Approche de sélection des motifs pertinents basées sur l'aide à la décision multicritère

Chaigne, Aras Maicher, Nathan Pavageau, Hadrien

February 2024

Remerciements

Nous tenons à remercier Fatima Zahra El Mazouri pour son accompagnement et son soutien tout au long de notre projet. Son expertise et ses conseils ont été inestimables et ont fortement contribué à la bonne réalisation de ce papier de recherche et à la rédaction de ce rapport.

Un grand merci également à Marc Gelgon pour sa coordination les projets de recherche.

Nous sommes également très reconnaissants envers José Martinez pour son aide précieuse dans la structuration de nos rendus.

Enfin, un remerciement spécial à tous les anciens étudiants qui ont partagé leurs travaux. L'accessibilité à leurs précédents rendus a été une aide considérable et a enrichi notre projet.

Table des matières

1 Introduction

Le data mining - en français la fouille de données - consiste en l'extraction automatique ou semi-automatique d'un savoir à partir de vastes ensembles de données. Il s'agit d'un domaine interdisciplinaire, sous-domaine de la data science, qui intègre des concepts issus de la statistique, de l'apprentissage automatique, des systèmes de bases de données, de l'entreposage de données, du calcul intensif, ainsi que de la visualisation et de l'interaction homme-machine. Ce champ d'étude a émergé en réponse à l'expansion remarquable des données dans diverses activités humaines, motivée par la nécessité économique et scientifique d'extraire des informations utiles à partir des données collectées. En effet, cette technique a été utilisée notamment dans l'éducation pour comprendre comment les étudiants apprennent [1], dans la santé pour détecter des causes d'infections [2], ou encore dans l'agriculture [3] Le terme data mining est un peu erroné puisqu'il consiste en l'extraction des modèles et motifs d'un ensemble de données et non l'extraction des données. Les motifs sont recherchés pour extraire des informations utiles ou des connaissances cachées dans les données.

Généralement, la fouille de données a deux objectifs : la prédiction et la description. Le premier répond à la question "quoi", tandis que le second répond à la question "pourquoi". Autrement dit, pour la prédiction, le critère clé est l'exactitude du modèle dans la réalisation de prédictions futures ; la manière dont la décision de prédiction est prise peut ne pas être importante. Pour la description, le critère clé est la clarté et la simplicité du modèle décrivant les données, dans des termes compréhensibles par les êtres humains. Il existe parfois une dichotomie entre ces deux aspects de la fouille de données dans le sens où le modèle de prédiction le plus précis pour un problème peut ne pas être facilement compréhensible, et le modèle le plus facilement compréhensible peut ne pas être très précis dans ses prédictions. Il est crucial que les motifs, les règles et les modèles découverts soient valides non seulement dans les échantillons de données déjà examinés, mais qu'ils soient généralisables et restent valides dans de nouveaux échantillons de données futurs. Seules les règles et les modèles obtenus peuvent être considérés comme significatifs. Les motifs découverts devraient également être novateurs et non déjà connus des experts ; sinon, ils apporteraient très peu de nouvelles connaissances. Enfin, les découvertes devraient être utiles ainsi que compréhensibles.

1.1 Enoncé du problème

Un algorithme de data mining peut nous donner de grandes quantités de motifs en manipulant de grandes tables de données, surtout si elles contiennent beaucoup d'objets (attributs). Parmi ces motifs, certains sont plus pertinents que d'autres. Plus un jeu de données possède de variables, plus il contient de combinaisons d'attributs et donc de motifs possibles : c'est le phénomène de "**patterns explosion**". Le terme "explosion" semble alors provenir du fait que le nombre de combinaisons possibles d'attributs augmente de manière expo-

nentielle avec le nombre d'attributs d'un jeu de données. Par conséquent, le nombre de motifs, qui sont des combinaisons de variables, augmente de manière exponentielle également.

1.2 Objectifs

L'objectif principal de cette recherche est de découvrir des patterns dans un ensemble de données en utilisant des algorithmes appropriés. La détection de ces patterns est essentielle car elle permet de découvrir des tendances cachées, des corrélations et des relations qui ne sont pas immédiatement visibles en observant directement les données.

Pour atteindre cet objectif, il va être important de choisir un algorithme d'exploration de données efficace et capable de gérer l'ensemble de données impliqué. De plus, nous allons devoir touver un moyen de classer les différents patterns extraits afin de faciliter l'identification des patterns les plus importants.

1.3 Travail accompli

Ce document présente une approche méthodique et innovante pour la détection et la sélection de motifs pertinents dans le cadre du data mining, en utilisant des méthodes multicritères pour l'aide à la décision et en appliquant ces méthodes à un ensemble de données spécifique lié au diabète.

Le travail accompli a impliqué une série d'expérimentations méthodiques. Différents seuils de confiance et de support ont été testés pour extraire les règles d'association via l'algorithme de data mining Apriori. Ces règles ont par la suite été soumises à un processus de sélection utilisant ELECTRE II. Cette approche a permis de trier efficacement les motifs extraits, en mettant en évidence ceux qui offrent le meilleur équilibre entre le support, la confiance, et le lift, offrant ainsi une perspective plus nuancée et précise des associations significatives au sein des données. Les résultats ont par la suite été nuancés par une étude de robustesse.

1.4 Organisation du rapport

Dans un premier temps, nous présenterons un état de l'art à la fois des méthodes multicritères et des algorithmes de fouille de données. Dans un deuxième temps, nous présenterons notre proposition d'expérimentation. Enfin, nous exposerons nos résultats et conclurons vis-à-vis de la pertinence de notre approche. Le nettoyage et la préparation des données a été effectué par Nathan Maicher. Par la suite, Hadrien Pavageau s'est occupé de mener une étude comparative entre les algorithmes Apriori et FP-Growth. Enfin, Aras Chaigne s'est occupé de l'application de la méthode multicritère afin d'établir un classement des patterns extraits.

2 Etat de l'art

2.1 Méthodes multicritère

Les méthodes d'aide multicritère à la décision deviennent de plus en plus populaires pour établir des décisions, notamment dans les domaines du développement durable [4] [5], dans l'établissement de stratégies [6] ou encore dans le domaine de la santé [7].

Les familles de méthodes ELECTRE et PROMETHEE se basent sur la notion de surclassement qui se définit comme suit. Une action A surclasse B lorsque A est au moins aussi bonne que B sur un critère donné. Pour la plupart des méthodes ELECTRE, le surclassement repose sur les concepts de concordance et de discordance. La concordance mesure dans quelle mesure deux actions sont en accord sur l'ensemble des critères considérés. Il existe différentes façons de définir la concordance, mais généralement, elle implique la comparaison des performances relatives des actions sur chaque critère. Plus la concordance entre deux actions est élevée, plus elles sont similaires dans leurs performances globales par rapport à tous les critères. La discordance mesure les écarts ou les différences significatives entre les performances de deux alternatives sur un critère donné. Elle permet de détecter les critères pour lesquels une alternative peut être clairement préférable à une autre, même si elles sont relativement similaires sur d'autres critères. La discordance peut être exprimée numériquement en utilisant des mesures spécifiques, telles que des indices de discordance, qui évaluent l'intensité des différences entre les alternatives. Ensemble, la concordance et la discordance permettent de capturer la complexité des relations entre les alternatives dans un contexte multicritère. Ces informations sont souvent utilisées pour établir des ordres préférentiels plus nuancés et précis lors de la prise de décision. Les seuils de concordance et de discordance peuvent être fixés par les décideurs pour refléter leurs préférences et le contexte spécifique du problème.

Dans la sous-section suivante nous présenterons deux méthodes d'aide multicritère à la décision : ELECTRE et PROMETHEE, ainsi que quelques unes de leurs dérivées.

2.1.1 Méthodes ELECTRE

Dans le contexte des méthodes ELECTRE, les problèmes de décision sont généralement classés en trois catégories principales : Problème de type P.α : L'objectif est d'isoler le plus petit sous-ensemble A0 de l'ensemble d'actions A qui justifie l'élimination de toutes les actions appartenant à A sans A0. Ce type de problème concerne essentiellement la sélection d'un ensemble restreint d'options parmi un plus grand ensemble. Problème de type P.β : Cette catégorie de problème vise à attribuer chaque action à une catégorie prédéfinie appropriée en fonction de ce que l'on souhaite qu'elle devienne par la suite. Il s'agit essentiellement de classer les actions dans des catégories distinctes basées sur des critères définis. Problème de type P.γ : L'objectif est de construire un pré-ordre (partiel ou complet) aussi riche que possible sur un sous-ensemble A0 de l'ensemble

ELECTRE	I	IS	II	III	IV	A
possibilité de prendre en compte un seuil d'indifférence ou de préférence	non	oui	non	oui	oui	oui
surclassement basé sur la concordance et la discordance	oui	oui	oui	non	oui	non
nécessité d'une quantification de l'importance relative du critère (pondération)	oui	oui	oui	oui	non	oui
Résultat final	un noyau	un noyau avec des indices de consistence et de connexion	un préordre partiel	un préordre partiel	un préordre partiel	une affectation à des catégories prédéfinies

TABLE 1 – Comparaison des différentes méthodes ELECTRE

d'actions A qui semble être le plus satisfaisant. Ces catégorisations aident à déterminer quelle méthode ELECTRE est la plus appropriée pour un problème de décision donné, en fonction des caractéristiques et des exigences spécifiques du problème. Dans la section 4.2 intitulée "How to Choose among ELECTRE Methods" du document de Bernard ROY, ce dernier suggère pour un problème de type P.α d'utiliser ELECTRE I ou IS, avec une préférence pour ELECTRE I si une méthode simple est nécessaire. Dans le cas de P.β, aucun choix spécifique n'est mentionné, tandis que pour P.γ, ELECTRE II, III et IV sont en compétition. ELECTRE II est conseillé pour sa simplicité, et ELECTRE IV est recommandé si l'on souhaite éviter d'introduire des coefficients d'importance.

Le choix entre les différentes méthodes ELECTRE dépend également d'autres facteurs, notamment la complexité du problème de décision, les préférences des décideurs et les caractéristiques spécifiques du contexte. Si le nombre de critères est relativement petit et que le problème de décision n'est pas très complexe, ELECTRE I peut être une option appropriée en raison de sa simplicité. Si le problème de décision nécessite un classement plus nuancé des alternatives en tenant

compte à la fois de la concordance et de la discordance, les méthodes ELECTRE II ou ELECTRE III peuvent être plus appropriées. Si les décideurs sont à l'aise avec un certain niveau de subjectivité et sont prêts à fournir des seuils de concordance et de discordance, les méthodes ELECTRE II et ELECTRE III sont des choix possibles. Si les décideurs préfèrent une approche plus simple sans avoir à spécifier des seuils, ELECTRE I est préférable. Si la complexité computationnelle est un facteur limitant, ELECTRE I peut être privilégiée car elle est moins gourmande en ressources. Si ces dernières sont disponibles et que les décideurs sont prêts à gérer une complexité accrue, ELECTRE III peut être envisagée pour sa capacité à fournir un classement total. Les méthodes ELECTRE II et ELECTRE III permettent aux décideurs d'explorer l'impact des variations des seuils sur le classement final, offrant une approche plus flexible.

En résumé, le choix entre les méthodes ELECTRE dépend de la nature du problème, des préférences des décideurs et de la complexité souhaitée dans le processus de prise de décision. Il peut être utile de discuter avec les parties prenantes, de tester différentes méthodes sur des scénarios fictifs ou de réaliser une analyse de sensibilité pour évaluer la robustesse des résultats.

2.1.2 Méthodes PROMETHEE

Pour cet état de l'Art sur la méthode PROMETHEE, nous allons utiliser le livre de J.-P. Brans et B. Mareschal, "Promethee Methods" [8]. La méthode PROMETHEE (Preference Ranking Organization Method for Enrichment Evaluation) est une approche influente dans le domaine de la prise de décision multicritère développée par J.P. Brans en 1982. Cette méthodologie a été conçue pour aborder les complexités inhérentes aux décisions impliquant de multiples critères, en fournissant un cadre structuré pour évaluer et classer diverses alternatives. Les applications de la méthode PROMETHEE sont vastes et variées, couvrant des domaines tels que la santé, la finance, la gestion environnementale et bien d'autres. Sa capacité à gérer divers types de données et à fournir des résultats clairs et compréhensibles en fait un outil utile pour les décideurs dans de nombreux secteurs. Aujourd'hui, la méthode PROMETHEE propose 6 versions différentes dont nous allons voir les fonctionnements, les cas d'usages ainsi que les avantages et les inconvénients, en nous basant sur le livre " Multiple Criteria Decision Analysis : State of the Art Surveys " de Jean-Pierre Brans et Bertrand Mareschal.

Dans un premier temps, la méthode propose la version PROMETHEE I, qui est conçue pour fournir un classement partiel des alternatives basé sur un ensemble de critères évalués. Son objectif principal est d'établir un classement partiel des alternatives en prenant en compte les préférences du décideur sur plusieurs critères.Il est applicable dans des situations où il est nécessaire de comparer un ensemble d'alternatives sur la base de critères multiples, mais où un classement complet n'est pas indispensable ou possible en raison de l'incomparabilité de certaines alternatives.

PROMETHEE I utilise des flux de surclassement positifs et négatifs pour évaluer chaque alternative. Ainsi, on retrouve un flux positif $\phi+$ qui représente

la force ou la puissance d'une alternative, indiquant dans quelle mesure elle surclasse les autres alternatives et un flux négatif $\phi-$ qui indique la faiblesse ou la vulnérabilité d'une alternative, c'est-à-dire dans quelle mesure elle est surclassée par les autres. PROMETHEE I effectue des préférences par paires : les préférences sont évaluées en comparant les alternatives deux à deux sur chaque critère. Après cela, les préférences individuelles sont agrégées pour obtenir un score global de surclassement pour chaque alternative. Les alternatives sont classées en fonction de leurs flux de surclassement positifs et négatifs. Certaines peuvent ne pas être comparables si elles sont fortes dans des critères différents. PROMETHEE I respecte cette incomparabilité, ce qui conduit à un classement partiel plutôt qu'à un classement complet. Lorsque les flux positifs et négatifs d'une alternative sont équilibrés, cela peut conduire à l'indécision quant à sa position relative dans le classement. La méthode PROMETHEE I est transparente et facile à comprendre pour les décideurs. Elle peut gérer des préférences variées et des critères de nature différente (quantitative et qualitative). De plus, elle est convenable pour les situations où toutes les alternatives ne peuvent pas être comparées de manière absolue. Cependant, elle ne fournit pas un classement définitif de toutes les alternatives, ce qui peut être limitatif dans certaines applications décisionnelles. De plus, comme pour toutes les méthodes multicritères, les résultats dépendent de la pondération des critères. PROMETHEE I est particulièrement utile dans les contextes où un classement complet n'est pas essentiel ou lorsque l'incompatibilité entre les alternatives est une considération importante. Elle offre une approche équilibrée pour évaluer une série d'options en tenant compte des multiples facettes d'un problème décisionnel.

Ensuite, la version PROMETHEE II est une extension de la méthode PROMETHEE I, développée pour fournir un classement complet des alternatives en fonction de multiples critères. Cette méthode produit un classement complet de toutes les alternatives sur la base de plusieurs critères, et est utilisée dans des situations où un classement distinct et définitif des alternatives est nécessaire, en tenant compte de plusieurs critères.

PROMETHEE II se base sur le calcul d'un flux net pour chaque alternative. Le flux net ϕ est la différence entre le flux positif $\phi+$ et le flux négatif $\phi-$ pour chaque alternative. Un flux net élevé indique une alternative globalement préférable, tandis qu'un flux net faible indique une alternative moins souhaitable. Chaque alternative est évaluée en considérant l'ensemble des critères, plutôt qu'en se concentrant sur des comparaisons par paires. Les préférences individuelles sur chaque critère sont agrégées pour obtenir le flux net de chaque alternative. Contrairement à PROMETHEE I, PROMETHEE II fournit un classement complet, ordonnant toutes les alternatives du meilleur au moins bon. Toutes les alternatives sont comparables entre elles, permettant une décision claire et définitive. La méthode fournit un classement clair et facile à interpréter pour les décideurs, et est utile dans des contextes décisionnels où un classement définitif est nécessaire pour la sélection ou le rejet d'options. La méthode possède également des inconvénients : elle peut sur-simplifier la réalité en ne tenant pas compte de l'incomparabilité potentielle entre certaines alternatives et tout comme pour PROMETHEE I, les résultats dépendent fortement

des poids attribués aux différents critères. PROMETHEE II est utile dans les contextes où un classement complet est essentiel pour la prise de décision. Cette méthode possède une capacité à fournir une hiérarchie claire et décisive des alternatives, facilitant ainsi le processus de sélection ou d'élimination dans des situations multicritères complexes.

La version PROMETHEE III est une variante de PROMETHEE conçue pour gérer l'incertitude et l'imprécision dans les données d'évaluation. Elle a pour objectif principal de permettre un classement des alternatives basé sur des intervalles pour gérer l'‘incertitude ou l'imprécision des données et s'utilise dans les situations où les évaluations des alternatives sur les critères sont incertaines ou exprimées sous forme d'intervalles plutôt que de valeurs précises. Au lieu d'utiliser des valeurs ponctuelles pour les critères, PROMETHEE III utilise des intervalles pour représenter les évaluations. Cette approche permet de traiter efficacement les données imprécises ou incertaines, souvent rencontrées dans des situations réelles. Les alternatives sont comparées en utilisant des intervalles d'évaluation sur chaque critère puis les préférences sont déterminées en tenant compte de la manière dont les intervalles des différentes alternatives se chevauchent ou diffèrent. Les alternatives sont classées en fonction de la manière dont leurs intervalles se comparent sur l'ensemble des critères. Cette méthode fournit une vision plus nuancée des préférences, reflétant la variabilité ou l'incertitude inhérente aux données. Cette adaptabilité aux données incertaines rend cette méthode capable de gérer efficacement les situations où les données ne sont pas clairement définies. De plus, cette méthode fournit une représentation plus réaliste des situations de décision où les données sont naturellement imprécises. Cependant, elle peut être plus complexe à interpréter en raison de la nature des intervalles utilisés et la manipulation et le calcul des intervalles peuvent être plus exigeants que dans les versions plus simples de PROMETHEE. PROMETHEE III est particulièrement utile dans des contextes où les données d'évaluation ne sont pas absolues et contiennent des éléments d'incertitude ou de variabilité. Cette méthode permet aux décideurs de prendre en compte de manière plus réaliste l'imprécision inhérente à de nombreuses situations d'évaluation et de prendre des décisions plus informées en présence d'incertitude.

Dans un autre cas, PROMETHEE IV est spécialement conçue pour gérer les situations où les critères ne sont pas discrets ou catégoriels mais suivent plutôt un spectre continu. Cette méthode permet la mise en œuvre d'un classement des alternatives dans des cas où les critères sont évalués de manière continue, sans seuils prédéfinis. Les préférences sont déterminées en considérant l'ensemble du spectre des performances sur chaque critère. Les alternatives sont évaluées en considérant les variations continues dans les performances sur chaque critère, puis les préférences sont agrégées sur l'ensemble du spectre d'évaluation pour chaque critère, offrant une analyse détaillée et nuancée. PROMETHEE IV fournit un classement flexible des alternatives qui reflète les variations subtiles dans les performances sur les critères continus, ce qui permet une compréhension plus fine des nuances dans les performances des alternatives. Globalement, cette méthode est particulièrement adaptée aux situations où les critères ne peuvent être facilement catégorisés ou limités à des seuils spécifiques. De plus,

elle est capable de capturer des différences subtiles dans les performances, ce qui est essentiel dans des contextes où de légères variations sont significatives. Cependant, elle peut être plus complexe à mettre en œuvre et à interpréter en raison de la nature continue des critères. Elle nécessite également des données détaillées et précises sur les critères pour une analyse efficace. PROMETHEE IV est une méthode idéale pour les situations où les critères de décision ne sont pas facilement séparables en catégories discrètes et nécessitent une évaluation sur un continuum. Cette méthode est la plus utile dans des domaines où les nuances et les petits écarts dans les performances ou les résultats sont cruciaux pour la prise de décision.

La méthode PROMETHEE V permet la prise de décision multicritère avec des contraintes spécifiques. Elle adapte la méthodologie PROMETHEE pour sélectionner un sous-ensemble d'alternatives, tout en tenant compte de contraintes spécifiques. Elle est utilisée dans des scénarios où il est nécessaire de respecter certaines contraintes, telles que des limites budgétaires, des capacités, ou d'autres restrictions opérationnelles. PROMETHEE V permet d'intégrer des contraintes dans le processus de décision, en plus des critères de préférence habituels. Ces contraintes peuvent être de diverses natures (quantitatives, qualitatives, temporelles...) et sont incorporées directement dans le modèle de décision. Cela se fait en deux étapes : une première étape de classement sans contraintes où les alternatives sont classées en utilisant la méthode PROMETHEE II pour obtenir un flux net de surclassement, puis une deuxième étape où sont appliquées les contraintes et où un modèle de programmation linéaire 0, 1 est utilisé pour sélectionner le meilleur sous-ensemble d'alternatives qui respecte les contraintes. Le résultat final est un ensemble d'alternatives qui maximise la performance globale tout en respectant les contraintes définies. La méthode vise à obtenir un équilibre entre les préférences du décideur et les restrictions imposées par le contexte. Cela permet de modéliser des situations de décision réalistes où les choix sont limités par des contraintes extérieures et de gérer divers types de contraintes, rendant la méthode adaptable à de nombreux contextes décisionnels. Cependant, l'ajout de contraintes augmente la complexité du processus de décision. En plus de cela, l'efficacité de la méthode dépend de la précision avec laquelle les contraintes sont définies et intégrées. PROMETHEE V est particulièrement utile dans les contextes où les décideurs doivent naviguer dans un environnement avec des limitations ou des contraintes spécifiques. Cette méthode offre un cadre structuré pour intégrer ces contraintes dans le processus d'évaluation multicritère, permettant ainsi de prendre des décisions plus informées et réalisables dans des scénarios contraints.

Enfin, PROMETHEE VI est conçue spécifiquement pour l'analyse de sensibilité dans les décisions multicritères. Elle fournit un outil d'analyse de sensibilité pour étudier comment les variations dans les poids des critères affectent les classements des alternatives et est utilisée dans des contextes où les décideurs souhaitent comprendre l'impact de la modification des priorités ou des préférences sur le résultat final de la décision. PROMETHEE VI explore l'effet des changements dans la pondération des critères sur le classement des alternatives et permet de modéliser comment les décideurs peuvent envisager différents scé-

narios de pondération pour les critères. Ces derniers définissent des intervalles pour les poids des critères, reflétant leur incertitude ou leur flexibilité dans les préférences. La méthode explore alors différents ensembles de poids dans les intervalles définis pour voir comment ils influencent le classement. La méthode montre comment les classements des alternatives peuvent varier en fonction des modifications apportées aux poids des critères. Elle permet d'identifier les alternatives qui restent préférables dans un large éventail de scénarios de pondération. Cela aide à évaluer la robustesse des décisions face à l'incertitude ou à la variation des préférences et offre une approche flexible pour tester différentes hypothèses et comprendre l'impact des préférences sur le résultat de la décision. En revanche, cette méthode peut être complexe à mettre en œuvre, en particulier dans des situations avec un grand nombre de critères et d'alternatives. De plus, elle nécessite une analyse approfondie pour interpréter correctement les variations dans les classements. PROMETHEE VI est particulièrement utile pour les décideurs qui souhaitent comprendre en profondeur l'impact de leurs préférences sur les décisions finales. Cette méthode offre un moyen précieux de tester la sensibilité des résultats de décision aux changements dans les poids attribués aux différents critères, permettant ainsi une prise de décision plus informée et adaptable.

Pour conclure, chaque version de PROMETHEE a été conçue pour répondre à des besoins spécifiques dans le cadre de la prise de décision multicritère. Ces différentes versions permettent une grande flexibilité et une adaptation à divers contextes de décision, allant des situations avec des données précises à celles caractérisées par l'incertitude et les contraintes multiples. PROMETHEE est supportée dans des logiciels comme DECISION LAB, qui intègre également la méthode de compréhension graphique GAIA. De plus, une recherche continue peut toujours être observée autour de cette méthode, dans le but de l'améliorer et de l'appliquer dans de nouveaux domaines.

2.2 Algorithmes d'extraction de motifs

Dans le contexte de la découverte de motifs en data mining, un motif ("pattern") est une structure ou une tendance significative qui se répète dans un ensemble de données. Plus précisément, un motif représente une association, une corrélation ou une séquence récurrente d'éléments, d'attributs ou d'événements au sein des données.

Ce motif est représenté sous forme de règle d'association. Une règle d'association est de la forme de couple $\{Antecedent \Rightarrow Consequent\}$. Un example qui semble souvent repris pour illustrer ce concept est celui du supermarché [9] : La règle d'association $\{Oignons, Pommesdeterre\} \Rightarrow \{Viande\}$ trouvée dans les données de vente d'un supermarché signifie que les clients qui achètent des oignons et des pommes de terre ensemble ont tendance à aussi acheter de la viande. Une telle règle d'association pourrait alors être utilisée pour le marketing.

Ainsi, le processus de découverte de règles d'association s'appelle l'apprentissage de règles d'association.

Dans cet état de l'art, nous allons voir deux principaux algorithmes de fouille de données.

2.2.1 L'algorithme Apriori

L'algorithme Apriori [10] est une méthode de base en data mining pour extraire des règles d'association. Développé par Agrawal et Srikant en 1994, il a été conçu pour identifier des ensembles d'éléments fréquents dans de grandes bases de données, une étape préliminaire essentielle à la génération de règles d'association. L'objectif principal de l'algorithme est de découvrir des patterns dans des ensembles de données, ce qui est important pour la prise de décision dans divers domaines tels que le marketing ou le placement de produits.

L'algorithme Apriori est basé sur deux concepts clés : le support et la confiance . Le support mesure la fréquence à laquelle un ensemble d'éléments apparaît dans la base de données, tandis que la confiance mesure la fréquence à laquelle les règles d'association se produisent. Le processus se déroule en plusieurs étapes : l'algorithme commence par calculer le support de chaque élément de la base de données et élimine les éléments dont le support est inférieur au seuil minimum. Il crée ensuite des ensembles de deux éléments, calcule leur prise en charge et élimine les ensembles peu fréquents. Ce processus est répété pour les ensembles de taille croissante jusqu'à ce qu'aucun nouvel ensemble fréquent ne puisse être créé. A partir de l'ensemble des items fréquents, l'algorithme génère des règles d'association respectant le seuil minimum de confiance.

L'algorithme a un concept simple et direct, le rendant facile à mettre en œuvre et à utiliser dans une variété de contextes. Il peut être appliqué à tout type de base de données transactionnelle pour identifier des modèles fréquents. Cependant, la nécessité d'analyser la base de données plusieurs fois et de générer un grand nombre de candidats peut rendre l'algorithme lent et inefficace avec de très grands ensembles de données. La gestion de la mémoire peut devenir problématique à mesure que la taille de l'ensemble d'éléments augmente. L'algorithme Apriori reste un pilier dans le domaine du data mining, notamment dans l'extraction de règles d'association.

Bien qu'il existe des limites de performances avec de grands ensembles de données, son influence est indéniable, car elle a jeté les bases de nombreuses recherches et développements ultérieurs dans ce domaine.

2.2.2 L'algorithme FP-Growth

L'algorithme FP-Growth (Frequent Pattern Growth) [11] représente une avancée dans le domaine de l'exploration de règles d'association, en proposant une méthode efficace qui évite la génération de candidats, l'un des principaux inconvénients de l'algorithme Apriori. Développé par Han, Pei et Yin en 2000, FP-Growth utilise une structure arborescente compacte, appelée Frequent Pattern Tree (FP), pour compresser la base de données et extraire des ensembles d'éléments fréquents sans créer d'ensembles candidats.

L'algorithme FP-Growth procède en deux étapes principales : tout d'abord, l'algorithme crée un arbre qui résume les informations contenues dans les données de la base de données, en suivant les chemins de transaction de données courants. Chaque nœud représente un élément et stocke le nombre de transactions passant par ce nœud. Cela permet de compresser l'ensemble de données dans une structure compacte, facilitant ainsi les calculs ultérieurs. À partir de l'arborescence, l'algorithme extrait des ensembles d'éléments fréquents en commençant par les éléments les moins fréquents et remonte l'arborescence, en utilisant une technique appelée « pattern growth », qui consiste à combiner les préfixes communs des chemin pour créer des ensembles d'éléments fréquents.

L'algorithme FP-Growth réduit le besoin d'analyser plusieurs bases de données, améliorant considérablement la vitesse de traitement par rapport à Apriori, en particulier pour les grandes bases de données. De plus, grâce à sa structure arborescente compacte, FP-Growth gère l'utilisation de la mémoire plus efficacement, permettant une plus grande évolutivité pour gérer de gros volumes de données. L'algorithme évite l'étape coûteuse de génération de candidats présente dans Apriori, réduisant ainsi le coût de calcul global. Cependant, la construction d'une arborescence FP peut devenir complexe et gourmande en mémoire pour les bases de données comportant un grand nombre de transactions distinctes. En plus de cela, malgré que plus efficace qu'Apriori en termes de gestion de la mémoire, FP-Growth peut toujours rencontrer des limitations de mémoire avec des ensembles de données extrêmement volumineux ou très divers.

L'algorithme FP-Growth est une méthode puissante et efficace pour extraire des ensembles d'éléments fréquents, offrant une solution robuste aux limitations de l'algorithme Apriori, en particulier pour les grandes bases de données. La capacité d'extraire des modèles communs sans générer de candidats a ouvert la voie à des recherches et des applications plus efficaces dans l'exploration de données.

3 Proposition

3.1 Choix du jeu de données

3.1.1 Contexte

Le diabète figure parmi les maladies chroniques les plus répandues aux États-Unis, touchant des millions d'Américains chaque année et imposant un fardeau financier significatif sur l'économie. Le diabète est une maladie chronique grave dans laquelle les individus perdent la capacité de réguler efficacement les niveaux de glucose dans le sang, ce qui peut entraîner une diminution de la qualité de vie et de l'espérance de vie. Après la décomposition des aliments en sucres pendant la digestion, ces derniers sont libérés dans le flux sanguin. Cela signale au pancréas de libérer de l'insuline, qui permet aux cellules du corps d'intégrer le sucre pour être utilisé comme source d'énergie. Le diabète est généralement caractérisé soit par le fait que le corps ne produit pas suffisamment d'insuline, soit par une incapacité à utiliser l'insuline produite aussi efficacement que nécessaire.

Des complications telles que les maladies cardiaques, la perte de vision, l'amputation des membres inférieurs et les maladies rénales sont associées à des niveaux élevés de sucre restant de manière chronique dans le flux sanguin pour ceux atteints de diabète. Bien qu'il n'existe pas de remède pour le diabète, des stratégies telles que la perte de poids, une alimentation saine, l'activité physique et les traitements médicaux peuvent atténuer les dommages de cette maladie chez de nombreux patients. Un diagnostic précoce peut conduire à des changements de mode de vie et à un traitement plus efficace, rendant les modèles prédictifs du risque de diabète importants pour le public et les responsables de la santé.

Il est également important de reconnaître l'ampleur de ce problème. Le Centre pour le Contrôle et la Prévention des Maladies (CDC) a indiqué qu'en 2018, 34,2 millions d'Américains souffraient de diabète et 88 millions étaient en état de prédiabète. Cela correspond à plus d'un tiers de la population. De plus, le CDC estime qu'1 diabétique sur 5, et environ 8 prédiabétiques sur 10, ne sont pas conscients de leur risque. Bien qu'il existe différents types de diabète, le diabète de type II est la forme la plus courante et sa prévalence varie en fonction de l'âge, de l'éducation, du revenu, de la localisation, de l'ethnie et d'autres déterminants sociaux de la santé. Le diabète impose également un fardeau massif sur l'économie, avec des coûts diagnostiqués de diabète d'environ 327 milliards de dollars et des coûts totaux comprenant le diabète non diagnostiqué et le prédiabète approchant les 400 milliards de dollars annuellement.

3.1.2 Contenu

Le Behavioral Risk Factor Surveillance System (BRFSS) est une enquête téléphonique liée à la santé, collectée annuellement par le Centre pour le Contrôle et la Prévention des Maladies (CDC). Chaque année, l'enquête recueille les réponses de plus de 400 000 Américains sur les comportements à risque liés à la santé, les conditions de santé chroniques et l'utilisation des services de prévention. Elle est menée chaque année depuis 1984. Pour ce projet, le fichier source est un csv de l'ensemble de données disponible sur Kaggle pour l'année 2015. Ce dataset original contient les réponses de 441 455 individus et comporte 330 caractéristiques. Ces caractéristiques sont soit des questions posées directement aux participants, soit des variables calculées sur la base des réponses individuelles des participants.

Le dataset utilisé pour ce projet est `diabetes_012_health_indicators_BRFSS2015.csv`. C'est un ensemble de 253 680 réponses à l'enquête BRFSS2015 des CDC. La variable cible `Diabetes_012` a 3 classes. 0 est pour l'absence de diabète ou seulement pendant la grossesse, 1 est pour le prédiabète, et 2 est pour le diabète. Il y a un déséquilibre des classes dans cet ensemble de données. Ce dataset comporte 21 attributs.

- **Diabetes012** : 0 = no diabetes 1 = prediabetes 2 = diabetes
- **HighBP** : 0 = no high BP 1 = high BP
- **HighChol** : 0 = no high cholesterol 1 = high cholesterol
- **CholCheck** : 0 = no cholesterol check in 5 years 1 = yes cholesterol

check in 5 years
— **BMI** : Body Mass Index
— **Smoker** : Have you smoked at least 100 cigarettes in your entire life ? [Note : 5 packs = 100 cigarettes] 0 = no 1 = yes
— **Stroke** : (Ever told) you had a stroke. 0 = no 1 = yes
— **HeartDiseaseorAttack** : coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
— **PhysActivity** : physical activity in past 30 days - not including job 0 = no 1 = yes
— **Fruits** : Consume Fruit 1 or more times per day 0 = no 1 = yes
— **Veggies** : Consume Vegetables 1 or more times per day 0 = no 1 = yes
— **HvyAlcoholConsump** : Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes
— **AnyHealthcare** : Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
— **NoDocbcCost** : Was there a time in the past 12 months when you needed to see a doctor but could not because of cost ? 0 = no 1 = yes
— **GenHlth** : Would you say that in general your health is : scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
— **MentHlth** : Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good ? scale 1-30 days
— **PhysHlth** : Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good ? scale 1-30 days
— **DiffWalk** : Do you have serious difficulty walking or climbing stairs ? 0 = no 1 = yes
— **Sex** : 0 = female 1 = male
— **Age** : 13-level age category 1 = 18-24 9 = 60-64 13 = 80 or older
— **Education** : Education level scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8
— **Income** : Income scale scale 1-8 1 = less than 10,000 5 = less than 35,000 8 = 75,000 or more

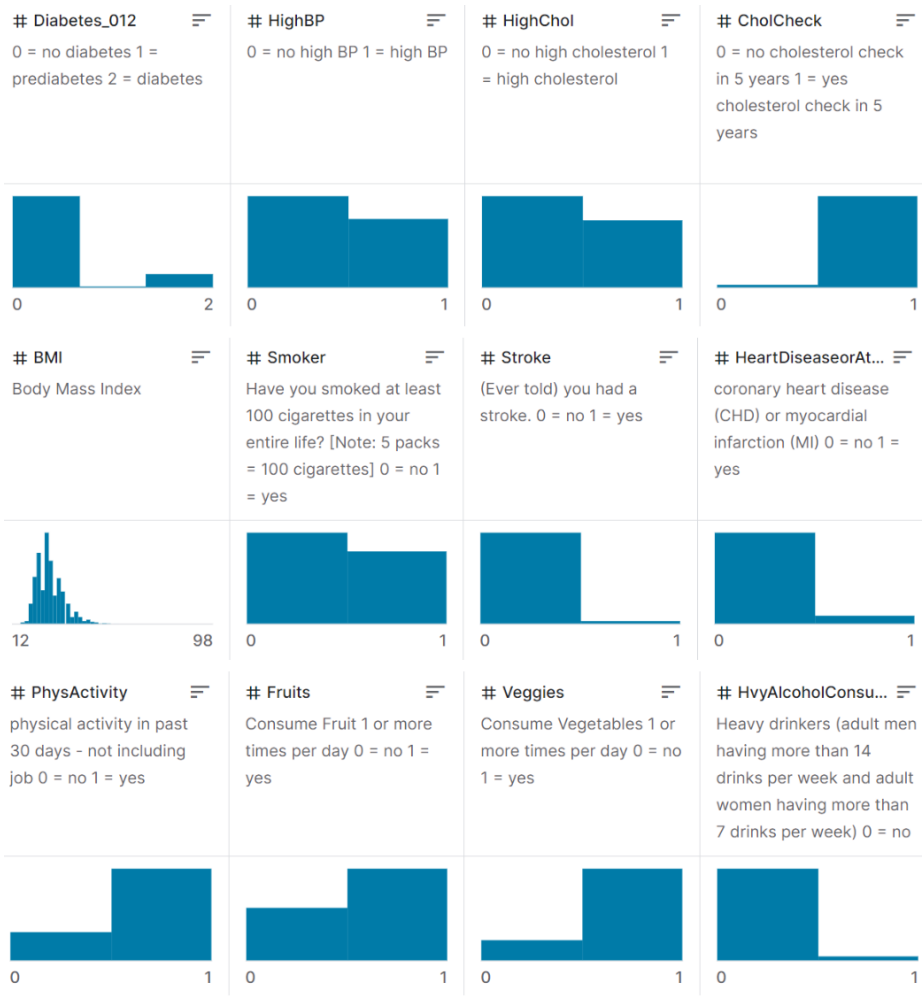

FIGURE 1 – Visualisation des distributions des données (1/2)

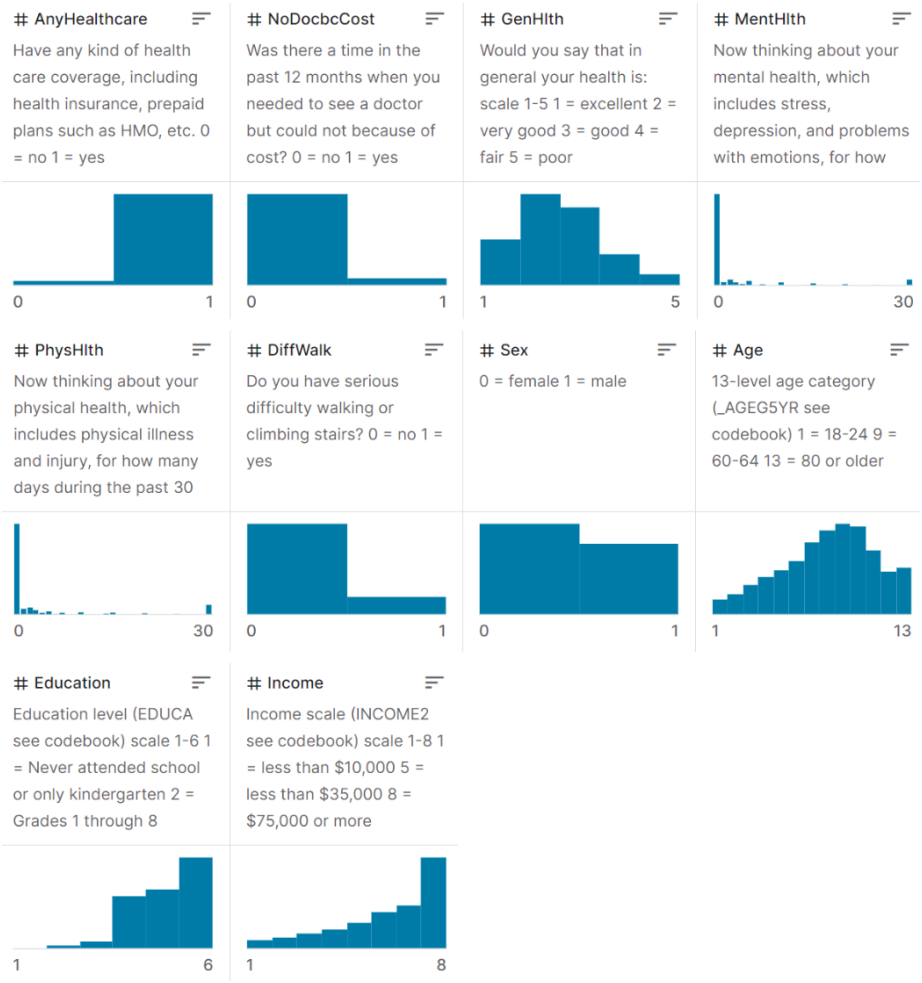

FIGURE 2 – Visualisation des distributions des données (2/2)

3.2 Choix de l'algorithme de fouille de données

Il existe d'autres algorithmes tels que AIS et SETM. L'algorithme Apriori semble le plus aproprié pour notre, étant donné que son temps d'exécution est linéaire est reste faible comparé aux algorithmes AIS et SETM. [12]. Les performances de Apriori étant suffisantes pour déterminer les règles d'association de notre jeu de données, nous l'utiliserons par la suite.

	Apriori	FP-Growth
Vitesse	Plus lent, le temps d'exécution augmente exponentiellement avec le nombre de jeux d'éléments.	Plus rapide, le temps d'exécution augmente linéairement avec le nombre de jeux d'éléments.
Mémoire	Importante, tous les candidats issus de l'auto-jonction sont stockés en mémoire.	Faible, stocke une version compacte de la table de transactions.
Candidats	Utilise l'auto-jonction pour la génération de candidats.	Pas de génération de candidats.
Motifs fréquents	Les motifs sont sélectionnés parmi les candidats dont le support est supérieur au minSup.	La croissance des motifs est réalisée par l'exploitation des arbres FP conditionnels.
Scans	Balaye la base de données à plusieurs reprises.	Seulement deux scans nécessaires.

TABLE 2 – Comparaison des algorithmes de data mining

3.3 Choix de la méthode multicritère

Afin d'effectuer un classement des motifs, nous utiliserons Le logiciel MCDA-ULaval. Le logiciel propose 6 méthodes multicritères :

— ELECTRE Tri-nC
— ELECTRE Tri-rC
— ELECTRE Tri-C (reformulated)
— ELECTRE Tri-B
— ELECTRE III
— ELECTRE II

ELECTRE II sera retenu pour ce projet en raison de sa simplicité et de la possibilité d'exploiter les résultats finaux de par leur forme qui est un classement. Par ailleurs, le logiciel utilisé ne propose pas PROMETHEE, ni quelques versions d'ELECTRE.

3.4 Choix des critères pour l'AMCD

Notre algorithme Apriori nous fournit en sortie un ensemble de règles d'association, qui chacune ont 3 attributs : Le Support, la Confiance et le Lift. Nous décrivons par la suite à quoi ces attributs correspondent. Soient X et Y des valeurs que peuvent prendre 2 champs différents du jeu de données :

$$Support(X) = \frac{Nombredetransactionscontenant X}{Nombretotaldetransactions}$$

Le support de X peut être interprété comme étant la popularité de X. On l'appelle également fréquence de X.

$$Confiance(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

La confiance correspond au ratio des transactions contenant à la fois X et Y par rapport à toutes les transactions où X existe. Elle peut également être interprétée comme étant la probabilité $P(Y|X)$.

$$Lift(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)}$$

Le lift est la vraisemblance de l'apparition de X quand Y apparaît. Si le lift est supérieur à 1, alors Y a tendance à apparaître si X apparaît (corrélation positive). Si le lift est inférieur à 1, alors Y a tendance à ne pas apparaître si X apparaît (corrélation négative). Si le lift est égal à 1, la corrélation entre X et Y est nulle.

$$Conviction(X \to Y) = (1 - Support(Y))/(1 - Confiance(X \to Y))$$

Par exemple, si $Conviction(X \to Y) = 0.32$, alors la règle $X \to Y$ est incorrecte 32% du temps si l'association entre X et Y était dûe à la chance.

L'inconvénient de la confiance c'est qu'elle ne prend en compte que la popularité de l'objet X, et pas vraiment celle d'Y. Si Y est aussi populaire que X, il y a de grandes chances que les transactions qui contiennent X contiennent aussi Y, ce qui augmente la confiance, mais ça n'implique pas forcément que X et Y ont tendance à être ensemble. Pour minimiser cet inconvénient, nous assignerons au lift un poids plus important que la confiance et que le support.

Nous ne conserverons que les critères Support, Confidence et Lift

Support	Confiance	Lift
1	1	2

TABLE 3 – Poids assignés à chaque critère

Tous les critères sont à maximiser.

4 Expérimentations et résultats

4.1 Préparation et formatage des données

Nous avons publié le travail de cette sous-partie sur https ://github.com/ArasCha/data-mining-mcda/blob/main/Notebook.ipynb

4.1.1 Préparation des données

Le jeu de données récupéré a déjà été préalablement nettoyé par Alex TEBOUL. Ce dernier a récupéré ces données sur le site des centres pour le contrôle

et la prévention des maladies (en anglais Centers for Disease Control and Prevention ou CDC) puis a effectué des étapes de nettoyage sur ces dernières (dimensions non pertinentes, valeurs non pertinentes / aberrantes), réduisant ainsi le nombre de dimensions de 330 à 22. Cette étape de nettoyage préalable est très utile dans le cadre de notre étude, puisque toutes les dimensions de base ne concernaient pas spécifiquement le diabète, et elle réduit également le nombre de valeurs possibles sur certaines des dimensions restantes. Nous allons cependant effectuer un autre nettoyage en important le jeu de données sur Python, sous forme de jeu de données Pandas.

Dans un premier temps, nous allons fusionner les dimensions "Fruits" et "Veggies". Ces dernières concernent en effet toutes deux le repas, et ainsi nous pouvons les regrouper en une seule nommée "Alimentation", avec pour valeur possible fruits, vegetables, fruits and vegetables ou enfin no fruits or vegetables.

Ensuite, nous allons catégoriser la valeur BMI (Body Mass Index). Cette dernière est une valeur continue, or nous voulons des valeurs catégoriques pour l'extraction de patterns. Nous allons donc définir des plages de BMI :

— < 1850
— $1850 <= BMI < 2500$
— $2500 <= BMI < 3000$
— $BMI >= 3000$

Après cela, nous supprimons les colonnes $'Diabetes_012', 'NoDocbcCost'$, $'PhysHlth', 'MentHlth', 'CholCheck', 'PhysActivity'$, ces dernières étant peu fiables ou inutiles (l'activité physique définie comme étant 1 fois les 30 derniers jours est-elle vraiment utile ?).

Enfin, nous allons remplacer les valeurs réduites (0, 1...) par leurs valeurs réelles en s'appuyant sur le travail de Alex TEBOUL et la documentation sur l'étude du CDC. Nous avons ainsi, à la fin de l'étape de nettoyage et de préparation de données, les dimensions et leurs valeurs possibles :

Dimension	Valeurs possibles
Blood Pressure	high blood pressure, normal blood pressure
Cholesterol	normal cholesterol, high cholesterol
Body Mass Index	obese, overweight, normal weight, underweight
Smoker	smokes, does not smoke
Stroke	never had stroke, had stroke
Heart Disease or Attack	no heart disease/attack, had/has heart disease/attack
Alimentation	vegetables, no fruits nor vegetables, fruits, fruits and vegetables
Alcohol Consumption	normal alcohol consumption, heavy alcohol consumption
Healthcare	covered by healthcare, not covered by healthcare
General Health	poor health, good health, very good health, fair health, excellent health
Difficulties To Walk	difficulties to walk, no difficulty to walk
Sex	female, male
Age	[60-64], [50-54], [70-74], [65-69], [55-59], 80+, [35-39], [45-49], [25-29], [75-79], [40-44], [18-24], [30-34]
Education	grade 12 or GED, college 4 years or more, grades 9 through 11, college 1 year to 3 years, grades 1 through 8, never attended school
Income	[15 000-19 999], <10000, 75 000+, [35 000-49 999], [20 000-24 999], [50 000-74 999], [10 000-14 999], [25 000-34 999]

TABLE 4 – Dimensions et leurs valeurs possibles

4.1.2 Formatage des données

Les algorithmes d'extraction de patterns, comme vu lors de l'état de l'art, s'effectuent sur des tables de données transactionnelles. Pour utiliser ce genre d'algorithme sur nos données, il va donc falloir convertir nos différentes colonnes en une table de transactions. On obtient alors une liste de transactions :

```
['high blood pressure', 'high cholesterol', 'obese', 'smokes', 'never had stroke', 'had / has heart disease / attack', 'fruits and v
['normal blood pressure', 'normal cholesterol', 'overweight', 'smokes', 'never had stroke', 'no heart disease / attack', 'fruits and
['high blood pressure', 'high cholesterol', 'overweight', 'does not smoke', 'never had stroke', 'no heart disease / attack', 'vegeta
['normal blood pressure', 'normal cholesterol', 'normal weight', 'smokes', 'never had stroke', 'no heart disease / attack', 'no frui
['high blood pressure', 'normal cholesterol', 'overweight', 'does not smoke', 'never had stroke', 'no heart disease / attack', 'frui
['high blood pressure', 'high cholesterol', 'obese', 'smokes', 'had stroke', 'had / has heart disease / attack', 'vegetables', 'norm
['high blood pressure', 'high cholesterol', 'overweight', 'smokes', 'never had stroke', 'had / has heart disease / attack', 'vegetab
['high blood pressure', 'high cholesterol', 'overweight', 'smokes', 'never had stroke', 'no heart disease / attack', 'fruits and veg
['high blood pressure', 'high cholesterol', 'obese', 'smokes', 'had stroke', 'no heart disease / attack', 'no fruits nor vegetables'
['high blood pressure', 'high cholesterol', 'normal weight', 'smokes', 'never had stroke', 'no heart disease / attack', 'no fruits n
```

FIGURE 3 – Liste de données converties en transactions

Nous allons maintenant pouvoir utiliser cette liste de transactions afin d'appliquer un algorithme d'extraction de patterns.

4.2 Application de l'algorithme de data mining

4.2.1 Comparaison des seuils de confiance et de support

Support Treshold	Confidence Treshold	Number of rules extracted
0.5	0.5	89
0.5	0.6	84
0.5	0.7	84
0.5	0.8	84
0.5	0.9	82
0.6	0.5	31
0.6	0.6	31
0.6	0.7	31
0.6	0.8	31
0.6	0.9	30
0.7	0.5	29
0.7	0.6	29
0.7	0.7	29
0.7	0.8	29
0.7	0.9	28
0.8	0.5	16
0.8	0.6	16
0.8	0.7	16
0.8	0.8	16
0.8	0.9	15
0.9	0.5	6
0.9	0.6	6
0.9	0.7	6
0.9	0.8	6
0.9	0.9	6
1	0.5	0
1	0.6	0
1	0.7	0
1	0.8	0
1	0.9	0

TABLE 5 – Experimentation de plusieurs seuils pour la confiance et le support

D'après la Table, le choix des seuils où minsup=0.9 conduit à l'extraction d'un nombre minimal de règles d'association (6), ce qui semble être un choix logique. Mais ces 6 règles sont extraites uniquement sur la base des deux mesures que sont le support et la confiance. Dans cette première expérience, les seuils minsup=0.6 et minconf=0.8 sont choisis. En effet, ces choix permettent à

l'algorithme Apriori d'extraire 31 règles d'association, ce qui confère un panel plus large à notre méthode d'AMCD. La méthode ELECTRE II sera utilisée pour choisir les meilleures règles parmi cet ensemble de 31 règles d'association, ce qui semble plus riche que la sélection dans un ensemble de 6 règles d'association. Ce choix ne doit en aucun cas altérer le résultat final, car les 6 règles d'association, extraites pour les choix minsup=0.9, sont implicitement incluses dans l'ensemble des règles d'association extraites pour les choix minconf=0.8 et minsup=0.6. Ce qui fait que ces 6 règles seront également examinées par la méthode ELECTRE II avec le reste des autres règles d'association afin de trouver les meilleures règles d'association.

4.2.2 Extraction des règles d'association

Désormais, nous assignons à chaque règle d'association un sigle, tel que $A1$, $A2$... Après application de l'algorithme Apriori pour les seuils minconf=0.8 et minsup=0.6 nous obtenons les 20 règles d'association suivantes :

ID	**Rule**	**Support**	**Confidence**	**Lift**
A1	{covered by healthcare} ⇒ {high blood pressure}	0.724	0.962	1.002
A2	{covered by healthcare} ⇒ {no difficulty to walk}	0.603	0.959	0.999
A3	{covered by healthcare} ⇒ {no heart disease/attack}	0.744	0.958	0.998
A4	{covered by healthcare} ⇒ {normal cholesterol}	0.644	0.961	1.001
A5	{never had stroke} ⇒ {high blood pressure}	0.674	0.895	0.986
A6	{normal alcohol consumption} ⇒ {high blood pressure}	0.735	0.976	1.000
A7	{no heart disease/attack} ⇒ {never had stroke}	0.731	0.805	1.036
A8	{never had stroke} ⇒ {normal cholesterol}	0.601	0.896	0.988
A9	{normal alcohol consumption} ⇒ {no difficulty to walk}	0.612	0.973	0.997
A10	{normal alcohol consumption} ⇒ {no heart disease/attack}	0.758	0.975	0.998
A11	{normal alcohol consumption} ⇒ {normal cholesterol}	0.654	0.976	1.000
A12	{never had stroke} + {covered by healthcare} ⇒ {high blood pressure}	0.648	0.861	0.988
A13	{covered by healthcare} + {normal alcohol consumption} ⇒ {high blood pressure}	0.707	0.939	1.002
A14	{never had stroke} + {covered by healthcare} ⇒ {no heart disease/attack}	0.700	0.901	1.034
A15	{covered by healthcare} + {normal alcohol consumption} ⇒ {no heart disease/attack}	0.726	0.935	0.997
A16	{covered by healthcare} + {normal alcohol consumption} ⇒ {normal cholesterol}	0.629	0.938	1.000

A17	{never had stroke} + {normal alcohol consumption} ⇒ {high blood pressure}	0.657	0.873	0.986
A18	{never had stroke} + {normal alcohol consumption} ⇒ {no heart disease/attack}	0.712	0.916	1.035
A19	{never had stroke} + {covered by healthcare} + {normal alcohol consumption} ⇒ {high blood pressure}	0.632	0.840	0.988
A20	{never had stroke} + {covered by healthcare} + {normal alcohol consumption} ⇒ {no heart disease/attack}	0.682	0.878	1.033

TABLE 6 – Extended Association Rules with their Support, Confidence, and Lift

4.3 Tri des motifs avec la méthode ELECTRE II

Avec nos 3 critères, Support, Confidence et Lift, la méthode ELECTRE II demande 9 paramètres à régler. Nous allons dans un premier temps utiliser ces paramètres de discordance pour chaque critère :

[Paramètre]	**Support**	**Confidence**	**Lift**
Poids	1.0	1.0	2.0
d_1	0.1	0.1	0.1
d_2	0.1	0.1	0.1

TABLE 7 – Poids et discordances pour chaque critère

Ainsi que ces paramètres pour la concordance :
— C- = 0.6
— C0 = 0.7
— C+ = 0.8

Voici les 4 classements fournis par ELECTRE II après exécution (Direct, Inverse, Final et Médian) :

Rang	Classement Direct	Classement Inverse	Classement Final	Classement Médian
1	A1, A6, A7, A10, A18	A6, A10, A18	A6, A10, A18	A6, A10, A18
2	A3, A4, A11, A13, A14	A1, A3, A14	A1	A1
3	A2, A9, A15, A16, A20	A4, A11, A13, A15, A20	A3, A7, A14	A3, A14
4	A5, A8, A12	A2, A5, A9, A12, A16	A4, A11, A13	A4, A11, A13
5	A17, A19	A7, A8, A17, A19	A15, A20	A7, A15, A20
6			A2, A9, A16	A2, A9, A16
7			A5, A12	A5, A12
8			A8	A8
9			A17, A19	A17, A19

TABLE 8 – Comparaison des classements

Voici les graphes de surclassement qui les accompagne :

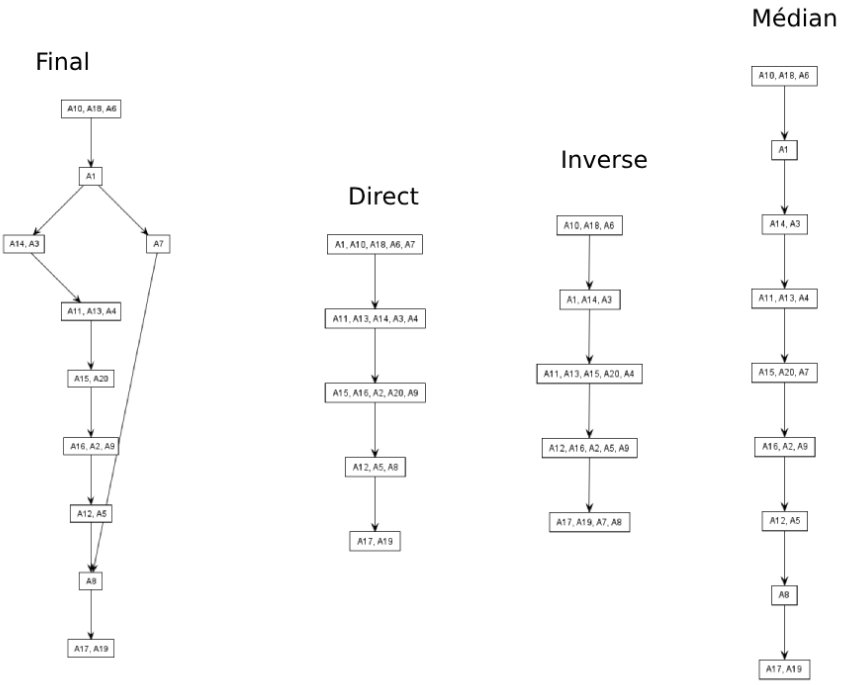

FIGURE 4 – Liste de données converties en transactions

4.3.1 Étude de robustesse

L'application de la méthode ELECTRE II repose sur plusieurs critères qui vont influer le résultat en sortie. Il est donc primordial de tester cette solution avec différents critères. L'objectif de l'étude de robustesse est de trouver les classements les plus stables, et donc les plus robustes. Dans cette étude de robustesse, nous allons effectuer deux applications supplémentaires de ELECTRE II, puis nous allons comparer les résultats en sortie afin de déterminer quels critères fournissent les meilleurs résultats.

Résultats avec les paramètres :

[Paramètre]	**Support**	**Confidence**	**Lift**
Poids	1.0	1.0	2.0
d_1	0.3	0.2	0.2
d_2	0.2	0.1	0.15

TABLE 9 – Discordances pour chaque critère Étude 2

Avec les concordances :
— C- = 0.7
— C0 = 0.8
— C+ = 0.9

Rang	Classement Direct	Classement Inverse	Classement Final	Classement Médian
1	A6, A7	A7	A7	A7
2	A1, A18	A18	A6, A18	A18
3	A13, A14	A1, A14	A1	A1
4	A20	A13, A20	A14	A6, A14
5	A4, A11	A4, A6	A13	A13
6	A10, A16	A11, A16	A20	A20
7	A2	A2, A10	A4	A4
8	A3	A3	A11	A11
9	A9	A9	A16	A16
10	A15	A15	A10	A10
11	A12	A12	A2	A2
12	A19	A19	A3	A3
13	A8	A8	A9	A9
14	A5	A5	A15	A15
15	A17	A17	A12	A12
16			A19	A19
17			A8	A8
18			A5	A5
19			A17	A17

TABLE 10 – Comparaison des classements de la seconde étude

Résultats avec les paramètres :

[Paramètre]	**Support**	**Confidence**	**Lift**
Poids	1.0	1.0	2.0
d_1	0.4	0.4	0.3
d_2	0.1	0.2	0.3

TABLE 11 – Discordances pour chaque critère Étude 3

Avec les concordances :
— C- = 0.75
— C0 = 0.80
— C+ = 0.85

Rang	Classement Direct	Classement Inverse	Classement Final	Classement Médian
1	A6, A7	A7	A7	A7
2	A1, A18	A18	A6, A18	A18
3	A13, A14	A1, A14	A1	A1
4	A20	A13, A20	A14	A6, A14
5	A4, A11	A4, A6	A13	A13
6	A10, A16	A11, A16	A20	A20
7	A2	A2, A10	A4	A4
8	A3	A3	A11	A11
9	A9	A9	A16	A16
10	A15	A15	A10	A10
11	A12	A12	A2	A2
12	A19	A19	A3	A3
13	A8	A8	A9	A9
14	A5	A5	A15	A15
15	A17	A17	A12	A12

TABLE 12 – Comparaison des classements de la troisième étude

Lors de la deuxième applicaction de la méthode ELECTRE II, on peut voir que les seuils de concordance et de discordance sont plus adaptés : on retrouve moins de règles d'alternatives sur chaque noeud du graphe de classement fourni en sortie par rapport aux deux autres applications faites.

Nous choisirons donc les critères de concordance et de discordance de la seconde étude.

4.4 Discussion

Dans les 3 études réalisées, on voit revenir très souvent en haut du classement les alternatives A6, A7 et A18. Ceci suggère que les règles d'association les plus pertinentes sont :

— {normal alcohol consumption} ⇒ {high blood pressure}
— {no heart disease/attack} ⇒ {never had stroke}
— {never had stroke} + {normal alcohol consumption} ⇒ {no heart disease/attack}

Les ajustements de ces paramètres à travers différentes études ont permis d'identifier les configurations les plus robustes. Il serait intéressant de comparer les résultats avec d'autres méthodes multicritères, afin de donner plus de robustesse à l'étude.

5 Conclusion

Ce projet sur la sélection de motifs pertinents basée sur l'aide à la décision multicritère met en lumière la capacité de l'approche multicritère à améliorer la sélection de motifs dans le contexte du data mining. En intégrant des méthodes d'aide à la décision multicritère, telles qu'ELECTRE II, pour évaluer et sélectionner des motifs d'association, notre étude a montré qu'il est possible d'identifier de manière plus efficace les motifs les plus pertinents à partir de vastes ensembles de données.

Cet article examine et discute du problème du diabète aux États-Unis, qui est préoccupant compte tenu de la proportion considérable de personnes concernées par cette maladie. Notre contribution à ce problème a été d'aider les décideurs à extraire des connaissances pertinentes sous forme de règles d'association. De plus, l'intégration de l'aide à la décision multicritère via ELECTRE II a résolu le problème du grand nombre de règles d'association extraites en ne sélectionnant que les plus intéressantes.

Le travail accompli a impliqué une série d'expérimentations méthodiques, où différents seuils de confiance et de support ont été testés pour extraire les règles d'association avant de les soumettre à un processus de sélection ordonnée utilisant ELECTRE II. Cette approche a permis de trier efficacement les motifs extraits, en mettant en évidence ceux qui offrent le meilleur équilibre entre le support, la confiance, et le lift, offrant ainsi une perspective plus nuancée et précise des associations significatives au sein des données.

Les résultats obtenus témoignent de l'efficacité de l'intégration des méthodes multicritères dans le processus de sélection des motifs. En particulier, les motifs finalement sélectionnés révèlent des associations intéressantes qui pourraient ne pas avoir été immédiatement évidentes sans cette approche analytique rigoureuse. Cela souligne l'importance de la prise en compte de multiples critères pour évaluer la pertinence des motifs, au-delà de la simple fréquence ou de la présence dans l'ensemble de données.

En conclusion, cette étude contribue de manière significative à la littérature sur le data mining et l'aide à la décision multicritère, en offrant une méthode robuste pour la sélection des motifs pertinents. L'approche peut être appliquée dans divers domaines pour améliorer la compréhension de données complexes et l'extraction d'interraltions. Il est envisagé de poursuivre cette recherche en explorant d'autres méthodes multicritères et en étendant l'application à d'autres types de données, ce qui pourrait ouvrir de nouvelles voies pour l'analyse avancée des données et la découverte de connaissances.

Références

[1] C. Romero and S. Ventura, "Data mining in education," *WIREs Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013. [Online]. Available : https ://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1075

[2] S. Brossette, A. Sprague, W. Jones, and S. Moser, "A data mining system for infection control surveillance," *Methods of information in medicine*, vol. 39, no. 4-5, p. 303—310, December 2000. [Online]. Available : http ://europepmc.org/abstract/MED/11191698

[3] A. Mucherino, P. Papajorgji, and P. M. Pardalos, "A survey of data mining techniques applied to agriculture," *Operational Research*, vol. 9, no. 2, pp. 121–140, Aug 2009. [Online]. Available : https ://doi.org/10.1007/s12351-009-0054-6

[4] J.-J. Wang, Y.-Y. Jing, C.-F. Zhang, and J.-H. Zhao, "Review on multi-criteria decision analysis aid in sustainable energy decision-making," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 9, pp. 2263–2278, 2009. [Online]. Available : https ://www.sciencedirect.com/science/article/pii/S1364032109001166

[5] I. B. Huang, J. Keisler, and I. Linkov, "Multi-criteria decision analysis in environmental sciences : Ten years of applications and trends," *Science of The Total Environment*, vol. 409, no. 19, pp. 3578–3594, 2011. [Online]. Available : https ://www.sciencedirect.com/science/article/pii/S0048969711006462

[6] G. Montibeller and A. Franco, *Multi-Criteria Decision Analysis for Strategic Decision Making*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2010, pp. 25–48. [Online]. Available : https ://doi.org/10.1007/978-3-540-92828-7_2

[7] V. Diaby, K. Campbell, and R. Goeree, "Multi-criteria decision analysis (mcda) in health care : A bibliometric analysis," *Operations Research for Health Care*, vol. 2, no. 1, pp. 20–24, 2013. [Online]. Available : https ://www.sciencedirect.com/science/article/pii/S2211692313000027

[8] J.-P. Brans and B. Mareschal, "Promethee Methods," in *Multiple Criteria Decision Analysis : State of the Art Surveys*, ser. International Series in Operations Research & Management Science. Springer, March 2005, ch. 0, pp. 163–186. [Online]. Available : https ://ideas.repec.org/h/spr/isochp/978-0-387-23081-8_5.html

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '93. New York, NY, USA : Association for Computing Machinery, 1993, p. 207–216. [Online]. Available : https ://doi.org/10.1145/170035.170072

[10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 1994, pp. 487–499.

[11] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *SIGMOD Rec.*, vol. 29, no. 2, p. 1–12, may 2000. [Online]. Available : https ://doi.org/10.1145/335191.335372

[12] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1994, p. 487–499.

[13] F. Z. El mazouri, M. C. Abounaima, and K. Zenkouar, "Data mining combined to the multicriteria decision analysis for the improvement of road safety : case of france," *Journal of Big Data*, vol. 6, 01 2019.

[14] F. Z. El Mazouri, M. C. Abounaima, and K. Zenkouar, "A selection of useful patterns based on multi-criteria analysis approach," 01 2018.