Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Caen Normandie

En cotutelle internationale avec Université de Mascara Mustapha Stambouli , ALGERIE

Exploration des données de la médecine personnalisée par des techniques de Data Mining

Présentée et soutenue par
HAFID KADI

Thèse soutenue le 17/12/2021 devant le jury composé de		
M. SOFIANE BOUKLI HACENE	Professeur des universités, Université de Sidi Bel Abbès	Rapporteur du jury
M. KHALIFA DJEMAL	Maître de conférences HDR, Université Evry Val D'Essonne	Rapporteur du jury
M. HUBERT CARDOT	Professeur des universités, Université de Tours François Rabelais	Membre du jury
M. BOUDJELAL MEFTAH	Professeur des universités, Université de Mascara Mustapha Stambouli	Membre du jury
MME SAMIRA CHOURAQUI	Professeur des universités, Université USTO	Président du jury
M. OLIVIER LEZORAY	Professeur des universités, Université Caen Normandie	Directeur de thèse

Thèse dirigée par OLIVIER LEZORAY, Groupe de recherche en informatique, image, automatique et instrumentation

UNIVERSITÉ CAEN NORMANDIE

GREYC
Laboratoire de recherche en sciences du numérique

Résumé

La médecine personnalisée est actuellement en fort développement, de par son adoption dans le monde entier et principalement dans les pays développés. Les profils des patients constituent en effet le point principal sur lequel est fondé le but de cette médecine. Cette dernière vise à aider les médecins et les praticiens de la santé à prévoir des maladies, à prendre des décisions précises et à individualiser les traitements d'une manière adéquate. De plus, le profil d'un patient peut comporter une variété importante de données que ce soient des données génétiques, des biomarqueurs clés, l'historique de traitements, les facteurs environnementaux et les préférences comportementales, des images (IRM, Radio, …), etc. L'exploration de ces données par les outils de la fouille de données nécessite une suite d'opérations pour former et extraire les connaissances cachées parmi ces données. L'intérêt d'un tel processus d'automatisation de la décision médicale et d'extraction des connaissances est généralement confirmé par sa précision. Il ne faut néanmoins pas éluder les contraintes liées à la rapidité de calcul de celles-ci, pour permettre leur usage pratique.

Ces travaux de thèse, intitulés « **Exploration des données de la médecine personnalisée par des techniques de Data Mining** », nous a conduit à la définition de deux activités importantes de la médecine personnalisée : la première porte sur la représentation de données et la seconde sur la prise de la décision médicale. Par conséquent, deux problèmes ont été identifiés. Le premier concerne la perte de données et d'information lors de la phase de représentation de l'information. Le deuxième concerne le choix de la série des traitements la plus appropriée à appliquer pour la prise de décision. La solution de la première problématique a été résolue par la proposition d'un modèle de représentation de données par région et par dispersion. Pour la deuxième problématique, nous avons proposé un modèle de prise de décision médicale réalisé reposant sur une classification de données issues de la médecine personnalisée. Ce modèle repose sur l'application de notre modèle de représentation de données et plusieurs suites de traitement et de classification. L'expérimentation de nos modèles et les résultats obtenus justifient l'utilité et la précision de nos approches.

Ces solutions avantageuses, en particulier le modèle de représentation de données, peuvent être utilisées comme une plateforme exploitable pour d'autres tâches telles que l'analyse de données médicales.

Mots clés : Médecine personnalisée; Data Mining; Représentation de données; Prise de décision médicale; Séries temporelles; Réduction de données; Classification; Clustering.

Abstract

Personalized medicine is currently in strong development, by its adoption all over the world and mainly in developed countries. The profiles of the patients are indeed the main point on which is based the purpose of this medicine. The latter aims to help doctors and health care practitioners to predict diseases, make accurate decisions and to individualize the treatment adequately. In addition, a patient's profile can include a wide data variety, among genetic data, key biomarkers, treatment history, environmental factors and behavioral preferences, images (MRI, X-ray, etc.), etc. The exploration of this data by data mining tools requires a series of operations to train and extract the knowledge hidden among this data. The advantage of such a medical decision automation process and knowledge extraction is usually confirmed by its accuracy. However, we must not omit the constraints related to the calculation speed of these, to allow their practical use.

This thesis work, entitled « **Exploration des données de la médecine personnalisée par des techniques de Data Mining** », has led us to define two important activities in personalized medicine: the first focuses on data representation and the second on making medical decisions. Therefore, two problems were identified. The first concerns the loss of data and information during the information representation phase. The second concerns the choice of the most appropriate treatment series to apply for decision making. The solution of the first problem was solved by the proposal of a model for the data representation by region and by

dispersion. For the second problem, we have proposed a model of medical decision-making based on a data classification from personalized medicine. This model is based on the application of our data representation model and several treatment and classification suites. Our models' experimentation and the obtained results justify the usefulness and accuracy of our approaches. These beneficial solutions, in particular the data representation model, can be used as an exploitable platform for other tasks such as medical data analysis.

Keywords: Personalized medicine; Data Mining; Data representation; Medical decision making; Time series; Data reduction; Classification; Clustering.

Remerciements

Ce travail de thèse a bénéficié d'un financement du PHC Tassili - projet Dermato.ai 19MDU212.

Je tiens à remercier l'Université Mustapha STAMBOULI de Mascara en Algérie, l'Université de Caen Normandie, et le laboratoire *GREYC UMR CNRS 6072* en France de m'avoir donné cette opportunité de poursuivre mes études afin de préparer mon doctorat.

Je remercie infiniment Monsieur le professeur MEFTAH Boudjelal, directeur de thèse au niveau de l'université Mustapha STAMBOULI de Mascara en Algérie, pour ses énormes efforts fournis et ses directives tout au long de notre travail.

Je remercie également Monsieur le professeur OLIVIER Lézoray, directeur de thèse au niveau de l'université de Caen Normandie en France, pour ses conseils et ses idées, ainsi que sa disponibilité régulière et permanente, et ce malgré que la majorité de notre travail est réalisé à distance.

Mes remerciements sont plus particulièrement adressés à Monsieur le docteur REBBAH Mohammed, co-directeur de thèse au niveau de l'université Mustapha STAMBOULI de Mascara en Algérie, pour son assistance, sa collaboration exceptionnelle et de ses encouragements.

Mes sincères remerciements exprimés à Madame la professeure CHOURAQUI Samira de l'université des Sciences et de la Technologie d'Oran Mohamed-Boudiaf USTOMB en Algérie, et à Monsieur le professeur CARDOT Hubert de l'université de Tours en France pour avoir accepté d'être examinateurs de ce travail.

Mes chaleureux remerciements prononcés à Messieurs le professeur BOUKLI Hacene Sofiane de l'université de Sidi Bélabes en Algérie, et le docteur DJEMAL Khalifa de l'université d'Evry Val d'Essonne en France, d'avoir accepté de rapporter cette thèse.

Je remercie également tous les membres du Comité de Suivi annuel de thèse en Algérie, et tous les membres du Comité de Suivi Individuel de thèse en France, pour

leurs participations à l'évaluation de notre travail annuel et leurs importantes remarques. Enfin, je remercie tous les professeurs qui m'ont enseigné tout au long de mon parcours d'études et de formation.

Je dédie cette thèse plus spécialement

A mon père, que Dieu ait pitié de lui,

A ma très chère mère,

A tous mes frères et mes sœurs,

A ma grande famille,

A tous mes amis.

Table des matières

Table des matières 1
Table des figures 6
Liste des tableaux 8
Introduction générale 11
Chapitre 1 : Qu'est-ce que la médecine personnalisée ? 17
1.1 Introduction 18
1.2 Définition 18
1.3 Historique de la médecine personnalisée 21
1.4 Médecine personnalisée et médecine traditionnelle 22
1.4.1 Médecine personnalisée versus médecine traditionnelle 22
1.4.2 Passage de la médecine traditionnelle vers la médecine personnalisée 22
1.5 Défis de la médecine personnalisée 23
1.5.1 Médecine personnalisée et la vue économique 23
1.5.2 Médecine personnalisée et l'aspect juridique 24
1.5.3 Médecine personnalisée et la vue éthique et sociale 25
1.6 Médecine personnalisée et outils informatiques 26
1.6.1 Médecine personnalisée et les données électroniques des patients 26
1.6.2 Exploitation de données de la médecine personnalisée 27
1.6.3 Prise de décision médicale et médecine personnalisée 28
1.6.4 Médecine personnalisée en temps réel 29
1.7 Opportunités de la médecine personnalisée 30
1.7.1 Bénéficiaires de la médecine personnalisée 30
1.7.2 Offres de la médecine personnalisée 31
1.7.3 Avenir de la médecine personnalisée 32
1.8 Conclusion 33
Chapitre 2 : Qu'est-ce que la représentation de donnés ? 39
2.1 Introduction 40
2.2 Data Mining 41
2.2.1 Etapes de processus DM 41

2.2.2 Tâches de Data Mining 44
2.2.3 Disciplines incorporées en data mining 46
2.2.4 Critères d'évaluation des modèles 49
2.2.5 Evaluation de la précision des modèles 50
2.3 Data Mining pour les séries temporelles (Time series data mining) .. 59
2.3.1 Séries temporelles (Time series) 59
2.3.2 Différentes tâches de data mining sur les séries temporelles 60
2.4 Big Data 64
2.4.1 Couches de Big Data 64
2.4.2 Chiffres et promesses en Big Data 65
2.4.3 Défis du Big Data 66
2.4.4 Techniques de Big Data 67
2.4.5 Types de Base de données NoSQL 68
2.4.6 Impact du Big data sur la médecine personnalisée 68
2.5 Conclusion 70
Chapitre 3 : Qu'est-ce que la représentation de donnés ? 75
3.1 Introduction 76
3.2 Représentation de données 77
3.2.1 Définition de représentation données 77
3.2.2 Transformation de données 78
3.2.3 Représentation des types de données de base 81
3.2.4 Représentation des types de données avancés 82
3.2.5 Représentation des séries de données temporelles 82
3.3 Réduction de données 87
3.3.1 Techniques de réduction de dimension 87
3.4 Classification de données 92
3.4.1 Distances en data mining 93
3.4.2 Techniques de segmentation (Clustering) 94
3.4.3 Indices de qualité du clustering 98
3.4.4 Techniques de classification 100
3.5 Conclusion 102
Chapitre 4 : Problématiques ! 113
4.1 Introduction 114

4.2 Problème de perte de données et de l'information 115
4.3 Problème de choix de série des traitements 120
4.4 Conclusion 121
Chapitre 5 : Représentation de Données de la MP par Région et Dispersion (DRRD). 125
5.1 Introduction 126
5.2 Description générale 126
5.3 Modèle proposé 127
5.3.1 Formulation du problème 128
5.3.2 Schéma détaillé 128
5.3.3 Réorganisation des données du type numérique et date (Etape A_1) 129
5.3.4 Partitionnement des événements numériques (Etape B_1) 131
5.3.5 Marquage des données numériques (Etape C_1) 134
5.3.6 Linéarisation des données numériques (Etape D_1) 136
5.3.7 Représentation des données du type nominal et booléen (Etape A_2) 137
5.3.8 Dispersion « Diffusion » des événements nominaux (Etape B_2) 138
5.3.9 Marquage des données nominales (Etape C_2) 139
5.3.10 Linéarisation des données nominales (Etape D_2) 140
5.3.11 Assemblage des résultats 140
5.4 Expérimentation 141
5.4.1 Description du dataset 142
5.4.2 Sélection, transformation et codification de données 142
5.4.3 Résultats et discussion 143
5.4.4 Exemple de représentation et évaluation 150
5.4.5 Exemple de vue par patient 152
5.5 Conclusion 153
Chapitre 6 : Prise de décision médicale basée sur l'exploration d'un dataset de la MP. 158
6.1 Introduction 159
6.2 Description générale 160
6.3 Modèle proposé 161
6.3.1 Représentation de données 162

6.3.2 Distance entre les patients (génération de matrice de distance).....164
6.3.3 Réduction de dimensionnalité.....165
6.3.4 Classification.....167
6.4 Résultats expérimentaux.....169
6.5 Discussion et évaluation.....175
6.5.1 Evaluation des résultats.....175
6.5.2 Comparaison.....178
6.6 Conclusion.....180
Conclusion Générale.....186
Publications.....191

Table des figures

Figure 1.1. Les quatre aspects de la médecine P4 (MP). 20

Figure 2.1. Étapes de processus de DM. (Fayyad et al., 1996). 42
Figure 2.2. Les disciplines incorporées en data mining. (Han et al., 2012) 47
Figure 2.3. La courbe ROC et l'aire sous la courbe ROC. 59
Figure 2.4. Type d'approches de détection des anomalies des séries temporelles. . 62
Figure 2.5. Revenu de Big Data et le marché analytique des affaires (Statista, 2018). 66
Figure 2.6. Appareils connectés médicales (Piwek et al., 2016). 69

Figure 3.1. Approches de représentation des séries temporelles. 83
Figure 3.2. Exemple illustratif de la technique PAA. 84

Figure 4.1. Représentation SAX de X1 et X2 sans normalisation. 118
Figure 4.2. Représentation SAX de X1 et X2 avec normalisation. 118
Figure 4.3. Représentation Multi-séries SAX de X1 et X2 sans normalisation. 119
Figure 4.4. Représentation Multi-séries SAX de X1 et X2 avec normalisation. 119

Figure 5.1. Modèle de représentation de données de la MP. 130
Figure 5.2. Processus de clustering d'un événement numérique E_i 132
Figure 5.3. Illustration des techniques de clustering appropriées correspondant aux statistiques de : (a) l'inertie intraclasse, (b) l'inertie interclasse. 146
Figure 5.4. Dispersion des événements nominaux : (a) "GENDER", (b) "TRIBE". 148
Figure 5.5. Partie d'une représentation par valeur réelle. 149
Figure 5.6. Partie d'une représentation binaire. 149
Figure 5.7. Partie d'une représentation par symbole. 149
Figure 5.8. Exemple de vue par patient "Patient Id = 75". 153

Figure 6.1. Modèle proposé pour de la prise de décision médicale. 162
Figure 6.2. Processus de représentation de données. 164
Figure 6.3. Processus de réduction de dimensionnalité.....167
Figure 6.4. Visualisation de la réduction 3D. (a)PCA, (b)KPCA, (c)MDS, (d)TSNE. 171
Figure 6.5. Visualisation de la réduction 3D du dernier fold de test. (a)PCA, (b)KPCA, (c)MDS et (d)TSNE. 172
Figure 6.6. Comparaison de FM des classifications sur les données LRTSNE en fonction des distances choisies. 177

Liste des tableaux

Table 2.1. Matrice de confusion. 51
Table 2.2. Matrice de confusion multi-classes. 53

Table 3.1. Breakpoints de division de l'espace sous la courbe gaussienne. 85

Table 5.1. Exemple de distances minimales. 133
Table 5.2. Exemple de table de notification. 134
Table 5.3. Marquage par valeur réelle. 135
Table 5.4. Marquage binaire. 135
Table 5.5. Marquage par symbole. 136
Table 5.6. Linéarisation par événement. 136
Table 5.7. Linéarisation par chronologie. 137
Table 5.8. Prototype de la vue par patient. 141
Table 5.9. Statistiques des événements numériques. 143
Table 5.10. Statistiques des événements nominaux. 143
Table 5.11. Statistiques des longueurs des séries maximales et du nombre de colonnes produites. 144
Table 5.12. Exemple de marquage des événements numériques. 145
Table 5.13. Exemple de marquage des événements nominaux. 145
Table 5.14. Statistiques globales de l'inertie intra et interclasse. 147
Table 5.15. Techniques et nombre de clusters choisis. 148
Table 5.16. Représentation par symbole du patient identifié par id = 75. 151

Table 6.1. Statistiques du dataset. 170
Table 6.2. Statistiques d'observation après l'étape de transformation. 170
Table 6.3. Résultats de FM pour toutes les classes. 173
Table 6.4. Résultats de FM globale. 174
Table 6.5. Comparaison de la classification en fonction des distances choisies. ..174

Table 6.6. Pourcentage du temps de classification écoulé sur les matrices LDMP et LRTSNE en millisecondes pour la classe AD. 175
Table 6.7. Approche proposée par rapport à la recherche actuelle.....179

Introduction générale.

Penser à offrir des traitements efficaces constitue l'un des objectifs principaux de la médecine. De ce point de vue, la proposition des nouvelles solutions et l'amélioration de celles existantes a constamment progressé au cours des décennies passées. Ainsi, la considération des signes symptomatiques ne reste pas l'axe principal de diagnostic des maladies et d'autres facteurs se sont imposés. Beaucoup de ces facteurs jouent un rôle primordial sur la prévention, l'identification et le développement des maladies chez les patients, mais également sur les prescriptions de traitements et de médicaments.

Avec le temps, l'innovation des nouvelles technologies a poussé l'amélioration des environnements de travail des praticiens de la médecine, et a permis de leur offrir des outils qui peuvent produire et collecter d'énormes ensembles de données concernant les patients. Ce grand volume de données se justifie par le niveau de détail des observations et des informations enregistrées. Cela inclut l'historique médical, l'environnement de vie, la démographie et les données génétique si elles existent. Généralement, le stockage de ces données repose sur des supports électroniques. Le volume de données et les variétés de leurs types implique la recherche et l'adoption des solutions applicatives efficaces. L'ensemble des données correspondant à chaque patient est appelé un profil de patient, et sa forme de stockage électronique est appelée **Electronic Health Records (EHR).**

La médecine personnalisée (en anglais **Personalized medicine (PM)**) est une alternative innovante et efficace de la médecine traditionnelle centrée sur une attitude commune pour tous les patient (**one fits for all**) à une attitude plus adaptée aux profils des patients (**the right drug for the right person**) (Becquemont et al., 2012; Fournier et al., 2021). Ces profils peuvent contenir des informations génétiques, des biomarqueurs clés, l'historique des traitements, les facteurs environnementaux et les préférences comportementales (Pfizer, 2015). A la lumière de ces informations, la médecine personnalisée a pour objectif la prescription de traitements spécifiques et thérapeutiques les mieux adaptés à chaque individu (Jain, 2015) avec la bonne dose, au bon moment pour la bonne durée (Barlesi et al., 2014). Plusieurs noms ont été attribués à cette médecine, notamment : « la médecine personnalisée, stratifiée,

individualisée ou même P4 (personnalisées, prédictives, préventives et participatives) » (Vukobrat et al., 2016) plus récemment.

L'exploration des données de la PM par des techniques de datamining est l'axe général de nos recherches. Notre thèse vise l'extraction des connaissances cachées derrière ces sources de données issues de la médecine personnalisée. Durant nos études, nous avons identifié plusieurs axes d'application de cette tâche. Ce sont : le rapportage de données des patients, l'identification des facteurs dominants de la maladie, la prédiction des maladies et des effets indésirables des médicaments prescrits, la stratification et la classification des patients. L'automatisation du processus de la prise de décision médicale à base de données de la PM constitue la branche commune qui englobe ces différents axes applicatifs.

Les sources de données de la PM peuvent inclure des données structurelles, semi-structurelles et non structurelles, avec des types de données numériques, nominales, booliennes et des dates. Elles peuvent inclure également des textes, des images, des enregistrements audios et vidéos. Cette hétérogénéité de données est le principal obstacle lors de l'exploration de ces sources de données. Les opérations de transformation de données généralement appliquées pour générer une représentation globale en fonction de plusieurs types de données conduisent souvent à la perte de données et d'information. Penser à résoudre une partie de cette difficulté constitue notre première contribution. Ceci nous pousse à développer un modèle de représentation de données de la PM afin de les préparer aux différentes tâches de classification.

En plus de la perte de donnés et de l'information, un autre problème peut être rencontré lors de la classification de données. La meilleure série des traitements à appliquer durant la modélisation de la classification est un autre défi qui nécessite une solution. Par l'application de la première contribution sur un ensemble de données de la MP, nous proposons dans une deuxième contribution le développement d'un modèle de prise de décision médicale basé sur la classification des patients. Cette fois notre processus vise à simplifier les calculs, minimiser la perte d'information et de

données et choisir la meilleure série des traitements appliqués sur la nouvelle représentation.

Le plan de cette thèse est le suivant :

- Le premier chapitre définit et introduit la médecine personnalisée et se concentre particulièrement sur leurs défis, leurs outils et leurs offres.
- Le deuxième chapitre est une vue globale sur les outils d'exploration de données. Il explique le domaine du Data Mining, les séries temporelles et le Big Data.
- Le troisième chapitre montre certaines opérations nécessaires pour nos travaux ultérieurs tels que la représentation, la réduction et la classification de données.
- Le quatrième chapitre présente les problématiques rencontrées en incluant le problème de la perte de données et de l'information, et le problème de choix de la série des traitements les plus appropriés.
- Le cinquième chapitre montre notre première approche destinée au problème de de la perte de données et de l'information. Il présente notre modèle de représentation de données de la MP par région et par dispersion.
- Le sixième chapitre est l'autre approche développée pour produire un système de prise de décision médicale basé sur l'exploration de données de la MP. Le modèle développé porte sur le choix et d'identification de la série des traitements adéquate.
- La dernière partie de cette thèse est la conclusion générale. Elle résume et conclut nos travaux et nos recherches et présente brièvement nos perspectives pour de futurs travaux.

Références

- Barlesi, F., Longeray, P. H., & Marquet, P. (2014). Ateliers 2014 Table ronde n°1 : Recherche translationnelle, Médecine de précision, médecine personnalisée, thérapie ciblée : science ou marketing?. www.ateliersdegiens.org/wp-content/uploads/Presentation-TR1.pdf .

- Becquemont, L., Bordet, R., & Cellier, D. (2012). La médecine personnalisée : comment passer du concept à l'intégration dans un plan de développement clinique en vue d'une AMM ?. *Therapies, Vol 67*(4), 339-348.

- Fournier, V., Prebet, T., Dormal, A., Brunel, M., Cremer, R., & Schiaratura, L. (2021). Definition of Personalized Medicine and Targeted Therapies: Does Medical Familiarity Matter?. *J. Pers. Med, Vol 11*(26).

- Jain, K. K. (2015). *Textbook of Personalized Medicine - Second Edition*. Humana Press, NY, USA.

- Pfizer Inc. (2015). Personalized Medicine. *Global Policy & International Public Affairs, Pfizer Inc*. https://www.pfizer.com/files/about/Position-Personalized-Medicine.pdf .

- Vukobrat, N. B., Rukavina, D., Pavelic, K., & Sander, G. G. et al. (eds.). (2016). *Personalized Medicine A New Medical and Social Challenge. Europeanization and Globalization, Vol. 2*. Springer Nature.

CHAPITRE 1

Qu’est-ce que la médecine personnalisée ?

Sommaire

1.1 Introduction 18
1.2 Définition 18
1.3 Historique de la médecine personnalisée 21
1.4 Médecine personnalisée et médecine traditionnelle 22
1.4.1 Médecine personnalisée versus médecine traditionnelle 22
1.4.2 Passage de la médecine traditionnelle vers la médecine personnalisée 22
1.5 Défis de la médecine personnalisée 23
1.5.1 Médecine personnalisée et la vue économique 23
1.5.2 Médecine personnalisée et l’aspect juridique 24
1.5.3 Médecine personnalisée et la vue éthique et sociale 25
1.6 Médecine personnalisée et outils informatiques 26
1.6.1 Médecine personnalisée et les données électroniques des patients 26
1.6.2 Exploitation de données de la médecine personnalisée 27
1.6.3 Prise de décision médicale et médecine personnalisée 28
1.6.4 Médecine personnalisée en temps réel 29
1.7 Opportunités de la médecine personnalisée 30
1.7.1 Bénéficiaires de la médecine personnalisée 30
1.7.2 Offres de la médecine personnalisée 31
1.7.3 Avenir de la médecine personnalisée 32
1.8 Conclusion 33

1.1 Introduction

La médecine actuelle emploie les facteurs technologiques, les recherches et les essais cliniques afin d'évaluer des méthodes de diagnostics ou de traitements, dont la plupart sont destinés à toucher une vaste population. Mais il y a beaucoup de cas où les solutions et les traitements ne donnent pas les résultats prévus sur l'ensemble des patients souffrants de la même maladie. Ceci nécessite le retour sur l'étude des autres facteurs tels que les types de diagnostics, les particularités biologiques et génétiques, l'environnement de vie et l'historique médicale de chaque patient, et conduit à la personnalisation individuelle des traitements. A cet effet, de nouvelles disciplines et sciences ont participé à l'évolution de la médecine généralement, et les recherches modernes ont réussi à participer efficacement aux soins de santé et à l'établissement des solutions inventé**e**s. À partir de ces développements et sur la base de données obtenues sur les patients et les participants, est apparue la médecine personnalisée (MP), considérée comme l'un des moyens actuels les plus importants pour faire face aux maladies, en particulier les plus difficiles, comme le cancer.

Tout au long de ce chapitre, nous essayons de définir la médecine personnalisée (MP) et de son historique, ses actualités, ses offres et d'autres éléments importants, y compris de son avenir.

1.2 Définition

La médecine personnalisée est une alternative innovante et efficace à la médecine centrée sur une attitude commune pour tous ***(one fits for all)*** à une attitude adapté**e** aux profils de patients ***(the right drug for the right person)*** (Becquemont et al., 2012; Fournier et al., 2021), dont chaque profil peut contenir des informations de compositions génétiques, des biomarqueurs clés, l'historique de traitements, les facteurs environnementaux et les préférences comportementales (Pfizer, 2015). Sur la base de ces informations**,** la médecine personnalisée essaie de prescrire les traitements

spécifiques et thérapeutiques les mieux adaptés à l'individu (Jain, 2015) avec la bonne dose, au bon moment, et pour la bonne durée (Barlesi et al., 2014). La médecine personnalisée est un concept à plusieurs noms parmi le**sq**uelles ***« la médecine personnalisée, de précision, stratifiée, individualisée ou même P4 (Personnalisées, Prédictives, Préventives et Participatives)*** » (Vukobrat et al., 2016 ; Gauld et al., 2021). La médecine personnalisée est orientée vers le profil du participant qui comporte des données et des informations individuelles, au lieu de s'orienter vers les symptômes qui sont généralement communes pour la plupart des patients (Rauch et al., 2019).

Les aspects décrits par le terme de la médecine P4 résument globalement le cadre fonctionnel de la médecine personnalisée (Leroy et Mauricio, 2012; Slim et al., 2021) :

- Personnalisées : Naturellement, il existe une différence importante des informations génétiques entres les personnes. Cette différence exige la prise en charge des participants d'une manière séparée. Par conséquent, la séparation pousse vers une approche de l'individualisation des traitements. Par l'aspect personnalisé, cette médecine dépasse la vue figée aux patients et l'attitude commune pour tous.
- Préventives : Pour une personne donnée, et par la détection et le suivi des éléments principaux conduisant à la possibilité d'apparition d'une maladie donnée ou son développement futur, la MP mesure les risques et permet parfois une prévention anticipée. En effet, la numérisation de données des participants et les connaissances antérieures des maladies jouent un rôle primordial pour le processus de prévention et la description des traitements adéquats.
- Prédictives : La description des médicaments avec la MP repose principalement sur les données individuelles des patients, en particulier les facteurs clés, qui incluent certains détails génétiques. L'utilité de cette implication est d'évaluer l'adéquation des traitements proposés et d'éviter tous résultats et effets indésirables.
- Participatives : Cette aspect implique la participation des patients à la gestion de leurs données. Cette gestion repose sur les connaissances théoriques médicales acquises en plus de leurs expériences et de l'auto-surveillance de leurs

vies et les événements en relation. Avec le développement technologique des systèmes de santé devenues plus efficaces en matière de contrôle et de surveillance des patients à distance, ce développement facilite la participation des patients par leurs informations à base de capteurs électroniques connectés et permet un meilleur suivi des habitudes quotidiennes des individus tel que le sommeil, la nutrition et les exercices sportifs.

Pour plus de lisibilité, la Figure 1.1 résume et montre les points essentiels des quatre aspects de la médecine P4 comme suit :

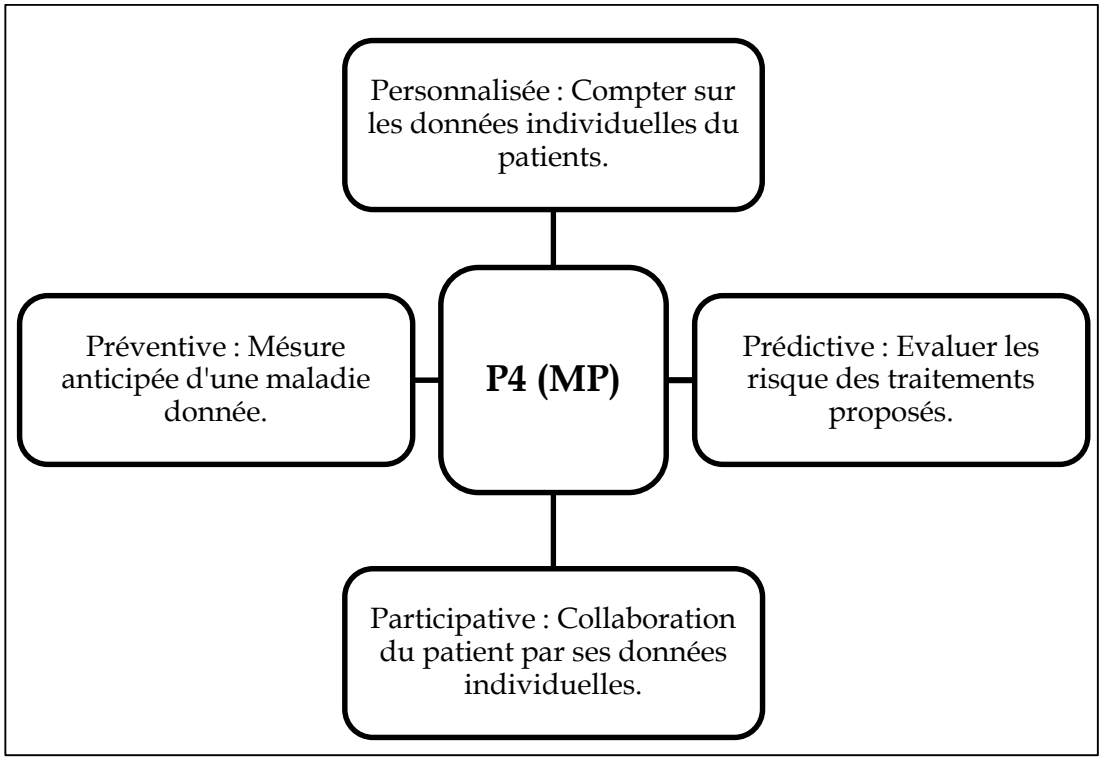

Figure 1.1. Les quatre aspects de la médecine P4 (MP).

Sur le plan de la nomination de cette médecine, la terminologie n'a pas encore été uniformisée (Jørgensen, 2019). Globalement, les deux nominations les plus utilisées sont "la médecine de précision" et "la médecine personnalisée". Certains spécialistes et praticiens préfèrent la terminologie "la médecine de précision". Leur justification repose sur la nécessité de confronter la vision parfois trop exagérée de l'individualisation du traitement et le dépassement de la vision de la personnalisation

à travers des traitements et des préventions destinées uniquement à un seul individu. Ainsi vus, ils impliquent que toutes les prescriptions de traitements sont personnalisées selon le cas traité. Cela implique l'intuition que toutes les médecines sont personnalisées, et la terminologie "la médecine de précision" semble être la meilleure description pour identifier cette médecine et mettre en évidence sa diversité. *Comme un point de consensus dans cette thèse, nous utilisons ici les deux termes comme synonymes et sans distinction.*

1.3 Historique de la médecine personnalisée

Beaucoup de concepts fondamentaux de la médecine personnalisée existent depuis longtemps, mais le terme en soit est apparu dans les travaux de professeur Kewal K. Jain en 1998 pour la première fois (Jain, 2015). Par exemple, après des siècles de médecine primitive qui interprétait les maladies selon des idées surnaturelles et proposait des traitements contenant de la magie et des incantations, les médecines égyptienne et mésopotamienne sont apparues. Ces deux dernières ajoutaient plusieurs rituels, et utilisaient également de la magie, des incantations mais également des médicaments extraits de la nature. La médecine indienne a été l'une des raisons de l'émergence du concept de soins de santé individualisés, elle s'est concentrée sur le diagnostic et la prévention, et adopte la méditation thérapeutique et les médicaments végétaux et animaux. La médecine traditionnelle chinoise qui existe depuis 3000 avant JC, prend au sérieux les variations entre les personnes. Elle peut appliquer des protocoles de traitement différents pour des individus qui ont les mêmes symptômes. D'autres concepts sont nés de l'émergence et du développement de la médecine grecque et arabe ancienne. En effet, l'émergence de la médecine empirique basée sur les expériences pratiques et la stratification des patients est une phase importante qui a offert d'autres outils pour la médecine personnalisée (European Society for Medical Oncology [ESMO], 2014). De manière générale, les concepts de la médecine personnalisée accompagnent le développement de la science médicale depuis l'Antiquité. Kewal K. Jain (Jain, 2015) a dressé un tableau descriptif sur l'évolution des

concepts et les repères en relation directe par la médecine personnalisée. La principale observation est que les progrès et les découvertes se sont développés rapidement au cours du siècle dernier. Cet avancement rapide a été poussé par les développements accélérés de la technologie et l'efficacité des méthodes d'analyse et de traitement des données de tailles importantes et leurs précisions, surtout le domaine de la génétique.

1.4 Médecine personnalisée et médecine traditionnelle

1.4.1 Médecine personnalisée versus médecine traditionnelle

D'une manière générale, la médecine traditionnelle est fondée sur la méthode essai-erreur. Selon les symptômes du patient, le médecin peut appliquer ou proposer des diagnostics et des traitements probablement adoptés pour une telle situation. S'il s'agit d'une prescription d'un médicament, le médecin précise la dose en fonction du poids du patient. Tant que le médicament n'a pas de résultats positifs, le médecin propose d'autres diagnostics, change le dosage et parfois prescrit d'autres médicaments. De l'autre côté, la MP est basée sur les diagnostics détaillés et les plus exacts de la maladie considérée chez le patient. Dans cette médecine, le profil du patient (incluant surtout des données génétiques) est le point focal du diagnostic et de la prescription du traitement (Vukobrat et al., 2016).

1.4.2 Passage de la médecine traditionnelle vers la médecine personnalisée

Le passage vers la MP a rencontré plusieurs obstacles qui ont quelque peu ralenti son expansion et son adoption en général. Parmi ces obstacles (Vukobrat et al., 2016) :

- Le succès remarquable des médicaments qui ont été développés selon une attitude commune pour tous (médecine pour tous), et son impact inhibiteur sur l'orientation vers l'industrie des médicaments personnalisés.

- Les difficultés de l'adoption de l'approche personnalisée à cause des réglementations imposées par le passé et la lenteur de leur modernisation pour l'adaptation avec cette médecine.
- L'opportunisme et l'économie irrationnelle reposent sur des examens médicaux et des médicaments coûteux. Cela pousse leurs acteurs à banaliser le rôle du diagnostic et de la prédiction des maladies.
- L'adoption de la médecine par essais et erreurs comme une habitude chez les médecins, au lieu de les adapter à la médecine personnalisée.

1.5 Défis de la médecine personnalisée

De nombreux facteurs ont contribué au succès de la médecine personnalisée et à l'expansion de son adoption, notamment le développement technologique, les besoins en matière de santé et d'autres facteurs. Tous ces facteurs n'éliminent pas le fait qu'il y a des défis que doit surmonter cette médecine.

1.5.1 Médecine personnalisée et la vue économique

La vue globale sur l'économie de la MP demande et exige des investigations importantes, et nécessite des coûts de santé de plus en plus élevés. Les investissements dans cette médecine rencontrent plusieurs problèmes tels que (Chule et al., 2021) :

- le choix des tests qui ont un effet positif sur le coût de traitement,
- les craintes du coût total des tests de diagnostic moléculaire même s'il est raisonnable individuellement,
- le problème de respect des standards appliqués au niveau des protocoles et l'assurance de fournir des soins adéquats à chaque patient en fonction des résultats de ses tests,
- les craintes de mauvaise utilisation des informations de test sur les patients durant l'étape de test et de développement et les nécessités en matière de l'infrastructure obligatoire.

Ces problèmes ont ralenti l'avancement rapide de la médecine personnalisée, mais les gouvernements des pays développés essaient toujours de les pousser par la politique de soutien des investisseurs et la fourniture des ressources fondamentales comme le financement des plateformes de séquençage génomique, et les concours des projets de recherches.

D'autre part, les tests diagnostiques pour la médecine personnalisée varient du coût favorable au coût plus élevé et en nombre de fois de l'application des procédures de test, ce qui impose un problème pour les fournisseurs et l'adoption de ces tests par les systèmes de remboursement. Par exemple, le test de risque de cancer de côlon nécessite une fréquence d'application de coloscopies trois fois élevé que la normale, mais il reste meilleur économiquement par rapport à un test de diagnostic moléculaire qui sera très cher pour chaque patient. Un autre exemple est le test diagnostic génique Oncotype Dx pour les patientes du cancer de sein, utilisé pour évaluer la probabilité de bénéficier un traitement Chimiothérapie. Malgré le coût élevé de ce test, mais avec la diminution du nombre des patients qui utilisent ce traitement, le revenu acquis peut rester important (Davis et al., 2009).

1.5.2 Médecine personnalisée et l'aspect juridique

Face à la modernité de la médecine personnalisée, il est devenu nécessaire d'actualiser et de réviser les lois qui organisent les systèmes de santé. La planification des révisions doit prendre en considération les avancements appliqués et les différences avec la médecine traditionnelle. En pratique, la médecine personnalisée demande plus de diagnostics surtout avec le processus de stratification des malades. De l'ensemble de ces diagnostics il en résulte un coût plus élevé. Les traitements nécessitent aussi des médicaments au coût par fois élevé. A l'inverse, la médecine traditionnelle comporte moins de diagnostics et les médicaments sont précédemment approuvés selon leurs efficacités sur un ensemble de malades par des tests selon un ensemble de normes. Les règlements qui mesurent les prix, les droits, les obligations et même les procédures de l'industrie des produits médicaux et pharmaceutiques doivent être reformulés et actualisés pour une bonne pratique de la médecine personnalisée. Puisqu'elle est une

médecine participative, elle doit avoir des solutions juridiques au problème de protection des données quel que soit la nature de ces dernières (Vukobrat et al., 2016). Pour devenir plus complète, l'aspect juridique doit prendre en compte le collecte, le traitement, la circulation des données au niveau international et au niveau local, y compris l'anonymisation des patients la plus sûre. De plus, toute modification de la politique du système de santé, notamment en ce qui concerne la confidentialité des données, nécessite des détails juridiques spéciaux, car cela inclut l'information du patient et son approbation des nouveaux amendements. La vision juridique doit couvrir aussi les autorisations et les exigences sur les analyses génétiques soit dans ou dehors le domaine médical, dans le travail et pour les assurances.

1.5.3 Médecine personnalisée et la vue éthique et sociale

La discrimination d'accès à la médecine personnalisée représente l'un de ses gros défis. Quand on parle de cette médecine, on doit parler de l'égalité sociale, puisque les situations socio-économiques des individus ou des patients sont différentes et l'implication des diagnostics, séquençage et traitements génétiques voire même l'éloignement géographique des centres et des plateformes spécialisées constituent un obstacle pour les patients pauvres. Un autre type de discrimination susceptible d'être posé, est la stratification des patients en groupes et la négligence de toucher uniquement quelques groupes par rapport aux autres, avec la justification de réponses négatives ou d'effets secondaires indésirables ou un nombre de cas très limité comme les maladies rares. Des inégalités d'accès à l'information de la médecine personnalisée peuvent se trouver au niveau de certaines régions territoriales. Cela impose la recherche de solutions pour diffuser l'information et prendre les préoccupations des patients et des citoyens en considération. A titre d'exemple, certains sites internet sont destinés au grand public. Ils transmettent les informations des examens médicaux, ce qui est un bon exemple de ces solutions. Le manque de connaissances et la crainte jouent aussi un rôle important dans le niveau culturel faible concernant cette médecine, et plus précisément dans le domaine de la génétique. Ceci impose l'amélioration des programmes éducatifs avec le renforcement des sujets et des matières en relation avec la santé. Les instances publiques incluent les associations, les

organisations, les agences, les chaines télévisées et autres qui doivent participer au perfectionnement et au succès de la politique culturelle couvrant la médecine personnalisée (Claeys et Vialatte, 2014).

1.6 Médecine personnalisée et outils informatiques

Plusieurs aspects reflètent la relation étroite entre l'outil informatique et la médecine personnalisée. Ces aspects peuvent toucher le stockage et le traitement des données, le bénéfice pratique, la prise de décision médicale, etc. A titre d'exemple et sans limitation, nous pouvons citer les aspects importants décrits dans les sections suivantes.

1.6.1 Médecine personnalisée et les données électroniques des patients

Les données de santé de chaque patient sont considérées comme un point de démarrage et dominant pour la médecine personnalisée. Tous les types de données et leurs origines que ce soit le profil génétique ou protéique, les informations cliniques et les diagnostics ou l'environnement de vie, peuvent améliorer une décision médicale quand elles s'associent entre elles. Cela permet de prédire des maladies lors de ces analyses anticipées, ou bien encore prédire la réaction de patient ou d'un organe avant la prescription d'un traitement. Les données d'un seul individu peuvent être grandes en termes de volume et de nombre, et hétérogènes en terme de genèse, de composition et de structure, et nécessitent parfois un partage entre les médecins et les professionnels de la santé pour bien intervenir.

Les facteurs qui accélèrent l'adoption de l'idée de dossier électronique médicale de patient « **Electronic Health Records (EHR).** » ou « **Personal Health Records (PHR).** » sont :

- L'énorme quantité de données et leurs qualités,
- les méthodes d'analyse,

- la nécessité d'intervention de plusieurs professionnels de santé dans des régions différentes et éloignées,
- l'apparition des systèmes d'informations incluent les réseaux informatiques,
- les recommandations des spécialistes.

Le dossier électronique médical (EHR) est le fichier informatisé contenant toutes les données médicales et parfois démographiques du patient. Il constitue une fenêtre sur l'historique clinique, médicamenteux, diagnostics, environnement de vie social et économique et géographique, même sur le profil génétique (Murdoch et Detsky, 2014; Pham et al., 2017). Cet ensemble de données doit disposer d'un accès sécurisé, partagé et maintenir une traçabilité pour une meilleure utilisation.

1.6.2 Exploitation de données de la médecine personnalisée

La médecine personnalisée vise à prédire l'apparition d'une maladie ou les effets secondaires d'un médicament donné pour l'un des participants. Nous entendons ici par les participants, toutes les personnes qui ont participé par leurs données, que ce soient des patients ou non. La stratification des participants en MP a pour but de pousser le maximum possible le partitionnement des individus pour former les plus petits groupes des personnes possédant les spécificités les plus similaires et qui peuvent être traitées de la même manière. Cette tâche peut conduire à la formulation d'un groupe qui ne possède qu'un seul participant ou patient, ce qui figure l'aspect d'individualisation des traitements. Pour faire cette stratification, la MP utilise globalement les données des participants et les biomarqueurs clés spécifiquement. L'exploitation d'un seul biomarqueur ne suffit généralement pas pour la réalisation de cette tâche. La sélection de sous-groupes de biomarqueurs clés, qui peut comporter des dizaines d'éléments parfois parmi un ensemble de centaines ou des milliers des biomarqueurs, nécessite des outils spécialisés. Globalement, ces derniers constituent tout l'outillage informatique nécessaire et cela inclut le matériel, les logiciels et les algorithmes sans oublier la bio-informatique et ses méthodes.

L'autre tâche basée sur les données de la MP est la production des modèles prédictifs ou de classification sur la base des profils des participants. Le niveau de détail important des données dans ces profils exige l'extraction et la réutilisation des connaissances cachées pour les réutiliser sur les nouveaux patients ou participants généralement. Les particularités de cette source de données EHR (tels que le volume, la qualité, le type et la structuration de données, ...) et les futurs besoins en résultats (la précision, le temps de repense, rapportage de données, ...) imposent des contraintes sur le choix des techniques appliquées que ce soit de Data Mining, de l'Intelligence artificielle, de statistique ou de Big Data.

La validation de tels modèles est nécessaire afin de confirmer leurs précisions et leurs exactitudes. Sur le même ensemble de données de construction de modèle, il est important d'appliquer une validation croisée, ce qui permet une évaluation globale du modèle sur les différentes parties de données crée successivement. De plus, la validation du modèle sur une partie des données indépendantes de celles qui ont été utilisées pour produire le modèle, met un autre accent sur l'exactitude et l'efficacité du modèle par rapport à tous les profils de cette source de données. La validation finale des modèles produits doit être examinée dans des études cliniques prospectives.

1.6.3 Prise de décision médicale et médecine personnalisée

Prendre un patient en charge par les professionnelles et les praticiens de la santé induit généralement à la prise d'une décision médicale. Nous pouvons définir la prise d'une décision médicale comme le processus conduisant à la formulation ou la planification des diagnostics ou des traitements à base des valeurs, des préférences et des croyances du décideur, et souvent l'incorporation des préférences du patient (Gellman et Turner, 2013).

Fournir plus des spécifications, des diagnostics, des informations et des connaissances médicales améliore de plus en plus la décision médicale voulue. Les qualités et la quantité de données disponibles à travers de la MP, ainsi que les recherches avancées à ce stade rendent une décision plus précise et plus

individualisée. Mais, pour la même maladie, les décisions médicales peuvent comporter des efficacités variables à cause des choix disponibles et de l'avis des professionnels et des spécialistes de la santé. Ces avis sont affectés directement par les qualités des formations et les expériences des équipes dès le début de carrière professionnelle.

Un autre point qui peut affecter la précision de cette décision est l'incorporation des préférences du patient, ce qui invoque la notion de la prise de la décision partagée. Cette dernière est définie comme le processus de collaboration, d'échange et de discussion entre le professionnel de la santé et le patient pour décider d'une décision médicale éclairée, et acceptée mutuellement dans un accord commun autour de la santé individuelle de ce patient, sans oublier ses données incluant ses préférences et ses attentes (Grad et al., 2017). L'implication des préférences du patient pour la prise d'une décision médicale repose sur la manière de les extraire, le positionnement et le rôle du professionnel de santé pour une contribution optimale avec le patient, jusqu'au moment où le patient peut lui-même participer à la prise de décision.

La collaboration entre les professionnels est un autre point important, puisqu'il y a des cas où des patients nécessitent l'intervention d'un autre professionnel ou même d'autres équipes de soins. La prise d'une décision dans telle situation nécessite l'observation, l'évaluation et l'adaptation du processus de collaboration d'une façon permanente et de rendre la communication interprofessionnelle optimale.

1.6.4 Médecine personnalisée en temps réel

L'évolution des algorithmes et les outils de l'intelligence artificielle (IA) ces dernières années a considérablement amélioré le temps des réponses médicales. Conjointement au développement de la technologie de la connectivité sans fil, de l'imagerie, des biocapteurs et du domaine du Big Data, la médecine personnalisée peut actuellement rendre des décisions médicales en temps réel. A titre d'exemple, les réseaux de neurones profonds permettent une détecter automatiquement certains changements

sur des images pour un organe donné avec des traitements en temps réel. (Biaoyang et Shengjun, 2021).

Malgré les avancements récemment atteints, plusieurs problèmes restent des obstacles pour une meilleure application de l'IA sur la MP. Par exemple, les données sur lesquels les entrainer les systèmes d'IA sont l'un de ces problèmes. Les résultats d'entrainement sont liés aux qualités et la complétude de ces données. La confidentialité de ces données et leur sécurité sont également d'autres problèmes. Elles rendent leur disponibilité réduite à certains organismes et dont la plupart des cas n'existe que dans les pays développés (Mehrabi M et al., 2019).

Généralement, les nécessités de développement des systèmes d'IA pour la MP et l'amélioration du temps de réponse médicale demande plus de travaux. Pour cette raison, les futures applications exigent une forte coordination entres des spécialistes de l'intelligence artificielle, les médecins et tous les professionnels des autres domaines impliqués par la MP.

1.7 Opportunités de la médecine personnalisée

La médecine personnalisée a prouvé son utilité exceptionnelle à l'échelle mondiale. Ses opportunités peuvent toucher de multiples bénéficiaires, présenter des offres spéciales et présager d'un bel avenir.

1.7.1 Bénéficiaires de la médecine personnalisée

Par l'adoption de la MP et de ses utilisations, plusieurs domaines en relation peuvent en bénéficier et se développer. Plus que le profit des sociétés de la production des équipements impliqués, il y a principalement trois bénéficiaires : le domaine de la pharmaceutique, les patients et le système de santé (Ghule et al., 2021).

La MP a ouvert une nouvelle porte de profit pour les sociétés de la pharmaceutique. Elle permet aux sociétés l'utilisation des données et des informations des patients pour innover et produire de nouveaux médicaments personnalisés. Cela permet également l'adoption de nouveaux protocoles et l'application des nouvelles

approches. Par conséquent, le domaine de la pharmaceutique sera l'un des principaux bénéficiaires, en particulier le commerce et le marché de ces sociétés.

Pour les patients, les traitements et les médicaments seront devenus plus sûr et leur utilisation pratique sera plus assurée. De plus, le patient sera plus confiant par le fait d'absence d'effets indésirables des traitements. En outre, les patients peuvent bénéficier de la possibilité de la prévention des maladies et l'amélioration de leurs vies surtout par l'augmentation des options des traitements.

Les systèmes de santé peuvent bénéficier de la rapidité de l'identification des traitements les plus appropriés pour un patient donné et leurs dosages parfaits. À travers les traitements guidés et les plus sûrs, la société peut bénéficier des réductions dépenses de santé par rapport aux frais supplémentaires générés avec la médecine traditionnelle.

1.7.2 Offres de la médecine personnalisée

Les avantages de la médecine personnalisée apparaissent sur les niveaux individuels et professionnels. Pour un individu, les traitements seront plus précis et plus efficaces, avec des soins plus performants. Pour ce faire, ils prennent les profils génétiques et biologiques en considération et même l'environnement de vie social, culturel et économique. Ainsi, le suivi d'un individu sur le temps évoque le rôle prédictif des maladies et guide le processus pour une thérapie personnalisée.

En exploitant la stratification des participants en sous-groupes selon leurs caractéristiques génomiques, ainsi que les résultats de la pharmacogénétique et l'étude des réactions des individus aux médicaments, cette médecine permet aux professionnels de la santé de mieux planifier les traitements et prescrire les médicaments les plus adaptés.

Comme un résultat direct de développement de la médecine personnalisée et les facteurs technologiques, une décision thérapeutique peut se révéler plus sûre avec un taux d'efficacité et de succès très élevé pour les spécialistes et les professionnels. Cela stimule leurs confiances en soi, et pousse vers une meilleure relation médecin-malade.

La médecine personnalisée apporte un changement important sur le plan de la recherche, et plus précisément elle a offert la reclassification des maladies basées sur les différences biologiques et moléculaires des organes au lieu de la classification basée sur les caractéristiques symptomatiques (Vukobrat et al., 2016). Ces nouvelles bases permettent aux chercheurs et aux professionnels de mieux comprendre les mécanismes de genèse des maladies et leurs développements.

Les avancements aux stades de la MP conduisent à l'apparition de débats sur la protection des données des participants, ce qui implique le développement et l'apparition de nouvelles approches qui touchent directement à la vie privée et à la confidentialité des particularités des patients. Ces avancements ont conduit aussi à l'innovation de nouveaux outils pour produire, analyser et partager des données médicales volumineuses, et simplifier l'accès aux données pour les praticiens, surtout avec l'adoption de la notion de dossiers électroniques de santé (EHR).

En parallèle, cette médecine a conduit à l'émergence de nouvelles alliances pour l'université en général, les professionnels du droit, les associations de défense des patients et les entreprises du secteur public et privé, y compris avec les sociétés de la pharmacogénétique.

1.7.3 Avenir de la médecine personnalisée

Durant les dernières décennies la médecine personnalisée a rencontré une évolution remarquable. Les sciences de la génomique et de la génétique ont été couplées avec la technologie, et les développements sont actuellement en progression rapide. Les signes de l'évolution remarqués sont très attirants, ce qui rend le futur de la médecine personnalisée encore plus brillant. A titre d'exemple, le coût et le délai de l'opération de séquençage d'ADN (**Acide Désoxyribonucléique**) ont connu une nette diminution ces dernières années, avec des études qui peuvent se réaliser avec des prix raisonnables et dans des délais très courts. Par la suite, ils facilitent l'acceptation de de cette opération par les patients et montrent aux praticiens de la santé plus d'informations sur le profil génétique, ce qui implique plus de personnalisation de

traitements. En outre, les chercheurs travaillent sur l'élaboration de schémas thérapeutiques proactifs, utilisant les génomes séquencés des individus et toutes les informations disponibles pour prédire la santé des nouveau-nés et résoudre les problèmes si nécessaire.

Au fil du temps, des systèmes intégrés sont mis en place pour suivre la santé de chaque personne séparément et pour maintenir sa sécurité, en disposant d'informations adéquates et abondantes. Ces systèmes se propagent rapidement et leur efficacité se développe de plus en plus grâce aux technologies modernes. En particulier le monde des technologies connectées permet de mieux suivre les activités des personnes, les systèmes alimentaires suivis et même leurs environnements de vie. En se référant aux données de ces systèmes et aux informations génétiques, la médecine personnalisée pourra assurer des décisions médicales efficaces. Par conséquent, les patients seront plus autonomes, plus maîtres et plus impliqués dans leur destin médical. Les technologies avancées sont les autres clés du futur de la médecine personnalisée, et leur évolution sera un facteur de motivation. Parmi eux, nous pouvons citer la nanotechnologie et la capacité computationnelle et de stockage de données. En général, tous les signes montrent que la médecine personnalisée occupera une place bien méritée dans les soins de santé des personnes au futur, pour la prédiction, le traitement et même la prédiction des maladies.

1.8 Conclusion

L'amélioration des soins de santé au fil du temps, motivée par la différenciation des caractéristiques des individus a conduit à l'émergence de la médecine personnalisée. La médecine personnalisée à travers ses autres noms de précision, stratifiée ou P4 (**Personnalisées, Prédictives, Préventives et Participatives**) a adopté le profil du patient individuel comme un point essentiel pour la prévention, la prédiction et le traitement des maladies.

Par son couplage à d'autres sciences et à d'autres technologies de pointe, cette médecine a fait de grands progrès pour bien préciser le traitement et mieux améliorer les performances médicales. Éviter les effets indésirables est également l'un des aspects cibles les plus importants sur lesquels cette médecine travaille. Outre les avantages généralement constatés sur la santé des patients et des participants impliqués dans la médecine personnalisée et la production de médicaments personnalisés, le côté économique de cette médecine est un autre de ses avantages, étant plus précis et plus efficace économiquement.

Les aspects éthiques, sociaux, juridiques et économiques sont les vues qui encadrent la médecine personnalisée et son adoption par tous. Il faut faire beaucoup plus pour mettre à jour les lois et moderniser les programmes éducatifs, ainsi que pour les promouvoir socialement et culturellement. Il est également nécessaire de faciliter son adoption et sa pratique institutionnelle. L'autre devoir est d'établir les bases et les structures de soutien de cette médecine, et de garantir leur distribution équilibrée afin de faciliter l'accès pour tous, que ce soit pour les patients, les médecins et même pour les étudiants. Les grands pas de la médecine personnalisée et les résultats obtenus confirment son importance, ce qui rend les défis restants comme une feuille de route pour atteindre les objectifs souhaités au futur.

L'EHR (**Electronic Health Records**) du patient au stade de la médecine personnalisée comporte toutes les données détaillées et parfois quantitatives et a une grande importance dans la prise d'une décision médicale. Mais les qualités et les quantités de données exigent l'intervention d'autres disciplines telles que l'informatique pour analyser et automatiser les décisions. Par la disponibilité de données de la médecine personnalisée et l'existence des solutions d'exploration et de datamining sur ces données, la décision médicale automatisée peut participer à l'amélioration de la vie des patients spécialement et de l'humanité globalement.

Références

- Barlesi, F., Longeray, P. H., & Marquet, P. (2014). Ateliers 2014 Table ronde n°1 : Recherche translationnelle, Médecine de précision, médecine personnalisée, thérapie ciblée : science ou marketing?. www.ateliersdegiens.org/wp-content/uploads/Presentation-TR1.pdf .

- Becquemont, L., Bordet, R., & Cellier, D. (2012). La médecine personnalisée : comment passer du concept à l'intégration dans un plan de développement clinique en vue d'une AMM ?. *Therapies, Vol 67*(4), 339-348.

- Claeys, A., & Vialatte, J. S. (2014). (Accessed April 2021). Rapport n° 306 (2013-2014) au nom de l'Office parlementaire d'évaluation des choix scientifiques et technologiques sur les progrès de la génétique : vers une médecine de précision ? Les enjeux scientifiques, technologiques, sociaux et éthiques de la médecine personnalisée. http://www.senat.fr/rap/r13-306/r13-3061.pdf .

- Davis JC et al. (2009). The microeconomics of personalized medicine: today's challenge and tomorrow's promise. *Nature Reviews Drug Discovery*, Vol 8, 279–286.

- European Society for Medical Oncology (ESMO). (2014). Guide ESMO à l'usage des patients, ESMO Médecine personnalisée. http://esmo.org/content/download/46498/855044/file/ESMO-Medecine-Personnalisee-Guide-Pour-les-Patients.pdf . Septembre 2014.

- Fournier, V., Prebet, T., Dormal, A., Brunel, M., Cremer, R., & Schiaratura, L. (2021). Definition of Personalized Medicine and Targeted Therapies: Does Medical Familiarity Matter?. *J. Pers. Med, Vol 11*(26).

- Gauld, C., Darrason, M., Dumas, G., & Micoulaud-Franchi, J. A. (2021). Personalized Medicine for OSA Syndrome in a Nutshell: Conceptual Clarification for Integration. *Chest, Vol 159*(1).

- Gellman, M. D., Turner, J. R (eds). (2013). *Encyclopedia of Behavioral Medicine,* Springer, NY, USA.

- Ghule, S., Gaikwad, A., & More, M. (2021). Review on personalized medicine: a multifaceted approach towards targeted therapies. *International journal of multidisciplinary educational research. Vol 10*(7).

- Grad, R., Légaré, F., Bell, N. R., Dickinson, J. A., Singh, H., Moore, A. E., Kasperavicius, D., & Kretschmer, K. L. (2017). Prise de décision partagée en soins de santé préventifs : Ce que c'est; ce que ce n'est pas. *Canadian family physician, Vol 63*(9), 377–380.

- Hood, L., & Flores, M. (2012). A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnology, Vol 29*(6), 613-624.

- Jain, K. K. (2015). *Textbook of Personalized Medicine - Second Edition*. Humana Press, NY, USA.

- Jørgensen, J. T. (2019). Twenty Years with Personalized Medicine: Past, Present, and Future of Individualized Pharmacotherapy. *The oncologist, Vol 24*(7), e432–e440.

- Lin, B., & Wu, S. (2021). Digital Transformation in Personalized Medicine with Artificial Intelligence and the Internet of Medical Things. *OMICS: A Journal of Integrative Biology, Vol 25*.

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *Cs.LG,* arXiv: 1908.09635.

- Murdoch, T. B., & Detsky, A. S. (2014). The Inevitable Application of Big Data to Health Care. *JAMA, Vol 309*(13).

- Pfizer Inc. (2015). Personalized Medicine. *Global Policy & International Public Affairs, Pfizer Inc*. https://www.pfizer.com/files/about/Position-Personalized-Medicine.pdf .

- Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. *Stat.ML,* arXiv:1602.00357v2.

- Rauch A et al. (2019). Médecine personnalisée. Bases pour la formation interprofessionnelle prégraduée, postgraduée et continue des professionnels de la santé. *Swiss academies communications, Vol 14*(6).

- Slim, K., Selvy, M., Veziant, J. (2021). Innovation conceptuelle : la médecine 4P et la chirurgie 4P. *Journal de Chirurgie Viscérale, Vol 158*(3), S13-S18.

- Vukobrat, N. B., Rukavina, D., Pavelic, K., & Sander, G. G. et al. (eds.). (2016). *Personalized Medicine A New Medical and Social Challenge. Europeanization and Globalization, Vol. 2*. Springer Nature.

CHAPITRE 2

Introduction au Data Mining

Sommaire

2.1 Introduction 40
2.2 Data Mining 41
2.2.1 Etapes de processus DM 41
2.2.2 Tâches de Data Mining 44
2.2.3 Disciplines incorporées en data mining 46
2.2.4 Critères d'évaluation des modèles 49
2.2.5 Evaluation de la précision des modèles 50
2.3 Data Mining pour les séries temporelles (Time series data mining) 59
2.3.1 Séries temporelles (Time series) 59
2.3.2 Différentes tâches de data mining sur les séries temporelles 60
2.4 Big Data. 64
2.4.1 Couches de Big Data 64
2.4.2 Chiffres et promesses en Big Data 65
2.4.3 Défis du Big Data 66
2.4.4 Techniques de Big Data 67
2.4.5 Types de Base de données NoSQL 68
2.4.6 Impact du Big data sur la médecine personnalisée 68
2.5 Conclusion 70

2.1 Introduction

La croissance du nombre des entreprises au niveau mondial et les évolutions de la technologie pour le stockage de données en terme de quantité et de qualité, ont conduit à l'apparition de techniques de statistique, d'analyse, et d'apprentissage automatique « ***machine learning*** ». Elles permettent d'explorer, de traiter et extraire des informations et des connaissances cachées et utilisables à partir de ces sources de données. Cet ensemble des techniques et de tâches constituent la base sur laquelle repose le domaine de découverte des connaissances dans les bases de données « ***Knowledge Discovery in Databases (KDD)*** » (Shapiro, 1990 ; Fayyad et al., 1996).

Avec les défis d'exploration de données et conformément à l'étape de Data Mining supposée importante, le processus KDD s'est popularisé sous le nom Data Mining (DM). Généralement, ce processus suit certaines étapes afin d'effectuer des tâches bien précises. Les tâches de DM peuvent varier d'une personne à un autre, car différentes contraintes sont adoptées par les chercheurs. D'autre part, par l'application du DM, un modèle sera produit. Par conséquent, la solution nécessite une autre étape importante d'évaluation afin de confirmer les résultats avant la délivrance finale.

Plusieurs formes et types de données existent. Les séries temporelles sont l'une de ces formes et nécessitent parfois des traitements spécialisés. Pour extraire leurs connaissances cachées, le processus de DM peut intervenir. Généralement, les besoins et les objectifs exigés déterminent la liste des différentes tâches possibles.

Au fil du temps les changements sur les données sont devenus importants et fréquents. L'explosion volumique, l'hétérogénéité, la complexité des données, poussés par les évolutions qui accompagnent le réseau mondial d'internet et la production des données structurés et non structurés par les entreprises commerciales et organisation gouvernementales, ont conduit à l'apparition du concept de « ***Big Data (BD)*».**

Plus que l'introduction et la conclusion, ce chapitre sera constitué de trois parties principales. La première partie 2.2 discute le DM, ses étapes, ses tâches et d'autres

éléments essentiels. La deuxième partie 2.3 parlera du DM des séries temporelles. Finalement, la partie 2.4 présentera un résumé général de domaine du BD.

2.2 Data Mining

En 1989 et durant l'atelier de la conférence IJCAI-89, le terme KDD a été inventé pour la première fois (Shapiro, 1990), ce terme réfère au processus itératif (plusieurs passes) d'identification, de reconnaissance et d'extraction des modèles, des expressions ou des structures cachés et inconnues, valables et utilisables sur les données (Fayyad et al., 1996). Alternativement au KDD beaucoup de gens utilisent un autre terme pour désigner ce processus par « ***Data Mining (DM)***» (Roiger, 2017). Il ne constitue qu'une seule étape du processus KDD, mais comme elle est très importante, avec le temps ce terme est devienu le plus utilisé.

Le DM utilise un ensemble d'algorithmes et de techniques afin d'identifier ou de produire les connaissances visées. Pour ce faire, toutes les techniques d'exploration de données utilisent des observations spécifiques pour des sources de données bien déterminées. Le résultat final sera un modèle généralisé qui peut découvrir d'autres sources de données, avec une application pour des nouvelles situations (Roiger, 2017).

2.2.1 Etapes de processus DM

Le processus de DM repose sur une série d'étapes (Fig 2.1), dont chacune est essentielle pour la suivante. Ceci va du démarrage par les données brutes jusqu'à l'obtention des connaissances comme résultat final.

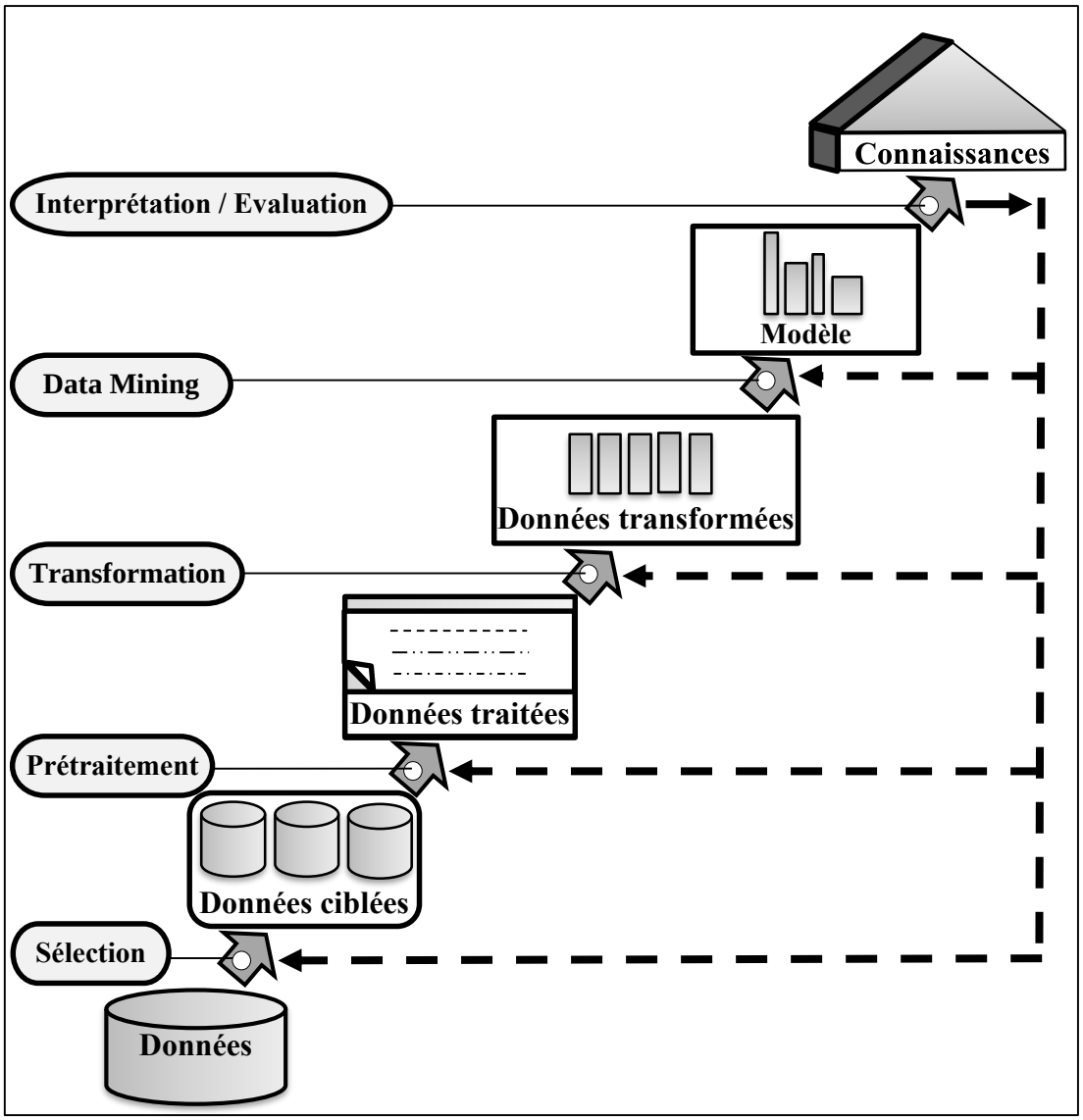

Figure 2.1. Étapes de processus de DM. (Fayyad et al., 1996).

A. Sélection / Intégration et Prétraitement (Preprocessing) : Une fois les objectifs fixés, l'étape de sélection des sources de données et leur ***intégration*** commence par la collecte des données issues d'un ou plusieurs emplacements. Leur taille peut varier de dizaines, centaines, des milliers à des millions d'enregistrements. Ce n'est pas forcément l'existence d'une grosse source de données qui permet qu'un algorithme d'exploration fonctionne bien, mais parfois des sources avec des centaines ou des milliers d'entrées pertinentes suffisent. Les formes de sources de données peuvent varier dans leurs implantations, ce qui implique une variation dans les méthodes d'accès, parmi lesquelles en trouve : les entrepôts de données **« data**

warehouse », la base de données relationnelle, les fichiers plats (Fichier texte) ou les feuilles de calcul et les serveurs de données dans les environnements distribués (Roiger, 2017).

Les données collectées nécessitent parfois une phase de ***nettoyage***. Elle touche les valeurs manquantes, les bruits tels que les valeurs invalides et les erreurs de saisie, et l'élimination des doublons inutiles. Chacun de ces cas a une stratégie de traitement. Par exemple, il y a des situations où il faut supprimer ou négliger les instances qui contiennent des valeurs manquantes, ou les remplir par les valeurs des instances les plus présentes. Ce traitement diffère selon les buts, les qualités et exigences posées (Han et al., 2012).

L'***enrichissement*** des données est considéré comme une autre phase. Elle résulte en l'ajout des nouveaux champs, et garde la plupart du temps le même nombre d'enregistrements. Ce retour justifie généralement par les besoins d'optimisation des stratégies de DM et d'amélioration de l'efficacité et la qualité du résultat final.

B. Transformation de données : C'est l'opération de codage des données, dont les choix sont soumis aux nécessités des algorithmes des traitements futurs. Le processus passe par plusieurs phases lors de cette étape. Citons la ***conversion ou changement des types*** (ex : Date naissance vers Age), la ***normalisation des données*** par l'ajustement des valeurs selon une fonction de transformation, la ***réduction des données***. Cette dernière peut être appliquée par plusieurs stratégies dont le but est de réduire le volume de données avec le maintien de l'intégrité des données d'origine. Par exemple l'utilisation d'un échantillonnage aléatoire ou de segmentation (Clustering) sur les enregistrements, ou par suppression des attributs redondants ou moins pertinents après une analyse de corrélation entre eux ou par le regroupement des valeurs pour certains attributs discrets ont un nombre considérable des valeurs (ex : « jeune » et « adulte » en « senior »).

C. Data Mining : C'est une étape essentielle, coûteuse et souvent difficile à appliquer. Elle implique le choix d'une ***technique*** ou d'un ***algorithme*** précis pour les pratiquer sur une portion de l'ensemble de données de l'étape précédente appelé

« ***Training Set*** ». Une autre portion de données nommée « ***Test Set*** » est utilisée pour tester le modèle développé lors de cette étape. Un troisième sous ensemble de données nommé « ***Validation Set*** » est employé dans quelques applications afin d'estimer les paramètres de classification.

D. Interprétation / Evaluation : L'examen et l'explication des résultats de l'étape de Data Mining est en plein cœur de cette étape. Ils dépendent de la nature de la tâche choisie, et nécessite l'intervention des experts ou l'utilisation de méthodes statistique, ou les deux en même temps. Dans le cas des réponses négatives, le retour à l'étape précédente est possible avec de nouveaux enregistrements et/ou attributs. Le choix des outils appropriés de visualisation et d'analyse a un effet important sur l'interprétation des résultats et sur les décisions prises. La simplicité de l'interprétation graphique et sa qualité, la taille et le nombre des informations présentées, sont tous des facteurs qui peuvent influencer nos choix (Olson et Delen, 2008).

A la fin de processus de KDD et comme résultats, l'ensemble des connaissances découvertes et les modèles construits, peuvent les utiliser sur d'autres données pour aider à la prise de décision, ou bien pour la compréhension et l'interprétation des phénomènes.

2.2.2 Tâches de Data Mining

Il existe beaucoup d'algorithme et méthodes en data mining, techniquement ils sont rassemblés en des ensembles de stratégies selon les critères d'utilisation.

A. Selon l'objectivité : Basé sur l'objectivité comme premier critère d'affiliation des stratégies, nous distinguons quatre techniques (Sagar et al., 2017) :

- **Classification (Discrimination) :** C'est un aspect de l'analyse de données, qui utilise les méthodes d'apprentissage supervisés et les données étiquetées

à l'avance. Elle vise à construire un modèle qui permet de labéliser d'autres données non étiquetées (Sagar et al., 2017).

- **Prédiction :** C'est une technique d'apprentissage supervisée. Elle concerne l'utilisation d'un ensemble d'attributs de données pour prédire d'autres dont les valeurs peuvent êtres inconnues ou pour une future évaluation (Roiger, 2017).

- **Association :** Ce sont des techniques d'exploration par apprentissage non supervisé, afin de découvrir et d'extraire les relations cachées dans les données, et interprétables par la corrélation entre les attributs (Roiger, 2017).

- **Segmentation (Clustering) :** C'est l'opération descriptive des données par apprentissage non supervisé. Elle est basée sur l'utilisation des fonctions d'évaluation et de tests de similarité entres les instances d'une population, afin de construire des groupes homogènes (Roiger, 2017).

B. Selon le type d'apprentissage : L'utilisation d'un type d'apprentissage comme un critère d'affiliation, implique le rassemblement des techniques de data mining en deux mécanismes de travail (Roiger, 2017) :

- **Apprentissage supervisé :** C'est le processus adopté dans une stratégie qui s'appuie sur l'utilisation d'un ensemble d'instances possédants des données d'entrée et de sortie connues, pour apprendre des modèles qui n'utilisent que des données d'entrées pour produire les sorties. C'est la technique employée par les méthodes de classification et de prédiction (Shmueli et al., 2017).

- **Apprentissage non supervisé :** C'est le processus qui permet d'apprendre un modèle adapté à l'ensemble de toutes les données des instances sans séparation (données d'entrées/ données de sorties). Ce type d'apprentissage

est utilisé par les méthodes d'association et de segmentation (Gorunescu, 2011).

C. Selon le caractère du modèle : Le troisième critère basé sur le caractère de modèle final construit, permet aussi de subdiviser les techniques de data mining en deux catégories (Kantardzic, 2011) :

- **Modèle prédictif :** C'est l'application du processus de data mining lors d'une classification ou prédiction, qui utilise les techniques d'apprentissage supervisé sur un ensemble des données avec des résultats connus, pour fournir des modèles décrits par cet ensemble et qui permettent de prévoir les résultats des autres.

- **Modèle descriptif :** C'est l'utilisation d'un ensemble de données, dans un processus de data mining lors d'une application d'association ou segmentation, avec un apprentissage non supervisé, pour fournir des descriptions contenantes des informations nouvelles et essentielles sur l'ensemble de données.

2.2.3 Disciplines incorporées en data mining

Le data mining comporte plusieurs disciplines qui touchent directement notre vie pratique, ce qui démontre son utilité applicative. Ces disciplines impliquent de nombreux domaines de recherche, et exigent l'existence et la disponibilité d'un ensemble de données de problème à étudier. Parmi ces disciplines nous pouvons citer les statistiques, l'apprentissage automatique, les bases de données, la recherche d'information et autres disciplines (Fig 2.2) (Han et al., 2012).

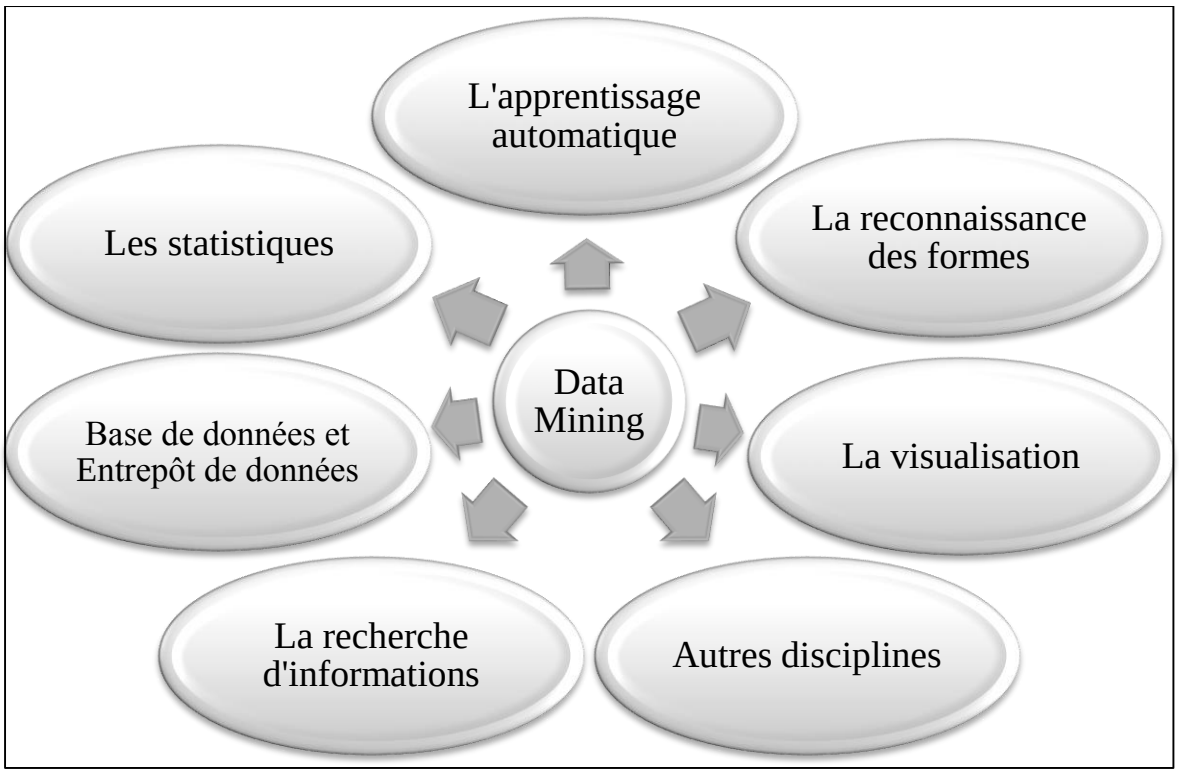

Figure 2.2. Les disciplines incorporées en data mining. (Han et al., 2012).

La statistique est destinée pour l'analyse, l'interprétation et la représentation des données collectées. Cependant, le modèle statistique est l'outil résultant de cette étude. Il comporte l'ensemble des fonctions mathématiques qui essaient de décrire le comportement des éléments d'une classe donnée parmi l'ensemble des données à étudier. Cette description est effectuée en fonction des distributions de probabilité associées aux variables aléatoires de ces éléments. Les recherches du domaine de la statistique utilisent des modèles statistiques sur les données pour produire des outils de prédiction et de vérification des résultats du processus de data mining. Parfois la complexité des modèles statistiques et les quantités de données exigent l'introduction de quelques techniques de data mining pour répondre à certaines nécessités surtout les besoins de réponse en temps réel.

L'apprentissage automatique est également largement associé au data mining. Il est considéré comme un domaine d'étude pour doter les ordinateurs par des solutions qui les rendent capables d'apprendre différentes situations ou pour perfectionner leurs performances. Pour ce faire, il exploite l'ensemble de données et des expériences

déjà acquises précédemment. L'apprentissage supervisé, non supervisé, semi-supervisé et actif sont des types d'apprentissage automatique, exploités par le data mining dans des modèles de traitement de données. Le data mining et l'apprentissage automatique se concentrent souvent sur l'exactitude et la précision des modèles. Mais le data mining adresse en plus l'évolutivité des techniques d'exploration pour des ensembles de données de taille importante et des outils de traitement des types de données complexes.

Les systèmes de bases de données sont conçus pour l'administration des ensembles de données parfois volumineuses. Parfois, le bon usage des technologies évolutives des systèmes de base de données est exploité en data mining pour améliorer son efficacité et son évolutivité sur des ensembles de données qui ont des tailles importantes. Ces technologies peuvent être utilisées lors de l'exploration de données pour gérer le flux de données qui peut être en temps réel. Les entrepôts de données participent à la simplification du processus d'exploration de données. Par l'intégration et la consolidation de données sous forme des cubes de données matérialisés, l'entrepôt de données peut pousser à l'adoption d'approches d'exploration sur les données multidimensionnelles.

La recherche d'information (RI) est une autre discipline incorporée en data mining. Elle se définit comme la science cernée par l'étude des processus et des manières de localisation d'une information ou un document dans un corpus. Ce dernier est formé par un ensemble des documents, dont chacun peut être de nature textuelle ou multimédia. Principalement, les informations recherchées dans un processus de RI sont non structurées et les requêtes de recherche ne comportent que des simples mots clés. La RI implique des modèles probabilistes durant les traitements. Le sac de mots d'un document et les similitudes avec les représentations des autres est un des exemples de ces modèles. Les techniques de recherche et d'analyse d'information (surtout pour les contenues sur le web) ont doté le data mining de nouvelles solutions. Par conséquent, l'intégration de la RI et du data mining est devenue importante et parfois inséparable.

2.2.4 Critères d'évaluation des modèles

Les modèles et méthodes de data mining sont soumis à une collection de critères pour garantir un bénéfice maximal dans leurs applications pratiques. Parmi ces critères nous pouvons citons les éléments suivant (Han et al., 2012 ; Coelho et Ebecken, 2002):

A. La précision de classification (Classification Accuracy) : En général, ce terme est utilisé pour désigner la capacité de bien classifier de nouvelles données par le modèle ou la technique exploitée pour différentes tâches de classification.

Elle mesure le pourcentage des instances de tests bien classés par le modèle de classification en évaluation.

B. La vitesse (Speed) : C'est un critère qui indique l'évaluation autour du coût computationnel lors de la production ou de l'application d'un modèle de classification en data mining.

C. La robustesse : C'est l'aptitude de la classification correcte de données bruitées par le modèle ou la technique en évaluation.

D. La scalabilité (Evolutivité) : Elle désigne l'évaluation de la qualification du modèle de classification construit pour les données de grande quantité. L'évaluation de ce critère se fait sur des ensembles de données avec des volumes croissants.

E. L'interprétabilité : Elle désigne le degré de compréhension et le niveau de conscience informationnelle explicative que le modèle fournit.

F. La fiabilité (Reliability) : Elle désigne l'aptitude de fonctionnement d'un modèle de classification sous certaines conditions et dans une durée déterminée.

G. L'utilité : C'est une manière de mesurer la capacité du modèle de classification à produire des informations ou des connaissances utiles.

2.2.5 Evaluation de la précision des modèles

L'évaluation de la précision de classification de différents modèles en data mining se base sur une collection des métriques à calculer. Bien sûr, ces métriques sont estimées en fonction d'un ensemble de données de test pour le modèle ou la technique de classification déjà entrainés sur un ensemble de données d'apprentissage. Les facteurs de genèse de ces ensembles peuvent toucher directement les qualités des modèles produits. Parmi eux nous pouvons citer la façon d'échantillonner des ensembles de données d'apprentissage et de test, leurs tailles, leurs compositions, leurs qualités. Pour pallier le mauvais usage de ces facteurs pendant l'exploration de données, diverses méthodes d'évaluation ont été proposées et certaines entres eux ont rencontré des succès de part de leurs bons résultats pratiques.

A. Métriques d'évaluation : Elles sont les mesures et les méthodes d'estimation des performances lors de la phase de l'évaluation de la précision et l'exactitude des modèles d'exploration de données. De plus, elles sont parfois utilisées pour la comparaison de la précision retournée. Il existe une collection de métriques d'évaluation et chacune possède une stratégie d'estimation. Ces stratégies emploient les indicateurs résultants de l'application du modèle de classification sur l'ensemble de données de test. La catégorisation des instances de test est un élément essentiel pour l'évaluation. La première catégorie considérée comme **positive** comporte l'ensemble de l'instance étiquetée par la catégorie remarquée comme principale. La deuxième est la catégorie **négative** et représente les instances restantes. Ces deux catégories sont définies pour une classification binaire. En fonction de ces deux classes les quatre indicateurs sont définit comme suit :

- **Vrai positif (True positives "TP") :** Il comprend toutes les instances de test de la catégorie positive qui ont été classées comme positives lors de l'évaluation. Le nombre des instances vraies positives est indiqué par TP.

- **Vrai négatif (True negatives "TN") :** Il comprend toutes les instances de test de la catégorie négative qui ont été classées comme négatives lors de l'évaluation. Le nombre des instances vraies négatives est indiqué par TN.

- **Faux positif (False positives "FP") :** Il comprend toutes les instances de test de la catégorie négative qui ont été classées comme positives lors de l'évaluation. Le nombre des instances fausses positives est indiqué par FP.

- **Faux négatif (False negatives "FN") :** Il comprend toutes les instances de test de la catégorie positive qui ont été classées comme négatives lors de l'évaluation. Le nombre des instances vraies positives est indiqué par FN.

La matrice de confusion (Tab 2.1) c'est une table de statistiques qui englobe tous les indicateurs afin de montrer plus de lisibilité sur l'évaluation. A la base de deux catégories positive (+) et négative (-), les lignes de cette structure figurent les statistiques des instances de test selon la distribution réelle, et les colonnes exposent les statistiques des instances de test selon la distribution résulté par la classification.

Classes		Distribution de la classification		Total
		(+)	(-)	
Distribution réelle	(+)	TP	FN	P=TP+FN
	(-)	FP	TN	N=FP+TN
Total		P'=TP+FP	N'=FN+TN	T=P+N Or T=P'+N'

Table 2.1. Matrice de confusion.

Parmi les métriques les plus célèbres, nous pouvons citer les suivantes (Han et al., 2012 ; Berger et Guda, 2020) :

- **Précision (P) :** Elle compte le taux des instances réellement positives et correctement classées TP par rapport à toutes les instances classées comme positives.

$$\boldsymbol{P = \frac{TP}{TP+FP}} \quad (2.1)$$

- **Rappel (R) :** Elle compte le taux des instances vraiment positives et correctement classées TP par rapport à toutes les instances réellement positives. Le taux de vrais positives (True Positive rate (TPR)) et la sensibilité sont d'autres noms correspondants.

$$\boldsymbol{R = \frac{TP}{TP+FN}} \quad (2.2)$$

- **Spécificité (S) :** Elle compte le taux des instances vraiment négatives et correctement classées TN par rapport à toutes les instances réellement négatives. Le taux des vrais négatifs est un autre nom correspondant.

$$\boldsymbol{S = \frac{TN}{FP+TN}} \quad (2.3)$$

- **F-Mesure (FM) (F-Score) :** C'est une estimation qui combine la précision et le rappel dans une moyenne harmonique.

$$\boldsymbol{FM = \frac{2*P*R}{P+R}} \quad (2.4)$$

- **Taux de reconnaissance (acc) :** Il estime le pourcentage des instances correctement classées (TP + TN) parmi tous les éléments de test T.

$$\boldsymbol{acc = \frac{TP+TN}{TP+FP+TN+FN}} \quad (2.5)$$

- **Taux d'erreur (err) :** Il estime le pourcentage des instances mal classées (FP + FN) parmi tous les éléments de test T.

$$\boldsymbol{err = \frac{FP+FN}{TP+FP+TN+FN}} \quad (2.6)$$

Il existe d'autres métriques pour la classification binaire (Han et al., 2012; Grandini et al., 2020), mais la classification multi-classes utilise aussi un ensemble des métriques plus générales pour ses évaluations. Pour plus de lisibilité la Table 2.2 représente la matrice de confusion multi-classes. Toutes les instances réellement considérées comme des membres de la classe *k* et correctement classées sont indiqués par TP_k. Sur la colonne de cette classe *k*, et sauf les instances TP_k les instances restantes sont incorrectement classées et sont désignées par FP_k. Sur la ligne de la classe *k*, et sauf les instances TP_k les instances restantes sont incorrectement classées et sont désignées par FN_k. A l'exception de la classe *k*, les instances des autres classes bien classées et qui positionnent sur le diagonal de cette matrice sont indiquées par TN_h et $h \neq k$.

		Distribution de la classification						Total (Par ligne)
	Classes	1	…		*k*	…	K	
Distribution réelle	1	TN^1			FP^1			FN_1
	2		TN^2		FP^2			FN_2
	…			…	…			…
	k	FN^1	FN^2	…	TP_k	…	FN^K	$FN_k=FN^1+FN^2+\ldots+FN^K$ sauf TP_k
	…				…	…		…
	K				FP^K		TN^K	FN_K
Total (Par colonne)		FP_1	FP_2	…	$FP_k=FP^1+FP^2+\ldots+FP^K$ sauf TP_k	…	FP_K	$T= TP_1+TP_2+\ldots+TP_K + FP_{1+} FP_2+\ldots+FP_K$ Et $T= TP_1+TP_2+\ldots+TP_K + FN_{1+} FN_2+\ldots+ FN_K$

Table 2.2. Matrice de confusion multi-classes.

Parmi les métriques utilisées pour l'évaluation de la classification multi-classes il existe des mesures basées sur la notion de **macro** et **micro Average.** Les métriques de macro Average mesurent chaque classe séparément puis calculent leur moyenne. Elles sont utilisées lors de l'évaluation de la performance des modèles par la même importance pour toutes les classes. Tandis que les métriques de micro Average rassemblent toutes les apports de toutes les classes afin de calculer la mesure moyenne. Elles sont utilisées lors de l'évaluation de la performance des modèles avec la même importance de toutes les instances. Les citations suivantes ne sont que quelques-unes des exemples des métriques multi-classes les plus célèbres (Han et al., 2012 ; Grandini et al., 2020) :

- **Précision par classe (P_k) :** Elle calcule la précision pour chaque classe *k*.

$$\boldsymbol{P_k} = \frac{\mathbf{TP_k}}{\mathbf{TP_k + FP_k}} \qquad (2.7)$$

- **Rappel par classe (R_k) :** Il compte le rappel pour chaque classe *k*.

$$\boldsymbol{R_k} = \frac{\mathbf{TP_k}}{\mathbf{TP_k + FN_k}} \qquad (2.8)$$

- **Macro Average Precision (P_{Ma}) :** Il calcule la précision moyenne par classe. Appelé par fois précision moyenne.

$$\boldsymbol{P_{Ma}} = \frac{\sum_{k=1}^{K} \mathbf{P_k}}{\mathbf{K}} \qquad (2.9)$$

- **Macro Average Recall (R_{Ma}) :** Il calcule le rappel moyen par classe. Appelé par fois rappel moyen.

$$\boldsymbol{R_{Ma}} = \frac{\sum_{k=1}^{K} \mathbf{R_k}}{\mathbf{K}} \qquad (2.10)$$

- **Micro Average Precision (P_{Mi}) :** Il calcule la précision des apports rassemblés de toutes les classes.

$$P_{Mi} = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + \sum_{k=1}^{K} FP_k} = \frac{\sum_{k=1}^{K} TP_k}{T} \quad (2.11)$$

- **Micro Average Recall (R_{Mi}) :** Il calcule le rappel des apports rassemblés de toutes les classes.

$$R_{Mi} = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + \sum_{k=1}^{K} FN_k} = \frac{\sum_{k=1}^{K} TP_k}{T} \quad (2.12)$$

- **Macro Average FM :** Il existe deux définitions (Berger et Guda, 2020) :

  - Macro FM (FM_{MaF}) : Il calcule le FM moyen par classe.

$$FM_{MaF} = \frac{\sum_{k=1}^{K} FM_k}{K} \quad (2.13)$$

  - FM Macro (FM_{FMa}) : Il calcule le FM moyen par classe.

$$FM_{FMa} = \frac{2 * P_{Ma} * R_{Ma}}{P_{Ma} + R_{Ma}} \quad (2.14)$$

- **Micro Average FM (FM_{Mi}) :** Il calcule le FM des apports rassemblés de toutes les classes. Il est clair que les P_{Mi} et R_{Mi} sont égaux, et par l'exploitation de la formule de la moyenne harmonique FM_{Mi} va générer la même valeur comme suit :

$$FM_{Mi} = \frac{2 * P_{Mi} * R_{Mi}}{P_{Mi} + R_{Mi}} = P_{Mi} = R_{Mi} \quad (2.15)$$

- **Taux de reconnaissance moyen (Average Accuracy) (acc_A) :** Il calcule le taux de reconnaissance moyen par classe.

$$acc_A = \frac{\sum_{k=1}^{K} \frac{TP_k + TN_k}{TP_k + FP_k + TN_k + FN_k}}{K} \quad (2.16)$$

- **Taux d'erreur moyen (Average Error Rate) (err_A) :** Il calcule le taux d'erreur moyen par classe.

$$err_A = \frac{\sum_{k=1}^{K} \frac{FP_k + FN_k}{TP_k + FP_k + TN_k + FN_k}}{K} \quad (2.17)$$

B. Méthodes d'évaluation :

- **Méthode de Holdout** (Han et al., 2012) **:** C'est une méthode qui partitionne aléatoirement le dataset en deux parties, Training set et Test set. La partie Training set est utilisée pour l'apprentissage du modèle, tandis que la partie Test set est utilisée pour mesurer la performance de la classification du modèle à base des métriques d'évaluation. Généralement, beaucoup de travaux adoptent une partition de données par deux tiers pour le Training set et un tiers pour le Test set. Mais, la détermination de la partie d'entrainement Training set du modèle dès le début est le principal inconvénient de cette méthode.

- **Sous-échantillonnage aléatoire (Random subsampling)** (Han et al., 2012) **:** Elle est considérée comme une variante de la méthode Houldout, puisqu'elle la répète k fois. Par conséquent, l'évaluation globale consiste à calculer la moyenne générale des performances obtenues durant les k itérations. L'échantillonnage des instances des ensembles de tests peut être vu parfois comme un inconvénient à cause de la nécessité d'une distribution indépendante sur tout l'ensemble de données initiale.

- **Cross-Validation (K-Fold Cross-Validation)** (Han et al., 2012) **:** C'est une méthode qui consiste à partitionner le dataset en k groupes 'k-fold' de même taille. Pour chaque itération i ($1 \leq i \leq k$), cette méthode utilise le groupe G_i comme un ensemble de test, et les autres groupes comme un ensemble de Training set. Parmi les variations reconnues de cette méthode nous pouvons citer :

  - **Leave-P-Out Cross-Validation :** C'est une méthode qui utilise une partition de P instances comme un Test set et le reste des instances pour l'apprentissage. Cette opération sera répétée pour couvrir toutes les

combinaisons possibles de cette partition. **Leave-one-out Cross-Validation** considéré comme un cas particulier de cette méthode, où le nombre des instances P est égale à un.

- **Stratified K-Fold Cross-Validation :** C'est une méthode qui préserve le pourcentage des instances de chaque catégorie lors de l'initialisation du partitionnement.
- **Repeated K-Fold Cross-Validation :** C'est une méthode qui applique le K-fold Cross-Validation plusieurs fois. La performance globale sera calculée en fonction de la moyenne de tous les résultats obtenus lors de la réexécution. A travers de cette façon, cette méthode vise à réduire l'erreur d'estimation de performance du modèle.

- **Bootstrap (Echantillonnage par replacement)** (Han et al., 2012) **:** Sur le même l'ensemble de données (Dataset), cette méthode répète plusieurs fois (*K*) le processus du choix aléatoire de *N* instances (Tirage avec remise) pour l'ensemble d'apprentissage (Training set) et le reste des instances sera utilisé pour le test (Test set). Cela aboutit parfois à choisir les mêmes instances pour former le Training set, par conséquent d'autres instances n'auront aucune participation à l'entrainement du modèle. Le « **.632 Bootstrap »** est l'un des variantes les plus utilisés de cette méthode. Elle calcule le taux de reconnaissance finale ***acc*** du modèle par :

$$\boldsymbol{acc} = \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} (\mathbf{0.632} * \boldsymbol{acc_Ts_k} + \mathbf{0.368} * \boldsymbol{acc_Tr_k}) \qquad (2.18)$$

dont, $\boldsymbol{acc_Ts_k}$ ($\boldsymbol{acc_Tr_k}$) est le taux de reconnaissance du modèle entrainé lors de l'itération *k* sur l'ensemble de Test set (respectivement Training set).

- **Receiver-operating characteristic (ROC)** (Roiger, 2017; Zou et al., 2007) **:** Il s'agit d'une technique graphique permettant d'évaluer et de comparer les performances des modèles de classification. Cet outil a été développé initialement durant la deuxième guerre mondiale à des fins militaires. En vue

de l'importance des analyses du graphe ROC et les possibilités de ses utilisations, il a été considérablement impliqué dans les domaines médicaux, en particulier pour la prise des décisions.

Techniquement, le graphe ROC trace la courbe expliquée par le **taux des vrais positives (TPR au-dessus)** sur l'axe Y en fonction du **taux des faux positives (False Positive rate FPR)** sur l'axe X :

$$\boldsymbol{FPR} = \frac{\mathbf{FP}}{\mathbf{FP+TN}} \tag{2.19}$$

La courbe ROC tracée en superposition avec le diagonal selon les deux points (0,0) et (1,1) montre la classification aléatoire du modèle. Tandis que la performance idéale du modèle sera tracée par la courbe ROC sur le segment (0,0), (0,1) et le segment (0,1), (1,1).

L'aire sous la courbe ROC (**Area Under the ROC Curve (AUC)**) est une technique basée sur le calcul de l'espace sous la courbe ROC. Les mesures AUC varient entre 0 et 1. Une valeur de 1 pour AUC signifie que toutes les classifications sont correctes et le modèle a la capacité de distinguer les catégories de différentes instances parfaitement, et la valeur 0 signifie que toutes les classifications sont incorrectes et le modèle prédit les catégories d'une manière inversée. Tandis que les valeurs supérieures à 0.5 expliquent la capacité de catégorisation correctement les instances de test par le modèle en validation. Finalement, la valeur 0.5 de l'AUC montre la classification aléatoire de toutes les données (Roiger, 2017; Zou et al., 2007). La Figure 2.3 montre la un exemple de courbe ROC et l'espace AUC.

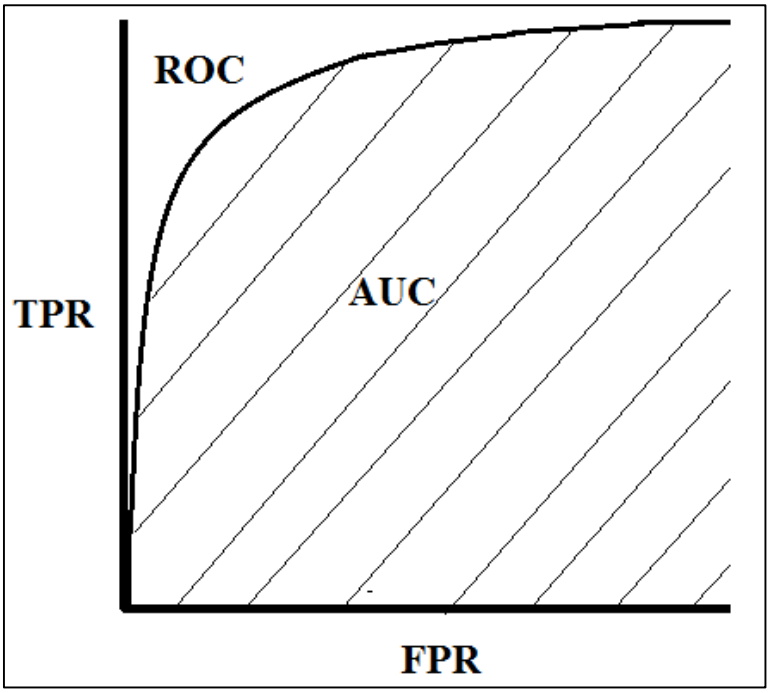

Figure 2.3. La courbe ROC et l'aire sous la courbe ROC.

Une fois que l'évaluation s'est terminée avec des résultats acceptables, les modèles produits peuvent être utilisés sur d'autres données de même composition. À titre d'exemple, les données de la médecine peuvent contenir des images de la radiologie, des rapports textuel, l'historique médicale, etc. Elles incluent occasionnellement les prélèvements sanguins, les dates d'hospitalisation, les données capturées sous surveillance et d'autres. Convenablement à notre vision ultérieure et tant qu'une composition inévitable pour les sources des données médicales, nous étudions brièvement dans la partie suivante l'exploration de toutes données formant des séries temporelles.

2.3 Data Mining pour les séries temporelles (Time series data mining)

2.3.1 Séries temporelles (Time series)

Une série temporelle ou une série chronologique est une suite d'observations numériques ordonnées en fonction du temps et capturées afin de mesurer une variable

(Evénement) bien précise. Plus que les valeurs prises, leurs ordonnancements constituent une autre information qui peut servir la tâche principale. Les enregistrements des séries temporelles peuvent êtres fait d'une manière périodique sur le temps, ce qui produit des séries temporelles régulières « ***Regular time series*** ». A l'inverse, la capture des observations sans périodicité forme des séries temporelles irrégulières « ***Irregular time series*** ». De plus, la série temporelle qui comporte les observations d'une seule variable est appelée une série temporelle uni-variée « ***Univariate Time Series*** ». Tandis que, la série temporelle qui admet plusieurs variables est appelée une série temporelle multivariée « ***Multivariate Time Series*** ». À titre d'exemple la série temporelle qui capture la température et l'humidité périodiquement est une série multivariée régulière à deux variables, et la série des retraits d'argent d'un compte bancaire est une série uni-variée irrégulière à une seule variable (André-Jönsson, 2002).

2.3.2 Différentes tâches de data mining sur les séries temporelles

Selon l'objectivité des traitements, les tâches sur séries temporelles peuvent êtres subdivisées en (Ratanamahatana et al., 2009) :

A. Indexation (Requête par contenu) : C'est la tâche de la recherche dans un dataset des séries temporelles les plus similaires à une série lancée comme une requête. Une mesure de similarité/dissimilarité doit s'appliquer afin d'achever cette opération. Généralement, les correspondances entre les séries temporelles sont calculées soit par comparaison de la série complète ou par des segments sous forme de fenêtres glissantes.

B. Clustering : Il s'agit du processus de recherche de regroupements de séries temporelles qui peuvent sembler naturellement similaires en fonction d'une mesure de similarité/dissimilarité bien précis. Par conséquent les groupes formés doivent différer entre eux autant que possible, et chacun ne doit comporter que les séries les plus homogènes.

C. Classification : C'est l'opération de catégorisation d'une série temporelle non étiquetée par rapport à un ensemble de données prédéfinies en utilisant des techniques de l'apprentissage supervisé.

D. Prédiction : C'est l'action de la prévision des valeurs d'une série temporelle pour une évolution future en fonction des observations enregistrées précédemment.

E. Caractérisation (Summarization) : C'est une façon de réduire ou de simplifier de la représentation d'une série temporelle extrêmement longue. Elle consiste à créer une autre série approximative de taille simple et qui conserve les caractéristiques essentielles de la série originale.

F. Détection des anomalies : C'est une tâche de recherche de toutes anomalies dans une série temporelle que ce soit par une sous séquence, une seule donnée ou par la série elle-même. Généralement, cette tâche évalue le comportement de la série testée par rapport au comportement modélisé des séries temporelles normales. Sur le plan technique il existe différentes méthodes destinées à détecter les anomalies. On trouve des méthodes basées sur la similarité, soit entre les portions produites par le processus de segmentation, soit entre les structures de données produites par le processus de clustering ou encore par la considération du traitement via des plus proches voisins. Les autres méthodes se basent généralement sur les modèles prédictifs et la modélisation séquentielle par Hidden Markov model (HMM). Par la suite, nous pouvons catégoriser toutes ces techniques en deux types (Fig 2.4).

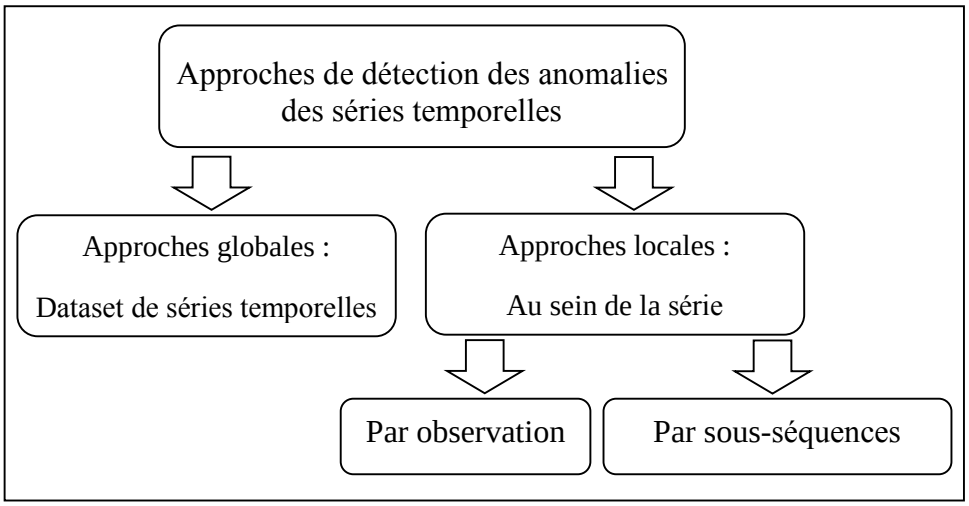

Figure 2.4. Type d'approches de détection des anomalies des séries temporelles.

- **Approches de détection globales :** Ces approches sont destinées à détecter les séries temporelles qui ont un comportement anormal vis-à-vis d'un ensemble de séries considérées possédants un comportement normal.

- **Approches de détection locales :** A l'inverse du type précédent, ce type comporte les approches travaillant sur la détection des comportements anormaux dans une série temporelle. En revanche, ces approches se différentient entres elles selon l'unité de traitement adoptée :

  - Approches de détection ponctuelles : Ces approches visent la découverte de toutes les observations apparues en tant que des valeurs anormales dans une série temporelle. Elles traitent un seul point à la fois.

  - Approches de détection sectionnelles : Ces approches travaillent sur la détection des sections (les sous-séries ou les séquences) anormales au sein d'une seule série temporelle.

G. Segmentation : C'est une façon de partitionner d'une série temporelle sous forme des sections homogènes afin d'effectuer des analyses ou de produire des

représentations pour d'autres tâches. Globalement, cette tâche regroupe trois catégories d'algorithmes (Keogh et al., 2004) :

- **Fenêtres coulissantes (Sliding Windows) :** C'est une famille d'algorithmes qui se base sur l'agrandissement du segment initialisé par le premier point gauche. La croissance par point est l'activité de base qui se répète tant que le segment génère une erreur moins que le seuil définit au début. D'une manière itérative le long de la série et par la même façon, la construction du nouveau segment démarre en fonction du premier point à droite du segment précédant.

- **De haut en bas (Top-Down) :** Ce sont des algorithmes qui se basent sur la partition de la série temporelle en sous sections d'une manière récursive tant que l'erreur approximée sur les segments obtenus dépasse un seuil défini au début.

- **De bas en haut (Buttom-Up) :** Ce type d'algorithme démarre par une partition maximale de la série temporelle, qui comporte deux points seulement par segment. Par la suite, le coût de la fusion initiale de tous les segments consécutifs sera recalculé. Tant que le coût minimal est inferieur par rapport au seuil défini au début, ce processus va fusionner leurs deux segments. Les coûts de la fusion de la nouvelle séquence produite avec ses segments adjacents seront recalculés afin d'indiquer à nouveau le coût minimal.

Les développements technologiques et scientifiques rapides, en particulier dans le domaine de la médecine, ont défini de nouveaux défis qui dépassent les capacités du data mining. Le traitement des données gigantesques, la typologie complexe et non structuré de données, la gestion des provenances de ces données sont des exemples qui ont exigé la recherche d'une solution innovante. A cet effet, l'environnement

appelé « Big Data » a été défini comme un nouveau concept destiné à surmonter la plupart des défis cités précédemment.

2.4 Big Data.

Le terme « ***Big Data*** » a été utilisé pour la première fois dans les travaux de la NASA en 1997 (Cox et Ellsworth, 1997). Il indique toutes les collections de données ***volumineuses***, ***complexes*** et très difficiles à traiter par les méthodes traditionnelles (Natarajan t al., 2017), et qui nécessitent parfois des actions en temps réel (Sawant et Shah, 2013). La complexité est imposée par la ***variété*** « variation des sources de données structurées et non structuré tel que le texte, l'audio, la vidéo, etc. », la ***vélocité*** « vitesse d'arrivage, de production, de partage, d'analyse et de traitement des données et leurs fréquences de changement », la ***véracité*** « la qualité de données » (Márquez et Lev, 2017).

Pour définir le Big Data, il existe beaucoup de travaux qui utilisent la notion en « V ». « V3 » caractérise les définitions utilisent **« Volume, Variété, Vélocité ». « V4 »** pour **« Volume, Variété, Vélocité, Valeur »,** le terme ***valeur*** est ici pour pointer vers l'utilité et les avantages stratégique et économique. **« V5 »** désigne **« Volume, Variété, Vélocité, Véracité, Valeur »** (Wamba et al., 2015).

2.4.1 Couches de Big Data

Afin d'exploiter le Big Data comme un service, les recherches utilisent une description par quatre couches (Marr, 2017) :

A. Collecte des données : La couche de collecte de données appelée aussi couche de sources de données est la phase d'identification des besoins en données. Elle est caractérisée par le choix et la spécification des endroits de provenance des données, dont leurs sources peuvent être internes et/ou externes. Les sources des données peuvent fournir des données différentes, ce qui pointe vers les outils et les systèmes

utilisés pour capturer ou pour générer ces données tel que les capteurs sur les appareils, les vidéos, les réseaux sociaux, etc. (Marr, 2017).

B. Stockage des données : Cette couche concerne les besoins de stockage de données. Les serveurs d'entreprise et les disques durs apparurent comme des solutions internes et classiques pour un stockage de données dont le volume est petit ou moyenne. D'autres solutions sont utilisées pour les volumes importants de données tel que les systèmes de stockage à base de cloud, les entrepôts et lacs de données (Marr, 2017 ; Sawant et Shah, 2013).

C. Analyse et traitement des données : C'est la couche d'application du processus d'extraction des connaissances à la base des données collectées. Il s'agit d'une préparation des données qui inclut le nettoyage et la transformation, l'utilisation des techniques analytiques sur les données afin de conclure par des décisions sur les connaissances et les résultats obtenus (Marr, 2017).

D. Accès aux données et communication : C'est la couche concernée par la mise en place des infrastructures et les outils nécessaires pour permettre l'accès aux données à soit des personnes ou des machines. La communication de données doit être entourée par les stratégies des authentifications, de contrôle d'accès, de permissions et de sécurité en général (Marr, 2017 ; Sawant et Shah, 2013).

2.4.2 Chiffres et promesses en Big Data

Une estimation de 281 exaoctets (10^9*281 gigaoctet) de données numérique stockées en 2007, avec une croissance remarquable qui peut atteindre 57% dans les organisations possédants un rythme rapide de développement (Kale, 2017). L'an 2011 s'est achevé par l'envoi de 118 milliards d'emails par jour et la création de 12 zettaoctets (10^{12}*12 gigaoctet) de données. Par exemple ***The Library of Congress*** a enregistré 235 téraoctets (10^3*235 gigaoctet) de données collectées (Kone, 2013). La Figure 2.5 (Statista, 2018) explique le développement des revenues du Big Data et de marché analytique des affaires au niveau mondial. En 2017 les contenues partagés de

Facebook dépassent les 500 millions par mois, et comportent des photos, des notes, des blogs, des liens internet et d'autres. Twitter rencontre plus de 10 téraoctets de données générées par jour à la base de 140 caractères. Les vidéos visionnées sur YouTube dépassent les 4 milliards d'heures (Kale, 2017). Le revenu de 2017 approximé par 150.8 milliard de dollars et les prévisions indiquent un chiffre de 210 milliards de dollars en 2020. Un autre exemple concerne le télescope « ***Square kilometers away*** », son estimation de production de données en 2024 atteindra un téraoctet par minute (Kone, 2013). Les résultats de la Table 2.5 montrent la croissance d'utilisation de Big Data, et son importance dans les stratégies de développement pour les organisations et les entreprises.

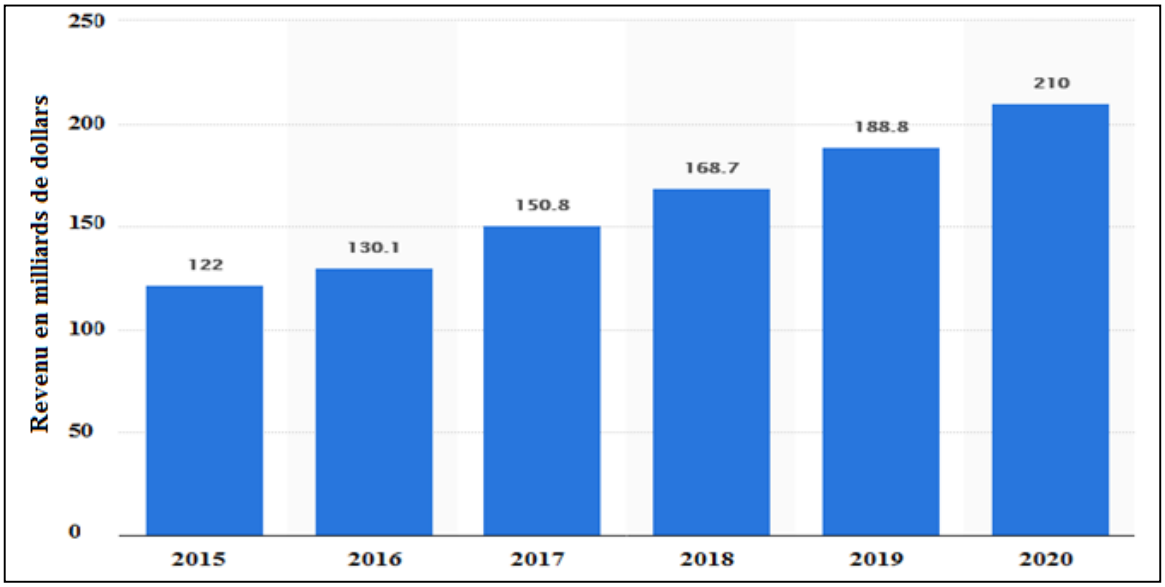

Figure 2.5. Revenu de Big Data et le marché analytique des affaires (Statista, 2018).

2.4.3 Défis du Big Data

Il y a beaucoup de défis que l'on doit prendre en considération lors d'un travail en Big Data. 90% des données brutes contenants des bruits, ce qui nécessite une étape de compréhension et d'hiérarchisation afin de les filtrer et de les classer. De plus, la confidentialité des données est une obligation légale mais l'incapacité de les analyser par des ressources internes oblige souvent de se tourner vers des services externes tels que les clouds. Ce défi impose un compromis entre la confidentialité, le coût

d'installation d'une solution propriétaire et le coût d'utilisation d'un service externe. La croissance du volume de données, la suppression des données non pertinents, la durée de stockage et les coûts impliqués représentent d'autres défis. À ceci s'ajoute le problème de disponibilité des compétences et les qualités demandées aux personnels qualifiés pour la mise en marche des systèmes Big Data (Sawant et Shah, 2013).

2.4.4 Techniques de Big Data

A. Distribution des traitements : Les traitements parallèles selon le principe « ***Diviser pour régner*** » sont utilisés par les plates-formes de Big Data. Elles exploitent des clusters contenants des nœuds sur lesquels les données seront partitionnées. Le nombre de ces nœuds peut varier de dizaines à des milliers dépondant de la quantité des données, les performances demandées et les contraintes d'évolutivité (Kale, 2017).

B. Base de données NoSQL « Not Only SQL database » :

C'est une base de données plus adaptée au traitement distribué des données volumineuses, et utilisée dans le cas où il nécessite parfois des réponses rapides pour les requêtes. Ces bases de données sont apparues comme des solutions aux problèmes des bases de données relationnelles telle que la diminution de performance pour les données massives, et le problème des représentations de données et leurs réplications pour le traitement (Kale, 2017).

C. **Stockage de données en mémoire :** C'est une technique pour les transactions et les analyses des requêtes en temps réel qui permet le stockage et le traitement des données massives au niveau de la mémoire RAM de taille importante. Elle bénéfice de la vitesse de la RAM qui peut être 100000 fois plus rapide qu'un accès au disque dur (Kale, 2017).

2.4.5 Types de Base de données NoSQL

A. Clé/Valeur : C'est le plus simple type, les clés sont uniques et chaque un forme un couple avec une valeur qui peut être un texte, XML, ou un autre objet. Exemple : Amazon DynamoDB, Voldemort, Redis, Memcached, Riak (Hu et al., 2014).

B. Document : Ce type utilisé pour le stockage des documents semi-structurés formatés en JSON ou XML, et chaque valeur présente un document. Exemple : MongoDB, SimpleDB, CouchDB (Hu et al., 2014).

C. Colonnes : Le stockage ou le traitement des données se fait par colonnes, au lieu d'être réalisé en ligne. Ce type de base permet d'introduire un nombre de colonnes dynamique, et procède la répartition des lignes et des colonnes sur plusieurs nœuds. Exemple : Bigtabl (Google), Cassandra (Facebook), HBase (Apache) (Hu et al., 2014).

D. Graphe : Les données sont stockées sous forme d'une structure de graphe, dont chaque nœud est caractérisé par un identificateur et comporte un ensemble d'attributs, et lié avec les autres nœuds par des relations figurées sur leurs attributs (Kale, 2017).

2.4.6 Impact du Big data sur la médecine personnalisée

La médecine personnalisée produit un ensemble de données importantes pour chaque patient, ces données auront diverses natures et diverses sources. Le EHR fournira alors toutes les informations du patient allant de la démographie de la personne, en passant par les données génétiques, biologiques et d'autres jusqu'aux traitements prescrits, les effets indésirables et le développement de la situation médicale. Cependant, la collecte de données des participants se fait généralement de deux manières. La première se réalise à travers les médecins et les professionnels de la santé. Tandis que la deuxième se réalise à travers l'individu lui-même. Cette dernière est expliquée par tous les appareils portables du monde connecté actuel. La Figure 2.6

(Piwek et al., 2016) englobe la plupart de ces appareils utilisées particulièrement à des finalités médicales.

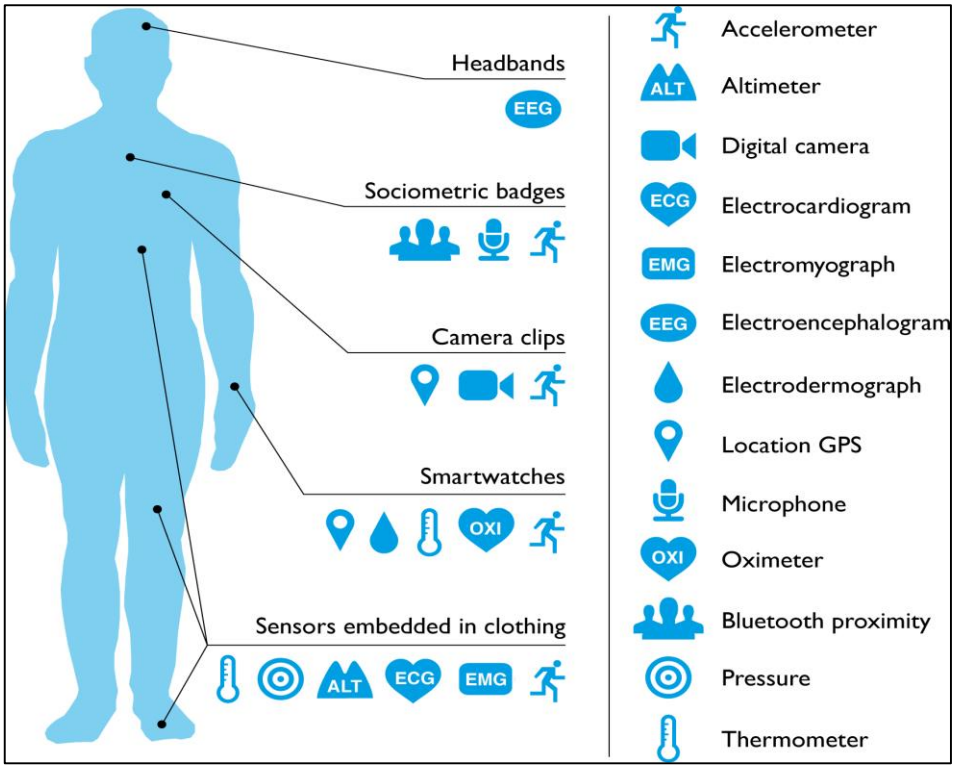

Figure 2.6. Appareils connectés médicales (Piwek et al., 2016).

Toutes les sources de données citées ne sont décrites pour qu'un seul patient, ce qui montre une vue préalable de leur volume par personne. Avec l'ensemble de tous les participants de la médecine personnalise dans une population donnée, le volume produit peut dépasser les capacités traditionnelles de traitements. Avec le temps, le Big data a été de plus en plus impliqué dans les projets de la médecine personnalisée. Par conséquent, la médecine personnalisée trouve dans le Big data un environnement adéquat pour réussir ses tâches. Généralement, la prévention des maladies, la prédiction des effets secondaires des médicaments, la personnalisation des traitements, l'intervention dans les meilleurs délais même sur les sites éloignés, la stratification des patients, le diagnostic à distance et d'autres activités sont devenues plus facilement réalisables. D'une manière générale, le Big data et ses évolutions

constituent un bon support pour la médecine personnalisée et cela peut l'accompagner vers d'autres succès.

2.5 Conclusion

L'exploration de données par les techniques du data mining nécessite une série d'opérations quel que soit le domaine d'application. Dans une première partie de ce chapitre, nous avons parlé brièvement de ce processus, des tâches et ses disciplines associées. Etant donné l'importance de l'activité d'évaluation des modèles produits, nous l'avons expliqué plus en détail que les autres phases, mais cela ne diminue en rien de l'importance de toutes les tâches.

Généralement, l'historique médical des patients de la médecine personnalisée comporte des données sous la forme de séries temporelles. A cet effet, nous avons dressé, en deuxième partie, une simple introduction sur le data mining des séries temporelles. Plus que de définir formellement les séries temporelles, nous avons cité diverses tâches d'exploration pouvant être appliquées à ce type de données.

Dans la troisième partie, et étant donné que la masse du volume de données, leur gestion et leur analyse constituent des défis de la médecine personnalisée, nous avons parlé du Big data. Cet environnement technologique utilisé dans plusieurs pratiques a été également engagé dans la médecine personnalisée et est l'un des facteurs qui contribuera de manière significative à l'adoption de cette médecine et son évolution. Bien que le Big data n'a été pas pratiqué dans cette thèse par manque de temps, nous l'avons détaillé car nous avions des ambitions de le pratiquer dans une contribution future.

L'exploration de données de la médecine personnalisée peut impliquer plusieurs activités tel que la catégorisation des patients. Cette dernière nécessite l'application de traitements particuliers. A cet effet, le chapitre suivant va détailler plus précisément des tâches de data mining dédiées et notamment l'activité de la représentation de données et de la classification.

Références

- André-Jönsson, H. (2002). *Indexing strategies for time series data.* Department of Computer and Information Science, Linköpings universitet.

- Berger, A., & Guda, S. (2020). Threshold optimization for F measure of macro-averaged precision and recall. *Pattern Recognition. Vol 102*.

- Coelho, P. S. S., & Ebecken, N. F. F. (2002). *A comparison of some classification techniques. In : Data Mining III, Zanasi et al. (Eds).* WIT Press, Ashurst Lodge, Southampton, UK.

- Cox, M., & Ellsworth, D. (1997). Application-Controlled Demand Paging for Out-of-Core Visualization. *In Proceedings of the 8th IEEE Visualization '97 Conferences, Arizona, USA*, 235-244.

- Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, Vol 17*(3).

- Gorunescu, F. (2011). Data Mining Concepts,Models and Techniques. *Intelligent Systems Reference Library. Vol 12*.

- Grandini, M., Bagli, E., & Visani G. (2020). Metrics for multi-class classification : an overview. *Stat.ML,* arXiv :2008.05756v1.

- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques, Third edition.* Morgan Kaufmann, MA, USA.

- Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward Scalable Systems for Big Data Analytics- A Technology Tutorial. *IEEE Access, Vol 2*, 652-687.

- Kale, V. (2017). *Big Data Computing A Guide for Business and Technology Managers.* CRC Press, FL, USA.

- Kantardzic, M. (2011). *Data mining Concepts, Models, Methods, and Algorithms, Second edition.* John Wiley & Sons Inc, Hoboken, New Jersey.

- Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). *Segmenting time series : A survey and novel approach. In : Data Mining in Time Series Databases.* WorldScientific, 1–21.
- Kone, A. (2013). Big data (rapport de stage).
- https://www.memoireonline.com/05/14/8890/Big-data-rapport-de-stage.html#_Toc369718828.
- Márquez, F. P. G., & Lev, B. (2017). *Big Data Management.* Springer Nature, Gewerbestrasse, Switzerland.
- Marr, B. (2017). *Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things.* Kogan Page, London, UK.
- Natarajan, P., Frenzel, J. C., & Smaltz, D. H. (2017). *Demystifying Big Data and Machine Learning for Healthcare*. CRC Press, FL, USA.
- Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques.* Springer-Verlag Berlin Heidelberg.
- Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The Rise of Consumer Health Wearables: Promises and Barriers. *PLoS Med, Vol 13*(2).
- Ratanamahatana, C. A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M,. & Das, G. (2009). *Mining Time Series Data. In: Maimon O., Rokach L. (eds.) Data Mining and Knowledge Discovery Handbook 2nd ed*. Springer Science+Business Media, 1049–1077.
- Roiger, R. J. (2017). *Data Mining A Tutorial-Based Primer, Second edition.* CRC Press, FL, USA.
- Sagar P., Prinima, P., & Indu, I. (2017). Analysis of Prediction Techniques based on Classification and Regression. *International Journal of Computer Applications, Vol163*(7).
- Sawant, N., & Shah, H. (2013). *Big Data Application Architecture Q & A. A Problem-Solution Approac.* Appress. 1–139.

- Shapiro, G. P. (1990). Knowledge Discovery in Real Databases : A Report on the IJCAI-89 Workshop. *AI Magazine, Vol 11*(5).

- Shmueli, G., Bruce, P. C., Stephens, M. L., & Patel, N. R. (2017). *data mining for business analytics Concepts, Techniques, and Applications with JMP Pro*. John Wiley & Sons Inc, Hoboken, New Jersey.

- Statista. (Accessed 2018). Revenue from big data and business analytics worldwide from 2015 to 2020 (in billion U.S. dollars). https://www.statista.com/statistics/551501/worldwide-big-data-business-analytics-revenue/.

- Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact : Findings from a systematic review and a longitudinal case study. *Int. J. Production Economics. Vol 165*, 234-246.

- Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation. Vol 115*, 654–657.

CHAPITRE 3

Qu’est-ce que la représentation de donnés ?

Sommaire

3.1 Introduction 76
3.2 Représentation de données 77
3.2.1 Définition de représentation données 77
3.2.2 Transformation de données 78
3.2.3 Représentation des types de données de base 81
3.2.4 Représentation des types de données avancés 82
3.2.5 Représentation des séries de données temporelles 82
3.3 Réduction de données 87
3.3.1 Techniques de réduction de dimension 87
3.4 Classification de données 92
3.4.1 Distances en data mining 93
3.4.2 Techniques de segmentation (Clustering) 94
3.4.3 Indices de qualité du clustering 98
3.4.4 Techniques de classification 100
3.5 Conclusion 102

3.1 Introduction

L'exploration de données de n'importe quel domaine par les techniques de data mining nécessite une étape de préparation de données, qui vise à transformer cet ensemble de données dans un format conforme aux structures de données utilisables par la technique de classification choisie. Les types et les qualités de données varient d'un domaine à un autre. La médecine personnalisée génère et utilise des données des patients, dont les formats peuvent êtres homogènes ou hétérogènes, simples ou complexes, structurés ou non-structurés. Pour avoir de bons résultats, la modélisation du problème exige le traitement de plusieurs types de données à la fois. Ce traitement peut appliquer des opérations de transformation dont le but est de produire une représentation globale. Par conséquent, le choix des opérations varie selon la nature et le type de données exploitées.

La quantité de données à explorer est un élément important et doit être pris en compte lors de la modélisation. Sous certaines contraintes le traitement de données volumineuses nécessite parfois certaines opérations de réduction de dimension. En pratique, le processus de DM essaye d'utiliser la technique la plus appropriée visant à optimiser l'ensemble des contraintes fixées dès le début.

Généralement, un processus de classification de données nécessite un ensemble d'outils pour atteindre son but. Cet ensemble d'outils peut comporter les formules de distance entres les instances, des algorithmes de classification ou de clustering, et des mesures d'évaluation relatives à la tâche programmée.

Conformément à nos modélisations expliquées ultérieurement, nos explications vont s'attendre spécialement sur les données structurées. Par conséquent, ce chapitre sera présenté en trois parties principales. La première s'intéresse à la représentation et aux transformations de données médicales. Un résumé sur la réduction de dimension sera présenté dans une deuxième partie. Enfin, la troisième partie explique partiellement le clustering et la classification de données, avec la présentation de certaines techniques populaires.

3.2 Représentation de données

Dans un processus d'exploration de données médicales, les trois natures de données non structurées, semi structurées et structurées dirigent généralement les phases de traitement à appliquer, dont surtout la représentation de données. L'importance de cette dernière vient de sa présence au début du processus de DM. Dans la suite, nous résumons les principes importants relatifs à la représentation de données structurées. De plus, des détails sont donnés sur certaines techniques qui seront appliquées ultérieurement dans la suite du manuscrit.

3.2.1 Définition de représentation données

« *Comment les données sont représentées. (How data is represented.) ?* » Nous reprenons ici la définition de la représentation de données donnée par Nettleton (2014). Du point vu global, nous pouvons l'expliquer comme un processus qui comporte des opérations de reformulation, de transformation et de restructuration de données élémentaires (types de données primaires) ou composées (types de données avancés) afin de simplifier leur compréhension et leur analyse, mais également les préparer au processus d'extraction de connaissances. Ce processus peut s'avère complexe de par la réutilisation de modèles développés pour différentes modalités et finalités. Conformément au processus d'exploration de données (Chapitre 2 DM & Big Data), la représentation de données englobe deux tâches que sont le prétraitement (preprocessing) et la transformation de données.

Pour clarifier le processus de représentation des données et son utilité, il est nécessaire d'analyser et d'expliquer plus en détails certains concepts qui servent le processus d'exploration des données médicales en général.

3.2.2 Transformation de données

La préparation des données pour l'analyse ou la classification impose parfois le passage par une opération de transformation de données. Généralement, elle est vue en tant que processus de conversion de données. Les objectifs derrières ce processus varient d'une approche à l'autre (García et al., 2015) :

A. Lissage des données (Data Smoothing) : C'est l'opération d'élimination des valeurs aberrantes (outliers) existant probablement dans les données. Afin de démarquer mieux les motifs importants, cette opération applique certains algorithmes tel que ***Simple Exponential***, ***Moving average*** et ***Random Walk*** (Han et al., 2012).

B. Agrégation de données : A cause de différentes raisons tels que les diversités de présentation et les différentes échelles de données dans les sources de ces dernières, l'opération d'agrégation transforme et combine les données issues de multiples sources dans un format unique pour les stocker et les traiter ultérieurement.

C. Discrétisation de données : C'est le processus de conversion de données quantitatives sous forme d'un ensemble d'intervalles labélisés par des étiquetés catégoriques ordinales. Généralement, elle permet de limiter les états possibles des données observées. Le nombre d'intervalles et leurs limites sont quelques défis qui doivent être résolus pendant la transformation. Parfois, la discrétisation est utilisée pour appliquer certains algorithmes qui n'acceptent que des attributs catégoriques.

D. Généralisation de données : C'est une tâche de conversion de données marquées par des valeurs de bas niveau en d'autres données décrites par des valeurs de haut niveau, en suivant une opération d'abstraction bien déterminée. La température du corps des patients est un exemple très simple. Nous pouvons transformer la température du 35° (38°, 41° respectivement) à un niveau conceptuel plus élevé comme *Bas* (*Normal*, *Elevé* respectivement).

E. Normalisation : C'est un processus de transformation de données reconnu comme une étape de prétraitement appliquée pour unifier les échelles ou les plages

de données observées. Généralement, les nouvelles plages de données ne varient que dans des intervalles petits comme de -1 à 1 ou 0 à 1. Par conséquent, cette opération peut simplifier le processus d'exploration de données et le rendre plus rapide. Parmi les méthodes de normalisation les plus célèbres en peut citer les trois suivantes :

- **La normalisation Min-Max (*Min-Max Normalization*) :** C'est la transformation linaire qui produit une échelle de données variant entre 0 et 1 selon la formule suivante :

$$\boldsymbol{V'} = \frac{\boldsymbol{V - \mathrm{Min}(A)}}{\boldsymbol{Max(A) - Min(A)}} \left(\boldsymbol{Max(A')} - \boldsymbol{Min(A')}\right) + \boldsymbol{Min(A')} \qquad (3.1)$$

Tel que V est la valeur à normaliser de l'attribut originale A. Min(A) et Max(A) sont le minimum et le maximum de l'attribut A et représentant les bornes inférieure et supérieure de cette attribut respectivement. Tandis que V′ est la nouvelle valeur normalisée du nouvel attribut A′ qui varie sur la plage bornée par Min(A′) et Max(A′).

- **La normalisation Z-score (Z-score Normalization) :** La présence de valeurs aberrantes peut affecter la normalisation Min-Max. La solution est la normalisation Z-score. C'est la transformation de toutes les valeurs V_i de données capturées par rapport à la moyenne μ arithmétique et l'écarte-type σ selon la formule suivante :

$$\boldsymbol{V_i}' = \frac{\boldsymbol{V_i - \mu}}{\boldsymbol{\sigma}} \qquad (3.2)$$

dont :

$$\mu = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{V_i} \qquad (3.3)$$

et :

$$\boldsymbol{\sigma} = \sqrt{\frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{V_i} - \mu)^{\boldsymbol{2}}} \qquad (3.4)$$

Ceci revient à centrer les données et est essentiel dans certaines méthodes comme l'Analyse en Composantes Principales.

- **Normalisation de l'échelle décimale (*Decimal Scaling Normalization*) :** C'est la transformation qui divise les valeurs numériques V_i par rapport à la puissance de dix.

$$\boldsymbol{V_i'} = \frac{\boldsymbol{V_i}}{\boldsymbol{10^J}} \qquad (3.5)$$

La puissance J est le nombre de chiffres de la partie entière de la valeur absolue maximale dans les données. Par exemple, si V_i=13,5 est la valeur absolue maximale alors J=2 parce que nous avons deux chiffres dans la partie entière égale à 13.

F. Construction d'attributs : Il s'agit d'une transformation de données qui consiste à ajouter un autre attribut basé sur d'autres existants auparavant. Cette opération vise à améliorer la précision du modèle et parfois contrôle mieux la régularité des données. A titre d'exemple, dans un ensemble de données nous pouvons ajouter un attribut de *Surface* en basant sur la *Longueur* et la *Largeur*. Le contrôle sur la Surface rend plus facile la détection des instances non appropriés.

La conversion de données **nominales en binaires** est une variante de ce type de transformation. Il est utile lorsqu'il y a des attributs nominaux dans le dataset et qui seront classées par des techniques qui ne prennent pas le type de données nominal en compte. Par conséquent, ce type se base sur un seul attribut à la fois.

Comme l'on a vu précédemment lors de l'ajout de la surface calculée, les nouveaux attributs peuvent être calculés en fonction de combinaison de plusieurs attributs. Pour bien le comprendre, si on suppose qu'on a un sous-ensemble d'attributs $A_1,A_2,\ldots,A_p$ formé à la base de l'ensemble d'attributs $A_1,A_2,\ldots,A_n$, ces combinaisons peuvent être obtenues par :

- **Transformation linéaire :**

$$\boldsymbol{V_i'} = \boldsymbol{r_1 B_1 + r_2 B_2 + \cdots + r_p B_p} \qquad (3.6)$$

dont : $r_1, r_2, \ldots, r_p$ sont des coefficients estimés ou calculés

- **Transformation quadratique :**

$$V_i' = r_{1,1}B_1^2 + r_{1,2}B_1B_2 + \cdots + r_{p-1,p}B_{p-1}B_p + r_{p,p}B_p^2 \qquad (3.7)$$

dont : $r_{1,1}, r_{1,2}, \ldots, r_{p-1,p}, r_{p,p}$ sont des coefficients estimés ou calculés.

- **Transformation à base de réduction de données :** C'est une transformation qui vise à réduire la taille de données avec le maintien de certaines caractéristiques dedans tel que l'intégrité de données et les informations portées sur les données. L'analyse des composantes principales (Principal Component Analysis (PCA)) est l'une des méthodes les plus célèbres. Nous verrons plus loin dans le manuscrit de la réduction de données.

On peut encore citer d'autres transformations comme **Rank Transformations** et **Box-Cox Transformations** (García et al., 2015), mais nous ne les détaillons pas car elles sont peu utilisées du côté famille de données médicales.

3.2.3 Représentation des types de données de base

Ce sont les types de données de base constituants le contenu des attributs à classifier. Principalement, leur utilisation est destinée pour représenter une variable liée à une instance dans une population d'individus (Nettleton, 2014). Ils comportent les données numériques, catégoriques ordinales (Valeurs nominales dont l'ordre comporte un sens significatif), catégoriques nominales (Valeurs nominal, mais sans aucune signification portée sur leur ordre), Binaire et Date. Après certains traitements de toutes les données, la plupart des données traitées et transformées peuvent prendre une ou plusieurs formes de ces types.

3.2.4 Représentation des types de données avancés

A. Type hiérarchique : Les données sont organisées dans une structure arborescente, qui est généralement parcourue de haut en bas ou à l'inverse de bas en haut. Par exemple, l'implémentation d'un fichier XML qui contient les données sur les types des maladies peut se représenter sous une forme arborescente.

B. Réseaux sémantiques : Ce type est utilisé généralement pour figurer les informations linguistiques sur les données ou sur les objets à représenter globalement. Par exemple on peut citer l'utilisation d'une ontologie pour décrire les maladies, les médicaments et les procédures.

C. Données graphiques : Cette représentation utilise un ensemble des nœuds pour montrer la présence des objets, et figure les autres informations de communication entres eux par des arêtes. Par exemple les réseaux sociaux peuvent être représentés par un graphe, dont les nœuds sont les personnes et les arêtes présentés par une relation d'amitié.

D. Données floues : Suite à les données d'un item, l'appartenance de cet item à plusieurs catégories peut être réalisée d'une manière incertaine. Par conséquent, cette appartenance sera estimée par des valeurs sur un intervalle de zéro à un, dont la somme de ces estimations sera « 1 ». L'appartenance totale à une catégorie prend la valeur « 1 », et à l'inverse la non appartenance prend un « 0 ». Par exemple on peut citer le profil d'un client dans une banque qui sera évalué sur quatre catégories de risque « Aucun, Léger, Modéré, Elevé », et dont le résultat de l'évaluation donne un risque modéré avec une estimation de 0.3 et un risque élevé par 0.7.

3.2.5 Représentation des séries de données temporelles

Les données temporelles sont des suites d'observations capturées et ordonnées chronologiquement (Fu, 2011). Il existe beaucoup de techniques et propositions pour

leurs représentations, ces approches sont classées généralement en quatre types illustrés sur la Figure 3.1 (Aghabozorgi et al., 2015 ; Bagnall et al., 2006 ; Özkoç, 2021):

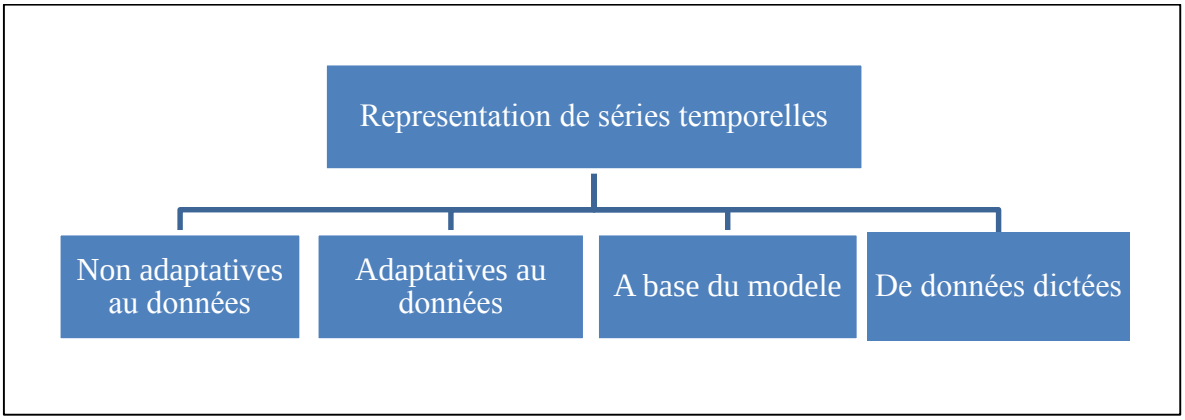

Figure 3.1. Approches de représentation des séries temporelles.

A. Approches non-adaptatives aux données : Ces approches sont mieux utilisées sur les séries temporelles dont les segments ont la même taille. Elles appliquent les mêmes paramètres de transformations sur toutes les séries. Par exemple :

- **Piecewise Aggregation Approximation (PAA) :** C'est une technique de réduction et de représentation de données par moyenne sur des séries normalisées. Elle commence par le partitionnement de la série temporelle X= (x_1, x_2, …, x_n) de longueur n en N segments dont chacune est de longueur n/N et N ≤ n. Par la suite, le vecteur S= (s_1, s_2,..s_N) de la nouvelle représentation sera généré en fonction de la moyenne de chaque segment selon l'équation suivante (Wilson, 2017):

$$s_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j \qquad (3.8)$$

A titre d'exemple, si on considère la série X= (0.5,-0.5,-1,0,0.5,1,0.5,0.5,0,-1) de longueur n=10 et N=2 alors la série S générée doit être S= (-0.1,0.2). La Figure 3.2 montre la vue graphique de cet exemple.

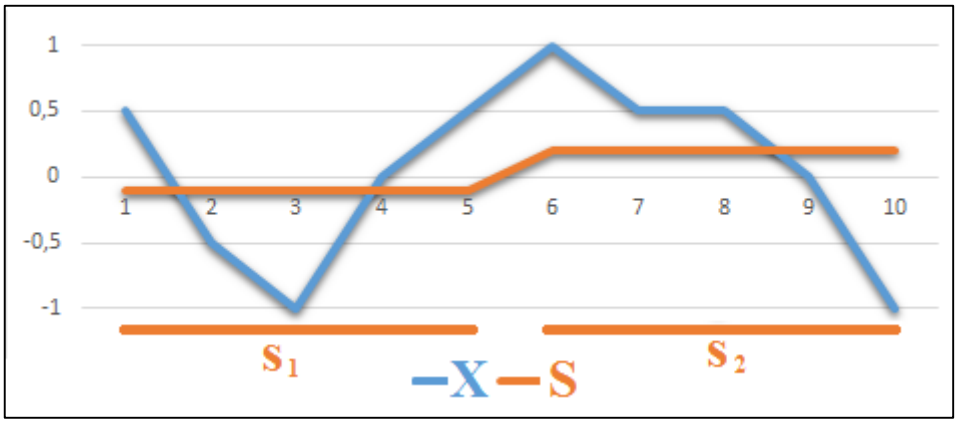

Figure 3.2. Exemple illustratif de la technique PAA.

- **Représentations à base de transformation :** Utilisent plusieurs techniques de transformation telle que la Transformée de Fourier Discrète, la Transformée en Cosinus Discrète, la Décomposition en Valeurs Singulières afin d'extraire des représentations pour les séries temporelles (Bettaiah et Ranganath, 2014).

B. Approches adaptatives aux données : Ce sont des approches qui utilisent les séries temporelles dont les segments ont des tailles différentes. Globalement, les paramètres de transformation sur les séries changent selon les données disponibles.

- **Symbolic Aggregate Approximation (SAX) :** Consiste à normaliser la série temporelle, puis appliquer la méthode PAA. Elle utilise les symboles d'un alphabet afin de représenter chaque segment (Wilson, 2017). Selon la description utilisée dans la technique PAA, nous avons la série X (|X|=n) originale et la série S (|S|=N) produite par approximation. L'alphabet A est prédéfini par $A=\{a_1,a_2,\ldots,a_m\}$, où a_i $(1 \leq i \leq m)$ sont des symboles et chaque symbole a_i représente un intervalle sur lequel varie la moyenne s_i. Les limites

des intervalles sont définis en fonction de points de rupture « Breakpoints (B) » B=<B_1, B_2,… B_{m-1}> déterminant les m espaces égaux sous la courbe gaussienne. Les auteurs de la technique SAX (Patel et al., 2002) donnent un exemple de ces points B_i pour $(3 \leq i \leq 10)$ selon la Table 3.1.

m	3	4	5	6	7	8	9	10
B1	-0,43	-0,67	-0,84	-0,97	-1,07	-1,15	-1,22	1,28
B2	0,43	0	-0,25	-0,43	-0,57	-0,67	-0,76	-0,84
B3		0,67	0,25	0	-0,18	-0,32	-0,43	-0,52
B4			0,84	0,43	0,18	0	-0,14	-0,25
B5				0,97	0,57	0,32	0,14	0
B6					1,07	0,67	0,43	0,25
B7						1,15	0,76	0,52
B8							1,22	0,84
B9								1,28

Table 3.1. Breakpoints de division de l'espace sous la courbe gaussienne.

Par conséquent, la nouvelle représentation R= $r_1, r_2, \ldots r_m$ est produite selon la correspondance entre les symboles a_i et la moyenne s_i et définit par:

$$\boldsymbol{Si\ B_{i-1} < s_j \leq B_i \Longrightarrow r_j = a_j} \qquad (3.9)$$

La classification peut être l'un des aspects de l'utilisation de la technique SAX. A titre d'exemple, la modélisation dans l'approche SAX-SVM (Senin et Malinchik, 2013) (SAX and Vector Space Model) est composée de deux phases. La première consiste à appliquer la technique SAX sur les séries temporelles afin de les représenter. Par la suite, elle transforme toutes les collections trouvées en vecteur de fréquence des termes en fonction des poids générés par la technique Tf-Idf. Puis une deuxième phase classifie les vecteurs représentants des séries non étiquetées par rapport aux vecteurs de la première phase en fonction de la distance cosinus maximale.

- **Piecewise Linear Approximation (PLA) :** Représente chaque segment par une ligne droite entre ses deux limites, et qui soit la meilleure parmi les lignes trouvées par l'application de la technique de régression linéaire (Bettaiah et Ranganath, 2014).

- **Adaptive Piecewise Constant Approximation (APCA) :** Les segments ont des longueurs différentes, et sont générés relativement au nombre de données durant un intervalle temporel. La longueur de chaque segment correspond au nombre d'activités dans cet intervalle du temps (Wilson, 2017).

C. Approches à base du modèle : C'est l'ensemble des approches qui génèrent des représentations en fonction d'un modèle de base, tel que la technique Hidden Markov Model (HMM). L'invocation de ces approches implique la recherche des paramètres adéquats aux modèles correspondants.

D. Approches de données dictées : Ce sont les approches qui utilisent des coupures de données temporelles afin de les représenter. Certains paramètres de ces approches sont déterminés automatiquement, comme le taux de réduction de dimension sur les séries. Le travail de Ratanamahatana et al., (2005) est un exemple de ces approches qui utilisent la méthode Clipped Data.

Actuellement, différents domaines tels que l'industrie, le web et les réseaux sociaux, génèrent et utilisent des données volumineuses. Généralement, la dimensionnalité de leurs données est originalement élevée. Pour ces domaines, et face aux contraintes applicatives, il est nécessaire de satisfaire certains critères tel que le temps maximal de la réponse. De plus, les méthodes de Machine Learning sont sujettes à la malédiction de la dimensionnalité et leurs performances peuvent décroitre sur des données volumineuses. En conséquence, dans la partie suivante, nous parlerons de la réduction de dimension et des techniques associées.

3.3 Réduction de données

Les données de la médecine personnalisée, comme d'autres domaines, peuvent être volumineuses. Elles comportent un grand nombre de participants qui sont décrits parfois par un ensemble d'attributs de taille importante. Les attributs peuvent apparaître dans l'ensemble de données comme une source de bruit si elles ne portent aucune information. Tantôt, ils sont présentés d'une manière redondante ou corrélée. De plus, certains processus décisionnels ne considèrent que les attributs les plus importants et ignorent les autres lors de l'implémentation. Outre que la nécessité de représentation et de transformation de données, les critères supposés et les objectifs fixés lors de la modélisation montrent le rôle important de la réduction de dimension.

La réduction de dimension est le processus de transformation d'un ensemble de données considéré de grande dimension (nombre d'attributs important) en un autre ensemble de petite dimension (généralement, 2 ou 3 attributs, surtout à des fins de visualisation). Durant les traitements, ce processus tente de garder les informations portées par les observations initiales. Généralement, les tâches de traitement, visualisation, analyse et interprétation en fonction de données réduites deviennent plus simples, mais les résultats diffèrent selon la stratégie de la technique adoptée.

3.3.1 Techniques de réduction de dimension

Étant donné l'importance de la réduction de la dimension, de nombreuses méthodes ont étés développées dans la littérature. Elles sont classées en deux groupes (Liu, 2011 ; Kira et al., 2020 ; Ayesha et al., 2020 ; Murty et Devi, 2015) :

A. Techniques de sélection des attributs : Ce sont des méthodes basées sur le choix du sous-ensemble des attributs les plus pertinents et qui accompagnent obligatoirement la meilleure solution du problème posé. Par conséquent, la suppression d'un attribut pertinent implique la détérioration de la précision du modèle. Cependant, les attributs non pertinents éliminés ne doivent pas nuire la performance du processus. La réduction par sélection des attributs est utilisable

surtout dans les applications nécessitantes la conservation des attributs originaux. Globalement, les approches de cette classe peuvent être catégorisées en :

- **Méthodes de filtrage :** Utilisent des critères (tel que la variance) pour la sélection des attributs indépendamment de la technique de classification appliquée. Dans la plupart des cas, ces méthodes sont plus rapides que celles des autres catégories.

- **Wrappers :** Ils adoptent une politique de choix des attributs par recherche et l'évaluation de la collection sélectionnée par la technique de classification déterminée. La stratégie de la recherche est appliquée par l'élimination des attributs non pertinents dans une phase de retour en arrière (Backward elimination) et par adjonction dans un ensemble plus large tous les attributs prometteurs lors de la phase d'avancement (Forward selection). Généralement, les wrappers demandent plus de temps lors des calculs parce qu'ils exécutent plusieurs fois la technique de classification pour évaluer chaque collection d'attributs.

- **Méthodes intégrées (Embedded) :** Pour surpasser la lourdeur du calcul important effectué par les méthodes Wrappers et l'indépendance entre la sélection des attributs et la technique de classification de méthodes de filtrage, cette catégorie combine les deux catégories précédentes dans un seul processus.

B. Techniques d'extraction des attributs : Regroupent les méthodes visent la transformation de l'ensemble des attributs de base en un nouvel ensemble réduit. Elles appliquent des traitements et des combinaisons sur tous les attributs afin de créer des attributs limités en nombre. Ces derniers préservent le maximum possible les informations des données originales dans une tâche de représentation fidèle à la nature des observations de la source.

Plusieurs critères de catégorisation des méthodes de réduction de dimension ont étés mentionnés par Lee and Verleysen, (2007). Mais le critère de la linéarité de la solution a été largement utilisé (Kira et al., 2020 ; Ayesha et al., 2020). Par conséquent, la réduction linéaire et la réduction non-linéaire sont devenues les catégories les plus adoptées lors du regroupement des techniques de la réduction de la dimension.

C. Réduction linéaire de dimension : Cette catégorie tire parti de la simplicité de représentation réalisée par la transformation linéaire sur les données originales. Plusieurs méthodes de réduction linéaires ont été développées. Le travail de Ayesha et al., (2020) est une étude comparative sur la réduction linéaire de dimensionnalité de données. Il explique et résume certaines techniques de réduction linéaire. Parmi ces techniques : Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Multidimensional Scaling (MDS), Latent Semantic Analysis (LSA) et d'autres. Les deux techniques PCA et MDS ont étés largement utilisées dans le domaine de la médecine.

- **L'Analyse en Composantes Principales :** En anglais elle s'appelle Principal component analysis (PCA). C'est une technique linéaire non supervisée utilisée pour l'extraction des attributs. A l'origine, elle a été introduite par Pearson, (1901), mais plus tard elle a été développée par Hotelling, (1933). L'ACP vise à maximiser la variance des nouvelles données dans l'espace de projection. Les nouveaux attributs non corrélés sont connus par les composantes principales (CP). Les d composantes principales possédants les variances maximales (VM) seront représentés dans le nouvel espace de projection réduit déterminé par les d axes principaux (Chaque CP_i possède VM_i et $VM_1 \geq VM_2 \geq \ldots \geq VM_d$).

  Les travaux de Shukla et al., (2019) et Raymond et al. (2019) sont des exemples d'utilisation de la technique ACP. Le premier travail (Shukla et al., 2019) utilise l'ACP pour la réduction de données dans un processus de classification des caractéristiques de la dysphonie autour de la maladie de Parkinson. Tandis que le deuxième travail (Raymond et al., 2019), est une étude qui

applique l'ACP pour la visualisation et l'interprétation des résultats de la maladie Lupus érythémateux systémique « Systemic Lupus Erythematosus ».

- **Positionnement multidimensionnel :** En anglais elle s'appelle Multidimensional Scaling (MDS). C'est une technique de réduction de dimension linéaire non supervisée. Elle a été introduite par Kruskal et Wish, (1978). Cette technique vise la représentation de données par un autre espace faible dimension, mais sous la contrainte de préservation de distances entres les paires de données. Bengio et al., (2003) ont proposé un algorithme général qui synthétise plusieurs techniques de réduction, y compris la technique MDS. Un exemple d'utilisation de cette technique a été réalisé par Vital et al., (2019). Ils ont utilisé la technique MDS pour réduire les dimensions de données, visualiser, analyser et rapporter les résultats.

D. Réduction non-linéaire de la dimension : Cette catégorie englobe les techniques de réduction de dimension produisant une représentation qui ne suit pas une correspondance linéaire entres les données de l'espace de grand dimension en entrée et l'espace de faible dimension (Réduit) de la sortie. Généralement, les techniques de réduction non linéaires adoptent les idées d'utilisation des noyaux (Kernelisation) et la découverte des structures de données pour représenter les données. Les techniques Kernel Principal Component Analysis (KPCA), Isomap, Locally Linear Embedding (LLE), Self-Organizing Map (SOM), t-Stochastic Neighbor Embedding (t-SNE) et autres sont des exemples de cette catégorie (Ayesha et al., 2020 ; Yang et al., 2015). Les techniques KPCA et t-SNE sont largement utilisées dans nombreuses disciplines, y compris le domaine médical.

- **Analyse en Composantes Principales à Noyaux :** En anglais elle s'appelle Kernel Principal Component Analysis (KPCA). C'est une technique proposée par Schölkopf et al. (1997). En effet, c'est la version kernelisée de la technique PCA. L'idée initiale de cette analyse est la conversion de l'espace de représentation originale dans un autre espace kernelisé. Elle applique l'ACP

sur les données transformées par une fonction noyau (Kernel) qui est considérée la plus adéquate afin de mettre en évidence les caractéristiques non linéaires des données (Ayesha et al., 2020 ; Fauvel et al., 2006). Par conséquent, plusieurs noyaux ont été proposés et les exemples suivants sont largement utilisés (Hofmann et al., 2008 ; Vert et al., 2004) :
En supposant l'ensemble des instances (Observations) X= $\{x_1, x_2, \ldots, x_n\}$, dont x_i est l'instance numéro i et n représente la taille de X.

Noyau polynomial (Polynomial kernel):

$$\boldsymbol{k(x_i, x_j) = (x_i . x_j + q)^d} \quad (3.10)$$

Noyau de fonction de base radiale (Radial basis function kernel (RBF)) :

$$\boldsymbol{k(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)} \quad (3.11)$$

dont q est une constante pour l'espace a noyau et d représente le degré de nouveau polynôme. σ est un paramètre à ajuster.

La technique KPCA a été beaucoup utilisée pour le domaine médical. A titre d'exemple l'approche du Lu et al., (2019) essaye de résoudre le problème de l'inséparabilité linéaire des données. Elle applique la technique KPCA avec d'autres techniques sur des sous-ensembles d'expression de gènes de plusieurs maladies. Dans un autre travail (Jiang et al., 2019) propose, pour la classification de maladies, de réduire la dimensionnalité par KPCA et d'appliquer une technique de classification.

- **t-Stochastic Neighbor Embedding (t-SNE) :** C'est une technique de réduction de dimension et de visualisation de données développée par Maaten et Hinton, (2008). Elle est basée sur la distribution des données et leurs similitudes pour les représenter dans un nouvel espace de faible dimension

(deux ou trois). La nouvelle distribution représente plus étroitement les données ayant une forte probabilité de similitude. Ainsi, les données les plus dissemblables sont représentées plus loin les unes des autres.

L'utilisation de la technique t-SNE sur les données médicales a été le sujet de plusieurs travaux. Par exemple, pour définir les sous-types de la maladie de Parkinson et après la représentation des données et le calcul de la similarité, Zhang et al., (2019) ont utilisé la technique TNSE pour réduire la dimensionnalité suivie par une application d'une technique de clustering. Pour évaluer l'efficacité de l'apprentissage profond (Deep learning) sur des données cliniques temporelles, Workman et al., (2018) proposent un processus de classification qui applique une réduction t-SNE sur la représentation générée par deep learning et le complètent par une tâche de classification.

Généralement, l'invocation des techniques de réduction de dimensionnalité ne représente qu'une seule étape dans un processus d'analyse, visualisation ou de classification de données. Les besoins derrière ces activités obligent le développement des outils adéquats qui facilitent le processus d'exploration de données et de prise de décisions. La partie suivante discute globalement de techniques de la classification.

3.4 Classification de données

En vue du chapitre précédent, la segmentation (ou clustering) est la tâche non supervisée qui a pour but la construction de groupes d'instances les plus homogènes. Les groupes produits sont homogènes et les plus hétérogènes entres eux. La classification est la tâche basée sur un apprentissage supervisé qui vise à prédire la catégorie ou le groupe d'une instance donnée. Plusieurs techniques ont été développées pour accomplir ces deux tâches. Cependant, le calcul de la distance entres les données est un élément essentiel pour la plupart de ces techniques. À cet effet, et avant de décrire les techniques de segmentation et de classification les plus célèbres, nous décrivons certaines distances qui sont reconnues de par leur utilisation fréquente en data mining.

3.4.1 Distances en data mining

Les distances suivantes sont utilisées si les données ne comportent que des valeurs numériques (Lengyel et Botta-Dukát, 2021 ; Kabasakal et Soyuer, 2021 ; Han et al., 2012). En supposant qu'on a deux instances X= $x_1,x_2…x_m$ et Y= $y_1,y_2…,y_m$ dont $x_i \in R$ et $y_i \in R$, on définit :

- **Distance Euclidienne :**

$$\boldsymbol{d(X,Y) = \sqrt{\sum_{i=1…m}(x_i - y_i)^2}} \quad (3.12)$$

- **Distance Manhattan (City block) :**

$$\boldsymbol{d(X,Y) = \sum_{i=1…m}|x_i - y_i|} \quad (3.13)$$

- **Distance Minkowski:**

$$\boldsymbol{d(X,Y) = \sqrt[p]{\sum_{i=1…m}|x_i - y_i|^p}} \quad (3.14)$$

- **Distance cosinus :** Elle est basée sur le calcul de la similarité.

$$\boldsymbol{Sim(X,Y) = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} x_i{}^2}\sqrt{\sum_{i=1}^{m} y_i{}^2}}} \quad (3.15)$$

$$\boldsymbol{d(X,Y) = 1 - Sim(X,Y)} \quad (3.16)$$

D'autre part, il existe des distances qui sont utilisées seulement avec les données non numériques. On suppose alors qu'on a deux instances X= $x_1,x_2…x_m$ et Y= $y_1,y_2…,y_m$ dont x_i et y_i sont des valeurs catégoriques.

- **Distance de Jaccard :** Elle est basée sur le calcul de l'index de Jaccard *J*. Cet index calcule le nombre des valeurs x_i et y_i communes par rapport à l'union des valeurs de ces deux instances (Levandowsky et Winter, 1971).

$$\boldsymbol{J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}} \quad (3.17)$$

$$\boldsymbol{d(X,Y) = 1 - J(X,Y)} \quad (3.18)$$

- **Distance de Levenshtein :** C'est l'une des distances utilisées pour le calcul de la proximité entre deux chaines de caractères. Elle est compte le nombre des opérations élémentaires d'insertion, de suppression et de substitution à appliquer sur la chaine X afin de la transformer en Y (Behara et al., 2020).

- **Distance de Hamming :** Elle calcule le nombre des composantes xi et yi inégales, en tenant compte de leurs positions i dans X et Y (Yang et Wang, 2007).

$$\boldsymbol{d(X,Y) = \sum_{i=1\ldots m} 1_{x_i != y_i}} \quad (3.19)$$

3.4.2 Techniques de segmentation (Clustering)

Globalement, les stratégies adoptées par les techniques du clustering pour former les groupes (Classes ou clusters) permettent de les organiser en (Han et al., 2012) :

A. Méthodes de partitionnement : Elles représentent l'ensemble des techniques qui produisent k clusters (Partitions), et chaque cluster comporte au moins une seule instance. De plus, la vue d'un seul niveau de toutes les données doit accompagner le processus de production de groupes c.-à-d. elle ne procède pas à une hiérarchisation de données. Ces méthodes empêchent l'appartenance multiple des instances à plusieurs groupes. Techniquement, la plupart sont des méthodes basées sur la distance. Par conséquent, chaque cluster ne comporte que les instances les plus proches. Inversement et comparativement entre les clusters, la distribution des instances sur les différents groupes représente des données qui sont les plus éloignées

d'un groupe à l'autre. Les méthodes k-means (Hartigan et Wong, 1979) et k-medoids (Kaufmann et Rousseeuw, 1987) sont deux exemples célèbres de méthodes de partitionnement.

B. Méthodes hiérarchiques : Ce sont les méthodes qui adoptent la formulation hiérarchique des groupes d'instances selon une vue par niveau. A chaque itération, leur processus applique une tâche de division ou d'assemblage des groupes d'instances. En effet, deux types d'approches correspondent à ces deux tâches les approches ascendantes (agglomératives) et les approches descendantes (Divisives). Le premier type considère chaque instance come un cluster individuel. Itérativement, il fusionne les groupes tant qu'il y en a plusieurs ou que la condition d'arrêt n'a pas été atteinte. Par contre, le deuxième type considère l'ensemble des instances comme un seul groupe. Il divise itérativement les groupes jusqu'à ce qu'une seule instance par groupe soit atteinte ou qu'une condition d'arrêt soit satisfaite. Certaines méthodes hiérarchiques utilisent des distances et des critères de liaison entres les clusters comme des indicateurs de division ou de fusion. Parmi ces critères nous pouvons citer les suivant :

En suppose qu'en a deux clusters $C_i=\{X_1,X_2,\ldots,X_r\}$ et $C_j=\{Y_1,Y_2,\ldots,Y_p\}$

- **Distance minimum (Single-linkage clustering) :**

$$\boldsymbol{d_{min}(C_i, C_j) = \min_{X\in C_i, Y\in C_j} \{d(X,Y)\}} \quad (3.20)$$

- **Distance maximum (Complete-linkage clustering) :**

$$\boldsymbol{d_{max}(C_i, C_j) = \max_{X\in C_i, Y\in C_j} \{d(X,Y)\}} \quad (3.21)$$

- **Distance moyenne (Average linkage clustering) :**

$$\boldsymbol{d_{avg}(C_i, C_j) = \operatorname{avg}_{X\in C_i, Y\in C_j} \{d(X,Y)\}} \quad (3.22)$$

- **Distance minimum (Ward linkage clustering) :**

$$d_{ward}(C_i, C_j) = \frac{r*p}{r+p} d(G_i, G_j) \quad (3.23)$$

dont G_i et G_j sont les centres de gravité du clusters C_i et C_j respectivement.

C. Méthodes basées sur la densité : Ce type de méthodes se base sur la densité de données défini par le nombre d'instances dans un voisinage limité par un seuil supposé. Par conséquent, toutes les instances voisines à ce cluster et respectant ce critère seront incubées. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Zhang, 2019) et Ordering Points to Identify the Clustering Structure (OPTICS) (Ankerst et al., 1999) sont des exemples de telles méthodes.

D. Méthodes basées sur les grids : Ces méthodes divisent l'espace de données en une structure de Grid composé par des cellules. Par la suite, elles regroupent les cellules sous un ensemble des critères afin de former les clusters. Cette stratégie est basée initialement sur l'exploration de cellules plutôt que les instances, ce qui rend les calculs rapides car le nombre de cellules est très limité par rapport au nombre gigantesque des instances. STatistical INformation Grid (STING) (Wang et al., 1997) et CLustering In QUEst Method (CLIQUE) (Agrawal et al., 1998) sont des exemples de techniques de clustering basées sur les Grids.

E. Méthodes basées sur les graphes : C'est l'ensemble des méthodes basé sur l'utilisation de la théorie des graphes. Elles considèrent les instances comme des sommets du graphe et forment les arêtes par la relation entre les sommets éventuellement par la distance entres eux. Généralement, ce type de clustering vise à produire des sous graphes à partir de la proximité des sommets et la suppression de certaines arêtes non importantes. Shared Nearest Neighbor (SNN) (Ertöz et al., 2002), IncSNN-DBSCAN (Singh et Awekar, 2013) Minimum Spanning Tree (MST) (Zahn, 1971) sont des exemples d'algorithmes de cette type de clustering.

F. Méthodes basées sur les Modèles : Ces méthodes adoptent l'idée que les données ont une distribution statistique selon un modèle donné. Ce type de clustering permet l'affectation multiple d'une instance donnée à plusieurs clusters par une probabilité d'appartenance. De plus, ce type a la capacité de définir automatiquement le nombre optimal des clusters. SVM-based clustering et COBWEB ce sont des exemples de ces méthodes.

Généralement, la médecine utilise beaucoup les techniques de partitionnement pour l'automatisation de la prise de la décision médicale. Par exemple l'approche de Singh et al., (2021) utilise le k-means dans une application qui diagnostique la maladie du foie (le diagnostic de l'hépatite). L'étude du Peker, (2016) a pour but le perfectionnement de la précision de la classification. Après le prétraitement de données elle applique le clustering par la technique k-medoids et continue par d'autres tâches qui visent à affecter des poids aux attributs à étiqueter les données. L'expérimentation de cette étude a été faite sur les datasets des maladies cardiaques, de la maladie de Parkinson et les troubles hépatiques. L'approche de Hahn et al., (2021) est une étude sur l'insuffisance cardiaque (Heart failure). Sur les données de patients (incluant des données génétiques) les auteurs appliquent la technique PCA et le clustering hiérarchique. Ceci conduit à la découverte de certaines caractéristiques génétiques de cette maladie en plus de l'identification des sous-groupes de la maladie considérée.

L'étude de Leal et al., (2021) est un autre exemple qui applique les techniques de clustering. Elle applique k-means, Agglomerative hierarchical clustering, DBSCAN et Expectation-maximization (EM) (Dempster et al., 1977) sur les données des enregistrements d'électrocardiogramme (ECG) traités. L'étude de Mirmozaffari et al., (2017) sur la prédiction des crises cardiaques et les techniques les plus efficaces, a testé plusieurs méthodes du clustering. Canopy, Cobweb, EM, Hierarchical Clusterer, Make Density Based Clusterer (MDBC) et k-means sont les techniques testées. Globalement, le domaine de décision médicale est très riche en termes d'application

des techniques de clustering, et les exemples cités indiquent partiellement le cas applicatif correspondant.

3.4.3 Indices de qualité du clustering

Pour évaluer la qualité du clustering plusieurs indices d'estimation ont été proposés. Beaucoup d'être eux se basent sur le calcul de la distance. Parmi ces indices nous pouvons citer :

A. Les indices inertiels : Ils comportent l'indice de l'inertie intraclasse et l'indice de l'inertie interclasse (Lebart et al., 1995). L'utilisation de tels indices pour l'évaluation de la qualité du clustering est plus populaire que les autres.

L'inertie intraclasse (**Within-class Inertia *(IW)***) exprime l'homogénéité des instances de même cluster, et si elle est faible alors le cluster est homogène. Elle est mesurée par l'équation suivante :

$$\boldsymbol{IW = \frac{1}{n}\sum_{i=1}^{k}\sum_{X\in C_i} d^2(X, G_i)} \qquad (3.24)$$

L'inertie interclasse (**Between-class Inertia *(IB)***) exprime l'hétérogénéité entre les clusters formés, et si elle est élevée alors les deux clusters sont bien séparés. Elle est mesurée par l'équation suivante :

$$\boldsymbol{IB = \frac{1}{n}\sum_{i=1}^{k} n_i \,.\, d^2(G_i, G)} \qquad (3.25)$$

dont G représente le centre du gravité du nuage de toutes les instances, n est le nombre totale des instances, G_i est le centre du gravité du nuage des instances du cluster C_i et n_i est le nombre des instances du clusters C_i.

B. L'indice de Dunn : Cet indice vise à évaluer le résultat du clustering vis à vis de la compacité et la séparabilité des clusters. Par conséquent, il calcule le rapport entre la distance minimale entres les classes différentes sur la distance maximale

entres les instances de la même classe. L'équation suivante exprime cet indice clairement (Dunn, 1974) :

$$\boldsymbol{Dunn} = \frac{\min\limits_{1\le i<j\le k}\{d_{min}(C_i,C_j)\}}{\max\limits_{1\le l\le k} D_{max}(C_l)} \tag{3.26}$$

dont

$$\boldsymbol{d_{min}(C_i, C_j)} = \min_{X\in C_i, Y\in C_j} \boldsymbol{d(X,Y)} \tag{3.27}$$

et

$$\boldsymbol{D_{max}(C_l)} = \max_{X,Y\in C_l} \boldsymbol{d(X,Y)} \tag{3.28}$$

Il est clair que l'indice de Dunn vise à maximiser ce rapport (Equation 3.26) pour aboutir à un meilleur partitionnement.

C. L'indice de Davies-Bouldin (DB) : Par la prise en compte de la compacité exprimée par la moyenne des instances dans le même cluster M(C_i) et de la séparabilité exprimée par la distance entres les centres des clusters, cet indice calcule la qualité du clustering en fonction de l'équation suivante (Davies et Bouldin, 1979):

$$\boldsymbol{DB} = \frac{1}{k}\sum_{i=1}^{k} \max_{\substack{i\neq j \\ X,Y\in C_l}} \left\{\frac{M(C_i)+M(C_j)}{d(C_i,C_j)}\right\} \tag{3.29}$$

Plus la valeur évaluée de cet indice est faible alors est plus le partitionnement est bon.

D. L'indice Silhouette : C'est un indice basé sur l'évaluation du bon classement de chaque instance X_i dans son cluster en relation des autres clusters. L'indice de Silhouette de l'instance X_i noté par S(X_i) est calculé par (Rousseeuw, 1987) :

$$\boldsymbol{S(X_i) = \frac{a(X_i) - b(X_i)}{max\{a(X_i), b(X_i)\}}} \quad (3.30)$$

où $a(X_i)$ représente la distance moyenne entre l'instance X_i et les autres instances Y_i du même cluster :

$$\boldsymbol{a(X_i) = \frac{1}{|C^i| - 1} \sum_{Y_i \in C^i, X_i \neq Y_i} d(X_i, Y_i)} \quad (3.31)$$

et C^i ici est le cluster supposé qui inclut l'instance X_i. $b(X_i)$ représente la valeur minimale parmi les distances moyennes de l'instance X_i avec les instances Y_l de chaque cluster (C^l) séparément du cluster C^i auquel appartient l'instance X_i, et C={C_1, C_2,…,C_k}.

$$\boldsymbol{b(X_i) = \min_{C^j, C^i \in C, C^j \neq C^i, X_i \in C^i} \left\{ \sum_{Y_l \in C^j} \frac{d(X_i, Y_l)}{|C^j|} \right\}} \quad (3.32)$$

L'estimation de l'indice Silhouette $S(X_i)$ varie dans l'intervalle [-1,1]. Plus la valeur de $S(X_i)$ est proche de 1, meilleure est l'évaluation du classement de l'instance X_i. Lorsque la valeur de $S(X_i)$=0, alors l'affectation de l'instance X_i est moins fiable. Tandis que si la valeur de $S(X_i)$ proche de -1, X_i est mal classé.

Par conséquent, le nombre des clusters optimal peut être estimé en fonction de la mesure connue par la largeur de la Silhouette. Cette dernière est la moyenne de tous les indices $S(X_i)$ calculés :

$$\boldsymbol{S = \frac{1}{n} \sum_{i=1}^{n} S(X_i)} \quad (3.33)$$

Par la suite, 1-≤S≤1 et le nombre de clusters optimal correspond à la largeur de Silhouette S maximale.

3.4.4 Techniques de classification

Pour la catégorisation et la prédiction de données plusieurs techniques de classification ont été développées. Les techniques suivantes ne représentent que des exemples connus par leurs utilités dans différente domaines et particulièrement au

stade de la prise de la décision et l'exploration de données médicales (Jayatilake et Ganegoda, 2021; Babar et Mahoto, 2018).

A. Naïve Bayes (NB) : Est un modèle probabiliste de classification supervisée basé sur le théorème de Bayes et suppose l'indépendance entre les attributs. Ce modèle conditionnel calcule la probabilité postérieure des catégories de classe pour les observations d'entrée. Les observations sont classées selon la classe avec la probabilité postérieure maximale (Aggarwal et Vig, 2019 ; Salmi et Rustam, 2019).

B. Support Vector Machine (SVM) : Est une méthode d'apprentissage supervisé basée sur la transformation par l'utilisation des noyaux et la séparation des données par la maximisation des marges pour produire des hyperplans séparateurs. La simplicité d'utilisation du SVM et ses fondements théoriques justifient son utilité dans plusieurs domaines (Noble, 2006).

C. k-Nearest Neighbor (KNN) : L'algorithme du KNN utilise un ensemble d'apprentissage étiqueté pour classer une instance donnée. Cet algorithme utilise une distance pour trouver les k données les plus proches de cette entrée. Par un vote majoritaire, la classe la plus présente parmi ces k exemples proches sera affectée à cette observation (Al Bataineh, 2019; Sarkar et Leong, 2000).

D. Random Forest (RF) : Il s'agit d'une technique de classification et de régression qui utilise une combinaison d'arbres de décision aléatoires. Semblable à la technique de bagging, le RF utilise l'agrégation moyenne pour la régression et le vote majoritaire pour la classification. L'idée principale de cette technique est d'entraîner des arbres de décision sur différents sous-ensembles de données et des variables choisies au hasard (Ishwaran et Lu, 2018; Kavzoglu, 2017).

E. Decision Trees (DT) : Est une technique de classification représentant le processus de la catégorisation de données sous forme d'un arbre graphique. Elle est composée par des nœuds connectés par des branches. Le nœud racine constitue la seule entrée de cette arbre. Les nœuds qui ne possèdent pas des nœuds fils sont

appelés des feuilles, et ce sont ceux qui comportent les résultats d'affectation finales des instances (Les catégories finales). Le reste des nœuds identifie les attributs à tester (Les variables prédictives) et les branches portent les réponses à suivre en correspondance avec les valeurs des attributs testés. Il existe plusieurs algorithmes de construction d'arbres de décision tels que CART, C4.5 et C5.0 (Mishra et al., 2019; Abspoel et al., 2021).

F. Artificial Neural Network (ANN) : Est un outil de la classification supervisée inspiré de la biologie humaine et du fonctionnement du cerveau plus précisément. Formellement, il comporte un ensemble des nœuds organisés sous forme d'une couche d'entrée, de sortie et d'une ou plusieurs couches cachées. Les sorties des neurones des couches cachées sont obtenues par une moyenne pondérée par la connexion entre deux couches. Généralement, les ANN nécessitent certains paramètres important tels que la fonction de combinaison et la fonction d'activation. Théoriquement, plusieurs techniques ont été proposées à titre d'exemple les Multi-layer perceptron (MLP), Generalized regression neural networks (GRNN) et radial basis function neural networks (RBF) (Jayatilake et Ganegoda, 2021 ; Sadiq et al., 2019 ; Kisi, 2004).

Bien sûr il existe d'autres techniques de classification, mais les techniques citées sont suffisantes pour disposer d'un panorama des techniques de classification considérées comme populaires pour l'exploration de données médicales. Les approches de Peker, 2016 (Applique SVM), Salmi et Rustam, 2019 (Applique NB), Al Bataineh, 2019 (Applique MLP, KNN, CART, BN, SVM) et de Mishra et al., 2019 (Applique SVM, KNN, RF, DT) ne sont que des exemples d'application de ces classifieurs.

3.5 Conclusion

Pour ouvrir la voie vers nos contributions autour de l'exploration de données de la médecine personnalisée, ce chapitre a parlé brièvement de la représentation de

données, de la réduction de dimension, du clustering et de la classification de données. Nous avons présenté tout d'abord le problème de la représentation de données hétérogènes à travers différentes opérations. Ces opérations peuvent inclure des transformations et des opérations sur des types de données simples et des séries temporelles. La réduction de données est considérée comme l'une de ces opérations de la représentation de données. Elle a été expliquée avec la description de quelques techniques. Nous avons continué avec quelques explications sur le clustering et la classification de données.

Tout au long de ces explications, nous avons essayé de citer un ensemble des travaux récents sur les données médicales pour présenter l'actualité applicative de la plupart des opérations et techniques discutées. Dans le chapitre suivant, nous allons nous focaliser sur les problèmes associés à la tâche de représentation des données.

Références

- Abspoel, M., Escudero, D., & Volgushev, N. (2021). Secure training of decision trees withcontinuous attributes. *In Proceedings on Privacy Enhancing Technologies,* 167–187.

- Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering – A decade review. *Information Systems, Vol 53*, 16-38.

- Aggarwal, G., & Vig, R. (2019). Acoustic Methodologies for Classifying Gender and Emotions using Machine Learning Algorithms. Amity *International Conference on Artificial Intelligence, Dubai, United Arab Emirates,* 672-677.

- Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *In: Proc. ACM SIGMOD'98*, 94–105.

- Al Bataineh, A. (2019). A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection. *International Journal of Machine Learning and Computing, Vol 9*, 248-254.

- Ankerst, M., Breunig, M. M., Kriegel, H., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record, Vol 28*(2), 49–60.

- Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion, Vol 59*, 44-58.

- Babar, A. H., & Mahoto, N. A. (2018). Comparative Analysis of Classification Models for Healthcare Data Analysis. *International Journal of Computer and Information Technology, Vol 7*(4), 170-175.

- Bagnall, A., Ratanamahatana, C., Keogh, E., Lonardi, S., & Janacek, G. (2006). A Bit Level Representation for Time Series Data Mining with Shape Based Similarity. *Data Mining and Knowledge Discovery, Vol 13*, 11-40.

- Behara, K. N. S., Bhaskar, A., & Chung, E. (2020). A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance. *Transportation Research Part C: Emerging Technologies, Vol 111*, 513-530.

- Bengio, Y., Paiement, J. F., Vincent, P., Delalleau, O., Roux, N. L., & Ouimet, M. (2003). Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. *In Proceedings of the 16th International Conference on Neural Information and Processing Systems; MA, United States,* 177-184.

- Bettaiah, V., & Ranganath, H. S. (2014). An Analysis of Time Series Representation Methods Data Mining Applications Perspective. *Proceedings of the 2014 ACM Southeast Regional Conference on - ACM SE '14*.

- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans Patt. Anal Machine Intell, Vol* PAMI-1, 224–227.

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Vol 39*(1), 1–38.

- Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics, Vol 4*(1), 95-104.

- Duwe. K., Lüttgau, J., Squar, J., & Kuhn, M. (2020). State of the Art and Future Trends in Data Reduction for High-Performance Computing. *Supercomputing Frontiers And Innovations, Vol 7*(1), 4-36.

- Ertöz, L., Steinbach, M., & Kumar, V. (2002). A new shared nearest neighbor clustering algorithm and its applications. *In: Workshop on clustering high dimensional data and its applications at 2nd SIAM international conference on data mining*. 105–115.

- Fauvel, M., Chanussot, J., & Benediktsson, J. A. (2006). Kernel Principal Component Analysis for Feature Reduction in Hyperspectrale Images Analysis. *Proceedings of the 7th Nordic Signal Processing Symposium - NORSIG 2006, Reykjavik, Iceland.* 238-241.

- Fu, T. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence, Vol 24*, 164-181.

- García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining.* Intelligent Systems Reference Library 72. Springer, Cham.

- Hahn, V.S., Knutsdottir, H., Luo, X., Bedi, K., Margulies, K.B., Haldar, S.M., Stolina, M., Yin, J., Khakoo, A.Y., Vaishnav, J., Bader, J. S., Kass, D. A., & Sharma, K. (2021). Myocardial Gene Expression Signatures in Human Heart Failure With Preserved Ejection Fraction. *Circulation, Vol 143*(2), 120–134.

- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques (Third Edition)*. Publisher: Morgan Kaufmann.

- Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society, Vol 28*(1), 100–108.

- Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics, Vol 36*(3), 1171-1220.

- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, Vol 24*(6), 417–441.

- Ishwaran, H., & Lu, M. (2018). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine. Vol 38*(4), 558-582.

- Jayatilake, S. M. D. A. C., & Ganegoda, GU. (2021). Involvement of Machine Learning Tools in Healthcare Decision Making. *Journal of Healthcare Engineering, Vol. 2021*.

- Jiang, J. L., Li. S. Y., Liao, M. L., & Jiang, Y. (2019). Application in Disease Classification based on KPCA-IBA-LSSVM. *Procedia Computer Science, Vol 154*, 109-116.

- Kabasakal, İ., & Soyuer, H. (2021). A Jaccard Similarity-Based Model to Match Stakeholders for Collaboration in an Industry-Driven Portal. *Proceedings, Vol 74*(1).

- Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids . *In I. D. Y & editor (ed.), North Holland , Elsevier* , 405–416.

- Kavzoglu, T. (2017). *Object-Oriented Random Forest for High Resolution Land Cover Mapping Using Quickbird-2 Imagery.* In: Pijush S, Sanjiban SR, Valentina EB. Handbook of Neural Computation. London, UK: Academic Press; 607-619.

- Kisi, Ö. (2004). Multi-layer perceptrons with Levenberg-Marquardt training algorithm for suspended sediment concentration prediction and estimation / Prévision et estimation de la concentration en matières en suspension avec des perceptrons multi-couches et l'algorithme d'apprentissage de Levenberg-Marquardt. *Hydrological Sciences Journal, Vol 49*, -1040.

- Kruskal, J. B., Wish, M. (1978). *Multidimensional Scaling. Sage University Paper Series on Quantitative Applications in the Social Sciences*. Sage Publications, Newbury Park. 07-011.

- Leal, A.,Pinto, M. F., Lopes, F.,Bianchi, A. M., Henriques, J., Ruano, M. G., de Carvalho, P.,Dourado, A., & Teixeira, C. A. (2021). Heart rate variability analysis for the identification of the preictal interval in patients with drug-resistant epilepsy. *Sci Rep, Vol 11*.

- Lebart, L., Morineau, A., & Piron, M. (1995). *Statistique exploratoire multidimensionnelle.* Dunod, Paris.

- Lee, J. A., & Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer, New York, pp 69–97.

- Lengyel, A., Botta-Dukát, Z. (2021). Review and performance evaluation of trait-based between-community dissimilarity measures. *bioRxiv*.

- Levandowsky, M., & Winter, D. (1971). Distance between Sets. *Nature, Vol 234*, 34–35.

- Liu, H. (2011). *Feature Selection. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning,* Springer, Boston, MA.

- Lu, H., Meng, Y., Yan, K., & Gao, Z. (2019). Kernel Principal Component Analysis Combining Rotation Forest Method for Linearly Inseparable Data. *Cognitive Systems Research, Vol 53*, 111-122.

- Maaten, L. V. D., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, Vol 9*, 2579-2605.

- Mirmozaffari, M., Alinezhad, A., & Gilanpour, A. (2017). Heart Disease Prediction with Data Mining Clustering Algorithms. *Int'l Journal of Computing, Communications & Instrumentation Engg, (IJCCIE), Vol. 4*(1).

- Mishra, V., Singh, Y., & Rath, S. K. (2019). Breast Cancer detection from Thermograms Using Feature Extraction and Machine Learning Techniques. *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Pune, India.

- Murty, M. N., & Devi, V. S. (2015). *Chapter 3: Feature Extraction and Feature Selection. In: Introduction to Pattern Recognition and Machine Learning.* IISc Lecture Notes Series 5, pp:75-110.

- Nettleton, D. (2014). *Commercial Data Mining Processing, Analysis and Modeling for Predictive Analytics Project.* Morgan Kaufmann, MA, USA.

- Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology. Vol 24*, 1565-1567.

- Özkoç, E. E. (2021). *Clustering of Time-Series Data.* In Derya Birant (Ed.). Data Mining. Rijeka: IntechOpen.

- Patel, P., Keogh, E., Lin, J., & Lonardi, S. (2002). Mining Motifs in Massive Time Series Databases. *In Proc. ICDM,* 370-377.

- Pearson, K. F. R. S. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Vol 2*(11), 559-572.

- Peker, M. (2016). A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM. *Journal of medical systems, Vol 40*(5).

- Ratanamahatana, C. A., Keogh, E., Bagnall, A. J., & Lonardi, S. (2005). A Novel Bit Level Time Series Representation with Implications for Similarity Search and Clustering. *In: Proceedings of 9th Pacific-Asian International Conference on Knowledge Discovery and Data Mining(PAKDD'05),* 771–777.

- Raymond, W. D., Eilertsen, G. Ø., & Nossent, J. (2019). Principal component analysis reveals disconnect between regulatory cytokines and disease activity in Systemic Lupus Erythematosus. *Cytokine, Vol 114*, 67-73.

- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, Vol 20*, 53-65.

- Sadiq, R., Rodriguez, M. J., & Mian, H. R. (2019). *Empirical Models to Predict Disinfection By-Products (DBPs) in Drinking Water: An Updated Review, in: Nriagu, J. (eds), Encyclopedia of Environmental Health (Second Edition).* Elsevier, 324-338.

- Salmi, N., & Rustam, Z. (2019). Naïve Bayes Classifier Models for Predicting the Colon Cancer. *IOP Conference Series: Materials Science and Engineering. Vol 546*(5).

- Sarkar, M., & Leong, T. Y. (2000). Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. *In Proceedings / AMIA ... Annual Symposium. AMIA Symposium,* 759-763.

- Schölkopf, B., Smola, A., & Müller, K. R. (1997). Kernel principal component analysis. *Artificial Neural Networks — ICANN'97,* 583-588.

- Senin, P., & Malinchik, S. (2013). SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model. *Data Mining (ICDM), 2013 IEEE 13th International Conference on Data Mining (ICDM),* 1175,1180.

- Shukla, A. K., Singh, P., & Vardhan, M. (2019). *Medical Diagnosis of Parkinson Disease Driven by Multiple Preprocessing Technique with Scarce Lee Silverman Voice Treatment Data.* In: Ray K et al. (eds) Engineering Vibration, Communication and Information Processing. Lecture Notes in Electrical Engineering, vol 478, 407-421. Springer, Singapore.

- Singh, A., Mehta, J. C., Anand, D., Nath, P., Pandey, B., & Khamparia. (2021). An intelligent hybrid approach for hepatitis disease diagnosis: Combining enhanced k-means clustering and improved ensemble learning. *Expert Syst, Vol 38*.

- Singh, S., & Awekar, A. (2013). Incremental shared nearest neighbor density-based clustering. *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management - CIKM '13*, 1533–1536.

- Vert, J. P., Tsuda, K., & Schölkopf, B. (2004). A primer on kernel methods. *Kernel methods in computational biology, Vol. 47*, 35–70.

- Vital, T. P., Kumar, K. D., Bhagya Sri, H. V., & Krishna, M. M. (2019). Analysis of Cancer Data Set with Statistical and Unsupervised Machine Learning Methods. Smart Intelligent Computing and Applications. *Smart Innovation, Systems and Technologies, Vol 104*, 267-276.

- Wang, W., Yang, J., & Muntz, R. (1997). STING: A Statistical Information Grid Approach to Spatial DataMining. *In: Proc. 23rd Int. Conf. on Very Large Data Bases, Athens, Greece, Publisher: Morgan Kaufmann*, 1997, 186-195.

- Wilson, S. J. (2017). Data representation for time series data mining: time domain approaches. *Wiley Interdisciplinary Reviews: Computational Statistics. Vol 9*(1).

- Workman, T. E., Hirezi, M., Trujillo-Rivera, E., Patel, A. K, Heneghan, J. A., Bost, J. E., Zeng-Treitler, Q., & Pollack, M. (2018). A Novel Deep Learning Pipeline to Analyze Temporal Clinical Data. *2018 IEEE International Conference on Big Data (Big Data). Seattle, WA, USA,* 2879-2883.

- Yang, H., & Wang, Y. (2007). A LBP-based Face Recognition Method with Hamming Distance Constraint. *In Proceedings of Fourth International Conference on Image and Graphics (ICIG 2007), Sichuan*, 645-649.

- Yang, J., Jin, Z., & Yang, J. (2015). *Nonlinear Techniques for Dimension Reduction.* In: Stan Z. Li., Anil K. Jain. (eds) Encyclopedia of Biometrics. Springer, Boston, MA.

- Zahn, C. T. (1971). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers, Vol C-20*(1), 68–86.

- Zhang, M. (2019). Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm to Identify Galaxy Cluster Members. *IOP Conf. Ser.: Earth Environ. Sci. Vol 252*(04).

- Zhang, X., Chou, J., , Liang, J., Xiao, C., Zhao, Y., Sarva, H., Henchcliffe, C., & Wang, F. (2019). Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. *Scientific Reports, Vol 9*(797).

CHAPITRE 4

Problématiques !

Sommaire

4.1 Introduction 114
4.2 Problème de perte de données et de l'information 115
4.3 Problème de choix de série des traitements 120
4.4 Conclusion 121

4.1 Introduction

Le dossier EHR de la médecine personnalisée est une source médicale importante de données. Par le marquage et le stockage électronique de données, il résume toute la vie du patient que ce soit médicale, démographique ou autres. La variété de données au sein d'un EHR considérée comme un point de richesse de conception et d'implémentation. D'un autre côté, la variété de données archivées elle-même pose un défi pour toutes les traitements basés sur l'EHR. Cette source hétérogène peut en effet comporter de données de la vidéo, audio, texte, tableaux, image, etc. Globalement, sont des données structurées, semi-structurées ou non structurées. Cependant, la détermination de la forme de données à traiter se repose sur l'avis du spécialiste et de l'utilisateur de la source de données sans oublier les nécessités autour des objectifs visés.

Le but global de cette thèse est d'explorer les données de la médecine personnalisée par les outils du data mining. Dans le cadre général, cette proposition est plus large dans le sens où il y a une vaste hétérogénéité de données. Cette exploration peut se modéliser en tant qu'une approche du traitement d'images, de textes, de tableurs ou une combinaison entres eux. Aux vues des approches d'extraction des caractéristiques des images et les approches de transformation des textes et leurs résultats qui peuvent être structurés, nous avons décidé de nous orienter vers l'exploration de données structurées. Par conséquent, les données considérées dans nos travaux sont toutes des observations structurées (Tableurs) comportant les types de données numérique, nominale, date et binaire.

Appliquer un processus d'exploration de données sur les quatre types de données cités ci-dessus simultanément implique plus de complications à cause de défis de l'hétérogénéité du type et la pénurie des classifieurs adéquats. A cet effet, penser à réduire le nombre des types à traiter peut lever certaines difficultés. Généralement, la suppression de données d'un type donné peut nuire à la qualité des modèles développés. Cependant, une opération de transformation de données peut révéler être une solution suffisante. Unifier le type de données par telles opérations de

transformation et la combinaison des résultats ne forme qu'un processus de représentation global.

En pratique, la production d'une représentation unifiée de données est l'une des plateformes initiales des outils du data mining pour l'exploration. L'automatisation du processus de la prise d'une décision médicale est l'un des axes importants des approches d'exploration. Cependant, les techniques composites de ce processus peuvent être le facteur qui fortifie ou qui banalise la décision médicale.

Dans l'ensemble, l'étude de l'état de l'art de l'exploration de données structurelles nous a permis de mettre en avant deux problèmes fondamentaux : la perte de données et d'informations lors de la production de la représentation de données et le choix de la meilleure série des traitements à adopter lors de la classification.

4.2 Problème de perte de données et de l'information

Plusieurs travaux autour de la représentation de données peuvent être cités. Deep Integrated Prediction (DIP) (Nezhada et al., 2018) est l'un des exemples des projets travaillants sur les méthodes d'exploration de données émergentes. Ce projet utilise l'apprentissage en profondeur pour représenter les attributs de type numérique et l'algorithme GloVe pour discriminer les attributs nominaux dans l'EHR des patients souffrants de la maladie cardiaque. Graph-based Attention Model (GRAM) est une approche qui utilise le marquage de chaque événement de visite médicale sur un graphe acyclique dirigé (Choi et al., 2017). Elle utilise les données de l'EHR et les ontologies médicales pour représenter ces événements sous une forme de matrice du poids. Ensuite, elle prédit les valeurs des nouvelles entrées par l'application des modèles de réseau de neurone en fonction de la représentation matricielle produite. Dans un autre travail, les auteurs Mallick et al. (2018) présentent une approche pour étudier l'interdépendance des gènes dans les cas de Cancer. Ces chercheurs

appliquent la logique floue pour calculer le gain d'information pour représenter les données des gènes vis-à-vis d'un graphe.

Dans la plupart des cas, l'EHR comporte une variété considérable de données qui inclut les données structurelles. Sans aucun doute, cette dernière peut contenir des données numériques chronologiques qui forment des séries de données temporelles. A titre d'exemple, l'approche de Bagattini et al. (2019) travaille sur la détection des effets indésirables des médicaments. Son objectif constitue l'un des axes sur laquelle intervient la MP. L'approche proposée est composée de trois phases : (i) la représentation symbolique des données, (ii) la génération de sous-séquences et (iii) la classification. Durant les traitements, leur processus ne considère que les données numériques et ignore toutes les autres données qui peuvent être portées sur l'EHR. Cependant, la phase de la représentation applique la technique Piecewise Aggregation Approximation (SAX) afin de produire des séquences de représentation symbolique pour toutes les séries temporelles. Indépendamment des limites de la technique SAX et des qualités des données traitées (les types de données), cette approche fournit un exemple opportun sur l'utilisation des données MP et met en avant le besoin de nouvelles techniques de représentation des données de l'EHR. Dans l'approche de Singh et al. (2015), une représentation par des fenêtres de marquage sur les événements médicaux a été proposée. La durée de ces fenêtres est fixée en fonction du facteur temps du diagnostic («time-Windows»). Le délai était initialement fixé à six mois. En pratique, cette approche provoque une éventuelle perte d'informations concernant les changements de comportement des événements durant cette durée. Des années plus tard, Zhao et al. (2017) proposent une approche différente pour le traitement de la temporalité des données tel qu'il est appliqué aux événements cliniques et à l'extraction de la représentation. Pour le processus de transformation, ils ont utilisé la méthode d'approximation symbolique agrégée (SAX) sur toutes les observations pour générer une représentation sous forme de chaînes de caractères. Par conséquent, c'est un processus qui peut également entraîner la perte d'informations durant les traitements.

Beaucoup d'approches utilisent la normalisation de données numériques comme un outil d'aide à l'unification des traitements et afin de rendre les résultats comparables. Ce processus génère probablement une perte d'information. Par conséquent, la qualité de la représentation de données finale porte les conséquences de la normalisation, y compris la perte d'informations. De plus, certains paramètres lors de la représentation peuvent constituer une autre source de perte de l'information et de la donnée elle-même.

Comme un exemple démonstratif, lors de l'application de la technique SAX, l'approche de Zhao et al. (2017) passe par la normalisation des valeurs des séries temporelles. Nous pouvons constater deux problèmes avec cette normalisation. Pour faciliter la description, nous utilisons deux séries temporelles qui mesurent la température réelle de deux patients : X1= {34.6, 34.6, 34.6, 34.6, 34.6, 34.0, 34.6, 34.6, 34.6, 34.6} et X2= {36.6, 35.7, 35.5, 35.9, 36.1, 36.1, 35, 8, 36.8, 36.7, 36.0}. Dans un traitement par la technique SAX, X1 et X2 seront considérés pour générer deux chaînes de neuf caractères à la base de trois symboles de représentation.

Le premier problème apparaît lors de l'application de la méthode SAX sur X1 et X2 d'une façon séparée. En effet, la représentation générée est la même pour les données normalisées et non normalisées. Cependant, le même symbole est utilisé dans les deux représentations qui peuvent correspondre à deux intervalles différents. Par exemple, sans normalisation, la technique SAX génère un symbole 'a' ∈]-∞; 34.462] pour X1 et 'a' ∈]-∞; 35.939] pour X2 comme le montre la Figure 4.1. Avec la normalisation, il génère 'a' ∈]-∞; 0.771] pour X1 et 'a' ∈]-∞; 0.338] pour X2 comme le montre la Figure 4.2. Les autres symboles ont le même comportement que le premier symbole «a».

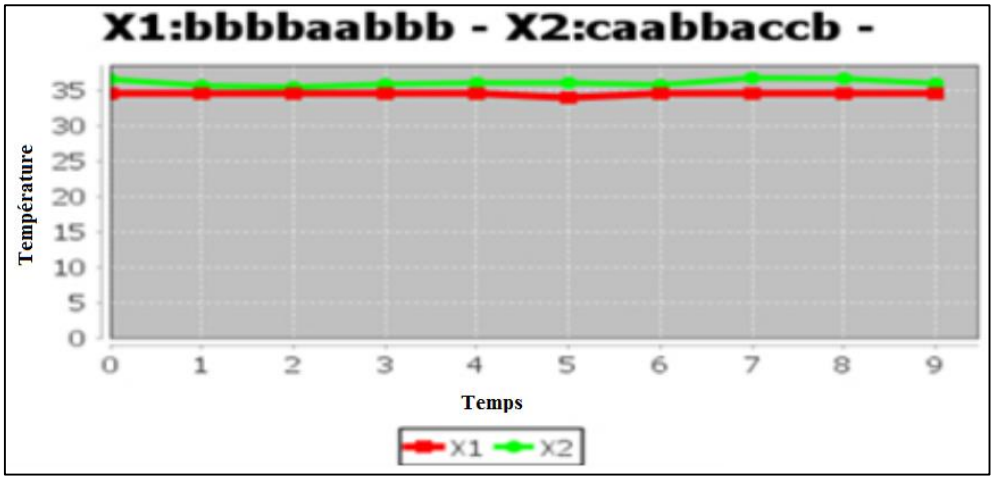

Figure 4.1. Représentation SAX de X1 et X2 sans normalisation.

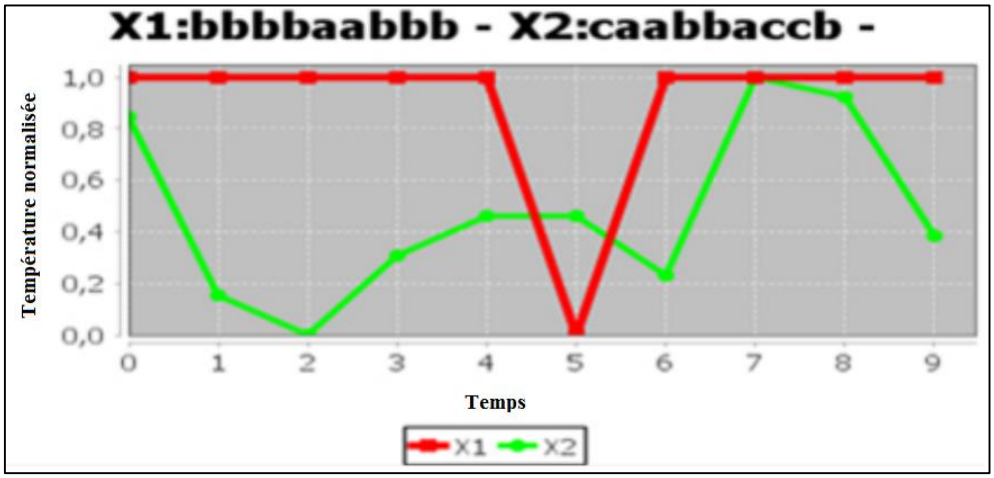

Figure 4.2. Représentation SAX de X1 et X2 avec normalisation.

Le deuxième problème est observé après l'application de la méthode SAX multi-séries, c'est-à-dire un traitement sur toutes les séries temporelles. En effet, SAX multi-séries se base sur l'utilisation des paramètres communs tels que la moyenne, la variance et l'écart-type. Les représentations générées ne sont pas identiques pour les séries normalisées et non normalisées, mais l'observation comportementale de ces séries montre qu'il y a une perte de sens après le processus de normalisation. Sur la Figure 4.3 qui montre le cas de la représentation sans normalisation, la courbe qui présente la série X1 est apparue totalement sous la courbe de X2, dans laquelle toutes

les valeurs de données de X1 sont inférieures aux valeurs de X2. Cela montre apparemment un comportement significatif avec l'ensemble de données non normalisées. Comparativement avec la Figure 4.4 qui présente le cas normalisé, la plupart des parties de la courbe X2, mais pas toutes, sont présentées sous les parties de X1, ce qui implique une perte de la signification des données et les niveaux capturés par rapport au comportement original des deux séries.

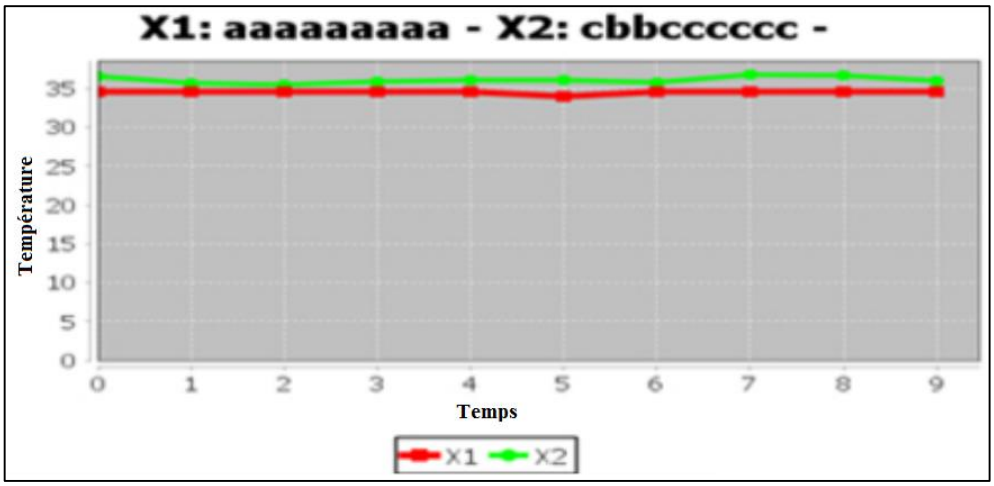

Figure 4.3. Représentation Multi-séries SAX de X1 et X2 sans normalisation.

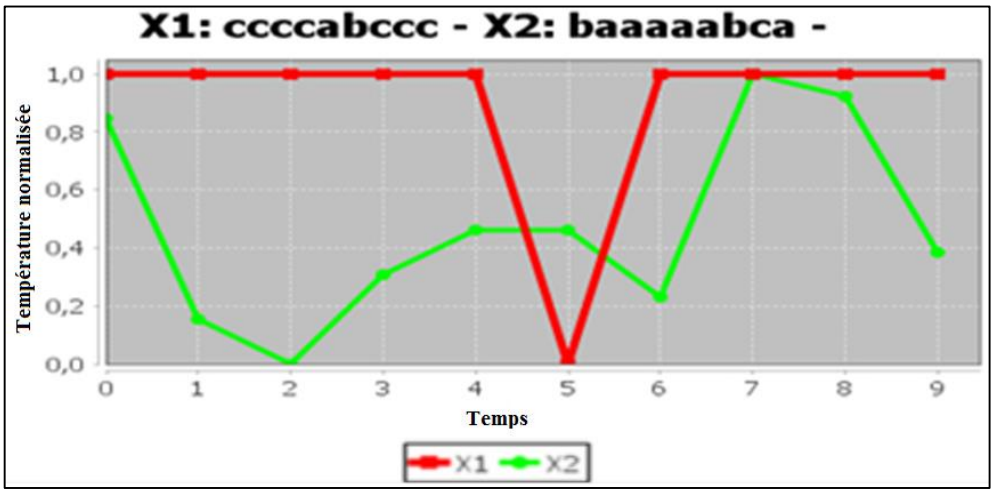

Figure 4.4. Représentation Multi-séries SAX de X1 et X2 avec normalisation.

En plus de la perte d'information lors de la normalisation, la technique SAX nécessite un paramètre d'entrée pour spécifier la longueur de la série de symboles résultante. Ce paramètre doit être inférieur ou égal à la longueur de la série à représenter; sinon, toutes les séries qui ont une longueur minimale seront perdues. L'impact de ce critère peut conduire à la perte de toutes les données si les données ne contiennent que de courtes séries.

4.3 Problème de choix de série des traitements

Les travaux sur l'exploration de données médicales montrent l'utilité des algorithmes de classification pour le domaine médical globalement et pour l'automatisation de la prise de la décision médicale particulièrement. Par exemple, Amin et al. (2019) ont appliqué sept algorithmes de classification pour la prédiction des maladies cardiaques : k-plus proche voisin (k-NN), arbre de décision (DT), Naïve Bayes (NB), régression logistique, SVM, réseau neuronal et vote (technique hybride avec NB et régression logistique). D'autres approches (Ayatollahi et al., (2019); Gultepe et Rashed, (2019)) prédisent les maladies coronariennes et cardiaques par les techniques de classification suivantes: ANN multicouche, SVM, NB et arbres de décision (C4.5). Dans Emre et al. (2019) et Vital et al. (2019), six techniques de classification sont considérées. La première approche (Emre et al., 2019) applique le classifieur NB, les arbres de décision (Classification and Regression Tree (CART), C4.5, C5.0, C5.0 boosté) et les algorithmes de forêt aléatoire (RF) pour analyser l'effet de la fièvre rhumatismale sur les maladies cardiaques dans l'enfance. La deuxième approche (Vital et al., 2019) applique l'arbre de décision alterné, DT (C4.5), NB, BayesNet, K-Star et RF pour prédire la maladie cancéreuse et analyser les performances de l'ensemble de données.

Les comportements de la sélection des classifieurs appliqués pour ces quatre derniers travaux varient selon les approches soit d'une façon injustifiée, soit du plus populaires et au plus recommandé. Ce constat n'exprime aucune forme de choix déterminé, et tout professionnel peut adopter n'importe quelle mesure subjective. De plus, le nombre de classifieurs testés pour chaque approche peut varier entre deux et

sept, une estimation qui peut exprimer une exagération ou une insuffisance autour de l'évaluation des approches. Dans l'ensemble, il n'existe pas une méthode déterministe à appliquer pour une sélection optimale des algorithmes et les méthodes à suivre. Implicitement, Quel classifieur dois-je utiliser?, Combien de classifieurs dois-je tester? Ce sont les questions qui se posent avec la plupart des approches de classification.

Sur la même échelle d'importance et de même genre, d'autres questions peuvent se poser sur le choix des techniques de réduction de données, de transformation, de clustering et de distances adoptées entres les instances à représenter et à classifier. En résumé, un point d'interrogation se pose toujours autour de séries de traitement qui doivent êtres appliquées ou testées.

4.4 Conclusion

Ce chapitre a expliqué certains caractères de l'EHR afin d'identifier certaines difficultés rencontrées et nos choix concernant la forme structurelle de données et les types considérés. De plus, il explique les deux problématiques principales lors de la représentation de données dans le contexte encadré par nos choix. Nous avons expliqué brièvement l'obstacle de la perte de données et de l'information, et ainsi abordé le problème du choix de la meilleure série de traitements. Cette explication ne constitue qu'une simple introduction pour signaler les objectifs fixés et les défis qui doivent êtres passer. Par conséquent, dans le reste de cette thèse, nous essayons de présenter nos travaux face au développement d'un modèle pour la représentation de données de la MP, et de la composition des techniques du data mining pour produire un modèle de prise de décision médicale en se basant sur notre représentation de données produite.

Références

- Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics. Vol 36, 82-93.

- Ayatollahi, H., Gholamhosseini, L., & Salehi, M. (2019). Predicting coronary artery disease: a comparison between two data mining algorithms. *BMC Public Health, Vol 19*.

- Bagattini, F., Karlsson, I., Rebane, J.& Papapetrou, P. (2019). A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Med Inform Decis Mak, Vol 19*(7).

- Choi, E., Bahadori, M. T., Song, L., Stewarty, W. F., & Sun, J. (2017). GRAM: graph-based attention model for healthcare representation learning. *International Conference on Knowledge Discovery and Data Mining, Vol 23*, 787–795.

- Emre, I. E., Erol, N, Ayhan, Y. I., Ozkand, Y., & Erole, C. (2019). The analysis of the effects of acute rheumatic fever in childhood on cardiac disease with data mining. *International Journal of Medical Informatics, Vol 123*, 68-75.

- Gultepe, Y., & Rashed, S. (2019). The Use of Data Mining Techniques in Heart Disease Prediction. *International Journal of Computer Science and Mobile Computing, Vol 8*(4), 136-141.

- Mallick, P., Seth, P., & Ghosh, A. (2018). Entropy-based fuzzy hybrid framework for gene prediction network – an application to identify and rank the biomarkers for human lung adenocarcinoma. *International Journal of Computers and Applications, 41*(1), 62–77.

- Nezhada, M. Z., Zhub, D., Sadatia, N., & Yanga, K. (2018). A Predictive Approach Using Deep Feature Learning for Electronic Medical Records: A Comparative Study. *Cs.LG*, arXiv:1801.02961v1.

- Singh, A., Nadkarni, G., Gottesman, O., Ellis, S. B., Bottinger, E. P., & Guttag, J. V. (2015). Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of Biomedical Informatics, Vol 53*, 220–228.

- Vital, T. P., Krishna, M. M., Narayana, G. V. L., Suneel, P., & Ramarao, P. (2019). Empirical Analysis on Cancer Dataset with Machine Learning Algorithms. Soft Computing in Data Analytics. *Advances in Intelligent Systems and Computing, Vol 758*, 789-801.

- Zhao, J., Papapetrou, P., Asker, L., & Bostrom, H. (2017). Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics, Vol 65*, 105–119.

CHAPITRE 5

Représentation de Données de la MP par Région et Dispersion (DRRD).

Sommaire

5.1 Introduction 126
5.2 Description générale 126
5.3 Modèle proposé 127
5.3.1 Formulation du problème 128
5.3.2 Schéma détaillé 128
5.3.3 Réorganisation des données du type numérique et date (Etape A_1) 129
5.3.4 Partitionnement des événements numériques (Etape B_1) 131
5.3.5 Marquage des données numériques (Etape C_1) 134
5.3.6 Linéarisation des données numériques (Etape D_1) 136
5.3.7 Représentation des données du type nominal et booléen (Etape A_2) 137
5.3.8 Dispersion « Diffusion » des événements nominaux (Etape B_2) 138
5.3.9 Marquage des données nominales (Etape C_2) 139
5.3.10 Linéarisation des données nominales (Etape D_2) 140
5.3.11 Assemblage des résultats 140
5.4 Expérimentation 141
5.4.1 Description du dataset 142
5.4.2 Sélection, transformation et codification de données 142
5.4.3 Résultats et discussion 143
5.4.4 Exemple de représentation et évaluation 150
5.4.5 Exemple de vue par patient 152
5.5 Conclusion 153

5.1 Introduction

Nous avons présenté, dans le 4eme chapitre, les axes principaux de nos recherches et nous avons décrit aussi le cadre général autour des problèmes rencontrés. Ce cadre place l'exploration de données structurelles de la MP comme le champ primordial de ce travail. Les données de type numérique, nominal, date et binaire fait partie des données structurées et sont indiquées pour préciser la vision adoptée vers les types de données à traiter. L'aspect chronologique de données numériques qui forment des séries temporelles est bien éclairé pour simuler la configuration des événements médicaux capturés. Nous avons indiqué les deux défis majeurs nécessitants des solutions. Particulièrement, le problème de la perte de données et d'information a été montré par des exemples démonstratifs. En particulier, a comparé deux cas de représentation simple et multi-séries SAX en fonction des traitements normalisés et non-normalisés. Le premier cas a montré une différence entre les valeurs estimées pour le même symbole de représentation. Tandis que, le deuxième cas a montré une perte de comportement de données après la normalisation vis-à-vis les informations portées sur les séries.

Ce chapitre va présenter notre approche développée comme solution destinée au problème de perte de données et d'information, et globalement de la représentation de données structurelles de la MP. Pour ce faire, les parties ci-dessous vont détailler clairement cette proposition et le modèle produit.

5.2 Description générale

L'idée de la proposition a débuté par une recherche destinée à la présentation de données structurelles du type numérique formant des séries temporelles. Avec le temps, les données de type date ont également été prises en compte en passant par la transformation en données de type numérique. Le modèle proposé a été présenté durant la conférence **CITIM'2018** (Kadi et al., 2018). Face à la richesse de la MP en matière de données et le défi de la perte des types de données entières, sans ignorer

l'ambition de traiter le maximum possible de toutes les données structurées, notre modèle a subi plus de développement. Généralement, le modèle de la représentation de données numérique a été le noyau de développement des autres extensions. Finalement, la dernière version du modèle de base nous a permis de traiter les quatre types de données structurelles numérique, date, nominal et binaire. Cette dernière extension a été l'objet de la proposition du papier intitulé « **A Data Representation Model for Personalized Medicine** ». Ce manuscrite a été publié au journal ***International Journal of Healthcare Information Systems and Informatics (IJHISI)***. (Kadi et al., 2021a).

5.3 Modèle proposé

Le modèle proposé peut représenter les données structurées, temporelles et / ou non temporelles et leurs différents types, y compris numérique, nominal, date et binaire. Les observations temporelles existent sous une forme tridimensionnelle (3-D: Patients, Evénements, Temps). Afin d'unifier le processus de traitement des données non temporelles qui se trouvent réellement en deux dimensions (2D: Patient, Data), nous les considérons sous une forme à trois axes (Patient, Data, Time-1) de telle sorte que "Time-1" prend une seule valeur, c'est-à-dire "1". Un tel modèle conserve autant que possible les données, même pour les séries temporelles courtes. Un processus de clustering pour les données numériques et un autre de dispersion pour les données nominales seront appliqués. En conséquence, ce travail aboutira à une représentation simplifiée avec seulement deux dimensions, et facile à l'explorer. Avec de tels résultats, il est clair que l'attention a maintenant été accordée aux défis qui nous avons mentionnés, y compris l'hétérogénéité des types de données, la couverture maximale des ressources de données pendant le traitement, la minimisation de la perte d'information et de données pendant le processus de transformation.

5.3.1 Formulation du problème

Pour formaliser notre modèle, nous utilisons D pour noter l'ensemble des dossiers de données EHR qui comprend un ensemble de patients $P = \{P_1, P_2, ..., P_n\}$ et un ensemble d'événements $E = \{E_1, E_2, ..., E_m\}$. La chronologie des observations capturées pour chaque événement E_i sera présentée selon le temps $T = \{T_1, T_2, ..., T_q\}$.

Soit e_{ijr} la valeur d'observation de l'événement E_i pour le patient P_j avec la chronologie T_r. Nous utilisons ensuite $E_i.T$ pour présenter la chronologie de la série la plus longue de événement E_i. Soit NuE l'ensemble d'événements numériques et NoE l'ensemble d'événements nominaux tel que :

$$\forall E_i \subset NuE \wedge \forall e_{ijr} \in E_i \Rightarrow e_{ijr} \in \mathbb{R} \quad (5.1)$$

$$\forall E_i \subset NoE \wedge \forall e_{ijr} \in E_i \Rightarrow e_{ijr} \in \text{String Type} \quad (5.2)$$

L'équation 5.1 signifie que toutes les observations d'événements numériques sont des valeurs de type réel, tandis que l'équation 5.2 signifie que toutes les observations d'événements nominaux sont des valeurs de type chaîne de caractères.

Nous utilisons IW_{ki} pour l'inertie intraclasse (**Within-class Inertia *(IW)***) et IB_{ki} pour l'inertie interclasse (**Between-class Inertia *(IB)***) par rapport à l'événement E_i et au nombre de clusters k_i, avec IT_{ki} (**Total Inertia (*IT*)**) leur inertie totale correspondante (Choukri et al., 2019).

5.3.2 Schéma détaillé

La Figure 5.1 schématise notre modèle de représentation de données par région et dispersion « ***Data Representation model per Region and Dispersion* (DRRD)** ». Par rapport aux activités effectuées, nous pouvons le diviser verticalement en deux parties principales. La première partie traite les données de type numérique et date tandis que la deuxième partie traite les données du type nominal et binaire. Les parties sont composées par deux suites de modules (étapes), y compris la représentation et/ou la

transformation des données, le marquage des événements et la linéarisation des résultats. L'indication des modules se fait par l'ordre alphabétique et le numéro de la partie déclenchée. Par exemple, A_1, B_1, C_1, D_1 (A_2, B_2, C_2, D_2, respectivement) sont les étapes de la première partie (Deuxième partie respectivement). A la suite de ces deux parties, le modèle comporte une autre partie qui rassemble leurs résultats dans une seule table, mais sous deux modules différents notés par E_1 et E_2.

Les traitements sur la partie concernée par les données numériques et dates se déroulent selon la suite modulaire suivante :

5.3.3 Réorganisation des données du type numérique et date (Etape A_1)

Principalement, ce module intervient pour la représentation de données de type numérique ou date. Sur le modèle, il montre les premières activités sur ce genre de données. Plus que la sélection de données numérique, il comporte un processus de transformation du type de données date en numérique.

Pour tout patient P_j, la transformation des valeurs e_{bj1} de l'événement date de naissance E_b est effectuée en calculant l'âge du patient. Les autres valeurs du type date e_{ijr} sont transformées par rapport à la date de naissance e_{bj1}, en calculant l'année d'apparition de l'observation selon la formule suivante :

$$DateToNumeric(e_{ijr}/e_{bj1}) = NbYears(e_{ijr} - e_{bj1}) \qquad (5.3)$$

dont la fonction $NbYears(e_{ijr} - e_{bj1})$ calcule la différence entre e_{ijr} et e_{bj1} en termes de nombre d'années.

Toutes les données simples ou temporelles sont présentées comme des événements temporels numériques, cette tâche devant être présentée comme une tâche de représentation de séries temporelles. En vue de la chronologie des événements, les résultats seront présentés sous une forme de réorganisation tridimensionnelle (P, NuE, T).

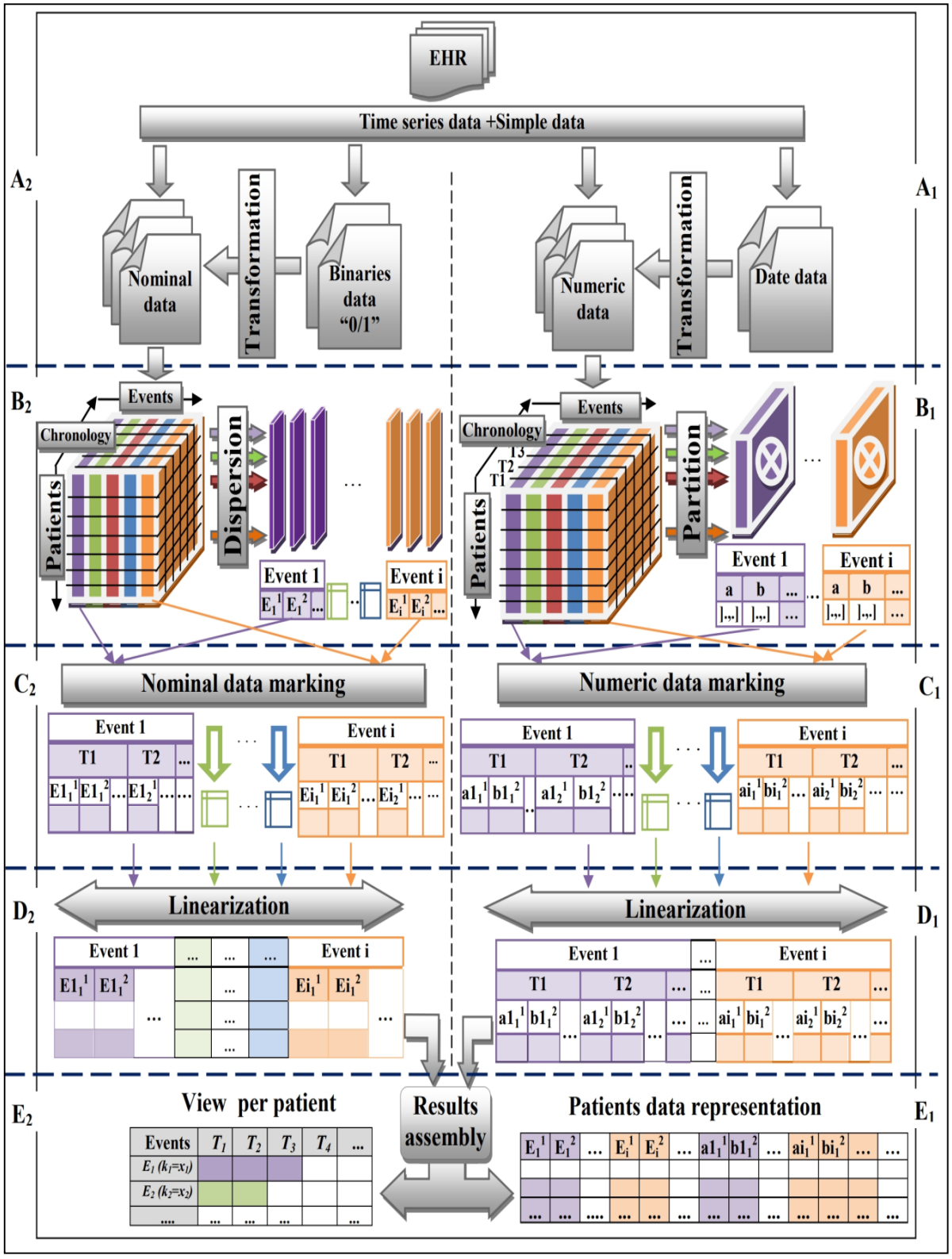

Figure 5.1. Modèle de représentation de données de la MP.

5.3.4 Partitionnement des événements numériques (Etape B_1)

Le clustering au sein de ce module vise à former des groupes similaires (clusters ou classes) basés sur les données ordonnées du même événement. Ces clusters seront les régions d'appartenance et les unités de comparaison de données de chaque patient sur cet événement.

Dans l'ensemble, pour chaque événement E_i dans le cube de données réorganisé avec la négligence du facteur du temps, nous appliquons une technique de clustering pour produire k_i clusters (Avec k_i>1) et la liste ordonnée de leurs centres (C_1, C_2, ..., C_p, ..., C_u, ..., C_{ki}) en fonction de leurs valeurs comme indiqué dans l'équation 5.4.

$$\forall p < u \, \Lambda \, \forall u \leq k_i \Rightarrow C_p < C_u \qquad (5.4)$$

La Figure 5.2 est un exemple démonstratif qui schématise les traitements détaillés de cette étape pour un événement donné.

Les centres des clusters sont ordonnés pour leur utilisation appropriée avec les symboles d'un alphabet S à définir ultérieurement. Les traitements appliqués à chaque événement E_i comprennent séquentiellement la sélection de données, le tri, le clustering et enfin le tri des centres de cluster résultants. L'association des données pertinentes à la région du centre le plus proche informe sur la distribution des données des régions en fonction des centres ordonnés. Par distribution, nous entendons la manière de diffusion, d'organisation et d'affichage des données dans l'espace de présentation des données.

Pour répondre à la question relative au choix du meilleur algorithme du clustering à appliquer, nous allons tester une variété de techniques connues. Cependant, pour ne pas exagérer dans le nombre des techniques à tester, nous allons évaluer les quatre méthodes reconnues suivantes : PAA (Wilson, 2017), k-means (Hartigan et Wong, 1979), EM (Dempster et al., 1977) et MDBC basé sur la classification hiérarchique (Witten et Frank, 2005).

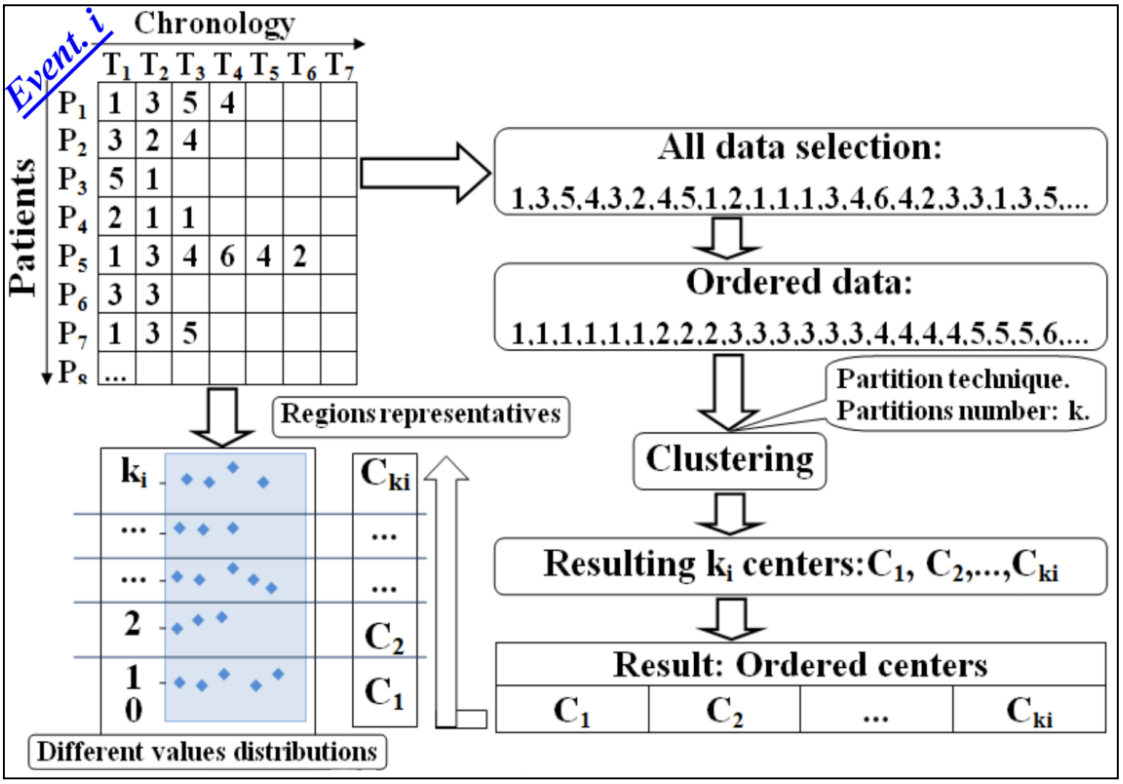

Figure 5.2. Processus de clustering d'un événement numérique E_i.

Chacune de ces techniques sera appliquée pour former k_i clusters sur chaque événement E_i. Plusieurs valeurs de k_i seront testées et le meilleur résultat sera considéré. Chaque événement E_i doit avoir une liste de différentes valeurs (L_i), avec une longueur $length(L_i)$ supérieure ou égale à k_i :

$$\forall e_{ijr}, e_{ils} \in L_i \Rightarrow e_{ijr} \neq e_{ils} \Lambda\, length(L_i) \geq k_i \qquad (5.5)$$

Une fois le clustering des événements effectué, une tâche de préparation de données au module suivant se déclenche. Correspondant à chaque événement E_i partitionné, elle consiste à créer une table de notification de données $MarkingTable_i$ qui possède le nombre des colonnes $ColumnsNumber(E_i)$, tel que :

$$ColumnsNumber(E_i) = k_i * length(E_i.T) \qquad (5.6)$$

où, la fonction *length*($E_i.T$) signifie la taille de la série T la plus longue dans l'événement E_i.

La notification pour chaque événement E_i consiste à indiquer par X une cellule pour chaque valeur de chaque série sur la table de notification $MarkingTable_i$. Cette indication doit être pointée devant la cellule correspondant au centre de cluster le plus proche, dont :

$$\forall e_{ijr} \in E_i \,, \forall C_{NC} \in \{C_1, \dots, C_{k_i}\} \,/\, Distance(C_{NC}, e_{ijr}) = \underset{p=1,\dots,k_i}{\text{MIN}} (Distance(C_p, e_{ijr})) \Rightarrow$$
$$MarkingTable_i[j][((r) * k_i) + NC] = \text{X} \qquad (5.7)$$

Distance(C_{NC}, e_{ijr}) est la fonction de soustraction entre le centre C_{NC} et l'observation e_{ijr}.

A titre d'exemple, nous supposons le patient P_3 qui possède la série suivante : e_{230}=3, e_{231}=6, e_{232}=2 sur l'événement E_2. Pour ce dernier, nous supposons aussi k_i=3, les centres C_1=2, C_2=5, C_3=8, et la longueur de la série maximale est *length*($E_2.T$)=3.

A cet effet, les distances minimales seront comme suit (Table 5.1) :

$\boldsymbol{Distance(C_{NC}, e_{ijr}) = \underset{p=1,\dots,k_i}{\text{Min}} (Distance(C_p, e_{ijr}))}$	**C1=2**	**C2=5**	**C3=8**
e230=3	1		
e231=6		1	
e232=2	0		

Table 5.1. Exemple de distances minimales.

Il est clair que le centre C_1 est le plus proche de la valeur de e_{230}, et de e_{232}, et C_2 est le plus proche de e_{231}. Par conséquent, pour chaque chronologie T_1 , T_2 et T_3 correspondants aux e_{230}, e_{231}, e_{232} respectivement nous notifions les centres C_1, $C2$, C_1 par des X. La table de notification sera comme suit (Table 5.2) :

	E_2								
	T_1 (e230=3)			T_2 (e231=6)			T_3 (e232=2)		
	C_1=2	C_2=5	C_3=8	C_1=2	C_2=5	C_3=8	C_1=2	C_2=5	C_3=8
P_1									
P_2									
P_3	X				X		X		
.....									

Table 5.2. Exemple de table de notification.

5.3.5 Marquage des données numériques (Etape C_1)

Ce module comporte principalement un processus de marquage (voir Algorithme 5.1) qui remplit les cellules indiquées sur les tables de notification. Pour assurer plus de richesse autour des représentations générées, nous proposons trois (3) types de marquage.

ALGORITHM 5.1. Marquage des évènements numériques.

EventMarking(P, E_i, $Centers_i$, MarkingType){

$MarkingTable_i$=new Table[n][k_i* *Length(E_i. T)*];

if(*MarkingType* == *Symbol*)// Define an ordered alphabet *S*.

Create an alphabet *S* of ordered symbols $S_0,S_1,...,S_{ki}$;

for all P_j in *P* do{

int *h=0;*

for all T_r in E_i.*T* do{

NC=*NearestCenter*($Centers_i$, e_{ijr}); // Nearest center index

if(*MarkingType*= Real Value)// Marking per real value.

$MarkingTable_i$[j][(h* k_i)+NC]=e_{ijr};

if(*MarkingType*= Binary) // Binary marking.

$MarkingTable_i$[j][(h* k_i)+NC]= 1;

if(*MarkingType*= Symbol) // Marking per symbol.

MarkingTable$_i$[*j*][(*h** k_i)+*NC*]= S_{NC};

h++;

}

}

return MarkingTable$_i$;

}

Le premier type est le *marquage par valeur réelle* tel qu'il est illustré sur la Table 5.3. Il comprend la réécriture de la valeur de données dans la cellule indiquée (X = e_{ijr}).

E_i						
T_1			T_2			...
C_1	...	C_{ki}	C_1	...	C_{ki}	...
					3.4	...
3						
…						…

Table 5.3. Marquage par valeur réelle.

Le deuxième type, le *marquage binaire* (Table 5.4), est une forme de signalisation de présence. Il remplit par 1 les cellules notifiées (X = 1) et met un 0 pour les autres (X = 0).

E_i						
T_1			T_2			...
C_1	...	C_{ki}	C_1	...	C_{ki}	...
0		0	0		1	...
1		0	0		0	
…						…

Table 5.4. Marquage binaire.

Tandis que, le dernier type est le *marquage par symbole* (Table 5.5). Puisque les centres sont ordonnés, nous pouvons associer à chacun un symbole d'un alphabet prédéfini et ordonné $S = \{S_0, S_1, \ldots, S_{ki}\}$. Par la suite, le processus place le symbole qui porte le même ordre avec le centre utilisé lors de la notification (X = S[NC]).

E_i						
T_1			T_2			...
C_1	...	C_{ki}	C_1	...	C_{ki}	...
					c	...
a						
…						…

Table 5.5. Marquage par symbole.

5.3.6 Linéarisation des données numériques (Etape D_1)

Séquentiellement, le module D_1 est la dernière tâche de la première partie. Il consiste à réorganiser les tables de marquage en une seule table. Deux arrangements sont proposés :

- La première est la linéarisation des événements numériques (Table 5.6). Dans une table qui rassemble tous les événements, on arrange (noté par ∏) les tables de marquage les uns après les autres,

où,

$$EventLinearization = \prod_{i=1,..,m}(MarkingTable_i) \quad (5.8)$$

E_1			E_2			...	E_m		
T_1	T_2	...	T_1	T_2	...	...	T_1	T_2	...

Table 5.6. Linéarisation par événement.

- La seconde est la linéarisation chronologique (Table 5.7). Dans une seule table, elle joint (noté ∑) toutes les colonnes T_1 de tous les tables de marquage dans une première itération. La deuxième itération joint toutes les colonnes T_2 au premier résultat, ainsi de suite jusqu'au parcours de toutes les colonnes :

$$ChronologicalLinearization = \sum_{j=1,\dots,SizeOf(E_1.T)}^{i=1,\dots,m}(MarkingTable_{iT_j}) \qquad (5.9)$$

E_1	E_2	...	E_m	E_1	E_2	...	E_m	...
T_1	T_1	...	T_1	T_2	T_2	...	T_2	...

Table 5.7. Linéarisation par chronologie.

Par imitation de la première suite des traitements, la deuxième partie concerne les données nominales et booléennes et sera composée par la série des modules donnés dans les sections suivantes :

5.3.7 Représentation des données du type nominal et booléen (Etape A_2)

Ce module considère les données de type nominal et booléen. Il comporte la sélection des données nominales et un processus de transformation des données booléennes temporelles et non-temporelles en données symboliques.

Le processus de conversion consiste à remplacer toutes les valeurs du "0" par "F" et les valeurs du "1" par "Y" :

$$\begin{cases} BooleanToNominal(e_{ijr}) = \text{"F" if } (e_{ijr} = 0) \\ \qquad\qquad and \\ BooleanToNominal(e_{ijr}) = \text{"Y" if } (e_{ijr} = 1) \end{cases} \qquad (5.10)$$

Le résultat sera un ensemble de données contenant des données nominales temporelles et non temporelles. Dans cette nouvelle réorganisation nominale des événements et suivant leurs chronologies, les données sont représentées selon trois axes (*P*, *NoE*, *T*).

5.3.8 Dispersion « Diffusion » des événements nominaux (Etape B_2)

Le module de dispersion des événements nominaux sélectionne l'ensemble de différentes valeurs définies L_i de chaque événement E_i et crée une nouvelle table vide correspondante nommée $DispersionTable_i$.

Nous associons $L_i.length$ cellules pour chaque valeur de la plus longue série $E_i.T$ de cet événement E_i. Les nombres de colonnes dans la table $DispersionTable_i$ doivent être égaux :

$$\forall P_j \in P, DispersionTable_i[P_j].length = L_i.length * length(E_i.T) \quad (5.11)$$

où,

$DispersionTable_i[P_j].length$ est le nombre de colonnes dans la table $DispersionTable_i$ avec tous les patients P_j. L'algorithme 5.2 décrit cette étape :

Algorithm 5.2. Dispersion des événements nominaux.

EventDispersion(E_i){
Set L_i = *DistinctDataSelection*(E_i); //Distinct data values selection.
$DispersionTable_i$=new Table[n][$L_i.length$ * *length*($E_i.T$)];
Y=0;
For all z in (0,..., *length*($E_i.T$)-1) do {
    For each value v in L_i do { //Labeling columns.
    $DispersionTable_i.Column[(z* L_i.\ indexOf(v))+Y].Name = E_i.ID$+"_"+$z$+"_"+$v$;
        Y++;
        }
    }
return $DispersionTable_i$ and L_i;
}

5.3.9 Marquage des données nominales (Etape C_2)

La notification de ce module utilise le résultat de la tâche précédente. Nous utilisons la valeur de chronologie *r* de chaque observation e_{ijr} comme indice pour notifier toutes les données nominales sur les tables de dispersion :

$$\forall e_{ijr} \in E_i, DispersionTable_i[j][(r * L_i.length) + L_i.indexOf(e_{ijr})] = \text{X} \qquad (5.12)$$

La fonction $L_i.indexOf(e_{ijr})$ renvoie l'indice de l'observation e_{ijr} dans l'ensemble L_i de différentes valeurs de l'événement E_i.

Adéquatement au marquage de données numérique, les trois types de marquages *valeur réelle, binaire, par symbole* décrits ci-dessus sont réutilisés.

L'algorithme 5.3 décrit cette étape.

Algorithm 5.3. Marquage des événements nominaux.

EventMarking(*P*, E_i, L_i, $DispersionTable_i$) {
$MarkingTable_i$= $DispersionTable_i$;
if(*MarkingType*$_i$= *Symbol*) {// Define an ordered alphabet *S*.
    w = L_i.*length*-1;
    Create *S* of an ordered symbols S_0,S_1,...,S_w ;
}
for all P_j in *P* do {
    for all T_r in E_i.*T* do {
        SI= L_i.*indexOf*(e_{ijr});
        RI=(*r** L_i.*length*)+*SI*;
        if(*MarkingType*= *Real Value*) // Marking per real value.
            $MarkingTable_i$[*j*][*RI*]= e_{ijr};

```
            if(MarkingType= Binary)              // Binary marking.
                  MarkingTable_i[j][RI]= 1;
            if(MarkingType= Symbol)              // Marking per symbol.
                  MarkingTable_i[j][RI]= S_SI;
            }
      }
return MarkingTable_i;
}
```

5.3.10 Linéarisation des données nominales (Etape D_2)

Méthodologiquement, ce module utilise la même logique que la linéarisation de données numériques pour collecter les résultats du marquage des événements nominaux dans une seule table. Il y a toujours deux types de réarrangement, par événement nominal et par la chronologie des observations nominales. Par rapport aux étapes précédentes, la fin d'exécution du module actuel signale la terminaison de la suite applicative de la deuxième partie.

Une fois les deux parties verticales terminées, la troisième partie de l'assemblage des représentations résultantes peut démarrer.

5.3.11 Assemblage des résultats

Entouré par les finalités d'utilisation visées ci-dessous, nous proposons deux approches pour l'assemblage des résultats produits lors de l'exécution de deux parties verticales précédentes :

- La première approche (Etape E_1) produit une représentation globale pour tous les patients. Selon le type de marquage défini, elle consiste à regrouper les résultats de la linéarisation numérique et de la linéarisation nominale en une seule table. L'avantage de cette opération est qu'elle aboutit à une représentation par valeurs réelles, par symbole ou à une représentation binaire. Structurellement, ces résultats sont des tables à deux dimensions, ce qui met en évidence la simplicité de notre représentation finale préparée pour des finalités d'exploration.

- La deuxième approche (Etape E_2) produit une vue par patient (Table 5.8). Cette proposition vise à fournir une assistance aux professionnels de santé, en donnant une idée claire des variations des observations de chaque événement pour un patient donné. La vue par patient comprend l'association d'une table de données au patient P_j. Les lignes correspondent aux événements capturés vis-à-vis de la représentation du patient P_j. Pour chaque k_i cellules de chaque instant T_r, on ne prend que la cellule non nulle qui contienne le symbole de représentation. La table créée doit contenir q colonnes de sorte que T_q est la chronologie maximale de la série la plus longue $E_i.T$ pour l'événement E_i.

Evènements	T_1	T_2	T_3	T_4	...
E_1(On k_1 Clusters)					
E_2(On k_2 Clusters)					
E_3(On k_3 Clusters)					
... .	...	...	...		...

Table 5.8. Prototype de la vue par patient.

5.4 Expérimentation

Pour évaluer le modèle DRRD développé, nous l'appliquons à un ensemble de données EHR réel et nous comparons les représentations obtenues avec les résultats produits par la technique SAX.

5.4.1 Description du dataset

Nous utilisons les données gratuites du système "Open Medical Record System" (OpenMRS, 2018). La plate-forme OpenMRS dans OpenMRS Project, (2004) est une application pour les EHR personnalisés. Basé sur 2 528 concepts médicaux, ce dataset stocke 476 973 observations pour 5 284 patients. Deux (2) maladies sont observées dans cet ensemble de données «Human Immunodeficiency Virus (HIV)» et «Tuberculosis (TB)».

5.4.2 Sélection, transformation et codification de données

Pour nos expériences, nous n'utilisons que les observations sur la maladie TB. Seuls les concepts qui ont plus d'une observation sont sélectionnés. Nous avons obtenu 21 événements numériques, deux (2) événements nominaux et quatre (4) événements de type date nécessitant une opération de transformation. Cependant, cet ensemble de données n'inclut pas les observations de type nominal temporel et booléen.

Dans la première étape et après la transformation, les données ont été réorganisées sous une forme tridimensionnelle (3D), ce qui induit au repositionnement de vingt-cinq (25) événements numériques.

La Table 5.9 présente les statistiques sur les événements numériques ordonnés par leurs identifiants.

Pour la deuxième étape, il n'existe que deux observations nominales. Par conséquent, la forme tridimensionnelle (3D) des événements nominaux va comporter seulement les données des observations "GENDER" et "TRIBE. La Table 5.10 présente leurs statistiques.

N°	Concept ID / Nom du concept	Nombre de différente instances	Fréquence
1	21 / HEMOGLOBIN	82	282
2	654 / SERUM GLUTAMIC-PYRUVIC TRANSAMINASE	207	278
3	678 / WHITE BLOOD CELLS	81	282
4	729 / PLATELETS	204	282
5	730 / CD4%	52	476
6	790 / SERUM CREATININE	154	160
7	851 / MEAN CORPUSCULAR VOLUME	58	282
8	853 / CD8 COUNT	396	470
9	952 / ABSOLUTE LYMPHOCYTE COUNT	190	275
10	980 / BODY SURFACE AREA	3	3
11	1113 / TUBERCULOSIS DRUG TREATMENT START DATE	7	9
12	1279 / NUMBER OF WEEKS PREGNANT	19	38
13	5085 / SYSTOLIC BLOOD PRESSURE	31	2. 144
14	5086 / DIASTOLIC BLOOD PRESSURE	21	2. 143
15	5087 / PULSE	113	2. 269
16	5088 / TEMPERATURE (C)	64	2. 274
17	5089 / WEIGHT (KG)	216	2. 262
18	5090 / HEIGHT (CM)	79	135
19	5092 / BLOOD OXYGEN SATURATION	25	2. 267
20	5096 / RETURN VISIT DATE	2. 049	2. 265
21	5242 / RESPIRATORY RATE	15	105
22	5314 / HEAD CIRCUMFERENCE	31	122
23	5497 / CD4 COUNT	330	478
24	5599 / DATE OF CONFINEMENT	6	7
25	5919 / BIRTH YEAR	803	824

Table 5.9. Statistiques des événements numériques.

N°	Concept ID / Nom du concept	Nombre de différente instances	Fréquence
1	992843 / GENDER	2	824
2	992844 / TRIBE	3	824

Table 5.10. Statistiques des événements nominaux.

5.4.3 Résultats et discussion

Conformément à la contrainte de longueur minimale (Equation 5.5) vis-à-vis de la liste des différentes valeurs à traiter pour les événements numériques à partitionner, nous avons constaté qu'un seul cas généré par l'événement «980» pose problème. Cet événement possède trois (3) valeurs différentes uniquement, ce qui implique une

exécution du processus de clustering avec k = 2 et k = 3 seulement. Les autres cas de k = 4, 5 et 6 ne sont pas applicables.

Les colonnes sont nommées en fonction de l'identifiant d'événement, du numéro de cluster et du numéro d'observation dans la série. Par exemple, si on prend k = 2 pour partitionner l'événement «21 / HEMOGLOBIN», sachant que la série maximale de cet événement est observée sur le patient identifié par le numéro "3618" et cette série comporte trois (3) valeurs, alors la représentation sera générée sur six (6) colonnes "21_C0_S0, 21_C1_S0, 21_C0_S1, 21_C1_S1, 21_C0_S2, 21_C1_S2".

La Table 5.11 résume les statistiques des longueurs des séries maximales et le nombre de colonnes produites (c.-à-d. la longueur des tables de notifications vides à créer) lors du partitionnement de chaque événement.

N°	Concept ID	Longueur maximale de la série plus longue	Nombre de colonnes produites				
			k=2	k=3	k=4	k=5	k=6
1	21	3	6	9	12	15	18
2	654	2	4	6	8	10	12
3	678	3	6	9	12	15	18
4	729	3	6	9	12	15	18
5	730	3	6	9	12	15	18
6	790	2	4	6	8	10	12
7	851	3	6	9	12	15	18
8	853	3	6	9	12	15	18
9	952	3	6	9	12	15	18
10	980	1	2	3			
11	1113	3	6	9	12	15	18
12	1279	4	8	12	16	20	24
13	5085	10	20	30	40	50	60
14	5086	10	20	30	40	50	60
15	5087	10	20	30	40	50	60
16	5088	10	20	30	40	50	60
17	5089	10	20	30	40	50	60
18	5090	4	8	12	16	20	24
19	5092	10	20	30	40	50	60
20	5096	9	18	27	36	45	54
21	5242	4	8	12	16	20	24
22	5314	4	8	12	16	20	24
23	5497	3	6	9	12	15	18
24	5599	2	4	6	8	10	12
25	5919	1	2	3	4	5	6

Table 5.11. Statistiques des longueurs des séries maximales et du nombre de colonnes produites.

La dispersion des données n'affecte que les deux (2) concepts nominaux sélectionnés. Le premier "992843 / GENDER" génère une table de dispersion avec deux (2) colonnes "992843_0_M et 992843_0_F". Le deuxième concept "992844 / TRIBE" génère une table de dispersion avec trois (3) colonnes "992844_0_Luo, 992844_0_Luhya, 992844_0_Unknown".

La table 5.12 montre le marquage de données de l'événement numérique "790 / SERUM CREATININE" correspondant au patient identifié par "5126". Nous utilisons les trois (3) types de marquage et la technique k-means avec k=2. L'alphabet S = {a, b} des symboles de marquage n'a que deux symboles vis-à-vis du nombre de clusters k.

Type de marquage	790_C0_S0	790_C1_S0	790_C0_S1	790_C1_S1
Valeur réelle		27. 000,0	51,9	
Binaire		1	1	
Symbole		b	a	

Table 5.12. Exemple de marquage des événements numériques.

De plus, la table 5.13 montre le marquage de l'événement nominal "992843 / GENDER" pour le même patient.

Type de marquage	992843_0_F	992843_0_M
Valeur réelle	F	
Binaire	1	
Symbole	a	

Table 5.13. Exemple de marquage des événements nominaux.

Une meilleure partition possède l'inertie intraclasse la plus faible, et l'inertie interclasse la plus élevée (Yang et al., 2018). Par conséquent, afin de comparer les techniques utilisées, nous évaluons les résultats de notre représentation des événements numériques sur la base de l'inertie intra et interclasse. Les graphes radar illustrés sur la Figure 5.3.a et la Figure 5.3.b ont été générés en fonction des résultats

de l'inertie intraclasse et interclasse respectivement. Visuellement, ces graphes mettent en évidence les statistiques des techniques de clustering utilisées de PAA, k-means, EM et MDBC et illustrent la technique la plus appropriée vis-à-vis de l'événement considéré.

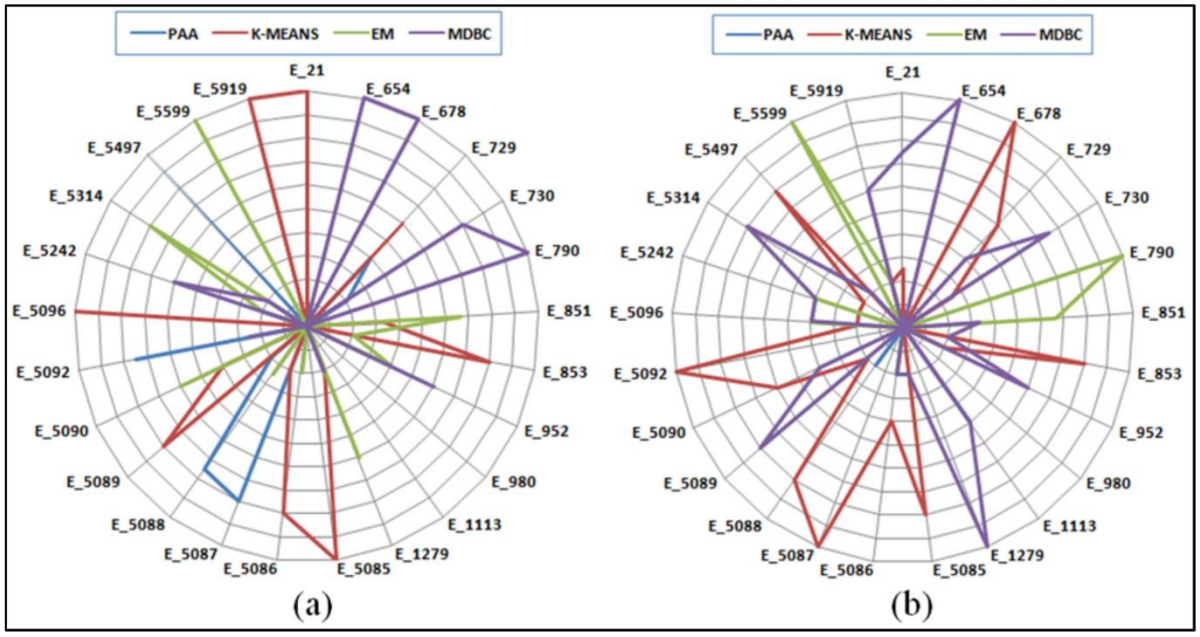

Figure 5.3. Illustration des techniques de clustering appropriées correspondant aux statistiques de : (a) l'inertie intraclasse, (b) l'inertie interclasse.

Pour chaque technique, nous associons un compteur de meilleurs cas. Ce compteur sera initialisé à zéro pour chaque événement. Par la suite, pour chaque événement et pour chaque k_i (k_i = 2, ..., k_i = 6), nous ajustons d'une façon incrémentale le compteur de la technique correspondant à la meilleure inertie. Si plusieurs techniques ont la même meilleure inertie dans une partition k_i donnée, on élimine ces résultats.

Dans l'ensemble, la table 5.14 résume la situation globale des techniques utilisées. Pour vingt-trois (23) cas observés sur vingt-cinq (25) événements dans l'évaluation d'inertie intraclasse, la technique k-means s'avère supérieure à toutes les autres approches, suivi par MDBC, EM, et finalement la technique PAA. Le même résultat est obtenu lors de l'évaluation de l'inertie interclasses, et pour vingt et un (21)

cas observés sur vingt-cinq (25) événements, la technique k-means reste toujours la meilleure.

Inertie	PAA	k-means	EM	MDBC
Intra class	4/23	8/23	5/23	6/23
Inter class	0/21	10/21	3/21	8/21

Table 5.14. Statistiques globales de l'inertie intra et interclasse.

Par apport à ses résultats, la technique k-means peut être utilisée sur tous les événements. Mais les qualités de chaque événement, notamment la répartition des valeurs, nous poussent à choisir la technique de clustering la plus adaptée à chaque cas. Pour choisir la technique et le nombre de clusters k_i pour chaque événement E_i, on calcule pour tous les k_i et toutes les techniques l'inertie totale IT_{ki} :

$$IT_{k_i} = IW_{k_i} + IB_{k_i} \tag{5.13}$$

La technique et le k_i correspondant à IT_{ki} maximum seront utilisés pour la représentation de l'événement E_i. La Table 5.15 présente les résultats de cette évaluation.

Pour les statistiques de dispersion de données nominales, aucune perte d'information ne s'est produite. La somme des instances dans chaque table de dispersion de l'événement "GENDER" (Fig 5.4.a) et de l'événement "TRIBE" (Fig 5.4.b) est égale au nombre total de patients (824).

N°	Concepts ID	Technique choisie	k_i choisi
1	21	MDBC	4
2	654	PAA	5
3	678	EM	2
4	729	MDBC	3
5	730	EM	2
6	790	PAA	2
7	851	PAA	2
8	853	MDBC	2
9	952	EM	2
10	980	PAA	2
11	1113	PAA	4
12	1279	MDBC	4
13	5085	MDBC	4
14	5086	MDBC	3
15	5087	MDBC	4
16	5088	MDBC	3
17	5089	MDBC	4
18	5090	MDBC	2
19	5092	k-means	6
20	5096	MDBC	5
21	5242	MDBC	2
22	5314	k-means	4
23	5497	MDBC	2
24	5599	PAA	3
25	5919	MDBC	2

Table 5.15. Techniques et nombre de clusters choisis.

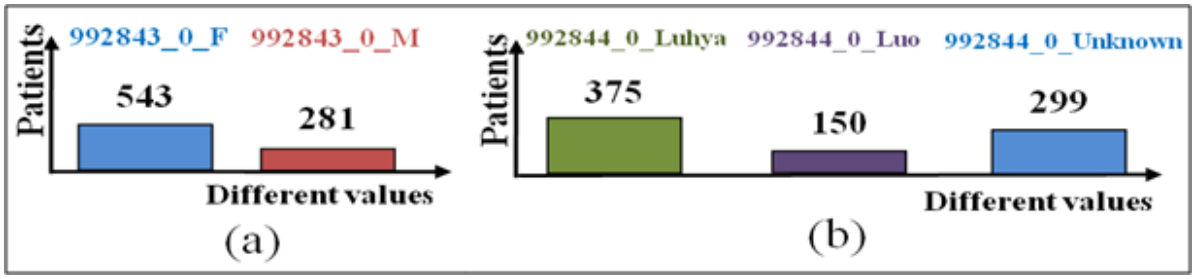

Figure 5.4. Dispersion des événements nominaux : (a) "GENDER", (b) "TRIBE".

Finalement, la représentation globale a été générée sur 431 colonnes. Parmi celles-ci, 426 colonnes sont associées à des événements numériques dont plusieurs d'entre eux sont composés par des séries temporelles. Les 5 colonnes restantes sont associées à des événements nominaux. Ces résultats prouvent la capacité des traitements des séries temporelles via notre modèle.

En vue de la représentation globale, les Figures 5.5, 5.6, 5.7 respectivement, représentent trois exemples de représentation globale générée selon les types de marquage par valeur réelle, binaire et par symbole respectivement.

E_5088_0	E_5088_1	E_5088_2	E_5088_3	E_5088_4	E_5088_5	E_5088_6	E_5088_7	E_5088_8
	36.40			36.30				
	35.80							
		39.50			38.40			39.70
	36.10			35.80				37.30
	36.00			36.10				36.80
	36.50				36.90			36.90

Figure 5.5. Partie d'une représentation par valeur réelle.

E_5088_0	E_5088_1	E_5088_2	E_5088_3	E_5088_4	E_5088_5	E_5088_6	E_5088_7	E_5088_8
	1			1				
	1							
		1			1			1
	1			1				1
	1			1				1
	1				1			1

Figure 5.6. Partie d'une représentation binaire.

E_5088_0	E_5088_1	E_5088_2	E_5088_3	E_5088_4	E_5088_5	E_5088_6	E_5088_7	E_5088_8
	b			b				
	b							
		c			c			c
	b			b				c
	b			b				c
	b				c			c

Figure 5.7. Partie d'une représentation par symbole.

Par rapport au type de données, les représentations globales binaire et par symbole n'utilisent qu'un seul type de données, ce qui fournit des représentations homogènes et élimine l'hétérogénéité dans les types de données initiaux des événements "Numérique, Nominal, Date, Booléen".

5.4.4 Exemple de représentation et évaluation

Pour évaluer notre approche, nous calculons trois (3) représentations symboliques de données du patient identifié par id = 75. Les résultats de la Table 5.15 ont été retenus pour cette proposition. La première représentation illustre notre approche DRRD dont chaque représentation d'événement est prise séparément des autres. La stratégie consiste à concaténer les symboles de représentation de chaque événement par rapport à l'ordre chronologique des observations correspondantes. La deuxième et la troisième représentation appliquent la technique SAX avec deux (2) valeurs différentes du paramètre de nombre de segments *W*. *W* indique la longueur de la représentation résultante, qui doit être inférieure ou égale à la longueur de la série en traitement. $W = 1$ est utilisé ici comme la longueur minimale acceptable par SAX dans la deuxième représentation nommée SAX1. Cependant, $W = 10$ est utilisé comme la longueur maximale correspondant à la série la plus longue pour la troisième représentation nommée SAX10. La technique SAX nécessite la précision du nombre de symbole d'entrée *Z* comparable au nombre de clusters. A cet effet, nous appliquons ensuite SAX1 et SAX10 sur chaque événement avec $Z_i = k_i$.

La Table 5.16 montre trois (3) représentations symboliques d'événements pour mettre en valeur les points essentiels de la comparaison.

Les cellules vides indiquent des événements non observés sur ce patient (par exemple, l'événement identifié par 790). Les cellules contenant "/" sont présentées uniquement pour la technique SAX, et sont informées par les conditions de longueurs minimales non respectées pour la série à présenter (par exemple, l'événement identifié par 21). Statistiquement, DRRD et SAX1 génèrent de nouvelles représentations pour dix-neuf (19) événements. De l'autre côté, SAX10 génère des nouvelles représentations pour huit (8) événements seulement. De plus, notre modèle représente chaque observation, quelle que soit la longueur de la série. En outre, la technique SAX ne génère des résultats de même longueur que si la série a passé la condition de longueur

minimale *W*. Plus généralement, la technique SAX élimine toutes les séries de longueur inférieure à *W*, ce qui constitue une perte d'information qui affecte directement la qualité de la représentation obtenue. Par exemple, la représentation de patients n'ayant que des séries temporelles de longueurs inférieures ou égales à un échouera directement avec la technique SAX et son paramètre $W \geq 2$.

Evénement Id	$Z=k_i$	DRRD	SAX1 (W=1)	SAX10 (W=10)
21	4	c	d	/
654	5	a	c	/
678	2	a	b	/
729	3	a	a	/
730	2	a	a	/
790	2			
851	2	a	a	/
853	2	a	b	/
952	2	a	b	/
980	2			
1113	4			
1279	4			
5085	4	dcdcccccdd	c	dbdaadcbdd
5086	3	cccccccccc	b	cacaabbacc
5087	4	cccccccccc	c	bcabdcbbbd
5088	3	bbbbbbbbbb	c	bccccccccc
5089	4	cccccccccc	c	cbccccccccc
5090	2			
5092	6	adbbbbbabb	a	adaaaaaaaa
5096	5	ccccccccc	e	/
5242	2			
5314	4			
5497	2	a	a	/
5599	3			
5919	2	b	b	/
992843	2	b	b	b
992844	3	a	a	a

Table 5.16. Représentation par symbole du patient identifié par id = 75.

Cet exemple démontre la capacité du nouveau modèle DRRD proposé à préserver les informations portées sur les séries temporelles et à conserver autant que possible les données tout au long du processus de représentation et de transformation.

5.4.5 Exemple de vue par patient

Cette solution consiste à présenter une fenêtre sur les données et ses variations pour chaque patient pris individuellement. La Figure 5.8 montre une instance de cette fenêtre pour le patient identifié par Id = 75. Cette visualisation des données peut aider les praticiens à analyser les observations. Dans cet exemple, le patient n'a qu'une seule apparence de l'événement "HEMOGLOBIN". Le caractère « c » code cet événement comme il correspond au troisième niveau parmi leurs 4 niveaux (4 est le nombre de clusters). Le praticien peut également connaître la durée maximale de la série temporelle de cet événement en fonction du nombre de cellules colorées. Pour l'événement "HEMOGLOBIN", la série temporelle comprend un maximum de trois échantillons. Selon ces informations, un praticien peut facilement décider si le patient a un fort besoin aux autres échantillons ou diagnostics. Après certaines expériences sur les patients, les praticiens peuvent alors constater les événements clés pour une maladie donnée. Par conséquent, ils peuvent examiner directement l'événement souhaité sur ce tableau de bord pour les cas suspects. Globalement, cette solution présente une vue détaillée sur les données d'un patient donné à base de l'ensemble des données de tous les patients.

N	Concept Name	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
1	HEMOGLOBIN (On 4 Clusters)	c									
2	SERUM GLUTAMIC-PYRUVIC TRANSAMINASE (On 5 Clusters)	a									
3	WHITE BLOOD CELLS (On 2 Clusters)	a									
4	PLATELETS (On 3 Clusters)	a									
5	CD4% (On 2 Clusters)	a									
6	SERUM CREATININE (On 2 Clusters)										
7	MEAN CORPUSCULAR VOLUME (On 2 Clusters)	a									
8	CD8 COUNT (On 2 Clusters)	a									
9	ABSOLUTE LYMPHOCYTE COUNT (On 2 Clusters)	a									
10	BODY SURFACE AREA (On 2 Clusters)										
11	TUBERCULOSIS DRUG TREATMENT START DATE (On 4 Clusters)										
12	NUMBER OF WEEKS PREGNANT (On 4 Clusters)										
13	SYSTOLIC BLOOD PRESSURE (On 4 Clusters)	d	c	d	c	c	c	c	c	d	d
14	DIASTOLIC BLOOD PRESSURE (On 3 Clusters)	c	c	c	c	c	c	c	c	c	c
15	PULSE (On 4 Clusters)	c	c	c	c	c	c	c	c	c	c
16	TEMPERATURE (C) (On 3 Clusters)	b	b	b	b	b	b	b	b	b	b
17	WEIGHT (KG) (On 4 Clusters)	c	c	c	c	c	c	c	c	c	c
18	HEIGHT (CM) (On 2 Clusters)										
19	BLOOD OXYGEN SATURATION (On 6 Clusters)	a	d	b	b	b	b	b	a	b	b
20	RETURN VISIT DATE (On 5 Clusters)	c	c	c	c	c	c	c	c	c	
21	RESPIRATORY RATE (On 2 Clusters)										
22	HEAD CIRCUMFERENCE (On 4 Clusters)										
23	CD4 COUNT (On 2 Clusters)	a									
24	DATE OF CONFINEMENT (On 3 Clusters)										
25	BIRTH YEAR (On 2 Clusters)	b									
26	GENDER (On 2 Clusters)	b									
27	TRIBE (On 3 Clusters)	a									

Figure 5.8. Exemple de vue par patient "Patient Id = 75".

5.5 Conclusion

Dans ce chapitre, nous avons présenté en détail notre modèle proposé DRRD pour la représentation de données de la médecine personnalisée spécialement. DRRD traite les données structurelles du type numérique, nominal, date et booléen pour assurer une couverture maximale des ressources de données. Au moyen des transformations et des traitements sur les données nominales et numériques, notre modèle dépasse les problèmes d'hétérogénéité des types de données. Les données nominales sont traitées par dispersion, et les données numériques sont traitées par l'appartenance régional (Clustering et les centres les plus proches). Une telle résolution a une forte contribution sur la représentation des séries temporelles. Ce type de données est traité comme une partie des données numériques. En résultat, une seule représentation globale est obtenue pour tous les patients. Cette représentation inclut le détail des données de base même si les données n'ont qu'une seule observation. Il conserve les

informations portées sur les données, en particulier pour les séries temporelles, car il représente chaque observation par rapport à sa région d'origine.

Sur l'échelle individuelle du patient, les résultats de la vue par patient montrent d'autres détails au niveau de chaque patient, notamment l'absence et la présence d'événements médicaux, le nombre de tests et de prélèvements pour le patient et leurs niveaux par rapport à un global échelle (nombre de clusters) de l'événement examiné. La comparaison de notre approche vis-à-vis de la technique SAX et de ses paramètres montre la grande efficacité et la capacité de notre modèle proposé. En effet, il peut minimiser les pertes d'informations et de données de séries temporelles (par rapport à SAX10, où le nombre de segments *W* est grand) pendant la transformation. De plus, il fournit une forte description de données des patients (par rapport à SAX1, où *W* est petit). En bref, notre approche a la capacité de traiter plusieurs types de données en même temps alors que la technique SAX ne traite que les données numériques.

Notre travail simplifie la compréhension des données de la MP par les praticiens de la santé. Il fournit également un outil pour l'analyse et la comparaison des données entre les patients, et facilite le suivi des événements clés et des observations importantes par les praticiens. Le diagnostic et la planification du traitement peuvent bénéficier d'un tel outil. En particulier, nous pensons que la représentation que nous proposons facilite l'exploration des données EHR et peut encourager les praticiens à se tourner vers la MP ainsi qu'à utiliser des systèmes d'aide à la décision automatique qui explorent les données EHR.

L'autre point essentiel de notre modèle est la proposition de trois types de marquage par valeur réelle, binaire et par symbole, et la production d'une représentation simple, homogène et unifiée sous forme d'une table à deux dimensions pour chaque type de marquage. L'achèvement de cette solution ouvrira notre voie vers d'autres travaux futurs, tels que la prise de décision médicale, la classification et l'exploration de données. La richesse de cette représentation en termes de types de données, et la richesse des outils de data mining en termes de disponibilité des algorithmes de classification et de leurs capacités pour traiter différents types de données, nous donnent la possibilité de faire d'autres travaux qui prennent cette

représentation comme une base. Le test de notre modèle de représentation sur des domaines différents que la médecine et qui possédants des ressources de données similaires, est un autre objectif applicatif future.

Plus que les défis du chapitre de la problématique, les futurs objectifs ciblés et cités ci-dessus, en particulier l'exploration de données de la MP par des techniques de data mining et la production d'un modèle de prise de décision médicale, feront partie de l'approche applicative qui sera développée et expliqué dans le chapitre suivant.

Références

- Choukri, A., Hamzaoui, Y, Amnai, M., & Fakhri, Y. (2019). Classification Algorithm Based on Nodes Similarity for MANETs. *International Journal of Online and Biomedical Engineering, Vol 15*(05), 86-100.

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*.

- Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society, Vol 28*(1), 100–108.

- Kadi, H., Rebbah, M., & Meftah, B. (2018). A data presentation model for personalized medicine. International Conference on Multimedia Information Processing, CITIM'2018. Mascara, Algeria.

- Kadi, H., Rebbah, M., Meftah, B., & Lezoray, O. (2021a). A data presentation model for personalized medicine. *International Journal of Healthcare Information Systems and Informatics. Vol 16*(4), (in press).

- OpenMRS Project. (2004). (Accessed 2018). Retrieved from Online Web site: https://openmrs.org.

- OpenMRS Wiki. (2018). (Accessed 2018). Demo Data [Data file]. Retrieved from https://wiki.openmrs.org/.

- Wilson, S. J. (2017). Data representation for time series data mining: time domain approaches. *Wiley Interdisciplinary Reviews: Computational Statistics, Vol 9*(1).

- Witten, H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, SF, USA.

- Yang, C., Ye, Y., Li, X., & Wang, R. (2018). Development of a neuro-feedback game based on motor imagery EEG. *Multimedia Tools and Applications, Vol 77*(12), 15929–15949.

CHAPITRE 6

Prise de décision médicale basée sur l'exploration d'un dataset de la MP.

Sommaire

6.1 Introduction 159
6.2 Description générale 160
6.3 Modèle proposé 161
6.3.1 Représentation de données 162
6.3.2 Distance entre les patients (génération de matrice de distance) . 164
6.3.3 Réduction de dimensionnalité 165
6.3.4 Classification 167
6.4 Résultats expérimentaux 169
6.5 Discussion et évaluation 175
6.5.1 Evaluation des résultats 175
6.5.2 Comparaison 178
6.6 Conclusion 180

6.1 Introduction

L'émergence de la médecine personnalisée et ses avancées exceptionnelles révèlent de nouveaux besoins en matière de disponibilité de modèles de décision médicale adéquats. Par conséquent, pour faciliter la prise de décision médicale, on peut envisager une procédure assistée par ordinateur. Cette dernière repose généralement sur des techniques de classification qui fonctionnent sur des données d'apprentissage. La qualité de ces classifieurs et ces données est très importante pour obtenir d'excellents résultats. En effet, les qualités et les types des données considérées et les outils de représentation et de prétraitement des données peuvent être des facteurs d'influence décisive.

Cependant, le choix des meilleures techniques de classification concernant un problème médical spécifique est également difficile si l'on veut concevoir un système efficace de prise de décision médicale assistée par ordinateur. Différents résultats d'exploration pour le même ensemble de données peuvent être renvoyés en raison des stratégies de calcul de chaque technique. Le temps de calcul des décisions médicales est un autre élément crucial de ces systèmes, en particulier pour les cas urgents. Le volume de données traitées, les techniques de réduction de la dimension des données et la complexité des classifieurs utilisés sont les principaux facteurs affectant le temps de traitement nécessaire pour effectuer une décision médicale automatisée.

Ce chapitre représente une solution qui vise les défis expliqués dans le chapitre de la problématique que ce soit la perte de l'information ou le choix de la série la plus appropriée des traitements. Le temps du calcul et la précision du résultat sont les contraintes considérées lors de la formulation de ces choix. Principalement, cette solution a pour but de produire un outil d'aide à la prise de décision médicale automatisé en fonction de données de la MP.

Sur le plan de la solution proposée, nous avons ciblé les données structurelles en maximisant les types de données appliquées (quatre types) et en minimisant autant

que possible la perte de données et d'informations portées sur les séries temporelles lors du choix de l'outil de représentation de données. Statistiquement, nous avons testé trois distances entre les patients et quatre techniques de réduction de dimensionnalité. Quatre classifieurs différents sont testés et les évaluations globales prennent en compte à la fois les contraintes de performances et de temps de calcul.

En fonction des contraintes de classification considérées, nos travaux expriment trois scénarios pour appliquer les séries de traitements les plus appropriées. Un compromis final entre le temps de calcul et les performances obtenues nous a permis de déterminer une suite de traitement préférée pour une application pratique.

Dans l'ensemble, ce chapitre est organisé comme suit : Nous décrivons brièvement notre approche publiée pour cette solution dans la section 2. Nous détaillons le principe de notre modèle de classification dans la section 3. La section 4 décrit l'expérimentation et l'évaluation de notre proposition en utilisant un ensemble de données de «Alzheimer's Disease Neuroimaging Initiative». Les résultats et les performances atteints sont également discutés. La section 5 résume globalement nos travaux et indique les choix finaux sur les techniques de réduction et de classification dans une véritable application de prise de décision médicale.

6.2 Description générale

Initialement, notre recherche a été lancé autour de l'exploration de données de la MP. Avec les avancements, nous avons constaté qu'il y a deux parties où nous pouvons intervenir, la représentation et la classification de données de la MP. Nous avons commencé par la représentation de données puisque la classification elle-même nécessite une forme précise de données à explorer. Malgré l'évaluation significative de notre approche DRRD de représentation de données (Kadi et al., 2021a) expliquée dans le chapitre précédent, cela n'a laissé qu'une vague vision sur les possibilités réelles de l'application du modèle DRRD, et surtout pour la classification, par rapport à des données plus complètes, réalistes et crédibles. Face aux trois types de marquage

du modèle DRRD, nous avons dû mener des recherches supplémentaires pour déterminer notre nouvelle destination pour l'exploration. Finalement, nous avons choisi de continuer sur le champ de la prise de la décision médicale à base de la représentation symbolique de données et de l'utilisation des techniques de la classification supervisée. Lors de l'élaboration du nouveau modèle de classification, les questions sur les classifieurs idéaux et le nombre de classifieurs à tester se sont imposées comme une priorité pour ce travail. En conséquent, nous avons développé la nouvelle approche qui teste plusieurs techniques du traitement. L'approche intitulée par « **Medical decision-making based on the exploration of a personalized medicine dataset** » a été publiée dans le journal ***Informatics in Medicine Unlocked* (*IMU*)** (Kadi et al., 2021b).

6.3 Modèle proposé

Nous représentons l'ensemble des patients par $P = \{P_1, P_2, ..., P_n\}$ et le nombre total de patients par n. E est la notation de l'ensemble des événements médicaux, où $E = \{E_1, E_2, ..., E_m\}$ et m est le nombre total des événements médicaux.

Notre modèle de classification (Fig 6.1) comporte quatre tâches séquentielles.

A. Le modèle traite les données structurelles et leurs types numériques, date, nominaux et booléens pour les représenter d'une manière plus précise,

B. Les distances entre les patients sont calculées avec la distance de Jaccard (Levandowsky et Winter, 1971),

C. Une réduction de dimensionnalité est appliquée sur la matrice de distance produite en sortie de la tâche précédente, et

D. La classification est effectuée sur les données dans le nouvel espace obtenu.

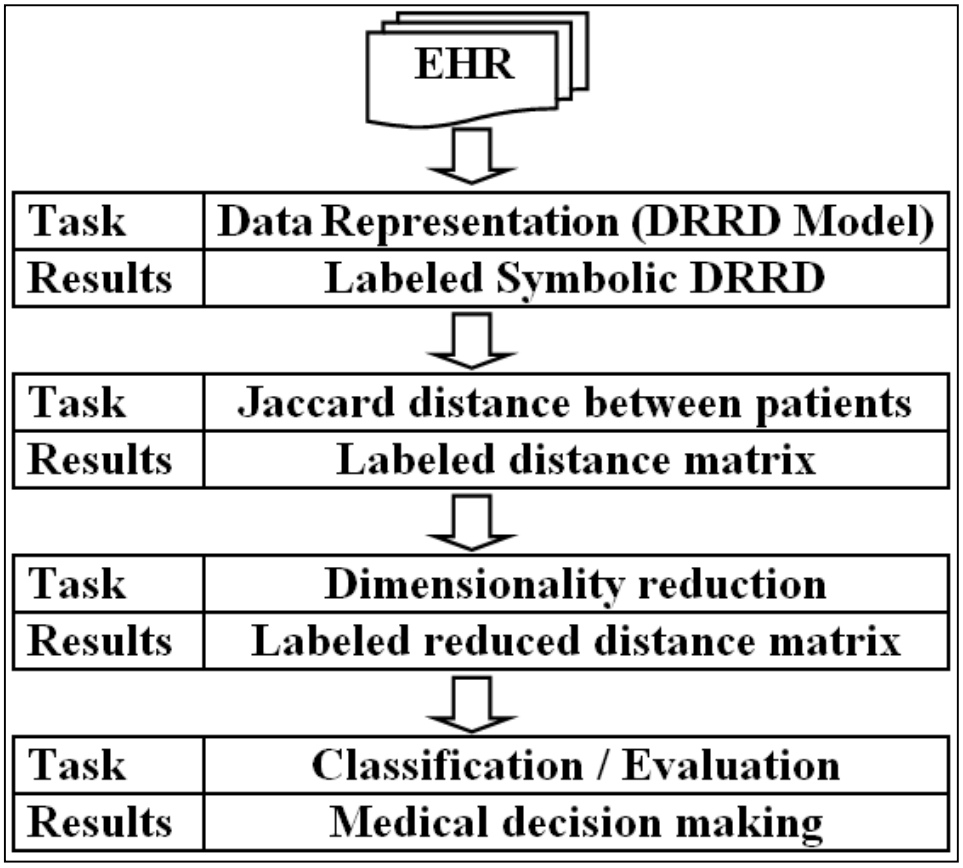

Figure 6.1. Modèle proposé pour de la prise de décision médicale.

Nous détaillons chacune de ces étapes dans la suite.

Parmi ces tâches, la première tâche présente le point du premier contact entre le modèle de classification proposé et les données de la MP. C'est une étape essentielle de cette solution.

6.3.1 Représentation de données

Cette tâche comprend le prétraitement, la transformation et la représentation des données. En raison des nombreux types de données et séries temporelles de données rencontrées dans l'EHR, nous utilisons notre modèle de représentation des données par région et dispersion (DRRD) précédemment proposé (Kadi et al., 2021a) pour générer la représentation symbolique de ces données. Nous rappelons brièvement son principe dans la suite.

La première phase du modèle DRRD est la représentation numérique des données par région, qui est basée sur le custering de données numériques. Nous notifions cette première phase par **DRR** (**Data Representation per Region**). DRR transforme les événements de type date en données numériques en calculant l'âge au moment de l'apparition des observations. Ensuite, il partitionne les données numériques de chaque événement et utilise les clusters comme régions d'appartenance d'observations. Chaque événement numérique sera représenté par une table. La linéarisation par jointure de toutes les tables de représentation produites génère une seule table de représentation globale. A base des opérations de notification et de marquage, la phase DRR générera les trois représentations suivantes : par valeur réelle, binaire et par symbole.

La deuxième phase tente d'imiter le plan du processus de la première phase, mais cette fois les événements de type binaire et nominal sont pris en compte. Cette phase commence par la transformation des données d'événements binaires en données nominales. Chaque valeur de «0» sera remplacée par «F», et chaque valeur de «1» sera remplacée par «Y». Pour chaque événement, E_i, une liste L_i est créée, où chaque liste ne doit inclure que les différentes valeurs observées pour l'événement correspondant. Par la suite, chaque liste L_i correspondant à l'événement E_i sera transformée en table T_i selon un algorithme proposé. Cette phase représente la dispersion de la liste L_i, c'est pourquoi nous la désignons sous le nom de représentation des données par dispersion (**Data Representation per Dispersion DRD**). Avec le même mécanisme de linéarisation que la première phase de DRR, le modèle DRRD génère une seule table de représentation globale pour tous les événements nominaux. Les opérations de notification et de marquage génèrent également trois représentations : par valeur réelle, binaire et par symbole.

Le modèle DRRD assemble les représentations du même type générées par les phases DRR et DRD pour former une seule représentation globale de type par valeur réelle, binaire, ou par symbole selon les besoins.

Dans ce chapitre, nous utilisons la représentation symbolique du modèle DRRD (SDRRD). La table SDRRD produite (équation 6.1) comprend $n = |P|$ lignes selon le

nombre de patients et q colonnes selon la nouvelle représentation des $m = |E|$ événements du dataset, tels que $q \geq |E|$.

$$\forall s_{ij} \in \text{SDRRD}\ (0 < i < |P| \ and\ 0 < j < q\ ) \Rightarrow \begin{cases} s_{ij} = \text{null} \\ \text{Or} \\ s_{ij} \text{ is a symbol} \end{cases} \quad (6.1)$$

Les paramètres du processus DRRD seront enregistrés pour une utilisation ultérieure afin de représenter les données des nouveaux patients. La Figure 6.2 résume toute la tâche de représentation des données.

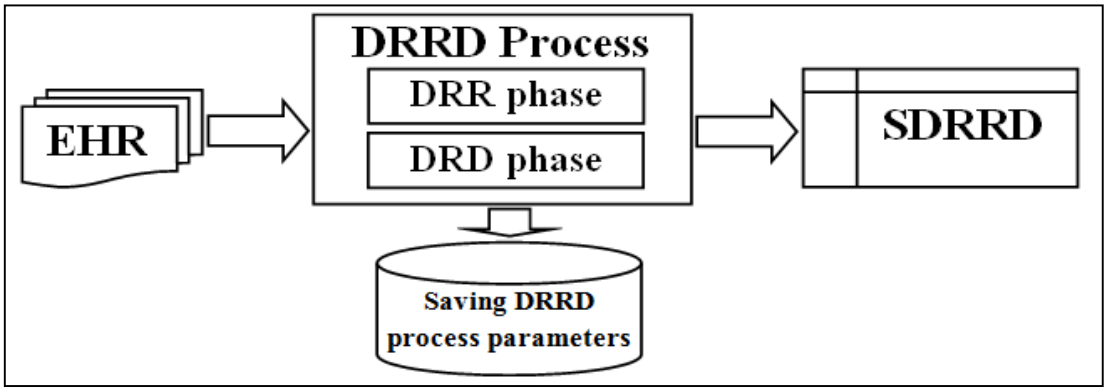

Figure 6.2. Processus de représentation de données.

6.3.2 Distance entre les patients (génération de matrice de distance)

La représentation SDRRD résultant de la tâche précédente constitue le point essentiel et l'entrée principale de la tâche en cours. Les lignes de la matrice SDRRD ont des tailles variables en termes de nombre de symboles exprimés en raison de la variabilité entre les patients. Cette variabilité dépend des événements capturés et de la longueur de la série temporelle enregistrée. A titre d'exemple, la fièvre en tant qu'événement E_f ne peut survenir que pour certains patients P_i, et le nombre de fois de mesures N_f peut également varier entre deux patients. Par conséquent, la représentation symbolique va générer des séquences symboliques de E_f pour ces patients uniquement, avec des longueurs N_f variables. Les patients restants n'auront aucune représentation pour E_f. En général, cette variabilité est la principale cause des valeurs manquantes présentes dans la matrice de représentation SDRRD.

Pour comparer les représentations des patients et éviter de rechercher une méthode de complétion des données manquantes, nous avons décidé de calculer la matrice de distance entre tous les patients (**Distance Matrix between all the Patients DMP**) en fonction d'une distance tenant compte de ce problème. La distance Jaccard (Levandowsky et Winter, 1971) a été choisie pour cette tâche. Pour tous les patients P_i et P_j, nous calculons l'indice de Jaccard J(P_i, P_j) selon l'équation 6.2. Cet indice calcule le pourcentage d'attributs communs par rapport à tous les attributs des patients P_i et P_j.

$$J(P_i, P_j) = {|P_i \cap P_j|}/{|P_i \cup P_j|} \tag{6.2}$$

Le complément de ce dernier indice donne la distance Jaccard DJ(P_i, P_j):

$$DJ(P_i, P_j) = 1 - J(P_i, P_j) \tag{6.3}$$

La matrice DMP générée est une matrice symétrique avec :

$$\forall\, v_{ij} \in \text{DMP}\ (0 < i < |P| \ and\ 0 < j < |P|\ ) \Rightarrow \begin{cases} v_{ij} = 0 \text{ if } i = j \\ \quad \text{Or} \\ v_{ij} \geq 0 \text{ if } i \neq j \end{cases} \tag{6.4}$$

La matrice DMP obtenue constitue une forte base pour la comparaison entre les patients. Elle forme une fenêtre pour l'évaluation de l'homogénéité de l'expression des observations captées chez les patients. Entourée par ses caractéristiques, telles que le volume de données et les variables principales en termes de valeur d'expressivité, cette matrice nécessite un traitement spécialisé avec la tâche suivante.

6.3.3 Réduction de dimensionnalité

Parfois, les EHRs incluent les données de milliers de patients, ce qui génère une matrice DMP étendue (n X n). En effet, la complexité de la technique de classification à appliquer dans la tâche suivante et la grande taille de la matrice DMP peuvent entraîner certaines difficultés de calcul. En plus de la diminution du temps de calcul,

l'objectif de réduction de dimension dans notre approche est également la prise en compte de la visualisation des données, et l'exploitation des variables pertinentes pour améliorer l'apprentissage de nos techniques lors de la prochaine tâche.

La tâche actuelle consiste à réduire la matrice DMP à trois dimensions (3D) seulement. Le résultat de cette réduction est une DMP réduite RDMP (**Reduced DMP**). Intuitivement, cette réduction de dimensionnalité donnera à chaque patient ses trois dimensions les plus proches. La sauvegarde des paramètres avancés de la réduction est une exigence inévitable pour les appliquer dans les futurs traitements avec les nouveaux patients. La préparation des données pour la prochaine tâche de classification nécessite l'insertion de la quatrième colonne dans la matrice RDMP, qui contient les informations d'étiquetage de chaque patient. Cette approche ne sera utilisée que pour l'apprentissage des classificateurs. Le remplissage des informations d'étiquetage basées sur l'ensemble de données EHR produit une matrice RDMP étiquetée LRDMP (Labeled RDMP) (Fig 6.3).

Il suffit d'appliquer une seule technique de réduction sur la matrice DMP dans notre modèle. Cependant, nous avons considéré plusieurs techniques de réduction à des fins de comparaison et d'analyse et pour la sélection de la meilleure technique. Sur la base de l'étude de comparaison réalisée par Ayesha et al., (2020) sur les techniques de réduction de dimension, nous choisissons quatre techniques à tester : PCA (Pearson, 1901; Hotelling, 1933), KPCA (Schölkopf et al., 1997), MDS (Kruskal et Wish, 1978) et TSNE (Maaten et Hinton, 2008). Chaque technique a sa propre stratégie de réduction ce qui nous donne les quatre matrices réduites suivantes : RPCA (Reduction of PCA), RKPCA (Reduction of KPCA), RMDS (Reduction of MDS), and RTSNE (Reduction of TSNE). Chaque stratégie comporte plusieurs lignes égales au nombre de patients et trois colonnes (réduction 3D). L'étiquetage de la réduction de chaque technique produit les matrices réduites étiquetées LRPCA (Labeled RPCA), LRKPCA (Labeled RKPCA), LRMDS (Labeled RMDS) et LRTSNE (Labeled RTSNE).

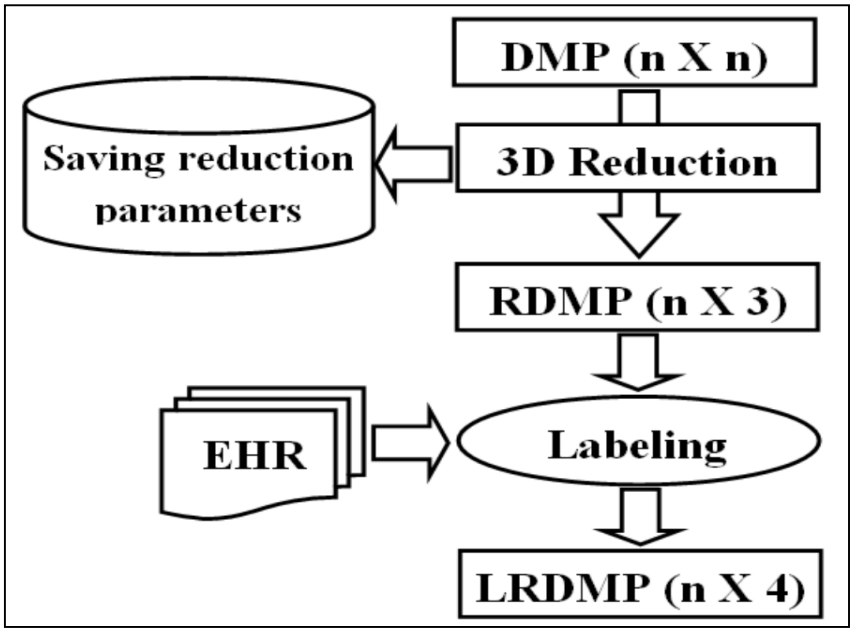

Figure 6.3. Processus de réduction de dimensionnalité.

De plus, réduire la dimensionnalité à un espace 3D permet la visualisation de tous les EHRs à la fois, ce qui peut être intéressant pour l'exploration interactive des données.

6.3.4 Classification

Initialement, la classification d'un nouveau patient *NP* nécessite la génération de sa représentation symbolique SDRRD, appelée NSDRRD (**New patient SDRRD**). Dans ce cas, nous utilisons les paramètres du processus DRRD déjà sauvegardés lors de la première tâche. Soit SDRRD(*NP*) la fonction qui renvoie la représentation symbolique du patient *NP* selon le modèle DRRD où:

$$\forall P_i \in P \ \text{SDRRD}_\text{i} = \text{SDRRD}(P_i) \ \wedge \text{NSDRRD} = \text{SDRRD}(NP) \qquad (6.5)$$

Par la suite, la distance de Jaccard est calculée entre ce *NP* et tous les patients en utilisant les représentations SDRRD (équation 6.6). Bien sûr, cette distance va générer une nouvelle ligne DMP_{NP} (**DMP de NP**) avec $|P|$ colonnes.

$$\forall P_i \in P\ \mathrm{DMP}_{NP}[i] = \mathrm{DJ}(NP, P_i) \text{ where}$$
$$\mathrm{DJ}(NP, P_i) = \mathrm{DJ}(\mathrm{NSDRRD}, \mathrm{SDRRD}_i) \wedge \mathrm{SDRRD}_i = \mathrm{SDRRD}(P_i) \quad (6.6)$$

Pour réduire la dimension de cette nouvelle ligne DMP_{NP}, nous réutilisons les paramètres de réduction générés et sauvegardés lors de la troisième tâche concernée par la réduction de dimension (équation 6.7). L'abréviation NRDMP (**New patient RDMP**) sera utilisée pour indiquer la ligne réduite de la ligne DMP_{NP} de nouveau patient.

$$\mathrm{NRDMP} = \mathrm{3DREDUCTION}(\mathrm{DMP}_{NP}) \text{ where}$$
$$\forall P_i \in \mathrm{P}\ \mathrm{RDMP}_i = \mathrm{3DREDUCTION}(\mathrm{DMP}_i) \quad (6.7)$$

DMP_i c'est la ligne numéro i dans la matrice DMP, et $\mathrm{3DREDUCTION}(\mathrm{DMP}_i)$ c'est la fonction qui renvoie la réduction de la ligne DMP_i selon la technique déployée.

Basé sur l'apprentissage des données de la matrice LRDMP, notre processus applique une technique de classification CT (**Classification Technique**) sur la nouvelle ligne réduite NRDMP pour trouver la catégorie du nouveau patient NPC (**New Patient Category**).

$$\mathrm{NPC} = \mathrm{CLASSIFY}(\mathrm{CT}, \mathrm{NRDMP}, \mathrm{LRDMP}). \quad (6.8)$$

De nouveau, la question de trouver la meilleure technique de classification se pose. Pour avoir une évaluation exhaustive, nous avons décidé le test de plusieurs classifieurs sur les quatre réductions étiquetées LRPCA, LRKPCA, LRMDS, LRTSNE, et sur la matrice DMP sans réduction étiquetée appelée LDMP (**Labeled DMP**). Par le parcours de différentes études (Pak et Teh, 2016; Babar et Mahoto, 2018; Gorade et al., 2017; Paul et Kumar, 2020), avec un nombre raisonnable de classifieurs et par un compromis entre la popularité, la précision et la recommandation des techniques, nous avons choisi de tester les quatre classifieurs suivants : NB (Aggarwal et Vig, 2019 ; Salmi et Rustam, 2019), SVM (Noble, 2006), KNN (Al Bataineh, 2019; Sarkar et Leong, 2000) avec k=3 (indiqué par 3NN) et RF (Ishwaran et Lu, 2018; Kavzoglu, 2017).

Les nouveaux patients peuvent être considérés comme des exemples hors échantillon «**out-of-sample**» qui n'appartiennent pas au dataset de l'apprentissage initiale. Leurs coordonnées dans l'espace réduit sont calculées par la projection en utilisant les techniques PCA et KPCA. Pour la technique MDS, nous utilisons l'extension hors échantillon proposée par Bengio et al., (2003), qui considère un noyau normalisé. Nous considérons l'algorithme de données hors échantillon proposé dans l'article de Gisbrecht et al., (2015) pour la technique du kernel TSNE comme extension pour intégrer les nouveaux patients.

6.4 Résultats expérimentaux

Le dataset de cette expérimentation est fourni par la base de données *The Alzheimer's Disease Neuroimaging Initiative* (ADNI) (adni.loni.usc.edu). ADNI a été lancé en 2003 en tant que partenariat public-privé dirigé par l'investigateur principal Michael W. Weiner, MD. L'objectif principal de l'ADNI a été le test de certains types d'imagerie médicale (tel que : Imagerie par résonance magnétique IRM). Les marqueurs biologiques et les évaluations cliniques et neuropsychologiques peuvent être combinées pour mesurer la progression de la déficience cognitive légère (Mild Cognitive Impairment MCI) et de la maladie d'Alzheimer précoce (early Alzheimer's Disease AD) (Alzheimer's Disease Neuroimaging Initiative [ADNI], 2019).

L'ensemble de données, en particulier la table ADNIMERGE_May15, contient 90 attributs. Parmi ces attributs, on peut citer les attributs suivants: AGE (patient age), FDG (average FDG-PET of angular, temporal, and posterior cingulate), PTETHCAT (ethnicity), PTEDUCAT (education), COLPROT (study protocol of data collection), Hippocampus_bl (UCSF hippocampus at baseline), APOE4 (apolipoprotein epsilon 4), PET (average PIB SUVR of frontal cortex; anterior cingulate; precuneus cortex; and parietal cortex), AV45 (average AV45 SUVR of frontal; anterior cingulate; precuneus; and parietal cortex relative to the cerebellum), CDRSB (clinical dementia rating scale – sum of boxes), ADAS11 (ADAS-Cog-with 11 tasks), GDP (average PIB SUVR of frontal cortex; anterior cingulate; precuneus cortex; and parietal cortex), MMSE (Mini-

Mental State Examination), RAVLT_immediate (Rey Auditory Verbal Learning Test immediate)[31] et autres. Les patients sont classés en cinq classes : Cognitively Normal (CN), Alzheimer's disease (AD), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Significant Memory Concern (SMC). A la base de ces classes, nous avons utilisé les données de 500 patients sélectionnés au hasard (100 patients par classe) pour évaluer notre modèle. La Table 6.1 décrit les statistiques du nombre de lignes de données dans cet ensemble de données, où chaque ligne peut contenir jusqu'à 90 valeurs selon le nombre d'attributs.

Classes	**AD**	**CN**	**EMCI**	**LMCI**	**SMC**	**ALL classes**
Nombre de lignes	456	1096	684	900	270	3406

Table 6.1. Statistiques du dataset.

Nous avons éliminé certains attributs qui n'ont aucune utilité pour cette expérimentation et peuvent biaiser les résultats, tels que l'ID du patient. Il ne reste que 87 attributs après cette élimination. Le processus DRRD déclenche directement l'opération de transformation des données de type date et binaires. Chaque ligne résultante de cette étape correspond à une seule observation (c'est-à-dire une valeur unique). La Table 6.2 résume les statistiques de résultat après cette transformation.

No. of patients	**Nb. des observation numériques**	**Nb. des observation nominales**	**Toutes les observations**
500	75,428	12,188	87,616

Table 6.2. Statistiques d'observation après l'étape de transformation.

Lorsque les deux traitements des représentations DRR et DRD se terminent, le processus DRRD génère la représentation symbolique SDRRD finale en assemblant les résultats. La table symbolique SDRRD représente les 500 patients en fonction de 1393 colonnes.

Le résultat direct de la deuxième tâche est une matrice symétrique DMP composée de cinq cents lignes et cinq cents colonnes.

La Figure 6.4 montre l'affichage 3D correspondant aux résultats de la réduction de dimensionnalité selon les quatre méthodes de réduction considérées. Pour montrer plus de détails sur la distribution de toutes les catégories, nous avons coloré chaque catégorie. Les classes AD, CN, EMCI, LMCI et SMC sont colorées respectivement en rouge, cyan, vert, magenta et orange.

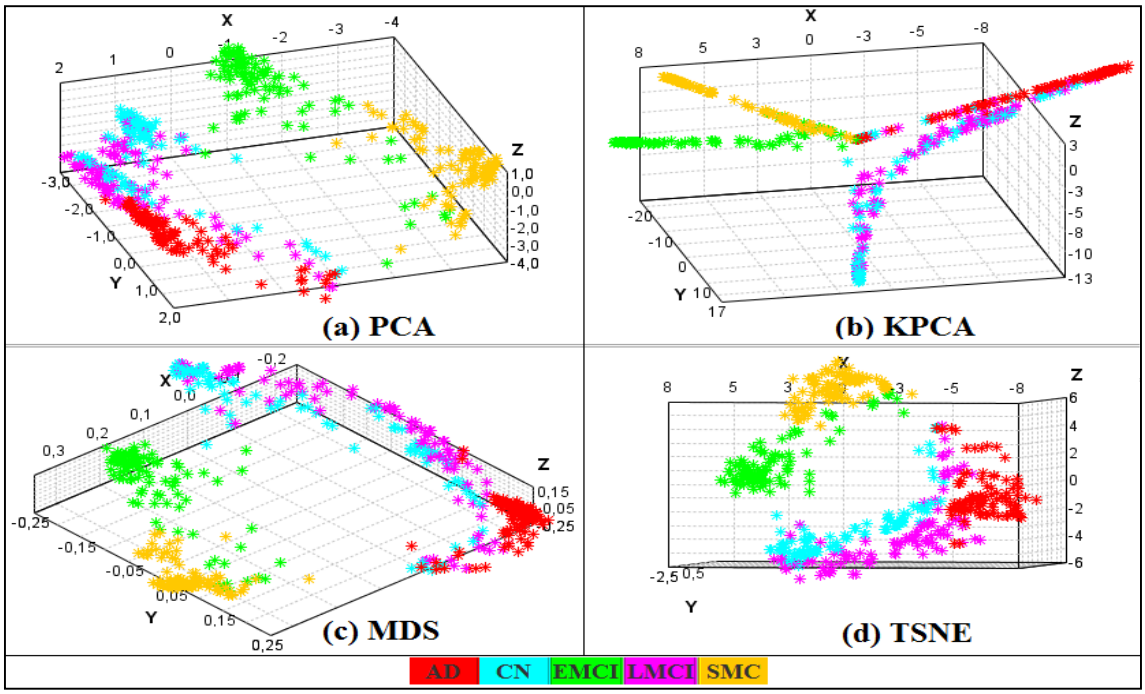

Figure 6.4. Visualisation de la réduction 3D. (a)PCA, (b)KPCA, (c)MDS, (d)TSNE.

Pour la validation de la classification, nous effectuons une validation croisée dix fois. Par la suite, nous utilisons les derniers résultats des tests pour la visualisation de réduction de nouveaux patients à titre de démonstration. Comme des abréviations, les premières lettres des noms de classe sont utilisées pour éviter le chevauchement des couleurs. Les exemples de test des classes AD, CN, EMCI, LMCI et SMC sont présentées respectivement par A, C, E, L et S. La Figure 6.5 visualise ces résultats.

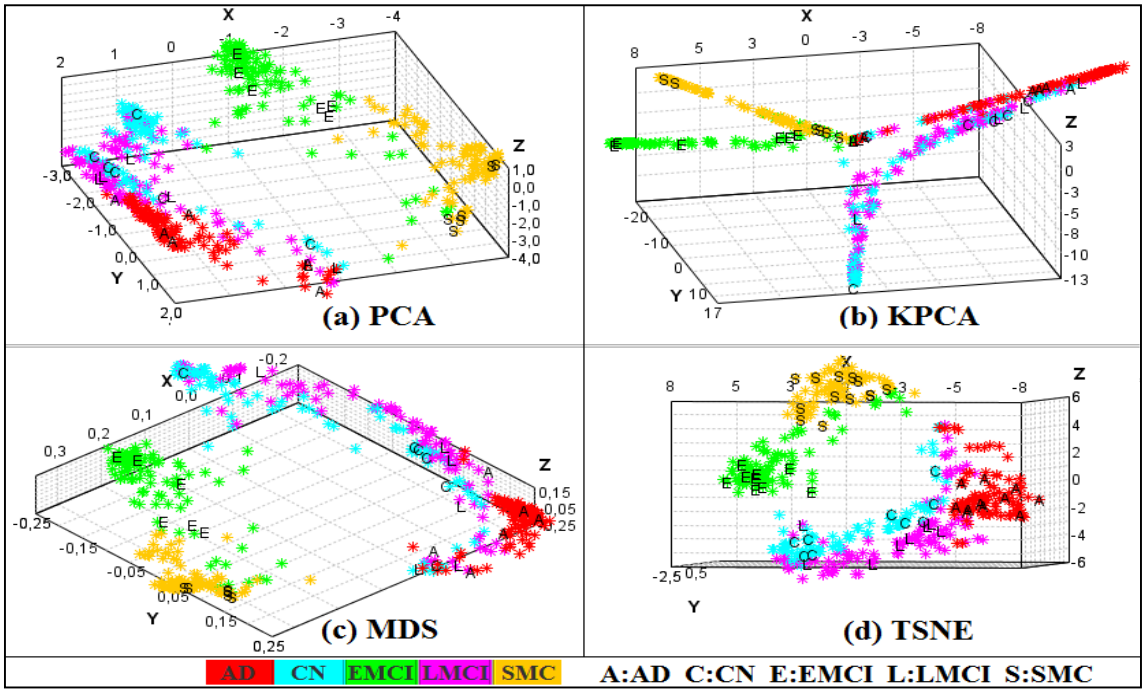

Figure 6.5. Visualisation de la réduction 3D du dernier fold de test. (a)PCA, (b)KPCA, (c)MDS et (d)TSNE.

Pour évaluer notre modèle de classification sur les données sans réduction et avec toutes les techniques de réduction, nous calculons F-measure (FM) par l'équation 6.9 :

$$FM_T = \frac{2*TP_T}{2*TP_T+FP_T+FN_T}. T \in \{NB, SVM, 3NN, RF\}. \quad (6.9)$$

Où, TP_T sont tous les patients testés de la catégorie considérée comme positive qui ont été classés comme positifs lors de l'évaluation de la technique T. Les FP_T sont tous les patients testés de la catégorie considérée comme négative qui ont été classés comme positifs lors de l'évaluation par la technique T. Les FN_T sont tous des patients testés de la catégorie considérée comme positive qui ont été classés comme négatifs lors de l'évaluation par l'évaluation de la technique T.

La Table 6.3 montre les résultats de l'évaluation de la classification de toutes les catégories. Les classifications sur les données réduites par la technique TSNE ont 19 meilleurs cas que les techniques PCA (0 cas), KPCA (1 cas) et MDS (0 cas). Par la suite,

contre les 9 cas pour la classification sans réduction LDMP, nous avons 11 cas pour la classification avec réduction TSNE.

Classes	Classifieur	FM				
		Sans réduction	Avec réduction			
		LDMP	LRPCA	LRKPCA	LRMDS	LRTSNE
AD	NB	0.814	0.82	0.702	0.814	**0.841**
	SVM	**0.985**	0.822	0.711	0.82	**0.823**
	3NN	0.908	0.849	0.792	0.884	**0.913**
	RF	0.919	0.905	0.839	0.906	**0.937**
CN	NB	0.694	0.578	0.59	0.579	**0.728**
	SVM	**0.983**	0.633	0.577	0.615	**0.715**
	3NN	0.882	0.724	0.545	0.763	**0.894**
	RF	0.869	0.737	0.636	0.854	**0.886**
EMCI	NB	**0.917**	0.905	0.877	0.916	**0.914**
	SVM	**1**	0.915	0.901	0.921	**0.931**
	3NN	**0.976**	0.923	0.873	0.96	**0.974**
	RF	0.944	0.93	0.894	0.956	**0.979**
LMCI	NB	**0.633**	0.524	0.461	0.504	**0.603**
	SVM	**0.968**	0.543	0.466	0.533	**0.604**
	3NN	0.813	0.599	0.437	0.659	**0.819**
	RF	0.809	0.675	0.541	0.792	**0.827**
SMC	NB	**0.934**	0.927	**0.936**	0.923	0.921
	SVM	**1**	0.939	0.928	0.936	**0.944**
	3NN	0.986	0.934	0.883	0.96	**0.987**
	RF	0.961	0.941	0.913	0.956	**0.987**
Nb. De meilleures classifications	**Cas de réduction**	/	0	1	0	**19**
	LDMP vs. LRTSNE	9	/	/	/	**11**

Table 6.3. Résultats de FM pour toutes les classes.

Globalement, pour les cinq catégories AD, CN, EMCI, LMCI et SMC, et correspondant à chaque classifieur NB, SVM, 3NN et RF, nous calculons le FM global par l'équation 6.10.

$$\mathrm{FM_T} = \underset{c \in C}{\mathrm{AVG}} \left(\mathrm{FM}_{\mathrm{T}}^{\mathrm{c}} \right). \text{ where}$$

$$C = \{\mathrm{AD}, \mathrm{CN}, \mathrm{EMCI}, \mathrm{LMCI}, \mathrm{SMC}\} \wedge \mathrm{T} \in \{\mathrm{NB}, \mathrm{SVM}, \mathrm{3NN}, \mathrm{RF}\}. \quad (6.10)$$

où, $\mathrm{FM}_{\mathrm{T}}^{\mathrm{c}}$ est l'évaluation FM du classifieur T correspondant à la classe c.

La Table 6.4 affiche ces résultats organisés selon les cas avec et sans réduction.

Pour les résultats de réduction des données dans la dernière table (Tab 6.4), la classification par la technique TSNE a généré les quatre meilleures moyennes. De plus,

TSNE génère 3 meilleures moyennes de FM avec les techniques NB (FM = 0,801), 3NN (FM = 0,917) et RF (FM = 0,923) que l'évaluation sans réduction, qui n'a renvoyé qu'un seul cas avec la technique SVM (FM = 0,987).

Classes		AVG de FM				
		Sans réduction	Avec réduction			
		LDMP	LRPCA	LRKPCA	LRMDS	LRTSNE
NB		0.798	0.751	0.713	0.747	**0.801**
SVM		**0.987**	0.77	0.717	0.765	**0.803**
3NN		0.913	0.806	0.706	0.845	**0.917**
RF		0.9	0.838	0.765	0.893	**0.923**
Nb. De meilleures classifications	Cas de réduction	/	0	0	0	**4**
	LDMP vs. LRTSNE	1	/	/	/	**3**

Table 6.4. Résultats de FM globale.

Pour étudier l'impact de la distance choisie sur le résultat de notre modèle, nous essayons de tester une autre distance au lieu de la distance de Jaccard (J. dist). Nous répéterons toute l'évaluation de notre modèle en calculant la distance de Hamming (Yang et Wang, 2007) (H. dist) et la distance de Levenshtein (Behara et al., 2020) (L. dist) entre les patients lors de la deuxième tâche. La Table 6.5 montre les résultats FM de cette évaluation en utilisant les données sans réduction LRDMP et avec réduction LRTSNE.

Classes	AVG de FM					
	LDMP basée sur			LRTSNE basée sur		
	J. dist	H. dist	L. dist	J. dist	H. dist	L. dist
NB	**0.798**	0.77	0.765	**0.801**	0.793	0.779
SVM	**0.987**	0.957	0.955	**0.803**	0.773	0.776
3NN	**0.913**	0.886	0.89	**0.917**	0.896	0.902
RF	0.9	**0.913**	0.907	**0.923**	0.907	0.906
Nb. De meilleures classifications	**3**	1	0	**4**	0	0

Table 6.5. Comparaison de la classification en fonction des distances choisies.

Pour les statistiques du cas sans réduction, les classifications basées sur la distance de Jaccard ont donné de meilleurs résultats (3 cas) que la distance de

Hamming (1 cas) et la distance de Levenshtein (0 cas). La technique SVM avec la distance Jaccard renvoie toujours les meilleures performances (FM = 0,987).

Comme mentionné précédemment, la réduction de dimensionnalité est utilisée pour minimiser les temps de calcul. Pour la comparaison du cas avec réduction et sans réduction, nous évaluons uniquement le temps écoulé par les classifications appliquées aux matrices étiquetées LDMP et LRTSNE. Pour chaque technique, ce temps du traitement est calculé par le temps de classification moyen des cinq catégories. La Table 6.6 affiche les pourcentages de temps des classifications avec réduction par rapport à celui sans réduction.

Techniques de classification	**Temps de classification (Millisecondes)**		
	LDMP (T1)	**LRTSNE (T2)**	**% (T2/T1)**
NB	109	1	0.90
SVM	118	32	27.11
3-NN	53	2	3.77
RF	443	76	17.15

Table 6.6. Pourcentage du temps de classification écoulé sur les matrices LDMP et LRTSNE en millisecondes pour la classe AD.

Le classifieur SVM a retourné les performances maximales des classifications sur les données sans réduction (FM = 0,987) en 118 millisecondes. D'autre part, les deux classifieurs 3NN et RF montrent d'excellentes performances sur la réduction TSNE (FM ≥ 0.917) avec la supériorité du RF, mais avec un temps court et une supériorité de 3-NN (Temps. 3NN : 2 millisecondes, RF : 76 millisecondes).

6.5 Discussion et évaluation

6.5.1 Evaluation des résultats

La prise de décision médicale automatisée doit faire face à de nombreux défis difficiles, car les résultats sont liés à la santé publique et aux cas individuels de patients et de personnes. L'un des défis auxquels sont confrontés les professionnels de la santé

et les praticiens est l'adoption de modèles performants. Notre contribution génère de nombreux résultats, et leur analyse nous permet d'argumenter nos choix et de discuter des priorités dans l'utilisation de différentes techniques.

L'excellente séparation des données visualisées sur les Figures 6.4 et 6.5 exprime le choix réussi du modèle de représentation des données appliqué et l'excellente qualité des données.

La Table 6.3 montre que la classification de la technique TSNE occupe presque tous les résultats et statistiquement montre une dominance complète. D'autre part, le classifieur SVM a montré des résultats impressionnants et des performances presque parfaites pour les données sans réduction. Simultanément, la technique RF a collecté toutes les meilleures performances de classement sur les données réduites par TSNE, sauf la catégorie CN.

Pour l'évaluation globale par la moyenne, la Table 6.4 confirme l'excellente évaluation de classification des données réduites par la technique TSNE.

La Table 6.3 et la Table 6.4 montrent que les variables pertinentes exploitées pour la représentation dans un espace 3D par la technique TSNE sont plus significatives et préférables pour une utilisation par les classificateurs, en particulier avec 3-NN et RF et des techniques qui adoptent globalement le vote majoritaire. L'ensemble complet des variables de la matrice LDMP est resté plus utile pour la transformation et la séparation par la maximisation des marges avec le classifieur SVM. Ces résultats justifient notre choix de classifieur SVM pour un scénario de classification sans réduction et le choix de la technique TSNE et du classifieur 3NN ou RF pour le scénario de classification avec réduction.

Par la suite, les performances obtenues selon les tests des trois mesures de distance de la Table 6.5 montrent clairement que la distance de Jaccard est la plus efficace avec la réduction des données par la technique TSNE dans notre modèle. Pour une meilleure lisibilité, la Figure 6.6 présente les courbes de FM en fonction des trois distances suivant la technique LRTSNE.

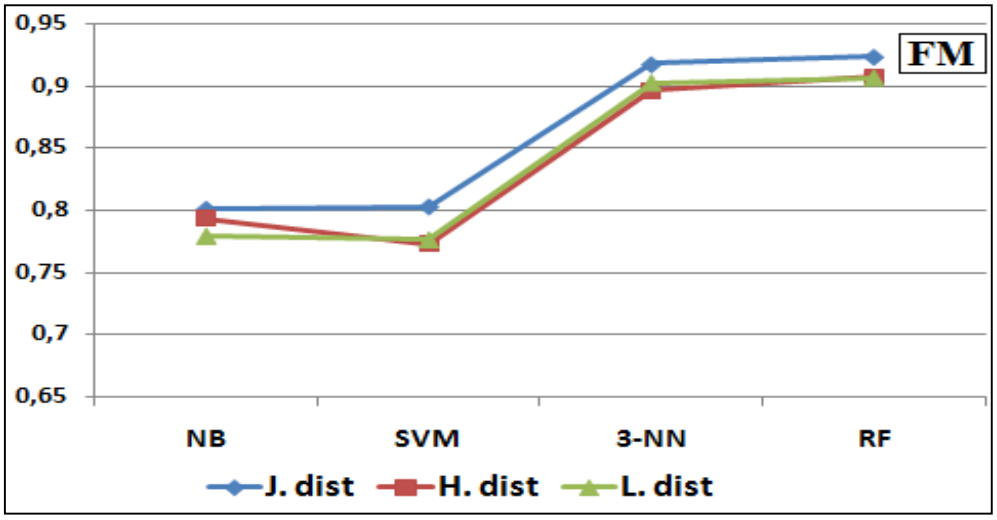

Figure 6.6. Comparaison de FM des classifications sur les données LRTSNE en fonction des distances choisies.

La Figure 6.6 conclut que la distance Jaccard est la meilleure et la plus appropriée et que les classificateurs 3NN et RF sont les plus performants avec les données réduites.

L'analyse de la Table 6.6 montre que le temps écoulé pour la classification sur la matrice LRTSNE est court et parfois négligeable par rapport au cas sans réduction de LDMP, ce qui est compatible avec les avantages de la réduction de données et les exigences globales de notre modèle.

À partir de cette analyse et en fonction des besoins d'utilisation, nous définissons trois scénarios d'application pour notre modèle. La représentation des données est une tâche commune entre eux. De plus, la production de matrice de distance selon la distance de Jaccard entre les patients est une autre tâche courante pour ces trois scénarios. Le premier scénario intervient si le facteur temps est significatif, ce qui invoque le classifieur 3NN. Cependant, si le temps est moins critique, nous utilisons le classifieur RF dans un second scénario. Bien sûr, le classifieur choisi sera appliqué à la matrice de données réduite par la technique TSNE pour ces deux scénarios. A l'inverse, si le facteur temps n'a aucune importance par rapport aux performances FM, le troisième scénario applique le classifieur SVM sur les données sans réduction.

Nous avons réussi le défi de choisir le modèle de représentation le plus approprié pour maximiser l'ensemble de données traitées et minimiser la perte de données et d'informations. Aussi, nous avons réussi le défi de choisir la distance de similarité, le plan de réduction et les classifieurs à appliquer. Les trois scénarios d'utilisation déterminés confirment notre succès face au problème du choix de la meilleure série de traitement.

Pour la maladie d'Alzheimer décrite par le dataset choisi, les performances résultantes démontrent l'importance de ces diagnostics pour juger l'état médical des patients. Ces diagnostics peuvent inclure des éléments faiblement associés à cette maladie, mais les performances atteintes indiquent que certains d'entre eux sont fortement corrélés aux différentes catégories s'il ne s'agit pas de l'ensemble complet.

6.5.2 Comparaison

Pour juger notre modèle et discuter ses caractéristiques par rapport à la recherche actuelle, nous avons examiné les travaux de (Joloudari et al., 2019; Terrada et al., 2020; Carvalho et al., 2020; Lu et al., 2020) qui adressent le même problème de classification en fonction des données des patients et des participants. Globalement, la Table 6.7 résume les quatre approches ciblées par rapport à notre contribution. Les points retenus dans cette table sont les caractéristiques considérées presque comme typiques parmi tous les travaux de l'état de l'art.

Joloudari et al., (2019) ont proposé un modèle pour prédire l'état des patients et la possibilité de souffrance de la maladie du foie. Leur processus a utilisé des données pour 583 patients collectées à partir de trois sources de données, décrit 14 attributs et testé par cinq classifieurs à des fins de comparaison. Avec une stratégie de sélection d'attributs, l'optimisation de l'essaim de particules (Particle Swarm Optimization) PSO -SVM est le meilleur classificateur qui atteint la performance FM = 0,958. Pendant le traitement, les valeurs manquantes, les types de données numériques et nominaux sont tous traités, mais les données de type booléen, date et de série temporelles ne sont pas traitées.

Terrada et al., (2020) ont proposé un processus automatique pour booster le diagnostic de l'athérosclérose. Les données de 835 patients ont été utilisées, y compris 29 attributs, et sept classifieurs ont été testés. Ce modèle a appliqué une stratégie de sélection d'attributs. Le classifieur ANN a généré la performance maximale de FM = 0,98. À l'exception des données numériques, nominales et booléennes, ce processus ne fonctionne pas sur le type de données date, les séries temporelles et les valeurs manquantes.

Approche		**Joloudari et al., 2019**	**Terrada et al., 2020**	**Carvalho et al., 2020**	**Lu et al., 2020**	**Notre modèle**
Patients		583	835	319	349	500
Attributs		14	29	Indéfini	49	87
Classifieurs		5	7	8	3	4
Meilleur classifiuer		PSO-SVM	ANN	A1DE	Log Reg	SVM
Meilleure FM		0,958	0,98	0,95	0,97	0,987
Catégories de sortie		2 Liver disease: Yes/No	2 Atherosclerosis: Yes /No	3 Diagnosis: D/AD/MCI	2 Ovarian Cancer/ Benign Ovarian Tumors	5 AD/CN/ EMCI/ LMCI/SMC
Stratégie		Sélection d'attribut	Sélection d'attribut	N'est pas applicable	Sélection d'attribut	Réduction de Dimension
Temps écoulé		Indéfini	Indéfini	145 min	Indéfini	32 millisec
Série temporelle		No	No	No	No	Yes
Valeurs manquantes		Yes	No	Yes	Yes	Yes
Types de données	**Numérique**	Yes	Yes	Yes	Yes	Yes
	Nominal	Yes	Yes	Yes	Yes	Yes
	Booléen	No	Yes	No	No	Yes
	Date	No	No	No	No	Yes

Table 6.7. Approche proposée par rapport à la recherche actuelle.

L'approche de Carvalho et al., (2020) est un modèle de décision dynamique basé sur un apprentissage supervisé. Pour l'expérimentation, un ensemble de données de 319 patients a été utilisé, mais cette approche ne spécifie pas le nombre d'attributs adopté pour évaluer les huit classifieurs testés. La meilleure performance obtenue était FM = 0,95 par le classificateur A1DE en 145 minutes. De plus, les séries

temporelles, les données booléennes et de date ne sont pas prises en compte dans cette approche.

Pour 349 patients concernés, Lu et al., (2020) ont classé les patients cancéreux en deux catégories : le cancer de l'ovaire et la tumeur bénigne de l'ovaire. Pour la sélection d'attributs pertinents parmi les 49 variables employées, les résultats de trois classifieurs sont comparés. Le classifieur de régression logistique (log reg) a renvoyé la meilleure performance FM = 0,97. Ce modèle traite uniquement les valeurs manquantes et les données numériques et nominales et ignore les données booléennes, date et les série temporelles.

Par rapport à notre contribution, nous avons utilisé le dataset le plus riche en termes de nombre d'attributs (87 attributs) sans compter l'approche de Carvalho et al., (2020) qui n'inclut pas ce détail. Notre modèle classe les données des patients en cinq catégories distinctes et teste quatre classifieurs par rapport à d'autres approches. La technique TSNE adoptée pour la réduction des données nous a permis de minimiser le temps de calcul jusqu'à 32 millisecondes. Bien que nous n'ayons pas utilisé le même ensemble de données que les autres approches, notre dataset a une qualité particulière (types et structures considérés), mais les excellentes performances obtenues indiquent que notre proposition réussit en termes de sélection des séries de traitement les plus appropriées. La variété complète des types de données traitées, la prise en compte des séries temporelles et des valeurs manquantes, les performances de traitement et le temps de traitement écoulé sont tous des facteurs qui mettent en évidence les bonnes caractéristiques de notre modèle et ses avantages par rapport à d'autres travaux. Cette notation modulaire nous permet de considérer notre modèle comme plus fiable et plus efficace.

6.6 Conclusion

La prise de décision médicale automatique basée sur un ensemble de données de la MP est l'objectif principal traité dans ce chapitre. Notre proposition applique une série

de traitements pour atteindre cet objectif. Les tâches appliquées traitent les données structurées de plusieurs types. Ils tiennent également compte les séries temporelles et les valeurs manquantes. Nous avons appliqué notre modèle de représentation de données DRRD développé dans le précédent chapitre. Nous avons testé trois mesures de distance pour calculer la similarité entres les patients : distance de Jaccard, distance de Hamming et distance de Levenshtein. Par la suite, nous avons testé quatre techniques de réduction : PCA, KPCA, MDS et TSNE. De plus, nous avons examiné quatre classifieurs NB, SVM, 3NN et RF pour catégoriser les patients. Par l'expérimentation sur un ensemble de données de la maladie d'Alzheimer, nous avons défini trois scénarios d'utilisation qui ont été formés en fonction à la fois des exigences de performance et du temps de traitement. L'évaluation de notre modèle a atteint FM = 0,987 selon le troisième scénario et 2 millisecondes de temps écoulé pour le premier scénario. Par un compromis entre performances et temps de calcul (FM = 0,923, temps = 76 millisecondes), nous recommandons le deuxième scénario constitué des tâches suivantes : SDRRD, distance de Jaccard, TSNE et RF. Par rapport à d'autres travaux, notre proposition a satisfait plus de facteurs caractéristiques et a démontré un potentiel plus large.

Pour les travaux futurs et plus que l'aspect pratique réel de cette proposition, nous visons à doter notre modèle par d'autres modules pour les diagnostics et les prescriptions de médicaments les plus significatifs. Cette extension vise à construire un système de prise de décision médicale pour contrôler la classification des patients, l'analyse des données et les traitements personnalisés sans effets secondaires.

Références

- Aggarwal, G., & Vig, R. (2019). Acoustic Methodologies for Classifying Gender and Emotions using Machine Learning Algorithms. *Amity International Conference on Artificial Intelligence, Dubai, United Arab Emirates,* 672-677.

- Al Bataineh, A. (2019). A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection. *International Journal of Machine Learning and Computing, Vol 9*, 248-254.

- Alzheimer's Disease Neuroimaging Initiative (ADNI) [Internet]. (2019). (Accessed 2019). Available from: http://adni.loni.usc.edu/.

- Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion, Vol 59*, 44-58.

- Babar, A. H., & Mahoto, N. A. (2018). Comparative Analysis of Classification Models for Healthcare Data Analysis. *International Journal of Computer and Information Technology, Vol 7*(4), 170-175.

- Behara, K. N. S., Bhaskar, A., & Chung, E. (2020). A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance. *Transportation Research Part C: Emerging Technologies, Vol 111,* 513-530.

- Bengio, Y., Paiement, J. F., Vincent, P., Delalleau, O., Roux, N. L., & Ouimet, M. (2003). Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. *In Proceedings of the 16th International Conference on Neural Information and Processing Systems; MA, United States,* 177-184.

- Carvalho, C. M., Seixas, F. L., Conci, A., Muchaluat-Saade, D. C., Laks, J., & Boechat, Y. (2020). A dynamic decision model for diagnosis of dementia, Alzheimer's disease and Mild Cognitive Impairment. *Computers in Biology and Medicine, Vol 126*.

- Gisbrecht, A., Schulz, A., & Hammer, B. (2015). Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing. Vol 147*, 71–82.

- Gorade, S. M., Deo, A., & Purohit, P. (2017). A Study of Some Data Mining Classification Techniques. *International Research Journal of Engineering and Technology, Vol 4*(4), 3112- 3115.

- Ishwaran, H., & Lu, M. (2018). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine, Vol 38*(4), 558-582.

- Joloudari, J. H., Saadatfar, H., & Dehzangi, S. S. (2019). Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. *Informatics in Medicine Unlocked, Vol 17*.

- Kadi, H., Rebbah, M., Meftah, B., & Lezoray O. (2021a). A data presentation model for personalized medicine. *International Journal of Healthcare Information Systems and Informatics, Vol 16*(4), (in press).

- Kadi, H., Rebbah, M., Meftah, B., & Lezoray, O. (2021b). Medical decision-making based on the exploration of a personalized medicine dataset. *Informatics in Medicine Unlocked, Vol 23*.

- Kavzoglu, T. (2017). Object-Oriented Random Forest for High Resolution Land Cover Mapping Using Quickbird-2 Imagery. *In: Pijush S, Sanjiban SR, Valentina EB. Handbook of Neural Computation. London, UK: Academic Press,* 607-619.

- Levandowsky, M., & Winter, D. (1971). Distance between sets. *Nature, Vol 234*, 34-35.

- Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., Li, J., Znati, T., Mi, Q., & Jiang, J. (2020). Using machine learning to predict ovarian cancer. *International Journal of Medical Informatics, Vol 141*.

- Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology, Vol 24*, 1565-1567.

- Pak, I., & Teh, P. L. (2016). Machine Learning Classifiers: Evaluation of the Performance in Online Reviews. Indian Journal of Science and Technology. Vol 9(45).

- Paul, Y., & Kumar, N. (2020). A Comparative Study of Famous Classification Techniques and Data Mining Tools. *In Proceedings of ICRIC 2019. Lecture Notes in Electrical Engineering; Springer, Cham, Switzerland,* 627-644.

- Salmi, N., & Rustam, Z. (2019). Naïve Bayes Classifier Models for Predicting the Colon Cancer. *IOP Conference Series: Materials Science and Engineering, Vol 546*(5).

- Sarkar, M., & Leong, T. Y. (2000). Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. *In Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 759-763.

- Terrada, O., Cherradi, B., Raihani, A., & Bouattane O. (2020). A novel medical diagnosis support system for predicting patients with atherosclerosis diseases. *Informatics in Medicine Unlocked, Vol 21*.

- Yang, H., & Wang, Y. (2007). A LBP-based Face Recognition Method with Hamming Distance Constraint. *In Proceedings of Fourth International Conference on Image and Graphics (ICIG 2007), Sichuan,* 645-649.

Conclusion Générale.

Nous avons commencé notre thèse par les définitions, les explications et les rappels nécessaires destinées à la clarification des axes impliqués dans nos recherches. La médecine personnalisée a été le premier chapitre abordé, pour présenter le cadre général de cette médecine et surtout la notion de qualité des données visées. Comme une nécessité dans le processus d'exploration des données de la MP, nous avons présenté dans le deuxième chapitre les domaines du Datamining, les séries temporelles et le Big data. Ensuite, nous avons continué ce chapitre avec la représentation de données comme un troisième axe. Ce dernier touche certaines opérations nécessaires lors de la présentation, le clustering et la classification de données. Le quatrième chapitre explique les problèmes associés, comme la perte de données et d'information ou le choix de la série des traitements la plus appropriée. Par conséquent, le cinquième chapitre a présenté notre première contribution. Cette dernière présente notre solution au problème de la perte de données et de l'information. D'autre part, le sixième chapitre explique notre deuxième approche concernant la prise de décision médicale à base de données de la médecine personnalisée. En utilisant du modèle de la première approche, la deuxième approche adresse spécialement le problème de choix de la série des traitements la plus approprié.

Depuis le début de cette thèse nous avons ciblé certains détails dont le but est de toucher les problématiques et les défis rencontrés. En pratique, la première solution dépasse le problème de la perte de données et de l'information lors de la représentation à l'hétérogénéité de types de données et l'unification des traitements. Cette solution prend en considération les types de données numérique, booléen, nominal et date et génère une seule représentation globale. Elle représente les données temporelles et non temporelles et conserve le maximum possible les données de ces séries et leurs informations portées. Généralement, l'idée du modèle résultant est la représentation de données de la médecine personnalisée par région et dispersion « Data Representation model per Region and Dispersion (DRRD) ». La comparaison des caractéristiques fonctionnelles du modèle DRRD par rapport à une autre approche de l'état de l'art (Symbolic Aggregate Approximation (SAX)) montre ses forces exceptionnelles et accentue son utilité.

Le modèle DRRD n'a été qu'une plateforme pour la préparation de données de la médecine personnalisée au processus d'exploration. La description des outils impliqués pour ce processus fût entamée depuis le début du manuscrit et surtout dans le deuxième et le troisième chapitre. De notre côté, nous avons visé la tâche d'exploration dans le but de la production d'un modèle de prise d'une décision médicale automatisée à base d'un ensemble de données de la médecine personnalisée. Ce but est dû à la nécessité de production d'un tel système par l'adaptation de telles sources issues de données médicales. La justification du choix et les qualités des résultats de l'application d'une opération sur l'ensemble de données est parfois difficile à cause du manque de résultats de référence. Alors, l'exécution d'une série composée des traitements implique plus de difficultés. Dans l'état de l'art, de nombreuses approches testent un ensemble d'opérations de traitement, mais parfois de manière non adaptée. Dans notre approche, nous avons testé un nombre adéquat de ces opérations. Les tests englobent les distances de similarité entres les patients, les techniques de réduction de données et les classifieurs expérimentés. Durant l'évaluation nous nous sommes concentrés sur deux contraintes d'application : la précision des résultats et la rapidité de calcul. Finalement, nous avons défini trois scénarios d'usage dont chacun exécute une série des traitements bien déterminée. Sur la base de certaines caractéristiques importantes et par rapport à des nouvelles approches, notre modèle de prise de décision médicale montre ses avantages et son utilité.

Relativement à la richesse des sources de données issues de la médecine personnalisée en particulier en matière de diversité, de nombreuses perspectives peuvent être envisagées. En effet il est envisageable d'exploiter des données textuelles issues rapports et de notes médicales, mais également les images, les vidéos et les interviews enregistrés avec les patients, etc. Tout comme les données que nous avons considérées, cela pose des problèmes relatifs à l'étude de leurs caractéristiques comme leur volume, leur pertinence et leur qualité. Enfin les données génétiques peuvent être intéressantes à exploiter mais sont souvent peu accessibles pour des raisons de confidentialité.

Conformément au développement rapide des technologies et la médecine personnalisée, nous pouvons envisager de doter notre modèle de représentation d'autres modules afin de l'étendre à l'ensemble de données possibles à traiter. Les extensions visées pourront alors augmenter la précision des systèmes de prise de décision et toutes les tâches basées sur notre modèle. Par contre, il est fort possible que les nouvelles extensions puissent alourdir notre processus de décision. A cet effet, il serait intéressant d'implémenter nos modèles sur des plateformes dédiées de Big Data. L'extension de notre approche de prise de décision médicale pour la prescription des traitements, la prédiction des effets secondaires et l'analyse de données sont également d'autres perspectives, puisque nous n'avons considéré que la classification comme application. Les maladies rares sont un autre domaine d'intérêt et peuvent être la cible de nos futures recherches, mais cet objectif est difficile car il est directement lié à leur faible disponibilité, souvent pour des raisons légales. A cet effet, la mise en place de plateformes en ligne qui permettent l'exploration de données de la médecine personnalisée semble être une alternative intéressante. Ceci peut avoir l'objectif d'apporter une aide aux organisations et institutions et contribuer à la propagation et l'adoption de la médecine personnalisée.

Publications.

- Kadi, H., Rebbah, M., & Meftah, B. (2018). A data presentation model for personalized medicine. International Conference on Multimedia Information Processing, CITIM'2018. Mascara, Algeria.
- Kadi, H., Rebbah, M., Meftah, B., & Lezoray, O. (2021a). A data presentation model for personalized medicine. *International Journal of Healthcare Information Systems and Informatics. Vol 16*(4), (in press).
- Kadi, H., Rebbah, M., Meftah, B., & Lezoray, O. (2021b). Medical decision-making based on the exploration of a personalized medicine dataset. *Informatics in Medicine Unlocked, Vol 23*.

Résumé

La médecine personnalisée est actuellement en fort développement, de par son adoption dans le monde entier et principalement dans les pays développés. Les profils des patients constituent en effet le point principal sur lequel est fondé le but de cette médecine. Cette dernière vise à aider les médecins et les praticiens de la santé à prévoir des maladies, à prendre des décisions précises et à individualiser les traitements d'une manière adéquate. De plus, le profil d'un patient peut comporter une variété importante de données que ce soient des données génétiques, des biomarqueurs clés, l'historique de traitements, les facteurs environnementaux et les préférences comportementales, des images (IRM, Radio, …), etc. L'exploration de ces données par les outils de la fouille de données nécessite une suite d'opérations pour former et extraire les connaissances cachées parmi ces données. L'intérêt d'un tel processus d'automatisation de la décision médicale et d'extraction des connaissances est généralement confirmé par sa précision. Il ne faut néanmoins pas éluder les contraintes liées à la rapidité de calcul de celles-ci, pour permettre leur usage pratique.

Ces travaux de thèse, intitulés « **Exploration des données de la médecine personnalisée par des techniques de Data Mining** », nous a conduit à la définition de deux activités importantes de la médecine personnalisée : la première porte sur la représentation de données et la seconde sur la prise de la décision médicale. Par conséquent, deux problèmes ont été identifiés. Le premier concerne la perte de données et d'information lors de la phase de représentation de l'information. Le deuxième concerne le choix de la série des traitements la plus appropriée à appliquer pour la prise de décision. La solution de la première problématique a été résolue par la proposition d'un modèle de représentation de données par région et par dispersion. Pour la deuxième problématique, nous avons proposé un modèle de prise de décision médicale réalisé reposant sur une classification de données issues de la médecine personnalisée. Ce modèle repose sur l'application de notre modèle de représentation de données et plusieurs suites de traitement et de classification. L'expérimentation de nos modèles et les résultats obtenus justifient l'utilité et la précision de nos approches.

Ces solutions avantageuses, en particulier le modèle de représentation de données, peuvent être utilisées comme une plateforme exploitable pour d'autres tâches telles que l'analyse de données médicales.

Mots clés : Médecine personnalisée; Data Mining; Représentation de données; Prise de décision médicale; Séries temporelles; Réduction de données; Classification ; Clustering.

Abstract

Personalized medicine is currently in strong development, by its adoption all over the world and mainly in developed countries. The profiles of the patients are indeed the main point on which is based the purpose of this medicine. The latter aims to help doctors and health care practitioners to predict diseases, make accurate decisions and to individualize the treatment adequately. In addition, a patient's profile can include a wide data variety, among genetic data, key biomarkers, treatment history, environmental factors and behavioral preferences, images (MRI, X-ray, etc.), etc. The exploration of this data by data mining tools requires a series of operations to train and extract the knowledge hidden among this data. The advantage of such a medical decision automation process and knowledge extraction is usually confirmed by its accuracy. However, we must not omit the constraints related to the calculation speed of these, to allow their practical use.

This thesis work, entitled « **Exploration des données de la médecine personnalisée par des techniques de Data Mining** », has led us to define two important activities in personalized medicine: the first focuses on data representation and the second on making medical decisions. Therefore, two problems were identified. The first concerns the loss of data and information during the information representation phase. The second concerns the choice of the most appropriate treatment series to apply for decision making. The solution of the first problem was solved by the proposal of a model for the data representation by region and by

dispersion. For the second problem, we have proposed a model of medical decision-making based on a data classification from personalized medicine. This model is based on the application of our data representation model and several treatment and classification suites. Our models' experimentation and the obtained results justify the usefulness and accuracy of our approaches. These beneficial solutions, in particular the data representation model, can be used as an exploitable platform for other tasks such as medical data analysis.

Keywords: Personalized medicine; Data Mining; Data representation; Medical decision making; Time series; Data reduction; Classification; Clustering.