REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 20 Août 1955 - Skikda

Faculté des Sciences - Département d'Informatique

Mémoire de fin d'études pour l'obtention du diplôme de

Master en informatique.

Option : Génie Logiciel Avancé et Applications (G.L.A.A)

Thème

Classification par arbre de décision pour la recherche de type de maladie d'un patient.

Réalisé par :

- KHEZZOUZ Kenza
- SALAH AIECH Dalel

Encadré par : A. MANSOUL

Année Universitaire 2021-2022

Remerciement

En préambule à ce mémoire nous remerciant ALLAH qui nous a aidé et nous a donné la patience et le courage durant ces longues années d'étude . Nous tenons à remercier notre encadreur A.Mansoul , pour l'orientation , la confiance , la patience qui a constitué un apport considérable sans lequel ce travail n'aurait pas pu être mené au bon port . Qu'elle trouve dans ce travail un hommage vivant à sa haute personnalité

Nous tenons à exprimer nos sincères remerciements à tous les professeurs qui nous ont enseigné et qui par leurs compétences nous ont soutenu dans la poursuite de nos études.

Enfin , on remercie tous ceux qui , de près ou de loin , ont contribué à la réalisation de ce travail .

Dédicace

Je dédie ce modeste travail

A ma très chère mère , affable , honorable , aimable , tu représentes pour moi la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et prier pour moi , ta prière et ta bénédiction m'ont été d'un grand secours pour mener à bien mes études .

A mon très cher père , rien au monde ne vaut les efforts fourmis jour et nuit pour mon éducation et mon bien être , ce travail est le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation .

Je vous dédie ce travail en témoignage de mon profond amour . Puisse dieu le tout puissant , vous préserver et vous accorder santé , longue vie et bonheur .

A mes jolies sœurs imen , Khaoula et Aicha ;

A mes chers frères Saad Eddine et Abde Allatif ;

A tous les membres de la famille : tentes , oncles , cousins maternelles et paternelles .

A mes cheres amies et camarades de travaille Dalel et Sara Je ne peux trouver les mots justes et sincères pou r vous exprimer mon affection et mes pensées , vous êtes pour moi des sœurs et des amies sur qui je peux compter . En témoignage de l'amitié qui nous uni et des souvenirs de tous les moments que nous avons passés ensemble , je vous dédie ce travail et je vous souhaite une vie pleine de santé et de bonheur .

Kenza

Dédicace

Je dédie ce travail :

À mes très chers parents, que Dieu tout puissant les protège
À mes frères et mes sœurs
À mon chère marie Ahmed et sa famille
À mes chers enfants :Tasnim ,Rahma,Abd allah et Mokim
A ma camarade dans ce travail Kenza
À tous mes enseignants

À mes amis ainsi que tous les gens qui m'ont aidé de près ou de loin à accomplir ce travail

Dalel

Résumé

Le data mining est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse des données informatique, c'est une étape très essentiel dans le processus d'extraction des connaissances.

Notre travail consiste à concevoir un système qui utilise la méthode classification de données dont le but est la recherche de type de maladies d'un patient à l'aide des arbres de décision.

D'abord nous appliquons la classification par arbre de décision pour classer les données en classes, afin que chaque classe contienne les données qui ont la même description, nous aurons crée le model. Pour l'atteindre nous utilisons la méthode J48 sous un environnement appelé WEKA destiné à la fouille de données.

Après nous utilisons un module que nous avons développé à partir du modèle construit par classification.

A la suite, nous expérimentons notre approche sur des données se rapportant aux maladies urinaires.

Enfin, nous évaluons notre approche.

Mots clés :

Extraction de connaissances, Fouille de données, Classification, maladies urinaires, arbres de décision, algorithme J48.

Table des matières

Table des figures

Liste des tables

Introduction générale

Depuis quelques années, une masse grandissante de données est générée de toute part et dans différents domaines. Les techniques usuelles analysant ces données sont insuffisantes d'où le besoin d'une nouvelle génération d'outils et de théories pour aider `a extraire les informations utiles (les connaissances) à partir des volumes de données numériques qui croissent rapidement. Ces théories et outils sont le sujet d'un nouveau domaine appel 'extraction de connaissances à partir de données dont le cœur est la fouille de données.

Le Data Mining est une technologie dont le but est la valorisation de l'information et l'extraction de connaissances d'un grand nombre de données. Cette technologie peut être utilisée dans tous les domaines, où il y a un besoin d'analyser une grande collection de données ou de prévoir l'évolution d'un processus ou d'un système donné. Parmi les domaines d'application de la fouille de données, on peut citer le commerce électronique, le marketing, la médecine, la biologie, la sécurité d'information, l'éducation et la télécommunication.

Dans notre projet on a choisi la spécialité des maladies urinaires dans le secteur médical qui consiste à la diversité des techniques hautement développés et difficiles a contrôler sans informatisation des données ,on va présenter tous le processus d'un ECD : à partir de l'ensemble des donnés (des donnés sur les symptômes des maladies) obtenu d' une base de donné et analysés par un environnement d' apprentissage WEKA, ce dernier permet de nous donner le modèle souhaité, jusqu'à l'étape d' extraire des connaissances (classification des maladies) avec une interface graphique développée en Java NetBeans.

Dans le premier chapitre intitulé « ECD », nous évoquerons les principales tâches et techniques utilisées en data mining, dont la classification qui a pour méthode les arbres de décision. Certains algorithmes utilisés dans la découverte de connaissance y sont aussi mentionnés.

Dans le deuxième chapitre, nous mettons en évidence la tâche de classification en général, en retraçant ses différentes étapes, ainsi que ses avantages et inconvénients et les critères qu'il faut pour une bonne classification. Puis les méthodes de classification .

Le troisième chapitre détaillera les méthodes que nous avons utilisé au cours de notre travail, à savoir, les méthodes de classification, en spécifiant leur Algorithme et leur principe de fonctionnement.

Le quatrième et dernier chapitre sera dédié à la présentation de notre application, en retraçant les différentes démarches utilisées, ainsi que les principaux outils qui ont servi à sa.

Nous finissons notre mémoire par une conclusion générale qui contient les perspectives possibles qui constituent une suite de recherche pour notre projet.

1.1.Introduction

Dans cette partie, nous allons donner un aperçu général sur le processus (ECD), la définition, les étapes (le nettoyage et l'intégration, le prétraitement des données, la fouille de données et l'évaluation et la présentation), notamment sur les taches principales e fouille de données (la description, le groupement, la classification, les règles d'association, la prédiction et la visualisation). Enfin nous terminons par la conclusion.

1.2.L'extraction des connaissances à partir des données

L'extraction des connaissances à partir des données (ECD) est un processus itératif qui met en œuvre un ensemble de techniques provenant des bases de données, de la statistique, de l'intelligence artificielle, de l'analyse des données, des interfaces de communication homme-machine. L'ECD vise à transformer des données (volumineuses, multiformes, stockées sous différents formats sur des supports pouvant être distribués) en connaissances.

Ces connaissances peuvent s'exprimer sous forme d'un concept général qui enrichit le champ sémantique de l'usager par rapport à une question qui le préoccupe. Elles peuvent prendre la forme d'un rapport ou d'un graphique. Elles peuvent s'exprimer comme un modèle mathématique ou logique pour la prise de décision. Les modèles explicites, quelle que soit leur forme, peuvent alimenter un système à base de connaissances ou un système expert.

La définition que nous venons de donner nous semble plus générale et mieux adaptée à l'usage du data mining moderne que celle proposée initialement par **Fayyad** en 1996

« L'extraction de connaissances à partir des données est un processus non trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données».

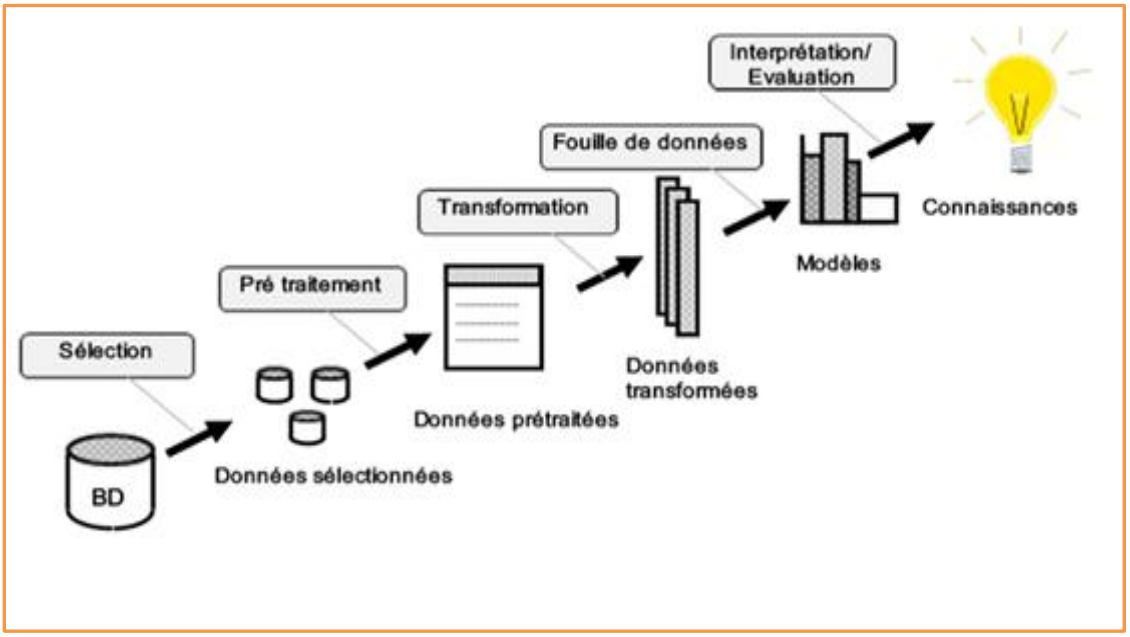

Figure 1.1 : Processus d'extraction des connaissances à partir de données. [1]

1.3.La fouille de données

" Le data mining, ou fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de bases de données informatiques (souvent grandes), de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données" .[2]

La fouille de données est un domaine qui est apparu avec l'explosion des quantités d'informations stockées, avec le progrès important des vitesses de traitement et des supports de stockage. La fouille de données vise à découvrir, dans les grandes quantités de données, les informations précieuses qui peuvent aider à comprendre les données ou à prédire le comportement des données futures. Le datamining utilise depuis sont apparition plusieurs outils de statistiques et d'intelligence artificielle pour atteindre ses objectifs.

La fouille de données s'intègre dans le processus d'extraction des connaissances à partir des données ECD ou (KDD : Knowledge Discovery from Data en anglais). Ce domaine en pleine expansion est souvent appelé le data mining.

La fouille de données est souvent définie comme étant le processus de découverte des nouvelles connaissances en examinant de larges quantités de données (stockées dans des entrepôts) en utilisant les technologies de reconnaissance de formes de même que les

techniques statistiques et mathématiques. Ces connaissances, qu'on ignore au début, peuvent être des corrélations, des patterns ou des tendances générales de ces données. La science et l'ingénierie modernes sont basées sur l'idée d'analyser les problèmes pour comprendre leurs principes et leur développer les modèles mathématiques adéquats. Les données expérimentales sont utilisées par la suite pour vérifier la correction du système ou l'estimation de quelques paramètres difficiles à la modélisation mathématiques. Cependant, dans la majorité des cas, les systèmes n'ont pas de principes compris ou qui sont trop complexes pour la modélisation mathématique. Avec le développent des ordinateurs, on a pu rassembler une très grande quantité de données à propos de ces systèmes. La fouille de données vise à exploiter ces données pour extraire des modèles en estimant les relations entre les variables (entrées et sorties) de ses systèmes. En effet, chaque jour nos banques, nos hôpitaux, nos institutions scientifiques, nos magasins, ... produisent et enregistrent des milliards et des milliards de données. La fouille de données représente tout le processus utilisant les techniques informatiques (y compris les plus récentes) pour extraire les connaissances utiles dans ces données. Actuellement, La fouille de données utilise divers outils manuels et automatiques. [3]

1.4.Les étapes du processus d'extraction des connaissance apartir des données

Le processus d'extraction de connaissance à partir de donnés comporte quatre étapes principales :

- ✓ Le nettoyage et l'intégration
- ✓ Le prétraitement des données
- ✓ La fouille de données
- ✓ L'évaluation et la présentation

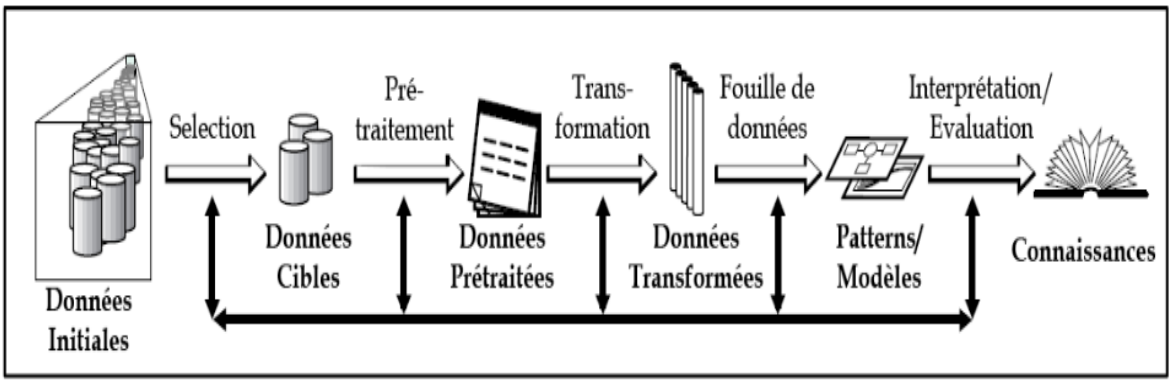

Figure 1.2 : Etapes du processus d'ECD . [5]

Mais avant de présenter ces quatre étapes nous allons d'abord voir qu'est-ce qu'une donnée, une donnée bruitée, une donnée incomplète et une données incohérente.

- **Donnée :** une donnée en data mining est un enregistrement au sens des bases de données, c'est-à-dire un ensemble de lignes caractérisées par un ensemble d'attributs.
- **Donnée bruitée :** les données contiennent des enregistrements erronés ou des aberrations.

Exemples :

- Distorsion de la voix transmise par téléphone dû à un problème matériel ou un problème de réseau
- Un salaire inférieur à zéro

- **Donnée incomplète :** Une donnée incomplète est une donnée ayant des valeurs ou des attributs manquants

Exemple :

- Ce tableau illustre un exemple d'enregistrement où certains attributs comportent des valeurs manquantes

Age	Statut	Genre	Salaire
35	Marié	F	735
40	Marié		M
20		546	
7	Célibataire	F	100

Table 1.1 : Exemple d'une donnée incomplète.

1.4.1 Le nettoyage et l'intégration

1.4.1.1. Le nettoyage des données consiste à retravailler ces données bruitées, soit en les supprimant, soit en les modifiant de manière à tirer le meilleur profit.

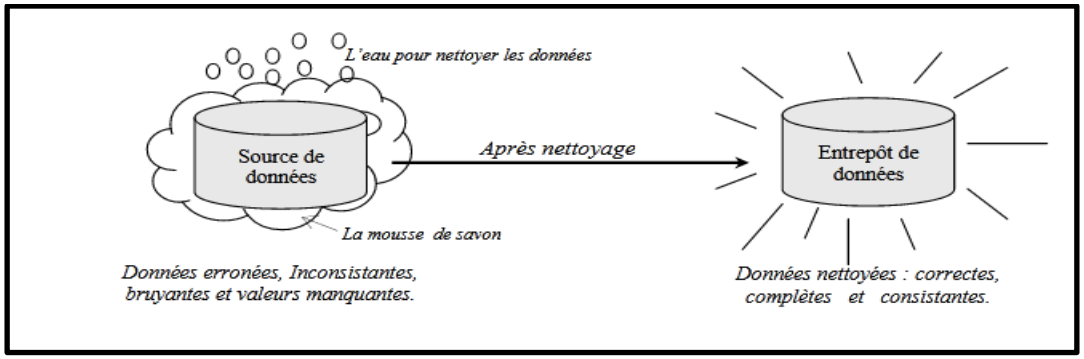

Figure 1.3 : Routine de nettoyage de données.[6]

1.4.1.2. L'intégration des données est la combinaison des données provenant de plusieurs sources (bases de données, sources externes, etc.). Le but de ces deux opérations est de générer des entrepôts de données et/ou des magasins de données spécialisés contenant les données retravaillées pour faciliter leur exploitation future. [4]

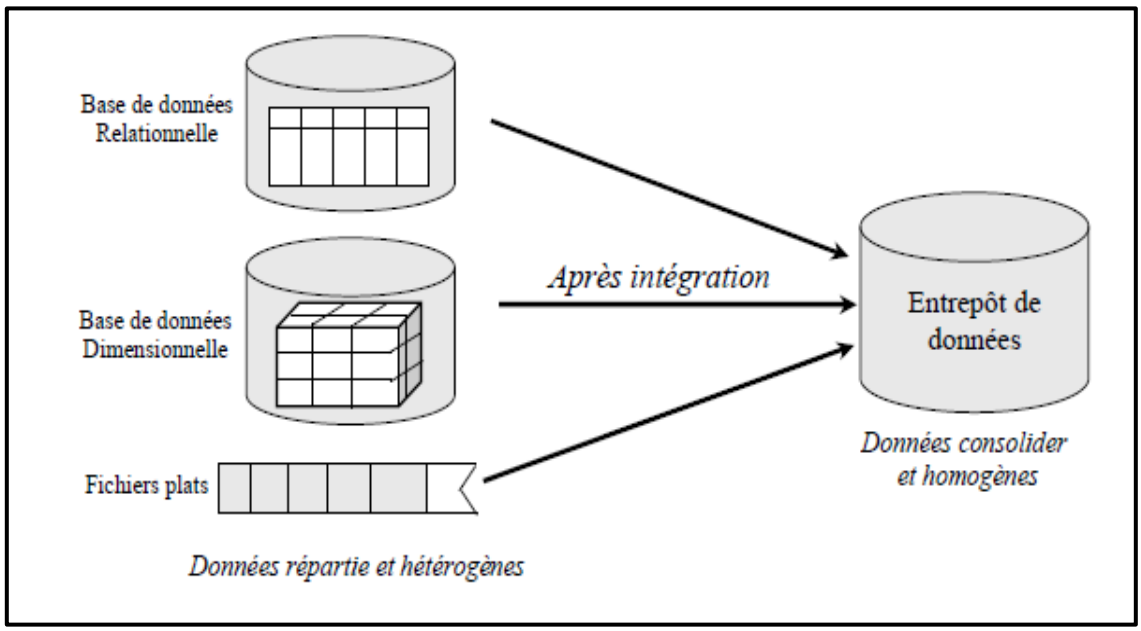

Figure 1.4: Phase d'intégration des données.[6]

1.4.2 Le prétraitement des données

Il peut arriver parfois que les bases de données contiennent à ce niveau un certain nombre de données incomplètes et/ou bruitées. Ces données erronées, manquantes ou inconsistantes doivent être retravaillées si cela n'a pas été fait précédemment. Dans le cas contraire, durant l'étape précédente, les données sont stockées dans un entrepôt. Cette étape permet de sélectionner et transformer des données de manière à les rendre exploitables par un outil de fouille de données.

Cette seconde étape du processus d'ECD permet d'affiner les données. Si l'entrepôt de données est bien construit, le prétraitement de données peut permettre d'améliorer les résultats lors de l'interrogation dans la phase de fouille de données. [4]

1.4.3 Le Data Mining

L'étape de Data Mining constitue le cœur du processus d'Extraction de Connaissances à partir de Données. Elle s'agit d'appliquer des méthodes intelligentes, spécifiques au Data

Mining, afin d'extraire à partir de données, dite données d'apprentissage, des modèles, des règles ou toutes autres formes compréhensibles et interprétables en connaissances utiles (voir la figure 1. 5).

Parmi les méthodes les plus connues du Data Mining, nous citons les méthodes de classification, les méthodes de Clustering et les méthodes de recherche de règles d'association.

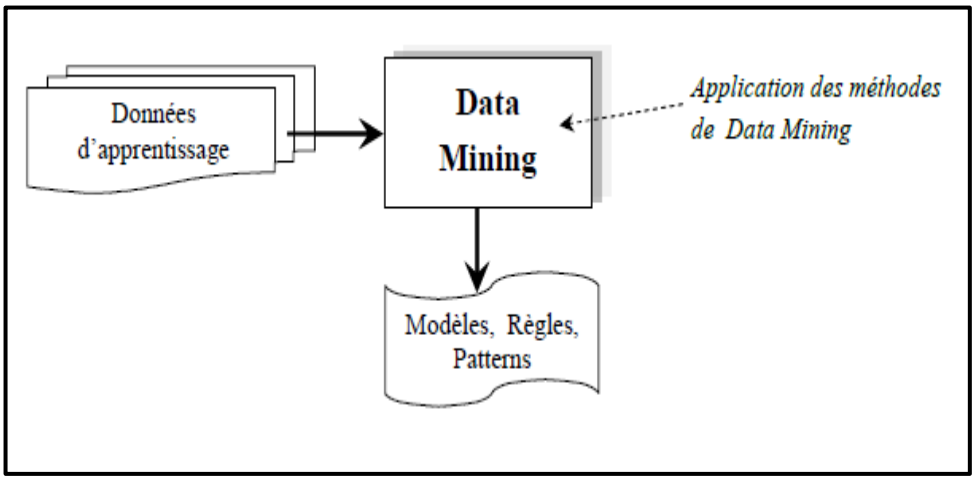

Figure 1. 5:Etape datamining du processus ECD.[6]

1.4.4 L'évaluation et la présentation

Cette phase est constituée de l'évaluation, qui mesure l'intérêt des motifs extraits, et de la présentation des résultats à l'utilisateur grâce à différentes techniques de visualisation. Cette étape est dépendante de la tâche de fouille de données employée.

En effet, bien que l'interaction avec l'expert soit importante quelle que soit cette tâche, les techniques ne sont pas les mêmes. Ce n'est qu'à partir de la phase de présentation que l'on peut employer le terme de connaissance à condition que ces motifs soient validés par les experts du domaine. [4]

1.5.Principales taches de fouille de données

Pour pouvoir extraire les données pertinentes à une entreprise parmi leur abondance, différentes méthodes sont mises en œuvre. Ces dernières se basent sur l'identification de liens logiques entre différents motifs et tendances, afin d'établir des statistiques.

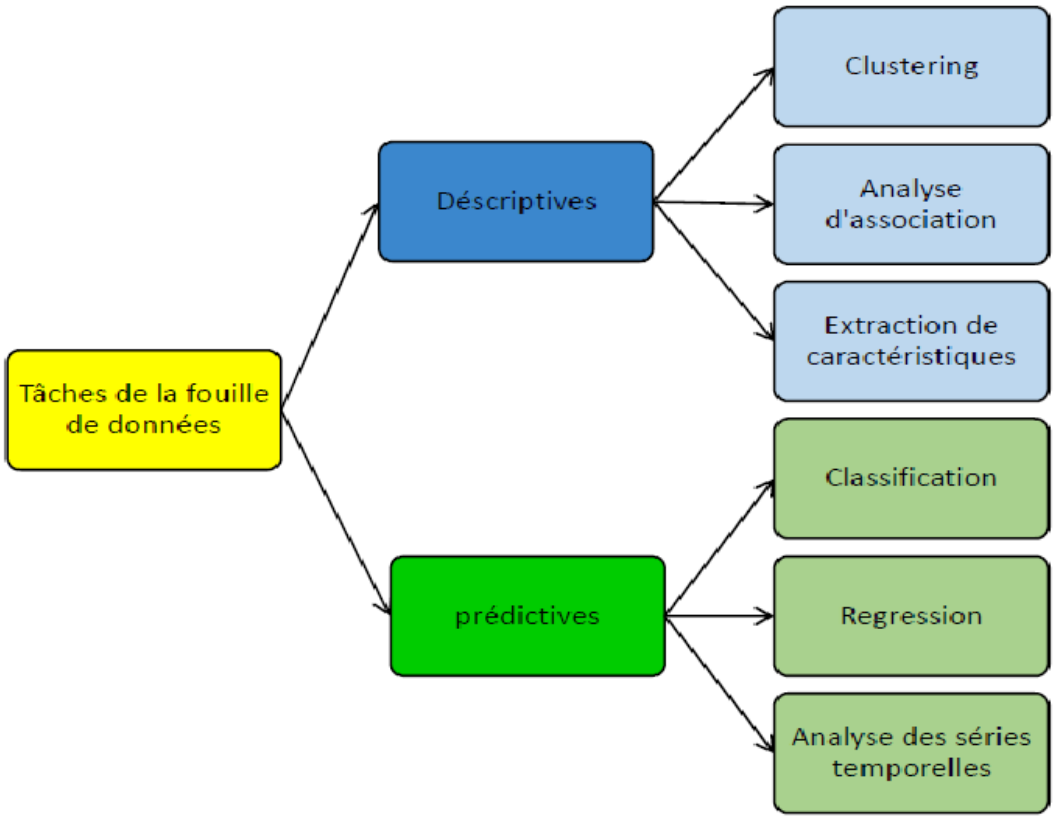

Figure 1.6 : Classification des tâches de la fouille de données.

1.5.1.La description

Parfois le but du Data Mining est simplement de décrire ce qui se passe sur une base de données compliquée en expliquant les relations existantes dans les données pour en premier lieu comprendre le mieux possible les individus, les produits et les processus présents sur cette base. Une bonne description d'un comportement implique souvent une bonne explication de celui-ci [7].

Dans la société Américaine nous pouvons prendre comme exemple comment une simple description, « les femmes supportent le parti Démocrate plus que les hommes », peut provoquer beaucoup d'intérêt et promouvoir les études de la part des journalistes, sociologistes, économistes et les spécialistes en politique. [8]

1.5.2.Le groupement (clustering)

Désigne le groupement des données, des observations ou des cas dans des classes d'objets similaires. Un cluster maximise la similarité des objets du même cluster et minimise la similarité des objets de clusters différents.

En effet, il n'y a pas de variable cible pour le clustering. La tâche de clustering ne cherche pas à estimer ou à prédire la valeur d'une variable cible. Mais plutôt à segmenter l'ensemble des données en sous-groupes relativement homogènes à l'aide de mesures de distances.

La segmentation est une tâche d'apprentissage « non supervisée » car on ne dispose d'aucune autre information préalable que la description des exemples. Après application de l'algorithme et donc lorsque les groupes ont été construit, d'autres techniques ou une expertise doivent dégager leur signification et leur éventuel intérêt.

Les algorithmes du clustering peuvent être appliqués dans des différents domaines, tel que :la découverte des groupes de clients ayants des comportements semblables.

La Classification des plantes et des animaux étant donné leurs caractéristiques, et la segmentation des observations des épicentres pour identifier les zones dangereuses. [4] (voir figure1.7)

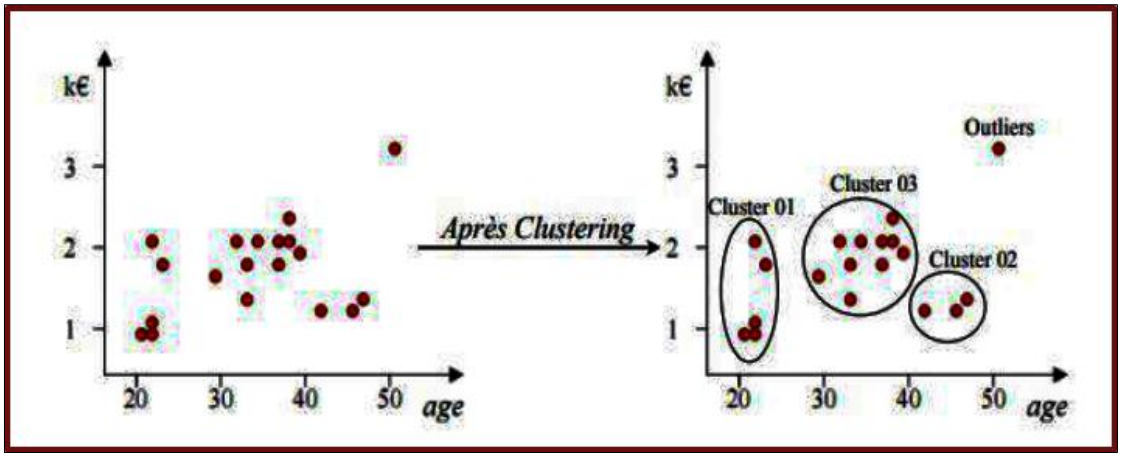

Figure 1. 7: Application de clustering .[6]

1.5.3. La classification :

La classification a pour rôle d'affecter un objet à une classe prédéfinie selon une mesure de proximité. Les techniques de classification commencent par définir un « plan d'expérience

» ou un ensemble de données d'apprentissage sur lequel on applique les méthodes de classification. [4]

Exemples :

Voici quelques exemples de l'utilisation des tâches de classification :

- Attribuer ou non un prêt à un client dans une banque ;
- Établir un diagnostic médical ;
- Accepter ou refuser un retrait dans un distributeur de billets ;
- Attribuer un sujet principal à un article de presse ; [5]

1.5.4. Les règles d'association :

Les règles d'associations ou groupement par similitude consiste à déterminer quels attributs "vont ensemble". C'est la tâche la plus répandue dans le monde du business, où elle est appelée l'analyse d'affinité ou l'analyse du panier du marché, est l'association des recherches pour mesurer la relation entre deux et plusieurs attributs. Les règles d'associations sont de la forme "Si antécédent, alors conséquent". [4]

Exemple :

Trouver dans un supermarché quels produits sont achetés ensemble et quels sont ceux qui ne s'achètent jamais ensemble, déterminer la proportion des cas dans lesquels un nouveau médicament peut générer des effets dangereux. [5]

1.5.5. La prédiction :

La prédiction est semblable à la classification et l'estimation, sauf que pour la prévision, les résultats se situent dans le futur. Exemples de tâches de prédiction :

- Prédire, au vu de leurs actions passées, les départs de clients dans une banque
- Prévoir le champion de la coupe du monde en football en se basant sur la comparaison des statistiques des équipes. [4]

1.5.6. La visualisation :

Prolongement de la cartographie multidimensionnelle, la visualisation de données se fonde sur une analyse préalable, qualitative et/ou quantitative, de données brutes ou structurées (*datamining*). Il s'agit d'opérer un « point de vue » sur un sous-ensemble de résultats pertinents afin d'en faciliter la compréhension, sinon d'offrir la possibilité de les

afficher selon différents critères (langue, couleur, forme, support, terminaux, droits d'accès, etc.). L'idée d'opérer une transformation sémiotique entre les résultats d'une analyse de données (numériques, catégorielles, textuelles) et une représentation graphique s'est par la suite étendue et perfectionnée, notamment grâce à l'informatisation qui permet de traiter des quantités de données de plus en plus massives.

La visualisation demande l'exécution de trois processus préalables :

1) extraire des données (à partir de différentes sources, notamment les interfaces de programmation) ;

2) les transformer (notamment *via* des méthodes algorithmiques) ;

3) charger ces données transformées (notamment dans un système d'aide à la décision) ;

4) visualiser les données chargées ou traitées.[16]

1. 6. Conclusion

Dans ce chapitre, nous avons exposé le processus ECD et ses différentes étapes en général, la fouille de données et les techniques utilisées pour extraire l'information utile .

Dans notre travail nous intéressons aux techniques de la classification automatique, nous avons vu qu'elle permet de regrouper des objets (individus ou variables) en groupes ou classes. Les détails, feront l'objet du chapitre suivant.

2.1. Introduction

« Le seul moyen de faire une méthode instructive et naturelle, est de mettre ensemble les choses qui se ressemblent et de séparer celles qui différent les unes des autres. »

M. Georges Buffon, Histoire naturelle, 1749.

La classification est une tâche très importante dans le data mining, et qui consomme beaucoup de recherches pour son optimisation. La classification supervisée est l'une des techniques les plus utilisées dans l'analyse des bases de données. Elle permet d'apprendre des modèles de décision qui permettent de prédire le comportement des exemples futurs.

La classification supervisée consiste à inférer à partir d'un échantillon d'exemples classés une procédure de classification. Un système d'apprentissage effectue la recherche d'une telle procédure selon un modèle. Les systèmes d'apprentissage peuvent être basés sur des hypothèses probabilistes (classifieur naïf de Bayes, méthodes paramétriques) ; sur des notions de proximité (plus proches voisins) ; sur des recherches dans des espaces d'hypothèses (arbres de décision, réseaux de neurones).

La problématique de la classification consiste à affecter les objets d'un ensemble de données à des catégories ou classes prédéfinies. Ce type de question fait partie des problèmes rencontrés lors de la phase du groupement et la classification de données.

Dans ce chapitre nous présenterons un panorama des méthodes de classification les plus connues et qui font référence à l'existence de groupes ou classes de données .

2.2. Définition

La classification est une discipline relié de près ou de loin a plusieurs domaines, elle est connue aussi sous noms variés (classification, clustering, segmentation,. . .) selon les objets qu'elle traite et les objectifs qu'elle vise à atteindre.

Pour attribuer une définition au terme « classification », il faudrait d'abord définir ses racines, ça vient du verbe "classer" qui désigne plus une action qu'un domaine, ou plutôt une série de méthodes qu'une théorie unifiée.

En mathématique, On appelle classification, la catégorisation algorithmique d'objets. Elle consiste à attribuer une classe ou catégorie à chaque objet (ou individu) à classer, en se basant sur des données statistiques. Elle fait couramment appel aux méthodes d'apprentissage et est largement utilisée en reconnaissance de formes.

Il est important de noter qu'il ne faut pas confondre entre ces deux termes :
« classification » et « classement », au fait le mot classification en anglais signifie une chose, alors que le même mot en français ait une autre signification (utilité).

Dans un classement on affecte les objets à des groupes préétablis, c'est le but de l'analyse discriminante que de fixer des règles pour déterminer la classe des objets. La classification est donc, en quelque sorte, le travail préliminaire au classement, savoir la recherche des classes "naturelles" dans le domaine étudié, en anglais « Cluster Analysis ».[9]

Cette collision entre les termes peut se résumer comme suite :

Français	**Anglais**
Classification	Clustering
Classement	Classification

Table 2.1 : Les noms attributs à la classification en Français/Anglais.[9]

2.3. Le processus de classification

Le processus de classification des données implique les tâches d'apprentissage, de test et de calibration. Au premier lieu, les données d'apprentissage sont analysées par un algorithme de classification, afin de construire des règles ou des modèles de classification. Ensuite, d'autres données de test sont utilisées pour estimer l'exactitude de ces règles ou ces modèles. Si la précision est acceptable, ces règles ou ces modèles de prévision peuvent être appliqués aux nouveaux enregistrements de données, sinon le modèle va être réévalué ou calibré, afin d'atteindre la précision prédéfinie à l'avance.

Par exemple, la surveillance de la fraude par cartes de crédit, en surveillant des millions de comptes qui sont représentés par des enregistrements d'activités frauduleuses ou légales. Dans ce cas, l'algorithme d'apprentissage du classificateur utilise ces exemples préclassifiés pour déterminer l'ensemble des paramètres requis pour une discrimination appropriée.

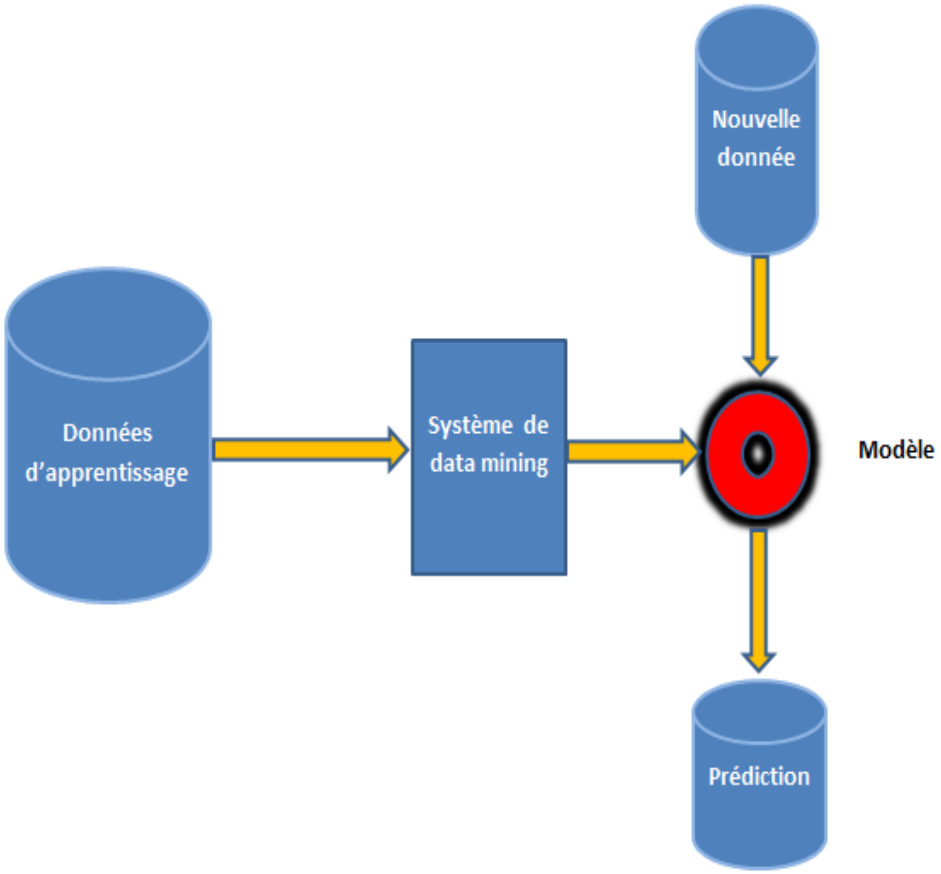

Figure 2.1 : Processus de classification.

2.4. Avantages et inconvénients de la classification

2.4.1. Les avantages

✓ La classification nous permet de voir des relations entre des choses qui peuvent ne pas être évidentes lorsqu'on les regarde dans leur ensemble.

✓ une pratique assez courante dans l'éducation.

✓ Stocker des quantités massives de données de manière non organisée est à la fois coûteux et risqué.

✓ Les organisations peuvent utiliser la classification des données pour maintenir la confidentialité, l'intégrité et la disponibilité de leurs données.

2.6.1. Les inconvénients

- ✓ les classifications elles-mêmes sont basées sur des jugements subjectifs, qui peuvent ou non être partagés par tous les participants. Cela conduirait à des différences de valeur perçue.
- ✓ Un autre problème est que ce sont les humains qui font tout ce classement et cette catégorisation, pour notre propre bénéfice ; le monde naturel ne fonctionne pas selon des catégories. Ainsi, nos propres connaissances peuvent parfois créer une limite artificielle à notre capacité à catégoriser correctement quelque chose. Ceci est démontré par les disputes fréquentes et la réorganisation des relations évolutives, ou par les fausses classifications causées par des recherches frauduleuses ou erronées.

2.5. Evaluation des modèles

La résolution d'un problème de classification s'effectue en comparant des modèles afin de choisir le plus apte à r résoudre le problème posé. L'́évaluation des modèles est donc un préalable inévitable à la classification. Elle est nécessaire pour connaıtre les performances d'un modèle et déterminer s'il est globalement significatif.

Dès lors, deux objectifs se dégagent: l'évaluation et la comparaison de modèles en vue de la sélection. La sélection du modèle idéal peut être envisagée :

• En comparant différentes méthodes de classification pour un même sous-ensemble de variables ;

• En comparant différentes méthodes de sélection de variables, pour une même méthode de classification ;

• En comparant simultanément des méthodes de classification et des méthodes de sélection de variables.[10]

➢ Critères d'évaluation

a)Taux de bon apprentissage

Parmi tous les exemples, quelle proportion est bien classée ?

b)Précision de la classe k

Parmi les exemples classés dans la classe k, quelle proportion est effectivement de la classe k ?

d)Rappel de la classe k

Parmi les exemples de la classe k, quelle proportion se retrouvent classés dans la classe k ?

e)Précision contre Rappel

f)Matrice de confusion : table de contingence

2.6. Méthodes de classification

Parmi les méthodes de classification, on peut citer :

- Les réseaux de neurones artificiels.
- Les arbres de décision.
- La classification bayésienne.
- La machine à vecteurs de support.

2.6.1. Les arbres de décision

L'arbre de décision est l'une des techniques de fouille de données les plus utilisées, car son modèle est facile à comprendre pour les utilisateurs. Elle est initialement utilisée dans la théorie de la décision et les statistiques, et devenue un outil très efficace dans d'autres domaines tels que la fouille de données, l'apprentissage automatique et la reconnaissance de formes [11]. Il existe plusieurs algorithmes d'induction d'arbre de décision, parmi ces algorithmes, on peut citer l'algorithme ID3[12] qui a été développé par John Ross Quinlan. Plus tard, il a présenté C4.5 [13], qui était le successeur de ID3. Les algorithmes ID3 et C4.5 adoptent une approche gourmande, dans lesquels il n'y a pas de retour en arrière et les arbres sont construits d'une manière récursive en utilisant une approche diviser pour régner de haut en bas.

Un arbre de décision est une structure qui inclut un nœud racine, des branches, des nœuds internes et des nœuds de feuille. Chaque nœuds interne représente une condition sur un attribut, chaque branche indique le résultat d'une condition et chaque nœud de feuille porte une étiquette de classe. Le nœud le plus haut dans l'arborescence est le nœud racine. Par exemple, l'arbre de décision suivant (figure 2.2) indique si un client d'une entreprise est susceptible d'acheter un ordinateur ou non. Chaque nœud interne représente une condition sur un attribut et chaque nœud de feuilles représente une valeur de variable cible, à savoir acheter un ordinateur ou non.

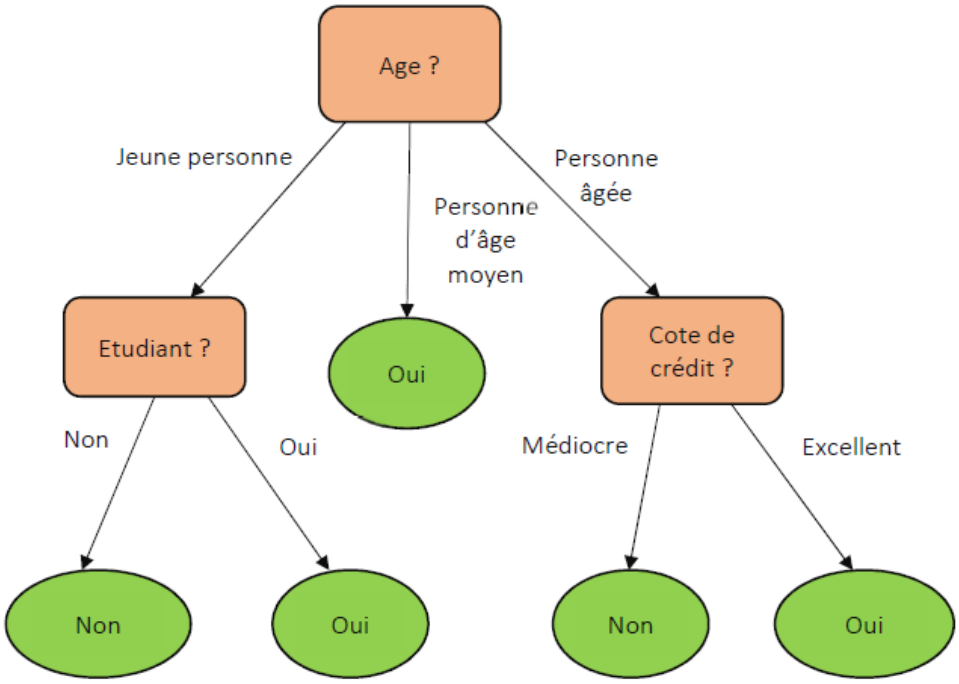

Figure 2.2 : Exemple d'un arbre de décision.

2.6.2. Les réseaux bayésiens

Le réseau bayésien est un modèle graphique probabiliste, utilisé pour représenter les indépendances conditionnelles entre un ensemble de variables aléatoires. Un réseau bayésien est représenté essentiellement par un graphe orienté acyclique DAG (Directed Acyclic Graph),les variables sont représentées par des nœuds et les relations entre ces nœuds sont représentées par des arcs. Dans ce graphe, nous associons à chaque nœud une table de probabilités conditionnelles qui représente la distribution de probabilité de la variable correspondante démontrant les dépendances entre cette variable et ses parents (les noeuds commençant des arcs vers cette variable). Donc, le processus d'apprentissage du réseau bayésien comprend l'apprentissage de la structure et des paramètres du réseau.

Les réseaux bayésiens ont été largement utilisés pour extraire les dépendances sous-jacentes entre un ensemble de variables aléatoires. Puis, utiliser ces modèles pour répondre à des questions de prévision ou pour trouver l'explication la plus probable d'une séquence d'événements. La séparation directionnelle (D-Séparation) est un concept important de réseaux bayésiens qui nous aide à découvrir les relations entre un ensemble de variables aléatoires.

Considérons X, Y et Z comme étant trois sous-ensembles de nœuds dans un DAG donné.

Nous disons que Z sépare X de Y (noté par X ⊥ Y | Z), si pour chaque nœud x appartenant à X,et un autre nœud y appartenant à Y, il existe un nœud z appartenant à Z qui convient à une des situations représentées à la figure 2.3 (un nœud ombré signifie que le nœud est observé) :

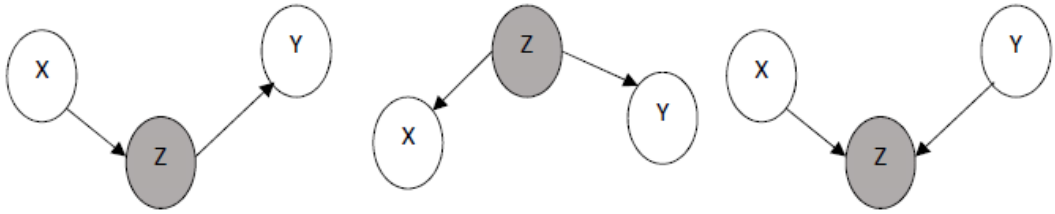

Figure 2.3 : Les règles de D-séparation

La probabilité conditionnelle de deux ensembles indépendants de variables aléatoires A et B, étant donné un autre ensemble C, peut être formulée comme suit :

$$P(A, B|C) = P(A|C).\ P(B|C)$$

La probabilité d'un ensemble de variables dans un réseau bayésien est le produit des probabilités conditionnelles de chaque variable de l'ensemble, compte tenu des valeurs de ses parents. Cette propriété peut être formulée comme suit :

$$P(x_1, x_2 \ldots., x_n) = \prod_{i=1}^{n} P(x_i|parents(x_i))$$

Les réseaux bayésiens sont caractérisés par les propriétés suivantes :

- La propriété de suffisance causale indique que les variables qui sont les parents des variables observées du domaine ne doivent pas être cachées.
- La propriété causale de Markov détermine l'indépendance d'une variable donnée par rapport à l'ensemble de ses non-descendants compte tenu de ses parents.

La propriété de fidélité indique que la structure du réseau bayésien et la distribution de probabilité s'accordent sur les relations de dépendance/indépendance découvertes entre les variables. Dans ce cas, on dit que la distribution de probabilité et la structure sont fidèles l'une à l'autre .

La Frontière de Markov B(x) est la plus petite famille d'une variable x donnée, qui se compose de ses parents, de ses partenaires (les autres parents de ses enfants) et de ses enfants.

Cette famille sépare (protège) le nœud x de tous les autres nœuds variables. En d'autres termes, disons que x est une variable de l'ensemble de variables U, donc la frontière de Markov est la couverture de Markov minimale de x, qui est l'ensemble minimal de variables BL(x) ⊆ U (il ne pouvait pas être unique), ce qui rend la variable x indépendante de toute autre variable y ∉ {BL(x) ∪ {x}}, et on note : B(x) = x ⊥ y | BL(x).

2.6.3. Les réseaux de neurones artificiels

Le réseau de neurones artificiels (RNA) est modelé sur le réseau neuronal biologique. Il représente une interconnexion de nœuds analogues aux neurones. Chaque réseau neuronal comporte trois éléments essentiels, à savoir le caractère du nœud, la topologie du réseau et les règles d'apprentissage. Le caractère de nœud détermine la façon dont les signaux sont traités par le nœud. Il est constitué du nombre d'entrées et de sorties associées au nœud, du poids associé à chaque entrée et sortie, et de la fonction d'activation. La topologie du réseau détermine la façon dont les nœuds sont organisés et connectés. Les règles d'apprentissage déterminent comment les poids sont initialisés et ajustés. Les sections suivantes expliquent chacune de ces composantes .

2.6.3.2.La structure d'un noeud de RNA

La structure de base d'un nœud RNA est illustrée à la figure 2.3. Chaque nœud reçoit sesentrées des nœuds « en amont » et fournit des sorties aux nœuds « en aval ». Chaque connexion entre 2 nœuds possède un poids. Lorsque la somme pondérée des entrées dépasse la valeur seuil du nœud, ce nœud active et transmet un signal via une fonction de transfert aux nœuds voisins. Ce processus peut être exprimé sous la forme d'un modèle mathématique [18]:

$$y = f(\sum_{i=0}^{n} w_i xi \quad < \text{T})$$

où y est la sortie du nœud, f la fonction de transfert, le poids de l'entrée xi , et T la valeur de seuil. La fonction de transfert a plusieurs formes. Par exemple, la fonction de transfert non linéaire est plus utile que la fonction linéaire , car seuls quelques problèmes sont séparables linéairement.

Parmi les fonctions de transfert les plus simples et les plus utilisées, il existe la fonction step

$$y = \begin{cases} 0 & if \quad \sum_{i=0}^{n} w_i xi \quad < T \\ 1 & if \quad \sum_{i=0}^{n} w_i xi \quad \geq T \end{cases}$$

La fonction sigmoïde est aussi souvent utilisée comme fonction d'activation, puisque cette fonction et sa dérivée sont des fonctions continues. Cette fonction est formulée comme suit :

$$y = \frac{1}{1 + \exp(-\beta x)}$$

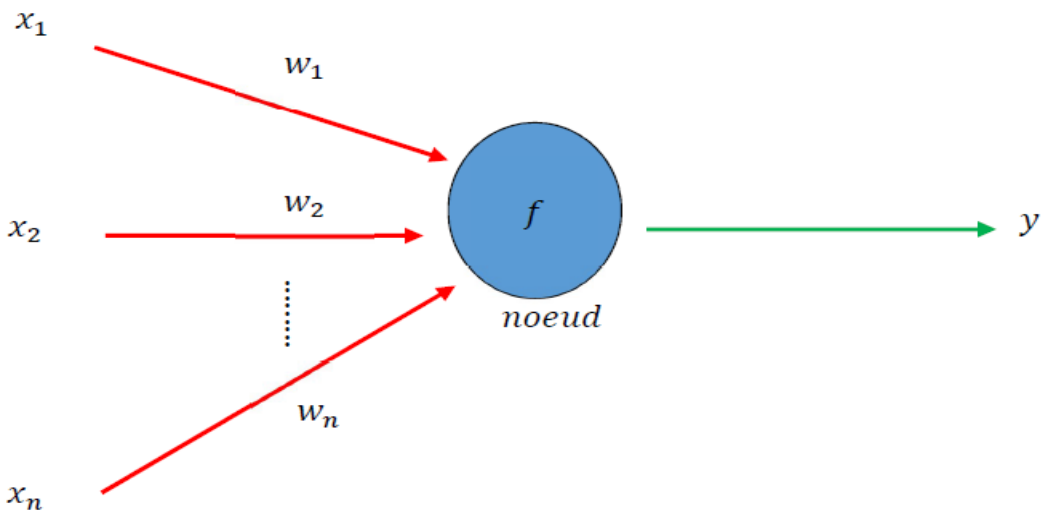

Figure 2.4 : Structure d'un nœud de réseau de neurones xi.
les entrées, wi : les poids, f : la fonction de transfert et : y la sortie

2.6.3.2.Les topologies de réseau de neurones

Selon la figure 2.5(a) représentant l'architecture générale d'un réseau de neurones artificiels, les nœuds sont organisés en tableaux linéaires, appelés couches. Généralement, il a une couche d'entrée, une couche de sortie, ainsi qu'aucune, ou plusieurs couches cachées. La

conception de la topologie du réseau implique de déterminer le nombre de nœuds à chaque couche, le nombre de couches dans le réseau et les différentes connexions entre les nœuds.

Ces facteurs sont d'abord déterminés par intuition et optimisés par de multiples cycles d'expériences. Aussi, quelques méthodes rationnelles peuvent être utilisées pour concevoir un réseau neuronal. Par exemple, le réseau neuronal génétique utilise un algorithme générique pour résoudre certains problèmes concernant la topologie du réseau, par exemple la sélectionde caractéristiques d'entrée [14].

Il existe deux types de connexions entre les nœuds. La première est une connexion unidirectionnelle sans boucle de retour. L'autre est une connexion en boucle dans laquelle la sortie des nœuds peut être l'entrée des nœuds de niveau précédent ou de même niveau.Sur la base du type de connexions susmentionné, les réseaux de neurones peuvent être classés en deux types, à savoir les réseaux proactifs (feedforward) et les réseaux récurrents.

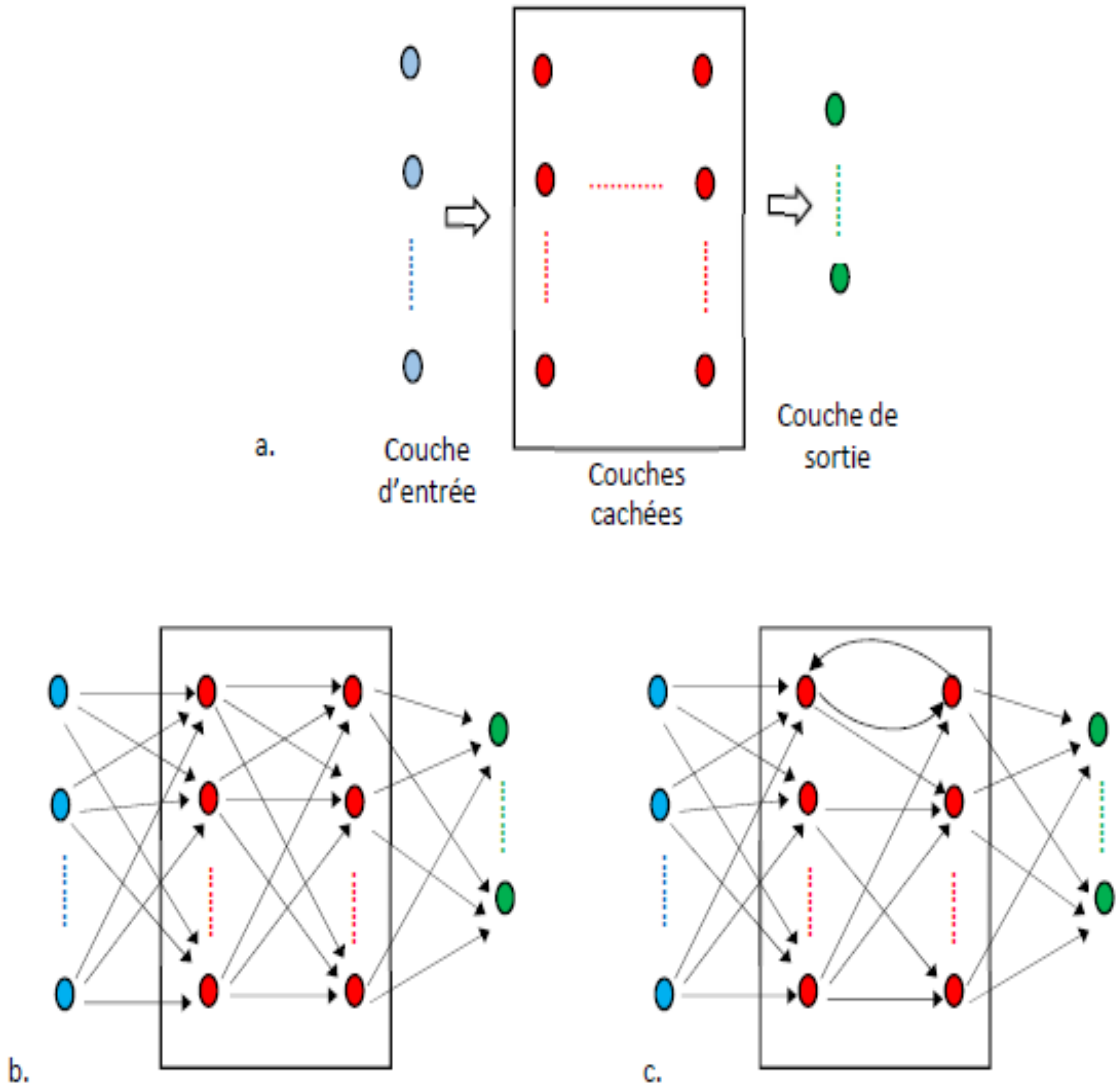

Figure 2.5 : (a) Structure générale des réseaux de neurones artificiels. (b) Réseau de neurones perception. (c) Réseau de neurones récurrent .

Dans les réseaux proactifs (réseau statique), le signal ne circule que dans un seul sens, c'est-à-dire qu'une entrée est associée à une sortie particulière. Tandis que, dans les réseaux récurrents (réseaux dynamiques), pour une entrée, l'état du réseau récurrent se modifie durant chaque cycle, jusqu'à ce qu'il atteigne un point d'équilibre, ainsi une entrée produit une série de sorties. Le réseau de neurones de types perception est un réseau proactif largement utilisé . Parmi les réseaux récurrents les plus connus, il existe le réseau de Hopfiel et les cartes autoorganisées de Kohonen.

2.6.4. La machine à vecteurs de support(MVS)

Le MVS est un classificateur dit linéaire, ça veut dire que, dans le cas parfait, les données (document texte dans notre cas) doivent être linéairement séparables. Ainsi notre corpus est représenté comme étant un espace vectoriel, où chaque document texte est représenté par un point dans ce dernier. La problématique maintenant est de trouver le meilleur séparateur (ligne, plan ou hyperplan) qui partage notre corpus en deux catégories. L'espace entre ces deux catégories est appelé marge, qui est définie par les points (Vecteurs de support) les plus proches du séparateur, de part et d'autre. Le but étant essentiellement de maximiser cette marge, plus elle est grande meilleurs est le résultat. Le classificateur se généralise bien avec les nouvelles données.

Toutefois, si les données ne sont pas linéairement séparables, la MVS peut être modifiée pour tolérer un minimum d'erreurs. Désormais, le but est de maximiser la marge et de minimiser l'erreur de classification. Une autre alternative pour parer à la non séparabilité des données, est de passer à un espace de dimension supérieur.
La figure (2.6) illustre la façon dont MVS départage les données dans le cas où elles sont linéairement séparables, de plus, elle choisit la séparation la plus optimale où la marge est maximale.[26]

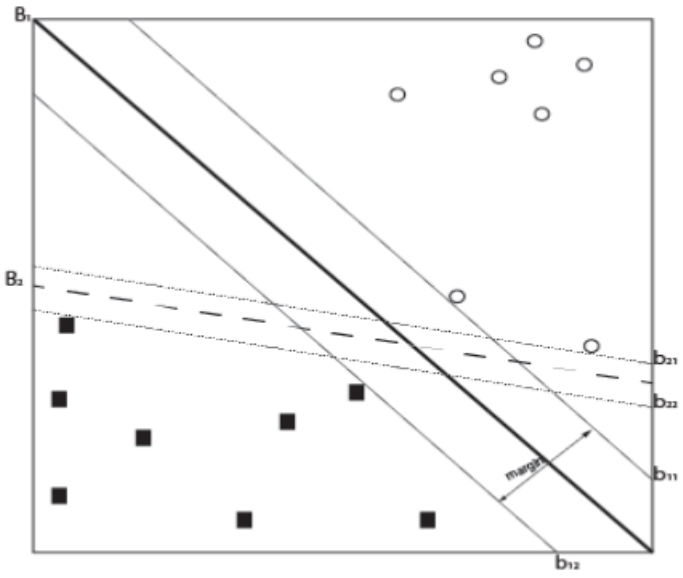

Figure 2.6: Séparation correcte(B2) et séparation optimale(B1). [15]

2.7.Conclusion

La classification de données est l'une des tâches importantes du data mining, dans ce chapitre nous présentons le principe, le processus et les différents méthodes concernant cette tache , nous avons également définit les avantages et les inconvénients de classification et comment évaluer les résultats de classification.

3.1. Introduction

Dans ce chapitre nous allons présenter une conception d'un système basée sur un processus ECD.Ce système permet la recherche du profils des malades par la méthode classification ,il suive des différents étapes depuis le téléchargement des données jusqu'à arriver à avoir des résultats fiables exploitables par un système de prédiction.

Nous proposons l'utilisation de l'algorithme j84 pour l'arbre de décision sous un environnement appelé WEKA destiné à la fouille de données.

3.2. La recherche de type de maladie par l'algorithme j48

3.2.1 Présentation générale de l'algorithme

WEKA propose une large gamme d'exemples de jeux de données pour appliquer des algorithmes d'apprentissage automatique. Les utilisateurs peuvent effectuer des tâches d'apprentissage automatique telles que la classification, la régression, la sélection d'attributs, l'association sur ces exemples de jeux de données, et peuvent également apprendre l'outil en les utilisant.

L'explorateur WEKA est utilisé pour exécuter plusieurs fonctions, à partir du prétraitement. Le prétraitement prend l'entrée sous forme de fichier .arff, traite l'entrée et donne une sortie qui peut être utilisée par d'autres programmes informatiques. Dans WEKA, la sortie du prétraitement donne les attributs présents dans le jeu de données qui peuvent être ensuite utilisés pour l'analyse statistique et la comparaison avec les étiquettes de classe.

WEKA propose également de nombreux algorithmes de classification pour l'arbre de décision. j48 est l'un des algorithmes de classification populaires qui génère un arbre de décision. À l'aide de l'onglet Classifier, l'utilisateur peut visualiser l'arbre de décision. Si l'arbre de décision est trop peuplé, l'élagage de l'arbre peut être appliqué à partir de l'onglet Prétraitement en supprimant les attributs qui ne sont pas requis et en relançant le processus de classification.

j48 c'est un algorithme pour générer un arbre de décision qui est généré par C4.5 (une extension d'ID3). Il est également connu comme un classificateur statistique. Pour la

classification de l'arbre de décision, il est besoin d'une base de données.(dans notre travail c'est le fichier urinaire.csv.)

L'algorithme C4.5 de Quinlan actualise J48 pour créer un arbre de décision C4.5 ajusté. Tous les aspects de l'information consiste à diviser en sous-ensembles mineurs pour fonder une décision.

j48 offre le gain de données standardisé qui donne vraiment les résultats diviser les informations en choisissant un attribut. Pour résumer, les données standardisées extrêmes d'attribut obtenues sont utilisé. Les sous-ensembles mineurs sont renvoyés par l'algorithme. Les stratégies de fractionnement s'arrêtent si un sous-ensemble a une place avec un classe similaire dans toutes les instances. J48 développe un nœud de décision utilisant les estimations attendues de la classe. J48 l'arbre de décision peut traiter des caractéristiques particulières, des estimations d'attributs perdus ou manquants des données et des variations coûts d'attribut. Ici, la précision peut être augmentée par l'élagage (Venkatesan, 2015).

3.2.2L'algorithme

Étape 1 : La feuille est étiquetée avec une classe similaire si les instances appartiennent à une classe similaire.

Etape 2 : Pour chaque attribut, les données potentielles seront chiffrées et le gain dans les données sera tiré du test sur l'attribut.

Étape 3 : Enfin, le meilleur attribut sera choisi en fonction du paramètre de sélection actuel.

3.3. Le système adopté

L'un des principaux inconvénients des algorithmes statiques de classification (par exemple, les arbres de décision) est qu'ils ne tiennent pas compte du temps auquel les données sont arrivées. Cela a incité les chercheurs à développer de nouvelles méthodes de classification incrémentale (par exemple, les arbres de décision incrémentales) qui sont mises à jour lorsque de nouvelles données arrivent , au lieu de réexécuter l'algorithme à partir de zéro.

Les algorithmes incrémentales semblables à l'algorithme illustré à la figure 3.2 mettent à jour leurs modèles prédictifs lorsque de nouvelles données apparaissent à différents moments dans le temps, contrairement aux algorithmes de classification habituelles (statiques)

qui construisent des modèles par lots (figure 3.1). De plus, la classification incrémentale permet d'extraire de nouvelles informations (modèles) à partir d'un ensemble de données futures qui seront disponibles ultérieurement sans perdre au préalable les connaissances du domaine étudié [17].

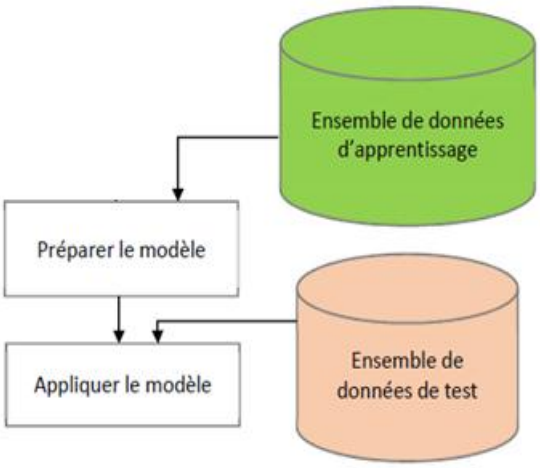

Figure 3.1 : Modèle de classification traditionnel .

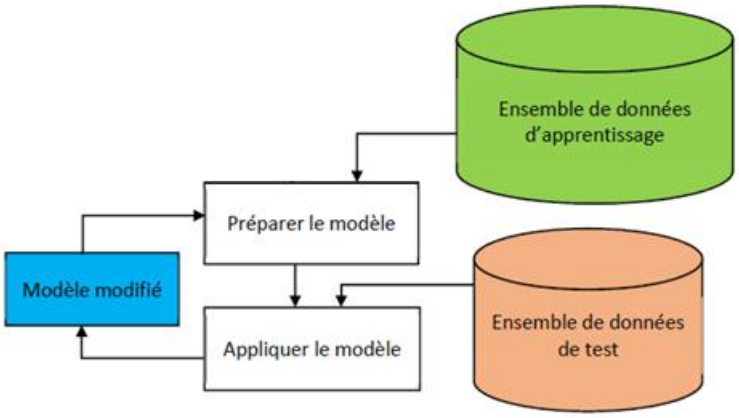

Figure 3.2 : Modèle de classification incrémentale.

Dans notre approche on va utiliser le Dataset **urinaire.csv** comme une source d'obtenir notre données des maladies pour qu'elles soient analyser par un environnement d'apprentissage WEKA, ce dernier va nous donner le modèle attendu qui sont les classes.

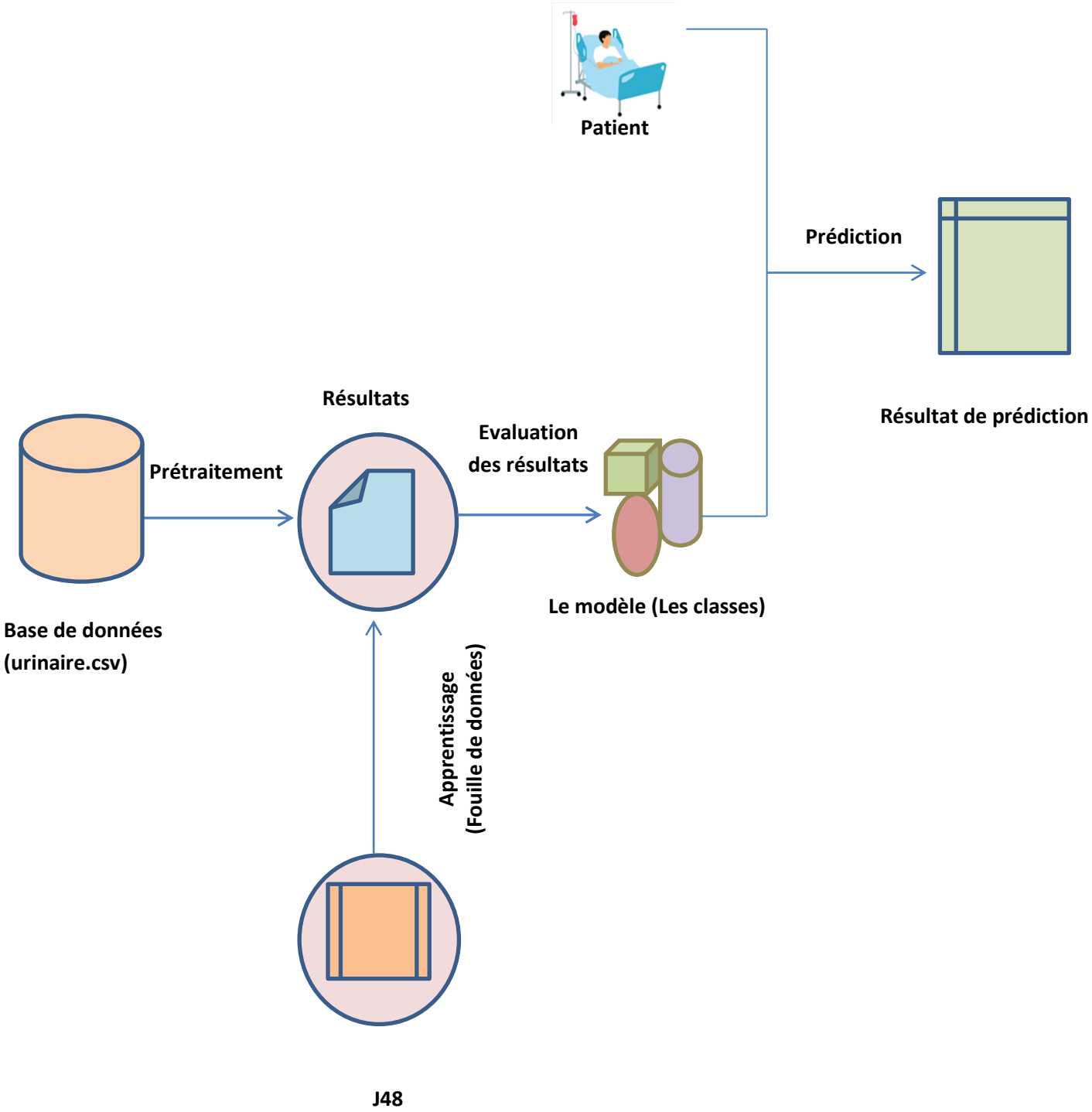

Figure 3.3 : Le système adopté .

3.4. Conclusion

Dans ce chapitre nous avons proposé une approche qui sert à extraire des nouvelles informations suivant un processus ECD , l'étape la plus importantes qui est destiné a construire des nouvelles connaissances est la fouille de données, cette dernière est considéré comme le coeur d'ECD qui utilise un ensemble des techniques dédiées à différentes tâches afin d'arriver à trouver des motifs valables exploitables. Dans le chapitre suivant, nous allons présenter la démarche suivie pour développer un système implémentant de cette approche.

4.1. Introduction

Dans ce chapitre nous présentons la dernière étape qu'est l'étape de réalisation, ainsi que le choix technique utilisé pour le développement et la préparation de l'algorithme du système expert, qui effectuera le diagnostic présomptif de deux maladies du système urinaire , La base pour la détection des règles était la théorie des ensembles approximatifs. Chaque instance représente un patient potentiel.

4.2. Le domaine d'application

Dans notre expérimentation la source des données utilisée est ASCII, les données ont été créées par un expert médical en tant qu'ensemble de données pour tester le système expert, qui effectuera le diagnostic présomptif de deux maladies du système urinaire.

4.3. L'environnement de l'expérimentation

4.3.1 Le logiciel WEKA

WEKA est un ensemble d'algorithmes d'apprentissage automatique qui peuvent être appliqués directement à un ensemble de données ou appelés à partir de votre propre code Java. WEKA contient des outils pour le prétraitement des données, la classification, la régression, le regroupement, les règles d'association et la visualisation.

Après l'installation , démarrer WEKA voir la figure 4.1.

Figure 4.1 : Fenêtre d'invite Weka 3.8.5.

Interface principale WEKA voir la figure 4.2

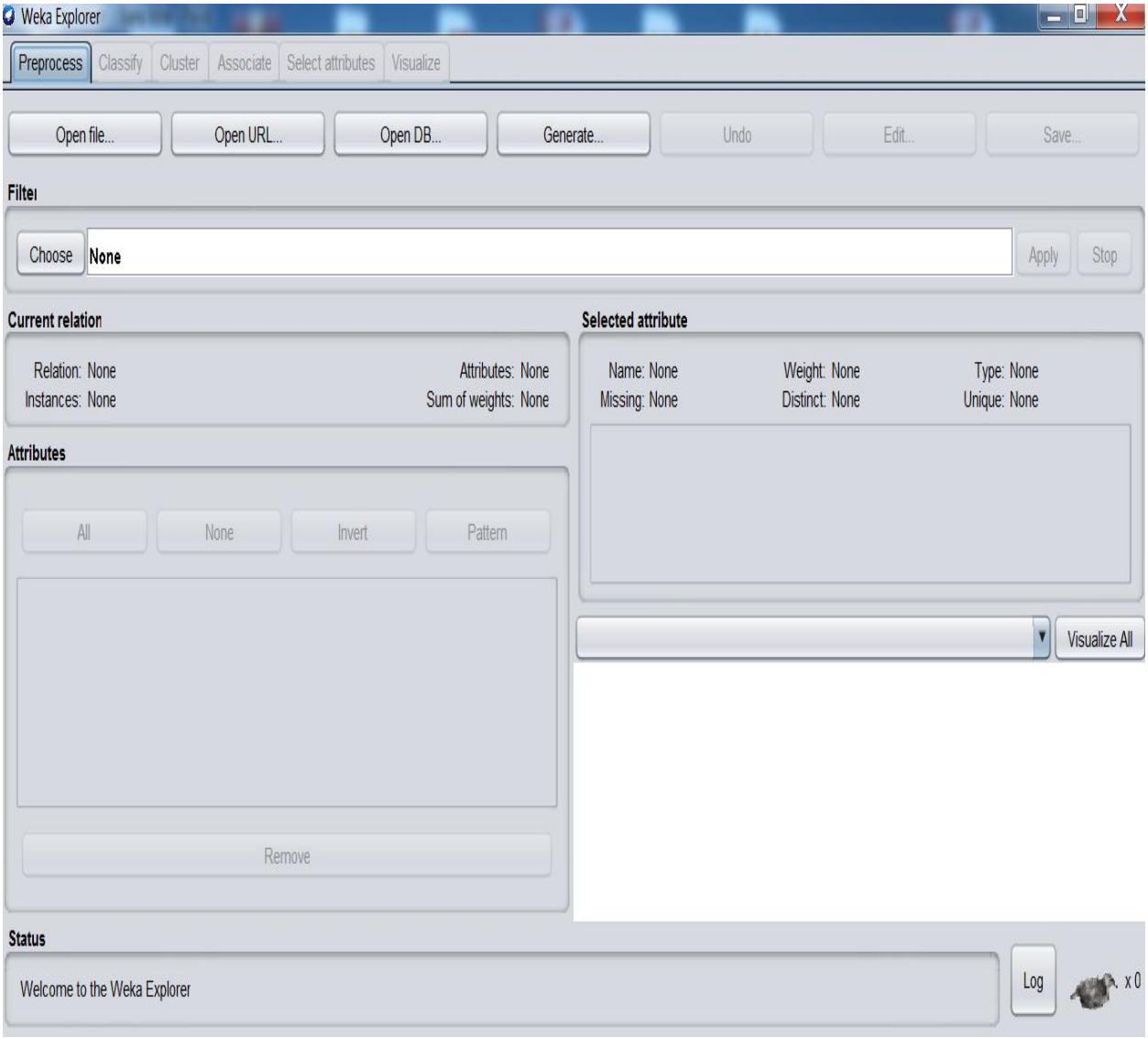

Figure 4.2 : Interface principale WEKA.

Les données destinées à la création de modèle sous le logiciel WEKA, ces données sont importées du Data Set au WEKA a l'aide de Microsoft EXCEL qui a réunit ces données dans un fichier CSV (urinaire. CSV) voir la figure 4.3.

4.3.2 Microsoft Excel (2007)

Est un logiciel tableur de la suite bureautique Microsoft Office développé et distribué par l'éditeur Microsoft. La version la plus récente est Excel 20191. Il est destiné à fonctionner sur les plates- formes Microsoft Windows, Mac OX, Androïde ou Linux (moyennant l'utilisation de Wine). Le logiciel Excel intègre des fonctions de calcul numérique, de représentation graphique, d'analyse de données (notamment de tableau croisé dynamique) et de programmation Microsoft Excel est utilisé dans notre expérimentation pour importer les données obtenu de Data Set UCI« represetory.»Ces données importées au niveau d'Excel vont être sauvegardé dans un fichier CSV pour qu'on puisse les importer au logiciel WEKA par la suite afin de réaliser l'étape de la création de modèle A partir de ces données, donc Microsoft Excel est l'intermédiaire entre le Data Set «urinaire.csv » et le logiciel WEKA.

1	35,5	no	yes	no	no	no	no	no		
2	35,9	no	no	yes	yes	yes	yes	no		
3	35,9	no	yes	no	no	no	no	no		
4	36	no	no	yes	yes	yes	yes	no		
5	36	no	yes	no	no	no	no	no		
6	36	no	yes	no	no	no	no	no		
7	36,2	no	no	yes	yes	yes	yes	no		
8	36,2	no	yes	no	no	no	no	no		
9	36,3	no	no	yes	yes	yes	yes	no		
10	36,6	no	no	yes	yes	yes	yes	no		
11	36,6	no	no	yes	yes	yes	yes	no		
12	36,6	no	yes	no	no	no	no	no		
13	36,6	no	yes	no	no	no	no	no		
14	36,7	no	no	yes	yes	yes	yes	no		
15	36,7	no	yes	no	no	no	no	no		
16	36,7	no	yes	no	no	no	no	no		
17	36,8	no	no	yes	yes	yes	yes	no		

Figure 4.3 : Echantillon des données de fichier(Urinaire.CSV).

Ces données vont être importées au WEKA avec ses 6 attributs (voir figure 4.4)

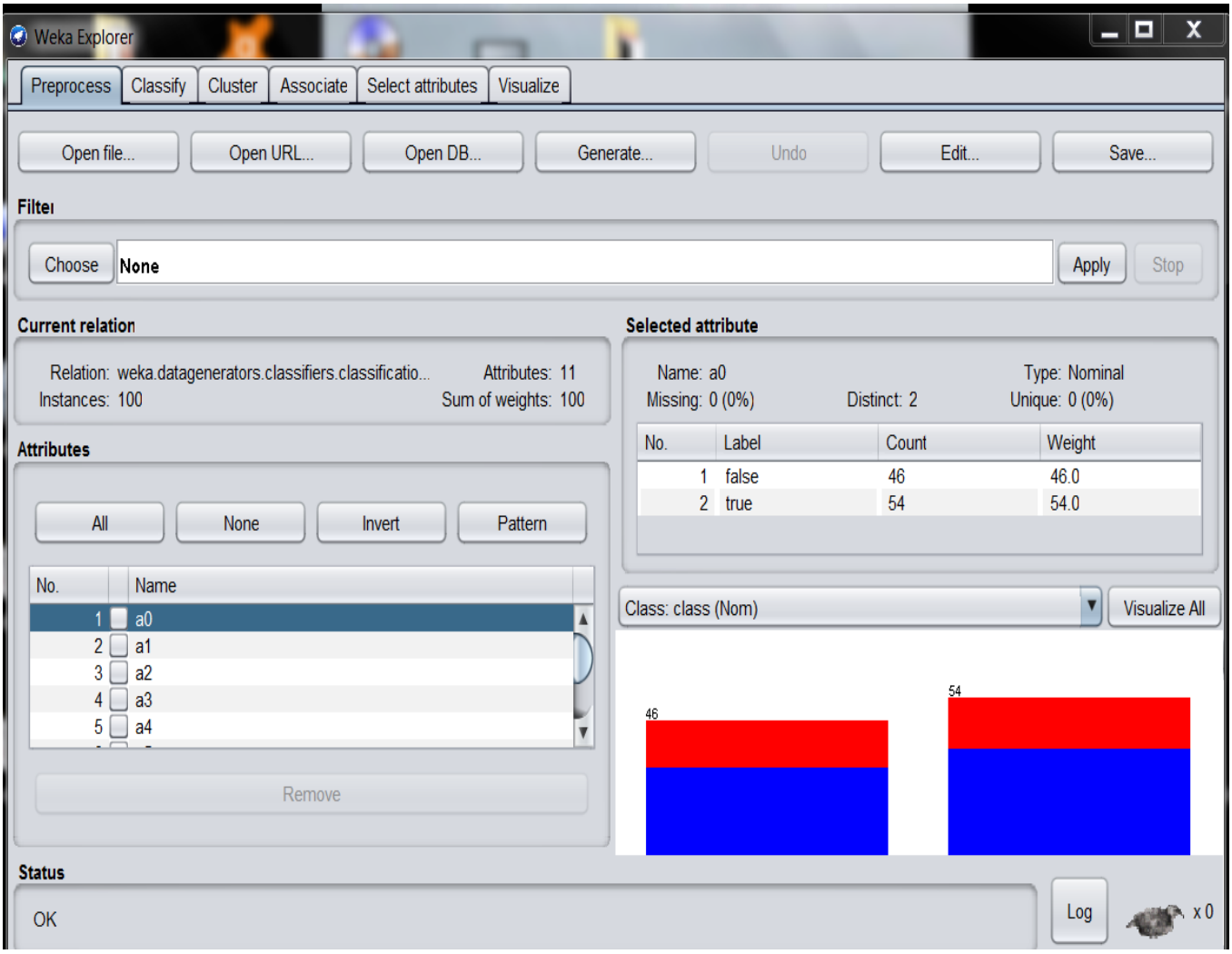

Figure 4.4 : Importation de fichier Urinaire. CSV au WEKA.

D'abord, vous avez à s'adapter à votre arbre de décision (j'ai utilisé le classificateur j48), de la façon habituelle. Dans la liste des résultats de panneau (en bas à gauche sur WEKA explorer), clique droit sur la sortie correspondante et sélectionnez "Visualiser l'arbre", comme illustré ci-dessous (voir figure 4.5et figure 4.6).

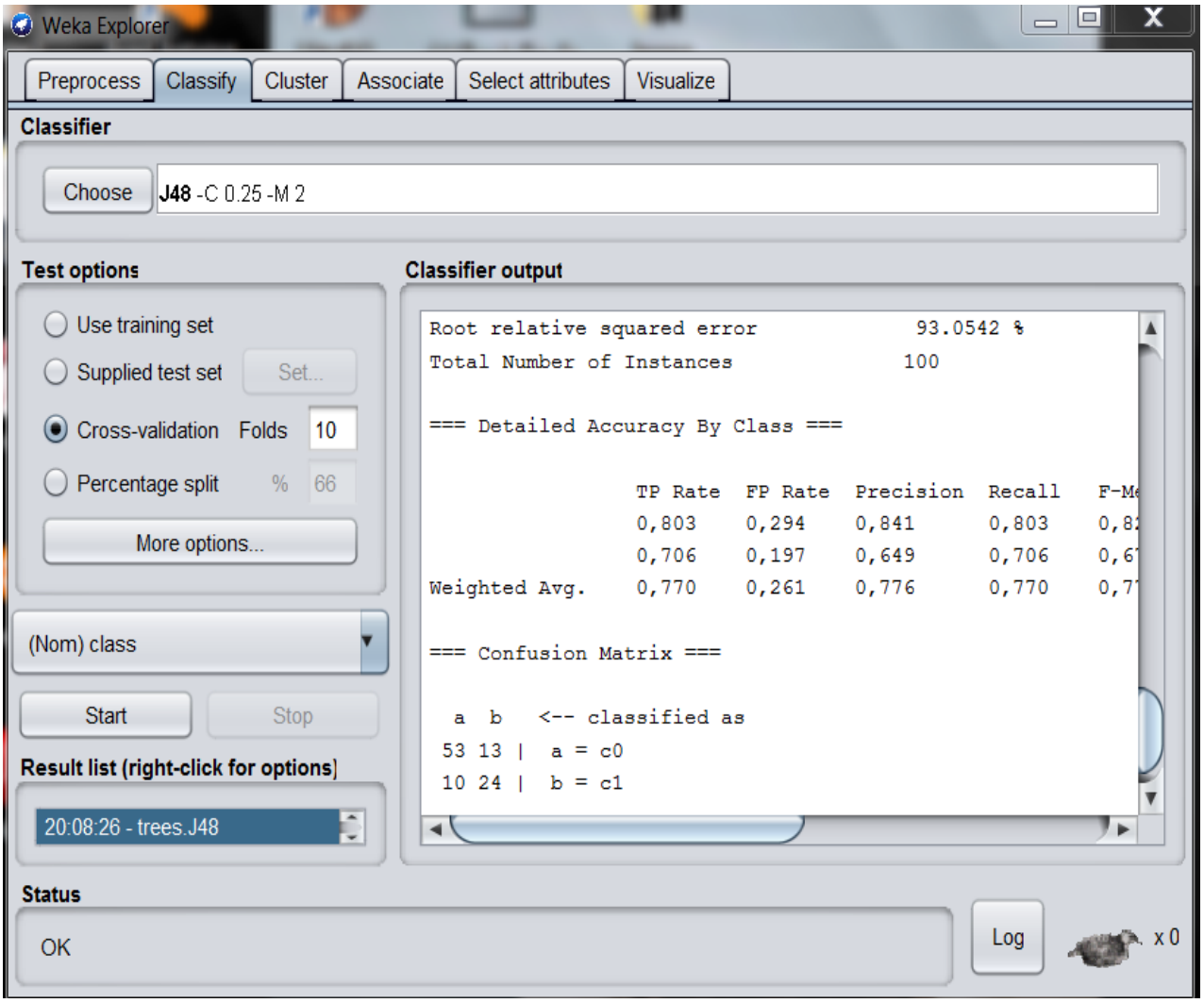

Figure 4.5 : Résultats de types j48 de classification produit par WEKA

L'onglet Classify permet à l'utilisateur d'estimer la précision du modèle prédictif, et de visualiser les prédictions erronées ou le modèle lui-même (si le modèle est sujet à Visualisation, comme un Arbre de décision).

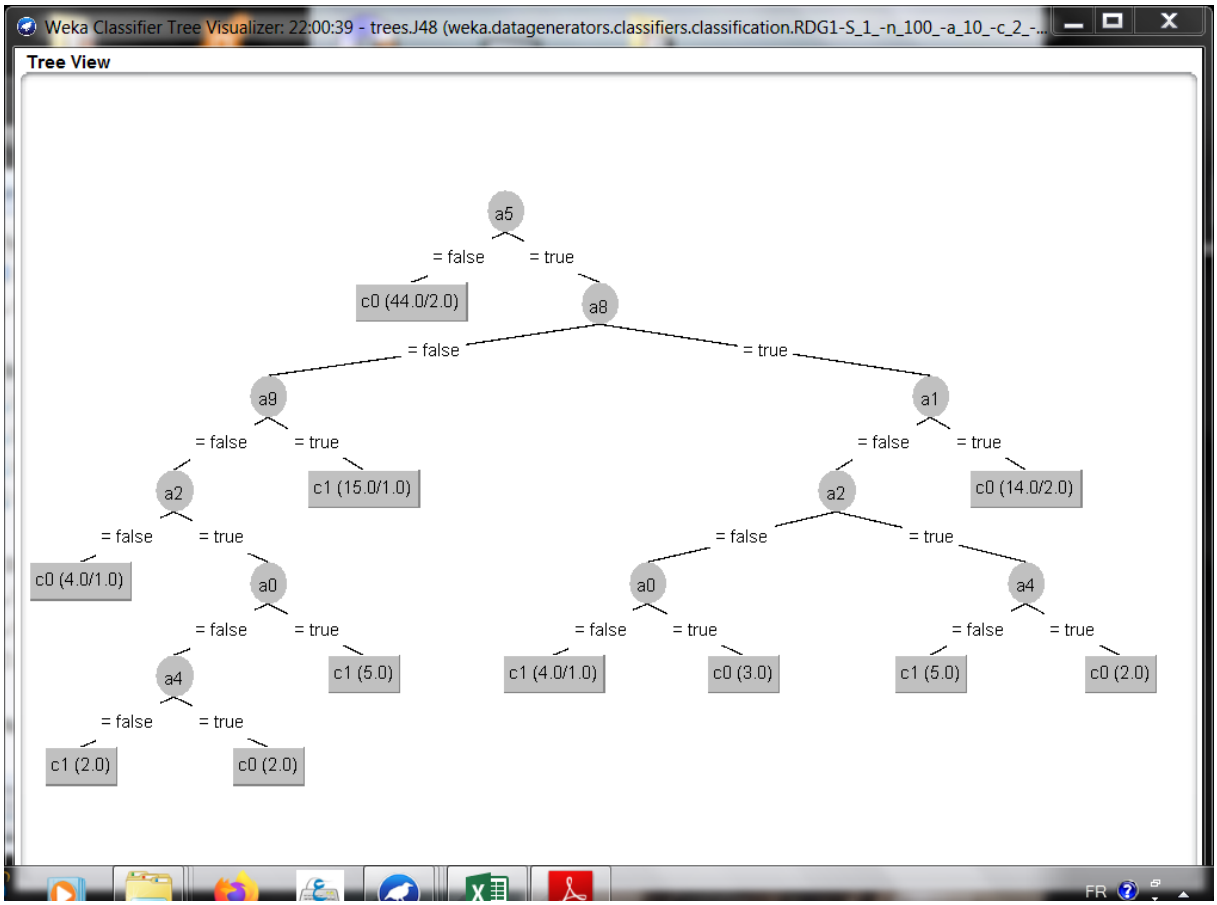

Figure 4.6 : Illustre les résultats de classification avec l'algorithme Arbre de décision.

4.4. L'application développée NetBeans

NetBeans (IDE 8.2). C'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 également de supporter différents autres langages,Comme C, C++, JavaScript, XML, Groovy, PHP et HTML de façon native ainsi que bien d'autres(Comme Python ou Ruby) par l'ajout de greffons. Il comprend toutes les caractéristiques d'un IDE Moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de Pages Web). Conçu en Java. NetBeans est disponible Sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X ou sousune version indépendante des systèmes d'exploitation (requérant une machine virtuelle

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

L'environnement de base comprend les fonctions générales suivantes :

- Configuration et gestion de l'interface graphique des utilisateurs .
- Support de différents langages de programmation.
- Traitement du code source (édition, navigation, formatage, inspection,…).
- Fonctions d'import/export depuis et vers d'autres IDE, tels qu'Eclipse ou JBuilder.
- Accès et gestion de bases de données, serveurs Web, ressources partagées.
- Gestion des tâches.
- Documentation intégrée.

La version utilisée dans notre projet est NetBeans IDE 8.2

La figure suivante illustre l'interface de l'environnement de développement NetBeans IDE 8.2

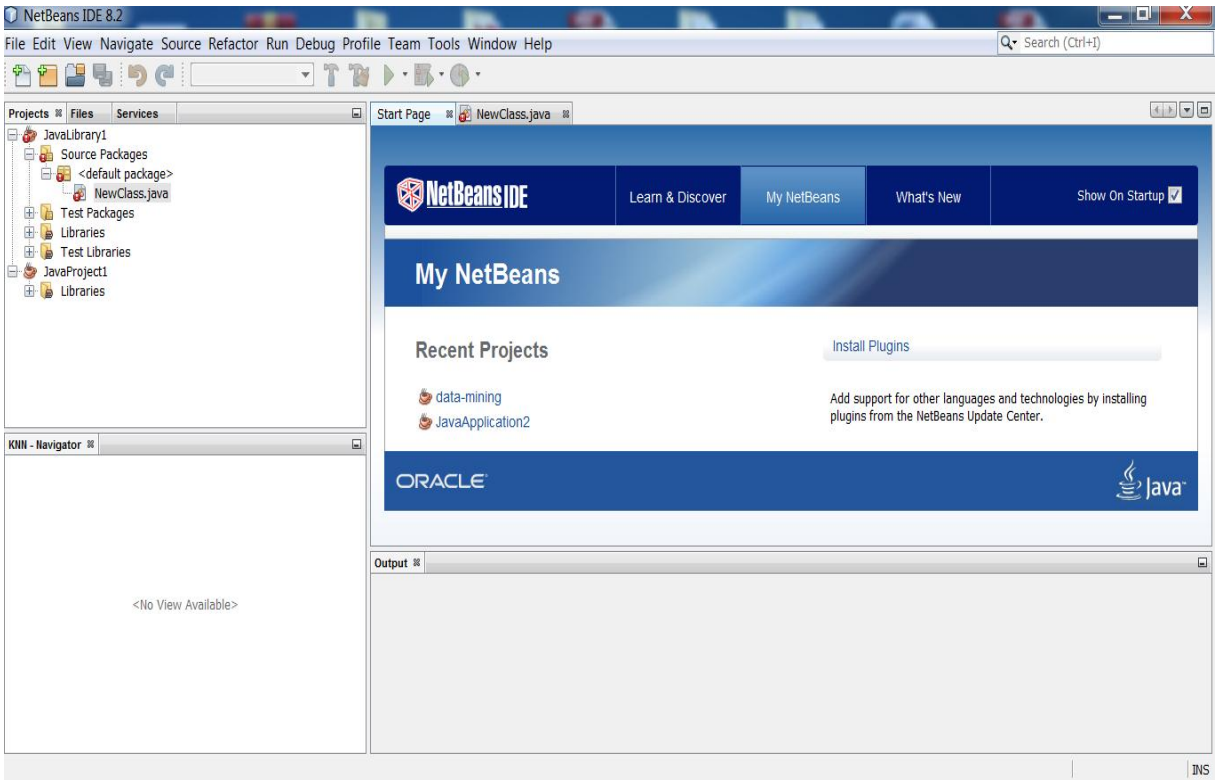

Figure 4.7: Interface NetBeans 8.2

4.5. Les interfaces du système

4.5.1 Les données expérimentales

Attribute Information:

a0 fièvre { oui, non }

a1 Température du patient {35C-42C}

a2 Apparition de nausées {oui, non}

a3 Douleur lombaire {oui, non}

a4 Poussée d'urine (besoin continu d'uriner) {oui, non}

a5 Douleurs mictionnelles {oui, non}

a6 Brûlure de l'urètre {oui, non}

a7 démangeaisons {oui, non}

a8 gonflement de la sortie de l'urètre {oui, non}

a9 Anorexie{ oui, non }

décision1 c0 : Inflammation de la vessie {oui, non}

décision2 c1 : Néphrite d'origine pelvienne { oui, non }

4.5.2 L'interface principale de système

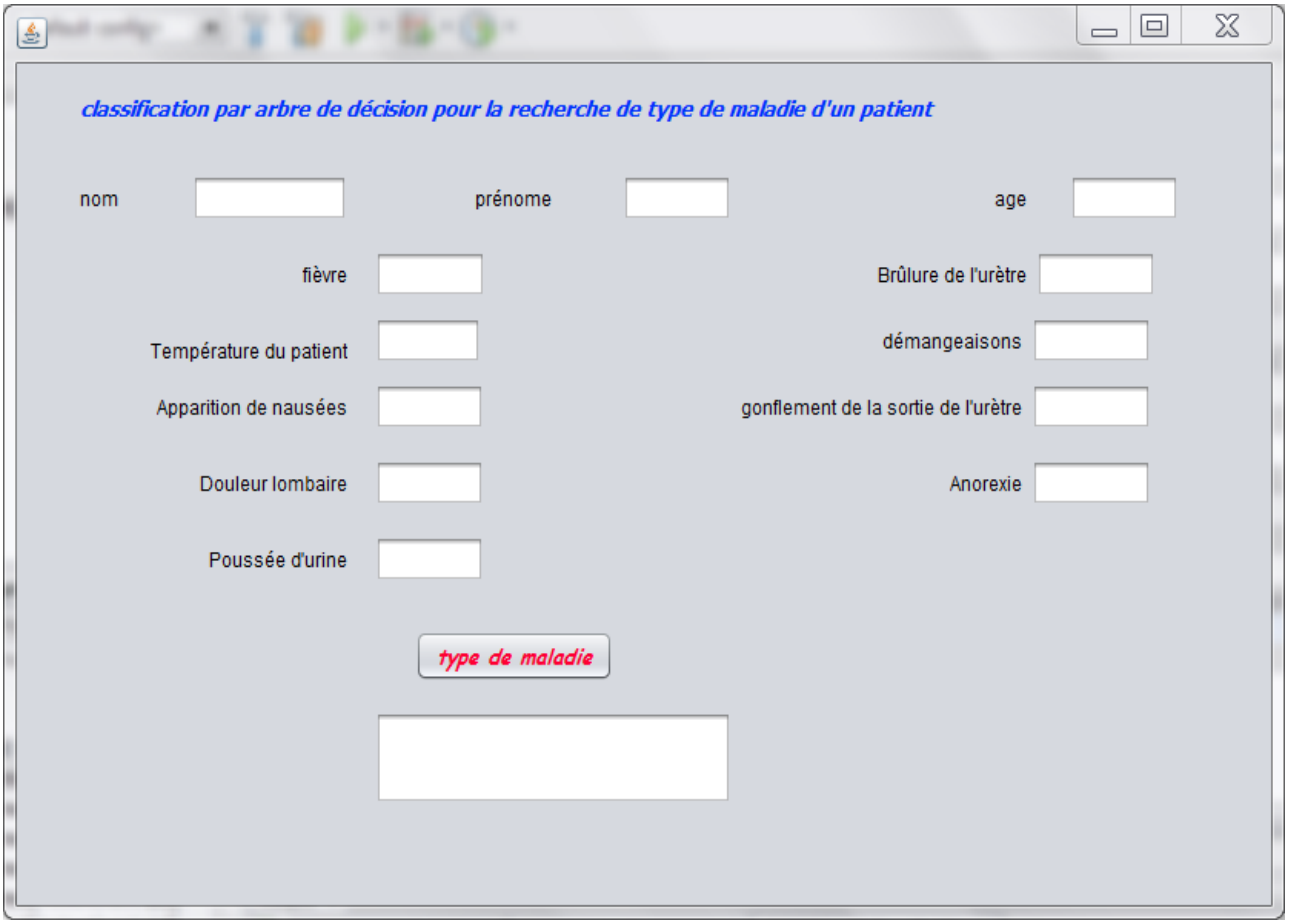

Figure 4.8 :L'interface principale de système.

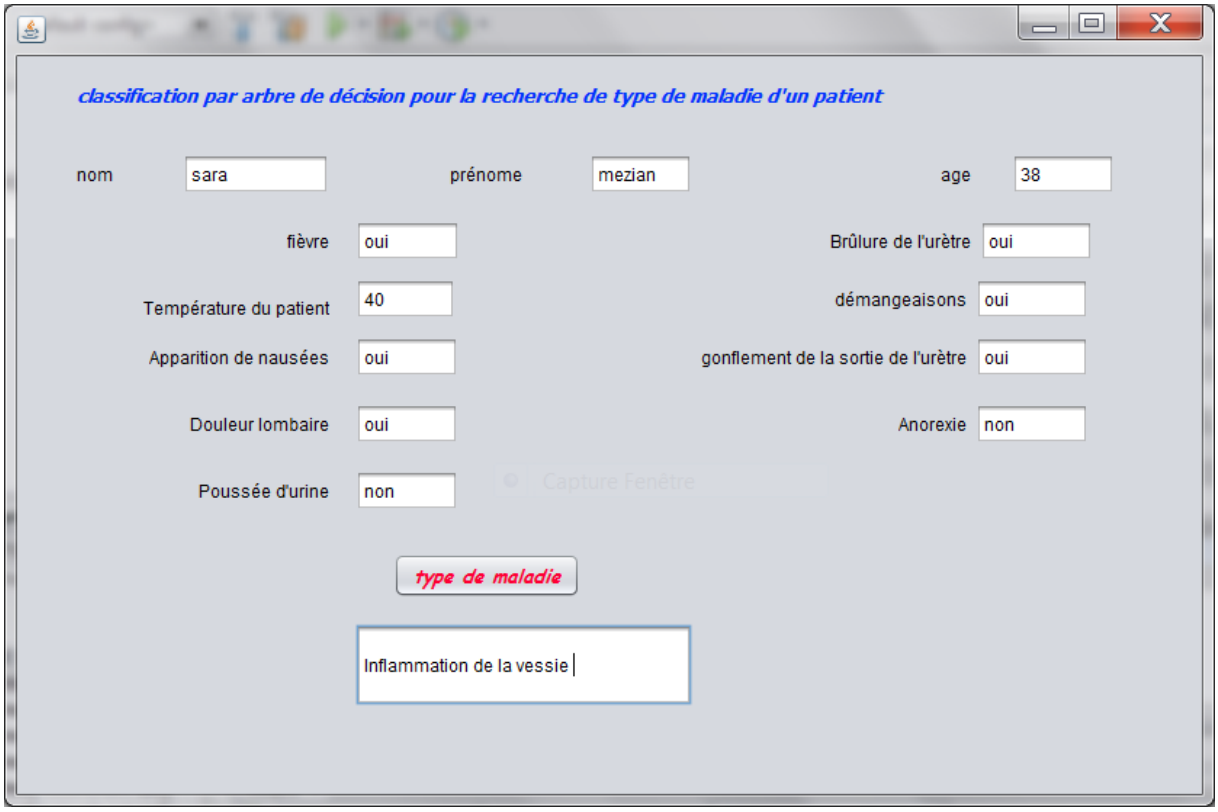

Figure 4.9 : Exemple d'un patient.

4.6. Conclusion

Les résultats de notre étude concluent que la prédiction les modèles peuvent être applicables pour, et le modèle J48 avait la meilleure précision de prédiction pour correctement classer les prédicateurs. Les précédentes tentatives de effectuera le diagnostic présomptif de deux maladies du système urinaire. Ils fournissent également des services centrés sur le patient approche vers des modèles nouveaux et cachés dans données, à partir desquelles les connaissances sont extraites généré. Cette connaissance peut aider à fournir des services médicaux et autres aux les patients. Les établissements de santé qui utilisent les données techniques minières ont la possibilité de prédire les exigences, les besoins, les désirs futurs, et l'état de santé des patients et de faire des décisions adéquates et optimales concernant leurs traitements. Les soins fondés sur des données probantes peuvent être mieux planifié et exécuté.

Conclusion générale

Au cours de notre mémoire nous avons répondu à la problématique concernant la classification des données par les arbres décision on utilisant la méthode J48, pour réaliser une application en nous aidant d'un environnement appelé Weka dédie à dataminig , qui nous a permis de créer notre propre modèle , et l'entrainer sur des données (Urinaire.CSV).

Dans ce mémoire nous avons constaté que les processus d'extraction des Connaissances a partir des données ont des spécificités qu'il faut prendre en compte, La première spécificité que nous avons constatée est que l'ECD est un processus et une suite d'étapes qu'il faut suivre commençant avec la sélection de données, la transformation, la création de modèle, l'évaluation de ce modèle pour extraire des nouvelles connaissances.

Nous avons constaté aussi que l' étape de création de modèle appelant « Fouille de données » est le coeur d' un processus ECD, c' est un ensemble des taches et techniques utilisées afin d'arriver aux nouvelles connaissances exploitables.

Nous sommes basées dans notre travail sur la tache de classification, cette dernière cherche à classifie une données dans une classe déjà connu, pour donner al fin le type de maladie urinaire.

Il existe plusieurs méthodes et techniques dédiées au classification par arbre de décision, dans notre travail on a choisi l'Algorithme J48 comme une méthode d'étude et pour réaliser notre système.

L'expérimentation effectuée nous a permis de confirmer que le Fouille de données peut toucher tous les domaines différents de vie, on a proposé un système dans le domaine médical qui peut donner le types de maladie d'un patient utilisant les techniques nécessaires.

Ce travail nous a permis d'approfondir nos connaissances sur les méthodes de classifications, et plus précisément sur les arbres de décision.

Nous espérons que ce mémoire éveillera une passion et un intérêt chez les promotions à venir et à toute personne intéressé par ce sujet afin qu'elle puisse persévérer dans ce domaine et améliorer ce modeste travail que nous avons entamé.

Bibliographies et références

[1] A. Ben Ali. *fusion et fouille de données par les conniassances. Thèse de doctorat.* http://thesis.univ-biskra.dz/5/1/fusion_et_fouille_de_donnees_guidees_par_les_connaissances.pdf

[2] M. Kantardzic M. *Data mining - concepts, models, methods,and algorithms*. IEEE Press ,Piscataway, NJ, USA, 2003.

[3] A. Djeffal. *Introduction au datamining.* https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjgwJjKqLH2AhVBtaQKHc6XAMsQFnoECEgQAQ&url=http%3A%2F%2Fabdelhamid-djeffal.net%2Fweb_documents%2Fintrofda 1516.pdf&usg=AOvVaw3wCLT-hx3Vhc-evlP-U9GW

[4] S. Ait Tahar. Les bases de données. Mémoire de licence. https://www.ummto.dz/dspace/bitstream/handle/ummto/12745/AitTaharSabrina.pdf?sequence=1

[5] El Moukhtar Zemmouri.*Représentation et gestion des connaissances dans un processus d'Extraction de Connaissances à partir de Données multi-points de vue* . https://tel.archives-ouvertes.fr/tel-00940780/document

[6] Benmammra El Djida. *Extraction et Gestion Parallèle et Distribuée des connaissances dans les Entrepôts de données. Mémoire de Magistère.* http://www.univ-bejaia.dz/xmlui/bitstream/handle/123456789/9534_/Extraction%20et%_20gestion%2_0par_al_l%c3%a8le%_____20et%20distribu__%c3%a9e%20des%20connaissances%20dans%20les%20entrep%c3%b4ts%20de%20donn%c3%a9 es .pdf? sequence=1&isAllowed=y

[7] B.AGARD, A.KUSIAK, *Exploration Des Bases De Données Industrielles À L'aide Du Data Mining perspectives.* https://docplayer.fr/storage/24/3001374/1656199002/U4RyS4u9TUhUF6-2rzWZog/3001374.pdf

[8] Outil de reporting décisionnel : DATA WAREHOUSE - Cube OLAP : Présentation conviviale des données chiffrées, Rapporté de http://www.sirom.fr/sycube.pdf

[9] E.Alpaydin, Introduction to Machine learning,2014.

[10] GAGAOUA Meriem. *Apprentissage et fouille de données par les algorithmes bio-inspirés : Application à la reconnaissance de caractères arabes manuscrits.* http://dspace.univ-setif.dz:8888/jspui/bitstream/123456789/2205/1/memoir%20gagaoua.pdf#page= 20&zoom=110,-68,527]

[11] L. C. Reddy and V. M, *"A Review on Data mining from Past to the Future," International Journal of Computer Applications.* https://www.researchgate.net/profile/Deepak-Bhardwaj/publication/304758913_rise_of_Data_Mining__Current_and_Future_Application_Areas/links/5779e0d008ae1b18a7e624f8/rise-of-Data-Mining-Current-and-Future-Application-Areas.pdf?origin=publication_detail

[12] J. R. Quinlan, *"Induction of decision trees,"* Machine Learning, vol. 1, pp. 81-106, 1986. https://link.springer.com/content/pdf/10.1007/BF00116251.pdf

[13] J. R. Quinlan, C4.5: Programs for Machine Learning: Morgan Kaufmann Publishers, 1993. https://books.google.dz/books?hl=fr&lr=&id=b3ujBQAAQBAJ&oi=fnd&pg=PP1&dq=J.+R.+Quinlan,+C4.5:+Programs+for+Machine+Learning:+Morgan+Kaufmann+Publishers,+1993.&ots=sR6sSSDtC8&sig=aPaagxG8_RTspMnaqUirauULSSM&redir_esc=y#v=onepage&q&f=false

[14] S. U. Amin, K. Agarwal, and R. Beg, *"Genetic neural network based data mining in prediction of heartdiseaseusingriskfactors,"* pp. 1227-1231, 2013. https://www.researchgate.net/profile/Syed-Amin-6/publication/261385302_Genetic__neural_network_based_data_mining_in_prediction_of_heart_disease_using_risk_factors/links/5a89838ca6fdcc6b1a4 23660/Genetic-neural-network-based-data-mining-in-prediction-of-heart-disease-using-risk-factors.pdf? origin=publication_detail

[15] LAHLOU Ouchiha;*Classification supervisee de documents étude comparative.* http://w4.uqo.ca/dii/etudMait.ise/uploads/73.pdf.

[16] https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-2.htm#pa4

[17] A. S. Al-Hegami, *"Classical and incremental classification in data mining process," International Journal of Computer Science and Network Security, vol. 7, 2007.*