble des structures ayant une activité autorisée en soins de suite ou de réadaptation [Age13]. Les informations recueillies dans le PMSI-SSR représentent les services suivants : maladies à évolution prolongée, convalescence, repos et régime, rééducation fonctionnelle et réadaptation, lutte contre la tuberculose et les maladies respiratoires, cures thermales, cures médicales, cures médicales pour enfants et postcures pour alcooliques. Les informations transmises par le PMSI-SSR incluent les caractéristiques socio-démographiques du patient, son type de soin et sa morbidité,

les données de transfert dans l'établissement de soins, les actes médicaux réalisés, la dépendance et les temps d'intervention.

A l'instar du PMSI-MCO, l'objectif du PMSI-SSR est de réduire les inégalités budgétaires en analysant l'activité médicale des établissements de soins. Les groupes de classification construits comprennent les périodes hebdomadaires de séjours et non pas tout le séjour hospitalier contrairement au PMSI-MCO. Ces groupes constituent une part du financement des établissements de soins. Les données de la semaine du séjour SSR sont sauvegardées dans des Résumés Hebdomadaires Standardisés (RHS) puis affectées à une Catégorie Majeure Clinique (CMC). Les CMC sont ensuite anonymisés et transformés en Résumé Hebdomadaires Anonymisés (RHA). Finalement, chaque RHA est affecté à un Groupe Homogène de Journées (GHJ). Les soignants qui procèdent au recueil des informations sont : les médecins (diagnostics et actes médicaux), les infirmières et aides-soignantes (scores de dépendance) et les kinésithérapeutes (temps de rééducation).

1.5.2.3 Hospitalisation à domicile (PMSI-HAD)

Le PMSI-HAD représente une structure de soins alternative à l'hospitalisation conventionnelle [Zol18]. Elle assure au domicile du patient des soins médicaux et paramédicaux importants, pour une période limitée mais renouvelable en fonction de l'évolution de l'état de santé du patient. Elle est considérée comme un mini réseau au sein d'un très grand réseau. Ce mini réseau est en coordination avec d'autres mini réseaux [Cha08]. Le PMSI-HAD a pour vocation de réduire les durées d'hospitalisation, le taux d'occupation des lits et les coûts de prise en charge des patients.

1.5.2.4 Psychiatrie (PMSI-PSY)

PMSI-PSY permet de décrire l'activité réalisée au bénéfice des malades dans le service de psychiatrie en hospitalisation complète, partielle ou ambulatoire [FAU21]. Les données collectées sont des informations liées au patient et des informations sur la consommation de ressources. Comme pour le PMSI-SSR, une unité d'hospitalisation est définie comme une semaine d'hospitalisation par patient. Chaque unité est ensuite classée dans des groupes homogènes de journées (GHJ). Les activités externes sont également sauvegardées dans des relevés à l'acte.

Afin d'inciter les hôpitaux français publics et privés à évaluer et à analyser leurs activités, le PMSI est rendu obligatoire depuis 1991 [FAU21]. Il a été généralisé en 1997 ce qui a permis de construire une base de données médicalisée décrivant l'activité hospitalière. Ceci a conduit à un nouveau système de gratification des hôpitaux basé sur leur activité selon des groupes de séjours doublement homogène

présentant une similitude économique et portant sur la même logique de prise en charge médicale. Ce nouveau système est nommé la tarification à l'activité (T2A) et est révélé en 2005 [Mor09]. La T2A est le mode de financement unique des établissements de santé, publics et privés. Le prix de chaque activité hospitalière est fixé chaque année par le ministre chargé de la santé en se basant sur les classes des Groupe Homogènes de Malades [san17b]. La T2A associe le paiement à l'activité réalisée qui est décrite à travers des groupes homogènes de malades (GHM) plutôt que selon les disciplines de services hospitaliers (ou spécialités). Les prix des GHM sont définis à l'avance (paiements prospectifs) et peuvent être fixés au niveau national, comme c'est le cas en France, ou au niveau local comme le cas en région des Hauts de France [RO09]. Le PMSI repose sur un certain mode de fonctionnement qui sera détaillé dans la section suivante.

1.5.3 Fonctionnement du PMSI

L'apparition de la loi n^o 91-748 du 31 juillet 1991 portant sur la réforme hospitalière incite et force les établissements de soins à transmettre les informations sur leurs moyens de fonctionnement et leurs activités [Lég21]. L'implémentation du PMSI oblige les hôpitaux à procéder à l'analyse de leurs activités médicales. Dans le cadre de nos travaux, nous nous concentrons sur les séjours hospitaliers effectués en médecine, chirurgie ou obstétrique (PMSI-MCO). Ce champ du PMSI est l'un des champs principaux car il accueille le plus grand nombre de patients. Les profils des patients sont variés en termes d'informations démographiques, médicales et administratives. Dans ce qui suit, nous allons décrire le fonctionnement du PMSI-MCO depuis la collecte d'informations jusqu'au calcul du budget médical.

Le PMSI produit une base de données médico-économique permanente normalisée à l'échelle nationale. Dans le cas d'un séjour hospitalier en MCO, des Résumés de Sortie Standardisé (RSS) sont établis pour chaque patient admis en MCO. Le RSS est produit par le Département d'Information Médicale (DIM). Ce RSS est composé d'un ou plusieurs Résumé d'Unité Médicale (RUM). Le nombre de RUM dans les RSS correspond au nombre d'unités médicales dans lesquelles le patient est admis. A la fin du séjour et après l'édition des RSS, un classement est réalisé en se basant sur les Catégories Majeurs de Diagnostics (CMD). Le CMD affecte chaque RSS à un Groupe Homogène de Malades (GHM). Les GHM représentent des profils patients construits sur la base des données médico-économiques des hospitalisations en secteur de soins de courte durée (médecine, chirurgie, obstétrique). Ces données rassemblent les informations administratives, les diagnostics et les actes médicaux [Elg05]. Dans chaque GHM, les caractéristiques médicales et la durée de séjour hospitalier sont homogènes. Cette durée de séjour est à son tour fortement corrélée au coût du séjour [Gra14]. Les données médico-économiques contenus dans les RSS sont transmises à

l'Agence Régionale de l'Hospitalisation (ARH). Elles sont ensuite anonymisées afin d'assurer la confidentialité et le secret médical liés aux données médicales [ATI03]. Elles sont alors transformés en Résumé de Sortie Anonymisé (RSA) et transmis à l'Agence Technique de l'Information Hospitalière (ATIH). D'autre part, afin de relier les séjours d'un même patient sans dévoiler son identité, un numéro d'anonymisation unique patient est créé et une procédure de chaînage des résumés de séjour a été mise en œuvre [ATI11]. Le principe du chaînage anonyme consiste en la création d'un numéro anonyme commun à toutes les hospitalisations d'un même patient, quel que soit le lieu de prise en charge hospitalière : public ou privé, MCO, soins de suite ou de réadaptation (SSR) ou psychiatrie. Au final, pour procéder à la tarification, chaque Groupe Homogène de Malades (GHM) est affecté à un Groupe Homogène de Séjours (GHS) et de ce fait, le tarif attribué à chaque séjour se base ainsi sur les GHM qui lui même se base sur les CMD. La figure 1.3 montre la procédure de collecte et de transformation des données issues du PMSI mais aussi le fonctionnement décrit dans cette section.

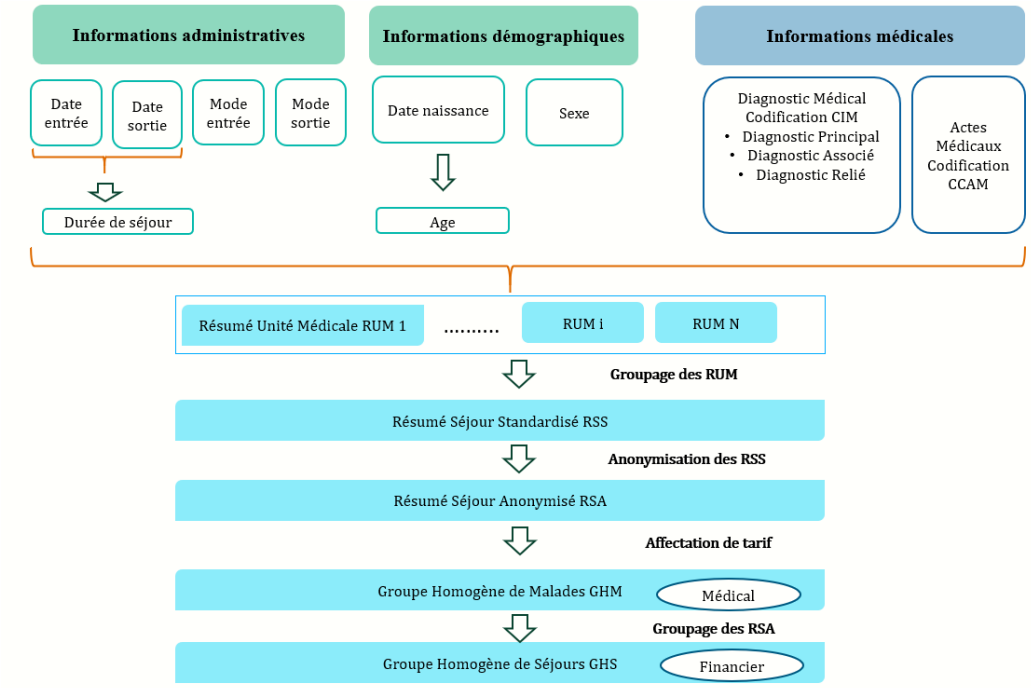


Fig. 1.3: Données issues du PMSI et son fonctionnement.

Les données collectées contiennent les informations démographiques du patient telles que sa date de naissance qui sera remplacée par son âge, son sexe, son code postal ou lieu de résidence. De plus, elles contiennent les informations administratives comme par exemple l'identifiant de l'établissement de soins (Finess), l'unité médicale d'admission du patient, la date et le mode d'entrée dans l'unité médicale ainsi que la date et le mode de sortie de l'unité médicale. Les informations médicales du patient enregistrées incluent le diagnostic médical et ses complications. Ils sont codés à l'aide de la Classification Internationale des Maladies (CIM). De plus, l'ensemble

des actes médicaux réalisés qui sont au moins en rapport avec le diagnostic médical sont enregistrés. La Classification Commune des Actes Médicaux (CCAM) est utilisée pour le codage de ces actes.

La Classification Internationale de Maladie en sa dixième version (CIM-10) a été crée par l'Organisation Mondiales de la Santé (OMS) en 1983 [Org21]. Elle définit un ensemble très large de maladies, des symptômes, des contextes sociaux ou encore des causes externes de maladies ou de traumatismes [OOr21]. Des milliers de codes sont générés et utilisés dans les recherches cliniques, l'allocation des ressources, la prise en charge et le suivi des patients. La CIM-10 est la référence pour les professionnels de santé, les chercheurs, les informaticiens pour représenter les différents diagnostics des patients. La CIM-10 permet aux établissements de soins de collecter des données standardisées et donc comparables entre les pays ou entre les régions.

La Classification Commune des Actes Médicaux (CCAM) fait référence à une liste ordonnée de libellés indiquant des actes techniques médicaux. Un acte médical correspond à « tout acte dont la réalisation par des moyens verbaux, écrits, physiques ou instrumentaux, est effectué par un membre d'une profession médicale, dans le cadre de son exercice et les limites de sa compétence » [san17a]. Sa structure est composée de 4 niveaux : les chapitres, sous-chapitres, paragraphes et sous-paragraphes. Chaque acte est constitué de 7 caractères alphanumériques couvrant des actes globaux et des gestes techniques complémentaires [Pap06]. Un groupement des actes médicaux peut être fait en utilisant la terminologie de l'acte afin de réduire le nombre de l'ensemble des actes médicaux existant et les utiliser à des fins de recherches académiques.

Les données présentes dans les SIH sont d'une grande richesse. Elles sont exploitées pour améliorer les systèmes de santé et la qualité des soins proposée aux patients. La Durée De Séjour hospitalier (DDS) est un indicateur clef d'évaluation de base des performances des systèmes de santé. Nous allons dans ce qui suit introduire la DDS et montré sa relation avec le PMSI.

1.6 Durée De Séjour hospitalier : indicateur clef dans les systèmes de santé

Face à l'augmentation du nombre des patients dans les hôpitaux, l'apparition de nouvelles maladies, la hausse des coûts des soins médicaux et le vieillissement de la population, les systèmes de santé sont confrontés à de nombreuses difficultés dont la surcharge du travail des professionnel de santé et l'insuffisance des ressources

matérielles. Le principal objectif d'un établissement de soins est d'améliorer la santé des individus et leur offrir une excellente qualité de soins. Le système de santé étant un système complexe, est souvent considéré comme mal structuré, mal dirigé, inefficacement organisé et insuffisamment financé [BBB20]. Il est donc nécessaire d'évaluer ses performances et d'en déduire les faiblesses afin de les corriger.

L'évaluation de l'efficacité et la qualité des soins s'impose dans les différentes institutions de santé. Il existe essentiellement trois indicateurs d'efficacité d'un système de santé : le taux de mortalité, le nombre de réadmissions et la durée de séjour hospitalier (DDS) [Bot+13]. Dans une autre étude menée par [Lin+18], les corrélations entre ces mesures a été démontrée afin de proposer une nouvelle méthode d'évaluation.

Dans nos travaux de recherche nous nous sommes intéressés à la Durée De Séjour hospitalier (DDS). La prédiction de la DDS d'un patient au moment de son admission et durant son séjour est un indicateur d'évaluation de base des services de santé [Mar+01]. Cette prédiction a fait l'objet de plusieurs études durant ces dernières années et contribue d'une manière sensible à l'optimisation des ressources des hôpitaux et du fonctionnement de leurs services tout en assurant leur qualité.

Les établissements de soins, les chercheurs universitaires et les entreprises industrielles cherchent à rassembler leurs efforts et à collaborer pour améliorer les performances des services des hôpitaux. La DDS permet de mieux comprendre le flux des patients et de suivre leur trajectoire. Ceci sert à l'évaluation des fonctions opérationnelles et cliniques des services hospitaliers [ABM17]. La prédiction de la DDS contribue à la planification et à l'organisation des activités de soins, ainsi qu'au management des lits réduisant leur occupation inutile. En effet, une DDS supérieure à la normale peut entraîner des pertes financières pour l'établissement de soins ainsi un allongement des délais d'attente des patients qui n'ont pas pu être hospitalisés par manque de places. Par ailleurs, l'étude de la DDS et sa prédiction produisent de meilleures conditions de travail des professionnels de santé [RIG09]. Une meilleure estimation de DDS facilite la planification de l'utilisation des salles, des lits et des opérations chirurgicales non urgentes en fonction de la disponibilité des lits [Gen+17]. Dans le cas de la planification des opérations chirurgicales, les journées d'hospitalisation inutiles sont considérablement réduites. La prédiction de la DDS permet également l'identification de la gravité des maladies [LT09]. Dans le cadre de l'utilisation des données issues du PMSI, la figure 1.4 illustre le fonctionnement d'un séjour en se basant sur le PMSI et la définition de la DDS pour ce séjour.

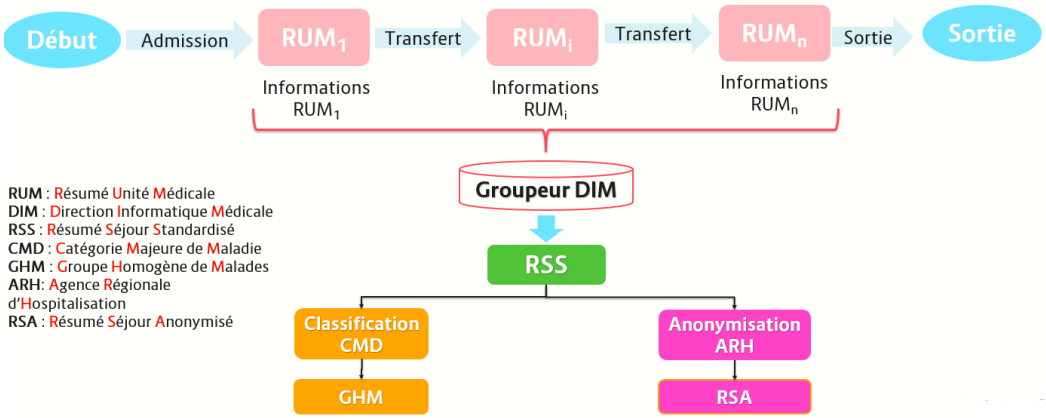


Fig. 1.4: Évaluation des systèmes de santé : DDS et PMSI.

La durée de séjour hospitalier est donc une composante essentielle dans l'évaluation des performances des systèmes de santé. Sa prédiction au moment de l'admission du patient et tout au long de son séjour hospitalier apporte une meilleure planification des activités de soins, une rapidité et une meilleure qualité des soins proposés. Cela incite beaucoup de chercheurs à la prendre en considération et la définir comme un élément primordial dans le contrôle des activités hospitalières. Plusieurs techniques issues de l'Intelligence Artificielle (IA) à savoir l'apprentissage automatique sont apparues pour la prédiction de la DDS et sont présentés dans la section suivante. Elles sont présentées en détails dans le chapitre 2.

1.7 Apprentissage automatique au service de la santé

Ces dernières années, l'Intelligence Artificielle (IA) est considérée comme l'une des innovations majeures. Ceci lui a permis d'occuper une place très importante dans plusieurs secteurs. Elle a été rapidement intégrée dans différents secteurs comme la finance, l'énergie ou la santé. Grâce aux techniques de l'Intelligence Artificielle, les machines sont maintenant en mesure d'effectuer des tâches humaines complexes telles que la reconnaissance automatique d'objets, le traitement automatique du langage naturel ou la prise de décision.

L'IA englobe plusieurs techniques comme par exemple l'apprentissage automatique, la vision par ordinateur, le raisonnement, la représentation des connaissances et la fouille de données. Ces techniques font partie des techniques les plus utilisées de nos jours dans les différents domaines de recherche. Les applications de l'IA s'étendent à des domaines que l'on pensait auparavant réservés aux experts humains. En effet, dans le domaine médical, et grâce aux progrès récents en matière d'acquisition de

données numérisées, d'infrastructure informatique et d'amélioration de la puissance et de la capacité de stockage des ordinateurs, le domaine médical est identifié comme l'un des domaines les plus promoteurs de l'IA. L'apprentissage automatique ou le Machine Learning (ML) en anglais, est une technique de l'IA largement employée dans les recherches cliniques. Elle est apparue dans les années 1950 avec Alan Turing quand il a écrit un article sur « Computing machinery and intelligence » [Tur50], dans lequel il a expliqué que pour démontrer l'intelligence d'une machine, elle doit être capable d'exécuter des tâches humaines de telle sorte que personne ne peut la différencier de celle d'un être humain.La figure 1.5 illustre les principales techniques de l'Intelligence Artificielle et leurs applications.

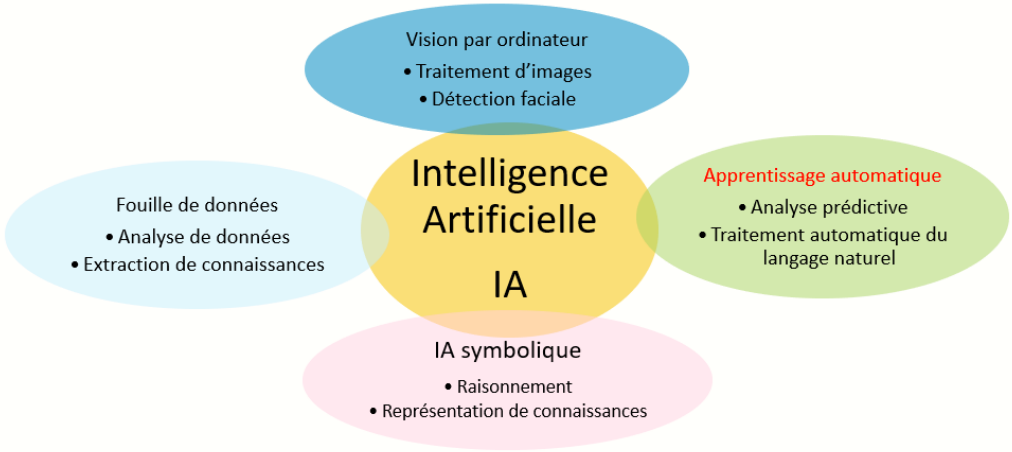


Fig. 1.5: Techniques de l'Intelligence Artificielle et leurs applications [YBK18 ; DK19]

L'apprentissage automatique consiste à doter les machines de capacités d'analyse, d'apprentissage et de généralisation à partir des données. L'objectif est de résoudre des problèmes pour lesquels il aurait été difficile de trouver une solution avec des approches informatiques traditionnelles. Il existe quatre types d'apprentissage automatique : l'apprentissage supervisé, l'apprentissage non-supervisé, l'apprentissage semi-supervisé et l'apprentissage par renforcement.

En imagerie médicale, la résolution de nombreux problèmes repose sur l'apprentissage automatique supervisé afin de diagnostiquer les maladies [YBK18]. Le diagnostic de la tuberculose pulmonaire et des maladies pulmonaires courantes par radiographie thoracique [RR17] et l'identification de masses mammaires à l'aide de mammographies [Are+], ont atteint une précision de niveau expert à l'aide des méthodes d'apprentissage automatique, de vision par ordinateur et de fouille de données. De plus, l'apprentissage automatique basé sur l'image permet l'identification de la cause de la maladie et le suivi de la trajectoire du patient au cours d'une maladie [Jia+17]. Une des applications de l'apprentissage automatique est le traitement automatique du langage naturel ou Natural Langage processing (NLP)

en anglais. Dans le domaine médical, le NLP est appliqué à la compréhension et la classification de la documentation clinique, l'analyse sémantique et syntaxique des notes textuelles sur des patients ainsi que des rapports médicaux, la transcription des interactions avec des patients et l'utilisation d'une IA conversationnelle [DK19]. Les techniques d'IA sont aussi mises en œuvre pour des fins d'organisation des établissements de soins et d'amélioration de la qualité des soins [KLH20]. Dans la logistique hospitalière, l'apprentissage automatique est utilisé afin de prédire les achats des produits en intégrant le délai de livraison actuel du fournisseur et la quantité économique d'approvisionnement pour chaque produit [JDJ19]. La prédiction des résultats cliniques comme les taux de mortalité, les réadmissions et la durée de séjour hospitalière est très prometteuse dans les travaux précédents en utilisant l'apprentissage automatique [Cai+16]. L'apprentissage automatique est également utilisé dans le suivi des parcours des patients et leur profilage pour évaluer l'efficacité des services des établissements de santé [Naj17].

Le plus grand défi de l'IA dans le domaine de la santé est de garantir l'adoption de ces techniques en pratique dans la vie quotidienne des établissements de soins. Une prédiction en temps réel, le respect de la confidentialité des données et la détection et l'analyse des profils rares des patients sont de vrais challenges à promouvoir. L'expertise médicale est cruciale dans ces cas pour atteindre les meilleures performances des systèmes médicaux intelligents se basant sur l'IA.

1.8 Conclusion

Dans ce chapitre, nous avons présenté les systèmes d'informations hospitaliers et leur description comme un moyen de sauvegarde des données médicales. Nous avons également présenté les source des données médicales. Les principales propriétés des données médicales dont la complexité, l'hétérogénéité, la confidentialité et l'aspect incrémentale sont également explorés. Le sous-système de production responsable sur la description médico-économique des activités des établissements de soins est présenté. Ce système est le Programme de Médicalisation des Systèmes d'Informations (PMSI). Les objectifs du PMSI sont mis en évidence. En plus, sa mise en place, son contexte économique et son mode de fonctionnement sont expliqués. Les méthodes d'apprentissage automatique issues du domaine de l'Intelligence artificielle sont principalement explorées. La prédiction de la durée de séjour hospitalier étant l'objectif de cette thèse, la durée de séjour est donc présentée comme un indicateur d'évaluation de base des qualité des soins.

Nous allons présenter, dans le chapitre suivant, un état de l'art sur les modèles de prédiction des durées de séjours hospitaliers, les facteurs qui impactent cette durée de séjour et les techniques de sa prédiction.

2

Modélisation et prédiction des durées de séjours hospitaliers : Etat de l'art

> *Have no fear of perfection, you will never reach it.*
>
> — **Marie Sklodowska-Curie**
> Prix Nobel de physique et de chimie.

2.1 Introduction

La prédiction des durées de séjours hospitaliers (DDS) a connu un intérêt croissant ces dernières années. Cette prédiction contribue à l'amélioration de l'organisation des hôpitaux, à l'optimisation des ressources et à la gestion des flux des patients. La construction d'un modèle de prédiction de DDS repose sur plusieurs éléments et doit intégrer le contexte complexe de l'environnement hospitalier. Plusieurs études ont été menées pour définir quels sont les éléments qui interviennent dans la DDS dans un environnement hospitalier et les facteurs qui l'impactent.

L'objectif de ce chapitre est double. D'une part, nous présentons l'état de l'art se rapportant aux divers modèles de Durée De Séjour hospitalier en dégageant ses caractéristiques. Cette analyse permet de déduire une modélisation de la DDS. Celle-ci sera utilisée dans la deuxième partie et dans laquelle nous exposons les différents algorithmes qui concerne la prédiction des DDS. Ensuite, nous nous focalisons sur les algorithmes basés sur l'apprentissage automatique. Cette étude sert à bien définir la problématique et à identifier la démarche méthodologique à entreprendre dans notre étude. De plus, cette étude met en évidence les forces et faiblesses des différentes méthodes utilisées dans la prédiction des DDS. Au final, les techniques d'apprentissage automatique employés dans nos travaux sont présentés afin d'introduire notre démarche méthodologique.

2.2 La Durée De Séjour Hospitalier (DDS)

Les établissements de soins cherchent sans cesse à optimiser le fonctionnement de leurs services tout en assurant la qualité des soins. La planification des activités et la gestion des ressources ont un impact important sur cet objectif. L'estimation de la durée de séjour d'un patient au moment de son admission et durant son séjour hospitalier est un indicateur d'évaluation de base des services de soins. La prédiction des Durées De Séjour hospitalier (DDS) contribue à l'amélioration de la qualité des soins ainsi qu'à renforcer l'efficacité de la charge de travail opérationnelle. Cela permet une planification précise des admissions et des réadmissions, une optimisation des coûts et une réduction du nombre de lits mal occupés. Afin de planifier les activités de soins de manière pertinente, des données de santé souvent sous format électronique sont utilisées pour l'analyse et la prédiction de la DDS. Plusieurs facteurs ont une incidence significative sur la DDS [CP14].

Dans la suite, nous commençons par définir la notion de DDS. Ensuite, une analyse permettra d'en déduire ses caractéristiques ce qui nous conduira à la définition d'un modèle générique de DDS.

2.2.1 Définition

L'objectif de chaque établissement de soins est de répondre à l'attente des patients en leur fournissant des services de soins efficients. Les hôpitaux étudient la DDS des patients pour évaluer les performances des systèmes de santé. La définition de la DDS varie selon l'objectif de chaque étude qui s'y intéresse. D'une manière générale, la DDS est définie comme l'intervalle de temps entre l'admission du patient et sa sortie de l'établissement de soins [Kho+16]. Elle représente le nombre de jours que le patient a passé pendant la même hospitalisation. L'unité de mesure de la DDS est la journée. Cette définition peut changer constamment dans d'autres contextes. Dans les services d'urgence et ambulatoire, l'admission et la sortie du patient sont réalisées dans la même journée. De ce fait, la DDS est égale à 0 jour. La valeur de la DDS est alors calculée en nombre d'heures et peut s'étaler sur 24 heures au maximum [Wre+05]. Plusieurs travaux ont étudié les flux des patients en service d'urgence en se basant sur l'estimation du nombre d'heures du séjour du patient dans ce service [Ben+19]. Nous distinguons donc deux définitions majeures de la DDS : la DDS dans des unités médicales dites « programmées » calculée en nombre de jours passés dans ces unités et la DDS dans des services dits « non programmés » calculée en nombre d'heures. Dans ces deux cas, la DDS est quantifiée par une valeur numérique discrète.

La DDS peut également être représentée par des catégories. La représentation de la DDS en catégories varie selon l'objectif visé. Dans l'étude menée par [Chu+16], deux classes de DDS sont définies : une DDS courte et une DDS longue. Les auteurs se sont intéressés à prédire les catégories de la DDS afin d'organiser le service de chirurgie de l'hôpital. La discrétisation de la DDS ou sa définition en catégories peut se baser sur la valeur de sa moyenne ou la valeur de sa médiane. Les catégories sont spécifiées selon le domaine d'application. Nous pouvons alors avoir 3 catégories de DDS : une DDS courte, une DDS moyenne et une DDS longue. Dans [Jon13], les auteurs se sont focalisés sur l'analyse de la moyenne de la DDS comme indicateur d'efficacité de la qualité des soins. Il est intéressant aussi d'étudier la moyenne de la DDS à des fins d'organisation et de planification. Elle est généralement mesurée en divisant le nombre total de jours passés par tous les patients hospitalisés au cours d'une année par le nombre d'admissions ou de sorties [OCD17]. A partir de cette définition, [OEC20] détermine la moyenne de DDS par groupe de diagnostic médical ou par motif d'hospitalisation. Cette moyenne est déduite également par groupe d'unités médicales.

Ces différentes définitions révèlent que la prédiction de la DDS nécessite une modélisation qui se rapproche le plus de la réalité et qui répond à un objectif bien précis soit la gestion des services d'urgences ou ambulatoire, l'étude de la moyenne de la DDS ou l'estimation du nombre de jours passés dans une unité médicale. Cependant, il est important de modéliser la DDS et identifier les facteurs qui l'impactent dans un environnement hospitalier. Dans ce qui suit, nous présentons une étude approfondie sur les facteurs impactant la DDS dans différentes unités médicales.

2.2.2 Facteurs impactants la Durée De Séjour hospitalier

La connaissance des facteurs qui déterminent une DDS dans un environnement hospitalier est l'objectif de nombreuses enquêtes [Row+07 ; Laf+15 ; Mah+18 ; Kho+16]. Afin de construire un modèle de prédiction de DDS, il est indispensable d'identifier tous les éléments qui le constituent. La modélisation des DDS en milieu hospitalier prend différentes formes que nous exposons dans ce qui suit.

La durée de séjour à l'hôpital dépend de plusieurs facteurs hétérogènes relatifs aux contextes cliniques et sociaux du patient et à l'organisation de l'établissement de soins. Dans [RIG09] , il a été montré que la prédiction de la durée de séjour dépend fortement du type de l'unité médicale dans laquelle le patient est admis. Les modèles de DDS dans un service d'urgence ou un service ambulatoire diffèrent de ceux d'autres unités médicales comme l'unité de cardiologie. De plus, dans [PK14], le type du service de l'hôpital est considéré comme un paramètre important pour la prédiction de la DDS. D'autres recherches antérieures ont tenté de grouper les patients en fonction de leur état de santé, en supposant que chaque maladie est associée à une DDS recommandée [She+95]. Dans ce cas, la gravité de la maladie impacte la DDS. Par conséquent, la durée de séjour est définie différemment d'un service hospitalier à un autre.

Nous avons étudié les différents modèles de DDS apparus dans la littérature selon un type de service bien défini. Nous avons constaté que les auteurs se focalisent souvent sur les services hospitaliers suivants : le service de cardiologie, le service de soins intensifs et le service de chirurgie générale [Mek+19]. Ces services requièrent plus de ressources que d'autres services et leur DDS dépend, non seulement des facteurs médicaux, mais aussi des facteurs économiques. Nous allons dans ce qui suit exposer les facteurs qui impactent une DDS dans ces différents service de soins.

2.2.2.1 Durée De séjour dans un service de cardiologie

Plusieurs travaux se sont intéressés à la prédiction des DDS dans les services de cardiologie. Dans l'étude menée par [Laf+15], 8 variables ont été sélectionnées parmi 36 pour concevoir le modèle de prédiction des DDS. D'une part des variables numériques comme l'âge du patient, le niveau d'émission de l'O2, le taux de sérum créatinine et la mesure de l'analyse du gaz dans le sang et l'hématocrite sont utilisées. D'autre part, des variables binaires sont employées telles que le sexe du patient, l'usage d'une pompe de ballon intra-aortique et l'emploi d'agents isotropes cardiaques positifs. Dans l'étude menée dans [Hac+13], les auteurs utilisent également des facteurs concernant l'état de santé du patient tels que la consommation de médicaments anticoagulants et de nitrate, la mesure de la pression artérielle diastolique et du cholestérol, la présence de douleur thoracique, la valeur de la fraction d'éjection, la densité de lipoprotéine, le taux d'hémoglobine et l'existence d'autres comorbidités. Des facteurs liés au comportement à risque comme la consommation du tabac sont également intégrés. Des données démographiques sont par ailleurs enrichies, comme l'âge du patient, son sexe et son état civil. Dans [Tsa+16], le modèle de prédiction de DDS au moment de l'admission du patient se base sur l'adresse du patient, son diagnostic, le type de l'intervention chirurgicale, la comorbidité et le mode de remboursement. C'est donc un modèle qui inclue des données médicales, administratives et économiques. Dans [Alm+16], les auteurs ont analysé les éléments qui influencent une longue durée de séjour dans le service de soins intensifs cardiaque. Ils ont recensé les variables démographiques (âge et sexe du patient), l'historique médical du patient, le nombre de complications, ou bien encore, la valeur de la fraction d'éjection ventriculaire gauche. Ces éléments sont tous disponibles au moment de l'admission du patient. La prédiction de la DDS dans ce cas se fait au moment de l'admission du patient.

2.2.2.2 Durée de séjour dans un service de soins intensifs

Plusieurs travaux ont tenté d'analyser et d'extraire les facteurs qui influencent la DDS dans les services de soins intensifs (SSI). L'objectif de l'analyse est de définir, d'un point de vue théorique, les variables les plus utilisées dans la prédiction des DDS. Dans [Gen+17], la base de données « Medical Information Mart for Intensive Care » (MIMIC) a été utilisée pour prédire les longues DDS. La DDS y est définie comme une variable binaire : longue ou courte. Le but visé est d'anticiper l'allocation des ressources dans les SSI. Pour cela, les variables suivantes sont collectées : les procédures d'admission, de transfert et de sortie du patient, les prescriptions médicales, les résultats de la prise de sang, le diagnostic médical en complément des facteurs démographiques. Ces variables sont par la suite utilisées dans le modèle de prédiction. Les données concernant 311 patients ont été utilisées dans [Mah+18]

pour identifier les facteurs qui impactent la DDS dans les SSI suite à une chirurgie cardiaque. Cette étude distingue les facteurs démographiques, le patient est-il fumeur ou non et l'historique médical du patient (maladie cardiaque, rénale et pulmonaire). De plus, les mesures du taux de créatinine, du rythme cardiaque et de la fraction d'éjection sont utilisés. Dans [EAH20], les auteurs ont montré que les informations suivantes : l'âge, le sexe du patient, la cause de l'admission et celle de l'intubation, l'historique médical et le rythme cardiaque sont utilisées afin d'examiner la DDS dans les SSI.

Nous avons relevé que le modèle de DDS dans un SSI est complexe et nécessite une bonne connaissance du domaine de la santé. Les experts médicaux doivent être impliqués afin d'accompagner les systèmes d'aide à la décision dans le processus de prédiction mais aussi de guider la recherche [Mek+20b]. L'expertise médicale est aussi déterminante dans la phase de validation des résultats. En effet, les SSI accueillent des patients dans un état de santé préoccupant, il existe de fortes possibilités d'avoir des complications ce qui va augmenter la DDS. Il est donc important de prédire la DDS dans ce type de service.

2.2.2.3 Durée de séjour dans un service de chirurgie

Différents travaux ont étudié la DDS dans les services de chirurgie. En particulier, ils ont mis en évidence deux phases : la phase pré-opératoire et la phase post-opératoire. Dans [Chu+16], les auteurs ont montré que dans un même service de chirurgie, les facteurs qui peuvent impacter la DDS sont différents pour une opération urgente et pour une opération non urgente. Ils ont également montré que ces facteurs incluent les informations démographiques du patient (âge, sexe), son historique médical, les mesures des signes vitaux (le rythme cardiaque, la respiration, la tension artérielle et la température), les résultats de l'analyse sanguine et les rapports médicaux. Dans [Kho+16], les facteurs qui influencent le plus la DDS concernent la situation familiale du patient, ses conditions de sortie du service hospitalier, le type du traitement et le mode de paiement. Cela montre qu'en dehors des données médicales, les données économiques et administratives sont importantes dans la prédiction de la DDS. L'organisation des services hospitaliers comme le service de chirurgie est aussi fortement liée à la DDS de ses patients. Les recherches dans [AK16] ont prouvé que certaines variables qui influencent la DDS sont liées à l'unité médicale. Ces variables comportent le type et le nombre d'opérations que le patient a subi, le temps entre l'ordre de sortie et la sortie effective du patient, les informations de transfert entre services, le nombre moyen de visites par jour, le nombre de consultations médicales, les hospitalisations précédentes du patient et le nombre d'examens réalisés par le service de chirurgie générale.

Dans ce dernier type de service médical étudié, les facteurs qui représentent les variables caractérisant la DDS que nous avons révélé sont liés à l'unité médicale de l'admission du patient. Dans la figure 2.1 , nous avons résumé les facteurs communs au services cités précédemment [Mek+19 ; AK16 ; Kho+16 ; Chu+16 ; Mah+18].

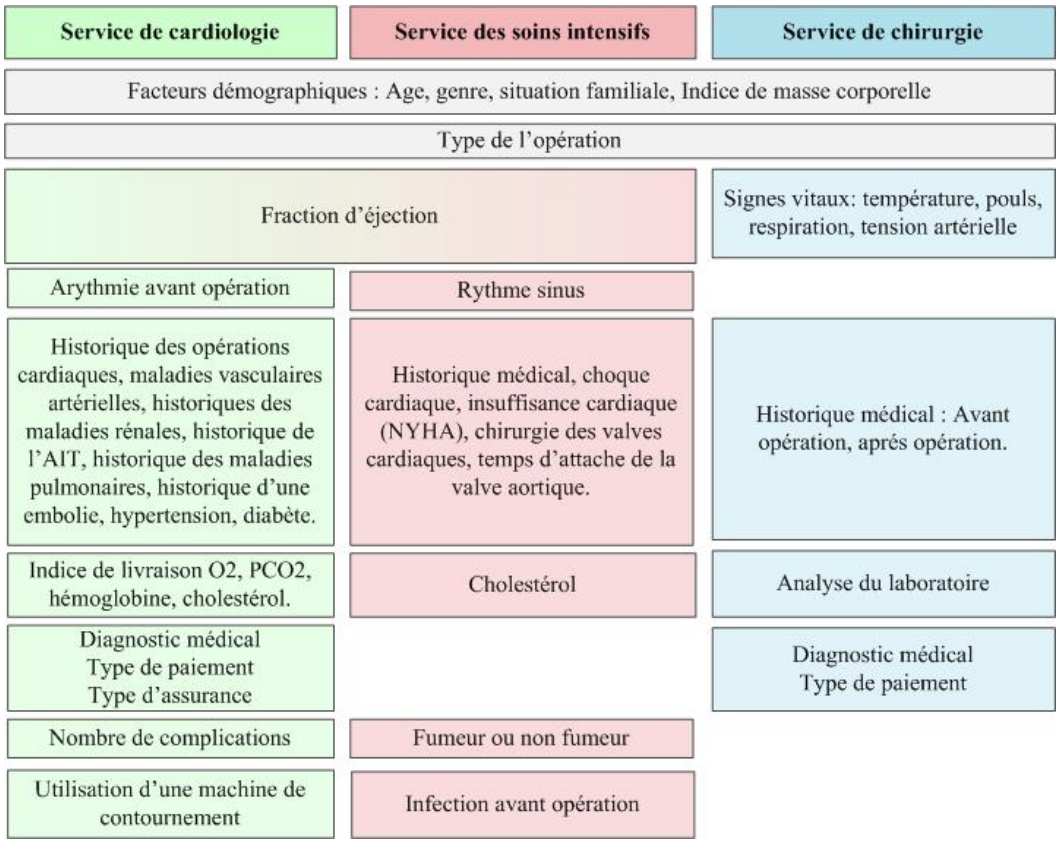


Fig. 2.1: Facteurs impactant la DDS dans trois unités médicales majeures [EAH20 ; Alm+16 ; AK16 ; Kho+16 ; Chu+16 ; Mah+18].

Toutes ces recherches [EAH20 ; Alm+16 ; AK16 ; Kho+16 ; Chu+16 ; Mah+18] ont révélé l'importance d'étudier les facteurs qui impactent une Durée de Séjour à l'hôpital. Différents facteurs sont identifiés et utilisés dans les modèles de prédiction. Quoique que ces modèles caractérisent la DDS dans les services les plus coûteux à l'hôpital, la limite de ces modèles réside dans le fait qu'ils sont restreints à une seule unité médicale. Il est plus intéressant de concevoir un modèle qui englobe différents facteurs liés à plusieurs unités et de les utiliser par la suite dans l'étape de prédiction de la DDS. Ce modèle caractérise la DDS dans différents services hospitaliers et s'adapte à l'environnement dynamique des établissements de soins. Cela permet une prédiction de DDS dans plusieurs types d'unités médicales. Dans la section suivante, nous présentons un modèle générique de DDS. Une de nos contributions est d'étendre l'étude de la DDS et de sa prédiction à plusieurs unités médicales.

2.2.3 Modèle générique de Durée de Séjour hospitalier

Nous définissons un séjour hospitalier est défini comme l'intervalle de temps entre l'admission du patient et sa sortie. Un séjour hospitalier est le passage d'un patient d'une unité médicale à une autre. Le patient peut être admis dans une seule unité médicale, dans ce cas, le séjour est dit séjour mono-UM. Dans le cas où le patient est transféré d'une unité médicale vers une autre, le séjour est appelé séjour multi-UM. Les informations concernant le séjour hospitalier s'alimentent d'une part, d'un passage d'une unité vers l'autre, et, d'autre part, au sein d'une même unité médicale. La figure 2.2 illustre un schéma représentant un séjour hospitalier. Par la suite, nous proposons un modèle générique de DDS dans le cas d'un séjour multi-UM en analysant les travaux antérieurs et les besoins quotidiens des hôpitaux.

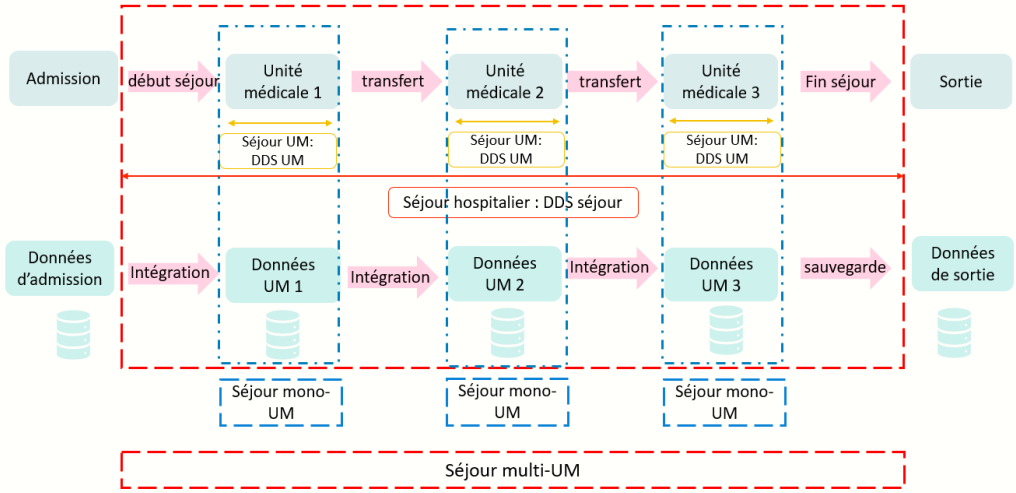


Fig. 2.2: Schéma d'un séjour hospitalier.

Dans la section 2.2.2, nous avons vu que les travaux antérieurs présentent la DDS comme une variable complexe et qu'elle dépend de plusieurs facteurs. En particulier, la DDS est très liée au type du service hospitalier dans lequel le patient est admis. La DDS dans les services d'urgence et ambulatoire diffère de la DDS dans d'autres services appelés services programmées. Nous avons également souligné que la majorité des études concernent la prédiction des DDS dans les unités de cardiologie, de soins intensifs et de chirurgie. Sur la base des travaux précédents, nous distinguons 4 classes d'informations qui concernent la DDS. Ces classes représentent l'ensemble des facteurs qui l'impactant.

- Les informations médicales : ces informations incluent le diagnostic principal du patient ou le motif d'hospitalisation, la comorbidité ou les complications médicales, son historique médical et les résultats des examens biologiques.

- Les informations démographique du patient constituent l'âge du patient, son sexe, son adresse du domicile et sa situation maritale.
- Les informations administratives englobent le type de l'organisation de l'hôpital, conditions d'admission et de sortie du patient, type de l'unité médicale qui l'accueille et les procédures de transferts entre unités médicales.
- Les informations économiques comportent le type de paiement, les procédures de remboursement et le type d'assurance du patient.

Un élément commun aux recherches cités dans l'état de l'art précédent est qu'elles représentent des modèles spécifiques à une étude et se basent le plus souvent sur des données issues d'une seule institution. Par conséquent, le niveau de service est également lié au système médical du pays.

Concernant les facteurs ayant un impact sur la DDS hospitalier et sa prédiction, nous avons remarqué que ces facteurs peuvent être déterminés soit par la revue de la littérature, soit par les experts du domaine médical. En effet, quoique toutes ces recherches, ont tenté d'étudier ces facteurs dans de multiple services, l'implication des experts dans le domaine médical est aussi fondamentale et, par conséquent, 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 l'objectif de l'étude. En plus, ces facteurs peuvent être déterminés selon le besoin de l'hôpital, sa région ou localisation et par d'autres facteurs occasionnels comme la période d'admission, la charge du travail et les ressources disponibles.

Nous proposons un modèle générique qui représente une Durée de Séjour hospitalier dans différentes unités de soins médicales programmés. Nous avons exclu les unités médicales d'urgence et ambulatoire car nous supposons que la DDS représente le nombre de jours que le patient a passé dans une unité médicale. De plus, pour des fins d'organisation des services hospitaliers, nous proposons de prendre en compte les informations liées au contexte médico-clinique de l'établissement de soins telles que la capacité d'accueil de l'unité médicale et la charge du travail des professionnels de santé. Nous avons également pris en considération des variables concernant la période d'admission du patient. Cela se justifie par le fait que les admissions et les sorties des patients sont moins fréquentes le week-end. Par exemple, dans le cas ou la sortie du patient est initialement prévue 4 jours après son admission et que ce jour coïncide avec un jour de week-end, la DDS sera prolongée d'un ou deux jours. Les périodes d'épidémies sont à prendre en compte également. Pendant ces périodes, la charge du travail est beaucoup plus importante et les ressources des hôpitaux sont fortement demandées. Elles sont consacrées aux patients admis en service concerné par l'épidémie. Par exemple, avec l'apparition du coronavirus en

décembre 2019 [Dor+20], plusieurs études se sont intéressées à la prédiction des DSS pour les patients affectés par le coronavirus.

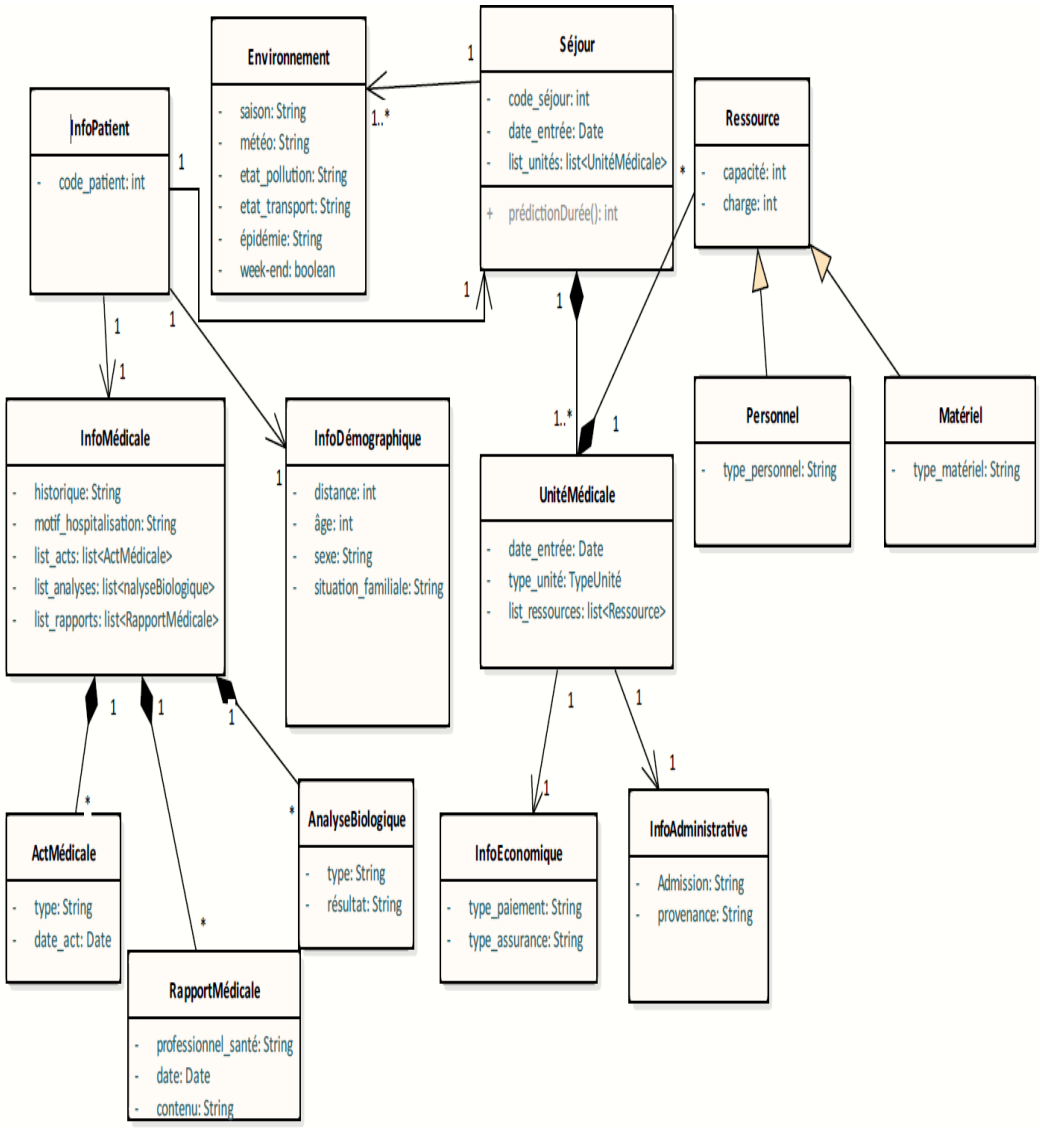


Fig. 2.3: Modèle générique de durée de séjour hospitalier.

Le digramme de classe de la figure 2.3 présente un modèle générique de DDS déduit et enrichi à partir de l'analyse des travaux précédents. Le modèle intègre également les contraintes liées à la vie quotidienne en termes d'organisation de l'hôpital et de la présence de forte demande pour raison de pandémie.

La prédiction de DDS se base sur le modèle de DDS défini. Nous allons dans ce qui suit, présenté les différentes méthodes de prédiction de DDS citées dans la littérature et montrer celles qui sont le plus souvent utilisées.

2.3 Méthodes de prédiction des Durées De Séjour hospitalier

La capacité à prédire la durée de séjour hospitalier au moment de l'admission du patient est essentielle pour une meilleure planification et allocation des ressources [WPS03]. En particulier, lorsque les ressources hospitalières sont limitées comme dans les unités de soins intensifs, les unités de cardiologie et les unités de chirurgie [Ver+07 ; Yan+10]. Des modèles de prédiction de DDS ont été présentés dans la littérature. Dans le contenu suivant, nous allons présenter les principales méthodes de prédiction de DDS citées dans la littérature.

2.3.1 Méthodes statistiques et Chaîne de Markov cachées

Le choix de la méthode de prédiction dépend essentiellement de l'objectif, de la taille, de la qualité et de la nature de la base de données utilisée. Dans l'étude menée par [Ngu+21], 1189 instances contenant des dossiers patients ont été analysées en utilisant des méthodes statistiques paramétriques et non paramétriques. La famille des distributions de Poisson et l'histogramme sont utilisés dans cette étude. Une analyse de régression multiple a également été réalisée pour modéliser la durée de séjour en fonction de plusieurs variables indépendantes. Dans [Alm+16], une régression logistique est utilisée pour déterminer les longues DDS suite à une chirurgie cardiaque. Une durée de séjour est reconnue comme longue si elle dépasse le 75 ème centile de la valeur de la DDS. De plus, les coefficients statistiques sont employés pour déterminer la DDS. Le t-tests est utilisé avec les variables continues, le coefficient de Mann-Whitney avec les variables dont la distribution est non gaussienne et le test du chi-2 avec les variables catégorielles.

Dans les années 1970, les premières études sur les DDS ont exploré des approches issues de la recherche opérationnelle, en utilisant des techniques telles que les modèles de Markov cachés. En effet, la modélisation de la DDS et l'estimation de sa moyenne en utilisant les chaines de Markov a montré son efficacité dans [FGP09]. Quoique ces méthodes ont abouti à des prédictions correctes, elles présentent plusieurs inconvénients. Elles sont d'une complexité temporelle importante et ne s'adaptent souvent pas à l'environnement hospitalier dynamique (données volumineuses, incertaines, complexes et incomplètes).

Les méthodes basées sur l'apprentissage automatique s'adapte plus souvent à l'environnement hospitalier et gère d'une meilleure façon les difficultés au-quelles il faut

faire face en utilisant les données médicales. Dans la suite, nous présentons un état de l'art sur ces méthodes dans la prédiction des DDS.

2.3.2 Méthodes basées sur l'apprentissage automatique

Une autre famille de méthodes de prédiction est celle issue de l'Intelligence Artificielle dont l'apprentissage automatique ou Machine Learning en anglais (ML) et la fouille de données ou le data mining en anglais. Ces techniques ont fait un retour fracassant dans le domaine de la santé, et particulièrement, dans la prédiction de DDS. Dans [Hac+13], les Réseaux de Neurones Artificielles (ANN), les Machines à Vecteurs de Support (SVM) et les arbres de décision sont utilisés dans la prédiction de la DDS pour des patients souffrants d'une maladie coronarienne. Les arbres de décision sont également utilisés pour l'extraction des règles de prédiction. Les SVM ont donné les meilleures performances pour l'ensemble des données utilisées dans cette étude. En plus des ANN, l'algorithme du Random Forest (RF) est utilisé dans [Gen+17] pour la classification de la DDS en deux groupes : DDS inférieure à 5 jours et DDS supérieure ou égale à 5 jours. La précision du modèle est de 80% ce qui est satisfaisant par rapport à son efficacité. Dans [Chu+16], les auteurs comparent les résultats obtenus avec les SVM, les arbres de décision et le Random Forest pour la prédiction de DDS longue suite à une chirurgie. Les résultats indiquent que l'algorithme RF constitue le modèle de prédiction le plus précis et le plus stable. Ces derniers algorithmes sont utilisés en complément de la régression logistique dans [CHL18] afin d'identifier les DDS prolongées. Dans le travail présenté par [Li+13] les ANN sont employés pour concevoir deux modèles de prédiction de DDS : un premier modèle comprenant toutes les variables et un autre avec un sous-ensemble de variables fortement corrélées à la DDS. La corrélation est étudiée à l'aide du coefficient Chi-2. Dans une autre étude [And19], le RF, les SVM, les ANN, les arbres de décision et le Gradient Boosting model (GBM) sont explorés pour la prédiction des DDS au moment de l'admission du patient. Les arbres de décision et le Random Forest ont abouti tous deux à un très haut degré d'interprétabilité et ils surpassent les autres algorithmes dans la prédiction. Le GBM a été exploré afin d'analyser si la méthode du boosting peut améliorer les résultats de prédiction. Un autre modèle d'apprentissage automatique se basant sur les GBM est proposé dans la prédiction des DDS des patients atteints du diabète. La précision de cet algorithme est de 80% ce qui est considéré comme un taux satisfaisant [AMB18].

Nous avons résumé dans le tableau 2.1 ci-dessous les principaux travaux concernant la prédiction de DDS. D'une part, les facteurs impactant les DDS sont mis en évidence. D'autres part, les différentes techniques d'apprentissage automatique sont reportées.

	Facteurs impactant DDS			
Algorithmes ML	***Démographiques***	***Médicaux***	***Administratives***	***Meilleurs résultats***
Arbres de décision Random Forest SVM [Chu+16]	oui	oui	non	Random Forest
Random Forest ANN [Gen+17]	oui	oui	oui	Random Forest
ANN SVM Arbres de décision [Hac+13]	oui	oui	oui	SVM
Random Forest Arbres de décision SVM [Hac+13]	oui	oui	oui	Random Forest
Random Forest GBM Arbres de décision SVM [And19]	oui	oui	non	Random Forest + Arbres de décision

Tab. 2.1: Méthodes de prédiction et facteurs impactant la DDS.

A partir de ces travaux, nous avons constaté que les algorithmes d'apprentissage automatique particulièrement celles d'apprentissage supervisé sont largement utilisées dans la prédiction des Durées De Séjours hospitaliers. Dans ce qui suit, nous présentons les principaux algorithmes issus de l'apprentissage automatique explorés dans notre étude.

2.4 Apprentissage automatique dans la prédiction des DDS

L'apprentissage automatique ou le machine learning en anglais est une branche de l'intelligence artificielle qui se focalise sur l'apprentissage à partir de données. Elle consiste à apprendre d'une manière autonome des motifs récurrents et effectue des tâches en améliorant leur performance à partir de ces données stockées numériquement. Une définition connue de ***Tom M. Mitchell*** de l'apprentissage automatique est la suivante : « On dit qu'un programme informatique apprend de l'expérience E par rapport à un type de tâches T et une mesure de performance P, si sa performance aux tâches de T, telle que mesurée par P, s'améliore avec l'expérience E » [MM97]. En machine learning, les algorithmes sont "entraînés" à trouver des modèles et des caractéristiques dans des quantités massives de données pour une prise de décision sur de nouvelles instances de données [IBM20]. Il existe 4 types d'apprentissage automatique : l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage semi-supervisé et l'apprentissage par renforcement. La figure 2.4 illustre ces types.

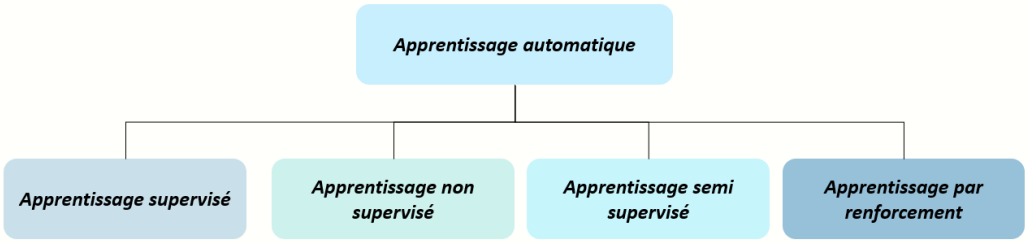


Fig. 2.4: Types d'apprentissage automatique.

- Apprentissage supervisé :

  Il s'agit de l'apprentissage automatique le plus couramment utilisé. Il apprend à partir des exemples. C'est un mécanisme qui déduit la relation qui existe entre les données observées (données en entrée) et une variable cible qui représente la donnée en sortie. Lors de l'apprentissage, l'algorithme tente de découvrir cette relation. Après l'apprentissage, un modèle est conçu et son objectif est de prédire l'étiquette correcte pour les données d'entrée nouvellement présentées. A son tour, l'apprentissage supervisé se divise en deux sous catégories : la classification et la régression. Nous allons aborder en détails ces deux sous-catégories par la suite dans la section 2.5.

- Apprentissage non supervisé :

  Dans l'apprentissage non supervisé (clustering en anglais), l'ensemble de données est divisé en sous-groupes homogènes pour obtenir une représentation simplifiée de l'ensemble de départ. Les données provenant de groupes différents sont bien séparées. Le but de cette approche est de ressortir la véritable structure des données pour extraire des connaissances.

- Apprentissage semi-supervisé :

  Ce type d'algorithmes est la combinaison entre l'apprentissage supervisé et l'apprentissage non supervisé. Ces algorithmes sont capables d'apprendre à partir d'ensembles de données partiellement étiquetées. En effet, ils utilisent une combinaison d'une petite quantité de données étiquetées et d'une grande quantité de données non étiquetées pour former des modèles. Une des applications les plus connues des algorithmes d'apprentissage semi-supervisé est la classification des documents textuels. En effet, l'étiquetage d'un ensemble volumineux de documents consomme énormément de temps. Une solution est de présenter un sous-ensemble étiqueté et d'utiliser l'apprentissage semi-supervisé dans le processus de classification.

- Apprentissage par renforcement :

  Dans ce type d'apprentissage, un ensemble d'agents interagissent les uns avec les autres dans un environnement dynamique pour atteindre leurs objectifs.

L'apprentissage des agents est effectué à partir d'une série de renforcements, de récompenses ou de punitions contrairement à l'apprentissage supervisé. Comme ces agents sont dotés d'un processus de renforcement, ils peuvent évoluer en apprenant de leur environnement [Bis06].

La plupart des problèmes de prédiction des DDS évoqués dans la littérature implémentent les algorithmes d'apprentissage supervisé car ils ont montré leur efficacité. Notre variable cible étant la durée de séjour, nous allons utiliser ces algorithmes pour trouver une relation entre les variables en entrées ou variables indépendantes et la variable en sortie ou la variable dépendante DDS. Les variables indépendantes sont celles qui caractérisent le modèle de DDS. Nous allons, dans ce qui suit, présenter la classification et la régression.

2.5 Apprentissage supervisé et durée de séjour hospitalier

L'apprentissage supervisé est une forme de l'apprentissage automatique. En apprentissage supervisé, l'entrée se compose d'un ensemble de données étiquetées. Cela signifie que les données sont décrites par un ensemble de variables dépendantes et dune variable cible à prédire. Il y a N individus ou instances et $P+1$ variables. Un ensemble de K classes est construit suivant les modalités des $P+1$ variables afin d'attribuer une valeur à la variable cible lorsque celle-ci est inconnue. Il existe deux grandes familles d'algorithmes d'apprentissage automatique supervisé :

- Les algorithmes d'apprentissage paresseux : ces algorithmes stockent simplement les données. Ils débutent le processus de classification lorsqu'ils reçoivent des données de tests. Par conséquent, ils prennent plus de temps dans la phase de tests que dans l'apprentissage du modèle. Des exemples de ce type d'algorithme sont le k-plus proches voisins et le raisonnement à partir de cas.
- Les algorithmes d'apprentissage désireux : l'apprentissage commence lorsque les données sont disponibles. Ils n'attendent pas de données de test alors pour apprendre. La phase d'apprentissage prend plus de temps que celle de test. Plusieurs algorithmes rentrent dans cette catégorie comme les arbres de décision et les réseaux de neurones.

Les données utilisées dans l'apprentissage supervisé peuvent être structurées ou non. Les données structurées contiennent les variables numériques, catégorielles et catégorielles multivaluées. Concernant les données non structurées, sont sous format

texte, sous format image et les données provenant du web. Dans le cadre de la prédiction de la DDS, ces données peuvent être déduite à partir du modèle générique de la DDS proposé dans la figure 2.3. Les données en entrée sont introduites dans un modèle d'apprentissage automatique et une fonction de perte est à optimiser. Le modèle après apprentissage est évalué à l'aide d'une fonction d'évaluation pour en déduire le meilleur modèle de prédiction.

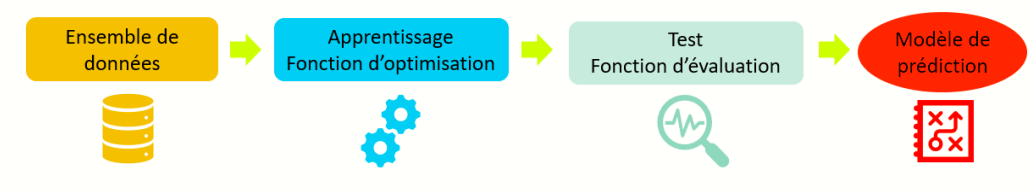


Fig. 2.5: Éléments d'une méthode d'apprentissage automatique.

Le problème de l'apprentissage se constitue essentiellement de 3 éléments : une représentation des données, une fonction d'optimisation et une fonction d'évaluation. La figure 2.6 illustre les fonctions d'optimisation et les fonctions d'évaluation pour la classification et la régression.

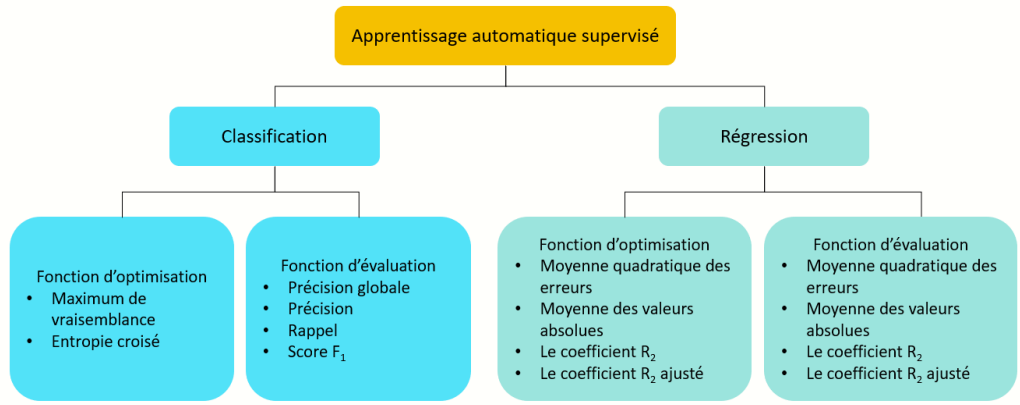


Fig. 2.6: Types d'apprentissage automatique.

Dans ce qui suit, nous allons présenter en détails la classification et la régression.

2.5.1 Classification

L'objectif de la classification est de prédire une classe ou catégorie à partir d'un ensemble de variables. La prédiction consiste à déterminer une fonction approximative qui lie les variables d'entrée à la variable qualitative de sortie. Cette fonction sert à catégoriser les nouvelles données. Les méthodes de classification sont de 3 types :

- La classification binaire : La valeur à prédire est binaire.

- La classification multi-classe : La valeur à prédire contient plus de deux sorties. Dans ce type de classification, chaque individu est affecté à une et une seule classe.
- La classification multi-label : Chaque individu est affecté à un ensemble de catégories ou de classes.

Dans un problème de classification, on cherche à optimiser une fonction de perte afin d'améliorer les performances du modèle. Pour modéliser le problème de classification, nous illustrons *Y* comme l'indicateur de classe d'un groupe de *K* classes mutuellement exclusives, représentées par un vecteur *X* contenant *P* variables. L'idée est d'estimer la probabilité que les données représentées par le vecteur *X* : $<x_1, x_2, ..., x_p.>$ appartiennent à la classe y_k en utilisant les paramètres *W* [Yer19]. Deux solutions sont possibles : l'estimation du maximum de vraisemblance ou la minimisation de l'entropie croisée (cross-entropy loss).

- Maximum de la vraisemblance : La classe avec la probabilité la plus élevée est affectée à chaque nouvelle instance de données. L'avantage de cette méthode paramétrique est qu'elle prend en compte la variance et la covariance dans la distribution des classes. De ce fait, pour des données avec une distribution normale, la méthode est plus performante que d'autres [OB10]. La formule ci-dessous représente la forme générale de la fonction du maximum de la vraisemblance.

$$p(y|x,w) = \prod_{k=1}^{K} \mu_k^{y^k} \tag{2.1}$$

où μ est la variance et :

$$p(y_k = 1|x,w) = \mu_k, avec : 0 \leq \mu_k \leq 1 et \sum_k \mu_k = 1 \tag{2.2}$$

- Entropie croisée :

  Cette métrique, connue comme la fonction de perte ou log loss, mesure les performances d'un modèle de classification dont la sortie est une valeur de probabilité. Sa valeur augmente lorsque la probabilité prédite s'écarte de l'étiquette réelle. Un modèle parfait aurait une valeur de log loss égal à 0. *N* étant le nombre d'instances dans l'ensemble de données, la formule de l'entropie croisé est la suivante [Yer19] :

$$\sum_{n=1}^{N} \sum_{k=1}^{K} y_{k,n} log(\mu_{k,n}) \tag{2.3}$$

Afin d'évaluer les méthodes de classification, des mesures de performances sont calculées. La matrice de confusion, la précision globale, la précision, le rappel et le score F_1 sont utilisés.

- Matrice de confusion :

  Une matrice de confusion est une matrice telle que chaque ligne représente les valeurs prédites et chaque colonne représente les valeurs réelles de la variable cible. Pour un problème de classification binaires, deux classes à prédire sont déterminées. Nous illustrons dans la figure 2.7 les terminologies nécessaires pour la compréhension des mesures d'évaluation des méthodes de classification. Plus leur valeur tend vers 1, plus le modèle est performant.

		Valeurs réelles	
		Positives	Négatives
Valeurs prédites	Positives	**Vraies positives (TP)**	**Fausses positives (FP)**
	Négatives	**Fausses négatives (FN)**	**Vraies négatives (TN)**

Fig. 2.7: Matrice de confusion.

- Précision globale : elle représente le nombre d'instances correctement prédites sur le nombre total d'instances. Cette mesure est utilisée quand les classes de l'ensemble de données sont bien équilibrées et non biaisées.

$$\text{précision globale} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

- Précision : la précision est le nombre d'exemples positifs correctement classés divisé par le nombre total d'exemples qui sont classés comme positifs.

$$\text{précision} = \frac{TP}{TP + FP} \tag{2.5}$$

- Rappel : le rappel représente le nombre d'exemples positifs correctement classés, divisé par le nombre total d'exemples positifs réels dans l'ensemble de tests.

$$rappel = \frac{TP}{TP + FN} \tag{2.6}$$

- Le score F_1 : Cette mesure comprise entre 0 et 1 représente la moyenne harmonique de la précision et du rappel.

$$F_1 = 2 * \frac{\text{précision} * rappel}{\text{précision} + rappel} \tag{2.7}$$

2.5.2 Régression

A la différence des problèmes de classification, les problèmes de régression tentent de prédire une valeur quantitative. La variable cible est une valeur numérique "expliquée" par un ensemble de variables numérique, catégorielles ou catégorielles multivaluées. Les modèles de régression cherchent à minimiser une erreur qui est l'écart entre la valeur de la variable cible et la valeur prédite retournée par le modèle. Le choix de ces fonctions de perte et des mesures d'évaluations dépendent de l'objectif de l'étude et la qualité des données utilisées. Nous citons dans ce qui suit les mesures que nous avons utilisées dans notre étude, la moyenne quadratique des erreurs (*MSE*) et la moyenne absolue des erreurs (*MAE*) comme fonctions de perte et d'évaluation des méthodes de régression. De plus, nous présentons le coefficient de détermination R_2 et le R_2 ajusté.

- Moyenne quadratique des erreurs (*MSE*) : elle représente la mesure la plus employée dans les problèmes de régression. Elle est la somme des écarts au carré entre les valeurs réelles et les valeurs prédites. Elle est plus simple à utiliser mais moins robuste face à la présence des données aberrantes. Elle aboutit à une solution plus stable [Gro18]. *N* est le nombre d'instances, $\hat{y}$ représente la valeur prédite et *y* représente la valeur réelle. La *MSE* est calculée comme suit :

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{2.8}$$

- Moyenne des valeurs absolues des erreurs (*MAE*) : Elle est utilisée quand l'ensemble de données contient du bruit ou des données aberrantes. Elle représente la somme des différences absolues entre les valeurs réelles et les valeurs prédites [Gro18].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{2.9}$$

- Le R_2 : Le R_2 ou le coefficient de détermination définit la variation dans la variable dépendante (ou cible) en utilisant les variables indépendantes (caractéristiques). Plus le R_2 est proche de 1, meilleur est le modèle. Un

inconvénient majeur de cette mesure est qu'elle suppose que chaque variable indépendante explique la variation de la cible. Ce n'est pas toujours vrai car il arrive d'utiliser des variables qui sont peu importantes dans la prédiction.

$$R_2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y}_i - y_i)^2} \tag{2.10}$$

Une alternative est d'utiliser une version améliorée du R_2 qui est le R_2 ajusté [Sin19].

- Le R_2 ajusté : similaire au R_2, le R_2 ajusté explique la variation dans la variable cible en utilisant uniquement les variables indépendantes qui contribuent à la création du modèle de prédiction. Par conséquent, contrairement au R_2, le R_2 ajusté donne moins de poids aux variables non nécessaires pour la prédiction. Comme pour le R_2, cette valeur doit être proche de 1 [Sin19]. *N* représente le nombre d'instances et *P* représente le nombre de variables indépendantes dans l'ensemble de données.

$$R_2 \textit{ ajusté } = 1 - \frac{(1 - R_2)(N - 1)}{(N - P - 1)} \tag{2.11}$$

Afin de mieux comprendre les algorithmes utilisés dans notre étude, une brève présentation est donnée dans ce qui suit.

2.6 Algorithmes d'apprentissage automatique supervisé

Il existe une panoplie d'algorithmes d'apprentissage automatique supervisé. Dans cette section, nous présentons l'ensemble d'algorithmes d'apprentissage automatique supervisé utilisés dans notre étude. Ces algorithmes peuvent être utilisés pour des problèmes de classification et pour des problèmes de régression. Nous avons choisi ces algorithmes en analysant les travaux précédents et compte tenu de la qualité et de la complexité des données médicales. Afin d'introduire les méthodes d'apprentissage ensembliste sur la base des arbres de décision, nous expliquons d'abord le concept général et la démarche des arbres de décision pour la prédiction.

La figure 2.8 résume les différentes méthodes d'apprentissage automatique que nous allons présenter dans ce qui suit.

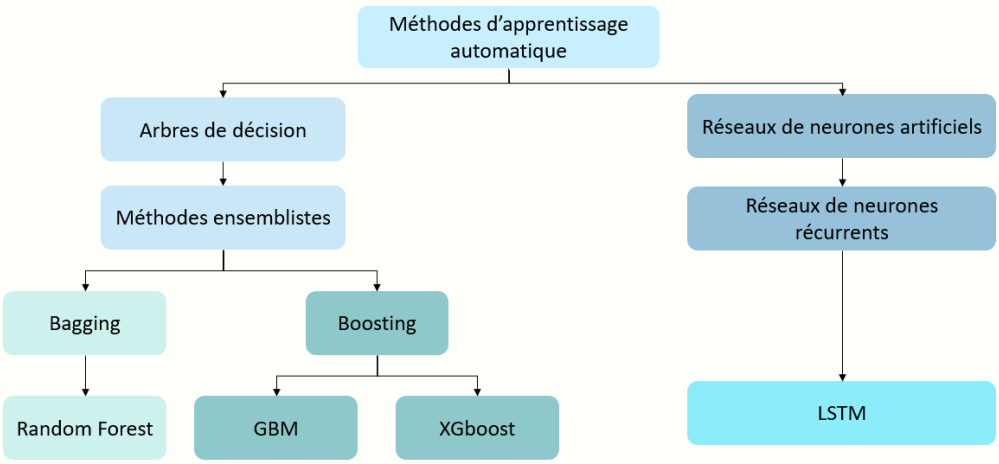


Fig. 2.8: Méthodes d'apprentissage automatique.

2.6.1 Arbres de décision

Un arbre est un ensemble de nœuds et d'arêtes organisés de manière hiérarchique. Chaque nœud stocke une fonction de test à appliquer aux données entrantes et chaque feuille stocke la réponse finale (la valeur prédite). Les arbres de décision ont été largement utilisés depuis plusieurs années et ont prouvé leur efficacité pour un grand nombre de problèmes de ML [Yan19]. Le concept d'un arbre de décision implique le choix des variables à prendre dans le modèle de la racine aux feuilles de l'arbre et aussi le choix des conditions à vérifier pour effectuer la division et l'arrêt du calcul. Les arbres de décision sont utilisés dans la prédiction et l'analyse des variables impliquées dans les modèles de prédiction. Ils ont été améliorés pour obtenir de meilleures performances. Pour cela, les méthodes d'apprentissage ensemblistes sont apparues. L'idée de ces méthodes est de combiner plusieurs méthodes dans un même modèle. Ceci permet d'améliorer les performances des méthodes classiques avec un algorithme individuel. Principalement, il existe deux techniques pour l'implémentation des méthodes ensemblistes : bagging et boosting. La figure 2.9 introduit la principale différence entre ces deux méthodes et les arbres de décisions classiques que nous allons détaillé dans ce qui suit.

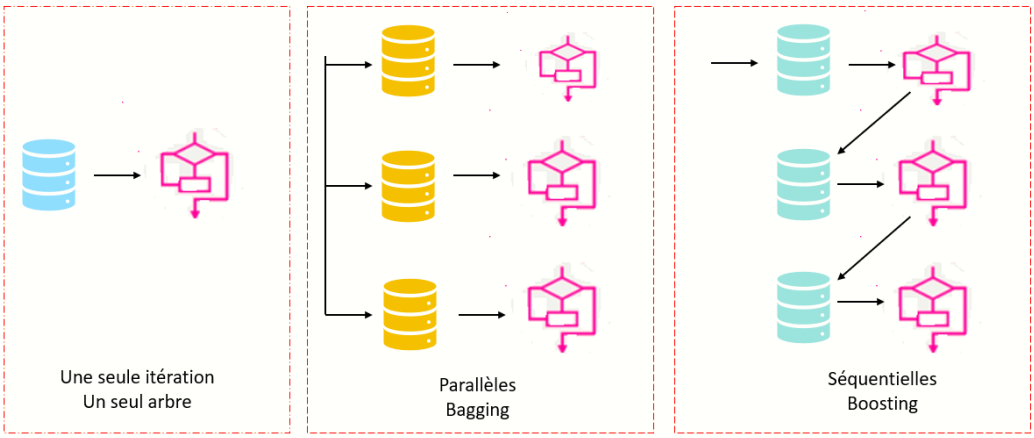


Fig. 2.9: Différence entre le bagging et le boosting.

Nous définissons ces deux approches puis nous explorons des algorithmes fondés sur ces approches.

- Bagging : Le bagging a été introduit par Leo Breiman en 1996 [Bre96]. Cette technique vise à réduire l'erreur d'apprentissage par la mise en œuvre d'un ensemble d'algorithmes d'apprentissage automatique homogènes. Ceci est réalisé en utilisant plusieurs algorithmes d'apprentissage de base formés séparément avec un ensemble de données d'apprentissage aléatoires. Le modèle produit est plus stable et précis. Il se construit soit par vote majoritaire dans le cadre de la classification soit par la moyenne dans le cadre d'une régression. Le processus du bagging est simple à produire. N sous-ensembles de données sont extraits à partir de l'ensemble original. Ils sont utilisés dans l'apprentissage de N algorithmes. L'apprentissage est donc réalisé d'une manière parallèle et les erreurs sont calculées en modifiant les poids des variables [Lóp21].
- Boosting : Cette technique génère des hypothèses d'une façon séquentielle. Chaque hypothèse tente d'optimiser la fonction de perte d'une étape à l'autre [Sch03].Les erreurs sont calculées en minimisant une fonction de perte ce qui diffère du bagging [Lóp21].

2.6.2 Les forêts d'arbres décisionnels

Les forêts d'arbres décisionnels (Random Forest - RF), est un ensemble d'arbres de décision formés d'une façon aléatoire [CSK11]. Le processus d'apprentissage est répété indépendamment pour chaque arbre. Le choix aléatoire des arbres est utilisé dans la phase d'apprentissage, tandis que la phase de test est complètement déterministe. La puissance des RF réside dans le fait que différents arbres produisent une précision beaucoup plus élevée sur l'ensemble de test [Bre01]. Ce mécanisme

est nommé la généralisation. Elle réduit la corrélation entre les arbres indépendants et offre, ainsi, une grande robustesse face aux données erronées [CSK11]. Cet algorithme se base sur la technique du bagging décrite ci-dessus.

2.6.3 L'amplification du gradient

L'algorithme de l'amplification du gradient (Gradient Boosting Model - GBM) est un algorithme basé sur les arbres de décision et appartenant à la famille des méthodes ensemblistes. Il se base sur le boosting. Il forme N arbres de manière graduelle, additive et séquentielle. Chaque nouvel arbre est un ajustement sur une version modifiée de l'ensemble de données originales. Le GBM cherche à optimiser une fonction de perte utilisée dans l'apprentissage et qui mesure la qualité de l'ajustement des coefficients du modèles à chaque étape d'apprentissage. Le passage séquentiel d'un arbre à un autre, doit assurer la minimisation de cette mesure [Sin18].

2.6.4 Extreme Gradient Boosting Model

Extreme Gradient Boosting Model (XGboost) est l'un des algorithmes d'apprentissage supervisé les plus performants. C'est une version améliorée du GBM qui prend en compte des paramètres afin d'améliorer le modèle de prédiction. Il est connu pour son efficacité car il assure le parallélisme sur l'ensemble des micro-processeurs de la machine et la réduction du temps de calcul. De plus, il introduit la notion de régularisation qui permet d'éviter le sur-apprentissage. Le modèle sera donc facilement généralisable par la suite. Il gère également les données manquantes et capture les relations non linéaires entre les données. Enfin, cet algorithme se base sur le calcul de la deuxième dérivée dans la réduction du taux d'erreur de la fonction de perte afin de mieux guider la direction du gradient.

2.6.5 Les réseaux de neurones récurrents

Les réseaux de neurones artificiels (Artificial Neural Network - ANN) ont montré leur efficacité dans la prédiction des DDS. Il existe plusieurs types de réseaux de neurones. Dans cette thèse, nous nous sommes intéressés particulièrement aux réseaux de neurones récurrents (Reccurent Neural Network - RNN). Ce type d'algorithme rajoute une dimension temporelle aux données en entrée. L'entrée des RNN est une série temporelle et chaque instance de données est ordonnée chronologiquement [BD02]. L'architecture de ces modèles permet d'alimenter les sorties d'un neurone itérativement [MVM11]. De ce fait, Les neurones dépendent donc de l'état d'entrée ainsi que de leur état interne [Bod01].

D'une manière simplifiée, un ANN est un graphe orienté avec des arêtes. L'architecture la plus simple des ANN est la feed-forward qui est un réseau multicouches ou chaque unité est connectée à celle d'après sans cycle. Les sorties de chaque couche d'unités de calcul sont transmises à la couche supérieure suivante. Les données d'entrée sont dirigées à travers les couches intermédiaires jusqu'à ce qu'elles atteignent la couche de sortie et le résultat de prédiction est donc retourné [JJ20]. Les ANN se compose d'une seule couche d'entrée, une seule couche de sortie et d'une ou plusieurs couches cachées.

Un exemple des RNN est les LSTM (Long-Short Term Memory). Cette technique surgie à travers des dépendances à long terme où chaque cellule mémorise l'information pour une longue durée. Lors de l'entraînement d'un réseau de neurones classique en utilisant la descente du gradient et la méthode de rétro-propagation [Mic19], la valeur du gradient tend vers une valeur très petite. Ceci empêche la modification des poids et pose un problème car les calculs se basent sur des valeurs numériques à précision finie. Les LSTM dans ce cas permettent au gradient de garder les flux inchangés. Tandis qu'il existe plusieurs type de LSTM, une unité d'un LSTM est composée d'une porte d'entrée, une porte de sortie et une porte d'oubli [BH19]. Ces portes sont capables d'apprendre quelles données dans une séquence est à garder garder ou à supprimer. Ainsi, les informations pertinentes sont transmises dans la longue chaîne de séquences pour faire des prédictions.

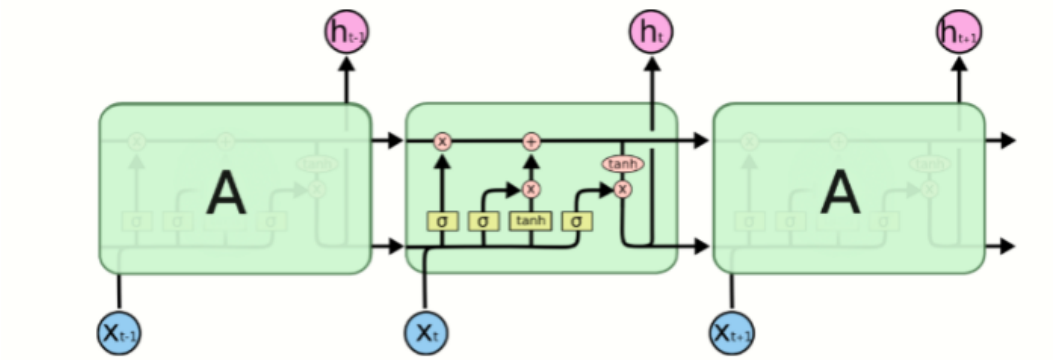


Fig. 2.10: Architecture classique d'un modèle LSTM.

Le concept de base des LSTM est la mémoire de la cellule et les différents portes. La porte d'entrée décide si l'entrée doit modifier le contenu de la cellule. La porte de sortie décide s'il faut remettre à 0 le contenu de la cellule et celle d'oubli décide si le contenu de la cellule doit influer sur la sortie du neurone. La figure 2.10 illustre l'architecture d'un LSTM classique.

Tout d'abord, la porte d'oubli décide quelles informations doivent être oublier ou conserver. Les informations de l'état caché précédent et celles de l'entrée courante

sont passées par une fonction sigmoid qui remet les valeurs entre 0 et 1. Tout nombre multiplié par 0 est égal à 0, ce qui fait que les valeurs sont oubliées. Tout nombre multiplié par 1 correspond à la même valeur, par conséquent, cette valeur est gardée. Ensuite, l'état de la cellule est mis à jour à l'aide de la porte d'entrée. L'état caché précédent et l'entrée courante sont passés par une fonction sigmoid. Ils sont également passés dans une fonction tanh qui transforme leurs valeurs entre -1 et 1. Les sorties de ces deux fonctions sont multipliées pour décider quelles informations faire passer. Par la suite, la sortie de la porte d'oubli est multipliée par les valeurs de la cellule courante. Une addition avec le vecteur de la porte d'entrée est ensuite réalisée. Au final, la porte de sortie opère une fonction sigmoid avec les valeurs de l'état caché précédent et l'état actuel. Le résultat est ensuite injecté dans une fonction tanh. Les deux résultats obtenus des deux fonction sigmoid et tanh sont multipliés pour décider quelles informations doivent être passées.

2.7 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur les définitions de la DDS, les facteurs qui l'impactent et les méthodes utilisées pour sa prédiction. Dans la littérature, les travaux qui concernent la prédiction des DDS sont très souvent limité à une unité médicale précise ou liée à une pathologie. Ces travaux ont une faible capacité à généraliser la prédiction des DDS à plusieurs unités médicales. De plus, les techniques présentées se limite à la prédiction de DDS au moment de l'admission du patient et n'intègrent pas les données disponibles après l'admission du patient. Cependant, il nous semble plus intéressant de concevoir un modèle qui généralise la prédiction de DDS à plusieurs unités médicales et met à jour la valeur de DDS en incluant de nouvelles données disponibles pendant le séjour hospitalier.

Dans notre travail, nous avons étudié d'une manière approfondie les éléments apparus dans la recherche bibliographique et analysé les besoins quotidiens des hôpitaux en terme d'organisation, charge du travail et ressources disponibles. A partir de cela, une nouvelle modélisation de la DDS est proposée. Elle considère plusieurs facteurs présents dans le milieu hospitalier.

Concernant la phase de prédiction, plusieurs méthodes de prédiction sont utilisées à savoir, les approches statistiques, les chaine de Markov cachées et les méthodes d'apprentissage automatique. Nous avons proposé une solution qui assure la prédiction de DDS au moment de l'admission et tout au long le séjour hospitalier dans différentes unités médicales. En outre, l'application de nos méthodes est réalisée sur un ensemble de données réelles issues du Programme de Médicalisation des Systèmes d'Informations (PMSI) présenté dans le chapitre précédent. Cet ensemble

de données présentent toutes les difficultés et propriétés des données médicales. C'est sur ces points que porte notre travail dans les prochains chapitres.

3

Modèle statique de prédiction des Durées De Séjour hospitalier

> *If we knew what it was we were doing, it would not be called research, would it ?*
>
> — **Albert Einstein**
> Prix Nobel de physique.

3.1 Introduction

Les établissements de soins cherchent sans cesse à optimiser le fonctionnement de leurs services tout en assurant la qualité des soins et la sécurité des patients. L'environnement hospitalier fait intervenir différents acteurs qui peuvent le modifier comme les soignants, les médecins, les financiers, les logisticiens et les patients. La planification des activités et la gestion des ressources ont un impact important sur cet environnement. L'estimation de la Durée De Séjour hospitalier (DDS) d'un patient au moment de son admission et durant son séjour hospitalier constitue un indicateur clé d'évaluation des hôpitaux et des services de soins et de santé. La DDS est identifiée comme une variable complexe dépendant de plusieurs facteurs liés au contexte médical du patient, aux conditions de son admission et à l'organisation de l'hôpital ou du service hospitalier.

Dans ce chapitre, nous proposons un premier modèle de prédiction de Durées De Séjour hospitalier au moment de l'admission d'un patient que nous appelons modèle statique de prédiction de DDS. Ce modèle se base sur des données disponibles au moment de l'admission. Nous allons ainsi détailler nos différentes méthodes et solutions de prédiction en partant des travaux de l'état de l'art présentés dans le chapitre précédent. Ce chapitre est structuré comme suit. D'abord, nous définissons notre démarche méthodologique et nous expliquons chaque étape de celle-ci. Ensuite, nous introduisons notre méthode de prédiction de DDS en se basant sur des algorithmes d'apprentissage automatique supervisé et non supervisé. Nous présentons notre procédure d'implémentation algorithmique de cette proposition. Enfin, nous terminons le chapitre par une conclusion qui récapitule les principaux éléments de nos contributions pour le modèle statique de prédiction de DDS.

3.2 Méthodes de prédiction de Durée De Séjour hospitalier

L'environnement hospitalier est un environnement complexe qui regroupe plusieurs acteurs. D'une part, de spécialité médicale tels que les médecins, les infirmiers et les biologistes. D'autre part nous retrouvons les administratifs, les financiers et les logisticiens. Dans ce contexte institutionnel et organisationnel la définition du séjour hospitalier ainsi que la Durée De Séjour hospitalier (DDS) doit prendre en compte cette dynamique et interaction entre plusieurs acteurs. Il est nécessaire de positionner la DDS au sein du séjour hospitalier afin de mieux l'analyser. Afin d'y répondre, nous avons commencé par définir le périmètre d'étude représentant la zone sur la quelle nous nous concentrons dans l'établissement de soins. Ensuite,

nous avons présenté un modèle qui caractérise la DDS et qui englobe les facteurs qui l'impactent. Enfin, un processus de prédiction est mis en place. Ce processus se base sur des techniques de fouille de données et d'apprentissage automatique. Nous exposons en détails ces étapes et nous justifions le choix des méthodes utilisées dans ce projet de thèse.

3.2.1 Périmètre d'étude et Durée De Séjour

Une première étape de notre démarche est de cerner le périmètre d'étude pour mettre en évidence la relation entre la définition du séjour hospitalier et de la DDS. Le périmètre d'étude représente alors le secteur où la DDS sera considérée. La définition du périmètre d'étude permet d'identifier l'ensemble des facteurs qui impactent la DDS et de savoir comment la DDS a été abordée dans les travaux de la littérature.

Dans le cadre de cette thèse, nous nous sommes focalisés sur les services médicaux hors urgence et ambulatoire ce qui constitue notre hypothèse de départ. Dans ce cas, la DDS ne sera pas calculée sur une journée (en nombre d'heures) mais en nombre de jours. Comme les objectifs de la prédiction de la DDS est de faciliter la planification des activités des établissements de soins et d'améliorer leur organisation, la structure de l'hôpital doit être prise en considération dans la définition de la DDS. Afin de généraliser les modèles de prédiction, nous avons défini la DDS dans plusieurs sites hospitaliers pour que les modèles de prédiction soient opérationnels par la suite.

La définition du périmètre d'étude concernant la DDS est en effet en relation avec la définition du séjour hospitalier. Dans le Programme de Médicalisation des Systèmes d'Informations, un séjour hospitalier est composé d'un ou de plusieurs passages par des unités médicales. Nous avons montré d'un autre côté que la DDS est fortement liée à l'unité médicale dans laquelle le patient est admis [RIG09 ; PK14]. Lors d'un passage par une seule unité médicale, la DDS est définie comme l'intervalle de temps entre la date d'admission du patient et sa date de sortie de l'unité médicale concernée. Dans le cas de plusieurs passages par plusieurs unités médicales, cette DDS est représentée par la somme des DDS de chaque unité médicale qui constitue le séjour hospitalier. Par conséquent, afin d'approfondir notre étude, nous nous sommes intéressés à plusieurs unités médicales avec des profils patients différents. Le service de cardiologie, le service de médecine polyvalente, le service de pédiatrie et le service de néonatologie ont été sélectionnés. Le choix de ces unités médicales s'est fondé sur la quantité et la diversité des données. Ces données contiennent les informations sur différents patients, différentes unités médicales et différents sites hospitaliers. La figure 3.1 présente les unités médicales choisies dans le cadre du PMSI.

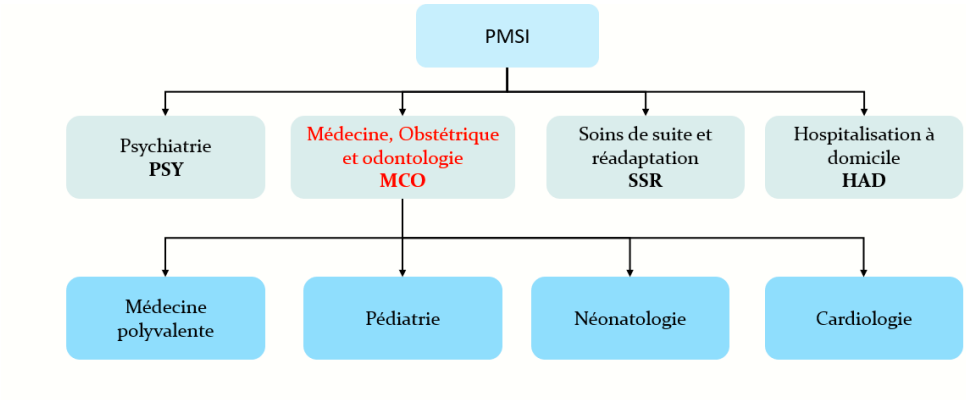


Fig. 3.1: Périmètre d'étude : choix des unités médicales.

Une fois le périmètre d'étude est déterminé, nous présentons une modélisation de la DDS dans pour ce périmètre et nous soulignons les différents facteurs ou variables qui la définissent au moment de l'admission d'un patient.

3.2.2 Modélisation de la Durée De Séjour

La modélisation de la DDS pour un modèle de prédiction statique concerne l'étude des différents facteurs qui l'impactent au moment de l'admission du patient. A partir du modèle générique présenté dans la figure 2.3 et la définition de la DDS présentée dans la section précédente 3.2.1, nous avons identifié un modèle qui caractérise une DDS au moment de l'admission du patient. A la différence du modèle générique, ce modèle se base sur des données disponibles au moment de l'admission du patient. A partir du modèle générique de DDS de la figure 2.3, nous avons alors instancié uniquement les variables disponibles au moment de l'admission du patient. Pour mieux montré le lien entre le modèle générique de la DDS et celui au moment de l'admission du patient, nous avons repris le modèle de la figure 2.3 et éliminer les variables non instanciées (encadrées).

Ce modèle constitue une étude qualitative sur les variables qui peuvent participer dans la prédiction de DDS. Cette étude est suivie d'une autre étude quantitative afin de confirmer ou rejeter les hypothèses posées pour la conception de ce modèle. Le diagramme de classe de la figure 3.2 illustre ce modèle.

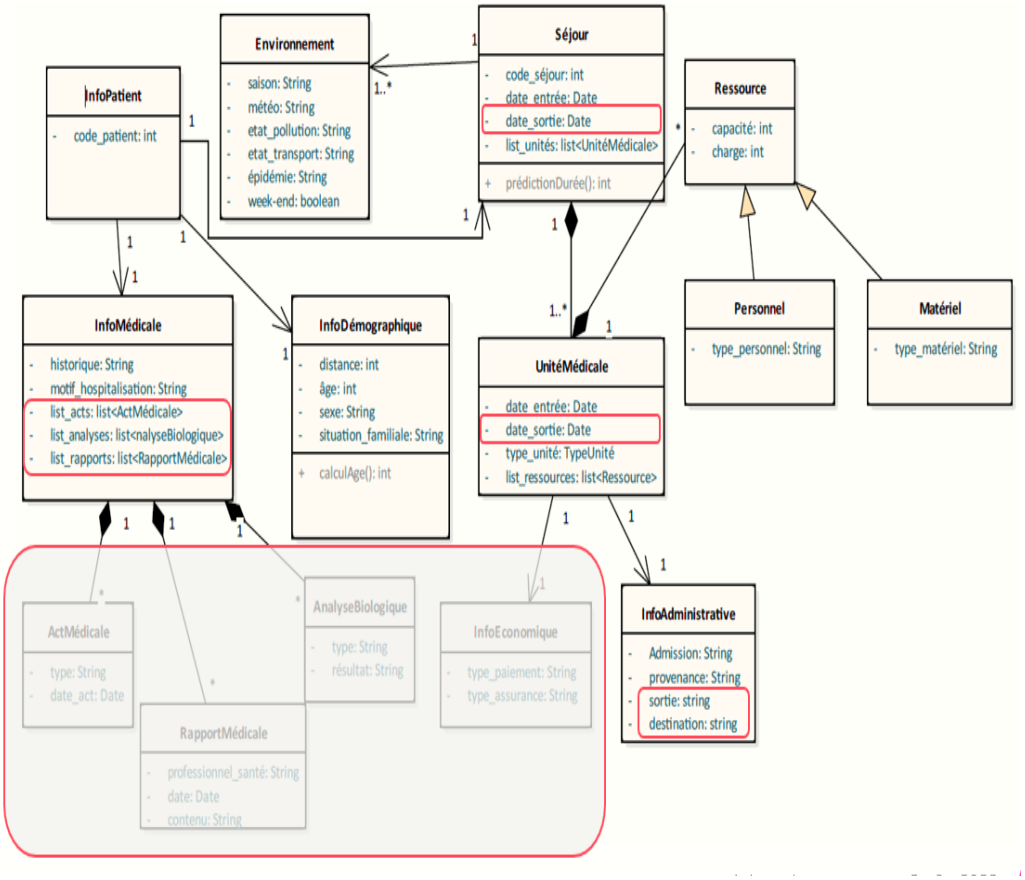


Fig. 3.2: Modélisation de la DDS au moment d'admission d'un patient.

Au moment d'admission du patient, et pour une seule unité médicale, les données disponibles englobent les données démographiques du patient : son nom, son prénom, son identifiant, son sexe, sa date de naissance, sa situation familiale et son adresse. Quoique ces données sont disponibles au moment de l'admission du patient, elles ne peuvent pas êtres utilisées dans le modèle de prédiction ou doivent être modifiées. Les informations considérées à caractères personnel sont protégées et restent confidentielles. Un identifiant anonyme est donné à chaque patient et les informations telle que le nom et le prénom su patient sont éliminées. La date de naissance est remplacée par l'âge et l'adresse par la distance entre le lieu de résidence et l'établissement de soins. En plus de ces informations, le motif d'hospitalisation et les antécédents médicaux figurent dans la description de l'état de santé du patient. Concernant les données administratives, nous retrouvons les conditions d'admission du patient, le type de l'unité médicale et le type de la structure de l'hôpital.

L'ensemble de ces données est l'entrée à notre processus de prédiction. Nous présentons en détails les étapes de ce processus dans ce qui suit.

3.2.3 Processus de prédiction

Le processus de prédiction que nous avons implémenté se base d'une part, sur des techniques de fouille de données pour la phase d'analyse et d'extraction des informations pertinentes. D'autre part, il se base sur les algorithmes d'apprentissage automatique pour la phase de prédiction.

La figure 3.3 montre les étapes essentielles d'un processus de prédiction en se basant sur des techniques d'apprentissage automatique.

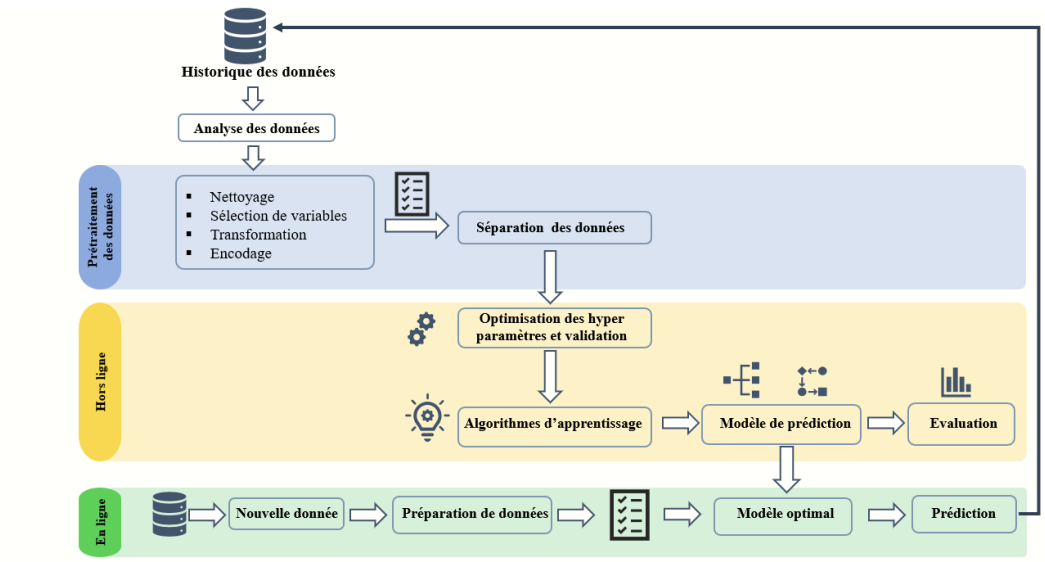


Fig. 3.3: Processus de prédiction de DDS avec les techniques d'apprentissage automatique.

Une première étape du processus de prédiction de DDS en se basant sur les techniques d'apprentissage automatique est la collecte des données. A partir des bases de données médicales, l'historique des données est récupéré. Cette étape est suivie par une analyse des données. Ensuite, une phase de pré-traitement de données est réalisée. Elle inclut le nettoyage des données, la sélection de variables, la transformation et l'encodage des données. L'ensemble de données résultat est séparé en 3 sous-ensembles : ensemble d'apprentissage, ensemble de validation et ensemble de test. L'ensemble d'apprentissage sert à l'apprentissage du modèle, l'ensemble de validation à la validation des résultats et l'ensemble de test pour l'évaluation des résultats obtenus. Enfin, nous distinguons deux parties : *hors ligne* et *en ligne*. La partie *hors ligne* sert à l'optimisation des hyper-paramètres des algorithmes d'apprentissage automatique, à la validation des résultats et à l'apprentissage du modèle. Le résultat de la partie *hors ligne* est un modèle de prédiction qui sera évalué sur certains critères jusqu'à obtention d'un modèle de prédiction optimal. Ce modèle de prédiction optimal est utilisé dans la partie *en ligne* dans la prédiction de nouvelles

instances de données en temps réel. Le résultat de prédiction est sauvegardé dans la base de données après validation.

Chaque étape du processus de prédiction de DDS à l'aide des algorithmes d'apprentissage automatique est décrite dans ce qui suit en mettant en évidence les difficultés rencontrées en utilisant les données réelles issues du PMSI.

3.3 Apprentissage automatique dans la prédiction de Durées De Séjour

L'entrée clef d'un algorithme d'apprentissage automatique est la donnée. Cependant, la phase de préparation de données est importante car elle détermine les performances des algorithmes d'apprentissage automatique. D'abord, nous présentons les méthodes que nous avons implémenté lors des pré-traitements des données. Ensuite, pour la phase de prédiction, différents algorithmes de prédiction sont explorés.

3.3.1 Collecte et analyse des données

Les données de santé proviennent de multiple sources. Elles sont donc hétérogènes et très volumineuses. La phase de collecte de données a pour but de recueillir toutes les informations nécessaires à la conception du modèle de prédiction. Ces informations sont présentes dans les Systèmes d'informations Hospitalier en général. En particulier, les données auxquelles nous nous intéressons sont sauvegardées dans le PMSI. Les bases de données relationnelles sont utilisées pour stocker ces données. Elles se composent de nombreuses tables inter-connectées par des liens. Ces liens permettent de retrouver les informations entre les différentes tables et de construire une instance représentant un séjour hospitalier.

Une instance est caractérisée par un ensemble de M variables indépendantes (X_1, X_2, ..., X_M) et une variable dépendante à prédire Y. La valeur du domaine de chaque variable nous permet de déterminer leur propre type. Lorsque cette valeur appartient au sous-ensemble des nombres réels, elle est de type numérique. Tandis que, lorsque cette valeur appartient à un sous-ensemble fini, elle est de type catégorielle. Les valeurs d'une variable catégorielle sont nommées modalités. Il arrive qu'une variable catégorielle contient plus d'une seule modalité. De ce fait, c'est une variable catégorielle multivaluée. La variable représentant *l'historique médical du patient* est un exemple de variable catégorielle multivaluée dans le cas où le patient possède plusieurs antécédents médicaux.

Dans cette étude, nous avons collecté les données issues du PMSI. La figure 3.4 décrit le contenu des données PMSI utilisées.

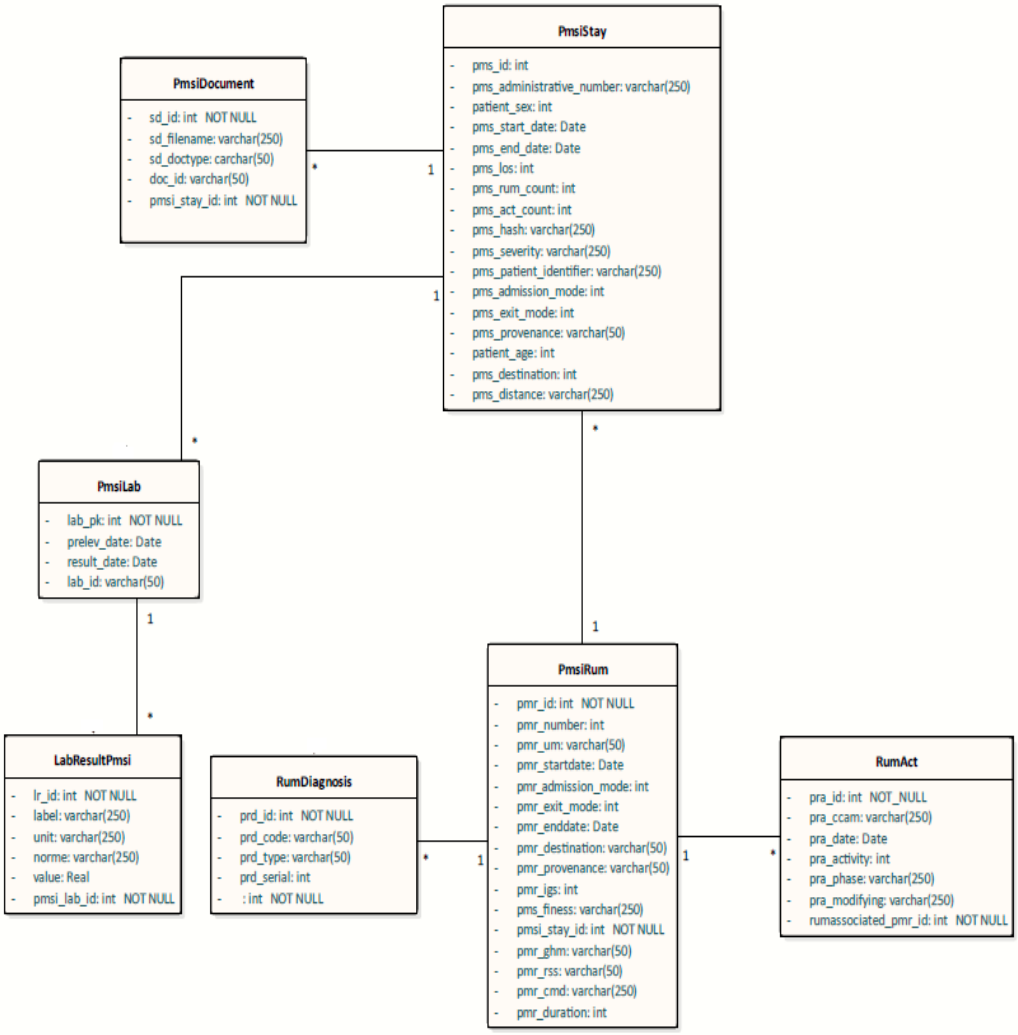


Fig. 3.4: Schéma relationnel des données du PMSI.

Les informations concernant chaque unité médicale sont stockées dans la table $PmsiRum$. Celle-ci est liée à la table $PmsiAct$ qui contient les actes médicaux réalisés et à la table $PmsiDiagnosis$ qui comporte les différents diagnostics du patient. Un résumé du séjour global du patient contenant les informations de toutes les unités médicales dans lesquelles le patient a été admis est sauvegardé dans la table $PmsiStay$. Le reste des tables contiennent les données des analyses du laboratoires et les rapports médicaux.

En plus de ces données, nous avons déduit les informations socio-clinique citées dans le chapitre précédent comme par exemple : la charge du travail au moment de l'admission du patient, la moyenne de la DDS par unité médicale ou la moyenne de

la DDS par diagnostic principal. En outre, pour analyser l'impact du jour d'admission sur la DDS, nous avons distingué les jours de la semaine de ceux du week-end.

L'ensemble des données est transformé de la forme relationnelle à la forme tabulaire avec : l'ensemble des lignes représentent les séjours hospitaliers dans différents unités médicales et l'ensemble des colonnes représentent les variables qui la caractérisent.

L'analyse de données consiste à représenter graphiquement les variables de l'ensemble des données ou à utiliser des mesures statistiques pour les décrire. Cette analyse permet d'extraire des informations nécessaires et décrire les variables par leur type et leur contenu.

Une analyse uni-variée des différentes variables est d'abord réalisée. Nous avons tracé les histogramme, la distribution et les boites à moustaches des données numériques. Pour les valeurs catégorielles, nous avons tracé leur fréquence et avons déduit leur modalités. Ensuite, une analyse bi-variée est effectuée pour extraire les relations existantes entre les différentes variables. Cette analyse aide à mieux comprendre quelles méthodes appliquer dans la prochaine étape qui est le pré-traitement de données.

3.3.2 Pré-traitements des données

Le pré-traitement des données est un mécanisme de préparation des données brutes pour les adapter à un modèle d'apprentissage automatique. C'est une étape cruciale d'un processus d'apprentissage automatique. La qualité des données et l'information utile qui peut en être dérivée ont une incidence directe sur la capacité d'apprentissage du modèle. Les ensembles de données recueillies de la vie quotidienne et particulièrement celles du domaine médical contiennent énormément de bruit et leur structure est souvent inadéquate pour les algorithmes d'apprentissage automatique. Le pré-traitement de données englobe le nettoyage de données, la sélection de variables, la transformation, la normalisation et la codification des données. La qualité des données résultantes de cette phase de pré-traitement participe largement aux performances des algorithmes employés par la suite. L'ensemble de données initialement utilisées dans cette étude est constitué de l'ensemble des variables suivantes : *l'identifiant du patient*, *l'identifiant du séjour*, *genre et âge du patient*, *la distance entre le domicile du patient et l'établissement de soins*, *les code CIM-10 et leur type*, *type de l'unité médicale*, *mode d'admission*, *mode de provenance* et la *DDS*.

3.3.2.1 Nettoyage de données

Les données médicales présentent énormément de bruit car elles proviennent de plusieurs sources. Une première étape à mener est de nettoyer les données brutes avant de procéder à d'autres traitements. Le nettoyage des données concerne la suppression ou la correction des erreurs syntaxiques ou sémantiques. Une erreur syntaxique est souvent due aux erreurs de sauvegarde comme par exemple les erreurs d'orthographes. Quand aux erreurs sémantiques représentent une valeur non définie pour une variable. L'élimination des caractères spéciaux et les espaces blancs inutiles font partie également du nettoyage de données.

Cette phase inclut également le traitement des données aberrantes et des données manquantes. Les données aberrantes sont les valeurs extrêmes qui sont distantes des autres valeurs. Par exemple, en service de pédiatrie, la valeur de l'âge du patient ne peut dépasser 16 ans. Les données manquantes représentent des erreurs structurelles telles que les cases vides. Nous avons proposé une solution pour traiter les données aberrantes et les données manquantes issues du PMSI.

Traitement des données aberrantes

Les données médicales sont recueillies de plusieurs façons : les questionnaires, les enquêtes, les rapports et les dossiers médicaux. Des erreurs d'acquisition, de communication et d'orthographes peuvent apparaître. Dans le domaine médical particulièrement, ces valeurs peuvent représenter des profils rares. Ces profils dits atypiques se produisent exceptionnellement et ne reflètent pas souvent la réalité. Il est nécessaire alors de distinguer entre ces profils rares et les données aberrantes. A partir de l'analyse des données, nous avons déterminé les données aberrantes. Pour mieux analyser la distribution des données, le coefficient d'asymétrie de $Fisher$ et $Pearson$ [KZ00] est calculé. A partir de cela, nous avons déduit si la variable est relativement, moyennent ou fortement asymétrique. Une distribution est dite asymétrique si les valeurs observées se répartissent de façon non uniforme autour des trois valeurs centrales : la moyenne, le mode et la médiane [Com21].

Traitement des données manquantes

Les données du monde réel comportent souvent des valeurs manquantes. Elles sont causées par la corruption des données et l'échec de leur enregistrement. Le traitement des données manquantes est très important car de nombreux algorithmes d'apprentissage automatique ne les supportent pas. Les valeurs manquantes peuvent être traitées en supprimant les lignes ou les colonnes ayant ces valeurs. Elles peuvent

être remplacées par la moyenne ou la médiane pour les variables numériques et par le mode pour les variables catégorielles.

3.3.2.2 Sélection de variables

La modélisation de la Durée De Séjour a dévoilé l'ensemble des facteurs qui peuvent impacter la DDS dans un environnement hospitalier. La modélisation de la DDS est une étude qualitative qui s'appuie sur des travaux antérieurs dans le domaine, la définition des besoins des hôpitaux et l'expertise médicale. Une étude quantitative complémentaire en se basant sur des données stockées les SIH est réalisée.

La sélection de variables permet de réduire l'ensemble des variables et de réduire la complexité temporelle et spatiale des algorithmes d'apprentissage automatique. Elle permet aussi d'améliorer les performances de ces algorithmes en éliminant les corrélations entre les variables indépendantes entre elles et garder que les variables corrélées à la DDS. La performance des algorithmes d'apprentissage automatique est grandement déterminée par leur entrée et peut être pénalisée si cette entrée représente mal la variable à prédire. Cette étape permet d'éviter le problème du sur-apprentissage des modèles, de réduire leur temps de construction et d'améliorer leur capacité de généralisation [CS14 ; GE03]. Les approches de sélection de variables peuvent êtres classées en 3 sous-catégories qui sont les suivantes :

- *Approches filtre* : Les approches filtres sont généralement utilisées comme une étape de pré-traitement de données. La sélection des variables dans ce cas est indépendante de l'algorithme d'apprentissage. La corrélation entre les variables est calculée en se basant sur un score statistique [Bom+20].
- *Approches wrapper* : la sélection de variable se base sur un algorithme d'apprentissage automatique spécifique que nous appliquons sur l'ensemble des données initial. Elle effectue une recherche exhaustive en essayant toutes les combinaisons possibles des variables et en évaluant la performance de l'algorithme utilisé selon un critère défini [Ana20].
- *Approches hybrides* : Les approches hybrides combinent les qualités des deux approches filtres et wrapper. Elles sont utilisées par les algorithmes qui incluent la sélection de variables dans leur implémentation [Ana20].

Dans notre étude, nous avons adopté les approches filtres pour sélectionner les variables en corrélations avec la variable cible DDS et pour éliminer les variables indépendantes corrélées entre elles. Pour cela, nous avons utilisé les tests statistiques en particulier *le coefficient de Spearman*, *l'indicateur de Theil* et *l'information mutuelle*.

Le coefficient de Spearman est utilisé pour mesurer la relation monotone entre les variables numériques. Il est utilisé dans le cas où les variables ne suivent pas une loi normale [Ako18]. Ce problème de distribution non normale des données numériques dans le domaine médical est omniprésent. L'indicateur de Theil également appelé coefficient d'incertitude ou coefficient d'entropie mesure la dépendance entre deux variables catégorielles distinctes [Sha48]. Les approches basées sur l'information mutuelle mesure toute dépendance arbitraire entre deux variables aléatoires X et Y. Elle calcule la quantité d'information de X fournie par Y (et inversement de Y fournie par X) [Ami+11].

Une fois le sous-ensemble de données est déterminé, nous passons à la transformation des données numériques et à l'encodage des données catégorielles.

3.3.2.3 Transformation et encodage des données

Pour améliorer les performances des algorithmes d'apprentissage automatique, une transformation et une standardisation des variables numériques sont réalisées. Les variables catégorielles doivent êtres codifiées en valeurs numériques pour qu'elles soient supportées par les algorithmes d'apprentissage automatique. Ainsi, dans cette phase de transformation, de standardisation et d'encodage des données nous avons procéder comme suit :

- *Variables numériques* : La distribution symétrique décrivant une loi normale n'est pas présente dans des données médicales dans la réalité du terrain. Une des techniques employées pour rendre la distribution des données numériques aussi normale que possible est la transformation logarithmique [Mek+21]. C'est la transformation la plus populaire parmi les autres types de transformations [Fen+14]. Elle consiste à transformer la valeur d'une variable X en *log(X)*. Dans notre étude, nous avons transformer les données extrêmement bruitées en *log(x+1)* pour soulever le problème des valeurs indéfinies. La formule suivante représente la transformation logarithmique.

$$X = log(X + 1) \tag{3.1}$$

  Comme les variables numériques sont souvent mesurées avec des unités différentes, il est nécessaire de standardiser ces variables. La standardisation a pour but de remettre toutes les variables sur la même échelle et avoir une représentation structurée cible. Pour chaque valeur, la fonction z-score est calculée et qui est couramment utilisée. Elle consiste à soustraire la moyenne

de la variable et la diviser par son écart-type [Los09]. La fonction Z-score est donnée comme suit [Mek+20a] :

$$X_i = \frac{X_i - \mu}{\sigma} \tag{3.2}$$

Avec X_i représente une instance de donnée, μ est la moyenne de la variable X et σ est son écart type.

- Variables catégorielles : Plusieurs techniques d'encodage des variables catégorielles existent dont l'encodage séquentiel, l'encodage par étiquettes et l'encodage « *One-hot-enconding* ». Dans notre étude, la technique « *One-hot-enconding* » est appliquée. Cette technique est la mieux adaptée à l'encodage des données catégorielles et elle est largement utilisée pour ignorer la relation d'ordre qui est naturellement présente entre les valeurs entières [Bro17]. Contrairement à l'encodage séquentiel et l'encodage par étiquettes, la relation d'ordre est prise en compte dans les algorithmes d'apprentissage automatique. D'abord, les modalités des variables catégorielles sont transformées en valeurs entières. Ensuite, ces valeurs entières sont converties en valeurs binaires. Le tableau 3.1 montre l'application de cette technique sur la variable $rcount$ qui représente le nombre d'admission du patient au cours des 180 jours précédents. Cette variable comporte 6 modalités : 0, 1, 2, 3, 4, 5+.

rcount	0	1	2	3	4	5+
rcount 0 : 0 readmissions	1	0	0	0	0	0
rcount 1 : 1 readmissions	0	1	0	0	0	0
rcount 2 : 2 readmissions	0	0	1	0	0	0
rcount 3 : 3 readmissions	0	0	0	1	0	0
rcount 4 : 4 readmissions	0	0	0	0	1	0
rcount 5+ : 5 or more than 5 readmissions	0	0	0	0	0	1

Tab. 3.1: Application de la technique « *One-hot-enconding* » sur la variable *rcount*

3.3.3 Algorithmes d'apprentissage

Le problème de prédiction de DDS est défini comme une estimation d'une valeur à partir d'un ensemble de variables. Les algorithmes d'apprentissage automatique supervisé sont utilisés. L'ensemble de données est divisé en ensembles d'apprentissage, de validation et de test. Les ensembles d'apprentissage et de validation comportent tout les deux les différentes caractéristiques du séjour hospitalier ainsi que la variable cible qui est la DDS. Ils sont utilisés dans la partie *hors ligne* pour l'entraînement du modèle et sa validation. Contrairement à l'ensemble de test qui contient uniquement les différentes caractéristiques du séjour et est utilisé dans la partie *en ligne* qui est l'évaluation des modèles d'apprentissage.

La sélection des algorithmes d'apprentissage automatique s'appuie sur le type, la qualité, la quantité des données et sur l'objectif visé. Dans le domaine médical en général et dans la prédiction des DDS en particulier, il est important d'interpréter les résultats obtenus en lien avec l'activité médicale. Nous avons choisi les algorithmes basés sur les arbres de décision car ils sont très intuitifs et facile à expliquer aux équipes techniques et aux parties prenantes. De plus, ils capturent les relations non linéraires entre les données. Ils sont aussi efficaces face à la complexité des données médicales. Il a été montré dans de nombreuses études antérieures que la fusion de plusieurs algorithmes d'apprentissage abouti à de meilleurs résultats [Mek+21]. Les algorithmes des méthodes ensemblistes sont alors utilisés. Le Random Forest (RF), le Gradient Boosting Model (GBM) et le Extreme Gradient Boosting (XGboost) sont comparés.

Pour chaque algorithme, une étape d'ajustement des hyper-paramètres est menée. La validation des résultats obtenus est également mise en place. Nous détaillons les méthodes utilisées dans ce qui suit.

3.3.4 Optimisation des hyper-paramètres et validation

Un hyper-paramètre est un paramètre dont la valeur est définie avant le début du processus d'apprentissage. Il ne fait pas partie des paramètres de l'algorithme et il ne peut pas être directement déduit à partir des données. L'optimisation de ces hyper-paramètres est souvent appelée « la recherche de l'espace optimal des valeurs des hyper-paramètres ». L'espace optimal de ces valeurs représente la combinaison qui reproduit les meilleurs performances de l'algorithme d'apprentissage automatique et assure leur efficacité. En général, cette phase inclut les tâches suivantes :

- Définition de l'algorithme d'apprentissage et de la plage des valeurs de chaque hyper-paramètre.
- Définition de la méthode d'échantillonnage des valeurs des hyper-paramètres.
- Définition d'un critère d'évaluation du modèle d'apprentissage.
- Définition d'une méthode d'optimisation de l'architecture du modèle (différentes valeurs des hyper-paramètres).

Il existe plusieurs méthodes d'optimisation des hyper-paramètres. Nous avons implémenté les méthodes d'approche bayésienne (Sequential Model Based Optimization ou SMBO en anglais). Dans l'approche bayésienne, on capitalise des résultats obtenus précédemment pour chercher le prochain jeux d'hyper-paramètres qui représente la nouvelle configuration du modèle. Un modèle probabiliste est créé en liant les

fonctions qui symbolisent les valeurs des hyper-paramètres à la fonction de perte à évaluer sur un ensemble de données de validation [Ber+11]. L'évaluation et la recherche de l'optimum se fait itérativement en équilibrant entre l'exploration des hyper-paramètres pour lesquels le résultat est le plus incertain et l'exploitation des hyper-paramètres attendus proches de l'optimum jusqu'à convergence. A chaque itération est calculé donc un score et la configuration du modèle est mise à jour [HHL11]. Le pseudo code de cet algorithme est décrit dans la figure 3.5.

SMBO(f, M_0, T, S)

1 $\quad \mathcal{H} \leftarrow \emptyset$,
2 $\quad$ For $t \leftarrow 1$ **to** T,
3 $\quad\quad x^* \leftarrow \operatorname{argmin}_x S(x, M_{t-1})$,
4 $\quad\quad$ Evaluate $f(x^*)$, $\quad \triangleright$ *Expensive step*
5 $\quad\quad \mathcal{H} \leftarrow \mathcal{H} \cup (x^*, f(x^*))$,
6 $\quad\quad$ Fit a new model M_t to $\mathcal{H}$.
7 $\quad$ **return** $\mathcal{H}$

Fig. 3.5: Pseudo-code de l'algorithme Sequential Model Based Optimization (SMBO) [HHL11].

Les paramètres de cet algorithme sont décrit comme suit :

- H : observation postérieure.
- T : le nombre d'essais.
- f : fonction de perte.
- M : fonction d'acquisition à optimiser avec le processus gaussien.
- S : nouvelle configuration d'hyper-paramètres.
- x^* : l'instance de données où M est égal au minimum.

D'autres méthodes d'optimisation des hyper-paramètres existent. Nous citons la recherche par grille et la recherche aléatoire [Cap19]. Dans la pratique, il a été démontré que l'optimisation bayésienne permet d'obtenir de meilleurs résultats avec moins d'évaluations par rapport à la recherche de grille et à la recherche aléatoire, en raison de la capacité à raisonner sur la qualité des expériences avant leur exécution [Lav20].

En optimisant les hyper-paramètres des algorithmes d'apprentissage automatique, il est essentiel d'évaluer la capacité de généralisation de ces algorithmes sur les nouvelles données. Si aucune technique de généralisation n'est utilisée, la capacité d'évaluer réellement le modèle et son fonctionnement sur de nouvelles données est perdue. Pour atténuer cette perte, l'ensemble de données est divisé en trois sous-ensembles : données d'apprentissage, données de validation et données de

test. L'introduction d'un ensemble de validation permet d'évaluer le modèle sur des données différentes de celles sur lesquelles le modèle d'apprentissage a été formé. La méthode de validation croisée est employée. Une des techniques qui applique la validation croisée est la technique K-Folds [Kum21]. Elle est simple à comprendre, et particulièrement populaire. L'ensemble de données est séparé d'une manière aléatoire en K sous-ensembles. En général, la valeur de K choisie est comprise entre 3 et 10. Elle permet d'utiliser l'intégralité du jeu de données pour l'entraînement et pour la validation en alternant entre les différents sous-ensembles [RWM19].

L'utilisation d'un ensemble de validation prévient le modèle du problème de sur-apprentissage (over-fitting en anglais) [Yin19] ou celui du sous-apprentissage (under-fitting en anglais) [Bas+20]. Le sur-apprentissage se pose quand le modèle de prédiction ajuste bien les données d'apprentissage, mais il ne parvient pas à généraliser aux données non rencontrées pendant l'apprentissage. Ainsi les résultats de prédiction sont très mauvais. Le sous-apprentissage est présent quand le modèle n'arrive pas à bien décrire les les données d'apprentissage.

3.4 Classification des Durée De Séjour

Dans cette section nous présentons une solution de prédiction de Durée De Séjour hospitalier. La DDS est dans ce cas représentée par 3 catégories : DDS courte, DDS moyenne et DDS longue. La classification est donc utilisée.

3.4.1 Définition du problème

Dans le PMSI, l'utilisation des Groupes Homogènes de Malades (GHM) pour le classement des patients a été introduite pour la mesure des points d'Indice Synthétique d'Activité (ISA). Les points ISA permettent de déduire la tarification appliquée dans un établissement de soins. Les groupes au sein des GHM sont hétérogènes du fait des différences des profils des patients, de la disparité des pratiques des médecins et de la différence de la structuration et l'organisation des établissements de soins. Cette hétérogénéité induit à différentes distributions de DDS et à différents coûts. En effet, comme les points ISA dépendent uniquement des classements des GHM, il arrive que ces points sont identiques même si la DDS du patient est égale à 2 ou 15 jours. Il est donc nécessaire de repérer les séjours où les patients restent plus longtemps ou moins longtemps que la normale. Cependant, ça permet de déduire les coûts des points ISA et ainsi augmenter l'allocation budgétaire de l'établissement de soins.

L'objectif de la classification des DDS est d'attribuer une valeur catégorielle à une instance. L'instance représente un ensemble de caractéristiques d'un séjour hospitalier dans une unité médicale. Ces caractéristiques regroupent des informations médicales du patient, des informations démographiques et des informations administratives. Une première étape de la classification est de définir les différentes catégories des DDS.

3.4.2 Stratégie de modélisation

La définition de la DDS comme des catégories a été abordée dans la littérature de diverse façons. La procédure de transformation de la DDS d'une valeur numérique à une valeur catégorielle est appelée discrétisation.

Notre stratégie de modélisation de la DDS par des catégories s'appuie sur la distribution de cette variable en analysant les mesures de sa moyenne, son minimum, son maximum, son premier quartile et la médiane. De plus, cette stratégie de modélisation s'appuie sur la fréquence des valeurs de la DDS dans l'ensemble de données utilisées. Le but est de construire des catégories plus au moins équilibrées tout en répondant aux besoins organisationnels de l'établissement de soins. La moyenne de la DDS est égale à 5,84 jours. Les valeurs du minimum, maximum et le premier quartile sont égales respectivement à 0 jours, 28 jours et 3 jours. La médiane est équivalente à 4 jours. Nous avons donc déterminé les catégories de la DDS comme suit :

- DDS courte : de 0 à 2 jours.
- DDS moyenne : de 3 à 5 jours.
- DDS longue : de 6 à 28 jours.

La modélisation nous a permis de définir les catégories de la DDS. La séparation des valeurs des DDS en plusieurs catégories offre la possibilité de distinguer entre les DDS au-dessus et celles en dessous de la moyenne. Le problème de classification concerne alors la prédiction d'une des catégories à partir des caractéristiques utilisées en entrée.

3.4.3 Méthodes proposées pour la classification

Pour la classification des DDS, nous avons proposé deux méthodes. La première méthode repose sur des techniques d'apprentissage supervisé et la deuxième méthode combine les techniques d'apprentissage supervisé et d'apprentissage non supervisé.

3.4.3.1 Classification à base d'apprentissage supervisé

La classification est une technique qui prend une place importante dans l'apprentissage automatique. Dans cette étude, les méthodes ensemblistes s'appuyant sur les arbres de décision sont utilisées. Les deux méthodes de bagging et boosting décrites dans le chapitre précédent sont comparées. La méthode du bagging est implémentée avec l'algorithme Random Forest (RF). Le boosting est implémenté avec les algorithmes Gradient Boosting Model (GBM) et Extreme Gradient Boosting (Xgboost). Nous avons suivi les étapes décrites dans la section 3.2.3 pour la mise en place de notre solution. L'ensemble de variables qui caractérisent la DDS sont utilisées comme descripteurs de notre variable cible DDS. Dans nos expérimentations, nous avons utilisé la mesure de l'entropie croisée comme fonction de perte à minimiser. Les mesures de la précision globale, la précision, le rappel et le score F_1 sont utilisés pour l'évaluation et la comparaison entre les performances des modèles de classification.

3.4.3.2 Classification à base d'apprentissage supervisé et non supervisé

Plusieurs travaux antérieurs ont montré que l'usage des techniques de l'apprentissage non supervisé ou le clustering peut améliorer la précision des modèles de prédiction [PR14 ; AJM12] et augmenter leur performance [YHK20]. L'un des critères couramment utilisés pour ces regroupements dans les soins de santé consiste à regrouper les patients en fonction des similitudes dans leur dossier médical [Xin+07]. Le clustering peut réduire la complexité du modèle de prédiction en formant des groupes homogènes d'instances. Par conséquent, afin d'évaluer notre modèle de prédiction de façon plus complète, nous avons combiné les deux techniques d'apprentissage supervisé et non supervisé dans la démarche de prédiction. Les séjours hospitaliers représentés par les variables sélectionnées sont d'abord partitionnés puis le résultat du clustering est intégré dans l'algorithme de prédiction. Une information supplémentaire est donc ajoutée à l'ensemble de données en entrée de l'algorithme de prédiction. Cette information représente la catégorie retournée par l'algorithme du clustering.

Il existe plusieurs méthodes de clustering : les méthodes hiérarchiques, les méthodes de partitionnement et les méthodes à base de modèles [Naj17]. Les algorithmes de clustering hiérarchiques ne sont pas adaptés aux ensembles de données volumineux regardant leur complexité temporelle et computationnelle. Dans le cas où les données sont larges, les algorithmes à base de partitionnement sont les mieux adaptés. Ces algorithmes se basent sur le calcul d'une mesure de similarité dans leur réalisation. En considérant les différents types de données présents dans les données du PMSI, l'algorithme K-prototype est utilisé dans cette thèse. Cet algorithme sup-

porte les données numériques, catégorielles et les données catégorielles multivaluées [Lia+12].

Une étape fondamentale pour tout algorithme de partitionnement est de déterminer le nombre optimal de groupes dans lesquels les données seront partitionnées. La méthode du coude est l'une des méthodes les plus populaires pour déterminer cette valeur optimale de k. Les étapes de cette méthode s'articulent autour des points suivants [Hab21] :

- La réalisation de l'algorithme K-prototype en variant la valeur du paramètre K.
- La mesure de la variance intra-classes pour chaque groupe.
- Visualisation du graphe de chaque valeur de cette variance en fonction du paramètre K.
- Une position du coude apparaît sur le graphe et sa projection sur l'axe des X représente la valeur optimal du paramètre K.

La figure 3.6 illustre la position du coude et la déduction de la valeur optimal du paramètre K.

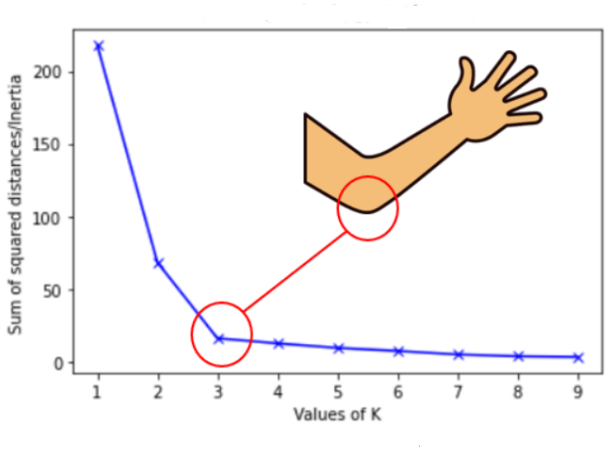


Fig. 3.6: Application de la méthode du coude sur l'algorithme K-prototype [Ban21]

.

3.5 Prédiction des DDS : régression

Dans la section précédente, nous avons abordé la prédiction des DDS comme un problème d'affectation d'une catégorie à une instance de variables. Dans ce qui suit nous définissions la DDS comme le nombre de jours passés dans un service hospitalier par le patient.

3.5.1 Définition du problème

L'un des objectifs de la prédiction de la Durée De Séjour est l'optimisation des ressources matérielles et humaines de l'établissement de soins. La prédiction de DDS contribue à la gestion logistique des lits, à la prévision des flux des patients et à la planification des activités des services hospitaliers. Une prédiction la plus exacte au moment de l'admission du patient de la valeur de la DDS parait plus bénéfique afin d'atteindre ces objectifs. Nous avons donc défini la DSS comme le nombre de jours passés dans une unité médicale. La DDS est représentée par une valeur entière. Des recherches récentes [Mek+20a ; Mek+21] ont exploré les méthodes de régression issues de l'apprentissage automatique supervisé afin de construire des modèles de prédiction pour cette tâche. Comme pour la tâche de classification, ces modèles utilisent également des données historiques dans l'apprentissage du modèle, sa validation et son évaluation. La démarche à suivre afin de réaliser le modèle de prédiction à base de régression est identique à celle de la classification. Contrairement aux problèmes de classification, les méthodes de régression tentent de prédire une valeur numérique. De ce fait, nous cherchons à minimiser un taux d'erreur qui représente l'écart entre la valeur prédite et la valeur réelle.

3.5.2 Stratégie de modélisation

La Durée De Séjour est représentée par le nombre de nuits passées dans une unité médicale précise. Elle est donc l'intervalle de temps la date d'admission du patient et sa date de sortie. Nous exposons les résultats des modèles en jours. Les valeurs prédites sont arrondies. L'ensemble des facteurs qui impactent la DDS et le périmètre d'étude pour la tâche de régression restent les mêmes que celle pour la tâche de classification. L'objectif est d'estimer une valeur de la DDS qui se rapproche le plus de sa valeur réelle afin de mieux gérer l'organisation de l'hôpital. Nous avons proposé deux méthodes de prédiction de DDS dans le cadre de la régression que nous allons présenter dans ce qui suit.

3.5.3 Méthodes proposées pour la régression

Suivant la nature, la qualité des données et l'objectif de l'étude, nous avons défini les différents algorithmes à implémenter, la fonction de perte à optimiser et les mesures d'évaluation de nos modèles d'apprentissage. Nous avons suivi le même processus de prédiction que celui pour la tâche de classification en utilisant le même ensemble de variables pour décrire la variable cible DDS. Nous avons comparé les algorithmes suivants : Random Forest (RF), Gradient Boosting Model (GBM) et Extreme Gradient

Boosting (Xgboost). De plus, nous avons proposé une pondération de la fonction de perte pour l'algorithme Xgboost.

Le processus de prédiction décrit auparavant est mis en oeuvre en adaptant la définition de la DDS et les algorithmes d'apprentissage automatique au contexte de la régression. Nous allons détailler nos solutions dans ce qui suit.

3.5.3.1 Minimisation de la MSE

Tout algorithme d'apprentissage tente d'optimiser une fonction de perte et utilise une fonction d'évaluation afin d'estimer la qualité de la prédiction. Nous présentons ces deux fonctions pour la première approche.

- Fonction de perte : dans le cadre de la régression, nous recherchons à minimiser un écart entre la valeur prédite de la DDS et sa valeur réelle. La moyenne des erreurs quadratiques (MSE) est la fonction la plus utilisée. La MSE est sensible envers les données aberrantes donc elle est souvent employée lorsque il est important de pénaliser ces données aberrantes. Le carré est nécessaire pour éliminer tous les signes négatifs. Il donne également plus de poids aux grands écarts. La MSE est utilisée en supposons que notre variable dépendante DDS conditionnée par l'ensemble des variables indépendantes, est normalement distribuée.

- Fonction d'évaluation : en estimant une durée, on cherche à minimiser son écart en valeur absolue de sa valeur réelle. La moyenne des écarts absolus (MAE) est la fonction d'évaluation adaptée dans ce contexte. En effet, nous cherchons à comprendre par exemple un écart de plus au moins la valeur réelle de la valeur prédite.

L'étude de la distribution de la variable DDS a révélé que cette dernière est asymétrique gauche. La distribution asymétrique est une situation dans laquelle les valeurs des variables se produisent à des fréquences irrégulières. En plus, la moyenne, la médiane et le mode se produisent à différents points. En effet, les données réelles dans de nombreux domaines d'application et particulièrement les données du domaine médical présentent une forte asymétrie et des fréquences très éloignées. Nous proposons une pondération de la fonction MSE utilisée dans l'algorithme Xgboost comme fonction de perte pour palier à ce problème.

3.5.3.2 Minimisation de la MSE pondérée

Un problème connu en apprentissage automatique est celui de la prédiction des valeurs rares. Il est très courant d'avoir des fréquences déséquilibrées dans les valeurs de la variable cible à prédire. Les algorithmes d'apprentissage automatique dans ce cas n'arrivent pas à bien prédire ces valeurs rares. Plusieurs solutions existent pour palier à ce problème notamment la transformation de la variable cible, l'application des méthodes de sur-échantillonage ou sous-échantillonage de l'ensemble de données et la création d'une nouvelle fonction de perte à optimiser. Dans ce travail de thèse, nous avons modifié la fonction de perte pour pondérer les valeurs de la DDS en se basant sur leur fréquences d'apparition dans la base de données utilisée dans l'étude.

Nous avons proposé une pondération de la fonction MSE décrite dans la section précédente. La nouvelle formule de la fonction de perte devient comme suit :

$$\textit{MSE pondérée} = \frac{1}{N} * \sum_{i=1}^{n} [(\frac{1}{[\frac{freq(y_i)}{N}]}) * (\hat{y} - y_i)^2] \tag{3.3}$$

Le N représente le nombre d'instances dans l'ensemble de données. La valeur de freq(y_i) représente la fréquence d'une valeur y_i. La valeur de la fréquence(y_i) / N représente la proportion d'une valeur de la DDS que nous notons D. Si D est grand alors $1/D$ est petit. Par conséquent, si la fréquence d'une valeur de la DDS est grande, un poids plus petit lui sera affecté. Ceci force l'algorithme à prédire les valeurs rares d'une manière plus robuste.

3.6 Conclusion

Dans ce chapitre nous avons proposé des méthodes de prédiction de DDS au moment de l'admission du patient. Les méthodes qui se basent principalement sur les techniques de fouille de données et d'apprentissage automatique sont explorées. D'abord, nous nous sommes intéressés à un périmètre d'étude bien précis afin de cerner la zone de l'hôpital dans laquelle le séjour se déroule. Ensuite, un modèle caractérisant la DDS au moment de l'admission du patient est exposé.

Nous avons proposé deux solutions pour le modèle statique de prédiction de DDS : la classification et la régression. Pour chacune des solutions, deux méthodes sont explorées. Dans le cadre de la classification, nous avons d'abord utilisé les algorithmes d'apprentissage supervisé. Puis, nous avons combiné les techniques d'apprentissage

supervisé et non supervisé. D'autre part, dans le cadre de la régression, nous avons utilisé les algorithmes d'apprentissage supervisé et avons proposé une pondération de la fonction de perte de l'algorithme Xgboost.

Nous avons nommé ce modèle comme modèle statique de prédiction de DDS car il considère des données disponibles au début du séjour hospitalier. Souvent les données médicales arrivent au cours du séjour hospitalier. Pour inclure de nouvelles données qui arrivent après l'admission du patient, nous proposons un nouveau modèle séquentiel de prédiction de DDS que nous allons présenter dans le chapitre suivant.

4

Modèle séquentiel de prédiction des Durées De Séjour hospitalier

> *Research is to see what everybody else has seen, and to think what nobody else has thought*
>
> — **Albert szent-Györgyi**
> Prix Nobel de physiologie ou médecine

4.1 Introduction

La prédiction de Durée De Séjour en milieu hospitalier a été étudiée sans prendre en compte l'évolution des données nécessaires à sa définition. Cependant, en situation réelle, les données utilisées pour prédire la DDS ne sont disponibles qu'au fur et à mesure que le patient progresse dans son parcours. La Durée De Séjour devait donc évoluer pour tenir compte de la disponibilité progressive des données médicales.

Dans ce chapitre, nous allons présenter une solution pour le calcul de la DDS et capable de tenir compte de l'évolution des données médicales. Nous commençons par définir le contexte, la problématique et l'objectif que nous visons ici de prédire la DDS avec des données incrémentales. Ensuite, nous proposons une modélisation de la DDS qui permet d'explorer des données incrémentales. Nous présentons deux méthodes de prédiction qui se basent sur les algorithmes d'apprentissage automatique.

4.2 Problématique

La Durée De Séjour hospitalier dépend de plusieurs facteurs. L'estimation d'une Durée De Séjour hospitalier se base sur les données stockées dans les Systèmes d'Informations Hospitalier (SIH). Ces données ne sont pas nécessairement disponibles en temps réel. La DDS est impactée par d'autres facteurs qui peuvent apparaître pendant le séjour. Par exemple, il est fréquent qu'il y ait un certain temps entre la collecte des données et leur présence dans les SIH. Un exemple de cette configuration est l'apparition des complications médicales suite à une opération chirurgicale. D'autres évènements imprévus peuvent également faire évoluer la DDS d'un patient. Par exemple, une opération médicale non prévue entraîne une modification de la DDS. le modèle à élaborer doit être capable de prendre en compte l'évolution des données médicales.

Le modèle statique de prédiction présenté dans le chapitre précédent ne suffit donc pas pour inclure de nouvelles données et raffiner la DDS prédite au moment de l'admission. D'autres part, en général, les données mémorisées dans les SIH contiennent des informations sur des processus du système de santé. l'enregistrement de ces processus génère une énorme quantité des données. Toutefois, dans le domaine de la santé, ces processus sont trop flexibles et présentent une grande variabilité [Naj17]. Cette propriété des données médicales ajoute une autre difficulté à leurs traitements. Nous proposons une codification des données qui arrivent d'une manière incrémentale dans les SIH. Ensuite, nous proposons deux méthodes de prédiction de DDS capables d'explorer les données incrémentales.

4.3 Méthodes de prédiction de DDS avec données incrémentales

Les données sur les séjours hospitaliers se présentent sous la forme d'un ensemble de variables difficiles à traiter et à gérer. Cet ensemble contient plusieurs types de données (numériques, catégorielles et catégorielles multivaluées). En plus de cette hétérogénéité, ces données sont souvent limitées et incomplètes. Cependant, nous avons adapté le processus de prédiction utilisé dans la conception du modèle statique de prédiction de DDS qui s'appuie sur les méthodes d'apprentissage automatique.

Pour rappel, le processus qui conduit à la prédiction de DDS et qui a été présenté dans le chapitre précédent se compose des étapes suivantes :

- La définition du périmètre d'étude.
- La modélisation de la Durée De Séjour.
- Le processus de prédiction de la Durée De Séjour qui se base sur les méthodes d'apprentissage automatique.

Le processus de prédiction de la Durée De Séjour s'articule autour des étapes suivantes :

- La collecte des données.
- L'analyse des données.
- Le pré-traitement des données.
- La sélection de variables.
- La transformation et encodage de données.
- La mise en place des algorithmes d'apprentissage.
- L'optimisation des hyper-paramètres et la validation croisée.

A la différence du modèle de prédiction statique de prédiction de DDS présenté dans le chapitre précédent, la modélisation de la DDS change pour le modèle séquentiel de prédiction de DDS. Cette modélisation intègre de nouvelles données qui arrivent au fur et à mesure du séjour hospitalier. Dans le cadre de notre travail, nous avons considéré la réalisation des actes médicaux codifiés à l'aide de la Codification Commune des Actes Médicaux (CCAM) comme de nouvelles données disponibles après l'admission du patient et tout au long de son séjour.

Nous avons commencé par regrouper l'ensemble des actes médicaux pour un patient P pendant un séjour S et récupérer la date de réalisation de chaque acte. Une analyse des différentes valeurs des codes CCAM est ensuite effectuée afin de comprendre leur structure et de détecter les différentes erreurs. La phase de pré-traitement de données englobe le nettoyage des codes CCAM (enlever les caractères spéciaux et les espaces blancs). Nous avons considéré les données concernant les actes CCAM comme données catégorielles multivaluées car pour chaque acte médical, il peut y avoir aucun ou plusieurs actes. Nous avons proposé une solution pour codifier ces données en les transformant en données séquentielles.

Une première étape est de définir le périmètre d'étude de la DDS. Nous avons défini le périmètre d'étude comme étant une unité médicale avec long séjour (hors urgence et ambulatoire). Nous avons étendu l'étude sur 4 unités médicales différentes comme pour le modèle statique de prédiction de DDS. Une nouvelle modélisation de la DDS est proposée afin d'inclure l'arrivée des actes médicaux. Dans ce qui suit, cette proposition est exposée en détails.

4.4 Modélisation de la DDS avec des données incrémentales

L'environnement hospitalier est un environnement très dynamique. Il connaît énormément de changements, de part, dans son organisation, et, d'autres part, dans l'apparition de nouveaux évènements. Une analyse des facteurs qui peuvent s'ajouter tout au long du séjour et qui impactent la DDS permet d'extraire des informations pertinentes. La modélisation de la DDS dans ce cas nécessite une compréhension des différents facteurs qui émergent au cours du séjour hospitalier. Ces facteurs sont les données collectées de l'admission du patient à sa sortie d'une manière différée. Ils sont enregistrés au fur et à mesure dans les SIH. Chaque information est sauvegardée à un instant T. Elle est représentée par le couple (information, temps de disponibilité). A partir de cette représentation, nous avons considéré la disponibilité d'une donnée à un instant T comme l'arrivée d'un évènement. A chaque apparition d'un nouvel évènement, la DDS initialement prédite au moment de l'admission est mise à jour.

Un modèle de DDS avec des données incrémentales est liée à la notion de temps. Il est considéré comme une succession de suite d'évènements. Dans ce contexte, nous définissons un évènement comme l'arrivée d'une information à un instant T_i. Au moment d'admission du patient (l'instant $T_{i=0}$), les données démographiques du patient, son motif d'hospitalisation ou son diagnostic principal et son historique

médical sont disponibles. Au fil du séjour hospitalier jusqu'à la sortie du patient (de l'instant $T_{i=0}$ jusqu'à l'instant $T_{i=N}$ avec N est égal à la DDS du patient), de nouvelles données apparaissent. Les patients admis dans une unité médicale sont suivis jusqu'à leur sortie et de ce fait de nombreuses informations médicales s'ajoutent. De plus, les nouvelles données concernent également les transferts entre les unités médicales qui ont une incidence directe sur l'organisation de l'hôpital. Le diagramme d'activité de la figure 4.1 suivant explique le séjour d'un patient dans un établissement de soins et qui passe par plusieurs unités médicales. La réalisation des actes médicaux est considérée comme-étant des événements dans une unité médicale précise. Dans cette étude, nous définissons la DDS comme l'intervalle de temps entre l'admission du patient et sa sortie d'une unité médicale. La partie encadrée de la figure 4.1 représente un cas d'étude de réalisation d'actes médicaux dans une seule unité médicale.

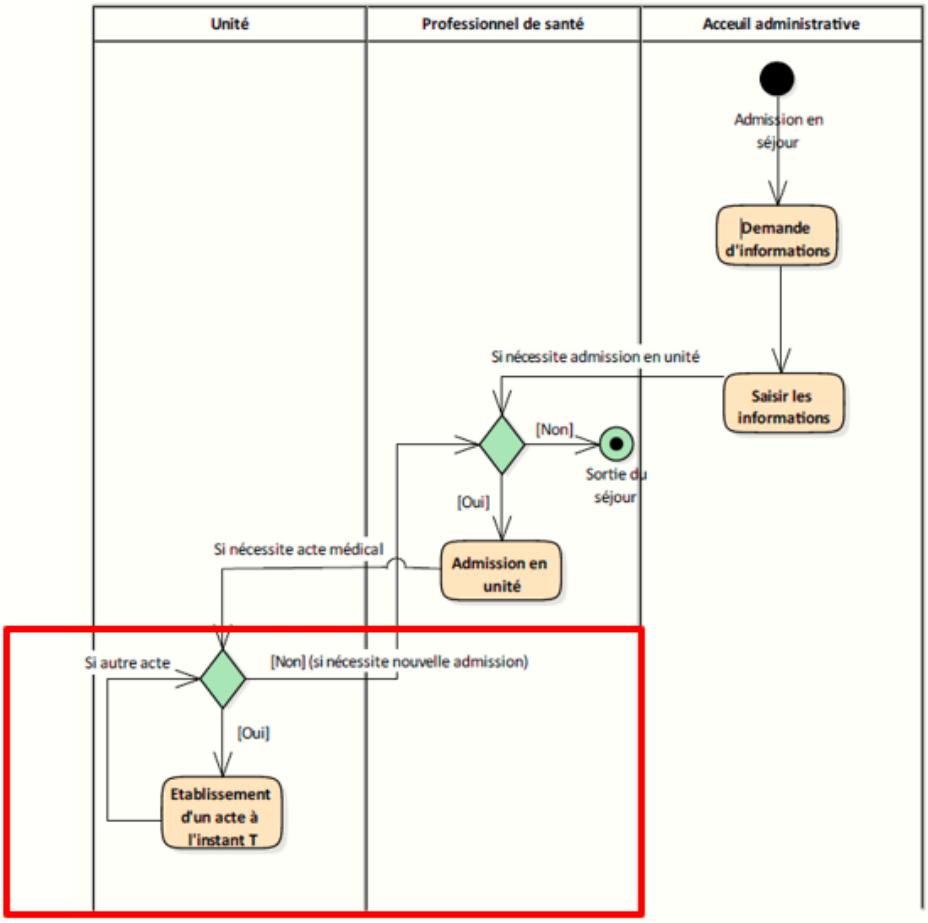


Fig. 4.1: Diagramme d'activité : description d'un séjour hospitalier avec réalisation d'actes médicaux.

Une configuration possible qui illustre un séjour hospitalier en considérant les données incrémentales et ainsi l'aspect temporel est la suivante :

- Au moment de l'admission du patient ($T = T_0$), les données démographiques sont récupérées. Ces données rassemblent l'identifiant du patient, son âge, son sexe, sa situation familiale et son adresse. Les informations concernant le motif d'hospitalisation et l'historique médical du patient et qui décrivent son état médical sont également disponibles à cet instant. Certaines données administratives, comme par exemple, le type de l'unité médicale où le patient est admis et le mode d'admission à cette unité sont recueillies.
- A l'instant $T_i > 0$, avec $i = <0,..N>$, un ensemble d'actes médicaux tels que les analyses biologiques, les scanners, les radiographies sont réalisés. De ce fait, les résultats de ces actes médicaux sont accessibles. De plus, les complications médicales survenues suite à une opération chirurgicale par exemple sont de nouvelles données qui peuvent se produire dans cet intervalle de temps. Ces complications peuvent étendre la DDS à plusieurs jours par rapport à ce qui a été prévu initialement.
- A la sortie du patient et à l'instant T_N, les rapports médicaux de sortie sont rédigés par les professionnels de santé et les données économiques telles que les informations de paiement, le type de tarification et le type d'assurance maladie appliquée sont aussi sauvegardés dans les SIH.

La figure 4.2 illustre une suite d'évènements observés durant un séjour hospitalier dans une unité médicale. Aux instants $T = i$, $T = j$ et $T = z$ différents actes médicaux sont réalisés. Ces actes peuvent être réalisé à tout instant du séjour hospitalier.

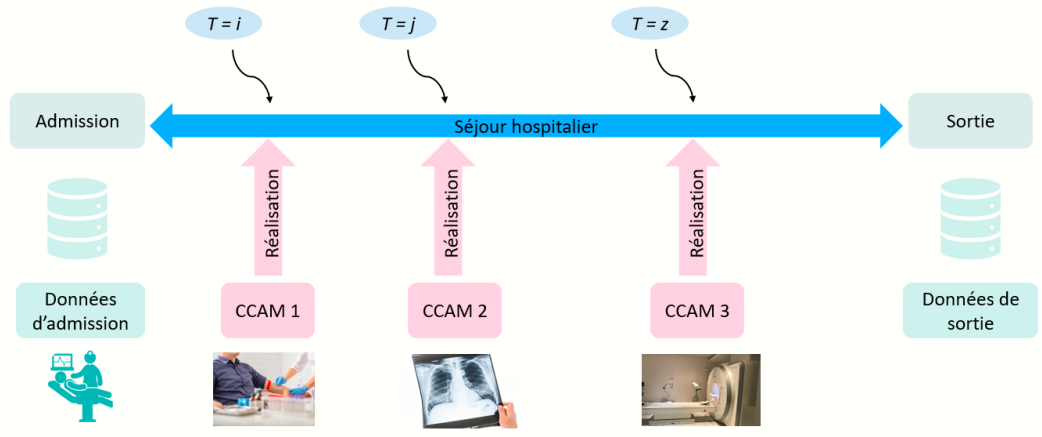


Fig. 4.2: Séjour hospitalier : réalisation d'actes médicaux comme évènements.

Les données provenant du Programme de Médicalisation des Systèmes d'Informations présentent bien cet aspect temporel. Ces données proviennent de plusieurs sources et s'enrichissent en fonction de la réalisation des actes médicaux et l'accès aux résultats des prélèvements sanguins. En plus des données utilisées dans la conception du modèle statique de prédiction des DDS présenté dans le chapitre précédent, les actes

médicaux sont intégrés au fur et à mesure de leur disponibilité. Les actes médicaux sont représentés à l'aide de la Codification Commune des Actes Médicaux (CCAM) introduite dans le chapitre 1.

Pendant le séjour hospitalier, plusieurs actes médicaux peuvent se produire. Chaque acte est codé par un code alphabétique. L'ensemble de tout les actes médicaux réalisés pour un patient est la successions de tout les actes qui a subi. C'est une chaîne de caractères de l'ensemble des actes séparés par un caractère blanc. Le type de cette variable est catégorielle multivaluée. La complexité derrière l'utilisation de ce type de données dans la prédiction des DDS est d'une part, leur représentation numérique pour les introduire dans un modèle d'apprentissage automatique et d'autre part, la préservation de la chronologie d'arrivée des actes.

La section suivante est consacrée à la démonstration de notre solution proposée dans le cadre du modèle séquentiel de prédiction des DDS.

4.5 Prédiction de DDS avec données incrémentales

L'intégration des données incrémentales dans le processus de prédiction de DDS doit prendre en compte le séquencement de ces données. Dans le cadre de notre travail, nous avons considéré la réalisation des actes médicaux comme une donnée disponibles après l'admission du patient. La date de réalisation de chaque acte médical est sauvegardée, nous avons proposer une solution permettant de structurer les données incrémentales tout en respectant l'ordre de leur arrivée et en mettant à jour la DDS. C'est l'une des principales contributions de cette thèse.

Dans ce qui suit, nous détaillons cette solution.

4.5.1 Structuration des données incrémentales

Le problème de l'extraction de modèles prédictifs à partir des données séquentielles a été discuté dans la littérature, avec et sans prise en compte de l'aspect temporel. Les données issues du PMSI telles qu'elles sont sauvegardées dans les SIH requièrent des pré-traitements pour valoriser leur contenu. La solution proposée prend en entrée les informations des différents séjours hospitaliers dans une unité médicale et retourne la DDS mise à jour pour chaque patient. Les différents évènements représentés par la réalisation des actes médicaux survenus au cours de son séjour sont ajoutés.

Pour chaque séjour d'un patient dans une unité médicale, il peut y avoir aucun à plusieurs actes médicaux. Le nombre d'actes médicaux réalisés est différant d'un patient à un autre. Une variable catégorielle multivaluée contenant tout les codes CCAM du patient est prise en compte. La méthode que nous avons établi est de définir un ensemble de processus pour chaque paire (séjour S, patient P). Le processus est représenté par une suite d'évènements chronologiques ou d'informations disponibles à des instants différents du séjour. Ces enregistrements sont perçus comme un journal d'évènements. En effet, lorsque nous travaillons sur des processus, le format de données le plus souvent rencontré est le journal d'événements. Ils regroupent des instances appelées traces pour lesquelles chaque évènement programmé est listé par ordre de leur arrivée. Chaque évènement est décrit par une seule ou plusieurs variables [Oli20].

L'intégration de nouvelles données au fur et à mesure dans le modèle d'apprentissage automatique afin de raffiner la prédiction des DDS doit prendre en compte l'ordre d'arrivée de ces données et leur occurrence. Nous devons donc construire les instances de chaque couple (séjour S, patient P) tout en respectant ces deux contraintes : l'ordre d'arrivée d'un acte et son occurrence. Un acte médical peut être répété plusieurs fois à des instants différents. Un exemple est de réaliser plusieurs scanners et suivre l'évolution de leurs résultats suite à un traitement donné. La datation de la réalisation de chaque acte est importante pour calculer le temps entre les actes. De plus, la fréquence de réalisation de chaque acte permettra d'attribuer des poids au différent actes accomplis. Cette pondération permet de donner plus d'importance à des actes qu'à d'autres. Par conséquent, les actes qui apparaissent le plus auront plus de poids que les actes par les moins fréquents. Nous avons travaillé la structure des données utilisées. Notre démarche de structuration de données est la suivante :

- Récupérer pour chaque couple (séjour S, patient P), l'ensemble de tous les actes et leurs dates de réalisation.
- Grouper les instances par couple de (séjour S, patient P) et ordonner chaque groupe par la date de réalisation des actes médicaux.
- Éclater chaque couple (séjour S, patient P) en N lignes. N représente le nombre d'actes concernant le couple (séjour S, patient P). L'apparition des instances est listée par ordre chronologique croissant de la date de réalisation de l'acte le plus ancien à celle du plus récent.
- Modifier la valeur de la variable cible à prédire DDS pour montrer la relation entre l'évolution du séjour hospitalier et la réalisation des actes chronologiquement. Une nouvelle variable est créée aussi représentant le nombre de jours déjà passés à l'hôpital. Elle servira comme marqueur par la suite et est intégrée

comme nouvelle information dans le modèle de prédiction. Sa valeur est égale à celle de la variable *time step* qui représente la taille des séquences utilisées dans les réseaux de neurones récurrents de type LSTM utilisés par la suite. La DDS à prédire est mise à jour. Elle est obtenue en soustrayant la valeur de la DDS courante (nombre de jours passés à l'hôpital) de la valeur de la DDS initialement prédite avant la réalisation de l'acte médical.

La figure 4.3 illustre un exemple de la démarche décrite ci-dessus.

Date d'entrée	Date de sortie	Acte CCAM	Date de l'acte	DDS courante	Time step	DDS restante	Données +
07/01/2017	12/01/2017	CCAM 1	07/01/2017	0	0	5	X
07/01/2017	12/01/2017	CCAM 1	10/01/2017	3	3	2	X
07/01/2017	12/01/2017	CCAM 2	11/01/2017	4	4	1	X
07/01/2017	12/01/2017	CCAM 3	12/01/2017	5	5	0	X

Fig. 4.3: Exemple de structuration des actes médicaux CCAM.

Le séjour du patient débute le 07/01/2017 et s'achève le 12/01/2017. La DDS de ce patient est donc égale à 5 jours. Le patient a subi 4 fois des actes médicaux. L'acte $CCAM1$ est répété deux fois : le premier jour du séjour et le troisième jour du séjour. Nous avons ajouté la variable $timestep$ comme indicateur du nombre de jours passés à l'hôpital. Après 3 jours de séjour, l'acte $CCAM2$ est réalisé, la DDS restante est mise à jour en soustrayant 3 de 5. Ce calcul est répété autant de fois qu'il y a des actes CCAM jusqu'à la sortie du patient de l'hôpital.

Quoique la succession des instances montre bien l'aspect temporelle d'arrivée des données, la liaison entre les différentes instances n'est pas formelle ou explicite. Or, pour utiliser les algorithmes d'apprentissage automatique dans la prédiction, il est important de montrer cette liaison au sein de chaque instance. Ces algorithmes se basent sur chaque instance indépendemment des autres pour détecter les différents motifs présents dans la base de données. De ce fait, en déduire les valeurs de la DDS pour les instances qui arrivent par la suite. Pour palier ce problème, nous avons représenté les données comme des séquences d'évènements. Pour chaque séjour S d'un patient P, chaque acte médical pour une instance d'indice i est codifié par la

succession des actes médicaux des instances qui la précèdent en ajoutant la valeur de l'acte médical de cette instance d'indice i. La figure 4.4 illustre un exemple de cette représentation séquentielle des actes médicaux.

Date d'entrée	Date de sortie	Acte CCAM	Date de l'acte	DDS courante	Time step	DDS	Données +
07/01/2017	12/01/2020	CCAM 1	07/01/2017	0	0	5	X
07/01/2017	12/01/2020	CCAM 1, CCAM 1	10/01/2017	3	3	2	X
07/01/2017	12/01/2020	CCAM 1, CCAM 1, CCAM 2	11/01/2017	4	4	1	X
07/01/2017	12/01/2020	CCAM 1, CCAM 1, CCAM 2, CCAM 3	12/01/2017	5	5	0	X

Fig. 4.4: Exemple de représentation séquentielle des actes médicaux CCAM.

L'acte $CCAM1$ est réalisé le premier jour du séjour. Ensuite, et après 3 jours de séjour, l'acte $CCAM1$ a été réalisé une deuxième fois successivement. De ce fait, Après 3 jours de séjour, nous retrouvons la présence de l'acte $CCAM1$ deux fois de suite. Le dernier jour du séjour, le patient a subi la liste des actes médicaux suivants : $CCAM1$, $CCAM2$ et $CCAM3$, avec répétions de l'acte $CCAM1$ deux fois. L'aspect séquentiel apparaît dans cette représentation et nous gardons trace de tout les actes médicaux qui ont été réalisés du début du séjour hospitalier jusqu'à sa fin.

Les valeurs de la variable qui symbolise les actes médicaux est donc de type catégoriel multivalué. Les variables catégorielles multivaluées comportent de un à plusieurs modalité en même temps. Par conséquent, ce type de données est plus difficile à gérer et n'est pas utilisable directement par les algorithmes d'apprentissage automatique. La méthode que nous avons employé dans l'encodage des données catégorielle "one-hot-encoding" n'est plus adapté à ce type de données. En effet, cette méthode considère toutes les différentes valeurs que peut prendre une variable catégorielle comme une seule modalité. Nous proposons une solution pour l'encodage de la variable représentant les actes médicaux. Cette solution respecte l'aspect séquentiel des données. Nous la détaillons dans la section suivante.

4.5.2 Encodage des actes médicaux

Plusieurs techniques existent pour l'encodage des données catégorielles multivaluées et l'encodage des données catégorielles à une seule modalité. Parmi ces techniques, nous pouvons citer la codification séquentielle, la codification binaire et la méthode du "one-hot-encoding". Avec ces techniques, les données catégorielles multivaluées sont transformées en valeurs numériques et chaque séquence de valeurs catégorielles multivaluées est considérée comme une valeur catégorielle à une seule modalité. Pour représenter les différents actes médicaux, nous avons projeté notre représentation des données à une autre description souvent utilisée en recherche d'information textuelle appelée « sac des mots » (Bag Of Words en anglais). L'ensemble des différents actes médicaux existants dans la base de données constituent ce « sac de mots » [FSF13].

En recherche d'information textuelle, la description en « sac de mots » consiste à représenter un document textuel par un vecteur de mots. A chaque mot, une valeur numérique lui est affectée. Cette valeur peut représenter un code numérique, sa fréquence d'apparition dans le document ou bien une pondération calculée à l'aide des formules mathématiques. Supposons que nous ayons un corpus de documents textuels. Un document d_i est composé de plusieurs termes t_j. le « sac de mots » est constitué des différents termes présents dans le corpus. Le corpus de documents est représenté par une table à deux dimensions dont les lignes représentent les documents et les colonnes l'ensemble de tout les termes constituant le « sac de mots », comme le montre la table 4.2.

Document /Terme	***Terme 1***	***Terme 2***	***Terme j***	***Terme M***
Document 1	$w_{1,1}$	$w_{1,2}$	$w_{1,j}$	$w_{1,M}$
Document 2	$w_{2,1}$	$w_{2,2}$	$w_{2,j}$	$w_{2,M}$
Document i	$w_{i,1}$	$w_{i,2}$	$w_{i,j}$	$w_{i,M}$
Document N	$w_{N,1}$	$w_{N,2}$	$w_{N,j}$	$w_{N,M}$

Tab. 4.1: Représentation d'un corpus de document textuels.

A chaque terme t_j, est attribué un poids $w_{i,j}$. Une pondération souvent employée est la pondération $TF * IDF$ (Term Frequency-Inverse Document Frequency en anglais) dans laquelle nous avons deux champs : TF et IDF. La grandeur $TF * IDF$ vise à accorder une pertinence lexicale à un terme au sein d'un document. Le premier champ TF (Terme Frequency en anglais) se base sur la fréquence du mot dans le document. Cette fréquence sera comparée aux autres fréquences des mots restants du document. Le TF inclue également la proportion de tous les mots apparus dans un document. Le deuxième champ IDF détermine la fréquence inverse du document

(Inverse Document Frequency en anglais). Il complète l'évaluation et l'analyse du mot dans le document. En définitive, $TF * IDF$ s'obtient par :

$$\text{TF*IDF} = \frac{\log_2(Freq(i,j)+1)}{\log_2(N)} * \log_2(\frac{M}{L}+1) \tag{4.1}$$

N représente le nombre total des différents mots présents dans tous les documents. Ainsi, il représente la taille de la liste « sac des mots ». M représente le nombre total de document et L représente le nombre de document contenant le mot i.

Nous nous sommes inspiré de cette technique de représentation des termes dans les documents textuels pour codifier les actes médicaux CCAM de type catégoriel multivalué. Ainsi, nous avons considéré les informations d'un séjour hospitalier dans une unité médicale comme un document textuel et l'ensemble des actes médicaux comme des mots appartenant à ce document. Nous illustrons cela dans la table 4.2 ci-dessous.

Séjour/acte	***CCAM 1***	***CCAM 2***	***CCAM j***	***CCAM M***
Séjour 1	$w_{1,1}$	$w_{1,2}$	$w_{1,j}$	$w_{1,M}$
Séjour 2	$w_{2,1}$	$w_{2,2}$	$w_{2,j}$	$w_{2,M}$
Séjour i	$w_{i,1}$	$w_{i,2}$	$w_{i,j}$	$w_{i,M}$
Séjour N	$w_{N,1}$	$w_{N,2}$	$w_{N,j}$	$w_{N,M}$

Tab. 4.2: Représentation des actes médicaux CCAM.

Pour la codification des actes médicaux, nous avons dans un premier temps réduit l'ensemble des actes médicaux en les regroupant par catégories. Cette réduction se base sur les connaissance dans le domaine de santé et particulièrement dans la codification des codes CCAM. L'objectif est de réduire le nombre des codes CCAM et d'obtenir un ensemble significatif d'actes médicaux. Ensuite, nous avons défini la liste des différents actes médicaux. L'ensemble de ces actes symbolise le « sac des actes ». Chaque valeur d'un acte médical est transformée en un vecteur de taille N. Ce vecteur contient les poids $w_{i,j}$ mesurés à l'aide de la formule $TF * IDF$ décrite ci-dessus. Le format final des données est obtenu par ce processus montré dans la table suivante.

Séjour/acte	***Données démographiques***	***Données administratives***	***Données médicales***	***CCAM 1***	***CCAM 2***	***CCAM j***	***CCAM M***	***DDS***
Séjour 1	valeur	valeur	valeur	$w_{1,1}$	$w_{1,2}$	$w_{1,j}$	$w_{1,M}$	DDS 1
Séjour 2	valeur	valeur	valeur	$w_{2,1}$	$w_{2,2}$	$w_{2,j}$	$w_{2,M}$	DDS 2
Séjour i	valeur	valeur	valeur	$w_{i,1}$	$w_{i,2}$	$w_{i,j}$	$w_{i,M}$	DDS i
Séjour N	valeur	valeur	valeur	$w_{N,1}$	$w_{N,2}$	$w_{N,j}$	$w_{N,M}$	DDS N

Tab. 4.3: Représentation finale des données.

Cet ensemble de données aussi obtenu est ensuite utilisé par les algorithmes d'apprentissage automatique que nous avons employé dans le cadre du modèle séquentiel de prédiction de DDS.

4.6 les méthodes ensemblistes dans la prédiction des DDS avec données incrémentales

Un algorithme de prédiction se défini par ses entrées (données) et sa fonction objective à optimiser (maximiser une précision ou minimiser une erreur). Le modèle séquentiel de prédiction de Durée De Séjour hospitalier se base sur des données disponibles au moment de l'admission du patient ainsi que sur les nouvelles données disponibles tout au long du séjour hospitalier. Pour la mise en place d'un modèle de prédiction qui intègre ces nouvelles données, nous avons proposé une structure et codification des données médicales incrémentales. Ce modèle prend en compte l'aspect temporel des données. Le processus de prédiction s'appuie sur les méthodes d'apprentissage automatique comme c'est le cas pour le modèle de prédiction statique de DDS présenté dans le chapitre précédent. Nous avons étudié les méthodes ensemblistes basées sur les arbres de décision testées précédemment. L'objectif est de comparer entre les techniques du boosting et du bagging afin d'obtenir les meilleures prédictions. Les algorithmes Random Forest (RF), Gradient Boosting Model (GBM) et le Extrem Gradient Boosting (Xgboost) sont implémentés.

L'ensemble des données dans le cadre du modèle séquentiel de prédiction de DDS est défini par un tableau à N lignes et M colonnes. Le nombre d'instances N décrit les séjours hospitaliers dans différentes unités médicales pour différents patients. Chaque instance est indépendante des autres. La structuration des données proposée dans la section 4.5.1 est utilisée pour mettre en évidence l'aspect séquentiel du modèle. Les colonnes représentent les variables disponibles au moment de l'admission du patient utilisées dans le modèle de prédiction statique de DDS en plus du « sac d'actes médicaux ». Pour chaque couple (séjour S, patient P), il peut y avoir aucun à plusieurs actes médicaux. Si pour un patient P, un acte médical n'existe pas alors le poids de cet acte vaut 0. Dans le cas contraire, ce poids vaut la pondération $TF * IDF$ de l'acte. La DDS étant mise à jour d'une instance à l'autre, sa prédiction peut être généralisée à tout instant d'un séjour hospitalier. En effet, avec la structure de données proposée dans la section 4.5.1, nous ne sommes pas limités à des instants précis du séjour comme par exemple après 3 jours d'admission ou après 5 jours d'admission mais nous avons étendu l'étude à chaque arrivée d'un acte médical.

Le processus de prédiction qui se base sur les algorithmes (Random Forest), le Gradient Boosting Model (GBM) et le Extreme Gradient Boosting (Xgboost) se compose des étapes décrites dans la section 4.3. Ainsi, nous avons adopté les mêmes techniques de pré-traitement des données disponibles au moment d'admission du patient et nous avons alimenté le modèle progressivement au fil de l'arrivée de nouvelles données.

Après avoir défini l'ensemble des données utilisé par les algorithmes d'apprentissage automatique, nous avons défini la fonction de perte et la fonction d'évaluation de ces algorithmes. La DDS ajustée est quantifiée par le nombre de jours restants à passer à l'hôpital. Le problème de prédiction consiste alors à prédire une valeur numérique. La fonction de perte à minimiser est fixée comme la mesure de la moyenne des erreurs quadratiques (MSE). La fonction d'évaluation choisie est la moyenne des écarts absolus (MAE).

Le processus de prédiction en temps réel se présente comme suit dans le cadre du modèle séquentiel de prédiction de DDS :

- Au moment d'admission du patient : Aucun acte médical n'est encore enregistré. Les valeurs du vecteur qui comporte les poids $TF * IDF$ des codes CCAM valent 0.
- Réalisation d'un premier acte médical : Le poids $TF * IDF$ de cet acte est calculé. Il est ajouté dans le vecteur de « sac d'actes médicaux ».
- Réalisation des actes médicaux suivants : l'historique des actes médicaux CCAM réalisés pour un patient est récupéré. Les pondérations $TF * IDF$ sont mises à jour pour chaque acte médical présent dans le vecteur de « sac d'actes médicaux ».

La figure suivante montre le processus décrit ci-dessus.

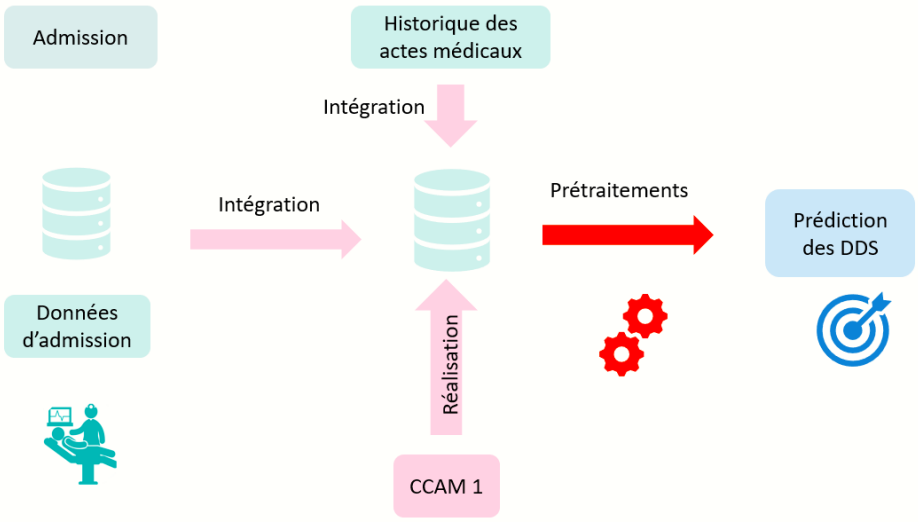


Fig. 4.5: Prédiction en temps réel des DDS avec des données incrémentales.

Un autre type d'algorithmes largement utilisé dans la prédiction des séries temporelles ou les données séquentielles est les réseaux de neurones récurrents. L'entrée à cet algorithme est une structure de données à 3 dimensions en intégrant une dimension temporelle. Dans cette thèse, nous avons exploré les LSTM pour Long-Short-Term-Memory comme un type des réseaux de neurones récurrents. L'objectif est d'utiliser la propriété de cet algorithme à inclure les prédictions antérieures comme entrée aux prochaines prédictions. Les LSTM ont fait leurs preuves dans la résolution des problèmes de séquences efficacement. Nous allons dans la section suivante détaillé notre proposition.

4.7 les réseaux de neurones récurrents dans la prédiction des DDS avec données incrémentales

Les réseaux de neurones récurrents (RNN) sont largement utilisés dans la résolution des problèmes à caractère temporel. Les données en entrée de ce type d'algorithme possède une troisième dimension représentant le temps. Les problèmes de séquences sont catégorisés comme suit :

- Séquence un à un : une seule entrée correspond à une seule sortie. La classification d'images est un problème de séquence un à un où à une image correspond une catégorie.

- Séquence plusieurs à un : l'entrée est une séquence de plusieurs variables et la sortie est une seule valeur soit numérique soit catégorielle. La majorité des problèmes de classification ou de régression se positionnent dans cette sous-catégorie.
- Séquence un à plusieurs : Il y a une seule variable d'entrée et la sortie se compose de plusieurs variables.
- Séquence plusieurs à plusieurs : l'entrée et la sortie se composent de plusieurs variables.

Le modèle séquentiel de prédiction de DDS est décrit par une séquence de variables qui définissent les caractéristiques en entrée et une seule sortie qui est la DDS. La DDS est représentée par une valeur numérique. Plusieurs type de RNN existent. Les LSTM sont largement utilisés dans la prévision des séries temporelles comme par exemple la prédiction des changements climatiques. Une série temporelle est représentée par des données séquentielles structurées comme suit :

Jour	Heures	Température
1	00H	38,5
	06H	39
	16H	38
	18H	38,5
2	00H	37,9
	06H	37,5
	16H	37,2
	18H	37

Fig. 4.6: Exemple des données de type série temporelle.

Pour un patient, sa température est prise deux jours successivement. Chaque 6 heures, cette température est enregistrée. La prédiction de la température du troisième jour se base sur celles des deux premiers jours.

Les données en entrée pour une architecture LSTM ont 3 dimensions :

- Le nombre d'instances représentants les séjours hospitaliers dans les unités médicales dans notre cas.
- Un indicateur d'intervalle de temps connu comme « time step » : cet indicateur représente la quantité de données stockée dans la mémoire du LSTM afin de réaliser la prédiction. Si ce « time step » est égal à N, alors la prédiction se base sur les données des N séquences précédentes.

- Le nombre de variables : à chaque « time step » correspond M variables caractérisant la variable cible. Dans notre cas, l'ensemble de ces variables représente d'une part des variables disponibles au moment de l'admission du patient telles que les informations démographiques du patient, les informations médicales et, d'autre part, les différents actes médicaux réalisés à des instants différentes.

Une étape primordiale pour les LSTM est de formater de manière séquentielle les données. La plus grande difficulté dans le processus concerne la transformation des données issues du PMSI sous forme tabulaire vers une représentation séquentielle. Dans la représentation tabulaire, chaque N lignes représente un séjour et les M colonnes représentent les variables indépendantes. La colonne $M + 1$ est la variable cible DDS. Dans la représentation séquentielle, une dimension temporelle est ajoutée. La taille de cette dimension est défini par le « time step ». La figure 4.7 illustre les deux représentations tabulaire et séquentielle.

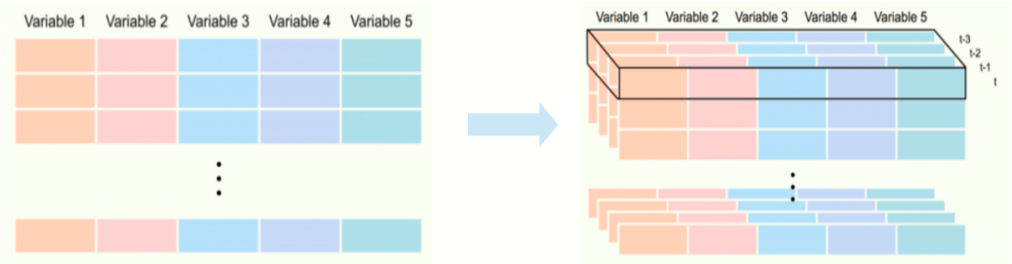


Fig. 4.7: Représentation tabulaire et séquentielle des données.

Les données issues du PMSI ne reflètent pas explicitement un processus d'évènements qui se déroule au fil du temps. Une des contributions essentielles de cette thèse est de proposer une solution pour préparer les données issues du PMSI pour les injecter dans une architecture de type réseaux de neurones LSTM.

Les réseaux de neurones récurrents dont les LSTM requièrent des entrées de même taille. Chaque entrée est représentée par les 3 dimensions (séjour, time step et variables). Concernant les données issues du PMSI, cette taille représente le nombre d'actes médicaux réalisés pour un patient P pendant un séjour S. Comme le nombre d'actes médicaux diffère d'un patient à un autre, la taille de chaque entrée est alors différente des autres entrées. Deux contraintes sont à respecter lors de la préparation des données issues du PMSI pour un algorithme type LSTM :

- L'information temporelle réside dans l'arrivée des actes médicaux. Le nombre d'actes médicaux réalisés diffère d'un patient à un autre. Avant la transforma-

tion des données de la forme tabulaire à la forme séquentielle, il faut d'abord normaliser la taille de chaque couple (séjour S, patient P) présent dans la base de données. La normalisation des tailles évite le mélange des informations de deux profils différents.

- La solution proposée doit inclure toutes les configurations possibles d'arrivée des actes. Chaque configuration possède une sortie représentant la valeur de la DDS. Avec ça, le processus de prédiction de DDS est possible à tout instant d'un séjour hospitalier.
- Le paramètre « time step » est un paramètre en entrée de l'algorithme LSTM. La solution proposée prends en compte la flexibilité et les différentes valeurs de ce paramètre dans la normalisation des tailles des données.

Il est primordial de normaliser la taille des données avant de procéder à leur transformation du format tabulaire au format séquentiel. La figure 4.8 illustre des données concernant la réalisation d'actes médicaux de deux patients. Les deux patients ont subi un nombre différent d'actes médicaux.

Patient	Jour	Actes médicaux CCAM
1	1	CCAM 1
	1	CCAM 2
	2	CCAM 3
	3	CCAM 4
2	3	CCAM 1
	4	CCAM 2
	5	CCAM 3

Patient 1 avec 4 actes

Patient 2 avec 3 actes

Fig. 4.8: Taille des données concernant les actes médicaux CCAM.

Une des méthodes les plus connues pour la normalisation dans ce contexte est nommé le remplissage ou le padding en anglais. Le remplissage vise à indiquer à l'algorithme que certains pas de temps « time step » dans une séquence sont manquantes et donc doivent être ignorés lors du traitement des données. L'information à ignorer est au début ou à la fin d'une séquence. Cette définition découle de l'obligation de coder les données de séquence en lots contigus. Ceci assure que toutes les séquences d'un lot correspondent à une longueur standard donnée. Il est alors impératif de tronquer ou de compléter certaines données [Ten21].

Pour un patient P, nous avons rassemblé ces informations et rajouté Z lignes supplémentaires au début des séquences. La valeur de Z est égale à la valeur du paramètre « time step ». L'algorithme suivant décrit les étapes que nous avons appliqué dans la normalisation de la taille des données :

Algorithme 1 : Remplissage de séquences

Entrées : Ensemble de données PMSI, taille de la séquence
Sorties : Ensemble de données PMSI après remplissage

1 **pour chaque** *Patient* P **faire**

- Récupération des données D disponibles au moment de l'admission du patient P.
- Création d'un vecteur vide pour la variable acte médical de taille égale au nombre total des actes différents de la base de données. Ce nombre représente simplement le nombre de mots appartenant au « sac des actes médicaux ».

pour chaque $i = 0$ *à (taille de la séquence - 1)* **faire**
Fusion des deux informations D et le vecteur des actes médicaux.
Ajout d'une nouvelle ligne au début de la séquence des données contenant ces informations.

Après la normalisation de la taille des données, la transformation nécessaire du format tabulaire à 2 dimensions vers le format séquentiel est effectuée. L'algorithme suivant présente en détails les étapes de cette transformation.

Algorithme 2 : Ajout de la dimension temporelle

Entrées : Ensemble de données PMSI après remplissage en 2 dimensions, taille de la séquence
Sorties : Ensemble de données PMSI en 3 dimensions avec taille = (nombre d'instances, taille de la séquence, nombre de variables)

1 **pour chaque** *Patient* P **faire**
2 **pour chaque** i*=0 à ((nombre d'instance pour le patient* P *- taille de la séquence)+1)*
faire
3 $j = i$;
$k = 0$;
tant que $K <$ *taille de la séquence* **faire**

- Ajout des informations dans le vecteur crée dans l'algorithme 1 ;
- $j = j + 1$;
- $k = k + 1$;

4 Ajout du vecteur résultat à 3 dimensions à la base de données.

A partir de cela, nous obtenons une structure de données à 3 dimensions. Une phase d'optimisation des hyper-paramètres de cet algorithme et une phase de validation des résultats à l'aide de la technique de validation croisée sont menées.

4.8 Conclusion

L'exploration des données stockées dans les Systèmes d'Informations Hospitalier (SIH) fait face à plusieurs difficultés. Les données médicales sont des données très particulières. Elles sont hétérogènes et incrémentales. Comme la prédiction des Durées De Séjour hospitalier se base sur les données du SIH, il est nécessaire d'inclure les données incrémentales dans le processus de prédiction. La conception d'un modèle séquentiel de prédiction de DDS a pour objectif de mettre à jour la valeur de la DDS si besoin.

Dans ce chapitre nous avons présenté une solution pour l'intégration des nouvelles données disponibles au fil du séjour hospitalier. Ces données concernent la réalisation des actes médicaux codifié à l'aide de la Codification Commune des Actes Médicaux (CCAM). Les codes CCAM sont de type catégoriel multivalué. Nous avons proposé une méthode d'encodage de ce type de données afin de préserver la séquentialité de leur arrivée. Plusieurs algorithmes d'apprentissage automatique ont été explorés : le Random Forest (RF), le Gradient Boosting Model (GBM), le Extreme Gradient Boosting (Xgboost) les Réseaux de Neurones Récurrents de type Long-Short Term Memory (LSTM). Nous avons également exposé une méthode de pré-traitement de données classiques pour les transformer sous forme séquentielle et les utiliser dans l'algorithme LSTM par la suite.

La suite de ce rapport de thèse présente les expérimentations et les résultats obtenus en suivant les processus décrits auparavant : le modèle statique de prédiction de DDS et le modèle séquentiel de prédiction de DDS. Ces expérimentations ont été faites sur des données réelles issues du Programme de Médicalisation des Systèmes d'Informations (PMSI) implémenté dans les établissements de soins français.

5

Implémentation et évaluation expérimentale : Données PMSI.

> *Everything is theoretically impossible, until it is done.*
>
> — **Robert A. Heinlein**
> Écrivain.

5.1 Introduction

Nous consacrons ce chapitre à la description de l'environnement matériel et logiciel utilisé lors de la mise en pratique de nos contributions. Nous allons exposer les résultats obtenus pour les deux modèles de prédiction à savoir le modèle statique de prédiction de DDS et le modèle séquentiel de prédiction de DDS. Ensuite, nous évaluons les résultats pour chaque méthode utilisée et nous discutons ces résultats.

5.2 Description de l'ensemble de données

L'étape initiale de la conception des processus d'apprentissage automatique est la définition des variables indépendantes et la définition de la variable dépendante. L'ensemble des données utilisées dans nos expérimentations est issu des données du Programme de Médicalisation des Systèmes d'Informations (PMSI). Particulièrement, les données du PMSI-MCO sont utilisées. Ces données proviennent du Groupement Hospitalier des Instituts Catholiques de Lille (GHICL). Elles concernent des séjours hospitaliers dans les unités médicales suivantes : cardiologie, pédiatrie, néonatologie et médecine polyvalente survenus entre le 01/01/2017 et le 31/12/2019.

Nous avons caractérisé un séjour hospitalier par un ensemble de variables indépendantes et une variable dépendante représentée par la Durée De Séjour hospitalier (DDS). Cet ensemble de variables indépendantes est construit en se basant sur le modèle générique proposé dans le chapitre 2.

L'ensemble des données est représenté par un tableaux de N linges et $M+1$ colonnes. Une ligne représente un séjour hospitalier au sein d'une unité médicale précise. Les M premières colonnes représentent les attributs caractérisant la DDS. La colonne $M+1$ est la variable cible DDS. Concernant les données PMSI, nous avons utilisé le schéma présenté dans la figure 3.4 pour récupérer les variables décrites dans le tableau 5.1.

Variable	*Description*	*Type*
ID séjour global	Identifiant du séjour hospitalier global	Numérique
ID séjour UM	Identifiant su séjour hospitalier dans une unité médicale	Numérique
ID patient	Identifiant du patient	Numérique
UM	Unité médicale visitée	Catégorie
genre	le sexe du patient	Catégorie
distance	La distance entre l'établissement de soins et l'adresse du patient. C'est une différence entre les codes postaux.	Numérique
âge	L'âge du patient	Numérique
date d'entrée	Date d'entrée à l'unité médicale	Date
date de sortie	Date de sortie de l'unité médicale	Date
mode d'admission	Mode d'admission du patient	Catégorie
mode de sortie	Mode de sortie du patient	Catégorie
provenance	Provenance du patient	Catégorie
destination	Destination du patient	Catégorie
code CIM_10	Codification de la maladie du patient	Catégorie
Type code CIM_10	Type du code à savoir, principal, associé ou relié	Catégorie
Acte CCAM	Acte médical réalisé codifié à l'aide du CCAM	Catégorie multivaluée
Date acte CCAM	Date de réalisation de l'acte	Date
DDS	Durée De Séjour hospitalier dans une unité médicale en jours	Numérique

Tab. 5.1: Description des données PMSI utilisées.

Pour prendre en compte les besoins quotidiens des hôpitaux dans le but d'optimiser leurs ressources et de suivre l'organisation de leurs services, nous avons également exploré des données supplémentaires. Ces données sont décrites dans le tableau 5.2

Variable	*Description*	*Type*
Jour d'admission	Si jour d'admission est en week-end alors 1, 0 sinon.	Catégorie
Jour de sortie	Si jour de sortie est en week-end alors 1, 0 sinon.	Catégorie
Charge	Nombre de lits occupés	Numérique
nombre d'UM	Nombre d'unités médicales visitées dans les séjours précédents	Numérique
antécédent DDS	Somme des DDS des séjours précédents	Numérique
moyenne de DDS par UM	Moyenne de la DDS par unité médicale	Numérique
Moyenne de DDS par DP	Moyenne de la DDS par motif d'hospitalisation	Numérique

Tab. 5.2: Description des données ajoutées.

En général, les admissions et les sorties sont moins fréquentes en week-end en raison du manque du personnel. C'est pour cela que nous avons ajouté le jour d'admission ainsi que celui de sortie. Pour la gestion des ressources, la variable $charge$ est calculée. Suivant le modèle générique de DDS proposé dans la figure 2.3, nous avons inclus les séjours hospitaliers précédents du patient à savoir, le nombre d'admission dans les séjours précédents et la DDS antérieure. De plus, pour quantifier la DDS par rapport à l'unité médicale et au motif d'hospitalisation, nous avons calculé la moyenne de la DDS par unité médicale et la moyenne de la DDS par motif d'hospitalisation. De plus, comme les codes « CIM 10 » peuvent représenter le diagnostic principal du patient, le diagnostic associé, le diagnostic relié et le diagnostic documentaire. Nous avons d'abord séparé entre les différents types des codes « CIM 10 » et avons garder que le code du diagnostic principal (DP), diagnostic associé (DAS) et diagnostic relié (DR). Au total, un ensemble de 16590 instances

sont alors prises en compte dans cette étude avec 25 variables. Cet ensemble de données est utilisé dans l'évaluation de nos processus de prédiction.

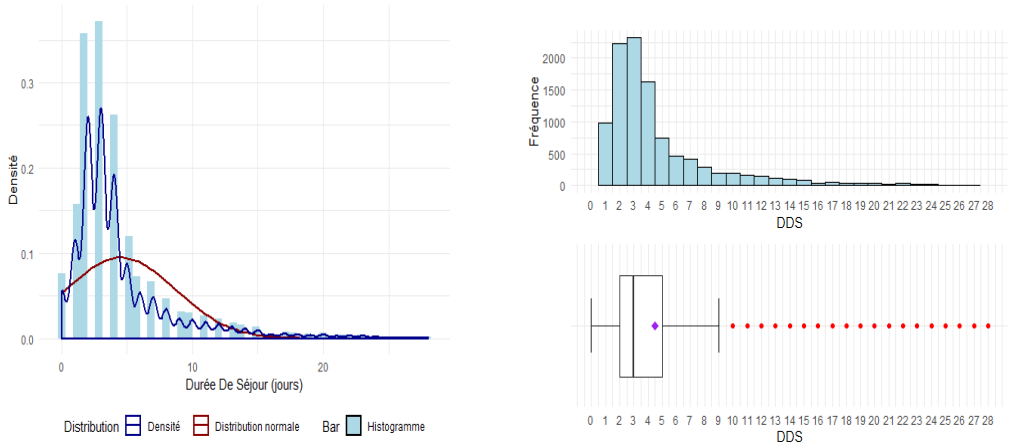
Nous allons maintenant exposer les résultats obtenus dans chaque phase de pré-traitements et illustrer les méthodes appliquées avec des exemples.

5.3 Prédiction des Durée De Séjour hospitalier

La prédiction des Durées De Séjour hospitalier (DDS) est calculée selon les étapes suivantes. Une première étape permet de collecter les données nécessaires issues du PMSI. Les données sont récupérées à partir des différentes tables de la base données. Elles sont transformées en forme tabulaire dont les lignes représentent les séjours hospitaliers et les colonnes les caractéristiques qui définissent ce séjour. Ensuite, les étapes d'analyse de données, de nettoyage, de sélection de variables, leur pré-traitement et la phase d'apprentissage sont successivement réalisées. Dans ce qui suit, nous exposons les résultats de chacune de ces étapes.

5.3.1 Analyse et nettoyage de données

Cette étape permet, d'une part, de décrire et de résumer le contenu des données, et, d'autre part, d'extraire des informations utiles et pertinentes à partir des données pour une prise de décision basée sur cette analyse. Dans notre étude, une analyse graphique et statistique uni-variée et bi-variée est menée. L'analyse uni-variée concerne chaque variable indépendemment des autres. Quant à l'analyse bi-variée, elle concerne la relation entre deux différentes variables. Nous avons calculé les coefficients statistiques selon le type de chaque variable. Pour les variables numériques, nous avons tracé l'histogramme, la densité ainsi que la boite à moustache. Cela permet de voir la distribution (symétrique ou asymétrique) des variables, de détecter la présence des données aberrantes et d'identifier les profils atypiques. Pour les variables catégorielles, les diagrammes à bâtons représentant les fréquences de chacune des modalités des variables catégorielles sont tracés. Ces diagrammes montrent les différentes modalités des variables catégorielles. La figure 5.1 montre le graphe de la variable *Durée De Séjour (DDS)* et la figure 5.2 montre la représentation graphique de la variable DDS.



(a) Densité de la DDS

(b) Histogramme et boite à moustache de la DDS

Fig. 5.1: Représentation graphique de la variable cible DDS.

Le graphe de la densité de la DDS et son histogramme montrent que la variable DDS est asymétrique. Cela signifie qu'il existe des valeurs moins présentes que d'autres dans la base de données. C'est un problème très fréquent dans les données médicales qui impacte négativement la performance des modèles de prédiction. C'est pour cela que nous avons proposé dans un premier temps, de regrouper les valeurs de la DDS par catégorie et considérer la prédiction comme un problème de classification. De plus, nous avons proposé une pondération de la fonction de perte dans le cadre de la régression en se basant sur la fréquence des valeurs de la DDS.

Dans notre étude, nous avons sélectionné 4 unités médicales différentes dont le service de cardiologie, le service de médecine polyvalente, le service de pédiatrie et le service néonatologie. Le graphe de la figure 5.2 montre que 40% des données concernent le service de cardiologie, 25% celles du service néonatologie, 23% celles du service de médecine polyvalente et enfin 18% du service de pédiatrie. Le choix de ce type d'unités médicales se justifie par la présence d'une grande variété dans les profils patients admis dans ces unités (nouveau né, adultes), différents motifs d'hospitalisation et aussi dans l'allocation des ressources.

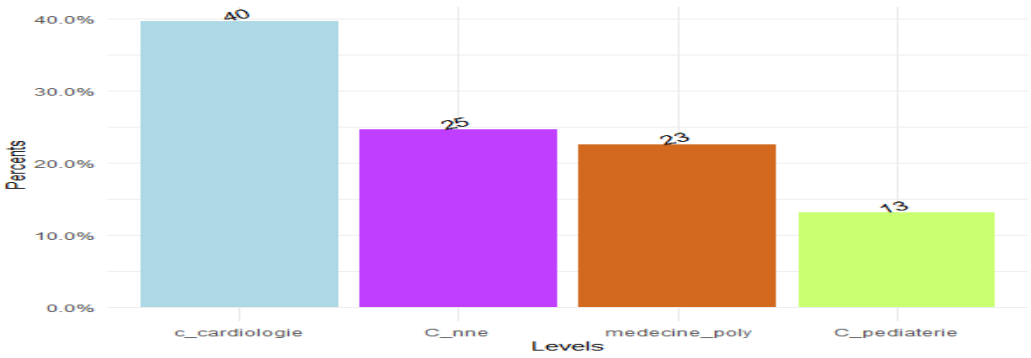


Fig. 5.2: Répartition des données en fonction du type de l'unité médicale.

Nous avons également effectué une analyse bi-variée sur les variables de l'ensemble des données afin d'en déduire la relation qui existe entre elles. L'exemple de la figure 5.3 illustre la distribution de la DDS par rapport à la variable *unité médicale*. Nous remarquons que la distribution de la DDS diffère d'une unité médicale à l'autre. Ceci confirme que la DDS dépend du type de l'unité médicale. C'est un résultat qui est confirmé dans la littérature. D'autre part, la figure montre la présence d'un grand nombre de données aberrantes pour chaque unité médicale. C'est une caractéristique connue des données médicales.

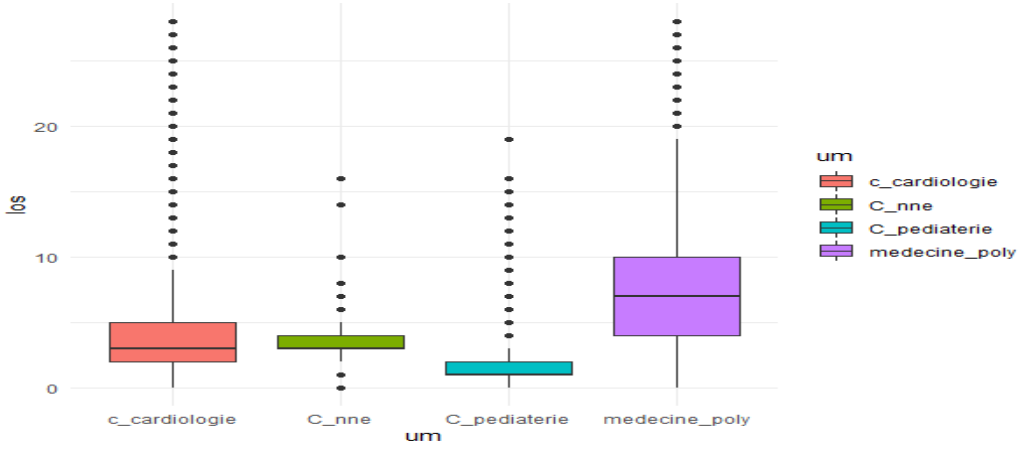


Fig. 5.3: Distribution de la DDS par unité médicale.

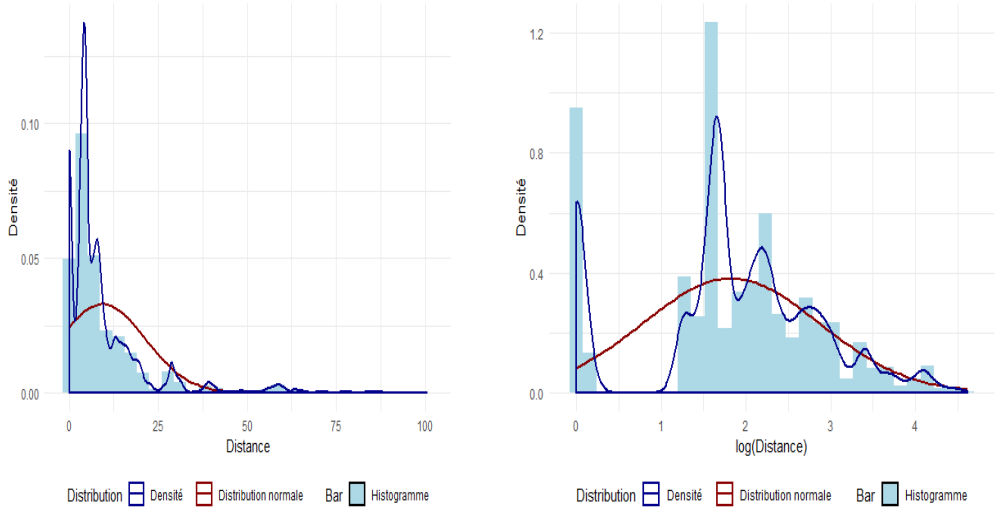
L'analyse de données a également permis de détecter les variables pour lesquelles les données sont manquantes. Dans notre cas, la variable $provenance$ comporte 49% de données manquantes. Nous avons remplacé ces valeurs par la valeur la plus fréquente de cette variable dans la base de données. De plus, nous avons appliqué une transformation logarithmique sur les variables qui ont une forte asymétrie. Le tableau 5.3 montre les résultats de cette transformation sur l'ensemble des variables numériques. Nous avons appliqué cette transformation sur les deux variables $distance$ et *antécédents UM*. Nous n'avons pas appliqué cette transformation sur la variable DDS étant donné qu'elle représente notre variable cible.

Variable	*asymétrie*	*interprétation*	*asymétrie Log(variable)*	*interprétation*
âge	-0,503	faible	-0,503	faible
moyenne DDS par DP	0,928	moyenne	0.928	moyenne
moyenne DDS par UM	0,916	moyenne	0.916	moyenne
distance	2,776	forte	-0.179	faible
antécédents UM	2,929	forte	0.877	moyenne
antécédents DDS	3,142	forte	0.477	faible

Tab. 5.3: Coefficient d'asymétrie des variables numériques.

Le coefficient d'asymétrie de la variable *âge* vaut -0,503 ce qui signifie que cette variable a une faible asymétrie. La variable $distance$ possède une forte asymétrie avec un coefficient d'asymétrie égal à 2,776. Après transformation logarithmique de la variable $distance$, son coefficient d'asymétrie est recalculé et il vaut -0,179.

La phase d'analyse de données résume le contenu de l'ensemble des données utilisées, le type de chaque variable, la présence des données aberrantes ainsi que les données manquantes. Elle donne une vue globale sur la relation existante entre les différentes variables dont la DDS. La figure 5.4 montre l'impact de la transformation logarithmique sur la variable *distance*.



(a) Distribution de la variable distance. **(b)** Distribution de la variable log (distance).

Fig. 5.4: Transformation logarithmique sur la variable distance

Le graphe 5.4a montre la distribution de la variable $distance$ avant transformation. Cette distribution est asymétrique. Nous remarquons sur le graphe 5.4b que la courbe de la distribution tend plus vers une courbe normale et donc moins asymétrique que la précédente.

Une fois que les données sont analysées et nettoyées, nous procédons à la sélection du sous-ensemble de données qui sera utilisé dans les algorithmes d'apprentissage automatique.

5.3.2 Sélection de variables

L'étape de sélection des variables a pour but de réduire l'ensemble des variables. Nous exposons dans cette section, pour les deux modèles de prédiction de DDS présentés dans ce mémoire, les sous-ensembles de variables.

Le modèle statique de prédiction de DDS concerne la prédiction au moment de l'admission du patient. A cet instant, nous disposons uniquement des variables liées aux conditions d'admission du patient, ses informations démographiques ainsi que le motif d'hospitalisation représenté par le Diagnostic Principal (DP). L'ensemble des variables au moment d'admission du patient englobe , les variables suivantes : ***ID séjour global***, ***ID séjour UM***, ***ID patient***, ***Unité Médicale (UM)***, ***genre***, ***distance***, ***âge***, ***date d'entrée***, ***mode d'admission***, ***diagnostic Principal (DP)***, ***moyenne de la DDS par Unité Médicale (UM)***, ***moyenne de la DDS par DP***, ***charge***, ***antécédents UM***, ***antécédents DDS*** et ***les antécédents médicaux***.

Les variables ***ID séjour globale***, ***ID séjour UM*** et ***ID patient*** sont éliminées car elles servent uniquement à l'identification des séjours et la collecte des données. Nous avons utilisé une combinaison de méthodes de sélection de variables pour définir d'une part, les variables corrélées avec la DDS, et d'autre part, celles qui sont corrélées entre elles. L'objectif étant de garder que les variables en forte relation avec la variable dépendante (DDS) et d'éliminer d'éventuelles liaison existantes entre les variables indépendantes elles-mêmes. En premier, nous avons tracé la matrice de corrélation du coefficient de Spearman des variables numériques en incluant la variable DDS. La figure 5.5 montre les résultats obtenus dans le cadre de la prédiction statique de DDS.

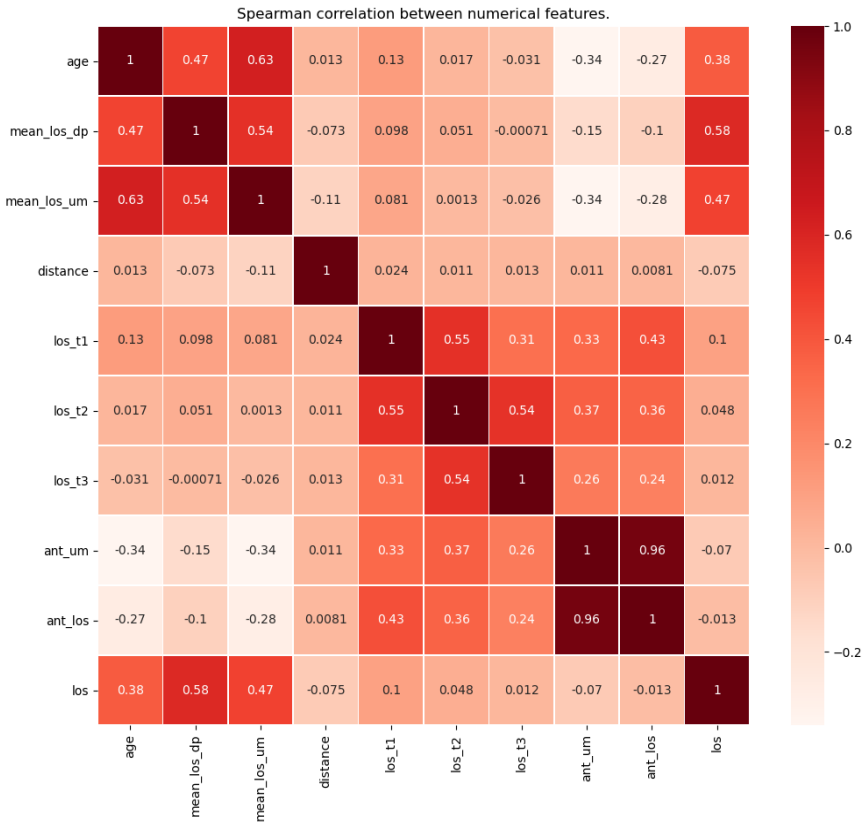


Fig. 5.5: Corrélation de Spearman des variables numériques.

A partir de la figure 5.5, nous concluons que la DDS est corrélée aux variables ***âge***, ***sa moyenne par DP*** et ***sa moyenne par UM***. De plus, la variable ***âge*** est corrélée aux deux variables ***moyenne de DDS par DP*** et ***moyenne de DDS par UM***. Aussi, les deux variables ***moyenne de DDS par DP*** et ***moyenne de DDS par UM*** sont positivement corrélées. Nous avons ainsi sélectionné parmi l'ensemble des variables numériques, ***l'âge du patient*** et ***la moyenne de DDS par DP*** en se basant sur le coefficient de Spearman. A l'aide de l'indicateur de Theil, nous remarquons que les variables ***admission mode*** et ***provenance*** sont fortement corrélées. Comme le

montre la figure 5.6. D'autre part, en nous appuyant sur l'information mutuelle, nous avons sélectionné l'ensemble des variables suivantes : ***DP***, ***âge***, ***moyenne de DDS par DP***, ***moyenne de DDS par UM***, ***UM***, ***distance*** et ***antécédents DDS***.

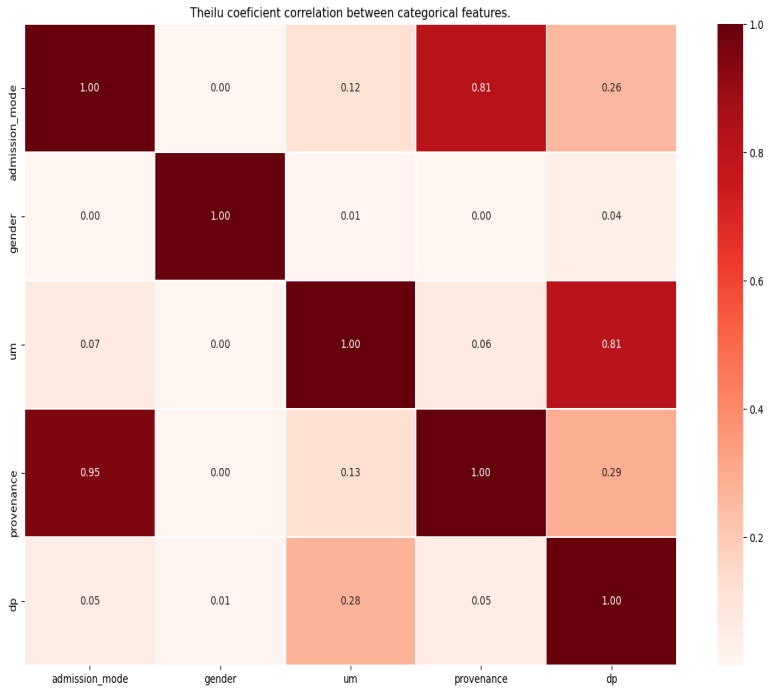


Fig. 5.6: Relation entre les variables catégorielles.

Le sous-ensemble de variables retenues pour les modèles d'apprentissage dans le cadre de la prédiction statique de DDS est résumé dans le tableau 5.4.

Variable	***Type***
âge	numérique
moyenne DDS par DP	numérique
UM	catégoriel
distance	numérique
Admission mode	catégoriel
antécédents DDS	numérique
DP	catégoriel

Tab. 5.4: Sous-ensemble des variables retenues pour le modèle statique de prédiction de DDS.

Les variables numériques sélectionnées sont standardisées. Les variables catégorielles sont codifiées à l'aide de la méthode « *one-hot-encoding* ». Nous présentons dans le tableau 5.5 un exemple de l'application de la méthode « *one-hot-encoding* » sur la variable ***UM*** qui représente l'unité médicale.

Modalités	*UM cardiologie*	*UM pédiatrie*	*UM médecine polyvalente*	*UM néonatologie*
cardiologie	1	0	0	0
pédiaterie	0	1	0	0
médecine polyvalente	0	0	1	0
néonatologie	0	0	0	1

Tab. 5.5: Exemple d'application de la méthode « *one-hot-encoding* » sur la variable unité médicale.

Concernant la sélection de variables pour le modèle séquentiel de prédiction de DDS, nous avons implémenté les mêmes techniques de sélection de variables. Pour rappel, le modèle séquentiel de prédiction de DDS explore les nouvelles données disponibles après l'admission du patient. Dans le contexte des données issues du PMSI, ces nouvelles données décrivent la réalisation d'un acte médical. Les actes médicaux étant codifiés à l'aide de la Codification Commune des Actes Médicaux ils, doivent êtres transformés en valeurs numériques pour que les valeurs des actes médicaux soient supportées par les algorithmes d'apprentissage automatique. Un patient peut subir un ou plusieurs actes médicaux tout au long de son séjour hospitalier. Par exemple, un patient peut subir une chirurgie, un scanner ou aucun de ces actes médicaux. La méthode de codification de ces actes a été présenté dans le chapitre 4.

Nous avons comparé différents algorithmes d'apprentissage automatique dont le Xgboost, le Random Forest et le Gradient Boosting Model. Nous avons également proposé une autre technique de codification des données pour les convertir en série temporelle et utiliser par la suite les réseaux de neurones récurrents (LSTM) dans la phase de prédiction. Nous présentons dans le tableau 5.6, les variables en entrée au modèle séquentiel de prédiction de DDS.

Variable	***Type***
âge	numérique
moyenne DDS par UM	numérique
UM	catégoriel
distance	numérique
Admission mode	catégoriel
antécédents DDS	numérique
DP	catégoriel
DDS courante	numérique
Antécédents UM	numérique
Actes CCAM	catégoriel multivalué

Tab. 5.6: Sous-ensemble de variables pour le modèle séquentiel de prédiction de DDS.

A ce stade, l'ensemble des données est défini, les variables numériques sont standardisées et les variables catégorielles et catégorielles multivaluées sont codifiées. Nous passons à l'implémentation de la phase d'apprentissage automatique et d'opti-

misation d'hyper-paramètres. Dans ce qui suit, nous exposons les différents hyper-paramètres que nous avons ajusté dans notre réalisation.

5.3.3 Ajustement des hyper-paramètres

Nous avons employé les méthodes de recherche bayésienne pour l'optimisation de chaque hyper-paramètre. Pour cela, nous avons défini l'espace de recherche des valeurs des hyper-paramètres pour chaque algorithme. Décrivons l'ensemble de ces hyper-paramètres. Les valeurs de ces hyper-paramètres sont variées et à la fin de cette étape la combinaison qui aboutit au meilleures performance de l'algorithme est sauvegardée pour l'utiliser dans l'évaluation des modèles.

- Random Forest :
  - N-estimators : est le nombre d'arbres de décision utilisé dans l'algorithme.
  - Max-depth : est la profondeur maximale de l'arbre.
  - min-samples-split : est le nombre minimal d'instances requis pour diviser un nœud interne.
  - min-samples-leaf : est le nombre minimal d'instances requis pour être à un nœud de feuille.
- Gradient Boosting Model :
  - N-estimators : est le nombre d'arbres de décision utilisé dans l'algorithme.
  - min-samples-leaf : est le nombre minimal d'instances requis pour être à un nœud de feuille.
  - min-samples-split : est le nombre minimal d'instances requis pour diviser un nœud interne.
  - max-depth : est la profondeur maximale de l'arbre.
  - learning-rate : est le contrôle sur la contribution de chaque arbre dans la prédiction.
  - subsample : est le pourcentage d'instances utilisé pour l'apprentissage des arbres de décision individuels.
- Xgboost :
  - eta : est le contrôle sur la contribution de chaque arbre dans la prédiction.
  - num-boost-round : est le nombre d'arbres de décision utilisé dans l'algorithme.
  - max-depth : est la profondeur maximale de l'arbre.
  - min-child-weight : est la somme minimale du poids d'une instance nécessaire dans un nœud.

- gamma : est la valeur minimale de la fonction de perte requise pour partitionner un nœud.
- subsample : est le pourcentage d'instances utilisé pour l'apprentissage des arbres de décision individuels.
- colsample-bytree : est le contrôle sur la partition des variables (colonnes).

- LSTM :
  - hidden-dim : est le nombre de neurones dans chaque couche du réseau.
  - Dropout : est la régularisation pour prévenir le sur-apprentissage.
  - learning-rate : est le contrôle sur la contribution de chaque arbre dans la prédiction.
  - batch-size : est le nombre d'instances utilisées par le réseau de neurones dans l'apprentissage du modèle.

Chaque hyper-paramètre de chacun des algorithmes cités ci-dessus aura un intervalle de valeurs dans lequel nous recherchons la meilleure combinaison par la méthode de recherche bayésienne.

5.4 Évaluation des modèles de prédiction de DDS

Cette section concerne les résultats obtenus suite à l'implémentation des différents processus pour l'apprentissage automatique décrits dans le chapitre 3 et le chapitre 4. En premier, nous présentons les résultats de la classification dans le cadre de la prédiction des DDS à savoir, la classification par les techniques d'apprentissage supervisé puis en combinant les deux techniques d'apprentissage supervisé et d'apprentissage non supervisé. Ensuite, les résultats de la régression sont présentés. Nous exposons les résultats du modèle statique de prédiction de DDS en tenant compte des données disponibles au moment de l'admission du patient et les résultats du modèle séquentiel de prédiction de DDS en intégrant de nouvelles données disponibles tout au long du séjour hospitalier.

5.4.1 Classification

La classification permet la prédiction d'une valeur catégorielle à partir d'un ensemble de variables indépendantes. Nous avons défini des intervalles pour convertir la variable DDS d'une valeur numérique en une valeur catégorielle. Nous avons comparé les algorithmes d'apprentissage automatique suivants : le Random Forest, le Xgboost et le Gradient Boosting model pour la classification. Ensuite, nous les avons évalué en employant les mesures de performance suivantes : la précision, le rappel, le score

F_1 et la précision globale du modèle (accuracy). De plus, le score nommé Cohen Kappa est utilisé pour prendre en compte l'aspect de classe déséquilibrées existant dans les données médicales réelles.

5.4.1.1 Apprentissage supervisé

La variable cible DDS est divisée en 3 groupes : DDS courte, DDS moyenne et DSS longue. Le tableau 5.7 présente les résultats obtenus pour chaque classe des différentes mesures utilisées, la précision, le rappel et le score F_1.

	classes	*Xgboost*	*Random Forest*	*Gradient Boosting Model*
Précision	DDS courte	0,86	0,83	0,81
	DDS moyenne	0,83	0,69	0,72
	DDS longue	0,89	0,69	0,75
Rappel	DDS courte	0,72	0,52	0,59
	DDS moyenne	0,86	0,70	0,76
	DDS longue	0,95	0,87	0,85
F_1 ***score***	DDS courte	0,78	0,63	0,68
	DDS moyenne	0,84	0,69	0,74
	DDS longue	0,92	0,77	0,80

Tab. 5.7: Mesures d'évaluation pour la classification des DDS.

Sur l'ensemble des algorithmes testés, l'algorithme Xgboost pour la classification a donné les meilleurs résultats. En effet, sur l'ensemble des 3 classes définies, les différentes mesures de la précision, du rappel et du score F_1 sont les plus élevés et sont proches de 1. Le score F_1 étant supérieur à 0,89, nous considérons la performance de cet algorithme dans la prédiction de DDS satisfaisante. Concernant les deux algorithmes Random Forest et Gradient Boosting model (GBM), nous avons remarqué que l'algorithme Gradient Boosting Model a légèrement dépassé le Random Forest en comparant les différentes mesures de précision, rappel et le score F_1 de chacune des classes. Comme les deux algorithmes Xgboost et GBM se basent sur la technique du boosting, nous notons que les méthodes ensembliste d'arbres de décision se basant sur la technique du boosting sont plus efficaces que celles se basant sur le bagging dans le cadre de cette étude.

Pour approfondir l'analyse de nos résultats, nous présentons dans le tableau 5.8 la précision globale des modèles ainsi que le score Cohen Kappa. Le Xgboost possède la plus grande valeur de ces deux mesures, avec une précision égale à 0,86 et un Cohen Kappa égal à 0,78. Ceci confirme que l'algorithme Xgboost est plus efficace dans la classification des DDS que le Random Forest et le GBM.

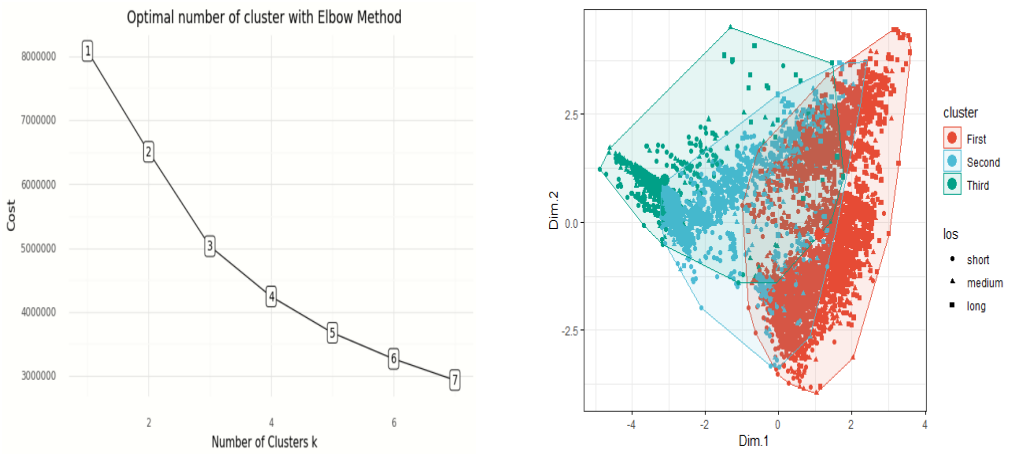
Algorithmes	*Gradient Boosting Model*	*Random Forest*	*Xgboost*
Précision globale	0,75	0,71	0,86
Cohen Kappa	0,61	0,55	0,78

Tab. 5.8: Évaluation du modèle de prédiction de DDS : classification sans clustering (précision globale et Cohen Kappa.

Une autre piste explorée est d'identifier des clusters ou groupes homogènes en utilisant les caractéristiques de la DDS. Chaque instance de donnée est affectée à un groupe. Chaque groupe possède une étiquette. Cette étiquette est ajoutée comme une nouvelle variable à l'ensemble de données en entrée des différents algorithmes d'apprentissage supervisé. Les résultats sont présentés dans la section suivante.

5.4.1.2 Combinaison de l'apprentissage supervisé et non supervisé

D'abord, nous présentons les résultats de l'application de la méthode du coude et de l'algorithme d'apprentissage non supervisé K-prototype. La figure 5.7 illustre ces résultats.



(a) Méthode du coude. **(b)** Clustering avec l'algorithme K-prototype.

Fig. 5.7: Apprentissage supervisé : K-prototype sur données PMSI.

La méthode du coude a révélé que la meilleure valeur du paramètre K de l'algorithme K-prototype est égale à 3. Ce paramètre K représente le nombre de partitions de l'ensemble de données utilisées. Nous avons appliqué le clustering avec K égal à 3. Nous distinguons 3 groupes colorés en rouge, bleu et vert respectivement. Nous remarquons que ces groupes ne sont pas parfaitement séparés et qu'il existe des chevauchements entre les clusters. Pour la suite, le résultat de ce clustering est injecté dans les différents modèles d'apprentissage automatique supervisé cités ci-dessus.

Le tableau 5.9 montre les mesures de performance (précision, rappel et le scrore F_1) en combinant les deux techniques d'apprentissage supervisé et non supervisé.

	classes	*Xgboost*	*Random Forest*	*Gradient Boosting Model*
Précision	DDS courte	0,84	0,81	0,82
	DDS moyenne	0,83	0,72	0,71
	DDS longue	0,88	0,74	0,73
Rappel	DDS courte	0,72	0,59	0,58
	DDS moyenne	0,85	0,74	0,74
	DDS longue	0,95	0,86	0,85
F_1 ***score***	DDS courte	0,78	0,68	0,68
	DDS moyenne	0,84	0,73	0,73
	DDS longue	0,91	0,80	0,79

Tab. 5.9: Mesures d'évaluation pour la classification des DDS : combinaison de l'apprentissage supervisé et non supervisé.

Sur l'ensemble des algorithmes, le Xgboost est celui qui obtient la meilleure précision, rappel et score F_1 des classes de la variable DDS. Le Random Forest est légèrement plus performant que le GBM dans ce cas.

Le tableau 5.10 montre que la précision globale du Xgboost est également plus élevée et vaut 0,85 que celle du Random Forest et le GBM qui sont égales à 0,70 et 0,74 respectivement. Sur le critère de la précision globale, nous remarquons aussi que le GBM est plus précis que le Random Forest.

Algorithmes	*Gradient Boosting Model*	*Random Forest*	*Xgboost*
Accuracy	0,74	0,70	0,85
Cohen Kappa	0,60	0,61	0,77

Tab. 5.10: Précision globale et Cohen Kappa du modèle de prédiction de DDS : combinaison de l'apprentissage supervisé et non supervisé.

Le comportement de ces algorithmes dans la prédiction des DDS est similaire en utilisant les algorithmes d'apprentissage supervisé ou en combinant l'apprentissage supervisé et l'apprentissage non supervisé. Le Xgboost a montré son efficacité en termes de précision, rappel, score F_1 et précision globale dans les deux cas. Il est alors intéressant de l'implémenter et d'utiliser ce modèle dans la prédiction en temps réel de la DDS dans les établissements de soins.

5.4.2 Régression

La régression est un problème qui consiste à trouver une relation entre des variables indépendantes numériques, catégorielles ou catégorielles multivaluées et une variable dépendante numérique représentée par la DDS. Nous avons comparé 4 différents algorithmes d'apprentissage.

Dans le cadre du modèle statique de prédiction de DDS, les algorithmes suivants : le Random Forest, le GBM, le Xgboost avec la moyenne des écarts carré (MSE) comme fonction de perte et le Xgboost avec la MSE pondérée comme fonction de perte sont comparés.

Concernant le modèle séquentiel de prédiction de DDS, le Random Forest, le GBM, le Xgboost et les Réseaux de Neurones Récurrents de type LSTM sont comparés. Nous présentons dans ce qui suit l'évaluation de ces modèles en nous basant sur la moyenne absolue des écarts (MAE), la variance expliquée par le modèle (R_2) et cette variance ajustée (Adjusted R_2).

5.4.2.1 Modèle statique de prédiction de DDS

La DDS est définie par le nombre de jours du séjour du patient dans l'établissement de soins. L'objectif principal de la prédiction de DDS est de savoir si le patient "respecte" sa DDS réelle par rapport à celle prédite. Nous cherchons donc à connaître l'écart entre une valeur prédite et une valeur réelle. La moyenne absolue des écarts est utilisée pour l'évaluation de nos modèles de prédictions. De plus, pour analyser le pouvoir des variables indépendantes à expliquer la variables dépendante DDS, nous avons calculé les deux mesures représentées par le R_2 et le R_2 ajusté.

Nous avons calculé le pourcentage des écarts égaux à 0 jour, 1 jour, 2 jours, 3 jours, 4 jours et 5 jours ou plus. Pour chaque algorithme d'apprentissage supervisé, nous avons tracé ces pourcentages à l'aide d'un diagramme à barres.

Le tableau 5.11 et la figure 5.8 illustrent les résultats obtenus.

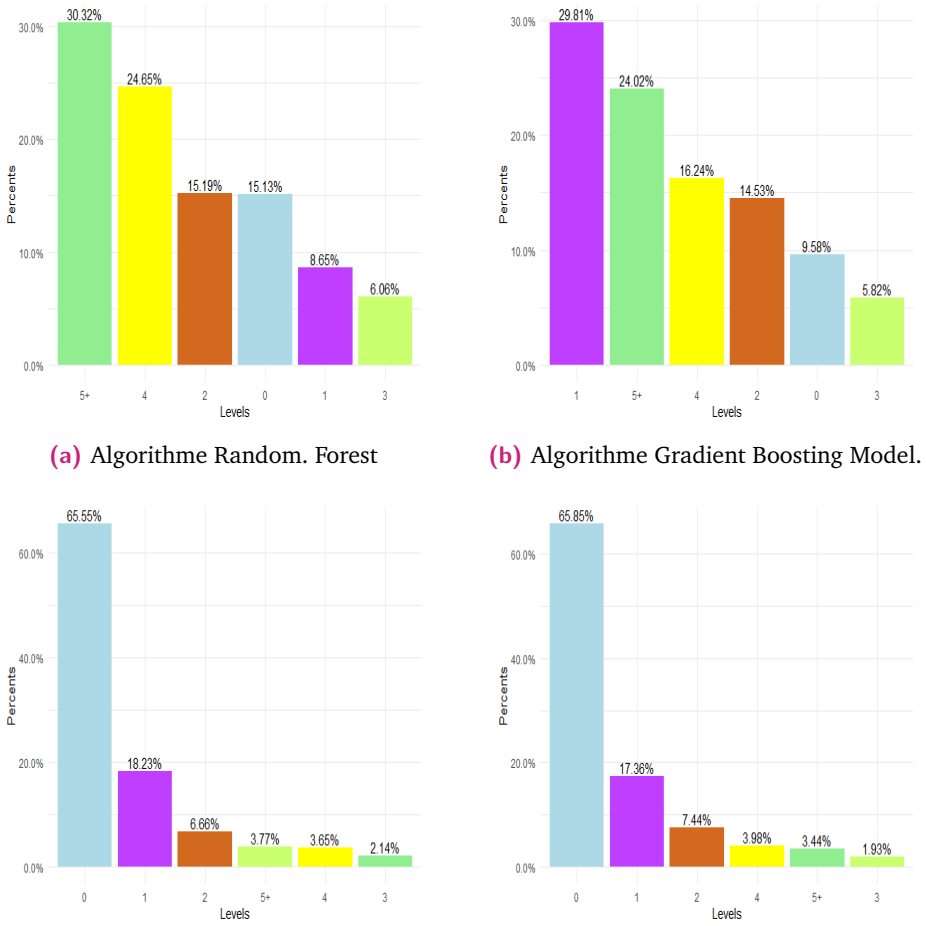
Algorithmes	***MAE***	R_2	R_2 ***ajusté***
Xgboost (MSE)	0,79	0,88	0,88
Xgboost (MSE pondérée)	0,79	0,88	0,87
Gradient Boosting Model	2,30	0,55	0,55
Random Forest	2,21	0,58	0,58

Tab. 5.11: Évaluation du modèle statique de prédiction de DDS : régression.

Le tableau 5.11 montre que l'algorithme Xgboost est le plus performant en comparant les mesures suivantes : MAE, R_2 et le R_2 *ajusté* parmi Random Forest et le GBM. Ceci est justifié par le fait que la valeur de la MAE est la plus petite et celle du R_2 et celle du R_2 *ajusté* sont plus élevée. Ceci est valable avec une fonction de perte égale à la *MSE* et celle égale à une *MSE pondérée*. Sur l'ensemble des données de test, en moyenne l'écart en valeur absolue entre la valeur prédite et la valeur réelle est égal à 0,79 jours. La pondération de la fonction de perte n'a pas eu d'effet sur les résultats de la MAE du modèle de prédiction avec le Xgboost. Plus cette valeur est

proche du 0, meilleure est la prédiction. Nous estimons que la valeur de la MAE est satisfaisante dans le cadre de notre étude et prenant en compte la complexité des données médicale du monde réel. Les coefficients R_2 et R_2 ajusté valent 0,88 tous les deux. Ceci veut dire que les variables indépendantes expliquent la variable DDS à 88%. Ce qui est un bon résultat. Le Random Forest, quant à lui, est plus précis que le GBM avec une MAE égale à 2,21 jours et R_2 et R_2 ajusté valant 0,58. En moyenne, l'erreur absolue est égale à 2,30 jours en utilisant le GBM et la variance expliqué est égale à 0,55. Afin de mieux comprendre ces résultats, nous présentons les erreurs de prédiction sous forme de pourcentage dans la figure 5.8. Chaque barre du graphe possède une couleur qui fait référence à la valeur de l'écart entre la valeur prédite et la valeur réelle comme suit :

- Bleu : écart = 0 jour.
- Violet : écart = 1 jour.
- Orange : écart = 2 jours.
- Vert clair : écart = 3 jours.
- Jaune : écart = 4 jours.
- Vert foncé : écart = 5 jours ou plus.



(a) Algorithme Random. Forest

(b) Algorithme Gradient Boosting Model.

(c) Algorithme Xgboost.

(d) Algorithme Xgboost avec pondération du MSE.

Fig. 5.8: Évaluation des modèles statiques de prédiction de DDS.

Le meilleur résultat est celui pour lequel la réalité coïncide avec la prédiction (écart =0). Pour l'algorithme Xgboost, et pour les deux métriques de la fonction de perte à savoir la *MSE* et la *MSE pondérée*, plus de 65% des données sont correctement prédites avec une erreur de prédiction égale à 0 jour. Concernant le Xgboost avec une *MSE* classique, un écart de 1 jour est noté sur l'ensemble de plus de 18% des données. En utilisant Xgboost avec une fonction de perte pondérée, plus de 17% des données ont un écart de prédiction de 1 jour. En terme de gestion hospitalière, une erreur de 1 jour est acceptable. Par conséquent, si nous tolérons une erreur de 1 jour, nous considérons le modèle avec une performance supérieure à 83% (65% + 18%). Pour le Random Forest et le GBM, le pourcentage des données avec un écart égal à 0 ou 1 jour est égal à 23,78% et 39,39% respectivement. Ceci reste loin des performances obtenues à l'aide du Xgboost. Nous concluons, que dans le cadre de la prédiction de la DDS comme valeur numérique, les méthodes ensembliste employant la technique du boosting sont plus efficaces en considérant les mesures de la MAE et le R_2 et le R_2 ajusté que celles employant la technique du bagging.

En résumé, les résultats que nous avons obtenu dans le cadre de la prédiction de la DDS en utilisant que les données disponibles au moment de l'admission, montrent que le Xgboost reste le meilleur algorithme à employer pour le modèle statique de prédiction de DDS.

5.4.2.2 Modèle séquentiel de prédiction de DDS

Le modèle séquentiel de prédiction de DDS tient compte des actes médicaux CCAM réalisés tout au long du séjour hospitalier. Nous avons codifié ces données de façon à préserver l'ordre de leur arrivée. Nous avons comparé les algorithmes cités auparavant en plus des réseaux de neurones récurrents de type LSTM afin d'exploiter leur propriété de gestion des données avec un aspect temporel. Nous avons utilisé les mêmes mesures d'évaluation que pour le modèle statique de prédiction de DDS à savoir la MAE, le R_2 et le R_2 ajusté. Pour les expérimentations de l'algorithme LSTM, nous avons défini la valeur du paramètre *time step* égale à 3 pour garder les informations concernant le patient des 3 séjours précédents. Nous exposons les résultats obtenus des différents modèles de prédiction dans le tableau 5.12. La figure 5.9 présente graphiquement les résultats obtenus et nous les commentons dans la suite.

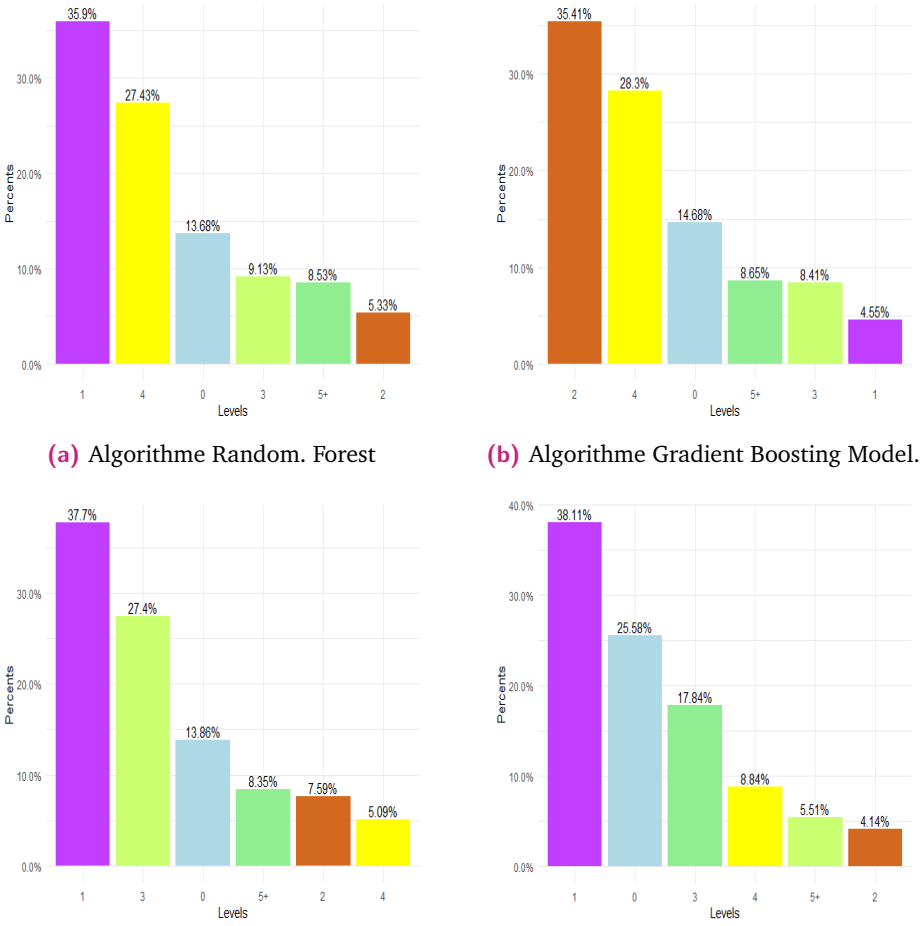
Algorithmes	***MAE***	R_2	***Adjuested*** R_2
Xgboost	1,56	0,57	0,56
Gradient Boosting Model	1,63	0,53	0,52
Random Forest	1,70	0,48	0,47
LSTM	2,47	0,06	0,05

Tab. 5.12: Évaluation du modèle statique de prédiction de DDS : régression.

L'analyse de la valeur de la *MAE* de chaque algorithme, montre que l'algorithme le plus performant est le Xgboost car la valeur de la MAE est la plus petite. En moyenne, la valeur absolue de l'écart entre la valeur prédite et la valeur réelle de la DDS est égale à 1,56 jours. Au deuxième rang, vient le GBM puis le Random Forest et finalement les LSTM. Avec le GBM, l'erreur absolue est égale à 1,63 jours. Avec le Random forest, elle est égale à 1,70 jours et 2,47 jours avec l'algorithme LSTM. En terme de variance expliquée par l'algorithme d'apprentissage automatique, le Xgboost reste aussi le meilleur algorithme en terme de performance avec une valeur de 57% pour le R_2 et 56% pour le R_2 ajusté.

La figure 5.9 montre que le nombre d'instances correctement prédites représente plus de 25% pour l'algorithme LSTM. Ce résultat est le meilleur pourcentage obtenu. En tolérant un écart de 1 jour, la précision des différents algorithmes est respectivement équivalente à 49,58%, 19,23%, 21,45%, 63,69% pour les algorithmes Random Forest, GBM, Xgboost et LSTM. En analysant ces résultats, nous concluons qu'en terme de

performance globale, l'algorithme LSTM est une piste intéressante à explorer pour améliorer les performances du modèle de prédiction. Au final, et pour conclure les différentes expérimentations menées en utilisant des données réelles issues du PMSI, nous discutons dans la section suivante les différents résultats obtenus en mettant en évidence le lien avec les différents modèles caractérisant la DDS dans un environnement hospitalier.



(a) Algorithme Random. Forest

(b) Algorithme Gradient Boosting Model.

(c) Algorithme Xgboost.

(d) Algorithme LSTM.

Fig. 5.9: Évaluation des modèles statiques de prédiction de DDS.

5.5 Discussion

Dans le cadre de la prédiction de la DDS avec des données réelles, les méthodes ensemblistes ont prouvé leur efficacité. En effet, les arbres de décision seuls permettent de capturer les relations non linéaires entre les données. De plus, nous avons remarqué que rassembler plusieurs algorithmes et les combiner améliore les performances en terme de précision et de taux d'erreurs de la prédiction. Ce résultat a été observé avec les deux techniques "bagging" et "boosting" implémentées dans

le Xgboost, le GBM et le Random Forest. De plus, nous avons exploré et démontré que les réseaux de neurones récurrents de type LSTM sont adapté à la prédiction des données médicales incrémentales.

Concernant les facteurs (démographiques, médicaux et administratifs) utilisés dans la prédiction de DDS que ce soit pour le modèle statique ou le modèle séquentiel, les recherches antérieures suggèrent d'utiliser des variables liées à une unité médicale particulière. Dans notre étude, nous avons utilisé des variables communes à plusieurs unités médicales (cardiologie, pédiatrie, médecine polyvalente et néonatologie). Ces facteurs sont démographiques telles que l'âge, le genre et l'adresse du patient. De plus, les informations médicales du patient sont considérées (motif d'hospitalisation, antécédents médicaux, complications, actes médicaux). Des informations administratives et économiques sont également ajoutées pour la gestion et l'organisation hospitalière. Nous avons également mené une étude et avons déduit de nouvelles variables liées aux besoins quotidiens des hôpitaux. Cet ensemble de variables a été analysé, nettoyé et traité pour une future utilisation comme entrée aux algorithmes d'apprentissage automatique.

Les méthodes de recherche bayésiennes ont permis d'ajuster les hyper-paramètres des différents algorithmes d'apprentissage automatique et la méthode de validation croisée pour la validation et la généralisation des modèles implémentés.

Comme les données utilisées dans cette étude proviennent des établissements de soins français, il est alors judicieux de privilégier des algorithmes interprétables et qui sont performants face à la complexité des données médicales tel que le Xgboost. Les résultats de cet algorithme sont facilement interprétables comme il se base sur un ensemble d'arbre de décision. De plus, le Xgboost a montré sa puissance face aux difficultés de la gestion des données médicales.

5.6 Environnement matériel et logiciel

Pour la mise en oeuvre de nos solutions, nous avons utilisé un microprocesseur ayant les caractéristiques suivantes :

- Un microprocesseur Intel(R) CoreTM i7 (8650U).
- Fréquence du CPU est de 2.11 GHz.
- Mémoire RAM 8.00 Go.
- Capacité disque dur 512 GO.
- Système d'exploitation Windows 10, 64 bits durant les tests.

Nous avons utilisé le langage de programmation Python, version 3.7. Ce langage est un langage de programmation orienté objet. Il existe plusieurs librairies sous python permettant la réalisation des algorithmes d'apprentissage automatique dont sklearn, xgboost, etc. Nous avons utilisé l'environnement de développement Pycharm.

La phase d'analyse de données a été réalisée à l'aide du langage de programmation R. R est un environnement logiciel libre pour l'analyse statistique des données. Le Système de Gestion de Base de Données (SGB) Postgress a été utilisé dans la phase de collecte de données pour gérer les bases de données relationnelles.

5.7 Conclusion

Dans ce chapitre, nous avons présenté les expérimentations que nous avons mené sur des données réelles issues du PMSI afin d'évaluer nos propositions.

D'abord, nous avons collecté les données nécessaires à partir du PMSI stockées dans les Systèmes d'Informations de Santé des hôpitaux. Ces données ont été extraites d'une base de données relationnelle et transformées au format tabulaire. Nous avons utilisé le modèle générique présenté dans la figure 2.3 du chapitre 2 pour compléter l'ensemble des données. Nous avons ensuite mis en place notre démarche méthodologique sur ces données. Différentes unités médicales ont été sélectionnées pour cette étude.

La première étape est l'analyse des données pour déduire des informations pertinentes et pour comprendre leur contenu. Après cette analyse, nous avons nettoyé les données en éliminant les données erronées et les données atypiques. Les données manquantes ont été traitées et les variables asymétriques ont été transformées. Cette mise en forme des données a été suivie d'une étape de sélection des variables représentatives. Pour cela, nous avons utilisé les coefficients statistiques. Les données représentatives sont celles qui sont corrélées à la DDS. Les corrélations entre les variables indépendantes ont été éliminées. Ensuite, l'étape de standardisation et de codification des variables a été réalisée. La dernière étape des traitements a été réalisée en utilisant les algorithmes d'apprentissage automatique supervisé suivant : Extreme Gradient Boosting (Xgboost), le Random Forest (RF), le Gradient Boosting Model (GBM), les réseaux de neurones récurrents de type LSTM. L'algorithme d'apprentissage non supervisé K-prototype a aussi été utilisé pour grouper les données.

L'étude empirique que nous avons effectué nous a permis de confirmer l'étude bibliographique sur les modèles caractérisant la DDS dans un environnement hospitalier.

De plus, en utilisant des données médicales réelles, nous avons montré les difficultés au-quelles nous devons faire face quant à la gestion de ce type de données. Les algorithmes d'apprentissage automatique ont montré leur efficacité dans la prédiction des DDS. Enfin, nous avons évalué les différents modèles proposés en se basant sur plusieurs mesures d'évaluation selon l'objectif de l'étude à savoir la classification ou la régression.

Conclusion générale

> *Change is the end result of all true learning*
>
> **— Leo Buscaglia**
> Écrivain, professeur d'éducation.

Résumé conclusif

Face à l'augmentation du nombre de patients dans les hôpitaux et à de nombreuses contraintes d'utilisation des ressources hospitalières, les établissements de soins sont toujours à la recherche d'une amélioration de la qualité des soins et de l'efficacité des services notamment en terme de gestion hospitalière et humaine. La Durée De Séjour hospitalier (DDS) est un indicateur d'évaluation des performances des hôpitaux. Dans cette thèse nous avons proposé différents modèles de prédiction des DDS qui se basent sur des données issues du Programme de Médicalisation des Systèmes d'Informations (PMSI) implémenté dans les établissements de soins Français. Ces données sont exploitées pour prédire la DDS au moment de l'admission du patient et pendant son séjour hospitalier. Nous avons montré dans cette thèse que les données stockées dans les Systèmes d'Informations Hospitalier (SIH) sont une source de connaissances importante en particulier pour la prédiction de DDS.

le travail présenté dans ce mémoire est essentiellement consacré à la proposition d'un modèle de prédiction de DDS. Cette proposition consiste dans un premier temps à prédire la DDS à partir des données disponibles lorsque le patient arrive à l'hôpital. Elle a été étendue de manière à prendre en compte également les données qui sont disponibles au cours du parcours du patient pendant son séjour. Cette proposition se base sur des techniques d'apprentissage automatique et de fouille de données. Nous avons étudié l'état de l'art pour cerner la problématique et proposer de nouvelles solutions pour la prédiction de la DDS.

Contributions

La Durée De Séjour hospitalier est une variable complexe qui dépend de nombreux facteurs. Ces facteurs sont liés à l'environnement dynamique de l'hôpital, à sa gestion, à son organisation, à l'état de santé du patient et son contexte social. Dans un premier temps, nous avons étudié l'ensemble des facteurs qui impactent la DDS. Ces facteurs caractérisent la DDS dans un milieu hospitalier et sont utilisés par la suite dans les modèles de prédiction. De ce fait, nous avons défini une approche méthodologique qui s'articule au tour des points suivants :

Définition d'un périmètre d'étude : la DDS a été définie par unité médicale ou par diagnostic du patient dans la littérature. Dans cette thèse, nous avons défini l'environnement hospitalier comme étant un ensemble de 4 unités médicales différentes : l'unité de cardiologie, l'unité de médecine polyvalente, l'unité de pédiatrie et l'unité

de néonatologie. Ces unités médicales se distinguent par rapport aux profils des patients, aux motifs d'hospitalisation et à leur gestion administrative et financière.

Modélisation générique de la DDS : une analyse approfondie des facteurs impactant la DDS a été réalisée. L'ensemble des facteurs démographiques (âge, sexe, situation familiale et adresse de résidence), médicaux (motif d'hospitalisation, antécédents médicaux et complications) et des informations administratives et financières (conditions d'admission et de sortie, type de remboursement et type de paiement) sont pris en compte. Nous avons regroupé les facteurs communs à diverses unités médicales et avons rajouté de nouvelles données en nous appuyant sur les besoins quotidiens des établissements de soins. Une modélisation générique de la DDS est proposé. Cette modélisation de la DDS est l'entrée aux processus de prédiction. Les processus de prédiction s'appuient sur les méthodes d'apprentissage automatique et de fouille de données.

Modèle statique de prédiction de DDS : nous avons présenté un modèle de prédiction de DDS au moment de l'admission du patient. Plusieurs techniques d'apprentissage automatique sont alors explorées dont la classification et la régression en se basant sur la définition de la DDS (catégorielle ou numérique respectivement). Une comparaison entre différents algorithmes d'apprentissage automatique est établie. Dans le cadre de la classification, nous avons proposé une solution qui se base d'abord sur des algorithmes d'apprentissage supervisé (Random Forest, Gradient Boosting Model et le Xgboost). Ensuite, nous avons combiné les techniques d'apprentissage supervisé et non supervisé. Regardant la nature des données médicales hétérogènes, l'algorithme K-prototype a été utilisé dans l'apprentissage non supervisé. Dans le cadre de la régression, en plus des algorithmes employés dans la classification , une pondération de la fonction de perte de l'algorithme Xgboost est proposée afin de palier au problème de prédiction des valeurs de DDS à basse fréquence. Pour chaque méthode proposée, une analyse approfondie des données médicales utilisées est mise en place suivie d'une étape de nettoyage et de pré-traitement. L'étape de pré-traitement des données est primordiale dans le domaine médical.

Modèle séquentiel de prédiction de DDS : Les données médicales ne sont pas toutes disponibles au moment de l'admission du patient. Nous avons proposé une méthode qui intègre les données disponibles au fur et à mesure que le patient évalue durant son séjour. Ce modèle permet d'affiner la valeur de DDS initialement prédite au moment de l'admission du patient en considérant les nouvelles informations. Pour mettre en place ce modèle, nous avons proposé une méthode de codification et structuration des donnés. Cette méthode respecte la chronologie de l'arrivée des données. Les mêmes algorithmes d'apprentissage automatique que ceux utilisés précédemment ont été implémentés et comparés. De plus, les réseaux de neurones

récurrents de type LSTM (Long-Short-Term-Memory) ont été implémenté en rajoutant la dimension temporel aux données. Cet algorithme a été implémenté pour étudier les propriétés séquentielles des données.

Nos propositions ont été expérimentées sur des données réelles issues du Programme de Médicalisation des Systèmes d'Informations (PMSI). Ce programme est utilisé dans les établissements de soins Français pour contrôler les contraintes et attribuer les allocations budgétaires aux différentes unités de soins. La prédiction de la DDS permet dans ce cas, de planifier les activités des soins et optimiser les ressources matérielles, logiciels et humaines.

Les résultats obtenus ont montré que les algorithmes d'apprentissage automatique sont performants en terme de précision et de taux d'erreurs dans la prédiction des DDS en milieu hospitalier. Les algorithmes issus des méthodes ensemblistes qui se basent sur les arbres de décision sont adaptés à la problématique de la prédiction de DDS. De part, les résultats de ces algorithmes sont faciles à interpréter, et, d'autre part, ils sont capable de gérer les difficultés rencontrées dans les données médicales.

Dans le cadre du modèle statique de prédiction de DDS, l'algorithme Extreme Gradient Boosting (Xgboost) a surpassé les algorithmes Random Forest (RF) et le Gradient Boosting Model (GBM) que ce soit pour les modèles de classification ou de régression. Dans la classification, en utilisant les algorithmes d'apprentissage supervisé, la meilleure précision vaux 86%. En combinant les algorithmes d'apprentissage supervisé et non supervisé, cette précision est égale à 85% en utilisant le Xgboost et le K-prototype.

Pour la régression, le taux d'erreur minimal est obtenu à l'aide de l'algorithme Xgboost et il est égal en moyenne absolue à 0.79 jour. La pondération de la fonction de perte n'a pas eu d'impact sur ce taux.

Concernant le modèle séquentiel de prédiction de DDS, en moyenne absolue, le meilleur modèle a donné une erreur de 1,56 jours en utilisant le Xgboost. Les réseaux de neurones récurrents de type LSTM (Long-Short-Term-Memory) ont été explorés et adaptés à la prédiction des DDS avec des données incrémentales.

L'approche que nous avons adopté présente des avancées dans la prédiction de la DDS dans un environnement hospitalier Elle intègre des données incrémentales et évolutives. Elle présente également des points à améliorer dans les futurs travaux. Ces points font partie de nos perspectives de recherche.

Perspectives

Le domaine de la prédiction de durées de séjour en milieu hospitalier est un domaine en plein essor. La prédiction s'implique dans l'organisation des établissements de soins mais aussi dans la planification de leurs activités et l'optimisation de leurs ressources. Notre travail réponds à une panoplie de questions dans ce domaine. Plusieurs pistes prometteuses ont également été identifiées.

Concernant le périmètre d'étude de la DDS, il serait intéressant d'explorer d'autres types d'unités médicales pour élargir l'étude. Les unités médicales de type urgence ou ambulatoire sont des exemples d'applications de la prédiction de DDS qui apportent des améliorations considérables pour l'organisation des établissements de soins.

Concernant la modélisation de la DDS, il serait intéressant d'impliquer les experts du domaine médical dans la sélection des facteurs qui impactent la DDS et leur validation.

Regardant la complexité des données médicales, l'expert médical doit aussi être impliqué dans l'analyse des profils atypiques pour les détecter et les distinguer des données aberrantes.

Dans la phase de sélection de variables, d'autres techniques qui se basent sur l'apprentissage automatique sont à explorer. Les arbres de décision, le Random Forest peuvent être utilisés dans cette phase.

Afin d'améliorer les performances des algorithmes d'apprentissage automatique, une piste serait d'enrichir l'ensemble de données utilisé dans l'apprentissage et celui utilisé dans la validation des processus de prédiction. L'ajout des nouvelles données et qui présentent une richesse dans les informations permet aux algorithmes d'apprentissage automatique d'apprendre sur plus de cas et de ce fait, ils aboutissent à des résultats plus précis.

Dans le modèle séquentiel de prédiction de DDS, une piste serait d'intégrer d'autres types de données incrémentales telles que les complications médicales qui peuvent survenir pendant le séjour hospitalier. Ces données supplémentaires représentent également l'ensemble des données disponibles pendant le séjour hospitalier.

Dans les Systèmes d'Informations H, d'autres types de données sont stockés. Par exemple, les documents textuels (les rapports médicaux) et les résultats d'imagerie médicales. Il serait pertinent de les intégrer dans la prédiction des DDS.

La prédiction de DDS s'avère un axe de recherche important dans le domaine médical. Les méthodes que nous avons proposé peuvent être généralisées à plusieurs établissements de soins. La prédiction de la DDS s'est révélée essentielle dans le quotidien des établissements de soins.

Bibliographie

[ABM17] Aya AWAD, Mohamed BADER–EL–DEN et James MCNICHOLAS. „Patient length of stay and mortality prediction : A survey“. In : *Health Services Management Research* 30.2 (2017), p. 105-120 (cf. p. 26).

[Age13] ATIH AGENCE TECHNIQUE DE L'INFORMATION SUR L'HOSPITALISATION. *Programme de médicalisation des systèmes d' information en soins de suite et de réadaptation (PMSI SSR)*. 2013 (cf. p. 21).

[Age18] MD. AGENCY FOR HEALTHCARE RESEARCH AND QUALITY, ROCKVILLE. *Data sources for health care quality evaluation*. 2018 (cf. p. 14).

[Age20] ATIH AGENCE TECHNIQUE DE L'INFORMATION SUR L'HOSPITALISATION. *Médecine, chirurgie, obstétrique, Chiffres clés.* Rapp. tech. 2020 (cf. p. 2).

[AJM12] Ali AZARI, Vandana P. JANEJA et Alex MOHSENI. „Predicting Hospital Length of Stay (PHLOS) : A Multi-tiered Data Mining Approach“. In : *2012 IEEE 12th International Conference on Data Mining Workshops*. 12. IEEE, déc. 2012, p. 17-24 (cf. p. 74).

[AK16] Samaneh AGHAJANI et Mehrdad KARGARI. „Determining Factors Influencing Length of Stay and Predicting Length of Stay Using Data Mining in the General Surgery Department“. In : *Hospital practice and research (HPR)* 1.2 (2016), p. 53-58 (cf. p. 36, 37).

[Ako18] Haldun AKOGLU. „User's guide to correlation coefficients“. In : *Turkish Journal of Emergency Medicine* 18.3 (2018), p. 91-93 (cf. p. 68).

[Alm+16] Ahmed ALMASHRAFI, Hilal ALSABTI, Mirdavron MUKADDIROV, Baskaran BALAN et Paul AYLIN. „Factors associated with prolonged length of stay following cardiac surgery in a major referral hospital in Oman : A retrospective observational study“. In : *BMJ Open* 6.6 (2016) (cf. p. 35, 37, 41).

[AMB18] Ayman ALAHMAR, Emad A. MOHAMMED et Rachid BENLAMRI. „Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes“. In : *International Conference on Big Data Innovations and Applications, Innovate-Data*. IEEE, 2018, p. 38-43 (cf. p. 2, 42).

[Ami+11] Fatemeh AMIRI, Mohammad Mahdi REZAEI YOUSEFI, Caro LUCAS, Azadeh SHAKERY et Nasser YAZDANI. „Mutual information-based feature selection for intrusion detection systems“. In : *Journal of Network and Computer Applications* 34.4 (2011). Advanced Topics in Cloud Computing, p. 1184-1199 (cf. p. 68).

[Ana20] ANALYTICS VIDHVA. *Introduction to Feature Selection methods with an example*. 2020 (cf. p. 67).

[And19] Olle ANDERSSON. *Predicting Patient Length Of Stay at Time of Admission Using Machine Learning*. Rapp. tech. 2019 (cf. p. 42, 43).

[Are+] John AREVALO, Fabio A. GONZÁLEZ, Raúl RAMOS-POLLÁN, Jose L. OLIVEIRA et Miguel Angel GUEVARA LOPEZ. „Convolutional neural networks for mammography mass lesion classification". In : *Annu Int Conf IEEE Eng Med Biol Soc*. T. 2015, p. 797-800 (cf. p. 28).

[ATI03] ATIH. *Agence technique de l'information sur l'hospitalisation*. 2003 (cf. p. 24).

[ATI11] ATIH. *Aide à l'utilisation des informations de chainage*. Rapp. tech. 2011, p. 1-22 (cf. p. 24).

[Ban21] Ankita BANERJI. *K-Means : Getting the optimal number of clusters*. 2021 (cf. p. 75).

[Bas+20] Daniel BASHIR, George D. MONTAÑEZ, Sonia SEHRA, Pedro Sandoval SEGURA et Julius LAUW. „An Information-Theoretic Perspective on Overfitting and Underfitting". In : *Advances in Artificial Intelligence* 12576.2 (2020), p. 347-358 (cf. p. 72).

[BBB20] Martin BEAULIEU, Omar BENTAHAR et Smail BENZIDIA. „The Evolution of Healthcare Logistics : The Canadian Experience". In : *Journal of Applied Business and Economics* 22.14 (2020), p. 1-8 (cf. p. 12, 26).

[BD02] Peter J. BROCKWELL et Richard A. DAVIS. *Introduction to Time Series and Forecasting*. 2. Springer-Verlag New York, 2002 (cf. p. 53).

[Ben+19] Sofia BENBELKACEM, Farid KADRI, Baghdad ATMANI et Sondes CHAABANE. „Machine Learning for Emergency Departement Management". In : *International Journal of Information Systems in the Service Sector* 11.3 (2019) (cf. p. 33).

[Ber+11] James BERGSTRA, Rémi BARDENET, Yoshua BENGIO et Balázs KÉGL. „Algorithms for hyper-parameter optimization". In : *Advances in Neural Information Processing Systems 24 : 25th Annual Conference on Neural Information Processing Systems, NIPS*. 2011, p. 1-9 (cf. p. 71).

[BH19] Carl BERGSTRÖM et Oscar HJELM. „Impact of Time Steps on Stock Market Prediction with LSTM". In : *Degree Project in Computer Science, Communication and Industrial Management*. 2019, p. 1-12 (cf. p. 54).

[Bis06] Christopher M BISHOP. *Pattern Recognition and Machine Learning*. 2006 (cf. p. 45).

[Bod01] Mikael BODÉN. „A Guide to Recurrent Neural Networks and Backpropagation". In : (déc. 2001) (cf. p. 53).

[Bom+20] Andrea BOMMERT, Xudong SUN, Bernd BISCHL, Jörg RAHNENFÜHRER et Michel LANG. „Benchmark for filter methods for feature selection in high-dimensional classification data". In : *Computational Statistics & Data Analysis* 143 (2020), p. 106839 (cf. p. 67).

[Bot+13] Alex BOTTLE, Steven MIDDLETON, Cor J. KALKMAN, Edward H. LIVINGSTON et Paul AYLIN. „Global comparators project : International comparison of hospital outcomes using administrative data“. In : *Health Services Research* 48.6 PART1 (2013), p. 2081-2100 (cf. p. 26).

[Bou+02] Mokrane BOUZEGHOUB, Bernadette Farias LÓSCIO, Zoubida KEDAD et Assia SOUKANE. „Heterogeneous data source integration and evolution“. In : *International Conference on Database and Expert Systems Applications* September 2002 (2002), p. 751-757 (cf. p. 18).

[Bre01] Leo BREIMAN. „Random forests“. In : *Machine learning* 45 (2001), p. 5-32 (cf. p. 52).

[Bre96] Leo BREIMAN. „Bagging predictions“. In : *Machine Learning* 24 (1996), p. 123-140 (cf. p. 52).

[Bro17] Jason BROWNLEE. *Why One-Hot Encode Data in Machine Learning ?* 2017 (cf. p. 69).

[Cai+16] Xiongcai CAI, Oscar PEREZ-CONCHA, Enrico COIERA et al. „Real-time prediction of mortality, readmission, and length of stay using electronic health record data“. In : *Journal of the American Medical Informatics Association* 23.3 (2016), p. 553-561 (cf. p. 29).

[Cap19] Guilherme CAPONETTO. *Random Search vs Grid Search for hyperparameter optimization*. 2019 (cf. p. 71).

[CDD01] Silvana CASTANO, Valeria DE ANTONELLIS et Sabrina Capitani Di DE VIMERCATI. „Global viewing of heterogeneous data sources“. In : *IEEE Transactions on Knowledge and Data Engineering* 13.2 (2001), p. 277-297 (cf. p. 19).

[Cha08] Salma CHAHED JEBALIA. „MODELISATION ET ANALYSE DE L' ORGANISATION ET DU FONCTIONNEMENT DES STRUCTURES D' HOSPITALISATION A DOMICILE“. Thèse de doct. Sciences de l'ingénieur [physics] Ecole centrale Paris, 2008 (cf. p. 22).

[CHL18] Mao-te CHUANG, Ya-han HU et Chia-lun LO. „Predicting the prolonged length of stay of general surgery patients : a supervised learning approach“. In : *INTERNATIONAL TRANSACTIONS INOPERATIONAL RESEARCH* 25.1 (2018), p. 75-90 (cf. p. 42).

[Chu+16] Mao Te CHUANG, Ya Han HU, Chih Fong TSAI, Chia Lun LO et Wei Chao LIN. „The Identification of Prolonged Length of Stay for Surgery Patients“. In : *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*. 2016, p. 3000-3003 (cf. p. 33, 36, 37, 42, 43).

[CNI19] CNIL. *La loi Informatique et Libertés |*. Rapp. tech. 2019, p. 1-51 (cf. p. 16).

[Com20] Valérie COMBE. *Évaluation du système d' information hospitalier (SIH) et du dossier patient selon le référentiel de certification*. 2020 (cf. p. 10).

[Com21] Hadrien COMMENGES. *Introduction à la statistique*. 2021 (cf. p. 66).

[CP14] Evelene M. CARTER et Henry Ww POTTS. „Predicting length of stay from an electronic patient record system : A primary total knee replacement example“. In : *BMC Medical Informatics and Decision Making* 14.1 (2014), p. 1-13 (cf. p. 32).

[CS14] Girish CHANDRASHEKAR et Ferat SAHIN. „A survey on feature selection methods“. In : *Computers & Electrical Engineering* 40.1 (2014). 40th-year commemorative issue, p. 16-28 (cf. p. 67).

[CSK11] Antonio CRIMINISI, Jamie SHOTTON et Ender KONUKOGLU. „Decision forests : A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning“. In : *Foundations and Trends in Computer Graphics and Vision* 7.2-3 (2011), p. 81-227 (cf. p. 52, 53).

[De +20] Edison Vitório DE SOUZA JÚNIOR, Gabriel Aguiar NUNES, Mariana Alves Soledade DE JESUS et al. „Hospitalizations and hospital costs for spontaneous abortion in Bahia, Brazil“. In : June. Juin 2020, p. 767-773 (cf. p. 19).

[Deg13] P. DEGOULET. „Les systèmes d'information hospitaliers“. In : *Informatique médicale, e-Santé*. 1998. Springer Paris, 2013, p. 307-330 (cf. p. 11, 12).

[DK19] Thomas DAVENPORT et Ravi KALAKOTA. „The Potential for Artificial Intelligence in Healthcare“. In : *Future Healthcare Journal* 6.2 (2019), p. 94-98 (cf. p. 28, 29).

[Dor+20] Neeltje van DOREMALEN, Trenton BUSHMAKER, Dylan H. MORRIS et al. „Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1“. In : *New England Journal of Medicine* 382.16 (2020), p. 1564-1567 (cf. p. 40).

[DRE15] DREES. *Données de santé : anonymat et risque de ré-identification*. Rapp. tech. 2015 (cf. p. 16).

[EAH20] Amr Mohamed ELSHARKAWY, Samir Mohamed AL-AWADY et Tamer Abdullah HELMY. „Postintubation Hypotension and its Association with Prolonged ICU Length of Stay and ICU Mortality“. In : *Egyptian Journal of Critical Care Medicine* 7.1 (2020), p. 26 (cf. p. 36, 37).

[Elg05] Haytham ELGHAZEL. *Initiation au PMSI*. Rapp. tech. 2005, p. 3-5 (cf. p. 11, 20, 21, 23).

[Evi89] Claude EVIN. „CIRCULAIRE n 275 du 6 janvier 1989 relative à l'informatisation des hôpitaux publics“. In : (1989) (cf. p. 19).

[FAU21] Emile FAURE. *Le programme de médicalisation des systèmes d 'information (PMSI)*. 2021 (cf. p. 12, 21, 22).

[Fen+14] Changyong FENG, Hongyue WANG, Naiji LU et al. „Log-transformation and its implications for data analysis“. In : *Shanghai Archives of Psychiatry* 26.2 (2014), p. 105-109 (cf. p. 68).

[FGP09] Malcolm FADDY, Nicholas GRAVES et Anthony PETTITT. „Modeling length of stay in hospital and other right skewed data : Comparison of phase-type, gamma and log-normal distributions“. In : *Value in Health* 12.2 (2009), p. 309-314 (cf. p. 41).

[FSF13] Mohammad Mehdi FARHANGI, Mohsen SORYANI et Mahmood FATHY. „Improvement the Bag of Words Image Representation Using Spatial Information“. In : *Advances in Computing and Information Technology*. Sous la dir. de Natarajan MEGHANATHAN, Dhinaharan NAGAMALAI et Nabendu CHAKI. Berlin, Heidelberg : Springer Berlin Heidelberg, 2013, p. 681-690 (cf. p. 91).

[GE03] Isabelle GUYON et André ELISSEEFF. „An Introduction to Variable and Feature Selection“. In : *Journal ofMachine Learning Research* 3 (2003), p. 1157-1182 (cf. p. 67).

[Gen+17] Thanos GENTIMIS, Ala' J. ALNASER, Alex DURANTE, Kyle COOK et Robert STEELE. „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)*. 2017, p. 1194-1201 (cf. p. 26, 35, 42, 43).

[Gra14] L. GRAMMATICO. „Intérêt et limites du Programme de médicalisation des systèmes d'information dans la surveillance des infections de prothèses orthopédiques“. Thèse de doct. Université Pierre et Marie Curie - Paris VI ; Université de Tours, 2014, S134 (cf. p. 20, 23).

[Gro18] Prince GROVER. *5 Regression Loss Functions All Machine Learners Should Know*. 2018 (cf. p. 49).

[Hab21] Adria Binte HABIB. *Elbow Method vs Silhouette Co-efficient in Determining the Number of Clusters*. Rapp. tech. June. 2021 (cf. p. 75).

[Hac+13] Peyman Rezaei HACHESU, Maryam AHMADI, Somayyeh ALIZADEH et Farahnaz SADOUGHI. „Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients“. In : *Healthcare Informatics Research* 19.2 (2013), p. 121-129 (cf. p. 35, 42, 43).

[HHL11] Frank HUTTER, Holger H. HOOS et Kevin LEYTON-BROWN. „Sequential Model-Based Optimization for General Algorithm Configuration“. In : *International Conference on Learning and Intelligent Optimization*. 2011, p. 507-523 (cf. p. 71).

[HIM] HIMSS. *EHR definition* (cf. p. 11).

[Hol15] Robert HOLCMAN. „Le programme de médicalisation du système d'information (PMSI)“. In : *Management hospitalier*. T. 22. 2015, p. 553-568 (cf. p. 20).

[HW14] Rebecca HERMON et Patricia WILLIAMS. „Big data in healthcare : What is it used for ?“ In : *Proceedings of the 3rd Australian eHealth Informatics and Security Conference*. 2014, p. 40-49 (cf. p. 16).

[IBM20] IBM CLOUD EDUCATION. *What is Machine Learning*. 2020 (cf. p. 43).

[INS16] INSERM. *Big data in health, Technical, human and ethical challenges*. 2016 (cf. p. 14).

[JDJ19] Kaio JORDON, Paul Eric DOSSOU et Joao Chang JUNIOR. „Using lean manufacturing and machine learning for improving medicines procurement and dispatching in a hospital“. In : *Procedia Manufacturing*. T. 38. 2019. Elsevier B.V., 2019, p. 1034-1041 (cf. p. 29).

[Jia+17] Fei JIANG, Yong JIANG, Hui ZHI et al. „Artificial intelligence in healthcare : Past, present and future“. In : *Stroke and Vascular Neurology* 2.4 (2017), p. 230-243 (cf. p. 28).

[JJ20] Daniel JURAFSKY et H. Martin JAMES. „Neural Networks and Neural Language Models“. In : *Speech and Language Processing* 3.7 (2020) (cf. p. 54).

[JK13] A.N. JACKSON et S. KOGUT. „Use of electronic personal health records to identify patients at risk for aspirin-induced gastrointestinal bleeding“. In : *Consult Pharm* 28.5 (2013), p. 688-692 (cf. p. 13).

[JO14] Václav JIRKOVSKÝ et Marek OBITKO. „Semantic heterogeneity reduction for big data in industrial automation“. In : *CEUR Workshop Proceedings* 1214 (2014), p. 1-10 (cf. p. 18).

[Jon13] Rod JONES. „Average length of stay in hospitals in the USA“. In : *British Journal of Health Care Management* 19.4 (2013), p. 186-191 (cf. p. 33).

[Kho+16] Omid KHOSRAVIZADEH, Soudabeh VATANKHAH, Peivand BASTANI et al. „Factors affecting length of stay in teaching hospitals of a middle-income country“. In : *Electronic physician* 8.10 (2016), p. 3042-3047 (cf. p. 33, 34, 36, 37).

[KLH20] Guilan KONG, Ke LIN et Yonghua HU. „Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU“. In : *BMC Medical Informatics and Decision Making* 20.1 (2020), p. 1-10 (cf. p. 29).

[Kum21] Ajitesh KUMAR. *K-Fold Cross Validation*. 2021 (cf. p. 72).

[KZ00] Stephen KOKOSKA et Daniel ZWILLINGER. *CRC Standard Probability and Statistics Tables and Formulae, Student Edition*. 2000 (cf. p. 66).

[Laf+15] Rocco J LAFARO, Suryanarayana POTHULA, Keshar Paul KUBAL et al. „Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre- Incision Variables“. In : *plos one* 10.12 (2015), p. 1-19 (cf. p. 34, 35).

[Lav20] Gupta LAVANYA. *Comparison of Hyperparameter Tuning algorithms : Grid search, Random search, Bayesian optimization*. 2020 (cf. p. 71).

[Lég21] LÉGIFRANCE. *Articles L. 6113-7 et L. 6113-8 du Code de la Santé Publique*. 2021 (cf. p. 23).

[Li+13] Jing Song LI, Yu TIAN, Yan Feng LIU, Ting SHU et Ming Hui LIANG. „Applying a BP neural network model to predict the length of hospital stay“. In : *Springer-Verlag Berlin Heidelberg 2013*. T. 7798 LNCS. 2013, p. 18-29 (cf. p. 42).

[Lia+12] Jiye LIANG, Xingwang ZHAO, Deyu LI, Fuyuan CAO et Chuangyin DANG. „Determining the number of clusters using information entropy for mixed data“. In : *Pattern Recognition* 45.6 (juin 2012), p. 2251-2265 (cf. p. 75).

[Lin+18] Hester F. LINGSMA, Alex BOTTLE, Steve MIDDLETON et al. „Evaluation of hospital outcomes : The relation between length-of-stay, readmission, and mortality in a large international administrative database“. In : *BMC Health Services Research* 18.1 (2018), p. 1-11 (cf. p. 26).

[Lóp21] Fernando LÓPEZ. *Ensemble Learning : Bagging & Boosting*. 2021 (cf. p. 52).

[Los09] David LOSHIN. „Data Consolidation and Integration“. In : *Master Data Management*. Elsevier, 2009, p. 177-199 (cf. p. 69).

[LT09] Apiradee LIM et Phattrawan TONGKUMCHUM. „Methods for Analyzing Hospital Length of Stay with Application to Inpatients Dying in Southern Thailand“. In : *Global Journal of Health Science* 1.1 (2009), p. 27-38 (cf. p. 26).

[Mah+18] Hamidreza MAHARLOU, Sharareh R. NIAKAN KALHORI, Shahrbanoo SHAHBAZI et Ramin RAVANGARD. „Predicting length of stay in intensive care units after cardiac surgery : Comparison of artificial neural networks and adaptive neuro-fuzzy system“. In : *Healthcare Informatics Research* 24.2 (2018), p. 109-117 (cf. p. 34, 35, 37).

[Mar+01] A. H. MARSHALL, S. I. MCCLEAN, C. M. SHAPCOTT, I. R. HASTIE et P. H. MILLARD. „Developing a Bayesian belief network for the management of geriatric hospital care“. In : *Health Care Management Science* 4.1 (2001), p. 25-30 (cf. p. 26).

[Mek+19] Rachda Naila MEKHALDI, Patrice CAULIER, Sondes CHAABANE et Abdelahad CHRAIBI. „Apports de lIntelligence Artificielle à la prédiction des durées de séjours hospitaliers“. In : *PFIA, IA & santé*. 2019, p. 1-7 (cf. p. 34, 37).

[Mek+20a] Rachda Naila MEKHALDI, Patrice CAULIER, Sondes CHAABANE, Abdelahad CHRAIBI et Sylvain PIECHOWIAK. „Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting“. In : *Advances in Intelligent Systems and Computing*. T. 1159 AISC. 2020, p. 202-211 (cf. p. 69, 76).

[Mek+20b] Rachda Naila MEKHALDI, Patrice CAULIER, Sondes CHAABANE et al. „Apprentissage automatique dans la prédiction des durées de séjour hospitalier“. In : *Conférence Francophone en Gestion et Ingénierie des Systèmes Hospitaliers*. 2020, p. 1-8 (cf. p. 36).

[Mek+21] Rachda Naila MEKHALDI, Patrice CAULIER, Sondes CHAABANE, Abdelahad CHRAIBI et Sylvain PIECHOWIAK. „A Comparative Study of Machine Learning Models for Predicting Length of Stay in Hospitals“. In : *JOURNAL OF INFORMATION SCIENCE AND ENGINEERING* 37.5 (2021), p. 1025-1038 (cf. p. 68, 70, 76).

[Mic19] Nielsen MICHAEL. *How the backpropagation algorithm works*. 2019 (cf. p. 54).

[MM97] Tom M. MITCHELL et Hill MCGRAW. *Machine Learning*. 1997 (cf. p. 43).

[Mor09] Michel MORKOS. *Le PMSI, qu'est-ce que c'est ?* 2009 (cf. p. 21, 23).

[MVM11] Nijolė MAKNICKIENĖ, Aleksandras VYTAUTAS RUTKAUSKAS et Algirdas MAKNICKAS. „Investigation of financial market prediction by recurrent neural network“. In : *Innovative Infotechnologies for Science, Busi- ness and Education* 2.11 (2011), p. 3-8 (cf. p. 53).

[Naj17] Ahmed NAJJAR. „Forage de données de bases administratives en santé“. Thèse de doct. 2017 (cf. p. 29, 74, 82).

[Ngu+21] Dung NGUYEN, Phuc HO, Thien NGUYEN et Van NGUYEN. „Statistical analysis on length of stay in hospital“. In : *Science & Technology Development Journal - Engineering and Technology* 3.SI3 (jan. 2021), SI97-SI106 (cf. p. 41).

[NQS17] Farnaz NOJAVAN A., Song S. QIAN et Craig A. STOW. „Comparative analysis of discretization methods in Bayesian networks“. In : *Environmental Modelling and Software* 87 (2017), p. 64-71 (cf. p. 19).

[OB10] J. R. OTUKEI et T. BLASCHKE. „Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms“. In : *International Journal of Applied Earth Observation and Geoinformation* 12.SUPPL. 1 (2010), p. 27-31 (cf. p. 47).

[OCD17] OCDE. *Average length of stay in hospitals*. 2017 (cf. p. 33).

[OEC20] OECD. *OECD Health Statistics 2020 Definitions , Sources and Methods*. Rapp. tech. 2020, p. 1-22 (cf. p. 33).

[Oli20] Hugo De OLIVEIRA. „Predictive modeling of patient pathways using process mining and deep learning“. Thèse de doct. 2020 (cf. p. 88).

[OOr21] OOREKA. *Classification internationale des maladies (CIM)*. 2021 (cf. p. 25).

[Org21] OMS ORGANISATION MONDIALES DE LA SANTÉ. *Site officiel de l'organisation mondiales de la santé.* 2021 (cf. p. 25).

[Pap06] P. PAPIN. „Classification commune des actes médicaux“. In : *Revue de Chirurgie Orthopedique et Reparatrice de l'Appareil Moteur* 92.5 SUPPL. (2006), p. 1-90 (cf. p. 25).

[PHS14] N. PEEK, J. H. HOLMES et J. SUN. „Technical challenges for big data in biomedicine and health : data sources, infrastructure, and analytics“. In : *Yearbook of medical informatics* 9 (2014), p. 42-47 (cf. p. 13, 14).

[Pic18] Devin PICKELL. *Structured vs Unstructured Data – What's the Difference ?* 2018 (cf. p. 18).

[PK14] Parag C. PENDHARKAR et Hitesh KHURANA. „Machine learning techniques for predicting hospital length of stay in pennsylvania federal and specialty hospitals“. In : *International Journal of Computer Science and Applications* (2014) (cf. p. 34, 59).

[PR14] V. PANCHAMI et N. RADHIKA. „A novel approach for predicting the length of hospital stay with DBSCAN and supervised classification algorithms“. In : *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)* (2014), p. 207-212 (cf. p. 74).

[RIG09] Magali RIGAL. „Management des lits et durée moyenne de séjour : Exemple de recherche d'optimisation au Centre Hospitalier d'Avignon“. Thèse de doct. Ecole des hautes études en santé publique, 2009 (cf. p. 26, 34, 59).

[RO09] Thomas RENAUD et Zeynep OR. *Principes et enjeux de la tarification à l'activité à l'hôpital (T2A) : Enseignements de la théorie économique et des expériences étrangères*. Rapp. tech. 2009, p. 29 (cf. p. 23).

[Row+07] Michael ROWAN, Thomas RYAN, Francis HEGARTY et Neil O HARE. „The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors“. In : *Artificial Intelligence in medecin* 40 (2007), p. 211-221 (cf. p. 34).

[RR17] K H REINERT et J H RODGERS. „Deep learning at chest radiography : Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks“. In : *Radiology* 284.2 (2017), p. 574-582 (cf. p. 28).

[RWM19] Christopher A. RAMEZAN, Timothy A. WARNER et Aaron E. MAXWELL. „Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification“. In : *Remote Sensing* 11.2 (2019) (cf. p. 72).

[san17a] Ministère des affaires sociales et de la SANTÉ. *CLASSIFICATION COMMUNE des Actes Médicaux. Descriptive à usage PMSI.* Rapp. tech. 2017 (cf. p. 25).

[san17b] Ministère des solidarité et de la SANTÉ. *Financement des établissements de santé*. 2017 (cf. p. 23).

[Sch03] Robert E SCHAPIRE. „The Boosting Approach to Machine Learning : An Overview“. In : *Nonlinear Estimation and Classification* 171.Chapter 9 (2003), p. 149-171 (cf. p. 52).

[Sha48] C. E. SHANNON. „A mathematical theory of communication“. In : *The Bell System Technical Journal* 27.3 (1948), p. 379-423 (cf. p. 68).

[She+95] Steven SHEA, Robert V. SIDELI, William DUMOUCHEL et al. „Computer-generated informational messages directed to physicians : Effect on length of hospital stay“. In : *Journal of the American Medical Informatics Association* 2.1 (1995), p. 58-64 (cf. p. 34).

[Sin18] Harshdeep SINGH. *Understanding Gradient Boosting Machines*. 2018 (cf. p. 53).

[Sin19] Aishwarya SINGH. *R-Squared vs. Adjusted R-Squared*. 2019 (cf. p. 50).

[sol02] Minsitère de l'emploi et de la SOLIDARITÉ. *PMSI généralités*. Rapp. tech. 2002, p. 1-22 (cf. p. 20).

[SS13] Liessman E. STURLAUGSON et John W. SHEPPARD. „Principal component analysis preprocessing with Bayesian networks for battery capacity estimation“. In : *IEEE Instrumentation and Measurement Technology Conference*. IEEE, 2013, p. 98-101 (cf. p. 19).

[SW73] Budd N. SHENKIN et David C. WARNER. „Giving the Patient His Medical Record : A Proposal to Improve the System“. In : *New England Journal of Medicine* 289.13 (1973), p. 688-692 (cf. p. 13).

[Ten21] TENSORFLOW. *Masquage et rembourrage avec Keras*. 2021 (cf. p. 98).

[Tsa+16] Pei Fang Jennifer TSAI, Po Chia CHEN, Yen You CHEN et al. „Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network“. In : *Journal of Healthcare Engineering* 2016 (2016), 1-11 pages (cf. p. 35).

[Tup+17] P. TUPPIN, J. RUDANT, P. CONSTANTINOU et al. „Value of a national administrative database to guide public decisions : From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France“. In : *Revue d'Epidemiologie et de Sante Publique* 65 (2017), S149-S167 (cf. p. 16).

[Tur50] Alan M. TURING. „COMPUTING MACHINERY AND INTELLIGENCE“. In : *MIND* LIX.236 (1950), p. 433-460 (cf. p. 28).

[Ver+07] Marion VERDUIJN, N. PEEK, F. VOORBRAAK, E. DE JONGE et B. A.J.M. DE MOL. „Modeling length of stay as an optimized two-class prediction problem“. In : *Methods of Information in Medicine* 46.3 (2007), p. 352-359 (cf. p. 41).

[VS08] J. VALACICH et C. SCHNEIDER. „Information Systems Today : Managing the Digital World“. In : 2008 (cf. p. 11).

[Wan17] Lidong WANG. „Heterogeneous Data and Big Data Analytics“. In : *Automatic Control and Information Sciences* 3.1 (2017), p. 8-15 (cf. p. 18).

[WPS03] Steven WALCZAK, Walter E. POFAHL et Ronald J. SCORPIO. „A decision support tool for allocating hospital bed resources and determining required acuity of care“. In : *Decision Support Systems* 34.4 (2003), p. 445-456 (cf. p. 41).

[Wre+05] Jesse WRENN, Ian JONES, Kevin LANAGHAN, Clare Bates CONGDON et Dominik ARONSKY. „Estimating patient's length of stay in the Emergency Department with an artificial neural network.“ In : *AMIA. Annual Symposium proceedings*. March 2014. 2005, p. 1155 (cf. p. 33).

[Xin+07] Kai XING, Dechang CHEN, Donald HENSON et Li SHENG. „A Clustering-Based Approach to Predict Outcome in Cancer Patients“. In : *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*. January 2008. IEEE, déc. 2007, p. 541-546 (cf. p. 74).

[Yan+10] Chin Sheng YANG, Chih Ping WEI, Chi Chuan YUAN et Jen Yu SCHOUNG. „Predicting the length of hospital stay of burn patients : Comparisons of prediction accuracy among different clinical stages“. In : *Decision Support Systems* 50.1 (2010), p. 325-335 (cf. p. 17, 41).

[Yan19] Feng Jen YANG. „An extended idea about decision trees“. In : *Science and Computational Intelligence*. 2019, p. 349-354 (cf. p. 51).

[YBK18] Kun Hsing YU, Andrew L. BEAM et Isaac S. KOHANE. „Artificial intelligence in healthcare“. In : *Nature Biomedical Engineering* 2.10 (2018), p. 719-731 (cf. p. 28).

[Yer19] Bhanu YERRA. *Objective Functions in Machine Learning*. 2019 (cf. p. 47).

[YHK20] Nooshin YOUSEFI, Farhad HASANKHANI et Mahsa KIANI. *Appointment scheduling model in healthcare using clustering algorithms*. 2020. arXiv : 1905 . 03083 [eess.SP] (cf. p. 74).

[Yin19] Xue YING. „An Overview of Overfitting and its Solutions“. In : *Journal of Physics : Conference Series* 1168.2 (2019), p. 1-6 (cf. p. 72).

[Zol18] Joan ZOLOT. „At-Home Hospital Care Reduces Readmissions and Length of Stay, Enhances Patient Satisfaction“. In : *American Journal of Nursing* 118.10 (2018) (cf. p. 22).

Table des figures

Liste des tableaux

Liste des publications

— **Mekhaldi, R. N.**, Caulier, P., Chaabane, S., Chraibi, A., & Piechowiak, S. (2020). A comparative study of machine learning models for predicting Length of Stay in hospitals. Journal of Information Science and Engineering, 37(5).

— **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. Advances in Intelligent Systems and Computing, 1159 AISC, 202–211.

— **Mekhaldi, R. N.,** Caulier, P., Chaabane, S., Piechowiak, S., Taillard, J., & Hansske, A. (2020). Apprentissage automatique dans la prédiction des durées de séjour hospitalier. GISEH.

— **Mekhaldi, R. N.,** Caulier, P., Chaabane, S., Piechowiak, S., & Chraibi, A. (2019). Apports de lIntelligence Artificielle à la prédiction des durées de séjours hospitaliers. IA & Santé, 1–7.

Glossaire

ARH Agence Régionale de l'Hospitalisation. 145

ATIH Agence Technique de l'Information Hospitalière. 145

CCAM Classification Commune des Actes Médicaux. 145

CIM Classification Internationale des Maladies. 145

CMC Catégorie Majeure Clinique. 145

CNIL Commission Nationale de l'Informatique et des Libertés. 145

DDS Durées De Séjour hospitalier. 145

DIM Département d'Information Médicale. 145

DME Dossier Médical Électronique. 145

DRG Diagnosis Related Groups. 145

ENC l'Echelle Nationale de Coût. 145

Finess l'identifiant de l'établissement de soins. 145

GHJ Groupe Homogène de Journées. 145

GHM Groupe Homogène de Malades. 145

HAD Hospitalisation A Domicile. 145

IA Intelligence Artificielle. 145

ISA Indice Synthétique d'Activité. 145

MCO Médecine, Chirurgie, Obstétrique et odontologie. 145

ML Machine Learning. 145

NLP Natural Langage Processing. 145

OMS Organisation Mondiales de la Santé. 145

PMSI Programme de Médicalisation des Systèmes d'Informations. 145

PSY Psychiatrie. 145

RGPD Règlement Général sur la Protection des Données. 145

RHA Résumé Hebdomadaires Anonymisés. 145

RHS Résumés Hebdomadaires Standardisés. 145

RSA Résumé Standardisé Anonymisé. 145

RSS Résumé Standardisé du Séjour. 145

RUM Résumé Unité Médicale. 145

SIH Systèmes d'Informations Hospitaliers. 145

SNDS Système National des Données de Santé. 145

SSR Soins de Suite et de Réadaptation. 145

T2A Tarification à l'activité. 145