République Algérienne Démocratique et Populaire

Ministère de L'enseignement Supérieur et de la Recherche Scientifique

Université Abbès Laghrour Khenchela

Faculté des Sciences et de la Technologie

Département de Mathématiques et Informatique


Mémoire pour obtenir le diplôme de

Master en Informatique

**Spécialité :** Génie logiciel et systèmes distribués

# Techniques d'apprentissage automatique pour la prédiction de la maladie de l'artère coronaire

**Présenté Par :**

- ✓ MELLAH Fouad

**Encadré par:** Dr. HAOUASSI Hichem

***Année universitaire : 2020-2021***

# Résumé

La fouille de données trouve son application dans différents domaines tels que la médecine, l'économie et le commerce ... Ces dernières années, en raison de l'utilisation du programme de gestion informatique et électronique, utilisée pour stocker des données dans le secteur de la santé, en particulier des hôpitaux les bases de données deviennent très riche en terme de données brutes. Donc, ces données peuvent les utilisées pour aider les médecins a diagnostiquer les patients atteints des maladies cardiaque ainsi que l'accélération dans le le processus du diagnostic. Dans le cadre de ce travail, nous sommes intéressés à utiliser les différents techniques d'apprentissage automatique a savoir la classification de données pour la prédiction de la maladie CAD.

La maladie coronarienne de l'artère (CAD) est l'une des maladies les plus courantes du monde entier. Un diagnostic précoce et précis de CAD permet une administration opportune de traitement approprié et contribue à réduire la mortalité. Dans ce document, nous appliquons plusieurs algorithmes d'apprentissage machines (classification) permettant une détection précise de la CAD et l'appliquer sur la base de données Z-Alizadeh Sani collectées de patients iraniens.

L'objectif de cette étude est de développer une application de prévision des patients en cas de CAD ou non, à travers ses informations médicales utilisées. Dans ce travail nous avons développé un système de diagnostic de CAD pour prédire la maladie cardiovasculaires à l'aide de jeux de données des malades. Nous avons utilisé le jeu de données Z-Alizadeh Sani contenant 303 patients, chacune dispose de 54 propriétés. À la suite de nos études, nous concluons que le meilleur modèle de classification trouvé est le modèle généré l'algorithme " Random_forest " avec un taux de classification de 85 %.

# Remerciements

*On remercie dieu le tout puissant de nous avoir donné la santé et la volonté d'entamer et de terminer ce travail.*

*Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de Mr HAOUASSI Hichem, on le remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce travail.*

*Nous exprimons toute notre reconnaissance aux membres de jury de nous avoir fait l'honneur de participer à nos jury. Nous remerciements Merci a qui n'ont cessé d'être pour Nous des exemples de persévérance, de courage et de générosité, a vous, nos parents, ma sœur.*

*Nous n'oublions pas de remercier tous mes collègues et amis sans exception.*

# Table des matières

Résumé ..... II

Remerciements ..... III

Table des matières ..... IV

Liste des abréviations ..... VIII

Liste des tableaux ..... IX

Liste des figures ..... X

Introduction générale ..... 11

Contexte ..... 11

Objectifs ..... 11

Méthodologie ..... 12

Résultats ..... 12

Structure du mémoire ..... 13

Chapitre 1 : La Fouille de données Médicale …. ..... 14

Introduction ..... 15

1.1 Définitions ..... 16

1.1.1 Fouille de données (Data mining) ..... 16

1.2 Processus du data mining ..... 16

1.3 Le Data mining, un processus dans l'ECD ..... 19

1.4 Les étapes du processus de data mining ..... 21

1.5 Application d'exploration de données ..... 22

1.5.1 Soins de santé ..... 22
1.5.2 L'éducation ..... 23
1.5.3 Gestion de la relation client (CRM) ..... 23
1.5.4 L'aspect médical ..... 23
1.6 Les taches de la fouille de données ..... 23
1.6.1 Classification ..... 24
1.6.2 La régression ..... 24
1.6.3 L'estimation ..... 25
1.6.4 Le groupement par similitude (règle d'association) ..... 26
1.6.5 L'analyse des clusters ..... 26
1.6.6 La description ..... 26
1.7 Fouille de données médicale ..... 27
1.7.1 Définition ..... 27
1.7.2 Objectif de la fouille de données médicale ..... 27
1.7.3 Domaine d'application ..... 28
1.7.4 Applications de la fouille de données dans le domaine de la santé ..... 30
1.8 Les logiciels de Data Mining ..... 31
1.8.1 Logiciels libres ..... 31
1.8.2 Logiciels commerciaux ..... 32
Conclusion ..... 33
Chapitre 2 : Classification et prédiction de données médicales ..... 34
Introduction ..... 35
1.1 Définitions ..... 36
1.1.1 Classification ..... 36
1.1.2 Classification supervisé ..... 36
1.1.3 Classification non supervisé ..... 36
1.1.4 Prédiction ..... 37
1.2 Les étapes de la classification ..... 37
1.2.1 Représentation des individus ..... 37

1.2.2 La proximité des individus ..... 37
1.2.3 Classification ..... 37
1.2.4 Abstraction des données ..... 38
1.2.5 Validation du résultat ..... 38
1.3 Classification supervisée dans le domaine médical ..... 38
1.4 Algorithmes de la classification supervisée ..... 39
1.4.1 Naïve Bayes ..... 39
1.4.2 Arbres de décision ..... 40
1.4.3 K plus proches voisins (KNN) ..... 41
1.4.4 Support Vector Machine (SVM) ..... 42
1.5 Classification des données médicales ..... 43
1.5.1 Digitalisation du domaine médical ..... 43
1.5.2 Evaluation des modèles de classification ..... 45
1.5.3 La matrice de confusion ..... 45
Conclusion ..... 47
Chapitre 3 : Etat de l'art sur la fouille de donnée du cardio-vasculaire ..... 48
Introduction ..... 49
1.1 Maladie du cardio-vasculaire ..... 50
1.2 Etat de l'art ..... 50
1.3 Data sets de la maladie cardio-vasculaire ..... 52
1.3.1 Data set Z-Alizadeh Sani ..... 52
1.3.1.1 Historique ..... 52
1.3.1.2 Définition ..... 52
1.3.1.3 Résultats obtenus dans la littérature ..... 54
1.3.2 Data set Cleveland ..... 55
1.3.2.1 Historique ..... 55
1.3.2.2 Définition ..... 55
1.3.3 Data set Hungarian ..... 56
1.3.3.1 Historique ..... 56

1.3.3.2 Définition ..... 56
1.3.3.3 Approches de la littérature qui utilisent la base Hungarian ..... 56
1.3.4 Data set Statlog ..... 57
1.3.4.1 Définition ..... 57
1.4 Machine Learning pour le diagnostic CAD ..... 57
1.4.1 Défis et inconvénients de l'utilisation des algorithmes ML ..... 59
Conclusion ..... 60
Chapitre 4 : Implantation et Réalisation ..... 61
Introduction ..... 62
1.1 Environnement de travaille et outils utilisés ..... 63
1.1.1 Java ..... 63
1.1.2 Weka ..... 64
1.1 Application ..... 66
1.1.1 Structure du notre système de diagnostic du CAD ..... 66
1.1.2 Fonctionnement du système développé ..... 69
1.1.2.1 Choix des données d'apprentissage ..... 69
1.1.2.2 Description des données de la base Z-alizadeh ..... 71
1.1.2.3 Choix et utilisation d'algorithmes de classification ..... 73
1.1.2.4 Diagnostic du CAD d'un nouveau patient ..... 75
Conclusion ..... 82
Conclusion générale ..... 83
Contribution ..... 83
Conclusion ..... 83
Travaux futurs ..... 83
Bibliographie ..... 85

# Liste des abréviations

| | |
|---|---|
| CAD | Coronary Artery Disease |
| SVM | Support Vector Machine |
| ECG | Électrocardiogramme |
| CRISP-DM | Cross-Industry Standard Process for Data Mining |
| TP | Taux Positif |
| TN | Taux Négatif |
| FP | Faux Positif |
| FN | Faux Négatif |
| UCI | University of California at Irvine |
| ML | Machine Learning |
| RH | Ressources Humaines |
| ACP | Analyse en Composantes Principales |
| OCDE | Organisation de Coopération et de Développement Economiques |
| ECD | Extraction de Connaissances à partir de Données |
| GRC | Gestion de la Relation Client |
| ADN | Acide Désoxyribonucléiques |

# Liste des tableaux

Tableau 1: Matrice de confusion pour une classification supervisée binaire ..... 46
Tableau 2 : Caractéristiques de la data set Z-Alizadeh Sani . ..... 54
Tableau 3 : Comparer les performances des algorithmes ..... 54
Tableau 4 : Résumé des méthodes de classification de CAD automatisées existantes utilisant diverses bases de données publiques ..... 56
Tableau 5 : Description de l'attribut de la StatLog Dataset de la maladie cardiaque ..... 57
Tableau 6 : La liste de l'article publié en utilisant une méthode d'apprentissage en profondeur pour la CAD ..... 58
Tableau 7 : Résultats de classification par l'algorithme J48 ..... 77
Tableau 8 : Matrice de confusion de la classification utilisent Algorithme J48 ..... 77
Tableau 9 : Liste des attributs influencent la performance des diagnostics ..... 78
Tableau 10 : Résultats de classification par l'algorithme One R ..... 78
Tableau 11 : Matrice de confusion de la classification utilisent l'algorithme One R ..... 78
Tableau 12 : Résultats de classification par l'algorithme PART ..... 79
Tableau 13 : Matrice de confusion de la classification utilisent l'algorithme PART ..... 79
Tableau 14 : Résultats de classification par l'algorithme Naive_Bayes ..... 79
Tableau 15 : Matrice de confusion de la classification utilisent l'algorithme Naive_Bayes .... 79
Tableau 16 : Résultats de classification par l'algorithme Random_Forest ..... 80
Tableau 17 : Matrice de confusion de la classification utilisent l'algorithme Random_Forest 80
Tableau 18 : Comparaison des résultats de classification de différents algorithmes ..... 81

# Liste des figures

Figure 1: Processus du data mining ..... 17
Figure 2 : Le Schéma du processus de data mining ..... 20
Figure 3 : SVM classification binaire ..... 42
Figure 4 : Schéma du cœur ..... 52
Figure 5 : Processus global de notre système de diagnostic du CAD ..... 66
Figure 6 : Fenêtre principale de l'application ..... 70
Figure 7 : Interface graphique de chargement le fichier d'apprentissage jeu de données Z-alizadeh (Fichier arff) ..... 71
Figure 8 : Interface graphique le contenu jeu de données Z-alizadeh (Fichier arff) ..... 72
Figure 9 : Interface graphique choix et utilisation d'algorithmes de classification ..... 73
Figure 10 : Exemple de appliqué l'algorithme J48 ..... 74
Figure 11 : Interface graphique résultats diagnostic du CAD d'un nouveau patient ..... 75
Figure 12 : Arbre de décision après l'utilisation de la classification de l'algorithme J48 ..... 83

# Introduction générale

## Contexte

Dans de nombreux domaines, il est nécessaire de prendre des décisions critiques, dans un contexte difficile et à un temps limité. Par exemple, un médecin qui doit prendre une décision rapide vis-à-vis un malade. Mais il ne peut pas se souvenir de tous les enregistrements qu'il a travaillés et étudiés depuis des années.

Les outils informatiques peuvent fournir une aide précieuse dans ce cas car elles peuvent prendre en compte un grand nombre de cas traités et les proposer pour un nouveau statut basé sur l'assemblage de tous les cas.

La fouille de données (Data Mining) est via de multiples techniques permettant de découvrir les connaissances cachées dans les données et de les utilisées pour prendre des décisions.

La fouille de données trouve son application dans différents domaines tels que la médecine, l'économie et le commerce ... Ces dernières années, en raison de l'utilisation du programme d'apprentissage et de la gestion informatique, dans le secteur de la santé, en particulier les hôpitaux lorsque, les tailles de données sont agrandies jour par jour. Donc, il est intéressant de les exploitées pour aider les médecins à prendre de décisions de diagnostique des maladies.

## Objectifs

La maladie cardio-vasculaire est l'une des maladies humaines les plus critiques du monde et effets très mal la vie humaine. Dans la maladie cardio-vasculaire, le cœur est incapable de pousser la quantité requise de sang vers d'autres parties du corps. Le diagnostic précis et à temps de la maladie cardio-vasculaire est important pour la prévention et le

traitement de l'insuffisance cardiaque. Notre application Faciliter le travail des médecins pour voir si la personne souffre d'une maladie CAD ou non.

L'objectif principal de ce travail est de développer une application de prédiction destiné aux patients en cas de maladie CAD ou non, et ceci à travers des techniques de classification de données apprentis sur les données de la base Z-Alizadeh Sani. Notre application contribue a facilite le processus de diagnostic par les médecins et cela pour gagner du temps dans le traitement du patient.

## Méthodologie

Pour mettre en place notre application, nous avons réalisé une application d'aide des médecins et patients à prédire que le patient est malade ou non, l'application est réalisé sous l'environnement NetBeans qui utilise les algorithmes offert par le Weka via une interface facile à utilisé. Donc, notre application appel les packages et les différentes classes nécessaire de Weka dans un programme Java afin d'assurer les fonctionnalités d'apprentissage, de test et de prédiction fixés par notre étude.

## Résultats

Dans ce travail nous avons développé un système de diagnostic pour la prédiction des maladies cardio-vasculaire en utilisant le jeu de données sur les maladies cardio-vasculaire. Nous avons utilisé data set Z-Alizadeh sani qui contient 303 patients, dont chacun a 54 caractéristiques. Selon les résultats d'évaluation, nous avons constaté que le modèle de classification qui donne le plus grand taux de classification était obtenu par l'algorithme " Random_forest " avec un taux de 85 % alors que nous avons constaté que le deuxième était obtenu par l'algorithme "Naive_Bayes" par un taux de 82 %. Alors que le modèle généré par l'algorithme " One_R" donne le plus faible taux avec un pourcentage de 45%. L'algorithme " J48 " donnent respectivement les taux 80 %. L'algorithme " PART " donnent le taux 81%. voir le tableau 18.

## Structure du mémoire

Notre mémoire contient une introduction, et 4 chapitres et une conclusion générale. Ce mémoire et organisé de la manière suivante :

**Chapitre 1: La fouille de données médicale**

Dans ce chapitre on a présenté la fouille de données (Data mining) qui est le domaine de notre étude, ses taches, ses techniques, et ses objectifs sont présentées dans ce chapitre.

**Chapitre 2 : Classification et prédiction des données médicale**

Dans ce chapitre on a présenté la classification, ses étapes, classification supervisée dans le domaine médical, les algorithmes de la classification supervisée et classification des données médicales.

**Chapitre 3 : Etat de l'art sur la fouille de données du cardio-vasculaire**

On a présenté dans ce chapitre la maladie cardio-vasculaire, un état de l'art sur les techniques utilisées, les différents data sets de la maladie cardio-vasculaire y a compris le jeu de données de notre travail Z-alizadeh sani, diagnostique du cardio-vasculaire.

**Chapitre 4 : Implémentation et réalisation**

L'objectif principal de ce travail est de développer une application de prédiction destiné aux diagnostics des patients, ce chapitre est consacré à la présentation de l'application et des techniques de fouille de données et data set utilisées, les résultats de cette étude Enfin, en conclure.

# *Chapitre 1*

# *La Fouille De Données Médicale*

# Introduction

Ces dernières années, nous avons assisté à une forte croissance des moyens de production et de collecte de données. Cela est principalement dû au développement de la technologie des supports de stockage, à ses capacités et à la réduction significative de ses coûts. En raison de l'informatique rapide des entreprises, des services, du commerce et des télécommunications, la quantité de données disponibles augmente très rapidement. Cependant, analyser et utiliser ces données reste très difficile. Les modèles traditionnels de recherche d'informations ne conviennent pas pour traiter d'énormes blocs de données et sont souvent hétérogènes "trop de données et aucune réaction". C'est ce constat qui a permis au concept de data mining d'émerger et de vulgariser les méthodes analytiques.

Ce chapitre a pour objet est de présenter dans un premier temps les concepts liées à la fouille de données. Dans un second temps, il présente les techniques de data mining qu’on peut utiliser pour l’extraction des connaissances à partir des données médicale.

## 1.1 Définitions

### 1.1.1 Fouille de données (Data mining)

La fouille de données n'est pas née à l'ère numérique. Le concept existe depuis plus d'un siècle, mais il s'est vraiment fait connaître dans les années 1980, et depuis lors, il a parcouru un long chemin. Les entreprises utilisent désormais l'exploration de données et l'apprentissage automatique pour accomplir une multitude de tâches, de l'optimisation du processus de vente à l'interprétation des données financières pour un investissement.

L'exploration de données est le processus d'analyse de quantités massives de données et de méga données sous différents angles pour identifier les relations entre les données et les transformer en informations exploitables. Cet appareil s'inscrit dans le cadre de la veille économique et vise à aider les entreprises à résoudre les problèmes, à atténuer les risques, à identifier et à saisir de nouvelles opportunités commerciales.

Le data mining est un processus indissociable de l'analyse du Big Data, de l'intelligence prédictive et de l'exploration de données [1].

## 1.2 Processus du data mining

Il est très important de comprendre que le data mining n’est pas seulement le problème de découverte de modèles dans un ensemble de donnée. Ce n’est qu’une seule étape dans tout un processus suivi par les scientifiques, les ingénieurs ou toute autre personne qui cherche à extraire les connaissances à partir des données. En 1996 un groupe d’analystes définit le data mining comme étant un processus composé de cinq étapes sous le standard CRISP-DM (Cross-Industry Standard Process for Data Mining) comme schématisé ci- dessous[2]:

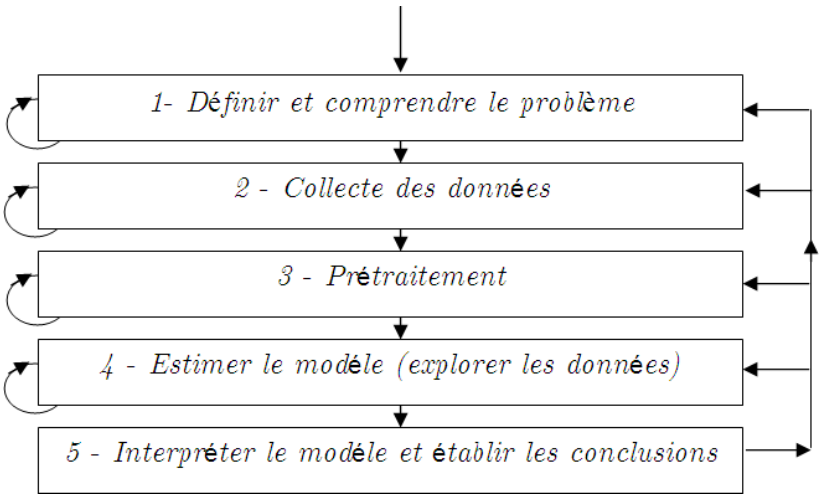


Figure 1: Processus du data mining [2]

Ce processus, composé de cinq étapes, n'est pas linéaire, on peut avoir besoin de revenir à des étapes précédentes pour corriger ou ajouter des données. Par exemple, on peut découvrir à l'étape d'exploration (5) de nouvelles données qui nécessitent d'être ajoutées aux données initiales à l'étape de collection (2). Décrivons maintenant ces étapes :

1. **Définition et compréhension du problème** : Dans la plus part des cas, il est indispensable de comprendre la signification des données et le domaine à explorer. Sans cette compréhension, aucun algorithme ne va donner un résultat fiable. En effet, Avec la compréhension du problème, on peut préparer les données nécessaires à l'exploration et interpréter correctement les résultats obtenus. Généralement, le data mining est effectué dans un domaine particulier (banques, médecine, biologie, marketing, ...etc) où la connaissance et l'expérience dans ce domaine jouent un rôle très important dans la définition du problème, l'orientation de l'exploration et l'explication des résultats obtenus. Une bonne compréhension du problème comporte une mesure des résultats de l'exploration, et éventuellement une justification de son coût. C'est-à-dire, pouvoir évaluer les résultats obtenus et convaincre l'utilisateur de leur rentabilité[2].
2. **Collecte des données** : dans cette étape, on s'intéresse à la manière dont les données sont générées et collectées. D'après la définition du problème et

des objectifs du data mining, on peut avoir une idée sur les données qui doivent être utilisées. Ces données n'ont pas toujours le même format et la même structure. On peut avoir des textes, des bases de données, des pages web, ...etc. Parfois, on est amené à prendre une copie d'un système d'information en cours d'exécution, puis ramasser les données de sources éventuellement hétérogènes (fichiers, bases de données relationnelles, temporelles,...).Quelques traitements ne nécessitent qu'une partie des données, on doit alors sélectionner les données adéquates. Généralement les données sont subdivisées en deux parties : une utilisée pour construire un modèle et l'autre pour le tester. On prend par exemple une partie importante (suffisante pour l'analyse) des données (80%) à partir de laquelle on construit un modèle qui prédit les données futures. Pour valider ce modèle, on le teste sur la partie restante (20%) dont on connaît le comportement[2].

3. **Prétraitement** : Les données collectées doivent être "préparées". Avant tout, elles doivent être nettoyées puisqu'elles peuvent contenir plusieurs types d'anomalies : des données peuvent être omises à cause des erreurs de frappe ou à causes des erreurs dues au système lui-même, dans ce cas il faut remplacer ces données ou éliminer complètement leurs enregistrements. Des données peuvent être incohérentes c-à-d qui sortent des intervalles permis, on doit les écarter où les normaliser. Parfois on est obligé à faire des transformations sur les données pour unifier leur poids. Un exemple de ces transformations est la normalisation des données qui consiste à la projection des données dans un intervalle bien précis [0,1] ou [0,100] par exemple. Un autre exemple est le lissage des données qui considère les échantillons très proches comme étant le même échantillon. Le prétraitement comporte aussi la réduction des données qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration. Une méthode largement utilisée dans ce contexte, est l'analyse en composantes principales (ACP). Une autre méthode de réduction est celle de la sélection et suppression des attributs dont l'importance dans la caractérisation des données est faible, en mesurant leurs variances. On peut même réduire le

nombre de données utilisées par le data mining en écartant les moins importantes. Dans la majorité des cas, le prétraitement doit préparer des informations globales sur les données pour les étapes qui suivent tel que la tendance centrale des données (moyenne, médiane, mode), le maximum et le mini- mum, le rang, les quartiles, la variance, ... etc. Plusieurs techniques de visualisation des données telles que les courbes, les diagrammes, les graphes,... etc, peuvent aider à la sélection et le nettoyage des données. Une fois les données collectées, nettoyées et prétraitées on les appelle entrepôt de données (data warehouse).

4. **Estimation du modèle** : Dans cette étape, on doit choisir la bonne technique pour extraire les connaissances (exploration) des données. Des techniques telles que les réseaux de neurones, les arbres de décision, les réseaux bayésiens, le clustering, ... sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat. Dans le reste de ce rapport on va détailler les différentes techniques utilisées dans l'exploration des données et l'estimation du modèle.
5. **Interprétation du modèle et établissement des conclusions** : généralement, l'objectif du data mining est d'aider à la prise de décision en fournissant des modèles compréhensibles aux utilisateurs. En effet, les utilisateurs ne demandent pas des pages et des pages de chiffres, mais des interprétations des modèles obtenus. Les expériences montrent que les modèles simples sont plus compréhensibles mais moins précis, alors que ceux complexes sont plus précis mais difficiles à interpréter[2].

## 1.3 Le Data mining, un processus dans l'ECD

Durant les dernières décennies, l'informatisation a facilité la génération, la manipulation et le stockage d'un grand nombre d'informations. En découle un volume de données toujours plus important, qui a entrainé avec lui une adaptation des outils statistiques nécessaires à leur exploitation. Parmi eux, l'Extraction de Connaissances à partir de Données (ECD) est déjà

largement répandue dans les domaines de l'industrie et des finances et tend à se populariser dans le domaine de la santé publique. L'objectif de l'ECD est d'identifier, dans des volumes importants de données, des relations jusqu'à lors inconnues, «cachées», pour aboutir à une connaissance nouvelle. Cette approche se distingue donc fondamentalement des statistiques dites «classiques», qui consistent en la vérification d'hypothèses formulées a priori. L'ECD fonctionne selon un processus cyclique dans lequel on distingue généralement 5 étapes [3] :

1- **Sélection** : Permet de sélectionner les données pertinentes pour la tâche de data mining à accomplir.
2- **Pré-traitement** : Cette phase traite la présence de bruits, d'erreurs et de données manquantes.
3- **Transformation** : Les données sont transformées ou consolidées dans un format approprié à la tâche de data mining choisie.
4- **Data Mining** : Dans cette phase, des méthodes intelligentes sont utilisées afin d'extraire des modèles, règles, etc.
5- **Interprétation, évaluation** : Enfin, cette étape identifie les modèles intéressants représentant les connaissances, en se basant non seulement sur des mesures d'intérêt mais aussi sur l'avis de l'expert

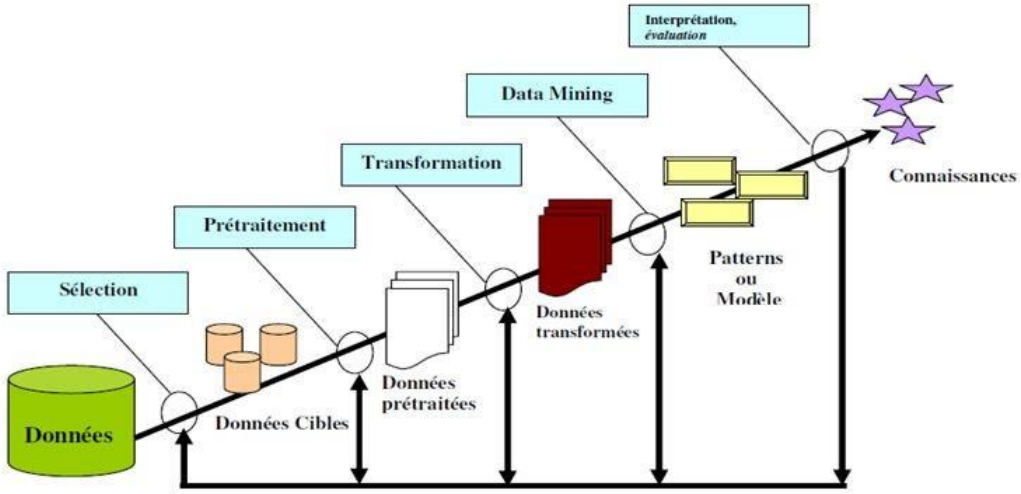


Figure 2 : Le Schéma du processus de data mining [4]

## 1.4 Les étapes du processus de data mining

L'exploration de données permet aux entreprises d'obtenir des informations intelligibles sur leurs données, qu'il s'agisse de données open source ou non. Cependant, le processus d'exploration de données est extensif, qui nécessite la combinaison d'un certain nombre d'étapes. Le processus d'exploration de données diffère d'un cas d'utilisation à l'autre et d'une entreprise à l'autre, mais ce guide d'exploration de données expliquera le processus d'une manière simple et basique. Cependant, la réponse à la question courante «combien d'étapes dans l'exploration de données» est qu'il y a sept étapes majeures dans l'exploration de données. Les étapes suivantes aident les utilisateurs à mieux comprendre comment démarrer l'exploration de données [5].

1- **Sélection des données :** La première étape du processus d'analyse d'exploration de données consiste à sélectionner les sources de données qui peuvent être utilisées pour extraire et obtenir des informations précieuses.

2- **Extraction de données :** La prochaine étape du processus d'exploration de données est la collecte et l'extraction de données. Un scientifique de données identifie les sources de données, les analyse et utilise le flux d'intégration pour consolider les données utiles.

3- **Transformer les données :** Une fois collectées, les données de différentes sources et de différents formats doivent être converties en un format commun pour pouvoir être utilisées.

4- **Données de nettoyage :** Une fois que les données sont transformées en un format commun, elles doivent être nettoyées afin de garantir que les données sont sans erreur, cohérentes et uniques. Le nettoyage des données consiste à minimiser la redondance des données, à les manipuler, à les organiser et à appliquer des règles de gouvernance pour que les données soient conformes aux normes de conformité.

5- **Stockage et gestion des données :** L'étape suivante consiste à stocker et à gérer les données dans différents entrepôts de données en fonction du type de données. Les données peuvent être transactionnelles, non opérationnelles ou des

métadonnées. Les données transactionnelles, qui incluent les opérations quotidiennes, sont stockées dans un emplacement distinct des données non opérationnelles. Les métadonnées concernent la conception de la base de données logique et sont également traitées séparément. Les données stockées sont ensuite mises à la disposition des analystes commerciaux à l'aide d'un logiciel d'application.

6- **Analyse et exploration de données :** Une fois les données collectées et chargées dans un entrepôt de données, le processus d'exploration de données proprement dit démarre. L'exploration et l'analyse nécessitent une combinaison d'algorithmes d'intelligence d'affaires et d'exploration de données. Comprendre l'entreprise permet aux scientifiques des données de produire plus facilement un modèle d'exploration de données pour l'analyse des données. Chaque algorithme d'exploration de données implique le processus d'identification des tendances dans un ensemble de données et l'utilisation des résultats obtenus pour définir des paramètres. Ces paramètres sont ensuite utilisés pour effectuer une analyse descriptive, une analyse diagnostique, une analyse prescriptive, une gestion des risques ou une analyse prédictive.

7- **Visualisation des données :** Après avoir obtenu les résultats du processus d'exploration de données, il est nécessaire de s'assurer que les données sont représentées visuellement sous une forme compréhensible. La visualisation de données permet aux entreprises de présenter les résultats générés à l'aide d'algorithmes d'exploration de données à l'aide de graphiques ou d'infographies[5].

## 1.5 Application d'exploration de données

L'exploration de données a des applications utiles dans différentes industries, telles que[5] :

### 1.5.1 Soins de santé:

L'exploration de données peut être utilisée dans le secteur de la santé pour réduire les coûts, détecter les activités frauduleuses et améliorer les résultats pour les patients.

### 1.5.2 L'éducation:

L'utilisation d'outils d'exploration de données dans l'éducation peut aider différents aspects de l'industrie de l'éducation, tels que l'identification de la façon d'encourager les besoins d'apprentissage des étudiants, la prédiction de la performance de certains étudiants aux examens et la prise de décisions opérationnelles efficaces.

### 1.5.3 Gestion de la relation client (CRM):

L'exploration de données peut aider à analyser les données client afin d'aider une entreprise à adopter des stratégies centrées sur le client et à établir des relations fructueuses, fidèles et durables avec leurs clients ou clients.

### 1.5.4 L'aspect médical:

L'importance de l'étude découle de l'utilisation de l'exploration de données dans le domaine médical pour aider à prendre la bonne décision avec la vitesse et la précision requises, car les analystes humains perdent beaucoup de temps à analyser les données. Utiliser l'approche d'analyse descriptive pour analyser et concevoir des systèmes tout au long du cycle de vie de développement des systèmes.

## 1.6 Les taches de la fouille de données

Beaucoup de problèmes intellectuels, économiques ou même commerciaux peuvent être exprimés en termes des six tâches suivantes:

- La classification et la prédiction.
- La regression
- L'estimation.
- Le groupement par similitude (règles d'association).
- L'analyse des clusters.

- La description.

Les trois premières tâches sont des exemples de la fouille supervisée de données dont le but est d'utiliser les données disponibles pour créer un modèle décrivant une variable particulière prise comme but en termes de ces données. Le groupement par similitude et l'analyse des clusters sont des tâches non-supervisées où le but est d'établir un certain rapport entre toutes La description appartient à ces deux catégories de tâche, elle est vue comme une tâche supervisée et non-supervisée en même temps [2].

### 1.6.1 Classification

La classification est la tache la plus commune de la fouille de données qui semble être une tâche humaine primordiale. Afin de comprendre notre vie quotidienne, nous sommes constamment obligés à classer, catégoriser et évaluer. La classification consiste à étudier les caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie. Les objets à classifier sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jours chaque enregistrement en déterminant la valeur d'un champ de classe. Le fonctionnement de la classification se décompose en deux phases. La première étant la phase d'apprentissage. Dans cette phase, les approches de classification utilisent un jeu d'apprentissage dans lequel tous les objets sont déjà associés aux classes de références connues. L'algorithme de classification apprend du jeu d'apprentissage et construit un modèle. La seconde phase est la phase de classification proprement dite, dans laquelle le modèle appris est employé pour classifier de nouveaux objets.

### 1.6.2 La régression

La régression est une méthode permettant de restituer les interactions des différents paramètres L'équation de régression est alors, selon Tomassone, "un outil de description permettant de préciser les relations entre les régresseurs, les x et d'analyser leur action sur une variable y". A des fins de simplicité, il faut dans un

premier temps tenter de rendre l'évolution d'un phénomène linéaire. Dans ce cas, les paramètres interviennent de façon linéaire, éventuellement après transformation.
Si les transformations ne permettent pas de se ramener à un modèle linéaire ou si le modèle linéaire n'est pas suffisamment robuste, on doit alors envisager de produire un modèle non-linéaire. Cette démarche méthodologique a pour but de déterminer le carré des sommes $R^2$ le plus élevé possible afin d'obtenir l'ajustement le plus parfait possible.

Ces analyses reposent en fait sur une décomposition fine d'un nuage de points multidimensionnel. Le fait de disposer de n variables détermine un espace d'analyse à n dimensions. La problématique de la régression multidimensionnelle est donc de situer un plan d'estimation passant aussi près que possible des différents points du nuage. Pour ce faire, on aura recours à la méthode des moindres carrés. Le coefficient de corrélation multiple ($R^2$) mesure la qualité de l'estimation.

Le paramètre $R^2$ mesure en fait la partie de la variance d'une variable dépendante qui puisse être expliquée par une combinaison linéaire des différentes variables indépendantes ou explicatives. Il en résulte que la variabilité de la variable analysée peut s'expliquer à concurrence de ($R^2$*100) % par l'influence combinée des variables explicatives.

Dans le cadre de relations complexes, il peut être intéressant de disposer de limites internes à chaque variable qui correspondent en ce sens au domaine de définition de la fonction. Ces limites peuvent être constituées de manière purement statistique -on prendra dans ce cas les paramètres centraux de la distribution en l'occurrence les déciles. Ou selon des considérations de seuils représentatifs.

### 1.6.3 L'estimation

L'estimation est similaire à la classification à part que la variable de sortie est numérique plutôt que catégorique. En fonction des autres champs de l'enregistrement l'estimation consiste à compléter une valeur manquante dans un champ particulier. Par exemple on cherche à estimer la lecture de tension systolique d'un patient dans un hôpital, en se basant sur l'âge du patient, son genre, son indice de masse corporelle et le niveau de sodium dans

son sang. La relation entre la tension systolique et les autres données vont fournir un modèle d'estimation. Et par la suite nous pouvons appliquer ce modèle dans d'autres cas [2].

### 1.6.4 Le groupement par similitude (règle d'association)

(Analyse des associations et de motifs séquentiels) Le groupement par similitude consiste à déterminer quels attributs "vont ensemble". La tâche la plus répandue dans le monde du business, est celle appelée l'analyse d'affinité ou l'analyse du panier du marché, elle permet de rechercher des associations pour mesurer la relation entre deux ou plusieurs attributs. Les règles d'associations sont, généralement, de la forme "Si <antécédent>, alors <conséquent>".

### 1.6.5 L'analyse des clusters

Le clustering (ou la segmentation) est le regroupement d'enregistrements ou des observations en classes d'objets similaires. Un cluster est une collection d'enregistrements similaires l'un à l'autre, et déférents de ceux existants dans les autres clusters. La déférence entre le clustering et la classification est que dans le clustering il n'y a pas de variables sortantes. La tâche de clustering ne classifie pas, n'estime pas, ne prévoit pas la valeur d'une variable sortantes. Au lieu de cela, les algorithmes de clustering visent à segmenter la totalité de données en des sous groupes relativement homogènes. Ils maximisent l'homogénéité à l'intérieur de chaque groupe et la minimisent entre les différents groupes.

### 1.6.6 La description

Parfois le but de la fouille est simplement de décrire se qui se passe sur une Base de Données compliquée en expliquant les relations existantes dans les données pour premier lieu comprendre le mieux possible les individus, les produit et les processus présents dans cette base. Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. Dans la

société Algériennes nous pouvons prendre comme exemple comment une simple description, "les femmes sup- portent le changement plus que les hommes", peut provoquer beaucoup d'intérêt et promouvoir les études de la part des journalistes, sociologues, économistes et les spécialistes en politiques [2].

## 1.7 Fouille de données médicale

### 1.7.1 Définition

Fouille de données médicale est un domaine qui vise à développer des méthodes pour explorer les types de données utiles qui nous aident à connaître le diagnostic réel d'un patient dans les établissements de santé et à utiliser ces méthodes pour mieux diagnostiquer ou anticiper les diagnostics des patients et les traitements que nous leur fournir. Les principales utilisations comprennent l'extraction de données médicales et la prédiction des résultats des tests et des investigations pour recommander des améliorations à la pratique médicale actuelle. L'exploration de données et l'analyse médicale sont un domaine connexe pour soutenir la décision fondée sur les données pour améliorer la pratique médicale.

### 1.7.2 Objectif de la fouille de données médicale

L'objectif de la fouille donnée médicales, en particulier dans la recherche sur les maladies cardiovasculaires. Bien que le nombre de patients ait diminué au cours des dernières décennies, les maladies cardiovasculaires restent la principale cause du taux de mortalité dans les pays de l'OCDE. Il apparaît donc important de comprendre et d'identifier les facteurs d'influence afin de diagnostiquer ou de prédire le pronostic des patients. Notre projet repose sur la prédiction des diagnostics des patients et l'utilisation d'algorithmes d'exploration de données pour mieux comprendre les facteurs opérant dans le contexte des maladies cardiovasculaires. Notre approche propose de mettre en œuvre différents algorithmes d'extraction de données pour prédire les diagnostics ainsi que des visualisations selon différentes métriques pour comprendre les indicateurs.

### 1.7.3 Domaine d'application

« **Selon Wikiversité**» Le Datamining est une approche d'analyse de donnée, adaptée et utilisée dans un large nombre de domaine d'activités.

**1- Assurances et santé**

- Découverte d'associations des demandes de remboursements
- Identification de clients potentiels de nouvelles polices d'assurances.
- Détection d'association de comportements pour la découverte de clients à risque.
- Détection de comportement frauduleux.

**2- Banques / Finances**

- Détection d'usage frauduleux de cartes bancaires.
- Gestion du risque lié à l'attribution de prêts bancaires par le scoring.
- Découverte de relations cachées entre les indicateurs financiers.
- Détection de règles de comportement boursier par l'analyse des données du marché.

**3- Vente, distribution / Marketing**

- La gestion de la relation client (GRC ou CRM) consiste en l'ensemble des activités visant à cibler, attirer et conserver les "bons" clients.
- Détection d'associations de comportements d'achat.
- Découverte de caractéristiques de clientèle.
- Prédiction de probabilité de réponse aux campagnes de mailing.

**4- Ressources Humaines**

Le Datamining est également utilisé dans les ressources humaines (RH) de certains ministères pour identifier les caractéristiques de leurs employés les plus performants. L'information obtenue (comme les universités fréquentées par des employés potentiels) peut contribuer aux efforts de recrutement des ressources humaines.

Ces dernières années, l'exploration de données a été largement utilisé dans les domaines de la science et de l'ingénierie, tels que la bioinformatique, la génétique, la médecine, l'éducation et l'énergie électrique.

**5- Médical / Pharmaceutique**

- Diagnostic assisté par ordinateur (CAD) par l'apprentissage de systèmes experts.
- Explication ou prédiction de la réponse d'un patient à un traitement.
- Identification des thérapies à succès (combinaison de prescriptions).
- Étude des corrélations entre le dosage dans un traitement et l'apparition d'effets secondaire.

**6- La génétique humaine**

Dans l'étude de la génétique humaine, le Datamining permet de répondre à l'objectif important de comprendre la relation de correspondance entre l'ADN et les maladies. En effet, il vise à savoir comment les changements dans la séquence d'ADN d'un individu affectent les risques de développer des maladies courantes telles que le cancer, qui est d'une grande importance à l'amélioration des méthodes de diagnostic, la prévention et le traitement de ces maladies. Le data mining peut contribuer de manière significative et avec succès à l'explication ou la prédiction de phénomènes complexes dans les domaines médical et pharmaceutique.

**7- Ingénierie électrique**

Dans le domaine de l'ingénierie électrique, les Datamining ont été largement utilisés pour la surveillance de l'état du matériel électrique à haute tension. Le but de surveillance de l'état est d'obtenir de précieuses informations par exemple, sur l'état de l'isolation (ou d'autres importantes des paramètres de sécurité).

**8- Aéro-spatiale**

Le Datamining est également intégré aux données spatiales. L'objectif final est de trouver des modèles dans les données relatives à la géographie. Jusqu'à présent, l'exploration de données et de systèmes d'information géographiques ont

existé en tant que deux technologies distinctes, chacune avec ses propres méthodes. L'immense explosion de données géo-référencées occasionnée par l'évolution de l'informatique, la cartographie numérique, la télédétection et la diffusion mondiale des systèmes d'information géographiques mettent l'accent sur l'importance de développer une analyse et une modélisation géographique plus fines.

### 1.7.4 Applications de la fouille de données dans le domaine de la santé

Fouille de données est l'un des outils les plus importants utilisés dans le domaine de la santé, les applications de la fouille de données dans le domaine de la santé sont concentrées dans :

1- Exploration et évaluation des conditions sanitaires en vigueur.
2- Rechercher les causes des maladies.
3- Explorer les comportements satisfaisants dans la communauté.
4- Contribuer à l'élaboration de plans et politiques médicaux et sanitaires appropriés.
5- Travailler pour limiter la propagation des maladies et des épidémies.

Partout où des bases de données médicales et sanitaires existent, l'exploration de données peut être utilisée pour étudier, analyser et explorer tout ce qui contribuerait à améliorer la situation sanitaire en général dans la société, à développer la performance des établissements de santé et à réduire les risques d'exposition aux maladies.

L'application de la fouille de données est également utilisée pour développer des politiques et des procédures de sensibilisation à la santé pour l'individu et la famille

L'application de la fouille de données aide à explorer et à caractériser les maladies les plus courantes dans des régions, des époques ou des circonstances et conditions spécifiques.

Ces applications visent généralement à développer des solutions adaptées et à prendre les précautions nécessaires pour limiter la propagation des maladies.

## 1.8 Les logiciels de Data Mining

« **Selon Wikipédia** » Les logiciels de fouille de données sont des programmes spécialisés dans l'analyse et l'extraction des connaissances à partir des données informatisées. Ce sont des logiciels qui aident l'analyste en exploration de données à trouver des motifs remarquables et intéressants. Il peut s'agir de logiciels commerciaux ou de logiciels libres.

### 1.8.1 Logiciels libres

Parmi les logiciels libres : KNIME et Weka ces deux logiciels sont décrits ci-dessous.

**KNIME :** Acronyme de Konstanz Information Miner, est un logiciel libre édité par un laboratoire de l'université de Constance dénommé Nycomed Chair for Bio-informatics and Information Mining. Il intègre notamment tous les modules d'analyse de Weka et permet de créer des scripts en langage R. KNIME s'exécute sur Linux, Windows et MacOS. Comme tous les logiciels libres, KNIME est extensible.

**Weka :** Est un logiciel libre de fouille de données développé en java et créé par l'université de Waikato (Nouvelle-Zélande). C'est une collection d'algorithmes d'apprentissage automatique mis en place pour effectuer des tâches d'exploration de données. Les algorithmes peuvent soit être appliqués directement à un ensemble de données soit être appelés directement par un code Java. Weka contient des outils pour les prétraitements des données, la classification, la régression, le clustering, les règles d'association et la visualisation. Comme KNIME, weka est un logiciel open source.

**RapidMiner :** Est un logiciel open source dédié au data mining. Il contient de nombreux outils pour traiter des données : lecture de différents formats d'entrée, préparation et nettoyage des données, statistiques, tous les algorithmes de data mining, évaluation des performances et visualisations diverses. C'est un logiciel puissant, il n'est pas facile à manipuler au premier abord, mais avec un peu de pratique, il permet de mettre en place rapidement une chaîne complète de traitement de données, de la saisie des données à leur classification.

### 1.8.2 Logiciels commerciaux

Les logiciels commerciaux sont édités par des sociétés bien connues sur le marché:

**KXEN :** Analytic Framework est un logiciel commercial édité par la société KXEN basée en Californie et fondée en 1998. Les modules de KXEN Analytic Framework per- mettent la prédiction, la segmentation, les associations, la fouille de textes et l'analyse des réseaux sociaux.

**SAS Enterprise Miner :** est un outil commercial édité par la société SAS Institute Inc. C'est un logiciel offrant toutes les facettes de l'exploration de données dont le processus est facilité par son interface homme-machine bien conçue.

**SPSS (Statistical Package for the Social Sciences) :** est un logiciel de statistiques, édité par la filiale d'IBM du même nom, qui se décompose en plusieurs modules dont SPSS Modeler pour le Data mining, SPSS Amos pour les modèles d'équation structurelle et Predictive Analytics pour l'analyse prédictive.

**CORICO :** est un logiciel commercial intégrant l' Iconographie des corrélations et les Interactions logiques, qui se prêtent bien à l'analyse multi relationnelle. Il intègre aussi une technique de modélisation prédictive fondée sur les modèles de régression multiple postulés et non postulés.

# Conclusion

Les développements scientifiques et la diffusion de la technologie dans divers aspects de la vie quotidienne ont augmenté la capacité de générer et de collecter rapidement des données à cette époque, et les progrès technologiques ont provoqué l'émergence de nouveaux types de données tels que des textes, des images, des vidéos et des systèmes de tâches dans en plus d'Internet qui contient d'énormes quantités de données sous toutes leurs formes. Tout cela a conduit à une augmentation sans précédent de la quantité de données stockées quotidiennement, révélant le besoin urgent de nouvelles technologies et d'outils intelligents qui peuvent aider à transformer cette énorme quantité de données en informations et connaissances utiles. Ceux représentés dans les outils d'exploration de données, et leur utilisation fournit aux entreprises et aux institutions dans tous les domaines civils et gouvernementaux, la possibilité d'explorer les informations les plus importantes et de se concentrer sur elles dans de grands blocs de données, et les techniques d'exploration se concentrent sur la découverte et la construction de prévisions futures et explorer les modèles, les corrélations, les comportements et les tendances, ce qui aide à évaluer les bonnes décisions. Et les prendre en temps opportun et développer des solutions appropriées aux problèmes, la planification, le développement et la modernisation dans tous les domaines.

# *Chapitre 2*

# *Classification et prédiction de données médicales*

# Introduction

Les données médicales souffrent de problèmes d'uniformisation ou d'incertitude, ce qui les rend difficilement utilisables directement par des logiciels médicaux, La classification de données médicales présente plusieurs challenges. D'une part, ces données présentent souvent une asymétrie au niveau de la pathologie à prédire. A travers ce qui précède, nous présenterons dans ce chapitre la classification, ses étapes, classification supervisée dans le domaine médicales, classification des données médicales ... etc. Enfin, nous détaillerons ses principales approches en étudiant et analysant quelques algorithmes.

## 1.1 Définitions

### 1.1.1 Classification

La classification consiste à étudier les caractéristiques d'un nouvel objet pour lui attribuer une classe prédéfinie. Les objets à classifiés sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification est caractérisée par une définition de classes bien précise et un ensemble d'exemples classés auparavant. L'objectif est de créer un modèle qui peut être appliqué aux données non classifiées dans le but de les classifiées [6].

### 1.1.2 Classification supervisé

La classification supervisée est une technique largement utilisée avec différentes applications dans la vie réelle. Elle permet de générer des règles de classification (modèle) à partir d'un jeu données classées à priori et d'un algorithme d'apprentissage automatique adéquat. Ces règles seront utilisées pour classer les nouvelles instances [7].

### 1.1.3 Classification non supervisé

Elle sert a établir des représentations des données dans des espaces à faible dimensions pour y lire des typologies d'individus tel que le nombre de classes n'est pas connu. Parmi les méthodes de classification non supervisée en trouve [8]:

- Analyse en composante principales (ACP)
- Analyse des associations.
- Analyse en clusters

### 1.1.4 Prédiction

La prédiction est la même que la classification, à part que dans la prédiction les enregistrements sont classés suivant des critères (ou des valeurs) numériques. La principale raison qui différencie la prédiction de la classification est que dans la création du modèle prédictif l'attribut prédit doit prendre une valeur numérique et non pas catégorique [9].

## 1.2 Les étapes de la classification

Le processus de classification comprend les étapes suivantes [4] :

- Représentation des individus.
- Définition d'une mesure de similarité appropriée aux données.
- Classification.
- Abstraction des données.
- Validation du résultat.

### 1.2.1 Représentation des individus

A pour but de déterminer les informations concernant les données : le nombre de classes désiré, le nombre d'individus disponibles, le nombre, le type et l'échelle des attributs de données. Ces informations sont utilisées dans l'algorithme de classification

### 1.2.2 La proximité des individus

Est souvent mesurée par une fonction de distance entre chaque pair d'individus. De nombreuses mesures de proximité sont proposées, se basant sur la nature de données.

### 1.2.3 Classification

Est une phase de groupement des individus dans les classes. Plusieurs algorithmes de classification sont proposés. La différence entre eux est la manière dont ils groupent les individus telles que la méthode hiérarchique, la méthode de partition…,

le type de données qu'ils traitent comme des données numériques, de catégorie, le flux de données…, la mesure de proximité des individus et des classes qu'ils utilisent, telle que le critère selon lequel on construit des classes [4].

### 1.2.4 Abstraction des données

Est un processus d'extraction d'une représentation simple et compacte pour un jeu de données. Typiquement, une abstraction des données est une description compacte de chaque classe, souvent en termes de prototypes des classes ou d'individus représentatifs des classes comme le centre des classes.

### 1.2.5 Validation du résultat

Vise à déterminer si les classes fournies sont significatives en utilisant un critère spécifique d'optimalité. Cependant, un tel critère est souvent subjectif, donc il y a peu de manière standard pour valider la classification sauf dans certains domaines bien décrits à priori.

## 1.3 Classification supervisée dans le domaine médical

Dans le domaine médical, les organisations de santé, telles que les hôpitaux, les cliniques et les laboratoires, souhaitent unifier leurs bases de données et coopérer ensemble pour concevoir des modèles de mines de données plus efficaces, sont un domaine d'études. Améliorer les connaissances en matière de santé. Pour ce faire, dépend spécialement de la collecte d'informations médicales.

La recherche médicale est indispensable pour le progrès médical. Vise à mieux connaître les gens, mieux:

- Piste ou contrôle (par exemple via des tests de diagnostic),
- La guérison ou les empêcher (avec des médicaments et des dispositifs médicaux…).

## 1.4 Algorithmes de la classification supervisée

Ils existent différents algorithmes pour l'extraction de la connaissance selon l'objectif et le type d'apprentissage. Bien que deux algorithmes d'apprentissage puissent différer dans leur type d'apprentissage, i.e. la nature de leur objectif, ils peuvent aussi se distinguer par la façon qu'ils accomplissent cet apprentissage. Nous allons passer en revue quelques approches populaires utilisées dans l'apprentissage automatique.

### 1.4.1 Naïve Bayes

Naïve Bayes est un classificateur probabiliste simple. Il calcule un ensemble de probabilités en comptant la fréquence et les combinaisons de valeurs dans un jeu de données. L'algorithme utilise le théorème de Bayes et suppose que tous les attributs sont indépendants compte tenu de la valeur de la variable classe. Cette hypothèse d'indépendance conditionnelle est rarement valable dans les applications du monde réel, d'où la caractérisation naïve. Cependant, l'algorithme tend à bien fonctionner et à apprendre rapidement dans divers problèmes de classification supervisée [10].

Compte tenu de la classe y et du vecteur de données (x1, x2, x3,…, xn), le théorème de Bayes énonce la relation suivante:

$$P(y|x1,x2,x3,\dots,xn) = \frac{P(y)P(x1,x2,x3,\dots,xn|y)}{P(x1,x2,x3,\dots,xn|y)} \quad (1)$$

$$P(y|x1,x2,x3,\dots,xn) = \prod_{i=1}^{n} P(xi|y) \quad (2)$$

Pour tous les xi cette relation est simplifiée comme suit :

$$P(y|x1,x2,x3,\dots,xn) = \frac{P(y)P\prod_{1}^{n} P(xi|y)}{P(x1,x2,x3,\dots,xn)} \quad (3)$$

[2]Puisque P (x1, x2, x3,…, xn) est constant pour l'entrée, nous pouvons utiliser la règle de classification suivante:

$$y = \arg\max(P(y)\prod_{i=1}^{n} P(xi|y)) \quad (4)$$

### 1.4.2 Arbres de décision

Les arbres de décisions sont des techniques très populaires par leur efficacité et leur simplicité dans le domaine de la classification supervisée. Ils fournissent une représentation graphique du modèle facilement interprétable [11].

Le modèle final est constitué d'un nœud racine et des nœuds intermédiaires, des branches et des feuilles. La racine est le point d'entrée à l'arbre. Les feuilles représentent les valeurs classes à prédire. Les branches représentent les résultats de test relatif à chaque nœud. Poureffectuer une classification, l'arbre est parcouru de la racine aux feuilles selon une série de tests à chaque niveau de l'arbre. La théorie de Shannon est à la base de partitionnement de plusieurs arbres de décision. Elle est définie comme suit :

La quantité d'information associée au nœud x est

$$I(x) = -\sum_{j} (x_j)\log P(x_j) \quad (5)$$

$$p_j = \frac{n_j}{n_s} \quad (6)$$

$n_j$ représente le nombre d'instances appartenant à la classe j et $n_s$ représente le nombre totald'instance du nœud s.

Le gain d'information est mesuré par la différence entre l'impureté du nœud parent s et la somme des impuretés des p nœuds fils obtenus grâce à un attribut X.

$$Gain(s,x) = I(s) - \sum_{i=1}^{p} \frac{n_i}{n} I(s_i) \quad (7)$$

$n_j$ représente le nombre d'instances total du nœud s et n représente le nombre d'instances total du nœud parent [12].

### 1.4.3 K plus proches voisins (KNN)

Le principe de la méthode KNN est de trouver k plus proches voisins, à partir de l'échantillon d'apprentissage, à une nouvelle instance qu'on cherche à classer. La classe de la nouvelle instance est la classe majoritaire (la plus représentée) parmi ces k voisins. Dansle cas d'une régression, la valeur de sortie est une valeur continue qui peut être, par exemple, la moyenne des valeurs des k voisins.

Il existe plusieurs fonctions pour calculer la distance entre deux voisins, notamment, la distance euclidienne, la distance de Manhattan, la distance de Minkowski, la distance de Jaccard, etc. Dans ce qui suit, noud définissons la distance euclidienne et la distance de manhattan.

Soit $x_i = (x_{1i}, x_{2i}, \ldots x_{pi})$ un vecteur représentant les variables prédicatrices de l'instance i. La distance euclidienne entre les deux instances $x_i$ $et$ $x_j$ est définie comme suit [13] :

$$d_{euclidienne}(x_i, x_j) = \sqrt{\sum_{s=1}^{p} (x_{is} - x_{js})^2} \quad (8)$$

$$d_{manhattan}(x_i, x_j) = \sum_{s=1}^{p} |x_{is} - x_{js}| \quad (9)$$

Il est à noter que, KNN n'a pas une phase d'apprentissage où un modèle Data Mining est généré. Les exemples d'apprentissage sont des vecteurs dans un espace multidimensionnel avec un label de classe d'appartenance, ils sont stockés en permanence dans la mémoire lors de la phase de classification.

### 1.4.4 Support Vector Machine (SVM)

Le classificateur SVM, développée par Vladimir Vapnik en 1995, est un classificateur puissant, il a fait ses preuves dans plusieurs domaines. Le principe est de projeter les données qui sont non linéairement séparables dans un autre espace de dimension plus élevée où elles peuvent le devenir, en utilisant différents noyaux. Le but du SVM binaire est de trouver un hyperplan optimal qui sépare les deux classes en maximisant la distance. Cette distance est appelée marge. Dans le cas d'une classification binaire, Figure 3, l'hyperplan est une droite. Les points les plus proches, qui seuls sont utilisés pour la détermination de la marge, sont appelés vecteurs de support.

L'hyperplan séparateur est représenté par l'équation

$$H(x) = W^T x + b \qquad (10)$$

w est un vecteur de m dimensions et b est un terme [14]. La fonction de décision, pour un exemple x, peut être exprimée comme suit :

$$\begin{cases} Classe = 1 Si\ H(x) > 1 \\ Classe = -1 Si\ H(x) < -1 \end{cases} \qquad (11)$$

Maximiser la marge revient maximiser $= \frac{2}{\|w\|}$ et qui vaut à minimiser $\|w\|$.

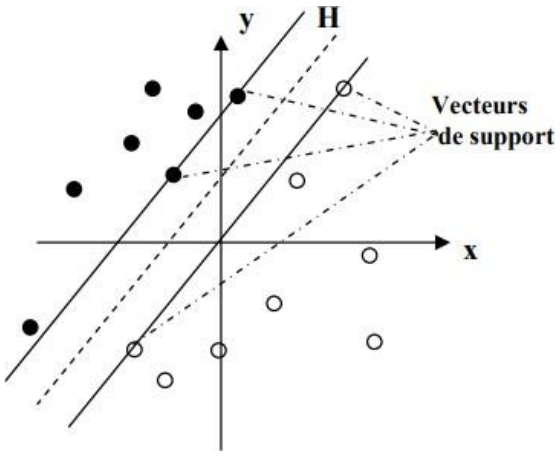


Figure 3 : SVM classification binaire

SVM réduit le problème multi classe à une composition de plusieurs hyperplans bi-classe permettant de tracer les frontières de décision entre les différentes classes. Il décompose l'ensemble d'exemples en plusieurs sous-ensembles représentant chacun un problème de classification binaire. A chaque fois un hyperplan de séparation est déterminé par la méthode SVM binaire. On construit lors de la classification une hiérarchie des hyperplans binaires qui est parcourue de la racine jusqu'à une feuille pour décider de la classe d'un nouvel exemple [14].

## 1.5 Classification des données médicales

### 1.5.1 Digitalisation du domaine médical

L'utilisation des nouvelles technologies de l'information dans le domaine médical a permisde muter la médecine traditionnelles vers d'autres pratiques modernes où elle est plus proactive, précise, personnalisée et rapide pour diagnostiquer et soigner les maladies et ce, àun coût compétitif. La digitalisation numérique a facilité l'accès, le transfert et le partage d'informations médicales qui sont essentielle pour la prise de décision Dumez et al. Elle a créé un moyen de communication et de coopération rapide entre les différents acteurs du domaine de la santé (médecins généralistes, médecins spécialistes, les chirurgiens, les radiologues, le personnel soignant, les pharmaciens, etc.). En conséquence, cela a permis une meilleure prise en charge des patients [7].

Conscient de ce changement incontournable et stratégique, tous les organismes de santés publiques et libérales, à savoir les hôpitaux, les cliniques, les laboratoires, les centres de radiologies, les centres d'analyse, etc., ont informatisé leurs documents et leurs processus. En effet, ils ont :

- Numérisé tous les documents papiers, surtout le dossier patient qui est le noyau de cette digitalisation.
- Acquis des ordinateurs, des serveurs et des solutions de stockage;
- Installé des réseaux informatiques;

- Intégré des outils et des équipements médicaux, numériques et connectés, pour bénéficier d'une plus grande précision et efficacité;
- Développé des systèmes d'information hospitaliers (SIH) pour faciliter la gestion des informations médicales et la gestion administratives.
- etc.

Mais en contrepartie, à l'instar des autres domaines, et grâce aux progrès technologiques qui ont réduits les coûts du stockage, la numérisation du domaine de la santé a provoqué une grande explosion des données. Cela est dû aux :

- Les SIH, qui stockent toutes les données relatives aux patients admis à un organisme médical. A savoir : tous les résultats des diagnostiques, les données des essais cliniques, les données génétiques, les données biocliniques, les données pathologiques, les prescriptions de pharmacie, les résultats de laboratoires, les radiologies, les scanners, etc.
- Les objets médicaux connectés (IoT) génèrent aussi un grand flux de données en permanence sur les patients et à n'importe quel endroit (au travail, à domicile, en promenade, pendant le sommeil, etc.). Différents indices de mesures (telles que le taux de sucre dans le sang, le rythme cardiaque, la température corporelle, la tension, l'alimentation, les déplacements, la prise des médicaments, le poids, les paramètres environnementaux, etc.) sont collectées et stockées dans de grande base de données.
- Etc.

De nos jours, avec le développement des techniques Data Mining, cette grande masse de données constitue un précieux trésor pour extraire de la connaissance, non visible, qui sera utilisée pour développer le secteur de la santé. Dans ce chapitre, nous allons décrire les avantages issus du croisement entre la fouille de données et la science de la médecine [7].

### 1.5.2 Evaluation des modèles de classification

L'apprentissage supervisé utilise une partie des données pour calculer un modèle de décision qui sera généralisé sur l'ensemble du reste de l'espace. Il est très important d'avoir des mesures permettant de qualifier le comportement du modèle appris sur les données non utilisées lors de l'apprentissage. Ces métriques sont calculées soit sur les exemples d'entrainement eux mêmes ou sur des exemples réservés d'avance pour les tests.

$$Taux\ de\ classification\ = \frac{(TP + TN)}{(TP + FP + TN + FN)} * 100 \quad (12)$$

$$Précision\ (Positif) = \frac{TP}{TP+FP} \quad (13)$$

$$Précision\ (Négatif) = \frac{TN}{TN+FN} \quad (14)$$

$$Rappel\ (Positif) = \frac{TP}{TP+FN} \quad (15)$$

$$Rappel\ (Négatif) = \frac{TN}{TN+FP} \quad (16)$$

$$F - score = \frac{Précision * Rappel}{Précision + Rappel} \quad (17)$$

La mesure F-score, appelée également F-mesure ou F1, fournit une mesure plus réaliste sur les performances en fonction des classes. Elle calcule la moyenne harmonique et pondérée entre la précision et le rappel. La précision, également appelée valeur prédictive positive, correspond à la proportion de résultats positifs réellement positifs. Le rappel, également appelé sensibilité, est la capacité d'un test à identifier correctement les résultats positifs pour obtenir le taux de réponse positif réel [7].

### 1.5.3 La matrice de confusion

Dans le contexte de la classification supervisée, la matrice de confusion, appelée aussi matrice de contingence, est un outil qui sert à évaluer les performances

d'un algorithme de classification. Elle synthétise les informations sur les classes réelles et les classes prédites par le modèle. Les colonnes de la matrice représentent les classes estimées et les lignes représentent les classes réelles des instances testées. Différentes métriques sont calculées à partir de la matrice de confusion. Nous citons quelques-unes, calculées dans le cas d'une classification binaire (Tableau 1), comme suit [7] :

| **Class actuelles** | **Classes prédites** | |
|---|---|---|
| | **Positif** | **Négatif** |
| Positif | Vrai positif (TP) | Faux négatif (FN) |
| Négatif | Faux positif (FP) | Vrai négatif (TN) |

Tableau 1: Matrice de confusion pour une classification supervisée binaire

# Conclusion

Nous avons généralement traité dans ce chapitre des concepts et définitions (classification, prédiction, classification supervisée, classification non supervisée, prédiction), puis discuté des étapes de classification, qui contient cinq étapes importantes, puis nous avons abordé la classification supervisée dans le domaine médical, puis nous avons examiné l'algorithme de classification supervisée a savoir naive_bayes, arbre de décision, .

# *Chapitre 3*

# *Etat de l'art sur la fouille de données du cardio-vasculaire*

# Introduction

Les chercheurs ont noté que les données explorent. Cela peut aider à déterminer ou à prédire une maladie cardiaque élevée ou faible. Ils ont trouvé par une expérience que le classeur SVM et les arbres de résolution sont efficaces dans le diagnostic de maladie cardio-vasculaire. A travers ce qui précède, nous présenterons dans ce chapitre la maladie cardio-vasculaire, Etat de l'art, les données de la maladie cardio-vasculaire (data set Z-Alizadeh sani, Clevland, …), les techniques utilisées pour le diagnostique du cardio-vasculaire.

## 1.1 Maladie du cardio-vasculaire

Les maladies cardio-vasculaires sont la première cause de mortalité dans le monde il meurt chaque année plus de personnes en raison de maladies cardio-vasculaires que de toute autre cause.

On estime à 17,7 millions le nombre de décès imputables aux maladies cardio-vasculaires, soit 31% de la mortalité mondiale totale. Parmi ces décès, on estime que 7,4 millions sont dus à une cardiopathie coronarienne et 6,7 millions à un AVC (chiffres 2015). Plus des trois quarts des décès liés aux maladies cardiovasculaires interviennent dans des pays à revenu faible ou intermédiaire. Sur les 17 millions de décès survenant avant l'âge de 70 ans et liés à des maladies non transmissibles, 82% se produisent dans des pays à revenu faible ou intermédiaire et 37% sont imputables aux maladies cardiovasculaires. Il est possible de prévenir la plupart des maladies cardiovasculaires en s'attaquant aux facteurs de risque comportementaux – tabagisme, mauvaise alimentation et obésité, sédentarité et utilisation nocive de l'alcool – à l'aide de stratégies à l'échelle de la population. Les personnes souffrant de maladies cardiovasculaires ou exposées à un risque élevé de maladies cardiovasculaires (du fait de la présence d'un ou plusieurs facteurs de risque comme l'hypertension, le diabète, l'hyperlipidémie ou une maladie déjà installée) nécessitent une détection précoce et une prise en charge comprenant soutien psychologique et médicaments, selon les besoins [15].

## 1.2 Etat de l'art

Les chercheurs ont observé que l'exploration des données. pourrait aider à identifier ou à prédire les maladies cardiaques à risque élevé ou faible. Ils ont constaté par l'expérimentation que le classificateur SVM et les arbres de décision sont efficaces pour diagnostiquer les maladies cardiovasculaires. Les chercheurs suggèrent aussi que les indicateurs tel que l'âge, le sexe, les douleurs à la poitrine, la tension artérielle, le cholestérol,

la glycémie à jeun, l'ECG au repos, la fréquence cardiaque maximale, etc., puissent être utilisés comme des indicateurs fiables pour prédire la présence d'une maladie cardiaque.

- Les chercheurs Kumar et al. [16], ont utilisé le classificateur Naïve Bayes et les Algorithmes Génétiques pour augmenter les performances du diagnostic des maladies cardiaques.
- Les chercheurs Kumar et al. [17] ont combiné les arbres de décision avec les algorithmes génétiques pour détecter la maladie du cœur. Ils ont obtenu une performance de 84%.
- Les chercheurs Lei et al. [18] ont utilisé le classificateur Naïve Bayes pour diagnostiquer la maladie coronarienne. Une maladie qui touche les artères ayant pour fonction d'alimenter le cœur en sang (artères coronaires). Elle est souvent causée par l'athérosclérose, une accumulation de plaques à l'intérieur de la paroi des artères. Cette accumulation rétrécit peu à peu l'intérieur des artères et ralentit le flot de sang [19].
- Les chercheurs Soni et al. [20] ont réalisé une étude comparative de performances entre différents classificateurs pour diagnostiquer les maladies cardiaques. Les arbres de décision et Naïve Bayes ont surpassé les classificateurs KNN et les réseaux de neurones. L'application des algorithmes génétiques a amélioré les performances des arbres de décision et de Naïve Bayes.
- Les chercheurs Kangwanariyakul et al. [21] ont utilisé des variantes des réseaux de neurones et le classificateur SVM pour diagnostiquer les maladies cardiovasculaires. Les chercheurs ont atteint une performance égale à 78.43% pour les réseaux de neurones contre 74.51% pour le classificateur SVM.

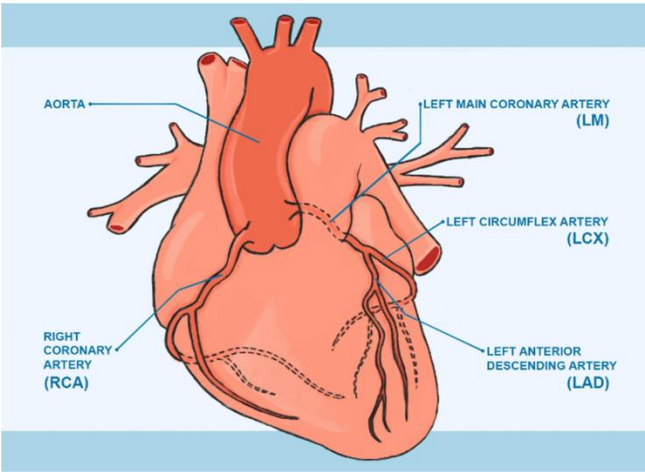


Figure 4 : Schéma du cœur [22]

## 1.3 Data sets de la maladie cardio-vasculaire

### 1.3.1 Data set Z-Alizadeh Sani

#### 1.3.1.1 Historique

Les informations ont été rassemblées de Shaheed Rajaei Centre cardiovasculaire, médical et de recherche, entre automne 2011 et hiver 2012. Le jeu de données inclus 303 patients [23].

#### 1.3.1.2 Définition

Le jeu de données Z-Alizadeh Sani contient les enregistrements de 303 j patients, dont chacun a 54 caractéristiques. Toutes les caractéristiques peuvent être considérées comme des indicateurs de coronaropathie pour un patient, selon la littérature médicale. Cependant, certains d'entre eux n'ont jamais été utilisés dans des approches basées sur l'exploration de données pour le diagnostic CAD. Les caractéristiques sont classées en quatre groupes: données démographiques, symptômes et examen, ECG et

caractéristiques de laboratoire et d'écho. Le tableau 2 présente les caractéristiques de l'ensemble de données Z-Alizadeh Sani ainsi que leurs plages valides, respectivement. Chaque patient peut être dans deux catégories possibles CAD ou Normal. Un patient est classé comme CAD, si le rétrécissement de son diamètre est supérieur ou égal à 50%, et sinon comme normal [24].

| Type de Caractéristique | Nom de Caractéristique | Varier |
|---|---|---|
| Demographic | Age | 30–86 |
| | Weight | 48–120 |
| | Sex | Male, female |
| | BMI (body mass index Kg/m$^2$) | 18–41 |
| | DM (Diabetes Mellitus) | Yes, no |
| | HTN (hyper tension) | Yes, no |
| | Current smoker | Yes, no |
| | Ex-Smoker | Yes, no |
| | FH (family history) | Yes, no |
| | Obesity | Yes if MBI > 25, no otherwise |
| | CRF (chronic renal failure) | Yes, no |
| | CVA (*Cerebrovascular Accident*) | Yes, no |
| | Airway disease | Yes, no |
| | Thyroid Disease | Yes, no |
| | CHF (congestive heart failure) | Yes, no |
| | DLP (*Dyslipidemia*) | Yes, no |
| Symptom and examination | BP (blood pressure: mmHg) | 90–190 |
| | PR (pulse rate) (ppm) | 50–110 |
| | Edema | Yes, no |
| | Weak peripheral pulse | Yes, no |
| | Lung rales | Yes, no |
| | Systolic murmur | Yes, no |
| | Diastolic murmur | Yes, no |
| | Typical Chest Pain | Yes, no |
| | Dyspnea | Yes, no |
| | Function class | 1, 2, 3, 4 |
| | Atypical | Yes, no |
| | Nonanginal CP | Yes, no |
| | Exertional CP (Exertional Chest Pain) | Yes, no |
| | Low Th Ang (low Threshold angina) | Yes, no |
| ECG | Rhythm | Sin, AF |
| | Q Wave | Yes, no |
| | ST Elevation | Yes, no |
| | ST Depression | Yes, no |
| | T inversion | Yes, no |
| | LVH (left ventricular hypertrophy) | Yes, no |
| | Poor R progression (poor R wave progression) | Yes, no |
| Laboratory and echo | FBS (fasting blood sugar) (mg/dl) | 62–400 |
| | Cr (creatine) (mg/dl) | 0.5–2.2 |
| | TG (triglyceride) (mg/dl) | 37–1050 |
| | LDL (low density lipoprotein) (mg/dl) | 18–232 |
| | HDL (high density lipoprotein) (mg/dl) | 15–111 |

| | | |
|---|---|---|
| | BUN (blood urea nitrogen) (mg/dl) | 6–52 |
| | ESR (erythrocyte sedimentation rate) (mm/h) | 1–90 |
| | HB (hemoglobin) (g/dl) | 8.9–17.6 |
| | K (potassium) (mEq/lit) | 3.0–6.6 |
| | Na (sodium) (mEq/lit) | 128–156 |
| | WBC (white blood cell) (cells/ml) | 3700–18,000 |
| | Lymph (Lymphocyte) (%) | 7–60 |
| | Neut (neutrophil) (%) | 32–89 |
| | PLT (platelet) (1000/ml) | 25–742 |
| | EF (ejection fraction) (%) | 15–60 |
| | Region with RWMA (regional wall motion abnormality) | 0, 1, 2, 3, 4 |
| | VHD (valvular heart disease) | Normal, mild, moderate, severe |

Tableau 2 : Caractéristiques de la data set Z-Alizadeh Sani [24].

#### 1.3.1.3 Résultats obtenus dans la littérature

Beaucoup d'approches dans la littérature ont été appliquées et testées sur la base Z-Alizadeh Sani avec des résultats différents. Le tableau 3 résume les résultats obtenus.

| Caractéristiques utilisée | Algotithme utilisé | Précision | Sensibilité | Spécificité |
|---|---|---|---|---|
| All features without three created features | Bagging SMO | 89.43 ± 6.78% | 91.67% | 83.91% |
| | Naïve Bayes | 47.84 ± 6.35% | 28.70% | 95.40% |
| | SMO | 89.76 ± 7.31% | 92.13% | 83.91% |
| | Neural Network | 85.43 ± 7.02% | 90.28% | 73.56% |
| All features and three created features | Bagging SMO | 90.10 ± 6.96% | 91.67% | 86.21% |
| | Naïve Bayes | 63.31 ± 8.01% | 50% | 96.55% |
| | SMO | 90.09 ± 6.49% | 91.67% | 86.21% |
| | Neural Network | 87.11 ± 6.05% | 91.67% | 75.86% |
| Selected features without three created features | Bagging SMO | 92.74 ± 6.43% | 95.37% | 86.21% |
| | Naïve Bayes | 55.37 ± 9.62% | 38.89% | 96.55% |
| | SMO | 93.39 ± 5.14% | 95.37% | 88.51% |
| | Neural Network | 87.13 ± 5.84% | 90.28% | 79.31% |
| Selected features and three created features | Bagging SMO | 93.40 ± 5.53% | 95.83% | 87.36% |
| | Naïve Bayes | 75.51 ± 10.32% | 67.59% | 95.40% |
| | SMO | 94.08 ± 5.48% | 96.30% | 88.51% |
| | Neural Network | 88.11 ± 6.17% | 91.20% | 80.46% |

Tableau 3 : Comparer les performances des algorithmes [24]

### 1.3.2 Data set Cleveland

#### 1.3.2.1 Historique

L'ensemble de données nommé "Cleveland Heart Disease Dataset" provient d'une étude menée en 1988 et provient du référentiel UCI Machine Learning. L'ensemble de données a été divisé en deux: un ensemble d'apprentissage et un ensemble de test. La tâche consiste à obtenir le meilleur prédicateur et à deviner si un patient a une maladie cardiaque [25].

#### 1.3.2.2 Définition

Le jeu de données est collecté par Detrano et al. [26] à partir de 303 échantillons de patients normaux et morts. L'original data set se compose de 76 variables; Cependant, nous considérons 13 variables comme d'autres travaux antérieurs. L'attribut d'étiquette de classe est normalisé en deux classes distinctes, c'est-à-dire oui (la présence de CHD) et non (l'absence de CHD) car dans l'ensemble de données d'origine, cinq valeurs entières allant de 0 (pas de CHD) à 4 (CHD sévère) exister.

Cleveland Heart Data Set est tiré de UCI. Ce jeu de données comprend 303 cas et 76 attributs / caractéristiques. 13 Les caractéristiques sont utilisées sur 76 caractéristiques. Deux essais avec trois algorithmes Bayes Net, la machine de vecteur de support et les arbres fonctionnels FT sont effectués à des fins de détection. Weka Tool est utilisé pour la détection. Après l'essai de maintien de la prise de vue Xperiming, une précision de 88,3% est atteinte à l'aide de la technique SVM. SVM et Bayes Net fournissent à la fois l'exactitude de 83,8%. La précision de 81,5% est atteinte après l'utilisation de FT. Bayes Net a atteint 84,5% de l'exactitude, SVM fournit une précision de 85,1% et une classification ft de 84,5% correctement.

### 1.3.3 Data set Hungarian

#### 1.3.3.1 Historique

L'ensemble de données nommé "Dataset Hungarian" provient d'une étude menée en 1988, L'ensemble de données hongrois a été collecté à l'Institut hongrois de cardiologie, Budapest d'Andras Janosi [26].

#### 1.3.3.2 Définition

L'ensemble de données transformé disponible dans le référentiel UCI. L'ensemble de données 210 comprend 13 caractéristiques d'entrée et un total de 294 observations. De plus, 106 patients sont identifiés comme des personnes atteintes de CDH, tandis que les autres sont en état normal (CDH n'est pas trouvé) [26].

#### 1.3.3.3 Approches de la littérature qui utilisent la base Hungarian

Dans la littérature n'y a pas beaucoup de travaux qui utilisent la base Hungarian pour le diagnostic de la maladie CAD. Le tableau 4 présente quelques approches dans la littérature qui utilise la bas Hungarian pour générer des modèles de diagnostic de la maladie CAD.

| Étude | Année | Base de données | Méthode | Précision (%) |
|---|---|---|---|---|
| Ahamed and ZahidHasan | 2017 | Hungarian | J48 | 72.10 |
| Subramaniyam et al. | 2019 | Hungarian | TGD (Taylorgradient descent)-based ACNN (actor critic neural network) | 82.55 |
| Saqlain et al. | 2019 | Hungarian | Fisher score-based feature selection &Forward feature selection & Reverse feature selection & RBF kernel-based SVM | 76.40 |

Tableau 4 : Résumé des méthodes de classification de CAD automatisées existantes utilisant diverses bases de données publiques [27]

### 1.3.4 Data set Statlog

#### 1.3.4.1 Définition

Le jeu de données Statlog comprend 270 échantillons, dont 120 échantillons ont une maladie cardiaque (présence) et 150 échantillons n'ont pas de maladie cardiaque (absence) sans aucune valeur manquante. Bien que le jeu de données consiste en 75 facteurs de risque de maladies cardiaques (caractéristiques), seules 13 caractéristiques distinctes ont été utilisées dans la littérature, y compris cet article, pour la prédiction de la maladie cardiaque, décrite dans le tableau 5. [28]

| Attribut | Type de données | Description de l'attribut |
|---|---|---|
| age | Real | Age (in years) |
| sex | Binary | 0—female, 1—male |
| cp | Nominal | Chest pain type (1—typical angina, 2—atypical angina, 3—nonangina, 4—asymptomatic) |
| restbps | Real | Resting blood pressure (mm of Hg) |
| chol | Real | Serum cholesterol (mg/dL) |
| fbs | Binary | Fasting blood sugar >120 mg/dL (1—true, 0—false) |
| restecg | Nominal | Resting electrocardiographic results (0—normal, 1—having ST-T wave normality, 2—probable/defined left ventricular hypertrophy) |
| thalach | Real | Maximum recorded heart rate |
| examg | Binary | Angina induced by exercise (1—yes, 0—false) |
| oldpeak | Real | ST depression tempted by workout comparative to rest |
| slope | Nominal | Slant of the peak exercise ST segment (1—upsloping, 2—flat, 3—downsloping) |
| ca | Real | Major vessels colored by fluoroscopy |
| thal | Nominal | 3—normal, 6—fixed defect, 7—reversible defect |
| class | Binary | Represent present or absence of heart disease (1—absence, 2—presence) |

Tableau 5 : Description de l'attribut de la StatLog Dataset de la maladie cardiaque [28]

## 1.4 Machine Learning pour le diagnostic CAD

Les techniques à base de ML ont été appliquées avec succès sur différents types de jeux de données CAD [29-30]. Ces algorithmes ont démontré des performances prometteuses dans la détection et le traitement de la CAD.

La détection de la CAD basée sur ML est un problème d'apprentissage de la machine pure. Bien que la simplicité, l'interprétabilité et la charge informatique sont des

facteurs importants, les médecins et les praticiens sont principalement préoccupés par la fiabilité et la performance globale du modèle dans la détection de la CAD. Plusieurs métriques, y compris la précision, la sensibilité, la spécificité et le score F, ont été rapportées dans la littérature pertinente pour l'évaluation du modèle. Indépendamment de la mesure que la métrique est prise en compte, la performance globale du modèle dépend de deux facteurs clés: (1) Data set et (2) Pipeline ML. Pour le jeu de données, l'analyse peut être basée sur la source des données, la taille de l'échantillon et le nombre de fonctionnalités. Le pipeline peut être étudié sur la base des méthodes de sélection ML et de fonctionnalités.

Les techniques utilisées pour le diagnostique cardio-vasculaire est définie dans le tableau suivant :

| Auteures | Année | Techniques |
|---|---|---|
| Shen et al. | 2019 | 3D fully convolutionalnetwork (FCN) |
| Acharya et al. | 2017 | 11- layer deepConvolutional Neural Network (CNN) |
| Betancur et al. | 2018 | 6-layer deep CNN |
| Betancur et al. | 2018 | Deep CNN |
| Abdolmanafiet al. | 2018 | 5-layer AlexNetarchitecture |
| Yeri et al. | 2018 | 6-layer deep CNN |
| Rubin et al. | 2017 | 6-layer deep CNN |
| Hamersvelt etal. | 2018 | 6-layer deep CNN |
| Sofian et al. | 2018 | 34 layers ResNet101architecture |
| Zreik et al. | 2017 | 8-layer deep CNN |
| Tan et al. | 2018 | 8-layer deep CNN |
| Acharya et al. | 2017 | 11-layer deep CNN |
| Acharya et al. | 2019 | 11-layer deep CNN |
| Allahverdi etal. | 2016 | Deep belief network(DBN) |

Tableau 6 : La liste de l'article publié en utilisant une méthode d'apprentissage en profondeur pour la CAD [22].

### 1.4.1 Défis et inconvénients de l'utilisation des algorithmes ML

Bien que ML techniques aient de nombreux avantages, ce ne sont pas des méthodes parfaites. Les facteurs suivants limitent leurs capacités dans certaines directions [31].

**a)** Selon le théorème sans lunch sans panne [32], différents algorithmes ML conviennent à leur problème particulier. Un algorithme peut bien fonctionner sur un ensemble de données spécifique alors qu'il ne peut pas montrer de bonnes performances sur certains autres. Ainsi, la sélection d'un algorithme approprié pour un ensemble de données spécifique est un grand défi dans la bioinformatique. Par conséquent, la sélection de bonnes fonctionnalités ou des algorithmes de classification est également un défi majeur dans ce domaine.

**b)** Les algorithmes ML ont généralement besoin de jeux de données massives à former. Ces jeux de données doivent être inclusifs et impartiaux de haute qualité. Les jeux de données ont également besoin de temps à collecter.

**c)** Les algorithmes ML ont besoin de temps pour être formés et testés suffisamment pour pouvoir générer des résultats avec une grande confiance. Ces algorithmes ont besoin de nombreuses ressources et équipements.

**d)** Les algorithmes ML font face à la vérification Problème. Il est difficile de prouver que la prédiction faite par eux fonctionne correctement pour tous les scénarios.5. L'interprétation correcte des résultats générés par ML Algorithmes est un autre défi que nous sommes confrontés.

**e)** Un autre inconvénient des algorithmes ML est leur sensibilité élevée d'erreur. S'ils sont formés avec des données biaisées ou incorrectes, elles se retrouvent avec des sorties imprécises. Cela peut conduire à une chaîne d'erreurs qui induisent en erreur les méthodes de traitement. Lorsque ces erreurs sont remarquées, il faut parfois diagnostiquer la source de ces erreurs et même avoir besoin de plus de temps pour les corriger [33].

# Conclusion

Nous avons traité dans ce chapitre la maladie cardio-vasculaires, On estime à 17,7 millions le nombre de décès imputables aux maladies cardio-vasculaires, soit 31% de la mortalité mondiale totale. Et ont ensuite discuté les données de cette maladie et mentionné (Z-alizadeh sani, Cleveland, ...) que notre data set pour nos études et data set Z-Alizadeh sani qui contient 303 patients, et ensuite discuté la manière diagnostiquer de cette maladie et finalement discuté les techniques utilisées dans le diagnostic sur cette maladie.

# *Chapitre 4*

# *Implantation et Réalisation*

# Introduction

L'objectif principal de la mise en œuvre de la demande de données après une série d'étapes dans le processus de développement consiste à développer des modèles de classification des patients pour la maladie cardio-vasculaire Coronary Artery Disease (CAD) à l'aide des différents algorithmes d'apprentissage supervisé fournis par l'outil Weka. Ces modèles nous permettent de prédire de nouveaux patients dans la même discipline en utilisant ses informations médicales.

Dans les chapitres précédents, nous avons fourni des concepts pour concevoir et mettre en œuvre la classification et la prévision du CAD. Ce chapitre donne un aperçu de notre système et de nos outils de développement. Dans ce chapitre, nous introduirons d'abord l'environnement de développement avec les différentes bibliothèques utilisées, en plus de la base d'apprentissage utilisées pour mettre en œuvre notre système de diagnostic du CAD et enfin, nous mettons fin à ce chapitre pour conclure.

## 1.1 Environnement de travaille et outils utilisés

### 1.1.1 Java

**« Selon Wikipédia »** Java est un langage de programmation orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld.

La société Sun a été ensuite rachetée en 2009 par la société Oracle qui détient et maintient désormais Java.

Une particularité de Java est que les logiciels écrits dans ce langage sont compilés vers une représentation binaire intermédiaire qui peut être exécutée dans une machine virtuelle Java (JVM) en faisant abstraction du système d'exploitation.

Éloigné. L'efficacité de Java n'est pas limitée au Web. Cela nous permet également de créer des programmes d'utilisation personnelle et professionnelle. Ces programmes sont exécutés à travers un certain nombre de programmes facilitant une application d'écriture telle que Netbean et Eclipse.

- **NetBeans**

« **Selon Wikipédia** » est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License) et GPLv2. En plus de Java, NetBeans permet la prise en charge native de divers langages tels le C, le C++, le JavaScript, le XML, le Groovy, le PHP et le HTML, ou d'autres (dont Python et Ruby) par l'ajout de greffons. Il offre toutes les facilités d'un IDE moderne (éditeur avec coloration syntaxique, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web). Compilé en Java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X ou sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle

Java). Un environnement Java Development Kit JDK est requis pour les développements en Java.

NetBeans constitue par ailleurs une plateforme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDE NetBeans s'appuie sur cette plate forme.

L'IDE Netbeans s'enrichit à l'aide de greffons.

**Version utilisée** La version utilisée pour le développement de notre application est la version IDE 8.2

### 1.1.2 Weka

Acronyme pour Waikato environment for knowledge analysis, en français : « environnement Waikato pour l'analyse de connaissances » est une suite de logiciels d'apprentissage automatique écrite en Java et développée à l'université de Waikato en Nouvelle-Zélande. Weka est un logiciel libre disponible sous la Licence publique générale GNU (GPL).

L'espace de travail Weka [34] contient une collection d'outils de visualisation et d'algorithmes pour l'analyse des données et la modélisation prédictive, allié à une interface graphique pour un accès facile de ses fonctionnalités. La version « non-Java » originale de Weka était un front-end en Tcl/Tk pour des algorithmes de modélisation (essentiellement tierces) implémentés dans d'autres langages de programmation, complété par un des utilitaires de préprocesseur de données en C, et un système à base de makefile pour lancer les expériences d'apprentissage automatique. Cette version originale était avant tout conçue comme un outil pour analyser des données agricoles [35-36] mais la version plus récente entièrement basée sur Java (Weka 3), pour laquelle le développement a débuté en 1997, est désormais utilisée dans beaucoup de domaines d'application différents, en particulier pour l'éducation et la recherche. Les principaux points forts de Weka sont qu'il :

- est libre et gratuit, distribué selon les termes de la licence publique générale GNU ;

- est portable car il est entièrement implémenté en Java et donc fonctionne sur quasiment toutes les plateformes modernes, et en particulier sur quasiment tous les systèmes d'exploitation actuels ;
- contient une collection complète de préprocesseurs de données et de techniques de modélisation ;
- est facile à utiliser par un novice en raison de l'interface graphique qu'il contient.

L'interface explorer possède plusieurs onglets qui donnent accès aux principaux composants de l'espace de travail. L'onglet préprocesseur a plusieurs fonctionnalités d'import de données depuis des bases de données, un fichier CSV et pour pré-traiter ces données avec une algorithme appelé filtering. Ces filtres peuvent être utilisés pour transformer les données (par exemple, transformer des attributs numériques réels en attributs discrets) et rendre possible l'effacement d'instances et d'attributs selon des critères spécifiques. L'onglet classifier permet à l'utilisateur d'appliquer des classifications et des algorithmes de régression (indifféremment appelés « classifiers » dans Weka) au jeu de données résultant, pour estimer la précision du modèle prédictif, et de visualiser les prédictions erronées, ROC curves, etc. ou le modèle lui-même (si le modèle est sujet à visualisation, comme un Arbre de décision).

## 1.1 Application

### 1.1.1 Structure du notre système de diagnostic du CAD

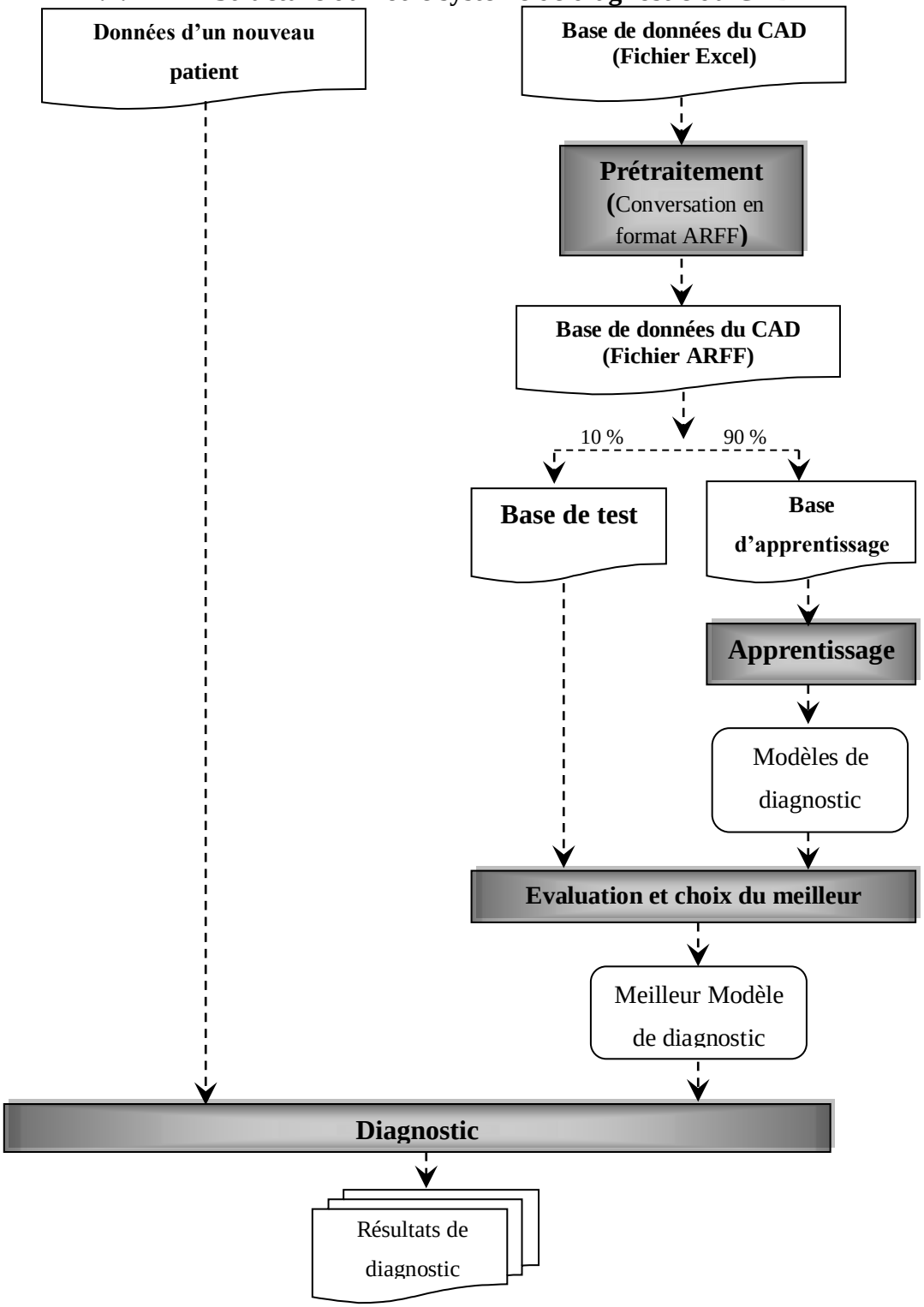


Figure 5 : Processus global de notre système de diagnostic du CAD

- **Prétraitement (Conversation en format ARFF)**

Convertir la base de données du CAD (Format Excel) en base de données du CAD (format ARFF).

- **Base d'apprentissage**

Weka traite des données contenues dans des fichiers respectant le format ARFF Attribute Relation File Format. Il s'agit de fichiers de type texte, décrivant des ensembles de "tuples" caractérisés par un certain nombre d'attributs communs.

- **Evaluation et choix du meilleur modèle**

Évaluation de performances et choit du meilleur modèle qui donne des bonnes performances sur les données de notre étude.

- **Diagnostic**

La prévision des résultats futurs d'un nouveau patient grâce à ses informations médicales

- Obtenir le fichier d

**Structure de la base d'apprentissage**

WEKA traite des données contenues dans des fichiers respectant le format ARFF Attribute-Relation File Format. Il s'agit de fichiers de type texte, décrivant des ensembles de "tuples" caractérisés par un certain nombre d'attributs communs.

**Format d'un fichier ARFF (Attribute-Relation File Format*)***

WEKA utilise (entre autres) le format de fichier arff pour enregistrer les données. Un fichier arff est composé d'une liste d'exemples définis par leurs valeurs d'attributs. Un fichier arff comprend toujours trois types d'informations: un nom pour la base de données, des attributs et des données. La chaîne de caractères @RELATION permet de donner un nom à la base de données. Par exemple, dans le cas du fichier Data.arff, le nom donné est Data. @RELATION Data. *La chaîne de caractères @ATTRIBUTE permet de définir un attribut. Un attribut peut être de 4 types :*

- Réel (NUMERIC ou REAL).
- Nominal (valeurs-possible) **par exemple :** @attribute Sexe FEM,MAL signifie que l'attribut Sexe peut avoir comme valeur soit Sexe-FEM ou soit Sexe-Mal.
- Chaîne de caractère (STRING).

Date (date [<date-format>] @data : suivi d'une instance par ligne. Les valeurs d'instance sont séparées par une virgule.

**Exemple de fichier ARFF**

@relation weather
@attribute outlook {sunny,overcast,rainy}
@attribute temperature numeric @attribute humidity numeric
@ attribute windy{TRUE,FALSE}
@attribute play{yes,no}
% Les données
% commencent ici
**@data**
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes

## But de l'application

Notre application a pour but:

- D'appliquer et comparer certains algorithmes de classification pour construire le modèle de diagnostic du CAD.
- Évaluation de performances et choit du meilleur modèle qui donne des bonnes

performances sur les données de notre étude.

•Utilisation du modèle pour prédire l'état du patient s'il a une maladie CAD ou non via ses données médicales.

### 1.1.2 Fonctionnement du système développé

#### 1.1.2.1 Choix des données d'apprentissage

Dans cette étude, nous avons choisi le jeu de données Z-Alizadeh Sani (fichier arff) qui contient les enregistrements de 303 j patients, dont chacun a 54 caractéristiques. Toutes les caractéristiques peuvent être considérées comme des indicateurs de coronaropathie pour un patient, selon la littérature médicale. Cependant, certains d'entre eux n'ont jamais été utilisés dans des approches basées sur l'exploration de données pour le diagnostic CAD. Les caractéristiques sont classées en quatre groupes:

- Données démographiques,
- Symptômes et examen,
- ECG
- Caractéristiques de laboratoire et d'écho.

Le tableau 2 présente les caractéristiques de la base de données Z-Alizadeh Sani ainsi que leurs plages valides, respectivement. Chaque patient peut être dans deux catégories possibles CAD ou Normal. Un patient est classé comme CAD, si le rétrécissement de son diamètre est supérieur ou égal à 50%, et sinon comme normal.

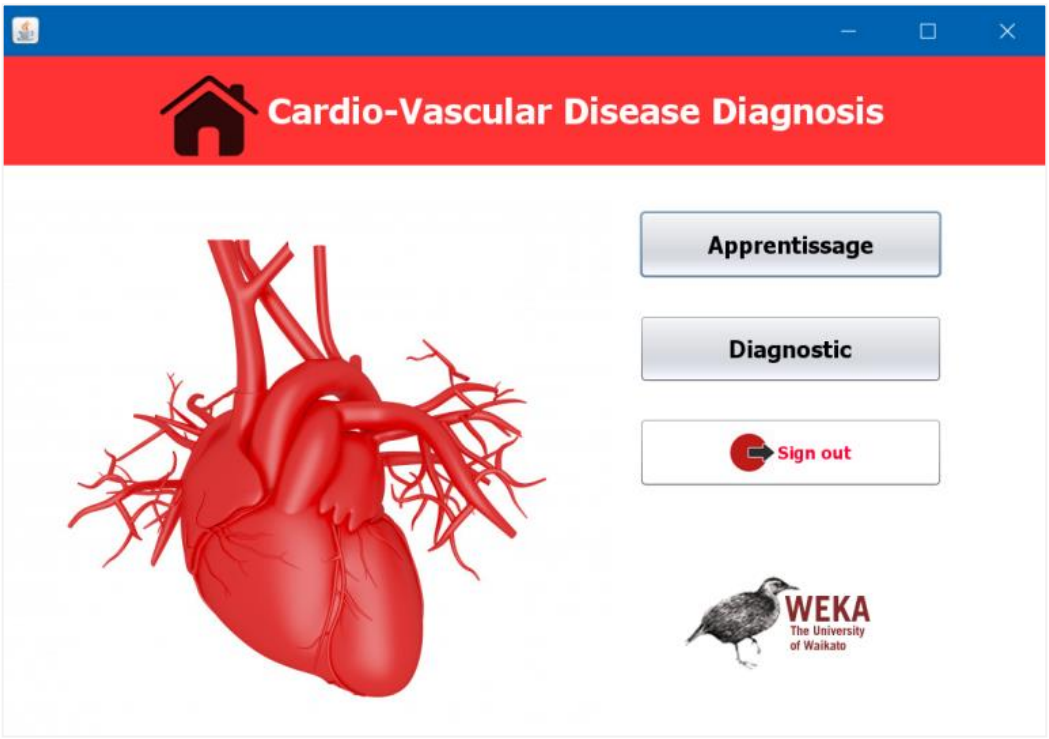


Figure 6 : Fenêtre principale de l'application

Cette fenêtre contient deux boutons principaux :

- Le bouton « **Apprentissage** » permet de sélectionner la base d'apprentissage et de lancer l'apprentissage pour construire le modèle de classification.
- Le bouton « **Diagnostic** » implémente une tâche d'entrée des informations médicales des patients pour découvrez les résultats de diagnostic d'un nouveau patient.

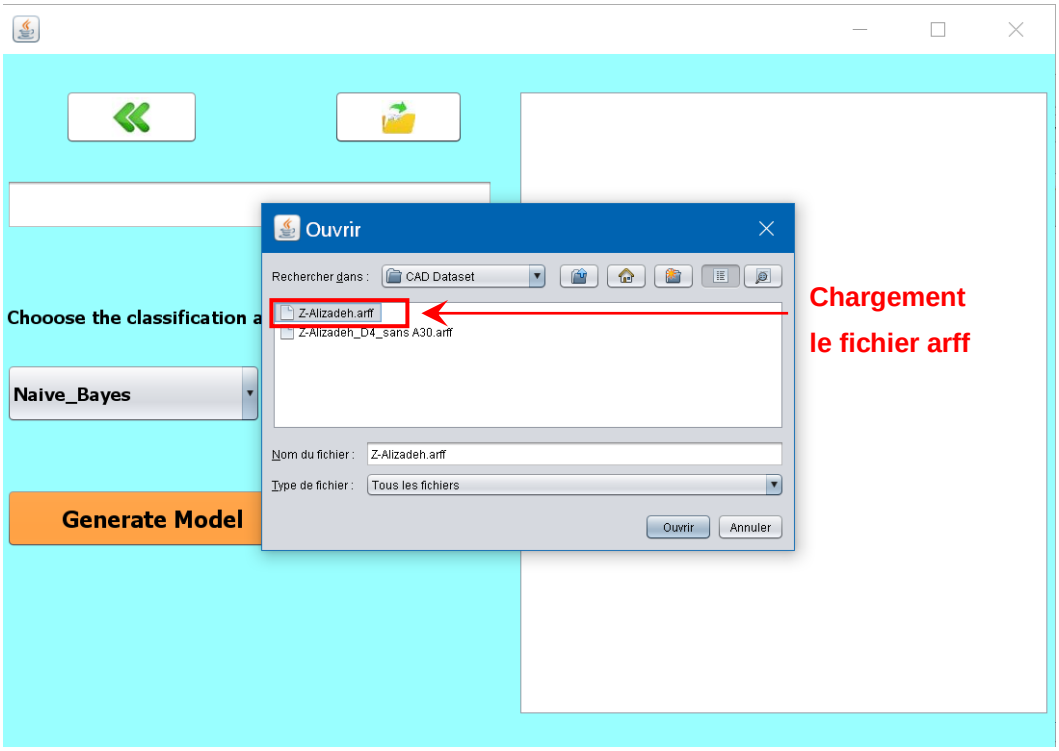


Figure 7 : Interface graphique de chargement le fichier d'apprentissage

jeu de donnée Z-alizadeh (Fichier arff)

### 1.1.2.2 Description des données de la base Z-alizadeh

En savoir plus sur données Pour développer un classifié basé sur des arbres pour le diagnostic de la CAD, l'ensemble de données rapporté par Alizadeh et al. [37], connu sous le nom de Dataset Z-Alizadeh Sani, est employé. La banque de données collectée comprend de la formation de 303 patients. Cette banque de données dispose de 55 patients indépendants et de classifications à une personne dans une classe normale ou de CAD. Le critère de classification d'une personne en tant que patient qui a CAD est son statut de rétrécissement de diamètre. Si le rétrécissement du diamètre est inférieur à 50%, le patient est classiqué comme normal, et autrement, comme la CAD affectée [38].

Les paramètres indépendants comprennent l'âge, le poids, la longueur, le gène, l'indice de masse corporelle (IMC), le diabète sucré (DM), la tension hyper

(HTN), le fumeur actuel, l'ex-fumeur, les antécédents familiaux (FH), OBE- sity, défaillance rénale chronique (CRF), accident cérébrovasculaire (CVA), maladie de la voie aérienne, maladie de la thyroïde, insuffisance cardiaque congestive (CHF), dys- lipidémie (DLP), pression artérielle (BP), impulsion (PR), œdème , faible pouls périphérique (WPP), taux de poumon, murmure systolique, murs diastolique, douleur typique de la chaîne, dyspnée, classe de fonctions, cp atypique, non plan- norme, angine thyroïdienne, bloc de branche de paquet (BBB), Q Wave, ST Elevation, ST Dépression, T inversion, Hypertrophie ventriculaire gauche (LVH), progression des vagues pauvres, glycémie à jeûne (SBF), creale (Cr), triglycérides (Tg), lipoprotéine à faible densité (LDL), en dentité élevée Lipoprotéine (HDL), azote de l'urée sanguine (Bun), taux de séchement de l'érythrocyte (ESR), hémoglobine (HB), potassium (k), CP exsertional, sodium (NA), globule blanc (WBC), lymphocyte, neutrophile, Platelet (PLT), Fonction d'éjection (EF), Abnor-Moteur Régional (région avec RWMA) et maladie cardiaque valvulaire (VHD).

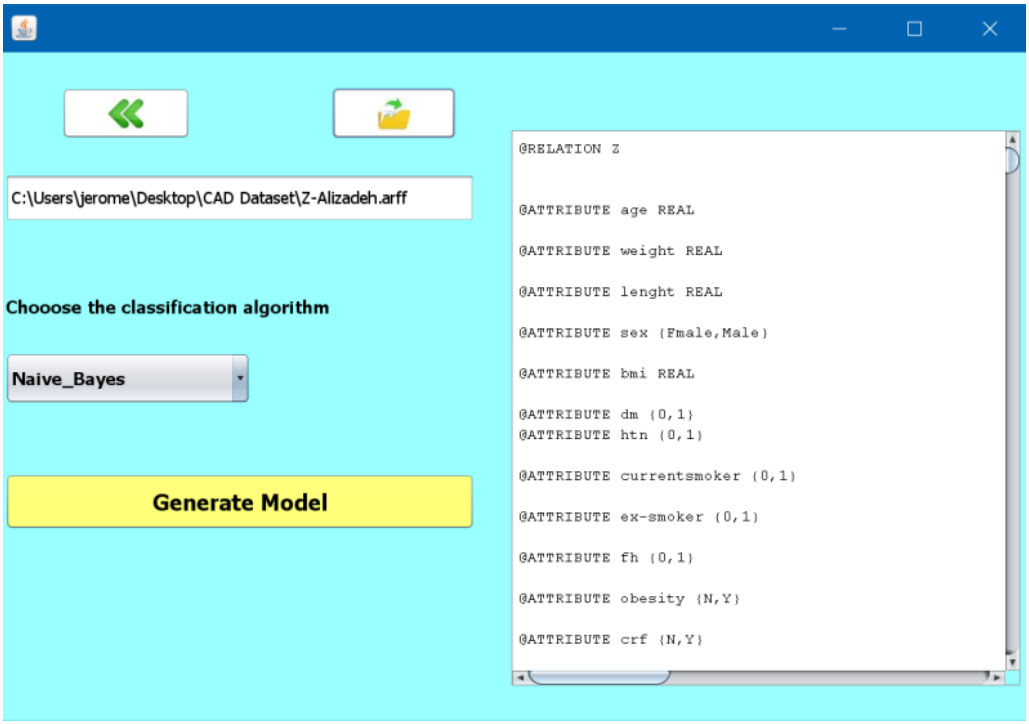


Figure 8 : Interface graphique le contenu jeu de données Z-alizadeh (Fichier arff)

### 1.1.2.3 Choix et utilisation d'algorithmes de classification

Dans cette travail, Après chargement le jeu de données de notre projet Z-alizadeh sani, Nous choisissons l'algorithme de classification pour appliquer sur notre jeu de données. Voir figure 12.

**Apprentissage** : Apprendre et générer un formulaire de classification.

- Application d'algorithmes de classification
- Évaluation des modèles générés

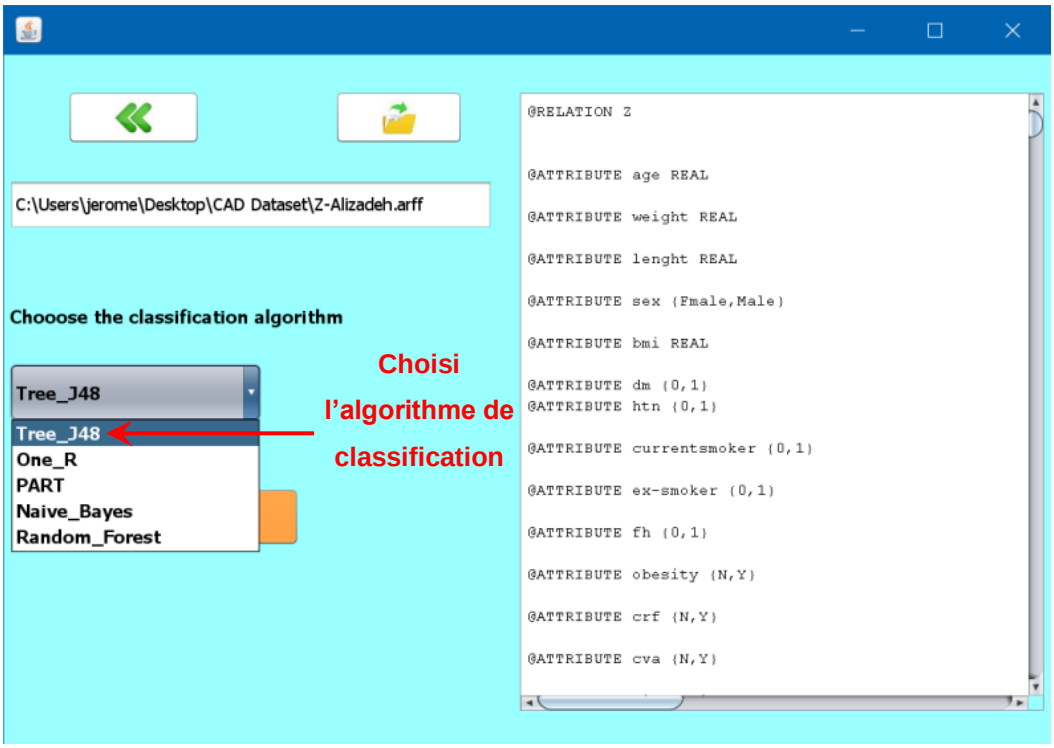


Figure 9 : Interface graphique choix et utilisation d'algorithmes de classification

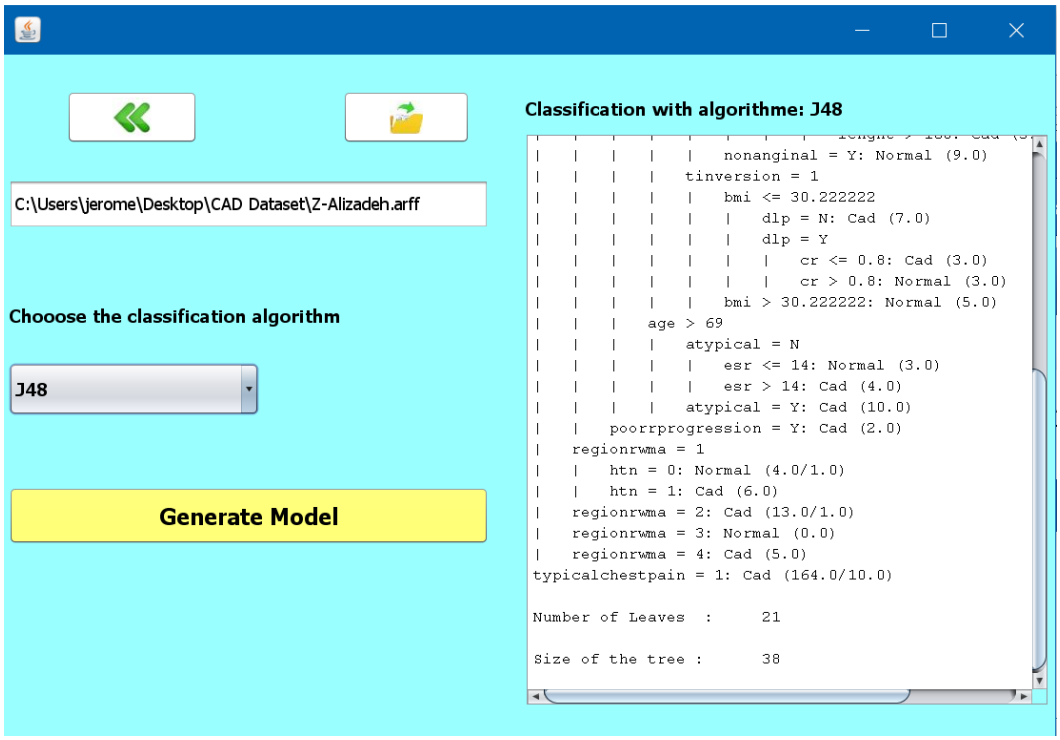


Figure 10 : Exemple de appliqué l'algorithme J48

#### 1.1.2.4 Diagnostic du CAD d'un nouveau patient

**Diagnostic** : Prévision des résultats futurs d'un nouveau patient grâce à ses informations médicales

- Obtenir le fichier de sortie contenant un résultat de diagnostic

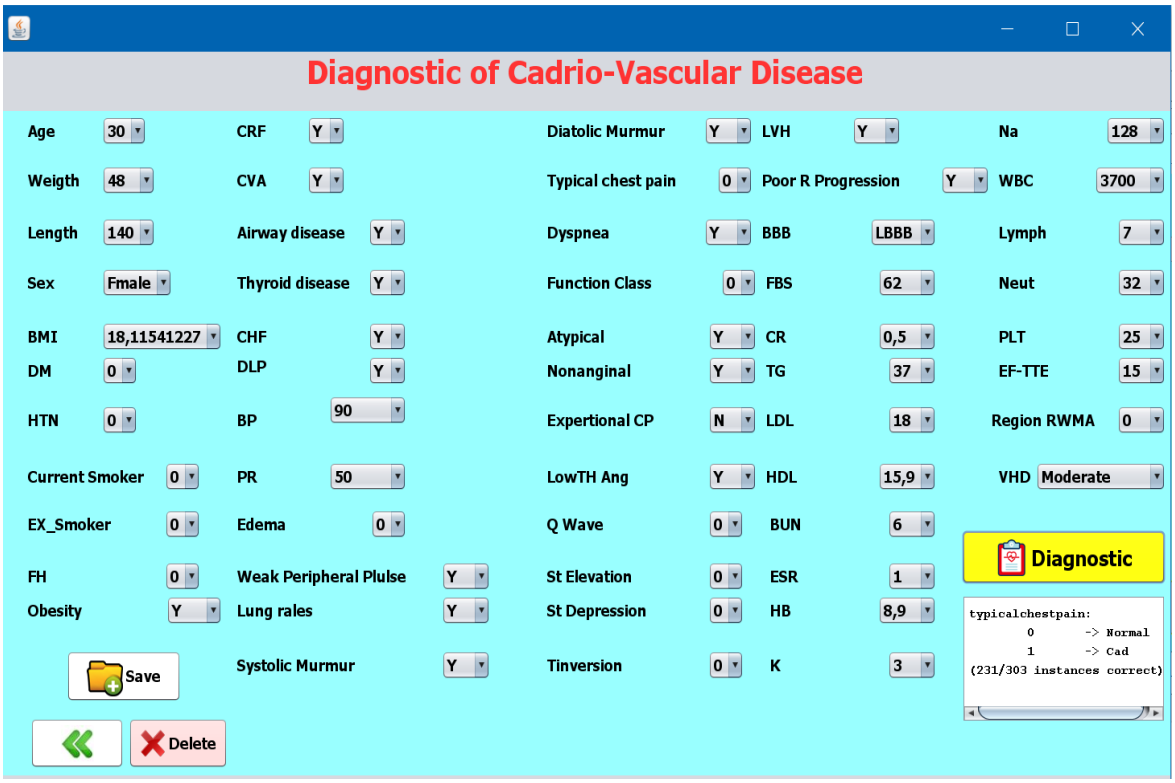


Figure 11 : Interface graphique résultats diagnostic du CAD d'un nouveau patient

Cette interface est une interface pour prédire le nouveau patient après avoir inséré ses informations médicales.

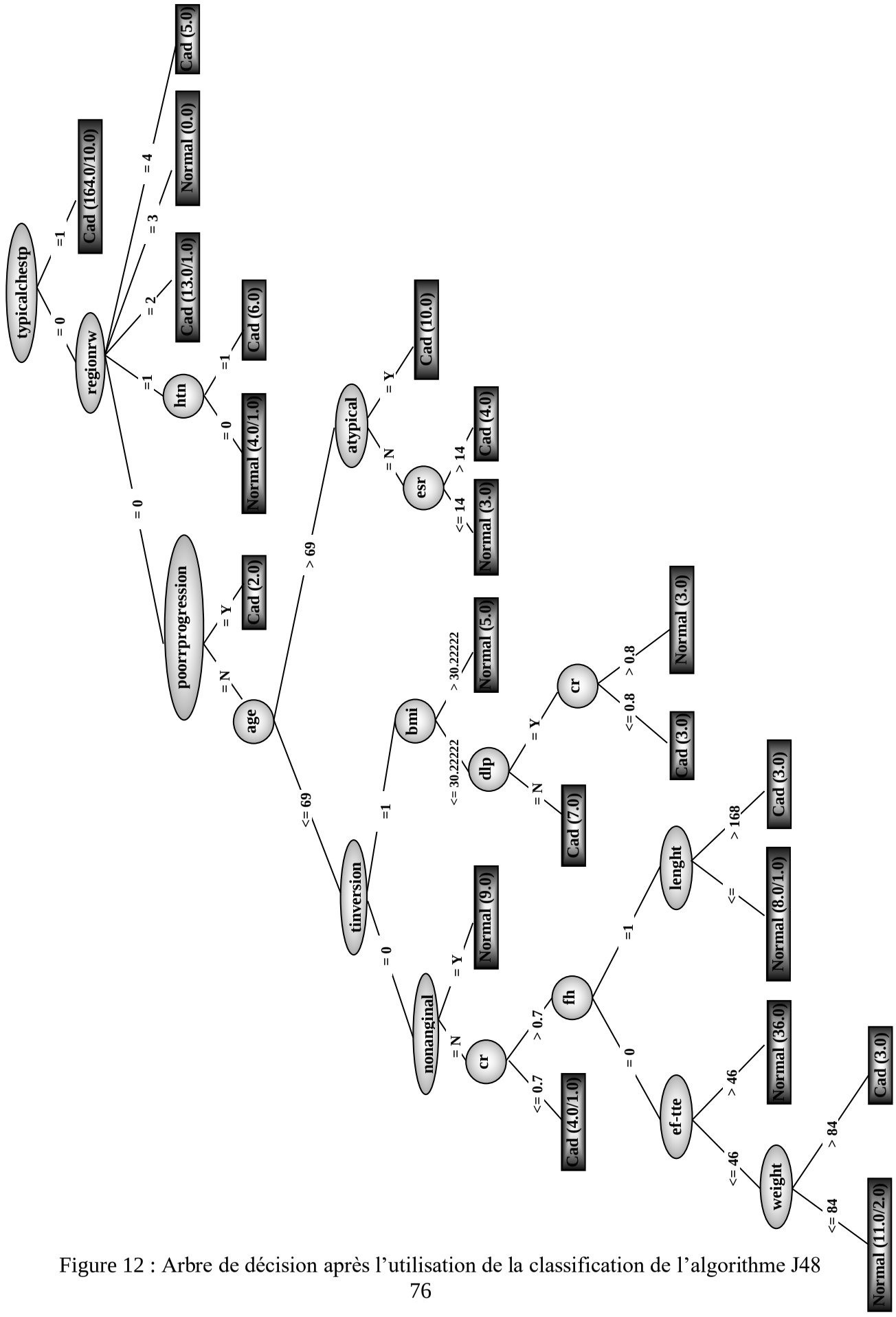


Figure 12 : Arbre de décision après l'utilisation de la classification de l'algorithme J48

| | **Nombre d'instances** | **Pourcentage** |
|---|---|---|
| Instances correctement classées | 245 | 80.8581% |
| Instances incorrectement classées | 58 | 19.1419% |
| **Total** | **303** | **100%** |

Tableau 7 : Résultats de classification par l'algorithme J48

**a) Matrice de confusion**

```
   a   b   <-- classified as
 189  27 |  a = Cad
  31  56 |  b = Normal
```

Tableau 8 : Matrice de confusion de la classification utilisent Algorithme J48

**b) Analyse de l'arbre de décision**

On a remarqué dans l'arbre de la figure 12 qu'il ya dix niveaux et 17 attributs présent dans l'arbre (typicalchestpain, regionrwma, poorrprogression, htn, age, tinversion, atypical, nonanginal, bmi, esr, cr, dlp, fh, cr, ef-tte, length, weight) ces attributs sont considérés comme les attributs les plus important et qui influencent la performance des résultats de diagnostic.

Dans chaque niveau on trouve un ou plusieurs attributs, et les attributs sont présent dans les différents niveaux selon leurs importance où le premier niveau est le plus important suivi par le deuxième niveau, …etc.

Donc les 17 attributs sont triés selon leur importance dans le tableau 9. Selon ce tableau on conclu que «typicalchestpain» est l'attribut le plus important qui influence les résultats des diagnostics des patients.

| Niveau | Identifient de l'attribut | Attributs |
|---|---|---|
| 1 | typicalchestpain | typicalchestpain |
| 2 | regionrwma | regionrwma |
| 3 | poorrprogression, htn | Poorrprogression, htn |
| 4 | age | age |
| 5 | tinversion, atypical | Tinversion, atypical |
| 6 | nonanginal, bmi, esr | Nonanginal, bmi, esr |
| 7 | cr, dlp | Cr, dlp |
| 8 | fh, cr | Fh, cr |
| 9 | ef-tte, length | ef-tte, length |
| 10 | weight | weight |

Tableau 9 : Liste des attributs influencent la performance des diagnostics

## L'algorithme One R :

| | Nombre d'instances | Pourcentage |
|---|---|---|
| Instances correctement classées | 211 | 45.3333 % |
| Instances incorrectement classées | 92 | 30.363 % |
| **Total** | **303** | **100%** |

Tableau 10 : Résultats de classification par l'algorithme One R

### a) Matrice de confusion

```
   a   b   <-- classified as
 160  56 |  a = Cad
  36  51 |  b = Normal
```

Tableau 11 : Matrice de confusion de la classification utilisent l'algorithme One R

## L'algorithme PART

| | Nombre d'instances | Pourcentage |
|---|---|---|
| Instances correctement classées | 248 | 81.8482 % |
| Instances incorrectement classées | 55 | 18.1518 % |
| **Total** | **303** | **100%** |

Tableau 12 : Résultats de classification par l'algorithme PART

**a) Matrice de confusion**

```
   a   b   <-- classified as
 186  30 |  a = Cad
  25  62 |  b = Normal
```

Tableau 13 : Matrice de confusion de la classification utilisent l'algorithme PART

## L'algorithme Naive_Bayes

| | Nombre d'instances | Pourcentage |
|---|---|---|
| Instances correctement classées | 251 | 82.8383 % |
| Instances incorrectement classées | 52 | 17.1617 % |
| **Total** | **303** | **100%** |

Tableau 14 : Résultats de classification par l'algorithme Naive_Bayes

**a) Matrice de confusion**

```
   a   b   <-- classified as
 185  31 |  a = Cad
  21  66 |  b = Normal
```

Tableau 15 : Matrice de confusion de la classification utilisent l'algorithme Naive_Bayes

## L'algorithme Random_Forest :

| | Nombre d'instances | Pourcentage |
|---|---|---|
| Instances correctement classées | 258 | 85.1485 % |
| Instances incorrectement classées | 45 | 14.8515 % |
| **Total** | **303** | **100%** |

Tableau 16 : Résultats de classification par l'algorithme Random_Forest

### a) Matrice de confusion

```
   a   b   <-- classified as
 209   7 |  a = Cad
  38  49 |  b = Normal
```

Tableau 17 : Matrice de confusion de la classification utilisent l'algorithme Random_Forest

## Comparaison de résultats

Nous avons fait une comparaison de résultats obtenues par l'application des cinq algorithmes (J48, One_ R, PART, Naive base, Random_Forest) pour voir quel est l'algorithme qui donne le meilleur modèle prédictif qui nous aide à prédure les résultats d'un nouveau patient, pour la comparaison nous avons choisi les deux critères les plus important qui sont :

- Instances correctement classées
- Instances incorrectement classées

| Critère / Algo | Instances correctement classées | Instances incorrectement classées |
|---|---|---|
| J48 | 80.8581% | 19.1419% |
| One_R | 45.3333 % | 30.363 % |
| PART | 81.8482 % | 18.1518 % |
| Naive_Bayes | 82.8383 % | 17.1617 % |
| Random_Forest | **85.1485 %** | 14.8515 % |

Tableau 18 : Comparaison des résultats de classification de différents algorithmes

Après la comparaison précédente nous constatons que l'algorithme Random_Forest est le meilleur parmi les autres algorithmes où il classifier 85 % des exemples correctement.

Pour la classification des patients a la matrice de confusion été une matrice diagonale ça nous confirme l'efficacité de cet algorithme et pour cela nous avons considérer l'arbre comme un modèle de diagnostic des patients.

# Conclusion

Dans ce chapitre, nous avons défini la base de données utilisée pour l'apprentissage et ses caractéristiques. Nous avons présenter en détaille ensuite le système développé. En fin les résultats d'apprentissage des différents algorithmes utilisés sont présentés avec une étude comparatif de ces algorithmes en termes de taux de classification.

Sur la base des résultats des expériences, nous avons constaté que l'algorithme Random_Forest donne des bons résultats par rapport aux autres algorithmes avec un taux de classification de 85.14%.

# Conclusion générale

## Contribution

L'application développé permette aux médecins de faire une diagnostique de la maladie CAD d'une façon rapide et précise utilisant plusieurs algorithmes d'apprentissage sans connaitre les détaille de l'informatique n'est de l'intelligence artificielle.

## Conclusion

Notre projet de fin d'études fait partie d'un grand projet qui consiste à utiliser les différentes techniques de l'intelligence artificielle pour la fouille de données et en particulier les données médicales. Notre travail consiste à concevoir et produire une application de diagnostique médicale qui permet de prédire la maladie CAD pour une personne utilisant seulement ses données médicale. Pour cela, nous avons présenté des concepts liés à l'extraction de connaissance à partir de données médicales.

Nous avons développé une application qui permette de générer des modèles de classification en utilisant les différentes techniques de classification offertes par l'outil Weka. Les modèles générés sont évalués et comparé afin de choisir le meilleur modèle et l'utilisé pour la diagnostique.

Selon les résultats d'évaluation, l'algorithme Random_Forest donne des bons résultats par rapport aux autres algorithmes avec un taux de classification de 85.14%.

## Travaux futurs

Pour les études futures, il est donc suggéré de collecter davantage de données de diverses ressources pour le diagnostic / classifications de CAD. En outre, la qualité

des jeux de données pour le diagnostic de la CAD peut être améliorée en tenant compte de plus de paramètres plus indépendants.

L'utilisation des nouvelles approches de classification comme les techniques de deep learning et la classification associative peuvent donner des bons résultats avec la base de données Z-Alizadeh Sani ou d'autres bases de données de la maladie CAD dans la littérature.

# Bibliographie


[1] https://www.talend.com/fr/resources/what-is-data-mining/

[2] Dr. Abdelhamid DJEFFAL Cours Fouille de données avancée Site web : www.abdelhamid-djeffal.net Année 2014/2015.

[3] Christophe GOETZ Thèse Doctorat En Médecine Apports d'une méthode de fouille de données pour la détection des cas incidents de cancer du sein dans les données du Programme de Médicalisation des Systèmes d'Information : Une analyse formelle des concepts sur les données 2001 du PMSI et du registre du cancer de l'Isère le 19 mai 2011

[4] BOUMAAZA Laid, ARAAR Mohamed Amine Mémoire de fin d'études Master en informatique thème Modèles de classification pour la prédiction des résultats des étudiants de MI en première année universitaire année :2019/2020.

[5] Un guide rapide sur l'exploration de données https://www.astera.com/fr/type/blog/a-quick-guide-to-data-mining/

[6] Mémoire de Magister en informatique Une plate forme orientée agent pour le data mining Melle. CHAMI Djazia Année 2009/2010.

[7] BOUDHEB Tarik THESE DE DOCTORAT en informatique Privacy priserving of biomedical data 2018/2019.

[8] Gordon S LINOFF et Michael JA BERRY. *Data mining techniques : for marketing, sales, and customer relationship management*. John Wiley, 2011.

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

[10] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. In proceedings of the International Journal of Computer Science and Applications, 6(2), 256-261.

[11] Santos, F. (2015). Arbres de décision.

[12] Taleb Zouggar, S. (2014) Contribution à l'apprentissage automatique par automate d'arbre

et mesure de sélection. Thèse de doctorat. Unviversité d'Oran. Extrait de : https://theses.univ-oran1.dz/document/15201425t.pdf

[13] Mulak, P., & Talhar, N. (2015). Analysis of distance measures using k-nearest neighbor algorithm on kdd dataset. International Journal of Science and Research, 4(7), 2101-2104.

[14] Djeffal, A. (2012). Utilisation des méthodes Support Vector Machine (SVM) dans l'analyse des bases de données. Thèse de doctorat, Université Mohamed Khider-Biskra.

[15] Les maladies cardiovasculaires https://www.who.int/fr/news-room/fact-heets/detail/cardiovascular-diseases-(cvds)

[16] Kumar, S., & Sahoo, G. (2015). Classification of heart disease using naive bayes and genetic algorithm. In Computational Intelligence in Data Mining-Volume 2 (pp. 269-282). Springer, New Delhi.

[17] Kumar, N., & Khatri, S. (2017). Optimizing Decision Tree Through Attributes Generation Using Genetic Programming for Clinical Data. Indian Journal of Science and Technology, 10(22).

[18] Lei, K., Zhang, L., Shen, Y., Huang, X., & Wu, J. (2017, March). Syndromes diagnostic model for coronary artery disease (CAD): An improved naïve Bayesian classification model based on attribute relevancy. In Big Data Analysis (ICBDA), 2017 IEEE 2nd International Conference on (pp. 897-902). IEEE.

[19] Ottawa. (2019a). Maladie coronarienne (Athérosclérose). Institut de cardiologie de l'université d'ottawa. Extrait de https://www.ottawaheart.ca/fr/maladie-du-c%C5%93ur/maladie-coronarienne-ath%C3%A9roscl%C3%A9rose.

[20] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011b). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[21] Kangwanariyakul, Y., Nantasenamat, C., Tantimongcolwat, T., & Naenna, T. (2010). Data mining of magnetocardiograms for prediction of ischemic heart disease. EXCLI journal, 9, 82.

[22] Roohallah Alizadehsani, Moloud Abdar, Parham Kebria, Saeid Nahavandi Article in Computers in Biology and Medicine · July 2019 Machine learning-based coronary artery disease diagnosis: A comprehensive review https://www.researchgate.net/publication/334227684

[23] R. Alizadehsani, J. Habibi, Mohammad Javad Hosseini, Reihane Boghrati, Asma Ghandeharioun, Behdad Bahadorian, Z. SaniArticle July 2012, Diagnosis of Coronary Artery Disease Using Data Mining Techniques Based on Symptoms and ECG Features https://www.researchgate.net/publication/265158683

[24] R.O. Bonow, D.L. Mann, D.P. Zipes, P. Libby, Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine, 9th ed., Saunders, New York, 2012. Article 2013 : A data mining approach for diagnosis of coronary artery disease www.intl.elsevierhealth.com/journals/cmpb

[25] https://www.kaggle.com/c/heart-disease-uci/overview/description

[26] Bayu Adhi Tama, Sun Im, Seungchul Lee, Article ID 9816142 / 2020, Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble https://doi.org/10.1155/2020/9816142

[27] Elham Nasariana, Moloud Abdar, Mohammad Amin Fahami, Roohallah Alizadehsani, Sadiq Hussaind, Mohammad Ehsan Basiri, Mariam Zomorodi-Moghadamf, Xujuan Zhoug, Paweł Pławiak, U. Rajendra Acharyaj, Ru-San Tan, Nizal Sarrafzadegan Article 2020, Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach https://www.researchgate.net/publication/339195364.

[28] Khalid Raza, January 2019 Article *in* Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule https://www.researchgate.net/publication/330049116

[29] R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z.A. Sani, A data mining approach for diagnosis of coronary artery disease, Computer Methods and Programs in Biomedicine, 111 (2013) 52-61.

[30] D. Pal, K.M. Mandana, S. Pal, D. Sarkar, C. Chakraborty, Fuzzy expert system approach for coronary artery disease screening using clinical parameters, Knowledge-Based Systems, 36 (2012) 162-174.

[31] T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom, A. Waibel, Machine learning, Annual reviewof computer science, 4 (1990) 417-433

[32] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE transactions on evolutionarycomputation, 1 (1997) 67-82.

[33] J.V. Tu, Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, Journal of Clinical Epidemiology, 49 (1996) 1225-1231.

[34] ( en) Ian H. Witten, Eibe Frank, et Mark A. Hall, Data Mining: Practical machine learning tools and techniques,3e édition, Morgan Kaufmann, 2011(ISBN 978-0-1237-4856-0), 629 pages.

[35] (en) G. Holmes, A. Donkin and I.H. Witten, « *Weka: A machine learning workbench* » *Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia*, 1994

[36] (en) S.R. Garner, S.J. Cunningham, G. Holmes, C.G. Nevill-Manning, and I.H. Witten, « *Applying a machine learning workbench: Experience with agricultural databases* » *Proc Machine Learning in Practice Workshop, Machine Learning Conference, Tahoe City, CA, USA*, 1995 (consulté le 25 juin 2007), p. 14–21.

[37] R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, et al., A data mining approach for diagnosis of coronary artery dis- ease, Comput. Methods Programs Biomed. 111 (1) (2013) 52–61.

[38] D.L. Mann, D.P. Zipes, P. Libby, R.O Bonow, Braunwald's Heart Disease E-Book: A Textbook of Cardiovascular Medicine, Elsevier Health Sciences, 2014.