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 « RANDOM TREE »

POUR DÉTERMINER LE TYPE D'UNE OPÉRATION MÉDICALE (TYPE D'ACCOUCHEMENT)

Réalisé par :

- Abidi Saad
- BOUCENNA Djamel Eddine

Encadré par :

A. MANSOUL

Année Universitaire 2021-2022

Remerciement

Avant tout nous remercions Allah le tout puissant qui nous a donné la force, la patience et le courage pour qu'on puisse accomplir ce modeste travail.

Nous tenons à exprimer toute nos reconnaissances à notre directeur de mémoire, Monsieur A. MANSOUL. Nous le remercions de nous avoir encadrés, orienté, aidé et conseillé.

Nous adressons nos sincères remerciements à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé nos réflexions et ont accepté de nos rencontrer et de répondre à nos questions durant nos recherches.

Nous remercions nos très chers parents qui ont toujours été là pour nous.

Enfin, nous remercions qui ont toujours été là pour nous. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

À tous ces intervenants, Nous présentons nos remerciements, nos respects et nos gratitudes.

Dédicace

Je dédie ce mémoire

A mes cher parents ma mère ***« BENHAMADA Nassima »*** *et mon père* ***« ABIDI Mourad »***

Pour leur patience, leur amour, leur soutien et leur

Encouragements.

A mes Grandes mères « ***Madaci Malika, moudjed el khamsa, gohtari louiza »***

A mes amis et mes camarades.

Sans oublier tout les professeurs qui ce soit du

Primaire, du moyen, du secondaire ou de

L'enseignement supérieur.

ABIDI SAAD

Dédicace

Je dédie ce mémoire

A mes cher parents ma mère ***« BOUSTIL Hadda »*** *et mon père* ***« BOUCENNA Mohamed »***

Pour leur patience, leur amour, leur soutien et leur

Encouragements.

A mes sœurs.

A mes amis et mes camarades.

Sans oublier tout les professeurs qui ce soit du

Primaire, du moyen, du secondaire ou de

L'enseignement supérieur.

BOUCENNA DJAMEL EDDINE

Résumé

La fouille de données, également connue sous le nom de Data Mining, est le noyau d'un processus d'extraction de connaissances à partir de grandes quantités de données. Son domaine d'application est extrêmement large.

Dans ce travail, nous présentons un modèle de prédiction permettant de localiser si une femme va accouchera un accouchement normale ou bien césarienne.

Pour atteindre cet objectif nous proposons un système qui va s'articuler autour de trois modules dont les tâches sont les suivantes:

1. Dans un premier temps nous employons la technique de la classification pour structurer les données en arbre de décision dont les nœuds sont plus ou moins proches en prédiction, c'est le modèle de connaissances que nous aurons construit. Pour ce faire, nous proposons l'utilisation de la méthode Random Tree sous un environnement appelé WEKA destiné à la fouille de données.

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

3. Dans une étape finale nous expérimentons notre approche sur des données se rapportant aux des femmes qui ont accouché déjà.

Le travail que nous présentons dans ce mémoire est très intéressant notamment dans recherche de l'information médicale. Ceci, permettra de contribuer au développement d'un système pour les gynécologues.

Mots clés :

Extraction de connaissances, Fouille de données, Classification, Prédiction, Accouchement, Random Tree.

ملخص

يعد استخراج البيانات، المعروف أيضًا باسم التنقيب عن البيانات، جوهر عملية استخراج المعرفة من كميات كبيرة من البيانات مجال تطبيقه واسع للغاية في هذا العمل نقدم نموذج تنبؤ يسمح بتحديد ما إذا كانت المرأة ستلد ولادة طبيعية أو عملية قيصرية.

لتحقيق هذا الهدف نقترح نظامًا يتمحور حول ثلاث وحدات تكون مهامها على النحو التالي:

1. نستخدم تقنية التصنيف لبناء البيانات في شجرة قرار تكون عقدها قريبة إلى حد ما من التنبؤ، وهذا هو نموذج المعرفة الذي سنقوم ببنائه. للقيام بذلك نقترح استخدام طريقة Random Tree في بيئة تسمى WEKA مخصصة لاستخراج البيانات.
2. في الخطوة الثانية، نستخدم وحدة نمطية قمنا بتطويرها لعمل التنبؤ من النموذج المبني حسب التصنيف.
3. في خطوة أخيرة قمنا بتجربة نهجنا على البيانات المتعلقة بالنساء اللواتي أنجبن بالفعل العمل الذي نقدمه في هذه الأطروحة ممتع للغاية وخاصة في البحث عن المعلومات الطبية. سيساهم هذا في تطوير نظام لأطباء أمراض النساء.

الكلمات المفتاحية:

* استخراج المعرفة، التنقيب عن البيانات، التصنيف، التنبؤ، الولادة،Random Tree .

Abstract

Data mining is the core process of extracting knowledge from large amounts of data. Its field of application is extremely broad.

In this work, we present a prediction model that allows locating whether a woman will give birth to a normal or a caesarean delivery.

To achieve this goal we propose a system that will be articulated around three modules whose tasks are the following:

1. First, we use the classification technique to structure the data in tree whose nodes are more or less close in prediction; this is the knowledge model that we will have built. To do this, we propose the use of the Random Tree method under an environment called WEKA for data mining.

2. In a second step we use a module that we have developed to make the prediction from the model built by classification.

3. In a final step we experiment our approach on data related to women who have already given birth.

The work we present in this thesis is very interesting especially in the search for medical information. This will contribute to the development of a system for gynaecologists.

Keywords:

Knowledge extraction, Data mining, Classification, Prediction, Delivery, Random Tree.

Sommaire

Résumé

Liste des figures

Introduction générale.....p. 1

Chapitre 01. L'Extraction de connaissance a partir de données

1. Introduction.....p. 4
2. L'extraction de connaissance à partir de données (ECD).....p. 4
3. La fouille de données.....p. 5
4. Etapes du processus de l'ECD.....p. 6
   - 4-1 Préparation des données.....p. 6
   - 4-2 Data mining (fouille de données).....p. 7
   - 4-3 Evaluation du résultat.....p. 7
5. Principales taches de fouilles de données.....p. 8
   - 5.1. La description.....p. 9
   - 5.2. La classification.....p. 9
   - 5.3. La segmentation.....p. 9
   - 5.4. La recherche d'association.....p. 10
   - 5.5. La prédiction.....p. 10
   - 5.6. La visualisation.....p. 10
6. Domaines d'application.....p. 11
7. Conclusion.....p. 13

Chapitre 02. La classification

1. Introduction.....p. 15
2. Définitions.....p. 15
3. Le processus de classification.....p. 16
   - 3.1. Construction du modèle.....p. 16
   - 3.2. Utilisation du modèle.....p. 16
4. Critères pour une bonne classification.....p. 17
5. Les méthodes de classification.....p. 17
   - 5.1. Classification supervisée.....p. 18
     - 5.1.1. Arbre de décision.....p. 19
   - 5.2. Classification non-supervisée.....p. 23
6. Evaluation des méthodes de classification.....p. 26
7. Les étapes d'une classification.....p. 26
8. Conclusion.....p. 27

Chapitre 03 : Approche d'arbre de décision pour déterminer le type d'opération médicale

1. Introduction.....p. 29
2. Les données expérimentable.....p. 29
3. Le Processus de fouille de données par classification (Random tree) pour déterminer le type d'opération médicale.....p. 29
4. Le processus général de Classification pour déterminer le type d'opération médicale.....p. 31
5. Conclusion.....p. 33

Chapitre 04 : Résultat d'évaluation et leur discussion

1. Introduction.....p. 35
2. Outil et environnement de développement.....p. 35
3. Le Domaine d'application.....p. 36
4. DataSet.....p. 36
5. L'environnement de l'expérimentation.....p. 37
6. Les interfaces du systéme.....p. 40
7. Conclusion.....p. 42

Conclusion Générale.....p. 44

Bibliographie

Liste des figures

Figure 01: La relation entre data mining et les autres technologies……..……..p. 5

Figure 02: Le processus d'ECD……..….. p. 6

Figure 03: Principales taches de fouille de données……..……. p. 8

Figure 04: visualisation des termes de recherche « data visualisation » …… p. 11

Figure 05: le processus de classification…….....……p. 16

Figure 06: Différents types de méthodes de classification……..…… p. 18

Figure 07: Arbre de décision…….……. p. 19

Figure 08: Exemple d'application de l'algorithme k-NN …… p. 21

Figure 09: Perceptron à trois couches (schéma type) ……..…… p. 22

Figure 10: Partitionnement basé sur k-Means…….…… p. 24

Figure 11: Partitionnement basé sur k-Medoids……..…… p. 25

Figure 12: Processus de création du modèle et son déploiement…… p. 30

Figure 13: Echantillon des données de fichier (caesarian.arff) ……..…… p. 36

Figure 14: Importation de fichier caesarian.arff au weka……..…… p. 37

Figure 15: résultat de l'algorithme Random Tree……..…… p. 38

Figure 16: visualisation de l'algorithme Random Tree……...…… p. 38

Figure 17.1: model de classification……p. 39

Figure 17.2: model de classification…….…….. p. 39

Figure 18: Interface d'application……...…….. p. 40

Figure 19: exemple du cas d'accouchement normal…….…… p. 41

Figure 20: exemple du cas d'accouchement césarien……p. 41

Introduction générale

Introduction générale

Les systèmes d'information deviennent de plus en plus compliqués et diversifiés, notamment grâce à l'émergence des nouvelles technologies. La croissance continue du volume de données numériques ainsi que la diversité croissante des sources de données, toutes deux de plus en plus hétérogènes, Conjugué aux besoins pressants des entreprises d'utiliser ces données dans une méthode d'aide à la décision. La prise de décision a entraîné l'émergence de nouvelles problématiques, tout comme les technologies émergentes.

Le processus d'extraction des connaissances des données et de stockage des données est toujours en cours examiner et tenter de trouver des solutions à leurs problèmes. Cela nécessite le développement de nouvelles architectures, de l'intégration et des approches de sécurité, d'optimisation, d'interrogation et de modélisation L'extraction de connaissance à partir de données (ECD) est un terme utilisé dans la communauté de l'intelligence artificielle pour décrire le processus d'identification des structures inconnues, légitimes et potentiellement exploitables dans les bases de données.

L'ECD propose un cadre général dans lequel sont regroupées les méthodes permettant de traiter les questions d'organisation et d'utilisation des données, en particulier le stockage et la récupération des données. L'objectif du stockage des données est d'organiser des quantités massives de données, de les classer et de les préparer à l'analyse. Elle se concentre sur les processus d'extraction, de transformation et de chargement des données (ETC).

Les données sont généralement stockées dans des entrepôts de données spécialisés, appelés Entrepôts des Données (ED). L'objectif de la Fouille des Données (FD) est d'extraire des connaissances des données en utilisant soit des méthodes de structuration (apprentissage non supervisé), soit des méthodes d'explication (apprentissage supervisé), une fois les données acquises et préparées. Comme la FD est étroitement liée au processus d'ECD, la majorité des projets de recherche utilisent les deux termes de manière interchangeable. [1]

L'objectif de notre étude est de présenter l'ensemble des processus d'un DPE, à savoir : à partir d'une collection de données (des donnes sur des accouchements) obtenues à partir d'une base de données et analysées par un environnement d'apprentissage WEKA, ce dernier permettant de nous fournir le modèle souhaité à l'aide d'une interface graphique.

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 de**s** méthodes de classification.
- Chapitre 03 : est consacre a 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 l'interface 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.

Chapitre 01

Extraction de connaissance à partir de données

1. Introduction

Les technologies informatiques actuelles permettent, depuis au moins deux décennies, la production et le stockage de quantités massives de données numériques. Si ces données sont généralement collectées pour fournir un service ou répondre à une requête spécifique, elles comportent également une multitude de connaissances sur les objets qui y sont inscrits.

Cependant, comme ces éléments de connaissance sont disséminés dans une quantité massive de données souvent complexes, le problème de l'accès à ces connaissances dépasse largement les capacités analytiques humaines.

La disponibilité de grandes quantités de données, d'une part, et l'incapacité à les exploiter pleinement, d'autre part, ont favorisé le développement d'une nouvelle discipline scientifique, l'extraction de connaissances à partir de données, qui a vu le jour au début des années 1990.

2. L'extraction des connaissances a partir des données (ECD)

L'ECD est un processus permettant de découvrir de nouvelles informations dans un champ d'application donné. Il se compose de plusieurs étapes, dont la plus essentielle est la fouille de données (Data Mining). Chaque étape du processus vise à accomplir une tâche spécifique et se réalise par l'utilisation d'une ou plusieurs méthodes spécifiques.

Fayad définit l'ECD comme "un processus non trivial qui permet d'identifier, dans des données, des patterns ultimement compréhensibles, valides, nouveaux et potentiellement utiles " Il s'agit de la définition la plus largement utilisée dans la communauté de l'extraction et de la gestion des connaissances .

L'ECD pourrait être considéré comme un besoin imposé par la nécessité pour les entreprises de valoriser les données qu'elles rassemblent dans leurs bases de données. En effet, avec l'augmentation des capacités de stockage et des vitesses de transmission des réseaux, les utilisateurs ont accumulé une quantité croissante de données.

L'ECD évolue et continue à évoluer, par l'intersection des recherches menées dans plusieurs disciplines à la fois comme les bases de données, apprentissage automatique, reconnaissance de formes, statistique, intelligence artificielle et raisonnement incertain, acquisition de connaissances pour des systèmes experts, visualisation de données [2].

- **Le processus ECD**

Le terme "processus" fait référence au fait que l'ECD implique plusieurs étapes, notamment la collecte des données, la préparation des données, l'apprentissage automatique, et enfin l'interprétation et l'évaluation des modèles trouvés, toutes ces étapes étant répétées plusieurs fois. Le terme "qualification non triviale" fait référence au fait que les calculs effectués à chaque étape du processus ECD sont des calculs complexes nécessitant de multiples opérations de recherche et d'induction.

En plus, les modèles découverts par le processus ECD doivent être valides sur de nouvelles données, sous une forme compréhensible par les utilisateurs, et ces modèles doivent apporter quelque chose de nouveau et potentiellement utile à l'utilisateur. L'objectif du processus DPE est de parvenir à des informations nouvelles et fiables. L'objectif du processus ECD est de générer des nouvelles connaissances, valides et potentiellement utiles.

Après tout, la connaissance n'est qu'un modèle suffisamment intéressant et sûr. L'exactitude d'un modèle est déterminée par des mesures appelées "mesures de certitude" ou "critères de certitude", qui sont spécifiées par les utilisateurs ou les experts du domaine. L'utilisation des mesures de certitude en conjonction avec la validité, la nouveauté, l'utilité et la simplicité du modèle permet à l'utilisateur de ne conserver que les modèles interprétables dans les connaissances utiles et d'écarter le reste.

En conséquence, le processus ECD produit une collection de modèles qui ne sont pas tous interprétables en connaissances utilisables par l'utilisateur.

3. La fouille de données

Les termes "fouille de données" et "ECD" sont souvent confondus et utilisés de façon interchangeable. Cependant, la définition la plus largement acceptée de l'exploration de données la considère comme une composante importante du processus ECD. L'équivalent Anglais de ce mot est "data mining", qui est l'une des composantes du ECD, mais incontestablement la plus importante. C'est un domaine multidisciplinaire qui combine la technologie des bases de données, l'intelligence artificielle, l'apprentissage automatique, les réseaux neuronaux, les statistiques, la reconnaissance des formes, l'acquisition de connaissances, le calcul à haute performance et la visualisation des données.

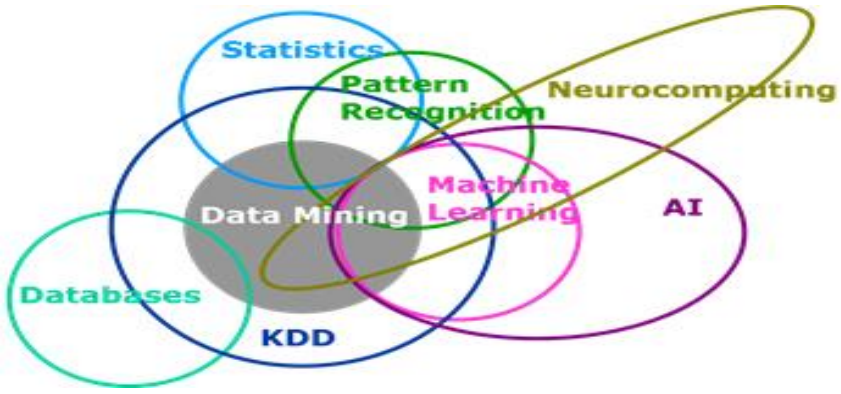


Figure 1 : La relation entre data mining et les autres technologies [3]

Elle permet d'extraire des informations de données réorganisées et préparées. Il est cependant nécessaire de passer par une étape d'évaluation avec l'aide d'un expert du domaine afin de déterminer la pertinence de ces données et leur contribution potentielle à la connaissance du métier des formes, de l'intelligence artificielle ou de l'apprentissage automatique.

Le processus de découverte de corrélations, de modèles et de tendances est connu sous le nom de data mining. Il peut être impartial et/ou interactif selon les objectifs à atteindre en analysant un vaste volume de données à l'aide de l'IA et de méthodes et techniques

mathématiques/statistiques pour : visualiser les données, corriger les données, découvrir des connaissances pour prendre de meilleures décisions [4]. Le terme "data mining" a été utilisé pour la première fois en 1990.

Les statistiques traditionnelles constituent le fondement des algorithmes de Data Mining : les concepts de distribution, de variance, d'analyse de régression, de déviation, d'analyse en grappes, d'analyse discriminante et d'intervalles de confiance sont la bible du domaine.

L'intelligence artificielle occupe la deuxième place. Ce domaine est basé sur l'heuristique plutôt que sur les statistiques, et il tente d'appliquer aux données un traitement qui se rapproche de la pensée humaine. L'IA n'a jamais connu l'ennui attendu en raison de la puissance de calcul phénoménologique requise. Le troisième et dernier grand domaine est l'apprentissage machine (ML), qui est un mariage entre les statistiques et l'IA. Cette discipline vise à donner à l'ordinateur la tâche d'apprendre par lui-même les données étudiées, en influençant le comportement et les décisions du programme.

4. Etapes du processus de l'ECD

Trois étapes principales peuvent être distinguées dans un processus d'ECD : préparation de données, application de techniques de data mining et interprétation des résultats; [4] ;[5].

De façon générale, le processus d'extraction de connaissances a partir de données illustré dans la Figure 2 consiste en une séquence itérative des étapes suivantes :

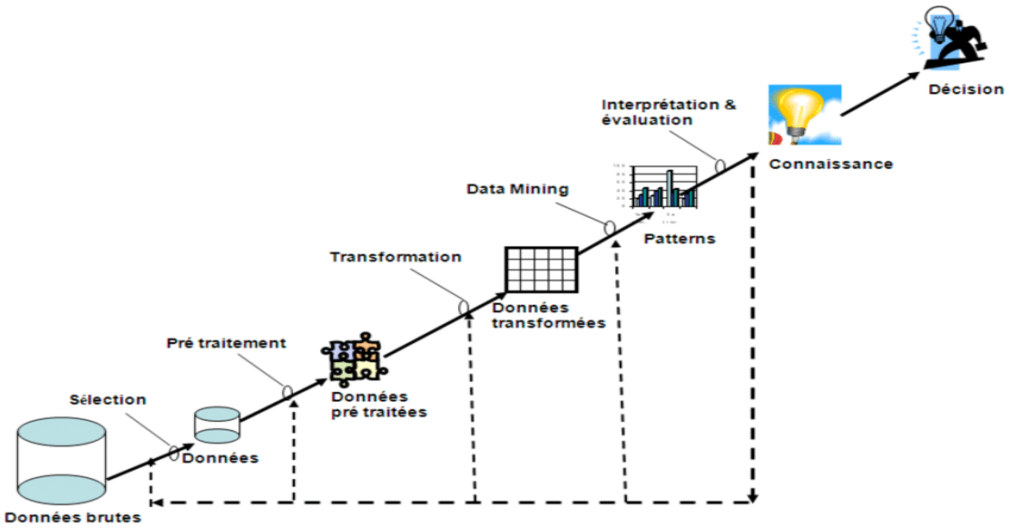


Figure 2 : Le processus d'ECD [6]

4.1. Préparation des données

La qualité des résultats d'un processus d'ECD est largement déterminée par la qualité des données utilisées, ce qui souligne la pertinence de l'étape de préparation des données [4].

Le prétraitement des données, selon [4], est toute activité effectuée sur les données avant l'application d'une approche de fouille de données (par exemple, la classification automatique, la classification et les règles d'association). Il s'agit essentiellement d'une transformation T qui

transforme les vecteurs de données brutes *Xik* en un ensemble de nouveaux vecteurs de données *Yij* qui sont plus utiles que *Xik*, en éliminant au moins un problème de *Xik* tout en préservant l'information. En fait, les données initiales peuvent être incomplètes, bruyantes, inconsistantes et incohérentes. Un processus de prétraitement et de préparation de données peut être décrit selon les étapes suivantes. [5]

- **La sélection**

Permet de sélectionner les données pertinentes pour la tâche de Data Mining sur laquelle vous travaillez. Ces données sont généralement issues de bases de données de production ou de référentiels de données. Les données sont organisées en champs typés (dans un domaine de définition).

- **Le nettoyage de données**

Cette tâche consiste en la détection et la suppression des erreurs, du bruit et de l'incohérence de données pour améliorer leur qualité [2] ;

- **L'intégration des données**

Cette tâche combine des données de sources multiples, détecte être sous des conflits de valeurs [5] ;

- **La transformation de données**

Cette tâche consiste à extraire/créer de nouvelles variables (features) afin de fournir une nouvelle représentation des données qui soit appropriée à l'application, au domaine et à l'objectif de l'étude. Plusieurs méthodes, telles que l'agrégation, la généralisation et la normalisation, permettent l'extraction de ces variables.

- **La réduction de données**

Le coût de la mesure et la précision des résultats de Data Mining sont les deux raisons les plus importantes pour maintenir le nombre de données aussi bas que possible. Cependant, cette réduction peut entraîner une perte d'informations. Il est donc nécessaire de trouver un accord. Nous mentionnons l'agglomération des données, la compression des données, la discrétisation des données et la génération de données comme stratégies de réduction des données.

4.2. Data Mining (fouille de données)

La fouille des données est l'étape cruciale du processus ECD, ainsi que son cœur. On peut dire que c'est la recherche de motifs valables qui, une fois validés, nous fournissent des connaissances exploitables.

4.3 Evaluation du résultat

Il s'agit de l'étape finale du processus d'extraction des connaissances à partir des données. Elle consiste à évaluer les résultats afin de déterminer si les modèles peuvent être considérés comme nouveaux et intéressants. Cette étape comprend également une interprétation des

résultats et une comparaison des modèles. En fait, la pertinence des connaissances nouvellement découvertes est évaluée à l'aide de critères de certitude fixés par les utilisateurs ou les experts du domaine.

Les modèles qui ont été identifiés comme des connaissances pertinentes seront ensuite testés sur d'autres ensembles de données ou sur différents systèmes. Les méthodes de validation seront déterminées par la nature de la tâche et du problème à résoudre.

La validation incombe principalement à l'expert, qui évaluera la pertinence des résultats à l'aide de méthodes et de critères qui seront utilisés en fonction des données utilisées et de l'objectif fixé au début de la méthode.

Pour valider les résultats d'une méthode d'exploration de données, les données sont séparées en trois ensembles : l'ensemble d'apprentissage (Training Data), l'ensemble de test (Testing Data) et l'ensemble de validation (Validation Data). 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.

5. Principales taches de fouille de données

La fouille de données est en fait un ensemble de techniques dédiées à différents tâches qui sont souvent divisées en deux catégories : les tâches descriptives et les autres tâches prédictives. [4]

Le premier groupe de tâches vise à décrire des phénomènes ou des tendances observées dans les données, tandis que le second groupe s'intéresse à l'estimation des valeurs futures des variables tout en tenant compte des valeurs historiques.

Présentons, dans les points suivants, les principales tâches que le DM est amené à accomplir, que nous avons résumé de [4], [5] (Figure 3).

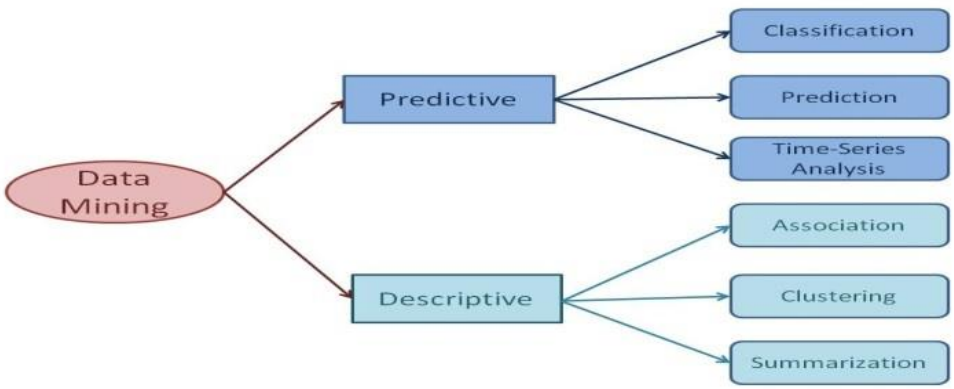


Figure 3 : Principales taches de fouille de données [7]

5.1. La description

Cette tâche vous permet de résumer les caractéristiques générales des objets dans un ensemble de données afin de générer des modèles sous forme de règles de caractérisation.

Les modèles doivent décrire des caractéristiques claires qui se prêtent à une interprétation et une explication intuitives.

Cette tâche peut être accomplie à l'aide d'une variété de stratégies qui diffèrent par leur degré de simplicité et de compréhension. Nous pouvons citer : analyse exploratoire, arbres de décision, méthodes de visualisation, réseaux neuronaux, etc.

5.2. La classification

Le but de la classification est d'identifier la classe d'objets appartenant à un groupe prédéfini, c'est- à - dire connu à l' avance, en examinant des propriétés ou des descripteurs particuliers. Exemples de tâches de catégorisation courantes : accorder ou refuser un crédit à un client en fonction de sa situation personnelle, établir un diagnostic médical en fonction de la description clinique d'un patient, lancer un processus d'alerte en fonction des signaux reçus par des capteurs, etc.

La classification est une tâche d'apprentissage supervisé dans laquelle le système apprend en généralisant une procédure de classification basée sur des exemples.

Une meilleure tâche de classification nécessitera une procédure avec un haut niveau de prédictibilité, c'est-à-dire la capacité de classer de nouveaux exemples qui n'ont pas participé au processus d'apprentissage. Par conséquent, la capacité de généralisation détermine la qualité de la catégorisation.

Cette tâche est fréquemment réalisée en deux étapes : une première étape d'apprentissage à partir d'exemples et de construction d'un modèle, suivie d'une seconde étape de test ou d'utilisation du modèle. Elle a fait l'objet de nombreuses études qui ont proposé diverses stratégies, chacune ayant ses propres avantages et inconvénients. Parmi les plus utilisées figurent les arbres de décision, les réseaux neuronaux, les classificateurs bayésiens naïfs et les algorithmes génétiques.

5.3. La segmentation

Similaire à la classification, la segmentation consiste à regrouper les objets entrants en groupes homogènes appelés segments, groupes ou clusters. Il s'agit d'une tâche d'apprentissage non supervisée car nous ne disposons d'aucune autre information que les descriptions des objets à segmenter. Par conséquent, les segments sont déduits automatiquement en fonction des données. Elle diffère de la classification en ce qu'il n'y a pas de segments prédéterminés.

Le regroupement des objets est basé sur les mesures de similitude existantes entre eux. L'idée est de maximiser ces mesures parmi les objets appartenant au même groupe et de les minimiser parmi ceux appartenant à des groupes différents. L'homogénéité interne et la séparation externe sont des termes utilisés de manière interchangeable. [5]

Cette similarité entre les objets s'exprime en termes d'une fonction de distance, où les objets à segmenter sont assimilés à des points de l'espace. Le choix de cette fonction est fait parmi plusieurs mesures disponibles, selon le type de données considérées (discrètes, réelles,…etc.) et le type de similarité recherchée. Rappelons qu'une distance doit vérifier les propriétés suivantes : (d(x, y) d est la distance entre les objets x et y)

1) d(x, y)>=0 ; 2) d(x, x)=0 ; 3) d(x, y)=d(y, x) ; 4) d(x, z) <=d(x, y) +d(y, z).

La distance euclidienne : $d(x,y)=\sqrt{\sum i(xi-yi)2}$
La distance de Manhattan : $d(x,y)=\sum i(\lvert xi-yi\rvert$

5.4. La recherche d'association

L'objectif d'une recherche d'association est de trouver des produits ou des éléments (également appelés "items") liés. Le type d'application de cette tâche est une analyse de panier, bien connue dans le monde du marketing, qui consiste à rechercher des corrélations possibles entre les produits commandés (souvent achetés ensemble) en examinant les enregistrements de transactions dans les bases de données des supermarchés.

Les règles d'association générées dans cette tâche sont les suivantes : si X, alors Y, alors note X Y, où X et Y sont des ensembles d'objets (souvent des produits). Les règles de ce type traduisent le fait que si les objets du groupe X sont présents dans une transaction, les objets de l'ensemble Y le sont aussi avec une certaine probabilité.

La recherche d'association peut être utilisée dans tout domaine où il est intéressant de trouver des associations d'objets ou des connexions de faits. Ce type d'analyse est largement utilisé dans divers domaines, notamment la banque, les télécommunications, la médecine, l'Internet, etc.

5.5. La prédiction

L'objectif de cette tâche est d'estimer les valeurs futures des variables, également appelées valeurs prédictives, à l'aide de divers algorithmes tout en tenant compte d'autres valeurs historiques des variables prédictives. Par exemple, avec cette tâche, on peut prévoir les valeurs futures des actions ou prédire, sur la base des actions précédentes, les départs des clients, etc.

Les prédictions sont réalisées à l'aide de diverses méthodes. Nous aborderons les méthodes statistiques telles que la régression (linéaire, multi variée ou logistique), les réseaux neuronaux, etc. [8]

5.6. La visualisation

La visualisation de données (également connue sous le nom de DataViz) est un terme général qui désigne tout effort visant à aider les gens à comprendre la signification des données en les plaçant dans un contexte visuel. Par conséquent, un logiciel de visualisation de données

peut aider à mettre en évidence et à identifier des modèles, des tendances et des corrélations qui pourraient autrement passer inaperçus dans des données textuelles.

Les outils actuels de visualisation des données vont au-delà des graphiques et des courbes typiques des feuilles de calcul Excel. Ils affichent les données de manière plus sophistiquée, notamment par le biais d'éléments infographiques, de cadrans et de jauges, de cartes géographiques, de graphiques de tendances (sparkline), de cartes thermiques (heat map) et de graphiques détaillés avec barres, secteurs et progression (fever chart).

Les images peuvent inclure des capacités interactives qui permettent à l'utilisateur de manipuler ou d'examiner les données à des fins d'interrogation et d'analyse.

Ces outils peuvent également inclure des indicateurs qui alertent l'utilisateur lorsque les données sont mises à jour ou lorsque des critères prédéfinis sont remplis. [9]

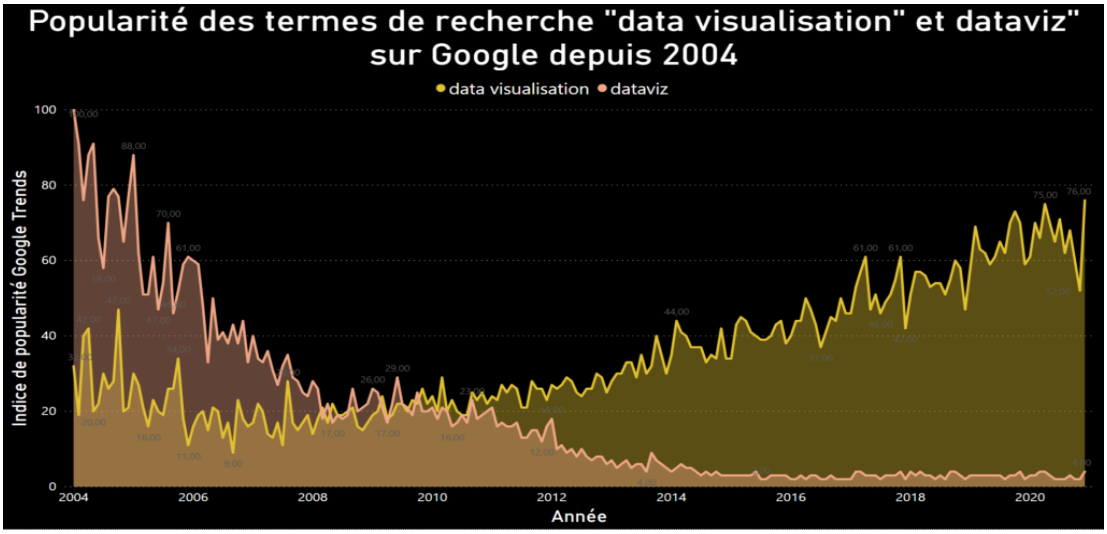


Figure 4 : Visualisation des termes de recherche « data visualisation » [10]

6. Domaine d'application

La technologie de data mining a une grande importance économique grâce aux possibilités qu'elle offre pour optimiser la gestion des ressources (humaines et matérielles). Les domaines d'application actuels du data mining sont les suivants :

➢ Scroring

Le scroring est une technique de marketing qui consiste à envoyer une note à un client ou à un prospect. L'objectif est de déterminer le profil du client par rapport à l'activité de l'entreprise, et ainsi de réduire le coût d'acquisition ou de fidélisation d'un client en concentrant les efforts marketing sur les profils les plus "réceptifs". Le scroring est utilisé dans de nombreux secteurs d'activité, notamment les assurances, les banques et les opérateurs téléphoniques. (Par exemple, refuser d'accorder un prêt à un client dont le profil de datamining indique un risque élevé de non-remboursement).

Par exemple, le datamining peut être utilisé pour déterminer les critères à prendre en compte pour considérer un client comme "réceptif". [11]

- **Prévention du crime**

Plusieurs expériences ont été menées dans ce domaine. Une utilisation aux USA a par exemple été d'identifier les associations de lieu et de plages horaires auxquelles les crimes se produisaient le plus, afin de renforcer la présence policière en conséquence. [11]

- **Détection de fraudes**

Dans les systèmes complexes comptant un grand nombre d'utilisateurs essentiels (comme les agences gouvernementales), un problème appelé fraude se pose fréquemment. La classification des données est utilisée dans le Data Mining. Ce mécanisme peut également être utilisé pour détecter des données qui sortent de l'ordinaire et qui n'auront pas le même impact qu'un comportement "normal". Certains comportements "normaux" peuvent s'écarter de la norme et donner lieu à des faux positifs lors de la détection d'une fraude. Il s'agit toutefois d'une méthode permettant d'identifier les cas de fraude potentiels. [11]

- **Le secteur bancaire**

En raison de ses vastes bases de données clients, le secteur bancaire est en tête de toutes les autres industries pour l'utilisation des outils d'exploration de données. Bien que les banques utilisent les outils d'analyse statistique avec un certain succès depuis plusieurs années, les modèles auparavant invisibles du comportement des clients deviennent plus clairs grâce aux nouveaux outils d'exploration de données.

Parmi les applications de l'exploration de données dans ce domaine, on peut citer :

- Prédire comment les clients réagiront aux changements de taux d'intérêt.

- Déterminer quels clients seront les plus réceptifs aux nouvelles offres de produits.

- Identifier les clients "fidèles".

- Déterminer quels clients sont les plus susceptibles de ne pas rembourser leurs prêts s'ils font une pause.[11]

- **La médecine**

Quelques exemples de l'usage médicaux et pharmaceutiques des techniques de Data Mining pour l'analyse de bases de données médicales.

• Prédiction de présence de maladies et/ou de complications.

• Le choix d'un traitement pour le cancer.

• Choix des antibiotiques pour des infections.

• Le choix d'une technique particulière (de sutures, matériel de suture, etc.) dans une des procédures chirurgicales. [11]

7. Conclusion

L'exploration de données est le processus d'extraction d'informations prédictives à partir de grands ensembles de données. Il s'agit d'une technologie nouvelle et puissante qui permet aux entreprises de se concentrer sur les données les plus importantes de leurs entrepôts de données. Les outils d'exploration de données peuvent prédire les tendances et les activités futures, ce qui vous permet de prendre les meilleures décisions possibles. C'est ce qui fait de l'exploration de données la technologie la plus cruciale.

Dans le prochain chapitre on va parler de la classification.

Chapitre 02

Classification

1. Introduction

Il est clair que le processus général de classification dans le domaine de l'informatique tente de l'appliquer à des données numériques (points, tableaux, images, sons, etc.), et que le travail général des méthodes de classification a été d'imiter et d'automatiser ce principe en utilisant et en inventant des outils appropriés (matériels-calculateurs, théories de classification, etc.).

Allons de ce principe, nous présenterons dans ce chapitre tout d'abord ce que c'est la classification, ses processus, et voir les critères pour une bonne classification, Ensuite nous allons présenter les méthodes de classifications.

2. Définitions

La classification est une tâche de fouille de données qui apparaît dans divers domaines, notamment la reconnaissance de formes, la reconnaissance de génomes, les statistiques, l'intelligence artificielle, l'aide à la décision multicritères, l'attribution de crédits, la prédiction de sites archéologiques, le diagnostic médical et la détection de fraudes fiscales, entre autres. En conséquence, les experts de ces domaines ont proposé plusieurs définitions :

- **Définition 01**

Selon Mari [12] "Effectuer une classification, c'est mettre en évidence des relations entre des objets, et leurs paramètres".

- **Définition 02**

Selon Henriet [22] "La classification consiste à affecter des objets, des candidats et des actions potentielles à des catégories ou des classes prédéfinies".

- **Définition 03**

"Le processus de classification cherche à mettre en évidence les dépendances implicites qui existent entre les objets, les classes entre elles, les classes et les instances. La classification recouvre les processus de reconnaissance de la classe d'un objet, et l'insertion éventuelle d'une classe dans une hiérarchie. Ce mode de raisonnement permet de reconnaître un objet en identifiant ses caractéristiques, relativement à la hiérarchie étudiée. La classification fait intervenir un processus de décision d'appartenance". [23]

Une classe est créée à partir d'un ensemble d'objets (attributs, caractéristiques ou critères) qui sont semblables entre eux et qui sont dissemblables à ceux d'autres classes. Les classes sont construites de façon à maximiser les similarités des objets qui sont dans la même classe et à minimiser les similarités de ces objets avec ceux des autres classes. Un « classificateur » est un algorithme qui, à partir d'un ensemble d'exemples, produit une prédiction de la classe de toute donnée.

➢ **Définition 04**

"La classification est l'action de regrouper en différentes catégories des objets ayant certains points communs ou faisant partie d'un même concept, sans avoir connaissance de la forme ni de la nature des classes au préalable, on parle alors de problème d'apprentissage non supervisé ou de classification automatique, ou l'action d'affecter des objets à des classes prédéfinies, on parle dans ce cas d'apprentissage supervisé ou de problème d'affectation".[24]

3. Le processus de classification

D'une manière générale, le processus de classification se divise en deux étapes principales :

3.1. Construction du modèle

La construction d'un modèle de classification est basée sur un ensemble d'exemples d'apprentissage (ensemble de formation). Chaque instance est supposée appartenir à une classe prédéfinie, la classe de cette instance étant déterminée par l'attribut classe. L'ensemble des instances d'apprentissage est utilisé pour former le modèle. Le modèle est représenté par des règles de classification, des arbres de décision, des formules mathématiques, des réseaux neuronaux, etc. Et le taux d'erreur du modèle sera calculé pour voir sa performance.

3.2. Utilisation du modèle

Lors de la classification des nouvelles instances, le modèle établi à l'étape précédente sera utilisé.

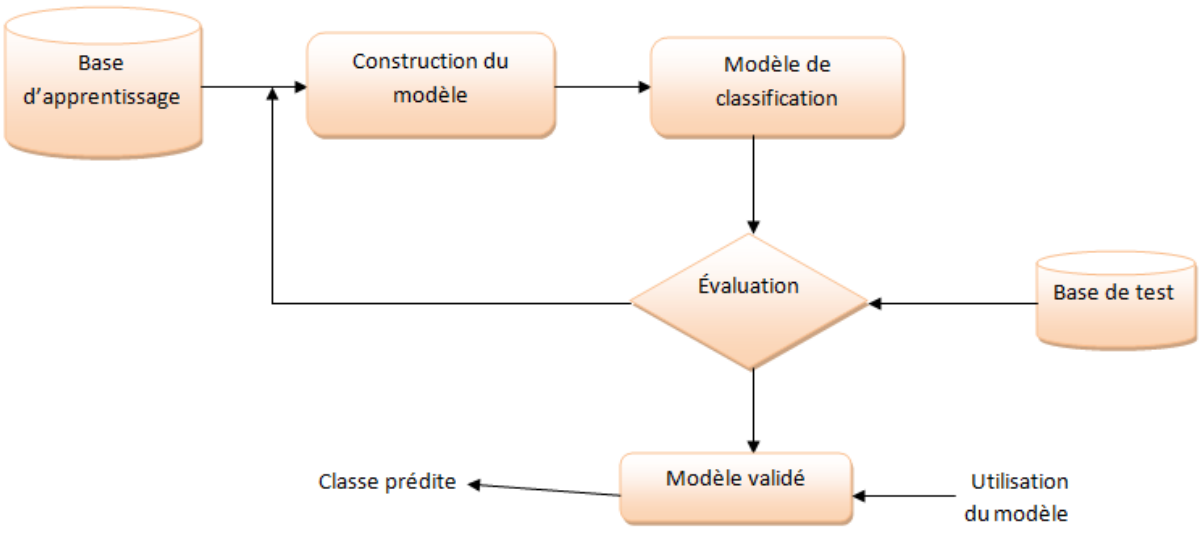


Figure 05 : Le processus de classification

4. Critères pour une bonne classification

L'objectif principal des techniques de classification est de trouver une partition dans laquelle les objets d'une même classe devraient être similaires (entre eux), tandis que les

objets de classes différentes devraient être distincts. Une bonne classification doit répondre aux critères suivants :

- **Validité**

Elle peut se définir par :

Chaque classe d'une partition doit être homogène : Les objets qui appartiennent à la même classe doivent être semblables. Les classes doivent être isolées entre elles : Les objets de différentes classes doivent être différents. La classification doit s'adapter aux données : La classification doit pouvoir expliquer la variation des données.

- **Interopérabilité**

Les classes doivent avoir une interprétation substantive c'est-à-dire qu'il est possible de donner des noms aux classes, dans le meilleur des cas les noms doivent correspondre aux types déduits d'une certaine théorie.

- **Stabilité**

Les classes doivent être stable ça veut dire que les petites modifications dans les données et dans les méthodes ne doivent pas changer les résultats.

- **D'autres critères**

Parfois la taille et le nombre de classes sont employés en tant que critères additionnels: le nombre de classes doit être aussi petit que possible, et la taille des classes ne doit pas être trop petite.

5. Les méthodes de classification

Pour les méthodes de classification, il existe deux approches principales : "supervisée" et "non supervisée".

Dans les deux cas, nous aurons besoin d'une mesure ou d'une caractéristique entre les données pour déterminer si elles appartiennent à une classe spécifique.

Dans le cas d'une classification supervisée, il y a également une phase de décision ou de généralisation.

Durant cette phase, si un nouvel élément est présent, le classificateur doit indiquer que cet élément appartient à quelle classe.

La Figure 06 : présente les différents types de méthodes, regroupés sous forme d'une hiérarchie par Jain et Dubes.

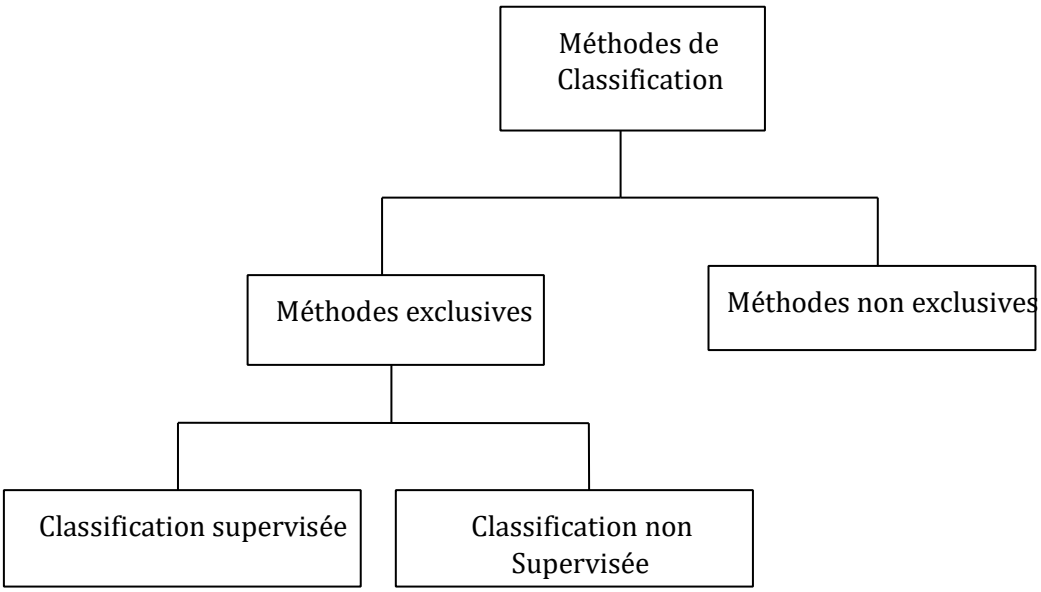


Figure 06 : Différents types de méthodes de classification

5.1. Classification supervisée

L'utilisation d'un ensemble d'exemples pour prédire la classification de nouvelles données est connue sous le nom de classification supervisée.

Il s'agit d'inventer une procédure de classification à partir d'un ensemble de données classées afin de prédire si un nouvel exemple appartiendra à une certaine classe.

Cette procédure de catégorisation est issue de recherches qui se déroulent dans un espace de modèles basés sur des hypothèses probabilistes, des concepts de proximité, des recherches structurelles, etc.

➢ Fonctionnement de la classification supervisée

Le fonctionnement de la classification supervisée se décompose en deux points :

- La phase d'apprentissage (modèle d'apprentissage) est celle où l'ensemble des informations apprises par l'algorithme d'apprentissage est représenté sous forme de règles de classification. Ces règles permettent de relier les objets à des classes de référence connues. L'algorithme de classification apprend du jeu d'apprentissage et construit le modèle.

  Il existe deux types de motivations pour l'apprentissage : La première est le raisonnement inductif, qui va du spécifique au général. Il implique de considérer le plus grand ensemble possible de règles de classification, puis de réduire cet ensemble à un plus petit ensemble de règles qui le résume au mieux. Le second raisonnement est déductif, du général au spécifique, et consiste à construire les règles une par une jusqu'à obtenir une bonne description de l'ensemble du processus d'apprentissage.

- La phase de test proprement dite, dans laquelle les données de test seront utilisées pour estimer la précision des règles de classification générées lors de la première phase. Si la précision du modèle est jugée satisfaisante, la règle peut être appliquée aux nouvelles

données. Le modèle d'apprentissage intégré est utilisé pour trier les nouveaux objets en catégories.

- **Les algorithmes de classification**

Parmi les algorithmes de classification supervisée les plus populaires dans la littérature On trouve: les arbres de décision, l'algorithme k-NN et les réseaux de neurones, Les réseaux bayésiens naïves.

5.1.1 Arbre de décision

Un arbre de décision est, comme son nom l'indique, un instrument de décision qui permet de classer un groupe d'objets en fonction de la valeur de leurs attributs. Voici une représentation graphique du processus de classification.

- Une feuille indique une classe ;
- Un nœud spécifie un test que doit subir un certain attribut ;
- Chaque branche sortant de ce nœud correspond à une valeur possible de l'attribut en question (Figure 07).

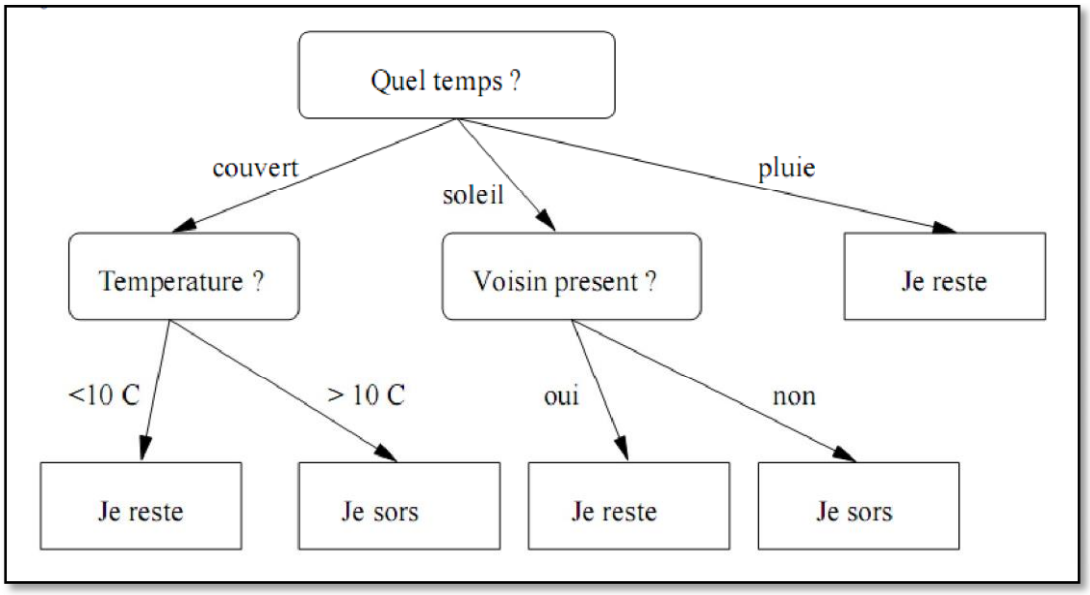


Figure 07: Arbre de décision [13]

Pour classer un nouvel objet, il faut suivre le chemin de la racine (nœud initial) à la feuille, en effectuant les différents tests d'attributs à chaque nœud. L'arbre permet de faire des prédictions sur les données en réduisant le domaine des solutions niveau par niveau.

La procédure générale de construction d'un arbre de décision se compose de deux étapes :

- Construction de l'arbre à partir des données (apprentissage) ;
- Elagage de l'arbre dans le but d'alléger l'arbre résultant souvent volumineux.

- **Construction de l'arbre**

Il existe une grande variété d'algorithmes pour construire des arbres de décision ; quelques-uns des plus répandus portent les noms de ID3 (*Inductive Decision-tree*) introduit par Quinlan et amélioré pour devenir C4.5, CART (*Classification And RegressionTrees)*, introduit par Breiman et al [14] et CHAID (*Chi-squaredAutomatic Interaction Detection*).

Le principe général part d'un arbre vide et procède à la construction de l'arbre de manière inductive (à partir des données) et récursive, en commençant par l'ensemble de la collection d'objets. Si tous les objets appartiennent à la même classe, une étiquette portant le nom de la classe est créée. Dans le cas contraire, la collection entière d'objets est divisée en sous-groupes sur la base de la valeur d'un seul attribut, qui sont tous soumis au même traitement.

Il ne suffit pas de traiter les attributs de manière séquentielle pour obtenir un arbre de décision concis et suffisant. Toute la richesse des arbres de décision, en revanche, vient de la sélection judicieuse des attributs éblouissants afin d'arriver au plus grand nombre d'objets d'une même classe par le chemin le plus court et le plus nécessaire.

Le choix des attributs peut se faire par plusieurs techniques, entre autres :

- Entropie (ID3, C4.5) [14].
- Indice de Gini (CART) [14].
- Table de Khi-2 (CHAID) [13].

- **Elagage de l'arbre**

L'opération d'élagage des arbres se divise en deux étapes : le pré élagage et le post élagage.

Le pré élagage consiste à établir un critère d'arrêt qui permet d'interrompre la construction de l'arbre pendant le processus de construction.

Le terme "post-élagage" désigne un traitement qui intervient après la construction de l'ensemble de l'arbre. Il s'agit de supprimer les sous arbres qui n'améliorent pas l'erreur de classification.

- **Les domaines d'application**

Cette méthode peut être utilisée dans plusieurs domaines tels que:

Les études (pour comprendre les critères prépondérants dans l'achat d'un produit, l'impactdes dépenses publicitaires), les ventes (pour analyser les performances par région, par enseigne, par vendeur), l'analyse de risques (pour détecter les facteurs prédictifs d'un comportement de non-paiement), Le domaine médical (pour étudier les rapports existant entre certaines maladies et des particularités physiologiques ou sociologiques) [15].

- **Critiques de la méthode**

❖ Les avantages

Les arbres de décision constituent un moyen très efficace de classification, et ce pour les avantages qu'elles présentent. Parmi ses avantages, on peut citer [14]:

- Facilité à manipuler des données catégoriques.
- Traitement facile des variables d'amplitudes très différentes.
- La classe associée à chaque individu peut être justifiée.
- Les attributs apparaissant dans l'arbre sont des attributs pertinents.
- Pour le problème de classification considéré.

❖ Les Inconvénients

Ces méthodes présentent tout de même des inconvénients dont les plus importants sont :

- La sensibilité au bruit et aux points aberrants.
- La sensibilité au nombre de classes (plus le nombre de classes est grand plus les performances diminuent).
- Le besoin de refaire l'apprentissage si les données évoluent dans le temps.

▪ Plus proches voisins (Classifieur Knn)

k-NN (k Nearest Neighbors) est un système de raisonnement à base de cas qui permet de prendre des décisions en recherchant des cas similaires déjà résolus. La décision consiste à rechercher les k échantillons les plus proches de l'objet et à affecter la classe la plus représentative dans ces k échantillons ("dis-moi qui sont tes amis, et je te dirai qui tu es").

L'approche la plus simple est de rechercher le cas le plus similaire et de prendre la même décision, on parle de 1-NN. Si cette approche peut fournir des résultats acceptables sur des problèmes simples pour lesquels les objets sont bien répartis en groupes denses de même classe, en règle générale, il faut considérer un nombre de voisin plus important pour obtenir de bons résultats.

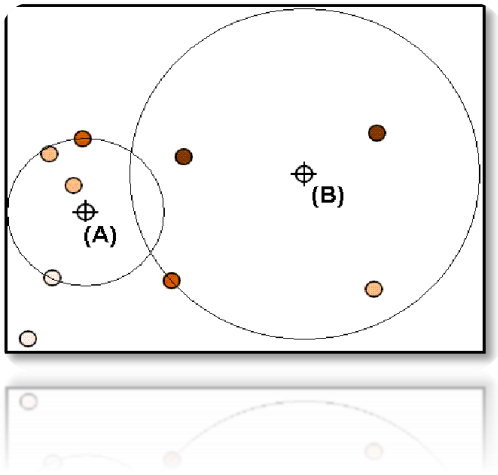


Figure 08: Exemple d'application de l'algorithme k-NN

L'algorithme k-NN a les avantages de :

Lors de l'introduction de nouveaux attributs, ne pas répéter le processus d'apprentissage ; Gérer tout type de données avec un grand nombre d'attributs ; Fournir des résultats clairs.

Cependant, le coût de la classification peut être élevé car :

- D'une part, le temps nécessaire pour calculer les voisins peut être prohibitif ;
- D'autre part, il faut toujours stocker le modèle durant toute l'opération de classification.

▪ Les réseaux de neurones

Un réseau de neurones est un système composé de plusieurs unités de calcul simples (nœuds) qui fonctionnent en parallèle, la fonction étant déterminée par la topologie du réseau et l'opération effectuée par les nœuds.

Le principe de fonctionnement est le suivant : on dispose d'une base de connaissances initiale constituée de couples de données (entrées / sorties), et on souhaite utiliser cette base de connaissances pour entraîner un algorithme à reproduire les liens découverts entre les entrées et les sorties de l'échantillon.

L'exemple le plus élémentaire de réseau de neurones est le "perceptron multicouches", qui est un cas particulier de réseau de neurones. (*Figure **09***).

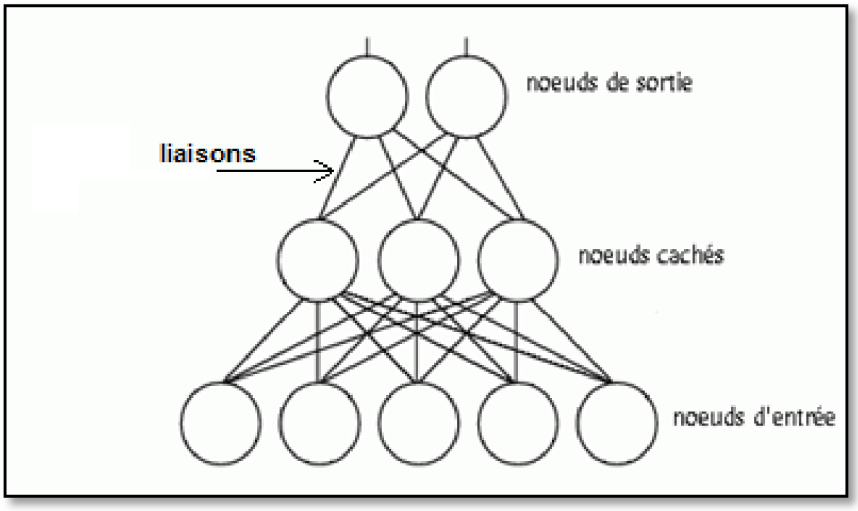


Figure 09 : Perceptron à trois couches (schéma type)

Pour un réseau de neurones avec *N* nœuds d'entrée, notées C(1),...,C(N) , et N poids affectées aux liaisons et notés w(1),...,w(N) l'entrée d'un nœud de la couche suivante sera généralement une somme pondérée des valeur de sortie des neurones précédents :

$$X = w(1)*C(1) + w(2)*C(2) + w(3)*C(3) + ... + w(N)*C(N)$$

Les poids sont des paramètres adaptatifs, dont la valeur est à déterminer en fonction du problème via un algorithme d'apprentissage (propagation, rétro-propagation…) [14].

Les réseaux de neurones peuvent être utilisés pour effectuer une classification supervisée floue de la manière suivante : chaque nœud d'entrée correspond à un attribut de l'objet (autant de nœuds d'entréeque d'attributs).

On peut prendre un neurone de sortie par classe ; la valeur de sortie est la valeur de la fonction d'appartenance (probabilité que l'objet appartienne à cette classe) [14]

- **Classificateur naïve bayésienne**

La classification bayésienne naïf est une classification probabiliste bayésienne simple basée sur le théorème de Bayes avec une forte indépendance (naïf) des hypothèses. Elle utilise un classificateur bayésien naïf, également appelé classificateur bayésien naïf, qui appartient à la famille des classificateurs de Linéaire.

En clair, un classificateur bayésien naïf estime que l'existence d'une caractéristique pour une classe n'est pas liée à l'existence d'autres caractéristiques. Si un fruit est rouge, arrondi, et mesure quelques centimètres de diamètre, il est considéré comme une pomme. Même si ces caractéristiques sont liées dans la réalité, un classificateur bayésien naïf déterminera que le fruit est une pomme sur la base des seules caractéristiques de couleur, de forme et de taille.

Les classificateurs bayésiens naïfs peuvent être entraînés efficacement dans un environnement d'apprentissage supervisé, selon la nature de chaque modèle probabiliste.

Ce classificateur est basé sur le théorème de Bayes, qui permet de calculer des probabilités conditionnelles. Ce théorème permet de calculer la probabilité conditionnelle d'une cause connaissant la présence d'un effet, sur la base de la probabilité conditionnelle de l'effet connaissant la présence de la cause et des probabilités a priori de la cause et de l'effet.

- **Critiques de la méthode**

  - **Avantages**
    - La facilité et la simplicité de leur implémentation.
    - Leur rapidité.
    - Les méthodes Naïve Bayes donnent de bons résultats.

  - **Inconvénients**
    - Performances limitées quand il s'agit d'une grande quantité à traiter.
    - Modèle qualifié de naïve ou simple à cause de l'hypothèse d'indépendance.

5.2. Classification non-supervisée

Les classes ne sont pas connues a priori dans la classification non supervisée, également appelée segmentation (clustering en anglais).

Elles sont construites en utilisant certaines règles ou critères de regroupement qui dépendent des données disponibles à un moment donné.

Comme les classes sont généralement basées sur la structure des données, il est plus difficile de déterminer la sémantique associée à chaque classe.

Cette fois, le but n'est pas d'estimer une fonction, mais de regrouper des objets qui partagent une caractéristique commune ; les objets utilisés comme données d'apprentissage sont présentés sans leurs catégories.

- **K-means**

J. MacQuenn a introduit l'algorithme k-Means (également connu sous le nom d'algorithme du centre mobile) et E. Forgy l'a mis en œuvre dans sa forme actuelle [16]. Il est le plus couramment utilisé dans les applications scientifiques et industrielles car il est le plus simple et le plus rapide.

Dans cet algorithme, chaque classe est représentée par la moyenne (centroide). k-Means est un algorithme itératif.

Il commence avec un ensemble de *k* individus de référence choisis de façon aléatoire.

Les individus de données sont ainsi partitionnés dans *k* classes ; un individu appartient à une classe si le centre de cette classe est le plus proche de lui (en terme de distance).

La mise à jour des centroides et l'affectation des individus de données aux classes sont réalisées pendant les itérations successives.

La Figure 10 montre un exemple de déroulement de l'algorithme k-Means sur un nuage d'objets bidimensionnels, avec $k = 3$.

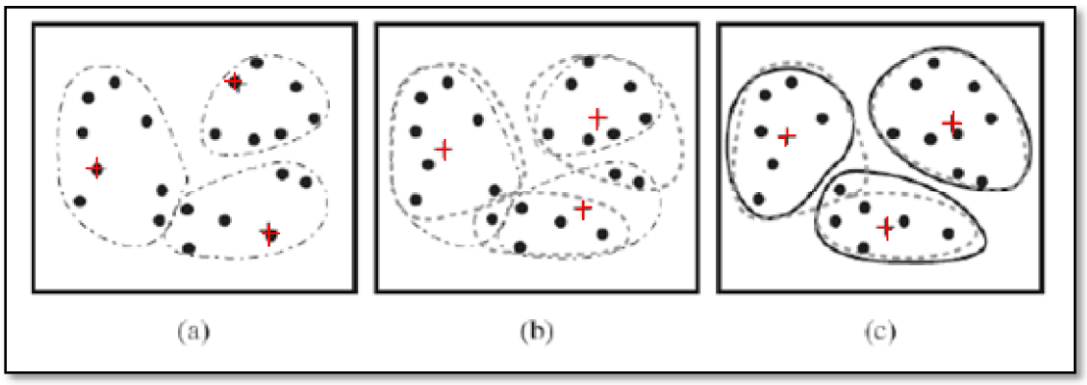


Figure 10 : Partitionnement basé sur k-Means [16].

Comme avantages de cet algorithme, on cite :

Il s'adapte bien pour des populations de tailles importantes ;

Il est relativement efficace ; si *n* le nombre d'objets, *k* le nombre de classes et *t* le nombre d'itération, l'algorithme converge, généralement, avec *k* et *t* suffisamment inférieur à *n* ($k, t << n$) ;

Il est indépendant de l'ordre d'arrivée des données.

Parmi les inconvénients de cet algorithme, on cite:

- ✓ Il est applicable seulement dans le cas où la moyenne des objets est définie ; o Le nombre de classes *k*, doit être spécifié à priori ;
- ✓ Il converge souvent vers un optimum local ;
- ✓ Il est sensible aux objets isolés (bruits) ;
- ✓ Il n'est pas adapté pour découvrir des classes avec des structures non-convexes, et des classes de tailles et formes différentes.

- **L'algorithme k-medoids**

Kaufman et Rousseew [16] fournissent une introduction. L'esquisse de l'algorithme k-Medoids est similaire à celle de k-Means, sauf que, contrairement à k-Means, où une classe est représentée par une valeur médiane, le centroide, dans k-Medoids, une classe est représentée par un de ses objets dominants, le médoide.

L'algorithme k-Medoids utilise une fonction objective qui définit la distance moyenne entre un objet et le médoide.

La Figure 11 : est une illustration du déroulement de l'algorithme k-Medoids sur un nuage d'objets bidimensionnels avec $k = 3$.

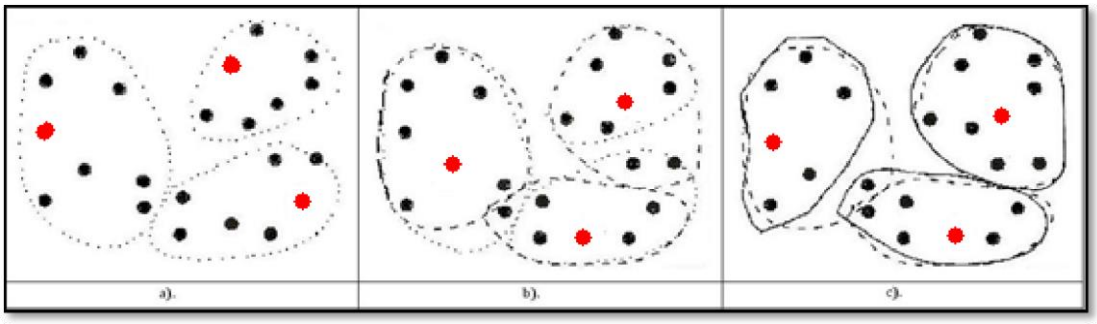


Figure 11 : Partitionnement basé sur k-Medoids [16].

Figure 11.a : Sélection des trois centres initiaux () et affectation de chaque objet restant dans le centre le plus proche

Figure 11.b : Calcul des nouveaux médoïdes pour chaque classe et redistribution des objets en fonction des nouveaux médoïdes (le médoïde étant l'objet le plus proche du centroides de la classe).

Figure 11.c: La technique est répétée jusqu'à ce que les classes soient stables.

- **L'algorithme PAM**

PAM (*PartitioningaroundMedoids*) a été développée par Kaufman et Rousseeuw [20]. L'idée de cet algorithme consiste à commencer avec un ensemble de k-médoides puis échanger le rôle entre un objet médoide et un non-médoide si cela permet de réduire la distance globale, ce qui revient à minimiser la fonction objectif [16].

Le principal inconvénient de cet algorithme est son coût total de calcul ; il a une complexité quadratique de l'ordre de O(k.(n-k)2) pour chaque itération, avec n le nombre d'objets et k le nombre de classes, ce qui le rend inadapté à un grand nombre d'objets.

6. Evaluation des méthodes de classification

La meilleure façon de résoudre un problème de classification est de comparer différents modèles et de choisir celui qui correspond le mieux à la situation. Par conséquent, l'évaluation des modèles est un préalable nécessaire à la classification. Elle est nécessaire pour comprendre la performance d'un modèle et déterminer si elle est statistiquement significative. Pendant ce temps, deux objectifs se dégagent : l'évaluation et la comparaison des modèles en vue de leur sélection. Voici quelques sélections possibles de modèles :

• 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.

7. Les étapes d'une classification

- **Choix des données**

Il est nécessaire de choisir les individus à classer ainsi que les caractéristiques qui seront utilisées comme critères de classification. Si vous voulez faire une classification basée sur le lieu de résidence, par exemple, stratifiez le fichier original en deux sous-fichiers.

On choisit les individus qui vivent en zone urbaine et ceux qui vivent en zone rurale, et on effectue la classification sur chacun de ces deux sous-fichiers.

- **Choix d'un algorithme de classification et exécution**

Sélection d'un algorithme en fonction des exigences (performances, représentation de la sortie, types de données) et ajustement des paramètres éventuels.

- **L'interprétation des résultats**

- évaluation de la qualité de la classification.
- description des classes obtenues.

8. Conclusion

La classification retient aujourd'hui l'attention des chercheurs. Plusieurs études ont été menées dans le but d'inventer et d'améliorer des méthodes de classification plus efficaces. On travaille toujours pour améliorer la classification et réduire la complexité.

Ces méthodes sont temporaires. Avec l'avancement rapide du matériel informatique, le second avantage semble être moins important maintenant, et la qualité et l'équité du produit semblent être plus importantes. Le dernier objectif est la classification.

Dans ce chapitre le principe de la classification a été présenté, ainsi que les méthodes utilisées pour évaluer la qualité de la classification.

Dans le chapitre suivant on va discuter de l’approche par arbre de décision pour déterminer le type d’accouchement d’une femme.

Chapitre 03

Approche d'arbre de décision pour déterminer le type d'accouchement

1. Introduction

Dans ce chapitre on va définir l'accouchement et donner leurs types, puis on va spécifier la méthode de classification utilisée dans notre système

2. Les données expérimentables

L'accouchement est la phase ultime de la grossesse, celle qui permet la sortie du nouveau-né. Il se déroule en 3 phases:

Le travail avec les contractions de l'utérus qui permettent la dilatation du col de l'utérus nécessaire à la sortie du bébé. Celui ci se présentant la tête en bas ou en siège.

L'expulsion est la sortie de l'enfant, l'accouchement proprement dit.

La délivrance survient à distance de la sortie du bébé, à la suite de nouvelles contractions utérines qui aboutissent à l'expulsion du placenta. [17]

- **Types d'accouchement**

Il y'a deux types d'accouchement :

- **Accouchement normale**

Dans sa définition la plus large, l'accouchement normal comporte un travail qui se déclenche de façon spontanée, généralement entre la 37^{e} et la 42^{e} semaine de gestation. L'accouchement normal comporte également le contact peau à peau après l'accouchement de même que l'allaitement dans la première heure de vie. [18]

- **Accouchement césarienne**

Lorsque l'accouchement ne se passe pas comme prévu, qu'il présente un risque pour la mère ou pour le bébé, la césarienne est parfois incontournable. Aujourd'hui bien maîtrisée, cette opération présente peu de risque. A tel point que certains dénoncent une hausse de césariennes "de confort". Le point sur cette méthode d'accouchement particulière. [18]

3. Le Processus de fouille de données par classification (Random Tree) pour déterminer le type d'opération médicale

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 opérations césarienne pour qu'elles soient analyser par un environnement d'apprentissage WEKA, ce dernier va nous donner le modèle attendu qui est l'arbre aléatoire.

Le schéma suivant explique les différentes étapes de cette approche :

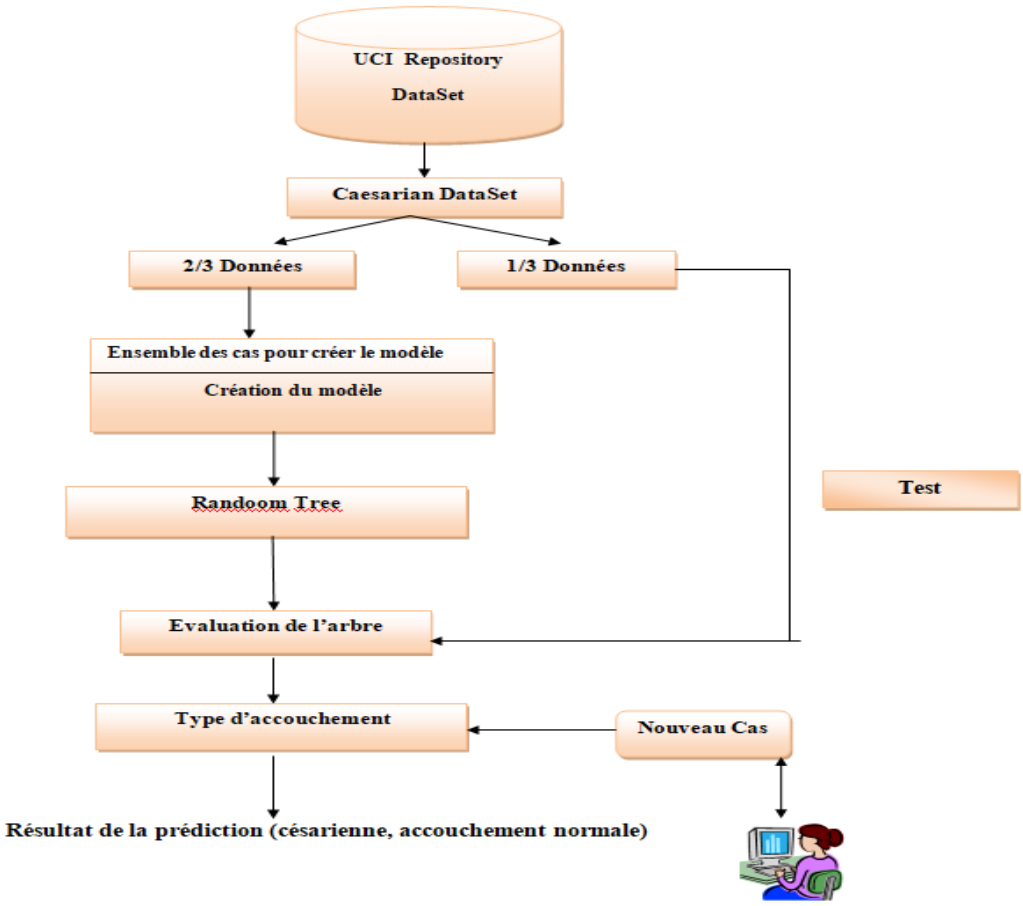


Figure 12 : Processus de création du modèle et son déploiement

Fonctionnement détaillé de l'approche

- **L'apprentissage**

Il s'agit de l'étape consistant à extraire de nouvelles informations d'un ensemble donné de données en utilisant diverses méthodes et stratégies dédiées à la classification.

Cette étape est réalisée dans l'environnement Weka 3.8, qui comprend une gamme d'outils de visualisation des données et d'algorithmes d'analyse des données et de modélisation prédictive.

Weka est un ensemble de logiciels d'apprentissage automatique écrits en Java et développés à l'Université de Waikato en Nouvelle-Zélande. Il s'agit d'un logiciel libre disponible sous la licence publique générale GNU.

La méthode de RandomTree est la principale méthode utilisée dans notre approche pour extraire l'arbre de décision.

Les deux tiers (2/3) des exemples sélectionnés sont destinés à la création du modèle, tandis que les autres exemples sont destinés à l'évaluation des performances.

- **L'Evaluation et Test**

Il s'agit de l'étape finale du processus d'Extraction de Connaissances à partir de Données. Elle consiste à évaluer les résultats de Weka afin de déterminer quels modèles

peuvent être considérés comme nouveaux et intéressants. Cette étape comprend également une interprétation des résultats et une comparaison des modèles. En fait, la pertinence des connaissances nouvellement découvertes est évaluée à l'aide de critères de certitude fixés par les utilisateurs ou les experts du domaine. Les modèles qui ont été identifiés comme des connaissances pertinentes seront ensuite testés sur d'autres ensembles de données ou sur différents systèmes.

Les méthodes de validation seront déterminées par la nature de la tâche et du problème à résoudre.

Le processus de validation relève principalement de la responsabilité du développeur, qui évaluera la pertinence des résultats à l'aide de méthodes et de critères qui seront utilisés en fonction des données utilisées et de l'objectif fixé au début de la méthode.

Pour valider les résultats d'une méthode d'exploration de données, les données sont divisées en trois ensembles distincts : les données d'apprentissage (Training Data), les données de test (Testing Data) et les données 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.

Par conséquent, tout en testant de nombreux modèles et en les comparant, on peut choisir le meilleur modèle sur la base des résultats globaux de ses tests, puis évaluer son caractère réellement erroné tout au long du processus de validation.

4. Le processus général de classification pour déterminer le type d'opération médicale

- **La méthode de Random Tree**

La classification est un algorithme d'extraction des connaissances a partir de données qui crée un guide étape par étape pour déterminer le résultat d'une nouvelle instance de données. L'arbre qu'il crée est exactement cela : un arbre où chaque nœud de l'arbre représente un endroit où une décision doit être prise en fonction de l'entrée. décision doit être prise en fonction de l'entrée, et de passer au nœud suivant et au suivant jusqu'à ce qu'une décision soit prise .le nœud suivant et le suivant jusqu'à ce que l'on atteigne une feuille qui indique la sortie prédite. Cela semble déroutant, mais c'est en fait assez simple.

Il y a également un débat pour savoir si les méthodes de classification de classification qui n'impliquent pas de modèle statistique peuvent être considérées comme "statistiques". D'autres domaines peuvent utiliser une d'autres domaines peuvent utiliser une terminologie différente.

L'écologie des communautés, le terme "classification" se réfère normalement à l'analyse des groupes, c'est-à-dire à un type d'apprentissage non supervisé, plutôt qu'à l'apprentissage supervisé. [19]

Random Tree

Les étapes spécifiques de la mise en œuvre de le Random Tree sont les suivantes :

L'algorithme Random Tree est représenté par E(K, X, D) où X représente un échantillon de données brutes, et K représente le nombre d'arbres de décision. Chaque arbre de décision produit un résultat de prédiction basé sur l'entrée de l'échantillon X = {x1x2, . . . xM} et obtient finalement une décision de classification selon la règle de vote.

1- Dans le modèle de classification de Random Tree, chaque classificateur de base utilise tous les échantillons d'entraînement (échantillons OOB) pour l'entraînement, en supposant que l'ensemble de données d'origine, D, le nombre d'échantillons, N, et le nombre de caractéristiques, M.

2- Générer un arbre de décision selon l'algorithme CART (Classification and Regression Tree). Dans le processus de division des nœuds, M Attributs sont sélectionnées de manière aléatoire parmi les M caractéristiques de chaque nœud de division. certaines catégories sont sélectionnées au hasard et placées dans une branche, et les catégories restantes sont placés dans une autre branche. Pendant ce temps, la valeur de division optimale de chaque nœud est calculée, et le fractionnement optimal de l'attribut est sélectionné ; aucune opération d'élagage n'est effectuée au cours du processus de fractionnement. Diviser de manière itérative les sous-ensembles à une valeur actuelle pour générer un arbre de décision.

3- Répétez les étapes (1) et (2) pendant K fois, et finalement, un modèle de Random Tree composé de K arbres de décision est généré.

4- Tester le modèle de Random Tree, formé à l'aide de données de test, et enfin générer le résultat de classification final par vote [20].

5. Conclusion

Dans ce chapitre nous avons utilisé 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 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.

Chapitre 04

Résultat d'évaluation et leur discussion

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 a partir des données d'accouchement.

2. Outil et environnement de développement

- **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 à 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.

- **Eclipse IDE**

Eclipse IDE est un environnement de développement intégré libre (le terme *Eclipse* désigne également le projet correspondant, lancé par IBM) extensible, universel et polyvalent, permettant potentiellement de créer des projets de développement mettant en œuvre n'importe quel langage de programmation. Eclipse IDE est principalement écrit en Java (à l'aide de la bibliothèque graphique SWT, d'IBM), et ce langage, grâce à des bibliothèques spécifiques, est également utilisé pour écrire des extensions.

La spécificité d'Eclipse IDE vient du fait de son architecture totalement développée autour de la notion de plug-in (en conformité avec la norme OSGi) : toutes les fonctionnalités de cet atelier logiciel sont développées en tant que plug-in.

Plusieurs logiciels commerciaux sont basés sur ce logiciel libre, comme par exemple IBM Lotus Notes 8, IBM Symphony ou Websphere Studio Application Developer.

3. Le Domaine d'application

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

Le domaine d'application de notre système est utilisé dans le domaine médical.

4. DataSet

- **Résume**

Cet ensemble de données contient des informations sur les résultats des césariennes de 80 femmes enceintes présentant les caractéristiques les plus importantes des problèmes d'accouchement dans le domaine médical. [21]

- **Informations sur les attributs**

Nous choisissons l'âge, le nombre d'accouchement, le temps d'accouchement, la tension artérielle et l'état cardiaque.

Nous classons le temps d'accouchement en trois catégories : prématuré, opportun et tardif. Nous considérons la tension artérielle selon trois états : faible, normal et élevé. Le problème cardiaque est classé comme apte et inepte. [21]

```
@attribute 'Age' { 22,26,28,27,32,36,33,23,20,29,25,37,24,18,30,40,31,19,21,35,17,38 }
@attribute 'Delivery number' { 1,2,3,4 }
@attribute 'Delivery time' { 0,1,2 }
@attribute 'Blood of Pressure' { 2,1,0 }

@attribute 'Heart Problem' { 1,0 }
@attribute Caesarian { 0,1 }


@data


22,1,0,2,0,0
26,2,0,1,0,1
26,2,1,1,0,0
28,1,0,2,0,0
22,2,0,1,0,1
26,1,1,0,0,0
27,2,0,1,0,0
32,3,0,1,0,1
28,2,0,1,0,0
27,1,1,1,0,1
36,1,0,1,0,0
33,1,1,0,0,1
23,1,1,1,0,0
20,1,0,1,1,0
29,1,2,0,1,1
25,1,2,0,0,0
25,1,0,1,0,0
20,1,2,2,0,1
```

Figure 13 : Echantillon des données de fichier (caesarian.arff)

5. L'environnement de l'expérimentation

- **Le logiciel WEKA**

Les 2/3 des données (53 données) sont les données destinées à la création de modèle sous le logiciel d'apprentissage automatique WEKA, ces données sont importées du DataSet au WEKA

Ces données (126 instances) vont être importées au WEKA avec ses 6 attributs (voire figure 14).

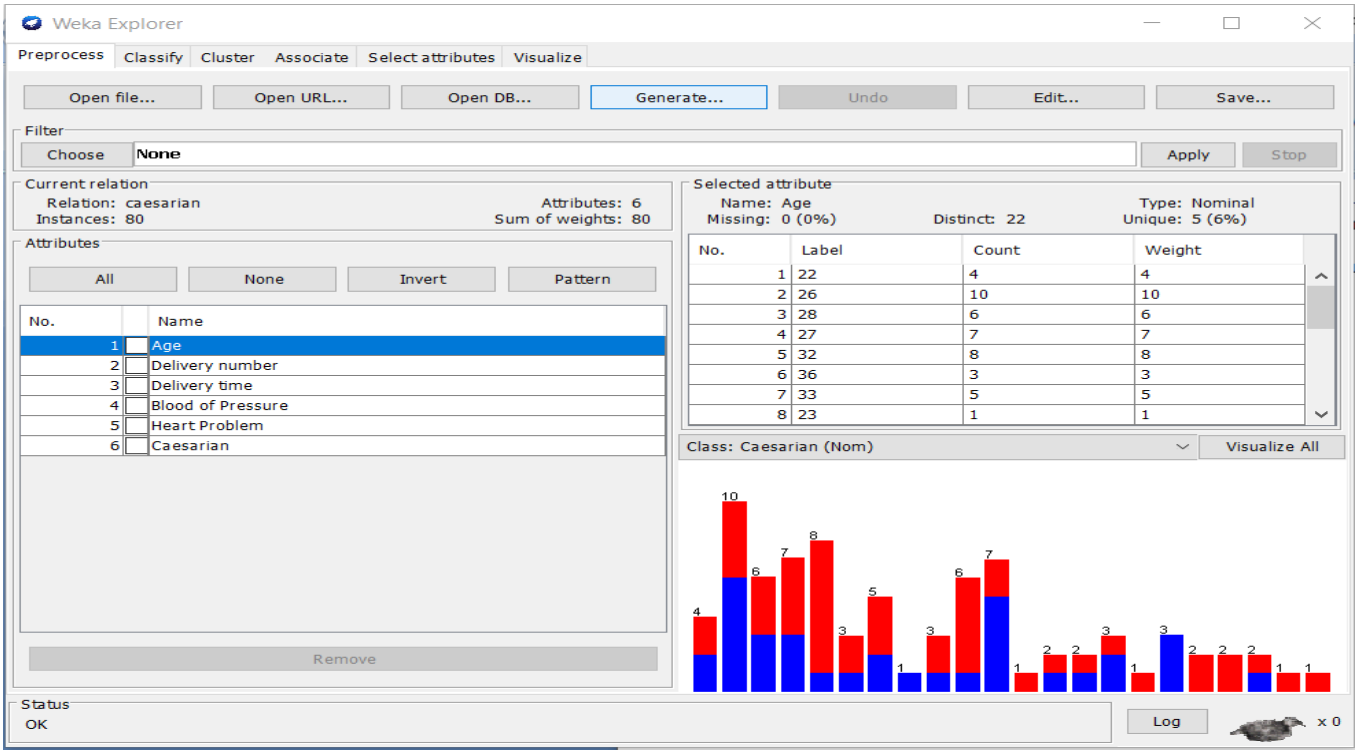


Figure 14 : Importation de fichier caesarian.arff au weka

La tache choisit dans notre expérimentation est « classify » utilisant l'algorithme d'arbre aléatoire

L'attribut caesarian est un attribut qui ne rentre pas dans le calcule car elle est la classe de prédiction, que 5 attributs suffisant pour la création de modèle

Les Résultats apparu par weka est deux classe (0,1) (figure 15)

La classe 0 ⟶ Accouchement normale.

La classe 1⟶ Accouchement par césarienne.

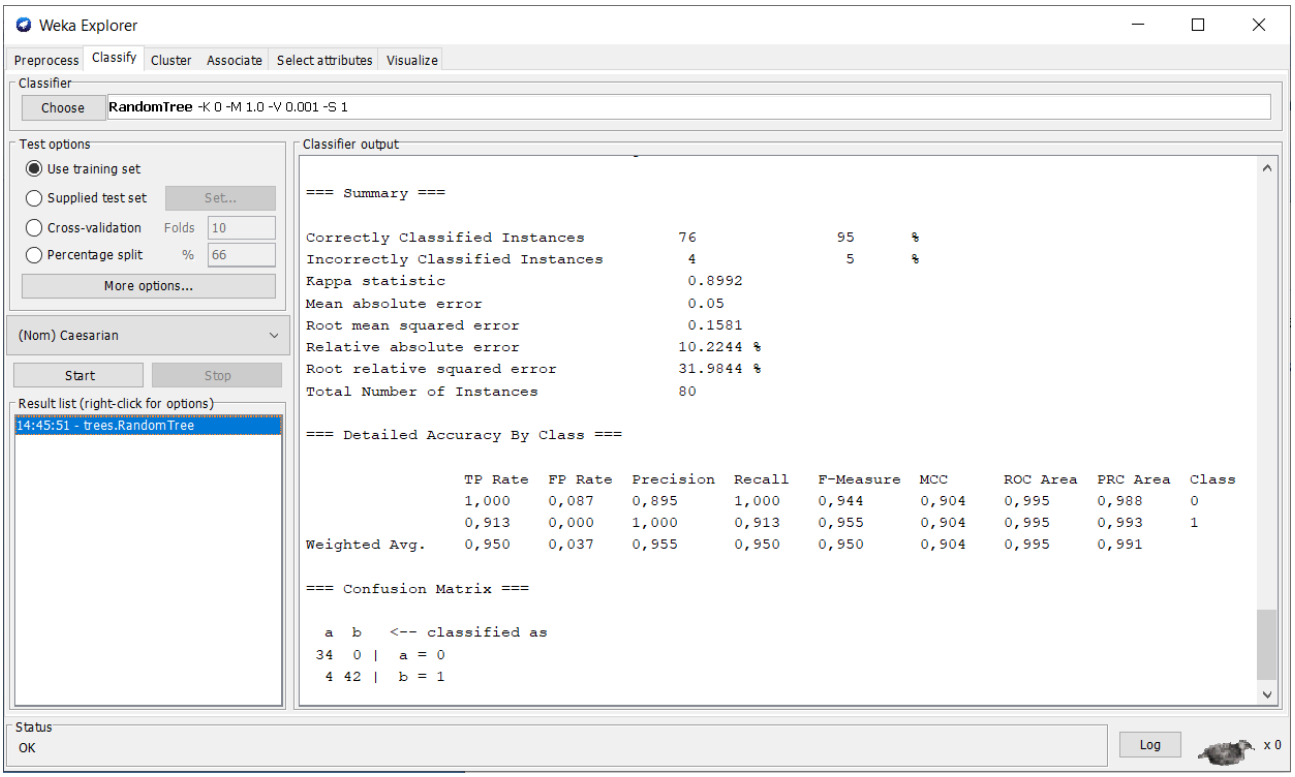


Figure 15: Résultat de l'algorithme Random Tree

Le logiciel WEKA permet de visualiser aussi l'arbre de décision (figure 16)

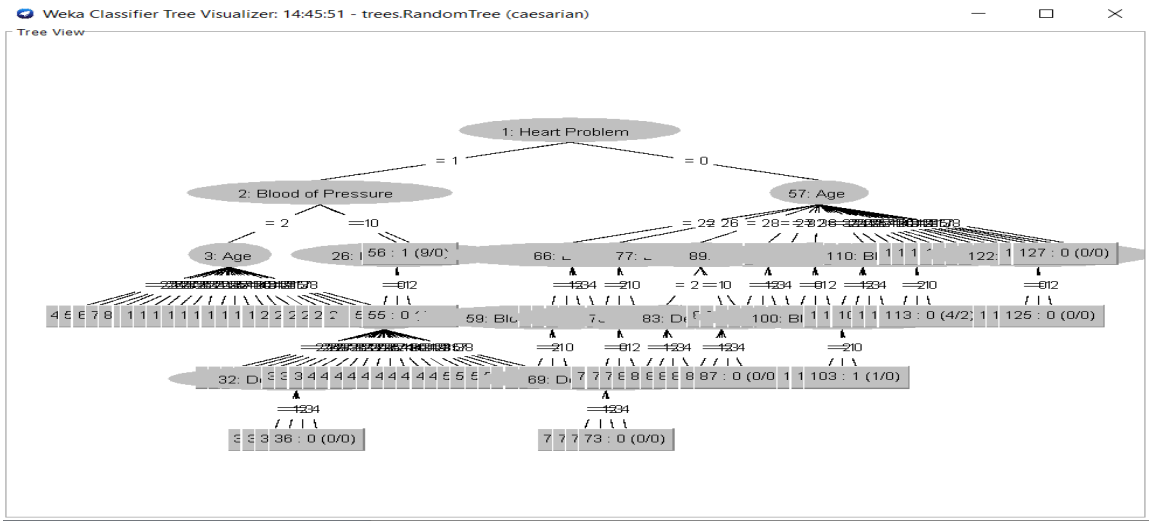
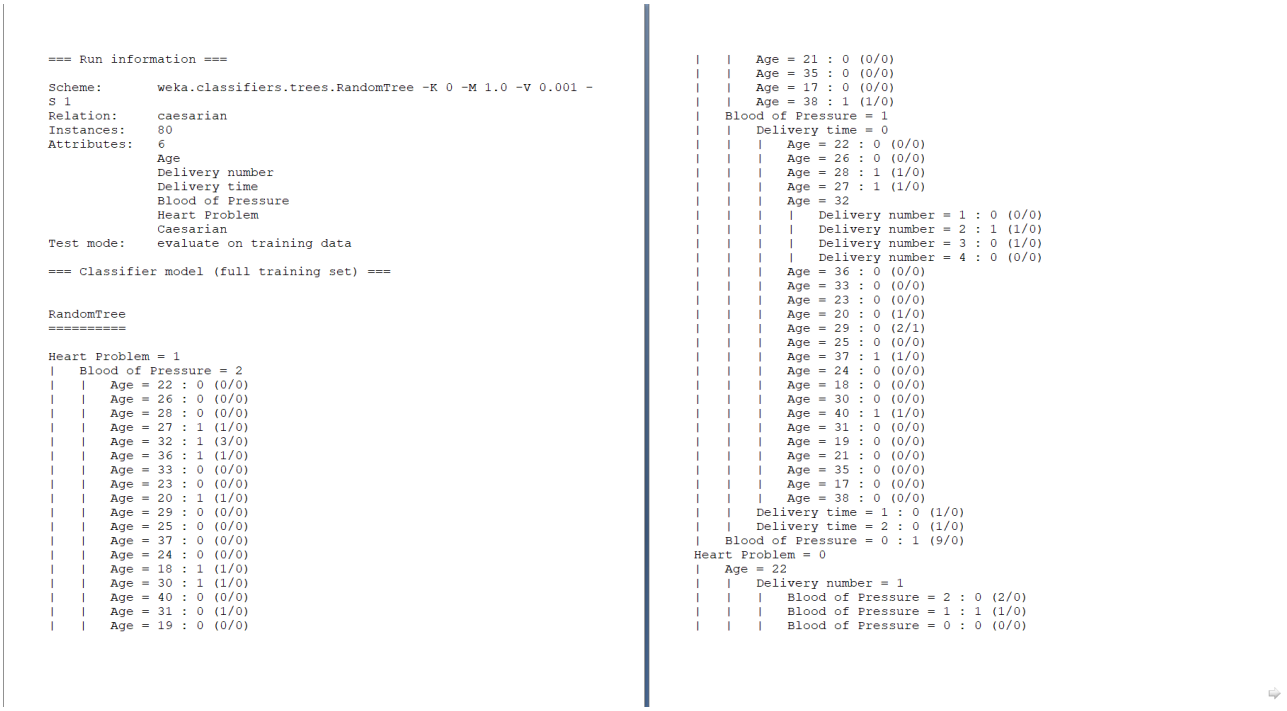


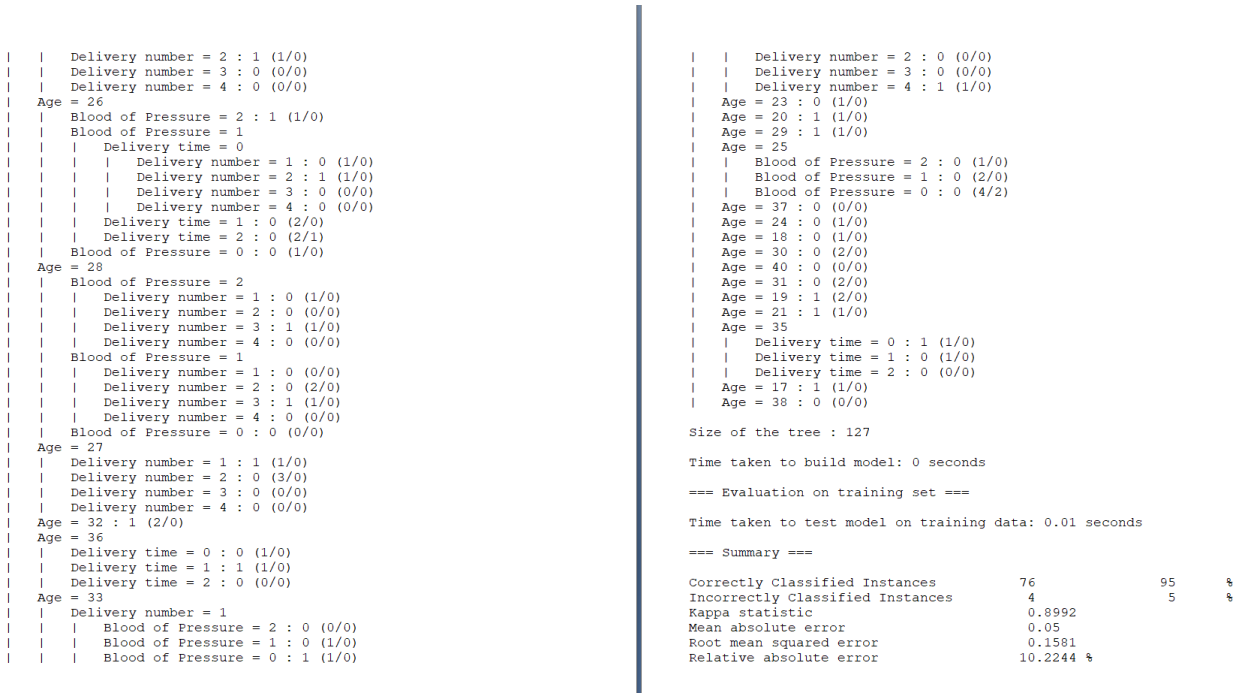
Figure 16 : Visualisation de l'algorithme Random Tree

Le logiciel WEKA permet aussi de donner le model de classification (figure 17.1, 17.2)



```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -
S 1
Relation:     caesarian
Instances:    80
Attributes:   6
              Age
              Delivery number
              Delivery time
              Blood of Pressure
              Heart Problem
              Caesarian
Test mode:    evaluate on training data

=== Classifier model (full training set) ===


RandomTree
==========

Heart Problem = 1
|   Blood of Pressure = 2
|   |   Age = 22 : 0 (0/0)
|   |   Age = 26 : 0 (0/0)
|   |   Age = 28 : 0 (0/0)
|   |   Age = 27 : 1 (1/0)
|   |   Age = 32 : 1 (3/0)
|   |   Age = 36 : 1 (1/0)
|   |   Age = 33 : 0 (0/0)
|   |   Age = 23 : 0 (0/0)
|   |   Age = 20 : 1 (1/0)
|   |   Age = 29 : 0 (0/0)
|   |   Age = 25 : 0 (0/0)
|   |   Age = 37 : 0 (0/0)
|   |   Age = 24 : 0 (0/0)
|   |   Age = 18 : 1 (1/0)
|   |   Age = 30 : 1 (1/0)
|   |   Age = 40 : 0 (0/0)
|   |   Age = 31 : 0 (1/0)
|   |   Age = 19 : 0 (0/0)
|   |   Age = 21 : 0 (0/0)
|   |   Age = 35 : 0 (0/0)
|   |   Age = 17 : 0 (0/0)
|   |   Age = 38 : 1 (1/0)
|   Blood of Pressure = 1
|   |   Delivery time = 0
|   |   |   Age = 22 : 0 (0/0)
|   |   |   Age = 26 : 0 (0/0)
|   |   |   Age = 28 : 1 (1/0)
|   |   |   Age = 27 : 1 (1/0)
|   |   |   Age = 32
|   |   |   |   Delivery number = 1 : 0 (0/0)
|   |   |   |   Delivery number = 2 : 1 (1/0)
|   |   |   |   Delivery number = 3 : 0 (1/0)
|   |   |   |   Delivery number = 4 : 0 (0/0)
|   |   |   Age = 36 : 0 (0/0)
|   |   |   Age = 33 : 0 (0/0)
|   |   |   Age = 23 : 0 (0/0)
|   |   |   Age = 20 : 0 (1/0)
|   |   |   Age = 29 : 0 (2/1)
|   |   |   Age = 25 : 0 (0/0)
|   |   |   Age = 37 : 1 (1/0)
|   |   |   Age = 24 : 0 (0/0)
|   |   |   Age = 18 : 0 (0/0)
|   |   |   Age = 30 : 0 (0/0)
|   |   |   Age = 40 : 1 (1/0)
|   |   |   Age = 31 : 0 (0/0)
|   |   |   Age = 19 : 0 (0/0)
|   |   |   Age = 21 : 0 (0/0)
|   |   |   Age = 35 : 0 (0/0)
|   |   |   Age = 17 : 0 (0/0)
|   |   |   Age = 38 : 0 (0/0)
|   |   Delivery time = 1 : 0 (1/0)
|   |   Delivery time = 2 : 0 (1/0)
|   Blood of Pressure = 0 : 1 (9/0)
Heart Problem = 0
|   Age = 22
|   |   Delivery number = 1
|   |   |   Blood of Pressure = 2 : 0 (2/0)
|   |   |   Blood of Pressure = 1 : 1 (1/0)
|   |   |   Blood of Pressure = 0 : 0 (0/0)
```

Figure 17.1 : Model de classification



```
|   |   Delivery number = 2 : 1 (1/0)
|   |   Delivery number = 3 : 0 (0/0)
|   |   Delivery number = 4 : 0 (0/0)
|   Age = 26
|   |   Blood of Pressure = 2 : 1 (1/0)
|   |   Blood of Pressure = 1
|   |   |   Delivery time = 0
|   |   |   |   Delivery number = 1 : 0 (1/0)
|   |   |   |   Delivery number = 2 : 1 (1/0)
|   |   |   |   Delivery number = 3 : 0 (0/0)
|   |   |   |   Delivery number = 4 : 0 (0/0)
|   |   |   Delivery time = 1 : 0 (2/0)
|   |   |   Delivery time = 2 : 0 (2/1)
|   |   Blood of Pressure = 0 : 0 (1/0)
|   Age = 28
|   |   Blood of Pressure = 2
|   |   |   Delivery number = 1 : 0 (1/0)
|   |   |   Delivery number = 2 : 0 (0/0)
|   |   |   Delivery number = 3 : 1 (1/0)
|   |   |   Delivery number = 4 : 0 (0/0)
|   |   Blood of Pressure = 1
|   |   |   Delivery number = 1 : 0 (0/0)
|   |   |   Delivery number = 2 : 0 (2/0)
|   |   |   Delivery number = 3 : 1 (1/0)
|   |   |   Delivery number = 4 : 0 (0/0)
|   |   Blood of Pressure = 0 : 0 (0/0)
|   Age = 27
|   |   Delivery number = 1 : 1 (1/0)
|   |   Delivery number = 2 : 0 (3/0)
|   |   Delivery number = 3 : 0 (0/0)
|   |   Delivery number = 4 : 0 (0/0)
|   Age = 32 : 1 (2/0)
|   Age = 36
|   |   Delivery time = 0 : 0 (1/0)
|   |   Delivery time = 1 : 1 (1/0)
|   |   Delivery time = 2 : 0 (0/0)
|   Age = 33
|   |   Delivery number = 1
|   |   |   Blood of Pressure = 2 : 0 (0/0)
|   |   |   Blood of Pressure = 1 : 0 (1/0)
|   |   |   Blood of Pressure = 0 : 1 (1/0)
|   |   Delivery number = 2 : 0 (0/0)
|   |   Delivery number = 3 : 0 (0/0)
|   |   Delivery number = 4 : 1 (1/0)
|   Age = 23 : 0 (1/0)
|   Age = 20 : 1 (1/0)
|   Age = 29 : 1 (1/0)
|   Age = 25
|   |   Blood of Pressure = 2 : 0 (1/0)
|   |   Blood of Pressure = 1 : 0 (2/0)
|   |   Blood of Pressure = 0 : 0 (4/2)
|   Age = 37 : 0 (0/0)
|   Age = 24 : 0 (1/0)
|   Age = 18 : 0 (1/0)
|   Age = 30 : 0 (2/0)
|   Age = 40 : 0 (0/0)
|   Age = 31 : 0 (2/0)
|   Age = 19 : 1 (2/0)
|   Age = 21 : 1 (1/0)
|   Age = 35
|   |   Delivery time = 0 : 1 (1/0)
|   |   Delivery time = 1 : 0 (1/0)
|   |   Delivery time = 2 : 0 (0/0)
|   Age = 17 : 1 (1/0)
|   Age = 38 : 0 (0/0)

Size of the tree : 127

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          76               95      %
Incorrectly Classified Instances         4                5      %
Kappa statistic                          0.8992
Mean absolute error                      0.05
Root mean squared error                  0.1581
Relative absolute error                 10.2244 %
```

Figure 17.2 : Model de classification

- **L'application développée**

Les données utilisées dans cette étape au niveau de l'environnement Eclipse IDE est le model produit par WEKA et les 1/3 des données (la base de teste), ces deux dernières sont des tables crées au niveau d'Access pour qu'on puisse les importer sur Eclipse IDE.

Le logiciel que nous avons réalisé permet de prédire le type d'accouchement d'une femme.

Ce schéma représente les principales Classes Java.

6. Les interfaces du système

Notre application contienne une interface qui permet a l'utilisateur d'entrer ces données dans un champ préciser, a fin d'entrer ces données le système va afficher la prédiction du la femme a l'aide du model distribué par WEKA.

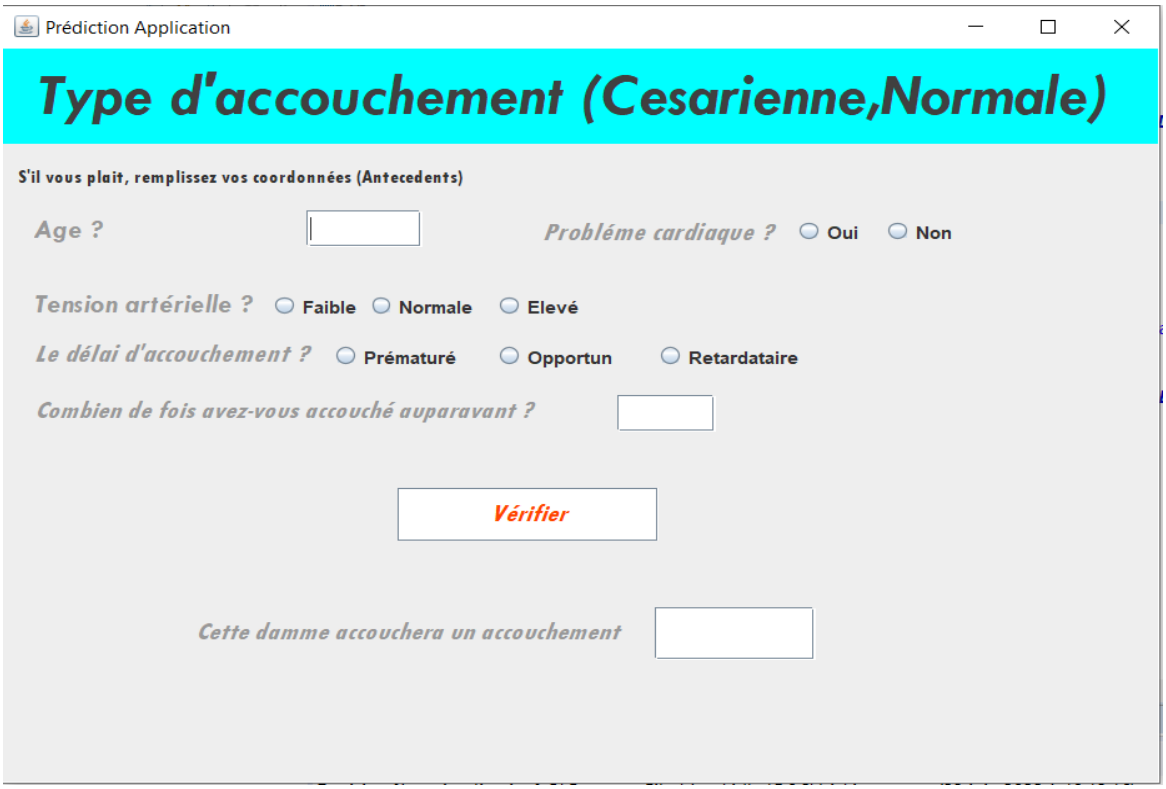


Figure 18 : Interface d'application

Dans la figure 19 on a fais un exemple d'une nouvelle données dans le cas d'accouchement normale

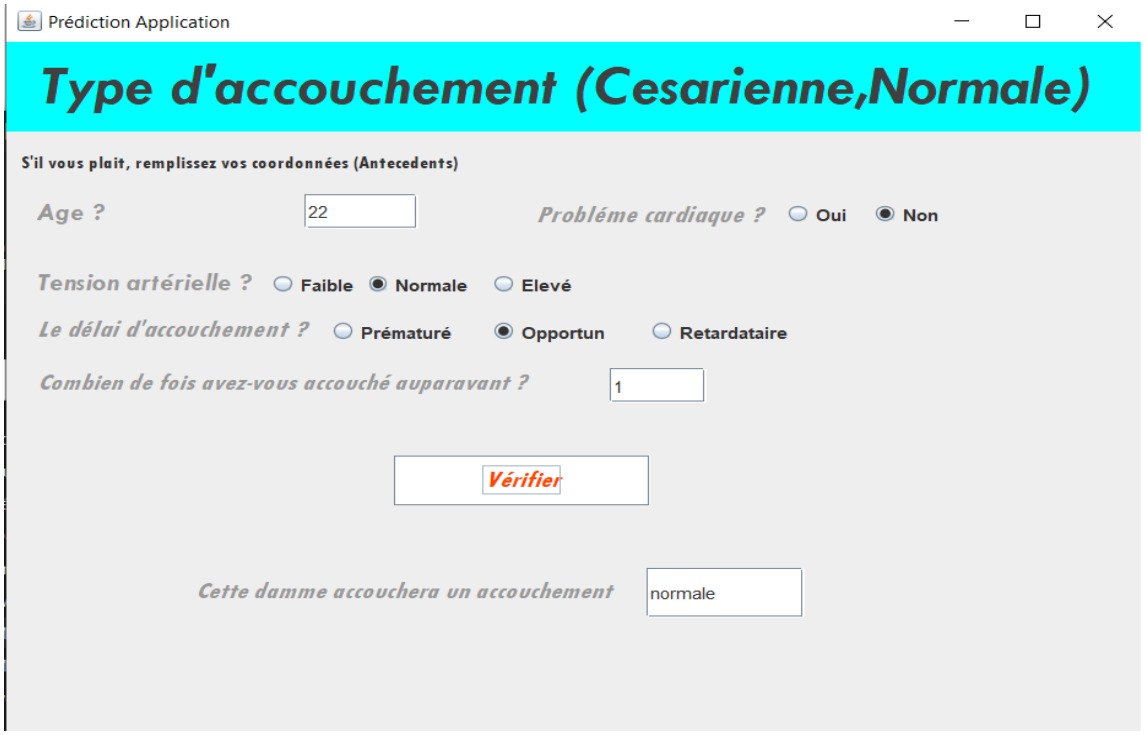


Figure 19: Exemple du cas d'accouchement normale

Dans la figure 20 on a fait un exemple d'un cas d'accouchement césarien.

Figure 20: Exemple du cas d'accouchement césarien

7. Conclusion

Dans ce chapitre, Nous avons définit un système qui permet à prédire le type d'accouchement d'une femme a travers des facteurs.

Notre résultat d'expérimentation basé sur les techniques de Fouille de données montre que notre système est performant de 95% de prédiction de types d'accouchement.

Conclusion générale

Conclusion générale

Le travail que nous avons effectué dans ce mémoire nous a permis de constater que les processus d'extraction de connaissances ont leur propre série de caractéristiques qui doivent être prises en compte. La première caractéristique distinctive que nous avons découverte est que l'ECD est un processus et une série d'étapes qui doivent être suivies, en commençant par la sélection des données, leur transformation, la création du modèle et l'évaluation du modèle afin d'extraire de nouvelles connaissances.

Nous avons également découvert que l'étape de "Fouille de données" de la création de modèles est au cœur d'un processus d'ECD ; il s'agit d'un ensemble d'outils et de procédures utilisés pour parvenir à de nouvelles connaissances utilisables.

Il existe plusieurs méthodes et techniques dédiées au Classification, dans notre travail on a choisit l'arbre de décision (ou DecisionTree) 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 qui peut prédire le type d'accouchement des femmes utilisant les techniques nécessaires.

Bibliographie

Bibliographie

[1] L, BRADJI. '' Adaptation des techniques de l'Extraction des Connaissances à partir des Données (ECD) pour prendre en charge la qualité des données '', Thèse de doctorat, Université Mentouri Constantine Faculté de sciences de l'ingénieur Département informatique 2012.

[2] M.JAMBU. "Exploration informatique et statistique des données". Edition Dunod (1989).

[3] R.CHAFI. Le machine Learning, https://medium.com/@redouanechafi/data-science-0-0-quest-ce-que-le-machine-learning-fde2b3c5f19f

[4] O. FAYYAD, U.M.,PIATETSKY-SHAPIRO, G., et SMYTH, P.. De la fouille de données à la découverte de connaissances: : Un aperçu, dans AI(DDM, AAAI/MIT P 1996.

[5] : R BRACHMAN, A. ANAND, Le processus de l'extraction de connaissance, en KDDM, AAAI/MIT P 1996.

[6] M. BENAYED. Le processus ECD, Journal d'Interaction Personne-Système. 6 octobre 2014.

[7] K. KAUR. La classification en Data Mining. En IEEE, En Dehradun, India.04-05 Septembre 2015

[8] S. OULAD NAOUI. Les taches en fouille de données.https://wikimemoires.net/2013/05/les-taches-en-fouille-de-donnees/

[9] TechTarget. Data visualization. https://www.lemagit.fr/definition/Data-Visualization-ou-DataViz#:~:text=La%20visualisation%20de%20donn%C3%A9es%20(Data,ci%20dans%20un%20contexte%20visuel.

[10] P. SCHWAB. Data visualisation. https://www.intotheminds.com/blog/data-visualisation/ .

[11] : Domaine d'application du Data Mining.http://igm.univ-mlv.fr/~dr/XPOSE2012/datamining/datamining-domaines-application.html

[12] J. MARI et A. NAPOLI. Aspects de la classification. Rapport technique 2909, INRIA.(1996).

[13] D. MASSE. ''*Arbres de décision*''. Cours en ligne, Master 2 ISC. Laboratoire d'Informatique des Systèmes Complexes, Université de Brest. (Consulté le 13.05.2010).http://www.lisyc.univbrest.fr/pages_perso/dmasse/cours/cours_decision.pdf

[14] F. MOUTARDE. ''*Brève introduction aux arbres de décision''*. Cours en ligne. Centre de Robotique (CAOR), Ecole des Mines de Paris. 2008. (Consulté le 13.01.2010). www.ensmp.fr/~moutarde/slides_AD.pdf

[15] C. DECAESTECKER, Les arbres de décision (decisiontrees), ULB, Marco Saerens, UCL, LINF2275.

Bibliographie

[16] J. HAN, M. KAMBER. *''Data Mining, concepts and techniques''*. Ouvrage. Edition Morgan Kaufmann Publishers. 2000.

[17] C. CHAINE. Accouchement.https://www.docteurclic.com/encyclopedie/accouchement.aspx

[18] M. BEN ALI. L'accouchement.
https://www.pregnancyinfo.ca/fr/birth/delivery/normalchildbirth/#:~:text=Dans%20sa%20d%C3%A9finition%20la%20plus,la%20premi%C3%A8re 20heure%20de%20vie.

[19] R. KALMEGH. Simple Cart et RandomTree pour la Classification.
http://ijiset.com/vol2/v2s2/IJISET_V2_I2_63.pdf

[20] Y. XU. RandomTree, https://www.mdpi.com/2076-3417/9/9/1728/htm .

[21] Cesarian Data Set.
https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset#

[22] Henriet L. Système d'évaluation et de classification multicritères pour l'aide à la décision, construction de modles et procédures d'affectation. (2000). Thèse de doctorat en science. Université Paris Dauphine.

[23] Bognar K. Aspects théoriques de la classification à base de treillis. Université Debrecen : Institut de mathématiques et informatique. . (2003).
[24] Preux p. Fouille de données notes de cours, université de lille. (2001).