UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE
APPLIQUÉES

PAR
PENGWENDE ABDOULAYE OUEDRAOGO

MÉTHODE D'ÉLAGAGE DES RÈGLES D'ASSOCIATION ET ESTIMATION DE LA PERTE D'INFORMATION DANS LES DONNÉES MÉDICALES

Juin 2021

Résumé

De nos jours, il existe de nombreux algorithmes proposés pour l'extraction des règles d'association, le plus connu et le plus simple est l'algorithme Apriori proposé par Agrawal en 1993. L'utilisation de l'algorithme Apriori dans le data mining permet de tester les différentes combinaisons possibles des éléments afin de déterminer les relations potentielles exprimées sous la forme de règles d'association.

D'autres algorithmes tels que AprioriTid ou FP-Growth se veulent être une amélioration de Apriori. Cependant, pour les utilisateurs, quel que soit l'algorithme utilisé pour l'extraction des règles d'association, le problème majeur lié au traitement du nombre pléthorique de ces règles revient toujours. En effet, le nombre important de règles qu'il est possible de générer rend leur utilisation humainement difficile, voire impossible. Ce qui entraine une difficulté au niveau de la visualisation de cesdites règles. La représentation visuelle de cette masse importante d'informations tient difficilement sur un écran et rend ainsi la lecture fastidieuse.

L'objet de recherche dans notre projet est d'explorer des solutions possibles pour réduire ce nombre de règles tout en évitant (minimisant) la perte d'information. Pour ce faire, nous avons utilisé des données médicales réelles provenant d'un hôpital. Nous proposons donc une méthode d'élagage qui

permettra d'éliminer les règles jugées inintéressantes tout en gardant les règles qui s'avéreront pertinentes pour une meilleure analyse.

Nos travaux démontrent que la méthode d'élagage proposée améliore la visualisation de plusieurs règles et facilite le traitement et l'analyse de gros corpus textuel.

Mots clés :

Data mining, règles d'association, algorithme Apriori, R, Rstudio, extraction de données, élagage, perte d'information

Remerciements

Enfin, ce moment est venu à la suite de deux années difficiles. Ce mémoire est dévoué avec un amour le plus profond et un respect éternel.

À mon père, ma mère, mon frère et à l'âme de ma sœur. À ma fiancée, mes amis bien-aimés et à tous ceux qui de près ou de loin m'ont apporté leurs aides.

À mes chers amis et partenaires de NACMO qui travaillent et se battent au succès de notre rêve commun.

À mon superviseur M. Ismail Biskri pour sa supervision, son incroyable patience et son soutien indéfectible.

Sans leurs soutiens et encouragements je n'aurais pas pu atteindre cette phase, alors MERCI.

Abdoulaye.

Table des matières

Liste des tableaux

Liste des figures

CHAPITRE 1 – Introduction

L'exploration de données, également appelée découverte de connaissances dans des sources de données, est un domaine de recherche important en informatique. Elle est largement utilisée dans les domaines des affaires (assurances, banques, système de détection de fraude par carte de crédit), dans la recherche scientifique (médecine, astronomie, analyse de données biologiques) et dans la sécurité du gouvernement (détection des criminels et des terroristes).

Cette discipline a connu une croissance importante ces dernières années grâce aux avancées récentes en intelligence artificielle et est maintenant appliquée dans plusieurs domaines. De nombreuses entreprises et chercheurs en informatique s'intéressent à l'analyse automatique du contenu. Cette discipline se retrouve au cœur des débats avec l'avènement du big data.

En 2001, un rapport de recherche du Groupe Gartner définit les enjeux inhérents à la croissance des données comme étant tridimensionnels selon la règle dite <<des 3V>> (volume, vélocité et variété). Ce modèle est encore largement utilisé aujourd'hui pour décrire ce phénomène [10].

En effet, le big data possède un important potentiel scientifique. Les chercheurs et autres professionnels cherchent aujourd'hui l'outil idéal leur permettant d'analyser automatiquement ces grandes masses de données hautement bruitées afin d'automatiser certaines tâches ou extraire l'information

enfouie dans ces grandes masses d'informations et ainsi développer des applications informatiques. Aussi, les moteurs de recherche, les systèmes de traduction automatique et les assistants personnels intelligents dans les téléphones cellulaires découlent tous des recherches effectuées dans ce domaine.

Une des plus importantes taches de l'exploration de données est de trouver les règles d'association et de découvrir les modèles et les relations les plus intéressantes et les plus utiles dans un large volume de données. Actuellement, il y a plusieurs algorithmes proposés pour l'extraction des règles d'association, le plus connu et le plus simple est l'algorithme Apriori proposé par Agrawal en 1993 [2]. Les règles d'association permettent de rechercher les relations entre les objets fréquemment utilisés ensemble. Les éléments fréquents désignent l'ensemble des éléments fréquemment rencontrés dans une base de données transactionnelles. Ces éléments fréquents sont utilisés pour générer l'ensemble des règles d'association. En d'autres termes, les règles d'association décrivent simplement le comportement d'une entité soumis à certaines conditions. Les éléments sont considérés comme fréquents s'ils se produisent dans la source de données pendant un certain temps supérieur ou égal à un seuil prédéfini. Par exemple: 68% des clients qui achètent des boissons gazeuses sont susceptibles d'acheter également un chocolat. Dans la base de données transactionnelles, supposons que les transactions qui contiennent les deux articles (boissons gazeuses et chocolat) représentent 19% de l'ensemble taille des transactions dans la base de données du

magasin. Les clients qui achètent des boissons gazeuses représentent l'antécédent de la règle d'association et ceux qui achètent des chocolats sont appelés une règle d'association qui en résulte. Les 68% de la règle d'association mentionnée ci-dessus dénotent la force de la règle et sont connus sous le nom de confiance de la règle, tandis que le 19% est une mesure statistique, connue sous le nom de support de la règle.

Dans le contexte du big data, les règles d'association apportent un atout d'analyse. Cependant, le nombre important de règles générées demeure un problème à résoudre. Notre défi est de trouver un moyen efficace d'élaguer les règles générées en ne gardant que les règles les plus intéressantes. Un nombre réduit de règles favorisera l'analyse et la prise de décision. Nous proposons donc une approche novatrice qui permettra aux scientifiques de minimiser la perte d'information tout en garantissant la pertinence des règles obtenues après élagage.

Le présent mémoire est composé de six chapitres. Dans le premier chapitre, nous mettons en avant la problématique et les objectifs de ce projet de recherche. Le second chapitre décrit les différentes étapes et techniques de l'exploration de données avec une mise en emphase sur la découverte des associations afin d'introduire les règles d'associations dans le chapitre 3. Ce chapitre est consacré à la définition et la description des concepts, des algorithmes et des techniques utilisées dans le domaine des règles d'association. Aussi nous passons en revue

les différentes techniques d'élagage proposées dans d'autres travaux. Le chapitre 4, aborde les aspects théoriques et conceptuels du système développé. Nous y présentons les outils, la méthodologie et le flux de travail de notre application. Les différentes fonctionnalités y sont développées et consolidées par des captures d'écran. Le chapitre 5 présente les expérimentations et les discussions sur les résultats obtenus. L'application traite des données médicales et fournit les résultats sous forme de graphe ou de tableau dans une interface intuitive et ergonomique. Enfin le chapitre 6 présente la conclusion générale, qui s'ouvre par la synthèse de notre travail aux débouchés et perspectives qui pourraient faire l'objet d'un travail de recherche ultérieur.

CHAPITRE 2 - Exploration de données

2.1 Introduction

L'expansion des nouvelles technologies a entrainé la collecte d'énormes volumes de données. Les organisations ont désormais accès à plus de données qu'elles n'en ont jamais eu auparavant. Ces données peuvent être structurées ou non structurées et leurs énormes volumes rendent leur compréhension difficile pour la mise en œuvre des améliorations à l'échelle de l'organisation. S'il n'est pas correctement traité, ce défi peut minimiser les avantages des données.

L'exploration de données est le processus par lequel les organisations détectent des modèles dans les données pour obtenir des informations pertinentes à leurs besoins commerciaux. C'est essentiel à la fois pour l'intelligence des affaires et la science des données.

Dans plusieurs travaux [28], [32], il est cité de nombreuses techniques d'exploration de données que les scientifiques peuvent utiliser pour transformer les données brutes en informations exploitables. De l'intelligence artificielle de pointe aux bases de la préparation des données, chaque étape est essentielle pour maximiser la valeur des investissements dans les données.

Dans ce chapitre nous présenterons ces différentes étapes qui interviennent très souvent dans le processus d'exploration de données.

2.2 Nettoyage et préparation des données

Le nettoyage et la préparation des données sont une partie vitale du processus d'exploration de données. Les données dans leurs formes brutes peuvent comporter des défauts ou des anomalies, les données doivent être nettoyées et formatées pour être utiles pour différentes méthodes analytiques. Certains éléments de modélisation tels que la transformation, la migration de données, l'intégration de données et d'agrégation sont des étapes nécessaires pour comprendre les caractéristiques et attributs de base des données afin de déterminer leur meilleure utilisation.

L'importance du nettoyage et de la préparation des données est évidente. Négliger cette première étape pourrait entrainer l'utilisation de données peu fiable en raison de leur qualité. Le principal objectif est d'obtenir des données reflétant au mieux la réalité, en extraire les résultats de son analyse et entreprendre des actions à partir de ces résultats.

Ces étapes sont également nécessaires pour la qualité des données et une bonne gouvernance des données.

Processus de préparation de données :

La figure ci-après (voir figure 2.1) illustre le processus de préparation d'un ensemble de données.

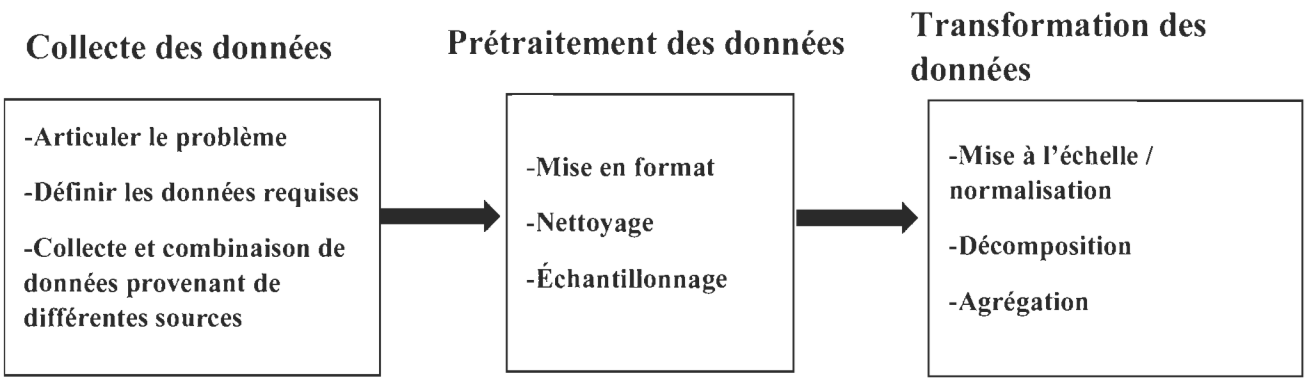


Figure 2.1. Exemple de processus de préparation de données

2.3 Modèles de suivi

Le suivi des modèles est une technique fondamentale d'exploration de données. Il s'agit d'analyser les données afin d'identifier les tendances ou les modèles. Ensuite on peut procéder à des inférences sur par exemple les résultats d'une organisation. Supposons détenir les données de vente d'une entreprise, en y trouvant les tendances de vente on pourrait mettre en place une base pour prendre des mesures afin de capitaliser sur cette information. Par exemple si un produit X se vend plus que d'autres pour un groupe démographique particulier {Y, X, Z}, on pourrait utiliser ces connaissances pour créer des produits ou services similaires X' ou X'' ou encore simplement mieux stocker le produit d'origine pour ce groupe démographique {X, Y, Z}.

2.4 Classification

Les techniques d'exploration de données de classification impliquent l'analyse des divers attributs associés à différents types de données. Une fois qu'on identifie les principales caractéristiques de ces types de données, on peut catégoriser ou classer les données associées. On pourra par exemple affecter un e-mail donné à la classe «spam» ou «non-spam», et l'attribution d'un diagnostic à un patient donné en fonction des caractéristiques observées du patient (sexe, tension artérielle, présence ou absence de certains symptômes, etc.) [39]. Cela peut être aussi essentiel pour identifier les informations personnellement identifiables que les organisations peuvent vouloir protéger ou supprimer des documents.

Exemple de modèle de classification:

La figure ci-après (voir figure 2.2) présente un modèle de classification.

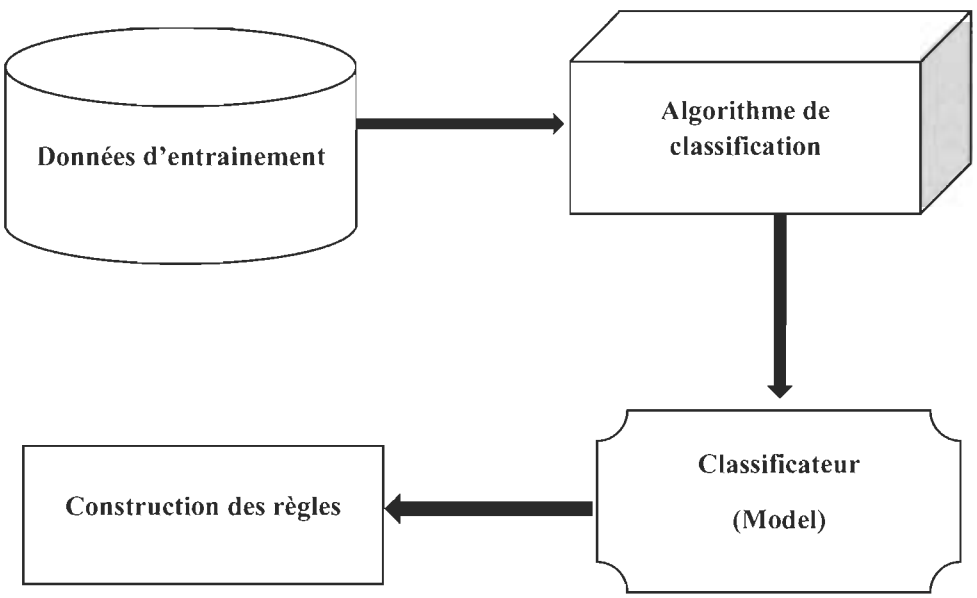


Figure 2.2. Exemple de modèle de classification

2.5 Association

L'association est une technique d'exploration de données liée aux statistiques. Il indique que certaines données (ou évènements trouvés dans les données) sont liés à d'autres données ou évènements basés sur les données. Elle est similaire à la notion de cooccurrence dans l'apprentissage automatique, dans laquelle la probabilité d'un évènement basé sur les données est indiquée par la présence d'un autre [40].

Le concept statistique de corrélation est également similaire à la notion d'association. Cela signifie que l'analyse des données montre qu'il existe une relation entre deux évènements de données: comme le fait que l'achat de hamburgers s'accompagne fréquemment de celui de frites.

Exemple d'associations :

La figure ci-après (voir figure 2.3) illustre à travers un graphe les associations des produits d'une épicerie.

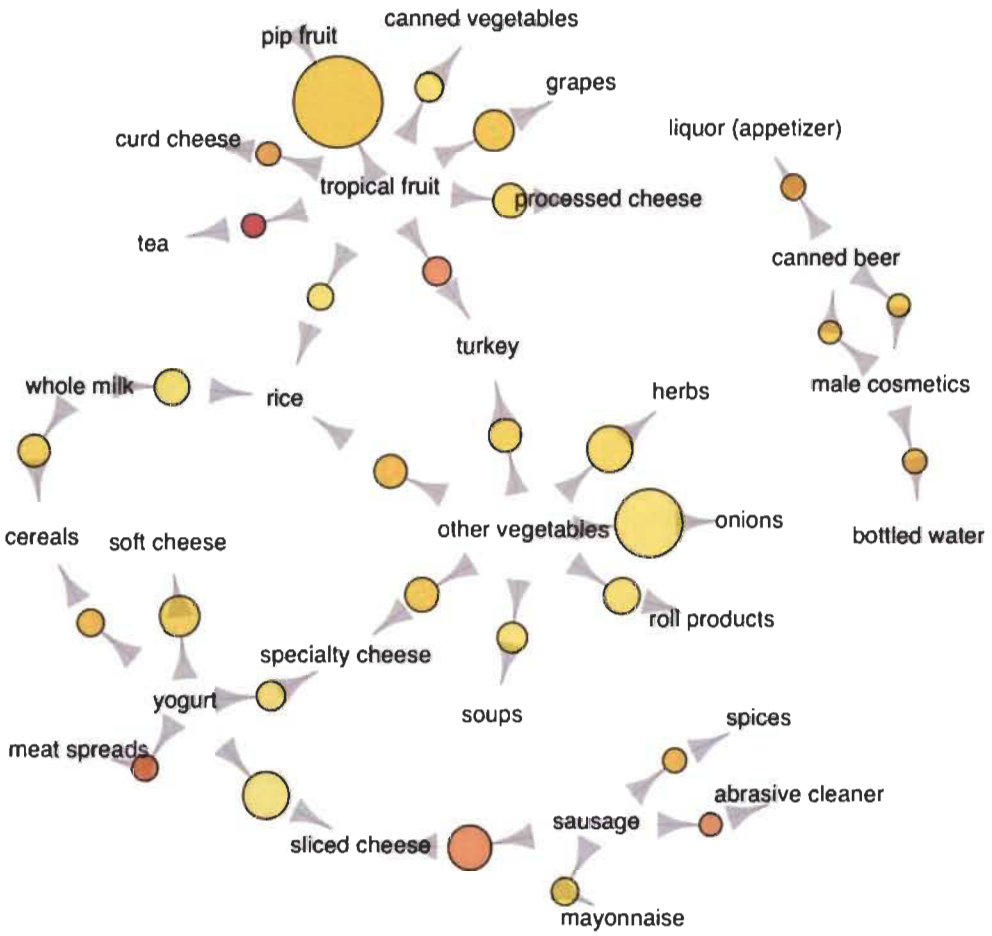


Figure 2.3. Exemple d'association [22]

2.6 Détection des valeurs aberrantes

La détection des valeurs aberrantes détermine toutes les anomalies dans les ensembles de données. En général, les éléments anormaux sont liés à des problèmes tels que la fraude bancaire, un défaut structurel, des problèmes médicaux ou des erreurs dans un texte. Les anomalies sont également appelées nouveautés, bruit, écarts et exceptions [41]. Dépendant du format des données on peut appliquer trois différents types de détection d'anomalies. Pour des données

de test non étiqueté et en considérant que l'ensemble des instances de l'ensemble des données sont normales on peut appliquer la technique de détection d'anomalies non supervisées [42]. Les techniques de détection d'anomalies supervisées nécessitent un ensemble de données qui a été étiqueté comme <<normal>> ou <<anormal>> et implique la formation d'un classificateur. Les techniques de détection d'anomalies semi-supervisées mettent en place un modèle possédant une caractéristique *normal* puis on soumet une instance de test à ce modèle [42].

Dans le domaine commercial, la détection des anomalies est très importante dans le sens qu'elle permet de comprendre pourquoi des anomalies se produisent et quelles décisions il faut prendre en conséquence pour atteindre au mieux les objectifs commerciaux. Dans le domaine bancaire par exemple, s'il y a un pic dans l'utilisation des systèmes transactionnels pour les cartes de crédit à un certain moment de la journée, et en y cherchant la cause on peut optimiser les ventes pendant le reste de la journée.

2.7 Regroupement

Le regroupement est une technique d'analyse qui repose sur des approches visuelles pour comprendre les données. Il consiste à regrouper des objets similaires dans le même groupe appelé ***cluster.*** Les algorithmes de regroupement sont en général très paramétrables, l'utilisateur peut par exemple y définir la

fonction de distance à utiliser, un seuil de densité ou le nombre de clusters attendus.

Les graphiques sont des outils majeurs dans les mécanismes de regroupement, afin de montrer où se situe la distribution des données par rapport aux différents types de métriques. Les couleurs sont fréquemment utilisées pour définir les différentes classes, ainsi l'utilisateur peut voir visuellement comment les données sont distribuées afin de prendre des décisions.

Exemple de regroupement de données :

L'illustration ci-après (voir figure 2.4) montre un exemple de regroupement de données avec des couleurs de distinction.

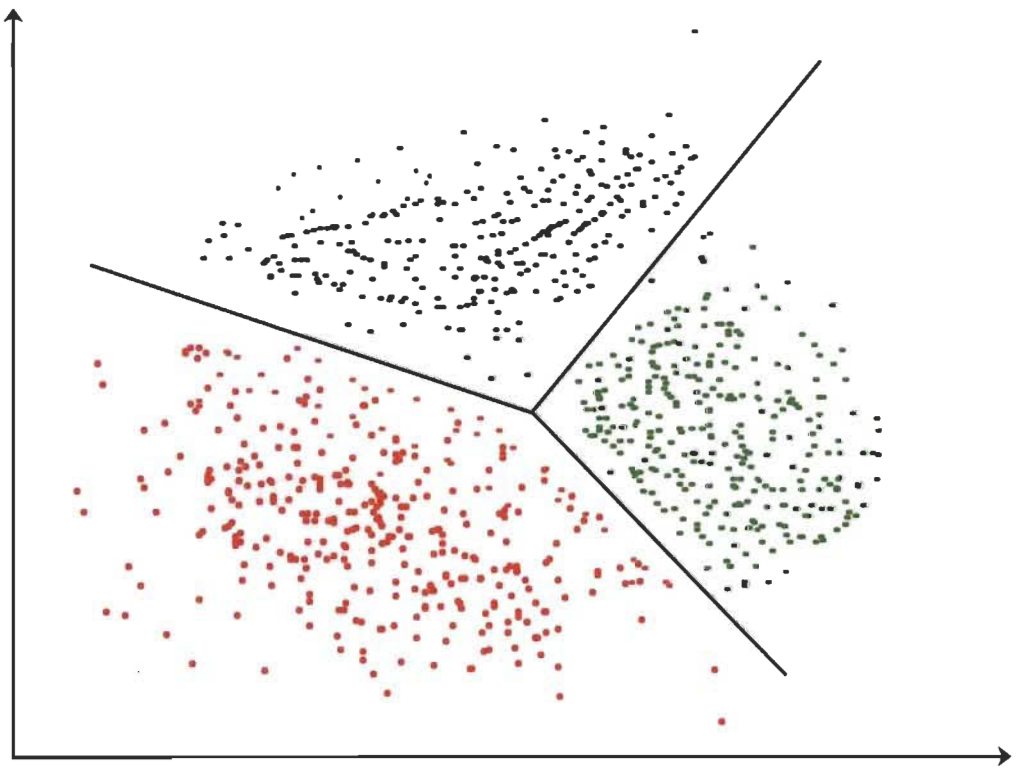


Figure 2.4. Exemple de regroupement de données [29]

2.8 Régression

Les techniques de régression sont utiles pour identifier la relation entre les variables d'un ensemble de données. La causalité ou la simple corrélation de ces relations sont souvent à déterminer pour une meilleure analyse de l'ensemble de données. C'est une technique simple, mais très utilisée dans les aspects de la prévision et de la modélisation des données.

Exemple de régression linéaire :

Voici ci-après un exemple (voir figure 2.5) de régression linéaire.

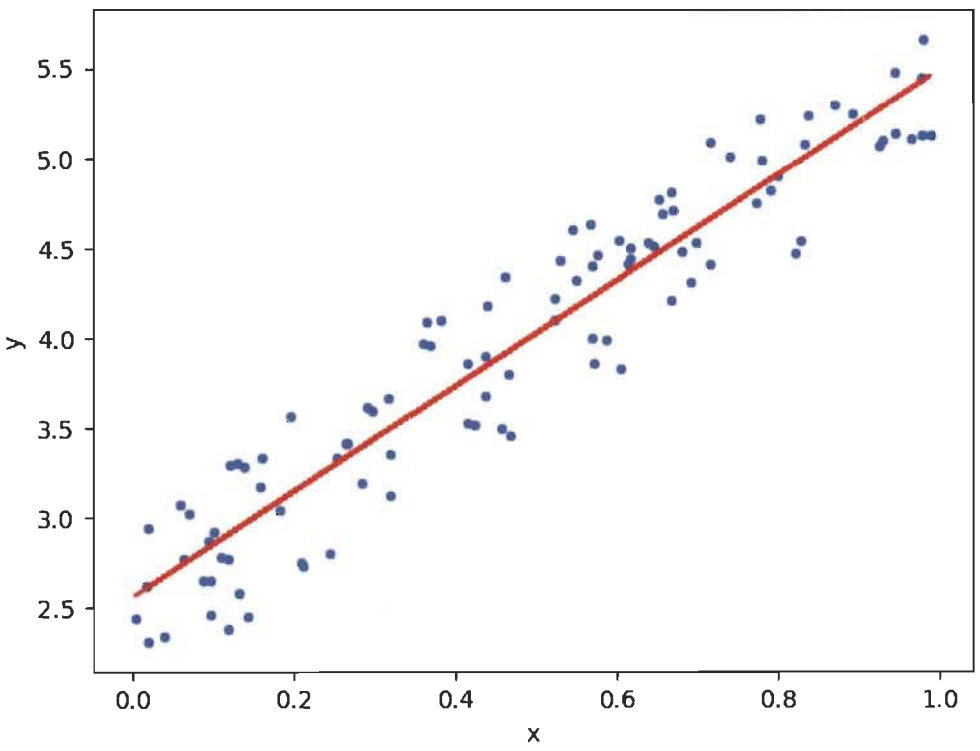


Figure 2.5. Exemple de régression linéaire [32]

2.9 Prédiction

La prédiction est un aspect très puissant de l'exploration des données qui représente l'une des quatre branches de l'analyse. L'analyse prédictive utilise des modèles construits à travers les données actuelles ou historiques afin de prédire le futur. Ainsi l'utilisateur peut avoir un aperçu des tendances qui se produiront dans les prochaines données.

De l'apprentissage automatique à l'intelligence artificielle, il existe plusieurs approches pour utiliser l'analyse prédictive. Elle peut également être facilitée par des approches statistiques plus simples.

2.10 Modèles séquentiels

Cette technique d'exploration de données se base sur la découverte d'une suite d'évènements qui se déroulent en séquence. C'est particulièrement utile pour les données transactionnelles d'exploration de données. Par exemple, pour une boutique de vente de vêtements on aimerait connaitre ce qu'un client serait susceptible d'acheter après un premier achat. On pourrait capitaliser sur cette information afin de recommander aux clients des articles supplémentaires pour stimuler les ventes.

2.11 Arbres de décision

Dans l'analyse prédictive, les arbres de décision sont un type de modèle prédictif. En raison de sa nature extrêmement simple, elle est catégorisée parmi les techniques d'apprentissage automatique en boite blanche.

Un arbre de décision permet aux utilisateurs de comprendre clairement la relation entre les entrées de données et les sorties. Lorsqu'on combine divers modèles d'arbre de décision, on obtient des modèles prédictifs d'analyse appelés forêts aléatoires. Même si les forêts aléatoires sont en général plus précises que les arbres de décisions seuls, il n'est pas toujours aisé de comprendre leurs sorties en fonction de leurs entrées.

Exemple d'arbre de décision :

Voici ci-après (voir figure 2.5) un modèle d'arbre de décision.

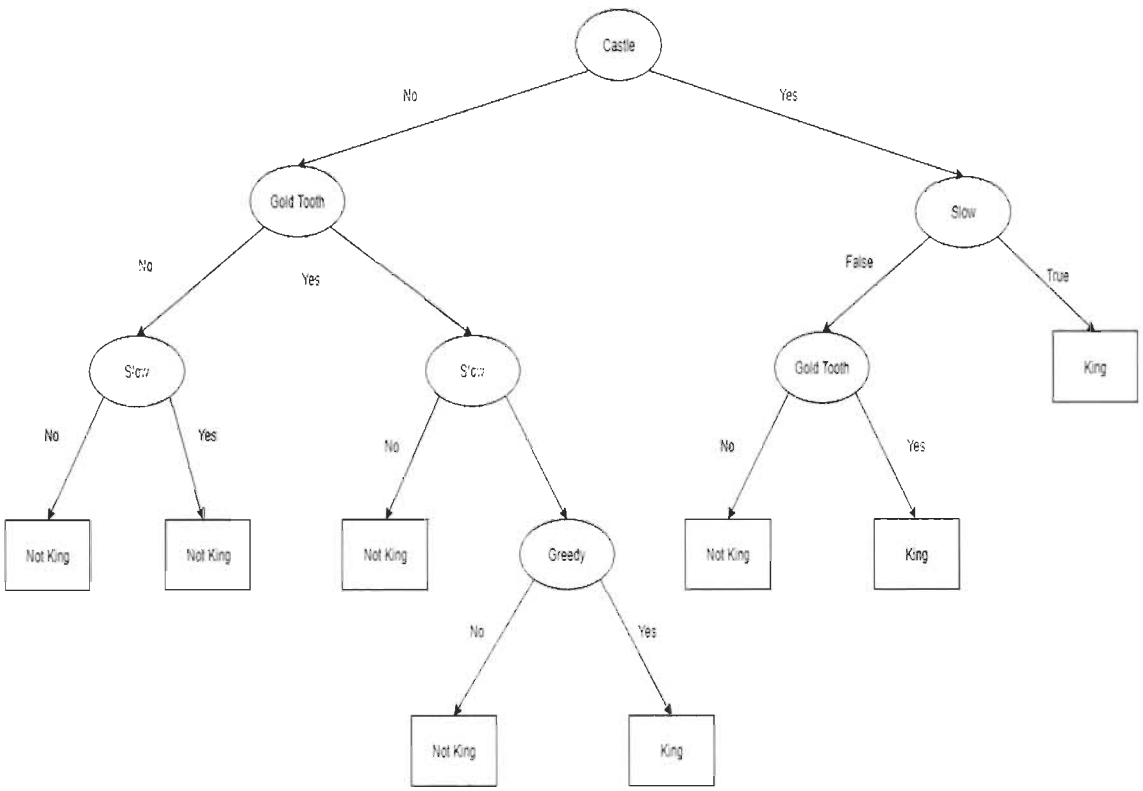


Figure 2.5. Exemple de modèle d'arbre de décision [25]

Exemple de forêt aléatoire :

L'illustration ci-après (voir figure 2.6) montre un exemple de forêt aléatoire.

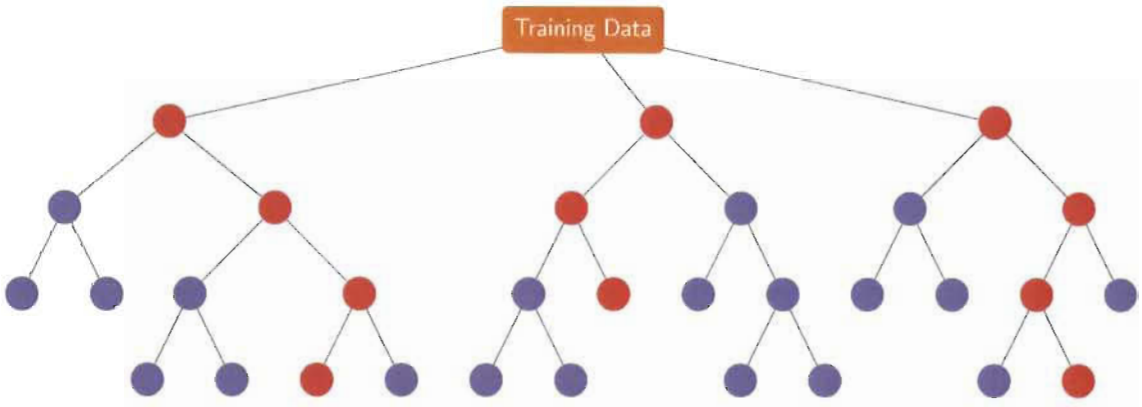


Figure 2.6. Exemple de forêt aléatoire [26]

2.12 Techniques statistiques

Les techniques statistiques sont le noyau de la plupart des analyses impliquées dans le processus d'exploration de données. Les modèles statistiques sont basés sur des approches statistiques, qui permettent de générer des valeurs numériques en réponse aux objectifs de l'utilisateur. Par exemple, les réseaux de neurones utilisent différentes mesures ou poids statistiques pour déterminer si une image est un chien ou un chat dans les systèmes de reconnaissance d'image.

Les modèles statistiques représentent l'une des deux principales branches de l'intelligence artificielle. L'apprentissage automatique apporte un dynamisme dans l'élaboration des modèles ainsi ils peuvent s'améliorer avec le temps, tandis que les modèles de certaines techniques statistiques sont statiques.

2.13 Visualisation

Les visualisations de données sont un autre élément important de l'exploration de données. Il permet aux utilisateurs d'obtenir un aperçu des données. Les visualisations les plus intéressantes sont dynamiques et utiles pour diffuser des données en temps réel et caractérisées par différentes couleurs qui révèlent différentes tendances et modèles dans les données.

Les interfaces utilisateurs ou encore tableaux de bord sont un moyen puissant pour mettre en évidence visuellement des modèles dans les données. En manipulant différentes métriques, les utilisateurs peuvent mieux représenter les sorties numériques des modèles statistiques.

2.14 Réseaux de neurones

Un réseau de neurones est un type spécifique de modèle d'apprentissage automatique qui est souvent utilisé avec l'intelligence artificielle et l'apprentissage profond [37]. Les paradigmes des réseaux de neurones ont été influencés par le fonctionnement du cerveau humain, on parlera de couches ou encore de neurones artificiels. Les réseaux de neurones sont l'un des modèles d'apprentissage automatique les plus précis.

Bien qu'un réseau de neurones puisse être un outil puissant dans l'exploration de données, les utilisateurs devront aussi considérer la difficulté à comprendre les sorties des réseaux de neurones. Ils devront adapter la technique d'exploration de données en fonction des objectifs visés.

Exemple de réseau de neurones :

Voir la figure ci-après (figure 2.7) illustrant le modèle d'un réseau de neurones.

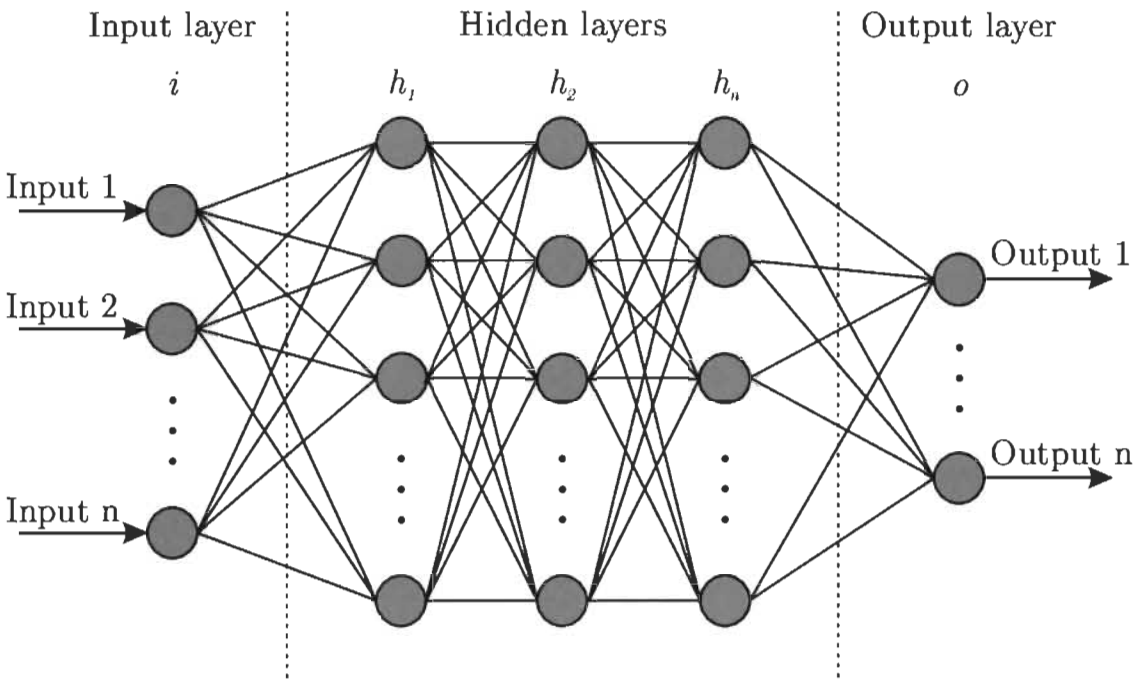


Figure 2.7. Exemple de perceptron multicouche [43]

2.15 Entreposage de données

Le stockage de données est une partie importante du processus d'exploration de données. Généralement, il s'agissait de stocker des données structurées dans des systèmes de gestion de base de données relationnelles pour des fins d'analyses. Aujourd'hui au-delà des données structurées, il existe des

entrepôts de données le nuage informatique capable de stocker des données semi-structurées et non structurées en utilisant des systèmes comme Hadoop ou Apache Spark. Les entrepôts utilisant les technologies du nuage informatique sont désormais capables de fournir une analyse approfondie et en temps réel des données.

Logo de Hadoop et Apache Spark :

Vous trouverez ci-après respectivement les logos de Hadoop et Apache Spark (voir figure 2.8 et figure 2.9).

Figure 2.8. Logo de Hadoop [18]

Figure 2.9. Logo de Apache Spark [18]

2.16 Traitement de la mémoire à long terme

Pour effectuer des analyses de données sur une période prolongée, les organisations utilisent les données historiques stockées dans les entrepôts de données afin d'être capables d'identifier des modèles qui pourraient être trop subtils à détecter. Cette capacité d'analyser les données sur de longues périodes fait référence au traitement de la mémoire à long terme. Par exemple, en analysant les différents mouvements des abonnés à un service sur une période de plusieurs années, une organisation peut trouver des indices subtils qui pourraient conduire à réduire le taux de désabonnement des finances.

2.17 Apprentissage automatique et intelligence artificielle

L'apprentissage automatique et l'intelligence artificielle représentent certainement les développements les plus avancés de l'exploration de données. Certaines formes avancées de l'apprentissage automatique telles que l'apprentissage en profondeur offrent des prédictions très précises lorsqu'on travaille avec des données à grande échelle. Par conséquent, ils sont utiles pour traiter les données dans les déploiements d'IA comme la vision par ordinateur, la

reconnaissance vocale ou l'analyse de texte sophistiquée à l'aide du traitement du langage naturel [49]. Ces techniques d'exploration de données sont utiles pour déterminer la valeur à partir de données semi-structurées et non structurées.

2.18 Conclusion

L'exploration de données dispose donc de techniques d'analyse variée laissée au choix de l'utilisateur. Il devra choisir les techniques les plus pertinentes pour son travail et disposer des outils appropriés pour optimiser au mieux ses analyses. Avec l'avènement du Big data, plusieurs outils sont mis à la disposition des scientifiques pour au mieux répondre aux besoins des problématiques. Certains outils de science des données tels que R, Python, Spark sont très utilisés pour l'analyse, la visualisation et le traitement de données.

L'analyse et la visualisation des données sont donc des étapes cruciales dans le processus d'exploration de données. Certaines techniques qui visent à mettre en emphase l'association entre les variables dans les données sont très utiles pour comprendre les données.

Dans ce travail de recherche, nous allons mettre l'accent sur cette approche. Nous travaillerons plus précisément, sur la recherche d'associations intéressantes dans les données en utilisant des mesures statistiques et des aperçus visuels.

CHAPITRE 3 - Les règles d'association

3.1 Introduction

Avec le volume énorme de données stockées dans les entrepôts de données, il est devenu plus qu'évident de se munir de différents outils d'analyse afin de traiter l'information et d'en extraire la connaissance.

Les chercheurs et les organisations utilisent diverses techniques pour découvrir des associations ou des corrélations intéressantes dans les bases de données. Une technique couramment utilisée est l'extraction des règles d'association.

Les règles d'association sont une méthode qui permet de découvrir des relations entre des données dans un ensemble de données. Par souci d'allègement du texte, nous utiliserons le terme ***ensemble de données*** pour qualifier aussi bien les bases de données que les autres référentiels de données. Les règles d'association permettent de rechercher les relations entre les objets fréquemment utilisés ensemble. Les applications des règles d'association sont l'analyse des données du panier, la classification, le marketing croisé, le regroupement, la conception de catalogues, l'analyse des pertes, diagnostics médicaux, etc. Par, exemple, un client qui achète du pain, achètera aussi probablement du beurre. Un

autre qui achète un ordinateur portable achètera aussi probablement une carte mémoire. Sous certains critères on dira donc que la relation Pain → beurre constitue une règle et ordinateur → carte mémoire en constitue une autre.

Les règles d'association utilisent deux critères de base : l'indice de support et l'indice de confiance. Ils identifient les relations et les règles générées en analysant les données pour les modèles fréquemment utilisés. Des règles d'associations sont généralement nécessaires pour satisfaire simultanément l'indice de support et l'indice de confiance minimal spécifié par l'utilisateur. Par conséquent, la force d'une règle d'association est mesurée par son indice de support et son indice de confiance.

Définissons un ensemble de transactions nommé T (provenant de la source de données) et deux éléments A et B tels que : $A \in T$, $B \in T$.

L'indice de support d'une règle A → B est défini par la proportion de transactions de T qui contiennent $A \cap B$ (à la fois A et B, et non A ou B), soit $Supp(A \cap B)$. Il s'agit donc d'une estimation de la probabilité $Pr(A \cap B)$.

L'indice de confiance d'une règle A→B est défini par la proportion de transactions de T contenant A qui contiennent aussi B soit A∩B.count / A.count. Il peut être vu comme une estimation de la probabilité conditionnelle $Pr(B|\grave{A})$.

Certains algorithmes utilisent une autre mesure appelée Lift. Le <lift> d'une règle A→B mesure l'amélioration apportée par la règle d'association par

rapport à un jeu de transactions aléatoire (ou A et B seraient indépendants). Il est défini par Supp(A∩B) / Supp(A).Supp(B). Un lift supérieur à 1 traduit une corrélation positive de A et B et donc le caractère significatif de l'association.

Ces mesures sont fréquemment utilisées dans les algorithmes d'extraction des règles d'association tels que Apriori, AprioriTID, et FP-growth [52]. La plupart de ces algorithmes génèrent un nombre pléthorique de règles d'association, alors pour pallier à cela plusieurs techniques d'élagage ont été proposées afin de ne garder que les règles intéressantes.

Dans ce chapitre on présentera un état de l'art sur les règles d'association, ainsi que les techniques d'extraction, les techniques d'élagage, les différents algorithmes et leurs variantes.

3.2 Notions et définitions sur les règles d'association

3.2.1 Représentation des données

Les données sont généralement représentées sous forme transactionnelles (Voir Tableau 3.1).

Tableau 3.1. Base de données transactionnelles [23]

Transaction	Items
t1	{T-shirt, Pantalon, Ceinture}
t2	{T-shirt, Veste}
t3	{Veste, Gants}
t4	{T-shirt, Pantalon, Veste}
t5	{T-shirt, Pantalon, Chaussure, Veste, Ceinture}
t6	{Pantalon, Chaussure, Ceinture}
t7	{Pantalon, Ceinture, Chaussure}

Dans le tableau 3.1 on peut observer sept transactions d'une boutique de vêtements. Chaque transaction montre des items achetés dans cette transaction. On peut représenter nos items comme un ensemble d'éléments :

$I=\{i1,i2,...,ik\}$

Dans notre cas cela correspond à :

$I=\{T\text{-}shirt, Pantalon, Ceinture, Veste, Gants, Chaussures\}$

On comprend donc qu'un item est un objet, article ou élément d'une base de données $(i1, i2, i3, ..., ik)$

Par exemple :

Item1 = T-shirt, Item2 = Pantalon, Item3 = Ceinture, etc.

Une transaction est représentée par l'expression suivante : $T=\{t1,t2,...,tn\}$

Par exemple :

$t1=\{T\text{-}shirt,Pantalon,Ceinture\}$

Alors, une règle d'association est définie comme une implication de la forme :

$X \rightarrow Y$, *ou* $X \subset I$, $Y \subset I$ *et* $X \cap Y = \emptyset$

Par exemple :

$\{T\text{-}shirt,Pantalon\} \rightarrow \{Ceinture\}$

Au cours du traitement ces informations sont représentées sous forme binaire. On obtient un tableau binaire (voir tableau 3.2) qu'on note par 0 l'évènement d'absence de chaque item et par 1 sa présence dans la transaction.

Tableau 3.2. Base de données binaire [23]

Transactions	T-shirt	Pantalon	Ceinture	Veste	Gants	Chaussures
T1	1	1	1	0	0	0
T2	1	0	0	1	0	0
T3	0	0	0	1	1	0
T4	1	1	0	1	0	0
T5	1	1	1	1	0	1
T6	0	1	1	0	0	1
T7	0	1	1	0	0	1

3.2.2 Support d'un Itemset

Le support indique la fréquence d'apparition de l'ensemble d'éléments dans l'ensemble de données.

$$supp(X \rightarrow Y) = \frac{|X \cup Y|}{n}$$

Figure 3.1. Support d'un Itemset

En d'autres termes, c'est le nombre de transactions avec à la fois X et Y divisé par le nombre total de transactions. Un itemset avec un faible indice de support n'est pas considéré comme pertinent. En se servant de l'exemple de la boutique de vêtements, on obtient :

supp(T-shirt → Pantalon) = 3/7 = 43%

supp(Pantalon → Ceinture) = 4/7 = 57%

supp(T-shirt → Ceinture) = 2/7 = 28%

3.2.3 Itemset Fréquent

On considère qu'un itemset est fréquent si et seulement si son indice de support est supérieur au support minimum défini par l'utilisateur.

3.3 Règles d'association

3.3.1 Définition

Une règle d'association est une application X→Y qui exprime une corrélation de cooccurrence [2].

Étant donné un ensemble de transactions T, trouver toutes les règles ayant un support supérieur ou égal au support minimal (*minsupp*) et une confiance supérieure ou égale à la confiance minimale (*minconf*) où minsupp et minconf sont des seuils définis par l'utlisateur.

3.3.2 Métriques d'une règle d'association

Les métriques des règles d'association sont généralement définies par l'utilisateur. C'est donc un choix arbitraire souvent influencé par la nature des données ou les objectifs visés. Des études de chercheurs comme Hajek, Havel et Chytil [16] ont introduit les notions de support et de confiance. Le lift est une autre métrique qui sera introduite bien plus tard.

3.3.2.1 Support d'une règle d'association

Le support d'une règle d'association s'exprime par le nombre de transactions qui contiennent les éléments de X et les éléments de Y divisé par le nombre total des transactions de l'ensemble de données.

Dans une base de données D, le support d'une règle d'association X →Y est le rapport du nombre de transactions qui contiennent X et Y par nombre total de transactions [67].

$$supp(X \rightarrow Y) = \frac{\text{Card}(X \cup Y)}{\text{Card(D)}}$$

Figure 3.2. Support d'une règle d'association

3.3.2.2 Confiance d'une règle d'association

La confiance fait référence à la probabilité qu'un article Y soit également acheté si l'article X est acheté. Il peut être calculé en recherchant le nombre de transactions pour lesquelles X et Y sont achetés ensemble, divisé par le nombre total de transactions pour lesquelles X est acheté.

$$conf(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp(X)}}$$

Figure 3.3. Confiance d'une règle d'association

Par exemple, la règle T-shirt → Pantalon a une confiance de 3/4, ce qui signifie que 75% des règles contenant l'item T-shirt sont corrects (75% de fois qu'un client achète, un T-shirt, des pantalons sont achetés aussi). D'autres exemples :

$$conf(Pantalon \rightarrow Ceinture) = \frac{4/7}{5/7} = 80\%$$

$$conf(T\text{-}shirt \rightarrow Ceinture) = \frac{2/7}{4/7} = 50\%$$

$$conf(\{T\text{-}shirt, Pantalon\} \rightarrow \{Ceinture\}) = \frac{2/7}{3/7} = 66\%$$

3.3.2.3 Lift d'une règle d'association

Le lift d'une règle est le rapport du support observé à celui attendu si X et Y étaient indépendants, et est défini comme suite :

$$lift(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)\text{supp}(Y)}$$

Figure 3.4. Lift d'une règle d'association

Un lift élevé indique des associations fortes. Par exemple :

$$Lift(T\text{-}shirt \rightarrow Pantalon) = \frac{3/7}{(4/7)(5/7)} = 1.05$$

$$Lift(Pantalon \rightarrow Ceinture) = \frac{4/7}{(5/7)(4/7)} = 1.4$$

$$Lift(T\text{-}shirt \rightarrow Ceinture) = \frac{2/7}{(4/7)(4/7)} = 0.875$$

$$Lift(\{T\text{-}shirt, Pantalon\} \rightarrow \{Ceinture\}) = \frac{2/7}{(3/7)(4/7)} = 1.17$$

3.4 Extraction des règles d'association

3.4.1 Algorithme d'extraction des règles d'association

L'exploration de règles d'association est le processus qui consiste à découvrir les modèles qui se produisent fréquemment dans une donnée, telle que les éléments fréquents dans le panier d'un client. Les éléments fréquents désignent l'ensemble des éléments fréquemment rencontrés dans une base de données transactionnelle [2]. Ces éléments fréquents sont utilisés pour générer l'ensemble des règles d'association. En d'autres termes, les règles d'association décrivent simplement le comportement d'une entité soumis à certaines conditions. Les éléments sont considérés comme fréquents s'ils se produisent dans la base de données pendant un certain temps supérieur ou égal à un seuil prédéfini appelé Support minimal. De nombreux algorithmes pour générer des règles d'association ont été proposés. Nous en présenterons quelques-uns à la suite de ce document.

3.4.1.1 Algorithme Apriori

L'exploration des règles d'association est considérée comme une approche en deux étapes:

1ère étape : Génération d'ensemble d'articles fréquents : rechercher tous les ensembles d'articles fréquents avec un support supérieur ou égal au support minimal prédéterminé

2nd étape : lister toutes les règles d'association des ensembles d'éléments fréquents. Calculer le support et la confiance pour toutes les règles. Élaguer les règles qui ne satisfont pas le support minimal et la confiance minimale.

La génération d'ensemble d'articles fréquents est l'étape la plus lourde en calcul, car elle nécessite une analyse complète de la base de données.

L'un des principes de l'algorithme Apriori est que tout sous-ensemble d'un ensemble d'éléments fréquent doit également être fréquent. En d'autres termes, aucun surensemble d'un ensemble d'éléments peu fréquent ne doit être généré ou testé [4].

Ce principe est très bien représenté dans Itemset ***treillis*** qui est une représentation graphique du principe de l'algorithme Apriori. Il se compose d'un nœud d'ensemble d'éléments k et d'une relation de sous-ensembles de cet ensemble d'éléments k (voir figure 3.5).

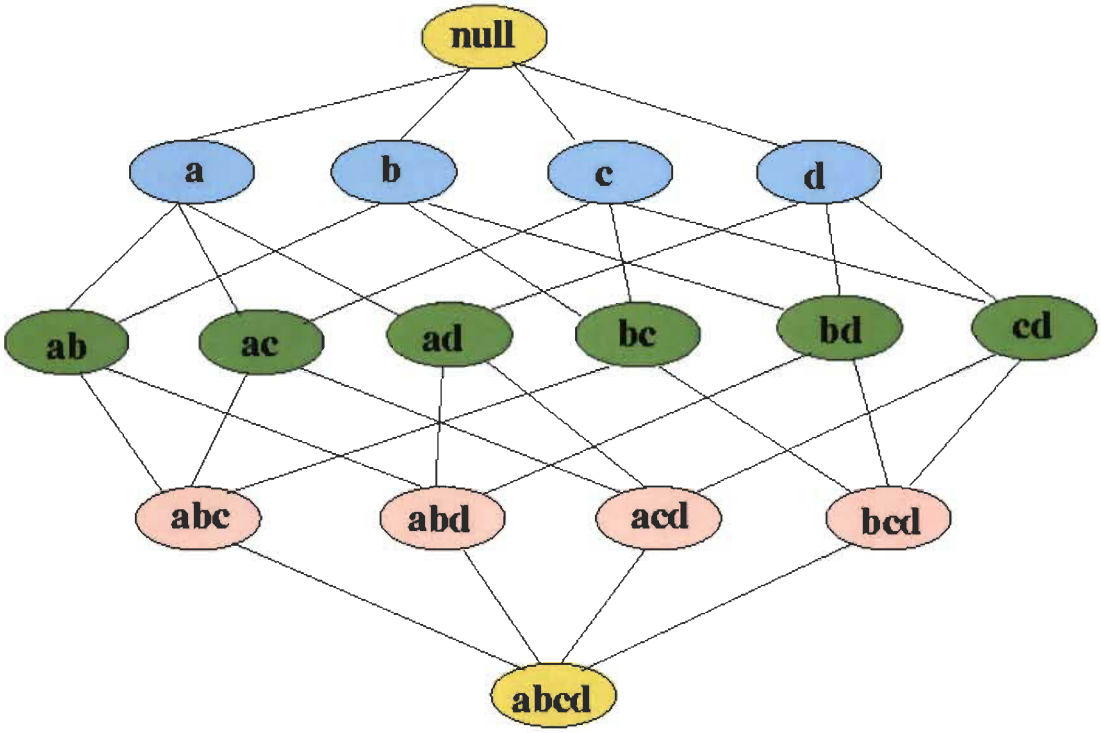
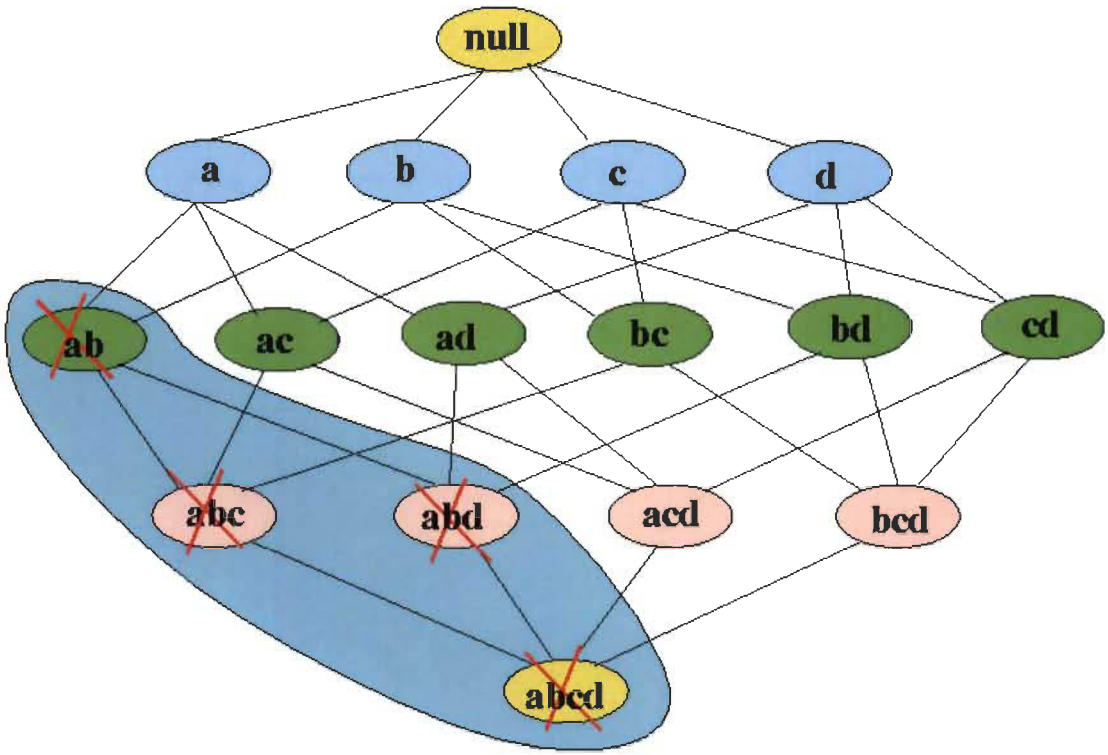


Figure 3.5. Graphe Itemset treillis [55]

Dans la figure 3.5 on observe que dans le bas se trouvent tous les éléments dans les données de transaction, puis en se déplaçant vers le haut, on crée des sous-ensembles jusqu'à l'ensemble nul. Cela montre à quel point il sera difficile de générer un ensemble d'éléments fréquents en trouvant un support pour chaque combinaison. La figure ci-après (voir figure 3.6) montre dans quelle mesure Apriori contribue à réduire le nombre d'ensembles à générer.



0.1 Figure 3.6. Graphe Itemset treillis premier passage [55]

Si l'ensemble d'items {a, b} est peu fréquent, nous n'avons pas besoin de prendre en compte tous ses sous-ensembles. Parcourons pas à pas un exemple afin d'observer l'efficacité de l'algorithme Apriori (voir figure 3.7) :

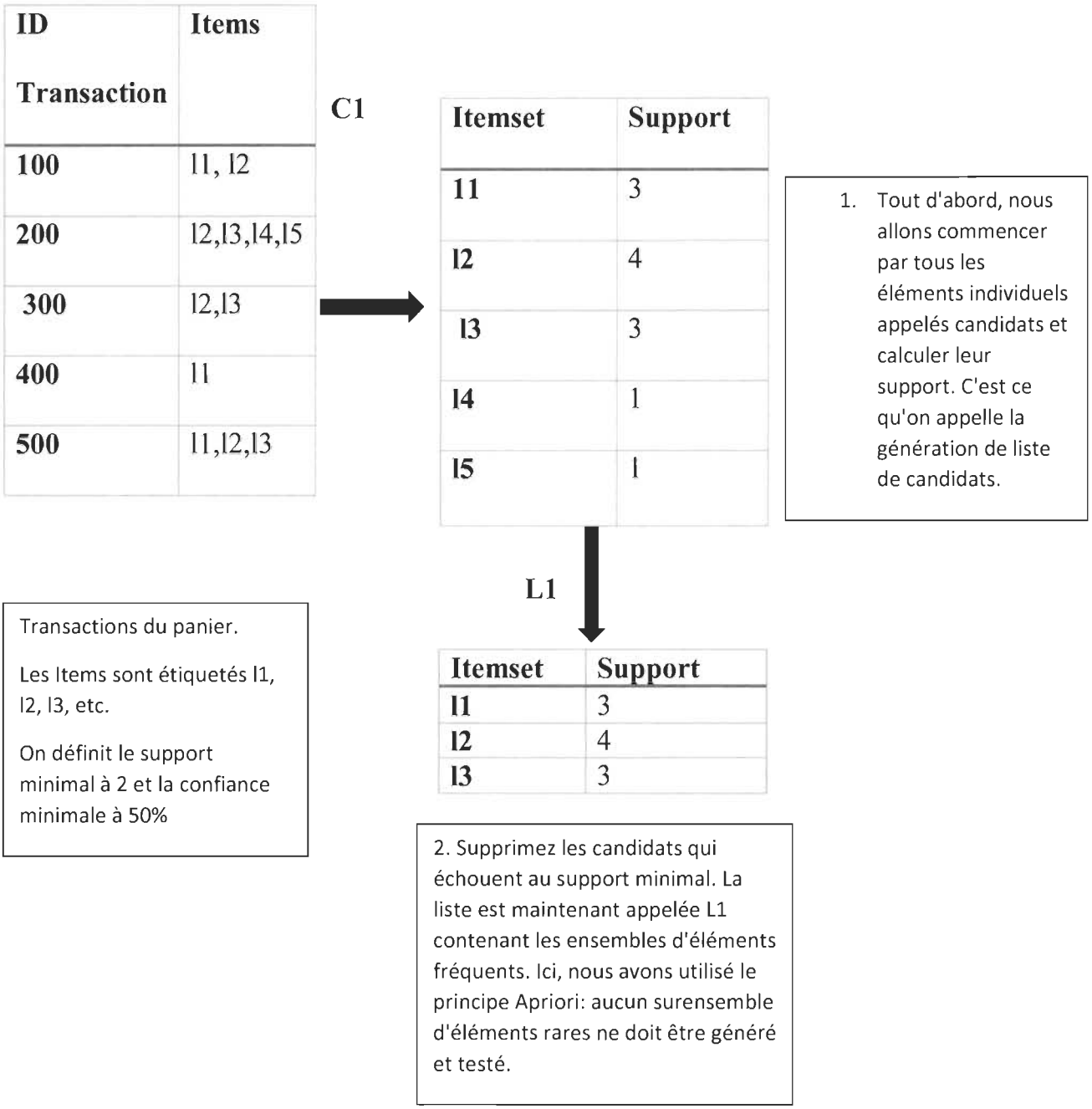


Figure 3.7. Première étape de l'algorithme Apriori [55]

Comme vous pouvez le voir, Apriori commence par créer une liste de candidats pour l'ensemble à 1 élément qui comprendra tous les éléments, qui sont présents dans les données de transaction, individuellement. En tenant compte des données

de transaction du monde réel, on se rend vite compte à quel point cette génération de candidats est couteuse en termes de calcul. Ici APRIORI joue son rôle et aide à réduire le nombre de la liste des candidats, et des règles utiles sont générées à la fin. Dans les étapes suivantes (voir figure 3.8), on verra comment nous arrivons à la fin de la génération de Frequent Itemset, c'est-à-dire la première étape de l'exploration des règles d'association.

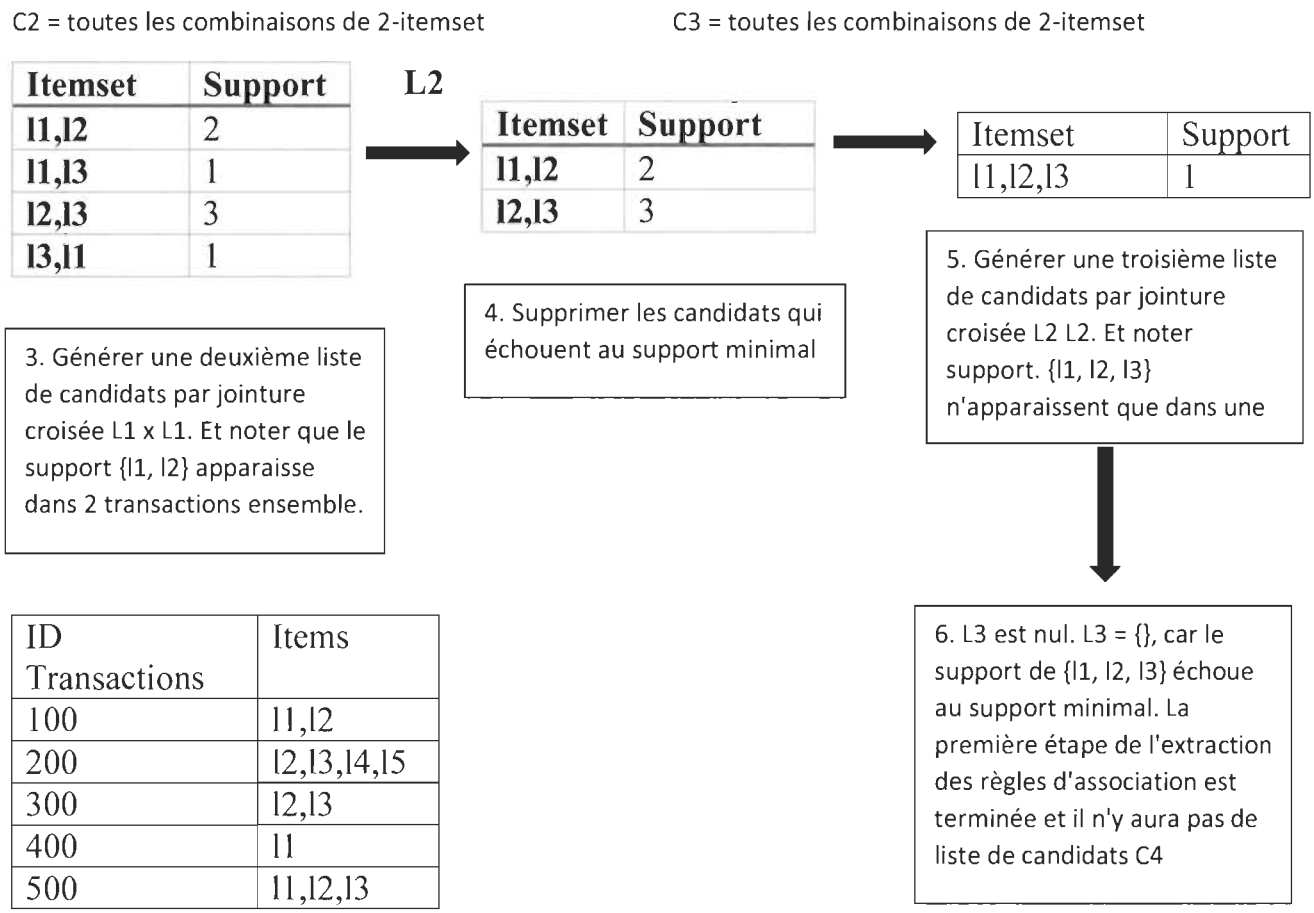


Figure 3.8. Seconde étape de l'algorithme Apriori [55]

La prochaine étape consistera à répertorier tous les ensembles d'éléments fréquents. On prendra le dernier ensemble d'éléments fréquents non vide, qui dans cet exemple est L2 = {I1, I2}, {I2, I3}. Ensuite, on rendra tous les sous-ensembles non vides des ensembles d'articles présents dans cette liste d'ensembles d'articles fréquents. Voir figure 3.9 ci-dessous :

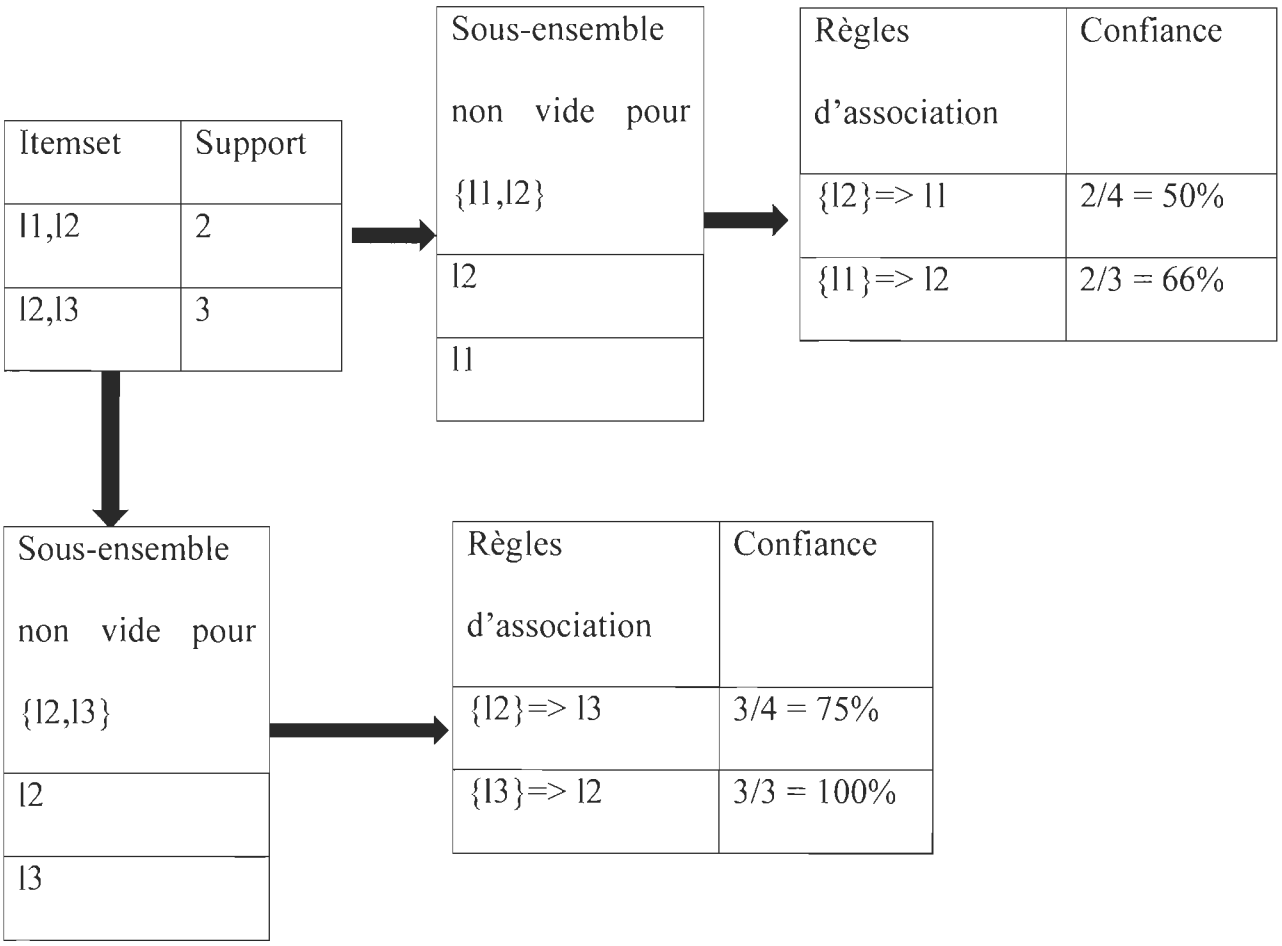


Figure 3.9. Troisième étape de l'algorithme Apriori [55]

Étant donné que la confiance minimale a été définie à 50%. La confiance est calculée comme suite :

$$C(A \rightarrow B) = P(AUB) / P(A) = n(AUB) / n(A)$$

À partir de la figure ci-dessus on remarque qu'il y a quatre règles fortes. Par exemple la règle l2→l3 ayant un indice de confiance égal à 75% dit que 75% des personnes qui achètent l2 achèteront aussi l3.

Voici un résumé de la procédure de l'algorithme Apriori (voir figure 3.10) :

```
Clk: Ensemble d'éléments candidats de taille k
Flk: éléments fréquents de taille k
Fl1 = {éléments fréquents};
Pour (k = 1; Flk! = Null; k ++) commencent
Clk + 1 = candidats générés à partir de Flk;
Pour chaque transaction t dans la base de données D do
Incrémenter la valeur de comptage de tous les candidats dans
Clk + 1 qui sont contenus dans t
Flk + 1 = candidats dans Clk + 1 avec min_support
Fin
Retournez Flk;
```

Figure 3.10. Procédure de l'algorithme Apriori [9]

L'algorithme Apriori présente deux inconvénients.

Le premier est le processus complexe de génération de candidats qui utilise la plupart du temps, de l'espace et de la mémoire. Un autre inconvénient est qu'il nécessite plusieurs analyses de la source de données.

3.4.1.2 Algorithme AprioriTID

AprioriTiD reprend un peu les grands principes d'Apriori seulement que dans son processus, la source de données n'est pas utilisée pour compter le support des jeux d'éléments candidats après le premier passage. Le processus de génération d'un jeu d'éléments candidats est identique à celui de l'algorithme Apriori. Un autre ensemble **C'**, est généré dont chaque membre possède le TID qui est une série de numéros identifiant de façon unique chaque transaction et les grands ensembles d'articles présents dans cette transaction. L'ensemble généré, c'est-à-dire **C'**, est utilisé pour compter le support de chaque jeu d'items candidat.

3.4.1.3 Algorithme FP-Growth

FP-growth nécessite la construction d'un arbre FP-tree. Pour cela, il faut deux passages. FPgrowth utilise la stratégie de division et de conquête. Deux analyses de la source de données sont nécessaires. Lors de la première analyse de la source de données, il calcule d'abord une liste d'éléments fréquents triés par fréquence en ordre décroissant (F-List). Lors de la seconde analyse, la base de données est compressée dans un arbre nommé FP-tree. Cet algorithme effectue une extraction sur l'arborescence FP-tree de manière récursive. Il existe un problème de recherche de jeux d'éléments fréquents, qui est converti en recherche et construction récursive d'arbres. Les ensembles d'éléments fréquents sont générés en seulement deux passages sur la source de données et sans processus

de génération de candidat. Il existe deux sous-processus de génération de modèles fréquents, à savoir: la construction de l'arborescence FP-tree et la génération des modèles fréquents à partir de l'arbre FP. FP-tree est construit sur le jeu de données en utilisant 2 passages qui sont les suivants:

Passage 1:

1) Scanner les données et trouver un support pour chaque élément.

2) Supprimer les articles peu fréquents.

3) Trier les éléments fréquents dans l'ordre décroissant, en fonction de leur indice de support.

En utilisant cet ordre, nous pouvons construire FP-tree, de sorte que les préfixes puissent être partagés.

Passage 2:

1) Ici, les noeuds correspondent aux items.

2) FP-Growth lit une transaction en une fois, puis l'enregistre sur un chemin.

3) Un ordre fixe est utilisé pour que les chemins puissent se chevaucher quand les transactions partagent les mêmes items.

3.4.1.4 Tableau récapitulatif

Numéro	Algorithme	Description	Mérites	Inconvénients
1	Apriori	Le processus en deux étapes consiste à rechercher d'abord des ensembles d'éléments volumineux, puis à analyser la source de données pour vérifier le nombre de supports pris en charge par les ensembles d'éléments correspondants.	Très efficace pour générer des ensembles d'articles fréquents.	Le premier inconvénient est le processus complexe de génération de candidats qui utilise la plupart du temps, de l'espace et de la mémoire. Un autre inconvénient est qu'il nécessite plusieurs analyses de la source de données.
2	AprioriTID	Il produit le même résultat que Apriori. Mais il utilise un mécanisme différent pour compter le support des jeux d'éléments.	Capable de produire des ensembles d'articles fréquents à partir de grandes sources de données.	Nécessite plusieurs analyses de la source de données.

3	FP-Growth	Il analyse la source de données seulement deux fois pour générer des ensembles d'éléments fréquents sans aucun processus d'itération. La première analyse construit l'arborescence FP-tree et l'analyse suivante génère des ensembles d'articles fréquents à partir de l'arbre en utilisant une procédure appelée FP-Growth.	Peut produire des ensembles d'articles fréquents impliquant seulement deux analyses de la source de données	L'inconvénient de FP-Growth est qu'il doit élaborer des bases de modèles conditionnels et construire de manière récursive des arbres. Il fonctionne mal dans les ensembles de données de modèles longs.

3.5 Élagage des règles d'association

Un certain nombre de techniques d'élagage ont été utilisées dans le contexte de l'extraction de données, dont certaines sont issues d'arbres de décision, d'autres de statistiques telles que l'estimation des erreurs pessimistes, les tests du khi-deux [1]. Ces techniques d'élagage sont utilisées soit pendant la phase de découverte des règles (pré-élagage), telle que le test du coefficient de corrélation de Pearson [1], soit pendant la phase de construction du classificateur (Postpruning), telle que la couverture de la base de données [1] ou élagage paresseux [1]. Une étape

d'élagage précoce a lieu avant que les règles ne soient générées en éliminant ces règles qui n'ont pas dépassé le seuil **minsupp** (support minimal) qui peut survenir lors du processus de recherche des règles fréquentes. Cette section présente les techniques d'élagage actuellement utilisées par les algorithmes de classification associative.

3.5.1 Techniques d'élagage de couverture de l'ensemble de données

Les techniques d'élagage de règles qui considèrent la règle comme une règle significative en fonction de sa capacité de couverture est appelée techniques de couverture de la source de données. Elles sont ensuite décrites dans la section suivante et décrivent leurs mécanismes de travail.

3.5.1.1 Couverture de l'ensemble de données

La couverture de la source de données est la première méthode heuristique appliquée dans CBA [1] pour sélectionner un sous-ensemble de règles afin de constituer le classificateur. La méthode de couverture de source de données est une méthode d'élagage simple et efficace; il évalue l'ensemble complet des règles générées par rapport au jeu de données d'apprentissage. La couverture de la source de données utilise une méthode heuristique dans laquelle, à partir de la règle la mieux classée, tous les cas d'apprentissage entièrement couverts par cette règle sont marqués pour être ensuite supprimés du jeu de

données d'apprentissage, enfin la règle est entrée dans le classificateur. Pour une règle ne couvrant aucun cas de formation (le corps de la règle ne correspond pas complètement à un cas de formation), la règle est supprimée. La méthode de couverture de la source de données se termine lorsque l'ensemble de données d'apprentissage est totalement couvert ou s'il n'y a plus de règles à ajouter. S'il ne reste plus de règles à évaluer, les cas de formation non couverts restants sont utilisés pour générer la règle de classe par défaut qui représente la classe de fréquence la plus grande (classe majoritaire) dans les cas non classés restants. Il convient de noter que la règle de classe par défaut est utilisée lors de l'étape de prédiction dans les cas où aucune règle de classificateur n'est applicable au scénario de test. Enfin, avant la fin de la couverture de la source de données, la première règle comportant le moins d'erreurs est identifiée comme règle de coupure. Toutes les règles après cette règle ne sont pas incluses dans le classificateur final, car elles génèrent souvent des erreurs [1]. La méthode de couverture de la source de données a été critiquée par [1] dans la mesure où elle élimine parfois des connaissances utiles. Alternativement, ils insistent sur le fait que les classificateurs riches fournissent souvent des connaissances utiles et riches pendant l'étape de classification.

3.5.1.2 Élagage paresseux (différé)

Les algorithmes associatifs paresseux [1] estiment que l'élagage devrait être limité aux règles qui couvrent de manière incorrecte les cas utilisés pour l'entrainement . Seules ces règles qui conduisent à une classification incorrecte sur les cas de test sont supprimées. La technique de couverture de source de données supprime toute règle qui ne permet pas de couvrir entièrement un cas de formation (cas utilises pour l'entrainement) et l'exactitude de la classe. L'élagage de règles paresseux est utilisé lorsque l'ensemble complet de règles est découvert et classé dans un ordre décroissant dans lequel des règles plus longues (règles avec plus d'éléments dans son antécédent) sont privilégiées par rapport aux règles générales. Pour chaque règle à partir de la règle la mieux classée, si elle couvre correctement un cas d'apprentissage, elle sera entrée dans les jeux de règles principaux et tous les cas correspondants seront supprimés de l'ensemble de données d'apprentissage. Par contre, si une règle de rang supérieur couvre correctement les cas d'entraînement, elle sera insérée dans les jeux de règles secondaires (jeux de règles Spare). Enfin, si la règle sélectionnée couvre à tort un cas de formation, elle sera supprimée. Le processus est répété jusqu'à ce que toutes les règles découvertes soient testées ou que le jeu de données

d'apprentissage soit vide. Cet élagage paresseux produira deux jeux de règles, un jeu primaire contenant toutes les règles couvrant correctement un cas de formation et un jeu secondaire contenant des règles qui n'ont jamais été utilisées lors de l'élagage, car certaines règles de rang supérieur ont été modifiées. La différence distinctive entre la couverture de la source de données et l'élagage paresseux réside dans le fait que les règles secondaires extraites par la méthode différée sont complètement supprimées par la couverture de source de données. En d'autres termes, le classificateur résultant de l'élagage de la couverture de source de données ne contient pas l'ensemble de règles secondaires de l'élagage différé, et sa taille est donc souvent inférieure à celle des algorithmes reposant sur ce dernier. Ceci est en effet un avantage, en particulier dans les applications où l'utilisateur final veut contrôler et gérer les règles. Des études empiriques réalisées dans [1] contre un certain nombre de jeux de données UCI ont révélé que l'utilisation d'algorithmes paresseux tels que L3 et L3G permettait parfois d'obtenir un taux de précision supérieur à celui des algorithmes CBA, à savoir CBA2 et MCAR. Enfin, nous concluons que les algorithmes basés sur l'élagage paresseux ont souvent des scores de performance élevés, mais faibles en termes d'efficacité en raison de la grande taille du classificateur qui prend plus de temps à générer des règles et à apprendre le classificateur.

3.5.1.3 Élagage des longues règles

Une méthode d'évaluation des règles qui supprime les règles longues (règles spécifiques) dont les valeurs de confiance sont supérieures à leur sous-ensemble (règles générales) a été introduite dans [1]. La règle générale avec la valeur de confiance la plus élevée est utilisée pour élaguer les spécifiques. En d'autres termes, il élimine la redondance des règles, car de nombreuses règles découvertes ont des valeurs d'attributs partagés communes dans leurs antécédents, ce qui entraîne souvent des règles redondantes, en particulier lorsque la taille du classificateur devient grande. CMAR [1] a été le premier algorithme à utiliser l'élagage à règles longues. L'ensemble de règles est d'abord classé en fonction de la confiance, du support et de la longueur des règles. Les règles sont ensuite stockées dans une structure CR-tree. Une requête de récupération sur l'arborescence est activée pour vérifier si une règle peut être supprimée. Le test du chi carré est appliqué à chaque règle R: ri -> c, pour déterminer si ri est positivement corrélé avec c ou non. L'algorithme sélectionne uniquement les règles qui sont corrélées positivement pour former le classifieur. Un certain nombre d'algorithmes AC utilisent ce type d'élagage, notamment ARC-BC et les règles négatives [1]. Les résultats d'expérimentation présentés dans [1] montrent que l'utilisation de cette technique d'élagage aura une incidence positive sur l'efficacité par rapport à d'autres techniques.

3.5.2 Techniques mathématiques

Certaines techniques d'élagage basées sur les mathématiques ont été proposées. La plupart d'entre eux ont tendance à mesurer la corrélation entre différents objets pour décider s'ils sont corrélés ou non.

3.5.2.1 Test de Khi-deux

Le test de khi-deux proposé par [1] sert à déterminer s'il existe une différence significative entre les fréquences observées et les fréquences attendues dans une ou plusieurs catégories. Elle est définie comme une hypothèse connue qui examine la relation entre deux objets [1]. L'évaluation utilisant $\chi 2$ pour un groupe d'objets afin de tester leur indépendance ou leur corrélation est donnée comme suit:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Figure 3.11. Évaluation khi-deux

Ou : ei sont les fréquences attendues et oi les fréquences observées. Lorsque les fréquences attendues et les fréquences observées sont notablement différentes; l'hypothèse selon laquelle ils sont corrélés est rejetée.

Cette méthode a été utilisée dans AC (Association classification) pour élaguer les règles à corrélation négative. Par exemple, un test peut être effectué sur chaque règle découverte, telle que r: x→ c, pour déterminer si la condition x est positivement corrélée à la classe c. Si le résultat du test est supérieur à une constante particulière, il existe une forte indication que x et c de r sont corrélés positivement et que, par conséquent, r sera stocké en tant que règle candidate dans le classificateur. Si le résultat du test indique une corrélation négative, r ne prendra aucune part à la prédiction ultérieure et sera écarté. L'algorithme CMAR [1] adopte le test du khi-deux dans son étape de découverte des règles. Lorsqu'une règle est trouvée, CMAR vérifie si son corps est positivement corrélé à la classe. Si une corrélation positive est trouvée, CMAR conserve la règle, sinon la règle est ignorée.

3.5.2.2 Précision de Laplace

La précision de Laplace [1] est une méthode post-élagage, qui est invoquée lors de la construction d'une règle. Pour estimer le taux d'erreur attendu d'une règle r: p1p2 ..... pn → c, la précision attendue pour une règle donnée r est calculé à l'aide de la formule suivante:

$$\text{Laplace}(r) = \frac{(p_c(r)+1)}{(p_{tot}(r)+m)}$$

0.2 Figure 3.12. Formule précision de Laplace

Ou :

p (r) indique le nombre de cas de formation couverts par r avec la classe c, ptot (r) est le nombre de cas de formation correspondant à la condition de r et m est le nombre de libellés de classe dans le domaine. La formule de Laplace a été adoptée en classification associative par le CPAR [1]. La méthode invoquée après la génération et le tri des règles, l'erreur attendue est calculé pour chaque règle de l'ensemble de règles potentielles. Si le résultat est supérieur à une valeur prédéfinie, la règle est ensuite supprimée, ce qui garantit que seules les règles ayant la meilleure précision attendue seront affichées.

3.5.2.3 I-Prune (Élément pruneau)

L'élément pruneau est une méthode récemment proposée dans [1]. I-Prune est une méthode de pré-élagage qui tend à marquer les articles sans intérêt en fonction de mesures d'intéressement (mesures de corrélation, telles que «Khi-deux», «Lift», «Odd ration»), de les supprimer et d'utiliser uniquement des

articles d'intérêt qui seraient utilisés dans la construction du modèle de classification. Par conséquent, cette étape d'élagage précoce réduira le nombre de règles générées, ainsi que le temps nécessaire à l'apprentissage du classificateur. Plusieurs algorithmes AC tels que CBA, CPAR, CMAR et MCAR considèrent un élément intéressant en fonction du nombre de supports. Alternativement, I-prune ne sélectionne que ceux qui sont fréquents et corrélés à une classe. Si un élément i est corrélé à la classe c, une mesure d'intérêt est donnée comme suit: si interestegness-mesureur (i, c)> seuil prédéfini, alors i est sélectionné sinon, l'élément est éliminé dès qu'il est détecté. Supposons que I soit un sous-ensemble d'éléments fréquents et corrélés par rapport à la classe c, seules les règles contenant des éléments intéressants sont générées. D'autre part, I-prune peut supprimer par inadvertance un élément susceptible de produire des règles de classification utiles dans les étapes ultérieures. Les résultats expérimentaux montrent que le khi-deux est la meilleure mesure de corrélation en ce qui concerne l'efficacité (voir [1] pour plus de détails).

3.6 Conclusion

Les règles d'association occupent une place très importante dans le domaine de l'analyse des données. On trouve plusieurs applications dans le marketing ou dans la recherche médicale. Les organisations travaillent à trouver les meilleurs placements de produits afin d'optimiser leurs ventes. Le corps médical pourrait

bénéficier d'un système de classification des symptômes et des maladies afin de mieux prédire la maladie d'un patient en fonction des symptômes de ce dernier.

Au-delà de ces exemples, les règles d'association peuvent être appliquées à divers secteurs d'activités pour lesquels il est intéressant de rechercher des corrélations potentielles entre des objets de diverses catégories. Les algorithmes tels que Apriori permettent d'extraire les règles d'association et des techniques d'élagage ont été proposées pour ne garder que les règles dites *intéressantes,* cependant le choix arbitraire des seuils avant l'étape d'élagage reste un défaut, car il devient difficile d'estimer la perte d'information. Dans le chapitre 4, nous présentons notre implémentation dans laquelle nous proposons des pistes de solution pour évaluer la perte d'information tout en intégrant aussi les notions introduites dans ce chapitre.

CHAPITRE 4 - Implémentation

4.1 Introduction

Dans ce chapitre nous mettons en emphase les moyens et la méthodologie utilisés pour la mise en œuvre de ce projet de recherche. Nous aborderons entre autres les outils, les technologies et les langages de programmation qui ont servi au développement du logiciel.

Pour une meilleure illustration, nous présenterons les principales interfaces de notre logiciel, tout en expliquant le fonctionnement et l'importance de chaque étape. Les possibilités de paramétrages du logiciel par l'utilisateur seront aussi présentées.

Les règles d'association étant au cœur de notre travail de recherche, nous présenterons les processus d'extraction des règles, la visualisation qui suit et une estimation de la perte d'information.

Notre système apporte un moyen efficace de parcourir et visualiser les règles d'association générées tout en optimisant le processus d'extraction et d'élagage afin de réduire le temps de calculs et de ressources mémoires.

4.2 Environnement logiciel et matériel de développement

4.2.1 Langage de programmation

R est un langage de programmation largement utilisé dans le domaine des statistiques et de la science des données. Il a été principalement écrit en C et est disponible sur tous les systèmes d'exploitation. C'est un langage très approprié pour l'analyse et l'exploration de données. Son atout majeur est le nombre pléthorique de bibliothèques qu'il embarque. Un autre de ses atouts et pas des moindres est sa forte communauté qui contribue à son développement. Pour toutes ces qualités, notre choix s'est tourné vers ce langage ayant déjà fait ses preuves dans le domaine de l'exploration de données.

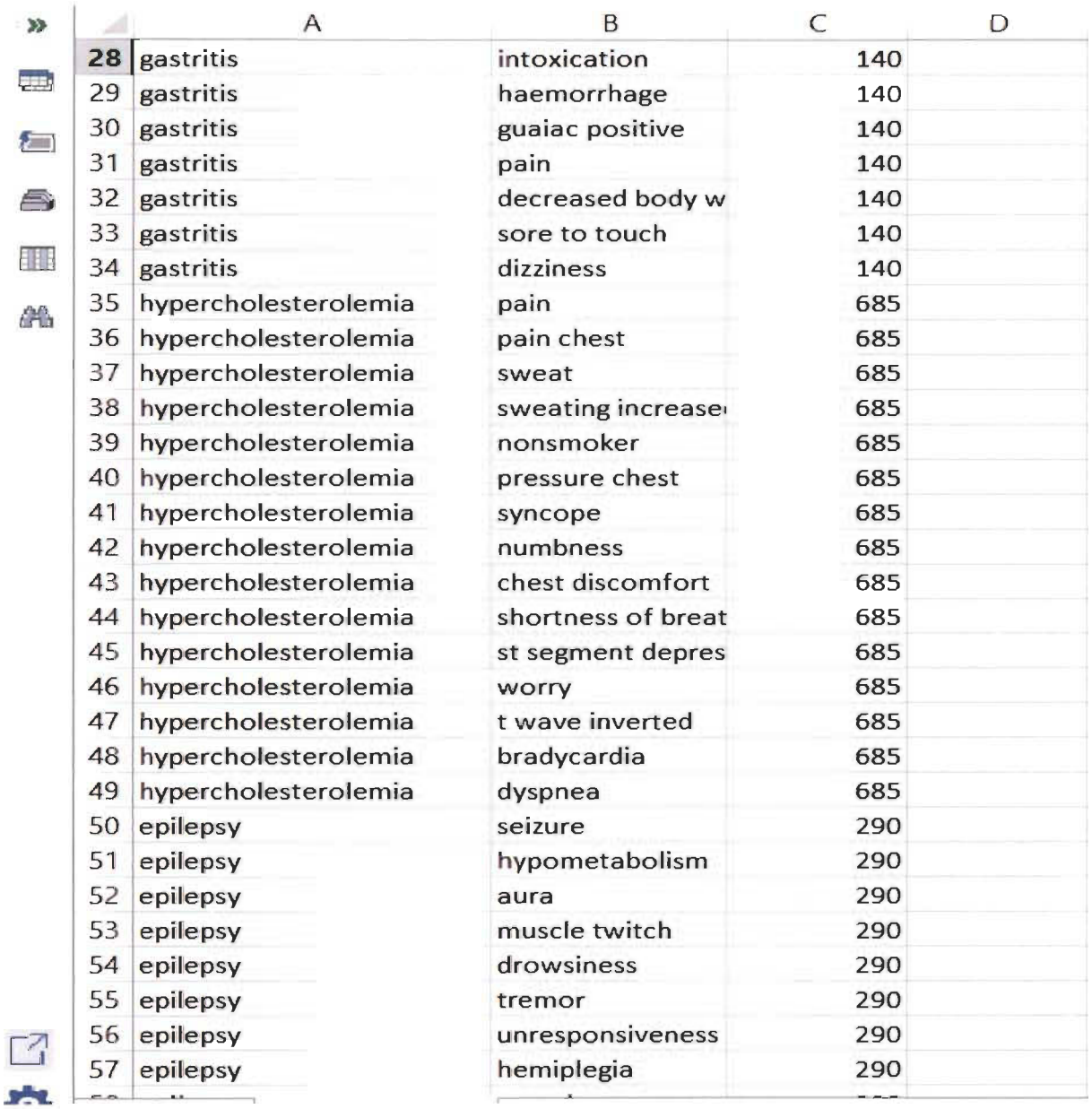
4.2.2 Environnement de développement

Rstudio est l'environnement de développement intégré (IDE) choisi pour notre travail de recherche. C'est un IDE qui intègre très bien les fondements du langage R, permettant ainsi de développer de solides applications tout en profitant des avantages visuels et des fonctionnalités d'un IDE. Nous avons aussi utilisé le framework Shiny permettant de réaliser des applications interactives depuis R.

4.3 Les données

L'ensemble de données utilisé dans ce travail de recherche contient des résumés textuels de sortie des patients du New York Presbyterian Hospital admis en 2004. La première colonne montre la maladie, la seconde les symptômes associés et la troisième montre un classement basé sur la force de l'association. La méthode a utilisé le système de traitement du langage naturel MedLEE (Medical language extraction and encoding system) pour obtenir des codes UMLS pour les maladies et les symptômes à partir des notes; puis des méthodes statistiques basées sur les fréquences et les cooccurrences ont été utilisées pour obtenir les relations. Une description plus détaillée de la méthode automatisée peut être trouvée dans [5]. Ensuite, nous traitons les données en éliminant la troisième colonne afin d'obtenir un format transactionnel pour l'extraction des

symptômes uniquement. À cette étape, notre ensemble de données contient 148 maladies et 805 symptômes (voir figure 4.1).



	A	B	C	D
28	gastritis	intoxication	140	
29	gastritis	haemorrhage	140	
30	gastritis	guaiac positive	140	
31	gastritis	pain	140	
32	gastritis	decreased body w	140	
33	gastritis	sore to touch	140	
34	gastritis	dizziness	140	
35	hypercholesterolemia	pain	685	
36	hypercholesterolemia	pain chest	685	
37	hypercholesterolemia	sweat	685	
38	hypercholesterolemia	sweating increase	685	
39	hypercholesterolemia	nonsmoker	685	
40	hypercholesterolemia	pressure chest	685	
41	hypercholesterolemia	syncope	685	
42	hypercholesterolemia	numbness	685	
43	hypercholesterolemia	chest discomfort	685	
44	hypercholesterolemia	shortness of breat	685	
45	hypercholesterolemia	st segment depres	685	
46	hypercholesterolemia	worry	685	
47	hypercholesterolemia	t wave inverted	685	
48	hypercholesterolemia	bradycardia	685	
49	hypercholesterolemia	dyspnea	685	
50	epilepsy	seizure	290	
51	epilepsy	hypometabolism	290	
52	epilepsy	aura	290	
53	epilepsy	muscle twitch	290	
54	epilepsy	drowsiness	290	
55	epilepsy	tremor	290	
56	epilepsy	unresponsiveness	290	
57	epilepsy	hemiplegia	290	

Figure 4.1. Exemple de données transactionnelles obtenues

Les données transactionnelles sont ensuite formatées comme un itemMatrix dans un format fragmenté avec 149 lignes (éléments / itemsets / transactions) et 548

colonnes (items) et une densité de 0,0269926. Le tableau ci-après (voir tableau 4.1) montre de statistiques importantes pour comprendre la distribution des données.

Tableau 4.1. Statistiques sur la distribution des données

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
1.00	12.00	14.00	14.79	17.00	29.00

4.4 Architecture du système développé

Le système développé a une architecture modulaire composée de six modules principaux à savoir :

Le module de récupération de données textuelles

Le module de traitement et de nettoyage du texte

Le module de génération des itemsets fréquents

Le module d'extraction des règles d'association

Le module d'élagage de règles d'association

Le module de génération de règles intéressantes

La figure ci-après représente (voir figure 4.1) l'architecture de notre système.

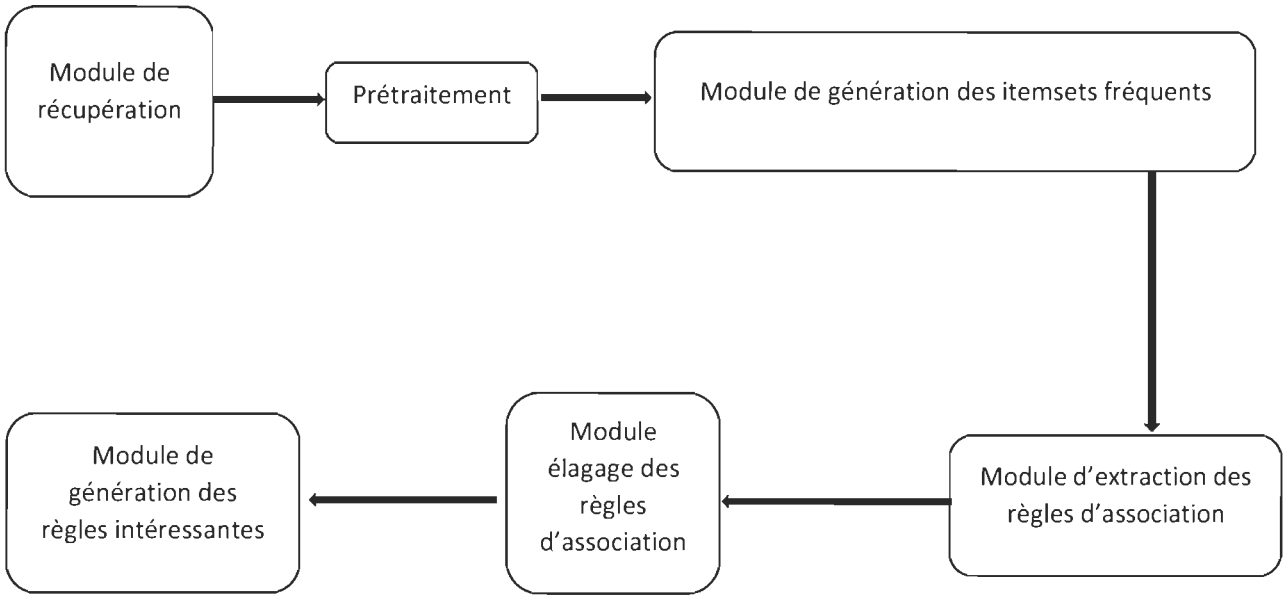


Figure 4.2. Architecture modulaire du système développé

4.5 Fonctionnement du système développé

Le système développé propose des fonctionnalités d'extraction et de visualisation des règles d'association, à partir des informations textuelles que fournira l'utilisateur.

Récupération des données médicales

La première étape de notre implémentation a été d'offrir une interface pour permettre à l'utilisateur de téléverser dans le logiciel les données qu'il souhaiterait traiter. L'utilisateur recherche le fichier (.csv) dans son ordinateur puis il peut paramétrer le logiciel sur le format du fichier à traiter ou l'information

qu'il veut voir s'afficher. En tâche de fond, le logiciel récupère les données de l'utilisateur qu'il formate en données transactionnelles. Voir figure ci-après : 3

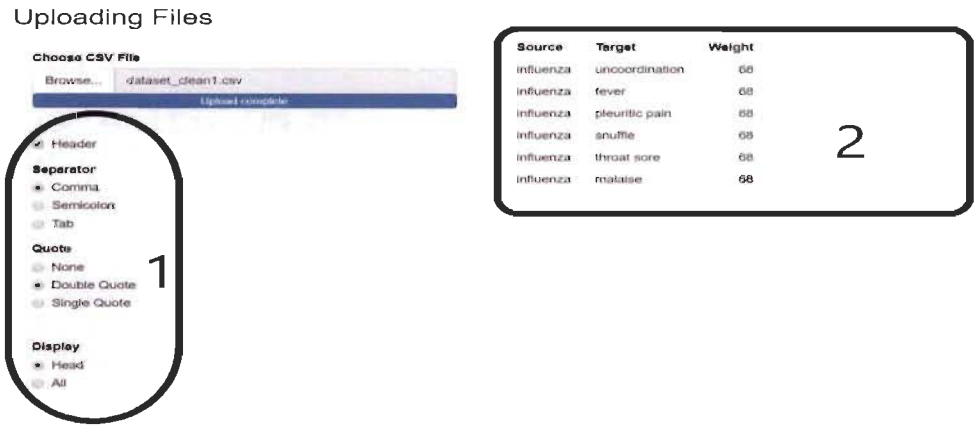


Figure 4.3. Récupération des données

Dans la section 1 se trouvent tous les paramètres que l'utilisateur peut utiliser pour renseigner le format de son fichier ou les informations qu'il veut voir s'afficher dans la section 2.

Prétraitement des données

Le prétraitement des données consiste au traitement et au nettoyage des données fournies par l'utilisateur. Il s'agira entre autres d'éliminer les mots vides qui sont considérés comme de l'information non pertinente. Les mots vides sont par exemples <<le>>, <<la>>, <<de>>, etc. Ensuite, on procède à l'élimination des ponctuations des chiffres et des caractères spéciaux qui sont aussi non pertinents. Enfin on écrit tout le texte épuré dans un fichier sous forme de données

transactionnelles qu'on réutilisera plus tard dans l'étape d'extraction des règles d'association.

Génération des règles

À cette étape du logiciel, l'utilisateur peut déjà observer les itemsets les plus fréquents. Il peut aussi avoir un aperçu résumant les propriétés de ses données. Ensuite il sera invité à paramétrer le support et la confidence avant de pouvoir générer les règles d'association.

Le bouton ***Generate*** lui permet de générer les règles d'association en fonction des valeurs du support et de la confiance définis au préalable. Après avoir généré les règles il sera à même de les visualiser en forme de graphe dans section ***visualization***. Dans la section ***Rules*** il pourra parcourir les règles dans un tableau avec la possibilité de paginer ou de trier les règles. Enfin la section ***Loss*** lui permettra de visualiser la perte d'information après l'élagage des règles.

L'image ci-après illustre les différentes sections (voir figure 4.3) :

Figure 4.4. Paramétrage et génération des règles

Élagage des règles

Dans ce travail de recherche, pour élaguer les règles significatives, nous avons utilisé une technique simple adaptée à notre contexte. Nous veillons à éliminer les règles redondantes. Cela recherche essentiellement toutes les règles plus spécifiques (même conséquence / RHS, mais plus d'éléments dans la LHS / antécédent) qui ont un lift, une confiance ou une autre métrique égale ou inférieure. Plus formellement, vous recherchez un sous-ensemble X 'de X qui présente une amélioration (ou du moins aucune diminution) en termes de lift ou de confiance (Voir figure 4.4).

$\exists X' \subset X \; \textit{conf}(X' \rightarrow Y) \geq \textit{conf}(X \rightarrow Y)$

Figure 4.5. Élagage des règles

En guise d'exemple, supposons un ensemble de règles comme suit :

	lhs		rhs	support	confiance	lift
1	{16058}	→	{16059}	0.01218522	**0.9375000**	67.886029
2	{16059}	→	{16058}	0.01218522	0.8823529	67.886029
3	{10049}	→	{10021}	0.01462226	**0.7826087**	34.406832
4	{10021}	→	{10049}	0.01462226	0.6428571	34.406832

lhs représente l'antécédent et rhs la conséquence. En considérant la confiance comme métrique d'amélioration, notre algorithme procédera à l'élimination des règles 2 et 4 et il ne restera que les règles 1 et 3. En effet, les couples de règles (1,2) et (3,4) sont dites acycliques, ainsi pour éviter une règle «circulaire» qui relie deux items avec deux règles, notre méthode d'élagage procède à l'élimination des règles ayant le plus faible indice de confiance.

Pour notre travail nous avons utilisé le lift comme mesure d'amélioration, avec cette approche le Y conséquent n'est pas le même.

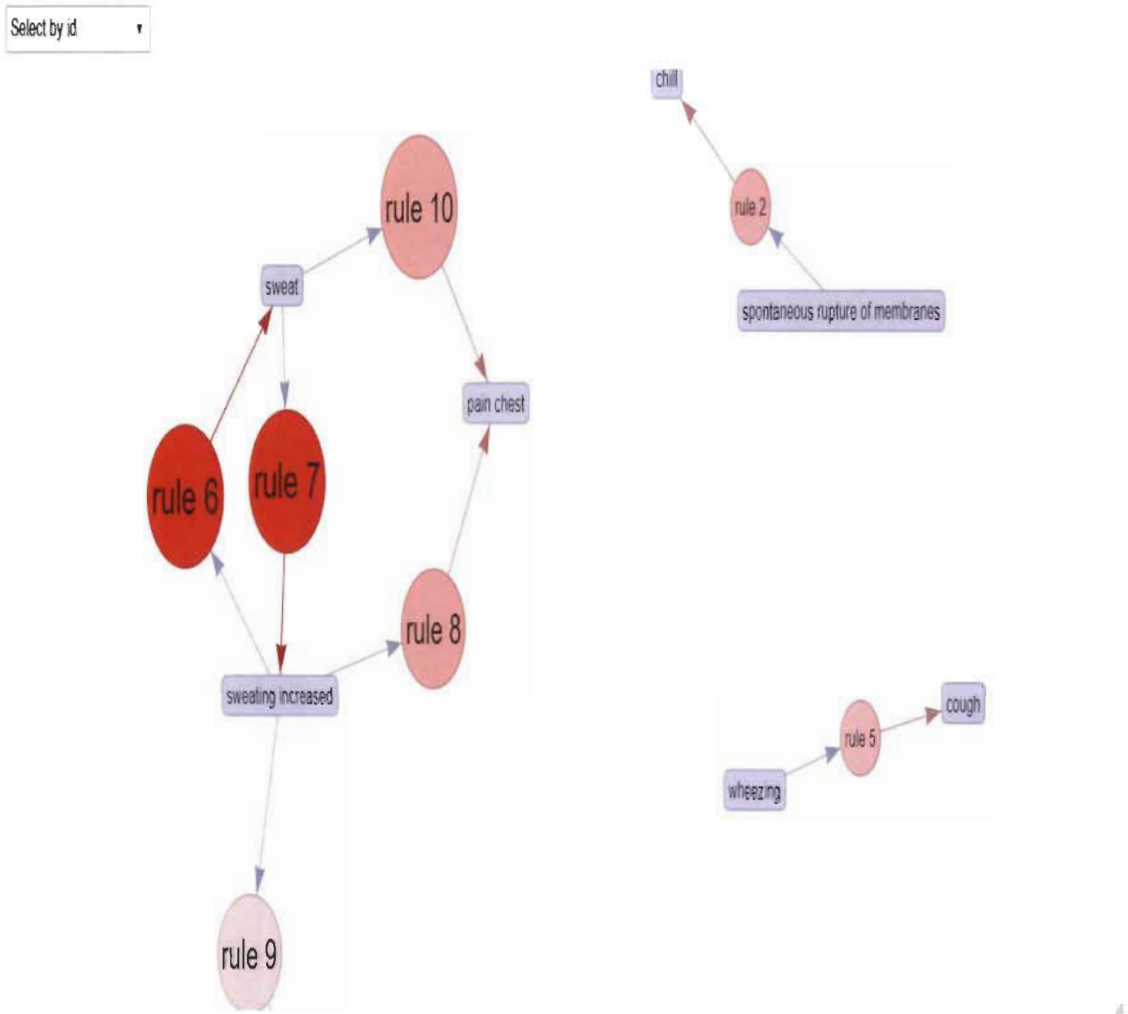


Figure 4.6. Dix premières règles avant l'élagage

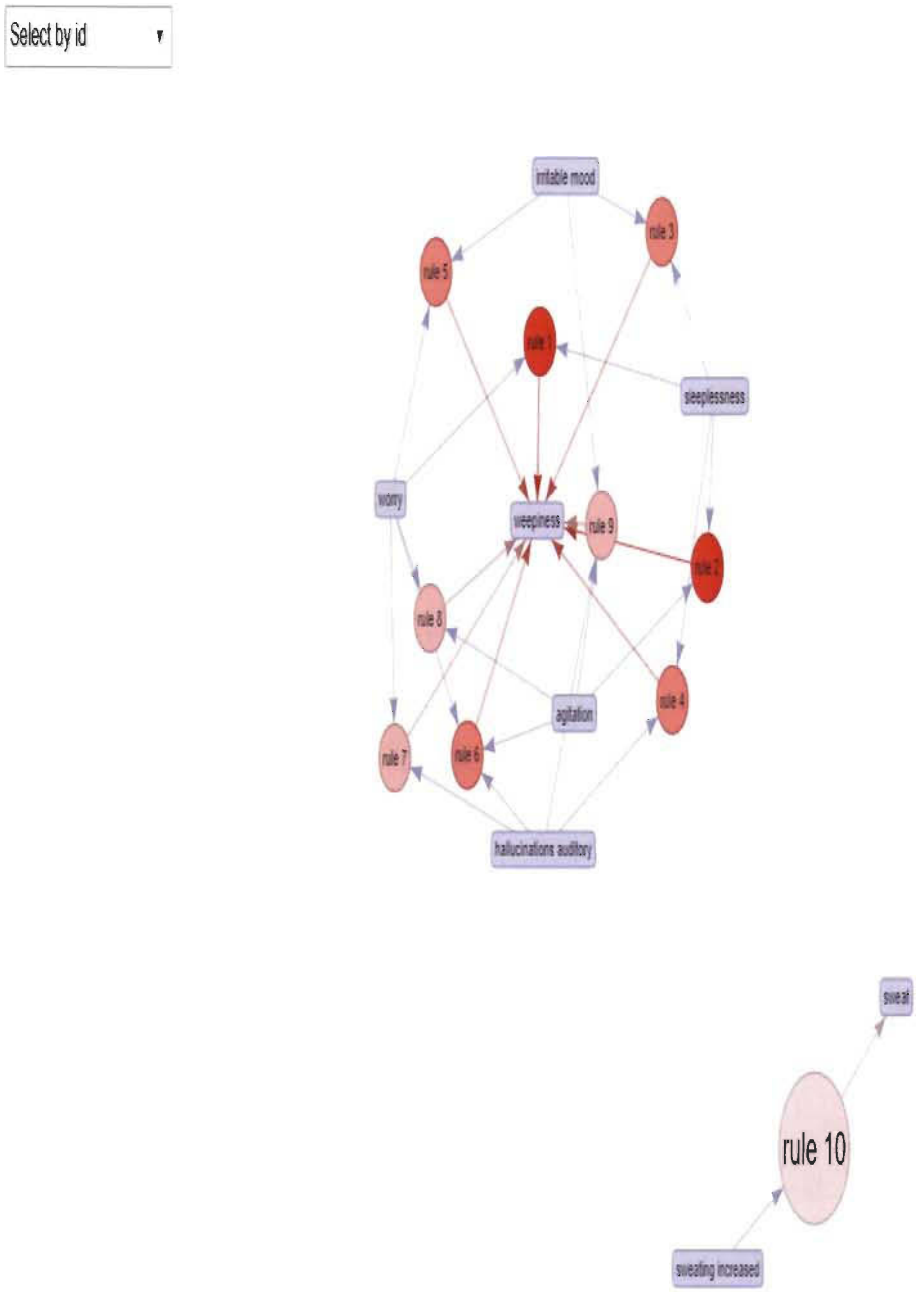


Figure 4.7. Dix premières règles après élagage

Sur la figure 4.5, nous remarquons ce que nous considérons comme des règles redondantes (règle 6 et règle 7). Après l'élagage (Fig 4.6), la règle 7 a été supprimée, car son lift est inférieur à celui de la règle 6 (désormais règle 10).

4.6 Modèle d'implémentation

Notre modèle d'implémentation se résume en quelques étapes qu'illustre la figure 4.7 ci-après :

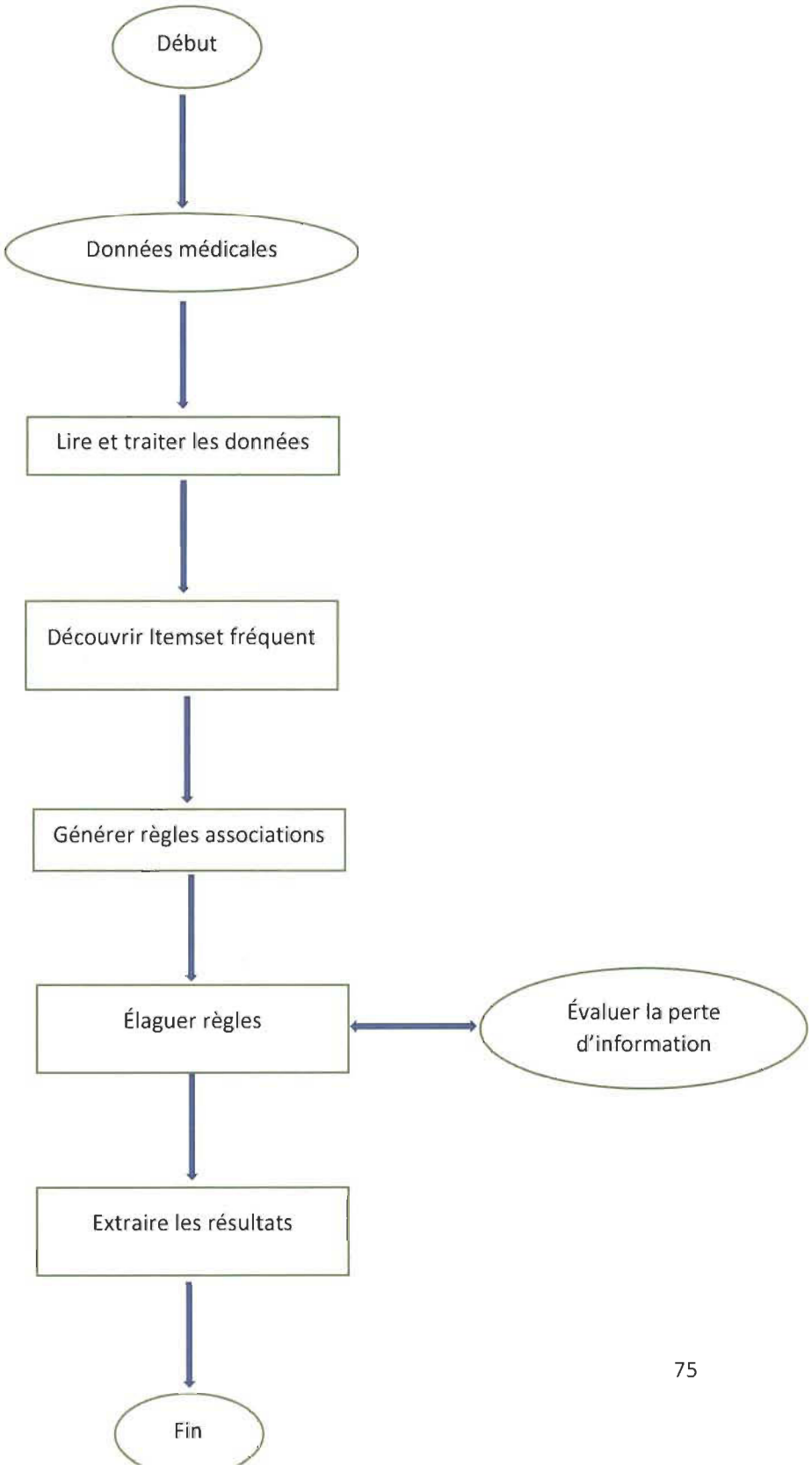


Figure 4.8. Modèle d'implémentation

4.7 Conclusion

Dans ce chapitre nous avons exposé le processus de traitement, l'implémentation, les différentes fonctionnalités et les paramètres de notre système.

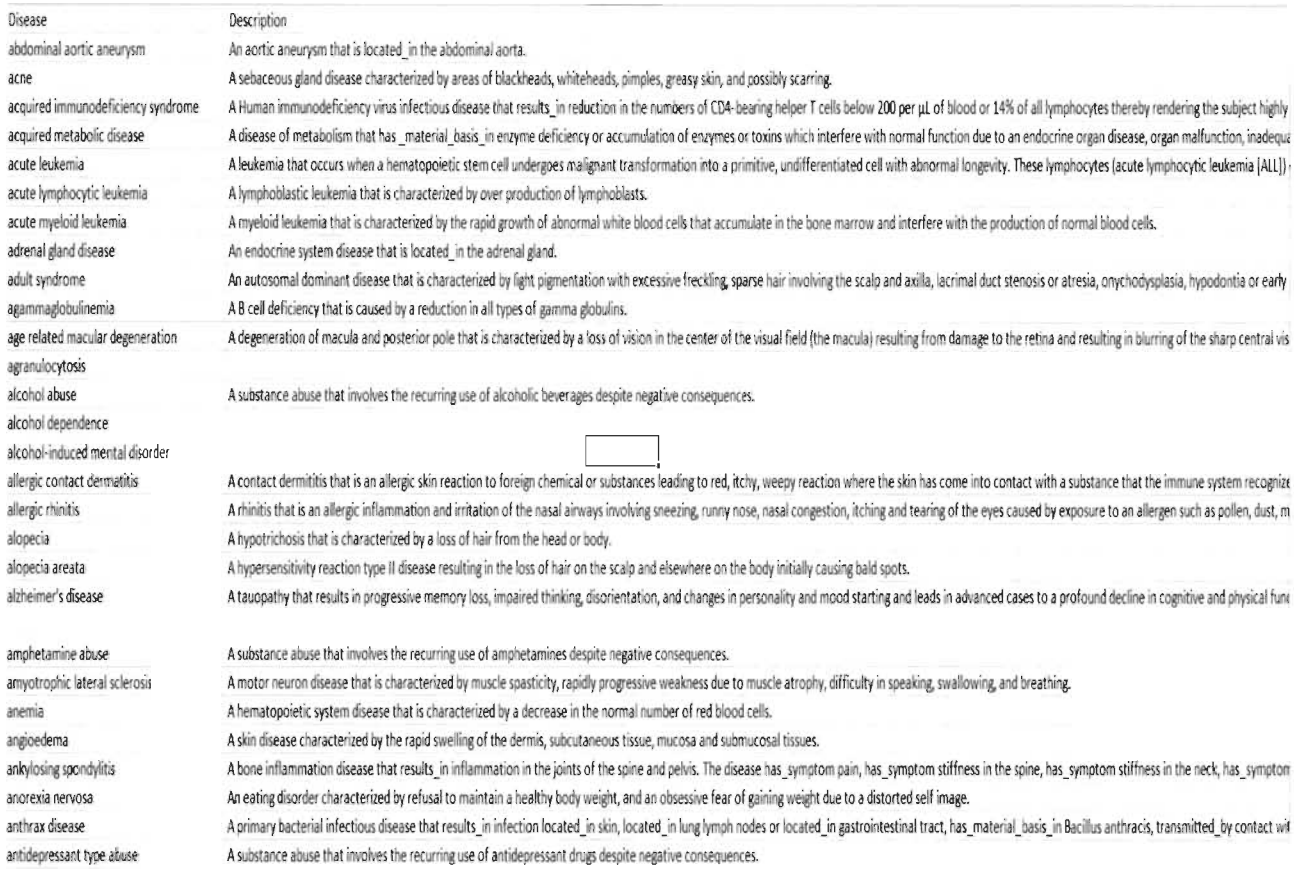
Dans le prochain chapitre, nous expérimentons notre système sur des données réelles en menant des interprétations et des discussions sur les résultats obtenus.

CHAPITRE 5 - Expérimentations et discussions

5.1 Introduction

Dans ce chapitre nous présentons les résultats des expérimentations menées sur notre système. Nous procédons à la mise en place de tests en utilisant des données médicales et en faisant varier les valeurs des indices de support et de confiance afin d'observer les différentes sorties.

Nos expérimentations se sont effectuées sur une source de données textuelles réelle (voir figure 5.1). La source des données et le traitement des données ont été énumérés dans le chapitre précédent (voir chapitre 4 section 4.3).



Disease	Description
abdominal aortic aneurysm	An aortic aneurysm that is located_in the abdominal aorta.
acne	A sebaceous gland disease characterized by areas of blackheads, whiteheads, pimples, greasy skin, and possibly scarring.
acquired immunodeficiency syndrome	A Human immunodeficiency virus infectious disease that results_in reduction in the numbers of CD4-bearing helper T cells below 200 per µL of blood or 14% of all lymphocytes thereby rendering the subject highly
acquired metabolic disease	A disease of metabolism that has_material_basis_in enzyme deficiency or accumulation of enzymes or toxins which interfere with normal function due to an endocrine organ disease, organ malfunction, inadequa
acute leukemia	A leukemia that occurs when a hematopoietic stem cell undergoes malignant transformation into a primitive, undifferentiated cell with abnormal longevity. These lymphocytes (acute lymphocytic leukemia [ALL])
acute lymphocytic leukemia	A lymphoblastic leukemia that is characterized by over production of lymphoblasts.
acute myeloid leukemia	A myeloid leukemia that is characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells.
adrenal gland disease	An endocrine system disease that is located_in the adrenal gland.
adult syndrome	An autosomal dominant disease that is characterized by light pigmentation with excessive freckling, sparse hair involving the scalp and axilla, lacrimal duct stenosis or atresia, onychodysplasia, hypodontia or early
agammaglobulinemia	A B cell deficiency that is caused by a reduction in all types of gamma globulins.
age related macular degeneration	A degeneration of macula and posterior pole that is characterized by a loss of vision in the center of the visual field (the macula) resulting from damage to the retina and resulting in blurring of the sharp central vis
agranulocytosis	
alcohol abuse	A substance abuse that involves the recurring use of alcoholic beverages despite negative consequences.
alcohol dependence	
alcohol-induced mental disorder	
allergic contact dermatitis	A contact dermitiis that is an allergic skin reaction to foreign chemical or substances leading to red, itchy, weepy reaction where the skin has come into contact with a substance that the immune system recognize
allergic rhinitis	A rhinitis that is an allergic inflammation and irritation of the nasal airways involving sneezing, runny nose, nasal congestion, itching and tearing of the eyes caused by exposure to an allergen such as pollen, dust, m
alopecia	A hypotrichosis that is characterized by a loss of hair from the head or body.
alopecia areata	A hypersensitivity reaction type II disease resulting in the loss of hair on the scalp and elsewhere on the body initially causing bald spots.
alzheimer's disease	A tauopathy that results in progressive memory loss, impaired thinking, disorientation, and changes in personality and mood starting and leads in advanced cases to a profound decline in cognitive and physical fun
amphetamine abuse	A substance abuse that involves the recurring use of amphetamines despite negative consequences.
amyotrophic lateral sclerosis	A motor neuron disease that is characterized by muscle spasticity, rapidly progressive weakness due to muscle atrophy, difficulty in speaking, swallowing, and breathing.
anemia	A hematopoietic system disease that is characterized by a decrease in the normal number of red blood cells.
angioedema	A skin disease characterized by the rapid swelling of the dermis, subcutaneous tissue, mucosa and submucosal tissues.
ankylosing spondylitis	A bone inflammation disease that results_in inflammation in the joints of the spine and pelvis. The disease has_symptom pain, has_symptom stiffness in the spine, has_symptom stiffness in the neck, has_symptom
anorexia nervosa	An eating disorder characterized by refusal to maintain a healthy body weight, and an obsessive fear of gaining weight due to a distorted self image.
anthrax disease	A primary bacterial infectious disease that results_in infection located_in skin, located_in lung lymph nodes or located_in gastrointestinal tract, has_material_basis_in Bacillus anthracis, transmitted_by contact wit
antidepressant type abuse	A substance abuse that involves the recurring use of antidepressant drugs despite negative consequences.

Figure 5.1. Texte d'origine pour l'extraction des règles.

Dans nos différentes expérimentations, nous récupérons les données qui sont constituées des maladies et de la description des symptômes, puis nous utilisons diverses techniques de traitement de langue afin de préparer les données pour l'extraction des règles d'association. Nous présenterons aussi la partie interprétation et discussion des résultats.

5.2 Résultats des expérimentations

L'expérimentation effectuée se déroule sur trois phases. À chaque phase nous changeons les valeurs des indices de support et de confiance, afin d'observer

les différences dans les résultats. Nous avons aussi divisé notre ensemble de données en deux. Une partie servira pour les besoins de l'expérimentation 1 et une autre pour les deux expérimentations restantes.

À chaque phase, on effectuera des traitements sur les données tels que :

- Éliminer les caractères spéciaux,
- Supprimer les espaces blancs du texte,
- Éliminer les mots vides,
- Réorganiser les données dans le format transactionnel

Nous discuterons les résultats des expérimentations en fonction de la qualité des règles générées, à quel point elles représentent au mieux la réalité et de leurs représentations visuelles. Nous observerons aussi la perte d'information à chaque phase.

Tous les cas expérimentaux sont mis en œuvre dans R en congestion avec les outils, les algorithmes et les stratégies de RStudio, ainsi que diverses techniques d'extraction de fonctionnalités, et s'exécutent dans un environnement avec un système ayant la configuration de la machine Intel Core i5, 2,30 GHz Windows 10 (64 bits) avec 8 Go de RAM.

5.2.1 Expérimentation 1

Dans cette expérimentation nous avons défini le support à 0.009 et la confiance à 0.8. Rappelons que le choix des valeurs des métriques est subjectif. Les règles sont sur la forme << X → Y >>:

La figure 5.2 ci-après fournit un résumé des données :

```
transactions as itemMatrix in sparse format with
 593 rows (elements/itemsets/transactions) and
 2030 columns (items) and a density of 0.0005000042

most frequent items:
     disease     locatedin characterized       system     diseasea       (Other)
         261           130           120           100           85          5323

element (itemset/transaction) length distribution:
sizes
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
 7 25 56 47 20 19 34 51 37 44 33 26 24 20 22 15 15 17 13 11  7  9  3  4  3  2  1  2
28 29 30 31 33 34 36 37 38 39 40 42 43 45 46 53 58
 1  5  2  2  1  2  1  2  1  1  1  1  1  2  1  1  1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    4.00    9.00   10.15   14.00   58.00

includes extended item information - examples:
     labels
1        al
2         a
3 abdominal
```

0.1 Figure 5.2 Résumé des données téléversées

On peut noter le nombre de lignes (593) et le nombre de colonnes (2030), les items les plus fréquents, et des statistiques sur les données. La figure ci-après (voir figure 5.3) présente une visualisation des cinq items les plus fréquents.

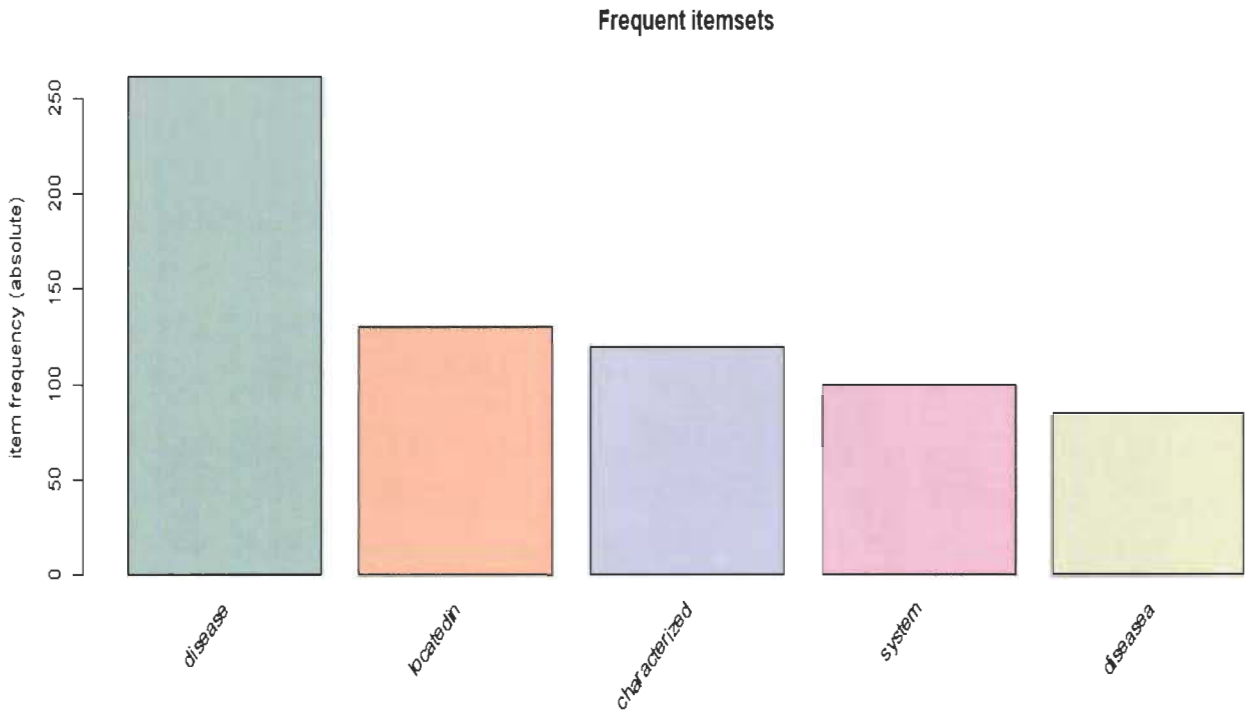


Figure 5.3 Les cinq items les plus fréquents expérimentation 1

On note que l'élément ***disease*** est le plus fréquent dans notre jeu de données. En exécutant l'algorithme d'extraction de règles on obtient 2037 règles extraites en 0.07s. En affichant les 10 premières règles en forme de graphe, on obtient le résultat suivant :

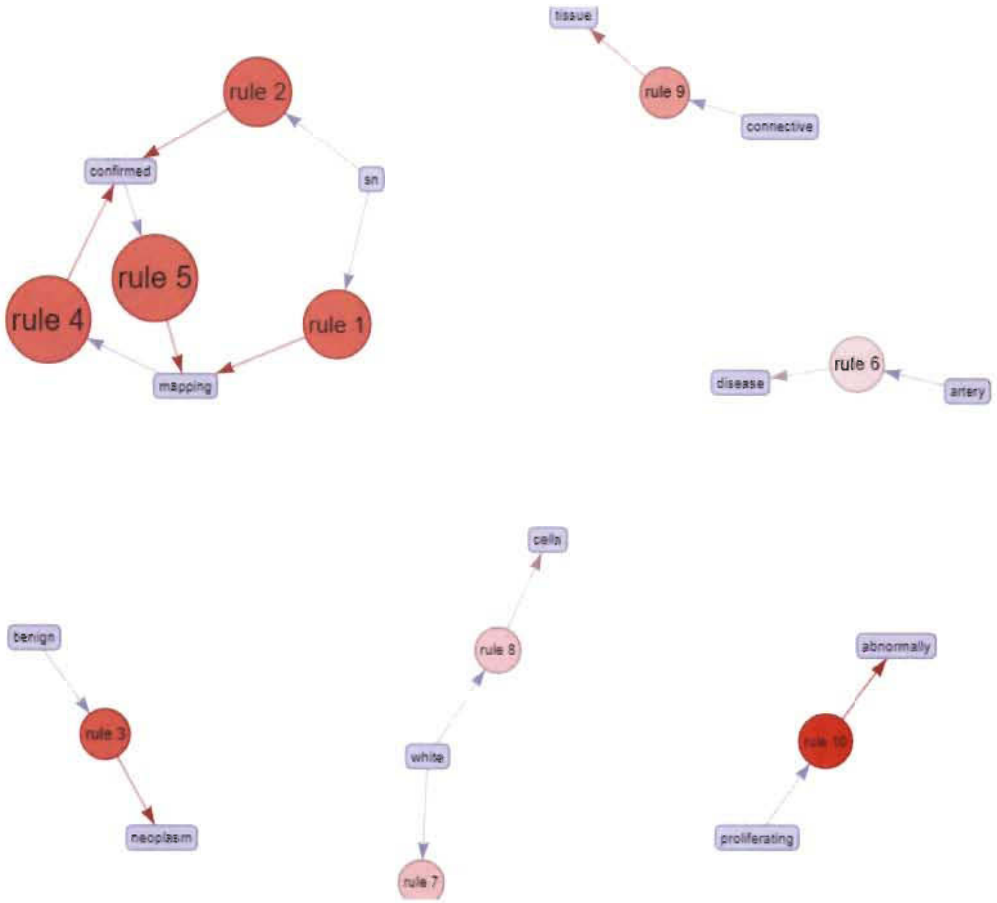


Figure 5.4. Graphe des 10 règles expérimentation 1

Les règles 1, 2, 4, 5 ne contiennent pas vraiment d'information cohérente à cause de l'item ***sn*** que nous soupçonnons d'être une abréviation. Par contre la règle 3 benign → neoplasm indique que la plupart des tumeurs cancéreuses sont de stade bénigne. Nous avons ensuite procédé à l'élagage des règles obtenues afin de ne garder que les règles intéressantes :

	lhs	rhs	support	confidence	coverage	lift
1	{abnormally,hasmaterialbasisin}	{proliferating}	0.0101180438448567	1	0.0101180438448567	98.8333333333333
3	{leukemia}	{leukemiaa}	0.0118043844856661	1	0.0118043844856661	74.125
5	{cancer,cellular}	{uncontrolled}	0.0101180438448567	1	0.0101180438448567	74.125
6	{cellular,locatedin}	{uncontrolled}	0.0101180438448567	1	0.0101180438448567	74.125
7	{locatedin,proliferation}	{uncontrolled}	0.0101180438448567	1	0.0101180438448567	74.125
8	{cellular,characterized,proliferation}	{uncontrolled}	0.0118043844856661	1	0.0118043844856661	74.125
9	{cancer,characterized,locatedin}	{uncontrolled}	0.0101180438448567	1	0.0101180438448567	74.125
12	{proliferating}	{abnormally}	0.0101180438448567	1	0.0101180438448567	59.3
13	{cellular,uncontrolled}	{proliferation}	0.0118043844856661	1	0.0118043844856661	59.3
14	{proliferation,uncontrolled}	{cellular}	0.0118043844856661	1	0.0118043844856661	59.3

Figure 5.5. Résultat des règles élaguées expérimentation 1

Les résultats de la figure 5.5 démontrent une forte relation entre ***{abnormally,hasmaterialbasisin}*** et ***{proliferating}*** avec un confiance de 100%. On retrouve aussi une relation forte entre ***{cancer,cellular}*** et ***{uncontrolled}*** avec une confiance de 100%.

Le professionnel analysant ces informations pourra déduire qu'une version anormale de ***hasmaterialbasisin (règle 1)*** qui est une maladie génétique prolifère de façon non contrôlée dans les cellules ***(règle 14).***

Dans cette expérience on décèle aussi du bruit et de l'information incohérente avec par exemple la règle {cellular, locatedin} → {uncontrolled}.

En fin d'expérimentation nous avons visualisé la perte d'information avant et après le processus d'élagage des règles et cela pour chaque métrique. Cela permet à l'utilisateur d'estimer la perte en information afin de redéfinir les métriques au besoin.

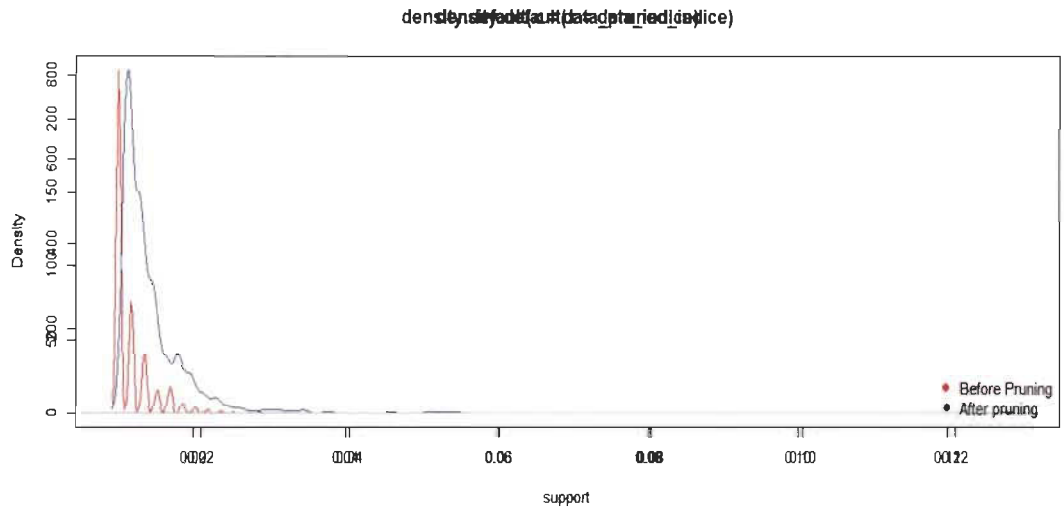


Figure 5.6. Tracé de la densité de l'indice du support expérimentation 1

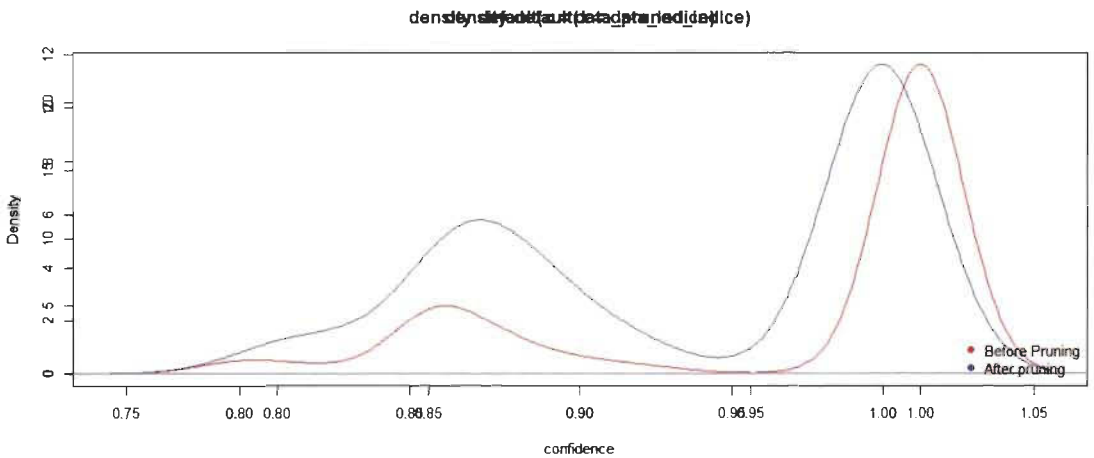


Figure 5.7. Tracé de la densité de l'indice de la confiance expérimentation 1

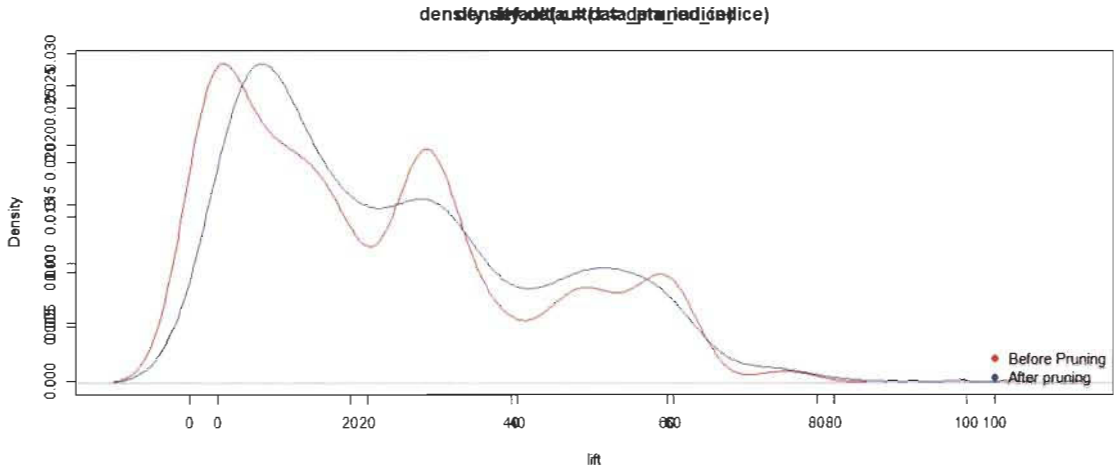


Figure 5.8. Tracé de la densité de l'indice du lift expérimentation 1

5.2.2 Expérimentation 2

Dans cette expérimentation nous avons utilisé la seconde partie de notre ensemble de données. Afin d'observer les changements au niveau des différentes expérimentations nous avons aussi changé les valeurs des métriques, de ce fait le support a été défini à 0.05 et la confidence à 0.8. Ci-après un résume des données téléversées dans notre système (voir figure 5.9) :

```
transactions as itemMatrix in sparse format with
 149 rows (elements/itemsets/transactions) and
 548 columns (items) and a density of 0.0269926

most frequent items:
shortness of breath                pain              fever            diarrhea
                 49                  43                 37                  29
     pain abdominal             (Other)
                 29                2017

element (itemset/transaction) length distribution:
sizes
 1  2  3  4  5  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 29
 1  1  1  1  1  2  2  4  8 11 11 18 14 12 18 11  7  3  6  3  4  2  4  1  1  2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   12.00   14.00   14.79   17.00   29.00

includes extended item information - examples:
                labels
1        abdomen acute
2   abdominal bloating
3 abdominal tenderness
```

Figure 5.9 Résumé des données expérimentation 2

On constate que le nombre de lignes est de 149 et le nombre de colonnes est de 548. La figure ci-après (voir figure 5.10) présente une visualisation des cinq items les plus fréquents.

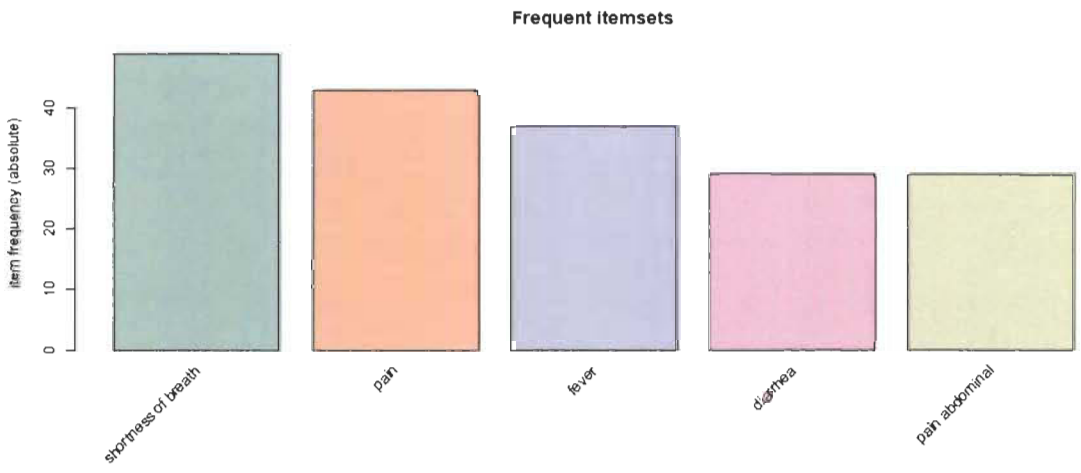


Figure 5.10. Les cinq items les plus fréquents expérimentation 2

On note que l'élément ***shortness of breath*** est le plus fréquent dans notre jeu de données. En exécutant l'algorithme d'extraction de règles, on obtient 313 règles extraites en 0.01s. En affichant les 10 premières règles en forme de graphe, on obtient le résultat suivant :

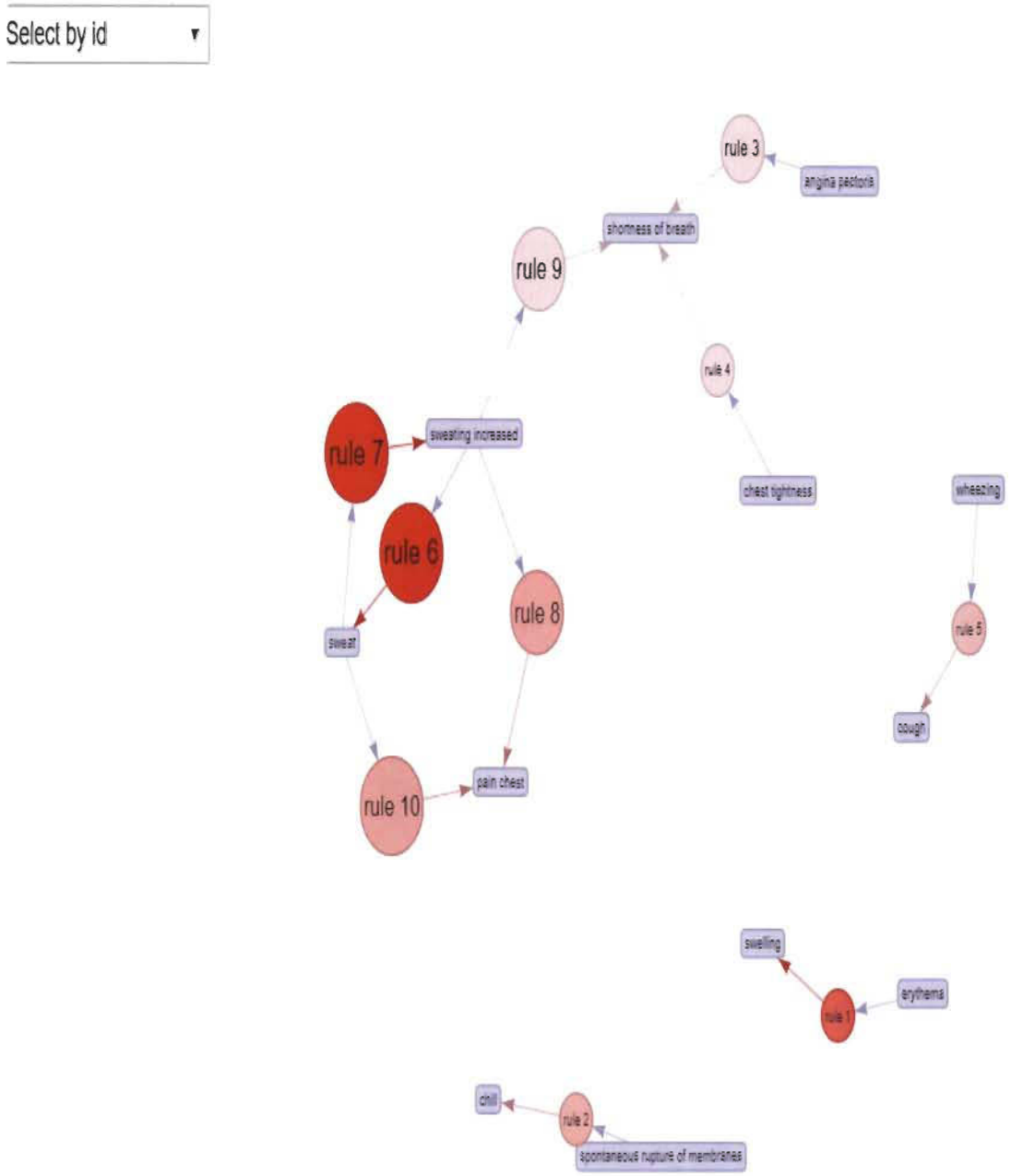


Figure 5.11. Graphe des 10 règles expérimentation 2

Les règles 6,7 représentent ce qu'on appelle des règles cycliques. L'une d'entre elles sera supprimée lors du processus d'élagage. La règle 5 {wheezing} →

{cough} établit la relation entre la toux et la respiration par sifflement, type de respiration des personnes atteintes d'asthme. Toutes les règles ci-dessus présentent une confiance de 100% et peuvent s'avérer très utiles au cours d'une analyse menée par un professionnel. Les 313 règles obtenues ont ensuite été élaguées, ce qui nous a laissés au total avec 139 règles. Le résultat est démontré dans la figure ci-après :

	lhs	rhs	support	confidence	coverage	lift
11	{sweat}	{sweating increased}	0.0738255033557047	1	0.0738255033557047	13.5454545454545
12	{weepiness}	{sleeplessness}	0.0536912751677852	1	0.0536912751677852	13.5454545454545
13	{feeling hopeless,irritable mood}	{sleeplessness}	0.0536912751677852	1	0.0536912751677852	13.5454545454545
14	{weepiness}	{worry}	0.0536912751677852	1	0.0536912751677852	12.4166666666667
15	{agitation,sleeplessness}	{worry}	0.0536912751677852	1	0.0536912751677852	12.4166666666667
16	{feeling hopeless,hallucinations auditory}	{sleeplessness}	0.0536912751677852	0.888888888888889	0.0604026845637584	12.040404040404
17	{irritable mood,worry}	{sleeplessness}	0.0536912751677852	0.888888888888889	0.0604026845637584	12.040404040404
18	{agitation,hallucinations auditory,worry}	{sleeplessness}	0.0536912751677852	0.888888888888889	0.0604026845637584	12.040404040404
19	{weepiness}	{irritable mood}	0.0536912751677852	1	0.0536912751677852	11.4615384615385
20	{hallucinations auditory,homelessness}	{irritable mood}	0.0536912751677852	1	0.0536912751677852	11.4615384615385

0.2 Figure 5.12. Résultats des règles élaguées expérimentation 2

La figure ci-dessus montre que la règle 19 {weepiness} → {irritable mood} indique que la mauvaise humeur est fortement liée aux pleurs. Une autre règle

avec une confiance de 100% est la règle 20 qui décrit la relation entre la mauvaise humeur et le couple hallucination auditoire, sans-abri. Ces données pourraient être interprétées pour des patients souffrant de troubles psychologiques. Ces personnes pourraient être des sans-abri présentant différents symptômes psychologiques.

La suite de l'expérimentation a consisté visualiser la perte d'information :

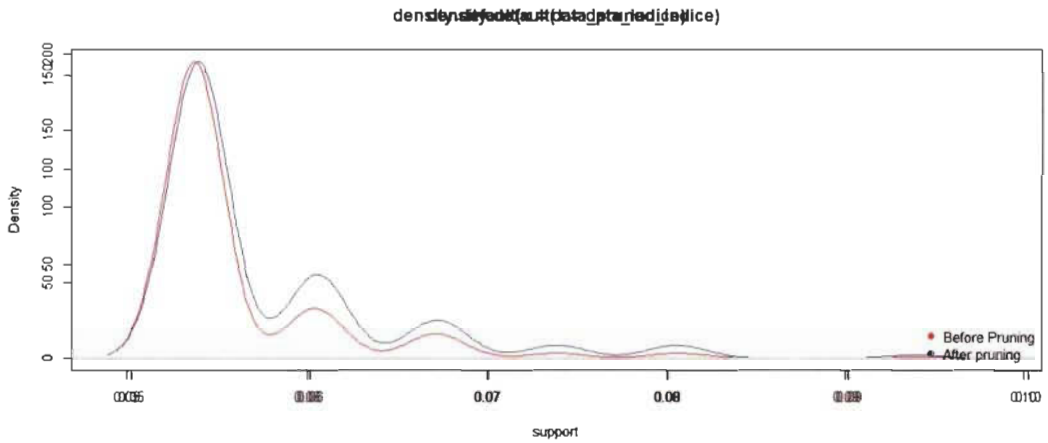


Figure 5.13. Tracé de la densité de l'indice de support expérimentation 2

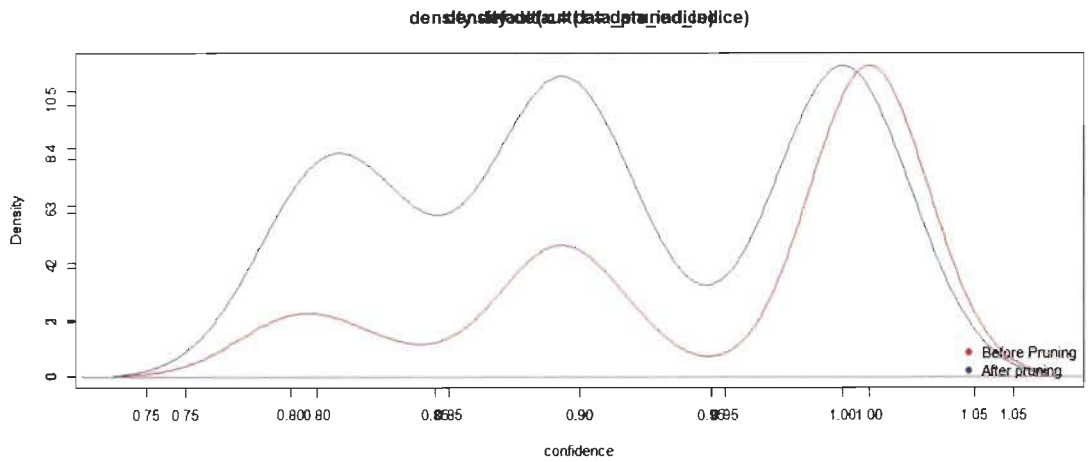


Figure 5.14. Tracé de la densité de la confiance expérimentation 2

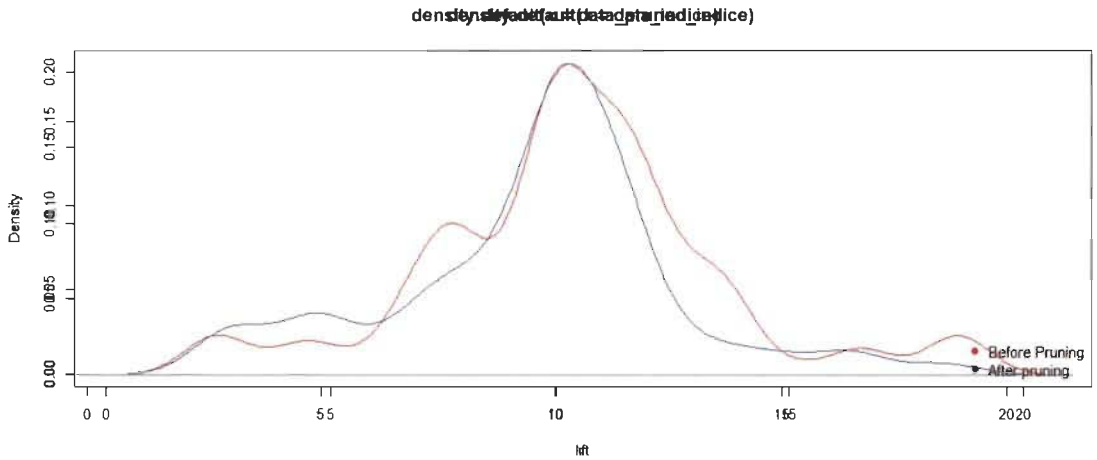


Figure 5.15. Tracé de la densité du lift expérimentation 2

5.2.3. Expérimentation 3

Au cours de cette expérimentation, notre principal but était de recueillir peu de règles d'association afin d'observer comment les résultats se présenteraient. Nous avons donc défini le support à 0.08 et la confiance à 0.5.

L'extraction de vingt règles d'association s'est déroulée en 0.04s, ci-après une visualisation des dix premières règles :

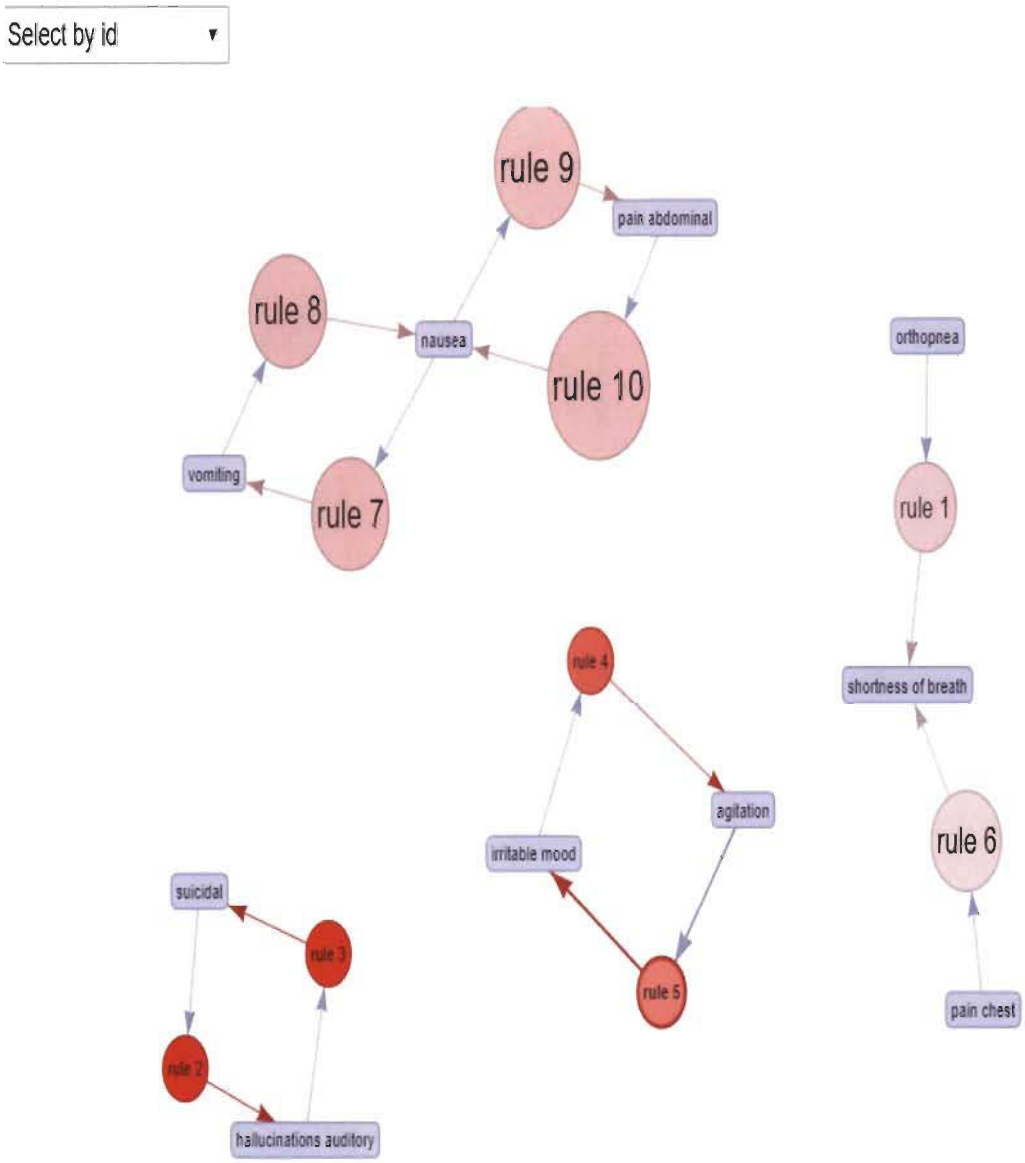


Figure 5.16. Graphe des dix premières règles expérimentation 3

Les dix règles visualisées sont toutes cohérentes, elles mettent en relation les symptômes et avec un algorithme de classification un spécialiste pourrait

découvrir la maladie correspondante. L'élagage des vingt règles nous permet d'observer les règles les plus pertinentes :

	lhs	rhs	support	confidence	coverage	lift
11	{pain abdominal}	{nausea}	0.114093959731544	0.586206896551724	0.194630872483222	3.63936781609195
12	{vomiting}	{pain abdominal}	0.100671140939597	0.576923076923077	0.174496644295302	2.96419098143236
13	{pain abdominal}	{vomiting}	0.100671140939597	0.517241379310345	0.194630872483221	2.96419098143236
14	{chill}	{fever}	0.120805369127517	0.72	0.167785234899329	2.89945945945946
15	{orthopnea}	{shortness of breath}	0.0939597315436242	0.933333333333333	0.100671140939597	2.83809523809524
16	{nausea}	{diarrhea}	0.0805369127516778	0.5	0.161073825503356	2.56896551724138
17	{dyspnea}	{shortness of breath}	0.120805369127517	0.782608695652174	0.154362416107383	2.37976929902396
18	{pain chest}	{shortness of breath}	0.100671140939597	0.75	0.134228187919463	2.28061224489796
19	{rale}	{shortness of breath}	0.0939597315436242	0.636363636363636	0.147651006711409	1.93506493506493
20	{pain abdominal}	{pain}	0.10738255033557	0.551724137931034	0.194630872483221	1.91178829190056

Figure 5.17. Règles élaguées expérimentation 3

En observant les deux règles les plus intéressantes qui sont la règle 15 {orthopena} → {shorteness of breath} et la règle 17 {dyspnea} → {shorteness of breath}, un non professionnel pourrait en déduire que ***l'orthopena*** et le ***dispnea***

sont toutes deux des maladies liées à la difficulté respiratoire. Par un professionnel ces informations pourraient permettre d'envisager l'apparition d'autres symptômes. Cette expérimentation a abouti à l'estimation de la perte d'informations :

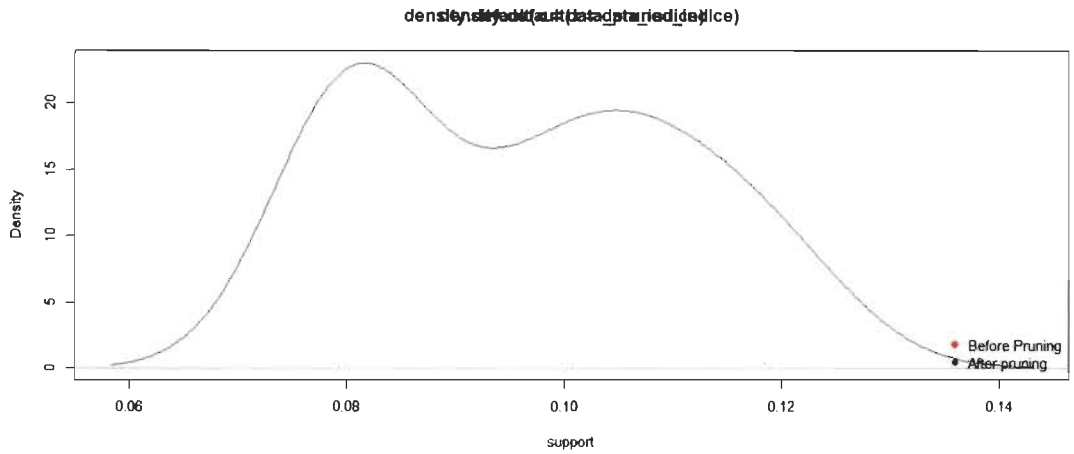


Figure 5.18. Tracé de la densité de l'indice de support expérimentation 3

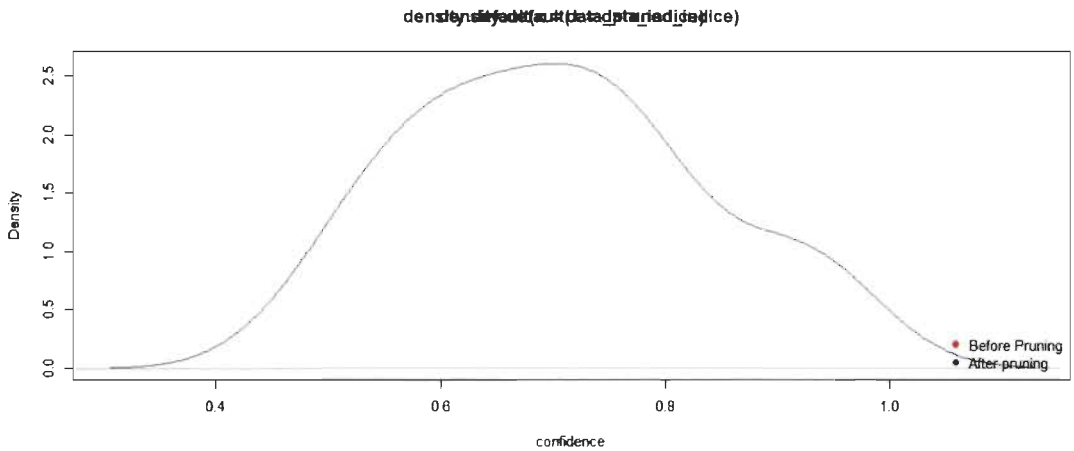


Figure 5.19. Tracé de la densité de l'indice de confiance expérimentation 3

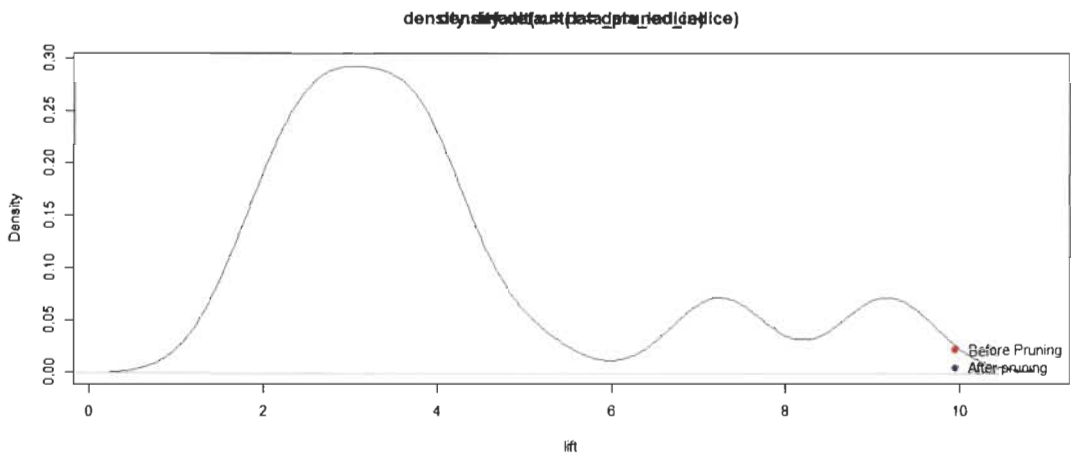


Figure 5.20. Tracé de la densité de l'indice de lift expérimentation 3

5.3 Discussion et interprétation des résultats

Les résultats obtenus lors des expérimentations reflètent les relations pouvant exister entre les symptômes et éventuellement les maladies.

En considérant la valeur de l'indice de confiance, on peut déterminer les règles les plus intéressantes en sortie. On retient donc des expérimentations les éléments suivants :

Les résultats obtenus sont en cohérence avec les données traitées et permettent de mieux comprendre les enjeux, dans notre cas ils représentent les symptômes les plus importants.

Dans les différentes expérimentations, nous avons observé que les règles intéressantes permettent d'obtenir des informations pertinentes, pouvant ainsi améliorer la compréhension des symptômes et la prise de décision. Dans

l'expérimentation 1, nous avons découvert des bruits qui sont la conséquence d'abréviations de certains mots ou des fautes d'orthographe. Par contre cela reste négligeable après élimination de ces bruits lors du processus d'élagage. Les graphes estimant la perte d'information nous convainquent du choix des valeurs des métriques. Par contre le grand nombre de règles obtenues peut relativement rendre l'analyse inconfortable. L'utilisateur devra prendre en compte les seuils définis afin de pallier ce problème.

Dans l'expérimentation 2, on retrouve un nombre acceptable de règles d'associations et ne contenant aucune information incohérente. Ces règles apportent des informations très pertinentes, sur l'état de certains patients en fonction de leurs conditions sociales. Les graphes de densité évaluant la perte d'informations avant et après l'élagage sont positifs et indiquent un faible ratio d'informations perdues après l'élagage.

L'expérimentation 3, montre des règles courtes qui ne représentent pas de l'information pertinente. Le peu de règles générées en est principalement la cause. On apprendra donc que pour notre ensemble de données, nous avons besoin d'un nombre de règles conséquent afin d'obtenir une meilleure analyse et interprétation des données. Les seuils définis ne sont pas non plus très judicieux, puisqu'il entraine une énorme perte d'information après l'élagage.

Les résultats obtenus se traduisent en informations exploitables selon l'appréciation de l'utilisateur. Elles sont en cohérence avec le sujet et le thème

des données. Ces informations peuvent s'avérer très importantes dans le domaine médical ou le besoin de prédire et de prévenir les maladies est très important. Ils peuvent notamment servir dans la recherche médicale pour trouver les associations entre différents symptômes afin d'élaborer de meilleurs vaccins et prédire les variantes de certaines maladies. Une classification des règles obtenues serait un moyen d'automatiser et de réduire l'effort d'analyse de cesdites règles. Le CBA (Classification Based on Associations) [9] est la technique la plus utilisée pour la classification des règles d'association.

5.4 Conclusion

Les règles d'association sont un excellent outil pour extraire de l'information dans une source de données structurées ou non structurées. Les algorithmes d'extraction de règles d'association tels que Apriori sont des moyens de découvrir les relations entre les différents items dans un ensemble de données. Le système que nous avons proposé tire avantage de cet algorithme afin de proposer des règles intéressantes pouvant permettre à l'utilisateur de comprendre ses données et faciliter ainsi la prise de décision.

Dans nos expérimentations, nous avons utilisé des données médicales pour démontrer la fiabilité de notre système, cependant tout type de donnée pourrait être utilisé pour des besoins d'analyses. La recherche dans le monde médical représentant des enjeux importants, nous avons trouvé judicieux de proposer un

outil permettant aux scientifiques d'exécuter des analyses de leurs données de façon efficace et optimale. Des modules de recherche et de report des règles d'association ont été intégrés à cet effet.

Au cours de ces différentes expérimentations, nous avons observé l'impact de notre méthode d'élagage sur les règles d'association obtenues :

Expérimentation	Nombre de règles avant élagage	Nombre de règles perdues	Nombre de règles après élagages	Nombre de règles intéressantes perdues	Indice de support	Indice de confiance
Expérimentation 1	2037	1034	1003	06	0.009	0.8
Expérimentation 2	313	8	305	0	0.05	0.8
Expérimentation 3	20	0	20	0	0.08	0.5

Tableau 5.1. Statistique des règles avant et après élagage

On note une perte de 1034 règles et de 08 règles respectivement dans la première et deuxième expérimentation. Plusieurs de ces règles ont des antécédents constitués d'abréviations ou de symboles d'alphabet grec. Nous considérons que ces règles sont bruitées et ne contiennent pas d'informations pertinentes. Certaines d'entre elles sont des règles redondantes qui ont été éliminées lors du processus d'élagage. Les 06 règles intéressantes perdues dans la première expérimentation sont des règles acycliques qui relient deux items avec deux règles. Notre méthode d'élagage a donc permis de réduire le nombre de règles

obtenues en garantissant la pertinence et l'unicité de ces règles. Les règles élaguées seront ainsi plus faciles à visualiser et apporteront une base de connaissance plus intéressante.

Les résultats de ces expérimentations sont satisfaisants et ont permis de mettre en lumière la pertinence des règles d'associations dans le domaine de l'exploration des données. Dépendant du volume de données, les utilisateurs devront adapter l'outil utilisé dans ce travail de recherche. En effet pour des volumes importants de données, la génération des règles peut prendre plus ou moins un temps conséquent de calcul dépendant des performances de calcul de la machine utilisée.

CHAPITRE 6 – Conclusion

Dans ce mémoire, nous avons traité la problématique d'élagage des règles d'association tout en proposant une méthode de représentation de la perte d'information. Nous avons proposé une méthode d'élagage qui permet de conserver les règles intéressantes et de visualiser la perte d'information en fonction des données statistiques avant et après élagage. Au cours de notre travail, nous avons obtenu des résultats intéressants qui pourraient permettre aux professionnels de mieux comprendre les relations résidents dans leurs données.

L'outil développé a été expérimenté en utilisant des données médicales, cependant tout type de données peut être utilisé. L'utilisateur aura toutefois la responsabilité de recueillir et stocker les données de manière à respecter le format demandé par l'application. Il devra s'assurer de ne pas avoir des données erronées qui pourraient biaiser son analyse.

L'impact des données est très important et trouver des associations dans les données médicales s'avère très utile dans la prise de décision. L'analyse de ces règles permet aux professionnels du corps médical de trouver des relations cachées entre les symptômes et les maladies ou entre les maladies elles-mêmes. Ce genre d'analyse peut participer au sauvetage de vie de certains patients qui pourraient présenter une maladie dans sa forme bénigne et indétectable au premier abord par certains professionnels.

Nous avons proposé un outil d'analyse qui assistera le professionnel dans ses diagnostics et qui permettra aux chercheurs de mieux comprendre les relations entre certaines maladies et leurs signes précurseurs. Ce travail qui met l'exergue sur le traitement de données médicales reste un sujet important à nos yeux face aux enjeux liés à la pandémie de la maladie à coronavirus qui bouleverse présentement le monde.

Les perspectives que nous proposons pourraient faire l'objet de travaux de recherche afin d'améliorer le système développé. La classification associative par exemple pourrait être un moyen d'améliorer l'expérience utilisateur des professionnels en leur permettant de prédire les maladies en fonction des symptômes du patient ou la recommandation des médicaments de soins en fonctions des données historiques. Une autre perspective serait de varier les données afin d'obtenir les relations entre les patients et optimiser ainsi le traitement des futurs patients qui présenteraient les mêmes maux. Notre méthode d'élagage pourrait ainsi se voir améliorer en intégrant, une technique de suppression des règles insignifiantes basée sur un indice de causalité des différents items.

Bibliographie et références

[1]. Abu-Mansour Hussein, Lee. McCluskey, Fadi thabta and Wael Hadi

(2010)Associative Text Categorisation Rules Pruning Method, Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention,De Montfort University, Leicester, UK. (pp. 39-44)

[2]. Agrawal R., Amielinski T. and Swami A. (1993). Mining association rule between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, (pp. 207-216). Washington, DC.

[3]. Agrawal R. and Srikant R. (1994). Fast algorithms for mining association rule.Proceedings of the 20th International Conference on Very Large Data Bases (pp.487-499), Santiago, Chile.

[4]. Alaa M. El-halees (2006). Arabic Text Classification Using Maximum Entropy.The Islamic University Journal (Series of Natural Studies and Engineering) Vol.15, No.1, pp 157-167, 2007, ISSN (pp.1726-6807)

[5]. Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical reports. AMIA Annu Symp Proc. 2008. p. 783-7. PMCID: PMC2656103.

[6]. Antonie M., Zaïane O. R. and Coman A. (2003). Associative Classifiers for Medical Images, Lecture Notes in Artificial Intelligence 2797, Mining Multimedia and Complex Data, (pp. 68-83), Springer-Verlag.

[7]. Antonie M. and Zaiane O. (2002). Text Document Categorization by Term Association, Proceedings of the IEEE International Conference on Data Mining (ICDM '2002), (pp.19-26), Maebashi City, Japan.

[8]. Baralis E., Chiusano S. and Garza P. (2008). A Lazy Approach to Associative Classification. IEEE Trans. Knowl. Data Eng. 20(2), (pp.156-171.

[9]. Baralis E., Chiusano S. and Garza P. (2004). On support thresholds in associative classification. Proceedings of the 2004 ACM Symposium on Applied Computing, (pp. 553-558). Nicosia, Cyprus.

[10]. Baralis E. and Paolo Garza.(2012). I-prune: Item Selection for Associative Classification. INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 27(pp. 279–299)

[11]. L. Cao, D. Luo, and C. Zhang, "Knowledge actionability: satisfying technical and business interestingness", *International Journal of Business Intelligence and Data Mining.*, Vol. 2, No. 4, pp. 496–514, 2007.

[12]. P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the Right Interestingness Measure for Association Patterns," In: *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min*, Edmonton, Alberta, Canada, pp. 32-41, 2002.

[13]. B. Liu, W. Hsu, S. Chen, and Y. Ma, "Analyzing the subjective interestigness of association rules", IEEE Intelligent Systems

[14]. S. O. Rezende, E. A. Melanda, M. L. Fujimoto, R. A. Sinoara, and V. O. de Carvalho, "Combining data-driven and user-driven evaluation measures to identify interesting rules", Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction, IGI Global, pp. 38–55, 2009.

[15]. R. Paul, T. Groza, J. Hunter, and A. Zankl, "Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain", Journal of Biomedical Semantics, Vol.5, No.1, pp.5-8, 2014.

[16]. L. Geng and H. J. Hamilton, "Interestingness measures for data mining", ACM Computing Surveys, Vol.38, No.3, pp. 1-32, 2006.

[17]. M. Velasquez and P. T. Hester, "An Analysis of Multi-Criteria Decision Making Methods", International Journal of Operations Research, Vol.10, No.2, pp.56-66, 2013.

[18]. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3, 2003.

[19]. A. S. Milani, A. Shanian, and C. El-Lahham, "Using different ELECTRE methods in strategic planning in the presence of human behavioral resistance", Journal of Applied Mathematics and Decision Sciences, Vol. 2006, pp. 1–19, 2006.

[20]. J. Wang, J. Han, and J. Pei, "Closet+: Searching for the best strategies for mining frequent closed itemsets", In: Proc. ninth ACM SIGKDD Int. Conf. Knowl. Discov. data Min, Washington, D.C, USA, pp. 236–245, 2003.

[21] M. J. Zaki and C-J. Hsiao, "CHARM: An efficient algorithm for closed association rule mining", In: Proc.of the 2nd SIAM Int. Conf. Data Min., Chicago, USA, pp.457-473, 1999.

[22]. Clifton, C., Cooley, R., Rennie, J.: TopCat: data mining for topic identification in a text corpus. IEEE Trans. Knowl. Data Eng. 16(8), 949–964 (2004)

[23]. Sirsat, S.R., Chavan, D.V., Deshpande, D.S.P.: Mining knowledge from text repositories using information extraction: A review. Sadhana 39(1), 53–62 (2014)

[24]. Madani, F.: Technology Mining bibliometrics analysis: applying network analysis and cluster analysis. Scientometrics 105(1), 323–335 (2015)

[25]. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49–56 (2008)

[26]. Clifton, C., Cooley, R.: TopCat: Data mining for topic identification in a text corpus. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 174–183. Springer, Heidelberg (1999)

[27]. Han, E.H., Karypis, G., Kumar, V., Mobasher, B.: Clustering based on association rule

hypergraphs. In: DMKD (1997)

[28]. A. Dutt, M. A. Ismail and T. Herawan, A Systematic Review on Educational Data Mining, vol. 3536, no. c, 2017.

[29]. Goh, D.H., Ang, R.P.: An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents. Behav. Res. Methods 39(2), 259–266 (2007)

[30]. Wong, P.C., Whitney, P., Thomas, J.: Visualizing association rules for text mining. In: 1999 IEEE Symposium on Information Visualization, 1999. (Info Vis' 99) Proceedings, pp. 120– 123. IEEE (1999)

[31]. Jayashankar, S., Sridaran, R.: Superlative model using word cloud for short answers evaluation in eLearning. Educ. Inf. Technol., 1–20(2016)

[32]. "Spatial data mining and geographic knowledge discovery-An introduction", Comput. Environ. Urban Syst., vol. 33, no. 6, pp. 403-408, 2009.

[33]. DePaolo, C.A., Wilkinson, K.: Get your head into the clouds: using word clouds for analyzing qualitative assessment data. TechTrends 58(3), 38–44 (2014)

[34]. Irfan, R., King, C.K., Grages, D., Ewen, S., Khan, S.U., Madani, S.A.& Tziritas, N.: A survey on text mining in social networks. Knowl. Eng.Rev. 30(2), 157–170 (2015)

[35]. Abdelghani Bellaachia and Erhan Guven Predicting Breast Cancer Survivability Using Data Mining Techniques"

[36]. Abdelghani Bellaachia and Erhan Guven Predicting Breast Cancer Survivability Using Data Mining Techniques"

[37]. Gaurav N. Pradhan & B. Prabhakaran, Associate rule mining in multiple, multidimensional time series medical data, IEEE. P 1–4, 2016.

[38]. Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 194-200, 2006.

[39]. Jyoti Soni, et al., "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975-8887), Volume 17-No.8, March 2011.

[40]. VAN RIJSBERGEN, C.J. (1977), "A THEORETICAL BASIS FOR THE USE OF CO-OCCURRENCE DATA IN INFORMATION RETRIEVAL", Journal of Documentation, Vol. 33 No. 2, pp. 106-119.

[41]. Williams G. J., Baxter R. A., He H. X., Hawkins S., Gu L., "A Comparative Study of RNN for Outlier Detection in Data Mining," IEEE International

Conference on Data-mining (ICDM'02), Maebashi City, Japan, CSIRO Technical Report CMIS-02/102, 2002.

[42]. Rosner B., On the detection of many outliers, Technometrics, 17, 221–227, 1975.

[43]. Two textbooks on neural networks are, C.M. Bishop, "Neural Networks for Pattern Recognition", Oxford University Press, Oxford (1995)

[44]. J. Dyer, Multiple Criteria Decision Analysis: State of the Art Surveys, Vol 78, Springer-Verlag, 2005.

[50]. Domingo-Ferrer, J., & Torra, V. (2001). Disclosure Protection Methods and Information Loss for Microdata. In P. Doyle, J. Lane, J. Theeuwes, & Z. L., Theory and Practical Applications for Statistical Agencies (pp. 91-110). Amsterdam

[51]. Zheng, Z., Kohavi, R., and Mason, L. (2001). Real World Performance of Association Rule Algorithms. In: Proc. 7th Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD'01). New York: ACM Press

[52]. Zheng, Z., Kohavi, R., and Mason, L. (2001). Real World Performance of Association Rule Algorithms. In: Proc. 7th Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD'01). New York: ACM Press

[53]. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., et al. (2012). Statistical Disclosure Control. Chichester, UK: John Wiley & Sons Ltd.

[54]. Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014, August 1). Introduction to Statistical Disclosure Control (SDC). Retrieved July 9, 2018, from http://www.ihsn.org/home/software/disclosure-control-toolbox.

[55]. Hafsa Jabeen, Market Basket Analysis using R August 21st, 2018, Datacamp

[56]. Antonie M. and Zaïane O. (2004). An associative classifier based on positive and negative rules. Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (pp. 64 - 69), Paris, France.