الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة سعيدة د. مولاي الطاهر

كلية الرياضيات والإعلام الآلي والاتصالات السلكية واللاسلكية

قسم: الإعلام الآلي

Mémoire de Master en informatique

Spécialité : Modélisation Informatique des Connaissances et Raisonnement

T h è m e

Exploration de dossiers de santé Electronique (EHR)

- **Présenté par :**
  Abdallah Hadjer
- **Dirigé par :**
  Dr. Kadi Hafid

Année universitaire 2025-2026

Résumé

Le domaine du data mining et des dossiers de santé électroniques connaît actuellement un développement rapide, notamment avec l'expansion des systèmes de santé numériques à travers le monde. Ce domaine vise à améliorer la qualité des soins de santé grâce à la collecte, l'organisation et l'analyse des données des patients afin de faciliter la prise de décision médicale et la prédiction des maladies. Les dossiers de santé électroniques (EHR) représentent une source riche d'informations médicales, car ils contiennent diverses données telles que les antécédents médicaux, les résultats d'analyses, les traitements, les images médicales et les examens cliniques, ce qui en fait un environnement adapté à l'application des techniques de data mining et d'extraction des connaissances.

L'importance du data mining réside dans sa capacité à analyser de grandes quantités de données médicales et à découvrir des relations et modèles cachés à l'aide de techniques d'apprentissage automatique et de classification telles que les arbres de décision, les réseaux de neurones, les machines à vecteurs de support (SVM) et les forêts aléatoires (RF). Ces techniques contribuent également à améliorer la précision du diagnostic et de la prédiction des maladies, tout en aidant les médecins à prendre des décisions thérapeutiques plus efficaces.

Ce mémoire intitulé : « **Exploration de dossiers de santé Electronique (EHR)** » étudie la relation entre les dossiers de santé électroniques et les techniques de data mining, en mettant l'accent sur l'utilisation de la représentation TF-IDF pour améliorer le processus de classification médicale

Abstract

The field of data mining and Electronic Health Records (EHR) is currently experiencing rapid development, especially with the widespread adoption of digital healthcare systems around the world. This field aims to improve the quality of healthcare by collecting, organizing, and analyzing patient data in order to support medical decision-making and disease prediction. Electronic Health Records (EHRs) represent a rich source of medical information, as they contain various types of data such as medical history, laboratory results, treatments, medical imaging, and clinical examinations, making them a suitable environment for applying data mining and knowledge discovery techniques.

The importance of data mining lies in its ability to analyze large amounts of medical data and discover hidden patterns and relationships using machine learning and classification techniques such as decision trees, neural networks, Support Vector Machines (SVM), and Random Forests (RF). These techniques also contribute to improving the accuracy of diagnosis and disease prediction while helping healthcare professionals make more effective treatment decisions.

This thesis, entitled « **Electronic Health Records (EHR) exploration** » studies the relationship between Electronic Health Records and data mining techniques, with a focus on using TF-IDF representation to improve the medical classification process.

ملخص

يشهد مجال التنقيب عن البيانات والسجلات الصحية الإلكترونية تطورًا متسارعًا، خاصة مع التوسع الكبير في استخدام الأنظمة الصحية الرقمية في مختلف أنحاء العالم. ويهدف هذا المجال إلى تحسين جودة الرعاية الصحية من خلال جمع بيانات المرضى، تنظيمها، وتحليلها بهدف دعم اتخاذ القرار الطبي وتسهيل تشخيص الأمراض والتنبؤ بها. تمثل السجلات الصحية الإلكترونية (EHR) مصدرًا غنيًا بالمعلومات الطبية، إذ تحتوي على بيانات متنوعة تشمل التاريخ المرضي، نتائج الفحوصات، العلاجات، الصور الطبية، والتحاليل السريرية، مما يجعلها بيئة مناسبة لتطبيق تقنيات التنقيب عن البيانات واستخراج المعرفة.

تتمثل أهمية التنقيب عن البيانات (Data Mining) في قدرته على تحليل الكم الهائل من البيانات الطبية واكتشاف الأنماط والعلاقات الخفية بينها، وذلك باستخدام مجموعة من تقنيات التعلم الآلي والتصنيف مثل أشجار القرار، الشبكات العصبية، آلات الدعم الشعاعي(SVM) ، والغابات العشوائية .(RF) كما تساهم هذه التقنيات في تحسين دقة التشخيص والتنبؤ بالحالات المرضية، إضافة إلى دعم الأطباء في اتخاذ قرارات علاجية أكثر فعالية.

تتناول هذه المذكرة المعنونة بـ: "استكشاف السجلات الصحية الإلكترونية" دراسة العلاقة بين السجلات الصحية الإلكترونية وتقنيات التنقيب عن البيانات، مع التركيز على كيفية تمثيل البيانات النصية الطبية باستخدام تقنية TF–IDF بهدف تحسين عملية التصنيف الطبي.

DÉDICACE

Je dédie ce travail avec tout mon amour et ma profonde gratitude.

À mon cher père, mon premier soutien,

À ma chère mère, source de tendresse et de courage,

À tous mes sœurs,

À toutes mes amies,

Merci pour votre soutien et vos encouragements.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Introduction générale

Les systèmes de santé modernes génèrent aujourd'hui une quantité massive de données médicales issues de sources diverses, telles que les dossiers médicaux papier, les dossiers médicaux informatisés DMI et, plus récemment, les dossiers de santé électroniques (Electronic Health Record- EHR). Ces données jouent un rôle fondamental dans l'amélioration de la qualité des soins, la prise de décision clinique et le développement de la médecine en général. Toutefois, leur exploitation efficace reste un défi majeur en raison de leur volume croissant, de leur hétérogénéité et de leur complexité.

Historiquement, le dossier médical papier représentait la principale ressource pour la gestion des informations des patients. Malgré son importance, ce type de dossier présente plusieurs limites, notamment la difficulté d'accès, le risque de perte d'informations et les contraintes liées au stockage et au partage. L'introduction du DMI a permis d'améliorer l'organisation et l'accessibilité des données. Cependant, ces systèmes restent confrontés à des problématiques liées à la sécurité, à la confidentialité et à l'interopérabilité.

Avec l'évolution des technologies de l'information, l'EHR s'est imposé comme une solution avancée permettant l'intégration, le partage et l'exploitation des données de santé à grande échelle. Il constitue aujourd'hui un pilier essentiel pour les systèmes d'aide à la décision médicale. Néanmoins, l'exploitation efficace des données issues des EHR nécessite des approches adaptées capables de gérer leur nature complexe, dynamique et hétérogène.

Dans ce contexte, l'exploration des données des EHR à l'aide des techniques de fouille de données (data mining) devient indispensable. Ces techniques permettent d'extraire des connaissances pertinentes à partir de grandes bases de données médicales, à travers des tâches telles que la classification, la prédiction, le clustering et l'analyse des associations. De plus, les données médicales incluent souvent des séries temporelles, représentant l'évolution de l'état de santé des patients, ce qui ajoute un niveau de complexité supplémentaire dans leur traitement et leur analyse.

Cependant, plusieurs défis persistent dans l'exploration des EHR. D'une part, certaines méthodes de représentation et de transformation des données peuvent entraîner une perte d'information, ce qui affecte la qualité des résultats. D'autre part, le choix des techniques de traitement et des algorithmes de classification reste souvent non justifié et dépend de décisions

empiriques. Ces limitations soulignent la nécessité de développer des approches plus efficaces et mieux adaptées.

Ce travail s'appuie sur les résultats d'un modèle développé précédemment et s'inscrit dans un effort visant à améliorer l'exploration des données issues des EHR. Par conséquent, notre proposition peut traiter des données hétérogènes, préserver au maximum l'information et optimiser les performances des méthodes de classification. Pour cela, différentes techniques de représentation, de traitement et d'analyse sont étudiées, comparées et intégrées dans une chaîne de traitement cohérente.

Enfin, les résultats de cette contribution démontrent une efficacité et une adaptabilité remarquables, et ouvrent des perspectives prometteuses pour l'amélioration du processus de prise de décision médicale en particulier.

Problématiques

L'automatisation de la prise de décision médicale constitue le principal axe de cette contribution. Bien que ce ne soit pas le premier travail sur cette branche, les décisions finales confiées aux systèmes automatisés restent le but visé.

En d'autres termes, l'exploration des données des EHR par les techniques de data mining peut améliorer le processus d'automatisation de la prise de décision médicale. Ce processus d'identification et d'extraction de connaissances cachées peut utiliser une variété d'algorithmes, tels que k-NN, SVM, Naïve Bayes, les arbres de décision et les réseaux de neurones. Cependant, le choix de ces algorithmes est directement affecté par les types des données traitées (numérique, nominal, date/heure ou booléen). L'impossibilité d'invoquer directement certains algorithmes, à cause de la qualité des données traitées, nécessite parfois une transformation par des algorithmes reconnus comme le *Term Frequency-Inverse Document Frequency* (TF-IDF). Par ailleurs, le reste des choix est souvent arbitraire ou basé sur leur popularité, sans justification méthodologique claire.

Généralement, on peut résumer les problèmes traités de cette contribution par :

- L'hétérogénéité des données
- Les types de données traitées
- Les transformations adoptées
- Les classifieurs choisis.

Par conséquent, la question posée est : « ***Comment améliorer la classification des patients à partir de données EHR hétérogènes en utilisant une représentation symbolique combinée à la pondération TF-IDF ?*** »

CHAPITRE 1

Electronic Health Record (EHR)

1. Introduction

La transformation numérique des systèmes de santé a profondément modifié les modes de gestion, de conservation et de circulation de l'information médicale. Du dossier médical traditionnel sur support papier au dossier médical informatisé (DMI), puis au dossier de santé électronique appelé en anglais Electronic Health Record (EHR), l'évolution des pratiques documentaires traduit une volonté d'améliorer la qualité, la continuité et la sécurité des soins.

Le dossier médical, autrefois simple outil de mémoire clinique, est progressivement devenu un instrument central du système de santé, à la fois support de la prise en charge thérapeutique, outil d'évaluation, moyen de coordination interprofessionnelle et élément de preuve juridique. Son informatisation s'inscrit dans une dynamique plus large de modernisation des infrastructures sanitaires, visant à optimiser le partage d'informations, à faciliter la prise de décision clinique et à renforcer l'efficience organisationnelle.

Cependant, si le dossier de santé électronique offre des avantages considérables en matière d'accessibilité, d'interopérabilité et de qualité des soins, il soulève également des enjeux majeurs relatifs à la sécurité des données, à la confidentialité des renseignements personnels et au respect du secret médical. La centralisation des informations, l'usage d'identifiants uniques, l'interconnexion des systèmes et les utilisations secondaires des données médicales posent des questions complexes d'ordre technique, juridique et éthique.

Ainsi, ce chapitre se propose d'analyser l'évolution du dossier médical vers le dossier de santé électronique, d'en présenter les caractéristiques, les apports et les limites, tout en mettant en lumière les risques liés à la protection des renseignements personnels dans les systèmes EHR.

2. Dossier médical (DM)

2.1. Historique

Autrefois, la mémoire du médecin suffisait à enregistrer les informations concernant les patients et à soutenir l'exercice de la médecine. Les données médicales étaient collectées sous forme d'articles scientifiques ou de registres à vocation épidémiologique, nosologique et administrative, tandis que le dossier médical se limitait à de simples notes destinées à guider la pratique.

Dans les années 1950, il a pris la forme d'une feuille de température, avant d'évoluer dans les années 1970 au sein des établissements hospitaliers, s'adaptant aux exigences de la médecine en matière de qualité, de sécurité et de continuité des soins, tout en répondant aux besoins de la recherche, de l'enseignement et de la santé publique. L'obligation de rendre compte devant la justice et la demande des patients d'accéder à leur dossier pour obtenir un second avis, accéder aux soins ou dans le cadre d'un litige ont également orienté cette évolution. Aujourd'hui, le dossier médical est encadré par une législation, parfois imprécise, notamment en ce qui concerne la propriété des données qu'il contient.

La bonne gestion d'un dossier médical requiert des moyens financiers, humains, et techniques, ainsi qu'un investissement en temps. L'exhaustivité des informations dépend de la participation active de tous les acteurs de santé, en particulier de ceux directement ou indirectement impliqués dans la prise en charge du patient. Dans les années 1980, la « médicalisation du système d'information » et la création des départements d'information médicale avaient principalement une finalité économique, sans viser à améliorer la qualité intrinsèque du dossier clinique.

Au cours des dernières décennies, le dossier médical a connu plusieurs transformations : il est devenu central dans le système de santé, son volume et son importance pratique ont augmenté, et le nombre de dossiers s'est considérablement multiplié. Ces évolutions ont conduit les pouvoirs publics à légiférer sur le dossier médical et à émettre des recommandations à travers certaines instances spécialisées. [1]

2.2. Définition

Le dossier médical constitue un ensemble de documents, qu'ils soient papiers ou numériques, retraçant les différents épisodes ayant eu un impact sur la santé d'un individu, tels que lettres, notes, comptes rendus, résultats de laboratoire, images radiologiques, etc.

Un dossier médical complet doit inclure les informations suivantes :

- **Informations administratives** : comprenant le nom et prénom du patient, sa date de naissance, son adresse, son numéro de téléphone, sa profession, ainsi que les dates d'admission et de sortie du service
- **Informations médicales** : regroupant l'ensemble des données collectées par les médecins et les praticiens généralistes.
- **Informations paramédicales** : rassemblant les données obtenues par le personnel infirmier et les autres professionnels paramédicaux. [2]

2.3. Contenu du dossier médical

Informations minimales devant figurer dans la première partie du dossier médical [3]

La première partie du dossier doit inclure au minimum les éléments suivants :

- La lettre du médecin ayant motivé la consultation ou l'admission.
- Les raisons de l'hospitalisation.
- Les antécédents médicaux et les facteurs de risque du patient.
- Les conclusions de l'évaluation clinique initiale.
- Le type de prise en charge prévu et les prescriptions effectuées à l'admission.
- La nature des soins dispensés et les prescriptions établies lors de la consultation externe ou du passage aux urgences.
- Les informations relatives à la prise en charge pendant l'hospitalisation, incluant l'état clinique, les soins reçus et les examens paracliniques, notamment l'imagerie.
- Les éléments relatifs à la démarche médicale.
- Le dossier d'anesthésie.
- Le compte rendu opératoire ou d'accouchement.
- Le consentement écrit du patient lorsque la loi ou la réglementation l'exige.
- Les actes transfusionnels réalisés et, le cas échéant, la copie de la fiche d'incident transfusionnel conformément.

- Les prescriptions médicales, leur exécution et les examens complémentaires associés.
- Le dossier de soins infirmiers ou, à défaut, les informations relatives aux soins infirmiers.
- Les soins dispensés par les autres professionnels de santé.
- Les correspondances échangées entre professionnels de santé.
- L'ensemble des résultats d'examens de laboratoire et les examens radiologiques ou d'imagerie (échographies, scanner, IRM, scintigraphies…) réalisés avant, pendant et après une intervention chirurgicale.

Informations formalisées établies à la fin du séjour à la sortie de l'établissement, le dossier doit inclure :

- Le compte rendu d'hospitalisation et la lettre de sortie.
- La prescription de sortie et les doubles des ordonnances.
- Les modalités de sortie (retour au domicile ou transfert vers une autre structure).
- La fiche de liaison infirmière.

Toute information médicale ou paramédicale mentionnée précédemment doit vous être communiquée lorsque vous demandez l'accès à votre dossier médical [2]. Cependant, les « notes personnelles » des professionnels de santé peuvent, dans certaines situations, ne pas être accessibles. Cela s'applique lorsque ces notes sont strictement à usage personnel du praticien et n'ont pas été, ou ne peuvent pas être, utilisées pour l'élaboration et le suivi du diagnostic, du traitement ou pour des mesures préventives.

2.4. Difficultés rencontrées avec le dossier médical papier [2]

- **Prescriptions illisibles** : pouvant entraîner des erreurs de la part du personnel soignant.
- **Confusion des documents** : inclusion de bilans ou clichés d'un patient dans le dossier d'un autre, exposant à des erreurs de diagnostic.
- **Perte de dossiers ou documents** : ce qui nécessite de les refaire, entraînant un surcoût financier et une perte de temps significative.
- **Difficultés pour les études épidémiologiques** : l'exploitation du contenu du dossier peut être complexe.
- **Coût élevé** : usage excessif de papier, pochettes, dossiers et ordonnanciers.

3. Le dossier médical informatisé (DMI)

3.1. Définition

Le dossier médical informatisé (DMI) est un élément clé d'un système d'information en réseau. Il ne se limite pas au suivi des diagnostics et des traitements, mais comprend aussi tous les échanges écrits entre les professionnels de santé. Il rassemble des informations administratives et médicales nominatives, formant ainsi une base de données complète. [3]

3.2. Utilisation d'un dossier patient informatisé (DPI)

Le DPI contient une grande variété de données. Il rassemble à la fois des informations administratives d'identification et des données médicales ou médico-sociales. Il inclut des observations cliniques, des éléments diagnostiques et pronostiques, ainsi que des décisions médicales relatives à la prévention, au diagnostic, au traitement, au pronostic ou au suivi du patient. Ce dossier couvre l'ensemble des informations liées aux soins, qu'elles soient médicales, infirmières, administratives ou relatives à l'Assurance-maladie.

Le DPI a pour objectif principal d'améliorer la qualité des soins. Il facilite la conservation et l'organisation des informations, favorise la communication entre les différents acteurs de soins et assure la continuité du suivi médical. Il permet également de centraliser les informations nécessaires à l'évaluation, à la recherche et à la planification. Conçu pour accompagner le patient tout au long de ses épisodes de soins dans les structures où il est pris en charge, il se distingue du dossier médical personnel (DMP), qui comprend non seulement des données de santé, mais aussi des informations sur le bien-être général et intègre l'avis du patient.

- **Dossier Médical Personnel (DMP)**

Le DMP centralise les informations médicales et améliore la qualité des soins en facilitant la coordination entre les professionnels de santé. Par conséquent, il permet à chaque patient de contrôler les informations contenues dans son DMP.

Les catégories de données regroupent l'ensemble des informations médicales essentielles du patient afin d'assurer un suivi efficace et une meilleure prise en charge par les professionnels de santé. Elles permettent de centraliser les données suivantes :

- **Données générales :** antécédents médicaux et chirurgicaux, historique des consultations, vaccinations, allergies.
- **Données de soins :** résultats d'examens biologiques, comptes rendus diagnostiques et thérapeutiques, pathologies en cours, traitements prescrits ou administrés.
- **Données de prévention :** facteurs de risque, actes de prévention, traitements préventifs, calendrier des vaccinations.
- **Données d'imagerie :** radiographies, scanner, IRM, échographie.
- **Espace personnel :** Permet au patient de communiquer des informations aux professionnels, comme ses souhaits concernant le don d'organes ou les contacts en cas d'urgence. [4]

3.3. Utilité d'un dossier patient informatisé

Le dossier patient informatisé occupe aujourd'hui une place centrale dans le système de santé, car il se situe au croisement de quatre évolutions majeures :

- Le développement de la médecine assistée par ordinateur.
- La recherche d'une meilleure efficacité sanitaire.
- Le besoin de transparence et d'accès à l'information.
- Le besoin de sécurité sanitaire.

Les informatisations du dossier de santé offrent plusieurs avantages :

- **Coordination des soins** : elle facilite la collaboration entre différents professionnels de santé et permet une prise en charge partagée du patient au sein d'un réseau de soins.
- **Organisation et accès rapide aux informations** : les outils de classification permettent de retrouver les données rapidement selon différents critères : nature des informations (cliniques, biologiques, imagerie), ordre chronologique, nom, âge, lieu de résidence ou type de pathologie.
- **Aide à la décision et à l'évaluation** : grâce à des protocoles de prise en charge prédéfinis basés sur des référentiels de pratiques, le dossier facilite l'évaluation de la qualité des soins, les études cliniques coopératives, les recherches épidémiologiques et la traçabilité du parcours du patient.

- **Accessibilité pour le patient** : via Internet, le patient peut consulter son dossier partout dans le monde et en plusieurs langues. Il peut aussi mieux s'impliquer dans sa santé grâce à des alertes automatiques, par exemple pour les vaccinations, les consultations annuelles ou les examens complémentaires à effectuer. [3]

3.4. Les inconvénients de DMI

- **Problèmes techniques** : pannes réseau, défaillances de serveur ou surcharge du trafic.
- **Accessibilité et convivialité** : certaines applications informatiques sont peu intuitives, ce qui nécessite une formation spécifique du personnel soignant.
- **Charge de travail** : la saisie d'un grand volume d'informations, souvent complexes et variées, peut être difficile pour un personnel médical et paramédical déjà fortement sollicité. [2]

4. Electronic Health Record (EHR)

4.1. Présentation du Dossier de Santé Électronique

La création d'un dossier de santé électronique (EHR) est considérée comme la solution privilégiée pour améliorer la circulation de l'information au sein du réseau de santé.

Ce dossier centralise les données essentielles sur la santé d'une personne et est relié à toutes les sources de soins au sein des structures de santé.

Il permet aux professionnels d'accéder rapidement aux informations nécessaires, au bon endroit et au bon moment, sans avoir à consulter les données dans différents dossiers, qu'ils soient papier ou numériques, et dans divers systèmes.

La principale différence entre le Dossier de Santé Électronique (EHR) et le Dossier Médical Électronique (DME) réside dans leur portée et leur contenu. Le DME est l'équivalent numérique du dossier médical traditionnel sur papier : il retrace la relation entre un patient et un professionnel de santé spécifique et appartient à un établissement ou à une clinique particulière, restant confiné à cette structure. Le DME contient des informations plus détaillées que l'EHR, notamment les diagnostics et les notes des professionnels de santé, qui ne sont pas systématiquement transférés dans l'EHR.

L'EHR se présente comme une vue intégrée et sécurisée de l'ensemble des dossiers médicaux d'un patient, provenant de tous les systèmes du réseau de santé. Il offre ainsi une vision globale des antécédents médicaux du patient, facilitant la continuité des soins et la coordination entre les professionnels de santé [5].

A terme, les informations saisies dans le DME seront automatiquement transférées dans l'EHR, permettant une circulation plus rapide et plus large des données, accessibles à tous les professionnels concernés. L'EHR adopte ainsi une logique centrée sur le partage efficace de l'information, garantissant que les soignants disposent toujours des données nécessaires pour assurer la meilleure prise en charge du patient.

4.2. Qu'est-ce que le dossier de santé électronique ?

1. Définition 1

Les EHR sont la version numérique des dossiers médicaux « papier » traditionnels. Ils sont gérés par les structures de santé et contiennent des informations importantes sur les patients, telles que :

- Les antécédents médicaux
- Les diagnostics
- Les médicaments prescrits
- Les traitements suivis
- Les notes cliniques des professionnels de santé
- Les signes vitaux

Ils sont généralement stockés dans une base de données numérique sécurisée, et vous pouvez y accéder via un réseau informatique sécurisé.
Vous pouvez également les utiliser dans divers milieux hospitaliers, y compris les hôpitaux publics et les cabinets privés. [6]

Et plus encore comme Les dossiers de santé génétiques électroniques (GEHRS) L'information génétique (IG) n'est pas encore intégrée de manière systématique dans les EHR à l'échelle mondiale, mais son inclusion devrait se généraliser prochainement, avec le développement de la génomique. Selon la législation locale et la politique des organismes de santé, l'EHR pourrait contenir différents types de données génétiques : [7]

- **Tests sur un seul gène** : pour détecter des maladies monogéniques spécifiques.
- **Tests sur panels de gènes** : analysent plusieurs gènes en même temps pour évaluer différents risques génétiques.
- **Séquençage complet du génome ou de l'exome** : permet d'examiner l'ensemble des régions codant pour les protéines.

Ces analyses aident à identifier des variations génétiques responsables de certaines maladies héréditaires ou susceptibles d'augmenter le risque de leur apparition, constituant ainsi un outil précieux pour la prévention et le suivi médical des patients.

2. Définition 2

Le Dossier de Santé Électronique (EHR) est un dossier longitudinal du patient qui centralise toutes les informations de santé générées lors de différentes consultations dans n'importe quel établissement de soins. Il inclut :

- Les données démographiques du patient,
- Les notes de suivi et d'évolution,
- Les problèmes de santé et diagnostics,
- Les médicaments prescrits,
- Les signes vitaux,
- Les antécédents médicaux et immunisations,
- Les résultats de laboratoire et les rapports de radiologie.

L'EHR permet d'automatiser et de simplifier le flux de travail des cliniciens. Les systèmes de dossiers de santé électroniques peuvent générer des comptes rendus complets des consultations et, grâce à une interface intégrant une aide à la décision fondée sur des données probantes, la gestion de la qualité et La préparation des rapports de résultats, soutenir directement ou indirectement d'autres activités liées aux soins. [8]

3. Définition3

Selon l'ISO TR 20514 :2005, l'EHR (Electronic Health Record En anglais OU Le Dossier de Santé Électronique) est défini comme un référentiel informatique d'informations sur l'état de santé d'un patient, pouvant être traité et utilisé par les professionnels de santé. [9]

Le système de EHR regroupe l'ensemble des composants nécessaires à la création, à l'utilisation, au stockage et à la récupération de ces dossiers électroniques. Il inclutnotamment:

- Les personnes impliquées dans la gestion et l'utilisation du EHR,
- Les données relatives aux patients et aux soins,
- Les règles et procédures régissant l'usage et la sécurité des informations,
- Les dispositifs de traitement et de stockage,
- Les outils de communication et de support assurant le bon fonctionnement du système.

4.3.Historique des EHR

Pour de nombreux hôpitaux et praticiens, les systèmes de EHR peuvent sembler être un phénomène nouveau qui a rapidement envahi le marché. Dans une certaine mesure, la mise en œuvre du EHR a presque doublé de 2007 à 2012, passant de 34,8% à plus de 71%, selon les Centres de contrôle et de prévention des maladies.

Voici 10 faits sur l'histoire des EHR. [10] [11]

1) **Années 1960** : Les premiers EHR apparaissent. En 1965, environ 73 hôpitaux ont lancé des projets d'information clinique et 28 projets de stockage et récupération de documents médicaux étaient en cours.
2) **Clinique Mayo** : Dès le début des années 1960, ce grand centre adopte un EHR pour améliorer la gestion des informations médicales.
3) **Eclipsys (1971)** : Conçu par Lockheed pour l'hôpital El Camino en Californie, ce système précurseur fait aujourd'hui partie d'Allscripts.
4) **DME (1972)** : Les premiers DME apparaissent à l'Institut Regenstrief d'Indianapolis, mais leur coût élevé limite leur diffusion aux hôpitaux publics.
5) **Programmes gouvernementaux (anneées 1970)** : Les EHR sont implantés au Ministère des Anciens Combattants et au Ministère de la Défense, donnant naissance à Vista.
6) **Health Level 7 (1987)** : Créé pour standardiser les EHR et faciliter l'échange d'informations entre systèmes, avec aujourd'hui des membres dans 55 pays.
7) **Objectif de l'IM (1991)** : Encourager l'usage de l'informatique par tous les médecins d'ici 2000. En 2001, seulement 18 % des médecins utilisaient un EHR.
8) **Interopérabilité (années 1990)** : Devient cruciale pour assurer la coordination des soins entre différents systèmes cliniques.

9) **Politique fédérale (2004)** : Le budget informatique en santé double sous George W. Bush et création du Bureau du Coordonnateur national des Technologies de l'information sur la santé, avec un objectif de EHR national.
10) **Responsabilité et EHR** : L'usage des EHR soulève des questions de vie privée et des risques liés aux erreurs de communication, influençant progressivement le cadre juridique et la responsabilité médicale.

4.4. Terminologie du EHR

A. Dossiers Médicaux Électroniques (EMR)

Les Dossiers Médicaux Électroniques (DMÉ) sont la version numérique des dossiers patients et contiennent beaucoup plus d'informations que les dossiers papier traditionnels. Alors que les dossiers papier se limitaient souvent aux antécédents médicaux (maladies, interventions, médicaments…), les DMÉ permettent :

- Saisie normalisée des codes de pathologie et des informations de facturation.
- Aide aux professionnels de santé pour le respect de la réglementation en vigueur.
- Analyse des données pour la recherche sur les maladies, les traitements et les pratiques hospitalières.
- Intégration d'images de haute qualité pour faciliter le diagnostic et réduire les erreurs dues à une écriture illisible. [12]

Il est essentiel de bien distinguer les dossiers médicaux électroniques (DMÉ) des dossiers de santé électroniques (EHR). Les DMÉ sont souvent limités à un seul hôpital ou établissement de santé et ne permettent pas toujours le partage d'informations avec d'autres institutions. À l'inverse, les EHR centralisent la gestion des données cliniques, intégrant les informations provenant des laboratoires, des pharmacies et des thérapeutes afin de constituer un dossier patient complet et de favoriser la coordination et la continuité des soins entre tous les professionnels concernés.

B. Dossiers de Santé Personnels (PHR)

Le PHR est un système électronique qui permet aux utilisateurs d'accéder à leur dossier médical à tout moment, 24h/24 et 7j/7. Il donne aux patients le contrôle de leurs informations de santé et leur permet de partager facilement et instantanément ces données avec tous les

professionnels de santé, qu'ils soient publics ou privés. Ce système facilite l'accès et le partage des données, aidant ainsi à prendre des décisions de santé plus éclairées.

Un dossier de santé personnel PHR fait référence à la collecte de documents médicaux d'une personne conservés par la personne elle-même ou par un soignant dans les cas où le patient est incapable de le faire lui-même. Ces informations personnelles comprennent des détails tels que :

- Les antécédents médicaux du patient,
- Les diagnostics correspondants,
- Les médicaments utilisés dans le passé et actuellement, y compris les traitements en vente libre et les thérapies alternatives,
- Les interventions médicales et chirurgicales antérieures Le statut vaccinal,
- Les allergies et les autres conditions médicales importantes pouvant influencer la prise en charge des soins d'urgence (par exemple le diabète de type 1),
- Le groupe sanguin,
- La personne à contacter en cas d'urgence,
- Les informations relatives à l'assurance,
- Les coordonnées des professionnels de santé qui suivent habituellement le patient.

Contrairement aux dossiers médicaux électroniques (DME) et aux dossiers de santé électroniques (EHR), qui sont généralement gérés par les médecins ou les hôpitaux pour fournir des soins et pour la facturation, le PHR peut être conservé sous forme physique ou, de plus en plus souvent, sous forme électronique.

Il peut contenir des données de santé déclarées et enregistrées par le patient lui-même, comme les problèmes de santé, les traitements, les signes vitaux et l'activité mesurés par des appareils personnels tels que les smartphones ou les montres intelligentes, ainsi que des informations nutritionnelles.

Certaines applications permettent de gérer un PHR et parfois d'intégrer ces données avec le DME ou l'EHR. L'objectif est de permettre au patient d'accéder facilement à ses informations de santé et de les partager avec les personnes impliquées dans ses soins, tout en garantissant la confidentialité et la sécurité des données. [13]

4.5. Facteurs influençant la mise en œuvre du EHR

A. Changements dans le flux de travail clinique

L'introduction d'un système de Dossier de Santé Électronique (EHR) au sein d'un établissement de santé entraîne généralement des modifications importantes dans l'organisation du travail clinique. Pour cette raison, il est préférable d'intégrer ce système dans la vision stratégique globale de l'organisation.

La conception et la mise en place d'un système de EHR doivent se faire avec la participation du personnel clinique, car ce sont eux qui utilisent le système au quotidien. Il est également essentiel de prendre en compte les politiques de l'organisation ainsi que les méthodes de travail déjà en place, afin que le système s'intègre de manière harmonieuse dans l'environnement existant.

Par ailleurs, même si un EHR peut être adapté pour répondre aux besoins d'une pratique médicale spécifique, les modes de travail diffèrent souvent d'une spécialité à une autre.

De ce fait, un système conçu pour un contexte médical particulier peut ne pas s'adapter facilement à d'autres domaines ou pratiques cliniques.

B. Confidentialité et Sécurité

Lors de la mise en place d'un système de EHR, il est très important de faire attention aux questions de confidentialité et de sécurité. Les professionnels de santé peuvent craindre que les informations du dossier de patients soient modifiées sans qu'ils le sachent. De leur côté, les patients ont souvent peur que leurs données personnelles soient consultées par des personnes non autorisées.

De plus, le système de EHR doit respecter les lois du pays qui protègent les données de santé. Cela permet de rassurer les patients et les professionnels de santé que les informations médicales sont bien protégées et conservées en sécurité.

Le système doit aussi enregistrer les accès aux dossiers médicaux et mettre des règles d'accès claires pour que seules les personnes autorisées puissent voir ou modifier ces informations. Cela aide à mieux protéger la confidentialité des données des patients.

C. Identification Unique

La duplication des dossiers de EHR d'un patient dans le même constitue un problème important d'utilisation. Ça arrive parce que les informations viennent de plusieurs hôpitaux ou cliniques, et chacun a sa propre façon d'inscrire les patients. Du coup, le même patient peut avoir plusieurs numéros ou identifiants différents

Quand toutes ces informations sont rassemblées, le système doit faire attention à bien tout relier pour que chaque patient ait un seul dossier complet avec toutes ses données.

D. Interopérabilité

Un système de EHR collecte l'ensemble des informations de santé d'un patient provenant de différents hôpitaux ou cliniques. Il doit donc être capable d'intégrer les données de tous ces systèmes. De plus, il doit permettre l'interopérabilité entre les différentes applications et systèmes de santé développés indépendamment et de manière systématique.

E. Utilisation cohérente des Normes

Pour permettre l'interopérabilité et le partage d'informations entre différentes applications et systèmes de santé, un système de EHR doit utiliser de manière cohérente des normes communes, comme des vocabulaires cliniques et des formats de données standardisés.

En effet, chaque application de santé propose souvent des fonctionnalités différentes et utilise des formats ou structures de données variés.

De plus, elles ne suivent pas toujours les mêmes règles de sécurité ou d'intégrité des données. L'EHR doit donc appliquer ces normes de façon uniforme et se mettre à jour régulièrement pour intégrer les nouvelles normes, afin d'éviter les problèmes et garantir un partage d'informations fiable. [14]

4.6. Importance du dossier de santé électronique (EHR)

Un système de EHR permet l'intégration de plusieurs logiciels de santé, notamment un système d'information hospitalier (SIH), un système de gestion de la pharmacie, un système d'imagerie et un système d'assurance maladie, offrant ainsi une vue d'ensemble cohérente des dossiers médicaux des patients. L'EHR joue ainsi un rôle essentiel dans l'amélioration de la qualité globale des soins, et plus précisément à travers les aspects suivants [14] :

a) Facilité de mise à jour des informations de santé des patients

Un système de EHR permet le traitement électronique des dossiers médicaux, éliminant ainsi le besoin de dossiers papier et réduisant l'espace nécessaire à la conservation des données de santé des patients. De plus, grâce à des politiques de sauvegarde appropriées, les EHR peuvent être conservés par un établissement de santé pendant une période prolongée, ce qui engendre des économies sur les coûts liés à la création, à la maintenance et au stockage des dossiers patients.

b) Efficace dans des environnements complexes

Au sein d'un vaste système de santé, on trouve de nombreuses entités spécialisées (services, laboratoires, centres de formation, centres de recherche, etc.). L'EHR optimise de nombreux processus cliniques et flux de travail associés à ces entités. L'EHR permet à plusieurs d'entre elles d'accomplir diverses tâches : un administrateur peut extraire les données nécessaires à la facturation ; un clinicien peut suivre l'évolution du traitement de son patient ; un infirmier peut consigner un événement indésirable ; et un chercheur peut étudier l'impact des médicaments sur les patients.

c) Meilleurs Soins aux Patients

Souvent, plusieurs professionnels de santé interviennent dans le traitement d'un patient. Un système de EHR permet le partage des informations relatives au patient entre eux.

De plus, les EHR permettent la saisie, la consultation et la modification des données à tout moment, offrant un accès immédiat aux données des patients, quel que soit l'établissement (clinique spécialisée ou hôpital). Cela permet aux professionnels de santé de prendre des décisions opportunes concernant les soins prodigués aux patients. L'accès aux informations de santé (antécédents médicaux, antécédents familiaux et vaccinations, par exemple) via l'EHR favorise la prévention et une meilleure prise en charge des maladies chroniques.

d) Améliorer la Qualité des Soins

L'EHR réduit le temps consacré à la saisie et à la rédaction des rapports pendant vos soins, contribuant ainsi à l'amélioration des normes de santé. L'EHR facilite également la gestion des risques en vous permettant de prendre de meilleures décisions concernant les diagnostics à poser lors de la consultation. Il permet de réduire les coûts des soins. Grâce à l'accès rapide aux informations de santé auprès de tous les organismes de santé concernés, lorsqu'un médecin

consulte la base de données ou le registre de son établissement pour accéder aux antécédents médicaux d'un patient, il n'est pas nécessaire de répéter les tests.

e) Meilleure sécurité

En assurant le contrôle d'accès, l'audit et l'autorisation de tous les systèmes impliqués dans la gestion des dossiers, les EHR offrent un niveau de sécurité pour les dossiers des patients qui n'était pas disponible auparavant dans les systèmes papier.

4.7. Les risques liés à la protection des renseignements personnels dans les EHR

A. Risques d'atteinte à la vie privée relatifs aux dossiers de santé électronique

L'informatisation et la numérisation des dossiers de santé ont entraîné une augmentation des risques liés à la protection de la vie privée. En effet, les dossiers médicaux sur support papier présentaient déjà certains risques, comme l'accès par des personnes non autorisées sans que le patient en soit informé. Toutefois, l'utilisation de systèmes informatisés permet de mieux tracer les accès aux dossiers, ainsi que les modifications, ajouts ou suppressions effectués.

Malgré ces avantages, plusieurs risques demeurent lors de la mise en place des dossiers de santé électroniques, notamment la perte ou le vol de dossiers, ainsi que l'utilisation secondaire des informations personnelles de santé. Ces éléments doivent donc être pris en compte lors de l'implantation de tels systèmes.

Par ailleurs, une note de recherche de la Chaire de recherche du Canada en sécurité, identité et technologie, intitulée « La sécurité précaire des données personnelles en Amérique du Nord » [15] , analyse différents incidents liés à la sécurité des données personnelles au Canada. L'étude indique que, parmi les 23 incidents signalés dans le pays, huit concernent le secteur de la santé. Ces incidents sont principalement dus à la négligence, au vol ou à la perte des données

Les auteurs de l'étude indiquent que certains secteurs en Amérique du Nord sont plus exposés aux incidents liés aux données personnelles.

Le domaine de la santé figure parmi les plus touchés et occupe la troisième place en termes de nombre d'incidents.

Ils expriment leur inquiétude face au développement des services publics en ligne (e-government), car les administrations ne semblent pas toujours capables de garantir la sécurité et l'intégrité des données personnelles ainsi que des transactions électroniques.

L'accès universel aux services publics conduit à la création de grandes bases de données contenant des informations sur toute la population, ce qui peut représenter une cible très attractive pour les fraudeurs.

Pour limiter ces risques, les auteurs estiment qu'il est nécessaire de réaliser d'importants investissements dans la sécurité des données, aussi bien au niveau des technologies utilisées que de la formation des employés aux bonnes pratiques de protection des informations. [15]

En 2007, à la suite du vol d'un ordinateur portable contenant des données de santé, la Commissaire à la vie privée de l'Ontario a exigé que les informations personnelles ne soient pas sorties des bases de données sans être cryptées. [16]

Malgré cette décision, un incident est survenu en 2009 : une clé USB contenant des données personnelles de personnes en attente du vaccin contre la grippe H1N1 a été perdue, et ces informations n'étaient pas cryptées. [17]

Ces événements montrent que la sécurité des données de santé est essentielle pour protéger la confidentialité des informations et maintenir la confiance des citoyens dans le système de santé.

B. Identifiant unique comme moyen de dépersonnalisation des renseignements personnels

En France, le nouvel identifiant unique permettant l'accès au Dossier Médical Personnel (DMP) est appelé « numéro d'identifiant santé ».

Au départ, il était prévu d'utiliser le numéro d'inscription au répertoire (NIR), connu comme le numéro de sécurité sociale.

Ce choix présentait plusieurs avantages, car il s'agit déjà d'un identifiant unique existant, simple à utiliser et rapide à mettre en place, ce qui évitait de créer un nouvel identifiant pour l'ensemble de la population française.

Le NIR est en effet considéré comme un numéro significatif, unique, stable et fiable. [18] Cependant, avec l'accord de la Commission nationale de l'informatique et des libertés (CNIL), ce numéro a été largement utilisé, ce qui a soulevé des inquiétudes concernant les risques de

croisement et d'interconnexion des données personnelles [19], De nombreuses critiques ont alors été formulées contre l'utilisation du NIR, jugé relativement facile à reconstituer, ce qui aurait pu permettre à des tiers d'accéder plus facilement au DMP. Par la suite, la CNIL s'est opposée à l'utilisation du NIR comme clé d'accès au DMP [20].

Finalement, un identifiant spécifique appelé numéro d'identifiant santé a été instauré et prévu par l'article L. 1111-8-1 du Code de la santé publique. [21]

Aux États-Unis, le docteur Richard Sobel met également en évidence les dangers liés à l'utilisation d'un identifiant unique, notamment en ce qui concerne la centralisation des informations médicales. Ce risque existe aussi au Québec, où les données sont centralisées à l'échelle régionale. Selon le docteur Sobel, cette situation pourrait faciliter l'accès aux bases de données médicales et soulever des inquiétudes importantes quant à la confidentialité des informations personnelles [22].

Il identifie notamment quatre types de risques principaux :

- L'augmentation des accès, autorisés ou non, aux informations par des tiers.
- L'utilisation ultérieure de ces données à des fins différentes de celles prévues lors de leur collecte.
- La possibilité d'utiliser cet identifiant pour effectuer des recoupements de données.
- Une éventuelle perte de confiance des usagers envers le système de soins de santé.

De son côté, la Commission d'accès à l'information insiste également sur la nécessité de faire preuve de vigilance concernant l'utilisation de cet identifiant unique. Elle souligne que la multiplication de ses usages pourrait favoriser les recoupements d'informations. Une telle situation faciliterait l'élaboration de profils détaillés sur les individus et pourrait conduire à un fichage de la population [23].

C. Utilisations secondaires des renseignements personnels de santé

Lors de la présentation des règles du Code fédéral américain, le secrétaire du Département de la Santé et des Services sociaux a rappelé que les individus ne peuvent partager des informations très personnelles que s'ils ont confiance dans la protection de leurs données. Selon lui, les violations de la vie privée diminuent la confiance des citoyens envers le système de santé et les institutions qui les servent, ce qui peut nuire à la qualité des soins reçus ainsi qu'à la situation financière des établissements de santé [24]

En principe, les renseignements personnels de santé sont collectés, utilisés et divulgués principalement pour permettre aux patients de recevoir des soins médicaux. Cette finalité constitue l'objectif principal de leur collecte.

Les dossiers de santé électroniques sont présentés comme un moyen d'améliorer la prise en charge des patients et l'organisation du système de santé, notamment dans ses aspects administratifs.

Dans son livre blanc sur la gouvernance de l'information dans le dossier de santé électronique interopérable, Inforoute Santé du Canada définit l'utilisation secondaire comme l'utilisation ou la divulgation de renseignements personnels pour des objectifs différents de ceux pour lesquels ils ont été collectés à l'origine [25],

L'organisme précise également que, même si les données du EHR sont d'abord recueillies pour les soins et l'administration du système de santé, il est probable qu'elles soient utilisées par la suite pour d'autres finalités, comme la surveillance de la santé publique, l'analyse du système de santé ou la recherche scientifique

Toutefois, ce type d'utilisation peut être critiqué, car il s'éloigne de l'objectif initial pour lequel le patient a accepté la collecte et l'utilisation de ses informations personnelles [25].

Pour rendre ces utilisations secondaires acceptables, il est nécessaire que les renseignements personnels de santé soient dépersonnalisés ou anonymisés, afin de protéger l'identité des patients [25]

Cependant, la dépersonnalisation ou l'anonymisation des données ne permet pas d'éliminer totalement les risques pour la vie privée. L'expérience menée par la professeure Latanya Sweeney en est une illustration claire.

Dans cette étude, elle a réussi à réidentifier certaines personnes à partir de données anonymisées provenant du groupe insurance Commission, l'organisme chargé de l'assurance médicale des employés de l'État. Pensant que ces données étaient suffisamment anonymisées, l'organisme avait fourni des copies à des chercheurs.

La professeure Sweeney a alors obtenu des informations provenant du registre électoral et a croisé certaines données communes, notamment le code postal, la date de naissance et le sexe. Grâce à ce recoupement entre les deux bases de données, elle a pu identifier certaines personnes et accéder, par exemple, aux informations du dossier médical du gouverneur du Massachusetts.

À la suite de cette expérience, l'auteure a proposé plusieurs méthodes visant à limiter les risques de réidentification par croisement de fichiers.

Ces mesures seraient particulièrement importantes à mettre en place si l'on souhaite permettre des utilisations secondaires des données de santé. [26]

5. Conclusion

L'informatisation des dossiers de santé, qui a évolué du support papier au dossier médical informatisé (DMI) puis au dossier de santé électronique (DSE ou EHR), traduit la profonde transformation numérique des systèmes de santé contemporains. Cette mutation ne se limite pas à un simple changement de support : elle reconfigure l'organisation, la gestion et la circulation de l'information médicale. Le dossier médical, initialement outil de mémoire clinique, est devenu un élément central du système de soins, améliorant la traçabilité, la coordination entre professionnels, la qualité des prises en charge et l'efficience administrative.

L'EHR offre une vision intégrée des données patient, facilite l'interopérabilité entre établissements et permet un accès rapide à l'information essentielle. Il soutient la continuité des soins, la prise de décision clinique, la réduction des erreurs liées aux prescriptions illisibles ou aux pertes de documents, et renforce l'implication du patient dans son parcours, qui peut désormais suivre et participer activement à ses soins.

Cependant, cette modernisation s'accompagne de défis majeurs. La centralisation des données, l'usage d'identifiants uniques et les possibilités d'utilisations secondaires exposent les patients à des risques accrus d'atteinte à la vie privée. Le consentement du patient, bien que fondamental, se trouve affaibli par de nombreuses exceptions prévues par la loi et par des finalités administratives périphériques à la prise en charge médicale. L'efficacité des outils permettant au patient de contrôler l'accès ou de masquer certaines informations dépend directement de la limitation de ces exceptions.

Ainsi, le succès du EHR repose sur un encadrement juridique clair et cohérent, des standards techniques fiables et une gouvernance qui respecte la sécurité, la transparence et les droits fondamentaux des patients. La confiance des usagers reste l'élément central pour garantir l'acceptabilité et la pérennité de ces systèmes numériques de santé.

CHAPITRE 2

Data mining

1. Introduction

La technologie du data mining intègre des concepts et des outils issus des statistiques, de l'intelligence artificielle et de l'informatique, et permet aux entreprises et aux organisations de tous les secteurs de traiter de grandes quantités de données et d'en extraire des connaissances — des informations auparavant inconnues — qui peuvent être utiles pour la prédiction et l'aide à la décision.

La croissance du nombre d'entreprises dans le monde, ainsi que les progrès réalisés en matière de capacité et de qualité des technologies de stockage de données, ont favorisé l'émergence de techniques statistiques, analytiques et d'apprentissage automatique. Elles permettent d'explorer, de traiter et d'extraire des informations et des connaissances cachées et utilisables à partir de ces sources de données. L'ensemble des techniques et de ces tâches constituent la base sur laquelle repose le domaine de découverte (ou extraction) des connaissances dans les bases de données « Knowledge Discovery in Data bases (KDD) » [27]

Dans ce contexte, la valorisation des données devient un enjeu stratégique majeur pour les organisations modernes. En effet, la capacité à analyser et interpréter efficacement les informations disponibles permet non seulement d'améliorer la performance opérationnelle, mais aussi d'anticiper les tendances, d'optimiser les processus et de soutenir la prise de décision. Ainsi, la maîtrise des méthodes et outils liés à la découverte de connaissances s'impose aujourd'hui comme un facteur clé de compétitivité et d'innovation.

Le chapitre suivant s'attache à définir le data mining ainsi que les techniques à mettre en œuvre.

2. Qu'est-ce que le DM ?

C'est Le processus inductif, itératif et interactif vise à découvrir, dans de grandes bases de données, des modèles de données qui soient valides, nouveaux, utiles et compréhensibles**.**

- **Itératif** : le processus nécessite plusieurs passes pour affiner et confirmer les modèles.
- **Interactif** : l'utilisateur participe activement au processus, prenant part aux choix et aux validations.
- **Valides** : les modèles découverts doivent rester fiables et applicables dans le futur.
- **Nouveaux** : ces modèles ne doivent pas être prévisibles à partir des connaissances existantes.

- **Utiles** : ils doivent aider l'utilisateur à prendre des décisions éclairées.
- **Compréhensibles** : les résultats doivent être présentés de manière claire et simple. [28]

2.1 Définition

Le Data Mining (DM) est un processus itératif visant à identifier, reconnaître et extraire des connaissances cachées, valides et utiles à partir d'un ensemble de données. Ces connaissances peuvent prendre la forme de modèles, de relations ou encore de structures inconnues. [29]

3. Le processus de data mining

Le processus de data mining consiste à analyser des données sous différentes perspectives afin d'en extraire des informations utiles. Ces informations peuvent ensuite être exploitées pour améliorer la prise de décision, optimiser les performances ou soutenir des objectifs stratégiques, tels que l'augmentation des revenus et la réduction des coûts.

Ce processus (Fig.1) peut être divisé en sept étapes principales [30] :

A. **Sélection des données** : On récupère uniquement les données pertinentes pour la tâche d'analyse.
B. **Intégration des données** : Les données provenant de plusieurs sources différentes sont combinées en une seule base cohérente.
C. **Nettoyage des données (Prétraitement)** : Cette étape permet de gérer les données bruitées, erronées, manquantes ou non pertinentes afin d'améliorer la qualité de l'analyse.
D. **Transformation des données** : Les données sont modifiées ou regroupées sous une forme appropriée, souvent par des opérations de synthèse ou d'agrégation.
E. **Exploration des données** : C'est l'étape clé où l'on applique des techniques ou des algorithmes pour découvrir des modèles et des motifs dans les données.
F. **Évaluation du modèle** : Cette étape consiste à examiner la qualité, la pertinence et l'utilité des connaissances obtenues, afin de s'assurer qu'elles sont fiables et exploitables pour la prise de décision.
G. **Présentation des connaissances** : Enfin, on utilise des techniques de visualisation et de représentation pour restituer les informations extraites de manière claire et compréhensible pour l'utilisateur.

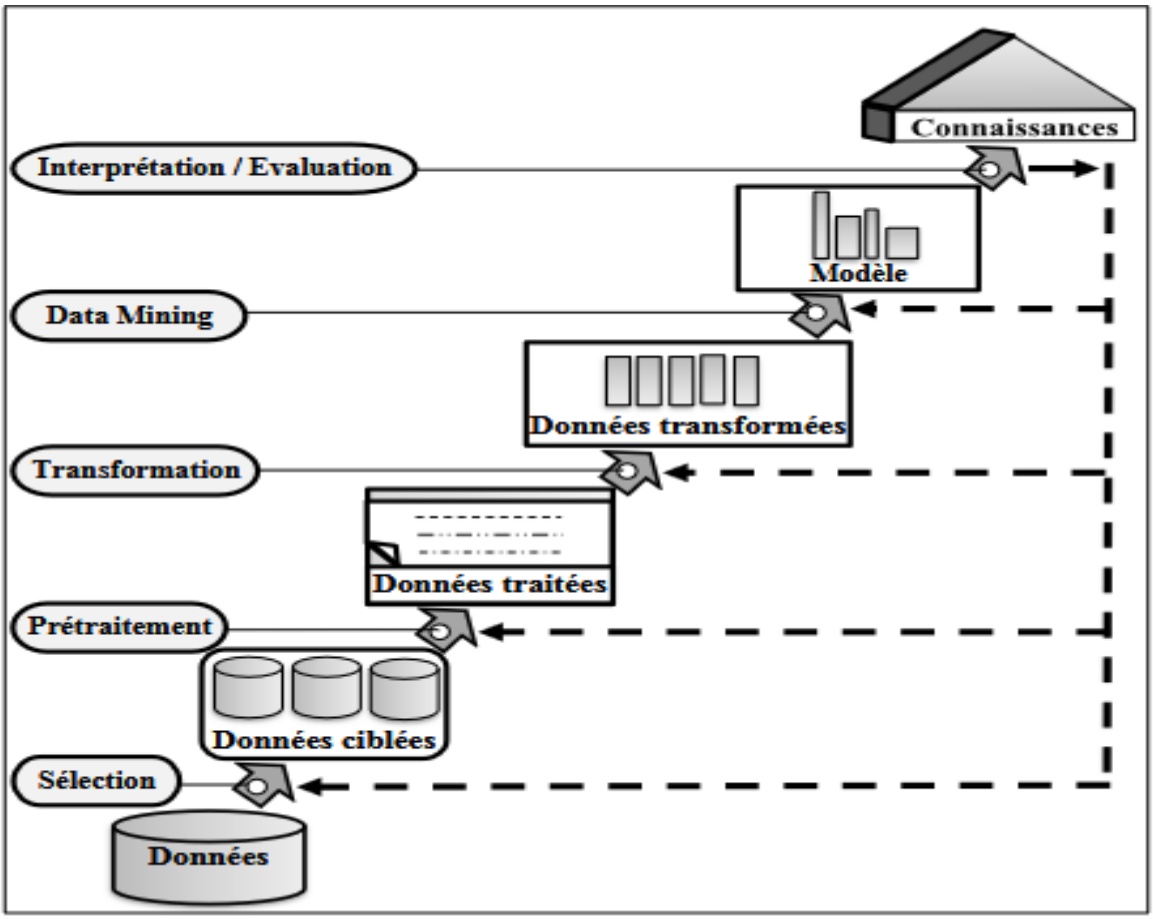

Figure 1 : Processus de data mining [27]

4. Les types de données en Data Mining

4.1 Les fichiers plats

Les fichiers plats sont aujourd'hui l'une des sources de données les plus courantes, surtout dans le cadre de la recherche. Ce sont des fichiers simples, en format texte ou binaire, dont la structure est connue à l'avance par l'algorithme de Data Mining qui sera appliqué.

4.2.Les bases de données relationnelles

Les algorithmes appliqués sur des bases relationnelles sont plus polyvalents que ceux conçus uniquement pour les fichiers plats. Ils peuvent exploiter la structure naturelle des bases relationnelles et utiliser le SQL pour sélectionner, transformer et consolider les données. Cela permet d'aller au-delà du SQL classique, par exemple pour la prévision, la comparaison ou la détection d'anomalies.

4.3.Les Data Warehouses

Un Data Warehouse regroupe des données provenant de sources multiples, souvent hétérogènes, dans une structure unifiée. Par exemple, une entreprise comme *OurVideoStore*, qui possède plusieurs magasins en Amérique du Nord, pourrait avoir des bases de données différentes dans chaque magasin. Si le directeur souhaite prendre des décisions stratégiques, il est préférable que toutes les données soient centralisées, nettoyées, transformées et intégrées afin de permettre une analyse interactive et homogène.

4.4.Les bases de données transactionnelles

Une base transactionnelle enregistre chaque transaction individuellement. Chaque transaction possède un identifiant unique et une liste d'items associés (par exemple, les achats d'un client lors d'une visite). Elle peut également inclure d'autres informations comme la date, l'identifiant du consommateur ou du vendeur. Ces bases sont souvent utilisées pour l'analyse des règles d'association, également appelée analyse du panier de la ménagère.

4.5.Les bases de données multimédia

Ces bases contiennent des documents audio, vidéo, images ou textes enrichis. Elles peuvent être stockées dans des bases orientées objet, relationnelles ou même simplement sur un système de fichiers. La taille importante des données multimédia rend le Data Mining plus complexe, nécessitant parfois des techniques de vision par ordinateur, infographie, interprétation d'images ou traitement du langage naturel.

4.6.Les bases de données spatiales

Ces bases contiennent, en plus des données classiques, des informations géographiques comme des cartes ou des positions mondiales/régionales. Elles représentent un défi supplémentaire pour les algorithmes de Data Mining en raison de la complexité de la gestion des données géospatiales.

4.7.Les bases de données de séries temporelles

Ces bases enregistrent des données liées au temps, comme les cours boursiers ou les activités enregistrées dans un système. Elles reçoivent souvent un flux continu de nouvelles données, ce

qui peut nécessiter une analyse en temps réel. Le Data Mining sur ces bases étudie les tendances, les corrélations et peut prédire l'évolution des variables dans le temps. Par exemple, une base sur le trafic automobile peut permettre de répondre à des questions comme : *« Quels sont les axes où la circulation est fluide le week-end ? »*.

4.8.Le World Wide Web

Le Web est une source de données extrêmement hétérogène et dynamique. Des milliers d'auteurs et d'éditeurs y contribuent chaque jour, et des millions d'utilisateurs accèdent aux ressources disponibles. Les données sont organisées dans des documents interconnectés qui peuvent être des textes, audio, vidéos ou applications. Le Web Mining s'intéresse à trois dimensions : le contenu du Web, la structure (liens entre documents) et l'usage (comment les ressources sont consultées). On peut également considérer la dimension dynamique, qui étudie l'évolution constante des documents. [31]

5. Les tâches du data mining

Il existe de nombreux algorithmes et méthodes en data mining. D'un point de vue technique, ces méthodes peuvent être organisées selon plusieurs critères de classification :

5.1.Selon l'objectivité

Le critère essentiel retenu dans cette analyse est l'objectivité, afin d'assurer une interprétation neutre et fiable des données. Ainsi, les résultats obtenus reposent sur des méthodes rigoureuses, indépendantes de toute subjectivité ou biais.

A. La classification : La classification est une technique d'apprentissage supervisé qui permet d'attribuer une nouvelle donnée à une catégorie prédéfinie, en se basant sur un ensemble de données déjà étiquetées. Les éléments à classifier sont souvent des enregistrements provenant d'une base de données. Le processus repose sur un ensemble d'exemples préalablement classés, appelé ensemble d'apprentissage, qui sert à construire un modèle capable de prédire la catégorie des nouvelles données.

L'objectif principal est donc de développer un modèle fiable pouvant être appliqué à des données non encore classifiées afin de prédire correctement leur catégorie [32] .

B. La prédiction : La prédiction ressemble à la classification et à l'estimation, mais elle ajoute souvent une dimension temporelle. Elle permet d'anticiper des valeurs ou des événements futurs en s'appuyant sur des données historiques. Contrairement à la classification classique, le modèle prédictif intègre l'évolution des variables dans le temps pour fournir des projections plus fiables [33].

Quelques exemples de l'utilisation des tâches de prédiction dans les domaines de recherche et commerce sont les suivants :

- Prévision du prix des actions dans les mois à venir [34].
- Prévision du vainqueur d'une compétition sportive à partir des statistiques des équipes
- Identification des clients susceptibles de changer de domicile dans les six prochains mois [33].

C. Association : Ce sont des techniques d'apprentissage non supervisé utilisées pour explorer les données et découvrir des relations cachées. Elles permettent de mettre en évidence des corrélations entre les attributs et d'interpréter ces relations de manière utile [35].

D. Le clustering (Segmentation) : Le clustering, ou segmentation, consiste à regrouper des enregistrements ou des observations en groupes d'objets similaires. Chaque groupe, appelé cluster, contient des éléments proches les uns des autres et distincts des éléments des autres clusters.

Contrairement à la classification, le clustering ne repose pas sur une variable cible. Il ne cherche pas à classer, prédire ou estimer une valeur précise. L'objectif est plutôt de diviser l'ensemble des données en sous-groupes relativement homogènes, en assurant une grande similarité à l'intérieur de chaque groupe et une grande différence entre les groupes. [34].

Exemples d'application :

- Classer des plantes ou des animaux selon leurs caractéristiques.
- Segmenter des observations d'épicentres pour repérer les zones à risque [36].

5.2.Selon le type d'apprentissage

A. Apprentissage supervisé

C'est un processus dans lequel on utilise plusieurs exemples dont les données d'entrée et les résultats sont déjà connus. À partir de ces exemples, on apprend un modèle qui peut fournir des résultats en utilisant seulement les données d'entrée. Cette technique est utilisée dans les méthodes de classification et de prédiction [37].

B. Apprentissage non supervisé

C'est un processus qui consiste à apprendre un modèle en utilisant toutes les données des instances sans distinguer entre les données d'entrée et les données de sortie. Ce type d'apprentissage est généralement utilisé dans les méthodes d'association et de segmentation [38].

5.3.Selon le caractère du modèle

Il existe un autre critère pour classer les méthodes de data mining, qui se base sur la nature du modèle obtenu après l'analyse des données. Selon le type de modèle construit, on peut distinguer deux grandes catégories de techniques : les modèles prédictifs et les modèles descriptifs [39]:

A. Modèle prédictif

C'est l'utilisation du data mining dans des tâches comme la classification ou la prédiction. Dans ce cas, on applique l'apprentissage supervisé sur des données dont les résultats sont déjà connus afin de construire un modèle capable de prévoir les résultats pour d'autres données.

B. Modèle descriptif

C'est l'utilisation du data mining pour analyser un ensemble de données à l'aide de méthodes comme l'association ou la segmentation. Dans ce cas, on utilise l'apprentissage non supervisé afin d'obtenir des descriptions qui révèlent des informations nouvelles et importantes sur les données.

6. Techniques de Data Mining

Dans ce travail, nous présentons uniquement un ensemble restreint de techniques parmi celles existantes en Data Mining. Ce choix s'explique par leur pertinence par rapport aux objectifs de notre étude

6.1. Les arbres de décision

Les arbres de décision sont des outils qui aident à prendre des décisions. Ils permettent de diviser une population en groupes homogènes en se basant sur des variables importantes et un objectif connu. Ces arbres sont très utilisés pour la classification et la prédiction.

Un arbre de décision utilise les données connues pour faire des prédictions en réduisant progressivement, niveau par niveau, l'ensemble des solutions possibles. Chaque nœud de l'arbre divise les éléments à classer de manière homogène selon une variable discriminante. Les branches représentent les différentes valeurs de cette variable, et les feuilles de l'arbre donnent le résultat final de la prédiction pour les données à classer [31].

La Table 1 présente les différents algorithmes d'arbre de décision :

Nom de l'algorithme	Développeur	Année
CHAID	Kass	1980
CART	Breiman,et al.	1984
ID3	Quinlan	1986
C4.5	Quinlan	1993
SLIQ	Agrawal,et al.	1996
SPRINT	Agrawal,et al.	1996

Table 1 : Les algorithmes d'inductions des arbres de décision [36]

6.2. Les réseaux de neurones

Un réseau de neurones est un modèle de calcul qui s'inspire du fonctionnement des neurones biologiques. Il est composé de nombreuses unités (appelées neurones), chacune ayant une petite mémoire et connectée aux autres par des canaux qui transmettent des données numériques. Chaque neurone ne peut agir que sur ses propres données et sur les informations qu'il reçoit de ses connexions.

Les réseaux de neurones (Fig.2), illustrant un exemple de structure permettent de prédire de nouvelles observations à partir d'observations existantes, après avoir appris à partir d'un ensemble de données. La phase d'apprentissage consiste en un processus itératif où le réseau ajuste ses « poids » pour améliorer ses prédictions sur les données d'entraînement. Une fois cette phase terminée, le réseau est capable de généraliser et de prédire correctement sur de nouvelles données [36].

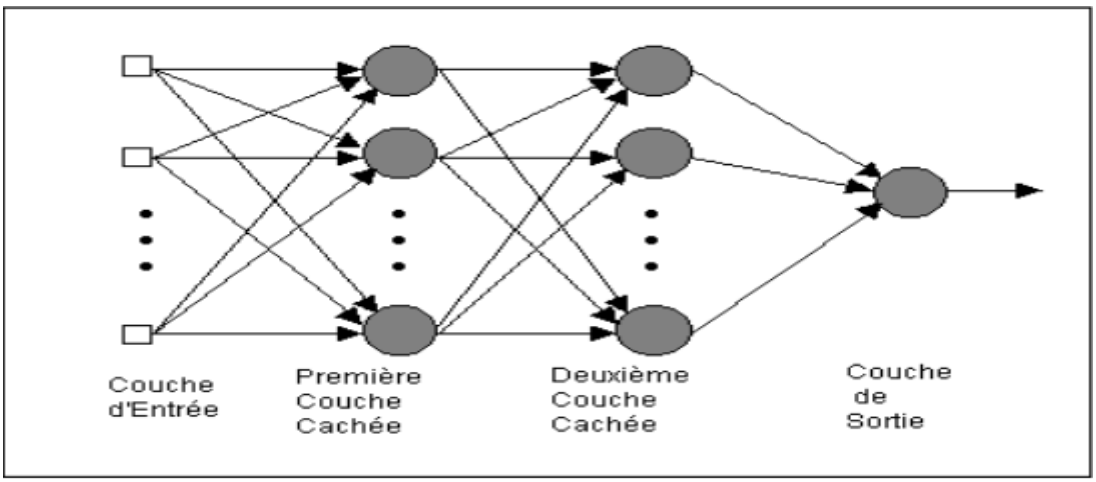

Figure 2: Réseau de neurones.

6.3. Machine à vecteur de support (SVM)

Les machines à vecteurs de support (ou SVM pour Support Vector Machines) sont des modèles d'apprentissage supervisé introduits par Corinna Cortes et Vapnik en 1993 [40].

Elles sont principalement linéaires et binaires, et peuvent être utilisées pour des problèmes de classification ou de régression. Ces modèles sont souvent appliqués dans des domaines comme la reconnaissance de formes ou la bioinformatique [41].

Un modèle SVM représente les données comme des points dans un espace multidimensionnel. Son objectif est de trouver un hyperplan (une « frontière ») qui sépare les points appartenant à différentes catégories en maximisant la marge, c'est-à-dire la distance entre cet hyperplan et les points les plus proches de chaque classe (appelés vecteurs de support).

Pendant la phase d'apprentissage, le SVM analyse les données étiquetées pour déterminer cet hyperplan optimal. Une fois entraîné, il utilise cette frontière pour classer de nouvelles données en fonction de leur position dans l'espace par rapport à l'hyperplan.

6.4. Les k-Plus proches voisins (k-NN)

L'algorithme des k plus proches voisins (k-NN) est une méthode de classification basée sur des exemples connus, et peut aussi être utilisé pour des tâches de régression (estimation). Son principe est de prendre des décisions en se basant sur des cas similaires déjà stockés en mémoire. Cette approche le distingue d'autres méthodes comme les arbres de décision ou les réseaux de neurones.

Contrairement à ces méthodes, le k-NN ne construit pas explicitement un modèle à partir des données d'apprentissage. En réalité, le « modèle » est constitué par l'ensemble des données d'apprentissage elles-mêmes, associé à une fonction de mesure de distance entre les points et à une règle de décision pour déterminer la classe à partir des voisins les plus proches. [42]

6.5. Les k-moyennes (k-means)

L'algorithme K-means est une méthode d'apprentissage non supervisé qui sert à regrouper des données sans étiquettes, c'est-à-dire sans catégories prédéfinies. Son objectif est de trouver des clusters en fonction de la similarité entre les données, le nombre de groupes étant Fixé à l'avance par la valeur de K.

Le fonctionnement est simple : on commence par initialiser K centroïdes, un pour chaque groupe, puis on les ajuste itérativement pour qu'ils représentent au mieux la structure des données. [43]

6.6. La classification Bayésienne

La classification bayésienne regroupe un ensemble de méthodes statistiques fondées sur le théorème de Bayes. Ces classifieurs permettent d'estimer la probabilité d'appartenance d'un individu (ou tuple) à une classe donnée.

On distingue principalement deux approches : le classifieur Naive Bayes, basé sur une hypothèse d'indépendance entre les variables, et les réseaux bayésiens (BayesNet), qui prennent en compte les dépendances entre ces variables.

a) **BayesNet (BN) :** Les réseaux bayésiens constituent un ensemble de méthodes statistiques basées sur les probabilités, utilisées pour le raisonnement et la prise de décision en situation d'incertitude, et reposant fortement sur la règle de Bayes. Ils permettent de modéliser des problèmes, d'extraire de l'information et d'aider à la prise de décision. Représentant un formalisme de raisonnement probabiliste, ils sont de plus

en plus utilisés dans divers domaines tels que l'industrie, la finance et le traitement d'images [44].

b) **Naive Bayes (NB) :** Le classifieur Naïve Bayes est un modèle probabiliste de classification supervisée basé sur le théorème de Bayes. Il repose sur l'hypothèse d'indépendance conditionnelle entre les attributs, ce qui signifie que chaque caractéristique contribue indépendamment à la classification.

Ce modèle calcule la probabilité a posteriori de chaque classe pour une observation donnée et attribue cette observation à la classe ayant la probabilité la plus élevée [45] [46].

6.7. Random Forest (RF)

Le Random Forest (RF) est une technique d'apprentissage supervisé utilisée pour la classification et la régression, basée sur un ensemble d'arbres de décision.

Semblable à la technique du bagging, le RF repose sur la moyenne pour la régression et sur le vote majoritaire pour la classification. Le principe consiste à entraîner des arbres de décision sur différents sous-ensembles de données et sur des variables sélectionnées aléatoirement. [47] [48]

7. Critères d'évaluation des modèles en data mining

Les modèles et les méthodes de data mining sont évalués selon plusieurs critères afin d'assurer leur efficacité et leur utilité dans les applications pratiques. Parmi les principaux critères, on peut citer les suivants [49] [50] :

7.1. La précision de classification (Classification Accuracy)

Ce critère correspond à la capacité d'un modèle à classer correctement de nouvelles données. Il représente généralement le pourcentage d'instances de test qui sont correctement classées par le modèle lors de son évaluation.

7.2. La vitesse (Speed)

La vitesse fait référence au coût computationnel nécessaire pour construire ou utiliser un modèle de classification. Elle permet d'évaluer le temps et les ressources nécessaires lors de l'apprentissage du modèle ou de son application.

7.3. La robustesse

La robustesse correspond à la capacité du modèle à effectuer une classification correcte même lorsque les données contiennent du bruit ou des informations imparfaites.

7.4. La scalabilité (Evolutivité)

Ce critère indique la capacité du modèle à traiter de grands volumes de données. Son évaluation se fait généralement en testant le modèle sur des ensembles de données de taille de plus en plus importante.

7.5. L'interprétabilité

L'interopérabilité représente le niveau de compréhension des résultats fournis par le modèle. Elle indique dans quelle mesure les informations produites par le modèle peuvent être facilement expliquées et comprises.

7.6. La fiabilité (Reliability)

La fiabilité désigne la capacité du modèle de classification à fonctionner correctement dans certaines conditions et pendant une période donnée.

7.7. L'utilité

Ce critère permet d'évaluer dans quelle mesure le modèle de classification peut produire des informations ou des connaissances réellement utiles pour l'utilisateur ou pour la prise de décision.

8. Catégorisation des systèmes du data mining

Les systèmes de data mining peuvent être classés selon différents critères. Parmi les classifications les plus courantes, on peut citer :

- **Selon le type de données à explorer :** Les systèmes sont regroupés en fonction du type de données qu'ils manipulent, comme les données spatiales, les séries temporelles, les données textuelles ou encore celles provenant du World Wide Web.
- **Selon les modèles de données avancés :** Cette classification se base sur le type de structures de données utilisées par le système, par exemple les bases de données

relationnelles, orientées objet, les entrepôts de données (data warehouses) ou les bases de données transactionnelles.

- **Selon le type de connaissance à découvrir :** Les systèmes peuvent être regroupés selon le type de connaissances ou les tâches qu'ils cherchent à extraire, comme la classification, l'estimation, la prédiction ou d'autres formes d'analyse.
- **Selon les techniques d'exploration utilisées :** Cette classification se fait en fonction des méthodes d'analyse employées, telles que la reconnaissance des formes, les réseaux de neurones, les algorithmes génétiques, les méthodes statistiques, la visualisation, ou encore les approches orientées base de données ou data warehouse. [31]

9. Evaluation de la précision des modèles

L'évaluation de la performance des modèles de classification en data mining repose sur plusieurs métriques calculées à partir d'un ensemble de test. Ces ensembles de test proviennent de modèles préalablement entraînés sur des données d'apprentissage, dont la manière de constitution — taille, composition et qualité — influence directement la précision des modèles. Pour limiter l'impact négatif de ces facteurs sur l'évaluation, différentes méthodes d'évaluation ont été développées et certaines se sont révélées particulièrement efficaces en pratique.

9.1.Métriques d'évaluation

Ces mesures et méthodes servent à estimer la performance des modèles de data mining pendant l'évaluation de leur précision et exactitude. Elles permettent également de comparer les résultats obtenus par différents modèles. Chaque métrique possède sa propre stratégie d'estimation, qui se base sur les résultats obtenus lorsque le modèle de classification est appliqué à un ensemble de test.

Dans le cadre d'une classification binaire, les instances de test sont réparties en deux catégories : la catégorie positive, regroupant les instances étiquetées avec la classe d'intérêt, et la catégorie négative, comprenant toutes les autres instances. À partir de cette distinction, quatre indicateurs principaux peuvent être définis pour évaluer la performance du modèle.

- **La matrice de confusion :**

La matrice de confusion (Tab. 2) est une table de statistiques qui englobe tous les indicateurs afin d'offrir une meilleure lisibilité de l'évaluation. A la base de deux catégories positive (+) et négative (-), les lignes de cette structure présentent les statistiques des instances de test selon la distribution réelle, et les colonnes exposent les statistiques des instances de test selon la distribution résultant de la classification

REEL	**Prédiction**	
	Positif	Négatif
Positif	Vrai positif (TP)	Faux négatif (FN)
Négatif	Faux positif (FP)	Vrai négatif (TN)

Table 2 : Matrice de confusion

- **Vrai positif (TP) :** Il s'agit de toutes les instances de test appartenant à la catégorie positive qui ont été correctement classées comme positives lors de l'évaluation. Le nombre de ces instances est indiqué par TP.
- **Vrai négatif (TN) :** Il s'agit de toutes les instances de test appartenant à la catégorie négative qui ont été correctement classées comme négatives lors de l'évaluation. Le nombre de ces instances est indiqué par TN.
- **Faux positif (FP) :** Il s'agit de toutes les instances de test appartenant à la catégorie négative qui ont été incorrectement classées comme positives lors de l'évaluation. Le nombre de ces instances est indiqué par FP
- **Faux négatif (FN) :** Il s'agit de toutes les instances de test appartenant à la catégorie positive qui ont été incorrectement classées comme négatives lors de l'évaluation. Le nombre de ces instances est indiqué par FN [49] [51].

A. La précision (P) : C'est la part des vrais positifs parmi tous les résultats prédits comme positifs (Voir Formule 1) [52] :

$$\boldsymbol{P} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (1)$$

B. Rappel (R) : C'est la part des vrais positifs et correctement classifiés par rapport à toutes les instances réellement positives (Formule. 2). On appelle aussi cela le taux de vrais positifs (TPR) ou la sensibilité.

$$R = \frac{TP}{TP+FN} \quad (2)$$

C. Spécificité (S) : C'est la part des vrais négatifs correctement identifiés par rapport à toutes les instances réellement négatives (Formule. 3). On l'appelle aussi le taux de vrais négatifs.

$$S = \frac{TN}{FP+TN} \quad (3)$$

D. F-Mesure (FM) ou F-Score : C'est une mesure qui combine la précision et le rappel en utilisant une moyenne harmonique (Formule 4), pour donner une idée globale de la performance du modèle.

$$FM = \frac{2*P*R}{P+R} \quad (4)$$

E. Taux de reconnaissance (acc) : Cela indique le pourcentage d'instances correctement classées (vrais positifs + vrais négatifs) par rapport à toutes les données de test (Formule 5).

$$acc = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

F. Taux d'erreur (err) : C'est le pourcentage d'instances mal classées (faux positifs + faux négatifs) par rapport à toutes les données de test (Formule 6).

$$err = \frac{FP+FN}{TP+FP+TN+FN} \quad (6)$$

Il existe d'autres métriques pour la classification binaire [49] [53], mais la classification multi-classes utilise aussi un ensemble de métriques plus générales pour ses évaluations

9.2. Méthodes d'évaluation

A. Holdout : Cette méthode consiste à diviser aléatoirement le dataset en deux parties : un Training set pour l'apprentissage du modèle et un Test set pour mesurer sa performance. Typiquement, deux tiers des données sont utilisés pour l'entraînement et un tiers pour le test. Son principal inconvénient est que la sélection initiale du Training set peut influencer fortement l'évaluation.

B. Sous-échantillonnage aléatoire (Random subsampling) : C'est une variante du Holdout qui répète le partitionnement k fois. La performance globale est calculée comme la moyenne des résultats obtenus lors de ces k itérations. Cette méthode peut présenter des limites si la distribution des instances dans le dataset n'est pas parfaitement indépendante.

C. Cross-Validation (K-Fold) : Le dataset est divisé en k groupes de taille égale. Pour chaque itération, un groupe sert de Test set et les autres constituent le Training set. Parmi les variations de cette méthode :

- **Leave-P-Out Cross-Validation** : P instances sont utilisées pour le test et le reste pour l'entraînement, répété pour toutes les combinaisons possibles. *Leave-One-Out* est le cas particulier avec P = 1.
- **Stratified K-Fold** : Maintient la proportion des classes dans chaque fold.
- **Repeated K-Fold** : Répète le K-Fold plusieurs fois et calcule la performance moyenne pour réduire l'erreur d'estimation.

D. Bootstrap (Échantillonnage avec remise) : Bootstrap (Échantillonnage avec remise) est une méthode de rééchantillonnage qui consiste à tirer aléatoirement, avec remise, N instances à partir d'un même ensemble de données afin de constituer un ensemble d'apprentissage. Cette opération est répétée plusieurs fois. Certaines instances peuvent être sélectionnées plusieurs fois, tandis que d'autres peuvent ne pas apparaître dans l'échantillon d'apprentissage. Les instances non sélectionnées servent généralement à tester le modèle. La variante « 0 .632 Bootstrap » combine l'erreur d'apprentissage et l'erreur de test en pondérant les résultats, en raison du fait qu'environ 63,2 % des observations sont utilisées pour l'entraînement à chaque tirage [49]. Cette méthode calcule le accuracy du modèle par la formule suivante :

$$\boldsymbol{Acc} = \frac{1}{k} \sum_{k=i}^{k} (0.632 \times acc\_Test + 0.368 \times acc\_Train) \quad \textbf{(7)}$$

E. Receiver-operating characteristic (ROC) : Le graphe ROC (Receiver Operating Characteristic) est une technique graphique utilisée pour évaluer et comparer les performances des modèles de classification. Initialement développé pendant la Seconde Guerre mondiale à des fins militaires, cet outil a par la suite connu une large adoption dans plusieurs domaines, notamment en médecine, où il contribue à la prise de décision.

D'un point de vue technique, la courbe ROC représente le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR). Le TPR est placé sur l'axe des ordonnées (axe Y), tandis que le FPR est positionné sur l'axe des abscisses (axe X).

Le taux de faux positifs est défini par la formule suivante :

$$FPR = \frac{FP}{FP+TN} \quad (8)$$

Modèle effectue une classification totalement aléatoire. En revanche, un modèle idéal se traduirait par une courbe ROC passant d'abord par le segment allant de (0,0) à (0,1), puis de (0,1) à (1,1), montrant ainsi une capacité parfaite à distinguer les différentes classes.

L'aire sous la courbe ROC, appelée AUC (Area Under the Curve), correspond à la surface sous cette courbe et permet d'évaluer globalement la performance du modèle. L'AUC prend des valeurs comprises entre 0 et 1. Une valeur de 1 signifie que toutes les classifications sont correctes et que le modèle distingue parfaitement les instances. À l'inverse, une valeur de 0 indique que le modèle classe toutes les instances de manière inversée. Les valeurs supérieures à 0,5 reflètent une capacité correcte du modèle à classer les instances, tandis qu'une valeur de 0,5 correspond à une classification purement aléatoire [54] [55].

La Figure 3 présente un exemple démonstratif de la courbe ROC et de l'espace AUC.

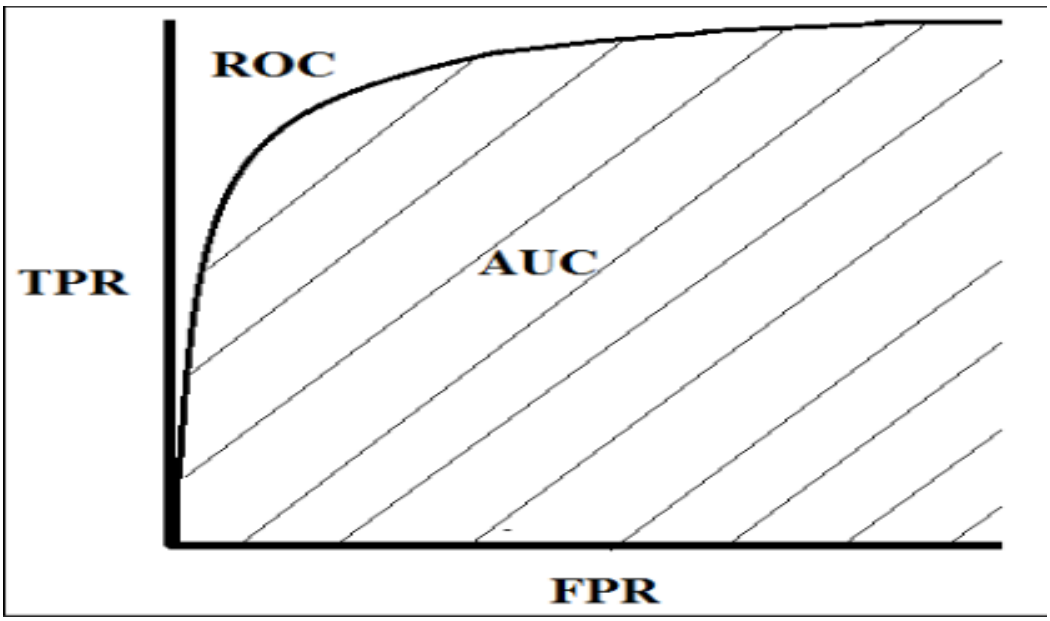

Figure 3 : La courbe ROC et l'aire sous la courbe ROC

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

10.1.Séries temporelles (Time series)

Une série temporelle, ou série chronologique, est une suite ordonnée d'observations numériques d'une même variable (Evénement) au cours du temp. L'agencement des valeurs apporte bien plus que de simples données statistiques : il enrichit la fonction principale grâce à l'évolution temporelle de ces valeurs.

Une série temporelle régulière est composée d'observations enregistrées à intervalles de temps réguliers, tandis qu'une série temporelle irrégulière est constituée d'observations relevées à des intervalles non réguliers.

De plus, une série temporelle univariée n'observe qu'une seule variable, tandis qu'une série temporelle multivariée admet deux variables ou plus. [56]

10.2.Principales tâches en analyse des séries temporelles

A. **Indexation (Requête par contenu) :** Cette tâche consiste à rechercher dans un dataset les séries temporelles les plus similaires à une série donnée considérée comme requête. Pour ce faire, il est nécessaire d'appliquer une mesure de similarité ou de dissimilarité. La comparaison peut se faire sur l'ensemble de la série ou sur des segments à l'aide de fenêtres glissantes. [57]

B. **Clustering :** Le clustering vise à regrouper des séries temporelles présentant des similitudes naturelles selon une mesure de similarité/dissimilarité. Chaque groupe doit contenir des séries homogènes, tout en étant distinct des autres groupes. [57]

C. **Classification :** C'est l'opération qui consiste à attribuer une série temporelle non étiquetée à une classe prédéfinie en utilisant des techniques d'apprentissage supervisé. [57]

D. **Prédiction :** La prédiction permet d'estimer les valeurs futures d'une série temporelle en se basant sur les observations passées.

E. **Caractérisation (Summarization) :** La caractérisation vise à simplifier ou réduire la représentation d'une série temporelle très longue. Elle produit une série approximative de taille réduite tout en conservant les principales caractéristiques de la série originale. [57]

F. Détection des anomalies : Cette tâche consiste à identifier les anomalies dans une série temporelle, qu'il s'agisse d'une sous-séquence, d'une seule valeur ou de la série entière. Elle compare généralement comportement de la série étudiée à celui des séries normales modélisées. Techniquement, il existe plusieurs approches pour la détection d'anomalies :

- Des méthodes basées sur la similarité, entre segments issus d'une segmentation, ou entre structures obtenues par clustering ou encore par des traitements utilisant la technique des plus proches voisins.
- Des méthodes basées sur des modèles prédictifs ou sur la modélisation séquentielle, par exemple avec les Hidden Markov Models (HMM). [57]

G. Segmentation : Cette technique permet de partitionner les séries temporelles en sections représentant des conditions homogènes, afin de les analyser ou de créer des représentations pour d'autres usages. On peut globalement classer cette tâche en trois catégories d'algorithmes. [58]

- **Fenêtres coulissantes (Sliding Windows)** : Il existe une famille d'algorithmes qui construisent un segment en partant du premier point le plus à gauche. Leur fonction principale, exécutée de manière itérative pour générer un nouveau segment, consiste à construire le segment suivant point par point jusqu'à ce que l'erreur produite soit inférieure à un seuil défini initialement. Chaque point de la série est construit à partir du point suivant et prend pour point de départ le premier point le plus à droite du segment précédent.
- **De haut en bas (Top-Down)** : Ces algorithmes fonctionnent en divisant la série temporelle en sous-séries individuelles plus petites de manière récursive jusqu'à ce qu'elles atteignent un seuil d'erreur défini au préalable.
- **De bas en haut (Bottom-Up) :** Ce type d'algorithme commence par diviser la série temporelle en segments très petits, contenant seulement deux points chacun. Ensuite, il calcule le coût pour fusionner chaque segment consécutif. Tant que le coût minimal est inférieur au seuil défini au départ, l'algorithme fusionne ces deux segments. Après chaque fusion, il recalcule les coûts pour la nouvelle séquence et ses segments voisins afin de trouver à nouveau le coût minimal.

11. TF-IDF (Term Frequency-Inverse Document Frequency)

Avec la croissance exponentielle des données, le domaine du text mining a connu un essor considérable afin de faire face au phénomène « d'explosion de l'information », rendant la gestion des données de plus en plus complexe.

La classification représente une approche efficace permettant d'organiser automatiquement des documents non structurés en catégories prédéfinies.

Cette technique facilite non seulement la gestion et l'accès à l'information, mais elle améliore également la compréhension rapide du contenu.

Dans ce contexte, l'approche basée sur le TF-IDF (Term Frequency – Inverse Document Frequency) occupe une place centrale. Elle est largement utilisée pour la pondération des termes dans les documents textuels. Le principe du TF-IDF repose sur deux composantes essentielles : la fréquence d'un terme dans un document (TF) et son importance inversement proportionnelle à sa présence dans l'ensemble des documents (IDF).

Le TF-IDF constitue une base solide pour les systèmes de classification automatique, offrant des résultats précis et fiables dans de nombreuses applications. [59]

11.1. Définition

TF-IDF (Term Frequency–Inverse Document Frequency) est une méthode statistique utilisée dans le traitement automatique du langage naturel et la recherche d'information pour évaluer l'importance d'un mot dans un document par rapport à un ensemble plus large de documents. TF-IDF combine deux composantes : [60]

A. **Fréquence de Terme (TF) :** Mesure la fréquence d'apparition d'un mot dans un document selon la Formule 9. Une fréquence plus élevée suggère une plus grande importance. Si un terme apparaît fréquemment dans un document, il est probablement pertinent par rapport au contenu du document.

$$\boldsymbol{TF(t,d)} = \frac{Nombre\ d'occurrences\ du\ terme\ t\ dans\ le\ document\ d}{Nombre\ total\ de\ termes\ dans\ le\ document\ d} \quad \textbf{(9)}$$

Où

t : terme

d : est un document appartenant au corpus D.

B. **Fréquence Inverse de Document (IDF) :** Réduit le poids des mots communs présents dans plusieurs documents tout en augmentant le poids des mots rares. Si un terme apparaît dans peu de documents, il est plus susceptible d'être significatif et spécifique, et on applique la Formule 10 suivante :

$$\boldsymbol{IDF(t)} = log\left(\frac{N}{df(t)}\right) \quad \textbf{(10)}$$

Où

t : terme

d : est un document appartenant au corpus D.

D : c'est l'ensemble de tous les documents, appelé corpus.

$\boldsymbol{N}$= nombre total de documents dans le corpus D

df (t) : document frequency (nombre de documents contenant le terme t)

Pour effectuer le classement ou l'évaluation dans un corpus, on utilise la Formule 11 de TF-IDF pour mesurer l'importance d'un élément dans un document par rapport à l'ensemble du corpus.

$$\textbf{TF-IDF}(\boldsymbol{t,d}) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t) \quad (11)$$

11.2. Mise en œuvre TF-IDF

Avant d'appliquer la technique TF-IDF, il est important de préparer les textes de manière à faciliter le calcul des poids. Cela passe par plusieurs étapes de traitement et de nettoyage. [59]

A. **Tokenisation :** La plupart des recherches en text mining utilisent les mots ou les phrases comme unité de traitement. Pour traiter les textes, il faut découper les phrases mot par mot et éliminer la ponctuation inutile, afin de simplifier le calcul des poids par la suite.

B. **Création de N-grammes :** Pour enrichir le dictionnaire et améliorer la précision, certaines approches créent également des n-grammes, c'est-à-dire des termes composés de plusieurs mots consécutifs.

C. **Suppression des doublons :** Tous les mots dupliqués dans le lexique sont supprimés pour réduire le temps de traitement et simplifier le calcul des poids.

D. **Suppression des mots vides (Stop Words) :** La suppression des mots vides est une technique courante en recherche d'information. Elle consiste à éliminer :

- Les mots très fréquents ou, dans certains cas, très rares,
- Les mots sans signification spécifique, par exemple : En anglais : “the”, “a”, “an”

E. **Filtrage selon la fréquence des termes :** Le filtrage selon la fréquence des termes est une technique utilisée en traitement automatique du langage naturel pour réduire la dimension du lexique. Elle consiste à éliminer les mots dont la fréquence d’apparition est trop faible dans le corpus.

F. **Suppression manuelle des mots invalides :** Après la génération des n-grammes, certains termes non pertinents ou invalides peuvent subsister. Une vérification manuelle est alors effectuée afin d’éliminer ces éléments indésirables.

G. **Implémentation TF-IDF :** Une fois le lexique finalisé, le poids de chaque terme est calculé à l’aide de la technique TF-IDF, citée précédemment (Formule 11).

12. Conclusion

Le Data Mining joue aujourd’hui un rôle central pour transformer de simples données en connaissances utiles, permettant aux organisations de mieux comprendre leur environnement et de prendre des décisions éclairées. En s’appuyant sur des méthodes analytiques, statistiques et informatiques, il met en lumière des tendances et des relations qui ne seraient pas visibles autrement, soutenant ainsi l’innovation et l’amélioration des processus internes.

Dans des secteurs sensibles comme la santé, ces techniques aident à affiner les diagnostics, à anticiper certains événements et à rendre les traitements plus efficaces, tout en garantissant des résultats fiables et de qualité.

La variété des applications et des systèmes de Data Mining montre à quel point ce domaine est riche et complexe. Maîtriser ces outils constitue un véritable atout pour valoriser les informations disponibles et développer des solutions adaptées aux besoins spécifiques des différents secteurs.

CHAPITRE 3

Classification des patients basée sur la représentation TF-IDF

Chapitre 3 Classification basée sur la représentation TF-IDF

1. Introduction

Avec la transformation numérique du secteur de la santé, les dossiers de santé électroniques (Electronic Health Records – EHR) occupent aujourd'hui une place centrale dans la gestion des informations médicales. Ces dossiers permettent de stocker, organiser et partager les données des patients de manière structurée, incluant les antécédents médicaux, les diagnostics, les traitements et les résultats d'examens.

L'exploitation de ces données représente une opportunité majeure pour améliorer la qualité des soins et faciliter la prise de décision médicale. En effet, les EHR offrent un volume important de données hétérogènes qui peuvent être analysées à l'aide de techniques d'intelligence artificielle, notamment les méthodes de classification, afin d'identifier des schémas pertinents et de prédire l'évolution de certaines maladies.

Cependant, l'utilisation des EHR pose plusieurs défis. La diversité des données (numériques, catégorielles, temporelles), leur complexité ainsi que la présence de données manquantes ou bruitées rendent leur traitement difficile. De plus, le choix des méthodes de représentation et d'analyse adaptées est crucial pour garantir des résultats fiables.

Dans ce contexte, l'intégration de techniques de prétraitement avant la classification devient essentielle pour exploiter efficacement ces données. L'objectif est de transformer les EHR en une source d'information exploitable permettant de soutenir les professionnels de santé dans leurs décisions, tout en améliorant la précision et la rapidité des analyses

2. Études antérieures

Plusieurs travaux de recherche se sont intéressés à l'exploitation des dossiers de santé électroniques (EHR) afin d'améliorer la prise de décision médicale à l'aide des techniques d'apprentissage automatique.

L'étude de Jamaluddin & Wibawa (2021) s'intéresse à l'exploitation des dossiers médicaux électroniques (EMR), qui représentent une source importante d'informations dans le domaine de la santé, notamment pour l'aide à la décision clinique (CDSS). Les EMR contiennent des données hétérogènes et souvent sous forme de texte non structuré, ce qui rend l'extraction de l'information plus complexe. Dans ce travail, les auteurs ont proposé une approche de classification des diagnostics à partir des EMR indonésiens, en se concentrant sur plusieurs maladies à forte prévalence, telles que la tuberculose, le cancer, le diabète, l'hypertension et les maladies rénales chroniques. La méthode TF-IDF a été utilisée pour l'extraction des

caractéristiques à partir des données textuelles, suivie par le classifieur SVM, connu pour son efficacité dans le traitement des données de grande dimension.Les auteurs ont également étudié l'impact de différents noyaux du SVM (linéaire, polynomial, RBF et sigmoïde), ainsi que l'effet de la suppression des mots vides (stopwords). Les résultats obtenus montrent que la combinaison TF-IDF et SVM permet une classification efficace, avec une amélioration des performances après suppression des stopwords. Les meilleures performances sont obtenues avec le noyau sigmoïde, atteignant une accuracy de 91.03 %, contre 89.91 % pour le noyau linéaire, 90.58 % pour le noyau polynomial et 90.75 % pour le noyau RBF.

L'Organisation mondiale de la santé (OMS) a identifié la COVID-19 comme une nouvelle pandémie en mars 2020. Ce virus mortel s'est propagé dans le monde entier, touchant de nombreux pays. De grandes quantités de données utiles ont été fournies par les plateformes de réseaux sociaux comme Twitter concernant cette épidémie, afin d'aider à l'évaluation des prises de décision liées à la santé. Rashmi et al (2023) ont appliqué cette étude pour prédire l'apparition de la maladie et fournir des alertes précoces, l'analyse des attitudes des utilisateurs a été réalisée à l'aide de techniques efficaces d'apprentissage profond (deep learning). Les tweets collectés ont été prétraités et classés en trois catégories : neutre, positif et négatif. Dans une deuxième étape, différentes caractéristiques ont été extraites des textes à l'aide de plusieurs techniques populaires afin de construire le jeu de données de caractéristiques, notamment TF-IDF, FastText et GloVe. La méthode TF-IDF implémentée est combinée avec des caractéristiques sémantiques (GloVe et FastText) afin de décrire plus précisément les publications et d'améliorer le processus de classification en utilisant des méthodes d'apprentissage profond. La performance de la méthode TF-IDF-CNN proposée est évaluée à l'aide d'un jeu de données Twitter. La méthode implémentée a obtenu de meilleures performances que les méthodes existantes de détection, notamment NLP-BiLSTM, TF-IDF simple, SVM et CF. Les résultats obtenus sont : 96 % de précision (accuracy), 0,89 de F-mesure, 0,93 de précision (precision) et 0,90 de rappel (recall).

L'étude de Pulluri et al (2023) porte sur l'exploration des dossiers médicaux électroniques (EHR) à l'aide des techniques de traitement du langage naturel (NLP) afin d'améliorer la prise de décision clinique, notamment pour la prédiction des réadmissions non planifiées. Les auteurs ont utilisé plusieurs méthodes, telles que TF-IDF pour l'extraction des termes, Word2Vec pour la représentation sémantique, NER pour l'identification des entités médicales, ainsi que LDA pour la détection des thématiques cachées. Les résultats obtenus montrent que TF-IDF offre de bonnes performances pour l'extraction des termes importants (précision = 0.85, rappel = 0.92),

tandis que NER permet une identification efficace des entités médicales avec un F1-score élevé. Par ailleurs, Word2Vec capture les relations sémantiques entre les termes, et LDA permet de révéler des structures thématiques dans les données.

L'étude de Nadya Mumtazah et al (2024) utilise des techniques de prétraitement du texte pour nettoyer et extraire des informations pertinentes à partir des dossiers médicaux. Les méthodes Bag of Words et TF-IDF ont été utilisées pour transformer le texte en caractéristiques numériques adaptées à la classification. Ils ont comparé les performances de deux algorithmes de classification : Naïve Bayes et K-Nearest Neighbors (KNN), chacun combiné avec Bag of Words et TF-IDF. Les résultats montrent que la combinaison KNN et Bag of Words a obtenu la meilleure précision avec 75 %, suivie de Naïve Bayes avec Bag of Words par 74 %. De plus, Naïve Bayes avec TF-IDF et KNN avec TF-IDF ont atteint une précision de 73 %. Cette classification vise à aider les médecins en fournissant une vue d'ensemble des maladies des patients.

À l'échelle mondiale, les registres de cancer traitent et encodent de grands volumes de rapports en texte libre afin d'assurer la surveillance des cancers et des tumeurs. Dans ce contexte, l'étude de Shivam Kalra et al (2019) vise à utiliser des techniques d'apprentissage automatique afin de prédire ou d'identifier le diagnostic principal (basé sur les codes ICD-O) à partir des rapports de pathologie. Des expériences ont été réalisées en extrayant des caractéristiques TF-IDF à partir des textes, puis en les classant à l'aide de trois algorithmes : SVM, XGBoost et régression logistique. Un nouveau jeu de données a été construit, comprenant 1 949 rapports de pathologie répartis en 37 catégories ICD-O, provenant de quatre sites principaux : poumon, rein, thymus et testicule. Les résultats montrent que la meilleure performance est obtenue avec XGBoost combiné à TF-IDF, atteignant une précision de 92 %, tandis que le SVM linéaire obtient 87 %. L'étude confirme également que TF-IDF permet de mettre en évidence les mots-clés importants des rapports, ce qui est utile pour la recherche sur le cancer et le processus de diagnostic.

L'objectif de l'étude de Michael et al. (2022) était d'intégrer les notes cliniques issues des dossiers médicaux électroniques (EMR) dans des modèles prédictifs de réhospitalisation à 30 jours (30-day rehospitalization) après une transplantation rénale. Il s'agit d'une étude rétrospective observationnelle portant sur des patients ayant reçu leur première greffe de rein dans un grand hôpital urbain du sud-est des États-Unis entre janvier 2005 et décembre 2015, en utilisant à la fois des données structurées (EMR) et non structurées (notes cliniques). Des

techniques de traitement automatique du langage naturel (NLP) ont été appliquées à huit types de notes cliniques afin d'extraire de nouvelles caractéristiques prédictives potentielles de la réhospitalisation à 30 jours après transplantation rénale. Ces caractéristiques ont ensuite été intégrées dans des modèles prédictifs construits à l'aide d'approches d'apprentissage automatique non supervisé et de techniques de fouille de texte basées sur la méthode TF-IDF (Term Frequency–Inverse Document Frequency). Plusieurs modèles prédictifs ont été développés, incluant des modèles basés uniquement sur les données structurées ainsi que des modèles combinant les données structurées et les notes cliniques. L'aire sous la courbe (c-statistic) a été utilisée pour évaluer et comparer la performance des modèles, et une validation croisée à 5 plis (5-fold cross-validation) a été utilisée pour tester leur performance.

Rupa et Ramani (2025) proposent une approche hybride complète pour résoudre les problèmes liés au résumé automatique de textes médicaux, en combinant des méthodes extractives et abstractive. Cette approche intègre la fréquence des termes et la fréquence inverse de documents (TF-IDF), issue du traitement automatique du langage naturel (NLP), ainsi que le modèle AutoModelForSeq2SeqLM basé sur les grands modèles de langage (LLM). Les performances de cette approche sont comparées à celles de méthodes existantes telles que BERT, TextRank, K-means, BART-Large-CNN de Facebook et GPT-2, en utilisant les métriques ROUGE-1, ROUGE-2 et ROUGE-L. Les résultats expérimentaux montrent que l'approche hybride surpasse les autres méthodes existantes. Le résumé automatique de textes médicaux permet ainsi d'extraire des informations importantes à partir de grands documents médicaux. Ce travail combine TF-IDF et AutoModelForSeq2SeqLM afin de produire de meilleurs résumés, avec des performances supérieures à celles de BERT et GPT-2 selon les scores ROUGE.

L'étude de Kumar et al. (2026) pour l'analyse des sentiments consiste à identifier et à catégoriser les émotions exprimées dans les avis et les contenus écrits publiés sur les sites web, en utilisant des technologies d'analyse de texte. Des recherches antérieures ont montré que l'analyse des sentiments dans les avis pharmaceutiques peut fournir des informations précieuses permettant aux organisations et aux professionnels de santé d'évaluer la sécurité des médicaments après leur mise sur le marché. Ces informations protègent les patients et renforcent leur confiance envers les prestataires de soins de santé. Les avis sont annotés à l'aide de deux lexiques d'émotions complets, SenticNet et TextBlob. Des techniques d'ingénierie des caractéristiques telles que TF (Term Frequency) et TF-IDF sont utilisées pour extraire les caractéristiques importantes. Enfin, les tâches de classification sont réalisées à l'aide de

modèles d'apprentissage automatique et de modèles d'apprentissage profond adaptés à la littérature biologique. Des métriques de performance sont utilisées pour évaluer l'efficacité de cette méthodologie combinée. Les résultats expérimentaux montrent que l'hybridation d'un modèle lexical et d'un modèle d'apprentissage médical basé sur les transformeurs produit de meilleures performances que l'utilisation de chaque méthode séparément. De plus, TextBlob affiche de très bonnes performances, atteignant 97 % de précision avec une combinaison LSTM et CNN. Un autre modèle, le transformeur médical BioBERT, obtient 95 % de précision avec TF et la régression logistique sur un jeu de données d'avis de médicaments. TextBlob atteint également 94 % de précision lorsqu'il est associé à TF et au modèle LSTM, et 97 % de précision lorsqu'il est combiné avec le modèle BioBERT basé sur les transformeurs sur un jeu de données issu de tweets.

3. Modèle proposé

On suppose l'ensemble P des n patients, tel que $P = \{P_1, P_2, \ldots, P_n\}$, et l'ensemble E des m événements médicaux, tel que $E = \{E_1, E_2, \ldots, E_m\}$

Les données des patients dans P regroupent les observations collectées à partir de l'ensemble E d'événements médicaux, comme la température ou le poids corporel. Ces observations peuvent être sous forme de « séries temporelles » (par exemple l'évolution d'une mesure dans le temps pour un patient donné) ou sous forme « non temporelle » (comme le sexe ou l'âge). La Figure 4 illustre le processus proposé pour la classification des données patients. Ce processus prend en compte quatre types de données : « numériques, nominales, Date/Time et booléennes ».

Le développement de cette approche s'organise selon deux axes principaux :

- Le premier vise à « construire le modèle de classification »,
- Le second concerne « la préparation des données et la classification des nouveaux patients ».

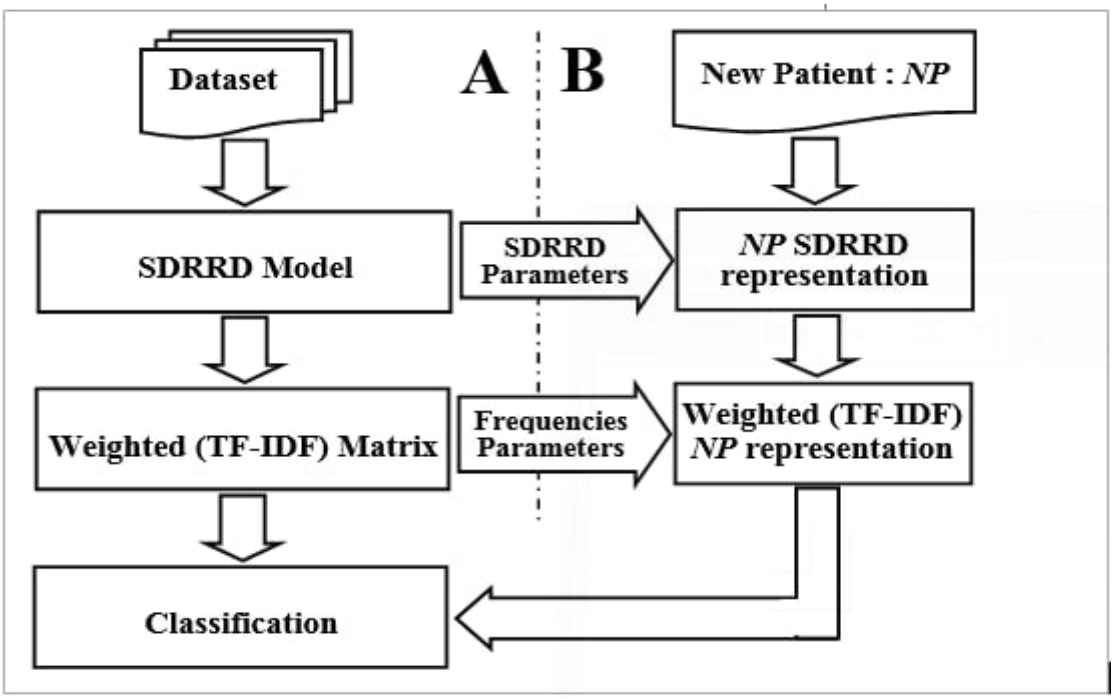

Figure 4 : Processus de classification

3.1. Construction du modèle de classification

Il s'agit de la tâche consistant à produire et entraîner un modèle de classification à partir de l'ensemble des données des patients. Elle se déroule en trois phases successives.

A. Préparation des données : C'est une phase de prétraitement et de représentation des données des patients issus de l'ensemble d'apprentissage. Étant donné que l'on traite des données structurées provenant des EHR (Electronic Health Records), celles-ci peuvent être sous forme temporelle ou non temporelle, et de types numérique, nominal, Date/Time ou booléen.

Nous utilisons le modèle proposé par Kadi et al. (2021) [61]. Pour le traitement et la préparation des données numériques, Date/Time, nominales et booléennes. Ce modèle est reconnu pour son efficacité dans le traitement des séries temporelles ainsi que pour sa robustesse face à la perte d'information lors de la représentation. Il se compose de deux phases principales : la représentation des données par régions (Data Representation by Region DRR) et la représentation des données par dispersion (Data Representation by Dispersion DRD). Globalement, ce modèle produit trois types de représentations, par valeurs, binaire et par symboles. Dans notre cas nous utilisons seulement la représentation symbolique.

La phase DRR est destinée au traitement des données numériques et Date/Time. En calculant l'âge des observations, le modèle transforme les données de type Date/Time en données numériques. Ensuite, les événements numériques sont traités par régions selon des techniques de clustering. L'étape de notification transforme chaque événement numérique E_j en un tableau T_j, où le nombre de colonnes dépend directement du nombre de régions (clusters) associées à cet événement ainsi que de la taille maximale des séries correspondantes. Tous les tableaux obtenus passent ensuite par une étape de linéarisation. Enfin, une opération de jointure permet de fusionner ces tableaux en une seule table regroupant toutes les représentations des événements numériques.

La phase DRD traite les données nominales et booléennes, selon un processus similaire à celui de la première phase. Les données booléennes sont d'abord converties en données nominales : les valeurs « 1 » sont remplacées par « Y » et les valeurs « 0 » par « F ». Cette phase considère la liste $L_i = \{O_{ij}\}$ des différentes observations O_{ij} associées à chaque événement nominal E_i. Elle construit un tableau T_i pour chaque événement, avec n lignes (correspondant aux patients) et $C_i = |L_i| \times |$*la plus longue série de* $E_i|$ colonnes, ce qui correspond à une représentation par dispersion. Les cellules de chaque tableau sont remplies en fonction des observations et de leur ordre d'apparition. Comme dans la phase précédente, tous les tableaux T_i sont ensuite fusionnés en une seule table finale par jointure.

Les deux phases DRR et DRD incluent une étape de marquage dans la table finale, qui peut se faire soit par valeur réelle, soit par présence, soit par symbole. Dans ce travail, seule la représentation par symbole est utilisée.

La dernière étape du modèle DRRD consiste à regrouper les deux représentations symboliques obtenues à partir de DRR et DRD en une seule représentation globale appelée SDRRD (Symbolic Data Representation by Region and Dispersion). En revanche, tous les paramètres de ce modèle doivent être conservés afin de pouvoir les réutiliser pour représenter les données de nouveaux patients.

Puisque la représentation SDRRD traite les observations une par une dans le temps (Formule 12), on l'appelle SDRRD par temps (SDRRD per Time -SDRRDT). Pour analyser l'impact de la manière dont les symboles sont organisés sur la tâche de classification, on utilise également une deuxième forme de représentation. Cette deuxième représentation est une réorganisation de la SDRRD par événement (SDRRD

per Event -SDRRDE). À partir de la table SDRRD, et pour chaque patient P_i, on regroupe (concatène) les symboles de représentation r_{ju} correspondant à chaque instant t_u d'un événement E_j dans une seule chaîne de représentation notée R_{ij} (Formule 13).

$$\forall R_{ij} \in SDRRDT\ (0 \leq i < |P|\ and\ 0 \leq j < NbColumnes(SDRRDT)\ ) \Rightarrow \begin{cases} r_{ij}=null. \\ Or \\ r_{ij}\ est\ un\ symbole. \end{cases} \quad (12)$$

$$\forall R_{ij} \in SDRRDE\ (0 \leq i < |P|\ Et\ 0 \leq j < |E|\ ) \Rightarrow \begin{cases} R_{ij}=null. \\ Ou \\ R_{ij}\ est\ une\ série\ de\ symboles. \end{cases} \quad (13)$$

	SDRRDT representation of time E_0												***SDRRDE representation of event E_0***	***Symbols set***
	t_0			***t_1***			***t_2***			***t_3***			**R_{ij} (By concatenation)**	
P_0	$r_{00}=a$				$r_{01}=b$		$r_{02}=a$				$r_{03}=b$		$R_{00}=r_{00}\ r_{01}\ r_{02}\ r_{03}$ $\Rightarrow R_{00}=abab$	$S_0=\{a,b,c\}$
P_1	$r_{00}=a$			$r_{01}=a$									$R_{10}=r_{00}\ r_{01}$ $\Rightarrow R_{10}=aa$	
P_2		$r_{00}=b$			$r_{01}=b$				$r_{02}=c$				$R_{20}=r_{00}\ r_{01}\ r_{02}$ $\Rightarrow R_{20}=bbc$	

Table 3 : Exemple illustratif des représentations SDRRDT et SDRRDE.

B. Génération de la matrice de pondération : Il s'agit d'une phase de transformation des représentations symboliques (qualitatives) obtenues en représentations quantitatives. Quelle que soit la représentation utilisée, SDRRDT ou SDRRDE, et afin de mesurer l'importance de chaque représentation R (R désigne r_{ju} ou R_{ij}) pour chaque patient P_i par rapport à l'ensemble des représentations des patients, nous calculons les poids Term Frequency–Inverse Document Frequency (TF-IDF) selon la formule suivante :

$$\boldsymbol{TF - IDF(R, P_i) = Occ(R, P_i) * Log \frac{|P|}{|P|_R}} \quad \textbf{(14)}$$

Où $|P|$ désigne le nombre de patients et Occ *(R, P_i)* représente le nombre d'occurrences de la représentation R pour le patient P_i. Étant donné que chaque colonne est différente des autres dans les deux matrices de représentation SDRRDT et SDRRDE.

Nous avons toujours Occ (R, P_i) = 1 pour R non null et Occ *(R, P_i)* = 0 dans le cas contraire (Formule 15).

$$\begin{cases} \forall\, R \in SDRRDT \\ Ou \\ \forall\, R \in SDRRDE \end{cases} \Rightarrow \begin{cases} Occ(R, P_i) = 0\ if\ (R = null). \\ Et \\ Occ(R, P_i) = 1\ if\ (R\ est\ un\ symbole)\ Or \begin{pmatrix} R\ est\ une\ série\ de \\ symboles \end{pmatrix}. \end{cases} \quad (15)$$

$|P|_R$ représente le nombre de patients (la fréquence) ayant la même représentation R dans la même colonne, c'est-à-dire au même instant d'un même événement pour SDRRDT, ou au niveau du même événement pour SDRRDE.

Les deux matrices de poids générées sont finalement étiquetées en ajoutant une colonne contenant les classes des patients. Le résultat donne deux matrices de poids étiquetées, la première pour le cas de SDRRDT (Labeled Weight SDRRT -LWT), tandis que la deuxième pour le cas de SDRRDE (Labeled Weight SDRRE -LWE).

C. Classification : Cette étape traite le résultat de la phase précédente à l'aide d'un algorithme de classification et permet d'entraîner un classifieur C applicable à de nouveaux patients. Afin de choisir le meilleur algorithme de classification sur lequel le classifieur C sera basé, plusieurs techniques de classification seront testées BayesNet (BN), Machine à vecteur de support (SVM), k-Plus proches voisins (k-NN) avec k=3 (3-NN) et Random Forest (RF).

3.2. Classification des nouveaux patients

Cette tâche utilise le modèle produit lors de la première phase ainsi que ses paramètres pour la préparation et la classification des nouveaux patients.

A. Préparation des données du nouveau patient : Elle consiste à utiliser les paramètres du modèle SDRRD construits et sauvegardés lors de la première phase de la première tâche pour générer la représentation symbolique SDRRD du nouveau patient. Selon les mêmes étapes, on obtient deux représentations : le SDRRD du nouveau patient par temps (New SDRRDT -NSDRRDT) et le SDRRD du nouveau patient par événement (New SDRRDE -NSDRRDE).

B. Génération des poids de représentation du nouveau patient : Les opérations de cette phase sont similaires à celles de la deuxième phase de la première tâche en utilisant la Formule 14, mais cette fois-ci nous utilisons les fréquences $|P|_R$ issues de l'ensemble d'apprentissage. Le résultat est une représentation SDRRD du

nouveau patient par temps (New LWT -NWT) et par événement (New LWE -NWE). Cette phase exclut l'ajout de l'information de labellisation, car c'est l'objectif de notre modèle.

C. **Classification du nouveau patient :** Il s'agit de la phase finale du processus appliqué au nouveau patient. Elle consiste à utiliser le classifieur C construit lors de la première phase afin de classifier et déterminer la catégorie des nouveaux patients sur la base de leurs représentations NWT et NWE.

4. Expérimentation

Le dataset utilisé dans cette expérimentation provient de la base de données Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu). Lancée en 2003, l'ADNI est une initiative publique-privée dirigée par le Dr Michael W. Weiner. Son objectif principal est l'étude de différents types d'imagerie médicale, notamment l'IRM, ainsi que l'analyse de biomarqueurs et d'évaluations cliniques et neuropsychologiques, afin de mieux comprendre la progression du trouble cognitif léger (MCI) et de la maladie d'Alzheimer à un stade précoce. [62]

Le dataset, en particulier la table ADNIMERGE_May15, contient 90 attributs. Parmi les plus importants on trouve l'âge des patients (AGE), des variables issues de l'imagerie PET telles que FDG, PET et AV45, ainsi que des indicateurs cliniques comme CDRSB, ADAS11 et MMSE. On retrouve également des informations démographiques et génétiques comme PTETHCAT, PTEDUCAT et APOE4, ainsi que d'autres mesures telles que le volume de l'hippocampe au baseline (Hippocampus_bl) ou les performances au test de mémoire RAVLT_immediate.

Les patients sont classés en cinq catégories : Cognitively Normal (CN), Alzheimer's Disease (AD), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI) et Significant Memory Concern (SMC). À partir de ces classes, un échantillon de 500 patients a été sélectionné aléatoirement, soit 100 patients par classe, afin d'évaluer le modèle proposé.

Lors de la première phase, la transformation génère 87616 observations (75428 numériques et 12188 nominales), et chaque observation possède un seul attribut dépendant d'un événement médical donné.

Les deux matrices de représentation « SDRRDT et SDRRDE » générées par cette phase comportent respectivement 1393 colonnes et 82 colonnes.

La phase de génération des matrices de poids produit deux matrices de type double. Après l'étape d'étiquetage, on obtient les deux matrices étiquetées « LWT et LWE », composées respectivement de 1394 colonnes et 83 colonnes.

5. Résultats et discussions

Nous utilisons les mesures suivantes : Kappa-Statistics (KS), F-Measure, (FM) et l'aire sous la courbe ROC (AUC) pour évaluer notre modèle de classification. La Table 4 présente les résultats obtenus.

	KS		FM		AUC	
Classifieur	**Per Event**	**Per Time**	**Per Event**	**Per Time**	**Per Event**	**Per Time**
BN	0,7850	0,8175	0,8242	0,8515	0,9733	0,9704
SVM	0,8450	**0,9825**	0,8760	**0,9860**	0,9608	0,9958
3-NN	0,7775	0,9150	0,8193	0,9314	0,9411	0,9842
RF	0,8925	0,9725	0,9140	0,9779	0,9888	**0,9981**

Table 4 : Résultats d'évaluation de la classification.

Pour analyser les performances obtenues, nous traçons les courbes correspondantes des trois mesures d'évaluation.

La Figure 5 présente le KS sous forme de courbes. Selon le type de performance, les résultats de la classification par temps sont meilleurs que par événement. Ainsi, les résultats des méthodes SVM et RF sont supérieurs ou égaux à 0,9725 pour le KS par temps. En revanche, aucun classifieur n'a atteint cette performance dans le cas par événement, et le meilleur résultat est de 0,8925 obtenu par la technique RF.

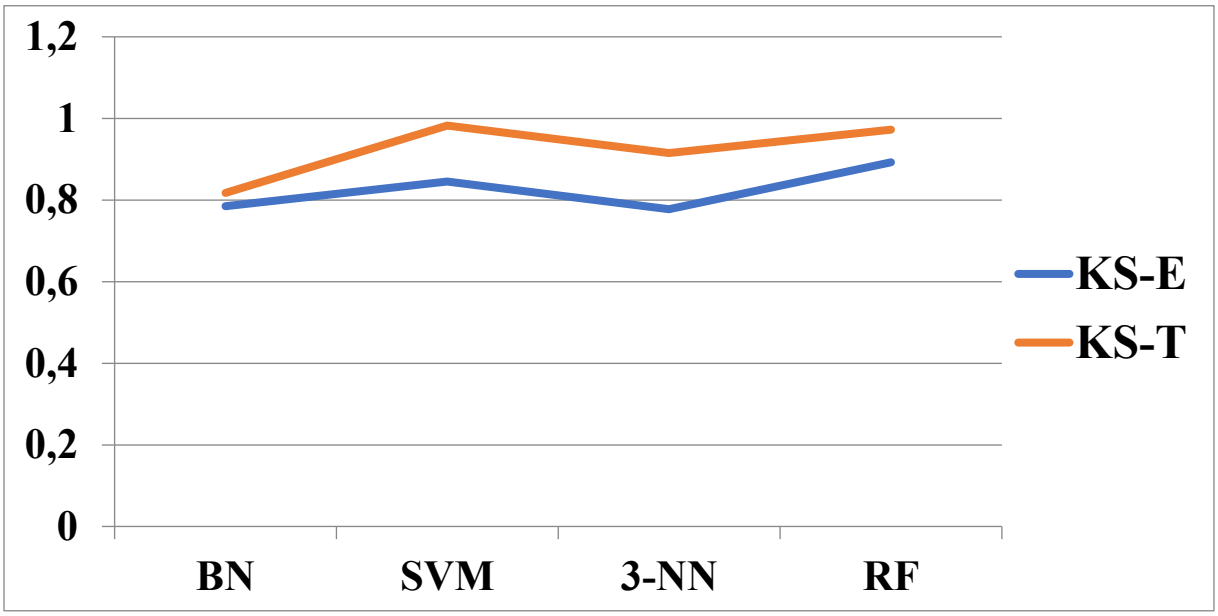

Figure 5 : Courbe d'évaluation de la Kappa-Statistics

Dans le même cadre expérimental, les deux courbes d'évaluation de FM présentées dans la Figure 6 suivent le même comportement. La courbe correspondant à la classification basée sur la représentation par temps est meilleure que celle basée sur la représentation par événement. Cette fois, à l'exception du BN, les classifieurs ayant des performances FM ≥ 0,9 sont SVM (FM = 0,9860), 3-NN (FM = 0,9314) et RF (FM = 0,9779) avec le cas par temps, et seul le classifieur RF dépasse ce seuil pour le cas par événement (FM = 0,9140).

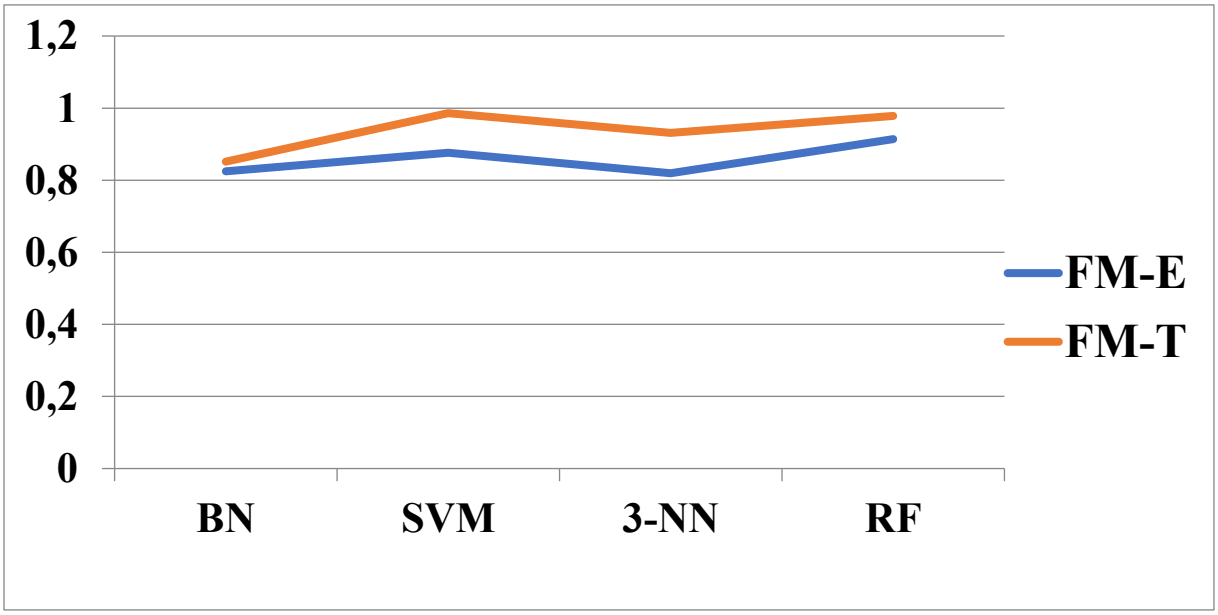

Figure 6 : Courbe d'évaluation de la F-mesure

La Figure 7 montre l'évaluation de la mesure AUC. Contrairement aux cas précédents, on observe de bonnes performances (AUC ≥ 0,9) quel que soit le classifieur appliqué aux deux cas par temps et par événement. Les performances obtenues par temps restent globalement meilleures que par événement, à l'exception du classifieur BN, qui donne de meilleurs résultats avec la représentation par événement (AUC = 0,9733).

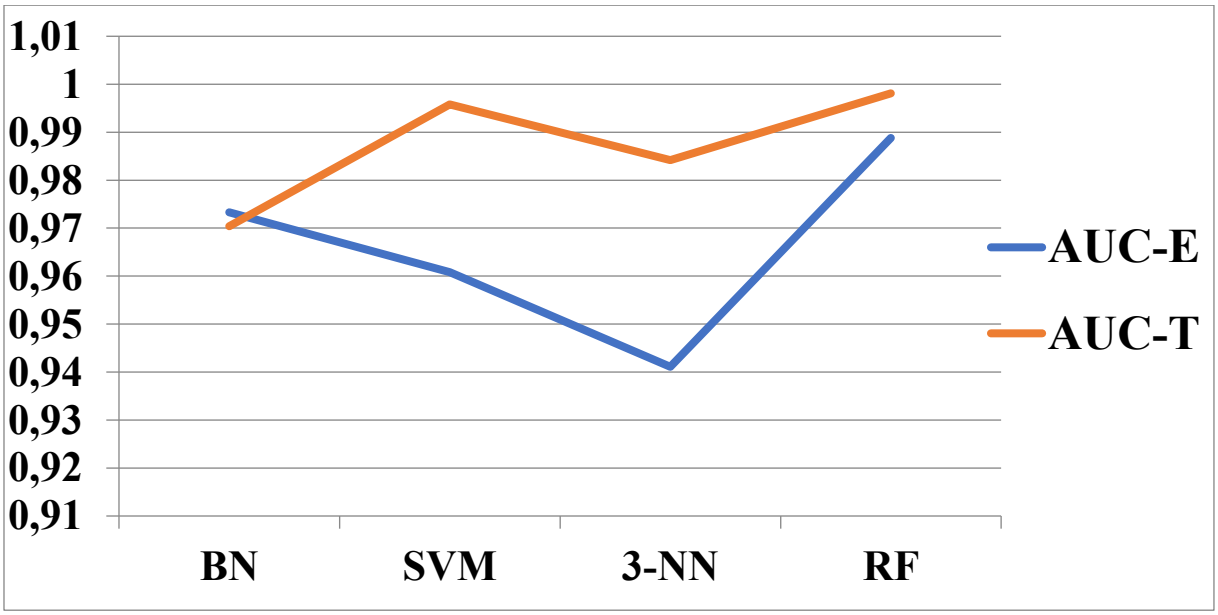

Figure 7 : Évaluation de l'AUC

Ces trois évaluations nous amènent à choisir la représentation par temps SDRRDT comme outil de représentation des données dans notre modèle.

Contrairement à l'évaluation précédente, nous essayons de sélectionner le meilleur classifieur en fonction des performances obtenues. La Figure 8 présente la comparaison entre ces classifieurs sous forme d'un histogramme explicatif. Nous nous concentrons sur la classification basée sur la représentation par temps, car elle donne toujours les meilleures évaluations.

Sur les trois cas des mesures KS-T, FM-T et AUC-T, les classificateurs NB et 3-NN sont les mauvais par rapport à tous les autres. De plus, le classifieur SVM obtient les meilleures évaluations pour KS-T et FM-T, tandis que le classifieur RF présente la meilleure performance en AUC-T.

Bien que la différence de mesure AUC-T entre les classifieurs SVM et RF soit presque négligeable (AUC = 0,9958 pour SVM et AUC = 0,9981 pour RF), nous considérons la

technique SVM comme le meilleur classifieur à appliquer. Ce choix est justifié par les meilleures performances obtenues par SVM selon les mesures KS et FM.

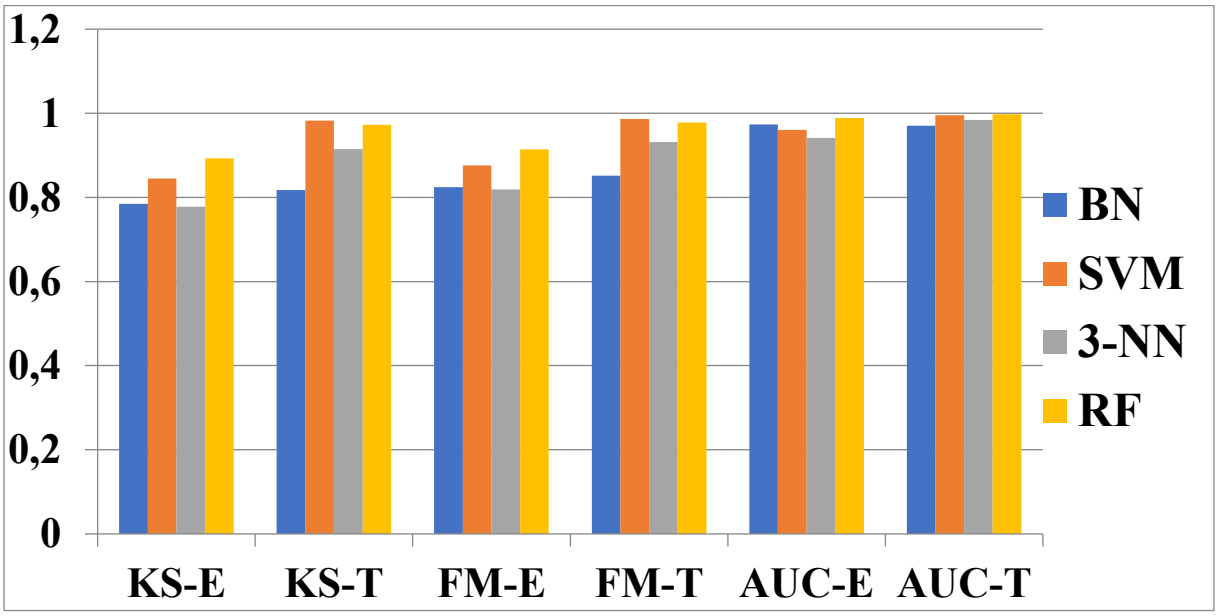

Figure 8 : Évaluation des performances des classifieurs.

Finalement, nous constatons que notre contribution touche huit points importants :

- **Expansion des données :** la représentation par région pour les données numériques et par dispersion pour les données nominales implique une expansion des données afin de former une nouvelle représentation plus riche.
- **Maximisation du nombre d'attributs traités :** traitement de tous les attributs sans contrainte, même ceux contenant un grand nombre de valeurs nulles.
- **Minimisation de la perte d'information :** réduction au maximum de la perte de données et d'informations, notamment pour les séries temporelles, grâce à l'utilisation du modèle DRRD pour la représentation des données.
- **Traitement de l'hétérogénéité des types de données :** prise en charge des différents types de données (numériques, nominales, Date-Time et booléennes) afin de maximiser l'ensemble des informations traitées.
- **Étude de l'impact des représentations élémentaires (par temps vs par événement):** analyse de l'effet de la concaténation symbolique lors de la représentation sur les résultats de classification, et choix de la meilleure représentation (par temps ou par événement) générée par le modèle DRRD.
- **Application de plusieurs classifieurs :** utilisation de différentes techniques de classification afin de sélectionner la plus adaptée.

- **Évaluation des performances selon plusieurs mesures :** évaluation des performances à l'aide de plusieurs mesures (Kappa-Statistics, F-Measure et AUC) afin de confirmer les résultats et faciliter le choix de la meilleure série de traitements.
- **Création d'un modèle de représentation à usages multiples :** la similarité des données médicales avec celles d'autres domaines en termes de structure et de type suggère que le modèle proposé peut être utilisé dans plusieurs domaines d'application.

6. Conclusion

Ce travail vise à proposer un modèle d'aide à la décision médicale basé sur l'exploitation des données issues des dossiers de santé électroniques (EHR), caractérisées par leur hétérogénéité et leur complexité. Dans ce contexte, une approche complète de prétraitement et de représentation des données a été développée afin de gérer différents types d'informations, notamment numériques, nominales, booléennes et temporelles, tout en préservant au maximum la richesse sémantique des données médicales.

Pour préparer les données et maintenir les informations contenues dans les données, nous avons introduit des techniques de représentation avancées basées sur la notion de région et de dispersion, permettant une meilleure restructuration des données médicales. L'objectif principal est de transformer des données brutes issues des EHR en une forme exploitable pour les algorithmes de classification.

Les expérimentations réalisées montrent que la représentation des données selon le temps, combinée au classifieur SVM, fournit de meilleures performances (F-measure, Kappa-Statistics et AUC). La représentation SDRRD par temps et le classificateur SVM seront les piliers de notre modèle pour la prise de décision médicale dans les cas pratiques, également pour de futurs travaux d'exploration sur d'autres sources de données hétérogènes et similaires.

Conclusion générale

Conclusion générale

Ce travail s'inscrit dans le cadre de l'exploitation des données issues des systèmes de santé modernes, en particulier les dossiers médicaux traditionnels, informatisés et les dossiers de santé électroniques (EHR). Dans un premier temps, nous avons présenté l'évolution des systèmes de gestion des données médicales, depuis le dossier médical papier jusqu'au dossier médical informatisé, en mettant en évidence leurs limites, notamment la perte d'information, les difficultés d'accès, ainsi que les problèmes liés à la sécurité et à la confidentialité.

Ensuite, nous avons introduit le concept des EHR comme une solution plus avancée permettant le partage, l'intégration et la gestion efficace des données de santé. Nous avons également discuté les facteurs essentiels de leur mise en œuvre tels que l'interopérabilité, la standardisation, la sécurité des données et l'identification unique des patients, tout en soulignant les risques liés à la protection des informations personnelles.

Dans une deuxième partie, nous avons abordé les fondements du data mining appliqué aux données médicales, en présentant les principaux types de données, les tâches d'exploration (classification, clustering, prédiction et association), ainsi que les techniques les plus utilisées telles que SVM, k-NN, BayesNet, Random Forest et les réseaux de neurones. Nous avons également mis l'accent sur les données de séries temporelles et les méthodes d'évaluation des modèles, ce qui est essentiel dans l'analyse des données EHR.

Sur la base de ces concepts, nous avons proposé un modèle de prise de décision médicale adapté aux données hétérogènes des EHR. Ce modèle repose sur une représentation avancée des données, permettant de gérer différents types d'informations tout en réduisant la perte de données. L'approche proposée combine plusieurs étapes de traitement, de calcul de similarité et la classification des patients.

Les résultats expérimentaux montrent que la représentation temporelle associée au classifieur SVM offre de meilleures performances en termes de précision et de robustesse. Cette combinaison s'avère donc la plus efficace pour la classification des patients et l'aide à la décision médicale.

Enfin, ce travail ouvre plusieurs perspectives intéressantes. Il est envisageable d'étendre ce modèle à d'autres types de données médicales telles que les textes cliniques, les images médicales et les données génétiques. De plus, l'intégration de ce modèle dans des plateformes Big Data et des systèmes intelligents en ligne pourrait améliorer significativement les systèmes

d’aide à la décision médicale, notamment pour la prédiction des maladies, la personnalisation des traitements et l’analyse des effets secondaires.

Bibliographie

[1] A. Lievre et G. Moutel, LE DOSSIER MÉDICAL : CONCEPTS ET EVOLUTIONS (DROITS DES PATIENTS ET IMPACT SUR LA RELATION SOIGNANTS-SOIGNÉS), 2010, p. 25.

[2] A. Atmane, AMIMEUR, *Conception et réalisation d'une application web pour la gestion des archives médicales Cas d'étude : CHU de Béjaia,* 2014/2015, p. 3.

[3] A.j. Meille, M. L. Dassie , Dr. L. Labreze, «Le dossier médical informatisé,» consulté le février 2026. [En ligne]. Available: https://www.caducee.net/DossierSpecialises/systeme-information-sante/dmi.asp#ctabs7.

[4] C. Daniel, J.-P. Jais, N. E. Fadly, et P. Landais, « Dossier patient informatisé à visée de recherche biomédicale », Presse Médicale, vol. 38, p. 1468.

[5] «INFOROUTE SANTÉ DU CANADA. DSE interopérable, s.d,» page consultée février 2026. [En ligne]. Available: http://www.infoway- inforoute.ca/lang-fr/about-infoway/approach/investment-programs/interoperable-ehr.

[6] «capsahealthcare,» [En ligne]. Available: https://fr.capsahealthcare.com/blog/allaitement/le-role-des-dossiers-de-sante-electroniques-dans-lamelioration-de-laccessibilite-des-soins-de-sante/. [Accès le page consultée le 18 Février 2026].

[7] «Integrating Genetic Information into the Electronic Health Record: The Case of Adolescents' Revelation of Misattributed Parentage - Journal of Bioethical Inquiry,» 2025. [En ligne]. Available: https://link.springer.com/article/10.1007/s11673-025104481#:~:text=The%20electronic%20health%20record%20(EHR)%2C%20accessible%20to%20patients%20through,particularly%20sensitive%20issue%20warranting%20attention..

[8] «The electronic health record as a primary source of clinical phenotype,» page consultée Février 2026. [En ligne]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2518657/#:~:text=The%20electronic%20health%20record%20(EHR, romote%20standardization%20of%20care..

[9] «Intemational standards for building electronic health record (ehr),» [En ligne]. Available: https://ieeexplore.ieee.org/abstract/document/1500373. [Accès le page consultée Février 2026].

[10] «Development of the Electronic Health Record,» 15 February 2011. [En ligne]. Available: https://journalofethics.ama-assn.org/article/development-electronic-health-record/2011-03.

[11] «A history of EHRs: 10 things to know - Becker's Hospital Review: Healthcare News & Analysis,» [En ligne]. Available: https://www.beckershospitalreview.com/healthcare-information-technology/a-history-of-ehrs-10-things-to-know/. [Accès le page consultée Février 2026].

[12] «oracle, Les DMÉ : une base pour des soins de santé coordonnés,» 19 mai 2025. [En ligne]. Available: https://www.oracle.com/ca-fr/health/electronic-medical-record-emr/.

[13] Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ, «Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption,» *J Am Med Inform Assoc,* vol. 13, n° %12, pp. 121-6, Mar-Apr 2006.

[14] Sinha, P. K., Sunder, G., Bendale, P., Mantri, M., and Dande, Electronic health record: standards, coding systems, frameworks, and infrastructures. John Wiley and Sons, 2012.

[15] DUPONT Benoît & Benoît GAGNON, «La sécurité précaire des données personnelles en Amérique du Nord, Note de recherche nº 1, Chaire de recherche du Canada en sécurité, identité, et technologie,» page consultée février 2026. [En ligne]. Available: http://www.benoitdupont.net/sites/default/files/securiteprecaire.pdf.

[16] «ORDER HO-004 de la Commissaire à la vie privée de l'Ontario,» page consultée février

2026. [En ligne]. Available: http://www.ipc.on.ca/English/Decisions-andResolutions/Decisions-and-Resolutions-Summary/?id=7616.

[17] «News Release, Commissioner Cavoukian expects health sector to encrypt all health information on mobile devices: Nothing short of this is acceptable,» page consultée 2026.

[En ligne]. Available: http://www.ipc.on.ca/english/Resources/News-Releases/News-Releases-Summary/?id=917.

[18] MANAOUIL Cécile, «Le dossier médical personnel (DMP): « autopsie » d'un projet ambitieux? Médecine & Droit,» pp. 24-41, 31, (2009).

[19] *CNIL, Recommandation du 29 novembre 1983, Commission Nationale de l'Informatique*

et des Libertés.

[20] «COMMUNIQUÉ DE PRESSE, CNIL, 20 FÉVRIER 2007,» 13 novembre 2019.

[En ligne]. Available: https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000017652041?.

[21] *CODE DE LA SANTÉ PUBLIQUE, article L1111-8-1:,* 2020.

[22] SOBEL Richard, «the demeaning of identity and personhood in national identifications systems,» p. 357, 2002.

[23] Commission d'accès à l'information, «Mémoire sur le Projet de loi n°83,» loi modifiant

la Loi sur les services de santé et les services sociaux et d'autres dispositions législatives Commission parlementaire des affaires sociales, janvier 2005.

[24] Federal Register, «GovInfo vol 65 pp 82,462; 82,467,» [En ligne]. Available: https://www.govinfo.gov/.

[25] «INFOROUTE SANTÉ DU CANADA , livre blanc sur la gouvernance de l'information dans le dossier de santé électronique (DSE) intéropérable,» mars 2007. [En ligne]. Available: http://www2.infowayinforoute.ca/Documents/Information%20Governance%20Paper%20Final_20070328_FR.pdf.

[26] SWEENEY Latanya, «« K-Anonymity: A model for protection privacy » International Journal on Uncertainty, Fuzziness and knowledge-based Systems,» vol. 10 , n° %15, 2002.

[27] Fayyad, U., Shapiro, G. P., & Smyth, P, «From Data Mining to Knowledge Discovery in Databases. AI Magazine,» vol. 17, n° %13, 1996.

[28] «Cours Data Mining,» Université Paris Descartes, Février 2026. [En ligne]. Available: https://helios2.mi.parisdescartes.fr/~lomn/Cours/DM/Material/ComplementsCours/Cours-Data-Mining.pdf.

[29] Roiger, R. J, Data Mining A Tutorial-Based Primer, Second edition éd., FL: CRC Press, 2017.

[30] HOGGUI Houdaïfa ,MANSOURI Ali, *Utilisation des techniques de data-mining pour La détéction de coincement de la garniture de forage,* UNIVERSITE KASDI MERBAH OUARGLA, 2017.

[31] Djazia CHAMI, *Une plate forme orientée agent pour le data mining Mémoire pour l'obtention du diplôme de Magister en informatiq,* Université HADJ LAKHDAR – BATNA, 2010.

[32] M. J. BERRY, G. S. LINOFF, *Mastering Data Mining: The Art and Scienceof Customer Relationship Management,* 2000.

[33] M. J. BERRY, G. S. LINOFF, *Data Mining Techniques For Marketing Sales, and Customer Relationship, Management,* S. Edition, Éd., 2004.

[34] D.T. LAROSE, *Discovering Knowledge In Data: An Introduction to Data Mining,* Central Connecticut State University, 2005.

[35] R. J. Roiger and M. W. Geatz, Data Mining: A Tutorial-Based Primer, S. edition, Éd., Boca Raton, FL: CRC Press, 2017.

[36] S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, P. New Age International (P) Ltd., Éd., New Delhi, 2007.

[37] G. Shmueli, N. R. Patel, P. C. Bruce, Data Mining for Business Analytics: Concepts, Techniques, and Applications in R, Second edition éd., Wiley, 2017.

[38] Gorunescu, Data Mining: Concepts, Models and Techniques, Springer, 2017.

[39] Kantardzic, M, Data Mining: Concepts, Models, Methods, and Algorithms, S. edition, Éd., Wiley-IEEE Press, 2011.

[40] Corina Cortes et Vladmir Vapnik, «Support-Vector Networks,» 1995.

[41] «SVM a marge dure,» [En ligne]. Available: https://reussirlem2info.wordpress.com/wp-content/uploads/2013/02/bekhelifi-okba-svm-resume.pdf. [Accès le mars 2026].

[42] R.GILLERON, M. TOMMASI, Dé, couverte de connaissances à partir de données, 2000.

[43] Napoleon D, Pavalakodi S, *A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set International Journal of Computer Applications,* 2011.

[44] Patro, V. Mohan & Patra, Manas Ranjan, «Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy,» *Department of Computer Science, Berhampur University, Berhampur-760007,TMLAI,* vol. 2 n (4).

[45] Sarkar, M., & Leong, T. Y, «Application of K-nearest neighbors algorithm on breast cancer diagnosis problem,» chez *AMIA Annual Symposium*, 2000.

[46] Al Bataineh, A, «A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection,» *International Journal of Machine Learning and Computing,* vol. 9, p. 248–254, 2019.

[47] Ishwaran, H. ; Lu, M., «Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival,» *Statistics in Medicine,* vol. 38, n° %14, p. 558–582, 2018.

[48] Kavzoglu, T, Object-Oriented Random Forest for High Resolution Land Cover Mapping Using Quickbird-2 Imagery,Handbook of Neural Computation, S. S. V. E. Pijush S, Éd., London: Academic Press, 2017, p. 607–619.

[49] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed éd., Waltham, MA, USA: Morgan Kaufmann, 2012.

[50] L. S. Coelho and N. F. F. Ebecken, Exploration of industrial databases using data mining techniques, in Proceedings of the 9th National PRIMECA Colloquium, April 2002.

[51] Berger, A., & Guda, S, Threshold optimization for F measure of macro-averaged precision and recall. Pattern Recognition, 102 éd., 2020.

[52] Chollet, François, Deep Learning with Python. Manning Publications, 2018.

[53] Grandini, M., Bagli, E., & Visani G, Metrics for multi-class classification : anoverview. Stat.ML, 2020.

[54] Zou, K. H., O'Malley, A. J., & Mauri, L, Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation, vol. 115, 2007, p. 654–657.

[55] Roiger, R. J, Data Mining A Tutorial-Based Primer, Second edition éd., CRC Press, 2017.

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

[57] Ratanamahatana, C. A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M,. & Das, G, Mining Time Series Data. In: Maimon O., Rokach L. (eds.) Data Mining and Knowledge Discovery Handbook. Springer Science+Business Media,, 2nd ed éd., 2009.

[58] Keogh, E., Chu, S., Hart, D., & Pazzani, M, Segmenting time series : A survey Segmenting time series : A survey and novel approach. In : Data Mining in Time Series Databases., W. Scientific, Éd., 2004, p. 1–21.

[59] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, «Automated Document Classification for News Article in Bahasa Indonesia based on Term Frequency Inverse Document Frequency (TF-IDF) Approach».

[60] «Understanding TF-IDF (Term Frequency-Inverse Document Frequency),» page consultée le 23 mars 2026. [En ligne]. Available: https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/.

[61] Kadi, H., Rebbah, M., Meftah, B., & Lézoray, O, *A data presentation model for personalized medicine. International Journal of Healthcare Information Systems and Informatics,* vol. 16, 2021.

[62] «Alzheimer's Disease Neuroimaging Initiative (ADNI),» (Accessed 2019). [En ligne]. Available: : http://adni.loni.usc.edu/.

[63] Shivam Kalra, Larry Li, H.R. Tizhoosh, «Automatic Classification of Pathology Reports using TF-IDF Features,» 05 mar 2019. [En ligne]. Available: https://www.semanticscholar.org/paper/58d984d5f23e6adfb0c4e5438a0bba2a89eee570.

[64] Matimpati Chitra Rupa, Kasarapu Ramani, «Hybrid Approaches for Advanced Medical Text Summarization: Combining TF-IDF, BERT, and Seq2Seq Models,» 2025. [En ligne]. Available: file:///C:/Users/HP/Desktop/1.+Matimpati+Chitra+Rupa.pdf, https://doi.org/10.26877/asset.v7i3.1427.

[65] Anuj Kumar, Rakesh Kumar, Shashi Shekhar, «HYBRID FRAMEWORK FOR SENTIMENT ANALYSIS OF PATIENT REVIEWS USING LEXICON BASED BIO-BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS,» 23 04

2026. [En ligne]. Available: https://journals.agh.edu.pl/csci/article/view/6461, https://doi.org/10.7494/csci.2026.27.1.6461.

[66] C. Rashmi, Dava Srinivas, S. Prabu, P. N. Santosh Kumar, A, «Natural Language Processing in Healthcare: Intelligent System for Extracting Insights from Clinical Text data,» 20-21 October 2023. [En ligne]. Available: https://www.semanticscholar.org/paper/Natural-Language-Processing-in-Healthcare%3A-System-Rashmi-Srinivas/43be665d4a2fe95b72487f6efb50452556589df8 et https://ieeexplore.ieee.org/document/10392668,DOI 10.1109/EASCT59475.2023.10392668.

[67] M. Jamaluddin, A. Wibawa, «Patient Diagnosis Classification based on Electronic Medical Record using Text Mining and Support Vector Machine,» 18-19 September 2021. [En ligne]. Available: https://ieeexplore.ieee.org/document/9573178/references#references, DOI 10.1109/iSemantic52711.2021.9573178.

[68] Pulluri Sreenivasgoud, M. K. Sharma, B. S. Kumar, Ginni Nijhawan, Hassan M. Al-Jawahry, R. Udhayakumar, «Natural Language Processing in Electronic Health Record Mining for Clinical Decision Support,» 29-30 December 2023.

[69] Nadya Mumtazah, Indra Waspada, K.N Dinar Mutiara, S .Adhy, «A Combination of Data Mining Methods for Disease Classification Using Patient-Perceived Symptoms from Medical Records,» 17-18 July 2024. [En ligne]. Available: https://www.semanticscholar.org/paper/A-Combination-of-Data-Mining-Methods-for-Disease-Mumtazah-Waspada/9cc27ea73c9829ec1ec07da86733fc4e4dd00f96, DOI 10.1109/ICICoS62600.2024.10636866.

[70] Michael Arenson, J. Hogan, Liyan Xu, Raymond J. Lynch, Hana Lee, Jin-Young Choi, Jimeng Sun, Andrew Adams, R. Patzer, «Prediction of Post-transplant Hospitalization Among Kidney Transplant Recipients With Clinical Notes and Electronic Healthcare Record Data,» vol. 106, n° %19S, p. 41, September 2022.