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 Avancées et Applications (G.L.A.A)

Thème

UTILISATION DE LA CLASSIFICATION POUR LA DETERMINATION DES PROFILS DES MALADES

Réalisé par :

- BOULGHAB Kenza
- AMEUR Zaimeche Ahlem

Encadré par : A. MANSOUL

Année Universitaire 2021-2022

Résumé

Le data mining, ou La fouille de données, constitue le cœur d'un processus d'extraction de connaissances à partir d'un large volume de données. Son champ d'applications est très vaste.

Dans le présent travail nous exposons un modèle de prédiction pour la recherche d'une donnée ou valeur et qui est en fait une valeur inconnue au départ et dont nous voulons prédire.

Pour atteindre cet objectif nous proposons un système qui va s'articuler autour de trois modules

Dans un premier temps nous employons la technique de la classification Pour une analyse de données qui extrait des modèles décrivant avec précision les catégories et les catégories de données importantes, C'est le modèle de connaissance que nous allons analyser. Pour ce faire, nous proposons l'utilisation de la méthode K-NN sous un environnement appelé WEKA destiné à la fouille de données.

Dans un deuxième temps nous utilisons un module que nous avons développé afin de faire la prédiction à partir du modèle construit par classification.

Dans une étape finale nous expérimentons notre approche sur des données se rapportant aux.

Malade de la colonne vertébrale. Dans ce type de maladie, le type de maladie est souvent reconnu par ses symptômes Enfin, nous évaluerons notre approche.

Le travail que nous présentons dans cette thèse est très intéressant, notamment dans la recherche des symptômes qui apparaissent sur le patient. Cela contribuera à l'identification du type de maladie.

Mots clés :

Extraction de connaissances, Fouille de données, Classification, Prédiction, Maladie.

Remercîment

En premier lieu, je remercie Dieu de m'avoir permis de mener ce travail.

Ce travail n'aurait pu aboutir sans la contribution d'un nombre de personnes, ainsi se présente l'occasion de les remercie.

Je tiens à remercier mon encadreur, Monsieur MANSOUL Abdelhak, qui a supervisé mon travail tout en me laissant une grande marge de liberté. Je le remercie pour son encadrement, sa disponibilité et la pertinence de ses remarques tout au long de la réalisation de ce projet.

Merci à mes parents, pour ses patience, ses conseils qui ont éclairé mon chemin, et soutenu tout au long de ma vie. À toi maman, tu m'as toujours poussé vers le sérieux et le travail, et maintenant c'est grâce à toi et pour toi que j'arrive là.

Ce mémoire n'aurait pas vu le jour sans la contribution de mon fiancé, Mohamed, qui été toujours à côté de moi dans les moments délicats. Je ne saurai assez le remercier, pour son soutien moral et sa présence. C'est grâce à ton aide et à ta patience avec moi que ce travail a pu voir le jour.

Je tiens à remercier toute ma famille, qui a toujours trouvé les mots pour m'encourager.

Ma plus profonde reconnaissance pour votre soutien.

Remercîment

Je remercie Dieu de m'avoir accordé le succès dans mon parcours académique
Et mes remerciements s'adressent à mon encadrante, Monsieur MANSOUL Abdelhak, maîtresse de conférences à l'Université 20 out Skikda, pour avoir accepté de diriger ce travail. Son soutien, ses compétences et sa clairvoyance m'ont été d'une aide inestimable.

Je tiens à remercier mes parents m'ont toujours encouragé et soutenu.

Mes remerciements les plus chaleureux vont à tous mes camarades

Table de matière

Introduction générale

Les systèmes d'information deviennent de plus en plus complexes et diversifiés en raison notamment de l'émergence de nouvelles technologies. L'accroissement continu de la volumétrie des données numériques ainsi que la multiplicité des sources de données de plus en plus hétérogènes, conjugués aux besoins pressants des entreprises à exploiter ces données dans un processus d'aide à la prise de décisions ont fait émergé de nouvelles problématiques que les technologies émergentes d'extraction des connaissances à partir des données et d'entreposage de données continuent à étudier et à leur chercher des solutions.

Ceci nécessite la définition de nouvelles approches pour les architectures, l'intégration, la modélisation, l'interrogation, l'optimisation L'Extraction des Connaissances à partir des Données (ECD) apparue dans la communauté de l'intelligence artificielle, a pour but l'identification de structures inconnues, valides, et potentiellement exploitables dans les bases de données.

L'ECD propose un cadre général dans lequel sont regroupées les méthodes qui permettent de faire face aux problèmes d'organisation des données et de leur exploitation, plus particulièrement : l'entreposage et la fouille des données. L'entreposage des données a pour objet d'organiser de très grands volumes de données, de les structurer et les préparer à l'analyse. Il est centré sur le processus d'Extraction, Transformation et Chargement de données (ETC).

Les données sont généralement stockées dans des bases des données spécialisées dites : Entrepôts des Données (ED). La Fouille des Données (FD) a pour but d'extraire des connaissances à partir des données par des méthodes de structuration (apprentissage non supervisé) ou par des méthodes explicatives (apprentissage supervisé), une fois les données acquises et préparées. Comme la FD est étroitement liée au processus d'ECD, la plupart des travaux de recherche utilisent ces deux termes de manière interchangeable.

L'objectif de notre étude est de faire présenter tous le processus d'un ECD : à partir de la collection des donnés (des donnés sur des verres) obtenu d'une base de donné et analysées par un environnement d'apprentissage WEKA, ce dernier permet de nous donner le modèle souhaité, jusqu'à l'étape d'extraire des connaissances (Classifier des verres) utilisant une interface graphiques

Notre mémoire est divisé en quatre chapitres, nous décrivons brièvement ici le contenu de chacun d'eux:

- le Chapitre 01, contient des généralités sur l'ECD.

- Chapitre 02, est consacré à la présentation de la tache classification et son processus avec ses différents algorithmes et les techniques d'évaluation des classifier.

- Chapitre 03 : est consacré à la présentation de l'architecture générale de l'approche avec sa modélisation

- Le chapitre 04, sera consacré à un exposé des différentes parties du processus expérimental que nous avons réalisé pour valider notre approche et nous présentons la plate-forme expérimentale réalisée en plus les essais et les résultats du système mis en œuvre sans oublier d'exposer les interfaces de notre système.

Enfin, nous terminons ce mémoire par une conclusion générale, qui récapitule les travaux réalisés, et ferrons le point sur un ensemble de perspectives envisagées.

1.1. INTRODUCTION

Ce premier chapitre présente les concepts de fouille de données, où les différentes étapes du processus d'extraction de connaissances à partir des données sont décrites. Nous insistons sur les différentes approches de mise en œuvre d'un modèle de fouille de données.

1.2. L'EXTRACTION DE CONNAISSANCES A PARTR DE DONNEES

L'ECD Est un processus pour la découverte de nouvelles connaissances sur un domaine d'application donné. Il consiste en de nombreuses étapes, dont la plus importante est la fouille de données (data mining). Chaque étape du processus vise l'achèvement d'une tâche particulière, et est réalisée par l'application d'une ou plusieurs méthodes particulières.

L'ECD est également défini par Fayad comme étant « un processus non trivial qui permet d'identifier, dans des données, des patterns ultimement compréhensibles, valides, nouveaux et potentiellement utiles ». Cette définition est la plus répandue dans la communauté Extraction et Gestion des Connaissances. [1]

Il est Inventée au premier workshop de KDD en 1989, l'expression « Extraction de Connaissances à partir de Données » désigne l'ensemble de « processus non trivial d'identification des modèles valides, nouveaux, potentiellement utiles et finalement compréhensibles à partir des données d'une grande base de données. En effet, les données représentent un ensemble de faits et le modèle l'expression, exprimée dans un langage de description d'un ensemble de données, qui décrit les corrélations entre ces données

L'extraction de connaissances à **partir** des **données** (ECD) se définit comme « l'acquisition de **connaissances** nouvelles, intelligibles et potentiellement utiles à **partir** de faits cachés au sein de grandes quantités de **données** »En fait, on cherche surtout à isoler des traits structuraux (*patterns*) qui soient valides, non triviaux, nouveaux, utilisables et si possible compréhensibles ou explicables Un processus d'ECD est constitué de quatre phases qui sont :

1. le nettoyage et intégration des données,
2. le prétraitement des données,
3. la fouille de données
4. l'évaluation et la présentation des connaissances [2].

La figure 1 récapitule ces différentes phases ainsi que les enchaînements possibles entre ces phases.

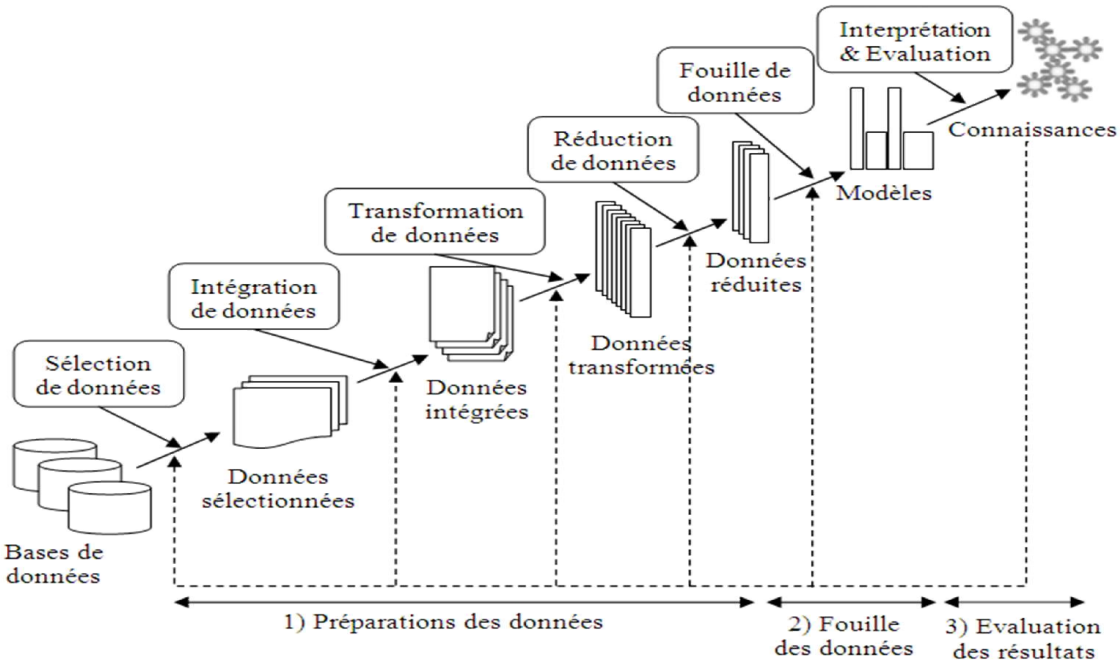


Figure 1.1: Processus d'extraction de connaissances à partir des données [3]

1.3.ETAPES DU PROCESSUS D'EXTRACTION DE CONNAISSANCES A PARTIR DE DONNEES

3.1.Nettoyage et intégration des données

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.

L'intégration est la combinaison des données provenant de plusieurs sources (base 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 leurs exploitations futures.

3.2.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.

3.3.Fouille de données (Data Mining)

La fouille de données (data mining en anglais), est le cœur du processus d'ECD. Il s'agit à ce niveau de trouver des pépites de connaissances à partir des données. Tout le travail consiste à appliquer des méthodes intelligentes dans le but d'extraire cette connaissance. Il est possible de définir la qualité d'un modèle en fonction de critères comme les performances obtenus, la fiabilité, la compréhensibilité, la rapidité de construction et d'utilisation et enfin l'évolutivité. Tout le problème de la fouille de données réside dans le choix de la méthode adéquate à un problème donné. Il est possible de combiner plusieurs méthodes pour essayer d'obtenir une solution optimale globale. Nous ne détaillerons pas d'avantage la fouille de données dans ce paragraphe car elle fera l'objet 'une section complète.

3.4.Evaluation et 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. Il y a principalement deux techniques de validation qui sont la technique de validation statistique et la technique de validation par expertise.

La validation statistique consiste à utiliser des méthodes de base de statistique descriptive. L'objectif est d'obtenir des informations qui permettront de juger le résultat obtenu, ou d'estimer la qualité ou les biais des données d'apprentissage. Cette validation peut être obtenue par :

- le calcul des moyennes et variances des attributs,
- si possible, le calcul de la corrélation entre certains champs,
- ou la détermination de la classe majoritaire dans le cas de la classification.

La validation par expertise est réalisée par un expert du domaine qui jugera la pertinence des résultats produits. Par exemple pour la recherche des règles d'association, c'est l'expert du domaine qui jugera la pertinence des règles.

Pour certains domaines d'application (le diagnostic médical, par exemple), le modèle présenté doit être compréhensible. Une première validation doit être effectuée par un expert qui juge la compréhensibilité du modèle. Cette validation peut être, éventuellement, accompagnée par une technique statistique [4].

Grâce aux techniques d'extraction de connaissances, les bases de données volumineuses sont devenues des sources riches et fiables pour la génération et la validation de connaissances. La fouille de données n'est qu'une phase du processus d'ECD, et consiste à appliquer des algorithmes d'apprentissage sur les données afin d'en extraire des modèles (motifs). L'extraction de connaissances à partir des données se situe à l'intersection de nombreuses discipline, comme l'apprentissage automatique, la reconnaissance de formes, les bases de données, les statistiques, la représentation des connaissances, l'intelligence artificielle, les systèmes experts, etc. (Figure 1.2) [5].

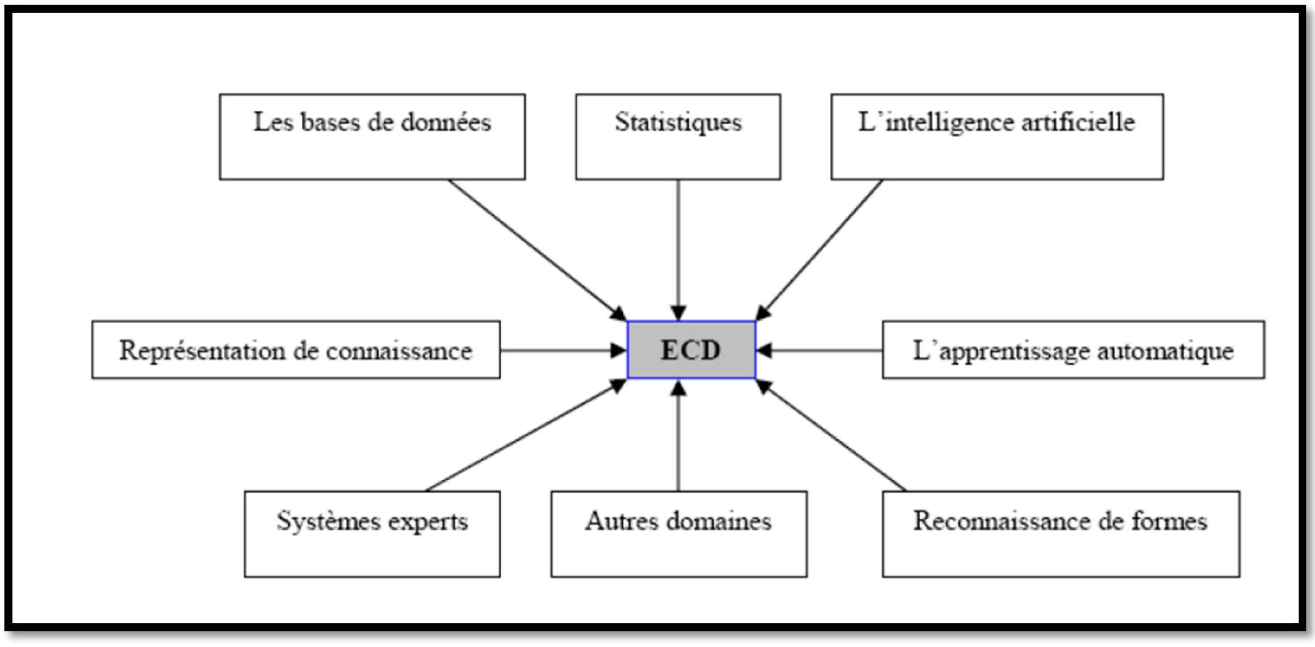


Figure 1.2 : L'Extraction de connaissances à partir des données à la confluence de nombreux domaines [5].

Les concepts de fouille de données et d'extraction de connaissances à partir de données sont parfois confondus et considérés comme synonymes. Mais, formellement on considère la fouille de

données comme une étape centrale du processus d'extraction de connaissances des bases de données.

1.4.LA FOUILLE DE DONNEES

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 été 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 son 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 datamining [6].

L'expression "data mining" est apparue vers le début des années 1960 et avait, à cette époque, un sens péjoratif. En effet, les ordinateurs étaient de plus en plus utilisés pour toutes sortes de calculs qu'il n'était pas envisageable d'effectuer manuellement jusque-là. Certains chercheurs ont commencé à traiter sans a priori statistique les tableaux de données relatifs à des enquêtes ou des expériences dont ils disposaient. Comme ils constataient que les résultats obtenus, loin d'être aberrants, étaient tout au contraire prometteurs, ils furent incites à systématiser cette approche opportuniste. Les statisticiens officiels considéraient toutefois cette démarche comme peu scientifique et utilisèrent alors les termes " data mining " ou "data mining" pour les critiquer.

Cette attitude opportuniste face aux données coïncida avec la diffusion dans le grand public de l'analyse de données dont les promoteurs, comme Jean-Paul Benzecri [7], ont également dû subir dans les premiers temps les critiques venant des membres de la communauté des statisticiens.

Le succès de cette démarche empirique ne s'est pas démenti malgré tout. L'analyse des données s'est développée et son intérêt grandissait en même temps que la taille des bases de données. Vers la fin des années 1980, des chercheurs en base de données, tel que Rakesh Agrawal, ont commencé à travailler sur l'exploitation du contenu des bases de données volumineuses comme par exemple celles des tickets de caisses de grandes surfaces, convaincus de pouvoir valoriser ces masses de données dormantes. Ils utilisèrent l'expression "data base mining" mais, celle-ci étant déjà déposée par une entreprise (Data base mining Workstation), ce fut "data mining" qui s'imposa. En mars 1989, Shapiro Piat et ski proposa le terme "knowledge discovery" à l'occasion d'un atelier

sur la découverte des connaissances dans les bases de données .Actuellement, les termes data mining et knowledge discovery in data bases (KDD, ou ECD en français) sont utilisés plus ou moins indifféremment. Nous emploierons par conséquent l'expression "data mining", celle-ci étant la plus fréquemment employée dans la littérature.

La communauté de "data mining " a initié sa première conférence en 1995 à la suite de nombreux ateliers (workshops) sur le KDD entre 1989 et 1994. La première revue du domaine " Data mining and knowledge discovery journal " publiée par "Kluwers " a été lancée en 1997.

« 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» [8].

La définition la plus communément admise de Data Mining est celle de [9] : «Le Data mining est un processus non trivial qui consiste à identifier, dans des données, des schémas nouveaux, valides, potentiellement utiles et surtout compréhensibles et utilisables».

En bref, le data mining est l'art d'extraire des informations (ou même des connaissances) à partir des données [10].

1.5. PRINCIPALES TACHES DE FOUILLE DE DONNEES

On dispose de données structurées. Les objets sont représentés par des enregistrements (ou descriptions) qui sont constitués d'un ensemble de champs (ou attributs) prenant leurs valeurs dans un domaine. De nombreuses tâches peuvent être associées au Data Mining, parmi elles nous pouvons citer:

5.1.La classification

La classification est la tâche 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 [11]

5.2.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.

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

- Estimer le nombre d'enfants dans une famille.
- Estimant le montant d'argent qu'une famille de quatre membres choisis aléatoirement dépensera pour la rentrée scolaire.
- Estimer la valeur d'une pièce immobilière.

Souvent, la classification et l'estimation sont utilisées ensemble, comme quand le Data Mining est utilisée pour prévoir qui va probablement répondre à une offre de transfert d'équilibre(de solde) de carte de crédit et aussi évaluer la taille de l'équilibre(du solde) à être transféré

5.3.La prédiction

La prédiction est la même que la classification et l'estimation, à part que dans la prédiction les enregistrements sont classés suivant des critères (ou des valeurs) prédites (estimées). La principale raison qui différencie la prédiction de la classification et l'estimation est que dans la création du modèle prédictif on prend en charge la relation temporelle entre les variables d'entrée et les variables de sortie.

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

- Prévoir le prix des actions dans les trois prochains mois.
- Prévoir le champion de la coupe du monde en football en se basant sur la comparaison des statistiques des équipes.
- Prévoir quels clients va déménager dans les 6 mois qui suivent.

5.4. La segmentation (Clustering).

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 différents à ceux existants sur les autres clusters. La différence 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 dessous groupes relativement homogènes. Ils maximisent l'homogénéité à l'intérieur de chaque groupe et la minimisent entre ces derniers.

Les algorithmes du clustering peuvent être appliqués dans des différents domaines, tel que :

- Découvrir des groupes de clients ayants des comportements semblables.
- Classification des plantes et des animaux étant donné leurs caractéristiques.
- Segmentation les observations des épicentres pour identifier les zones dangereuses.

5.5.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 produit et les processus présents sue cette base.

Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. 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 politiques [12].

Cette tâche, plus connue comme l'analyse du panier de la ménagère, consiste à déterminer les variables qui sont associées. L'exemple type est la détermination des articles (le pain et le lait, la

tomate, les carottes et les oignons) qui se retrouvent ensemble sur un même ticket de supermarché. Cette tâche peut être effectuée pour identifier des opportunités de vente croisée et concevoir des groupements attractifs de produit.

5.6.Les règles d'association.

Les règles d'association sont traditionnellement liées au secteur de la distribution car leur principale application est "l'analyse du panier de la ménagère (mark et basket analyses)" qui consiste en la recherche d'associations entre produits sur les tickets de caisse. Le but de la méthode est l'étude de ce que les clients achètent pour obtenir des informations sur "qui" sont les clients et "pourquoi" ils font certains achats. La méthode peut être appliquée à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services:

Services bancaires, services de télécommunications, par exemple. Elle peut être également utilisée dans le secteur médical pour la recherche de complications dues à des associations de médicaments ou à la recherche de fraudes en recherchant des associations inhabituelles.

Un attrait principal de la méthode est la clarté des résultats produits. En effet, le résultat de la méthode est un ensemble de règles d'association.

Cette tâche, plus connue comme l'analyse du panier de la ménagère, consiste à déterminer les variables qui sont associées. L'exemple type est la détermination des articles (le pain et le lait, la tomate, les carottes et les oignons) qui se retrouvent ensemble sur un même ticket de supermarché. Cette tâche peut être effectuée pour identifier des opportunités de vente croisée et concevoir des groupements attractifs de produit.

1.6.LES METHODES DE FOUILLE DE DONNEES

Pour tout jeu de données et un problème spécifique, il existe plusieurs méthodes que l'on choisira en fonction de :

- la tâche à résoudre,
- la nature et de la disponibilité des données,
- l'ensemble des connaissances et des compétences disponibles,
- la finalité du modèle construit,
- l'environnement social, technique, philosophique de l'entreprise,
- etc.

Quelque méthode a été élaborée pour résoudre des tâches bien définies. On trouve:

6.1.Les Plus Proches Voisins.

La méthode des plus proches voisins (PPV en bref, nearest neighbor en anglais) est une méthode dédiée à la classification qui peut être étendue à des tâches d'estimation. La méthode PPV est une méthode de raisonnement à partir de cas. Elle part de l'idée de prendre des décisions en recherchant un ou des cas similaires déjà résolus en mémoire.

6.2.Les Arbres de Décision.

Les arbres de décision est l'une des plus intuitives et des plus populaires du data mining, d'autant plus qu'elle fournit des règles explicites de classement et supporte bien les données hétérogènes, manquantes et les effets non linéaires. Pour les applications relevant du marketing de bases de données, actuellement la seule grande concurrente de l'arbre de décision est la régression logistique, cette méthode étant préférée dans la prédiction du risque en raison de sa plus grande robustesse. Remarquons que les arbres de décision sont à la frontière entre les méthodes prédictives et descriptives, puisque leur classement s'opère en segmentant la population à laquelle ils s'appliquent : ils ressortissent donc à la catégorie des classifications hiérarchiques descendantes supervisées.

6.3.Les Réseaux de Neurones.

Les réseaux de neurones sont apparus dans les années cinquante avec les premiers perceptrons, et sont utilisés industriellement depuis les années quatre-vingt. Un réseau de neurone "ou réseau neuronal" a une architecture calquée sur celle du cerveau, organisée en neurones et synapses, et se présente comme un ensemble de nœuds "ou neurones formels, ou unités" connectés entre eux, chaque variable prédictive continue correspondant à un nœud d'un premier niveau, appelé couche d'entrée, et chaque variable prédictive catégorique (ou chaque modalité d'une variable catégorique) correspondant également à un nœud de la couche d'entrée.

Le cas échéant, lorsque le réseau est utilisé dans une technique prédictive, il y a une ou plusieurs variables à expliquer ; elle correspondant alors chacune à un nœud (ou plusieurs dans le cas des variables catégorielles) d'un dernier niveau : la couche sortie. Les réseaux prédictifs sont dits "à apprentissage supervisé" et les réseaux descriptifs sont dits "à apprentissage non supervisé". Entre la couche d'entrée et la couche sortie sont parfois connectés à des nœuds appartenant à un niveau intermédiaire : la couche cachée. Il peut exister plusieurs couches cachées [13].

1.7. CONCLUSION

Dans ce premier chapitre, nous avons présenté les principaux concepts de fouille de données, les processus, les tâches et les méthodes les plus utilisés en data mining.

2.1.INTRODUCTION

La classification est le processus permettant d'identifier à quelle catégorie appartient une nouvelle observation à partir d'un ensemble de données d'apprentissage contenant des observations dont l'appartenance à une catégorie est connue.

La classification est la tâche la plus commune du Data Mining et qui semble être une obligation humaine. Afin de comprendre notre vie quotidienne, nous sommes constamment classifiés, catégorise et évalués.

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'un bas 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'exemple classés auparavant.

L'objectif est créer un modèle qui peut être appliqué aux données non classifiées dans le but de les classifiées, quelques exemples de l'utilisation des taches de classification dans les domaines de recherche et commerce sont les suivants :

- Déterminer si l'utilisation d'une carte de crédit est frauduleuse.
- Diagnostiquant si une certaine maladie est présente.
- Déterminer quels numéros de téléphone correspondent aux fax.
- Déterminer quelles lignes téléphoniques sont utilisées pour l'accès à internet [14].

2.2.LE PROCESSUS DE CLASSIFICATION

La classification est un processus à deux étapes : une étape de **construction du modèle** et une étape **d'utilisation du modèle**.

Dans construction du modèle : description d'un ensemble de classes prédéterminées

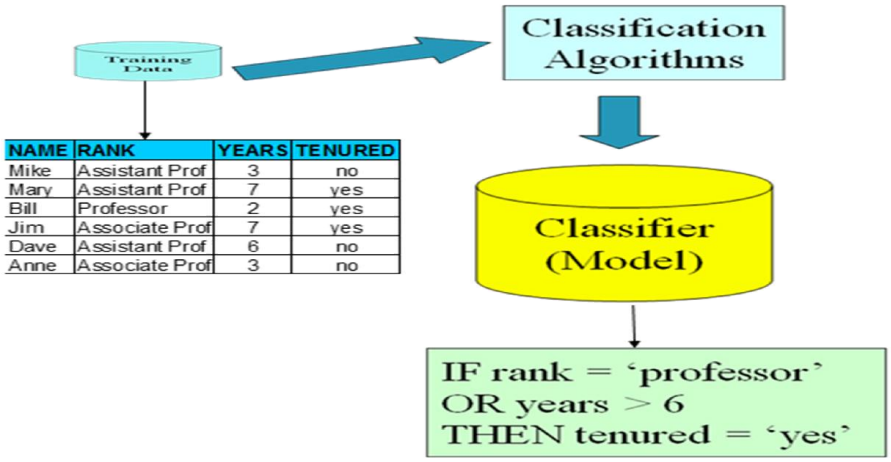


Figure 2.1 : construction de modèle [15]

- Chaque tulpe/échantillon est supposé appartenir à un classe prédéfinie, tel que détermine par l'attribue d'étiquette de classe.

- L'ensemble des tulples utilisés pour la construction du modèle : ensemble d'entrainement.

- Le modèle est représenté sous forme de règles de classification, d'arbre de décision, ou des formules mathématiques.

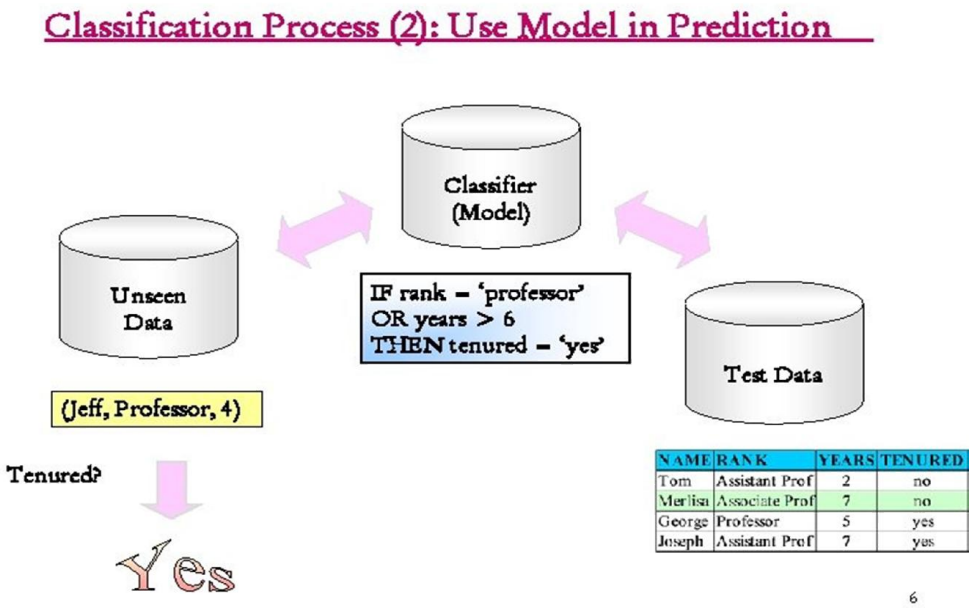


Figure 2.2 : utilisation de modèle [15]

- Estimer la précision du modèle.

- L'étiquette connue de l'échantillon à tester est comparée avec résultat classifié du modèle.

- Le taux de précision est le pourcentage d'échantillons de test qui sont correctement classé par le modèle.

L'ensemble de teste est indépendant de l'ensemble d'apprentissage, sinon sur-ajustement se produira [15].

2.3.AVANTAGES DE LA CLASSIFICATION

Avantages

- Inventorier vos données : Il faut tout d'abord réaliser un inventaire exhaustif de vos données. La découverte est essentielle dans cette phase d'identification car les données pourraient contenir des informations sensibles rangées au même niveau que des informations obsolètes.
- Respecter les normes : Les différentes réglementations selon votre secteur d'activité obligent votre entreprise à protéger les données spécifiques ; comme les informations de santé (HIPAA), les données financières (SOX) ou encore les données personnelles des résidents de l'UE (RGPD). La découverte et la classification des données vous aident à déterminer où se trouvent ces catégories de données afin de vous assurer que les contrôles de sécurité appropriés seront mis en œuvre pour respecter les exigences légales dans ce domaine.
- Sécuriser les données confidentielles : Pour protéger convenablement les données sensibles, vous devez être en mesure de savoir [16].

2.4.EVALUATION ET VALIDATION DES CLASSES (LE MODEL)

L'apprentissage d'un modèle de décision se fait à base de plusieurs paramètres, à savoir les exemples d'entrainement et leur nombre, les paramètres de la méthode utilisée, ...etc.

Le choix des valeurs de ces paramètres se fait à travers plusieurs essais et évaluation pour atteindre des performances satisfaisantes du modèle. Les paramètres optimaux pour un modèle donné sont les paramètres qui lui permettent de donner une précision de 100%.

Cette situation serait idéale si l'ensemble des exemples représentait parfaitement l'ensemble de tous les exemples possibles. Le modèle appris peut donner une très grande précision face aux exemples d'entrainement, mais se comporte très mal avec les nouveaux exemples.

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.

La métrique intuitive utilisée est la précision du modèle appelée aussi le taux de reconnaissance. Elle représente le rapport entre le nombre de donnée correctement classées et le nombre total des données testées. L'équation suivante donne la formule utilisée :

$$P=\frac{1}{N}\sum_{i=1}^{n} L(yi, \hat{f}\ (xi))$$

Avec :

$$L=\begin{cases}1, & siyi = \hat{f}(xi)\\ 0, & sinon\end{cases}$$

Généralement, la précision est donnée sous forme de pourcentage ce qui nécessite de multiplier la précision de l'équation précédente par 100.

(a) Matrice de confusion

La mesure précédente donne le taux d'erreurs commises par le modèle appris (100−précision) mais ne donne aucune information sur la nature de ces erreurs. Dans la plus part des cas d'application, il est très important de connaître la nature des erreurs commises :

Quelle classe est considérée comme quelle classe par le modèle ? Par exemple dans un modèle appris pour des objectifs médicaux, considérer un échantillon non cancéreux alors qu'il l'est, est beaucoup plus grave de considérer un échantillon cancéreux alors qu'il ne l'est pas. Dans le cas de classification binaire, le résultat de test d'un modèle peut être une possibilité parmi quatre :

$$\begin{cases}\hat{f}(x_i) = +1 \text{ et } y_i = +1 \text{ correcte positive}\\ \hat{f}(x_i) = +1 \text{ et } y_i = -1 \text{ fausse positive}\\ \hat{f}(x_i) = -1 \text{ et } y_i = -1 \text{ correcte négative}\\ \hat{f}(x_i) = -1 \text{ et } y_i = +1 \text{ fausse négative}\end{cases}$$

Si le modèle donne une classe positive pour un exemple d'une classe positive, on dit que c'est une classe correcte positive (CP). Si par contre l'exemple appartient à la classe négative on dit que c'est une classe fausse positive (FP). Si le modèle donne une classe négative pour un exemple d'une classe négative, le résultat est une classe correcte négative (CN), si, par contre, la classe de l'exemple est positive le résultat est qualifié de classe fausse négative(FN).

La matrice de confusion est une matrice qui rassemble en lignes les observations (y) et en colonnes les prédictions $\hat{f}(x)$. Les éléments de la matrice représentent le nombre d'exemples correspondants à chaque cas.

Observations (y)	Predictions ($\hat{f}$)	
	+1	-1
+1	CP	FN
-1	FP	CN

Table 2.1–Matrice de confusion pour la classification binaire [12].

Un modèle sans erreurs aura ses résultats rassemblés sur la diagonale de sa matrice de confusion (CP et CN). Dans le cas multi classes la matrice sera plus large avec les classes possibles au lieu des deux classes +1 et -1. La précision P du modèle peut être calculée à partir de la matrice de confusion comme suit:

$$P=\frac{CP+CN}{CP+FP+CN+FN}$$

Deux autre mesures sont utilisées dans la littérature : la sensitivité Sv et la spécificité Sp. La sensitivité représente le rapport entre les observations positives correctement prédites et le nombre des observations positives, et la spécificité représente le rapport entre les observations négatives correctement prédites et le nombre total des observations négatives.

$$\begin{cases} SV = \dfrac{CP}{CP+FN} \\ SP = \dfrac{CN}{CN+FP} \end{cases}$$

Une autre métrique calculée à base de la sensitivité et la spécificité est utilisée. C'est la moyenne harmonique.

$$\text{Moyenne harmonique}=\frac{2\times SV\times SP}{SV+SP}$$

Cela représente un phénomène très connu en apprentissage qui est le sur-apprentissage ou l'apprentissage par cœur. Le sur-apprentissage donne, généralement, des modèles à faible capacité de généralisation, et par conséquent la mesure de précision n'est pas suffisante pour qualifier les performances d'un modèle. Les méthodes d'évaluation permettent de tirer des conclusion sur le comportement d'un modèle face à tout l'espace d'exemples en limitant l'influence des exemples

d'entrainement et du bruit qui peut y exister (erreurs d'étiquetage, erreurs d'acquisition, ...) et leur ordre sur le modèle appris.

(b) Méthode HoldOut

La méthode HoldOut suppose que les données d'entrainement sont tout l'espace d'exemples. Elle consiste à diviser l'ensemble des données en deux parties, la première partie est utilisée pour l'entrainement et la deuxième pour les tests. Le test du modèle appris sur la partie de test permet de donner une idée sur son comportement en dehors des exemples d'entrainement et éviter le phénomène de sur apprentissage.

Le modèle qui maximise la précision pour tout l'espace d'exemple est donc celui qui la maximise pour la partie de test du fait que cette partie représente la majorité de l'espace.

Une question importante qui se pose pour cette méthode est comment choisir les deux parties puisque ce choix a une grande influence sur la qualité du modèle. Le pire est de mettre les exemples positifs dans une partie et les exemples négatifs dans l'autre. La méthode qui suit répond à cette question.

(c) Validation croisée.

Pour minimiser l'influence du choix du partitionnement de l'ensemble des exemples, la validation croisée subdivise l'ensemble d'entrainement initial en k sous ensemble disjointsd_1, d_2, .., dk de même taille. L'entrainement et le test sont effectués k fois. A l'itération i le sous-ensemble Di est réservé pour le test et le reste des exemples sont utilisés pour entrainer le modèle. La précision finale du modèle est égale à la moyenne des k précisions de test.

La méthode Le ave-One-Out est un cas particulier de la validation croisée où k = N. Á chaque itération, le modèle est entrainé sur N − 1 exemples et testé sur l'exemple exclu de l'entrainement. On obtient à la fin N précisions, la précision du modèle est égale à leur moyenne.

(d) Le Bootstrap.

La méthode de Bootstrap, appelée aussi échantillonnage par remplacement, entraine le modèle sur un ensemble de N exemples choisis aléatoirement de l'ensemble des exemples, des exemples peuvent être choisis plus d'une fois et d'autre ne se seront pas choisis du tout. Les exemples non choisis pour l'entrainement sont utilisés pour le test. Cette opération peut être répétée plusieurs fois pour obtenir une précision moyenne du modèle.

Parmi les méthodes de Bootstrap les plus utilisées, la méthode Bootstrap ".632" qui tire son nom du fait que 63.2 % des exemples contribuent à l'entrainement et les restants (36.8%) contribuent aux tests. A chaque prélèvement, un exemple a une probabilité 1/N d'être choisi et (1−1/N) de ne pas l'être, et puisqu'on répète le prélèvement N fois, chaque exemple aura une probabilité de (1−1/N) N de ne pas être choisi du tout dans un ensemble d'entrainement. Si N est grand cette probabilité approche de e−1 = 0.368. La méthode répète le processus k fois et la précision finale P est donnée par :

$$P=\sum_{i=1}^{k}(0.632 \times pi_{test}+0.368\times pi_{entr})$$

Où pi_{test} est la précision du modèle entrainé sur les exemples choisis dans l'itération i, appliqué sur les exemples de test dans la même itération. pi_{entr} Est la précision du même modèle appliqué sur les données d'entrainement [12].

2.5.LES METHODES DE LA CLASSIFICATION

Il existe plusieurs méthode de la classification Nous en énumérons quelques-uns :

(a) Classification hiérarchique

La classification hiérarchique consiste à effectuer une suite de regroupements en classes de moins en moins fines en agrégeant à chaque étape les objets ou les groupes d'objets les plus proches. Elle fournit ainsi un ensemble de partitions de l'ensemble d'objets [5]. Cette approche utilise la notion de distance, qui permet de refléter l'homogénéité ou l'hétérogénéité des classes. Ainsi, on considère qu'un élément appartient à une classe s'il est plus proche de cette classe que de toutes les autres. La figure 2.3 est une illustration du principe des méthodes hiérarchiques [7].

Dans cette figure, on représente la suite de partitions d'un ensemble *{a, b,* c, *d, e}* :

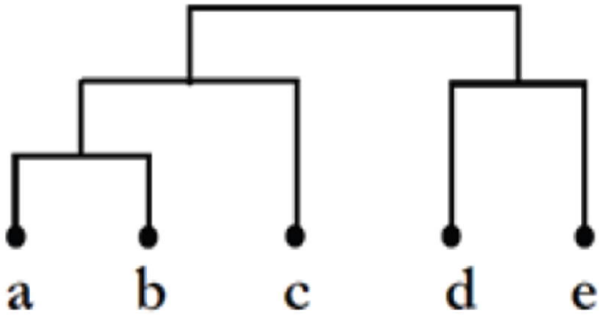


Figure 2.3 : La partition hiérarchique [6].

1. Stratégies d'agrégation sur dis similarités

Le problème est de définir la dis similarité entre la réunion de deux éléments et un troisième :

d (a-b, c).A chaque solution correspond une ultra métrique différente.

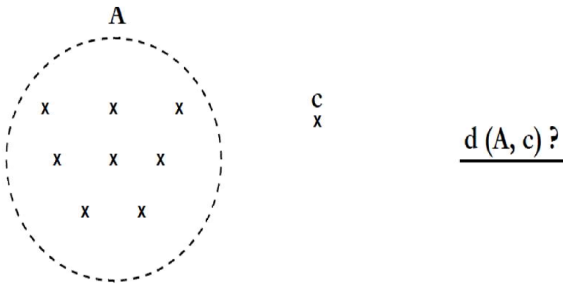


Figure 2.4 : stratégies d'agrégation sur dis similarité [6].

- **Le saut minimum**

Cette méthode (connue sous le nom de « single linkage » en anglais ») consiste à écrire que :

d (a-b, c) = inf {d (a-c); d(b, c)}

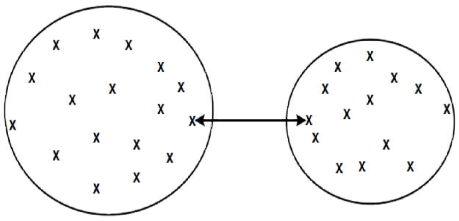


Figure 2.5 : le saut minimum [6].

La distance entre parties est donc la plus petite distance entre éléments des deux parties.

- **Le diamètre (« complète linkage »)**

On prend ici comme distances entre parties la plus grande distance entre deux éléments. d[(a, b); c] = sup [d (a, c), d (b, c)].

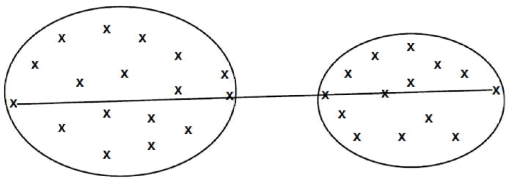


Figure 2.6 : le diamètre [6].

2. Stratégies diverses

- saut minimum (plus proche)
- diamètre
- moyenne des distances
- médiane des distance
- distance au centre de gravité.

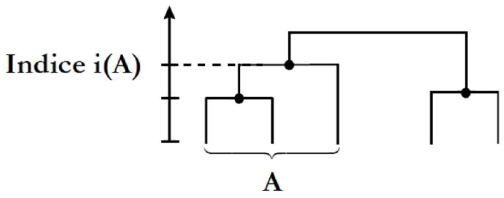


Figure 2.7 : stratégies diverses [6].

L'indice ou niveau d'agrégation est le niveau auquel on trouve agrégés pour la première fois tous les constituants de A.

3. La méthode de Ward pour distance Euclidienne

Si on peut considérer E comme un nuage d'un espace Rp, on agrège les individus qui font le moins varier l'inertie intra-classe.

A chaque pas, on cherche à obtenir un minimum local de l'inertie intercalasse ou un maximum de l'inertie interclasse.

L'indice de dis similarités entre deux classes (ou niveau d'agrégation de ces deux classes) est alors égal à la perte d'inertie interclasse résultant de leur regroupement.

Calculons cette perte d'inertie :

g_A =centre de gravité de la classe A (poidsp_A)

g_B =centre de gravité de la classe B (poidsp_B)

g_{AB} =centre de gravité de leur réunion

$$g_{AB} = \frac{p_A g_A + p_B g_B}{p_A + p_b}$$

L'inertie interclasse étant la moyenne des carrés des distances des centres de gravité des classes au centre de gravité total, la variation d'inertie interclasse, lors du regroupement de A et B est égale à :

$$p_A d^2(g_A, g) + p_B d^2(g_B, g) - (p_A + p_B) d^2(g_{AB}, g)$$

Elle vaut :

$$\delta(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$$

Remarque : Cette méthode entre dans le cadre de la formule de Lance et Williams généralisée :

$$\delta[(A, B) ; c] = \frac{(p_A + p_C)\,\delta(A,C) + (p_B + p_C)\delta(B,C) - p_C \delta(A,B)}{p_A + p_B + p_C}$$

On peut donc utiliser l'algorithme général.

On notera que la somme des niveaux d'agrégation des différents nœuds de l'arbre doit être égale à l'inertie totale du nuage, puisque la somme des pertes d'inertie est égale à l'inertie totale.

Cette méthode est donc complémentaire de l'analyse en composantes principales et repose sur un critère d'optimisation assez naturel. Elle constitue à notre avis la meilleure méthode de classification hiérarchique sur données euclidiennes.

Il ne faut pas oublier cependant que le choix de la métrique dans l'espace des individus conditionne également les résultats [6].

(b) Classification par la méthode de k-moyenne

Introduite par J. McQueen en 1971 et amélioré sous sa forme actuelle par E. Forgy, la méthode du k-moyen est considérée comme un outil de classification efficace qui perm et de diviser un ensemble de données en k classe homogènes. En effet, cette méthode initialise k clusters avec k vecteurs qui servent comme centres de gravité pour le reste des vecteurs à classifier. Chaque vecteur est ajouté dans ce cas, au cluster dont le centre est le plus proche. Les k clusters sont produits de façon à minimiser la fonction objective suivante [21,22]:

$$E=\sum_{r=1}^{k}\sum_{x_i c_r}(x_i-g_r)2$$

Où:

c_r: représente l'ensemble des classes.

x_i: Un point qui appartient à une classec_r.

g_r: Le point moyen de la classec_r.

Dans le domaine de la classification non supervisée, cet algorithme cherche à partitionner l'espace des données en classes isolées les unes des autres, et cela, en minimisant la variance entre ces derniers.

L'exécution de la méthode du K-moyen se déroule en trois étapes :

1. Initialisation de tous les groupes.
2. Faire une première allocation des entités aux groupes de l'étape 1.
3. En cas de besoin, Réallocation des entités aux groupes pour minimiser un critère quelconque.

Il est important de combiner la méthode K-moyen avec un autre algorithme pour fournir une estimation des m groupes à obtenir. Ensuite, pour améliorer le regroupement, K-moyen utilise des conditions de transfert pour reclasser les entités.

Parmi ces conditions, on peut citer :

1. La vérification si un transfert est possible.

2. Arrêt de l'algorithme si aucun transfert n'a eu lieu ou Si on a atteint le nombre maximum d'itérations.

On peut résumer le fonctionnement de l'algorithme K-moyen dans les étapes suivantes :

1. On choisit k objets au hasard qu'on considère comme des centres pour les classes initiales.

2. On affecte chaque objet au centre le plus proche pour obtenir une partition de k classes.

3. On recalcule les centres de chaque classe.

4. La répétition des étapes 2 et 3 jusqu'à la stabilité des centres.

La complexité de l'algorithme du K-moyen est de $O(lkn)$, où l est le nombre d'itérations, k le nombre des classes, et $k < n$.

On retrouve différents types de K-moyen qui se distingue, par la sélection initiale des k objets de bases, par la méthode de calcul des moyennes et des similarités entre les objets de l'espace. Parmi ces méthodes on peut citer, par exemple, la méthode des centres mobiles et celle des nuées dynamiques [23].

(c) Utilisation des réseaux neurone

Depuis quelques années, les réseaux de neurones ont commencé à prendre de plus en plus une grande place dans divers domaines tels que: le traitement des signaux au niveau des télécommunications, la cryptographie, ainsi que le traitement des langues naturelles. Le principe de fonctionnement de ces réseaux est directement inspiré du fonctionnement de vrais neurones humains.

Le fait d'accepter que le cerveau humain fonctionne d'une façon totalement différente de celle d'un ordinateur a eu un impact très important sur le développement des réseaux de neurones. En effet, La quantité énorme des travaux effectués pour comprendre le fonctionnement du cerveau humain a mené la représentation de ce dernier par un ensemble de composantes appelées neurones, interconnectées les unes avec les autres. Le cerveau humain a la capacité d'organiser ces neurones, selon une organisation très complexe, non linéaire et extrêmement parallèle, afin d'accomplir des tâches très élaborées.

Selon (Haykin, 1994) : «Un réseau de neurones est un processus distribué de manière massivement parallèle, qui a une propension naturelle à mémoriser des connaissances de façon expérimentale et de les rendre disponibles pour l'utilisation. Il ressemble au cerveau en deux points:

1. La connaissance est acquise au travers d'un processus d'apprentissage

2. Les poids des connexions entre les neurones sont utilisés pour mémoriser la connaissance».

Cette définition est considérée comme une base pour l'élaboration des réseaux de neurones artificiels.

De manière similaire à la nature, le fonctionnement d'un réseau de neurones est influencé par la connexion des éléments entre eux. On peut entraîner un réseau de neurones pour un rôle spécifique (traitement des signaux par exemple) en ajustant les valeurs des connections (ou poids) entre les neurones [24].

Pour pouvoir utiliser les capacités d'un réseau de neurones dans la classification, il faut premièrement le construire. Ce processus se déroule en quatre étapes :

1. Construire la structure du réseau.

2. Construire une base de données de vecteurs pour modéliser le domaine étudié, ce qui se fait en deux étapes: la première consiste en l'apprentissage du réseau et la deuxième aux différents tests de cet apprentissage.

3. La troisième étape consiste à paramétrer le réseau par apprentissage. Puisque les vecteurs de la base de données d'apprentissage sont présentés au réseau séquentiellement, un algorithme d'apprentissage interviendra pour ajuster les poids du réseau afin que les vecteurs soient correctement interprétés.

4. Reconnaissance: Cette phase consiste à utiliser une base de données de tests qui permettra de voir si les entrées de tests seront reconnues par le réseau construit.

Après l'exécution de plusieurs tests, si le réseau de neurones semble efficace dans le traitement des entrées, on peut l'utiliser pour de vraies applications.

(d) Les arbres de décision

Les arbres de décision sont considérés parmi les méthodes les plus populaires pour la classification textuelle. Parmi les algorithmes les plus connus, on peut citer ID3 (Quinlan, 1986) et C4.5 (Quinlan, 1993).

Le fonctionnement des arbres de décision se base principalement sur des exemples. En effet, si on veut classer des documents dans des catégories, on doit construire un arbre de décision par catégorie.

D'une manière générale, chaque nœud de l'arbre de décision exécute un test If *I* Then et les feuilles de l'arbre ont les valeurs de décision Oui ou Non. Les tests exécutés, observent les valeurs des attributs de chaque exemple. Pour un texte quelconque par exemple, l'attribut peut être un mot avec une valeur de 0 ou 1 selon que ce mot appartient à ce texte ou non [13].

(e) K plus proches voisine(KNN)

La méthode des plus proches voisins (noté parfois K-PPV ou K-NN pour -Nearest-Neighbor) consiste à déterminer pour chaque nouvel individu que l'on veut classer, la liste des plus proches voisins parmi les individus déjà classés. L'individu est affecté à la classe qui contient le plus d'individus parmi ces plus proches voisins. Cette méthode nécessite de choisir une distance, la plus classique est la distance euclidienne (voir chapitre3), et le nombre de voisins à prendre en compte.

Cette méthode supervisée et non-paramétrique est souvent performante. De plus, son apprentissage est assez simple, car il est de type apprentissage par cœur (on garde tous les exemples d'apprentissage). Cependant, le temps de prédiction est très long, car il nécessite le calcul de la distance avec tous les exemples, mais il existe des heuristiques pour réduire le nombre d'exemples à prendre en compte [8].

2.6. CONCLUSION

Dans ce chapitre nous avons passé en revue le principe de la tache Classification avec son processus et les différents algorithmes concernant cette tache ; nous avons également définit les avantages et les inconvénients du Classification et comment évaluer les résultats du Classification.

3.1.INTRODUCTION

l'Extraction de Connaissances à partir de Données est un processus constitué de plusieurs étapes ; elles sont répétées dans des itérations multiples (des feedbacks et des boucles récursives peuvent être observés durant le processus) et à chaque itération ou étape du ce processus une intégration des connaissances des expertes de domaine est nécessaire(le processus est en perpétuelle interaction avec les utilisateurs) pour découvrir de nouvelles connaissances cachées interprétables et utilisables.

Les étapes de ce processus consistent principalement en collection des données contenant dans les différentes sources opérationnelles de l'entreprise, la préparation des données nécessaires pour accomplir la ou les tâches de fouille de donnée souhaitées, l'application des méthodes de fouille de donnée nécessaires Pour résoudre ces tâches et enfin l'évaluation et la validation des résultats obtenus.

La finalité de processus de l'ECD est d'aboutir à des nouvelles connaissances valides et potentiellement utiles. En effet, la connaissance n'est qu'un modèle suffisamment intéressant et sûr. Dans ce chapitre nous allons Présenter une conception d'un système basée sur un processus ECD. Ce système suive des différents depuis le téléchargement des données jusqu'à arriver à avoir des résultats fiables exploitables par un système de prédiction.

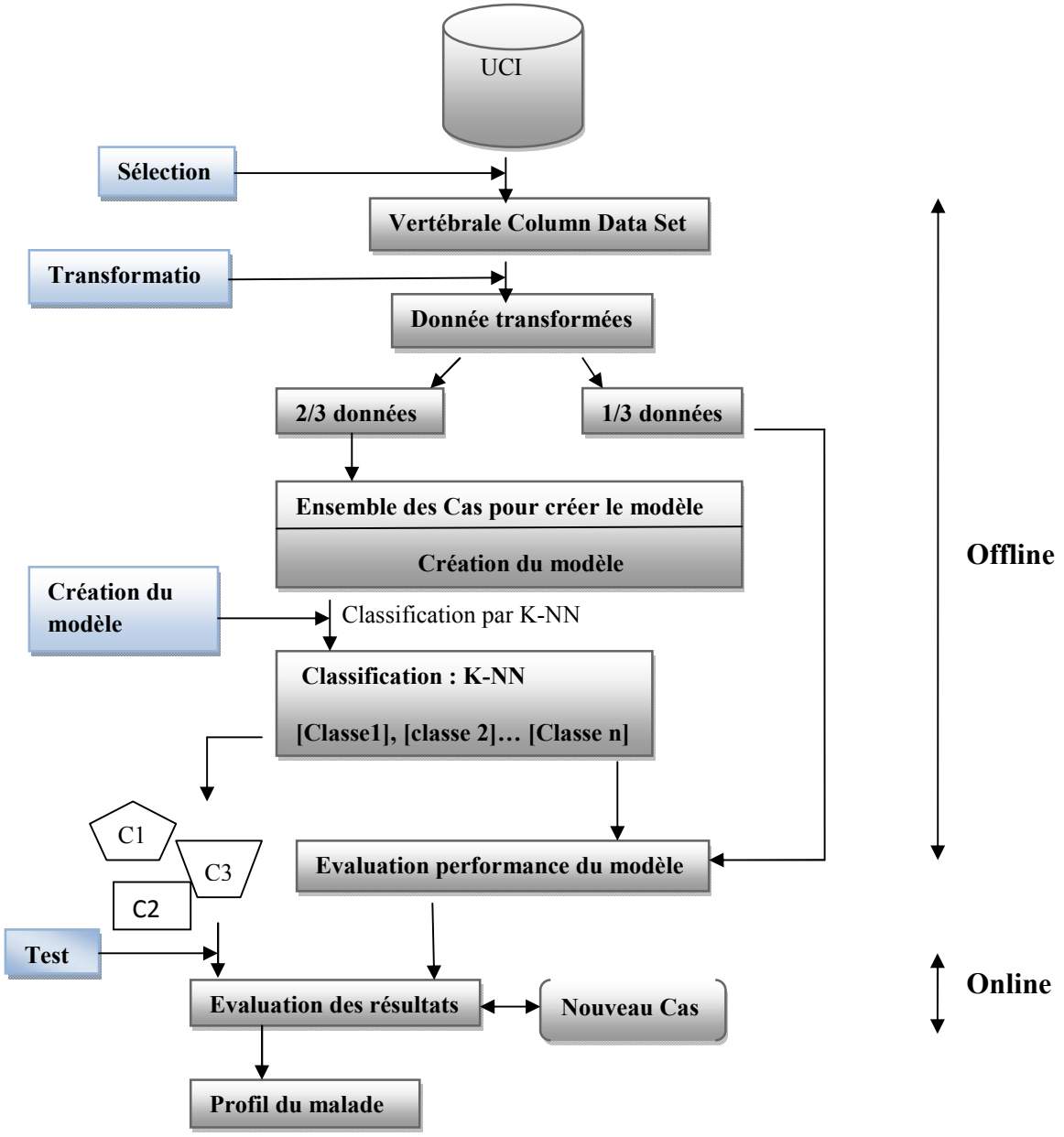
3.2.LE PROSUSE DE FOUILE DE DONNEE PAR CLASSIFICATION POUR LA DETERMINATION DES PROFILS DES MALADES

L'approche proposée dans ce travail est une approche basée sur le processus ECD, cette approche exprime les taches principales que peut un système suivre pour arriver à extraire des nouvelles connaissances.

Dans notre approche on va utiliser le Data set UCI Machine Learning Repository comme une source d'obtenir notre données des vertébral Colum pour qu'elles soient analyser par un environnement d'apprentissage WEKA, ce dernier va nous donner le modèle attendu qui sont les trois classifier des vertébral Colum avec ses trois Classe.

Les Classe de Classifier des verres vont être utilisées par une application de prédiction Java.

Le schéma suivant explique les différentes étapes de cette approche.



Fonctionnement détaillé de l'approche:

- **La sélection**

L'objectif de la sélection de donnés est de sélectionner et d'analyser l'état des données requises pour exécuter la ou les tâches de fouille de donnée souhaitées.

Il s'agit d'identifier les attributs (pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, grade of spondylolisthesis, Class labels) qu'ils sont les différents composants que peut un vertébral Colum contenir dans son structure.et/ ou les enregistrements (lignes) à utiliser pour le fouille de donnée.

Les données ont été collectées à partir de la source de données, UCI Machine Référentiel d'apprentissage car il s'agit d'une base de données couramment utilisée par Les chercheurs en apprentissage automatique avec les enregistrements les plus complets. La technique K-NN a été appliqués pour créer des modèles de prédiction pour cette expérience en utilisant l'ensemble de données préparé

- **La transformation**

Parmi les transformations des données utilisées dans cette approche est la réduction de nombres des données : parmi les 310 instances des vertébral Colum trouvées au Data set (2/3 pour la création de modèles (310 instances), 1/3 pour l'évaluation la détermination des profils des malades))

- **La création de modèle**

C'est La fouille de donné ou l'étape d'extraire des nouvelles informations utilisant des différents méthodes et techniques dédies au tache de Classification.

Cette étape est réalisée sous l'environnement Weka 3.8, ce dernier contient une collection d'outils de visualisation et d'algorithmes pour l'analyse de données et la Modélisation prédictive.

Weka est une suite de logiciels d'apprentissage automatique écrite en Java et développée à l'université de Waikato en Nouvelle-Zélande. C'est un logiciel libre disponible sous la Licence publique générale GNU.

La méthode principale utilisée dans notre approche pour extraire des Classe des vertébral Colum est la méthode K-NN.

Le Deux tiers (2/3) des donnés des vertébral Colum sélectionnés sont destinés pour la création de modèle, le reste des donnés sont pour l'évaluation des performances.

- **L'évaluation et Test**

C'est l'étape finale du processus d'Extraction de Connaissances à partir de Données. Elle consiste à évaluer les résultats obtenus du Weka pour déterminer quels modèles peuvent être considérés comme des connaissances nouvelles et intéressantes. Cette étape comporte aussi une interprétation des résultats et une comparaison des modèles. En effet, la pertinence de la connaissance découverte est estimée par des critères de certitude imposés par les utilisateurs ou les

experts de domaine. Ensuite, les modèles é numéros comme des connaissances pertinentes seront validés sur d'autres ensembles de données ou sur d'autres systèmes.

Les méthodes de validation vont dépendre de la nature de la tâche et du problème considéré. La validation est essentiellement du ressort de l'expert qui jugera de la pertinence des pertinences des résultats, Celui-ci utilisera des méthodes et critères qui s'appliqueront en fonction des données utilisées et de l'objectif fixé dès de le départ à la méthode.

Pour valider les résultats obtenus par une méthode de fouille de données on décompose les données en trois ensembles disjoints : un ensemble d'apprentissage (Training Data), un ensemble de test [Testing Data] et un ensemble de validation [Validation Data]. Au moins deux ensembles sont nécessaires : l'ensemble d'apprentissage permet de générer le modèle et l'ensemble de test permet d'évaluer l'erreur réelle du modèle sur un ensemble indépendant. Ainsi, lorsqu'il s'agit de tester plusieurs modèles et de les comparer, on peut sélectionner le meilleur modèle selon ses performances sur l'ensemble de test et ensuite évaluer son erreur réelle sur l'ensemble de validation.

Remarque

L'état Offline contient les étapes (la sélection, la transformation, La création de modèle, la validation et une partie de test qui est l'évaluation de performance).

L'état Online signifie que les résultats obtenus de chaque étape est reste localement chez l'expert qui est en train de faire une étude sur un échantillon des donnés contrairement l'état d'Online ou les résultats prédit est distribuées et utilisées aux domaines Aide A la décision.

3.3.LE PROCESSUS GENERALE DE CLASSIFICATION

(a) La méthode K-NN

La classification consiste à obtenir un modèle de classification, par apprentissage, et à utiliser un modèle de classification pour attribuer des objets de données de classes inconnues à une classe particulière

Les algorithmes de classification peuvent être classés en groupes qui sont, à savoir l'algorithme de classification hiérarchique, l'algorithme de classification de partitionnement, l'algorithme de classification arbre de décision, l'algorithme de classification k plus proche voisine

Ces algorithmes de classification donnent des résultats différents selon les conditions. Certaines techniques de classification sont meilleures pour les données volumineuses définir et certains donne un bon résultat pour trouver un classe avec arbitraire formes.

Les algorithmes qui sont sous l'exploration dans cette approche est K-NN.

K-NN C'est l'un des algorithmes de classification les plus simples et l'un des algorithmes d'apprentissage les plus largement utilisés.

L'algorithme KNN classe le point de teste en fonction des points d'apprentissage environnants, c'est-à-dire les voisins les plus proches du point de teste. Lorsque certains des points adjacents au point de test sont plus proches que le reste des points, le point de test sera considéré comme appartenant à la catégorie de ces points.

On peut dire que l'algorithme KNN dépend de la mesure de similarité du point teste avec ses points d'apprentissage les plus proches.

Les similitudes entre les cas sont calculées à l'aide d'échelles telles que l'échelle de distance euclidienne et la distance de Hamming.

Algorithme KNN :

Algorithme : prédiction de la classe d'une donnée par la méthode des k plus proches voisins

Nécessite 3 paramètres : un ensemble d'exemples X, une donnée x et k e {1...., k}

Pour chaque exemple x_i ∈X faire

Calculer la distance entre x_i et x : δ (x_i, x)

Fin pour

Pour j ∈ {I,..., k} faire

KNN[j] arg. Min ⟵ δ (x_i, x)

δ (x_i, x) ⟵ ∞

Fin pour

Déterminer la classe de x à partir de la classe des exemples dont le numéro est stocké dans le tableau KNN

(b) La prédiction

Dans notre contexte, la prédiction est définie comme étant un problème de classification supervisée où un algorithme de classification standard est soumis à des modifications afin qu'il soit capable à prédire correctement la classe des nouvelles instances.

Plus formellement, le but des algorithmes de classification est de prédire la classe des nouveaux cas à partir d'un ensemble de données classifiées.

Dans la phase d'apprentissage ces algorithmes visent à découvrir la structure complète du concept cible à partir la formation des classifier à la fois pures en termes de classe et distincts les instances dans chaque classifier doivent être le plus homogènes possibles et différentes de celles appartenant aux autres classifier).

A la fin du processus d'apprentissage, chaque groupe appris prend j comme étiquette si la majorité des instances qui le forme sont de la classe j.

Au final, la prédiction d'une nouvelle instance se fait selon son appartenance à un des groupes appris. Autrement dit, l'instance reçoit j comme prédiction si elle est plus proche du centre de gravité (le représentant / le Classe) du groupe de classe j (utilisations de la distance Euclidienne)

3.4.LES ALGORITHMES UTILISEES

Dans notre approche, La prédiction à base de Classification a été réalisée par un algorithme générale de prédiction comme ci-dessous :

***Algorithme* : Prédiction**

Débuts

Entrée : Base de Données Access

Un tableau de 3 Classe produit par Weka

Un tableau de 1/3 éléments de vertébral Colum

Sortie :Un tableau contient chaque élément avec son classe prédit

Lire le 1 er Elément de C1

Tant que *C1 n'est pas fini faire :*

Lire C2 pour le premier Classe

Lire C2 pour le deuxième Classe

Lire C2 pour le troisième Classe

A=Calculer la distance Euclidienne entre élément et c1

B=Calculer la distance Euclidienne entre élément et c2

C=Calculer la distance Euclidienne entre élément et c3 Calculer min (A, B, C)

Classe de Vertébral Colum prédit d'élément=classe de Vertébral Colum de min

Stuquer classe de Vertébral Colum prédit d'élément avec élément

Lire élément suivant de C1

Fin Faire

Fin

3.5.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 à construire des nouvelles connaissances est la fouille de donné, cette dernière est considéré comme le cœur d'ECD qui utilise un ensemble des techniques dédiées à différentes tâches afin d'arriver à trouver des motifs valables exploitables par un système de prédiction. Dans le chapitre suivant, nous allons présenter la démarche suivie pour développer un système implémentant de telle 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 de notre application et présenté les résultats de l'extraction de connaissance à partir des données de Vertébral Colum.

Les résultats obtenus seront enfin visualisés et validés.

Outil et Environnement de développement

Microsoft Excel (2010) : est un logiciel tableur de la suite bureautique Microsoft Office développé et distribué par l'éditeur Microsoft. La version la plus récente est Excel 2019[1].

Il est destiné à fonctionner sur les plates- formes Microsoft Windows, Mac OS X, Androïde ou Linux (moyennant l'utilisation de Wien). 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 « UCI represetory. » et le logiciel WEKA.

Microsoft Office Access(2010) : Est une base de données relationnelle éditée par Microsoft. Ce logiciel fait partie de la suite Microsoft Office, Depuis les premières versions, l'interface de Microsoft Access permet de gérer graphiquement des collections de données dans des tables, d'établir des relations entre ces tables selon les règles habituelles des bases de données relationnelles, de créer des requêtes avec le QBE (Quercy, ou directement en langage SQL), de créer des interfaces homme/machine et des états d'impression. Comme pour les autres logiciels Office, le VBA, (Visual Basic for Applications), permet de créer des applications complètes et en réseau local, y compris en utilisant, créant ou modifiant les fichiers (documents Word, classeurs Excel, instances Outlook, etc.) des autres logiciels de la suite sans quitter Access.

Le Microsoft Access est utilisé dans notre expérimentation pour la création d'une base de données contient des tables (Classe, Eléments à Prédire).Les données de ces tables sont importées par Access du l'Excel.

Ces tables vont être importées par la suite vers l'environnement de développement Netbeans pour faire réaliser notre système de prédiction.

WEKA « environnement Waikato pour l'analyse de connaissances »

C'est l'outil utilisé dans notre expérimentation pour la création de modèle, est une suite de logiciels d'apprentissage automatique écrite en Java et développée à l'université de Waikato en Nouvelle-Zélande. C'est un logiciel libre disponible sous la Licence publique générale GNU (GPL).

WEKA est un ensemble de classes et d'algorithmes en Java implémentant les principaux algorithmes de Fouille de données. Il est disponible gratuitement à l'adresse www.cs.waikato.ac.nz/ml/weka, dans des versions pour Unix et Windows.

Nous avons implémenté notre système avec le langage **java** qui est un langage orienté objet, Développé par **SUN** Microsystems. Les premières versions datent de 1991 et a réussi à intéresser Beaucoup de développeurs à travers le monde. C'est aussi un langage multi plate-forme disposant De JVM (Java Virtual Machine) lui permettant de s'exécuter dans des environnements hétérogènes En permettant une indépendance envers les réseaux et les systèmes d'exploitation, le langage Java à la particularité principal d'être portable sur plusieurs systèmes d'exploitation tels que Windows, MacOs ou Linux. C'est la plateforme qui garantit la portabilité des applications développées en Java.

NetBeans(IDE8.2) : C'est un environnement de développement intégré (EDI), placé en *open Source* par Sun en juin 2000 sous licence CDDL (Common Développement and Distribution License) Et GPLv2. En plus de Java, NetBeans permet également de supporter différents autres langages, Comme C, C++, JavaScript, XML, Groove, 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 sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java).

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

4.2.LE DOMAINE DAPPLICATION

Dans notre expérimentation la source des données utilisée est « UCI Machine Represetory » Référentiel d'apprentissage car il s'agit d'une base de données couramment utilisée par Les chercheurs en apprentissage automatique avec les enregistrements les plus complets.

4.3.LENVIRONNEMENT DE LEXPERIMENTATION

3.1. Le logiciel Weka

Les 2/3 des données (310 données) sont les données destinées à la création de modèle sous le logiciel d'apprentissage automatique WEKA, ces données sont importées du Data Set au WEKA à l'aide de Microsoft EXCEL qui a réuni ces données dans un fichier CSV (Data310. CSV) voir la figure 4.1.

39.05695098	10.06099147	25.01537822	28.99595951	114.4054254	4.564258645	Hernia
68.83202098	22.21848205	50.09219357	46.61353893	105.9851355	-3.530317314	Hernia
69.29700807	24.65287791	44.31123813	44.64413017	101.8684951	11.21152344	Hernia
49.71285934	9.652074879	28.317406	40.06078446	108.1687249	7.918500615	Hernia
40.25019968	13.92190658	25.1249496	26.32829311	130.3278713	2.230651729	Hernia
53.43292815	15.86433612	37.16593387	37.56859203	120.5675233	5.988550702	Hernia
45.36675362	10.75561143	29.03834896	34.61114218	117.2700675	-10.67587083	Hernia
43.79019026	13.5337531	42.69081398	30.25643716	125.0028927	13.28901817	Hernia
36.68635286	5.010884121	41.9487509	31.67546874	84.24141517	0.664437117	Hernia
49.70660953	13.04097405	31.33450009	36.66563548	108.6482654	-7.825985755	Hernia
31.23238734	17.71581923	15.5	13.51656811	120.0553988	0.499751446	Hernia
48.91555137	19.96455616	40.26379358	28.95099521	119.321358	8.028894629	Hernia
53.5721702	20.46082824	33.1	33.11134196	110.9666978	7.044802938	Hernia
57.30022656	24.1888846	46.99999999	33.11134196	116.8065868	5.766946943	Hernia
44.31890674	12.53799164	36.098763	31.78091509	124.1158358	5.415825143	Hernia
63.83498162	20.36250706	54.55243367	43.47247456	112.3094915	-0.622526643	Hernia
31.27601184	3.14466948	32.56299592	28.13134236	129.0114183	3.623020073	Hernia
38.69791243	13.44474904	31	25.25316339	123.1592507	1.429185758	Hernia
41.72996308	12.25407408	30.12258646	29.475889	116.5857056	-1.244402488	Hernia
43.92283983	14.17795853	37.8325467	29.7448813	134.4610156	6.451647637	Hernia
54.91944259	21.06233245	42.19999999	33.85711014	125.2127163	2.432561437	Hernia
63.07361096	24.41380271	53.99999999	38.65980825	106.4243295	15.77969683	Hernia
45.54078988	13.06959759	30.29832059	32.47119229	117.9808303	-4.987129618	Hernia
36.12568347	22.75875277	29	13.3669307	115.5771163	-3.237562489	Hernia
54.12492019	26.65048856	35.32974693	27.47443163	121.447011	1.571204816	Hernia
26.14792141	10.75945357	14	15.38846783	125.2032956	-10.09310817	Hernia
43.58096394	16.5088837	46.99999999	27.07208024	109.271634	8.992815727	Hernia
44.5510115	21.93114655	26.78591597	22.61986495	111.0729197	2.652320636	Hernia
66.87921138	24.89199889	49.27859673	41.9872125	113.4770183	-2.005891748	Hernia
50.81926781	15.40221253	42.52893886	35.41705528	112.192804	10.86956554	Hernia
46.39026008	11.07904664	32.13655345	35.31121344	98.77454633	6.386831648	Hernia
44.93667457	17.44383762	27.78057555	27.49283695	117.9803245	5.569619587	Hernia
38.66325708	12.98644139	39.99999999	25.67681568	124.914118	2.703008052	Hernia
59.59554032	31.99824445	46.56025198	27.59729587	119.3303537	1.474285836	Hernia

Figure 4.1 : Echantillon des données de fichier(Data310.CSV)

Ces données (310 instances) vont être importées au WEKA avec ses 7 attributs (voir figure 4.2)

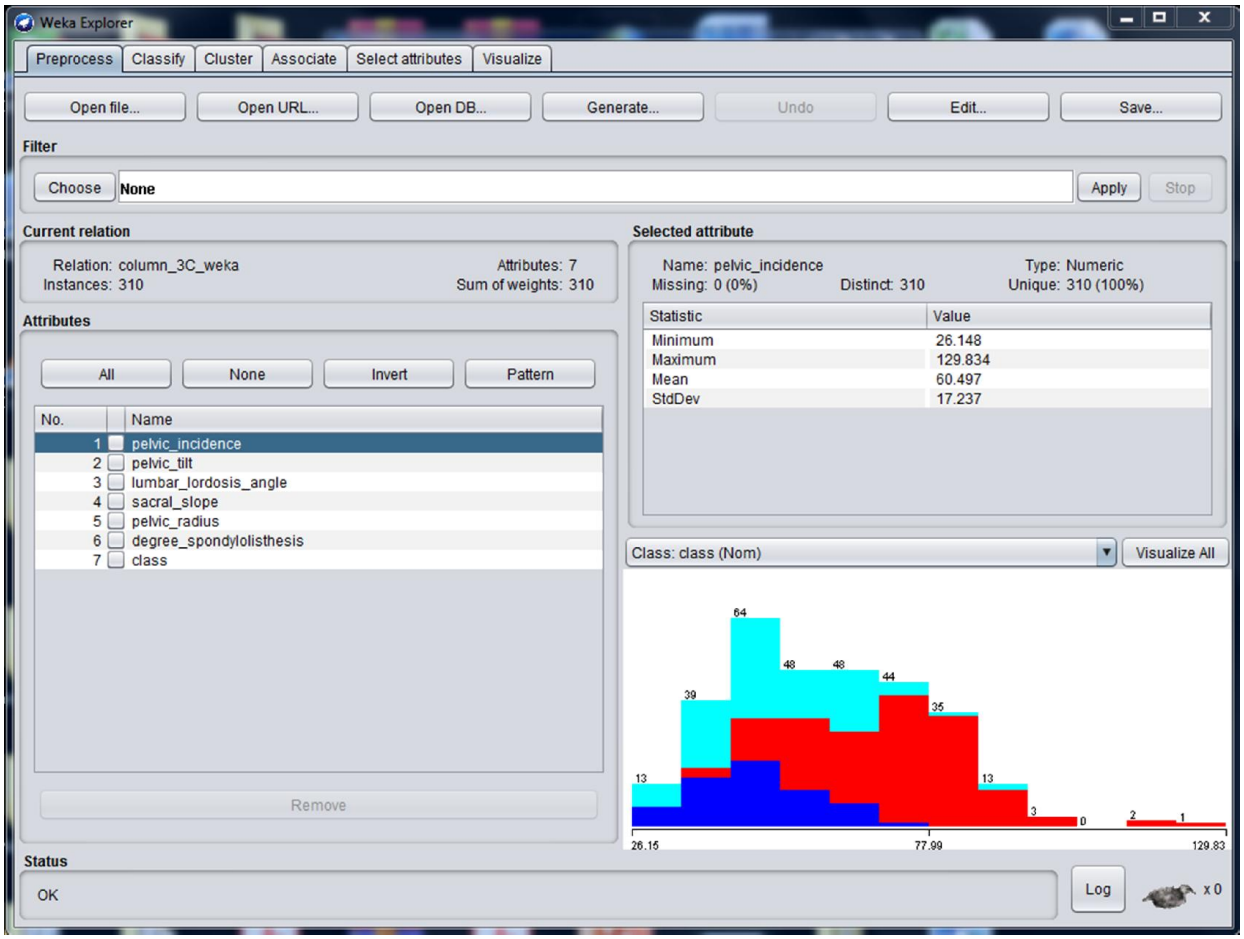


Figure 4.2 : importation de fichier Data310. CSV au WEKA

La tache choisit dans notre expérimentation est « le Classify » utilisant l'algorithme K-NN(IBK) sachant que K=3

Les classe de classifie sont :

Classe 1: hernia (class DH)

Classe2: Spondylolisthesis (class SL)

Classe3: Normal (class NO)

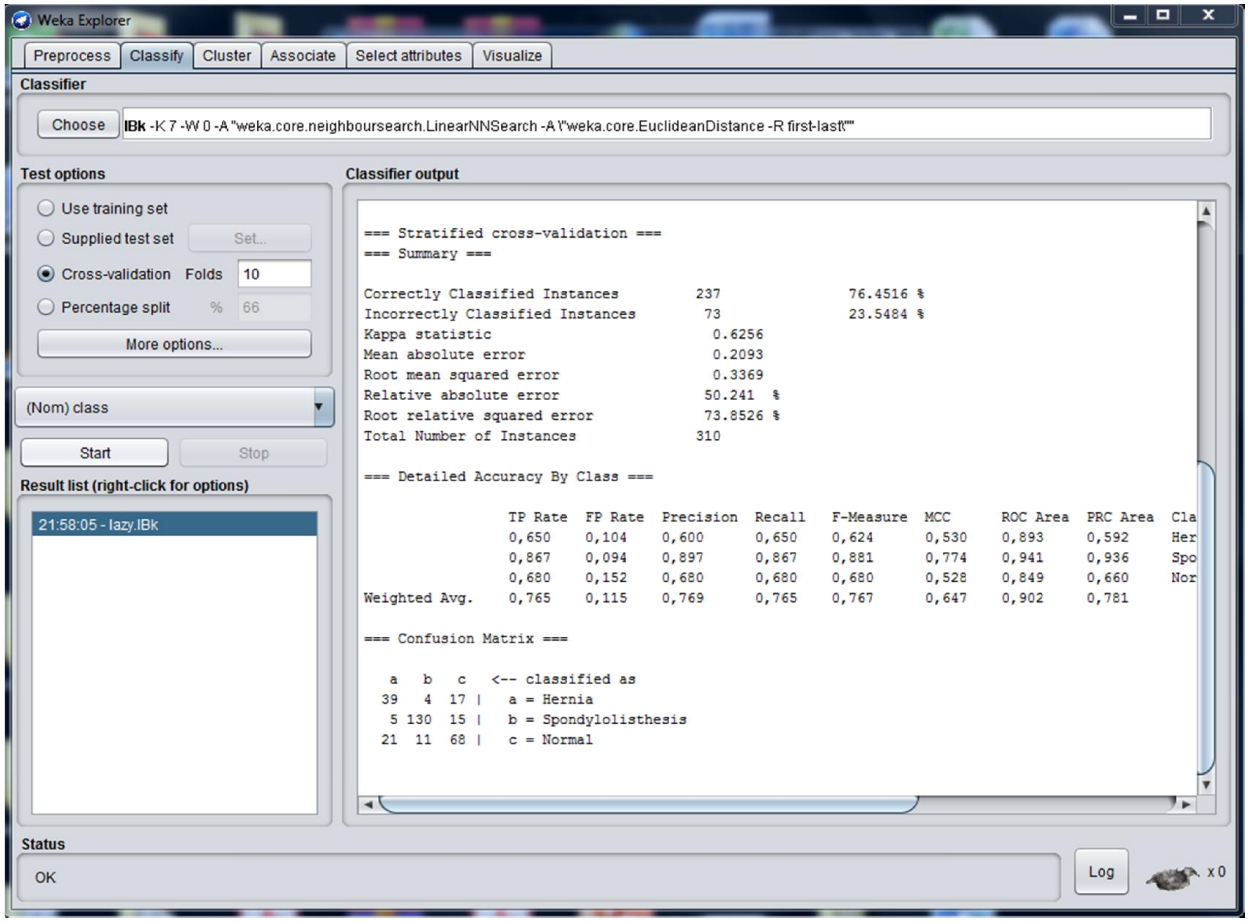


Figure 4.3: résultats de classe de Classify par WEKA

3.2. l'application développée

Les données utilisées dans cette étape au niveau de l'environnement NETBEANS est les 3 Classe produit par WEKA et les 1/3 Eléments dès la Vertébral Colum, ces deux dernières sont des tables crées au niveau d'Access pour qu'on puisse les importer sur NETBEANS.

Le logiciel que nous avons réalisé permet de prédire les classes des contrées (Vertébral Colum).

Ce schéma représente les principales Classes Java et les différentes interactions possibles entre elles

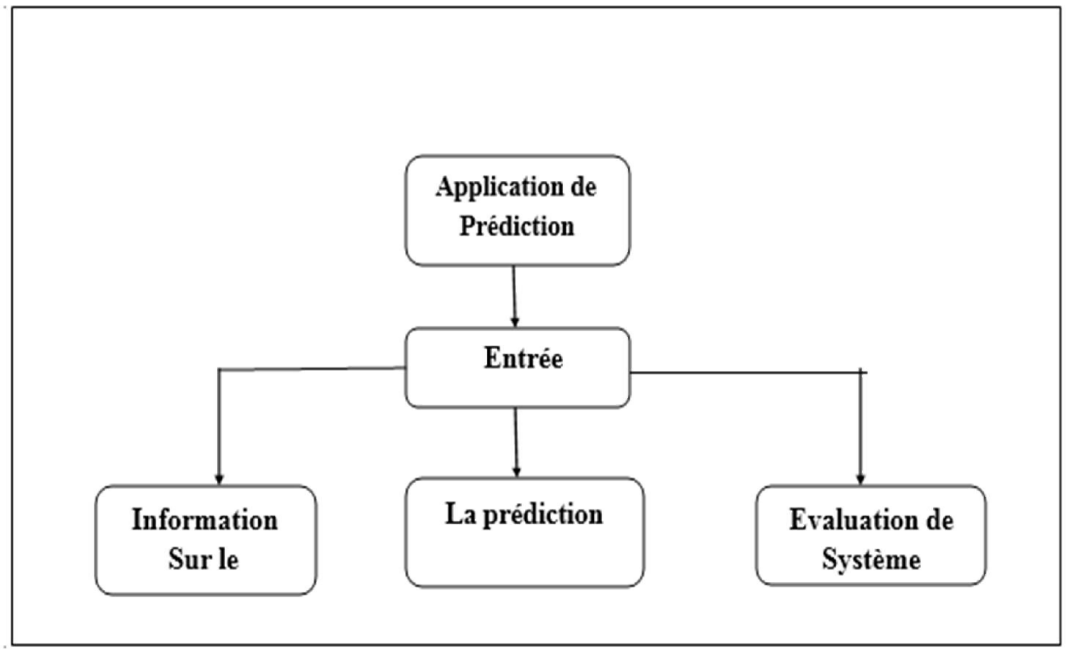


Figure 4.4: Architecture fonctionnelle de système

- Les différentes Classes

1- Classe Page Accueil

C'est la fenêtre principale qui réunit toutes les classes de système, elle contient 3 boutons chaque boutons considère comme une classe séparée. Voir la figure 4.5

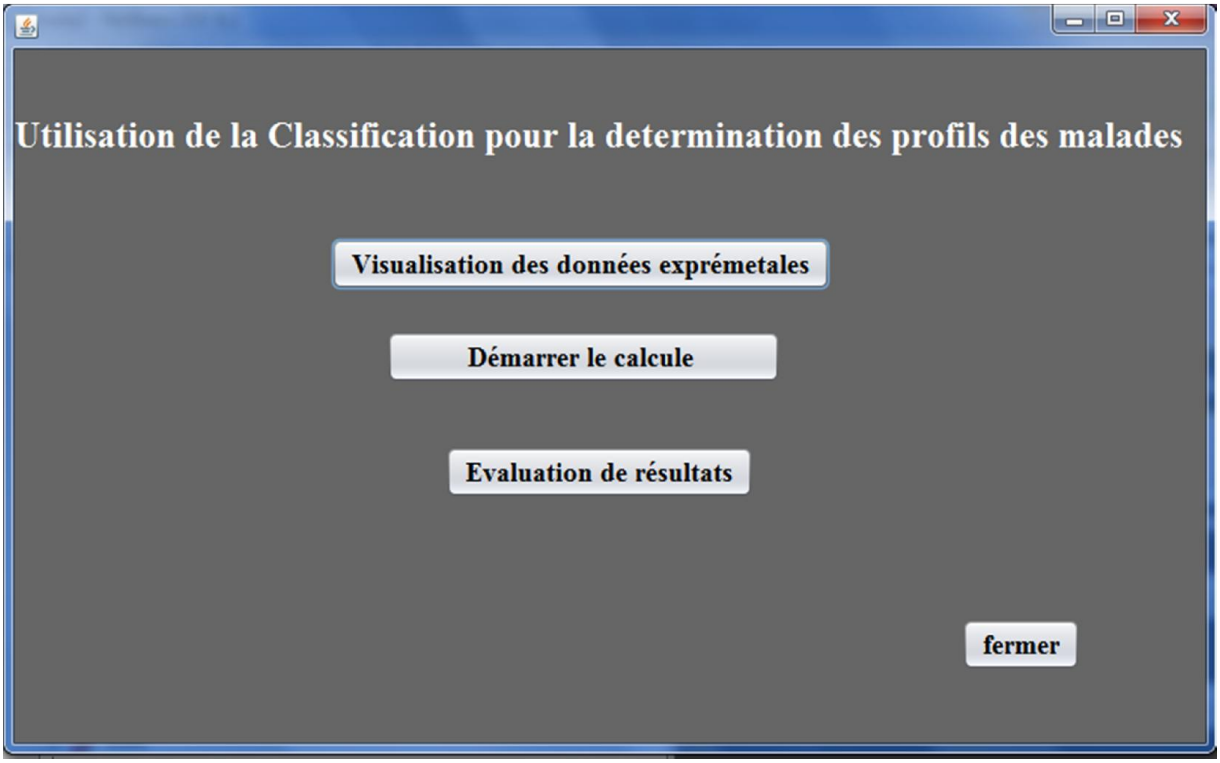


Figure 4.5 :L'interface principale de système

2- Classe Information Sur Le système

C'est une classe permet d'afficher les informations nécessaires pour la réalisation de système Elle contient deux J Tables dont ses informations sont importées de Microsoft Access

Le premier J Table contient les 3 Classe de Classify (1-2-3)

Le deuxième J Table contient les éléments à prédire(les 1/3 données dès la Vertébral Colum). Voir la figure 4.6

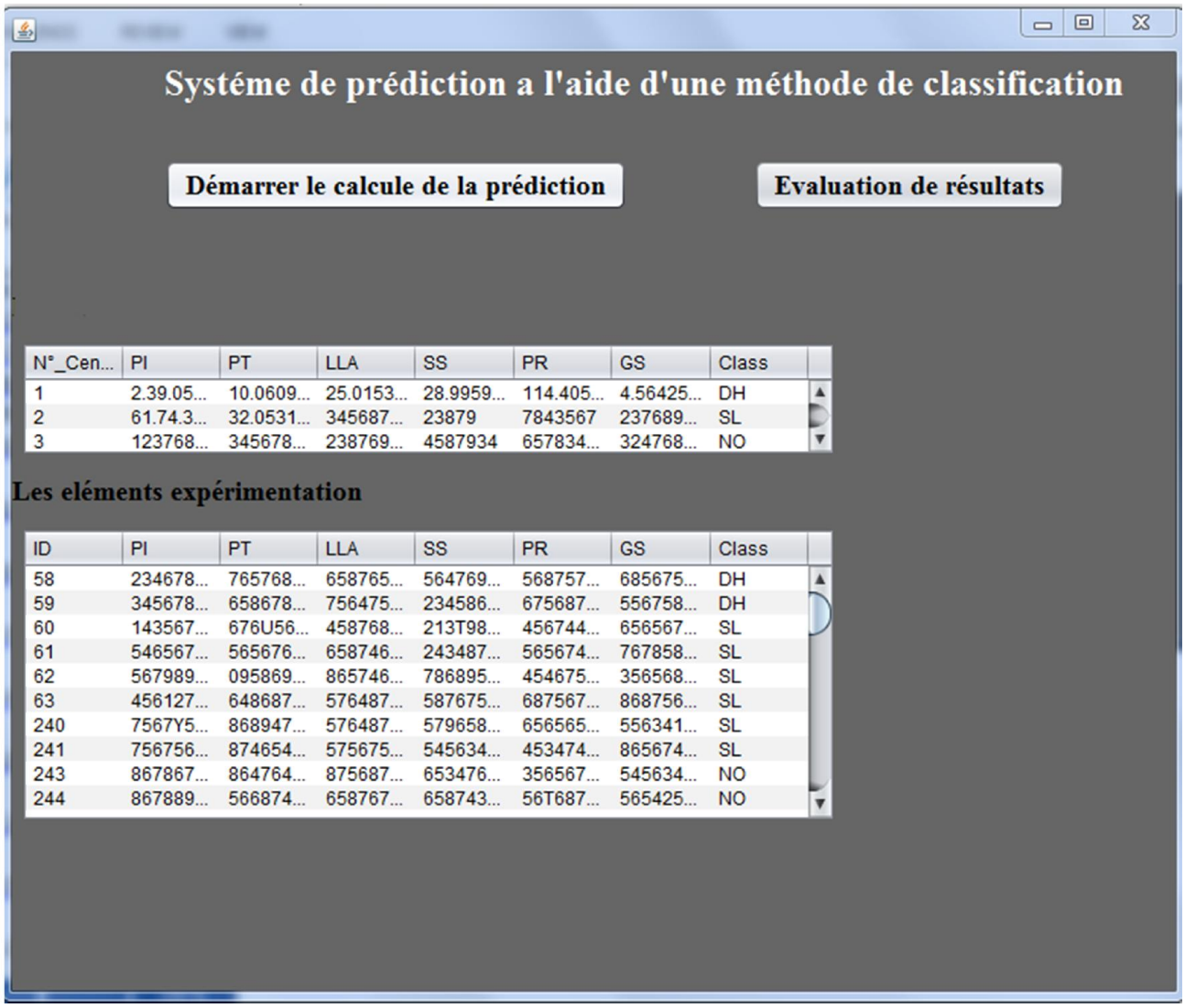


Figure 4.6 : les classe importées de WEKA et la base de teste

3-Classe Faire De Prédiction

C'est la classe la plus importante de système, elle contient l'algorithme de prédiction Qui permet de prédire les classes des éléments des vertébral colum à partir :

- Le calcul de Classe des 1/3 des éléments dès la Vertébral Colum: c'est le calcul de la distance

Euclidienne de chaque élément avec les 3 Classe utilisant cette formule :

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+\ldots+|x_{i_p}-x_{j_p}|^2)}$$

- Le calcul de la fonction minimum entre les 3 résultats de la distance Euclidienne entre chaque élément et les 3 Classe

4.4.LES DONNEES EXPRIMENTALES

Notre système utilise 310 instances des verres obtenus de la source de données « UCI Machine Learning Represetory» car il s'agit d'une base de données couramment utilisée par Les chercheurs en apprentissage automatique avec les enregistrements les plus complets.

Les données comportent 6 Attributs :

1. **PE :** pelvic incidence,
2. **PT :** pelvic tilt,
3. **LLA :** lumbar lordosis angle,
4. **SS :** sacral slope
5. **PR :** pelvic radius
6. **GS :** grade of spondylolisthesis

Les deux tires (2/3) de totale des données (310 données) sont importées par Excel vers le logiciel WEKA pour la création de modèle.

Les uns tiers (1/3) de totale de données sont utilisés pour la validation ou l'évaluation de modèle par la suite :

2.	10.060991	25.015378	28.995959	114.40542	4.5642586	Hernia
3.	22.218482	50.092193	46.613538	105.98513	-3.530317	Hernia
4.	24.652877	44.311238	44.644130	101.86849	11.211523	Hernia
5.	9.6520748	28.317406	40.060784	108.16872	7.9185006	Hernia
6.	13.921906	25.124949	26.328293	130.32787	2.2306517	Hernia
7.	15.864336	37.165933	37.568592	120.56752	5.9885507	Hernia
8.	10.755611	29.038348	34.611142	117.27006	-10.67587	Hernia
9.	13.533753	42.690813	30.256437	125.00289	13.289018	Hernia
10.	5.0108841	41.948750	31.675468	84.241415	0.6644371	Hernia
11.	13.040974	31.334500	36.665635	108.64826	-7.825985	Hernia
12.	17.715819	15.5	13.516568	120.05539	0.4997514	Hernia
13.	19.964556	40.263793	28.950995	119.32135	8.0288946	Hernia
14.	20.460828	33.1	33.111341	110.96669	7.0448029	Hernia
15.	24.188884	46.999999	33.111341	116.80658	5.7669469	Hernia
16.	12.537991	36.098763	31.780915	124.11583	5.4158251	Hernia

Figure 4.7 : Echantillon des données des vertébral Colum importées du Data Set

4.5.LALGORITHME UTILISE

Une présentation générale de l'Algorithme de Prédiction

***Algorithme* : Prédiction**

Débuts

Entrée : Base de Données Access

Un tableau de 3 Classe produit par Weka

Un tableau de 1/3 éléments de vertébral Colum

Sortie :Un tableau contient chaque élément avec son classe prédit

Lire le 1 er Elément de C1

Tant que *C1 n'est pas fini faire :*

Lire C2 pour le premier Classe

Lire C2 pour le deuxième Classe

Lire C2 pour le troisième Classe

A=Calculer la distance Euclidienne entre élément et c1

B=Calculer la distance Euclidienne entre élément et c2

C=Calculer la distance Euclidienne entre élément et c3 Calculer min (A, B, C)

Classe de Vertébral Colum prédit d'élément=classe de Vertébral Colum de min

Stuquer classe de Vertébral Colum prédit d'élément avec élément

Lire élément suivant de C1

Fin Faire

Fin

4.6.LES RESULTATS DE LEXPERIMENTATION

Après avoir programmé l'algorithme de prédiction en java notre système donne des résultats intéressantes, il a prédit pour chaque élément d'1/3 des Vertébral Colum un clase, Voir les figures 4.8

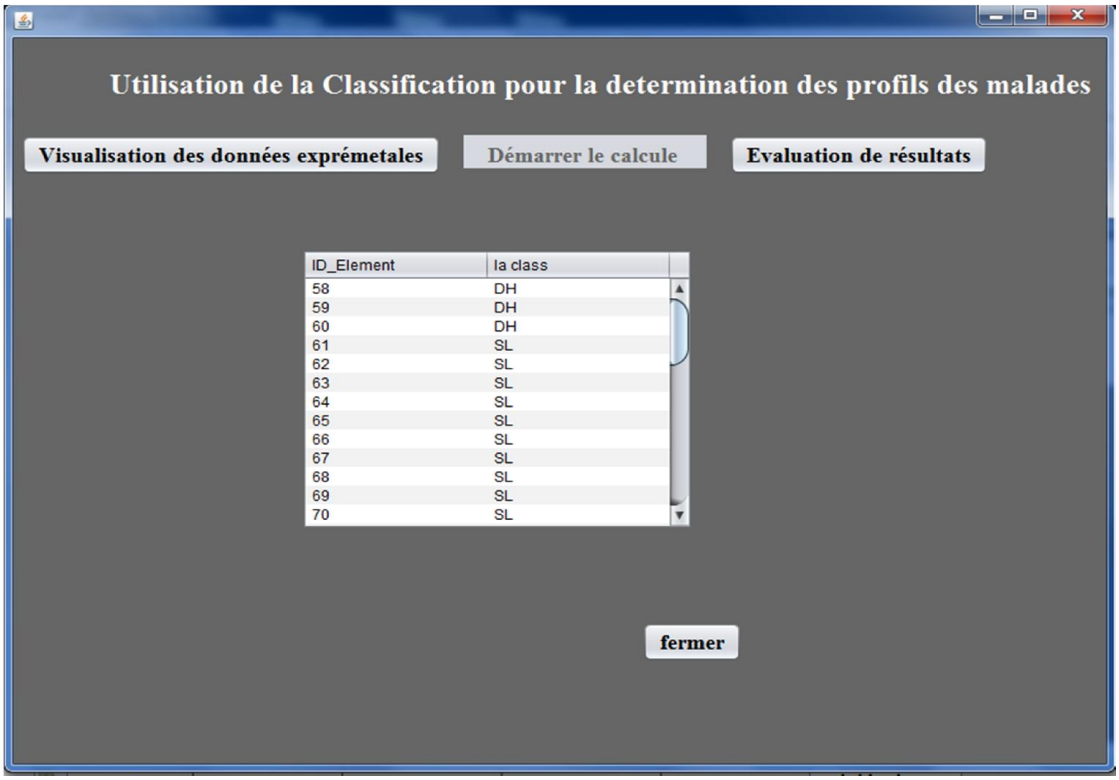


Figure 4.8. Résultats de prédiction

L'évaluation des résultats

Le système a donné pour chaque clase prédit une évaluation (Bonne prédiction/Mauvaise prédiction) comparant avec les clase réels de cette Vertébral Colum.

Enfin, il a affiché un pourcentage de performance de prédiction. Voir figure 4.9

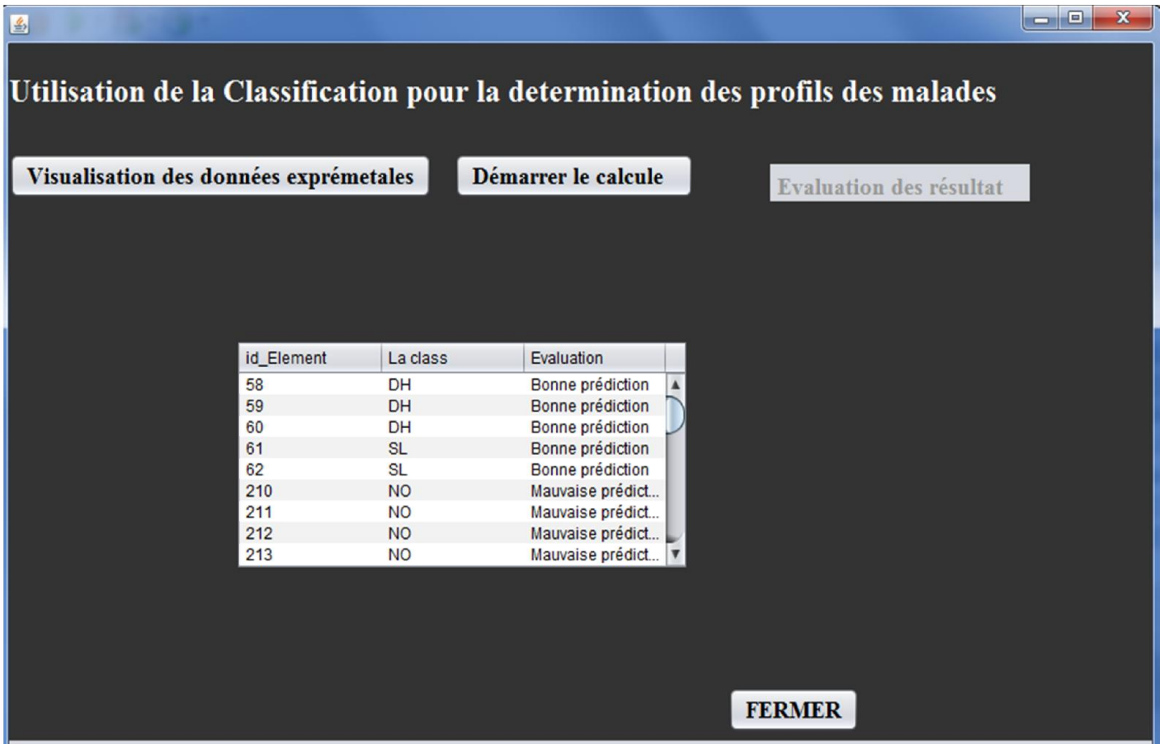


Figure 4.9. L'évaluation des résultats

4.7.CONCLUSION

Dans ce chapitre, nous définissons un système qui permet de prédire les catégories de chaque élément échantillonné dans le rachis lombaire ; cette étude apporte une grande contribution au domaine médical car elle permet de connaître le clase de maladie chez le patient ou s'il a cette maladie ou ne pas.

Notre résultat d'expérimentation basé sur les techniques de Fouille de données montre que notre système est performant de 52% de prédiction des clase dès la Vertébral Colum.

Conclusion générale

Le travail que nous avons mené dans ce mémoire a permis de constater 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 cœur d'un processus ECD, c'est un ensemble des taches et techniques utilisées afin d'arriver aux nouvelles connaissances exploitables.

Nous nous sommes basées dans notre mémoire sur la tache de Classification, cette dernière cherche à discerner une structure dans un ensemble de données non étiquetées. L'objectif est de trouver une typologie des individus en groupes distincts. Chaque groupe doit contenir les individus les plus homogènes possibles.

Il existe plusieurs méthodes et techniques dédiées au Classification, dans notre travail on a choisi l'Algorithme K-NN comme une méthode d'étude et pour réaliser notre système.

K-NN C'est l'un des algorithmes de classification les plus répandus. La méthode des K plus proches voisins (KNN) a pour but de classifier des points cibles (classe méconnue) en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori).

La classification k-NN est classée en positions en utilisant la distance euclidienne pour obtenir une similarité dans les données d'entraînement et de test.

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 qui peut prédire des classe des vertébral Colum utilisant les techniques nécessaires.

Bibliographie

Les figures

Les tableaux

Référence

[1] El Moukhtar, Z. (2013). Représentation et gestion des connaissances dans un processus d'Extraction de Connaissances à partir de Données multi-points de vue (Doctoral dissertation, Thèse de doctorat, Ecole Nationale Supérieur d'Arts et Métiers–Meknès).

[2] Crié, D. (2003). De l'extraction des connaissances au Knowledge Management. Revue française de gestion, 146(5), 59-79.

[3] Boulila, W. (2012). Extraction de connaissances spatio-temporelles incertaines pour la prédiction de changements en imagerie satellitale (Doctoral dissertation, Télécom Bretagne, Université de Rennes 1).

Bibliographie

[4] Brahimi, B. (2011). Extraction de connaissances à partir de données incomplètes et imprécises. Université de M'sila.

[5] Schoier, G. (2004). Introduction au Data Mining. In Notes de cours (p. 55). Université de Trieste.

[6] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[7] Zighed, A., Auray, J. P., & Duru, G. (1992). Sipina, méthode et logiciel. A. Lacassagne.

[8] Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[9] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

[10] Tufféry, S. (2005). Data mining et statistique décisionnelle: l'intelligence dans les bases de données. Editions Technip.

[11] Preux, P. (2011). Fouille de données, notes de cours. Université de Lille, 3.

[12] Kateb, N., & Guerram, T. (2011). Une Approche multi agents pour les datd mining.

[13] Tuffery, S. (2014). Cours de data mining. université de Rennes, 1.

[14] Macal, C. M. (2005, April). Model verification and validation. In Workshop on" Threat Anticipation: Social Science Methods and Models.

[15] Gorade, S. M., Deo, A., & Purohit, P. (2017). A study of some data mining classification techniques. International Research J. of Engineering and Technology (IRJET), 4.

[16] Kesavaraj, G., & Sukumaran, S. (2013, July). A study on classification techniques in data mining. In 2013 fourth international conference on computing, communications and networking technologies (ICCCNT) (pp. 1-7). IEEE.