UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À

L'UNIVESITÉ DU QUÉBEC A TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE

DE LA MAITRISE EN MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES

PAR

ABDERRAHMANE BOUDJEMA

EXPLORATION DES RÈGLES D'ASSOCIATON CAUSALES

Novembre 2022

Remerciement

Grace à dieu, le tout puissant qui nous a donné le courage, la volonté et la patience pour achever notre travail.

Nous tenons à remercier vivement nos chers parents qui nous ont guidés depuis notre enfance vers le chemin du savoir.

Nous exprimons nos vifs remerciements et notre profonde gratitude à notre directeur de recherche, Biskri Ismail, qui a consacré tout son temps pour nous transmettre un ensemble de connaissances avec une volonté exemplaire, son soutien, son aide, ses conseils et sa bienveillance durant l'élaboration de ce mémoire.

Nous tenons à remercier nos enseignants du département de mathématiques et informatique appliquées de l'Université du Québec à Trois-Rivières, qui ont tant donné pour nous transmettre le savoir.

Nous remercions également toutes les personnes qui nous ont aidé de près ou de loin, à réaliser ce travail de recherche.

Résumé

L'analyse des données est le processus qui consiste à examiner et à interpréter des données afin d'élaborer des réponses et des questions. Conformément aux progrès rapides de la science et de la technologie au cours de la dernière décennie, la génération massive des données a un impact considérable sur divers domaines (médical, économique, etc.). Les mégadonnées nécessitent des méthodes pour les explorer et les structurer, parmi ces méthodes nous citons : les règles d'association, la classification, etc.

La recherche des règles d'association est une méthode très utilisée. Ces associations sont efficaces à découvrir les relations entre les variables dans une grande base de données. Néanmoins une simple règle d'association s'avère insuffisante. Elle peut dissimuler d'autres significations. Pour savoir ce que ces significations représentent, les chercheurs veulent de plus en plus identifier quel objet (ou évènement, ou autre) a causé une situation particulière.

Une règle d'association se présente sous la forme d'un antécédent menant à un conséquent selon un certain score. Dans les travaux les plus récents, des chercheurs ont émis comme hypothèse que certaines des règles d'association expriment une relation de causalité. Ils ont pu la démontrer pour des données structurées.

La plupart des algorithmes de repérage de cas de causalité sont basés sur une approche bayésienne. Cependant, l'application de ces algorithmes est limitée en raison de la complexité du calcul. Elle identifie, de ce fait, moins de relations causales.

Dans notre travail nous optons pour une approche pour identifier les relations causales, se basant sur des règles d'association. Cette méthode permet de découvrir plus de règles d'association causales dans un temps raisonnable.

Afin de prouver l'efficacité de notre approche, nous l'avons testé sur un dataset, puis nous avons comparé les résultats avec ceux obtenus au moyen d'une approche bayésienne.

Abstract

Data analysis is the process of examining and interpreting data to develop answers and questions. In line with the rapid progress of science and technology over the past decade, the massive generation of data has a considerable impact on various fields (medical, economic, etc.). Big data requires methods for exploring and structuring them, among these methods we cite association rules, classification, etc.

The search for association rules is a widely used method, these associations are effective in discovering the relationships between variables in a large database. However, a simple rule of association is insufficient and may hide other meanings. And to know what these meanings stand for, researchers increasingly want to find what object (or event, or whatever) caused a particular situation.

An association rule takes the form of an antecedent leading to a consequent according to a certain score. In more recent work, researchers have hypothesized that some of the association rules express a causal relationship. They were able to prove it for structured data.

Most causal case finding algorithms are based on a Bayesian approach. However, the application of these algorithms is limited due to the computational complexity and therefore finds fewer causal relationships.

In our work we opt for an approach to collect causal relationships based on association rules. This method makes it possible to discover more causal association rules in a reasonable time.

In order to prove the effectiveness of our approach, we tested it on a dataset, then we compared the results with those obtained using a Bayesian approach.

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1 : Introduction générale

Durant la dernière décennie, le volume des informations numériques a atteint des niveaux gigantesques. Chaque jour, 2,5 quintillions d'octets de données sont créés en raison de la surutilisation des réseaux sociaux comme facebook, twitter, etc [1].

La capacité de générer les informations digitales n'a jamais été aussi importante depuis l'invention de la technologie de l'information. Par exemple, le 4 octobre 2012, le premier débat présidentiel entre le président Barack Obama et le gouverneur Mitt Romney a déclenché plus de 10 millions de tweets en seulement 2 heures [2].

Ces discussions en ligne comme les commentaires ou bien la diffusion directe (streaming) d'un évènement offrent un nouveau moyen de détecter les intérêts du public. Elles sont généralement plus attrayantes par rapport aux médias génériques, tels que la radio ou la télévision

Afin de concevoir une description riche de l'information à partir de mégadonnées, des outils d'intelligence artificielle sont nécessaires, tels que l'extraction de règles d'association, les arbres de décision, les réseaux de neurones, etc.

Les progrès de l'intelligence artificielle ne cessent de s'accélérer et son utilisation s'impose de plus en plus, comme en médecine, en météorologie, etc.

Plusieurs branches de l'intelligence artificielle sont dédiées à la prédiction du comportement humain, aux prévisions climatiques, et aussi aux prévisions de l'évolution des marchés financiers. Toutefois, l'un des inconvénients de certains outils est le biais des algorithmes pour faire ces prédictions. Nous pouvons citer un exemple concret de fausses prédictions liées aux données utilisées. Un logiciel de prévision de risque a été appliqué à environ 200 millions de personnes pour identifier quels patients devront bénéficier de soins médicaux supplémentaires en fonction de ce qu'ils sont susceptibles de coûter au système. Un apprentissage automatique prédictif et des algorithmes ont été utilisés pour améliorer leurs précisions en fonction des nouvelles données reçues. Mais comme l'ont montré [3], cet outil a eu un résultat surprenant. Les données montraient clairement que les patients noirs, qui avaient plus de maladies chroniques que les patients blancs, n'étaient pas identifiés comme nécessitant des soins supplémentaires.

Le problème était le fait que l'algorithme a utilisé des données de réclamation d'assurance pour prédire le futur besoin sanitaire des patients. Toutefois, les concepteurs de l'algorithme n'avaient pas pris en compte le fait que les dépenses des soins de santé des Américains noirs

sont, généralement, inférieures à celles des Américains blancs ayant des problèmes de santé similaires. L'utilisation des dépenses de soins de santé comme indicateur de maladie a conduit l'algorithme à faire des recommandations en faveur des patients blancs. Les chercheurs ont informé le producteur du logiciel, qui a effectué de nouveaux tests sur ses propres données. Il a confirmé le problème et a collaboré avec les experts pour supprimer le biais de l'algorithme.

Cette situation illustre l'un des inconvénients de l'IA. Les algorithmes prédictifs peuvent commettre des erreurs et se tromper entre la corrélation et la causalité. Par conséquent, il est nécessaire que les analystes des données considèrent également la causalité.

Le domaine de la causalité en intelligence artificielle évolue rapidement. Au fur et à mesure que son potentiel devient plus apparent, les chercheurs l'utilisent dans des domaines divers, démontrant ainsi son vaste potentiel.

Nous proposons dans le cadre de ce travail une approche qui permet de générer les relations causales. Notre méthode consacre les étapes suivantes :

- Appliquer l'algorithme « Apriori » pour générer des règles d'association à partir d'une base de données transactionnelle, puis, enregistrer ces règles dans un tableau.
- Analyser le tableau et regrouper toutes les relations qui ont des variables « contrôlées » en commun, puis calculer le rapport des côtes, qui est une mesure statistique pour chaque relation. Au moyen de cette mesure l'algorithme distingue les règles causales des relations non-causales

Notre mémoire se divise en quatre chapitres. Outre le premier chapitre introduisant notre travail de recherche, le deuxième chapitre est consacré à la causalité et à l'exploration des relations causales. Le lecteur y trouvera la définition de la causalité selon le point de vue de plusieurs philosophes et chercheurs, comme Aristote et Hume, ainsi que les approches utilisées pour l'extraction des relations causales.

Dans le troisième chapitre nous présentons l'approche des règles d'association. Nous introduisons les différents algorithmes utilisés dans l'extraction des associations.

Nous présentons au chapitre quatre notre approche méthodologique et l'architecture général de notre algorithme

Nous donnons dans le cinquième chapitre, une interprétation des résultats obtenus avec des explications basées sur la notion de causalité selon Pearl et Mackenzie [4].

Nous formulons, enfin une conclusion finale pour ce travail et proposons nos perspectives de recherche.

Chapitre 2 : La causalité et les règles causales

2.1. Introduction à la causalité

La causalité est la relation qui s'établit entre une cause et son effet. La cause est ce qui produit quelque chose, et en est à l'origine. L'effet est la conséquence.

Comprendre la causalité était au centre de l'intérêt de plusieurs philosophes dans le passé, tels qu'Aristote. Ce dernier a introduit la théorie de la causalité comme un moyen de comprendre l'expérience humaine de la nature physique.

Quant à Hume [5], il a défini la causalité comme suit : "Une cause est un objet antérieur et contigu à un autre, et tellement uni à lui, que l'idée de l'un détermine l'esprit à former l'idée de l'autre, et l'impression de l'un à former une idée plus vive que l'autre". À noter que lorsque Hume dit "objets", du moins dans le contexte du raisonnement, il se réfère aux objets de l'esprit, c'est-à-dire aux idées et aux impressions, puisque Hume adhère à la "voie des idées" des débuts de l'ère moderne.

Pearl et Mackenzie, dans le livre « The book of why » [4], définissent la causalité comme suit : « Une variable X est une cause d'une variable Y, si Y dépend de la valeur de X ».

La liste des philosophes et leurs opinions sur la causalité est très longue, nous avons pris juste quelques-uns en considération.

Depuis sa création, l'humanité a toujours voulu connaitre la raison de son existence sur la planète. Aussi, comprendre la causalité pourrait satisfaire la curiosité humaine et améliorer le besoin de la société.

Ces dernières années, le monde a connu une révolution dans le domaine de l'intelligence artificielle. Plusieurs machines autonomes ont été créées et les algorithmes de prédiction initient à perfectionner leurs résultats à chaque évènement. Toutefois, les chercheurs ont remarqué que ces systèmes manquent d'adaptabilité et d'efficacité. Lorsqu'un nouvel évènement se produit, les résultats sont infructueux. Par exemple, durant la pandémie de Covid-19 en 2020, les algorithmes utilisés dans les grands sites de commerce électronique, comme amazon et ebay, ont fait de mauvaises prédictions. Par conséquent, cette pandémie pourrait être un bon facteur pour que les grandes entreprises commencent à considérer la causalité, en l'intégrant dans leurs systèmes intelligents

Une autre cause qui révèle la défectuosité de ces systèmes autonomes est l'absence d'explication. [6] a mentionné que l'apprentissage automatique restera quasiment une boite noire. L'inaptitude du système à prouver des prédictions avec des explications peut diminuer la confiance de l'utilisateur face aux algorithmes intelligents.

Notons qu'il est complexe pour une Machine-Learning de comprendre la relation « cause à effet ». Selon Pearl, si une machine autonome surmonte ce problème, elle s'approchera de l'intelligence humaine.

2.2. L'échelle de causalité

L'échelle de causalité contient trois étapes [4].

La première étape est l'observation. Par exemple, nous pouvons remarquer que lorsqu'un client achète du beurre et du pain, il achètera aussi du lait et peut être du yaourt. L'IA, dans ce cas, peut nous donner des résultats satisfaisants et faire des prédictions sur les achats, en se basant sur des données observationnelles.

La deuxième étape est l'intervention. Elle se classe plus haut que l'association (observation) parce qu'il ne s'agit pas seulement de voir ce qui est, mais de changer ce que nous voyons. Si nous posons la question suivante : « Que se passera-t-il si on augmente le prix du lait ? » Dans ce cas, on n'aura pas assez de données pour prédire la réponse. En revenant dans la base de données et en trouvant ce qui a causé la hausse des prix, l'IA peut trouver une réponse à notre question. Par exemple, la demande pour le lait était à la hausse, donc les prix ont augmenté. Toutefois, une autre cause peut entrainer une augmentation des prix mais l'algorithme d'apprentissage l'ignore. Par exemple, la pandémie a causé l'augmentation des prix alimentaires. Dans ce cas, l'intelligence artificielle est incapable de savoir la vraie cause de l'augmentation des prix, à moins qu'un modèle soit déjà testé sous des conditions qui ont causé cette augmentation. Pearl suggère que la meilleure façon de prédire les résultats d'une intervention est d'expérimenter les données sous des conditions soigneusement contrôlées. Il a aussi mentionné que les meilleures prédictions des effets d'interventions peuvent être faites avec un modèle causal précis. Une machine intelligente, qui utilise des faits (données observationnelles) sans un modèle causal, ne peut répondre aux questions interventionnelles. Dans ce cas, la machine sera toujours limitée par les données utilisées durant le test.

Les contrefactuels sont placés au sommet de la hiérarchie parce qu'ils englobent les interventions et les questions associatives. Si nous avons un modèle qui peut répondre à des requêtes

contrefactuelles, nous pouvons également répondre sur des questions sur les interventions et les observations. Par exemple, la question interventionnelle : Que se passera-t-il si nous doublons le prix ? Quelle serait la réponse à la question contrefactuelle : qu'est-ce qui va arriver si le prix avait été le double de sa valeur actuelle?

Le tableau 1 représente l'échelle de causalité introduite par Pearl et Mackenzie

Degré	**Activité**	**Questions**	**Exemples**
Association	Voir	-Qu'est-ce que c'est ? -Comment le fait de voir X changerait-il ma croyance en Y ?	- Qu'est-ce qu'un symptôme me dit sur une maladie ? -Qu'est-ce qu'un sondage nous dit sur les résultats des élections ?
Intervention	Faire	-Et si je faisais X ?	-Si je prends de l'aspirine, est-ce que mon mal de tête sera guéri ? -Et si nous interdisions les cigarettes ?
Contrefactuel	Imaginer Rétrospection	Pourquoi ? Est-ce que c'est X qui a causé Y ? Et si j'avais agi Différemment ?	-Est-ce l'aspirine qui a arrêté mon mal de tête ? -Et si je n'avais pas fumé ces deux dernières années ?

Tableau 1 : Échelle de causalité [4]

Afin de bien illustrer l'idée de causalité, nous citerons un exemple intuitif du tabac et du cancer.

Le tableau 2 représente la proportion de fumeurs et de non-fumeurs chez les patients atteints du cancer du poumon et chez les patients témoins atteints de maladies autres que le cancer

Groupe des malades	Nombre de non-fumeurs	Nombre de fumeurs
Hommes : - Nombre de patients atteints de cancer du poumon (649) - Nombre de patients atteints d'autres maladies (649)	2 27	647 622
Femmes : - Nombre de patients atteints de cancer du poumon (60) - Nombre de patients atteints d'autres maladies (60)	19 32	41 28

Tableau 2 : Résultats médicaux concernant les fumeurs et les non-fumeurs atteints de cancer du poumon

Si Jean fume des cigarettes pendant plus de dix années, il aura plus de chance d'avoir un cancer des poumons et s'il n'en fume pas, les risques d'avoir un cancer des poumons sont réduits.

Les résultats dans le tableau 2 montrent que fumer cause le cancer des poumons. 647 patients fumeurs ont eu le cancer contre 2 patients non-fumeurs qui ont eu le cancer des poumons.

Dans la causalité, nous comparons toujours l'effet avec l'action (S) qui l'accompagne (qui cause cet effet) et aussi le résultat d'un événement si une action (S) est absente, comme dans l'exemple cité précédemment.

2.3. Les paralogismes liés à la causalité

L'utilisation de l'intelligence artificielle pour prédire le comportement d'une population peut entrainer des erreurs. Les systèmes d'informations médicales doivent intégrer les modèles de causalité qui expliquent mieux le comportement des individus. Analyser les données liées aux êtres humains sans considérer la causalité pourrait conduire au paralogisme, par exemple le paradoxe de Simpson et la fausse corrélation.

2.3.1 Fausse corrélation

Il s'avère que la corrélation n'est pas une relation de cause à effet. Ceci illustre le concept des fausses corrélations. L'indicateur de Superbowl Américain, suggérant qu'une victoire de l'équipe de la conférence de football américain signifie une baisse du marché boursier au cours de l'année, alors qu'une victoire de l'équipe de la conférence national de football indique une

hausse du marché financier, est un exemple de fausses corrélations. Il est clair que la victoire d'une équipe ou sa défaite et le marché boursier n'ont aucune relation de causalité, en dépit de leur corrélation. Cette relation pourrait être superflue ou serait le produit d'une troisième variable ignorée.

Un tel exemple nous rappelle que nous devons toujours être prudents lorsque nous travaillons avec la corrélation. Pour éviter une mauvaise interprétation des résultats, il faut chercher si une association est statistiquement significative. Plusieurs paramètres permettent de le savoir comme « p-value », « chi-square », « intervalle de confiance », etc.

2.3.2 Le paradoxe de Simpson

Le paradoxe de Simpson est l'expression d'un résultat contre-intuitif qui peut se produire dans les agrégations statistiques. Le paradoxe fait référence au fait que les résultats des comparaisons entre les groupes sont inversés lorsque les groupes sont combinés.

Prenons le cas de l'université de Berkeley, qui a été presque poursuivi pour discrimination sexuelle en 1973, à cause d'une mauvaise interprétation des données.

	Hommes		**Femmes**	
	Application	**Admission**	**Application**	**Admission**
Total	8442	44%	4321	35%

Tableau 3 : Taux d'application et d'admission à l'université de Berkeley pour les deux sexes en 1973

A partir de tableau 3, nous remarquons que le taux d'applications pour la session d'automne à l'université de Californie Berkeley était de 12,763 pour l'année 1973. Il y avait 8442 candidats et 4321 candidates, environ 44% d'hommes et 35% de femmes ont été admis.

Après une analyse profonde des dossiers pour comprendre ce biais, les enquêteurs ont obtenu des données plus précises (voir tableau 4). Il s'est avéré que quatre départements sur six ont accepté plus de femmes que d'hommes, d'où ce biais qui a été en faveur des femmes.

Département	Hommes		Femmes	
	Application	Admission	Application	Admission
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Tableau 4 : Taux des postulants et des admis dans chaque département pour les deux sexes

Les auteurs de [7] ont conclu que les femmes ont postulé pour des départements plus compétitifs avec un taux d'acception très bas, alors que les hommes ont choisi des spécialités moins compétitives avec un taux d'admission élevé.

Pour empêcher la reproduction de ce phénomène, les chercheurs ont suggéré plusieurs approches qui aident à explorer les relations causales d'une façon rapide et qui donnent des résultats efficaces, en utilisant les données observationnelles. Parmi ces méthodes, nous pouvons citer des méthodes basées sur des contraintes et d'autres basées sur le score.

Avant de parler des approches utilisées dans l'exploration des règles causales, nous définirons quelques notations et termes utilisés dans ces approches comme les graphes, les réseaux bayésiens, etc.

2.4. Définition et notation

2.4.1 Les Réseaux bayésiens

Les réseaux bayésiens sont des modèles graphiques probabilistes, dans lesquels les nœuds représentent des variables aléatoires, tandis que les arêtes représentent des hypothèses de dépendance conditionnelle. Les réseaux bayésiens fournissent une représentation compacte des distributions de probabilité associées [8]. Ils sont appliqués au domaine médical [9], à la recherche heuristique [10], etc.

2.4.2 Les graphes

Un graphe simple G se compose d'un ensemble V(G) infini de nœuds et un ensemble E(G) fini de paires distinctes non ordonnées nommées arêtes. Une arête est une liaison entre deux nœuds et habituellement notée par vw où v et w sont des nœuds. Deux nœuds connectés sont

aussi adjacents. La figure 1 représente un graphe simple G avec un ensemble de nœuds V(G)= {w, z, y, v} et un ensemble d'arêtes E(G) contient {wv, wz, yz, vz}

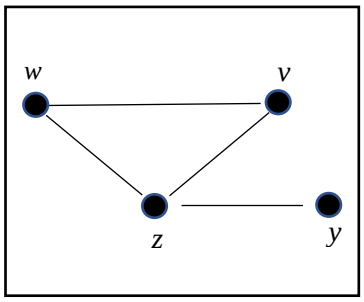


Figure 1: Un graphe simple

Un graphe direct est un graphe qui possède uniquement des arêtes directes entre les nœuds « → » comme dans la figure 2.1. Un graphe partiellement direct est un graphe ou il existe les arêtes directes et indirectes. La figure 2.2 illustre ce type des graphes.

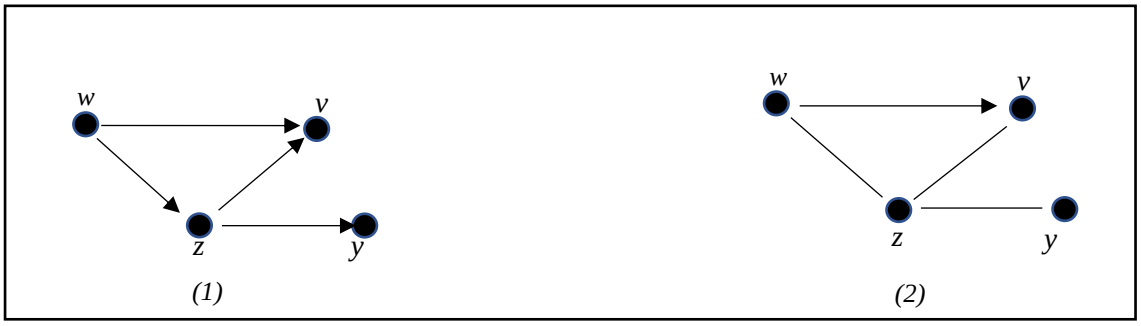


Figure 2 : (1) Graphe direct (2) Graphe partiellement direct

Un chemin dans un graphe est une séquence de nœuds adjacents, quelle que soit le sens de l'orientation. Un chemin direct se compose d'arêtes qui sont toutes orientées dans le même sens. Comme dans la figure 3 où la flèche (→) représente un chemin. « w → v → y » ou « w← v ← y » sont des chemins orientés ou directs. Dans « w→ v ←y », le chemin est non orienté ou indirect.

Un graphe avec un nœud w qui a un chemin direct entrant et un chemin sortant de ce nœud lui-même est un graphe orienté cyclique 3.a. Un graphe orienté acyclique est un graphe qui ne possède pas de cycles 3.b.

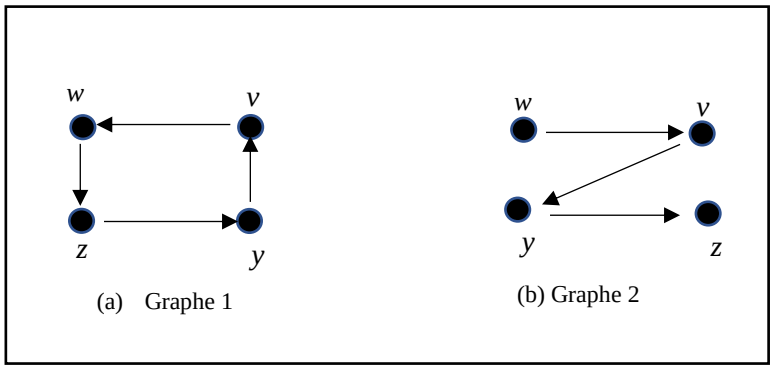


Figure 3 :: (a) graphe orienté cyclique (b) graphe orienté acyclique

2.5. Les algorithmes d'exploration des règles causales

Plusieurs algorithmes sont proposés pour implémenter l'apprentissage de réseaux bayésiens en suivant des différentes approches.

Dans le tableau 5 nous citons des algorithmes basés sur deux approches :

- Approche basée sur la contrainte
- Approche basée sur le score

Approches	**Algorithmes**
Approche basée sur la contrainte	- Algorithme de Peter and Clarck [11]
Approche basée sur le score	-Recherche locale [12]

Tableau 5 : Algorithmes d'exploration des règles causales

2.5.1 Approche basée sur la contrainte

2.5.1.1 L'algorithme de Peter et Clarck (PC algorithm)

L'algorithme PC est basé sur la condition causale de Markov [11]. Selon cette condition, une relation entre deux variables est directement causale si le facteur de confusion est absent et que ces deux variables ne sont pas conditionnellement indépendantes par un autre ensemble de variable.

Cet algorithme est basé sur deux étapes principales. La première étape permet d'analyser un graphe complet non orienté (voir figure 4.b). Un graphe contient des variables (nœuds) et un ensemble d'arêtes entre une paire de variables. Chaque arrête est considéré comme une association entre les variables. Dans la deuxième étape, l'algorithme oriente les arêtes pour former un graphe complet orienté (voir figure 4.d) [13] .

Exemple du fonctionnement de l'algorithme PC [14] :

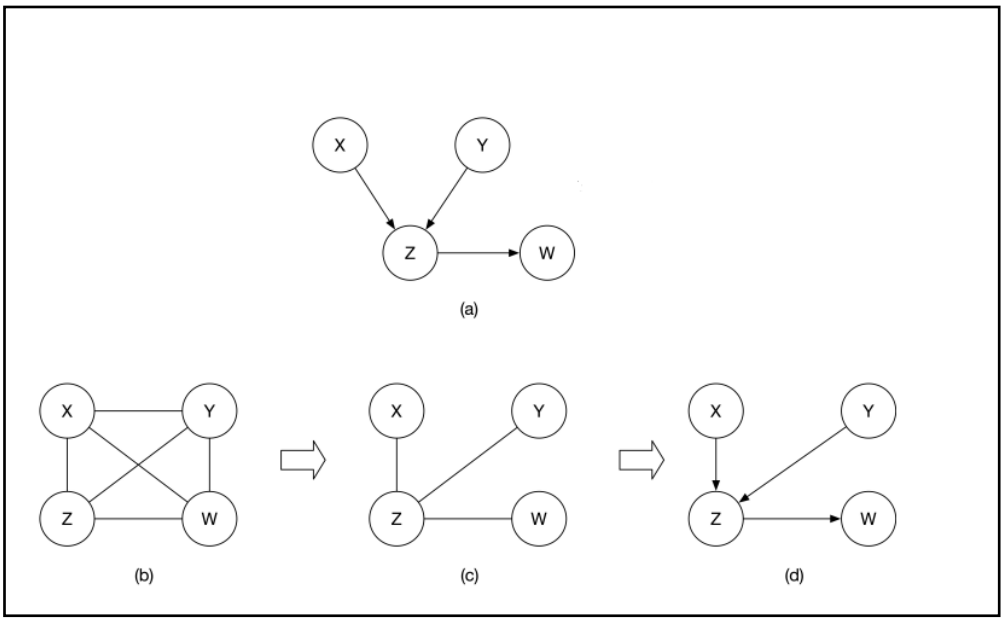


Figure 4 : Illustration graphique de déroulement de l'algorithme PC

Il faut bien considérer que la taille d'un ensemble de variables conditionnelles augmentera à chaque itération. Dans notre exemple, le graphe est simple. Donc il suffit d'une seule itération pour supprimer les arêtes adjacentes. La première itération consiste à supprimer les variables conditionnées par Z. Le fonctionnement de l'algorithme PC est comme suit :

- Identifier un graphe complet non orienté (figure 4.b), en se basant sur l'ensemble des nœuds.
- Éliminer les arêtes X-Y conditionnées par une variable Z.
  $X \perp\!\!\!\perp W \mid \{Z\}$ $X \perp\!\!\!\perp Y \mid \{Z\}$
  $Y \perp\!\!\!\perp W \mid \{Z\}$
- Orientation des arêtes, comme dans la figure 4.d.
- Le graphe (d) est la structure finale obtenue par l'algorithme PC.

Les inconvénients des algorithmes basés sur la contrainte surgissent lorsqu'ils sont appliqués à des données massives. Plusieurs versions de PC ont été suggérées pour améliorer la complexité de l'algorithme et pour augmenter la vitesse de découverte des relations causales.

2.5.2 Approche basée sur le score

Cette approche définit au préalable un critère pour évaluer l'adéquation d'un réseau bayésien aux données. Par la suite, elle cherche dans l'espace des graphes une structure qui a un score maximal.

Avant tout, l'approche basée sur le score est un problème de recherche et se compose de deux parties : la définition de la métrique du score et l'identification de l'algorithme de recherche.

Dans la littérature, il existe plusieurs algorithmes basés sur la méthode de score, tels que l'évaluation par l'entropie et l'information mutuelle [15], la matrice de confusion [16] et l'algorithme de recherche qui est la méthode la plus utilisé pour l'évaluation d'un graphe. Parmi ces algorithmes de recherche, nous citons la recherche locale [12] et l'algorithme glouton [17].

2.5.2.1 La recherche locale

C'est une méthode heuristique pour résoudre des problèmes combinatoires difficiles, tout en appliquant des algorithmes simples. La recherche locale est utilisée dans différents domaines qui exigent des calculs computationnels complexes comme la recherche opérationnelle, la bio-informatique et l'informatique.

Ces algorithmes sont utilisés dans l'apprentissage de la structure d'un réseau bayésien. Parmi ces algorithmes, il existe la recherche gloutonne [18].

Dans ce chapitre, nous avons présenté la causalité selon le point de vue de quelques philosophes et chercheurs. Nous avons mentionné aussi les différents algorithmes et approches basés sur les graphes bayésiens pour identifier si une règle est causale ou non causale dans un ensemble de données. Néanmoins, Les algorithmes cités se butent au temps de traitement des données lorsqu'il s'agit de faire des calculs sur des grandes bases de données, tel que l'algorithme de Peter et Clark. Quelques chercheurs ont tenté de perfectionner ces algorithmes pour accélérer le traitement de données. D'autres se basent sur des approches différentes comme dans [19] où ils proposent une méthode pour tester si une règle est causale ou non. Cette dernière est fondée sur des règles d'associations et des métriques statistiques. Dans le chapitre suivant, nous parlerons des règles d'association et leurs relations avec notre travail.

Chapitre 3 : Les règles d'association

3.1. Introduction

Les règles d'association nous permettent de visionner les itemsets qui cooccurrent et qui sont générés par les bases de données transactionnelles. Cependant, l'exploitation de ces données massives reste difficile pour l'homme et la machine. C'est la raison pour laquelle, divers travaux de recherche se sont intéressés à optimiser les techniques de fouille des données. Parmi ces travaux nous citons l'algorithme Apriori [20] et l'algorithme FP-growth [21]. Ces algorithmes permettent d'optimiser l'extraction des données et de générer des règles d'association pertinentes.

Soit un ensemble d'items tels que I = {i1, i2, i3, i4, …, im}. Un item ij est une donnée dont le type dépend du domaine étudié. Les données peuvent être de format textuel, image, audio ou autre [22].

Soit un ensemble de transaction tels que T = {t1, t2, t3, t4, …, tn}. Chaque transaction ti a un identifiant unique qui contient un ensemble d'items appartenant à I.

Une règle d'association peut se présenter sous la forme suivante :

$$A \rightarrow B$$

Où A et B sont des itemsets. Un itemset est un ensemble d'items appartenant à I. L'itemset A est l'antécédent d'une règle et l'itemset B représente le conséquent d'une règle. Pour extraire les associations importantes, des mesures de qualité devront être prédéfinies, tels que le support, la confiance et le lift.

Le support permet de calculer le pourcentage de transactions qui contiennent A et B. Il est calculé avec la formule suivante :

$$\text{Support}(A \rightarrow B) = \frac{nombre\ d'occurence(A \cup B)}{nombre\ total\ des\ transactions}$$

où *nombre d'occurrence* $(A \cup B)$ représente le nombre de transactions où les itemsets A et B cooccurrent.

La confiance permet de calculer le pourcentage de transactions qui contiennent les itemsets A et B par rapport à toutes les transactions qui contiennent l'itemset A, mais pas nécessairement B. Elle est calculée au moyen de la formule suivante :

$$\text{Confiance}(A \rightarrow B) = \frac{nombre\ d'occurence(A \cup B)}{nombre\ d'occurence(A)}$$

Ces mesures sont généralement complétées par d'autres mesures utiles, comme le lift.

$$\text{Lift}(A \rightarrow B) = \frac{Confiance(A \rightarrow B)}{Support(B)}$$

Sachant que le Support (B) correspond au rapport du nombre de transactions contenant B par le nombre total de transactions.

Le lift mesure la fréquence de cooccurrence de l'antécédent et le conséquent d'une règle. Si le lift équivaut 1 alors l'antécédent et le conséquent sont indépendants. Si le lift est supérieur à 1, alors A et B cooccurrent fréquemment. Si le lift est inférieur à 1, la cooccurrence de A et B est moins fréquente

Le but de ces calculs est de découvrir les règles d'association pertinentes. Une règle d'association est dite pertinente lorsque son support et sa confiance sont respectivement supérieurs à des seuils choisis par l'utilisateur. Si cette condition est vérifiée, la règle d'association est considérée comme une règle importante.

3.2. Les algorithmes d'extraction des règles d'association

3.2.1 Algorithme Apriori

C'est un algorithme (voir figure 5) classique. Il est basé sur l'approche « bottom-up »[1] et utilise aussi la recherche en largeur d'abord dans un arbre de recherche binaire [23]. Il sert à extraire les règles d'association fréquentes à partir de bases de données. Cet algorithme est très effectif dans plusieurs domaines professionnels (économie, médecine, etc.).

[1] La méthode « bottom-up » est une technique utilisée pour résoudre les problèmes de la programmation dynamique, elle aide à éviter la récursion et économiser la consommation d'espace mémoire.

Algorithme : Apriori

Entrée : Base de données

Sortie : Enregistrement pour item sets

L_1 = {Ensemble des item sets}

count = compteur pour compter les candidats dans un ensemble

Pour (k = 2; L_{k-1} ensemble vide; k++)

C_k = Apriori (L_{k-1}) // Génération des nouveaux candidats

Pour chaque transaction t ∈ D

C_{t} = sous-ensemble (C_k, t)

Pour chaque candidat c ∈ C_k

count= count+1

fin pour

L_k = {c ∈ C_k | count ≥ $Supp_{min}$} //Ajouter à l'ensemble L_k tous les Item sets qui ont un support supérieur au support minimal

Fin pour

Fin pour

Return L_k

Figure 5 : Algorithme Apriori

Pour bien comprendre le fonctionnement d'algorithme Apriori, donnons un exemple explicatif sur le panier d'épicerie (Voir tableau 6).

#Transaction ID	Transactions
1	{Oignon, Pomme de terre, Lait}
2	{Oignon, Burger, Lait}
3	{Oignon, Boisson gazeuse, pomme de terre}
4	{Pomme de terre, Burger}
5	{Pomme de terre, Lait}
6	{Burger, Lait, Boisson gazeuse}
7	{Oignon, Pomme de terre, Burger}

Tableau 6 : Ensemble des transactions

1. Analyser la base de données et extraire la liste des items dans les transactions pour former les itesemts. Dans la première itération la taille des itemsets égale à 1 (voir tableau 7)

Itemsets	Fréquence	Support	Confiance
Oignon	4	4/7 = 0,57	1
Pomme de terre	5	5/7 = 0,71	1
Burger	4	4/7 = 0,57	1
Lait	4	4/7 = 0,57	1
Boisson gazeuse	2	2/7 = 0,28	1

Tableau 7 : Articles fréquents dans une base de données

2. L'élimination des itemsets qui ont la valeur de support inférieure à 0,4 (seuil de support minimum fixé par l'utilisateur). Dans ce cas, l'itemset formé de l'item Boisson gazeuse sera supprimé, son support étant de 0.28.

Support (Boisson gazeuse) = $\frac{nombre\ d'occurence(boisson\ gazeuse)}{Nombre\ de\ transaction} = \frac{2}{7} = 0.28$

Confiance (Boisson gazeuse) = $\frac{nombre\ d'occurence\ (Boisson\ gazeuse)}{nombre\ d'occurence\ (Boisson\ gazeuse)} = \frac{2}{2} = 1$

3. Après avoir éliminé l'itemset avec le support inférieur au seuil, l'algorithme augmente de 1 la taille des itemsets. L'ordre dans lequel les items apparaissent dans les itemsets n'est pas important. Notons que l'item Boisson Gazeuse ne sera plus pris en considération ayant été éliminé en raison de son support.
4. Calcul pour chaque itemset de sa fréquence et de son support (voir tableau 8). Les seuls itemsets retenus seront ceux dont le support est supérieur au seuil établi à l'étape 2.

Itemsets	Fréquence	Support
{Oignon, Pomme de terre}	3	3/7 = 0.42
{Oignon, Burger}	2	2/7 = 0.28
{Oignon, Lait}	2	2/7= 0.28
{Pomme de terre, Burger}	2	2/7 = 0.28
{Pomme de terre, Lait}	2	2/7 = 0.28
{Burger, Lait}	2	2/7 = 0.28

Tableau 8 : Ensemble des itemsets fréquents

5. Après la génération de tableau 8, l'algorithme élimine les paires dont le support est inférieur au seuil minimum (0,4). Les itemsets restants sont {Oignon, Pomme de terre}.

6. L'algorithme tente d'augmenter une autre fois la taille des itemsets pour passer de 2 à 3. On remarque que dans notre cas ce ne sera pas possible puisqu'à l'étape précédente seul un itemset formé de deux items a été conservé. On ne peut donc former d'itemsets à trois items.

3.2.2 Algorithm FP-growth (Frequent pattern-growth)

C'est un algorithme qui permet d'extraire les règles d'association fréquentes, en se basant sur la méthode « diviser pour régner ». Il est considéré comme une amélioration de l'algorithme Apriori. Un modèle fréquent est généré sans passer par plusieurs itérations [21].

Le fonctionnement de l'algorithme est comme suit [24] :

Notant que Le support minimum dans cet exemple est égal à 3.

FP-growth commence par faire la compression de la base de données qui contient les itemsets fréquents. La représentation de ces itemssets sera sous forme d'un FP-tree, qui contiendra tous les itemssets ainsi que leurs informations. L'étape suivante est de diviser cette base de données compressée en base de données conditionnelles. Chacune de ces collections contiendra les itemsets fréquents. FP-growth fouillera chacune des bases de données. Cet algorithme est couteux au niveau de la complexité en espace.

Un exemple pour construire un « FP-tree » est donné à la figure 7. La Figure 7 (a) est un échantillon pris d'une base de données transactionnelles pour la création de « FP-tree », avec le support minium = 3.

Chaque ligne du tableau est composée d'un ensemble d'items apparaissant simultanément dans une transaction. Pour construire le FP-Tree avec cette base de données transactionnelle, nous devons, tout d'abord, trouver la fréquence des items. Pour les items fréquents, seuls les items avec un support égal ou supérieur au support minimum, qui est égal à 3, seront enregistrés dans la liste par ordre décroissant. Ces items seront utilisés pour la création du FP-Tree. La figure 7(b) représente la liste des items les plus fréquents en ordre décroissant. La figure 7(c) montre l'ensemble des transactions après la suppression des items avec un support inférieur à 3.

Transaction	Items
1	f, a, c, d, g, i, m, p
2	a, b, c, f ,l, m, o
3	b, f, h, j, o, w
4	b, c, k, s, p
5	a, f ,c, e, l, p, m, n

(a) Ensemble des transactions-dans une base de données

Items	Nombre d'item
f	4
c	4
a	3
m	3
p	3
b	3

(b) Liste des items plus fréquents par ordre

Transaction	Items
1	f, c, a, m, p
2	f, c, a, b, m
3	f, b
4	c, b, p
5	f, c, a, m, p

(c) Liste des transactions tronquée et ordonnée

Figure 6 : Exemple des transactions dans une base de données et un ensemble des transactions ordonné et compressé

Pour construire un arbre FP-Tree, l'algorithme ajoutera des items à chaque itération jusqu'à le parcours de toutes les transactions. Dans notre exemple, nous remarquons qu'il existe cinq transactions, l'algorithme, donc, doit faire cinq itérations. Dans la première itération, il ajoutera les items {f, c, a, m, p} dans l'arbre, comme illustré dans la figure 7(a), ensuite il passe à la deuxième itération, l'algorithme vérifie les items qui se répètent dans les deux transactions pour incrémenter de nombre d'occurence, si un nouvel item y apparait, alors un nouveau chemin sera créé, comme la figure 7(b) illustre l'ajout d'un item {b}. En comparant les transactions {#1, #2}, le nouvel item {b} y apparait après l'item {a} (Voir figure 6(c), transaction #2), donc l'algorithme créera un nouveau chemin en partant de nœud {a : 2}. La même logique est appliquée pour le reste des transactions.

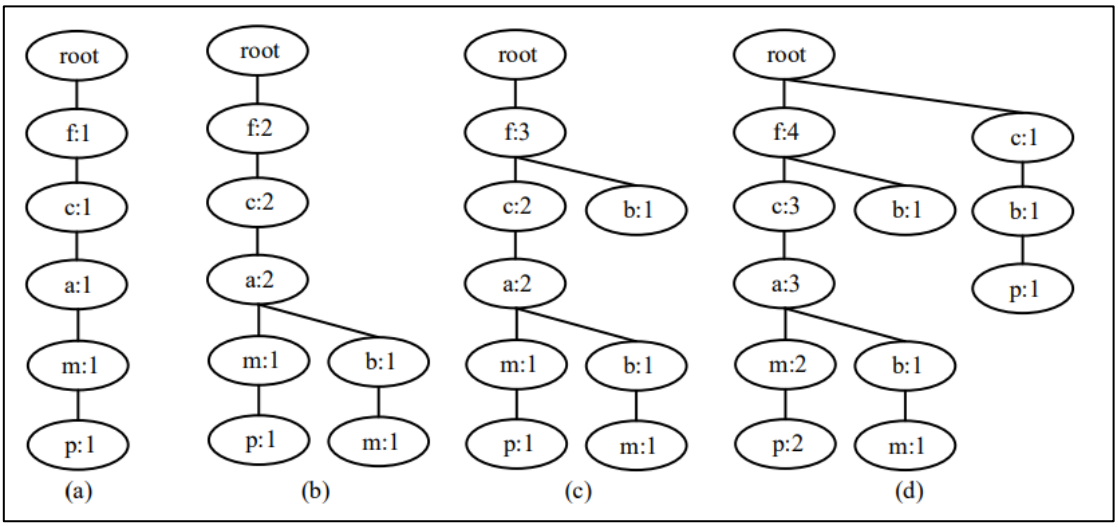
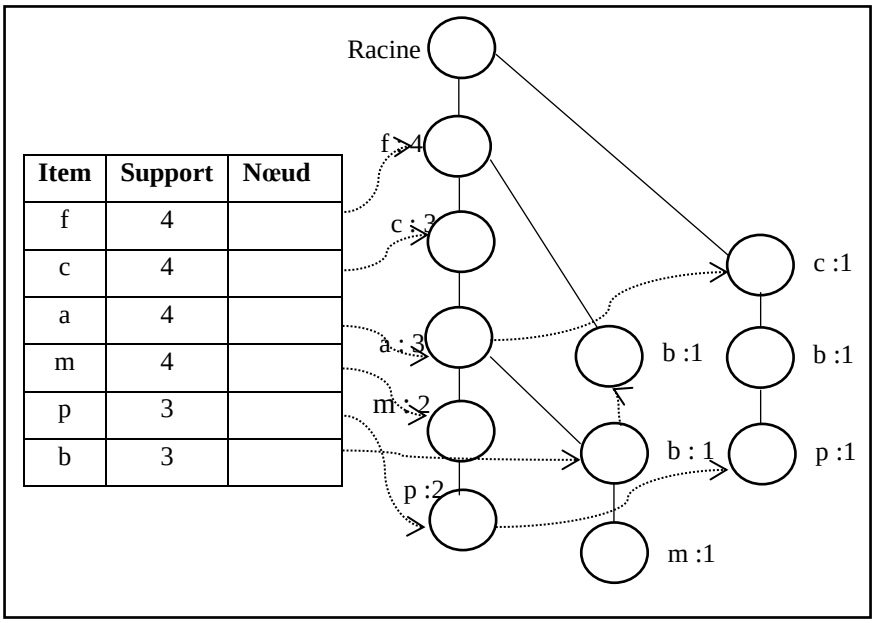


Figure 7 : Les étapes de construction de l'arbre FP-Tree

La figure 8 représente l'arbre final avec son tableau contenant les items fréquents



Item	Support	Nœud
f	4	
c	4	
a	4	
m	4	
p	3	
b	3	

Figure 8 : Construction d'un arbre FP-tree

Le tableau 9 montre le calcul de deux paramètres importants pour l'extraction des règles d'association fréquentes dans l'Algorithme FP-growth :

1. La base de modèle conditionnelle : cette mesure permet d'avoir la fréquence d'un item en trouvant tous les chemins qui mènent à un nœud qui a les informations de l'item recherché. Par exemple dans la figure 8, il existe trois chemins qui mènent au nœud (b :1). Alors nous sauvegardons ces trois chemins avec la fréquence de b. Ce qui donne [{f, b :1}; {c, b :1}; {f, c, a, b :1}]
2. Le FP-Tree conditionnel : Permet de connaitre les items communs aux deux itemsets extraits dans la base de modèle conditionnelle

Soit « Φ » représentant un ensemble vide.

Item	Base de modèle conditionnelle	FP-Tree conditionnel	Itemset fréquent
p	{{f, c, a, m : 2}, {c, b : 1}}	{c :3}	{<c, p : 3>}
m	{{f, c, a : 2}, {f, c, a, b : 1}}	{f, c, a :3}	{<f, m: 3>, <c, m: 3>, <a, m: 3>, <f, c, m: 3>, <f, a, m: 3>, <c, a, m :3>}
b	{{f, c, a : 1}, {f : 1}, {c : 1}}	Φ	{}
a	{{f, c : 3}}	{f, c : 3}	{<f, a :3>. <c, a :3>. <f, c, a :3>}
c	{f : 3}}	{f : 3}	{<f, c : 3>}
f	Φ	Φ	{}

Tableau 9: Extraction des item sets fréquents à partir d'un arbre FP-Tree conditionnel et base de modèle conditionnelle

Nous remarquons que les règles d'association {{f, m}, {a, m}, {f, a, m}} sont des règles redondantes. L'utilisateur doit supprimer les règles redondantes et conserver les autres règles.

L'extraction des règles d'association fréquentes et le calcul de support de chacune d'elles est exhaustif et augmente la complexité en temps et en espace. Par exemple, l'algorithme Apriori génère de nombreux itemsets, ce qui entraine la génération des règles d'association superflues.

Selon [25] il existe cinq règles générées pour l'item set « XYZ » et si la règle X→YZ répond aux exigences de support et confiance choisis par l'utilisateur alors les règles suivantes : XY→Z, XZ→Y, X→Z, X→Y, sont des règles redondantes.

Certains travaux portant sur ces limitations suggèrent des améliorations pour optimiser le calcul computationnel, pour économiser l'espace mémoire de la machine et simplifier l'implémentation des algorithmes. Comme dans [26], une nouvelle méthode est proposée pour améliorer l'algorithme FP-growth, vu que la première version de l'algorithme était difficile à implémenter à cause de la surutilisation des structures des données et aussi de la consommation énorme d'espace mémoire pour enregistrer les transactions.

3.3. Domaines d'Application

Les règles d'association sont utilisées dans de nombreuses applications pour trouver des informations pertinentes. Les exemples ci-dessous montrent les différentes applications des règles d'association :

1- Analyse du panier d'épicerie :

L'une des applications les plus connues des règles d'association est l'analyse du panier d'épicerie [27]. Elle permet, notamment, de découvrir la relation entres les articles achetés par les clients. L'amélioration de la technologie d'information permet aux détaillants d'obtenir des données de transactions quotidiennes à un coût très faible. Ainsi, une grande quantité de données utiles est extraite pour permettre aux détaillants de bien analyser le marché.

L'exploitation des données est utilisée pour extraire des informations précieuses à partir des grandes bases de données.

De nos jours, chaque produit est accompagné avec un code barre. Ces données peuvent être documentées rapidement par les grandes entreprises, comme ayant une énorme valeur dans le marketing. En précisant ceci, les organisations commerciales s'intéressent aux règles d'associations qui identifient les schémas d'achat, de telle sorte que l'apparition d'un produit dans un panier d'épicerie indique la présence d'un ou plusieurs autres produits. L'objectif de cette analyse aidera à recommander une combinaison d'articles pour des promotions, des ventes spéciales et concevoir un aménagement plus stratégique du magasin. L'analyse conduira également les gestionnaires vers une prise de décision stratégique, réelle et efficace.

2- Développement d'un système de transport intelligent

L'internet des objets, désigne le réseau collectif d'appareils connectés et la technologie qui facilite la communication entre les appareils et le cloud (qui est un serveur) ainsi qu'entre les appareils eux-mêmes. L'internet des objets et l'intelligence artificielle permettent la mise en place d'une nouvelle classe de systèmes de transport intelligent (ITS) pour la route, l'air, le train, et la mer.

D'après les auteurs de [28], ces solutions connectent les véhicules, les feux de circulation, les postes de péage et d'autres infrastructures pour contribuer à réduire les embouteillages, prévenir les accidents, réduire les émissions et rendre le transport plus efficace. Parmi les exemples, citons la gestion de flotte, la gestion intelligente de trafic, la recharge des véhicules électriques, etc.

Les conditions suivantes s'appliquent à l'application d'ITS :

3- Les informations collectées sur les routes et la circulation doivent être précises, complètes et en temps réel.

4- L'échange des informations entre l'installation de gestion de trafic et de gestion de route doit être efficace et en temps réel.
5- Les centres de gestion de trafic et de gestion des péages doivent être équipés de système informatique d'auto-apprentissage.
6- Les données enregistrées à partir du web

L'utilisation massive d'internet a achevé l'extraction automatique des connaissances à partir des fichiers web [29]. Les experts sont intéressés à toutes les techniques qui permettent d'étudier les informations d'un utilisateur sur internet. Ces techniques aident à améliorer la structure des sites web pour des fins promotionnelles.

La fouille de contenu sur le web est un apprentissage automatique qui permet l'obtention d'informations importantes.

L'exploitation du web s'articule autour de trois classes. Chacune de ces classes est créée pour une partie du web à exploiter. Les classes sont introduites comme suit : 1) la fouille de contenu ; 2) la fouille de la structure de web ; 3) la fouille de l'usage de web.

Des algorithmes tels qu'Apriori [20] et FP-growth [21] sont utilisés pour ce genre d'apprentissage et facilitent la tâche aux experts pour obtenir l'information.

3.4. Les règles d'association et les relations causales

Malgré l'avantage des règles d'association, prendre des décisions fondées sur les systèmes de l'intelligence artificielle mènent peut mener à de mauvais résultats, notamment lorsqu'il s'agit du domaine médical ou du domaine financier. En fait, les experts ignorent la causalité et parfois ils la confondent avec la corrélation, ce qui entraine une ambiguïté lors de l'interprétation des résultats générés par des algorithmes d'extraction des règles d'association.

Dans notre travail nous appliquerons une approche basée sur les règles d'association et des mesures statistiques, appelées « rapport de côtes et l'intervalle de confiance », pour extraire les relations causales. Pour savoir si le résultat de rapport de côtes est significatif nous calculons l'intervalle de confiance. Avant de passer aux calculs statistiques, il est nécessaire d'extraire l'ensemble des variables contrôlées.

3.4.1 Définition

3.4.1.1 Ensemble des variables contrôlées

L'ensemble des variables contrôlées est un ensemble qui exclut les variables suivantes :

- Les variables conséquentes : Ensemble des variables qui sont du côté droit d'une règle d'association.
- Les variables antécédentes : Ensemble des variables qui sont au côté gauche d'une règle d'association.
- Les variables non pertinentes (redondantes) : Les variables redondantes peuvent être extraites en analysant la corrélation entre les variables. Ces variables sont détectées en utilisant des méthodes statistiques comme le coefficient de corrélation. Plus le coefficient de corrélation est élevé, plus les variables sont fortement corrélées et l'une d'entre elle peut être supprimée.

3.4.1.2 Rapport de côtes (Oddratio)

Le rapport de côtes est le rapport de la probabilité que l'événement se produise dans un groupe exposé à une cause par la probabilité que l'événement se produise dans un groupe non exposé à la même cause (une cause peut être un test médical, une activité physique, un phénomène naturel, etc). Plus le rapport des côtes est élevé, plus la probabilité que l'événement se produise avec l'effet est élevée.

Les rapports de côtes inférieurs à 1 (RC < 1) signifient que l'événement a moins de chances de se produire avec la cause.

Les rapports de côtes égaux à 1 (RC = 1) signifient que la cause et l'effet n'ont aucune association causale.

Les rapports de côtes supérieurs à 1 (RC > 1) signifient que l'événement a plus de chances de se produire avec la cause.

Tableau 10 illustre la matrice de confusion pour calculer le RC

	Résultat/effet(présent)	**Résultat/effet(absent)**
Intervention / cause (présente)	a	b
Intervention / cause (absente)	c	d

Tableau 10 : matrice de confusion pour le calcul de RC

A partir de tableau 10 :

- a : nombre de cas exposés à une cause (intervention, évènement).
- c : nombre de cas non-exposés à une cause (intervention, évènement).
- b : nombre de cas contrôlés exposés a une cause (intervention, évènement).
- d : nombre de cas contrôlés non-exposés a une cause (intervention, évènement).

Le RC est calculé comme suit :

$$\text{RC } (p \rightarrow z) = \frac{a*d}{b*c} = \frac{support(p,z).support(\neg p,\ \neg z)}{support(p,\neg z).support(\neg p,\ z)}$$

Notant que :

- p et z sont des items de la règle d'association « p → z »
- Support (p, z) : Indique la présente de la cause et de l'effet dans une règle d'association
- Support (¬p, ¬z) : Indique l'absence de la cause et de l'effet.
- Support (p, ¬z) : Indique la présence de la cause et l'absence de l'effet.
- Support (¬p, z) : Indique l'absence de la cause et la présence de l'effet

Exemple :

Groupe des malades	Nombre de patients atteints de cancer du poumon (649) **(Effet présent)**	Nombre de patients atteints d'autres maladies (649) **(Effet absent)**
Nombre de fumeurs (cause présente)	647 (a)	622 (b)
Nombre de non-fumeurs (cause absente)	2 (c)	27 (d)

Tableau 11 : Nombre de fumeurs et non-fumeurs atteints de cancer du poumon

Calcule de RC en utilisant les données de tableau 11, dans cet exemple l'intervention ou la case c'est fumer :

$$\text{RC } (p \rightarrow z) = \frac{a*d}{b*c} = \frac{27*647}{2*622} = \frac{17469}{1244} = 14.04$$

Vue que le RC supérieure à 1 ceci indique que fumer est la cause de cancer du poumon et d'autres maladies

3.4.1.3 Intervalle de confiance :

Le but de prélever un échantillon à partir des données d'une population et faire des calculs statistiques est de faire des estimations de la moyenne de cette population. Savoir si l'estimation statistique est bonne et significative est toujours un problème dans le domaine de la recherche scientifique. L'intervalle de confiance résout ce problème car il fournit une plage de valeurs susceptibles de contenir le paramètre de population d'intérêt.

La formule ci-dessous montre le calcul d'intervalle de confiance

$$\text{IC}= exp\left(ln\ \omega \pm z'\sqrt{\frac{1}{supp(pz)}+\frac{1}{supp(p\neg z)}+\frac{1}{supp(\neg pz)}+\frac{1}{supp(\neg p\neg z)}}\right) = [\omega-, \omega+]$$

Où z' est un écart normal standard correspondant au niveau de confiance choisi par l'utilisateur (z = 1,96 pour une confiance de 95 %). ω− et ω+ représentent la borne inférieure et la borne supérieure, respectivement, des rapports de côtes. Si ω− > 1 et le rapport de côtes > 1, alors le rapport de côte est statistiquement significatif. Nous pouvons conclure qu'une règle est causale. Si par contre ω− < 1, alors le rapport de côtes n'a pas de signification statistique.

Notre travail ne se limite pas juste aux mesures statistiques (rapport de côtes et intervalle de confiance) pour savoir si une règle est une relation causale ou non-causale, mais nous voulons savoir si ces mesures mènent aux résultats qui conviennent à la notion de causalité de Pearl et Mackenzie [4].

Dans le prochain chapitre, nous introduirons notre algorithme avec des exemples explicatifs qui montrent le fonctionnement de notre système.

Chapitre 4 : Méthodologie

4.1. Introduction

Nous proposons, dans ce chapitre, un système qui permettra de générer des règles d'association, les règles causales ainsi que des paramètres statistiques comme le RC et l'intervalle de confiance. L'objectif est d'aider l'utilisateur à distinguer entre les règles d'association et les règles causales en se basant sur des paramètres statistiques générés. Pour cela, nous avons besoin de faire, en premier lieu, une architecture du système illustré sur la figure 9. Ensuite, nous citerons le pseudocode avec une explication de son fonctionnement.

4.2. Architecture de système

Nous avons développé un schéma pour bien montrer le fonctionnement de notre algorithme, qui est basé sur cinq processus :

- Processus d'extraction des données
- Processus de traitement des données
- Processus de génération des règles d'association
- Processus de génération des règles causales

La figure 9 représente l'architecture de notre système

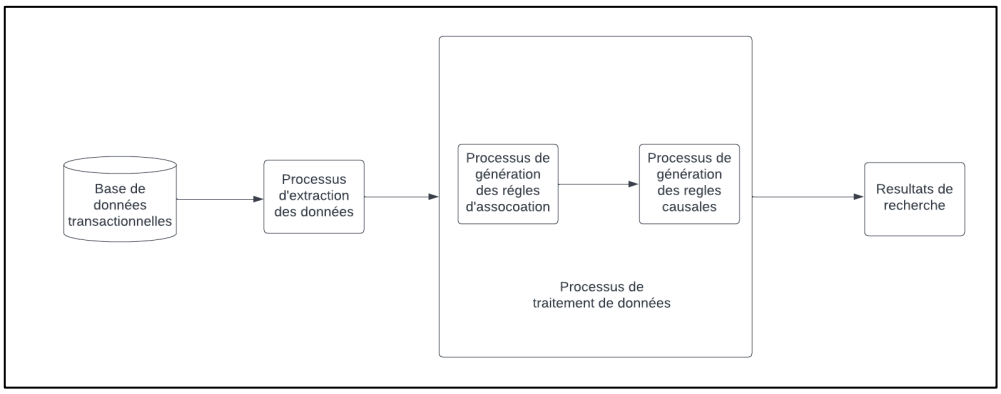


Figure 9 : Architecture de système implémenté

4.3. Description de l'architecture

- **Base de données transactionnelles :**

Une base de données transactionnelles est une base de données qui prend en charge les transactions ACID (Atomicité, cohérence, isolation et durabilité). Les bases de données transactionnelles alimentent une variété d'applications, des applications les plus simples aux services financiers critiques et aux applications de soins de santé.

Les transactions sont une fonctionnalité essentielle de la base de données. Selon la façon dont les données sont modélisées dans la base de données, des transactions à enregistrement unique peuvent être suffisantes ou des transactions à plusieurs enregistrements peuvent être nécessaires. Par exemple, les applications qui utilisent des bases de données relationnelles s'appuieront fréquemment sur des transactions multi-enregistrements, car les données associées sont généralement réparties sur plusieurs tables.

- **Processus d'extraction de données :**

L'extraction de données est le processus de collecte ou de récupération des types de données disparates à partir de diverses sources, dont plusieurs peuvent être mal organisées ou complètement non structurées. L'extraction de données permet de consolider, de traiter et d'affiner les données afin qu'elles puissent être stockées dans un emplacement centralisé dans le but d'être transformées. Ces emplacements peuvent être sur un site, basés sur le cloud ou un hybride des deux.

Dans notre cas, les données étaient déjà extraites d'une base de données et enregistrées sous forme d'un fichier « csv » pour l'utiliser dans l'implémentation.

- **Processus de traitement des données :**

La phase de traitement de données se constitue de deux étapes. Dans la première étape, nous avons utilisé l'algorithme « Apriori » pour générer les règles d'association, après avoir obtenu ces règles sous forme d'un tableau. À l'étape suivante, nous avons traité chaque règle obtenue pour savoir si elle est causale ou non causale, en utilisant des paramètres statistiques (rapport de côtes et intervalle de confiance).

- **Résultat de recherche :**

Dans cette étape finale, le résultat de notre algorithme sera généré sous forme d'un tableau qui contient toutes les règles avec leurs paramètres ; les règles causales seront distinguées à l'aide de paramètres statistiques générés.

4.4. Implémentation

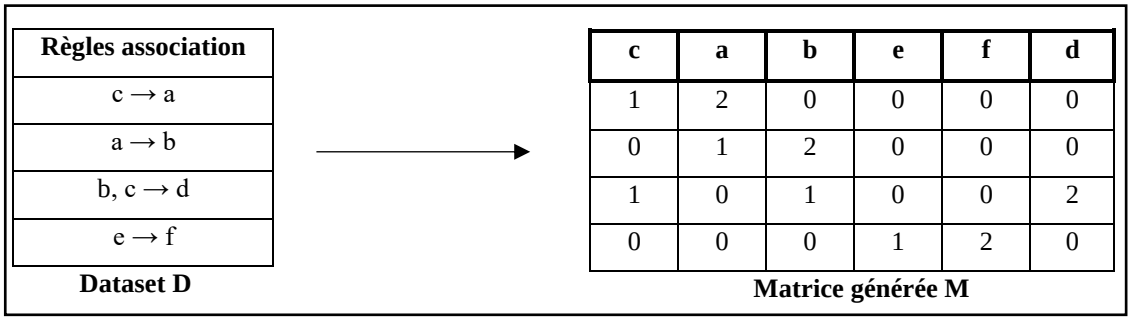
Dans cette partie, nous citerons les différents algorithmes implémentés dans notre système ainsi que des exemples de déroulement de chacun.

```
Algorithme 1 : Binariser les règles d'association
    Input: Dataset D
    Output: Matrice M
    L : Liste des variables
    Pour chaque règle r dans D :
        Pour chaque item i dans r :
            Si L ne contient pas item i :
            Ajouter i a la liste (
            L)
        Fin Pour
    Fin Pour
    k=0 ;
    j=0 ;
    Pour chaque règle r dans D :
        Pour chaque item i dans L :
            Si i ∈ antécédent(r) :
              M(k,j)=1
            Sinon si i ∈ conséquent(r) :
                M(k,j)=2
            Sinon M(k,j)=0
         j=j+1
        Fin pour
        k=k+1
Fin Pour
```

Figure 10 : Pseudo-algorithme pour la binarisation des règles d'association

Dans l'algorithme illustré dans la figure 10, nous avons numérisé les règles d'associations obtenues. L'algorithme a comme entrée un dataset D contenant des règles d'association. Pour chaque règle r dans D, l'algorithme remplit une liste L avec les items

de r, sans duplication. Après avoir fini avec la tâche précédente, l'algorithme parcourt chaque règle r dans D et attribue une étiquette à chaque item comme suit : la valeur 1 si l'item est un antécédent ; 2 si l'item est un conséquent ; 0 si l'item est absent de la règle r. À la fin, une matrice M sera générée comme le montre la figure 11.



Règles association
c → a
a → b
b, c → d
e → f

Dataset D

c	a	b	e	f	d
1	2	0	0	0	0
0	1	2	0	0	0
1	0	1	0	0	2
0	0	0	1	2	0

Matrice générée M

Figure 11 : Exemple de binarisation de règles d'association

Exemple :

Prenons l'Exemple de la règle « c → a » :

- « c » existe dans l'antécédent de la règle, alors l'algorithme lui attribue le numéro 1 (voir la matrice générée M, ligne 1)
- « a » existe dans le conséquent de la règle, alors l'algorithme lui attribue le numéro 2
- Les items « b, e, f, d » seront étiquetés par 0, car ils n'existent ni dans l'antécédent ni dans le conséquent de la règle « c → a » La même logique s'applique pour les règles restantes.

Algorithme 2 : Génération d'une liste fair-dataset, calcul de rapport de côtes et calcul d'intervalle de confiance.

Input : Matrice B, règles p->z

Output: Matrice C

Fairdataset=∅

controlvariable = ensemble des variables contrôlées

Pour chaque règle dans R

Controlvariable= x\(LHS(r), RHS(r), I)

Pour k = 1 à nbrlignes(B)

Pour i=2 à nbrlignes(B)

Si controlvariable(k) = controlvariable(i)

Fairdataset= Fairdataset+(B(k),B(i))

Fin Pour

Fin Pour

Fin Pour

Pour chaque ligne dans FariDataset :

SupportPZ= card(FD ∋ pz) //cardinalité

support¬P¬Z= card (FD ∉ pz)

supportP¬Z= card (FD ∋ p ∧ FD ∉ z)

support¬PZ= card (FD ∉ p ∧ FD ∋ z)

RC = (supportPZ * support¬P¬Z) / (supportZ*supportP)

Calcul d'intervalles de confiance à 95% [ω−, ω+]

IC= $exp\left(ln\ \omega \pm z' \sqrt{\frac{1}{supp(pz)} + \frac{1}{supp(p\neg z)} + \frac{1}{supp(\neg pz)} + \frac{1}{supp(\neg p\neg z)}}\right)$ = [ω−, ω+]

Fin Pour;

Figure 12 : Pseudo-code pour la génération d'une liste « fair-dataset »

L'ensemble des variables contrôlées ne contient pas les variables suivantes :

- les variables antécédentes
- les variables conséquentes,
- les variables non pertinentes.

L'algorithme illustré a la figure 12 permet de générer une liste, nous l'appelons fair-dataset, un rapport de côtes et un intervalle de confiance de chaque règle.

L'algorithme fonctionne comme suit : fixer la première ligne de la matrice et comparer la ligne avec les autres lignes, si deux lignes sont identiques au niveau des variables contrôlées, elles seront ajoutées à la liste fair-dataset qui est mentionnée dans le pseudo algorithme. Après avoir terminé la comparaison entre les lignes de la matrice, l'algorithme passe au calcul des variables suivantes : support (PZ), support (¬P¬Z), support(¬PZ), support(P¬Z), odd ratio et l'intervalle de confiance.

Exemple de déroulement d'algorithme :

Le tableau 12 représente un échantillon des règles générées par l'algorithme apriori, les variables « graduation, emploi, pays, salaire » sont un ensemble des variables qui forme une règle d'association.

Graduation = 9th	Emploi = tech_support	Graduation = 10ieme	Graduation = Master	Emploi = protection_serv	Pays = USA	Salaire = <=50000	Salaire = >50000
1	0	0	0	0	0	2	0
0	1	1	0	0	0	2	0
0	0	0	1	0	0	0	2
0	0	0	0	1	1	0	2

Tableau 12 : Présentation des données dans une matrice

Après avoir exécuté l'algorithme 1, une matrice sera générée (tableau 12) où les variables qui représentent l'antécédent sont numérotées par 1, les variables conséquentes sont numérotées par 2, les variables qui n'appartiennent pas à une règle sont distinguées par le zéro.

Cette matrice sera utilisée dans l'algorithme 2 pour générer la variable contrôlée de chaque règle, ensuite l'algorithme procède aux générations de fair-dataset, puis le calcul de RC de chaque règle dans la liste « fair-dataset ».

Prenons un exemple sur la règle suivante : « Graduation = 9th -> Salaire = <= 50k USD ». Nous démonterons comment l'algorithme extrait les variables contrôlées à partir de la matrice générée dans l'algorithme 2.

Admettons que notre matrice contient les données suivantes, nous cherchons une relation de la forme suivante : « 1 → 2 ».

	Gradua-tion= 9ieme	Emploi = protect_ser-vice	Pays = USA	Gradua-tion= mas-ter	Salaire = <=50k	Salaire = <=50k
#1	1	0	0	0	2	1
#2	0	2	0	0	2	0
#3	0	1	0	1	2	0
#4	1	0	0	0	2	1
#5	0	0	0	0	0	1

Tableau 13 : Exemple des variables contrôlées

Nous constatons que les variables : « pays = USA » et « Emploi = protect _service » sont des variables contrôlées et seront ajoutées dans la liste des variable contrôlées de la règle « Graduation = 9th → Salaire = <= 50k USD ».

S'il existe un antécédent et un conséquent avec les variables graduation et salaire, respectivement, ces variables seront ignorées par l'algorithme, car nous focalisons sur deux variables précises, colorées en gris dans le tableau 13, qui sont Graduation = 9th → Salaire = <= 50000 USD.

La tâche suivante de l'algorithme est la construction d'une liste « Liste Identique » qui contient les règles d'association avec des variables contrôlées identiques. Celles-ci sont étiquetées avec 0, à condition qu'elles n'appartiennent pas aux items de l'antécédent ni de conséquent. Par exemple, nous remarquons dans le tableau, qu'il y a deux variables de même type qui sont « graduation » et « salaire », nous sommes intéressés par la « graduation = 9ieme » et « salaire = <=50000 USD ». Nous voyons que les variables « pays = USA » et « emploi = protect_service » ont l'indice 0, elles n'appartiennent ni à l'antécédent ni au conséquent de la règle.

Donnant une règle « Graduation = 9th → Salaire = <= 50k USD », avec une liste des variables contrôlées L ou L= {pays= « USA », Emploi= « protect_service »}. Toutes les règles qui ont des variables contrôlées identiques seront ajoutées à la liste « fair-dataset ».

Les enregistrements (#1, #4) et (#2, #3) forment deux paires identiques, donc la liste « fair-dataset » (voir tableau 14) de la règle « Graduation = 9th → Salaire = <= 50k USD » contient les enregistrements suivants (#1, #2, #3, #4).

La liste « fair-dataset » sera représentée comme dans le tableau 14.

Id	**Graduation =9th**	**Emploi=protect_service**	**Pays=usa**	**Salaire=<=50k**
1	1	0	0	2
2	0	1	0	2
3	1	2	0	2
4	0	0	0	0

Tableau 14 : La liste « fair-dataset »

L'étape suivante est le calcul de RC et l'intervalle de confiance de la règle « Graduation= 9th → Salaire = <= 50k USD », la liste fair-dataset (tableau 14).

En appliquant la méthode pour le calcul de RC sur la règle « Graduation= 9th → Salaire = <= 50k USD » :

$$\text{RC (p}\rightarrow\text{z)} = \frac{support(pz).support(\neg p\ \neg z)}{support(p\neg z).support(\neg pz)} = \frac{2 \text{ x } 1}{1} = 2.$$

Ou p = « Graduation= 9th » et z = « Salaire = <= 50000 USD ». Si un support de la règle n'existe pas dans la liste « fair-dataset », nous lui attribuons «1 » pour éviter le problème mathématique de la forme indéterminée, comme dans le cas de $support(p\neg z) = 0$, car il n'existe pas un enregistrement de cette forme où la graduation existe et le salaire n'existe pas.

Après avoir calculé le RC, notre objectif était de savoir si le résultat représente une signification statistique. Pour cela, l'algorithme calcule l'intervalle de confiance.

En appliquant la formule d'intervalle de confiance sur la règle « Graduation= 9th → Salaire = <= 50k USD » :

$$\text{IC } (\omega-) = exp\left(ln\ (2)\ - 1.96\sqrt{\frac{1}{2}+\frac{1}{1}+\frac{1}{2}}\right) = 0{,}125.$$

$$\text{IC } (\omega+) = exp\left(ln\ (2) + 1.96\sqrt{\frac{1}{2}+\frac{1}{1}+\frac{1}{2}}\right) = 32.$$

L’intervalle de confiance à 95% est [0,128 ; 32] avec un RC = 2, vue que la borne inferieure d’IC est inférieure à 1 et RC supérieure à 1, ceci indique que le RC de la règle « Graduation= 9th →Salaire = <= 50k USD » ne représente aucune signification statistique. D’où la règle n’est pas causale, c’est-à-dire qu’une population avec une graduation = 9ieme n’est pas la cause d’avoir un salaire inférieur à 50000 USD.

La question qu’un utilisateur doit se poser : est ce que les mesures statistiques seules peuvent identifier qu’une règle est causale ou non ?

Dans le prochain chapitre nous interprétons les résultats obtenus et nous abordons cette problématique.

Chapitre 5 : Implémentation et expérimentation

5.1. Introduction

Afin de garantir une bonne recherche scientifique, nous devons prendre en considération les hypothèses et ses apports par rapport aux travaux qui sont faits. De plus, il est nécessaire d'effectuer un ensemble de tests sur l'approche suggérée, afin d'obtenir des résultats et voir s'ils conviennent aux résultats attendus. Dans le chapitre précédent, nous avons présenté notre approche de génération des règles d'association et les règles causales ainsi que l'architecture du système. Dans ce chapitre, nous parlerons d'abord des différents outils utilisés pour concevoir notre travail. Ensuite, nous présenterons l'interprétation des résultats obtenus après l'exécution de notre système.

5.2. Présentation des outils de travail

Toutes les expérimentations ont été exécutées et réalisées à l'aide d'un ordinateur muni d'un processeur de type Intel (R) Core (TM) i9-9900U CPU @ 3.60GHz 3.60 GHz avec une mémoire RAM de taille 16 Go.

5.3. Plateformes de développement

- **Rstudio**

  Rstudio est un environnement de développement gratuit, libre et multiplateforme pour R, un langage de programmation utilisé pour le traitement de données et l'analyse statistique. Il est disponible sous la licence libre AGPLv3 ou bien sous une licence commerciale, soumise à un abonnement annuel.

- **Python**

  Le langage de programmation Python est le plus approprié comme premier langage à apprendre pour les programmeurs débutants, car il dispose d'outils puissants qui reflètent la façon dont les gens pensent et la façon dont ils implémentent le code [30].

  De plus, il minimise les mots-clés supplémentaires qui sont nécessaires pour écrire un programme syntaxiquement correct. Cette approche semble plus productive que l'enseignement des langages C++ ou Java, qui ont beaucoup de termes et éléments liés aux spécificités d'un langage plutôt qu'à la concrétisation d'un algorithme.

Premièrement, nous avons appliqué l'algorithme Apriori sur notre dataset « Adult »[31] (voir tableau 15) en utilisant la librairie « arules » dans rstudio. Nous avons fixé le seuil de support minimal à 0.01 (le seuil est choisi au hasard). Après avoir obtenu les règles d'association sous forme d'un tableau, nous avons exporté le tableau qui contient les règles d'associations au langage python où nous avons implémenté une logique permettant de calculer le RC et l'intervalle de confiance de chaque règle.

Caractéristique de dataset « Adult »	
Caractéristiques des attributs	Catégoriels et nombres entier
Nombre d'instances	488842
Nombre d'attributs	14
Valeur manquantes (oui/non)	Oui
Nombre de clique sur le site web	2599305

Tableau 15 : les caractéristiques de dataset « Adult »

5.4. Interprétation et discussion

Dans cette partie, nous présentons d'abord les données générées par l'algorithme implémenté.

Le tableau 16 représente un échantillon de 167 règles avec des paramètres statistiques comme le RC, le support, la confiance et le lift. Le seuil de support minimum (min-supp).

Notons que TP (vrai positif) et FN (faux négatif) indiquent que les règles sont causales, alors que les règles avec un FP (faux positif) ou un TN (vrai négatif) sont des règles non causales.

RULE	**CAUSALITY**	**OR**	**SUP-PORT**	**CONF**	**Lift**	**FP/FN , TP/TN**
Graduation = Doctorate => Gende r= Male	[0.32685950091969657,39.908009431144464]	2.5427286356821 59	0.0100426891065999	0.791767554479419	1.183145632923 56	FP
Graduation = Doctorate => SkinColor = White	[0.23022520502351843,28.083236646732 84]	2.3771067415730336	0.0113325757808421	0.893462469733656	1.045874010533 42	FP
Graduation = Doctorate => country = United-States	[0.21524557680926 72,26.25204450003 2124]	1.2170212765957447	0.0100734006940819	0.794188861985472	0.886512976863523	FP
Graduation = 12th => CivicState = Private	[0.14191709959188017,10.436661909989313]	4.8027522935779 81	0.0102269586314917	0.769053117782 91	1.103328276706 44	FP
Graduation = 12th => Salary = <=50K	[1.07196657703129 92,21.517862671939 5]	1.0161919040479 76	0.0122846349927828	0.923787528868 36	1.216806056937	FN
Graduation = 12th => SkinColor = White	[0.11852695951214 68,8.712330005789 306]	3.5675105485232 07	0.0102883818064556	0.773672055427252	0.905649115500674	FP
Graduation = 12th => country = United-States	[0.79669036429470 3,15.975003695559 286]	1.9684290604792696	0.0112097294309143	0.842956120092379	0.940949407827492	FP
Graduation = 9th => Gender = Male	[0.54718554738109 1,7.0811683252312 41]	1.6591312772485063	0.0113632873683241	0.719844357976654	1.075670130338 59	FP

Graduation = 9th => CivicState = Private	[0.461423884790865,5.965700272303625]	7.22947416974169698	0.0118853843555173	0.752918287937743	1.080180312545486	TN
Graduation = 9th => Salary = <=50K	[2.7773216697500 26,18.818596830257174]	1.384930384930385	0.014956543103713	0.947470817120623	1.248001507939 51	FN
Graduation = 9th => SkinColor = White	[0.38532211374689473,4.977737074190648]	1.294591484464902 2	0.012376769755228 6	0.784046692607004	0.917793513013253	FP
Graduation = 9th => country = United-States	[0.36023864961903074,4.652380063666 46095]	470.6333333333333 33	0.012131077055373	0.768482490272374	0.857818250454534	TN
Graduation = Prof-school => Job = Prof-specialty	[155.27246564909, 1426.4971804145664]	34.464614154338264	0.013881637541844 5	0.784722222222222	6.171821323134 73	TN
Graduation = Prof-school => Salary = >50K	[19.365664083189216,61.335858336947844]	5.671878257244111	0.012991001504867 8	0.734375	3.049609026272 16	TN
Graduation = Prof-school => Gender = Male	[3.536724018554164,9.096045604980475]	3.965496728138013	0.014864083412672	0.840277777777778	1.255634911529 24	FP
Graduation = Prof-school => SkinColor = White	[2.4788235480487506,6.343801402585 3615]	3.703141853101840 7	0.015785755965725 9	0.892361111111111	1.044584776347 75	FP
Graduation = Prof-school => country = United-States	[2.315746722526153,5.921744140151606]	2.153891859774213	0.015417216915942 4	0.871527777777778	0.972842508475 222	TN
Country = Mexico => Gender = Male	[2.964928221645864,9.846376782320 58]	4.250963391136802	0.015263658978532 6	0.772939346811 82	1.155010466798 52	FP
Country = Mexico => CivicState = Private	[2.33534081134041,7.7379240173571535]	4.325746451297112	0.016983507877522 2	0.860031104199067	1.233850580887 64	FP
Country = Mexico => Salary = <=50K	[2.376249378958593 5,7.874629037929618]	3.125917767988252 4	0.018734068363993 7	0.948678071539658	1.249591694474 22	TP
Country = Mexico => SkinColor = White	[1.719116294694693,5.683944665279323]	3.993876211999708	0.018119836614354 6	0.917573872472784	1.074098463531 29	FP
Graduation = 7th-8th => Gende r= Male	[2.152978630467546 2,7.408827459338585]	1.515597147950089 2	0.014925831516231 1	0.752321981424149	1.124201745624 22	TN
Graduation = 7th-8th => CivicState = Private	[0.6948357566349836,3.305867167801 9875]	4.996282527881041	0.013021713092349 7	0.656346749226006	0.941633173314 592	TN
Graduation = 7th-8th => Salary = <=50K	[2.8045082275775384,8.900968395436 52]	3.959703254088686 7	0.018611222014065 9	0.938080495356037	1.235632646006 79	TP
Graduation = 7th-8th => SkinColor = White	[2.22511909285289,7.046476708056874]	2.617172999902315 2	0.016983507877522 2	0.856037151702786	1.002064484346 94	FP
Graduation = 7th-8th => country = United-States	[1.4138792022801325,4.844540113732128]	6.134989200863931	0.015325082153496 5	0.772445820433437	0.862242316048 445	TN
Job = Protective-serv => Gender = Male	[2.932234122459272,12.835977934514275]	3.587714731910463 4	0.017597739627161 3	0.882896764252696	1.319320860065 72	FP
Job = Protective-serv => Salary = <=50K	[1.6345270822736586,7.874875330705802]	2.797237706963973 6	0.013451675317097 1	0.674884437596302	0.888952757790 178	FP
Job = Protective-serv => SkinColor = White	[1.2754789638948674,6.134588661005949]	3.781900138696255	0.015939313903135 7	0.799691833590139	0.936107484668 123	FP
Job = Protective-serv => country = United-States	[1.8111561041203723,7.897038044667723]	3.678357487922705 4	0.018611222014065 9	0.933744221879815	1.042291587542 98	FP
Job = Tech-support => Gender = Male	[2.188862885959157 5,6.181435071036012]	4.987545787545788	0.017812720739535	0.625	0.933943322625 057	FP
Job = Tech-support => CivicState = Private	[3.120107018178870 4,7.972679410652079]	2.971677559912854	0.022603728386720 3	0.793103448275862	1.137832277904 05	TN

Job = Tech-support => Salary = <=50K	[1.77088024335848 87,4.986710735075 909]	3.615782 3129251 7	0.019808 9739258 622	0.69504 3103448 276	0.91550 5602402 076	FP
Job = Tech-support => SkinColor = White	[2.26648704204113 ,5.76834611976803 9]	3.534598 7345987 346	0.024753 5395104 573	0.86853 4482758 621	1.01669 3676053 47	FP
Job = Tech-support => country = United-States	[2.21586758561331 3,5.6381474668167 17]	5.200071 6845878 13	0.026104 8493596 634	0.91594 8275862 069	1.02242 6870426 63	FP
Graduation = 10th => Gender = Male	[3.47373143980547 2,7.7843512065993 85]	4.363848 4848484 85	0.019593 9928134 885	0.68381 5648445 874	1.02183 2094036 08	FP
Graduation = 10th => CivicState = Private	[2.91971552526699 3,6.5222702122574 14]	4.589630 3258145 37	0.021344 5532999 601	0.74490 8896034 298	1.06868 9573659 36	FP
Graduation = 10th => Salary = <=50K	[3.06995123068292 57,6.861576925751 523]	3.628330 7810107 197	0.026749 7926967 845	0.93354 7695605 573	1.22966 2075914 77	TP
Graduation = 10th => SkinColor = White	[2.43089382266346 ,5.41561467378512 7]	3.387225 4059216 81	0.023402 2296612 512	0.81672 0257234 727	0.95604 0706637 185	FP
Graduation = 10th => country = United-States	[2.27042746320340 47,5.053363798873 948]	3.523206 7510548 52	0.026043 4261846 995	0.90889 6034297 964	1.01455 4808802 74	FP
CivicState = Federal-gov => Gender= Male	[1.61793849583199 9,7.6720999238573 09]	2.602499 6189605 245	0.019808 9739258 622	0.67187 5	1.00398 9071821 94	FP
CivicState = Federal-gov => Salary = <=50K	[1.19639505572863 36,5.661177078807 594]	2.929465 9175522 915	0.018089 1250268 726	0.61354 1666666 667	0.80815 2516518 339	FP
CivicState = Federal-gov => SkinColor = White	[1.41741094975746 18,6.054539485227 588]	3.257107 0234113 71	0.022143 0545744 91	0.75104 1666666 667	0.87915 8315657 655	FP
CivicState = Federal-gov => country = United-States	[1.58371827108795 3,6.6986321719126 76]	6.210909 5396766 63	0.027210 4665090 139	0.92291 6666666 667	1.03020 5333676 15	TP
Job = Farming-fishing => Gender = Male	[4.34582393103919 35,8.876428940097 963]	4.659736 6597366 6	0.028531 0647707 38	0.93460 7645875 251	1.39659 2912223 22	TP
Job = Farming-fishing => Salary = <=50K	[3.25509732101331 65,6.670505854901 5765]	3.708010 3359173 13	0.026995 4853966 402	0.88430 5835010 06	1.16480 1063663 53	FP
Job = Farming-fishing => SkinColor = White	[2.59019453909842 76,5.308227024544 409]	3.378502 4906600 247	0.028101 1025459 906	0.92052 3138832 998	1.07755 0831303 61	FP
Job = Farming-fishing => country = United-States	[2.35014362296309 54,4.856843202205 957]	5.118019 5739781 23	0.026995 4853966 402	0.88430 5835010 06	0.98710 6009385 073	TN
SkinColor = Asian-Pac-Islander => Gende r= Male	[1.55687833046168 18,16.82477291071 0063]	4.274467 0542635 66	0.021283 1301249 962	0.66698 7487969 201	0.99668 5617061 274	FP
SkinColor = Asian-Pac-Islander => CivicState = Private	[1.30084344281314 75,14.04555536558 0677]	4.281234 9252291 36	0.021897 3618746 353	0.68623 6766121 27	0.98451 5127849 607	FP
SkinColor = Asian-Pac-Islander => Salary = <=50K	[1.30291990224815 6,14.067612639407 491]	3.920614 2256989 716	0.023432 9412487 331	0.73435 9961501 444	0.96729 3475179 956	FP
Graduation = Assoc-acdm => Gender = Male	[2.52242318343596 45,6.093829143218 929]	3.293468 5996260 766	0.019839 6855133 442	0.60543 5801312 09	0.90470 8358261 724	FP
Graduation = Assoc-acdm => CivicState = Private	[2.12195135365504 27,5.111773838754 218]	3.936506 3691122 74	0.022388 7472743 466	0.68322 3992502 343	0.98019 2827805 287	FP
Graduation = Assoc-acdm => Salary = <=50K	[2.55898359144974 4,6.0555614525384 63]	3.880256 5532626 883	0.024630 6931605 295	0.75164 0112464 855	0.99005 4761406 478	FP
Graduation = Assoc-acdm => SkinColor = White	[2.57429986410978 2,5.8487323598351 21]	3.870370 3703703 702	0.028101 1025459 906	0.85754 4517338 332	1.00382 8984363 44	FP
Graduation = Assoc-acdm => country = United-States	[2.57792960815560 47,5.810774179578 257]	30.86542 1273351 124	0.030158 7789072 817	0.92033 7394564 199	1.02732 6222297 05	FN

CivicState = Self-emp-inc => Salary = >50K	[17.1496091160189 8,55.550783923790 37]	5.353537 3608903 02	0.019102 6074137 772	0.55734 7670250 896	2.31447 4874510 83	TN
CivicState = Self-emp-inc => Gender = Male	[3.29381054753534 65,8.701278309978 676]	3.516375 9689922 482	0.030128 0673197 998	0.87903 2258064 516	1.31354 6092466 21	FP
CivicState = Self-emp-inc => SkinColor = White	[2.17961411012871 14,5.672976651163 263]	3.114120 8624354 69	0.031940 0509812 352	0.93189 9641577 061	1.09086 7997892 96	FP
CivicState = Self-emp-inc => country = United-States	[1.92131671947761 57,5.047449307833 295]	4.824712 6436781 61	0.030435 1831946 193	0.88799 2831541 219	0.99122 1617683 017	FP
Graduation = 11th => Gender = Male	[3.30635424195009 84,7.040338206573 5824]	4.440689 4750587 62	0.022818 7094990 94	0.63234 0425531 915	0.94491 2188882 271	FP
Graduation = 11th => CivicState = Private	[3.06122140095536 3,6.4417826844355 06]	4.934563 0215195 43	0.028346 7952458 463	0.78553 1914893 617	1.12696 9716287 06	FP
Graduation = 11th => Salary = <=50K	[3.41000862656096 6,7.1407186549862 93]	3.488868 6960472 51	0.034243 4200423 82	0.94893 6170212 766	1.24993 1660125 32	TP
Graduation = 11th => SkinColor = White	[2.40122634940865 08,5.069161756140 185]	3.442631 0380267 213	0.030005 2209698 719	0.83148 9361702 128	0.97332 9202846 67	FP
Graduation = 11th => country = United-States	[2.37793456012181 05,4.984034742897 985]	3.517171 7171717 174	0.032769 2638432 481	0.90808 5106382 979	1.01364 9610865 14	FP
CivicState = State-gov => Gender = Male	[2.02119813211746 1,6.1203781517020 25]	3.168386 4153396 164	0.024845 6742729 032	0.62326 6563944 53	0.93135 3032978 331	FP
CivicState = State-gov => Salary = <=50K	[1.85523085244782 32,5.411009882497 066]	3.113484 9362688 297	0.029022 4501704 493	0.72804 3143297 381	0.95897 3009259 952	FP
CivicState = State-gov => SkinColor = White	[1.84568372012252 14,5.252139541941 357]	3.602734 0823970 038	0.032615 7059058 383	0.81818 1818181 818	0.95775 1588359 871	FP
CivicState = State-gov => country = United-States	[2.15102439609818 9,6.0341913797022 74]	5.709954 2334096 11	0.037161 0208531 679	0.93220 3389830 508	1.04057 1634428 22	TP
Job = Handlers-cleaners => Gender = Male	[4.30343287241919 ,7.57617890512234 9]	4.911	0.037038 1745032 401	0.88029 1970802 92	1.31542 8492946 94	FP
Job = Handlers-cleaners => CivicState = Private	[3.70844090399300 57,6.503520380770 096]	4.566914 4981412 64	0.039095 8508645 312	0.92919 7080291 971	1.33308 0108009 64	FP
Job = Handlers-cleaners => Salary = <=50K	[3.45195522192500 85,6.041998430588 5555]	3.520588 2352941 176	0.039433 6783268 327	0.93722 6277372 263	1.23450 7476436 82	FP
Job = Handlers-cleaners => SkinColor = White	[2.66806074277180 3,4.6455244904262 83]	3.439208 6330935 25	0.034826 9402045 392	0.82773 7226277 372	0.96893 7008369 914	FP
Job = Handlers-cleaners => country = United-States	[2.60695208055297 3,4.5371589720345 41]	4.331896 5517241 38	0.036516 0775160 468	0.86788 3211678 832	0.96877 4263129 052	FP
Graduation = Assoc-voc => Gender = Male	[2.98985029056409 33,6.276343599564 991]	4.192902 3357594 79	0.027087 6201590 86	0.63820 5499276 411	0.95367 6423218 872	FP
Graduation= Assoc-voc => CivicState= Private	[2.92137850634075 62,6.017854228426 597]	4.646958 9533606 31	0.030865 1454193 667	0.72720 6946454 414	1.04329 3328494 1	FP
Graduation = Assoc-voc => Salary = <=50K	[3.24559528585484 2,6.6533950207322 09]	3.600877 1929824 563	0.031356 5308190 78	0.73878 4370477 569	0.97312 1273750 814	FP
Graduation = Assoc-voc => SkinColo r= White	[2.51509204890493 75,5.155404377580 018]	3.535876 6859344 892	0.037068 8860907 22	0.87337 1924746 744	1.02235 6314411 8	FP
Graduation = Assoc-voc => country = United-States	[2.47744932085820 76,5.046490288569 9106]	133.42	0.039587 2362642 425	0.93270 6222865 412	1.04113 2921587 96	TN
Job = Transport-moving => Graduation = HS-grad	[80.8104452956285 3,220.27964745001 756]	5.252402 7459954 24	0.025337 0596726 145	0.51659 3613024 421	1.60182 8838557 11	TN

Job = Transport-moving => Gender = Male	[4.044652813202843,6.820791766375207]	4.5164	0.0462823623353091	0.943644333124609	1.4100965181675 3	FP
Job = Transport-moving => CivicState = Private	[3.4851685363195837,5.8527640047905996]	4.2	0.0388808697521575	0.792736380713838	1.1373056614567 9	FP
Job = Transport-moving => Salary = <=50K	[3.2444183997769236,5.437029946942995]	3.2364705882352 94	0.0392186972144 59	0.799624295554164	1.0532591702078 9	TN
Job = Transport-moving => SkinColor= White	[2.50722643235586,4.1778204526464515]	3.1615827338129496	0.0417677589754614	0.851596743894803	0.9968666083534 18	FP
Job = Transport-moving => country = United-States	[2.449788216872735,4.080191632035852]	39.390495867768 59	0.0457909769355978	0.933625547902317	1.0421591177664 5	TN
Graduation = Masters => Salary = >50K	[25.8685852752095,59.98051877215106]	4.4576569184785795	0.0294524123951967	0.556587347649449	2.3113175139413	TN
Graduation = Masters => Gender = Male	[3.2526898510357607,6.109007041213127]	2.4969586374695862	0.0364546543410829	0.688914683691236	1.0294516299068 5	TN
Graduation = Masters => CivicState = Private	[1.7735510363216578,3.5154344642739708]	3.9853182022406 77	0.0274561592088695	0.518862449216483	0.7443902101223 96	TN
Graduation = Masters => SkinColor = White	[2.9638305197210326,5.358862818716711]	3.7977591946090 77	0.0472037099597678	0.892048752176436	1.0442191335784 1	FP
Graduation = Masters => country = United-States	[2.8312416030738845,5.09422257873674]	1	0.0468965940849482	0.886244921648288	0.9892705140140 52	TN
Job = Machine-op-inspct => Graduation = HS-grad	[43.02990824061435,100.72128793635316]	4.6274241591347 81	0.0314179539940 42	0.510989010989011	1.5844503558530 8	TN
Job = Machine-op-inspct => Gender = Male	[3.5787699348933812,5.983355940197416]	4.6608148354972 31	0.0445932250238015	0.725274725274725	1.0837847787824 8	FP
Job = Machine-op-inspct => CivicState = Private	[3.635047499979863,5.9760415595425975]	4.3306002074150 89	0.0587512668529836	0.955544455544456	1.3708795830535 3	FP
Job = Machine-op-inspct => Salary = <=50K	[3.381297123583667 3,5.546421231561895]	3.1444970414201 183	0.0538067012683886	0.875124875124875	1.1527079716400 1	TN
Job = Machine-op-inspct => SkinColor = White	[2.455613321231161,4.026636261503274]	3.0775555847914746	0.0494763674334326	0.804695304695305	0.9419644742660 27	FP
Job = Machine-op-inspct => country = United-States	[2.405843247082635,3.9368102593410046]	2.5082066105316074	0.0518104480820614	0.842657342657343	0.9406158976436 66	FP
CivicState = Local-gov => Gender = Male	[1.616003040053509,3.893000350362053]	2.7655535804977145	0.0386351770523018	0.601051122790253	0.8981562922980 01	FP
CivicState = Local-gov => Salary = <=50K	[1.8463955540042187,4.142279583601215]	3.3300393156862744	0.0453303031233684	0.705207835642618	0.9288945119886 44	FP
CivicState = Local-gov => SkinColor = White	[2.279688975653257,4.864330747062018]	3.5422482601344 71	0.0528239304689659	0.821786908743431	0.9619716542851 18	FP
CivicState= Local-gov => country= United-States	[2.431279459538398,5.160872267151041]	4.7342287758230 53	0.0600718651147078	0.934543717152413	1.0431840237984 1	TP
CivicState = Self-emp-not-inc => Gender = Male	[3.586423807863489 4,6.249379131570874]	2.8541873005392318	0.0657842203863518	0.842975206611 57	1.2596657045653 7	TN
CivicState = Self-emp-not-inc => Salary = <=50K	[2.161794882004340 2,3.7683432477212566]	3.3219047619047 62	0.0558029544547158	0.715072805981897	0.9418885774909 61	FP
CivicState = Self-emp-not-inc => SkinColor = White	[2.542004067672047 5,4.341083237239415]	3.2002242686460525	0.0719265378827432	0.921684376229831	1.0789101587007 3	FP
CivicState = Self-emp-not-inc => country = United-States	[2.427856799123918 4,4.218302897158819]	1.4982029685313727	0.0710359018457664	0.910271546635183	1.0160902238597 3	TN

SkinColor = Black => Gender = Male	[1.00664899098414 36,2.229786305871 908]	3.851933 6000367 868	0.048186 4807591 904	0.50224 0717029 449	0.75050 2982432 12	TN
SkinColor = Black => CivicState = Private	[2.81803583237045 7,5.2651539375961 525]	4.533871 6161886 89	0.066828 4143607 383	0.69654 2893725 992	0.99930 0897189 462	FP
SkinColor = Black => Salary = <=50K	[3.31617936242211 2,6.1986972312219 85]	3.019705 9813713 39	0.084057 6149381 162	0.87612 0358514 725	1.15401 9214951 37	TN
SkinColor = Black => country = United-States	[2.22963968811063 93,4.089729951684 1725]	32.34016 9632925 47	0.086975 2157489 021	0.90653 0089628 681	1.01191 3824079 52	TN
Job = Other-service => Gender = Female	[23.1373728991713 25,45.20334164315 835]	4.710149 3616666 21	0.055280 8574675 225	0.54628 2245827 011	1.65142 4770808 03	TN
Job = Other-service => CivicState = Private	[3.76746675613554 46,5.888706774406 982]	4.870970 8298524 09	0.084149 7497005 62	0.83156 2974203 338	1.19300 8547895 44	FP
Job = Other-service => Salary = <=50K	[3.92013733133368 1,6.0524300094356 79]	3.100254 4529262 086	0.096987 1932680 2	0.95842 1851289 833	1.26242 6128634 64	TN
Job = Other-service => SkinColor = White	[2.49227680711595 53,3.856545005533 0057]	3.505226 5722342 45	0.077485 3352169 774	0.76570 5614567 527	0.89632 3717138 813	FP
Job = Other-service => country = United-States	[2.83130903328612 37,4.339552192378 212]	4.278215 6418520 06	0.085286 0784373 944	0.84279 2109256 449	0.94076 6330802 168	FP
Job = Sales => Gende r= Male	[3.45756724837999 97,5.293643698980 2225]	4.568318 1167203 73	0.073308 5593194 312	0.65397 2602739 726	0.97723 7352813 594	FP
Job = Sales => CivicS- tate = Private	[3.72302711384218 07,5.605527377966 936]	3.393910 0795325 43	0.090353 4903719 173	0.80602 7397260 274	1.15637 3725863 23	TN
Job = Sales => Salary = <=50K	[2.75219971146971 34,4.185243381847 999]	3.346916 1740036 806	0.081907 8038143 792	0.73068 4931506 849	0.96245 2753025 668	FP
Job = Sales => SkinCo- lor = White	[2.74687846336502 2,4.0780282146465 18]	3.472882 8965308 285	0.099413 4086790 946	0.88684 9315068 493	1.03813 2749063 32	FP
Job = Sales => country = United-States	[2.85521472216247 ,4.22417113480057 8]	38.54530 5514157 974	0.103313 7802893 03	0.92164 3835616 438	1.02878 4536561 77	TN
Job = Adm-clerical => Gender = Female	[27.5415900313499 26,53.94534503957 813]	4.173565 8914728 68	0.077915 2974417 248	0.67294 4297082 228	2.03432 7291550 87	TN
Job = Adm-clerical => CivicState = Private	[3.29875366922117 9,5.2803737402369 24]	4.130590 7677589 095	0.087005 9273363 84	0.75145 8885941 645	1.07808 6569666 28	FP
Job = Adm-clerical => Salary = <=50K	[3.29864914893388 94,5.172353687935 996]	3.112797 1674253 92	0.100211 9099536 25	0.86551 7241379 31	1.14005 2867983 48	TN
Job = Adm-clerical => SkinColor = White	[2.48058778989191 7,3.9061331532046 806]	3.359149 6429164 59	0.094714 5357943 552	0.81803 7135278 515	0.95758 2224683 769	FP
Job = Adm-clerical => country = United-States	[2.70306822242940 2,4.1744733743213 85]	3.947844 5317861 377	0.105924 2652252 7	0.91485 4111405 836	1.02120 5509821 23	FP
Job = Exec-managerial => Gender = Male	[3.14236229685530 9,4.9597961580530 47]	4.320531 4477592 66	0.089278 5848100 488	0.71495 3271028 037	1.06836 1333544 93	FP
Job = Exec-managerial => CivicState = Private	[3.48558188669722 52,5.355488006843 17]	1.945101 6635859 52	0.082644 8819139 461	0.66182 9808165 273	0.94949 9488177 188	TN
Job = Exec-managerial => Salary = <=50K	[1.52623823497617 78,2.478918687123 5023]	3.630368 3507474 975	0.064432 9105371 457	0.51598 6227250 369	0.67965 3217860 003	TN
Job = Exec-managerial => SkinColor = White	[2.96068056760455 84,4.451535402474 202]	3.629141 0565160 61	0.111974 4479592 15	0.89670 4377766 847	1.04966 8940338 88	FP
Job = Exec-managerial => country = United-States	[2.96299294026991 02,4.445054400598 992]	5.035852 2452921 29	0.114707 7792451 09	0.91859 3212001 968	1.02537 9279259 38	TN

Job = Craft-repair => Gender = Male	[3.98417773396628 9,6.3651296527797 81]	3.221544 7154471 546	0.119068 8246675 47	0.94584 0448889 973	1.41337 8194415 16	FN
Job = Craft-repair => CivicState = Private	[2.51923302807122 7,4.1196468282139 636]	2.717351 1211122 72	0.098123 5220048 524	0.77945 8404488 9	1.11825 6305453 08	FP
Job = Craft-repair => Salary = <=50K	[2.12300738846717 96,3.478083569337 6457]	2.897673 3743088 885	0.097355 7323178 035	0.77335 9355940 473	1.01866 3187248 29	FP
Job = Craft-repair => SkinColor = White	[2.29863733691477 23,3.652821108112 8807]	2.882068 3956267 48	0.113448 6041583 49	0.90119 5413515 492	1.05492 6080654 22	FP
Job = Craft-repair => country = United-States	[2.30787730364585 3,3.5991160465717 52]	2.435101 3181616 987	0.113172 1998710 11	0.89899 9756038 058	1.00350 8092435 9	FP
Job = Prof-specialty => Gender = Male	[1.82301610151668 13,3.252696684784 923]	1.616434 4741845 84	0.080617 9171401 37	0.63405 7971014 493	0.94747 8733097 884	TN
Job = Prof-specialty => CivicState = Private	[1.18459943474284 61,2.205691082318 9737]	1.661832 4749254 54	0.071035 9018457 664	0.55869 5652173 913	0.80153 7236977 211	FP
Job = Prof-specialty => Salary = <=50K	[1.23330622261653 79,2.239255039886 0096]	3.350242 0695136 93	0.070053 1310463 438	0.55096 6183574 879	0.72572 8555962 041	TN
Job = Prof-specialty => SkinColor = White	[2.64393299866125 4,4.2452368989768 985]	3.174207 9401923 25	0.112128 0058966 25	0.88188 4057971 015	1.03232 0492220 1	FP
Job = Prof-specialty => country = United-States	[2.51191512585979 3,4.0111212133934 13]	3.741656 1792096 65	0.113417 8925708 67	0.89202 8985507 246	0.99572 6972818 013	FP
Graduation = Bachelors => Gender = Male	[2.98852352709768 8,4.6845844901257 97]	3.876798 1218108 782	0.114738 4908325 91	0.69766 5732959 851	1.04252 8404355 47	FP
Graduation = Bachelors => CivicState = Private	[3.15695847823941 33,4.760773314211 6295]	1.970068 0272108 844	0.109056 8471484 29	0.66311 8580765 64	0.95134 8436213 87	TN
Graduation = Bachelors => Salary = <=50K	[1.56062309378776 05,2.486934896252 67]	3.880620 9226467 847	0.096250 1151684 531	0.58524 7432306 256	0.77088 3561623 139	TN
Graduation = Bachelors => SkinColor = White	[3.19353527946574 5,4.7155322949190 275]	3.669116 4233909 714	0.143791 6525905 22	0.87432 3062558 357	1.02346 9702328 25	FP
Graduation = Bachelors => country = United-States	[3.02804385970210 2,4.4459116023906 54]	2.217605 2112011 506	0.146371 4259390 07	0.89000 9337068 161	0.99347 2541113 348	TN
Graduation = Some-college => Gender = Male	[1.81397168153786 01,2.711052726345 3634]	3.524402 4175544 95	0.137741 4698565 77	0.61514 1955835 962	0.91921 2355391 224	TN
Graduation = Some-college => CivicState = Private	[2.95110909250055 85,4.209065816112 867]	4.351528 9699570 82	0.156444 8266330 89	0.69866 9592648 471	1.00235 1983002 59	TN
Graduation = Some-college => Salary = <=50K	[3.64981195415347 5,5.1881589008515 5]	3.430519 8733510 895	0.181321 2124934 74	0.80976 5464271 019	1.06661 7042157 31	TN
Graduation = Some-college => SkinColor = White	[2.89591573030673 16,4.063815282432 365]	3.615369 6949412 97	0.190626 8235005 07	0.85132 3549581 676	0.99654 6811113 35	FP
Graduation = Some-college => country = United-States	[3.06292283991587 3,4.2674591278502 25]	5.781211 5322319 4	0.206996 0996283 9	0.92442 7376217 254	1.03189 1662564 62	TN
Salary = >50K => Gender = Male	[4.80082722083115 1,6.9618016318915 3]	3.457640 0185049 77	0.204600 5958047 97	0.84963 6525953 322	1.26961 9776115 93	TN
Salary = >50K => CivicState = Private	[2.86221496775998 44,4.176931024479 7895]	3.696412 9191675 5	0.152421 6086729 52	0.63295 4980232 113	0.90807 3982699 059	FP
Salary = >50K => SkinColor = White	[3.09929726892370 9,4.4085698412962 08]	3.577374 8330432 67	0.218574 3681090 88	0.90766 4838668 537	1.06249 9094473 91	FP
Salary = >50K => country = United-States	[3.00055852964926 65,4.265076174863 7676]	2.821657 9637895 243	0.220232 7938331 13	0.91455 1715342 431	1.02086 7960345 04	TN

Graduation = HS-grad => Gender = Male	[2.37229727847683 77,3.356136575652 4344]	2.886080 8106669 54	0.218390 0985841 96	0.67717 3602514 046	1.01190 6822921 52	FP
Graduation = HS-grad => CivicState = Private	[2.44672627030030 25,3.404329510336 967]	3.377061 9899288 072	0.238936 1506096 25	0.74088 1820778 973	1.06291 2097567 16	FP
Graduation = HS-grad => Salary = <=50K	[2.87902227938045 7,3.9612571828642 062]	2.580351 7283201 94	0.271060 4711157 52	0.84049 1381773 165	1.10708 8991986 89	TN
Graduation = HS-grad => SkinColor = White	[2.21336580668498 15,3.008185552399 5844]	2.690018 4464029 513	0.273455 9749393 45	0.84791 9245786 116	0.99256 1783219 791	FP
Graduation = HS-grad => country = United-States	[2.32205092579274 8,3.1162965297661 205]	3.449523 6391286 843	0.297963 8217499 46	0.92391 2008380 154	1.03131 6383437 31	TN
Gender = Female => CivicState = Private	[2.92200873602894 13,4.072271650042 646]	4.756445 1794921 69	0.238076 2261601 3	0.71971 0333302 386	1.03253 8251791 46	TN
Gender = Female => Salary = <=50K	[4.04913451676033 9,5.5873102392299 1]	2.923877 4184778 507	0.294585 5471269 31	0.89053 9411382 416	1.17301 1884062 41	TN
Gender = Female => SkinColor = White	[2.49273670217552 87,3.429587710095 307]	3.517498 0921083 37	0.265409 5390190 72	0.80233 9615634 574	0.93920 6939339 854	TN
Gender = Female => country = United-States	[3.01051903043149 1,4.1098537172217 88]	3.255900 9846214 34	0.297349 5900003 07	0.89889 5181505 895	1.00339 1361159 19	FP
Gender = Male => Ci-vicStat e= Private	[2.83035600124175 1,3.7454268003770 33]	3.866304 4170061 1	0.458953 9633303 65	0.68581 9183111 519	0.98391 6039006 617	TN
CivicState = Private => Gende r= Male	[3.34370744244720 3,4.4705794697219 5]	2.779559 1182364 73	0.458953 9633303 65	0.65844 2016214 311	0.98391 6039006 617	TN
Gender = Male => Sa-lary = <=50K	[2.42202851264196 85,3.189867027347 2507]	2.991020 8556289 315	0.464604 8954270 45	0.69426 3423588 802	0.91447 8613894 619	FP

Tableau 16 : Échantillon représentant les règles d'association générées par le système

Type de règle	Nombre des règles
Règles d'association	3497
Règles causales	336
Règles non causales	3161

Tableau 17 : Tableau représentant le nombre de chaque type de règle après l'exécution de l'algorithme

Règle causale / Règle non causale	Nombre des règles causales/ non causales
Vraie Positive (causale)	266
Vraie Négative (non causale)	846
Fausse Positive (non causale)	2895
Fausse Négative (causale)	70

Tableau 18 : Nombre de règles causales et non causales

Le tableau 17 donne les résultats après l'exécution de notre algorithme. Il existe 3947 règles d'association, 336 règles causales générées et 3161 règles non causales. En fait, le nombre de règles causales extraites est toujours inférieur au nombre de règles non causales.

Le tableau 18 illustre le nombre de règles causales et non causales. Si une règle est vraie positive ou fausse négative alors règle est causale, sinon, la règle est non causale

Afin de démontrer que les règles sont causales et statistiquement significatives selon l'approche [19], nous prenons un exemple de deux règles :

$$Country = Mexico \rightarrow Salary = <=50{,}000\ USD.$$

$$Graduation = 9th \rightarrow Salary = <=50{,}000\ USD.$$

Où les RC sont respectivement de 3.12 et 1.38. Les deux règles ont un RC supérieur à 1. Ceci indique que le pays ou la graduation d'un individu influence sur son salaire. De notre côté, nous voulons savoir si ces règles sont statistiquement significatives. L'intervalle de confiance permettra de le savoir. Nous remarquons que les deux règles ont un intervalle de confiance où la borne minimale supérieure est 1 et il n'y a aucun chevauchement avec le RC. Nous concluons donc que ces deux règles sont statistiquement significatives.

Certes l'algorithme proposé par [19] peut générer des règles causales qui ont une signification statistique, mais est-ce que ces règles causales conviennent à la notion de causalité définie par Pearl et Mackenzie ? Si la réponse est oui, le modèle de [19]est valide et considéré comme un modèle causal; sinon, cette approche n'aide pas à détecter les règles causales dans une base de données.

Pearl et Mackenzie, dans le livre « The book of why » [4], ont défini la causalité comme suit :

« Une variable X est une cause d'une variable Y; si Y dépend de la valeur X ».

Nous avons appliqué la notion de causalité sur un échantillon des règles et nous l'avons comparé avec les résultats générés par l'algorithme. Le tableau 19 montre le résultat obtenu.

RULE	FP/FN, TP/TN	Selon Notion de Pearl
Graduation = Doctorate => Gender = Male	FP	FP
Graduation = Doctorate => SkinColor = White	FP	FP
Graduation = Doctorate => country = United-States	FP	FP
Graduation = 12th => CivicState = Private	FP	FP
Graduation = 12th => Salary = <=50K	FN	FP
Graduation = 12th => SkinColor = White	FP	FP
Graduation = 12th => country = United-States	FP	FP
Graduation = 9th => Gender = Male	FP	FP
Graduation = 9th => CivicState = Private	TN	FP
Graduation = 9th => Salary = <=50K	FN	FP
Graduation = 9th => SkinColor = White	FP	FP
Graduation = 9th => country = United-States	TN	FP
Graduation = Prof-school => Job = Prof-specialty	TN	FP
Graduation = Prof-school => Salary = >50K	TN	FP
Graduation = Prof-school => Gender = Male	FP	FP

Graduation = Prof-school => SkinColor = White	FP	FP
Graduation = Prof-school => country = United-States	TN	FP
Country = Mexico => Gender = Male	FP	FP
Country = Mexico => CivicState = Private	FP	FP
Country = Mexico => Salary = <=50K	TP	TP
Country = Mexico => SkinColor = White	FP	FP
Graduation = 7th-8th => Gender = Male	TN	FP
Graduation = 7th-8th => CivicState = Private	TN	FP
Graduation = 7th-8th => Salary = <=50K	TP	FP
Graduation = 7th-8th => SkinColor = White	FP	FP
Graduation = 7th-8th => country = United-States	TN	FP
Job = Protective-serv => Gender = Male	FP	FP
Job = Protective-serv => Salary = <=50K	FP	FP
Job = Protective-serv => SkinColor = White	FP	FP
Job = Protective-serv => country = United-States	FP	FP
Job = Tech-support => Gender = Male	FP	FP
Job = Tech-support => CivicState = Private	TN	FP
Job = Tech-support => Salary = <=50K	FP	FP
Job = Tech-support => SkinColor = White	FP	FP
Job = Tech-support => country = United-States	FP	FP
Graduation = 10th => Gender = Male	FP	FP
Graduation = 10th => CivicState = Private	FP	FP
Graduation = 10th => Salary = <=50K	TP	FP
Graduation = 10th => SkinColor = White	FP	FP
Graduation = 10th => country = United-States	FP	FP
CivicState = Federal-gov => Gender = Male	FP	FP
CivicState = Federal-gov => Salary = <=50K	FP	FP
CivicState = Federal-gov => SkinColor = White	FP	FP
CivicState = Federal-gov => country = United-States	TP	FP
Job = Farming-fishing => Gender = Male	TP	FP
Job = Farming-fishing => Salary = <=50K	FP	FP
Job = Farming-fishing => SkinColor = White	FP	FP
Job = Farming-fishing => country = United-States	TN	FP
SkinColor = Asian-Pac-Islander => Gender = Male	FP	FP
SkinColor = Asian-Pac-Islander => CivicState = Private	FP	FP
SkinColor = Asian-Pac-Islander => Salary = <=50K	FP	FP
Graduation = Assoc-acdm => Gender = Male	FP	FP
Graduation = Assoc-acdm => CivicState = Private	FP	FP
Graduation = Assoc-acdm => Salary = <=50K	FP	FP
Graduation = Assoc-acdm => SkinColor = White	FP	FP
Graduation = Assoc-acdm => country = United-States	FN	FP
CivicState = Self-emp-inc => Salary = >50K	TN	FP
CivicState = Self-emp-inc => Gender = Male	FP	FP
CivicState = Self-emp-inc => SkinColor = White	FP	FP
CivicState = Self-emp-inc => country = United-States	FP	FP
Graduation = 11th => Gender = Male	FP	FP
Graduation = 11th => CivicState = Private	FP	FP
Graduation = 11th => Salary = <=50K	TP	FP
Graduation = 11th => SkinColor = White	FP	FP
Graduation = 11th => country = United-States	FP	FP
CivicState = State-gov => Gender = Male	FP	FP
CivicState = State-gov => Salary = <=50K	FP	FP

CivicState = State-gov => SkinColor = White	FP	FP
CivicState = State-gov => country = United-States	TP	FP
Job = Handlers-cleaners => Gender = Male	FP	FP
Job = Handlers-cleaners => CivicState = Private	FP	FP
Job = Handlers-cleaners => Salary = <=50K	FP	FP
Job = Handlers-cleaners => SkinColor = White	FP	FP
Job = Handlers-cleaners => country = United-States	FP	FP
Graduation = Assoc-voc => Gender = Male	FP	FP
Graduation = Assoc-voc => CivicState = Private	FP	FP
Graduation = Assoc-voc => Salary = <=50K	FP	FP
Graduation = Assoc-voc => SkinColor = White	FP	FP
Graduation = Assoc-voc => country = United-States	TN	FP
Job = Transport-moving => Graduation = HS-grad	TN	FP
Job = Transport-moving => Gender = Male	FP	FP
Job = Transport-moving => CivicState = Private	FP	FP
Job = Transport-moving => Salary = <=50K	TN	FP
Job = Transport-moving => SkinColor = White	FP	FP
Job = Transport-moving => country = United-States	TN	FP
Graduation = Masters => Salary = >50K	TN	FP
Graduation = Masters => Gender = Male	TN	FP
Graduation = Masters => CivicState = Private	TN	FP
Graduation = Masters => SkinColor = White	FP	FP
Graduation = Masters => country = United-States	TN	FP
Job = Machine-op-inspct => Graduation = HS-grad	TN	FP
Job = Machine-op-inspct => Gender = Male	FP	FP
Job = Machine-op-inspct => CivicState = Private	FP	FP
Job = Machine-op-inspct => Salary = <=50K	TN	FP
Job = Machine-op-inspct => SkinColor = White	FP	FP
Job = Machine-op-inspct => country = United-States	FP	FP
CivicState = Local-gov => Gender = Male	FP	FP
CivicState = Local-gov => Salary = <=50K	FP	FP
CivicState = Local-gov => SkinColor = White	FP	FP
CivicState = Local-gov => country = United-States	TP	FP
CivicState = Self-emp-not-inc => Gender = Male	TN	FP
CivicState = Self-emp-not-inc => Salary = <=50K	FP	FP
CivicState = Self-emp-not-inc => SkinColor = White	FP	FP
CivicState = Self-emp-not-inc => country = United-States	TN	FP
SkinColor = Black => Gender = Male	TN	FP
SkinColor = Black => CivicState = Private	FP	FP
SkinColor = Black => Salary = <=50K	TN	FP
SkinColor = Black => country = United-States	TN	FP
Job = Other-service => Gender = Female	TN	FP
Job = Other-service => CivicState = Private	FP	FP
Job = Other-service => Salary = <=50K	TN	FP
Job = Other-service => SkinColor = White	FP	FP
Job = Other-service => country = United-States	FP	FP
Job = Sales => Gender = Male	FP	FP
Job = Sales => CivicState = Private	TN	FP
Job = Sales => Salary = <=50K	FP	FP
Job = Sales => SkinColor = White	FP	FP
Job = Sales => country = United-States	TN	FP
Job = Adm-clerical => Gender = Female	TN	FP

Job = Adm-clerical => CivicState = Private	FP	FP
Job = Adm-clerical => Salary = <=50K	TN	FP
Job = Adm-clerical => SkinColor = White	FP	FP
Job = Adm-clerical => country = United-States	FP	FP
Job = Exec-managerial => Gender = Male	FP	FP
Job = Exec-managerial => CivicState = Private	TN	FP
Job = Exec-managerial => Salary = <=50K	TN	FP
Job = Exec-managerial => SkinColor = White	FP	FP
Job = Exec-managerial => country = United-States	TN	FP
Job = Craft-repair => Gender = Male	FN	FP
Job = Craft-repair => CivicState = Private	FP	FP
Job = Craft-repair => Salary = <=50K	FP	FP
Job = Craft-repair => SkinColor = White	FP	FP
Job = Craft-repair => country = United-States	FP	FP
Job = Prof-specialty => Gender = Male	TN	FP
Job = Prof-specialty => CivicState = Private	FP	FP
Job = Prof-specialty => Salary = <=50K	TN	FP
Job = Prof-specialty => SkinColor = White	FP	FP
Job = Prof-specialty => country = United-States	FP	FP
Graduation = Bachelors => Gender = Male	FP	FP
Graduation = Bachelors => CivicState = Private	TN	FP
Graduation = Bachelors => Salary = <=50K	TN	FP
Graduation = Bachelors => SkinColor = White	FP	FP
Graduation = Bachelors => country = United-States	TN	FP
Graduation = Some-college => Gender = Male	TN	FP
Graduation = Some-college => CivicState = Private	TN	FP
Graduation = Some-college => Salary = <=50K	TN	FP
Graduation = Some-college => SkinColor = White	FP	FP
Graduation = Some-college => country = United-States	TN	FP
Salary = >50K => Gender = Male	TN	FP
Salary = >50K => CivicState = Private	FP	FP
Salary = >50K => SkinColor = White	FP	FP
Salary = >50K => country = United-States	TN	FP
Graduation = HS-grad => Gender = Male	FP	FP
Graduation = HS-grad => CivicState = Private	FP	FP
Graduation = HS-grad => Salary = <=50K	TN	FP
Graduation = HS-grad => SkinColor = White	FP	FP
Graduation = HS-grad => country = United-States	TN	FP
Gender = Female => CivicState = Private	TN	FP
Gender = Female => Salary = <=50K	TN	FP
Gender = Female => SkinColor = White	TN	FP
Gender = Female => country = United-States	FP	FP
Gender = Male => CivicState = Private	TN	FP
CivicState = Private => Gender = Male	TN	FP
Gender = Male => Salary = <=50K	FP	FP

Tableau 19 : Tableau comparatif entre les règles basées sur l'approche [19] et la notion de Pearl et Mackenzie[4]

Dans le tableau 16, les règles causales sont des vrai-positives et faux-négatives.

Prenons l'exemple des règles :

Graduation = 12th → Salaire = <=50,000 USD.

Graduation = 9th → Salaire= <=50,000 USD.

En voyant ces deux règles, nous nous posons des questions : Est-ce que le salaire dépend vraiment de la graduation d'un individu ? Et si nous changeons la valeur de salaire Y (salaire = >50,000), la valeur X (graduation=9th) changera-t-elle ?

Dans la vie réelle, il existe plusieurs individus qui ont une graduation 9ieme ou 12ieme et peuvent toujours avoir un salaire supérieur à $50,000 après avoir finir leurs études. Ceci dépend de leurs expériences dans le domaine de travail. La graduation a un impact sur le salaire, mais elle n'est pas considérée comme une cause directe. Toutefois, si nous sélectionnons des variables externes reliées au salaire, comme l'expérience, cette dernière peut causer l'augmentation de salaire. Bien évidemment, une personne avec dix ans d'expérience est mieux payée qu'une personne qui vient de décrocher un emploi. Ainsi, d'après la notion de Pearl, les règles {« graduation=12th → salaire=<=50,000 », « graduation=9th → salaire=<=50,000 »} sont non-causales.

Prenons un autre exemple:

Job= Farming-Fish → Gender= Male.

Job= Craft-Pair → Gender= Male.

Cet exemple est presque similaire à l'exemple de paradoxe de Simpson mentionné dans le chapitre 2. Nous remarquons, à partir de ces deux règles, que le système relie l'emploi au sexe d'un individu. L'expert de données doit se poser des questions pour approuver que ces deux règles soient causales. Si le sexe était féminin, le type d'emploi serait changé ? La réponse à cette question est certainement non, car il existe des femmes et des hommes qui font l'élevage de poissons ou bien qui sont des artisans. Donc les deux règles sont non-causales.

Voyons une autre règle :

Country=Mexico → Salaire= <50,000 USD.

Il est possible de considérer cette règle comme une règle causale. Selon la définition de Pearl, la variable Y (salaire) dépend de la variable X (pays). Un individu qui habite au Mexique a moins de chance d'être payé plus de 50,000 USD qu'un individu qui habite aux États-Unis d'Amérique.

Ignorer les variables externes dans l'approche [7] diminue son efficacité à extraire les règles causales. Pour valider un modèle causal il faut toujours prendre en considération les variables

externes, le modèle proposé par Pearl et Mackenzie [6] qui est basé sur les graphes bayésiens, explique mieux le mécanisme de causalité dans un système, étant donné que les variables externes sont prises en considération. Une variable externe (exogène) est, par définition, une variable dont la valeur est entièrement causale et indépendante des autres variables du système.

Chapitre 6 : Conclusion

Avec le développement rapide de la technologie, l'explosion massive des données et surtout la progression des travaux de recherche en intelligence artificielle, rendent la compréhension de comportement de l'être humain importante pour le domaine médical et économique. Tirer profit de ces mégadonnées nécessite des outils afin de faciliter l'analyse de résultats pour un utilisateur, les outils peuvent être des règles d'association, la classification, des arbres de décision, etc.

Dans ce mémoire nous avons introduit les différentes méthodes pour fouiller les données et extraire les informations importantes, aussi appelées les règles d'association. Néanmoins, ces associations obtenues peuvent être inutiles. Il existe des algorithmes qui permettent de bien interpréter les règles d'association tout en basant sur des méthodes statistiques. Une simple analyse des résultats révèle qu'une règle d'association peut avoir plusieurs significations statistiques telles que la corrélation et la causalité.

Se concentrer sur l'extraction des règles d'association causales peut améliorer des divers domaines. Plusieurs chercheurs ont proposé des techniques pour savoir si une relation est causale ou non causale, des chercheurs comme Pearl et Mackenzie [4] se basent sur des approches bayésiennes ainsi que des méthodes statistiques pour concevoir un système causal et répondre aux requêtes causales. Notre travail est fondé sur l'approche [19] pour l'extraction des règles causales, tout en respectant la notion de la méthode [4]. Notre modèle peut générer des règles causales où l'antécédent (a) d'une règle peut influencer sur le résultant (b) ($a \rightarrow b$), toutefois cet antécédent ne représente pas une cause directe, d'où, ce système est moins efficace pour répondre aux requêtes causales.

Dans notre futur travail, nous œuvrons à améliorer l'approche [19] pour la rendre plus robuste et capable d'explorer les règles d'association causales, tout en considérant la notion de [4].

Bibliographies

[1] Wu Xindong, Zhu Xingquan, Wu Gong-Qing and Ding Wei. (2013). Data mining with big data. (Vol. 26, pp. 97-107).

[2] Mehta Ms Yesha and Buch Sanjay. (2015). Big Data Mining and semantic technologies: challenges and opportunities. (Vol. 3, pp. 4907-4913).

[3] Obermeyer Ziad, Powers Brian, Vogeli Christine and Mullainathan Sendhil. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. (Vol. 366, pp. 447-453).

[4] Pearl Judea and Mackenzie Dana. (2018). The book of why: the new science of cause and effect. Basic books.

[5] White Peter A. (1990). Ideas about causation in philosophy and psychology. (Vol. 108, pp. 3).

[6] Ribeiro Marco Tulio, Singh Sameer and Guestrin Carlos. (2016). " Why should i trust you?" Explaining the predictions of any classifier. (pp. 1135-1144).

[7] Bickel Peter J, Hammel Eugene A and O'Connell J William. (1975). Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. (Vol. 187, pp. 398-404).

[8] Murphy Kevin. (2001). An introduction to graphical models. (Vol. 96, pp. 1-19).

[9] Arroyo-Figueroa Gustavo and Sucar Luis Enrique. (2013). A temporal Bayesian network for diagnosis and prediction. (pp. 13–20).

[10] Nikovski Daniel. (2000). Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. (Vol. 12, pp. 509-516).

[11] Spirtes Peter, Glymour Clark N, Scheines Richard and Heckerman David. (2000). Causation, prediction, and search. MIT press.

[12] Dechter Rina. (2003). Constraint Processing.

[13] Glymour Clark, Zhang Kun and Spirtes Peter. (2019). Review of causal discovery methods based on graphical models. (Vol. 10, pp. 524).

[14] Le Thuc Duy, Hoang Tao, Li Jiuyong, Liu Lin, Liu Huawen and Hu Shu. (2016). A fast PC algorithm for high dimensional causal discovery with multi-core PCs. (Vol. 16, pp. 1483-1495).

[15] Bhattacharyya Arnab, Gayen Sutanu, Price Eric and Vinodchandran NV. (2021). Near-optimal learning of tree-structured distributions by Chow-Liu. (pp. 147-160).

[16] Cooper Gregory F and Herskovits Edward. (1992). A Bayesian method for the induction of probabilistic networks from data. (Vol. 9, pp. 309-347).

[17] Lachapelle Sébastien, Brouillard Philippe, Deleu Tristan and Lacoste-Julien Simon. (2019). Gradient-based neural dag learning.

[18] Kleinberg Jon and Tardos Eva. (2006). Algorithm design.

[19] Li Jiuyong, Le Thuc Duy, Liu Lin, Liu Jixue, Jin Zhou and Sun Bingyu. (2013). Mining causal association rules. (pp. 114-123): IEEE.

[20] Agrawal Rakesh, Imieliński Tomasz and Swami Arun. (1993). Mining association rules between sets of items in large databases. (Vol. 22, pp. 207-216).

[21] Han Jiawei, Kamber Micheline and Pei Jian. (2012). Outlier detection. (pp. 543-584).

[22] Rompré Louis and Biskri Ismail. (2018). Les « itemsets fréquents » comme descripteurs de documents textuels.

[23] Agrawal Rakesh and Srikant Ramakrishnan. (1994). Fast algorithms for mining association rules. (Vol. 1215, pp. 487-499): Santiago, Chile.

[24] Said Aiman Moyaid, Dominic PDD and Abdullah Azween B. (2009). A comparative study of fp-growth variations. (Vol. 9, pp. 266-272).

[25] Aggarwal, C Charu and S Philip. (2001). A new approach to online generation of association rules. (Vol. 13, pp. 527-540).

[26] Malik Kuldeep, Raheja Neeraj and Garg Puneet. (2011). Enhanced FP-growth algorithm. (Vol. 12, pp. 54-56).

[27] Shaukat Kamran , Zaheer Sana Zaheer and Nawaz Iqra (2015). Association Rule Mining: An Application.

[28] Hu Manjiang, Li Chongkang, Bian Yougang, Zhang Hui, Qin Zhaobo and Xu Biao. (2022). Fuel economy-oriented vehicle platoon control using economic model predictive control. (Vol. 23, pp. 20836-20849).

[29] Kamran Shaukat Sana Zaheer, Iqra Nawaz. (2015). Association Rule Mining: An Application Perspective.

[30] Bogdanchikov Andrey, Zhaparov Meirambek and Suliyev Rassim. (2013). Python to learn programming. (Vol. 423, pp. 012027): IOP Publishing.

[31] Dua, Dheeru, Graff and Casey. (2017). {UCI} Machine Learning Repository.