République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saâd Dahlab, Blida
USDB.

Faculté des sciences.
Département informatique.

Mémoire pour l'obtention du diplôme de
Master II Recherche en informatique
Option : Ingénierie logicielle

Sujet

Techniques de fouilles de données pour l'aide à la décision pour la recherche épidémiologique : Application au domaine cardiovasculaire

Réalisé par :
MOUNA KENOUI

Encadré par :
PR. SALIHA OUKID

Membres de Jury :
Mr Bala
Mr Hadj Yahia
Melle Bouyerbou Hafida

Promotion : 2010/2011

MA-004-103-1

Aux miens,

Remerciements

Je tiens en premier lieu à exprimer mes remerciements à ma promotrice, Madame OUKID Saliha, d'abord pour m'avoir fait confiance en me confiant ce sujet, ensuite pour son aide, son suivi et ses conseils tout au long de mon travail. Merci de m'avoir donné l'occasion de travailler sur un thème aussi intéressant.

Je remercie les membres du jury qui ont bien voulu juger mon travail.

Un merci particulier, sincère et chaleureux à Melle Ameur Khadidja, doctorante au Département Informatique de l'Université Saâd Dahlab - Blida - pour son aide précieuse, ses encouragements et son ouverture d'esprit.

C'est aussi pour moi l'occasion ici, de remercier toutes les personnes qui ont contribué, d'une manière ou d'une autre, à ce travail.

Enfin je remercie affectueusement mes parents, Salim et Nardjes pour leur présence et leurs encouragements dans toutes les démarches que j'entreprends.

Mouna.

Table des matières

Résumé

L'épidémiologie vise en particulier à la recherche des causes des maladies et à l'amélioration de leurs traitements et moyens de prévention. L'épidémiologie dite « clinique » utilise des données cliniques recueillies auprès de groupes de malades pour une meilleure prise de décision face à un malade donné.

De là, apparaît la nécessité en recherche épidémiologique, de disposer d'outils d'aide à la décision pour le pronostic, le diagnostic ou encore l'analyse de risques. En outre, ces outils manipulent des données cliniques, et devront par conséquent, prendre en compte l'analyse et l'exploration de ces dernières.

Nous voulons dans ce travail montrer l'apport des techniques de data mining pour l'aide à la décision en recherche épidémiologique, notamment pour la prédiction de facteurs de risques cardiovasculaires.

Nous examinons d'abord de près les données cliniques à considérer, avec les modalités de récolte et de recueil. Nous nous penchons, ensuite, sur l'utilisation des techniques de fouille de données dans le domaine médical, en particulier pour la prédiction des risques cardiovasculaires.

Nous mettons en place une approche hybride de data mining qui se décline en deux phases : dans la première, nous appliquons cinq méthodes de classification (C4.5, classification Tree, Logistic Regression, Naifes Bayes et CN2) pour identifier séparément les facteurs de risques de l'HTA et du diabète (deux facteurs majeurs des maladies cardiovasculaires). Dans la deuxième phase, nous introduisons uniquement les facteurs de risque communs obtenus dans l'étape précédente et nous appliquons la méthode Multivariate Adaptative Regression Splines (MARS) pour construire le modèle prédictif traitant de l'HTA et du diabète simultanément. Avec le modèle retenu, nous sommes capables de prédire l'HTA, le diabète ainsi que la coexistence des deux pathologies chez un même patient avec une précision de 97.75 % et une sensibilité de 96,87 %.

Mots clés : Fouilles de données (data mining), recherche épidémiologique, analyse exploratoire, données cliniques, maladies cardiovasculaires; hypertension artérielle; diabète; classification, modèle prédictif, facteurs de risque communs; MARS.

Liste des figures

Liste des tableaux

Liste des formules

Introduction générale

De nombreux domaines industriels, commerciaux, médicaux ... collectent et gèrent des masses de plus en plus volumineuses d'informations. La taille et le nombre de bases de données augmentent sans cesse. Généralement, ces bases de données sont très peu exploitées en vue d'extraire de nouvelles connaissances sur des phénomènes divers ou simplement éclairer des choix et/ou des décisions.

Alors que la découverte de connaissances dans les bases de données devient un enjeu stratégique afin de mieux cibler les consommateurs, de mieux évaluer les risques financiers, de mieux diagnostiquer des patients ..., une nouvelle préoccupation informatique est apparue depuis ces dernières années : **l'Extraction de Connaissances à partir de Données (ECD)**. Ainsi, une combinaison d'intérêts financiers, commerciaux, médicaux et scientifiques favorise et entretient cette nouvelle activité qui est la production d'outils de découverte de connaissances notamment pour des fins d'aide à la décision. La problématique de l'ECD exploite les principes de l'Apprentissage Automatique et utilise de façon privilégiée les méthodes d'apprentissage supervisé.

Dans un contexte médical, le recours à l'ECD et donc aux techniques de **Fouilles de données (Data Mining)** est de plus en plus envisagé afin de répondre à des besoins spécifiques de médecins, de cliniciens et d'épidémiologistes. Dans un cadre plus général, les systèmes d'aide à la décision médicale **(SADM)** qui sont « des applications informatiques dont le but est de fournir aux médecins et cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients » utilisent non seulement des outils d'analyse de données mais également des moyens appropriés de recueil et de récolte de données médicales nécessitant des solutions d'entreposage adaptées en raison de la nature de ces données, qualifiées souvent de complexes entre autres pour leur volume, leur hétérogénéité, leur disparité et leur évolution temporelle ...

C'est justement dans ce cadre médical que s'inscrit notre travail, qui vise à montrer l'apport des techniques de Data mining pour l'aide à la décision en recherche épidémiologique.

Ce travail est définit dans un contexte clinique algérien puisque les données manipulées sont réelles et sont issues d'une étude épidémiologique algérienne engagée par des médecins cardiologues s'intéressant aux facteurs de risques des pathologies cardiovasculaires et souhaitant aboutir à des connaissances approfondies de la population étudiée où un facteur de risque majeur de ces maladies, à savoir l'hypertension artérielle, est en pleine prolifération.

Ainsi notre travail s'articule autour de deux problématiques :

- Recueil et récolte de données cliniques en recherche épidémiologique
- Application des techniques de fouille de données pour le cardiovasculaire

Ce mémoire est alors organisé en quatre chapitres :

- ✓ Le premier chapitre est consacré aux systèmes d'aide à la décision médicale, il permet dans une première partie, d'introduire les notions de système d'aide à la décision, d'entrepôts de données, de magasins de données. Il détaille, dans une seconde partie, les systèmes d'aide à la décision médical et met l'accent sur les caractéristiques des données cliniques.
- ✓ Le deuxième chapitre porte sur les systèmes de récolte et de recueil de données cliniques en recherche épidémiologique, il fait un état de l'art des moyens de stockage et d'entreposage de ces données à des fins de prises de décision.
- ✓ Le troisième chapitre présente le procédé d'extraction de connaissances à partir des données et se focalise sur les méthodes d'apprentissage supervisé pour la prédiction des facteurs de risques cardiovasculaires. De nombreux travaux sont décris et étudiés pour situer les problématiques de la construction des modèles prédictifs.
- ✓ Le dernier chapitre met en œuvre la conception d'un modèle prédictif optimal pour une catégorie de malades à haut risque cardiovasculaire, en se basant sur une approche hybride de data mining, guidée par l'ECD allant ainsi de la compréhension du domaine étudié à l'implémentation de la solution. Plusieurs algorithmes de classification supervisée sont utilisés et comparés. L'approche est évaluée par de nombreux tests et les résultats sont interprétés.

CHAPITRE I :

Systèmes d'aide à la décision médicale

Introduction

La mondialisation et la concurrence qu'elle engendre rendent le pilotage d'une organisation de plus en plus complexe. Cette complexité est liée non seulement à l'augmentation du nombre de paramètres à prendre en compte mais également à la nécessité de prises de décisions rapides afin d'être réactifs à la demande des clients. L'efficacité de ces prises de décisions repose sur la mise à disposition d'informations fiables, pertinentes et d'outils facilitant cette tâche. Les systèmes traditionnels, dédiés à la gestion quotidienne d'une organisation, s'avèrent inadaptés à une telle activité **[1].** Face à ce besoin est né le secteur de l'informatique décisionnelle.

Dans ce chapitre, nous introduisons les notions essentielles dont il faut disposer pour comprendre les systèmes d'aide à la décision dans le milieu médical, leur importance pour les professionnels de la médecine et plus précisément pour les cliniciens afin d'établir le lien avec le thème de notre étude.

1. Aide à la décision

La modélisation systémique de toute organisation se décompose en trois sous-systèmes : Systèmes Opérant (SO), Système d'Information (SI) et Système de Pilotage (SP). Le SO représente l'activité productrice de l'organisation étudiée. Cette activité consiste à transformer les flux primaires (matières, finances, personnel …) pour répondre aux besoins des clients. Le SP regroupe l'ensemble du personnel d'encadrement qui effectue les tâches de régulation, de pilotage et d'adaptation de l'organisation à son environnement **[2].** Le SI permet de collecter, mémoriser, traiter et restituer des différentes données de l'organisation afin de permettre au SP d'effectuer ses fonctions tout en assurant son couplage avec le SO **[3].** L'activité du SO produit des informations stockées dans le SI ; après traitement, la transmission de ces informations vers le SP permet à ce dernier de connaître l'activité du SO (flèches « informations » dans la figure N° 1.1). Les décisions du SP seront répercutées vers le SI puis vers le SO pour permettre au SP d'en maîtriser le fonctionnement (flèches « décisions » dans la figure N° 1.1).

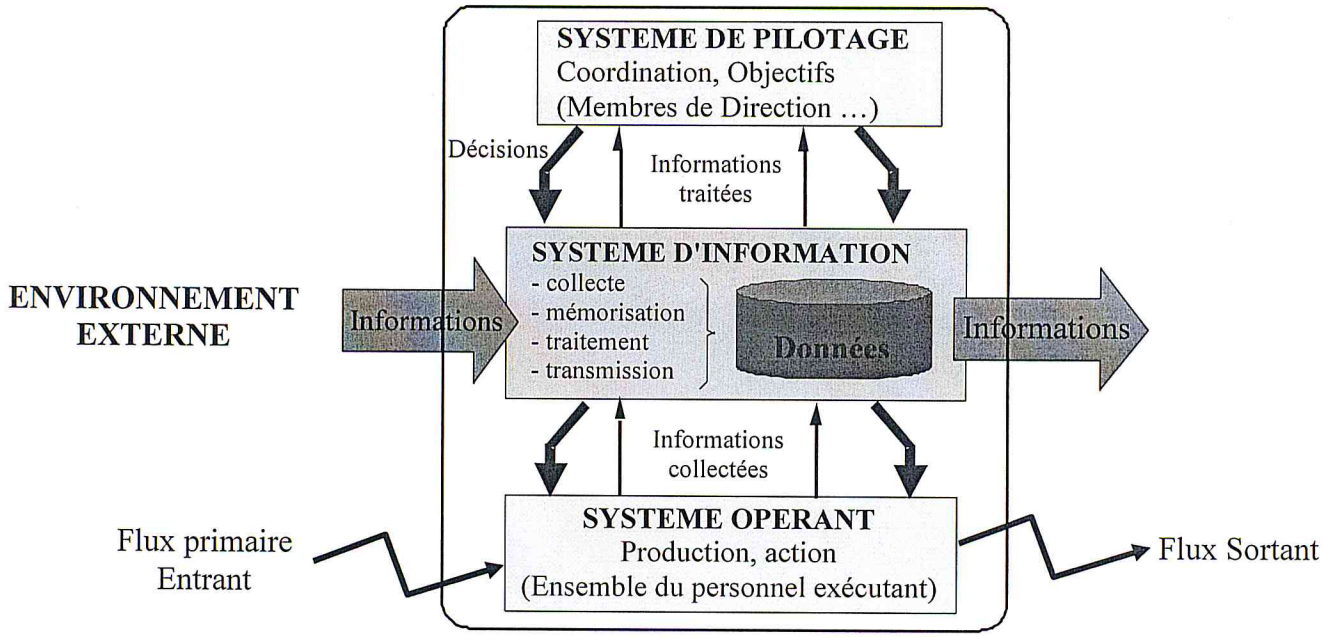

Figure N° 1.1 : Représentation systémique d'une organisation [2]

Pour répondre aux besoins des décideurs, il est nécessaire de synthétiser, réorganiser et historiser les données de production du SI afin d'en déterminer une sous partie relative à l'aide à la décision. La suite de ce mémoire se centre sur cet aspect. Notamment, dans les sections suivantes, nous définissons les concepts de système d'information d'aide à la décision, de système d'aide à la décision, d'entrepôts et de magasins de données **[1].**

1.1. Systèmes d'information d'aide à la décision (SIAD)

Nous proposons la définition du Système d'Information d'Aide à la Décision (SIAD) suivante :

Définition : Un SIAD est la partie d'un système d'information permettant d'accompagner les décideurs dans le processus de prise de décision. Les fonctions d'un SIAD permettent de :

- collecter, intégrer, synthétiser et transformer les données opérationnelles d'un SI,
- mémoriser de manière adaptée les données décisionnelles,
- traiter ces données (alimentation, rafraîchissement, pré-calculs…),
- restituer de manière appropriée ces données afin de faciliter la prise de décision.

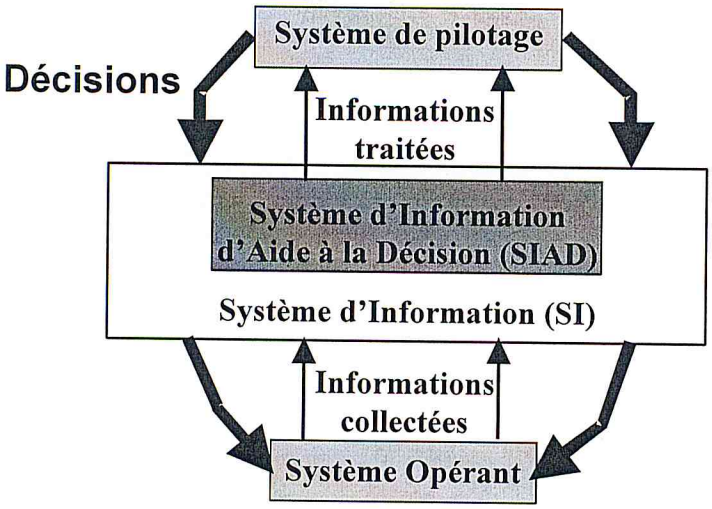

Figure N° 1.2 : Le SIAD dans le SI

De nos jours, l'ensemble des outils informatiques permettant de supporter un SIAD est qualifié de Business Intelligence (BI) ou de Système d'Aide à la Décision (SAD). Un SAD vise à exploiter les données opérationnelles d'une organisation afin de faciliter la prise de décision pour un pilotage éclairé. Afin d'être plus explicite, nous proposons la définition suivante :

Définition : Un Système d'Aide à la Décision (SAD) regroupe l'ensemble des outils informatiques (matériels et logiciels) permettant :

- D'extraire, de transformer et de charger les données opérationnelles,
- De constituer un ou des espaces de stockage de données décisionnelles,
- De manipuler ces données au travers d'outils d'analyse ou d'interrogation destinés au pilotage des organisations.

La plupart des travaux déclinent ces applications informatiques en trois catégories :

– Extraction, transformation et chargement (ou ETL acronyme de "Extraction Transformation Loading") des données opérationnelles (hétérogènes et disparates) pour alimenter et rafraîchir le système d'aide à la décision,

– Stockage (entreposage) et traitement des données décisionnelles,

– Restitution des données sous une forme adaptée aux utilisateurs (interrogations ou analyses décisionnelles) **[1]**

Nous pouvons schématiser ces différents outils dans la figure suivante :

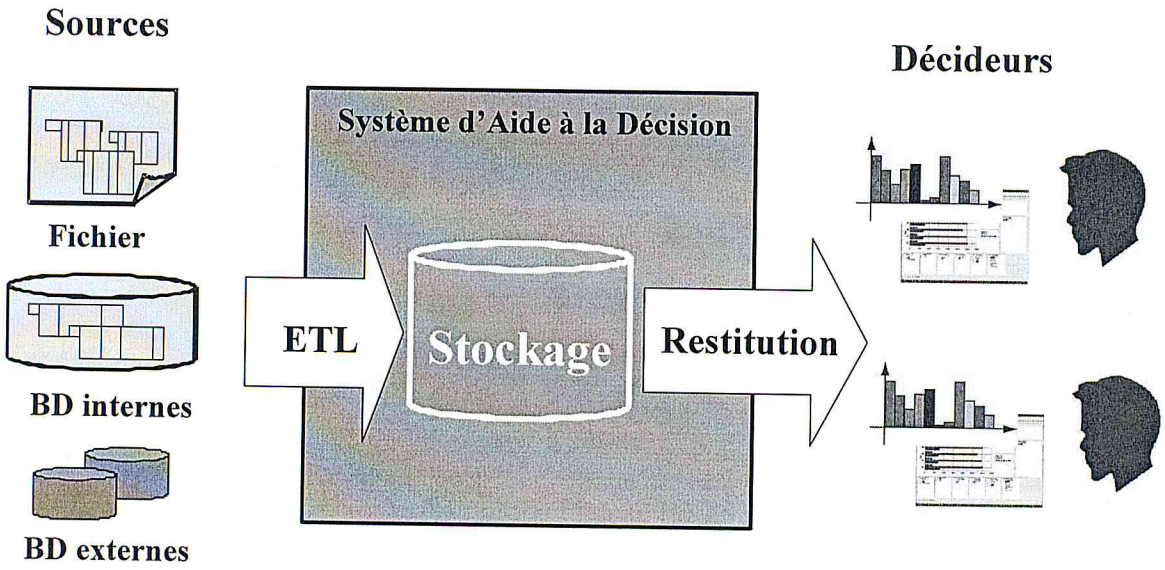

Figure N° 1.3 : Le système d'Aide à la Décision

1.2. Des systèmes d'aide à la décision aux entrepôts de données

De nos jours, les entrepôts de données constituent une solution adéquate pour construire un système d'aide à décision. Un entrepôt de données (ED) est défini comme étant "une collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse" **[4].**

Cette définition met l'accent sur les caractéristiques suivantes :

- **Intégrées :** les données alimentant l'entrepôt proviennent de sources multiples et hétérogènes. Les données des systèmes de production doivent être converties, reformatées et nettoyées de façon à avoir une vision globale dans l'entrepôt.
- **Orientées sujet :** contrairement aux systèmes de production structurant les données par processus fonctionnel, les données d'un ED s'organisent par thèmes d'analyse. L'intérêt de cette organisation est de disposer de l'ensemble des informations utiles sur un thème, le plus souvent transversales aux structures fonctionnelles et organisationnelles d'une entreprise. Cette orientation « sujet » permet de mettre en avant les indicateurs de performance pour chaque thème d'analyse.

– **Non volatiles :** après intégration, transformation et synthèse des données opérationnelles dans un ED, les seules actions que peuvent effectuer des décideurs sont des interrogations et des analyses décisionnelles (pas de mise à jour).

– **Historisées :** l'alimentation et le rafraîchissement d'un ED consistent en l'intégration des données opérationnelles à différents points d'extraction. Cette intégration de données à des dates différentes permet de conserver "l'historisation" des données qui est vitale pour toute prise de décision.

– **Résumées :** les informations issues des sources doivent être transformées mais surtout *agrégées* pour faciliter le processus de prises de décision.

– **Disponible pour l'interrogation et l'analyse :** afin d'améliorer les performances d'une organisation, les décideurs doivent pouvoir consulter et analyser les données contenues dans un ED au travers d'outils interactifs.

1.3. Entrepôts et magasins de données

D'après la définition de **[4]**, l'ED doit permettre d'extraire, de transformer et de stocker un grand volume de données opérationnelles et, en même temps, de répondre à des requêtes utilisateurs concernant un thème d'analyse spécifique. En fait, cette définition regroupe deux problématiques à prendre en compte :

– La gestion efficace des données (intégration des sources),

– La définition d'un sous-ensemble de données autour d'un thème particulier afin de répondre aux besoins spécifiques de décideurs.

Aussi, l'architecture des systèmes d'aide à la décision que nous pouvons mettre en exergue est basée sur une dichotomie d'espaces de stockage : l'entrepôt et le magasin de données **[5].**

Définition 1 : Un Entrepôt de Données (ED) est l'espace de stockage centralisé d'un extrait des sources pertinent pour les décideurs. Son organisation doit faciliter la gestion des données et la conservation des évolutions nécessaires pour les prises de décision.

Définition 2 : Un Magasin de Données (MD) est un extrait de l'ED adapté à un thème d'analyse particulier et organisé selon un modèle adapté aux outils d'analyse et

d'interrogation décisionnelles.

Dans la figure suivante, nous schématisons l'architecture des SAD telle que nous l'avons définie précédemment **[1]**

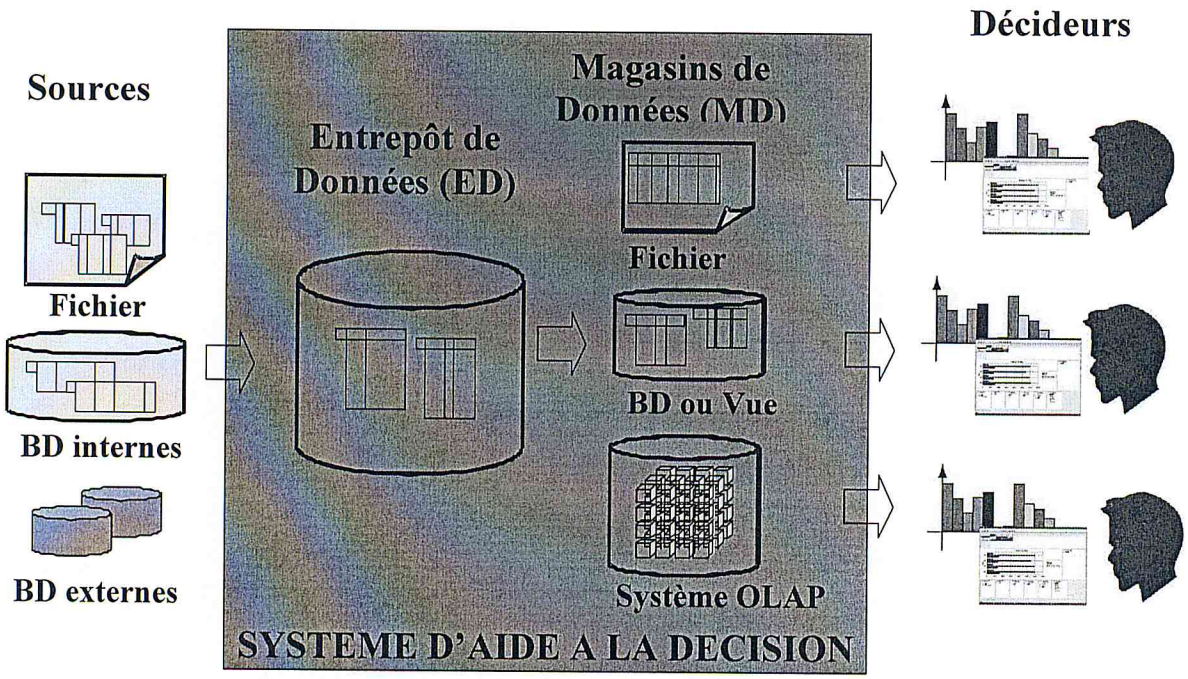

Figure N° 1.4 : Entrepôt et magasins de données

1.4. Synthèse

Dans cette première section, nous avons défini les différents concepts servant de support à notre travail. A partir de la représentation systémique d'une organisation, nous avons identifié le concept de SIAD qui est la partie d'un SI permettant d'accompagner un ou plusieurs décideurs dans le processus de prise de décision. Un Système d'Aide à la Décision (SAD), partie informatisée d'un SIAD, regroupe l'ensemble des outils informatiques capables d'extraire les données opérationnelles afin de les transformer en informations pertinentes pour les décideurs. D'un point de vue architectural, nous avons identifié deux espaces de stockage des données dans un SAD : l'entrepôt (espace de stockage centralisé) et les magasins de données (espace de stockage extrait d'un ED et centré sur un thème d'analyse particulier).

2. Aide à la décision médicale

Durant les dernières décennies, le développement de systèmes d'aide à la décision dans le domaine médical a été perçu avec beaucoup d'intérêt par les médecins. Le développement de ces systèmes est d'autant plus marqué que l'avancée dans la recherche médicale s'accompagne de l'utilisation de moyens informatiques de plus en plus puissants visant à faciliter la tâche du médecin.

Notre travail concerne la recherche épidémiologique pour les maladies cardiovasculaires, il traite plus précisément de la prédiction de facteurs de risque. Dans cette section, nous allons approcher les systèmes d'aide à la décision dans un milieu médical afin d'asseoir les caractéristiques de la prise de décision médicale (en recherche épidémiologique et clinique).

2.1. Problématiques d'un système d'aide à décision médicale

Les systèmes d'aide à la décision médicale, traitent globalement des problématiques suivantes :

- Amélioration de la sécurité, de la qualité et de l'efficience des soins.
- Couverture de tout ou d'une partie des activités cliniques (prévention, diagnostic, prescription médicamenteuse, prescription d'actes diagnostiques, pronostic ou de suivi des soins …).
- Méthodes de traitement et d'analyse du signal pour l'interprétation des données médicales à visée diagnostique (ECG, EEG, dosages et prélèvements biologiques).
- Méthodes de planification des traitements et de guidage des interventions par l'imagerie (dosimétrie en radiothérapie, reconstruction et modélisation 3D, réalité virtuelle) **[6]**.

2.2. Aide à la décision clinique («clinical decision support»)

Une définition de l'aide à la décision clinique, que nous pouvons donner est :

Définition : «fournir aux cliniciens ou aux patients des connaissances cliniques ou des données relatives au patient, filtrées intelligemment et présentées au moment opportun pour faciliter la démarche de soin.»

(Clinical Decision Support Implementers' Workbook, Osheroffet al.HIMSS 2004)

Quatre types d'objectifs des systèmes d'aide à la décision clinique :

• «Administratif»: aide au codage et à la documentation des actes et diagnostics cliniques, pour l'organisation des procédures de soins, des consultations et des prises de rendez-vous.

• «Gérer les situations cliniques complexes»: patients justifiant un traitement ou suivi à long terme ou en plusieurs séances (protocoles de chimiothérapie ou de recherche cliniques), traçabilité des prescriptions, des RV, rappels de mesures préventives.

• «Contrôles des coûts»: surveillance et estimation des erreurs de prescription, prévention des prescriptions inutiles, répétées ou redondantes

• «Prise de décision»: diagnostic, pronostic clinique et plan thérapeutique. Promouvoir les bonnes pratiques, la mise en œuvre de recommandations de pratiques concernant des pathologies particulières, les politiques de santé publique **[7].**

2.3. L'information dans un système d'aide à la décision clinique

D'abord nous donnons dans la figure ci-dessous le système d'aide à la décision dans le circuit d'informations médicales afin d'expliciter les interactions possibles.

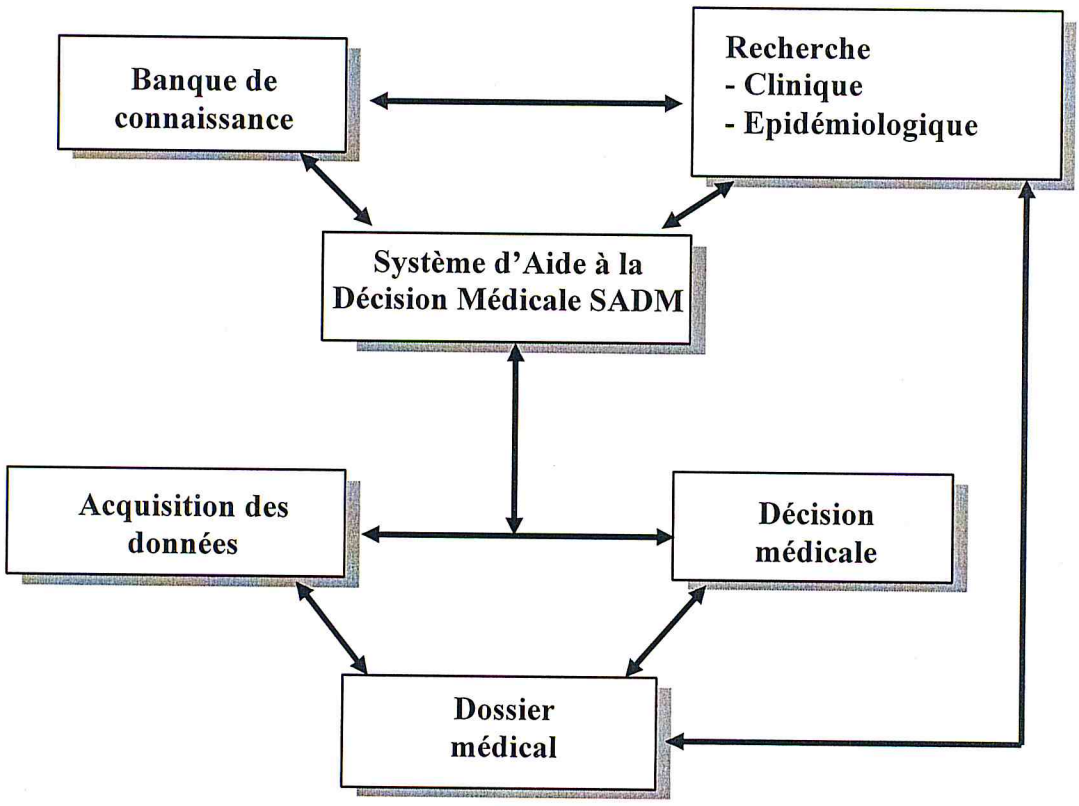

Figure N° 1.5 : Circuit d'informations médicales [7]

Les informations, extraites du dossier patient ou des bases de connaissances et rassemblées à partir des règles logiques du SADM, peuvent être ensuite organisées en fonction d'une tâche clinique donnée (diagnostic, prescription, suivi) ; elles peuvent être présentées au moyen d'une interface utilisateur permettant la consultation à partir d'un seul écran des informations utiles à la décision :

- Les données pertinentes du patient : histoire de la maladie, données cliniques, antécédents et allergies, résultats d'examens, traitements en cours,
- Les extraits des recommandations de pratique relatifs à la situation clinique,
- Les objectifs cliniques relatifs à un problème et un traitement particulier,
- Un accès aux textes de référence adaptés aux problèmes spécifiques du patient.

Ces données peuvent être aussi organisées sous forme de textes, de tableaux, de vues par spécialité/problème, ou encore de tableaux de bord pour permettre une décision médicale appropriée **[6]**.

Nous pouvons par conséquent, schématiser les composants d'un système d'aide à la décision médicale par la figure suivante :

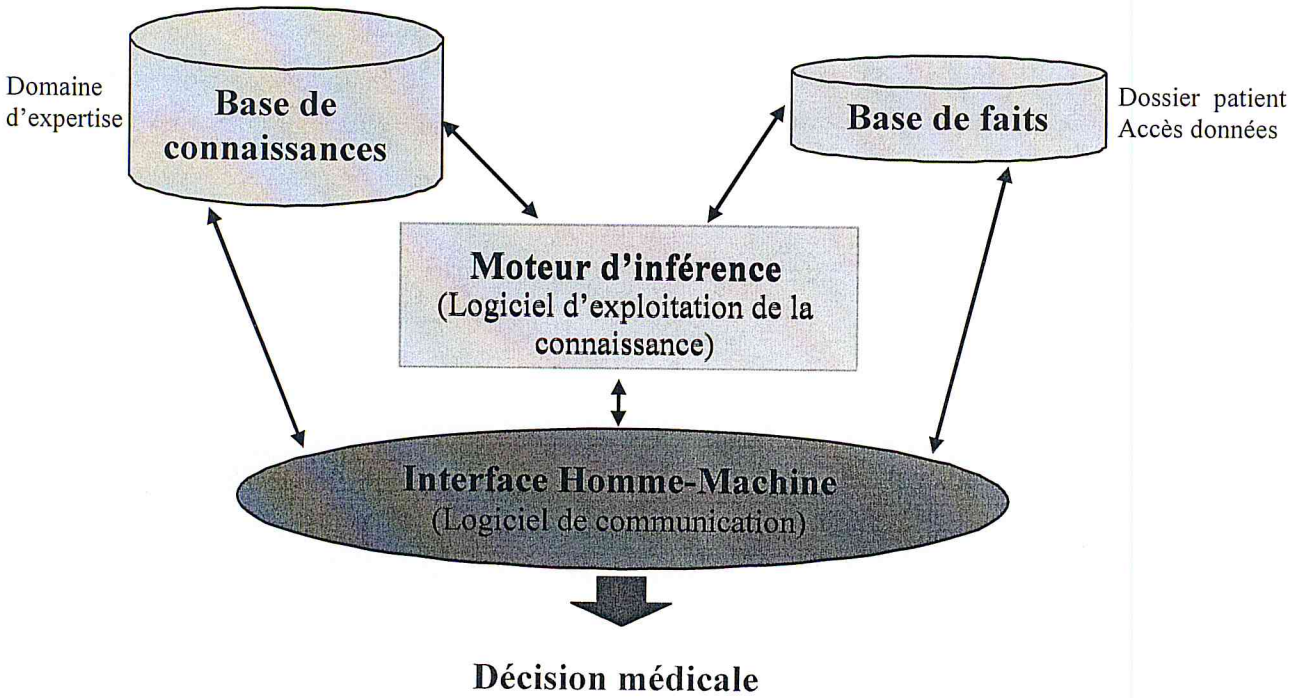

Figure N° 1.6 : Composants d'un système d'aide à la décision médicale [7]

2.4. Caractéristiques des données cliniques dans un SADM

Les données cliniques à intégrer dans un système d'aide à la décision en vue d'analyse sont, de part leur nature particulière, des données complexes à manipuler **[8]**.

2.4.1. Le volume des données

Les données que les études cliniques manipulent sont nombreuses. En effet, ces données peuvent, comprendre des fichiers volumineux (images médicales, signaux biologiques…). Indépendamment, des banques de données externes, connexes aux données cliniques des sujets étudiés sont des sources d'informations très riches dont la taille croit de manière exponentielle. En résumé, la quantité très importante des données à traiter est un critère à considérer dans le choix des méthodes d'analyse.

2.4.2. La qualité des données

Les données brutes sont de qualité très variable selon leur source : banques de données biologiques externes validées à la main ou non, mises à jour plus ou moins fréquemment, banques de données cliniques issues d'études plus ou moins précises (données manquantes ou insuffisantes…). Cette disparité est problématique pour l'intégration des données en un ensemble présentant une qualité homogène.

2.4.3. La localisation des données

En général, les données cliniques sont locales, en revanche, les données biologiques, la littérature scientifique, les bases de connaissances pharmacogénétiques sont dans des banques de données externes, publiques et disponibles librement à des localisations éparses sur internet. Il peut aussi s'avérer intéressant de recouper le contenu de plusieurs études cliniques dont les sources peuvent être éloignées physiquement.

2.4.4. L'hétérogénéité des données

La nature de ces données est aussi très variable. Les données des études cliniques

contiennent des informations générales sur le patient (âge, poids, taille, régime, pathologie…) ainsi que des données plus précises résultant de l'exploration biologique (dosages d'enzyme …). On peut également, avoir une information disponible directement (un diagnostic, un chiffre significatif) ou une information qu'il faudra traiter préalablement avant d'en extraire une connaissance (image médicale, signal biologique). Le contenu des banques de données externes est également très hétéroclite.

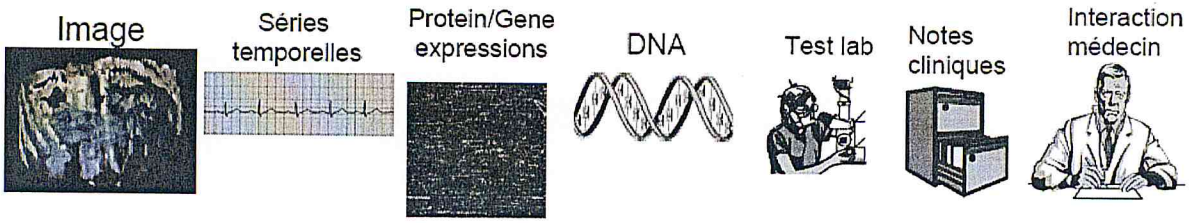

Figure N° 1.7 : Hétérogénéité des données médicales

2.4.5. L'évolutivité des données et les données temporelles

C'est une caractéristique majeure des données à intégrer. Les données des banques de données externes sont mises à jour régulièrement. Les données cliniques sur les patients d'une étude possèdent une dimension temporelle pour rendre compte de l'évolution de la pathologie et de l'effet du traitement au cours du temps.

2.5. Domaines cliniques d'application des systèmes d'aide à la décision médicale

La littérature récente **[6]** montrent que les SADM utilisés en pratique et ayant fait l'objet d'au moins une étude d'impact sur la qualité des soins couvrent :

- L'ensemble des activités médicales
- Les maladies chroniques, les affections aiguës et les urgences
- La plupart des spécialités médicales, cardiologie, cancer, les affections psychiatriques, la pédiatrie …

Le tableau suivant précise la prévalence des différents types d'activité clinique et de différentes maladies représentées dans la revue de Garg **[9]**; nous faisons ressortir le nombre

relatif aux maladies cardio-vasculaires ainsi que celui du diagnostic d'affections aigues et chroniques dont font partie les maladies cardiovasculaires, thème de notre étude :

Domaine d'application clinique	Nombre d'Etudes / 100
Prévention et dépistage	**21/100**
Dépistage des cancers	07
Vaccination	03
Combinaison dépistage/vaccination	11
Diagnostic d'affections aiguës ou chroniques	**10/100**
Prise en charge d'affections aigues ou chroniques	**27/100**
- Diabète	07
- **Maladies cardio-vasculaires**	13
- Autres maladies chroniques	05
- Affections aiguës	02
Prescription médicamenteuse	**31/100**
Combinaison de plusieurs domaines d'application	**11/100**

Tableau N° 1 : Répartition des 100 études rapportées dans la revue de Garg en fonction du domaine d'application clinique

Conclusion

Notre travail vise à exploiter des données cliniques issues d'études épidémiologiques avec comme objectif d'en extraire une connaissance sur la prédiction des facteurs de risque cardiovasculaire. Pour répondre à ce problème, nous devons nous pencher principalement sur l'utilisation des techniques de fouilles adaptées aux données épidémiologiques que nous souhaitons manipuler. Il était essentiel, avant de s'orienter vers ces techniques de voir le processus de prise de décision médicale en abordant d'abord les systèmes d'aide à la décision dans leur globalité, ensuite en se focalisant sur ces systèmes dans un milieu médical. Les caractéristiques des données cliniques nous ont permis de voir la complexité des données à traiter. Cela va avoir un impact sur le stockage et l'entreposage de telles données, nous traitons ces aspects dans le chapitre suivant.

CHAPITRE II :

Systèmes de récolte de données cliniques pour la recherche épidémiologique

Introduction

Au cours du changement profond survenu dans la culture médicale sur les vingt dernières années, on a reconnu que c'était en s'appuyant sur des démarches scientifiques rigoureuses — et non plus que sur l'expérience — que l'on devait décider de la meilleure politique de prévention ou de l'attitude thérapeutique la plus efficace vis-à-vis d'une maladie, ou même de la prise en charge d'un malade donné. Cet appel à la science est en pratique un appel à l'épidémiologie.

Ainsi, la médecine fondée sur la preuve scientifique (ou, en anglais, *Evidence Based Medicine*) devient la base culturelle de la médecine clinique. Dans ce chapitre, nous examinons les moyens de récolte et de recueil des données cliniques pour des fins de recherche épidémiologique. Nous trouvons intéressant de commencer par quelques détails concernant la recherche épidémiologique.

1. L'épidémiologie et son champ d'application

L'épidémiologie décrit les variations de fréquence des maladies dans les groupes humains, et recherche les déterminants de ces variations. Elle vise à la compréhension des causes des maladies, et à l'amélioration de leurs traitements et moyens de prévention **[10]** :

– L'épidémiologie concerne **l'ensemble** des maladies et situations pathologiques, et pas seulement les seules maladies transmissibles.
– La recherche épidémiologique de déterminants des maladies se fait en comparant les fréquences de ces maladies dans des **groupes** ou **sous-groupes** au cours du temps, en fonction du lieu, de la profession, des antécédents médicaux, de caractéristiques phénotypiques ou biologiques, etc.
– La recherche de ces « déterminants » vise en général à identifier des facteurs de risque, individuels ou collectifs, des maladies étudiées. L'épidémiologie permet alors de fournir des modèles prévisionnels reposant sur la connaissance de ces facteurs de risque.
– La recherche des « déterminants » des variations de fréquence des maladies mène à étudier parallèlement les variations à travers les groupes humains des **facteurs** intervenant dans la survenue des maladies : on parle alors d'épidémiologie de la tension artérielle, de la pollution atmosphérique, et même d'épidémiologie du médicament.

Sur le plan des méthodes, l'épidémiologie moderne s'appuie depuis longtemps sur la statistique ; réciproquement, elle a été, et est, à la source de la découverte de nouvelles méthodes statistiques, de fouille de données (data mining) …
De plus, depuis plusieurs années on assiste à une forte implication de nouveaux champs des mathématiques en épidémiologie, notamment calcul des probabilités, théories des systèmes complexes, analyse numérique et modélisation en général.

Contrairement a l'épidémiologie classique, qui étudie et interprète la maladie en tant que phénomène de groupe sur la base de multiples données individuelles, l'épidémiologie clinique utilise les informations épidémiologiques recueillies auprès des groupes de malades pour une meilleure prise de décision clinique face a un malade donné. Ainsi selon Feinstein, « l'épidémiologie clinique est une extension de l'épidémiologie classique (traditionnellement orientée vers les stratégies en santé publique et communautaire), c'est un champ d'activités qui recouvre l'ensemble des décisions cliniques telles qu'elles se présentent lors de la prise en charge de l'individu (patient) tant en ce qui concerne le diagnostic, le pronostic et les mesures thérapeutiques que d'autres jugements médicaux » (Feinstein, 1985) **[10]**.

2. Modalités de recueil de données cliniques pour la recherche épidémiologique

2.1. Dossier Patient Electronique DPE, (E*lectronic* Medical record) [11] [12]

Le Dossier Patient (Electronique) ou encore le Dossier Patient Informatisé (DPI) est à la fois, un instrument de conservation des données personnelles relatives à la santé du patient, mais aussi un instrument de travail du médecin qui sert de base à son travail diagnostic et thérapeutique. Ceci amène à la distinction entre les faits ou **données factuelles** et **les données d'exploitation**.

Les données figurant dans un dossier patient électronique (DPE) sont nombreuses **[11]**. Elles concernent des éléments d'identification d'ordre administratif et des données médicales et/ou médico-sociales. Le dossier patient comprend des faits d'observation, des éléments diagnostiques et/ou pronostiques et des décisions médicales de prévention, de diagnostic, de traitement, de pronostic ou de suivi. Donc, ce dossier recouvre un ensemble d'informations

associées aux soins, d'ordre médical, administratif et relatif à l'Assurance-maladie. Le DPE a pour objectif d'améliorer la qualité des soins. Il facilite la mémorisation des informations, la communication entre les partenaires de soins et favorise la continuité de soins et la prise de décisions médicales. Il permet de regrouper des informations destinées à faciliter l'évaluation, la recherche et la planification. Il est censé accompagner un patient au cours de ses épisodes de soins dans les institutions où il est pris en charge.

Donc la vocation principale du DPE, est de contenir en permanence la synthèse clinique sur le patient. Il est formé entre autres d'une base de documents médicaux. Dans ce contexte est née la standardisation pour le partage de documents médicaux (centré « patient ») entre structures de santé afin de permettre aux médecins une meilleure prise en charge du patient en collaborant entre eux même dans un contexte d'éloignement géographique.

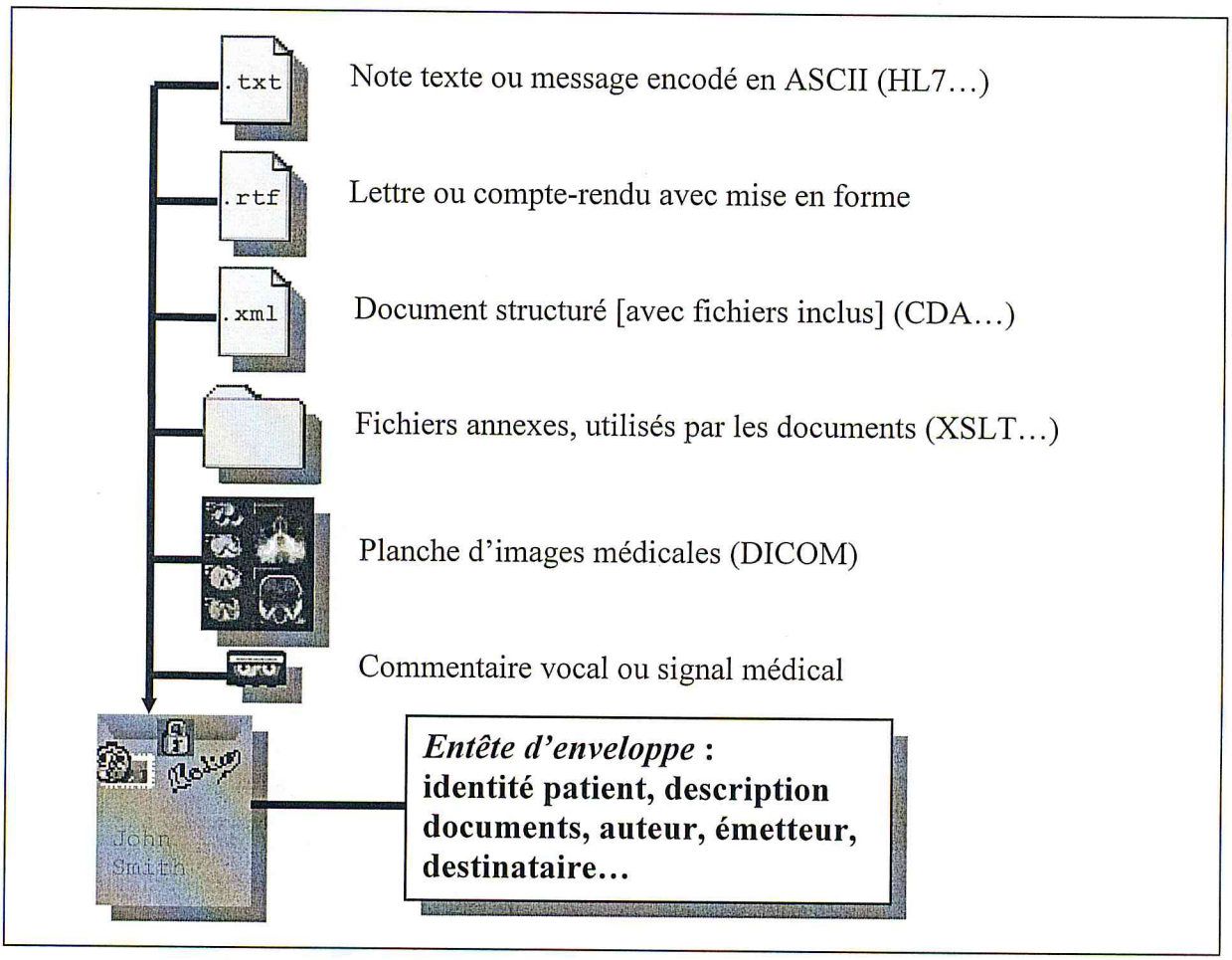

Figure N° 2.1 : Exemple de partage de documents médicaux en utilisant le concept d'enveloppe

Nous trouvons également dans la littérature de ce domaine la notion de ***« Document Repository »*** que nous pouvons traduire par « Dépôt ou encore Entrepôt de Documents». Ce dépôt contient les documents du patient (de formats quelconques .txt, .pdf, format image …). Le document patient peut être enregistré dans le dépôt de l'établissement patient comme il peut être renvoyé dans un autre dépôt (il y a un ou plusieurs dépôts de documents selon l'architecture globale du système médical déployé).

2.1.1. Les standards utilisés dans le partage de documents médicaux

En ce qui concerne la recherche épidémiologique ou la veille sanitaire, à l'instar des promoteurs de la recherche biomédicale, les instituts de veille sanitaire ont identifié l'importance d'un recueil et d'une transmission de données optimisés. Ainsi, en s'appuyant sur des normes courantes comme XML ou JPEG, la communauté santé s'est peu à peu retrouvée sur des standards communs. Voici les principaux :

- **IHE** (Integrating the Healthcare Enterprise), c'est une initiative internationale qui réunit les principales normes d'échange d'informations en santé.
- **IHE XDS** (Cross-Enterprise Document Sharing) c'est le profil de partage de documents, basé sur ebXML.
- **IHE XDS-I** (Cross-Enterprise Document Sharing for Imaging) qui permet le partage de document d'imagerie avec pour avantage de pouvoir connecter des systèmes (coûteux) déjà existants.
- **IHE PIX** (Patient Identifier Cross-referencing Integration Profile) qui permet le rapprochement des identités (identity patient)

2.1.2. Le DPE dans la recherche épidémiologique [11]

Les données cliniques recueillies lors d'une démarche de soin au sein d'un DPE (ou DPI pour Dossier Patient Informatisé) peuvent être très utiles pour la recherche biomédicale ou épidémiologique (Figure N° 2.2). Des initiatives visent à définir des solutions d'intégration des activités de soin et de recherche permettant d'optimiser le recueil et l'exploitation des données cliniques pour une réutilisation secondaire des données cliniques du DPE.

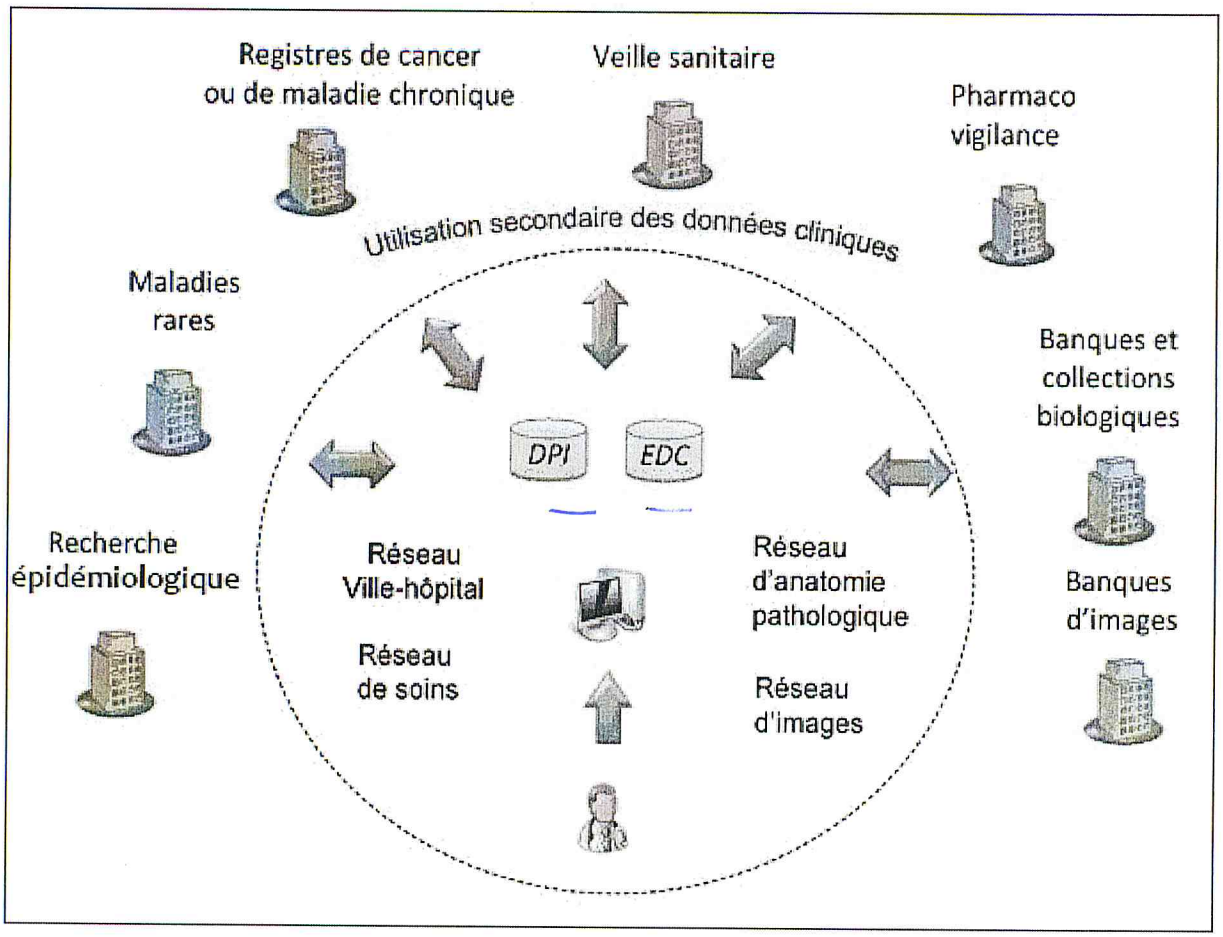

Figure N° 2.2 : Différentes situations d'utilisation secondaire de données cliniques à partir d'un dossier patient informatisé (DPI) ou dans un entrepôt de données cliniques

Concernant la recherche épidémiologique, l'étude du rôle du DPI dans le déroulement d'une étude a fait l'objet de nombreux travaux. Les données cliniques du DPI peuvent être utiles lors de différentes phases. Les données du DPI peuvent favoriser l'inclusion de patients lors de la phase d'instruction ou encore optimiser le recueil des données et faciliter le signalement spontané d'événements indésirables lors de la phase de mise en œuvre.

Même dans les hôpitaux universitaires ayant mis en œuvre des DPE arrivés à maturité, ces DPE sont complètement déconnectés des **systèmes de gestion des données cliniques** (SGDC) de la recherche épidémiologique. Dans la mesure où 30 % à 50 % des données recueillies dans un contexte de recherche biomédicale sont dans les DPI, les cliniciens sont donc souvent confrontés à une double saisie de données : d'abord dans le DPI, puis dans le SGDC de l'étude épidémiologiques ou biomédicale dont ils sont investigateurs.

En résumé, les données cliniques dans un DPE peuvent constituées une source de données pour les bases de données épidémiologiques ou encore dans les entrepôts cliniques que nous allons traiter dans ce qui suit. Néanmoins, ces données telle quelles sont dans les DPE s'apprêtent peu à l'analyse de données cliniques même si elles contribuent fortement à la prise de décision médicale en favorisant la collaboration et l'échange entre médecins et en facilitant la recherche d'informations liées à un patient en particulier.

2.2. Bases de données cliniques (Clinical Databases) [5, 13, 14, 15, 16, 17, 22]

Nous constatons en étudiant la bibliographie dans ce domaine, qu'il y a eu explosion de l'information susceptible d'être analysée en recherche épidémiologique. En effet, le nombre de variables mesurables en pratique par individu s'est accru de façon exponentielle. L'existence de gigantesques bases de données médicales construites initialement dans un but de gestion économique ou d'organisation des soins, laisse les spécialistes espérer pouvoir tirer une connaissance appropriée à leurs problèmes et aux situations cliniques des malades en utilisant ces mêmes bases de données.

En effet, les bases de données relationnelles pour la recherche épidémiologique permettent de stocker et de manipuler d'importants volumes de données concernant entre autres des données cliniques. L'exploitation de ces données pour l'analyse et la prise de décision, est réalisée souvent par des outils de reporting (requêtes), graphiques, tableaux...

Pour gérer au mieux ces données, les systèmes de gestion de bases de données (SGBD) basés sur le modèle **OLTP** (On Line Transaction Processing) sont utilisés. L'objectif de ce modèle est de permettre l'insertion et la modification de façon rapide et sûre. Les SGBD sont particulièrement adaptés pour répondre efficacement à des requêtes concernant une petite partie des données stockées dans la base. En ce sens, il est d'usage de dire que les SGBD permettent un traitement élémentaire des données.

Alors que le modèle OLTP est particulièrement efficace pour récolter et consolider les informations produites (bases de production), ce modèle transactionnel montre ses limites quant à l'analyse efficace et globale des données stockées. Partant de ce constat, les bases de données dites **multidimensionnelles** se sont alors imposées car elles pallient les limites des bases de données relationnelles. En effet, dans un contexte d'analyse décisionnelle concernant les données cliniques, le médecin expert a souvent besoin d'informations agrégées, résumées et observables sur plusieurs niveaux de précision. L'intérêt des médecins ne se situe plus que

dans l'analyse détaillée des individus mais également dans la recherche de tendances générales concernant un ensemble d'individus. Dès lors, un modèle de traitement OLTP pour réaliser de telles opérations est inadéquat car de nombreuses et coûteuses jointures seraient nécessaires et dégraderaient les performances du système.

En outres, toutes les caractéristiques propres aux données cliniques, abordées dans le premier chapitre, en font des données complexes. Pour en extraire la connaissance qui intéresse la recherche épidémiologique, il faudra trouver des méthodes d'analyse appropriées basé sur des systèmes exploitant efficacement les données cliniques pour des fins d'analyse et de prise de décision. Face à ce problème, la recherche épidémiologique fait recours à des systèmes d'aide à la décision spécifiques, basés sur l'approche multidimensionnelle **(OLAP)**.
En effet, contrairement aux modèles relationnels, entité/association ou orienté-objet, les modèles multidimensionnels sont les plus appropriés pour faire l'analyse et faciliter la prise de décision. Ils permettent d'observer des faits à travers des indicateurs (mesures) et des dimensions. Autrement dit, le modèle multidimensionnel se compose de **faits** contenant les *mesures* à analyser et de **dimensions** contenant les *paramètres* de l'analyse.
La modélisation multidimensionnelle est donc une technique qui vise à organiser les données de telle sorte que les applications OLAP soient performantes et efficaces.
Ainsi, les bases de données multidimensionnelles permettent d'analyser, de synthétiser les informations et représentent à ce titre une alternative intéressante aux systèmes relationnels classiques.

Le tableau N° 2.1 donne les principales différences entre les modèles OLTP et OLAP **[13]**.

	OLTP	**OLAP**
Utilisation	SGBD	Entrepôt de données
Opération typique	Mise à jour	Analyse
Type d'accès	Lecture / écriture	Lecture
Niveau d'analyse	Elémentaire	Global
Quantité d'information échangée	Faible	Importante
Taille de la base	Faible (quelques Go)	Importante (plusieurs To)
Ancienneté des données	Récente	Historique

Tableau N° 2.1 : Tableau récapitulatif des différences entres les modèles OLTP et OLAP

2.3. Référentiel de données cliniques, CDR (Clinical Data Repository) [20, 28]

Un dépôt ou référentiel de données cliniques (**CDR**) est défini comme une base de données (souvent relationnelle à base du modèle OLTP) en temps réel qui consolide les données d'une variété de **sources cliniques** afin de présenter une vue unifiée d'un seul patient. Il est optimisé pour permettre aux cliniciens de récupérer les données pour un seul patient plutôt que d'identifier une population de patients présentant des caractéristiques communes ou encore à faciliter la gestion d'un département clinique spécifique. Les données que nous trouvons dans un **CDR** comprennent : les résultats cliniques des tests de laboratoire, les données démographiques du patient, les données de son dossier médical, l'information sur les médicaments administrés, des rapports et images de radiologie, des rapports de pathologie, des résumés rédigés lors de l'hospitalisation du patient, et les notes d'évolution de ce dernier.

2.4. Entrepôt de données cliniques, CDW (Clinical Data Warehouse) [14, 20, 21, 22, 23, 24, 25, 26, 28, 30, 31]

Tous les travaux que nous avons examinés et qui distinguent le **CDR** de l'**entrepôt de données cliniques** *(Clinical Data Warehouse)* affirment que ces deux concepts ont comme même objectif, celui d'apporter aux décideurs une infrastructure pour intégrer les données permettant la prise de décision clinique. Ils expliquent, par ailleurs, qu'ils diffèrent dans leurs rôle et fonctionnalités.

Comme nous venons de le voir plus haut, le **CDR** contient des informations détaillées, centrées sur le *« Patient »*, mises à jour dans un environnement temps réel et organisées de façon à permettre la récupération rapide d'information le concernant. Les données tendent à être concentrées sur le plan clinique, fournissant ainsi aux cliniciens l'information dont ils ont besoin pour prendre des décisions notamment sur le traitement des patients.

L'entrepôt de données clinique **(CDW)** est par contre défini comme une collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse. En effet, il est optimisé pour l'analyse de telles données pour permettre la prise de décision et inclut les données cliniques, administratives et financières de l'institution de santé. Un entrepôt de données cliniques peut recevoir ses données à partir du référentiel de données cliniques (**CDR**), du dossier patient électronique

(**DPE**), des bases de données opérationnelles (**OLTP DB**) ou encore de toute autre source de données externes. Ce qui différencie l'entrepôt de données cliniques du **CDR**, c'est qu'il n'est pas centré sur le patient, mais les données sont plutôt combinées et organisées de telle façon à répondre à des questions spécifiques (traitement des tendances par exemple) ou pour fournir des retours (feedback) sur la performances/qualité au sein de l'organisation.

Les objectifs globaux d'un entrepôt de données sont identifiés comme suit:

1. L'entrepôt de données permet d'accéder à des données cliniques ou organisationnelles.
2. Les données de l'entrepôt sont cohérentes avec les données des systèmes sources.
3. Un entrepôt de données complet n'est pas seulement les données, mais aussi un ensemble d'outils pour interroger, gérer, analyser et présenter l'information.

Nous récapitulons dans le tableau suivant les principales différences entre les deux concepts **CDR** et **CDW** repris de l'article ***« Data warehouse and clinical data repositories »*** de *Alan Smith et Michael Nelson* **[28]**, la ressource la plus citée dans la bibliographie que nous avons étudiée :

Dépôt de données cliniques (CDR)	**Entrepôt de données cliniques (CDW)**
Les données orientées « détails », se focalise sur l'individu (un seul patient)	Les données sont agrégées, résumées au niveau décisionnel
Les utilisateurs peuvent lire et écrire dans la base (Accès en lecture/écriture)	Non volatile. Accès aux données en lecture seule
Mis à jour en temps réel à partir des systèmes opérationnels	Mis à jour périodiquement (statique) à partir des systèmes opérationnels
Données normalisées, pas de redondances	Les données sont parfois non normalisées, redondances de données.
Intégration de données cliniques	Intégration de données opérationnelles, cliniques et financières
Stockage des données dans leur plus récente forme. (dernières mises à jour de données)	Notion de temps, stockage des données et de leurs dates (permet d'exhiber les tendances)
Les données sont alimentées à partir des systèmes cliniques	Les données sont alimentées à partir des systèmes cliniques, financiers et administratifs.

Tableau N° 2.2 : Comparaison entre le CDR et l'entrepôt de données cliniques (CDW)

2.5. Magasins de données cliniques comme variante du CDW [18, 19, 20, 27, 28, 29, 30]

On note également, l'utilisation des « **Clinical Data Marts** », **CDM** ou « **magasins de données cliniques**» qui sont définis dans la littérature comme un sous-ensemble de l'entrepôt de données cliniques (voir le premier chapitre, section 1.3).

Les magasins de données cliniques contiennent des informations se rapportant à un contexte particulier de l'organisation de santé, souvent à un service donné de l'établissement hospitalier. Ils sont utilisés pour appréhender un nombre réduit de données cliniques (par rapport au **CDW**) et pour traiter des problématiques spécifiques pour les médecins et/ou cliniciens au sein du service en question. Ainsi leur utilisation peut procurer plus d'efficacité dans l'analyse de données, on a plus besoin d'explorer toutes les données contenues dans le **CDW** pour répondre à une analyse spécifique dans un service (on se réfère directement au sous-ensemble de données liées au service en question).

D'un autre coté, l'effort qu'on met pour la construction d'un **CDW** est beaucoup plus considérable que celui d'un **CDM**. A ce propos nous distinguons deux types de constructions d'un magasin de données cliniques selon les sources de données utilisées :

Magasin de données dépendant : il est alimenté directement à partir de l'entrepôt de données cliniques. Ce type de construction suppose que **CDW** existe déjà.

Magasin de données indépendant : il est alimenté à partir d'un ou plusieurs systèmes opérationnels (Bases de données, OLTP), de sources d'informations externes ou de données générées localement au sein d'un service (département) particulier.

Dans cette deuxième approche on construit l'entrepôt de données en rassemblant plusieurs magasins de données afin de regrouper l'ensemble de données relatives à l'organisation.

En effet dans les structures hospitalières, le processus de mise en oeuvre d'un entrepôt de données est généralement progressif. Souvent, on commence par la mise en place d'un datamart (un magasin de données relatif à un service/spécialité donné). Ensuite, deux types d'évolution sont possibles. Selon le choix organisationnel des managers, on se dirige soit vers la centralisation progressive des données stratégiques, soit vers autant de datamarts que de services spécifiques.

D'une manière générale, comparativement aux entrepôts de données, le magasin de données a les avantages suivants : cycle de construction court, moins d'investissements, faible risque.

Exemples de travaux utilisant les « Data Marts » pour les données cliniques

La majorité des travaux qui traitent de la récolte des données cliniques (plus généralement des données médicales) relatives à une maladie précise, emploient les « Data Marts » pour l'entreposage des données de la spécialité concernée. Ce choix est motivé souvent par le fait que le magasin de données se concentre sur les problématiques d'un contexte particulier directement lié à la maladie qui intéresse le médecin spécialiste. Ce magasin de données est alors mis en œuvre en analysant les besoins de l'expert. Nous présentons dans ce qui suit, quelques travaux qui utilisent un « Data Mart » dans un contexte clinique.

1) **Le magasin de données « Cardiovasculaire » dans le projet MAP ((Médecine d'anticipation personnalisée), [18] [19]**

L'entrepôt MAP est formé de plusieurs magasins interconnectés partageant les mêmes données sur les patients, les laboratoires, les médecins, etc. Chaque magasin contient également d'autres données de natures différentes (biologiques, biométriques, cardiovasculaires, psychologiques, etc.).
Le travail de [18], traite des problèmes de la modélisation multidimensionnelle de données complexes, en l'occurrence les données médicales du projet MAP (Médecine d'anticipation personnalisée). Une modélisation incrémentale a été proposée. L'idée principale était de modéliser le module le plus complexe **« Cardio-M »** dans l'entrepôt médical MAP afin d'extraire les différents concepts pour la création d'un méta modèle générique pour générer les autres modules de l'entrepôt MAP.

2) **Magasin de données basé sur le système d'information classique pour la chirurgie cardiaque, [20]**

Ce travail propose une approche « data mart » basé sur le système d'information hospitalier d'un institut cardiaque. Partant des données du **SIH** (de chirurgie, administrative, d'anesthésie et de laboratoire) ainsi qu'à partir de données de suivi et en utilisant un processus d'extraction, de transformation, et de consolidation, un magasin de données est construit.

Ce travail présente une étude comparative entre le **CDR**, **CDW** et le **Data Mart**. En se basant sur la comparaison de A. Smith et M. Nelson, 1999 déjà abordée, il donne les spécificités des « Data Marts » et leur bénéfice dans la recherche clinique.

3. Synthèse sur les modalités de récolte de données cliniques

Nous avons constaté en nous intéressant aux modalités de récolte des données cliniques que les systèmes d'informations cliniques sont en pleine mutation. Historiquement construits par des médecins cherchant à améliorer les outils informationnels à leur disposition, ils sont désormais des systèmes d'informations complexes, orientés processus, au cœur des systèmes hospitaliers.

Ainsi, nous avons vu que les systèmes d'information cliniques sont nés de l'informatisation des dossiers patients, longtemps considérés comme de simples archives informatisées des données médicales du patient, mais ayant engendré d'importantes évolutions dans le stockage et l'analyse des données cliniques.

Ces évolutions se sont accompagnées d'objectifs plus ambitieux que le simple stockage de l'information. Ces objectifs peuvent être résumés en deux grands axes : amélioration de la qualité des soins et amélioration de l'efficience de la production des soins. Ils incluent la qualité de la décision médicale, comme la prescription du traitement, mais également les activités de diagnostic et de pronostic. Ils incluent aussi la facilitation des prises en charge multidisciplinaires, avec des dossiers communs partagés (**centralisation du DPE**) et une meilleure transmission de l'information entre spécialistes (échanges de documents médicaux via des **standards**).

Au cours de notre étude, nous avons pu voir que l'utilisation du **modèle OLTP** dans les bases de données opérationnelles traitant des données cliniques s'avère insuffisant pour l'analyse de ses données, notamment à cause de leur nature particulière. La modélisation multidimensionnelle de ses données est alors employée souvent avec le **modèle OLAP**.

L'entreposage de données permet d'intégrer de telles **données complexes** et de les préparer pour des **analyses en ligne**, **statistiques** et/ou de **fouille de données**. Dans un contexte clinique lié à la recherche épidémiologique, nous avons identifiés différentes formes d'entreposage de données. En effet, **« les référentiels de données cliniques »**, **« les entrepôts de données cliniques »** et les **« magasins de données cliniques »** s'imposent comme

concepts prépondérants pour l'aide à la décision clinique mais avec des fonctionnalités différentes, le premier **(CDR)** est caractérisé par son orientation « Patient » utilisé souvent pour l'aide à la décision de la prescription médicamenteuse (traitement du patient), le second **(CDW)** s'oriente vers une prise de décision stratégique et managériale souvent transversale incluant outre les données cliniques, d'autres données organisationnelles de l'institution médicale. Les **« clinical Data Marts »** ou **« magasins de données cliniques »** s'apprêtent particulièrement aux données d'un service hospitalier. Ils peuvent inclure les données cliniques et d'autres données administratives. Leur construction rapide par rapport aux entrepôts de données favorise leur utilisation pour l'analyse de données liées à un contexte particulier et pour répondre à des problématiques spécifiques, ils sont très utilisés dans la recherche clinique notamment dans l'épidémiologie des maladies chroniques.

Dans l'étude que nous venons d'effectuer sur les modalités de recueil et de récolte de données cliniques pour la recherche épidémiologiques, et pour les travaux de fouille de données que nous allons aborder ultérieurement afin de prédire les facteurs de risques cardiovasculaires, le choix quant à l'organisation des données s'articule autours d'un « Data Mart » pour faciliter l'accès aux données.

Enfin, nous terminons cette partie avec un récapitulatif sous forme de schéma (figure ci-dessous) englobant l'organisation de données cliniques en montrant les interactions possibles entre les différents moyens de récoltes abordés.

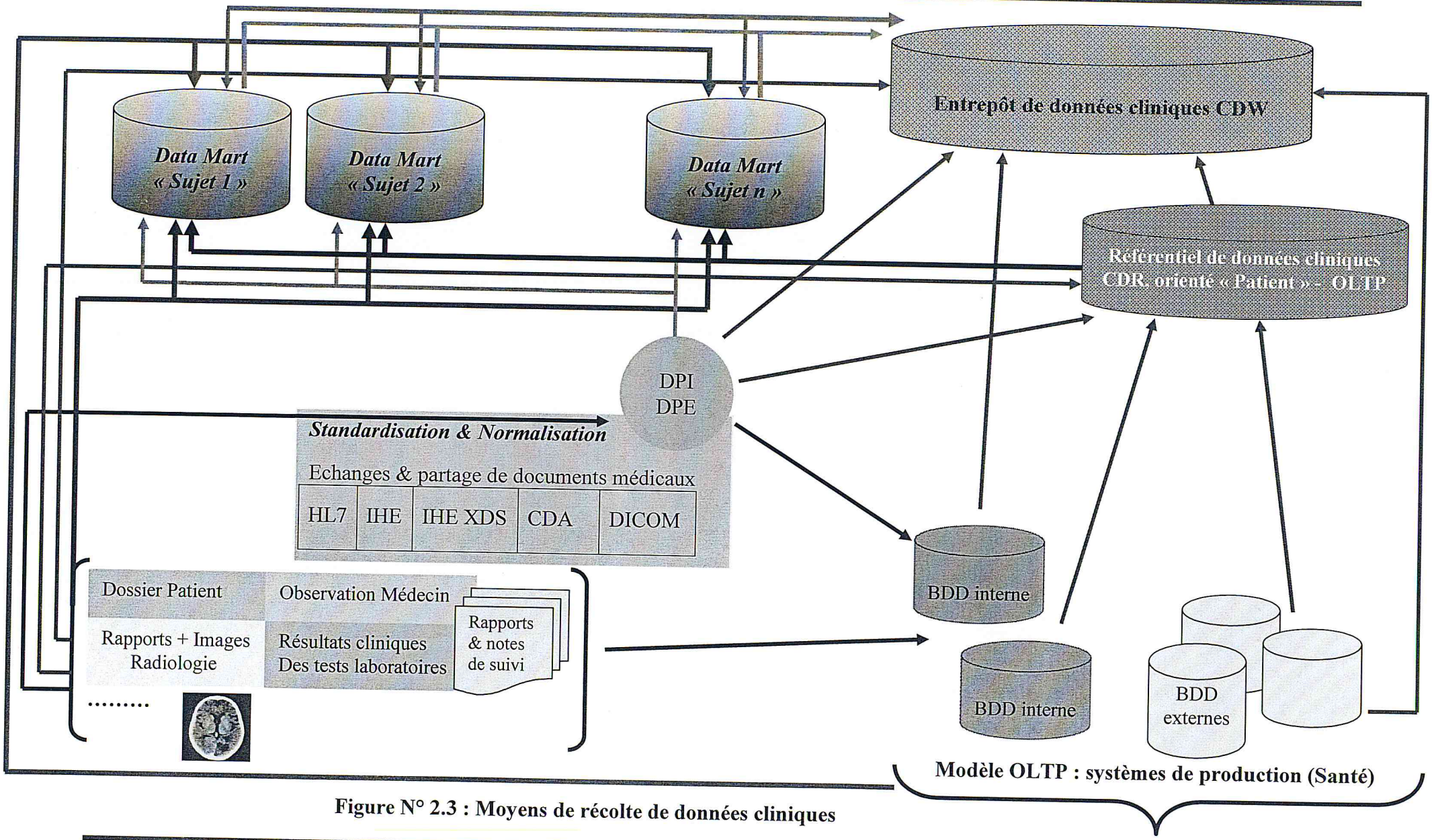

Figure N° 2.3 : Moyens de récolte de données cliniques

Conclusion

L'étude présentée dans ce chapitre traite de l'entreposage de données médicales en recherche épidémiologique et clinique, données qualifiées dans la majorité des travaux par une complexité due en partie à l'hétérogénéité et au volume important de ces données.

Plusieurs moyens de récolte et de recueil de données ont été identifiés pour appréhender les problématiques propres à ces données complexes. Ainsi, l'usage de dossier patient informatisé et de bases de données cliniques relationnelles présentent des avantages mais aussi des limites certains notamment en ce qui concerne l'analyse de données pour des fins d'aide à la décision. D'un autre coté, les référentiels de données cliniques sont une solution intéressante si la vision décisionnelle est orientée « Patient ». Nous avons pu voir aussi, en approchant de prêt, l'organisation des données dans un contexte clinique que l'entrepôt de données cliniques est organisé souvent sous forme d'une collection de magasins de données cliniques (DataMarts). Chaque magasin contient les données spécifiques, à une spécialité médicale. Les travaux s'intéressant à l'entreposage de données relatives aux maladies cardiovasculaires se sont ainsi logiquement orientés vert les magasins de données cliniques.

CHAPITRE III :

Extraction de connaissances et Data Mining dans le domaine cardiovasculaire

Introduction

Nous l'avons déjà vu, la médecine engendre un volume important de données. En grande partie, ces données restent inexploitables et l'on arrive peu à extirper les informations qui s'y cachent. Par ailleurs, l'**extraction de connaissances à partir des données** (ECD) est un processus complexe, qui vise à exploiter des techniques venant de différents domaines de recherches (intelligence artificielle, apprentissage automatique, statistique, analyse de données, visualisation d'informations, bases de données) pour l'exploration et l'analyse de données.

L'intérêt de la communauté scientifique pour le Data Mining est grandissant dans plusieurs domaines. L'industrie de santé s'oriente vers ces nouvelles techniques afin de tirer profit de la richesse des données en identifiant des corrélations et des associations permettant d'apporter des connaissances jusque là enfouies parmi l'ensemble des données. Nous allons dans ce présent chapitre présenter les concepts principaux de l'ECD, ensuite nous donnons une vue générale des différentes techniques de fouilles de données en apprentissage supervisé, notamment celles qui serviront de support à notre travail.

1. Extraction de connaissance [36]

L'extraction de connaissances à partir des données est un domaine qui prend de plus en plus de l'ampleur en raison principalement du besoin d'apprendre des grandes masses d'informations stockées et de la limite des techniques alors utilisées pour tirer parti du potentiel en terme de connaissances de données.

De toutes les définitions possibles de L'ECD (plusieurs définitions ont été données et varient selon le domaine de compétence de leurs auteurs), on peut retenir que l'ECD est : « un processus non trivial d'identification de connaissances, valides, nouvelles, potentiellement utiles et compréhensibles à partir de données » **[32]**.

La non-trivialité réfère au fait que, contrairement à la statistique qui est confirmatoire, la fouille de données est plutôt exploratoire **[33]**. En d'autres termes, avec le KDD « knowledge discovery in the databases » terme anglais utilisé pour l'ECD, on ne sait pas à priori ce qu'on pourrait apprendre des données. Cette non connaissance à priori caractérise les résultats mis à nus qui sont plutôt cachés (données nouvelles). La non-trivialité se justifie également par le fait que la découverte des connaissances passe par plusieurs étapes.

Les résultats d'une fouille de données devraient être non seulement utiles mais compréhensibles par les utilisateurs du domaine. En effet, les résultats devraient servir de support au processus de décision.

Une autre caractéristique essentielle de l'ECD est son utilisation sur de larges ensembles de données souvent issus d'entrepôts de données; large dans le sens qu'un entrepôt contient un volume important de données, et que ces données sont décrites par plusieurs attributs.

Enfin, il est important de noter que l'ECD est un processus; c'est-à-dire un ensemble d'étapes et d'actions dont la finalité est l'extraction de tendances et corrélations au sein des données. Contrairement aux idées reçues, l'ECD ne se limite pas exclusivement à la fouille de données qui en constitue toutefois la partie visible. Ce dernier se compose en effet d'un ensemble d'étapes allant de la compréhension du domaine d'étude, à l'exploitation des résultats de la fouille en passant par la fouille de données elle-même. Il est important de noter que le processus d'extraction de connaissances est particulier en ce sens qu'il est interactif et itératif **[32]**.

- Par itératif, il faut entendre que le processus d'extraction de connaissances n'est pas un processus linéaire où chaque étape est appliquée une seule fois pour aboutir à la fin au modèle de connaissance recherché. Cela pourrait être le cas dans le meilleur des scénarii possibles mais cela arrive assez rarement dans la pratique.
- L'ECD est un processus « human-centric » c'est-à-dire que l'Homme est au cœur du processus d'où le qualificatif interactif. Il est important de relativiser le qualificatif automatique à tord attribué à la fouille. En effet, comme le note **[35]**, les outils de fouille ne sont pas des robots qui seuls doivent parcourir de larges ensembles de données afin d'y extraire quelques informations utiles au domaine d'intérêt. Bien au contraire, il s'agit d'un ensemble d'interactions entre l'utilisateur et les outils de fouille afin que les résultats obtenus au bout du processus soient non seulement compréhensibles mais utiles.

1.1. Etapes de L'ECD

Longtemps réduit à la tâche de fouille de données, l'ECD a été érigée, comme on l'a déjà vu, en un processus regroupant diverses étapes allant de la compréhension du domaine et des données, au déploiement des résultats de la fouille de données.

Trois étapes principales peuvent être distinguées dans le processus de l'ECD : *préparation de données*, *application de techniques de data mining* et *interprétation des résultats*. Le procédé adopté par l'ECD est explicité par la figure ci-dessous :

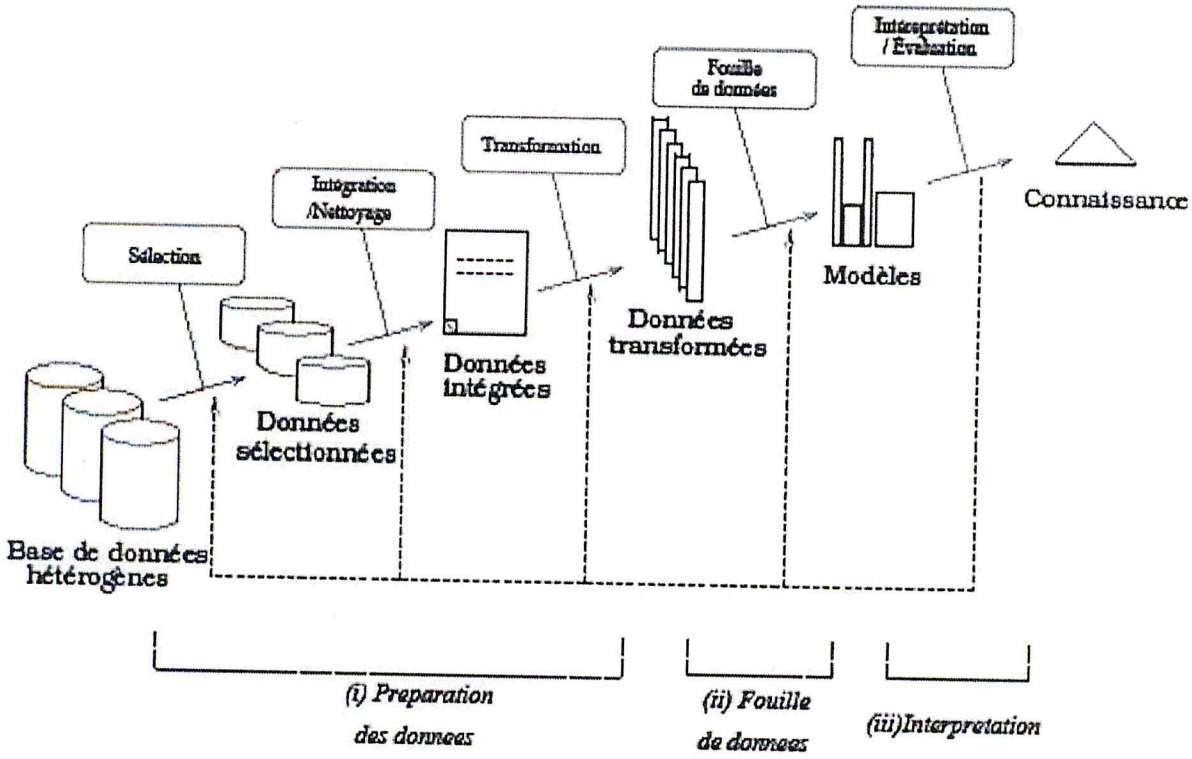

Figure N° 3.1 : Les étape du processus de l'ECD

1.1.1. Préparation des données [36]

Les données ne se présentent pas toujours de manière à ce qu'elles soient directement exploitables par la fouille de données. Le problème est surtout lié à : la présence des bruits (erreur introduite dans la mesure d'une variable) dans les données, la présence de valeur inconnue pour certains attributs (données manquantes), l'inconsistance des données, l'incohérence des données et le volume élevé des données.

La qualité des résultats d'un processus d'ECD dépend en grande partie de la qualité des données utilisées, d'où l'importance de l'étape de préparation de ces données **[32]**. Le prétraitement des données et donc leur préparation consiste en toute action effectuée sur les données avant l'application d'une technique de data mining (exemple : la classification automatique, le classement et les règles d'association). C'est essentiellement une transformation des données initiales en des données plus utiles. Cette transformation devra éliminer au moins un problème des données initiales tout en préservant l'information. La

phase de préparation des données est celle qui requiert le plus d'effort, environ 60 à 70% de l'ensemble du processus **[34]**. Elle regroupe essentiellement les opérations suivantes :

- **L'intégration des données :** ce processus permet de regrouper les données provenant de plusieurs sources et d'uniformiser le format.

- **Le nettoyage des données :** cette tâche consiste à détecter et à supprimer des erreurs, corriger les incohérences et supprimer les données inconsistantes, développer entre autres des stratégies pour traiter les valeurs manquantes et enfin retravailler les données bruitées, soit en les supprimant, soit en les modifiant de manière à en tirer le meilleur profit.

- **La transformation des données :** le but est de mettre les données sous un type plus approprié aux algorithmes de traitement, il s'agit entre autres d'uniformiser les données ayant la même sémantique. En effet, cette tâche consiste à extraire et/ou construire de nouvelles variables pour fournir une nouvelle représentation des données adéquates à l'application, au domaine et à l'objectif de l'étude. Elle intègre les opérations de :

  Normalisation : elle consiste à transformer les valeurs d'un attribut afin que celles-ci puissent appartenir à un domaine.
  Binarisation : elle consiste à transformer les données de type numérique (réel ou entier) ou symbolique (caractères ou énuméré) en données binaires.
  Discrétisation : elle consiste à transformer les données continues d'un attribut en valeurs discrètes.
  Agrégation : ou encore construction de caractéristiques, elle consiste à créer de nouveaux attributs à partir des données existantes en appliquant des opérations ou en définissant des fonctions.
  Projection : elle consiste à ajouter à chaque exemple un attribut supplémentaire (appelée projection) dont la valeur est calculée comme les sommes des carrés des autres attributs. Le but est d'éviter la dépendance du modèle à l'attribut dominant.

- **La réduction des données :** est couramment utilisée pour diminuer la dimensionnalité des données brutes. elle consiste à éliminer les redondances dans les données, regroupe ces données et fournit les moyens permettant de sélectionner les attributs pertinents et d'éliminer les attributs non pertinents. Ce processus d'optimisation est aussi appelé **sélection de caractéristiques**.

L'existence d'un entrepôt de données peut aider à diminuer sensiblement l'effort dépensé au niveau de cette phase. En effet, les données seront déjà passées par la phase d'Extraction-Transformation-Chargement avant d'être stockées dans l'entrepôt.

1.1.2. Data Mining

Cette étape est de loin la plus connue du processus global de découverte de connaissances et constitue le cœur de celui-ci **[32]**, étape sur laquelle bien des acteurs se focalisent d'emblé au détriment des autres étapes qui sont toutes aussi importantes.
La fouille de données en elle même consiste en l'utilisation des techniques et algorithmes afin d'extraire des connaissances implicites et potentiellement utiles, pouvant servir de support au processus de décision. En effet, la fouille de données vise à extraire des régularités (ou des irrégularités) de l'ensemble de données préparées **[37]**. Nous pouvons catégoriser ces techniques et algorithmes selon divers points de vue. Ainsi donc, vue sous l'angle de la finalité, ces algorithmes et techniques peuvent être regroupés en quatre grandes catégories que sont la classification, l'estimation, l'association et le clustering. Tandis que d'un autre point de vue, ces algorithmes et techniques peuvent être regroupés en méthodes supervisées ou non selon que l'algorithme ait besoin d'une connaissance préalable des données manipulées.
La fouille de données repose, pour une partie, sur une recherche exploratoire, dépourvue de préjugés concernant les relations entre les données. Pour cette phase, il s'agit essentiellement de choisir la/les technique(s) de fouille qui sied au problème posé et de calibrer efficacement les paramètres **[36]**.

Nous nous appesantirons davantage sur ce point, notamment pour la construction de modèles prédictifs, au niveau de la section dédiée à la classification supervisée.

1.1.3. Evaluation et interprétation

Il s'agit de l'évaluation de la qualité des résultats issus de la phase précédente. Au cas, où les résultats obtenus ne sont pas satisfaisants, on pourrait retourner dans une étape précédente. Cette phase est d'une importance capitale dans le processus d'ECD. Il ne s'agit pas seulement de produire des résultats ayant un certain niveau de précision mais il faut que ces résultats soient d'une interprétation facile pour les utilisateurs et les décideurs **[36]**.

Ainsi, l'étape d'interprétation du processus d'ECD (également appelée le post-traitement), consiste en la prise en charge des résultats bruts de la fouille de données, en leur transformation et validation en unités de connaissance **[38]**. Cette étape demande une implication de l'analyste (souvent un expert métier). La forme des résultats est différente selon la méthode de fouille utilisée : motif fréquent, concept formel, règle d'association, cluster par exemple.

Afin de faciliter l'interprétation, les résultats sont transformés pour faire l'objet d'une visualisation graphique. Les connaissances attendues par l'analyste orientent l'interprétation. L'analyste peut alors être amené à filtrer parmi les connaissances extraites celles qu'il juge triviales, redondantes, sans intérêt, fausses en comparaison de ce qu'il souhaite trouver.

1.2. Caractéristiques de la fouille de données

Dans cette section, nous présentons les différentes caractéristiques de la fouille de données. Notons toutefois que les différentes caractéristiques peuvent avoir des points communs dans la mesure où certaines techniques ou algorithmes possèdent plus d'une caractéristique **[36]**.

1.2.1. Les méthodes supervisées/prédictives

Il s'agit de méthodes qui requièrent la définition d'une variable cible dont on veut par exemple prédire la valeur.

On peut également subdiviser les méthodes supervisées/prédictives en deux classes d'algorithmes: les algorithmes de classification et ceux de régression. La classification consiste en la prédiction de catégories de valeurs discrètes, exemple : quelle sera la réponse d'un client à une offre ? La réponse pouvant être « j'accepte sans réserve », « j'accepte avec réserve », « je refuse », etc. Contrairement à la régression qui prédit plutôt des valeurs continues (numériques), exemple : quelle sera la valeur d'une maison ou le revenu d'une

personne. La classification et la régression diffèrent également quant à leur mode d'évaluation du résultat de la prédiction.

1.2.2. Les méthodes non supervisées/descriptives

Contrairement aux méthodes supervisées, les non-supervisées n'utilisent pas de cible. Elles fonctionnent plutôt sur la base de recherche de structures intrinsèques, de relations, ou affinités dans le jeu de données fourni en entrée. En d'autres termes, il s'agit de trouver des tendances et corrélations qui résument les relations entre données. La plus connue des tâches d'apprentissage non supervisé est le clustering. Le caractère descriptif de ces méthodes réside dans le fait qu'elles décrivent de manière concise et résumée un jeu de données en présentant les propriétés intéressantes de ces données.

1.2.3. Tâches de la fouille de données

- ***Classification - Estimation***

Le terme classification pose souvent une ambiguïté en raison de la confusion possible avec regroupement ou clustering. La classification est de loin l'une des tâches de fouille de données la plus utilisée car intervenant dans plusieurs domaines d'activité : l'attribution de crédit bancaire, la reconnaissance de gènes, la prédiction de sites archéologiques, le diagnostic médical, etc. La finalité d'une tâche de classification est d'assurer la prédiction d'un attribut cible nominal (catégoriel) sur la base d'une connaissance préalable des données qui leur sont fournies en entrée (données d'apprentissage).
L'estimation ressemble beaucoup à la classification à la différence que la variable cible est numérique au lieu d'être nominale comme dans le cas de la classification. Comme algorithme, on note la régression linéaire simple ou multiple. Les réseaux de neurones peuvent aussi être utilisés à cette fin.

Un exemple concret d'une tâche de classification est la prévision météo. En effet, pour prévoir le temps qu'il fera dans une zone géographique donnée, on prend en considération un certain nombre de paramètres nommés variables explicatives pouvant être l'humidité, la vitesse du vent, la température de la veille…et éventuellement la position de la zone géographique étudiée par rapport aux pôles. Ces paramètres ainsi que la variable à prédire (la

température par exemple) pourraient être modélisés dans une fonction. Chaque fois qu'on aura alors besoin de prédire la température, il suffira de passer à cette fonction ou modèle les paramètres appropriés pour obtenir les résultats désirés.

- ***Association***

Encore connue sous le nom d'analyse d'affinité, la tâche d'association en fouille de données vise à voir quelles sont les variables qui vont ensemble **[39]**. Il s'agit de trouver des règles du type si X alors Y avec un certain niveau de probabilité. Deux métriques sont utilisées pour caractériser généralement la qualité d'une règle d'association : le support et la confiance. Le support décrit la probabilité d'existence de X et Y au sein du jeu de données. La confiance décrit quant à elle la probabilité d'existence de Y dans l'ensemble de données contenant X.

À titre d'exemple, on peut mettre à profit les techniques d'association pour déceler les liens éventuels entre les différents produits vendus dans un supermarché. On peut ainsi noter que chaque fois que de la viande hachée est achetée, à 80% les pâtes sont également achetées. On note donc une certaine association entre les produits Viande et Pâtes avec 80 % de taux de confiance. On pourra entreprendre comme action de disposer le rayon « Pâtes » à proximité de celui concernant la « Viande » afin d'amener le client à ne pas fournir d'effort pour aller dans le rayon « Pâtes » ou que celui-ci se rappelle qu'il doit acheter des pâtes au cas ou il aurait oublié.

- ***Clustering***

L'idée de clustering renvoie tout simplement à l'utilisation de mesures de similarité (ou dissimilarité) entre les entités de sorte à regrouper ensemble celles similaires et celles dissimilaires dans un autre groupe. En d'autres termes, il s'agit d'une organisation des données en un ensemble de groupes homogènes : les clusters regroupés de telle sorte à minimiser la variance intra-classe et à maximiser celle interclasse.

À titre d'exemple, on peut grâce au clustering, se pencher sur l'étude de la typologie des étudiants – toutes années confondues – d'une université; avec comme question de fond : est ce que la qualité de l'enseignement a baissée ou augmentée ? Ou est ce plutôt le niveau des étudiants qui a baissé ou non?

1.3. Démarche statistique versus démarche de fouille de données

La fouille de données est un processus d'analyse dont l'approche est différente de celle utilisée en statistique. En règle générale, cette dernière présuppose que, dans un premier temps, l'on se fixe une hypothèse que, dans un second temps, les données vont nous permettre d'infirmer ou de confirmer celle-ci. Au contraire, la fouille de données adopte une démarche sans à priori et essaie de faire émerger, à partir des données brutes, des inférences que l'expérimentateur peut ne pas soupçonner, et dont il devra valider la pertinence **[39]**.

2. La classification supervisée pour le cardiovasculaire

Cette partie a pour objet la présentation de la classification supervisée, des méthodes de résolution des problèmes de classification supervisée incluant un état de l'art sur l'utilisation de ces méthodes dans un contexte médical, en l'occurrence pour la prédiction des facteurs de risque de maladies cardiovasculaires.

2.1. Classification supervisée [40]

Le mot *Classification* (en français) désigne à la fois la classification supervisée (classification en Anglais) et la classification non supervisée (Clustering en Anglais). Elle est dite automatique lorsqu'elle est faite par une machine, c'est-à-dire qu'elle est réalisée par un programme informatique. La classification est supervisée si :

Les données sont fournies sous forme de couple (x, y) où x est le vecteur d'attributs et chacun de ses composants a pris une valeur permettant de décrire un objet ou une situation, et y est l'étiquette de la classe de l'objet x.

L'objectif de la classification est de construire un modèle capable d'associer la classe y au vecteur x pour chaque couple (x, y).

Quant à la classification non supervisée, les valeurs du composant y des vecteurs ne sont pas fournies ; on cherche à construire un modèle qui regroupe les objets similaires par catégorie.

Notant que :

Les attributs du vecteur x sont appelés variables indépendantes ou variables explicatives.

L'attribut y est appelée variable dépendante ou étiquette de la classe cible.

L'objet désigné par le couple (x, y) est appelé observation, individu ou encore exemple.

Nous utilisons dans ce travail le terme classification pour désignée la classification supervisée.

Ainsi, nous pouvons aussi définir la classification comme étant le processus de recherche d'un ensemble de modèles (ou fonctions) qui décrivent et distinguent les classes d'objets; ces modèles construits à partir d'un ensemble d'objets de classe connue seront capables de prédire la classe des autres objets.

Les applications de la classification sont nombreuses, certaines touchent directement à notre vie quotidienne **[59]** :

1. Déterminer la viabilité d'un client sollicitant un crédit à partir de ses caractéristiques (âge, type d'emploi, niveau de revenu, autres crédits en cours …)
2. Pour une enseigne de grande distribution, cibler les clients qui peuvent être intéressés par tel ou tel type de produit.
3. Discerner les facteurs de risque de survenue d'une maladie cardiovasculaire chez des patients (exemples : l'âge, le sexe, le tabac, l'alcool …) ;
4. Quantifier le risque de survenue d'un sinistre pour une personne sollicitant un contrat d'assurance.

Un processus de classification comprend plusieurs étapes **[40]** : le prétraitement éventuel, la construction du modèle, l'évaluation du modèle construit et le classement. Le modèle défini peut être représenté sous diverses formes selon la technique de data mining utilisée : les réseaux de neurones, les arbres de décisions, les k-plus proches voisins, les réseaux bayésiens, les séparateurs à vaste marge (SVM), etc.

2.1.1 Prétraitement de données

Incluant entre autres les opérations d'intégration, de nettoyage, de transformation et de réduction de données (voir la section 1.1.1 dédiée à la préparation des données dans l'ECD)

2.1.2 Construction du modèle

La construction du modèle de classification est la phase principale d'une tâche de classification. Après le prétraitement, diverses méthodes supervisées peuvent être utilisées pour traiter ces données afin de proposer une aide à la prise de décision. Les algorithmes et techniques de classification diffèrent suivant le modèle construit pour classer les exemples et

la manière de le construire. Deux phases sont distinguées : l'apprentissage du modèle et sa validation. En effet, la construction du modèle se fait sur les **données d'apprentissage** qui doivent être distinctes des **données de test**.

Les données d'apprentissage et de test sont généralement issues à partir de la même base de données mais elles ne comprennent pas les mêmes données. L'apprentissage prend généralement 70% à 80% des enregistrements, la base de test étant constituée du reste **[37]**.

2.1.3 Validation du modèle

Plusieurs techniques utilisées pour diviser l'ensemble de données en ensembles d'apprentissage et de test sont proposées dans la littérature. Parmi ces techniques, nous pouvons citer :

1. La validation croisée d'ordre k. Elle consiste à partitionner les exemples en k sous-ensembles S1, S2, ..., Sk disjoints de taille égale ou presque; à l'itération i, le sous-ensemble Si est utilisé pour faire le test du modèle construit et appris avec les k-1 autres sous-ensembles. Afin d'évaluer le modèle prédictif, la moyenne des résultats des k tests est ainsi calculée.
Une variante de la validation croisée est le "leave-one-out" où k est égal à la cardinalité de l'ensemble d'apprentissage.

2. Le « holdout ». Contrairement à la validation croisée, l'ensemble des exemples est divisé aléatoirement en deux parties ; une partie permet de faire l'apprentissage du modèle et l'autre partie est utilisée pour tester le modèle. L'évaluation du modèle est celle obtenue sur l'ensemble de test.

3. La re-substitution. Le même ensemble est utilisé pour construire et pour tester le modèle. Elle est utilisée lorsque la taille de l'ensemble d'apprentissage est assez réduite.

2.1.4 Classement

Pendant l'étape de classement, le modèle obtenu à la phase d'apprentissage puis validé par la suite est utilisé pour prédire la classe de nouveaux exemples. Le classement est la finalité d'une classification supervisée.

2.2. Les techniques de classification

Parmi les techniques de classification supervisée les plus utilisées dans la littérature :

2.2.1. Arbres de décision

Les arbres de décision font partie des méthodes les plus populaires en apprentissage supervisé **[41]**. En effet, ils constituent une technique puissante de data mining, souvent utilisée pour des tâches de classification et de prédiction.
Un arbre de décision est une représentation arborescente d'une procédure de classification. Il permet de classer un objet selon les valeurs de ses attributs **[38]**. Dans cet arbre, un nœud représente un attribut, une feuille désigne une classe et les branches représentent un test sur l'attribut. La figure suivante constitue un exemple :

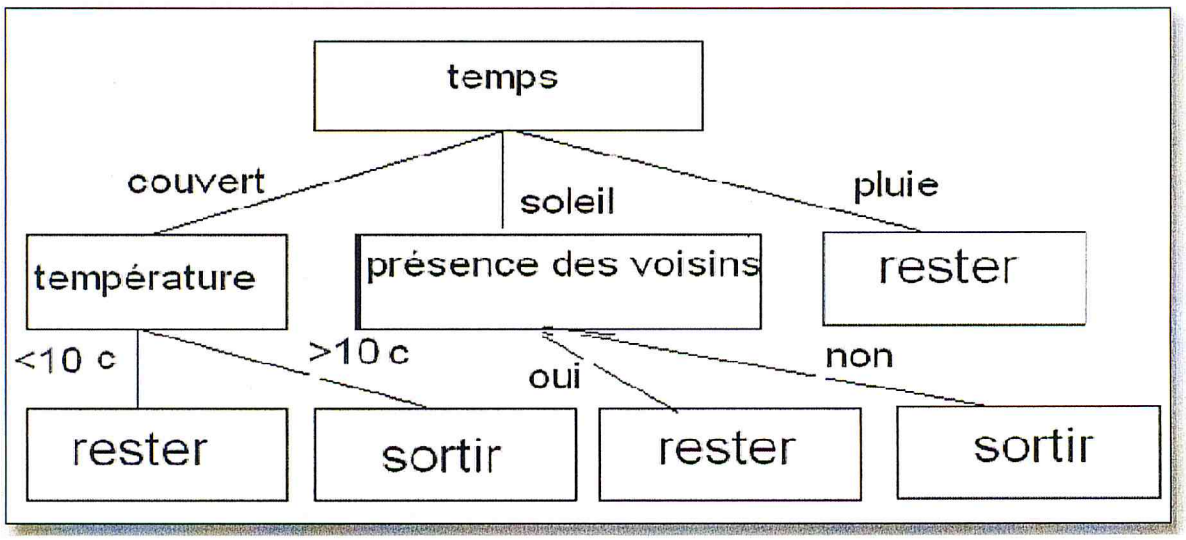

Figure N° 3.2 : Exemple d'un arbre de décision [38]

Pour classer un nouvel objet, il suffit de suivre le chemin partant de la racine à une feuille en effectuant les différents tests d'attributs à chaque nœud. La feuille où aura atterri cet objet, représentera sa classe.

La construction de l'arbre de décision se fait de façon récursive, en découpant successivement l'ensemble d'apprentissage E, comme suit **[38]** :

1. Si tous les exemples forment une seule classe, alors créer une feuille pour cette classe ;
2. Sinon, choisir le meilleur sélecteur (un attribut) qui représentera un nœud, puis :
- Fixer un test (le plus discriminant possible) sur cet attribut ;
- Découper l'ensemble d'exemples suivant ce test selon les valeurs possibles de cet attribut, ce qui représente les branches du nœud ;
- Pour chaque nouvel ensemble, construire un sous arbre de décision (refaire les étapes 1 et 2).

Il s'agit donc de trouver, à chaque nœud de l'arbre, la segmentation la plus intéressante au regard d'un critère donné. L'arbre est ainsi construit récursivement de la racine vers les feuilles. Les points les plus importants sont : le choix du critère de partitionnement, la détermination de la bonne taille de l'arbre, et l'assignation d'une classe à une feuille **[41]**.

Les principaux avantages qu'offrent les arbres de décisions sont : **[38, 42]**

- ✓ Applicables à tous types de données ;
- ✓ Ils sont efficaces avec un nombre de tests limité au nombre d'attributs représentés dans l'arbre ;
- ✓ Ils permettent de générer des règles de type Si-Alors, ce qui permet de justifier la classe associée à chaque objet et d'interpréter facilement les résultats obtenus;
- ✓ Les attributs apparaissant dans l'arbre sont des attributs pertinents pour le problème de classification considéré.

Ainsi, les problèmes de classification dans plusieurs domaines, notamment en médecine, sont souvent résolus par cette technique **[42, 43, 44, 45, 46, 47, 48]**.

Cependant, cette méthode rencontre quelques inconvénients, tels que **[38]** :

- ✓ Plus le nombre de classes est grand, plus l'arbre devient grand et plus la performance de ces algorithmes diminue ;
- ✓ Le bruit affecte la construction de l'arbre, ce qui peut donner un arbre non fiable ;
- ✓ L'arbre de décision devra être reconstruit (refaire l'apprentissage) si les objets initiaux changent, donc il est non incrémental.

Un autre inconvénient, est l'utilisation fréquente de variables moins pertinentes pour l'étape de construction de l'arbre engendrant le sur-apprentissage. Ce problème est néanmoins résolu par l'étape d'élagage qui consiste à supprimer les sous-arbres superflus ou trop liés aux données, dans le but d'améliorer l'aspect prédictif de l'arbre d'une part, et réduire sa complexité d'autre part **[49]**.

Plusieurs algorithmes ont été proposés pour la construction des arbres de décision. Un des premiers est CHAID qui utilise un critère de partitionnement basé sur la statistique du χ^2. L'ouvrage fondateur CART propose d'utiliser l'indice de Gini, et introduit la notion d'élagage. R. Quinlan proposera ensuite ID3 puis C4.5, basés sur l'entropie de Shannon, le premier avec le gain d'information, et le second avec le gain ratio **[49]**.
La différence principale entre eux est la façon dont ils choisissent le meilleur attribut. Selon **[47]**, les algorithmes ID3 et C4.5 sont considérés comme très performants : ils construisent rapidement l'arbre de décision qui prédit avec une assez grande fiabilité la classe de la nouvelle donnée.
Au regard des maladies cardiovasculaires auxquelles nous nous intéressons dans ce travail, nous avons constaté un usage presque systématique de cette technique de fouille de données pour la construction de modèles prédictifs.

L'article **[50]** traite des facteurs de risque de la pathologie cardiovasculaire et bâti un modèle basé sur l'algorithme C4.5 à partir de dix variables explicatives en exploitant 236 individus pour l'apprentissage, les plus importants facteurs identifiés par l'arbre sont : l'hypertension artérielle, diabète, consommation d'alcool, obésité abdominale et l'absence d'exercice physique (sédentarité). En conclusion, ce travail fait un constat important concernant d'abord les variables utilisées jugées insuffisantes pour avoir un modèle efficace pour la prédiction de la maladie. En effet, dix variables seulement sont utilisées et sont d'ordre général, à titre d'exemples pour l'hypertension artérielle, la variable est HBP (High Blood Pressure presence) avec deux valeurs possibles. Le nombre d'exemples pour l'apprentissage du modèle est aussi réduit, notamment le nombre d'individus ayant la maladie.
Les auteurs donnent des recommandations pour pallier à ces manques : s'intéresser à la fois au mode de vie des patients et à des mesures cliniques plus fines pour les variables à considérer, prendre en compte les aspects de co-morbidité et l'emploi de plusieurs techniques de classification puisqu'elles sont complémentaires pour identifier les facteurs de risque de maladie (se baser sur une seule technique n'est pas toujours pertinent).

Un autre travail **[65]** utilise quatre techniques de classification dont les arbres de décision avec l'algorithme C4.5 pour la prédiction de ces pathologies. Il fait également une comparaison des performances des méthodes utilisées. L'algorithme C4.5 se place en troisième position (après les réseaux de neurones, et SVM) avec un taux de précision de 0.82%.

Une autre étude comparative de modèles utilise trois techniques dont le C4.5 qui construit le modèle le plus performant avec une précision de 95.29 % **[42]**. Les variables explicatives utilisées sont plus nombreuses que les deux travaux précédents.

Ainsi, l'algorithme C4.5 est l'un des plus utilisés dans la littérature du domaine, d'autres travaux appliquent d'autres algorithmes : dans **[43]** par exemple, les auteurs s'intéressent à un facteur de risque significatif des maladies cardiovasculaires : l'hypertension artérielle et utilisent plusieurs techniques supervisées dont les arbres de décision avec deux algorithmes : CART et CHAID.

2.2.2. La régression logistique

La régression logistique a été utilisé comme technique de data mining appliquée au domaine médical dans plusieurs travaux comme **[52, 53, 54, 55, 56, 57]**.

La régression logistique fait partie des méthodes d'apprentissage supervisé et peut être utilisée quand la variable dépendante est catégorielle. En effet, elle est une généralisation de la régression linéaire. Elle permet ainsi, de calculer la probabilité d'un évènement à prédire comme une combinaison linéaire de variables explicatives.

La variable cible peut représenter par exemple le statut d'un patient : "ayant la maladie" ou " n'ayant pas la maladie". La formule du modèle logistique calcule donc la probabilité de la maladie sélectionnée y (y=0 si le sujet ne souffre pas de la maladie et y=1, le cas échéant) comme une fonction des valeurs des facteurs de risque prédictifs. Si le sujet souffre de la maladie, la probabilité conditionnelle P (y = 1 | X) = P (X) est donnée par :

$$P\,(y=1) = \frac{1}{1 + e^{-(\beta 0 + \sum_{i=1}^{n} \beta i\, xi)}} \qquad (1)$$

où xi, i = 1, n sont les variables indépendantes, βi sont les coefficients correspondants à la régression et β0 est une constante, qui contribuent tous à la probabilité.

L'équation précédente se réduit à un modèle de régression linéaire et le résultat positif de y, est donné par la formule suivante :

$$y = \log\left[\frac{P(X)}{1-P(X)}\right] = \beta 0 + \beta 1\, x1 + \beta 2\, x2 + \ldots + \beta k\, xn, \quad (2)$$

L'approche de la régression logistique sélectionne les k facteurs de risque significatifs (variables indépendantes contribuant significativement à la maladie) lors de la production du modèle qui se base sur le rapport de vraisemblance maximum **[45, 53]**.

En s'intéressant à la régression logistique dans le domaine médical, nous avons rapidement constaté le nombre important des travaux qui ont recours à cette technique. Le cardiovasculaire, en est un parfait exemple.
En effet, **[52]** s'intéresse à un facteur de risque majeur des évènements cardiovasculaires en Thaïlande, à savoir le diabète sucré. Ce travail utilise un modèle prédictif basé sur la régression logistique et montre que la combinaison de l'âge, l'IMC et la pression artérielle systolique pourrait aider à identifier les individus thaïlandais à haut risque de diabète non diagnostiqué. **[45]** traite entre autres de la dyslipidémie qui constitue un autre facteur cardiovasculaire, cette étude identifie par le modèle de régression logistique six variables significativement liées à la maladie parmi lesquelles : le cholestérol total, le sexe et les triglycérides. Enfin, **[43]** utilise cette méthode pour la prédiction des facteurs de risques de l'hypertension artérielle et compare les performances du modèle obtenu avec les autres techniques utilisées à savoir CHAID, CART, MARS et les réseaux de neurones.

2.2.3. Classificateur Naïf Bayes

Naïf Bayes est une technique utilisant des probabilités obtenues par le théorème de Bayes (Bayes, 1763; Laplace, 1820) avec de fortes hypothèses d'indépendance des données **[36]**. Cette méthode a fait ses preuves pour des problèmes d'apprentissage supervisé, nous citons à titre d'exemples en médecine, les travaux de : **[44, 47, 58]**.

Selon le théorème de Bayes, la probabilité qu'un exemple de données Xt appartienne à C est :

$$P(C|Xt) = \frac{P(C)P(Xt|C)}{P(Xt)}, \quad (3)$$

Sur la base de la formule ci-dessus, le classificateur calcule les probabilités conditionnelles qu'une instance appartienne à chaque classe, et sur la base de ces données de probabilités conditionnelles, l'instance est classée comme la classe de la plus forte probabilité conditionnelle **[42]**.

Dans le domaine cardiovasculaire plusieurs travaux appliquent cette technique pour la prédiction de facteurs de risque, nous citons dans ce qui suit quelques uns d'entre eux.
[42] par exemple utilise cet algorithme pour la prédiction des pathologies vasculaires cérébrales, il compare les résultats de l'évaluation du modèle à ceux du C4.5 et aux réseaux de neurones. **[44]** utilise aussi naïf bayes pour construire un modèle identifiant les facteurs de risque du diabète. Enfin nous mentionnons le travail de **[65]** qui utilise cette technique et la compare aux autres méthodes utilisées, en l'occurrence C4.5, SVM et les réseaux de neurones.

2.2.4. Règles de classification [60]

L'induction de règles de classification fait partie des approches « separate-and-conquer » qui produisent de manière incrémentale un ensemble de règles de la forme :

Si Condition **Alors** Conclusion

Où condition représente une suite de conjonctions de couples « **attribut-valeur** », et conclusion la **classe d'affectation**.
Concernant la construction des règles de classification, nous mentionnons deux approches possibles :
1. Méthode directe : consiste à extraire l'ensemble de règles directement des données. Nous citons à titre d'exemples : RIPPER et CN2.
2. Méthode indirecte : consiste à extraire l'ensemble de règles à partir d'un autre modèle de classification. Comme exemples, nous pouvons mentionner la génération des règles à partir des arbres de décision.

Ces règles peuvent ou bien être ordonnées (appelé souvent liste de décision) ou indépendantes :

2.2.4.1. Les listes de décision

La méthode produit une base de règles ordonnées. Lors du classement d'un individu, la première règle est évaluée. Si elle n'est pas déclenchée, on passe à la suivante, etc. Si aucune règle n'est activée, une règle par défaut est utilisée. Le modèle prend donc la forme suivante :

Si Condition 1 Alors Conclusion 1
Sinon Si Condition 2 Alors Conclusion 2
Sinon …
Sinon (Règle par défaut) Conclusion M

L'énorme avantage de cette représentation est qu'il ne peut pas y avoir de collision entre les règles, une et une seule sera activée lors du classement d'un individu.

2.2.4.2. Les règles indépendantes

Les règles indépendantes sont apparues peu de temps après les listes de décision, dans le but de faciliter leur *interprétation*. En effet, lorsque les règles sont ordonnées, pour lire correctement la règle n° i, nous devons tenir compte des (i-1) règles précédentes. L'interprétation devient difficile dès que le modèle contient un nombre important de règles. Pour ce deuxième type, les règles s'écrivent de la manière suivante :

Si Condition 1 Alors Conclusion 1
Si Condition 2 Alors Conclusion 2
…
(Règle par défaut) Conclusion M

Le classement d'un nouvel individu, se fait en testant toutes les règles du modèle. Si aucune d'entre elles n'est activée, la règle par défaut sera déclenchée.
Il arrive que parfois plusieurs règles, avec des conclusions contradictoires, peuvent être déclenchées dans ce cas on s'intéresse au support de ces règles afin de déterminer la classe de

l'exemple. La stratégie adoptée pour gérer ces collisions est celle que les auteurs (P. Clark and T. Niblet) proposent : si lors du classement d'un individu deux règles se déclenchent, nous effectuons l'addition des fréquences observées pour chaque règle, et nous affectons à l'individu la classe majoritaire. Nous citons ici les deux algorithmes les plus connus :

- **L'algorithme CN2 (Clark & Niblett, 1989) [61]**

CN2 (Clark & Niblett, 1989) est décrit comme l'un des algorithmes les plus performants pour l'induction des règles, il est capable de traiter des données imparfaites et bruitées **[62]**. Il est considérer comme un algorithme hybride puisqu'il incorpore les aspects de deux algorithmes ID3 et AQ pour générer ces règles. Il a été développé dans le but de combiner les avantages de ces deux algorithmes.
Ainsi, pour construire un ensemble de règles H, et pour une classe c, le système adopte la stratégie « separate-and-concquer ». Il construit une première règle R pour la classe c qu'il ajoute à H en utilisant l'ensemble des exemples E. il retire de E tous les exemples de la classe c couvert par la première règle et construit une deuxième règle à partir des exemples restants. Il itère ainsi jusqu'à ce que l'ensemble E ne possède plus aucun exemple de classe c. cette stratégie est adopter pour chacune des classes d'apprentissage.
Cet algorithme a fait l'objet de plusieurs travaux, notamment pour le diagnostic médical **[62, 63, 64]**.

- **RIPPER (Repeated incremental pruning to produce error reduction algorithm),**

Développé en **1995** par **Cohen**, **RIPPER** construit un ensemble de règles indépendantes selon toujours l'approche « separate-and-conquer », sa particularité c'est qu'il ajoute une heuristique de post élagage sur les règles, cette heuristique est appliquée au modèle comme une phase d'optimisation.

Parmi les travaux auxquels nous nous sommes intéressé et qui utilisent l'induction de règles pour le domaine cardiovasculaire, nous mentionnons à titre d'exemple l'étude effectuée par **[42]**, qui une fois le modèle optimal (se basant sur l'algorithme **C4.5)** est sélectionné, seize (16) règles prédictives sont extraites pour la prédiction de maladies vasculaires cérébrales.
Un autre travail **[65]** applique l'algorithme **RIPPER** dans une étude comparative incluant trois autres techniques (C4.5, SVM et les réseaux de neurones) pour la prédiction de maladies

cardiovasculaires. En termes d'efficacité du modèle, RIPPER se classe en deuxième position après SVM avec un taux de précision de 81,08%.

2.2.5. Méthode MARS

Friedman est le premier à introduire la méthode MARS *« Multivariate Adaptative Regression Splines »* en 1991 **[45]**. Il s'agit d'une méthode de régression non-paramétrique dans laquelle aucune hypothèse n'est faite quant à la relation fonctionnelle entre les variables dépendantes et indépendantes. Au lieu de cela, MARS construit cette relation à partir d'un ensemble de coefficients et de fonctions de base (FB), qui sont à leur tour fortement influencés par la régression des données. Donc cette technique se base sur les fonctions de base pour construire le modèle prédictif. Chaque fonction représente les informations contenues dans une ou plusieurs variables indépendantes. Parce que MARS est facile à utiliser et sélectionne les variables automatiquement, cette approche est largement utilisée dans plusieurs domaines, notamment pour des fins de prédiction en médecine **[45, 66]**. Le modèle général de MARS est donné par :

$$y = f\,(X) = \beta 0 + \sum_{m=1}^{M} \beta m\, hm\,(X), \qquad (4)$$

y est la variable à prédire au moyen de la fonction $f\,(X)$ qui se compose d'une constante initiale $\beta 0$ et la somme de M termes qui représentent le nombre de FB. En effet, chaque terme est constitué du coefficient βm et de la fonction de base $hm\,(X)$. $\beta 0$ et βm sont des paramètres qui fonctionnent comme les coefficients de régression linéaire.
La construction du modèle optimal se fait en deux étapes :

- D'abord un large modèle est définit englobant un nombre maximum de fonctions de base.
- Ensuite, le modèle est affiné par la suppression des fonctions de bases qui contribuent le moins à la performance global du modèle ainsi développé.
  Cette contribution est calculée par la validation croisée généralisée (GCV, Generalized Cross Validation). Si le retrait d'une FB réduit significativement l'erreur donnée par GCV, alors les variables définis par la fonction sont d'importants facteurs explicatifs et la BF ne peut être enlevée. GCV est donnée par :

$$GCV\ (M) = \frac{1}{n}\ \sum_{i=1}^{N}\left(yi - \hat{f}M\right)^2 \frac{1}{\left(1-\frac{C(M)}{N}\right)^2}\ , \qquad (5)$$

où C (M) est la fonction de coût de complexité à l'utilisation de Mième FB. GCV est donc la valeur résiduelle moyenne quadratique de l'ajustement aux données multiplié par une pénalité pour tenir compte de l'augmentation de la variance associée à l'augmentation de la complexité du modèle.

MARS est appliquée dans plusieurs travaux traitant de la prédiction de facteurs de risque des maladies cardiovasculaires. C'est le cas de **[43]** par exemple, qui est cité plus haut, à plusieurs reprises, puisqu'il procède à la comparaison de cinq techniques différentes y compris la méthode MARS. Une autre étude **[67]** s'oriente vers le concept du syndrome métabolique et l'influence de ce dernier dans la survenue de maladies cardiovasculaires, ce travail expose quelques limites du modèle CART et propose l'utilisation de MARS comme alternative. Enfin, nous citons le travail **[68]** qui s'intéresse à une affection cardiovasculaire connue, l'infarctus aigu du myocarde. Ce travail utilise plusieurs techniques : la régression logistique, les arbres de décision et enfin MARS.

2.2.6. Séparateurs à Vaste Marge (SVM) [65, 69, 70]

Les SVM (Support Vector Machine ou Séparateurs à Vaste Marge) (Vapnik, 1995) sont une méthode plus récente de classification supervisée. Ils ont pour objectif de rechercher le meilleur hyperplan de séparation des données en deux classes. La classification d'un nouvel individu x est donnée par sa position par rapport à l'hyperplan.
En effet, cette technique permet la classification à la fois des données linéaires et non linéaires. Le principe des SVM consiste à projeter les données de l'espace d'entrée (appartenant à deux classes différentes) non linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques de façon à ce que les données deviennent linéairement séparables (voir la figure ci-dessous). Dans cet espace, on construit l'hyperplan optimal séparant les classes tel que :

- Les vecteurs appartenant aux différentes classes se trouvent de différents côtés de l'hyperplan.
- La plus petite distance entre les vecteurs et l'hyperplan (la marge) soit maximale.

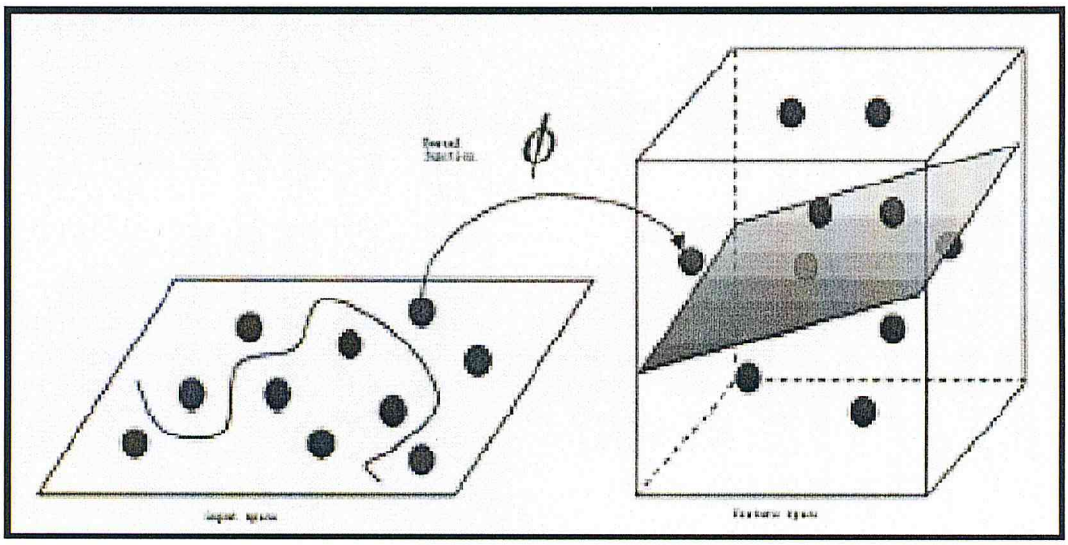

Figure N° 3.3 : principe de SVM pour séparer les données

Ainsi, la méthode SVM effectue des tâches de classification en maximisant la marge qui sépare les deux classes, tout en minimisant les erreurs de classification. Les SVM ont montré leur efficacité dans de nombreux domaines d'applications comme la reconnaissance de chiffres manuscrits, la classification de textes ou la bioinformatique et ce même sur des ensembles de données de très grandes dimensions. Mais leurs résultats ne sont pas facilement interprétables, ce qui constitue un inconvénient majeur. L'utilisateur sait qu'il peut classifier de manière efficace ses données grâce à l'hyperplan de séparation mais il est par exemple très difficile d'expliquer ce qui fait qu'un individu est dans la classe C1 plutôt que dans la classe C2.

Pour le domaine cardiovasculaire, tous les travaux que nous avons identifié plus haut : **[65]** à titre d'exemple ou encore **[51]**, utilisent la méthode SVM dans un contexte de prédiction de facteurs de risque mais ne montrent dans leurs articles respectifs que l'évaluation du modèle basée sur le résultat de la sortie du modèle, sans jamais discuter les facteurs significatifs de la maladie, le problème de l'interprétation des modèles SVM est alors clairement posé. En effet, d'autres travaux explicitent cette problématique et tentent de la solutionner, c'est le cas par exemples de **[69]**.

2.2.7. Réseaux de neurones [40, 71]

Les Réseaux de Neurones est une technique inspirée du fonctionnement du cerveau humain. Ils utilisent les liens entre les neurones pour apprendre à partir d'information observée. Cette méthode d'apprentissage en parallèle est robuste et résistante à l'introduction d'information erronée. Elle nécessite habituellement une bonne quantité de données d'apprentissage, ce qui peut être coûteux en temps machine. Mais, une fois le modèle entraîné, ces réseaux fournissent des prédictions suffisamment rapides pour fonctionner en temps réel dans un milieu industriel (en ligne de production par exemple). Enfin, les réseaux de neurones sont appropriés si la compréhension de la fonction apprise par le réseau n'est pas essentielle. Avec un arbre de décision par exemple, l'utilisateur peut toujours visualiser l'arbre et ***comprendre*** comment le modèle décide. Avec un réseau de neurone, des techniques de visualisation existent, mais elles demandent généralement plus d'expertise que l'analyse d'un arbre de décision.

En milieu médical, notamment pour la prédiction des facteurs de risque des maladies cardiovasculaires, beaucoup de travaux recourent à l'usage des réseaux de neurones, c'est le cas des travaux déjà cités plus haut **[42, 43, 51, 65]**. De même que pour les SVM, ces travaux s'intéressent uniquement à l'évaluation du modèle de prédiction, souvent dans une démarche comparative. Ainsi l'intérêt est porté sur le résultat de la sortie du réseau de neurones et pas sur les variables influençant le résultat de cette sortie.

2.3. Evaluation de modèles pour la prédiction de maladies

Les situations réelles d'apprentissage supervisé sont généralement des problèmes à deux classes, où l'une est la classe d'intérêt. Les individus de cette classe, généralement minoritaire, sont appelés individus positifs. Cette section présente quelques indices d'évaluation d'un modèle de prédiction dans le cas à deux classes. L'outil de base de l'évaluation d'un modèle est la ***matrice de confusion.*** Cette matrice croise la prédiction des individus avec leur classe réelle : tous les indices de performance d'un modèle sont basés sur cette matrice **[72]**.

Le tableau 3.1 montre un exemple de matrice de confusion, souvent utilisée dans la prédiction de maladie ;

Prédiction	Situation Réelle	
	Souffre de la maladie	***Ne souffre pas de la maladie***
Souffre de la maladie	(vrais positifs) VP	(faux positifs) FP
Ne souffre pas de la maladie	(faux négatifs) FN	(vrais négatifs) VN

Tableau N° 3.1 : Matrice de confusion pour la prédiction de maladie

Les indicateurs les plus importants pour l'évaluation de modèles sont **[42, 45]**:

- **Sensibilité et spécificité :**

La sensibilité et la spécificité, très utilisées dans le domaine médical, sont deux indices à observer conjointement.
Le taux de sensibilité est la probabilité qu'un individu positif soit effectivement classé positif par le modèle.

$$Sensibilité = \frac{VP}{(VP + FN)}\ , \qquad (6)$$

De même, le taux de spécificité est la probabilité qu'un individu négatif soit classé négatif par le modèle.

$$Spécificité = \frac{VN}{(VN + FP)}\ , \qquad (7)$$

- **Précision :**

Définit le taux total d'individus correctement classés.

$$Précision = \frac{(VP + VN)}{(VP + FP + VN + FN)}\ , \qquad (8)$$

Conclusion

Ce chapitre nous a permis de faire le tour du processus d'extraction de connaissance à partir de données (ECD). Plusieurs étapes le constituent, chacune d'elles est importante et contribue significativement à la pertinence de l'information extraite. Le prétraitement permet d'améliorer la qualité des données brutes et fournit aux algorithmes de Data Mining, des données plus adaptées à la nature de chaque technique utilisée. Le traitement de ces données peut se réaliser par des méthodes supervisées ou non-supervisées pour accomplir différentes tâches de fouille, il peut s'agir de classification, d'estimation, d'association ou encore de regroupement. Enfin, les résultats de cette fouille sont interprétés afin d'être traduit en unités de connaissances. Nous l'avons perçu dans la littérature du domaine, le succès de ce processus repose sur un pivot central : *la compréhension du domaine d'intérêt*, ainsi que sur un acteur indispensable : *l'expert métier*. Ces deux éléments représentent deux conditions impératives dans la conduite d'un projet de Data Mining pour espérer le réussir et apporter l'aide à la décision escomptée.

Nous nous sommes focalisés dans la deuxième partie de ce chapitre sur la prédiction des facteurs de risques de maladies cardiovasculaires. Cet objectif est un problème de classification supervisée. Une fois ce problème clairement défini, nous avons identifié plusieurs techniques permettant de construire des modèles prédictifs à savoir : les arbres de décision, la régression logistique, naïf bayes, l'induction de règles de classification, la méthode MARS, les séparateurs à vastes marges et enfin les réseaux de neurones. Un état de l'art complet sur l'utilisation de ces techniques pour la prédiction de maladies cardiovasculaires a été donné. Le nombre de travaux dans ce domaine est important, les facteurs de risques les plus souvent traités à part entière dans des articles dédiés sont : l'hypertension artérielle, le diabète sucré, la dyslipidémie ou encore le syndrome métabolique. C'est au regard de cette partie que nous allons décrire nos travaux dans le chapitre suivant dédié principalement à construire un modèle prédictif optimal pour prévenir la survenue d'un évènement cardiovasculaire en identifiant des patients à haut risque cardiovasculaire.

CHAPITRE IV :

Analyse et Application

Introduction

Les maladies cardiovasculaires sont la première cause de mortalité en Algérie. Ces pathologies sont à l'origine de 58% des décès, selon l'étude épidémiologique internationale Cepheus, à laquelle l'Algérie a participé en 2011. Cette étude a parfaitement mis en évidence la nécessité de développer une stratégie de prévention adéquate à cette situation.

L'hypertension artérielle et le diabète ont une prévalence croissante en Algérie. En tant que pathologies à l'origine de maladies cardiovasculaires graves, elles contribuent à une morbidité et mortalité considérables, elles ont par conséquent un lourd impact sur la vie d'un patient, a fortiori s'il y a coexistence des deux pathologies chez un même malade. Les problématiques liées aux maladies cardiovasculaires sont multiples, complexes, elles s'y réfèrent à la fois au diagnostic et à la prise en charge et occasionnent une énorme surcharge du système de santé algérien.

Afin de conduire une politique de santé basée sur le préventif, il est nécessaire de construire des modèles prédictifs à la fois pour établir le plutôt possible, un diagnostic précis, et dans le but de maitriser les facteurs de risque de ces pathologies en vu d'améliorer le traitement. L'objectif est aussi de permettre aux cliniciens de disposer d'outils d'aide à la décision exploitables sur les données cliniques qu'ils manipulent à différentes étapes de la prise en charge de leurs patients.

Le but de ce chapitre est alors de fournir une méthodologie utilisant le processus de l'ECD pour répondre à la problématique de notre travail, à savoir la prédiction des risques cardiovasculaires. Nous adoptons une approche hybride de data mining que nous appliquons et évaluons sur des données réelles en utilisant plusieurs techniques de fouille de données et en exploitant un environnement "Open Source" pour fournir un outil d'aide à la décision.

1. Démarche générale et problématique

Comme nous l'avons déjà mentionné, le but principal de notre travail est l'exploitation de données cliniques pour la prédiction de facteurs de risques cardiovasculaires dans un cadre de recherche épidémiologique. Nous voulons démontrer l'apport des techniques de fouille de données dans ce contexte en manipulant des données réelles issues d'étude épidémiologique algérienne.

Pour réaliser ce besoin, nous avons rapidement compris que nous ne pouvons pas procéder directement à la fouille de données. Ainsi notre travail est conduit par le procédé de l'extraction de connaissance à partir des données (ECD) dans sa globalité.

1.1. Démarche suivie

Notre premier choix concerne l'utilisation du modèle de référence **CRISP-DM** pour entreprendre les travaux de fouille de données auxquels nous nous intéressons. CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining, est une méthode mise à l'épreuve sur le terrain permettant d'orienter les travaux de Data mining.
Ainsi, nous organisons notre démarche en six composantes principales, incluant :

- ✓ La compréhension du domaine d'intérêt
- ✓ La compréhension des données utilisées
- ✓ La préparation des données
- ✓ La fouille de données
- ✓ Evaluation et interprétation des résultats
- ✓ Exploitation et déploiement des résultats

Chacune de ces composantes a été décortiquée pour donner lieu à plusieurs étapes permettant de fournir un résultat exploitable par une autre composante. Il est important de mentionner à ce stade que la démarche est aussi itérative : plusieurs interactions entre les composantes ont été opérées mais aussi des retours en arrière pour revoir des phases antérieures insatisfaisantes ou incomplètes.
Les deux premières composantes nous ont permis de situer la problématique d'une manière plus explicite et de définir des objectifs plus précis pour entreprendre la prédiction de maladies cardiovasculaires. Les trois composantes suivantes nous ont permis de construire un modèle prédictif optimal pour des patients à haut risque cardiovasculaire à partir de données réelles. Le déploiement se fait par l'outil open source "Orange 2.0", nous essayons de régénérer du code pour adapter cet outil à nos besoins spécifiques.

La figure ci-dessous montre le procédé que nous adoptons dans ce travail :

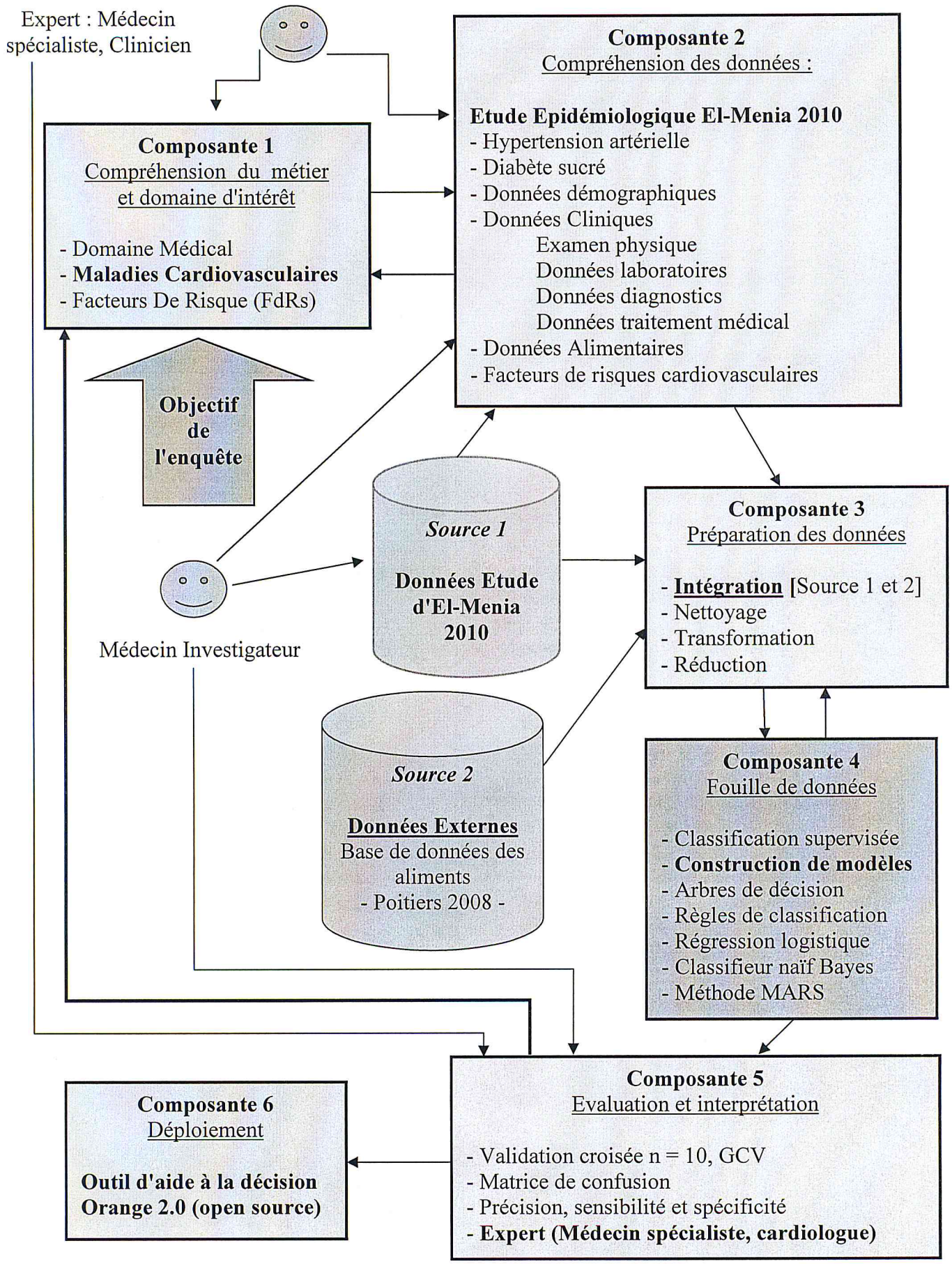

Figure N° 4.1 : Démarche suivie par notre travail basée sur CRISP-DM

1.2. Formulation de la problématique

Pour identifier clairement la problématique de notre travail, nous avons entre autres détaillé les deux premières composantes. Nous nous basons également sur les acquis du chapitre précédent, notamment sur l'état de l'art des modèles prédictifs traitant des facteurs de risque cardiovasculaire pour orienter notre travail.

1.2.1. Compréhension du domaine cardiovasculaire

Dans cette partie, nous avons étudié les maladies cardiovasculaires et leurs facteurs de risque dans la littérature médicale. Ce travail d'**analyse** concernant le domaine auquel nous nous intéressons nous a permis de :

- Identifier les différentes formes de pathologies définies par le terme cardiovasculaire
- Identifier la prévalence des maladies cardiovasculaires dans le monde ;
- Identifier la prévalence de ces pathologies en Algérie ;
- Situer les maladies cardiovasculaires comme un problème de santé publique dans le monde et notamment en Algérie ;
- Différencier la prévention primaire de la prévention secondaire des maladies cardiovasculaires ;
- Identifier les maladies associées au risque cardiovasculaire dans la littérature médicale ;
- Distinguer une association particulière « Hypertension artérielle (HTA) et diabète sucré » définissant une catégorie de patients à haut risque cardiovasculaire ;
- Identifier le besoin des médecins spécialistes de se baser sur des modèles prédictifs traitant des facteurs de risque cardiovasculaire.

Notre but n'étant pas de détailler les aspects purement médicaux, mais nous essayons de donner ici les éléments qui ont motivé l'orientation de notre travail.

En effet, les maladies cardiovasculaires, en tant que cause majeure de mortalité, et certaines pathologies (telles que : l'hypertension artérielle ou encore le diabète sucré) conduisant à des complications cardiovasculaires parfois irréversibles et fortement invalidantes, présentent des prévalences d'allure épidémique de part le monde entier. En 2005, L'OMS **[73]** rapporte des

chiffres inquiétants, 35 millions de personnes ont succombé à des maladies chroniques, ce qui représente 60% du nombre total de décès (58 millions) enregistrés cette année-là. Sur l'ensemble des décès liés à des maladies chroniques, 30% étaient dus à des maladies cardiovasculaires (MCV). On sait en outre que 80% des cardiopathies, des accidents vasculaires cérébraux et des diabètes de type 2, pourraient faire l'objet de mesures préventives efficaces.

L'Algérie en est au même constat, en participant à l'étude Cepheus, menée auprès de 1236 patients recrutés sur l'ensemble du territoire du pays, conclue dernièrement que 58% des décès sont à imputer aux maladies cardiovasculaires **[74]**. Elle rappelle, la forte charge que fait supporter, à de nombreux systèmes de santé, la prévalence de ces pathologies. Ainsi, elle souligne l'importance d'une prévention prenant en charge les facteurs de risques en vue de les contrôler et d'établir un diagnostic précoce et précis permettant une meilleure prise en charge du malade, d'autant plus que les évènements cardiovasculaires sont connus pour leur caractère éruptif, les rendant encore plus dangereux et difficilement décelables.
L'association de l'hypertension artérielle et du diabète pose au clinicien quelques problèmes particuliers. Ces deux pathologies constituent chacune de son côté un facteur de risque cardiovasculaire avec un effet cumulatif **[76,77]**. Là encore, les statistiques algériennes sont alarmantes, l'examen des différents types de maladies chroniques selon l'enquête MICS3 établie par l'office national des statistique (ONS) montrait déjà en 2006, que l'hypertension artérielle (HTA) occupait la 1ère place de ces maladies, la deuxième étant attribuée au diabète **[75]**. Depuis, ces chiffres ne cessent d'augmenter et ne font qu'étayer la nécessiter de s'appuyer sur des modèles prédictifs afin de prévenir les éventuelles complications cardiovasculaires observées chez ces malades, notamment ceux qui présentent une association des deux pathologies.

Concernant les modèles prédictifs pour exhiber les facteurs de risques de ces maladies, nombreux sont les travaux qui se sont intéressé à une pathologie spécifique, tel que nous l'avons détaillé dans le chapitre précédent. Par contre, les modèles traitant de facteurs de risques communs sont moins fréquents, nous n'avons retrouvé dans les travaux étudiés qu'une seule référence **[45]** qui traite communément de l'hypertension artérielle et de la dyslipidémie.

Ainsi, nous voulons démontrer dans cette étude l'apport des techniques de Data mining dans ce contexte particulier. Nous nous situons dans un cadre de prévention primaire, dans laquelle

on traite de patients n'ayant pas forcément une maladie cardiovasculaire avérée (c'est le champ de la prévention secondaire) donc notre but n'est pas de répondre à la question : ce patient a-t'il une maladie cardiovasculaire « x » ou pas ? L'intérêt est alors de traiter des malades définis par la médecine comme patients à haut risque cardiovasculaire. Notre choix s'est porté sur l'association « Hypertension artérielle – Diabète sucré » pour les raisons déjà citées, notamment le manque de modèles prédictifs traitant de deux pathologies concomitantes affectant le système cardiovasculaire, ainsi que pour d'autres raisons liées à l'objectif de l'enquête épidémiologique et aux données cliniques que nous traitons.

En effet, l'objectif principal de notre travail est de produire un modèle prenant en compte à la fois la prédiction de l'HTA et du diabète. Ainsi, au lieu de s'intéresser à chaque pathologie séparément, l'essentiel de cette étude est d'extraire les facteurs de risques individuels afin d'identifier les facteurs communs qui vont nous servir à construire le modèle prédictif traitant conjointement de l'HTA et du diabète puisqu'il est établi que le risque cardiovasculaire chez un patient souffrant de ces deux maladies est plus important que chez un individu avec une mono-pathologie.

1.2.2. Association de l'HTA et du diabète [76, 77]

En approchant le domaine médical, nous avons identifié l'association de l'hypertension artérielle et le diabète sucré, nous montrons dans cette section l'influence de ces deux maladies sur quelques pathologies cardiovasculaires pour étayer notre choix :
En effet, l'HTA et le diabète coexistent fréquemment dans la population générale. Ces deux pathologies représentent toutes deux des facteurs majeurs de risque cardiovasculaire et rénal. Par ailleurs, le diabète favorise la survenue d'une HTA, par divers mécanismes complexes, tandis que l'HTA est également reconnue comme un facteur de risque de survenue d'un diabète de type 2.
L'analyse des relations entre HTA et diabète démontre à la fois une grande complexité sur le plan physiopathologique et une hétérogénéité importante des situations rencontrées en pratique clinique.
Il est clairement identifié que le diabète et l'HTA sont les plus grands facteurs de risque d'athérosclérose. La prévalence de la cardiopathie ischémique est fortement augmentée dans le diabète et peut atteindre 55%, contre 2 à 4% dans la population en général. En présence à la fois d'un diabète et d'une HTA, le risque de celle-ci est nettement plus élevé qu'en présence

d'un seul de ces facteurs. Diabète et hypertension sont également les causes les plus fréquentes d'insuffisance rénale terminale. Un diabétique sur trois développera une néphropathie diabétique qui peut évoluer vers la néphropathie manifeste et est un marqueur de la présence d'une maladie cardiovasculaire. Avec l'hypertension, le risque d'insuffisance rénale augmente considérablement. En outre, le risque de maladie vasculaire cérébrale est multiplié par 1,5 à 4 dans le diabète tout comme dans l'hypertension. Chez les hypertendus, le diabète augmente de plus de 2 fois le risque d'accident vasculaire cérébral.

Par ces quelques exemples, il est évident que le patient diabétique et hypertendu est par définition un patient à haut risque cardiovasculaire.

1.2.3. Trois maladies liées au risque cardiovasculaire

Comme on l'a en partie mentionné plus haut, plusieurs maladies sont répertoriées comme facteurs prédisposant à des évènements cardiovasculaires.

Hypertension artérielle : L'hypertension artérielle est une élévation permanente de la pression du sang dans les artères au dessus des chiffres normaux. Une pression artérielle est dite « normale » lorsque la pression artérielle systolique (PAS) est inférieure à 140 mm Hg et la pression artérielle diastolique (PAD) est inférieure à 90 mm Hg. Il existe une relation continue entre le niveau de risque cardiovasculaire et les chiffres de pression artérielle, dès 115/75. En pratique, on peut se reporter aux valeurs seuils indiquées par la Société Européenne d'Hypertension, basées sur les définitions de l'Organisation Mondiale de la Santé (Tableau 4.1) pour matérialiser ce risque **[78]**.

Catégories	Valeurs seuils		
	Systolique		Diastolique
Optimale	< 120	et	< 80
Normale	120-129	et/ou	80-84
Normale haute	130-139	et/ou	85-89
HTA grade 1	140-159	et/ou	90-99
HTA grade 2	160-179	et/ou	100-109
HTA grade 3	≥ 180	et/ou	≥ 110

Tableau N° 4.1 : Classification de l'HTA, définition OMS

Diabète sucré : Le diabète sucré est défini par une glycémie à jeun supérieure à 1.26 g/l sur deux prélèvements successifs, les valeurs normales de glucose dans le sang étant de 0.7 à 1.1 g/l. Les patients ayant une hyperglycémie (dite intolérance au glucose) se situant entre 1.1 et 1.25 g/l doivent être pris en charge, à défaut 20 % d'entre eux développent un diabète dans les deux ans et quasiment 100 % dans les 10 ans à venir **[79]**.

Dyslipidémie : La dyslipidémie se réfère à des taux élevés de cholestérol total et/ou de triglycérides dans le sang. Ces derniers sont transportés au niveau plasmatique par deux principaux types de lipoprotéines : les lipoprotéines de faible densité (LDL) et les lipoprotéines de haute densité (HDL). Depuis 2005, plusieurs recommandations internationales ont proposé le concept de patients à « haut risque cardiovasculaire » pour lesquels on recommande une cible plus basse du cholestérol LDL. L'hyper-triglycéridémie est un facteur de risque cardiovasculaire, dont le rôle athérogène est intimement lié à la baisse du cholestérol HDL, qui lui est d'ailleurs souvent associée **[80]**.

1.2.4. Compréhension des données de l'enquête épidémiologique

Les données sur lesquelles nous appliquons notre méthodologie sont issues de l'étude d'El-Menia 2010 *« Prévalence de l'HTA & les facteurs de risque cardiovasculaires »* **[81]**. Ayant comme but principal l'évaluation de la prévalence de L'HTA dans l'oasis d'El Menia, les médecins cardiologues responsables de l'investigation, se sont également penchés sur la prévalence du diabète parmi la population étudiée. Ainsi, le lien entre les deux maladies a été fortement établi. L'objectif de l'enquête est aussi de décrire les facteurs de risques associés.
Cette enquête a été menée auprès de 726 individus et inclut outre les données démographiques quatre types de données :

- ✓ Les données issues de l'examen physique,
- ✓ Les données issues de l'examen de sang (laboratoires),
- ✓ Les données issues du diagnostic pathologique
- ✓ Les données des habitudes alimentaires.

Au total, la base de données inclut 133 items établis par le questionnaire qui a servi l'enquête, les médecins n'ont pu analyser qu'un sous-ensemble des données. Dans le prolongement des

résultats de cette enquête, l'ambition de notre travail est aussi de procéder à une fouille de données sur toutes les variables prises en charge y compris sur les données alimentaires.

Pour comprendre les données à manipuler nous avons étudié dans la littérature médicale plusieurs éléments, les plus importants sont :

- ✓ Obésité : l'indice de masse corporelle (IMC), l'indice de masse grasse (IMG), le rapport hanche-taille (RHT), le tour de taille et l'obésité abdominale.
- ✓ L'hypertension artérielle : la pression artérielle systolique (PAS), la pression artérielle diastolique (PAD), la fréquence cardiaque (FC) et les prescriptions médicamenteuses associées à l'HTA.
- ✓ Le diabète : notamment la glycémie et les prescriptions médicamenteuses pour la maladie.
- ✓ La dyslipidémie : le cholestérol total, l'LDL cholestérol, l'HDL cholestérol, les triglycérides et les prescriptions médicamenteuses associées.
- ✓ Néphropathie : créatinémie et fonction rénale.
- ✓ Syndrome métabolique : définit par une association de trois caractéristiques cliniques et métaboliques parmi les cinq suivantes :
    - Augmentation du périmètre abdominal
    - Elévation du taux de triglycérides
    - Diminution de l'HDL cholestérol
    - Hypertension artérielle
    - Augmentation de la glycémie à juin
- ✓ Alimentation : glucides, lipides, protéines, sel (sodium alimentaire) et nutrition équilibrée.

Dans cette partie, nous avons appris pour chaque variable étudiée, à distinguer les valeurs normales des valeurs anormales ainsi que l'influence de chaque variable sur le système cardiovasculaire.

En analysant les données de l'étude, pour chaque individu participant à l'enquête nous pouvons savoir s'il est hypertendu, s'il est diabétique, ou encore s'il a déjà eu un évènement cardiovasculaire (historique du malade) mais rien ne permet d'identifier s'il est atteint ou pas d'une maladie cardiovasculaire permanente. C'est encore une autre raison qui a guidé notre choix pour étudier l'HTA et le diabète comme facteurs de risque majeurs des évènements

cardiovasculaires et de produire un modèle prédictif traitant des facteurs de risques communs en se situant ainsi en prévention primaire de maladies cardiovasculaires.

1.2.5. Problématique et objectifs

À partir de tous les éléments étudiés dans cette partie, nous formulons la problématique de notre travail comme ceci :

- Il s'agit d'un problème de **classification supervisée** pour conduire une tâche de prédiction ;
- Les variables explicatives sont définies en partie par l'enquête épidémiologique, le reste est à introduire par **la composante de préparation de données** ;
- Deux variables dépendantes sont à prédire simultanément, toutes les deux catégorielles, donc il s'agit de classification et non de régression.

L'objectif de data mining que nous donnons est alors :

Construire un modèle prédictif optimal traitant des facteurs de risque communs à l'HTA et au diabète à partir des données de l'étude d'EL-Menia.

La démarche de fouille de données que nous suivons :

Notre démarche se décline en deux phases **[42, 45]** :

- Dans la première, nous appliquons plusieurs méthodes de classification (algorithmes capables de prédire une seule variable cible) pour identifier séparément les facteurs de risques de l'HTA et du diabète,
- Dans la deuxième phase, nous introduisons uniquement les facteurs de risque communs obtenus dans l'étape précédente et nous appliquons une technique fournissant deux variables cibles simultanément afin de construire le modèle prédictif traitant à la fois de l'HTA et du diabète.

Nous allons dans ce qui suit décrire les travaux que nous avons entrepris pour atteindre notre objectif. Il s'agit de détailler les trois composantes : « préparation des données », « fouille de données » et enfin « évaluation et interprétation ».

2. Conception du modèle prédictif

2.1. Approche hybride de data mining

La construction du modèle optimal est réalisée par une approche hybride de data mining en cinq phases :

Phase 1 : Prétraitement des données et sélection de variables;
Phase 2 : Construction d'un ensemble de modèles prédictifs par maladie, en utilisant cinq algorithmes de classification : C4.5, Classification Tree, régression logistique, naïf Bayes et CN2, deux modes d'entrée pour les attributs (sans et avec les attributs alimentaires) et deux jeux distincts de données (continues et normalisées).
Phase 3 : Evaluation des performances pour valider la pertinence des facteurs de risques individuels, choix de jeux de données à utiliser et déduction des facteurs communs.
Phase 4 : Construction de deux modèles MARS à partir des facteurs déduits lors de la phase précédente en distinguant les deux modes d'entrée pour les attributs.
Phase 5 : Comparaison des performances de classification et détermination du modèle prédictif optimal.

2.1.1. Choix des algorithmes utilisés

Nous avons utilisé cinq techniques de classification parmi les sept étudiés dans le chapitre précédent, à savoir : les arbres de décision, la régression logistique, naïf bayes, les règles de classification et la méthode MARS. Nous avons ainsi écarté les SVM et les réseaux de neurones (malgré leur efficacité reconnue dans plusieurs travaux) à cause de la difficulté dans l'interprétation de leurs résultats.
Pour les arbres de décision, nous utilisons deux variantes : C4.5 et classification Tree. Quant aux règles de classification, nous avons choisi CN2.
Concernant le traitement des facteurs communs, la fouille est réalisée par la méthode MARS qui supporte la prédiction de deux variables cibles.

2.1.2. Dichotomie des données : Apprentissage et tests

Pour apprendre les modèles, nous utilisons la validation croisée avec n = 10 pour distinguer les données d'apprentissage et les données de tests.

2.2. Procédure guidée par le data mining

2.2.1. Prétraitement des données et sélection de variables

Les données de l'étude d'El-Menia que nous exploitons, fournissent 726 enregistrements et 133 éléments d'enquête. Outre les données se rapportant à l'investigation (4 items) et les données démographiques des individus admis par l'enquête (7 items), nous disposons de :

- 27 éléments liés à l'examen physique.
- 25 éléments liés à l'examen de sang dont la glycémie, le bilan lipidique, la fonction cardiaque et la fonction rénale.
- 33 items du diagnostic pathologique dont 27 portants sur les prescriptions médicamenteuses.
- et enfin 37 éléments sur les prises des principaux repas ainsi que les habitudes alimentaires.

La préparation des données, étant cruciale dans un processus de fouille de données, nous avons par conséquent utilisé quatre techniques de prétraitement afin d'améliorer la qualité des données pour contribuer ainsi à améliorer la précision du modèle à construire. Ces techniques sont appliquées sur les variables, sur les observations (individus) ou les deux à la fois.

Nettoyage de données, nous avons utilisé les opérations de :

a- Remplissage de valeurs manquantes quand l'information disponible le permet par déduction à partir d'autres attributs.

b- Remplissage de valeurs manquantes par substitution (exemples : moyenne, maximum ou minimum selon les différents cas …) quand cela est pertinent.

c- Correction de valeurs aberrantes (éliminer l'incohérence)

Transformation de données, nous avons employé :

a- La *normalisation* pour discrétiser des données continues selon les standards et normes connus (connaissances du domaine médical), plusieurs catégories sont définies pour représenter : les valeurs normales (N), en-dessous des valeurs normales (L) ou supérieure aux valeurs normales (H) et parfois nous avons défini des paliers du type : H1, H2, H3

selon des interprétations médicales approuvées (Tableau N° 4.1, constitue un exemple de ce type).

b- L'*agrégation* de variables, nous a permis de construire 19 variables agrégées à partir de 57 items. Les attributs listés dans le tableau ci-dessous sont transformés à partir des éléments des examens physique et sanguin. Dix autres variables sont construites en suivant une procédure plus élaborée faisant usage des masses d'aliments (en appliquant aussi la technique d'intégration de données).

Variables agrégées	**Fonctions de transformation**
	Les items utilisés à partir de l'étude d'El-Menia
AGE	$date\ Examen - date\ naissance$
IMC	$Poids/(taille)^2$
RHT	$tour\ de\ hanche/tour\ de\ taille$
IMG	$(1{,}2 * IMC) + (0{,}23 * Age) - (10{,}8 * Sexe) - 5{,}4$ Sexe = 0 : Femme | Sexe = 1 : Homme (Deurenberg et al 1991)
TAB	$Tabac\ ou\ TabacCh\ ou\ (AntecédentTabac\ et\ ArrêtTabac \leq 3)$ TAB prend deux valeurs VRAI ou FAUX
SED	$Marche(1,2,3) + Sport(2: oui, 0: Non) + auto(0: V, 1: F)$
PAS	$\frac{\sum_{i=1}^{k} PASi}{k},\ k = 3\ à\ 6\ (selon\ cas)$
PAD	$\frac{\sum_{i=1}^{k} PADi}{k},\ k = 3\ à\ 6\ (selon\ cas)$
FC	$\frac{\sum_{i=1}^{k} FCi}{k},\ k = 3\ à\ 6\ (selon\ cas)$

Tableau N° 4.2 : Variables agrégées à partir des données de l'examen physique et l'examen de sang

Intégration de données, nous avons utilisé comme source externe, la banque des aliments (académie de Poitiers, disponible sur internet) afin d'exploiter les informations sur la masse des aliments organiques et minéraux pour estimer en moyenne la consommation journalière d'un individu en terme de : glucides, lipides, protéines et sodium alimentaire (NA). Ainsi, nous avons construit dix variables alimentaires, la procédure de calcul nécessite plusieurs opérations et fait recours à la fois à des opérations d'intégration (plusieurs sources de

données), de transformation (fonctions de calcul) et de réduction de données (sur la dimensionnalité).

Les variables alimentaires que nous avons construits sont présentées dans le tableau suivant :

Variables agrégées	Colonnes utilisées (source locale)	Données utilisées (source externe)
Glucide végétal Lipide végétal Protéine végétal	Consommation de pain Quantité de pain consommée par jour Consommez-vous les légumes ? Citez les plus consommés Consommez-vous les fruits ? Citez les plus consommés Nombre de date par jour ? Consommez-vous les féculents ? Citez les plus consommés	Masse des aliments : pain, carotte, courgettes, haricot banc, fèves, oignons, abricot, raisins, dattes … Exemple : 100 g de pain : 56g de glucides 1 g de lipides 7,5 g de protéines 650 mg de NA
Glucide animal	Consommez-vous des œufs ? Œufs consommés par semaine ? Consommez-vous le lait ? Combien buvez-vous ? Consommez-vous le petit lait ? Combien buvez-vous ? Consommez-vous les yaourts ? Combien de pots par semaine ? Consommez-vous le beurre ? Combien ? Consommez-vous du fromage ? Portions ? Nombre portions ? Klila ? Fondu ?	Masse des aliments : œufs, lait, petit lait, yaourt, beurre, fromage fondu, fromage en portions … Exemple : 100 g d'œuf : 0,3 g de glucides 10,5 g de lipides 12,5 g de protéines 133 mg de NA
Lipide animal Protéine animal	La viande consommée le plus ? Quantité de viande par jour ? Consommez-vous des œufs ? Œufs consommés par semaine ? Consommez-vous le lait ? Combien buvez-vous ? Consommez-vous le petit lait ? Combien buvez-vous ? Consommez-vous les yaourts ? Combien de pots par semaine ? Consommez-vous le beurre ? Combien ? Consommez-vous du fromage ? Portions ? Nombre portions ? Klila ? Fondu ?	Masse des aliments : ovin, chameau, poulet, œufs, lait, petit lait, yaourt, beurre, fromage fondu, fromage en portions … Exemple : 100 g d'œuf : 0,3 g de glucides 10,5 g de lipides 12,5 g de protéines 133 mg de NA Ou encore 100 g de poulet : pas de glucide 6,8 g de lipides 20,8 g de protéines 88 mg de NA

Tableau N° 4.3 : Variables agrégées à partir des données alimentaires et par intégration d'une source externe

Quatre autres variables sont agrégées comme suit :

- Glucide global : somme de glucide végétal et animal ;
- Lipide global : somme de lipide végétal et animal ;
- Protéine global : somme de protéine végétal et animal ;
- Enfin NA total qui est construite à partir de toutes les colonnes citées ci-dessus.

Réduction de données, nous avons opérer comme suit :

a- Suppression des items non significatifs pour l'objectif de l'étude ou à cause du nombre important de leurs valeurs manquantes.

b- Sélection de caractéristiques : le choix de chaque variable repose sur l'analyse des pratiques médicales et cliniques ainsi que le recours à des médecins spécialistes.

c- Suppression des observations incluant un nombre important de variables non disponibles (valeurs manquantes après nettoyage et transformation).

Le prétraitement nous a permis de valider 488 individus et 11 variables physiques, 9 variables de sang, 16 variables alimentaires et 4 variables pathologiques. Le tableau suivant liste l'ensemble de ces variables ainsi que les abréviations utilisées.

Variables diagnostic pathologique		*Variables de l'examen physique*		*Variables examen de sang*		*Variables alimentaires*	
Attributs	*Code*	*Attributs*	*Code*	*Attributs*	*Code*	*Attributs*	*Code*
Hypertension artérielle	HTA	Age	AGE	Cholesterol total	CHOL-T	Glucide végétal	GluV
Diabète sucré	DIAB	Sexe	SEX	Cholestérol de haute densité	LDL	Glucide animal	LipV
Dyslipidémie	DYSL	Situation de famille	SM	Cholestérol de basse densité	HDL	Glucide global	GluG
Maladie cardiovasculaire	MCV	Niveau intellectuel	NI	Triglicérides	TRIG	Lipide végétal	LipV
		Indice de masse corporelle	IMC	Glycémie	GLYC	Lipide animal	LipA
		Rapport hanche Taille	RHT	Créatinemie	CREA	Lipide global	LipG
		Indice de masse graisse	IMG	Pression artérielle systolique	PAS	Protéine végétal	ProV
		Tour de taille	TRT	Pression artérielle diastolique	PAD	Protéine animal	ProA
		Couleur	CLR	Fréquence cardiaque	FC	Protéine global	ProG
		Tabagisme	TAB			NA total	NA
		Sédentarité	SED			Epices	EPIC
						Type eau	TE
						Quantité Eau été	QEE
						Quantité Eau hiver	QEH
						Quantité Thé Jour	TJR
						Quantité Café Jour	CJR

Tableau N° 4.4 : Sélection de 40 variables et codes utilisés

2.2.2. Modes d'entrée des attributs

Les 40 variables sélectionnées sont divisées en deux groupes :

- G1 : variables de l'examen physique + variables de l'examen de sang + variables du diagnostic pathologique ;
- G2 : variables de l'examen physique + variables de l'examen de sang + variables alimentaires + variables du diagnostic pathologique.

2.2.3. Jeux de données

Nous utilisons deux jeux de données distincts au sein de G1:

- J1 : Données continues pour les variables quantifiables ;
- J2 : Données normalisées par transformation pour ces mêmes variables. La normalisation guidée par l'expertise médicale dans ce cas, permet d'introduire implicitement la connaissance médicale dans les données.

La classification se fait en deux itérations en utilisant tour à tour les deux jeux de données (les tests empiriques nous ont permis de juger l'effet complémentaire des jeux de données ainsi utilisés).

Pour G2, les données continues sont maintenues pour deux raisons :

- La discrétisation de la consommation en glucides, lipides et protéines se base sur le sexe, l'âge, la corpulence et l'activité de l'individu. En l'absence d'information détaillée sur l'activité et l'effort journaliers de la population étudiée, nous n'avons pas normalisé les variables alimentaires
- Les comparaisons de l'efficacité des modèles de G1, nous permettent de choisir le jeu de données continues (J1)

2.2.4. Modèles prédictifs par maladie

Nous avons utilisé la validation croisée avec (n=10) pour déterminer les données d'apprentissage et les données de tests, avec les cinq algorithmes de classification supervisée : C4.5, classification Tree, régression logistique, naïf bayes et CN2. 30 modèles prédictifs (15 par maladie) ont été construits en trois sous-groupes :

A1 : En entrainant les données d'apprentissage avec le mode G1 et le jeu de données J1 par les cinq algorithmes : 5 modèles prédictifs ont été produits pour chaque maladie.
A2 : En entrainant les données d'apprentissage avec le mode G1 et le jeu de données J2 par les cinq algorithmes : 5 autres modèles prédictifs ont été produits pour chaque maladie.
A3 : En entrainant les données d'apprentissage avec le mode G2 et le jeu de données J1 par les cinq algorithmes : nous avons élaboré 5 autres modèles pour chaque maladie.
L'évaluation de ces modèles se fait à chaque fois par l'usage des données de tests en validation croisée.

2.2.5. Facteurs de risque par maladie

L'extraction d'un ensemble de facteurs de risque par maladie se fait par les règles suivantes :
1. Une variable est déclarée comme facteur de risque de la maladie dans un sous-groupe (A1, A2 ou A3) si elle est au moins sélectionnée par trois techniques différentes.
2. L'ensemble des facteurs de risques individuels pour une maladie est la combinaison des facteurs identifiés au sein des trois sous-groupes A1, A2 et A3, vu leur rôle complémentaire.

2.2.6. Facteurs communs aux deux maladies

Deux listes de facteurs commmuns pour l'HTA et le diabète sont répertoriées (FCG1 et FCG2) en considérant les modes G1 et G2 pour l'intégration ou pas des facteurs alimentaires. Nous avons voulu démontrer l'influence de l'alimentation sur l'association des deux maladies.
Suivant des considérations médicales et diététiques ainsi que les précaunisations de médecins spécialistes, FCG2 s'est vu augmenté par trois facteurs de risque supplémentaires (NA,ProA et GluG). En effet, ProA et GluG représentent des facteurs individuels très importants (ProA pour l'HTA et GluG pour le Diabète), selectionnés par quatre techniques parmi les cinq utilisées. Enfin pour NA, l'ONS évoque l'importance de cet élément pour les deux pathologies.

2.2.7. Modèles prédictifs pour l'association HTA-Diabète

Deux modèles prédictifs ont été construits par la méthode MARS avec les éléments suivants :
- Les variables indépendantes sont tour à tour les facteurs de risque FCG1 et FCG2;
- Les variables dépendantes sont : HTA et DIAB. Nous avons définit la variable "maladie" qui représente l'associaction des deux maladies, ayant les valeurs suivantes :

maladie = 1 : si HTA = DIAB = FAUX

maladie = 2 : si HTA = VRAI; DIAB = FAUX

maladie = 3 : si HTA = FAUX; DIAB = VRAI

maladie = 4 : si HTA = DIAB = VRAI

2.2.8. Performance de classification et modèle optimal

Les même paramètres ont été définit lors de la construction des deux modèles MARS avec un seuil de 0.0005, le même nombre maximum de fonctions de base et le même degré d'interaction ont été fixés.

L'évaluation des modèles est basée sur la matrice de confusion, pour calculer la précision, la sensibilité et la spécificité des modèles. La comparaison nous permet de déterminer le modèle prédictif optimal et de sélectionner également le meilleur mode d'entrée pour les attributs.

2.3. Tests et résultats

Cet étude utilise l'outil open source "Orange 2.0" pour la construction des modèles prédictifs des facteurs individuels des maladies. Les facteurs communs sont traités par le module MarsSplines de Statistica 8.0.

Après prétraitement, les données incluent 189 hypertendus, 43 diabétiques et 119 individus souffrant simultanément de l'HTA et du diabète.

2.3.1 Facteurs de risque individuels par maladie

Le tableau ci-dessous montre le résultat de l'extraction des facteurs individuels par maladie, correspondant aux étapes 2.2.2 à 2.2.5 de la procédure adoptée pour la construction du modèle optimal, décrite dans la section précédente :

Dans ce tableau, les cinq techniques de data mining utilisées lors de ces étapes, C4.5, classification Tree, régression logistique, naïf bayes et CN2 sont notées respectivement : A, B, C, D et E. Ce tableau cite pour chaque facteur de risque (FDR) le nombre de techniques qui le référencent noté NBTech (règle adoptée : au moins trois sur cinq soit 60% des techniques utilisées) et lesquelles parmi l'ensemble des 15 modèles construits par maladie.

FDR	Techniques Data mining faisant référence au FDR			
	HTA		DIAB	
	NBTech	Techniques	NBTech	Techniques
AGE	5	A, B, C, D et E	4	B, C, D et E
NI			3	A, B et C
IMC	4	A, C, D et E	5	A, B, C, D et E
IMG	3	C, D et E		
TRT	4	B,C,D et E	4	A, B, D et E
SED	4	B,C,D et E	5	A, B, C, D et E
PAS	5	A, B, C, D et E	5	A, B, C, D et E
PAD	5	A, B, C, D et E	5	A, B, C, D et E
FC	4	A, C, D et E	5	A, B, C, D et E
CHOL-T	5	A, B, C, D et E	4	A, C, D et E
HDL	3	A, Bet E	3	A, B et E
LDL	5	A, B, C, D et E	4	A, C, D et E
GLYC	5	A, B, C, D et E	5	A, B, C, D et E
TRIG			5	A, B, C, D et E
CREA	4	B, C, D et E	5	A, B, C, D et E
GluV	3	A,C et E		
GluG			4	B, C, D et E
LipG	4	B, C, D et E	4	B, C, D et E
ProV			3	B, C et E
ProA	4	C, D et E		
EPIC	3	A, C et D		
TE			3	C, D et E
HTA			4	A, C, D et E
DIAB	3	A, D et E		
DYSL			3	B, D et E
MCV	4	A, C, D et E		

Tableau N° 4.5 : Extraction des facteurs individuels

- ***Facteurs de risque de l'HTA :***

Les facteurs de risque de l'HTA identifiés par les premières étapes de l'approche hybride adoptée sont donc : AGE, PAS, PAD, CHOL-T, LDL, GLYC, IMC, TRT, SED, FC, CREA, LipG, ProA, MCV, IMG, HDL, GluV, EPIC, DIAB.

Ainsi, à partir de 39 variables indépendantes (et une variable cible) , les modèles construits ont permis d'extirper 19 facteurs déterminants pour prédire la maladie.

- ***Facteurs de risque du diabète :***

De même, 20 facteurs déterminants pour prédire le diabète ont été identifiés par le même procédé d'extraction à partir de 39 variables indépendantes (et une variable cible), ces facteurs de risque sont : IMC, SED, PAS, PAD, FC, GLYC, TRIG, CREA, AGE, TRT, CHOL-T, LDL, GluG, LipG, HTA, NI, HDL, ProV, TE, DYSL.

2.3.2 Evaluation des modèles prédictifs par maladie

Les tableaux **4.6** et **4.7** rassemblent les indicateurs de l'évaluation des modèles prédictifs produits par la classification au sein des différents sous-groupes A1, A2 et A3, soient les trois indices : précision, sensibilité et spécificité. La moyenne de chaque indicateur est aussi calculée.

Algorithme de Classification	Evaluation du modèle construit								
	A1 : (G1,J1)			*A2 : (G1,J2)*			*A3 : (G2,J1)*		
	Précision	*Sensibilité*	*Spécificité*	*Précision*	*Sensibilité*	*Spécificité*	*Précision*	*Sensibilité*	*Spécificité*
C4.5	91.59	92.53	90.00	91.18	91.23	91.11	90.77	92.53	87.78
Classification tree	89.55	91.56	86.11	92.02	92.21	91.67	86.89	90.58	80.56
Régression Logistique	88.94	89.29	88.33	90.97	91.56	90.00	86.70	86.36	87.22
Naïf Bayes	87.70	89.29	85.00	91.17	91.56	90.56	87.49	90.26	82.78
CN2	89.14	93.51	81.67	90.36	92.86	86.11	89.75	93.18	83.89
Moyenne	89.38	91.24	86.22	91.14	91.88	89.89	88.32	90.58	84.45

Tableau N° 4.6 : Evaluation de la classification de l'HTA

Algorithme de Classification	Evaluation du modèle construit								
	A1 : (G1,J1)			*A2 : (G1,J2)*			*A3 : (G2,J1)*		
	Précision	*Sensibilité*	*Spécificité*	*Précision*	*Sensibilité*	*Spécificité*	*Précision*	*Sensibilité*	*Spécificité*
C4.5	91.58	85.80	94.48	77.88	64.81	84.36	92.20	85.80	95.40
Classification tree	93.64	88.27	96.32	79.12	68.52	84.36	91.81	87.04	94.17
Régression Logistique	85.26	56.79	99.39	80.14	72.84	83.74	84.64	54.94	99.39
Naïf Bayes	89.15	80.86	93.25	79.10	72.22	82.52	84.85	75.93	89.26
CN2	94.67	86.42	98.77	76.03	59.26	84.36	94.67	85.80	99.08
Moyenne	90.86	79.63	96.44	78.45	67.53	83.87	89.63	77.90	95.46

Tableau N° 4.7 : Evaluation de la classification du diabète

En analysant ces deux tables concernant les méthodes de data mining utilisées, les arbres de décision par les deux techniques C4.5 et classification tree ainsi que les règles de classification de l'algorithme CN2 (basé entre autres sur l'algorithme ID3) sont globalement plus performants que les algorithmes de la régression logistique et naïf bayes.

En terme de jeux de données, les modèles basés sur des variables normalisées sont plus efficaces dans la prédiction de l'HTA que dans la prédiction du diabète. En effet, une dégradation des performances a été notée pour le sous-groupe A2 dans le cas du diabète (moyenne = 78.45) par contre la meilleure précision pour l'HTA a été observée par

l'algorithme "classification tree" dans ce même sous-groupe. La normalisation peut donc contribuer à améliorer les performances des modèles et est sensible à la nature des données. Ces tests empiriques ont permis d'arrêter le choix sur les données continues (variables quantifiables) pour la construction des deux modèles MARS.

Un autre point est à noter dans l'évaluation des modèles de prédiction par rapport à la sensibilité, si l'on compare ces modèles notamment dans les sous-groupes A1 et A3, on constate d'abord qu'en terme de précision les taux moyens sont très proches : 89.38 et 88.32 pour L'HTA contre 90.86 et 89.63 pour le diabète. Par contre, en terme de sensibilité la différence est plus significative : 91.24 % et 90.58 % pour l'HTA contre 79.63 et 77.90 pour le diabète. Les modèles de l'HTA ont donc une plus grande facilité à détecter la maladie que les modèles du diabète, cela s'explique en partie par le problème d'asymétrie constatée dans la classe "Diabète". En effet sur les 488 individus, 126 seulement sont diabétiques et 326 ne le sont pas.

2.3.3. Facteurs de risque communs

Treize facteurs communs à l'HTA et au diabète ont été déterminés par l'approche adoptée, le tableau suivant les résume. Trois autres ont été intégrés (voir 2.2.6): GluG, ProA et NA.

FDR	**Maladie**	
	HTA	***Diabète***
Communs	AGE, IMC, TRT, SED, PAS, PAD, FC, CHOL-T, HDL, LDL, GLYC, CREA, LipG	
Individuels	IMG, GluV, ProA, EPIC, DIAB, MCV	NI, TRIG, TE, GluG, ProV, HTA, DYSL

Tableau N° 4.8 : Facteurs de risque de l'HTA et du Diabète

2.3.4. Evaluation des modèles MARS

L'évaluation des deux modèles construits par la méthode MARS est réalisée par comparaison des indices calculés à partir de leur matrice de confusion. En effet, les tableaux **4.9** et **4.10** montrent respectivement la matrice de confusion du premier modèle MARS construit en utilisant le mode d'attributs G1, et la matrice de confusion du second modèle construit en utilisant cette fois-ci le mode d'attributs G2.

TABLE I. MATRICE DE CONFUSION A PARTIR DE FCG1

Prédiction	Situation Réelle	
	Souffre de la maladie	*Ne souffre pas de la maladie*
Souffre de la maladie	338	0
Ne souffre pas de la maladie	13	137

Tableau N° 4.9 : Matrice de confusion à partir de FCG1

Prédiction	Situation Réelle	
	Souffre de la maladie	*Ne souffre pas de la maladie*
Souffre de la maladie	340	0
Ne souffre pas de la maladie	11	137

Tableau N° 4.10 : Matrice de confusion à partir de FCG2

Ces deux matrices indiquent que le second modèle est plus performant que le premier. En effet, c'est ce modèle qui minimise le plus l'erreur GCV égale à 0,2449. En plus, il donne une précision égale à 97,75 %, une sensibilité de 96,87 % et une spécificité de 100 % contre 97,33 % de précision, 96,3% de sensibilité et 100% de spécificité pour le premier modèle.

Ces résultats permettent de sélectionner le second modèle MARS comme modèle prédictif optimal pour la prédiction simultanée de l'HTA et du Diabète.

Le modèle obtenu nous permet de classer de nouveaux patients en ce basant sur leurs données cliniques et de prédire ainsi simultanément l'HTA et le diabète avec une performance de **97.75 %** en précision et une sensibilité de **96.87 %**. Ainsi nous avons amélioré la prédiction par rapport au traitement des facteurs individuels en adoptant une approche hybride traitant à la fois des facteurs individuels et des facteurs communs.

3. Travaux de déploiement de la solution

Un travail conséquent nous a permis d'adopter une approche de data mining qui se décline en deux phases afin d'aboutir au modèle prédictif optimal minimisant les erreurs de prédiction et capable de prédire la coexistence de l'HTA et du diabète chez un même patient. Nous avons alors suivi plusieurs expérimentations pour évaluer et valider la méthodologie mise en œuvre.

Plusieurs outils existent pour le data mining, nous mentionnons à titre d'exemple Tangara, Weka ou encore Orange pour ne citer que les outils open source les plus connus. Plusieurs autres solutions commerciales existent. Généralement, les outils sont génériques et traitent des

projets de data mining de domaines différents, ils incluent aussi plusieurs techniques permettant de répondre à plusieurs tâches de fouille de données.
Quand nous regardons de prêt ces outils, nous constatons dans Weka ou encore Tangara un manque dans la visualisation des résultats engendrant une difficulté d'interprétation. La force de l'outil Orange, en plus de sa simplicité, c'est surtout sa capacité à visualiser les données manipulées, les modèles construits, les matrices de confusion et les résultats de l'évaluation. Ainsi, nous l'avons choisi pour les expérimentations, tests et enfin nous essayons de l'adapter pour proposer un outil final d'aide à la décision au médecin spécialiste afin de traiter les données cliniques cardiovasculaires.

Un outil open source pour la fouille de données :

Orange est un logiciel de data mining qui comporte des outils de modélisation et d'exploration des données. L'université de Lubiana ainsi que l'institut Jozef Stefan en Slovénie, ont développé une plateforme de data mining nommé ML* EN 1996, qui est devenu le projet Orange. Le code source est écrit en C++, il nécessite un compilateur Python.
Orange permet la définition du cheminement de traitement, en construisant et en reliant les différents widgets (icônes) nécessaires pour la construction du modèle de data mining. La manipulation de la version 2.0 nous a permis de construire tous les modèles prédictifs mis en œuvre dans la partie traitant des facteurs individuels.

Ainsi nous procédons à la régénération du code d'Orange 2.0 pour proposer un outil simple que nous déclinant en trois modules :
Module 1 : Données et sélection de variables
Module 2 : Modèle prédictif
Module 3 : Evaluation

Le premier module inclut trois opérations : la lecture du fichier de données, la visualisation de ces données et enfin la sélection de variables;
Le second module permet deux opérations : la construction de l'arbre de décision basé sur l'algorithme C4.5 et l'affichage de cet arbre;
Enfin, le dernier module définit deux opérations : visualiser la performance et enfin visualiser la prédiction (c'est la visualisation de la matrice de confusion)

Voici un exemple de traitement de données au sein de l'outil proposé :

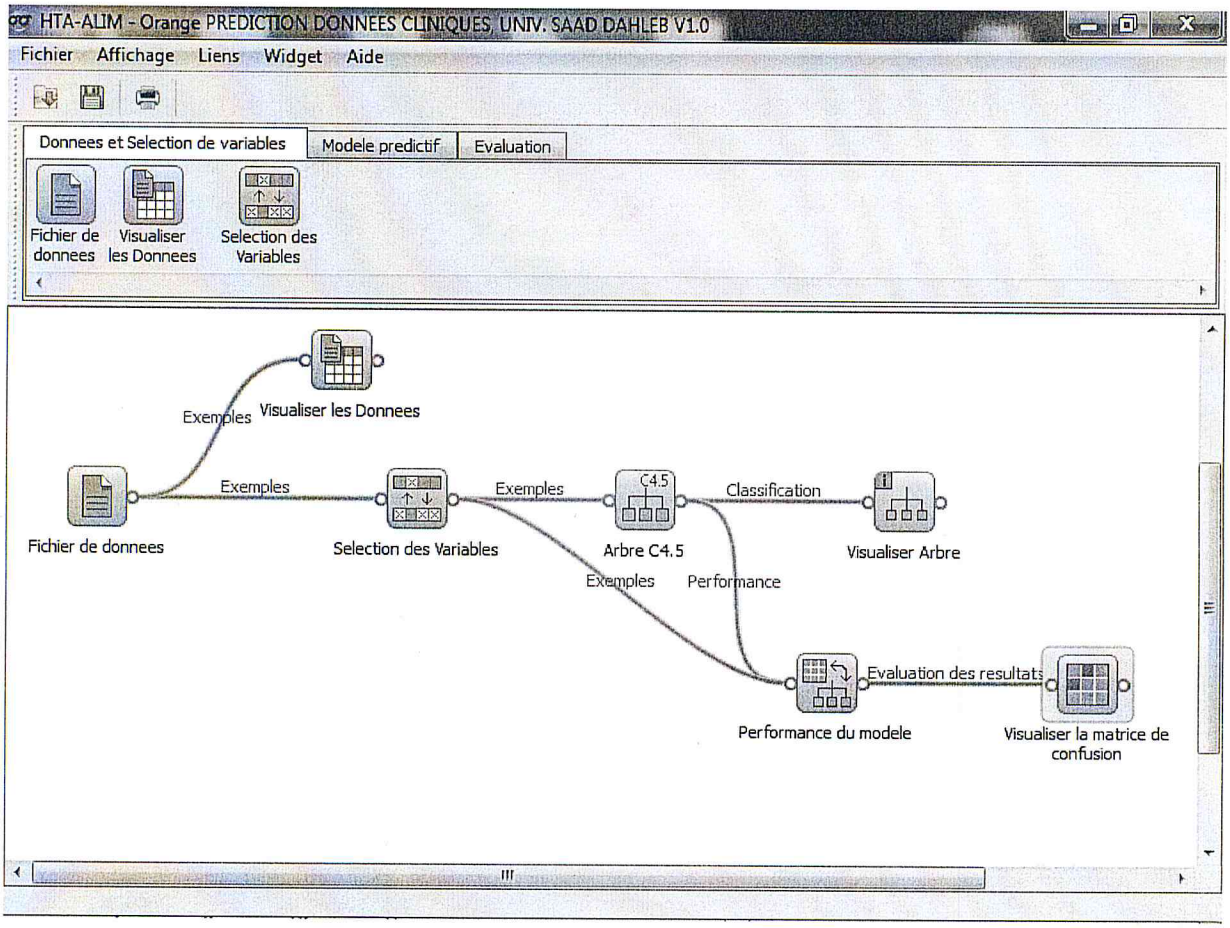

Figure N° 4.2 : Aperçu de l'outil proposé

Conclusion

Dans ce chapitre, nous avons donné un procédé basé sur l'ECD pour construire un modèle prédictif optimal prenant en compte l'association de deux maladies : l'hypertension artérielle (HTA) et le diabète et nous avons exploité les données de l'étude épidémiologique d'El Menia pour l'apprentissage et le test des modèles produits dans ce travail.

Notre méthodologie se base sur l'application de cinq techniques de data mining (C4.5, Classification Tree, Logistic regression, Naive bayes et CN2) pour extraire, séparément, les facteurs de risques individuels les plus pertinents de l'HTA et du diabète. Ensuite, nous avons identifié, les facteurs de risques communs aux deux maladies et nous leurs avons appliqué la

méthode "Multivariate Adaptative Regression Splines" (MARS) afin d'obtenir un modèle prédictif optimal traitant simultanément de l'HTA et du diabète.

Les résultats que nous avons obtenus donnent un taux de précision de 97.75% avec une sensibilité de 96,87%. Le modèle retenu permet de prédire la coexistence ou non des deux maladies chez un même patient à partir de 16 facteurs de risques déterminants dont quatre facteurs alimentaires et ce au lieu des 133 items que l'enquête à initialement introduit.
L'expérimentation et les tests empiriques nous ont permis de retenir notre démarche et de la proposer au médecin spécialiste qui a initié l'enquête, nous attendons de sa part une validation médicale afin de pouvoir ensuite proposer un outil complet d'aide à la décision supportant toute la méthodologie. Dans ce sens, nous avons commencé à mettre en place un outil simple basé sur les arbres de décision (algorithme C4.5) pouvant guider un médecin dans le traitement des données cliniques.

Conclusion et perspectives

Dans ce travail, nous nous sommes fixé comme objectif de démonter, l'apport des techniques de data mining pour l'aide à la décision en recherche épidémiologique en traitant le cas des maladies cardiovasculaires. Nous pouvons dire, pour clôturer notre étude, que cet apport est non seulement possible mais aussi significatif à la lumière des résultats de prédiction que nous avons obtenus en utilisant plusieurs techniques de fouille de données.

Ce que nous retenons également est que pour espérer tirer profit des données médicales d'une manière générale, appliquer un algorithme de fouille de données n'est certainement pas suffisant. En effet, la compréhension du domaine et donc pour notre part des maladies cardiovasculaires ainsi que des données à manipuler est indispensable. Ensuite la préparation des données, le choix des techniques à mettre en place, le calibrage des paramètres d'algorithmes, les méthodes d'évaluation qui répondent au mieux à la quantité des données, la possibilité d'interprétation et de visualisation des résultats sont autant d'éléments qui concourent à l'extraction de connaissances pertinentes à partir des données, en particulier dans le domaine cardiovasculaire que nous avons traité.

Pour atteindre notre objectif, dans un premier temps, nous avons étudié les systèmes d'aide à la décision médicale, pour asseoir les caractéristiques les plus importantes des données cliniques afin de les comprendre et les utiliser. Nous avons alors distingué pour ces systèmes trois composantes : « l'acquisition, l'intégration et les outils ETL des données médicales », « le stockage et l'entreposage des données médicales » et enfin « l'analyse et l'exploration des ces données ».

Ainsi, nous avons étudié les moyens de récoltes des données cliniques en recherche épidémiologique. Nous avons identifié plusieurs systèmes possibles entre autres les DPE, les bases de données multidimensionnelles et surtout les référentiels de données cliniques, les entrepôts de données cliniques et les magasins de données cliniques utilisés pour des fins d'aide à la décision de manière différente selon le besoin. Pour recueillir des

données cliniques d'un service cardiovasculaire, nous préconisons de se baser sur un magasin de données cliniques facilitant l'intégration des données issues de différentes sources et se concentrant sur des analyses spécifiques aux maladies cardiovasculaires.

Le cœur de notre travail se situe dans les deux dernières parties consacrées à l'étude des techniques supervisées pour la prédiction des facteurs de risques cardiovasculaires et au processus ECD. Ainsi, l'approche que nous adoptons permet de :

- Poser les fondements d'une méthodologie data mining pour l'aide à la décision en prévention primaire.
- Construire un modèle prédictif optimal prenant en charge des patients à haut risque cardiovasculaire.
- Traiter les facteurs communs aux deux maladies en plus des facteurs individuels.
- Améliorer les performances des modèles basés sur les facteurs individuels.
- Traiter des données cliniques y compris des variables alimentaires.

Comme perspectives, nous proposons deux extensions à notre travail :

- ✓ Construire un data mart pour le service cardiovasculaire incluant les données cliniques que nous avons manipulées et supportant les opérations de préparation des données que nous avons mis en œuvre dans ce travail.
- ✓ Développer un outil complet d'aide à la décision exploitant les données du data mart cardiovasculaire et supportant la méthodologie mise en place après validation du médecin.

Bibliographie

[1] Franck Ravat, « Modèles et outils pour la conception et la manipulation de systèmes d'aides à la décision », Habilitation à diriger des recherches de l'Institut de recherche en informatique de Toulouse, Décembre 2007.

[2] Jacques Mélèse, « L'analyse modulaires des systèmes de gestion, AMS », Edition Hommes et Techniques - Paris, 1972.

[3] D. Nanci, B. Espinasse, « Ingénierie des systèmes d'informations : Merise, Deuxième génération », quatrième édition, Vuibert, 2001.

[4] Williams. H. Inmon, « Building the Data Warehouse », New York, 1996.

[5] Olivier Teste, « Modélisation et manipulation d'entrepôts de données complexes et historisées », Thèse de Doctorat de l'université de Paul Sabatier, Décembre 2000.

[6] Jean-Louis Renaud-Salis, Philippe Lagouarde, Stephan Darmoni, sous la direction de Pierre-Henri Comble, « Etude des systèmes d'aide à la décision médicale ». Etude commanditée par la haute autorité de santé et réalisé par Cegedim-Activ, Livrable 2, Version du 12 juillet 2010.

[7] Isabelle Colombet, « Aide à la décision et évaluation informatisée des soins », Faculté de Médecine René Descartes, Université Paris 5, Hôpital Européen Georges Pompidou (HEGP)

[8] Adrien Coulet, Marie Dominique Devignes, Malika Smail, « Extraction de connaissances pharmaco-génomiques à partir d'études cliniques : problématique », LORIA, INRIA, Université Henri Poincare-Nancy.

[9] Lobach DF, Kawamoto K, Anstrom KJ, Russell ML, Woods P, Smith D. « Development, deployment and usability of a point-of-care decision support system for chronic disease management using the recently-approved HL7 decision support service standard». Stud Health Technol Inform. 2007.

[10] Alain-jacques Valleron, « L'épidémiologie humaine, Conditions de son développement en France, et rôle des mathématiques » - Académie des sciences 2006.

[11] Christel Daniel, Jean-philippe Jais, Naji el Fadly, Paul Landais « Dossier patient informatisé à visé de recherché biomédical », INSERM UMRS 872 & Service de biostatistiques et informatique médicale Université Paris Descartes - Press Med 2009.

[12] Hallvard Lærum, « Evaluation of electronic medical records - A clinical task perspective », Doctoral thesis.

[13] DJEDDAOUI Mohamed, « Elaboration d'un système multidimensionnel », mémoire de Magister - Spécialité : Informatique Option : Informatique industrielle, Université MOHAMED BOUDIAF de M'sila. Septembre 2006.

[14] Houssem Jerbi, Franck Ravat, Olivier Teste, and Gilles Zurfluh, « Applying Recommendation Technology in OLAP Systems », IRIT, Institut de Recherche en Informatique de Toulouse.

[15] Yoann Pitarch, « Résumé de Flots de Données : Motifs, Cubes et Hiérarchies », Thèse de Doctorat en Informatique - Monpellier 2, mai 2011

[16] Rémy Choquet, Christel Daniel, Omar Boussaid, Marie-Christine Jaulent. « Etude méthodologique comparative de solutions d'entreposage de données de santé à des fins décisionnelles ». INSERM UMR_S 872 eq.20, centre de recherche des cordeliers, Laboratoire ERIC, Université de Lyon 2, ICSSHC 2008.

[17] Lael Dickinson, Modeling Transactional data into decision making information in the medical industry : The case of Laberman General Hospital, Rensselaer Polytechnic Institute TROY, NEW YORK, U.S.A. Irina Ilovici, Ph.D.

[18] Midouni Sid Ahmed Djallal, « Modélisation multidimensionnelle des données complexes : application aux données médicales » - 2005.

[19] Jérôme Darmont, « Approche de modélisation multidimensionnelle des données complexes : Application aux données médicales », Université de Lyon (ERIC lyon 2) – 2008.

[20] Bert Arnrich, « Data Mart Based Research in Heart Surgery», Printed on acid-free paper, 2006.

[21] Puhr C. « The clinical Data Warehouse ». PHD thesis (MIAS student) Univ. of Wien, 2002.

[22] Robert L. Leitheiser, « Data Quality in Health Care Data Warehouse Environments », University of Wisconsin – Whitewater

[23] Torben Bach Pedersen et Al, « Research Issues in Clinical DataWarehousing » - Center for Health Informatics

[24] LI Ping, WU Tao, CHEN Mu, ZHOU Bin and XU Wei-guo, « A study on building data warehouse of hospital information system Medical informatics » - Chin Med J 2011 ; 124 (15) : 2372-2377

[25] Philip Burrowes, Jason Oliveira, « CP-Nexus: A Clinical Data Warehouse at Columbia-Presbyterian Medical Center », Columbia-Presbyterian Medical Center, N.Y. (Homepage, http://www.cpmc.columbia.edu/ais/resources/ic/)

[26] Yasuo Takahashi & Yayoi Nishida & Satoshi Asai. « Utilization of health care databases for pharmacoepidemiology », June 2011

[27] Tony R. Sahama and Peter R. Croll. « A Data Warehouse Architecture for Clinical Data Warehousing », Faculty of Information Technology - Queensland University of Technology – Brisbane.

[28] Smith, A. and Nelson, M. « Data Warehouses and Clinical Data Repositories ». In Ball, M., Douglas, J., and Garets, D., editors, Strategies and Technologies for Healthcare Information, pages 17–31. Springer (1999).

[29] Catherine COMBES, « Couplage simulation à événements discrets et Data Mart appliqués aux établissements de soins : une application au service de chirurgie », Laboratoire d'Analyse des Systèmes de Santé, Université Claude Bernard Lyon I UMR 5823 du CNRS.

[30] Emilie GUÉRIN, « Intégration de données pour l'analyse de transcriptome : Mise en œuvre par l'entrepôt Gedaw (Gene Expression Data Warehouse) », Thèse Doctorat de l'université de rennes 1, 2005.

[31] Eric Zapletala, Nicolas Rodona, Natalia Grabarab, Patrice Degouletab. « Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case », A Department of Medical Informatics, Georges Pompidou University Hospital, Paris, France

[32] Fayyad, Usama, Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Intelligent Systems, Octobre 1996, Vols. vol. 11, no. 5, pp. 20-25.

[33] Wijsen J, Data Mining et Data Warehousing. 2001

[34] Fayyad, Usama, Data Mining and Knowledge Discovery. Kluwer Academic Publishers. Janvier 1998, Vol. Volume 2, Issue 1, pp. Pages: 5 - 7.

[35] Larose, Daniel T. Discovering Knowledge in Data: An introduction to data mining. Hoboken, New Jersey : John Wiley & Sons, Inc, 2005.

[36] mamadou Ouattara, « Fouille de données : vers une nouvelle approche intégrant de façon cohérente et transparente la composante spatiale ». Mémoire présenté à la Faculté des études supérieures de l'Université Laval dans le cadre du programme de Maîtrise en sciences géomatiques pour l'obtention du grade de maître des science (M.SC.) – 2010.

[37] MENOUER Tarek, DERMOUCHE Mohamed. « Application de techniques de data mining pour la classification automatique des données et la recherche d'associations ». Mémoire de fin d'études Diplôme d'ingénieur – Ecole Nationale d'informatique – 2010.

[38] Kellou Kenza, Mokhtari Abdeldjalil, « Réalisation d'une plateforme d'expérimentations et de tests d'algorithmes de data mining (www.ESIMiner.com) ». Mémoire de fin d'études Diplôme d'ingénieur – Ecole Nationale d'informatique – 2011.

[39] Sahar BAYAT MAKOEI, « Etude de l'accès à la transplantation rénale en Lorraine par méthodes biostatistiques conventionnelles et par fouille de données », thèse de doctorat, Epidémiologie et Santé Publique, Université Henri Poincaré, Nancy 1, 2008

[40] Norbert Tsopze, « Treillis de Galois et réseaux de neurones : une approche constructive d'architecture des réseaux de neurones », thèse de doctorat de l'Université d'Artois et de l'Université de Yaoundé I.

[41] « Arbres de décisions », Optimisation pour ECA - RSTI-RIA-ECA pages 686-702, 2002

[42] Duen-Yian Yeh, Ching-Hsue Cheng, Yen-Wen Chen. « A predictive model for cerebrovascular disease using data mining», Expert Systems with Applications 38 (2011) 8970–8977.

[43] Mevlut Ture, Imran Kurt, A. Turhan Kurum, Kazim Ozdamar. « Comparing classification techniques for predicting essential hypertension ». Expert Systems with Applications 29 (2005) 583–588

[44] Yue Huang, Paul McCullagh, Norman Black, Roy Harper. « Feature selection and classification model construction on type 2 diabetic patients' data». Artificial Intelligence in Medicine (2007) 41, 251—262

[45] Cheng-Ding Chang, Chien-Chih Wang b, Bernard C. Jiang. « Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors ». Expert Systems with Applications 38 (2011) 5507–5513

[46] Rashedur M. Rahman □, Fazle Rabbi Md. Hasan. « Using and comparing different decision tree classification techniques for mining ICDDR,B Hospital Surveillance data ». Expert Systems with Applications 38 (2011) 11421–11436

[47] Damrongrit Setsirichok et Al. Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naive Bayes classifier and a multilayer perceptron for thalassaemia screening. Biomedical Signal Processing and Control 7 (2012) 202– 212.

[48] Mevlut Ture, Fusun Tokatli, Imran Kurt. Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Systems with Applications 36 (2009) 2017–2026.

[49] Mansour Mededjel, Hafida Belbachir. « Post-élagage Indirect des Arbres de Décision dans le Data Mining » Université Abdelhamid Ibn Badis – Mostaganem et Université des Sciences et de la Technologie - Mohamed Boudiaf – Oran, Algérie. SETIT 2007 – 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications March 25-29, 2007 – TUNISIA

[50] Dan-Andrei SITAR-TAUT, Adela SITAR-TAUT. Cardiovascular Attributable Risk and Risk Factors Evaluations as a Matter of Statistics and Data Mining Confluences. Informatica Economică vol. 14, no. 4/2010.

[51] K.Srinivas, Dr. G.Raghavendra Rao and Dr. A.Govardhan. SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.3, May 2011.

[52] Chatlert Pongchaiyakul, Praew Kotruchin, Ekgaluck Wanothayaroj, Tuan V. Nguyen. An innovative prognostic model for predicting diabetes risk in the Thai population. Diabetes researchand clinical practice (2011) 193-198.

[53] Biswanath Samanta et Al. Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms. Artificial Intelligence in Medicine (2009) 46, 201—215

[54] Baha S_en a, Emine Ucar b, Dursun Delen. Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications 39 (2012) 9468–9476.

[55] Ting-Ting Lee et Al. Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. international journal of medical informatics 8 0 (2 0 1 1) 141–150.

[56] Marco Roman et Al. Serum seleno-proteins status for colorectal cancer screening explored by data mining techniques - a multidisciplinary pilot study. Microchemical Journal 2011.

[57] Tim Van den Bulcke et Al. Data mining methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data. Journal of Biomedical Informatics 44 (2011) 319–325.

[58] S. Aruna, Dr S.P. Rajagopalan and L.V. Nandakishore. Knowledge based analysis of various statistical tools in detecting breast cancer.

[59] [Ricco Rakotomalala, « Pratique de la Régression Logistique : Régression Logistique Binaire et Polytomique » Version 2.0

[60] Somia RAHMOUN, « Méthodes d'apprentissage pour améliorer la QoS d'une flotte de logiciels embarqués », Mémoire de fin d'étude Master en Informatique, Option : Modèle Intelligent et Décision(M.I.D) – Université Abou Bakr Belkaid– Tlemcen 2010-2011

[61] Chapitre de livre, Supervised Learning: Decision Trees, Rule Algorithms, and Their Hybrids

[62] Saso Dzeroski, nada Lavrac. Rule induction and instance-based learning applied in medical diagnosis. Technology and health care 1996.

[63] Chang Sik Son, , Byoung Kuk Jang, Suk Tae Seo, Min Soo Kim and Yoon Nyun Kim, A hybrid decision support model to discover informative knowledge in diagnosing acute appendicitis. BMC Medical Informatics and Decision Making 2012.

[64] Nada Lavrac, Chapter 52 : "DATA MINING IN MEDICINE" Joief Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia, Nova Gorica Polytechnic.

[65] Milan Kumari, Sunila Godara. Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction Department of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India. 2011.

[66] Hui-Yi Lin et Al , « Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP–SNP interactions and their application in prostate cancer , 2008.

[67] Cynthia J et Al. An Exploratory Analysis of Criteria for the Metabolic Syndrome and Its Prediction of Long-term Cardiovascular Outcomes, The Hoorn Study. American Journal of Epidemiology - by the Johns Hopkins Bloomberg School of Public Health 2005.

[68] MARGUERITE ENNIS et Al. A COMPARISON OF STATISTICAL LEARNING METHODS ON THE GUSTO DATABASE. STATISTICS IN MEDICINE Statist. Med. 17, 2501-2508 (1998).

[69] Thanh-Nghi Do & François Poulet. Interprétation des résultats de SVM. ESIEA - Parc Universitaire de Laval-Changé.

[70] Mohamadally Hasan, Fomani Boris. SVM : Machines a Vecteurs de Support ou Separateurs a Vastes Marges - ISTY3.

[71] Guillaume CALAS. Études des principaux algorithmes de data mining, SCIA, EPITA 2009.

[72] Simon Marcellin. Arbres de décision en situation d'asymétrie, thèse de Doctorat en Informatique. Université Lumière Lyon II, Septembre 2008.

[73] Rapport du forum et de la réunion technique OMS : « RÉDUIRE LES APPORTS EN SEL AU NIVEAU DES POPULATIONS » 5–7 octobre 2006, Paris, France

[74] Rym Nasri, Maladies cardiovasculaires : Première cause de mortalité en Algérie, soir d'algérie, quotidien indépendant DIMANCHE 20 NOVEMBRE 2011 - 24 DOU AL-HIJA 1432 - N° 6413.

[75] « Suivi de la situation des enfants et des femmes, Enquête national à indicateurs multiples Rapport principal », Ministère de la santé, de la population et de la réforme hospitalière et l'Office National de statistiques ONS (Fonds des Nations Unies pour l'enfance,

Fonds des Nations Unies pour la population, Système des Nations Unies pour le Développement pour l'Algérie, ONUSIDA) - Décembre 2008.

[76] Rolf Stöcklia et Al. Hypertension et diabète , Lukas Zimmerlib, UniversitätsSpital, Zürich Forum Med Suisse 2009;

[77] A.J. Scheen, J-C. Philips, J-M. Krzesinski. Hypertension et diabète : à propos d'une association commune mais complexe. Rev Med Liège 2012; 67 : 3 : 133-138

[78] Julie MARTINONI. Evaluation de la prescription des antihypertenseurs chez le sujet âgé. Thèse doctorat en médecine 2011

[79] Philippe PASSA, Service de diabetologie, hôpital Saint-Louis - Paris. Le plan santé diabète : pourquoi et comment ? Vigilance et santé

[80] Hermans Louvain Med. Traiter la dyslipidémie du diabétique de type 2 : place des fibrates M.P. 2011; 130 (3): S31-35.

[81] F. Hamida, M. Temmar, A. Chibane, N. Guendouz, A. Bouamra, MT. BOUAFIA. Enquête Epédémiologique: " prévalence de l'HTA et les facteurs de risques cardiovasculaires Etude el-Menia Mars -Avril 2010.