République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Abderrahman Mira - Bejaia -
Faculté des Sciences Exactes
Département d'Informatique

Mémoire de fin de Cycle
En vue de l'obtention du diplôme de Master en Informatique
Option : Génie Logiciel

Thème

Application de l'Analyse Statistique Implicative dans le domaine de la santé

Réalisé par :

Melle Katia AIT MOUHOUB
Melle Liticia ABBOU

Encadré par :

Mme Hayette KHALED M.C.B

Soutenu le : 01/07/2025

Devant le Jury composé de :

Présidente	**Mme**	**Karima AIT ABDELOUAHAB**	**M.C.B**	**Université de Béjaïa**
Examinateur	**Mr**	**Nabil DJEBARI**	**M.C.B**	**Université de Béjaïa**
Examinatrice	**Mme**	**Samia CHIBANI**	**M.C.A**	**Université de Béjaïa**
Examinatrice	**Mme**	**Sarah CHAABANE**	**M.C.B**	**Université de Béjaïa**

Année Universitaire : 2024-2025

* *Remerciements* *

En premier lieu, nous tenons à remercier le bon Dieu de nous avoir accordé la santé, le courage et la volonté nécessaires pour mener à bien ce travail de fin de cycle.

Nous exprimons notre profonde gratitude à toutes les personnes qui ont, de près ou de loin, contribué à la réalisation de ce mémoire.

Nous adressons nos remerciements les plus sincères à notre encadrante, Mme KHALED HAYETTE, pour sa disponibilité, sa patience, et l'accompagnement précieux qu'elle nous a offert tout au long de cette recherche. Ses conseils avisés et son soutien constant ont été déterminants dans l'aboutissement de ce travail.

Nos remerciements vont également aux membres du jury, qui nous font l'honneur de juger ce mémoire, pour l'intérêt qu'ils y ont porté et pour le temps qu'ils y ont consacré.

Nous exprimons aussi notre reconnaissance à l'ensemble des enseignants du département d'informatique, pour la qualité de leur enseignement, leur engagement et leur contribution à notre formation.

Enfin, nous remercions du fond du cœur nos parents, pour leur soutien moral, matériel et affectif tout au long de notre parcours universitaire. Leur confiance et leur présence ont été un véritable moteur pour avancer.

Nous sommes pleinement conscients que ce travail n'aurait pu voir le jour sans l'aide et le soutien de chacun. À toutes et à tous, nous adressons nos plus sincères remerciements.

* *Dédicaces* *

À nos chers parents,
pour leur amour inconditionnel, leur soutien moral et financier, et leurs encouragements constants tout au long de notre parcours universitaire. Ce travail est avant tout le fruit de leurs sacrifices et de leur confiance.

À nos frères et sœurs,
pour leur présence réconfortante, leurs mots d'encouragement et leur soutien indéfectible, même dans les moments les plus difficiles.

À nos ami(e)s,
qui ont partagé avec nous les joies, les doutes, les efforts et les réussites de ces années d'études. Leur amitié a été un véritable moteur dans cette aventure.

À tous les enseignants du département d'informatique de l'Université Abderrahmane Mira de Béjaïa, pour la qualité de leur enseignement, leur disponibilité et leur dévouement tout au long de notre formation.

À toutes celles et ceux qui, de près ou de loin, ont contribué à la réalisation de ce mémoire, nous vous dédions ce travail avec gratitude et reconnaissance.

Table des matières

Table des figures

Liste des abréviations

ASI	Analyse Statistique Implicative
IERM	Institut de Recherche sur l'Enseignement des Mathématiques
CHIC	Classification Hiérarchique Implicative et Cohésitive
EPS	Éducation Physique et Sportive
IRSB	Inventaire des Rôles Sexués de Bem
OPS	Organisation Panaméricaine de la Santé
CRAN	Comprehensive R Archive Network
MSD	Merck Sharp et Dohme
OMS	Organisation Mondiale de la Santé

Introduction générale

1 Introduction générale

À l'ère de l'information, les organisations, les chercheurs et les entreprises génèrent et accumulent des volumes massifs de données. Toutefois, ces données brutes nécessitent un traitement et une analyse approfondis afin d'en extraire des informations significatives. **La Fouille de Données (ou Data Mining)**, qui constitue un domaine central de l'analyse des données, a précisément pour objectif d'explorer ces ensembles volumineux afin d'en dégager des connaissances exploitables. Elle permet de mettre en évidence des modèles cachés, des tendances récurrentes et des relations invisibles à première vue, en mobilisant des techniques statistiques, des algorithmes d'apprentissage automatique (machine learning) ainsi que des méthodes de modélisation mathématique.

Plusieurs méthodes ont été développées pour extraire des connaissances à partir des données. Parmi elles, une approche moderne se distingue par son orientation innovante et sa capacité à révéler des structures directionnelles complexes : il s'agit de **l'Analyse Statistique Implicative (ASI)**, fondée et développée par **Régis Gras et son équipe.**

L'ASI se concentre sur l'extraction de connaissances implicites et de règles inductives non symétriques, permettant ainsi de découvrir des relations fines entre les variables et les objets. Contrairement à d'autres techniques qui se limitent à des associations statistiques globales, l'ASI cherche à mettre en évidence des invariants au sein des données — des règles qui se maintiennent avec cohérence dans différents contextes. Elle évalue la solidité des relations observées à l'aide d'une métrique spécifique, qui mesure l'étonnement d'observer un faible nombre de contre-exemples à une règle. Cette approche permet ainsi de distinguer les implications les plus robustes au sein des données. Par son alliance entre rigueur mathématique et interprétabilité, l'ASI ouvre de nouvelles perspectives dans l'analyse des données complexes.

Dans ce contexte, une question centrale guide notre travail :

En quoi l'Analyse Statistique Implicative peut-elle représenter une alternative pertinente aux méthodes traditionnelles du Data Mining, et comment peut-elle contribuer à extraire des connaissances directionnelles plus fines à partir de données complexes ?

Ce mémoire poursuit plusieurs objectifs complémentaires :

• Valoriser l'Analyse Statistique Implicative, une méthode encore peu connue et peu utilisée en Algérie, bien qu'elle présente un fort potentiel d'amélioration dans des domaines tels que l'éducation, la médecine, etc.

• Mettre en évidence les limites des métriques d'évaluation classiques utilisées en Data Mining, notamment dans le cadre des règles d'association, afin de justifier l'intérêt d'une méthode alternative.

• Présenter les fondements théoriques de l'ASI, en décrivant ses principes de fonctionnement, ses critères spécifiques, ainsi que les outils informatiques qui lui sont associés (CHIC et RCHIC).

• Explorer les domaines d'application actuels de l'ASI à travers une synthèse de travaux existants, afin de mieux cerner ses usages dans des contextes variés tels que l'éducation, la médecine ou la psychologie.

• Mettre en œuvre l'ASI sur un jeu de données médicales, dans le but d'identifier les facteurs significatifs liés au diabète, à l'aide des graphes implicatifs générés par RCHIC.

Pour répondre à ces objectifs, ce mémoire est structuré en trois chapitres :

Chapitre 1 : Limites du Data Mining et Présentation de l'Analyse Statistique Implicative
Dans ce chapitre, nous présentons les principales métriques utilisées en Data Mining, avant d'en exposer les limites. Nous introduisons ensuite l'ASI comme une méthode alternative, en expliquant son fonctionnement, ses critères, ainsi que les outils CHIC et RCHIC utilisés pour son traitement.

Chapitre 2 : État de l'art sur les domaines d'application de l'Analyse Statistique Implicative
Dans ce chapitre, nous exposons une synthèse des travaux ayant mobilisé l'ASI, afin de mettre en évidence ses usages actuels dans différents domaines (éducation, santé, psychologie, etc.) et d'identifier des pistes pertinentes pour notre propre application.

Chapitre 3 : Application de l'Analyse Statistique Implicative
Dans ce dernier chapitre, nous appliquons l'ASI à un jeu de données médicales, dans le but d'identifier les facteurs les plus significatifs liés au diabète, à l'aide des graphes implicatifs générés par RCHIC.

En fin de mémoire, une conclusion générale viendra synthétiser les principaux apports de notre étude et ouvrir des perspectives pour de futures recherches.

2 Data Mining et Présentation de l'Analyse Statistique Implicative (ASI)

2.1 Introduction

Dans ce premier chapitre, nous allons d'abord présenter le Data Mining et les règles d'association, en expliquant leurs principes et critères d'évaluation. Nous mettrons ensuite en évidence certaines limites de cette approche.

Nous introduisons ensuite une approche plus avancée : l'Analyse Statistique Implicative. Cette méthode sera étudiée en détail, car elle constitue le cœur de notre projet. Nous présenterons notamment les outils CHIC et RCHIC, qui permettent d'exploiter l'ASI de manière efficace pour analyser et interpréter les données.

2.2 Data Mining

Le **"Data Mining"** que l'on peut traduire par **"Fouille de Données"** apparaît au milieu des années 1990 aux États-Unis comme une nouvelle discipline à l'interface de la statistique et des technologies de l'information : bases de données, intelligence artificielle, apprentissage automatique (≪ machine Learning ≫). Ses premières applications furent menées sur l'analyse du panier de la ménagère (en anglais Market Basket Analysis). Au départ, Data Mining s'est donc intéressée aux bases de données des supermarchés afin d'identifier les améliorations possibles des ventes d'articles grâce à des décisions stratégiques.[1]

L'objectif poursuivi par le Data Mining comme illustré dans la figure (1.1), est donc celui de la valorisation des données contenues dans les importantes bases de données. En effet, pour exploiter un volume important de données brutes, une étape de traitement est nécessaire afin de les mettre sous un format adéquat. Elles sont ensuite étudiées pour retrouver des éléments fréquents dans la base de données ou des règles. Ces derniers constituent des connaissances de valeur pour une prise de décision par la suite. La fouille de données s'apparente généralement à deux notions fondamentales qui sont : **les motifs fréquents et les règles d'association.** Ces deux notions sont fondamentales et font la réussite et l'extension de Data Mining dans divers domaines. Ainsi, grâce à son analyse, le Data Mining est utilisée souvent à des fins de classification, de prédiction et d'apprentissage.[1]

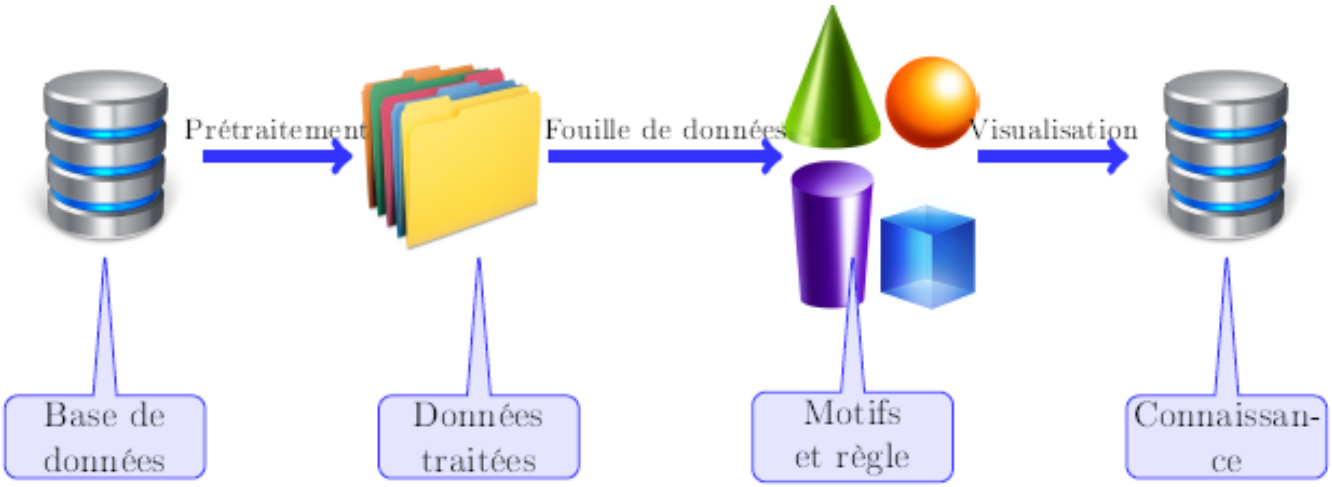

Figure (1.1) : Processus d'extraction de connaissances. [1]

2.2.1 Les règles d'association et les métriques classiques utilisées

Les règles d'associations sont introduites par Agrawal et al. au début des années 90 pour exprimer simplement des tendances implicatives entre les attributs d'une table relationnelle. Une règle d'association est de la forme : « Si Condition alors Résultat », notée condition → résultat. Ces règles signifient que si un enregistrement de la table vérifie la condition, alors il vérifie sûrement également le résultat. Les règles sont dotées de plusieurs mesures de qualité. Les plus utilisées sont le support, la confiance et le lift. [2] [3]

- **Support**

– Une règle donnée : « Si A → B », le support de cette règle se définit comme le numéro de fois ou fréquence (relative) avec laquelle A et B figurent ensemble dans une base de données transactionnelle.

– Support peut être défini individuellement pour les items, mais aussi peut être défini pour la règle.

– La première condition nous pouvons imposer pour limiter le nombre de règles est d'avoir un seuil de support minimum. [4]

$$\textbf{Support}(A \rightarrow B) = \frac{\text{Nombre de transactions contenant A et B}}{\text{Nombre total de transactions}}$$

• **Confiance**

– Une règle donnée « Si A $\rightarrow$ B », la confiance de cette règle correspond au quotient du support de la règle (A et B) par le support de l'antécédent A uniquement.

– La confiance mesure la précision de la règle. Elle indique la proportion d'entités vérifiant le conséquent B parmi celles qui vérifient la prémisse A.

– La deuxième condition que nous pouvons imposer pour limiter le nombre de règles est d'avoir un seuil de confiance minimum. [4]

$$\textbf{Confiance}(A \rightarrow B) = \frac{\text{Support(A} \rightarrow \text{B)}}{\text{Support(A)}} = \frac{\text{Nombre de transactions contenant A et B}}{\text{Nombre de transactions contenant A}}$$

• **Lift**

– Est défini de la manière suivante :

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

– Lift = 1 ou très proche de 1 indique que la relation est produite au hasard.

– Lift supérieur à 1 traduit une corrélation positive de A et B, et donc le caractère significatif de l'association.

– Lift < 1 indique une relation réellement faible.

– Malheureusement n'existe pas de valeurs critiques pour déterminer c'est quoi « loin de 1 » ou au dessous de 1. [4]

2.2.2 Limites des métriques classiques

Les différentes techniques de Data Mining se basent sur le support et la confiance pour l'extraction de règles de la forme (A $\rightarrow$ B) alors que ces deux métriques ne sont pas suffisantes pour assurer une bonne qualité des règles extraites pour les raisons qui suivent :

• Problème du support élevé

Le support d'une règle correspond à la fréquence d'apparition de A et B ensemble dans la base de données. Si on fixe un seuil de support trop élevé, on élimine automatiquement les règles qui apparaissent rarement. Pourtant, ces règles peu fréquentes peuvent être très fiables (grande confiance), ce qui signifie qu'elles sont intéressantes malgré leur faible occurrence. Ces règles rares mais précieuses sont appelées "pépites de connaissance". [5]

• Problème de la confiance

La confiance d'une règle (A $\rightarrow$ B) est le pourcentage de cas où B est vrai parmi ceux où A est vrai. La confiance augmente lorsque le nombre de contre-exemples (c'est-à-dire les cas où A est vrai mais pas B) diminue. Cependant, cette augmentation suit un rythme fixe quel que soit le nombre total de sujets. Cela montre que la confiance peut être trompeuse, car elle ne tient pas compte de l'importance relative des données. [5]

• Problème de la confiance élevée

Une confiance élevée signifie que lorsque A est vrai, B l'est aussi souvent. Mais si B est très courant dans l'ensemble des données, cela ne prouve pas que a un vrai effet sur. Dans ce cas, la règle est trop évidente on l'appelle **règle triviale** et donc peu utile. [5]

• Sensibilité aux contre-exemples

La confiance diminue si on trouve plus de contre-exemples (des cas où A est vrai mais pas B). Cependant, tous les contre-exemples n'ont pas la même importance, et certains peuvent être dus au bruit (des erreurs ou des anomalies). Cela signifie que la confiance est une mesure fragile face aux données imparfaites. [5]

• Limitation aux données binaires

Les techniques de Data Mining classiques fonctionnent surtout sur des données binaires (présence/absence d'un élément). Elles ne sont pas adaptées aux autres types de données, ex : données numériques ou catégorielles complexes. [5]

Vu la nécessité de compléter le support et la confiance par d'autres mesures d'intérêt, et dans le but de pallier aux limites des autres métriques, Régis et al ont proposé une mesure qui rapproche la règle de l'implication logique. Cette mesure tient compte des règles d'association transactionnelles, formulées ainsi : "si des articles (a) sont présents dans le panier, alors d'autres articles (b) y sont généralement aussi." Contrairement à l'implication logique qui exige une stricte égalité, cette contrainte n'est pas requise dans les règles d'association. Elle évalue l'invraisemblance d'un faible nombre de contre-exemples (n_{ab}) par rapport à l'hypothèse d'indépendance entre (a) et (b). Cette mesure prend en compte la non-satisfaction de l'implication (liée aux contre-exemples) et est asymétrique. Cet indice est appelé **l'intensité d'implication**. Ce dernier est un indice de quasi-implication développé par Gras qui est au fondement d'une méthode d'analyse exploratoire des données nommée **Analyse Statistique Implicative (ASI)**. [3]

2.3 Origine et définition de l'Analyse Statistique Implicative (ASI)

L'ASI, à l'origine développée par **Régis Gras et ses collaborateurs**, est apparue suite aux difficultés rencontrées pour évaluer le niveau des élèves dans un test Mathématique. Régis a enseigné les Mathématiques à tous les niveaux d'enseignements en France et même en Afrique francophone, au Moyen-Orient et en Amérique Latine dans le cadre des missions des affaires étrangères, ce qui lui a permis de percevoir les différences de représentation des notions mathématiques. [5]

En 1969, à l'ouverture de l'Institut de Recherche sur l'Enseignement des Mathématiques (I.R.E. M) pour l'étude des problèmes de formation continue des enseignants et de changement de programmes d'enseignement, Régis qui a participé à cet institut, il a rencontré des difficultés d'apprentissage tant au niveau de l'école primaire que du collège et du lycée ainsi que chez l'adulte plus au moins chevronné. Il a utilisé des méthodes permettant de formaliser des énoncés tels que « quand l'élève réussit ceci alors, on générale il réussit cela» ou l'inverse. Afin d'accorder une mesure à cette quasi-implication et structurer hiérarchiquement l'ensemble des réussites on catégories aucune méthode statistique connue ne permettait de répondre de façon globale et symétrique, **d'où les premiers pas de l'ASI (1979)**. Depuis, elle est toujours en développement par lui-même, ses collaborateurs et d'autres chercheurs. [5]

L'analyse statistique implicative est une méthode **non symétrique** d'analyse de données conçue par Régis Gras et qui a un impact significatif sur divers domaines allant de la recherche pédagogique et psychologique à l'exploration de données. Elle présente un véritable cadre paradigmatique de traitement statistique de la causalité et de la complexité. [6]

Le fondement théorique de l'ASI, repose sur le concept de **quasi-implication** qui est représentée par la relation **(si a, alors généralement b)**. Contrairement aux implications logiques strictes, les quasi-implications tolèrent la présence de contre-exemples. L'ASI s'intéresse à des règles **asymétriques**, et se focalise sur les cas où cette implication (si a alors b) n'est pas vérifiée qui, apparait dès que (a) étant vrai, (b) est faux. C'est sur ces **contre-exemples** que reposent les mesures de qualité des règles implicatives. [3]

2.3.1 Fonctionnement de L'ASI

Notons A et B les sous ensembles respectifs de E d'individus qui vérifient respectivement les variables booléennes a et b. Soient $\overline{A}$ et $\overline{B}$ les ensembles complémentaires de A et B respectivement dont les cardinaux de A et B sont : $card(E) = n, \quad card(A) = n_a, \quad card(B) = n_b, \quad card(\overline{A}) = n_{\overline{a}} = n - n_a, \quad card(\overline{B}) = n_{\overline{b}} = n - n_b.$

Pour une règle quelconque règle a $\rightarrow$ b, observée dans E, l'ASI prend en considération la non satisfaction de cette implication, qui apparait lors a est vrai, b est faux. Elle représente le nombre de contres exemples $n_{a\wedge\overline{b}}$ à cette règle observée dans l'intersection $A \cap \overline{B}$. L'ASI consiste à comparer le nombre de contres exemples $n_{a\wedge\overline{b}}$ avec le nombre de contres exemples qui apparaitraient lors d'un choix aléatoire et indépendant de deux parties de même cardinaux respectifs que A et B (figure (1.2)) . Pour formaliser l'hypothèse que a et b sont indépendants, les auteurs ont considéré comme I.C .Lerman deux parties quelconques X et Y de E, choisies aléatoirement et indépendamment (absence de lien a priori entre ces deux parties) et de même respectifs que A et B. Soit $\overline{Y}$ et $\overline{B}$ les complémentaires respectifs de Y et de B dans E de même cardinal. Soit α un réel quelconque de l'intervalle [0,1]. [3]

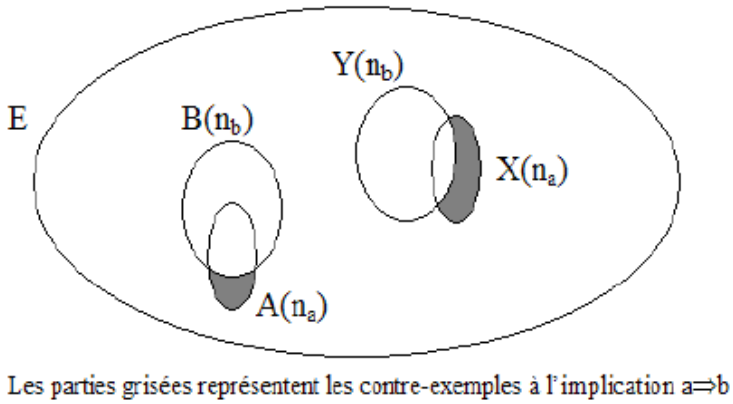

Figure (1.2) : Représentation par les diagrammes d'Euler. [3]

Définition 1 : la quasi-implication a $\rightarrow$ b est admissible au niveau de confiance 1- α si et seulement si [3] :

$$\Pr\left[card(X \cap \overline{Y}) \leq card(A \cap \overline{B})\right] \leq \alpha$$

Définition 2 : on appelle l'intensité d'implication de la quasi règle a $\rightarrow$ b, le nombre

$$\varphi(a,b) = 1 - \Pr\left[card(X \cap \overline{Y}) \leq card(A \cap \overline{B})\right] si \quad n_b \neq n \quad et \quad \varphi(a,b) = 0 \quad si \quad n_b = n$$

Cet indice permet de mesurer l'étonnement du au fait que le nombre de contres exemples à la règle a $\rightarrow$ b est petit par rapport aux grands nombres d'instances, alors que a et b sont supposés indépendants.[3]

2.3.2 Les critères de l'ASI

L'Analyse Statistique Implicative repose sur plusieurs critères permettant de mesurer la force et la qualité des relations implicatives entre les données. Initialement, l'intensité d'implication constituait le critère principal utilisé pour détecter et valider ces relations. Cependant, certaines limites de ce premier indicateur ont conduit au développement de critères complémentaires, tels que l'intensité entropique, l'implifiance et la confiance combinée à l'intensité d'implication . Ces outils permettent d'affiner l'analyse en tenant compte de la fiabilité, de la stabilité et de la précision des implications mises en évidence.

• L'intensité d'implication

L'intensité d'implication est une mesure probabiliste plutôt qu'une simple fréquence. Elle permet de déterminer si une relation de quasi-implication entre deux variables binaires a et b doit être retenue ou non. Ce modèle de quasi-implication est particulièrement utile pour évaluer l'étonnement lié à la rareté des contre-exemples par rapport au nombre élevé de cas où l'implication est observée. Il s'agit ainsi d'un indicateur de la pertinence inductive et du pouvoir informatif de l'implication. [7]

Par conséquent, si la règle est triviale, par exemple lorsque B est très étendu ou coïncide avec l'ensemble E, cet étonnement diminue fortement. D'ailleurs, Gras (1996) a démontré que cette trivialité entraîne une intensité d'implication très faible, voire nulle : lorsque n_a est fixé et que A est inclus dans B, alors si n_b tend vers n (autrement dit, si B $\ll$ croît $\gg$ jusqu'à couvrir E),

alors intensité d'implication $\varphi(a,b)$ tend vers 0. C'est pourquoi, par continuité, il définit cette intensité comme étant nulle lorsque $n_b = n$.

De même, dans le cas où A est inclus dans B, l'intensité $\varphi(a,b)$ peut rester inférieure à 1 si la confiance inductive, évaluée à travers l'étonnement statistique, s'avère insuffisante. [7]

Limites de la mesure d'intensité d'implication

L'intensité d'implication présente l'inconvénient d'être peu discriminante quand les cardinaux étudiés sont grands, car ses valeurs peuvent être souvent proches de 1 alors que A n'est pas inclus dans B. D'où la nécessité d'adapter le concept d'intensité à des situations où les populations en jeu deviennent très importantes. Pour résoudre ce problème, Gras et al ont proposé dans de moduler les valeurs de l'intensité d'implication par un indice de quasi-implication fondé sur l'entropie de Shannon : l'indice d'inclusion. L'indice formé s'appelle **intensité entropique**. [3]

Les utilisateurs de l'implication entropique ont apprécié la capacité à accepter plus facilement la grande taille de l'échantillon des sujets considérés. D'où son intérêt pour ce que l'on appelle les « big data ». Ce dernier présente aussi un caractère jugé trop ad-hoc par les familiers de l'ASI. Ceci a motivé les auteurs à créer un nouvel indice appelé **implifiance**. Tous ces indices prennent en compte la contraposée $\overline{B} \to \overline{A}$ qui permet de renforcer l'affirmation de la relation implicative de a sur b. Elle pourrait également contribuer à répondre aux problèmes de l'approche support-confiance puisque si on a un support très petit avec une confiance très élevée c'est-à-dire si A et B sont petits relativement à E leurs complémentaires seront grands et réciproquement. [3]

• L'intensité entropique

C'est la version améliorée de l'intensité d'implication pour le traitement de données volumineuses basée sur une pondération par l'entropie de Shannon. L'intensité entropique de la règle a $\to$ b est définie par : $\Psi(a,b) = (\varphi(a,b) \times \tau(a,b))^{1/2}$

Où $\varphi(a,b)$ est **l'intensité d'implication** et $\tau(a,b)$ est **l'indice d'inclusion**. [8]

Definition 1 : L' indice d'inclusion de I_a, support de a, dans I_b, support de b, est le nombre qui intègre l'information délivrée par la réalisation d'un faible nombre de contre-exemples, d'une part à la règle a $\to$ b et, d'autre part, à la règle $\overline{b} \to \overline{a}$. [8]

$$\tau(a,b) = \left(1 - h_1^2(t)\right) \times \left(1 - h_2^2(t)\right)$$

• L'implifiance

C'est une mesure de l'implication statistique qui tient compte à la fois de l'implication directe (a $\rightarrow$ b), de sa contraposée ($\overline{b} \rightarrow \overline{a}$) et du degré de confiance associé à cette relation. En intégrant ces trois éléments, elle permet d'évaluer de manière plus complète la pertinence et la fiabilité d'une relation entre deux variables binaires, offrant ainsi une vision plus équilibrée et précise de l'implication statistique. [7]

Sa valeur est donnée par la formule suivante [3] :

$$\phi(a,b) = \varphi(a,b) \times \left[C_1(a \rightarrow b) \times C_2(\overline{b} \rightarrow \overline{a})\right]^{\frac{1}{4}}$$

où :

– $\varphi(a,b)$ est la **force d'implication directe** $a \rightarrow b$,

– $C_1(a \rightarrow b)$ est le **degré de confiance de la règle directe**,

– $C_2(\overline{b} \rightarrow \overline{a})$ est le **degré de confiance de la contraposée logique**, c'est-à-dire : si b est absent, alors a l'est aussi.

Ces deux degrés de confiance sont définis comme suit [3] :

$$C_1(a,b) = Fr[Y \mid X] = \frac{card(X \cap Y)}{card(X)} = \frac{n_{a \wedge b}}{n_a}$$

$$C_2(\overline{b} \rightarrow \overline{a}) = Fr[\overline{X} \mid \overline{Y}] = \frac{card(\overline{X} \cap \overline{Y})}{card(\overline{Y})} = \frac{n_{\overline{a} \wedge \overline{b}}}{n_{\overline{b}}}$$

• L'indice d'implication combiné avec la confiance

C'est un critère utilisé en Analyse Statistique Implicative pour améliorer la lecture et l'interprétation des graphes d'implication. Il associe la force du lien entre deux éléments (mesurée par l'indice d'implication) et la fiabilité de cette relation (mesurée par la confiance). L'ajout de la confiance à chaque règle permet de mieux distinguer l'importance des liens, tandis que l'utilisation d'un seuil de confiance rend le graphe plus lisible en ne conservant que les relations les plus fiables. [9]

2.3.3 Logiciel CHIC

Le logiciel CHIC constitue la réalisation informatique des travaux menés sur l'analyse implicative. Initialement développé en Pascal, il a ensuite été réécrit en C++ sous Windows, avec d'importants ajouts fonctionnels et une interface plus conviviale. Depuis, il a fait l'objet de nombreuses évolutions, tant sur le plan pratique que théorique, intégrant divers nouveaux modes de calcul. [10]

CHIC permet d'effectuer plusieurs traitements statistiques fondés sur le principe de l'étonnement statistique, notamment l'analyse des similarités et l'analyse implicative. Il est capable de traiter rapidement de grands tableaux de contingence, allant jusqu'à une taille de 200 × 100000, cette capacité dépendant des ressources matérielles de l'ordinateur (puissance de calcul et mémoire). [11]

Le logiciel permet également de sauvegarder les calculs intermédiaires, ce qui optimise les traitements lors d'analyses répétées sur un même jeu de données. Il prend en charge différents types de variables : binaires, fréquentielles, modales ou encore en intervalles, ce qui en fait un outil polyvalent, adapté à des analyses où les variables ne sont pas homogènes. [10]

• Fonctionnalités principales du logiciel CHIC

– Fourniture de statistiques descriptives : moyennes, écarts-types, coefficients de corrélation.

– Réalisation d'une classification hiérarchique des similarités, selon l'algorithme de la vraisemblance de I.C. Lerman.

– Exécution d'une analyse implicative basée sur la méthode de R.Gras à savoir l'ASI, avec la possibilité de choisir entre la méthode classique et la méthode entropique.

– Manipulation des variables : ajout, suppression, conjonction et disjonction.

– Calcul des intensités d'implication, des similarités, des corrélations linéaires et des croisements deux à deux des variables.

– Génération d'un graphe implicatif selon différents seuils, avec identification des sujets contributeurs aux chemins significatifs (selon les travaux de M. Bailleul).

– Production d'une classification cohésive en arbre, avec ses niveaux significatifs, ainsi que la contribution des sujets et des catégories de sujets. [12]

La version actuelle de ce logiciel, appelée **RCHIC**, est implémentée sous **R** et a été développée par **Raphaël Couturier**. [3]

2.3.4 Logiciel RCHIC

RCHIC est un **package** pour **R** qui implémente la plupart des outils de l'Analyse Implicative Statistique, fonctionne sous Windows, Linux et MacOS conçu à partir de la version en C++. RCHIC subie régulièrement des mise à jour, ce qui le met au même niveau avec les déférents développements théorique de l'ASI. [3]

• Les données traitées par RCHIC

Les données sont disposées sous forme d'un tableau numérique, dans lequel à chaque variable que nous souhaitons évaluer, nous faisons correspondre le résultat de l'évaluation de chaque objet ou individu à cette variable. [10]

Les variables à étudier peuvent avoir différents types, à savoir : binaire, modale et fréquentielle, quantitative ou intervalle. De plus elles peuvent être principales, c'est-'a-dire qu'elles interviennent directement dans tous les calculs ou elles peuvent être supplémentaires comme il est fait en analyse factorielle. [10]

Les variables modales et fréquentielles doivent avoir une valeur réelle comprise entre 0 et 1. Les valeurs des variables quantitatives sont normalisées dans l'intervalle [0-1] en divisant toutes les valeurs par la valeur maximum obtenue par la variable. Il faut effectuer cette manipulation à l'aide d'un tableur à ce stade du traitement.[10]

.

Les variables-intervalles sont automatiquement découpées en différents intervalles par un algorithme approprie, de type « nuées dynamiques »qui, à partir d'un nombre d'intervalles choisi par l'utilisateur, constitue des intervalles tout en maximisant la variance inter-classe. Ayant formaté les données, il faut sauvegarder le fichier avec le type « **CSV**» qui est un format standard, chaque champ étant séparé par un point virgule, où les variables sont disposées en colonne [3] (ici Affectueux, Agile, Agressif, Angoissant, Attirant, Beau, Bete, Blanc, Bon) et les individus en ligne (ici Aigle, Ane, Autruche, Baleine, Bouc, Canard, Chamois, Chat, Chien, Cigale) et à chaque variable nous faisons correspondre le résultats de l'évaluation de chaque individu à cette variable comme montré sur la figure (1.3).

	X	Affectueux	Agile	Agressif	Angoissant	Attirant	Beau	Bete	Blanc
1	Aigle	0	1	0	1	1	1	0	0
2	Ane	0	0	0	0	0	1	1	0
3	Autruche	0	0	1	1	0	0	1	0
4	Baleine	0	0	0	1	1	1	0	1
5	Bouc	0	1	1	0	1	0	1	0
6	Canard	0	0	1	0	0	1	1	0
7	Chamois	0	1	0	0	0	1	0	0
8	Chat	1	1	0	0	1	1	1	0
9	Chien	1	0	0	0	0	0	1	0
10	Cigale	0	0	0	0	0	1	0	0
11	Corbeau	0	0	1	1	0	1	0	0
12	Couleuvre	0	0	0	1	0	0	0	0
13	Crocodile	0	0	1	1	0	0	0	0
14	Crotale	0	1	1	1	0	0	0	0
15	Dauphin	0	0	0	0	1	1	0	0
16	Fourmi	0	0	0	0	1	0	0	0

Figure (1.3) : Extrait du jeu de données sous type.csv.

Ce logiciel a pour objectif de découvrir les implications les plus pertinentes entre les variables d'un ensemble de données avec l'algorithme apriori et calcule pour chaque implication : le nombre d'occurrences, le support, la confiance, l'indice d'implication, l'indice entropique, etc. Toutes ces informations sont enregistrées sous forme d'un tableau dans un fichier appelé **transaction.out**. [3] [5] Voici un extrait de ficher (figure (1.4)).

hyp -> con	occurrence(hyp)	occurrence(con)	support(rule)	confidence	classical index	entropic index
Vif -> Laid	24.000000000000000	21.000000000000000	58.5365853658536537	45.8333333333333286	28.7979028077989376	18.2252390572530025
Laid -> Vif	21.000000000000000	24.000000000000000	51.2195121951219505	52.3809523809523867	25.9924170995981747	17.7134148825696940
Vif -> Mechant	24.000000000000000	21.000000000000000	58.5365853658536537	50.0000000000000000	39.0601490037602730	26.1210959571991310
Mechant -> Vif	21.000000000000000	24.000000000000000	51.2195121951219505	57.1428571428571388	37.4091227480263484	26.9388020010602069
Vif -> Craintif	24.000000000000000	20.000000000000000	58.5365853658536537	50.0000000000000000	45.7468982097617243	31.1250428189317141
Craintif -> Vif	20.000000000000000	24.000000000000000	48.7804878048780495	60.0000000000000000	44.8235759571666392	33.6505921217859552
Vif -> Beau	24.000000000000000	20.000000000000000	58.5365853658536537	50.0000000000000000	45.7468982097617243	31.1250428189317141
Beau -> Vif	20.000000000000000	24.000000000000000	48.7804878048780495	60.0000000000000000	44.8235759571666392	33.6505921217859552
Vif -> Sournois	24.000000000000000	20.000000000000000	58.5365853658536537	58.3333333333333357	68.2689333472636122	50.7569670798263743
Sournois -> Vif	20.000000000000000	24.000000000000000	48.7804878048780495	70.0000000000000000	72.1052533635089645	59.1530794053556264
Vif -> Gentil	24.000000000000000	19.000000000000000	58.5365853658536537	33.3333333333333286	15.5864469365624370	8.5585255999764804
Gentil -> Vif	19.000000000000000	24.000000000000000	46.3414634146341484	42.1052631578947327	10.3322237830777102	6.4151007680884460
Vif -> Discret	24.000000000000000	19.000000000000000	58.5365853658536537	50.0000000000000000	52.3427262577863956	36.1404178824720717
Discret -> Vif	19.000000000000000	24.000000000000000	46.3414634146341484	63.1578947368421026	52.9890737190067824	41.3695701873921706
Vif -> Agressif	24.000000000000000	19.000000000000000	58.5365853658536537	41.6666666666666714	31.2479599633716845	19.4955586993561205
Agressif -> Vif	19.000000000000000	24.000000000000000	46.3414634146341484	52.6315789473684177	26.8364409099145504	18.9320125931368786
Vif -> Bete	24.000000000000000	18.000000000000000	58.5365853658536537	33.3333333333333286	19.9546674047052939	11.2617720073725991
Bete -> Vif	18.000000000000000	24.000000000000000	43.9024390243902474	44.4444444444444429	13.4641074337442319	8.8063150495508218
Vif -> Malin	24.000000000000000	18.000000000000000	58.5365853658536537	50.0000000000000000	58.6821996156977974	41.0360072368708941
Malin -> Vif	18.000000000000000	24.000000000000000	43.9024390243902474	66.6666666666666572	61.6825194390336122	49.9890899448439541
Vif -> Mysterieux	24.000000000000000	18.000000000000000	58.5365853658536537	58.3333333333333357	78.5981220318533644	59.5587051087599875

Figure (1.4) : Extrait du fichier transaction.out

.

.

RCHIC propose d'organiser les implications selon trois modes principaux de représentation (voir la figure (1.5)). Le graphe implicatif qui donne une classification orientée ainsi que l'arbre des similarités et l'arbre hiérarchique (ou cohésif) qui fournissent une classification non orientée.[5]

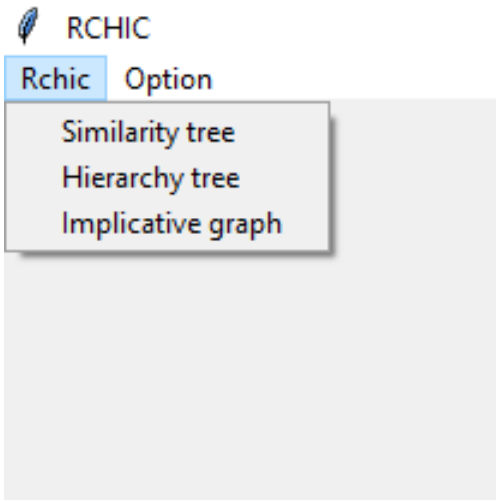

Figure (1.5) : Les modes de représentation proposés par RCHIC.

• **L'arbre des similarités**

L'algorithme utilisé est l'algorithme de la vraisemblance du lien (AVL) de Lerman (1981), il calcule pour chaque paire de variables la similarité entre celles-ci. Puis il agrège des classes qui sont établies à leur tour par d'autres classes. Sur l'arbre de la figure (1.6) les variables Fort et Puissant sont dans un premier temps les variables les plus similaires. Ensuite l'algorithme forme la classe (Gros, Lourd), puis à l'itération trois il forme la classe (Fin, Raye), et dans l'itération suivante apparait la classe (Grand, Violent), ainsi de suite jusqu'à la fin du graphe. Les niveaux marqués par un trait rouge sont les niveaux les plus significatifs par rapport aux autres niveaux. [10]

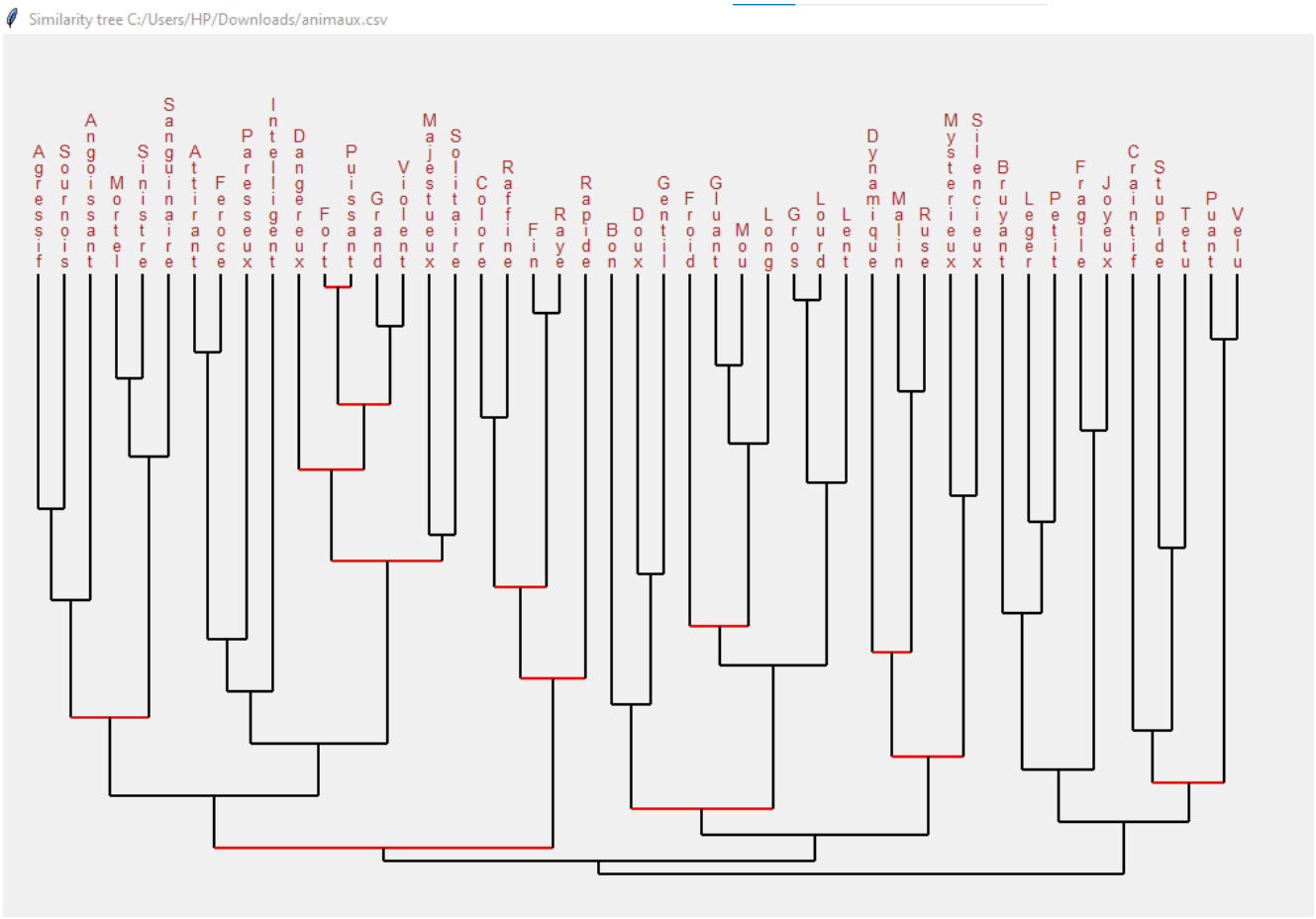

Figure (1.6) : Exemple d'arbre des similarités.

Par défaut toutes les variables impliquées dans le graphe sont représentés dans la zone de travail. Lors de l'interprétation, l'utilisateur peut identifier les variables les plus pertinentes et supprimer celles qui sont moins utiles. Cela peut être fait via une boîte de dialogue spécifique qui permet de mettre à jour le graphique en temps réel. L'utilisateur a la possibilité d'ajouter ou de retirer des variables à tout moment, selon les besoins de l'analyse en cours.[3] La figure (1.7) illustre cette boîte de dialogue.

Figure (1.7) : Exemple d'une boîte de dialogue.

• L'arbre cohésif (hiérarchique)

Dans cet arbre, des classes de variables ou de règles entre variables sont constituées à partir des implications entre celles-ci. L'algorithme agrège à chaque étape les variables conduisant à la cohésion la plus forte à cette étape, la figure (1.8) représente un exemple d'un arbre cohésif. Au premier niveau de la hiérarchie, on remarque que la classe (Ruse, Malin).Elle représente le fait que la variable « Ruse » implique la variable « Malin » avec une intensité plus forte que tous les autres couples de variables. Ce premier niveau de la hiérarchie est d'ailleurs significatif comme l'indique la flèche rouge (en gras sur la figure). Au deuxième niveau, la classe (Puissant, Grand) est formée. Au troisième niveau, la classe (Agressif, Sournois) est formée. Au quatrième niveau, la classe (Silencieux, Mystérieux) est formée. Au cinquième niveau, la classe (Sanguinaire,(Agressif, Sournois)) est formée. Cette classe à trois composantes admet la plus forte cohésion parmi celles de toutes les classes possibles à trois composantes. L'Algorithme arrête son processus de construction dès que la cohésion entre les variables ou entre règles devient trop faible. [10]

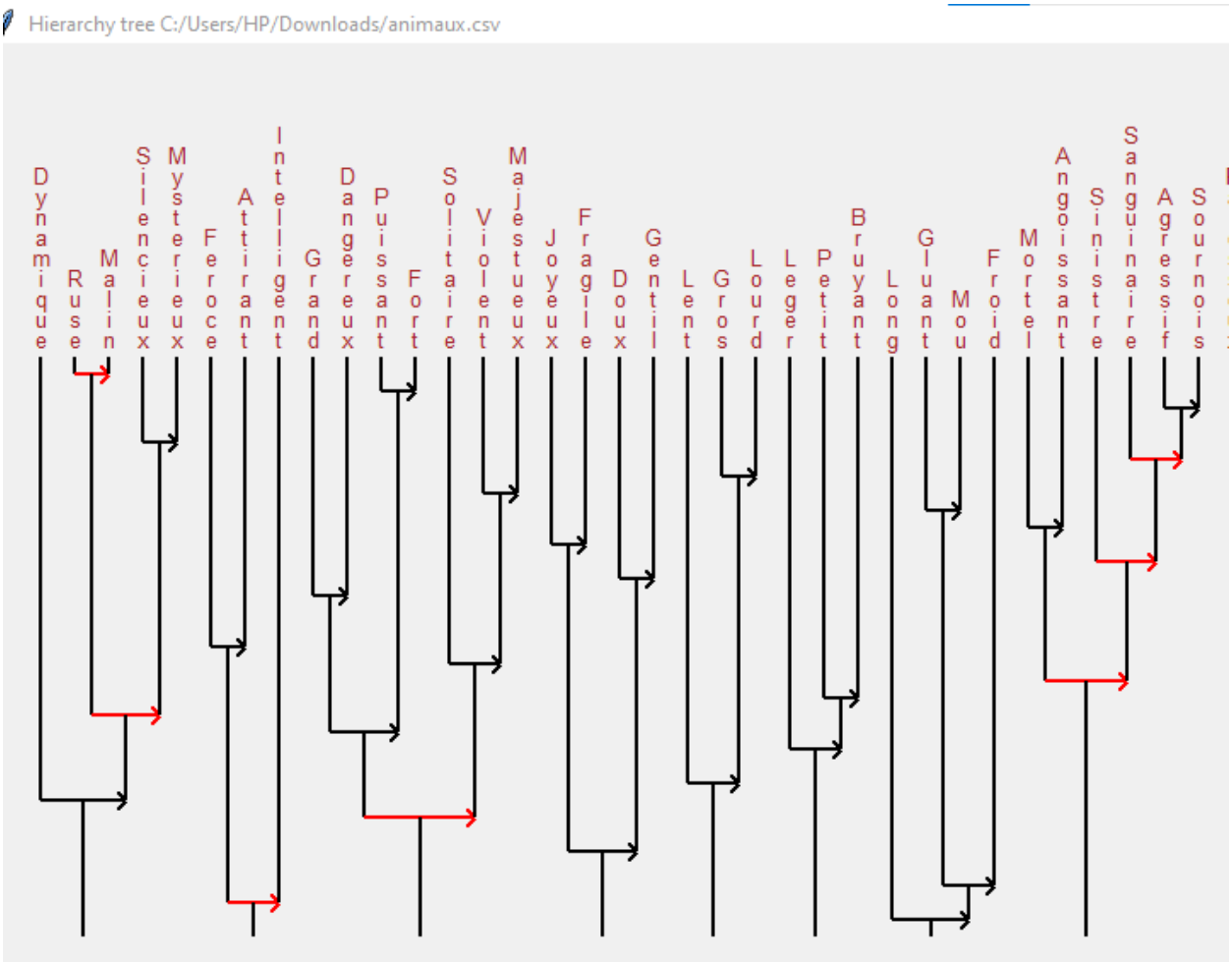

Figure (1.8) : Exemple d'arbre cohésif (hiérarchique).

• **Le graphe d'implication**

C'est un graphe sur lequel les variables qui possèdent une Intensité d'implication supérieure à un certain seuil sont reliées par une flèche représentant l'implication. La figure (1.9) représente un exemple de graphe implicatif. RCHIC offre 4 seuils d'implications différents identifiés par différentes couleurs. L'utilisateur peut disposer les valeurs comme il le souhaite. Nous avons par exemple la variable Ruse qui implique la variable Malin avec l'intensité d'implication 0.99 (flèche rouge) et la variable Gros qui implique la variable Lourd avec l'intensité d'implication 0.95 (flèche verte), etc. [10]

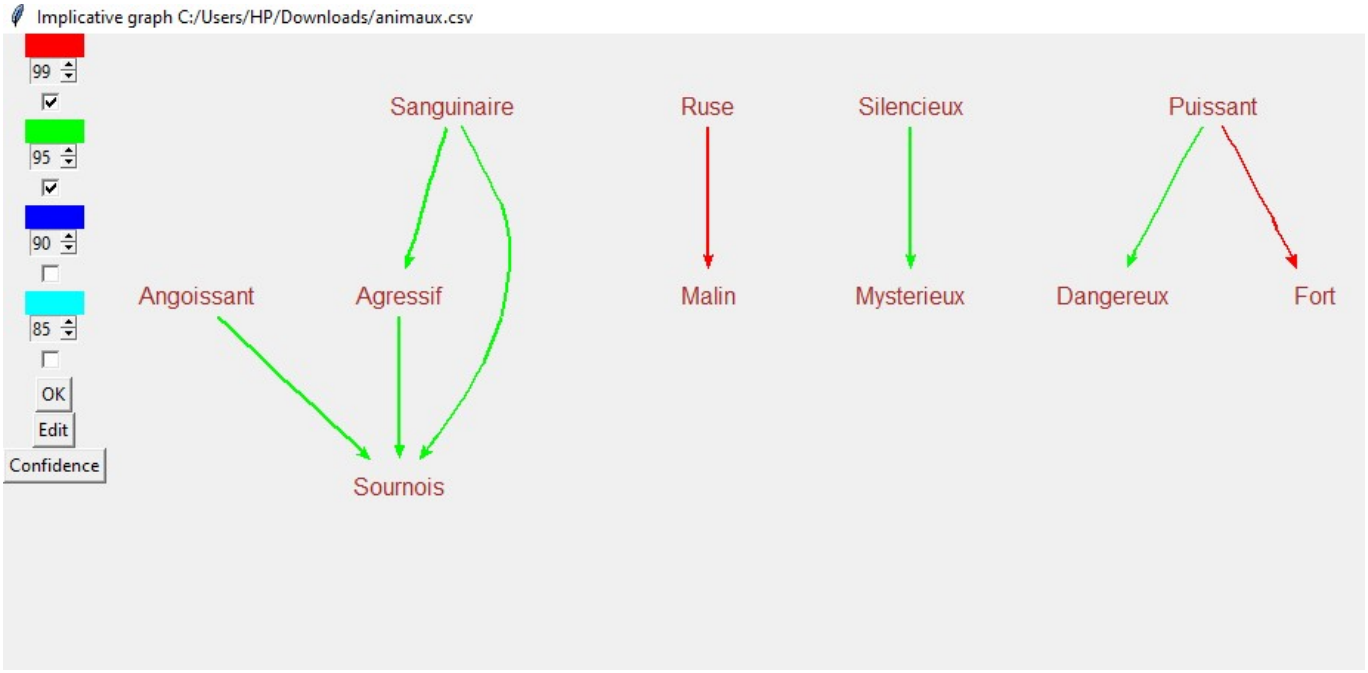

Figure (1.9) : Exemple de graphe d'implication.

RCHIC offre aux utilisateurs plusieurs modes pour le calcul, les modes existants sont : indice classique(intensité d'implication), indice entropique, confiance combinée à l'intensité d'implication et implifiance [5] (voir la figure (1.10)).

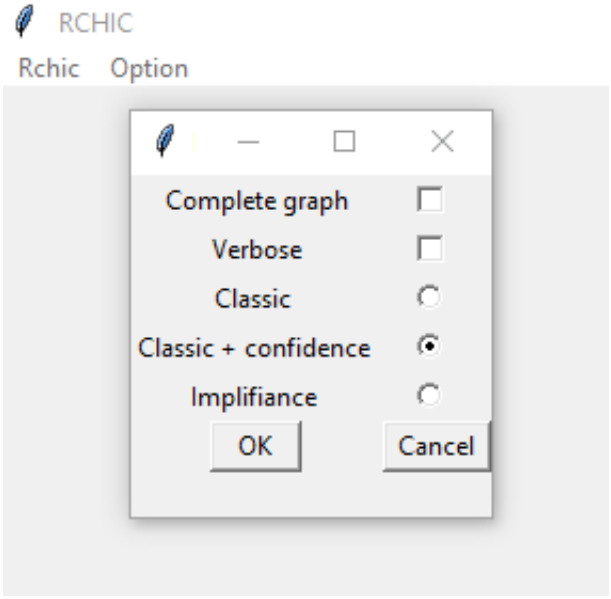

Figure (1.10) : Les différents modes de calcul.

2.4 Conclusion

Dans ce chapitre nous avons présenté les fondements du Data Mining et ses limites, puis nous avons introduit la méthode d'analyse de données non symétrique ASI comme nouvelle approche pour pallier ces limites. De plus nous avons présenté le logiciel de traitement utiliser CHIC et sa version RCHIC , dans laquelle nous avons détaillé les fonctionnalités que nous pouvons trouver sur ce logiciel.

Dans le deuxième chapitre nous allons présenter les domaines d'application de l'ASI.

3 État de l’art sur les domaines d’application de l’Analyse Statistique Implicative (ASI)

3.1 Introduction

L'Analyse Statistique Implicative a été utilisée dans de nombreux domaines de recherche. Dans ce deuxième chapitre, nous présenterons quelques-uns de ces domaines, tels que la psychologie, la médecine ou encore l'éducation. À travers des exemples concrets, nous mettrons en évidence la diversité des usages de cette méthode et son intérêt dans des contextes scientifiques variés.

3.2 Domaine de la psycho-sociologie de l'éducation

Dans cette section, nous proposons d'explorer deux thématiques majeures au sein de la psycho-sociologie de l'éducation. La première porte sur l'application d'une méthode implicative pour analyser les représentations sociales et les dynamiques sexuées dans le cadre de l'éducation physique et sportive (EPS). La seconde aborde les impacts récents de la pandémie de Covid-19, en particulier les effets de l'enseignement à distance sur la santé mentale des étudiants issus de milieux multiculturels. Ces deux axes permettent d'éclairer des enjeux contemporains importants qui façonnent les expériences éducatives des élèves et étudiants.

3.2.1 Application d'une méthode implicative à l'analyse des représentations sociales et des dynamiques sexuées en EPS [13]

L'éducation physique et sportive (EPS) a longtemps été un domaine où les différences sexuées se sont manifestées de manière marquée, tant dans les pratiques que dans les représentations des élèves. Cette étude s'inscrit dans la continuité des travaux en psycho-sociologie, notamment ceux de Davisse (1991), Scraton (1992) et Penney (2002), qui ont analysé les inégalités de genre en EPS. Ces travaux, qu'ils soient issus de la tradition française ou anglo-saxonne, ont permis de mettre en évidence des mécanismes sociaux et culturels influençant les comportements et les représentations des élèves, selon leur sexe et leur genre. À partir de ce cadre théorique, cette étude s'intéresse spécifiquement à l'enseignement du volley-ball dans les lycées agricoles et cherche à comprendre comment les représentations de cette activité sportive, en lien avec les notions de sexe et de genre, peuvent orienter les comportements des élèves en EPS.

- **Problématique**

Comment les représentations du volley-ball chez les filles et les garçons, en lien avec leur orientation de genre, influencent-elles leurs comportements en classe d'EPS ?

• Méthodologie

L'enquête a été menée auprès d'élèves de première (filières générales, technologiques et professionnelles) dans plusieurs lycées agricoles de la région Midi-Pyrénées. Elle s'appuie sur une méthode mixte combinant différents outils :

– BSRI (Bem Sex Role Inventory) : pour mesurer l'orientation de genre des élèves à partir de traits stéréotypés (ex. : confiance en soi, douceur, ambition, empathie, etc.).

– Questionnaire EPS : permettant de cerner les préférences sportives, les attitudes face à la mixité ou encore l'intérêt pour la discipline.

– Test d'association de mots : en réponse au mot inducteur volley-ball, les 2386 mots récoltés ont été regroupés en 20 catégories thématiques (ex. : "aspect collectif", "attaque", "peur/douleur", etc.).

– Différenciateur sémantique : inspiré du modèle d'Osgood, il permet d'évaluer les connotations affectives associées à l'activité volley-ball, en opposant des adjectifs pairs (ex. : agréable/désagré facile/difficile, etc.).

• Résultats

Les analyses montrent que des réseaux implicatifs structurent les représentations des élèves autour du volley-ball. Si l'orientation de genre joue un rôle, c'est surtout la variable sexe biologique qui apparaît comme déterminante dans la manière dont les élèves perçoivent cette activité.

Les garçons tendent à projeter des représentations valorisant l'engagement, la puissance ou la performance, alors que les filles associent davantage le volley à la coopération, la technique ou la peur de l'erreur. Ces représentations différenciées influencent directement leur implication, leur rapport au jeu et leurs dynamiques d'apprentissage en EPS.

Les inégalités sexuées en EPS sont complexes, car elles résultent d'interactions entre sexe, genre et représentations sociales. Le volley-ball, en tant qu'activité collective, technique et codifiée, cristallise certaines différences de perception selon les sexes.

Mieux comprendre ces représentations – notamment à travers des outils comme le IRSB, la différenciatrice sémantique ou l'association de mots – permettrait d'adapter les pratiques pédagogiques pour limiter les inégalités en EPS, en prenant en compte les mécanismes sociaux qui façonnent les comportements des élèves.

Dans cette perspective, une analyse plus approfondie des données issues du test d'association de mots a été conduite à l'aide du logiciel CHIC, afin de mettre en lumière les réseaux implicites de représentations qui structurent les perceptions du volley-ball. Cette analyse permet de croiser les effets du sexe biologique et de l'orientation de genre mesurée par l'IRSB, pour affiner la compréhension des dynamiques à l'œuvre.

• Analyse des réseaux de représentations : effets croisés du sexe et du genre (IRSB)

L'analyse implicative réalisée avec CHIC a permis d'identifier trois réseaux distincts de représentations du volley-ball, chacun regroupant des mots associés de manière significative selon la méthode des implicites. Ces réseaux peuvent être interprétés comme structurant différentes visions de l'activité, différenciées en fonction des caractéristiques sexuées et genrées des élèves.

– Le réseau A, composé majoritairement de termes connotés positivement (e.g., "équipe", "plaisir", "entraide", "réussite"), semble traduire une représentation valorisante et coopérative du volley-ball. Ce réseau est significativement associé aux filles, mais également aux individus présentant un score élevé de féminité sur l'IRSB, tous sexes confondus. Cela suggère que les représentations positives et relationnelles de l'activité sont davantage portées par des élèves socialement ou psychologiquement situés du côté des attributs féminins, indépendamment de leur sexe biologique.

– Le réseau B regroupe des termes neutres ou descriptifs (e.g., "filet", "ballon", "passe", "terrain"), renvoyant à une représentation plus technique et décontextualisée du volley-ball. Ce réseau n'est significativement corrélé à aucune des variables de sexe ou de genre, ce qui pourrait indiquer une forme de représentation consensuelle, plus scolaire et normée, moins influencée par les dimensions identitaires.

– Le réseau C, enfin, est constitué de mots à connotation plus négative ou compétitive (e.g., "stress", "perte", "erreur", "frappe", "puissance"). Il est significativement lié aux garçons et aux individus ayant un score élevé de masculinité. Cette association tend à confirmer l'hypothèse selon laquelle les représentations plus conflictuelles, centrées sur la performance et la confrontation, sont davantage présentes chez les élèves s'identifiant à des traits masculins.

Ces résultats mettent en évidence l'importance des dimensions de sexe et de genre dans la structuration des représentations sociales du volley-ball. Si le sexe biologique joue un rôle, notamment dans les pôles A et C, l'influence du genre (mesuré via l'IRSB) s'avère tout aussi déterminante, voire plus discriminante dans certains cas. On observe ainsi que la masculinité et la féminité psychologiques modulent fortement les perceptions de l'activité, suggérant que les pratiques pédagogiques en EPS gagneraient à intégrer une réflexion plus fine sur les identités de genre et leurs effets sur les expériences et représentations des élèves.

3.2.2 Application à l'analyse des impacts de la pandémie de Covid-19 et de l'enseignement à distance sur la santé mentale des étudiants multiculturels [14]

La crise sanitaire due au Covid-19 a engendré des conséquences psychosociales majeures, notamment l'isolement social, facteur de stress, d'anxiété et de dépression (Barbosa et al., 2021). L'enseignement à distance, adopté au Brésil pendant la pandémie, a aggravé ces effets, particulièrement pour les étudiants vulnérables socio-économiquement. Les obstacles incluent :

– Le manque d'accès à des technologies adaptées (ordinateurs, internet stable),

– L'absence d'un environnement propice à l'étude à domicile,

– L'exposition de réalités sociales précaires via les cours en ligne (Santos de Aquino et al., 2021).

Cette étude, menée à l'Institut fédéral d'éducation, de science et de technologie du Sertão de Pernambouc (IFSertãoPE) dans le cadre d'une thèse sur l'enseignement scientifique en contexte multiculturel, combine trois approches : les théories du multiculturalisme (Candau Moreira, 2008), l'analyse des rapports de l'OPS et de la School Board Association, et l'Analyse Statistique Implicative (Gras et al., 2017) pour examiner les données quantitatives.

• Problématique

Quels sont les impacts psychologiques de la pandémie et de l'enseignement à distance sur des étudiants issus de cultures diversifiées (indigènes, quilombolas, sertanejos, urbains) ?

- **Méthodologie**

Pour construire des données , Un questionnaire exploratoire basé sur la triade informative interculturelle (Kidman et al., 2013) a été administré par vidéoconférence à 14 élèves multiculturels (5 sertanejos, 4 indigènes, 4 urbains, 1 quilombola) âgés de 16 à 18 ans. Bien que l'anxiété et la dépression ne fussent pas initialement ciblées, ces thèmes ont émergé spontanément dans leurs témoignages sur les impacts de la pandémie et de l'enseignement à distance. Ces données ont été traitées par l'Analyse Statistique Implicative (ASI) via le logiciel CHIC v.7 (2014), méthode adaptée aux petits échantillons. Cette approche a identifié des relations significatives (seuil ¿0.70) entre variables culturelles ("culture", "connaissances traditionnelles"), socio-économiques ("difficultés", "usage du portable") et psychologiques ("anxiété", "dépression"), révélant leurs interdépendances dans un graphe implicatif.

- **Résultats**

On constate que l'impact de la pandémie sur la santé mentale des étudiants diffère selon les cultures. La figure (2.1) présente le pourcentage d'étudiants ayant déclaré avoir développé ou intensifié des problèmes d'anxiété et de dépression.

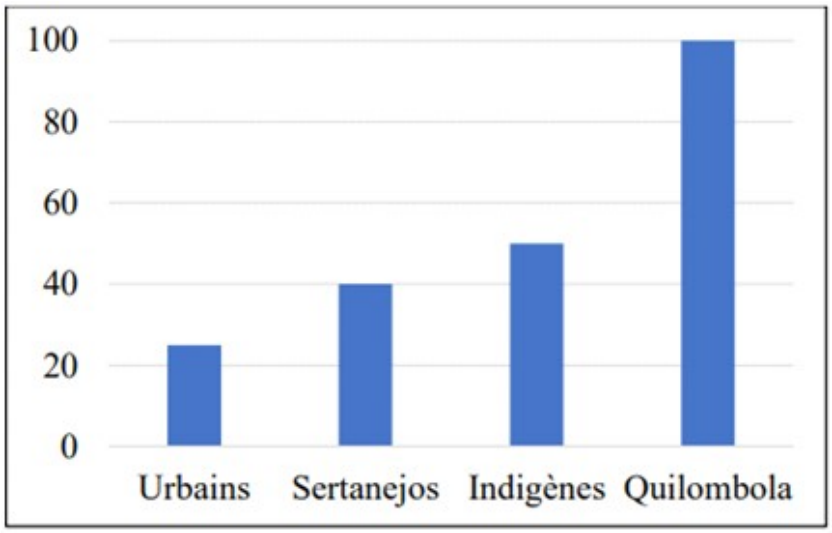

Figure (2.1) : Pourcentage de rapports de détresse psychologique par culture.

Nous avons remarqué que les étudiants urbains présentaient le pourcentage le plus faible de déclarations de développement ou d'augmentation de l'anxiété et de la dépression, alors que ces souffrances psychologiques étaient plus fréquentes chez les étudiants de l'arrièrepays, les indigènes et les quilombolas, respectivement.

Une autre donnée pertinente a été observée à travers la relation entre les sexes, les filles étant plus susceptibles de souffrir psychologiquement que les garçons. Comme le montre la Figure (2.2).

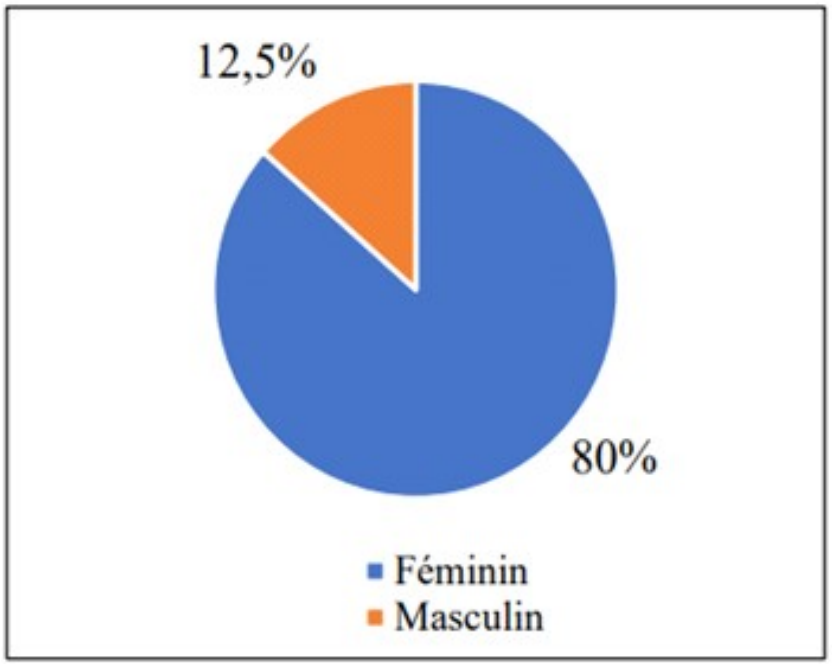

Figure (2.2) : Pourcentage de rapports de détresse psychologique par sexe.

Le graphique implicatif présenté dans la figure (2.3) confirme les données en pourcentage et présente d'autres relations avec des variables importantes qui contribuent à une analyse Holistique de l'objet d'étude. On y remarque que les étudiants ont déclaré conjointement la dépression et l'anxiété qui tendent à l'implication mutuelle (Anxieté Dépression) avec un indice d'implication de 0,90 (vecteur rouge), ce qui explique pourquoi toutes les autres variables impliquent conjointement la dépression et l'anxiété.

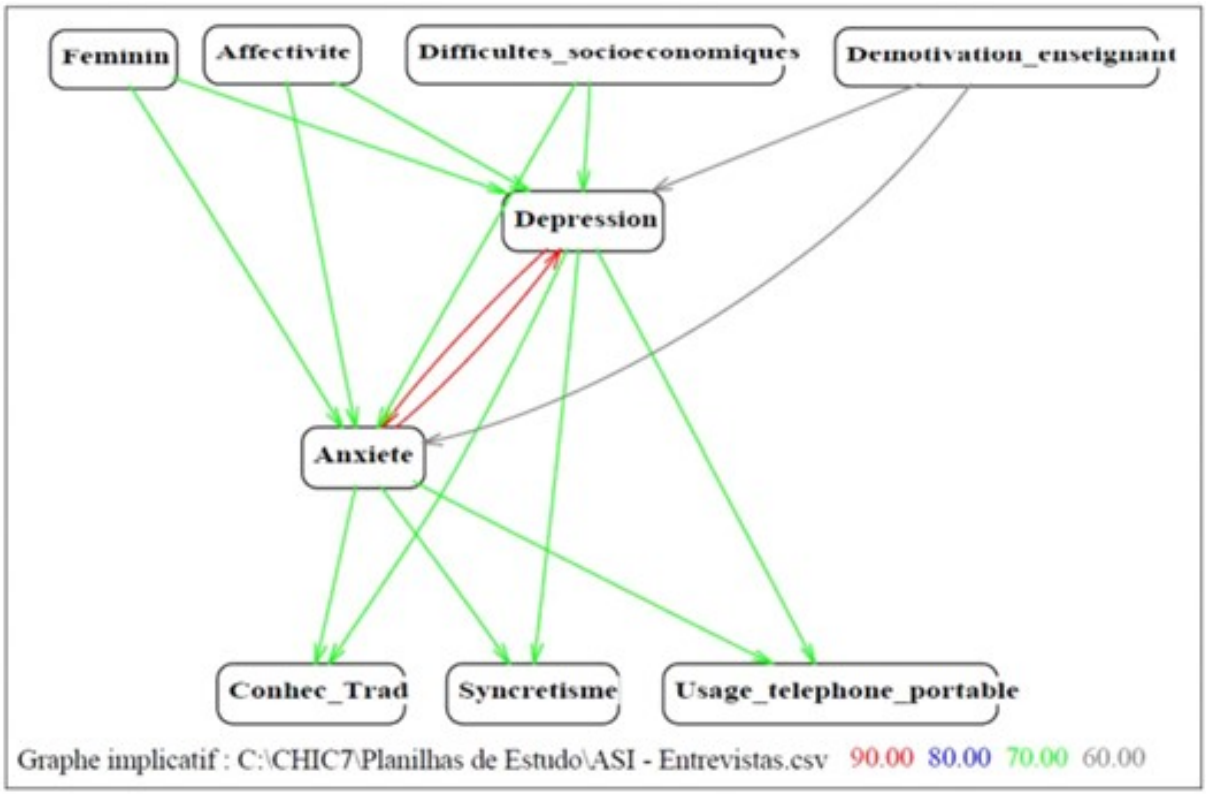

Figure (2.3) : Graphique implicatif relatif aux variables « anxiété » et « dépression ».

• **Analyse des résultats**

L'étude montre que les étudiants pauvres souffrent plus de dépression et d'anxiété. Leur seul outil pour étudier pendant la pandémie était souvent un vieux téléphone portable, partagé avec toute la famille. Ce manque d'accès à de bons outils technologiques a rendu les cours à distance plus difficiles et a aggravé leur stress. L'étude révèle des liens importants :

– Les étudiants sensibles aux relations affectives (besoin d'attention, de respect) sont plus vulnérables à l'anxiété et la dépression. Ceci est particulièrement vrai pour les étudiants indigènes et quilombolas.

– La démotivation causée par les enseignants aggrave les problèmes mentaux. Une approche pédagogique plus bienveillante et interculturelle pourrait aider.

– Les étudiants qui utilisent leurs connaissances traditionnelles en classe ont plus de risques de souffrir mentalement, montrant un conflit entre leur culture et l'école.

– Enfin, le sexe féminin tend à développer et à potentialiser l'anxiété et la dépression (Feminin→Anxieté ; Feminin → Dépression). La pression sociale exercée sur les filles s'ajoute aux effets de la pandémie, l'isolement social et l'éloignement de l'éducation nuisant plus fortement aux étudiantes.

Cette étude avec la méthode de l'ASI met en lumière comment la pandémie de Covid-19 a affecté de manière inégale la santé mentale des populations vulnérables de Salgueiro, au Brésil. Les résultats révèlent que les communautés culturellement minoritaires - notamment les indigènes, les quilombolas et les sertanejos - ainsi que les femmes issues de milieux défavorisés, ont été particulièrement touchées par des troubles psychologiques comme l'anxiété et la dépression.

3.3 Domaine de la médecine

Dans le domaine médical, la compréhension des mécanismes à l'origine des maladies repose souvent sur l'analyse de plusieurs variables interdépendantes. L'ASI offre une approche pertinente pour mettre en évidence des relations significatives entre ces variables cliniques. C'est dans cette optique qu'elle a été appliquée à l'analyse des données d'échocardiographie de stress, afin d'identifier les facteurs associés à l'apparition d'un état de stress chez les patients.

3.3.1 Application à l'analyse de données issues de l'échocardiographie de stress [9]

Cette étude est consacrée à l'analyse des causes potentielles d'un état de stress chez les patients à partir de données d'échocardiographie, en s'appuyant sur l'Analyse Statistique Implicative via le logiciel RCHIC. Elle repose sur un jeu de données médicales collecté par Frank Harrell à l'Université de Vanderbilt, comprenant 558 patients et 31 variables cliniques et physiologiques. Nous allons présenter ici les résultats décrits dans le papier de Ghanem Souhila et al., illustrant la pertinence de l'approche implicative pour mettre en évidence les relations significatives entre différentes variables médicales. Cette application mobilise en particulier le critère de l'intensité de l'implication combinée avec la confiance, développé par Ghanem Souhila et Raphaël Couturier, pour renforcer la robustesse de l'analyse des relations entre variables.

• **Problématique**

Comment l'ASI permet-elle d'identifier les facteurs cliniques et physiologiques associés à l'apparition d'un état de stress chez les patients à partir des données d'échocardiographie de stress ?

• **Méthodologie**

Les données ont été préparées sous format .csv, comme illustré dans la figure (2.4), les individus (patients) en lignes et les variables en colonnes. Plusieurs variables numériques ont été partitionnées automatiquement en trois niveaux (faible, moyen, élevé) selon l'algorithme des nuées dynamiques (Diday, 1971). Cela permet une dichotomisation binaire (0/1) utilisée par RCHIC pour construire les règles d'implication.

	bhr p	basebp p	basedp p	pkhr p	sbp p	dp p	dose p	maxhr p	pctMphr p	mbp p	dpmaxdo p	dobdose p	age p	mal	female	baseEF p	d
1	92	103	9476	114	86	9804	40	100	74	121	12100	40	85	1	0	27	
2	62	139	8618	120	158	18960	40	120	82	158	18960	40	73	1	0	39	
3	62	139	8618	120	157	18840	40	120	82	157	18840	40	73	1	0	39	
4	93	118	10974	118	105	12390	30	118	72	105	12390	30	57	0	1	42	
5	89	103	9167	129	173	22317	40	129	69	176	22704	40	34	1	0	45	
6	58	100	5800	123	140	17220	40	123	83	140	17220	40	71	1	0	46	
7	63	120	7560	98	130	12740	40	98	71	130	12740	40	81	0	1	48	
8	86	161	13846	144	157	22608	40	144	111	157	22608	40	90	0	1	50	
9	69	143	9867	115	118	13570	40	113	81	151	17063	40	81	0	1	52	
10	76	105	7980	126	125	15750	40	126	94	125	15750	40	86	1	0	52	
11	105	134	14070	171	182	31122	40	171	108	182	31122	40	61	0	1	52	
12	72	112	8064	127	95	12065	30	125	80	101	12625	20	63	1	0	53	
13	90	120	10800	169	184	31096	40	169	126	184	31096	40	86	1	0	54	
14	81	110	8910	110	130	14300	40	110	58	130	14300	40	29	0	1	55	
15	84	176	14784	110	194	21340	40	110	74	194	21340	40	71	0	1	55	

Figure (2.4) : Extrait du jeu de données.

Parmi les 31 variables utilisées figurent la fréquence cardiaque de base (bhr), la tension de base (basebp), la pression artérielle maximale (mbp), le dosage de dobutamine (dobdose), les antécédents médicaux (hxofHT, hxofDM, newMI), le sexe (gender), ou encore le diagnostic ECG (ecg), entre autres. Ces variables ont servi à construire des règles d'implication mettant en relation différents états cliniques et la présence de pathologies.

• Résultats obtenus en utilisant un seuil de confiance égale à 80

Avec un seuil de confiance fixé à 80%, seules les règles d'implication les plus stables et fortement impliquées sont conservées, comme le montre la figure (2.5).

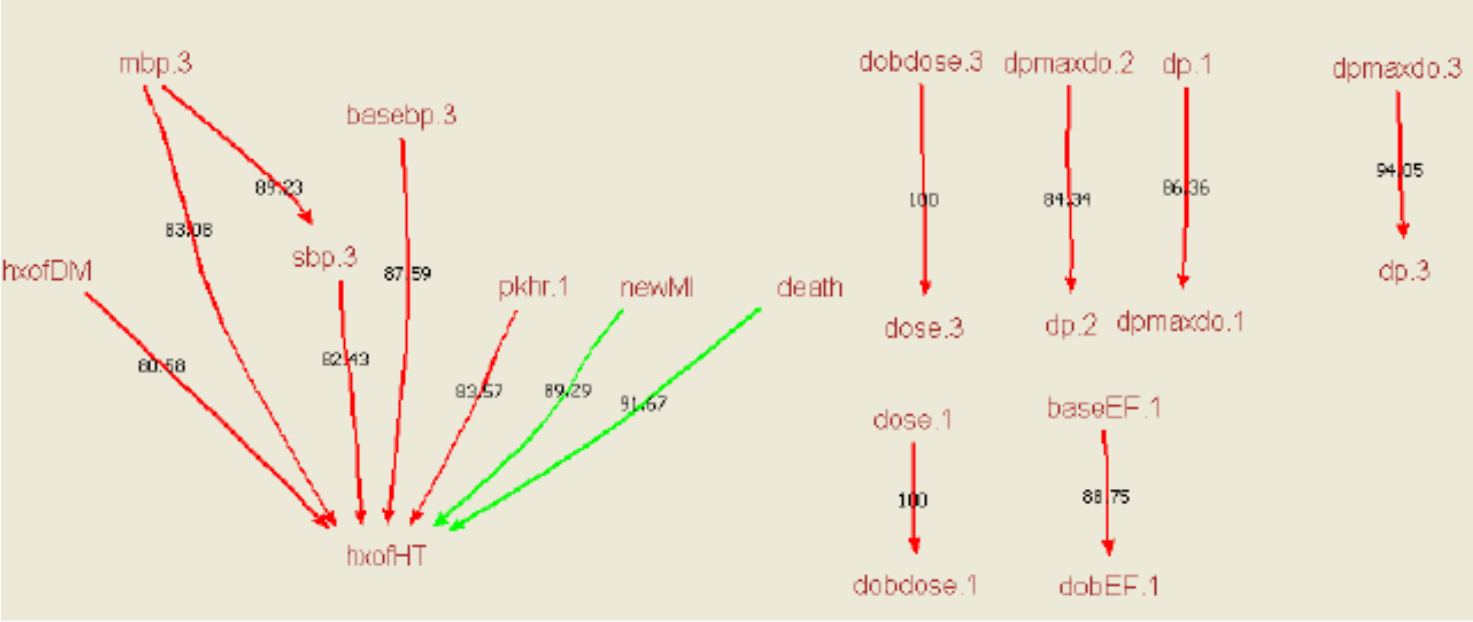

Figure (2.5) : Graphe implicatif avec un seuil de confiance égal à 80.

Principales implications

– Le diabète (hxofDM) implique un état de stress,

– Une fréquence cardiaque de base faible (bhr.1) implique un état de stress,

– Une pression de base ou maximale élevée (basebp.3, mbp.3) implique un état de stress,

– Une crise cardiaque récente (newMI) implique un état de stress.

Ces implications sont cliniquement interprétées comme des indicateurs de stress cardiovasculaire latent, souvent médié par l'hypertension (hxofHT), utilisée comme variable cible dans le graphe d'implication.

les principaux facteurs de stress identifiés sont :

– le diabète,

– une fréquence cardiaque faible,

– une pression sanguine maximale très élevée,

– une tension artérielle de base élevée,

– une nouvelle crise cardiaque.

La robustesse de ces associations est confirmée par leurs valeurs de confiance élevées, renforçant leur pertinence dans l'évaluation du stress à partir des données d'échocardiographie.

• Résultats obtenus en utilisant un seuil de confiance égale à 70

Avec un seuil de confiance fixé à 70%, d'autres implications apparaissent en complément de celles observées à 80%. Ces nouvelles règles, bien que légèrement moins stables, révèlent des tendances intéressantes :

– une dose très élevée de dobutamine est administrée dans 71% des cas chez les femmes,

– plus de 70% des personnes non fumeuses sont des femmes,

– 73% des individus présentant une anomalie de mouvement de la paroi au repos sont également des femmes,

– les personnes très jeunes tendent à présenter un électrocardiogramme (ECG) normal.

Ces résultats suggèrent l'influence du sexe et de l'âge sur certaines caractéristiques cliniques associées à l'état de stress.

• Résultats obtenus en utilisant un seuil de confiance égale à 65

Avec un seuil de confiance abaissé à 65%, de nouvelles associations, plus nombreuses mais statistiquement moins robustes, sont mises en évidence :

– 66% des individus présentant une anomalie de mouvement de la paroi au repos ont un diagnostic ECG normal.

– 68% des patients ayant une fraction d'éjection cardiaque initiale très élevée, et 69% de ceux présentant une fraction d'éjection élevée sous dobutamine, présentent également une anomalie de mouvement de la paroi au repos.

Ces résultats mettent en lumière des relations plus fines entre les paramètres fonctionnels du cœur et certains signes cliniques, possiblement indicateurs d'un état de stress cardiovasculaire latent.

3.4 Domaine de l'éducation

Dans le domaine de l'éducation, l'analyse des parcours académiques repose souvent sur l'étude de multiples variables pédagogiques interdépendantes. L'Analyse Statistique Implicative (ASI) constitue une approche pertinente pour mettre en évidence des relations significatives entre les performances des étudiants dans différents modules d'enseignement. C'est dans cette perspective qu'elle a été mobilisée afin d'explorer les résultats académiques des étudiants en informatique à l'Université de Béjaïa.

3.4.1 Application à l'analyse des liens entre modules à l'Université de Béjaïa [15]

L'article intitulé « Analysis of Bejaia University Computer Science students' marks through the CHIC software and Statistical Implicative Analysis », rédigé par Hayette Khaled, Souhila Ghanem (Université de Béjaïa, Algérie) et Raphaël Couturier (Université de Franche-Comté, France), a été publié en 2015. Cette recherche s'inscrit dans une démarche visant à démontrer l'intérêt de l'Analyse Statistique Implicative, développée par Régis Gras, pour l'étude des performances académiques des étudiants.

• Problématique

Comment l'ASI permet-elle d'identifier les relations entre les différents modules d'enseignement afin de mieux comprendre les parcours académiques des étudiants en informatique à l'Université de Béjaïa ?

• **Méthodologie**

Les auteurs ont appliqué la méthode ASI à travers le logiciel CHIC (Classification Hiérarchique Implicative Cohésive) afin d'explorer les relations implicatives entre les notes obtenues par les étudiants en informatique de l'Université de Béjaïa au cours de trois promotions successives : 2010-2011, 2011-2012 et 2012-2013. L'analyse porte sur les niveaux licence 2 et licence 3, et met en évidence des règles d'implication récurrentes entre différents modules d'enseignement, révélant ainsi des liens pédagogiques forts, utiles tant pour l'orientation des étudiants que pour l'organisation des cursus. Les notes sont converties en fichiers .csv (voir la figure (2.6)), puis les variables (notes) sont réparties en intervalles grâce à un algorithme de classification dynamique. Des graphes d'implication sont générés à partir de ces données.

	ARCH p	DSTR1 p	IS p
0809TMI02	11,25	9,88	11,5
09MI0034	9,38	11,38	7,33
09MI0590	10,63	9,38	8,67
09MI0061	11,38	14,69	10,5
...	...	...	...

Figure (2.6) : Extrait du jeu de données de type .csv.

• **Interprétations des résultats selon le niveau Licence 2 (L2)**

L'analyse des graphes implicatifs pour les promotions 2010-2011 (Figure 2.6), 2011-2012 (Figure 2.7) ainsi que 2012-2013 (Figure 2.8) en deuxième année (L2) fait apparaître un certain nombre de règles d'implication récurrentes, soulignant des relations pédagogiques fortes entre certains modules.

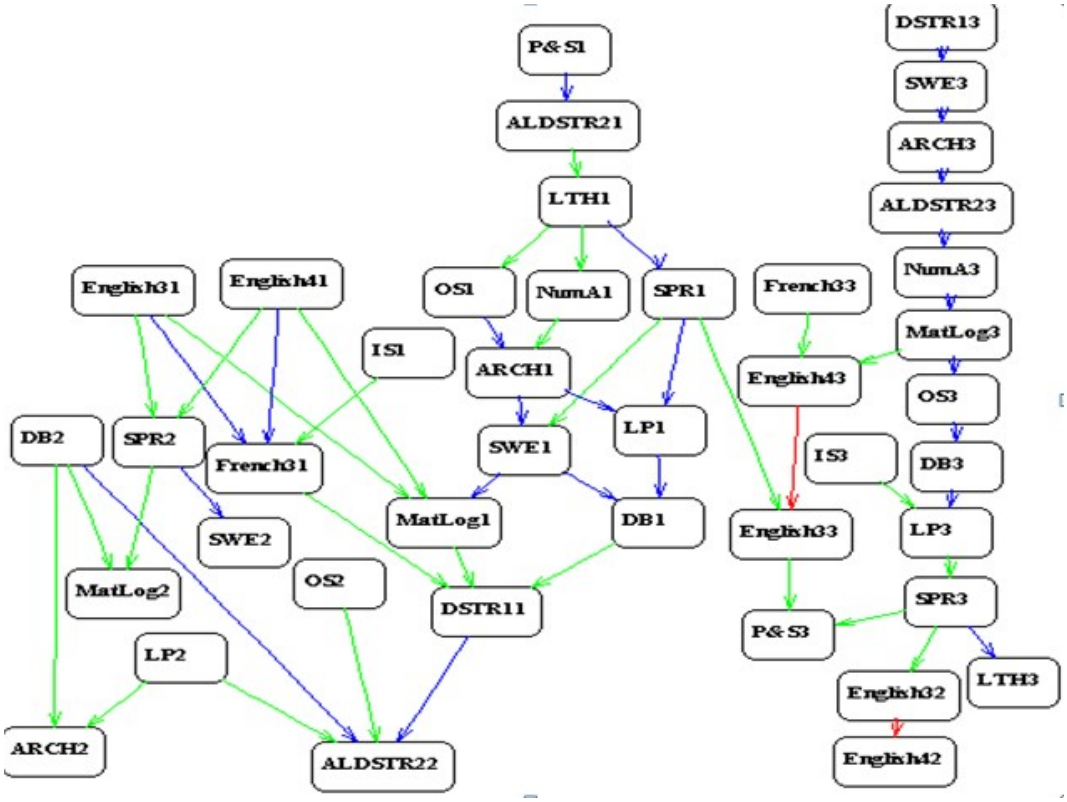

Figure (2.7) : Graphe implicatif Licence2 2010-2011.

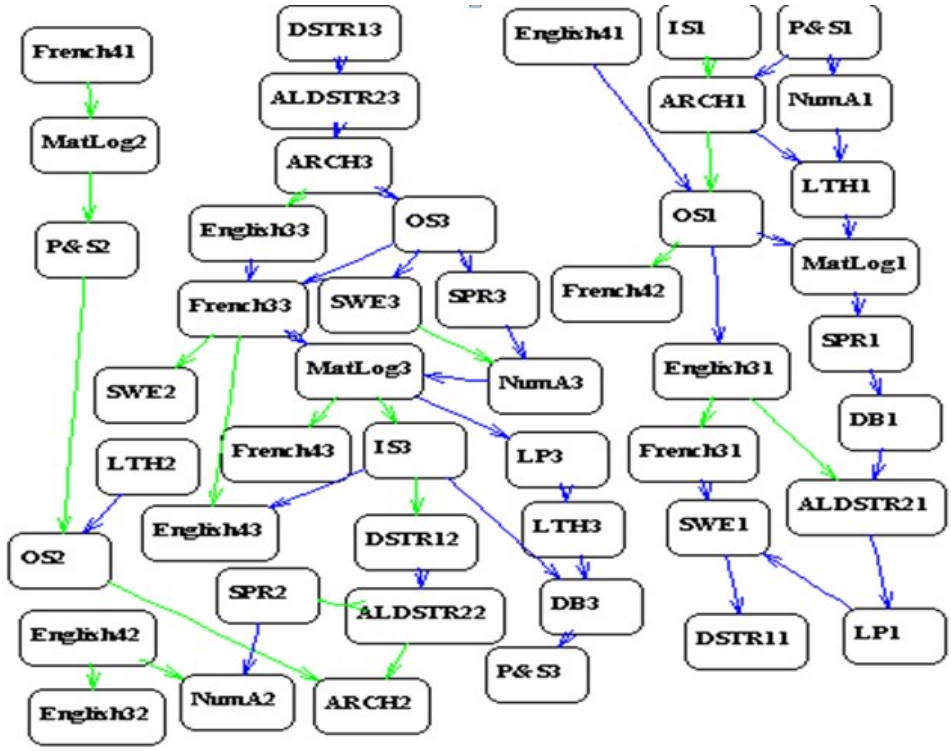

Figure (2.8) : Graphe implicatif Licence2 2011-2012.

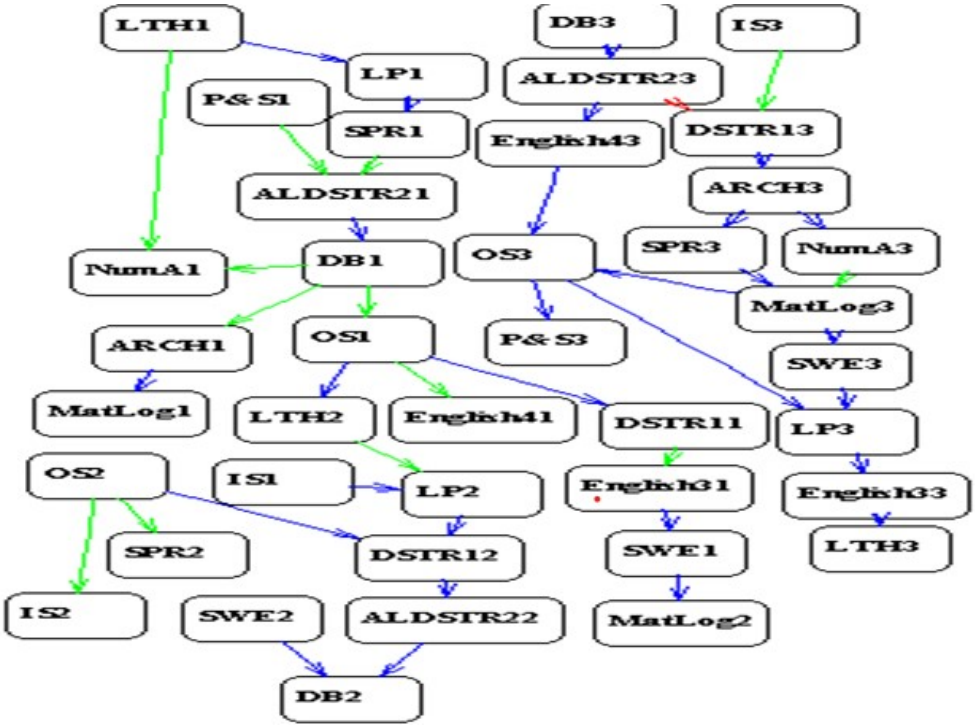

Figure (2.9) : Graphe implicatif Licence2 2012-2013.

Principales implications

– NumA3 → MatLog3
Les étudiants maîtrisant bien l'analyse numérique réussissent aussi en logique mathématique. Ces deux modules reposant sur des compétences mathématiques similaires, cette implication est cohérente,

– DSTR1 → ALDSTR2
Le module « Structures de Données » du premier semestre est un prérequis direct pour « Algorithmes et Structures de Données 2 » du second semestre. Cette continuité de contenus justifie l'implication forte,

– SWE → DB
Les étudiants bons en « Ingénierie logicielle » réussissent également en « Base de données ». Les deux modules nécessitent une compréhension conceptuelle de la modélisation logicielle,

– LTH → OS
Des liens bidirectionnels sont observés entre « Théorie des Langages » et « Systèmes d'Exploitation ». La compréhension des graphes et automates, communs aux deux modules, explique cette relation,

– Anglais → Français
Les compétences en langues sont transférables : les bons (ou mauvais) résultats en anglais sont souvent liés à ceux en français,

– OS → English (2011)
Cette implication spécifique à 2011 s'explique par un contenu d'anglais centré cette année-là sur le vocabulaire des systèmes d'exploitation.

Ces résultats soulignent la cohérence structurelle des programmes, mais également le rôle de certains modules comme indicateurs de réussite ou d'échec dans d'autres.

- **Interprétations des résultats selon le niveau Licence 3 (L3)**

Pour les promotions 2010-2011, 2011-2012 et 2012-2013, l'analyse en troisième année (L3) révèle également des implications pédagogiques stables sur plusieurs années.

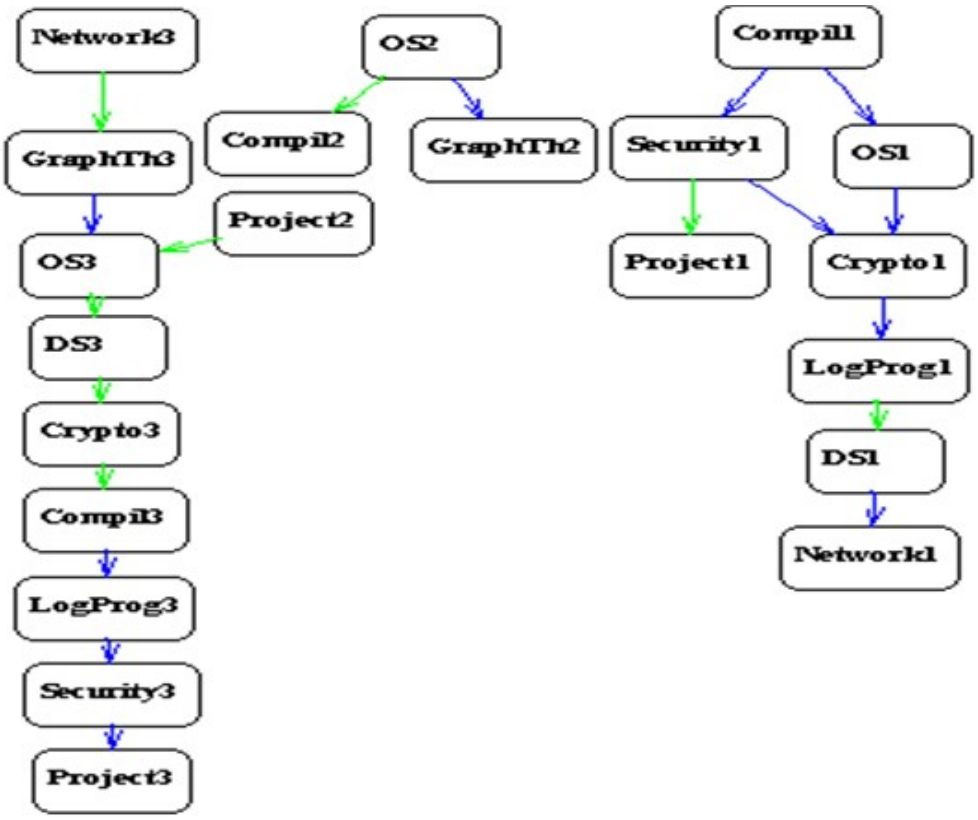

Figure (2.10) : Graphe implicatif Licence3 2010-2011.

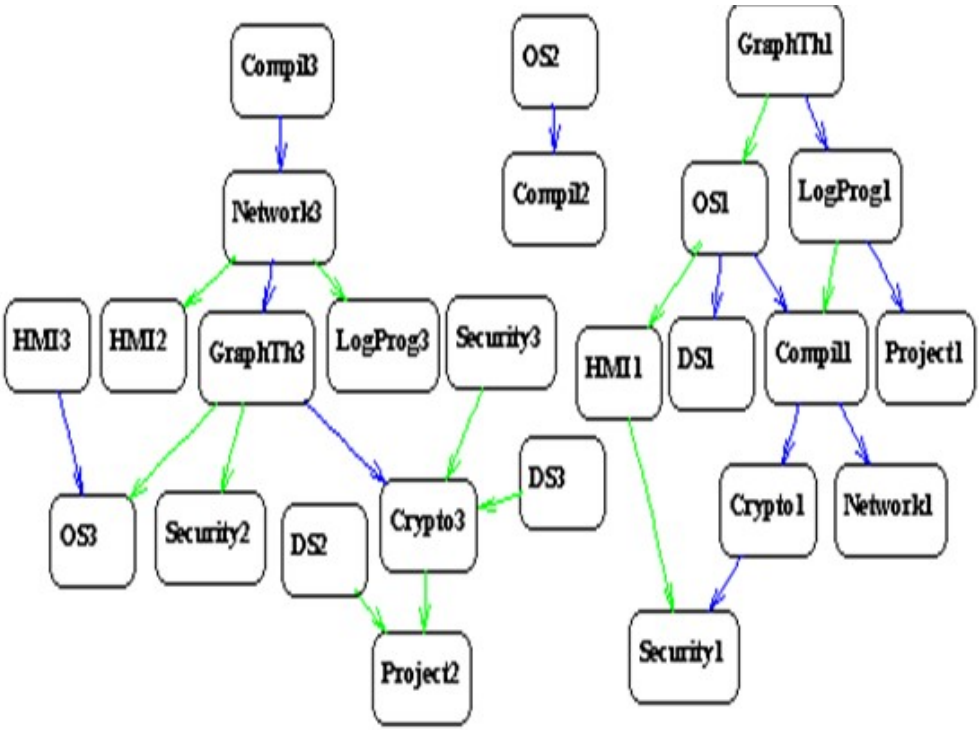

Figure (2.11) : Graphe implicatif Licence3 2011-2012.

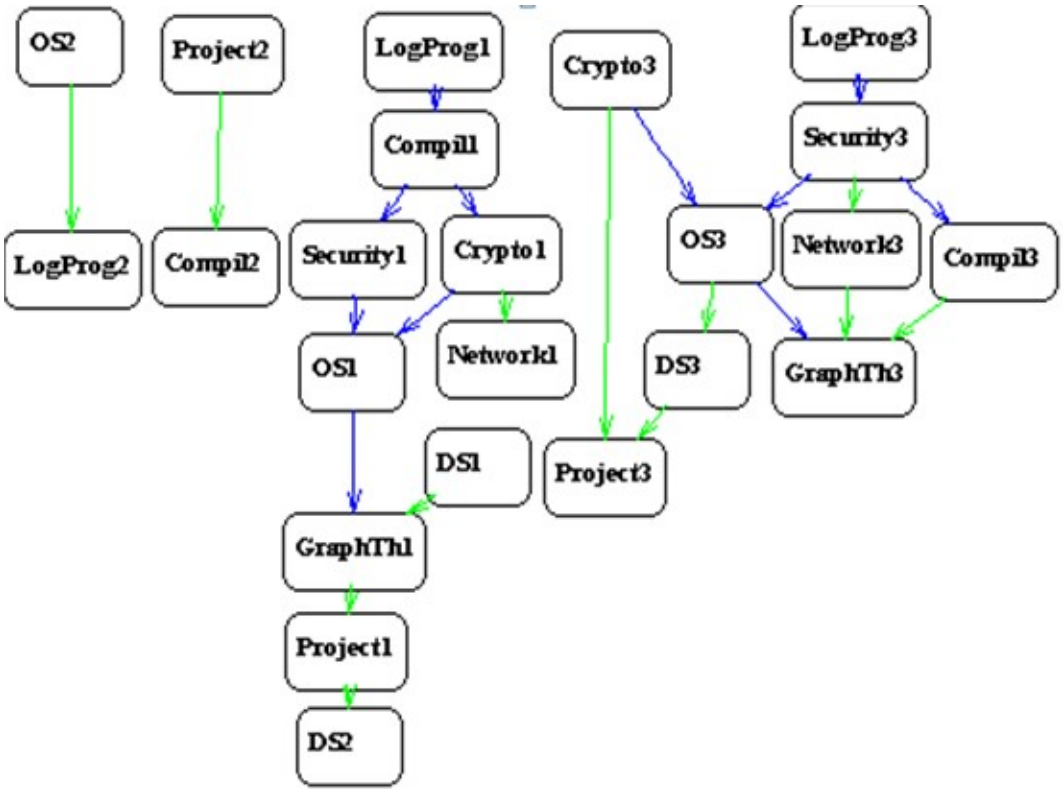

Figure (2.12) : Graphe implicatif Licence3 2012-2013.

Principales implications

– OS → DS
Le module « Systèmes d'Exploitation » conditionne la réussite en « Systèmes Distribués », qui en est une extension naturelle. Cette relation est logique sur le plan didactique.,

– LogProg → Compil
La logique de programmation et la compilation sont fortement liées, souvent enseignées par le même enseignant, avec des compétences mobilisées similaires,

– Security → Crypto
La cryptographie est intégrée comme chapitre dans le module « Sécurité », d'où des implications dans les deux sens selon les promotions,

– OS → GraphTh
Les algorithmes de gestion dans les systèmes d'exploitation font appel à des compétences en théorie des graphes, expliquant les liens croisés.

Ces relations implicatives confirment que certains modules constituent des nœuds pédagogiques stratégiques. Leur maîtrise semble favoriser la réussite dans des enseignements plus spécialisés, renforçant ainsi l'idée d'un parcours logique de compétences.

– Elles peuvent aider à mieux planifier les parcours étudiants, en identifiant les modules qui jouent un rôle central dans la réussite globale. Un module qui conditionne la réussite dans plusieurs autres peut être considéré comme stratégique, et donc prioritaire dans l'accompagnement pédagogique.

– Elles permettent aussi de détecter les modules à prérequis forts. Si la réussite dans un module dépend systématiquement d'un autre, cela signifie qu'il existe un lien de dépendance pédagogique important. Cette information est précieuse pour éviter les échecs en aval.

– Ces résultats peuvent conduire à réorganiser l'ordre des modules dans le cursus. En effet, si un module A prépare clairement à un module B, mais qu'il est placé après lui dans le programme, cela peut nuire à la progression des étudiants.

– Enfin, les implications relevées peuvent révéler l'impact d'un changement d'enseignant ou de méthode. Par exemple, si une relation implicative stable disparaît soudainement une année, cela peut signaler une modification dans les pratiques pédagogiques ou dans le contenu

du cours, méritant une attention particulière.

En somme, les règles d'implication mises en évidence chaque année ne sont pas seulement des constats statistiques : elles sont aussi des indicateurs précieux pour mieux organiser l'enseignement, accompagner les étudiants et renforcer la cohérence des formations.

3.5 Constat issu de l'état de l'art

À travers l'étude des différentes recherches consacrées à l'analyse statistique implicative, nous avons pu constater que cette méthode présente un impact positif dans plusieurs domaines d'application, tels que l'éducation, la médecine ou encore la psychologie. Ces travaux montrent que l'ASI permet de mettre en évidence des structures relationnelles implicites au sein de données complexes, offrant ainsi une lecture fine et pertinente des phénomènes étudiés.

Nous avons également relevé que plusieurs critères ont été proposés pour qualifier les relations implicatives. Parmi eux, le critère combinant l'intensité d'implication et la confiance, développé récemment par Souhila Ghanem (Université de Béjaïa, Algérie) et Raphaël Couturier (Université de Franche-Comté, France), publié en 2015, apparaît comme une avancée méthodologique notable.[9]

Ce critère présente un intérêt particulier pour notre étude, car il permet de mieux qualifier les relations implicatives en tenant compte à la fois de la force de l'implication et de la fiabilité des données observées. Pour cette raison, nous avons choisi de l'utiliser dans notre analyse, présentée dans le chapitre suivant, portant sur un jeu de données médicales connu, appelé Pima Indian Diabetes.

3.6 Conclusion

Dans ce chapitre, nous avons exploré plusieurs domaines dans lesquels l'Analyse Statistique Implicative a été appliquée avec succès, notamment en psychologie, en médecine et en éducation. Ces exemples ont permis de montrer la souplesse d'utilisation de la méthode, ainsi que sa capacité à faire émerger des relations implicatives pertinentes dans des contextes très variés.

Dans le prochain chapitre, nous présenterons notre propre environnement de travail, puis nous appliquerons à notre tour l'ASI afin de mettre en évidence les relations structurelles présentes dans notre jeu de données.

4 Application de l'Analyse Statistique Implicative (ASI)

4.1 Introduction

Ce dernier chapitre présente la mise en œuvre pratique de notre démarche. Après avoir posé les bases théoriques de l'Analyse Statistique Implicative, nous détaillons ici l'environnement technique utilisé et le jeu de données analysé.

Nous introduirons d'abord les outils mobilisés : le logiciel R, l'environnement RStudio, et le package RCHIC, qui permet de réaliser une ASI de manière efficace. Nous expliquerons brièvement le choix de ces outils et les principales fonctionnalités exploitées dans notre travail.

Nous présenterons ensuite le jeu de données, en précisant son origine, sa structure, les variables utilisées, ainsi que les traitements préalables nécessaires.

Enfin, nous appliquerons l'ASI aux données préalablement préparées, puis analyserons et interpréterons les résultats obtenus.

4.2 Présentation de R

R est un logiciel libre et open source, conçu pour le traitement des données, l'analyse statistique et la représentation graphique. Il repose sur un langage de programmation interprété dérivé du langage S, utilisé notamment dans le logiciel S-PLUS, et intègre des fonctionnalités avancées comme la gestion de données simples et structurées, les opérations d'entrée-sortie, les branchements conditionnels, les boucles et la récursivité. Créé par Ross Ihaka et Robert Gentleman, R est développé depuis plus de vingt ans par une communauté internationale de chercheurs et développeurs. Publié sous licence GNU GPL, il est aujourd'hui largement employé dans les milieux académiques, scientifiques et professionnels, notamment par les statisticiens, data miners et data scientists pour le développement de logiciels statistiques et l'analyse approfondie des données. [16]

• Les principales fonctionnalités de R en analyse statistique

– Un système performant pour la manipulation et le stockage des données,

– Une grande variété d'opérateurs pour le calcul sur tableaux, notamment les matrices,

– Un vaste ensemble d'outils pour l'analyse statistique et l'exploration des données,

– Des moyens graphiques avancés pour une visualisation efficace des résultats. [16]

Grâce à son approche par objets, sa flexibilité et la richesse de ses packages, R s'impose comme un outil incontournable pour toute démarche d'analyse statistique approfondie.

• Les avantages de l'utilisation de R

– c'est un logiciel multiplateforme, qui fonctionne aussi bien sur des sytèmes Linux, Mac OS X ou Windows,

– c'est un logiciel libre, développé par ses utilisateurs et modifiable par tout un chacun,

– c'est un logiciel gratuit,

– c'est un logiciel très puissant, dont les fonctionnalités de base peuvent être étendues à l'aide de plusieurs milliers d'extensions,

– c'est un logiciel dont le développement est très actif et dont la communauté d'utilisateurs ne cesse de s'élargir,

– les possibilités de manipulation de données sous R sont en général largement supérieures à celles des autres logiciels usuels d'analyse statistique,

– c'est un logiciel avec d'excellentes capacités graphiques et de nombreuses possibilités d'export.[16]

4.2.1 Installation de R

Pour installer R sous Windows, il suffit de se rendre à la page suivante [16] :
https ://drive.google.com/file/d/11j-ZYyCAC5N4wKWhFAGB9BWri2fhAtB_/view

Il convient ensuite de télécharger le fichier d'installation. Une fois le programme téléchargé et exécuté, l'installation s'effectue en suivant les instructions de l'assistant d'installation.

Une fois R correctement installé, il est possible de lancer le logiciel et d'accéder à son interface principale. La figure (3.1) ci-dessous illustre cette interface telle qu'elle apparaît au démarrage du logiciel.

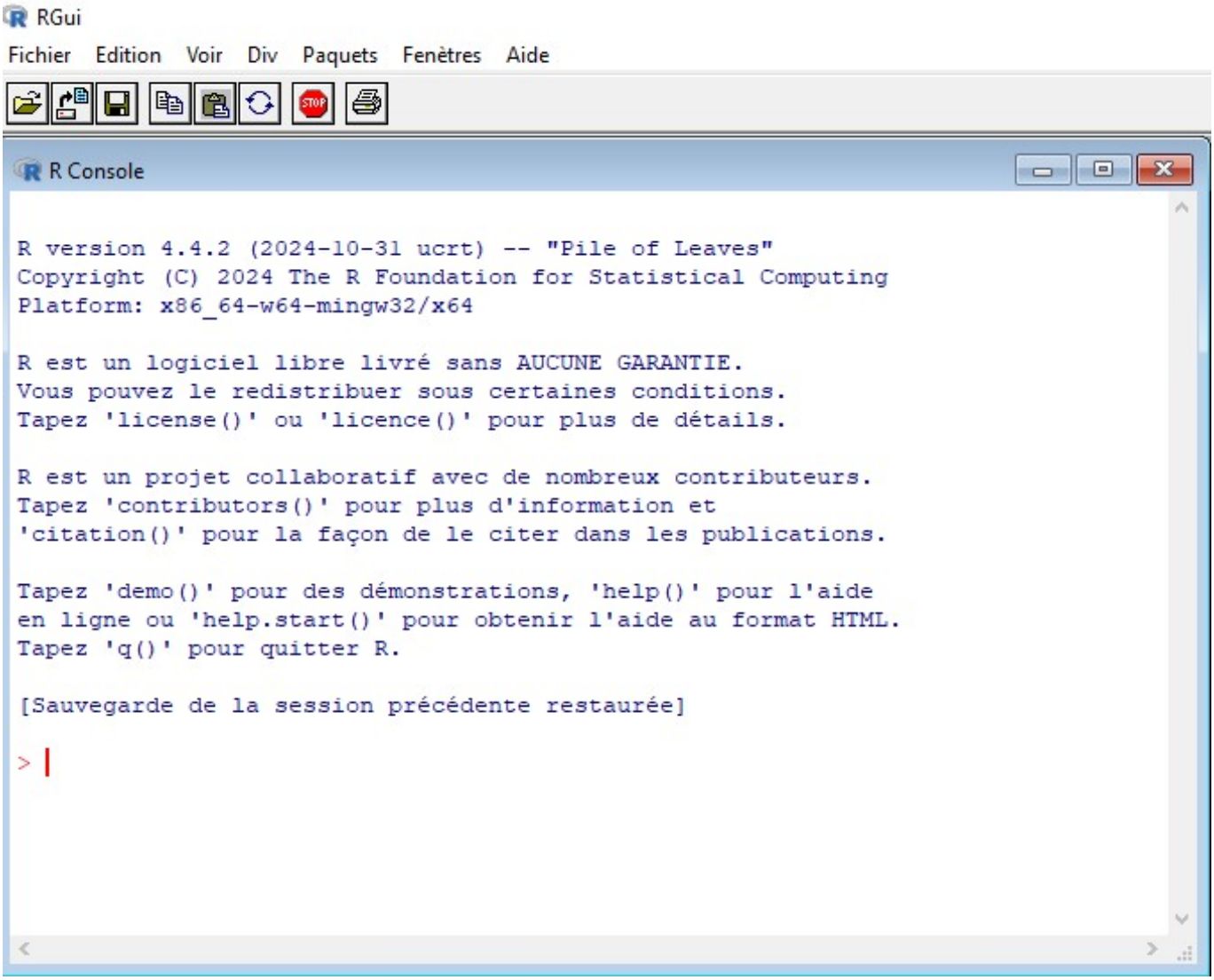

Figure (3.1) : Interface de R sous Windows.

R repose sur un système d'extensions appelées packages, qui permettent d'enrichir considérablement ses fonctionnalités. Ces packages, développés par une large communauté d'utilisateurs et de chercheurs, couvrent un vaste éventail de domaines : statistiques, visualisation, data mining, machine learning, etc. On en dénombre aujourd'hui plusieurs milliers, accessibles via le réseau de diffusion officiel CRAN (Comprehensive R Archive Network). Chaque utilisateur peut installer facilement ces packages selon ses besoins. La liste complète des extensions disponibles est consultable à l'adresse suivante : http ://cran.r-project.org/web/packages/.[16]

4.3 Présentation de RStudio

RStudio est un environnement de développement intégré (IDE) libre et gratuit, compatible avec Windows, Mac OS X et Linux. Il complète R en offrant un éditeur de script doté de la coloration syntaxique, de l'autocomplétion, et de nombreuses fonctionnalités facilitant l'écriture, l'édition et l'exécution du code. Il permet un affichage simultané de plusieurs éléments essentiels à l'analyse de données : le code, la console R, les fichiers, les graphiques, ainsi que les pages d'aide. RStudio prend également en charge la gestion des extensions (packages), l'intégration avec des systèmes de contrôle de version comme Git, et la création de rapports dynamiques via R Markdown. En développement actif, il s'enrichit régulièrement de nouvelles fonctionnalités. Son principal inconvénient reste l'absence de traduction française de l'interface, disponible uniquement en anglais.[16]

4.3.1 Installation de RStudio

Pour installer RStudio, il est nécessaire d'avoir d'abord installé le logiciel R, car RStudio fonctionne en s'appuyant sur ce dernier. Une fois R correctement installé, on RStudio en se rendant à l'adresse suivante [16] :
https ://drive.google.com/file/d/1WWt0stNQER37ruWj2jLI65JJNB6aU6SK/view ?usp=sharing

Il suffit ensuite d'exécuter le fichier téléchargé et de suivre les instructions pour finaliser l'installation. Une fois le processus terminé, le lancement de RStudio donne accès à son interface principale, comme illustré par la figure (3.2) ci-dessous.

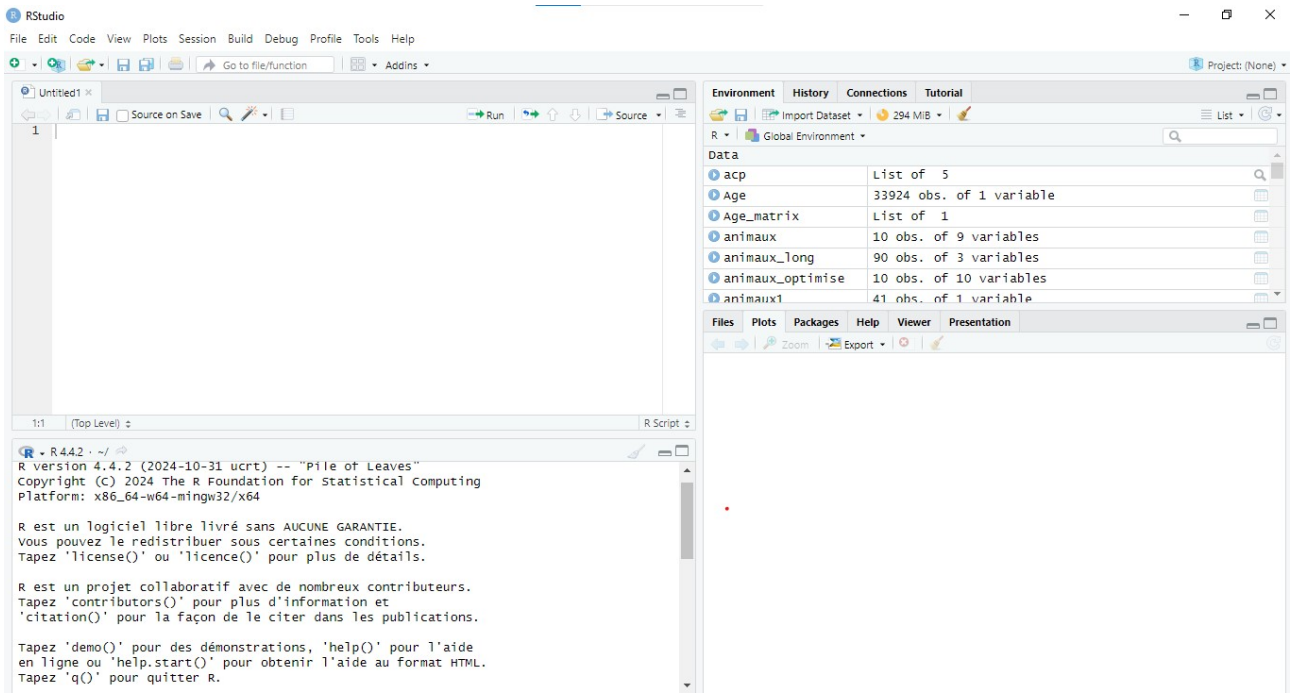

Figure (3.2) : Interface de RStudio sous Windows.

- **Présentation de l'interface de RStudio**

– le quadrant supérieur gauche est dédié aux différents fichiers de travail,

– le quadrant inférieur gauche correspond à ce que l'on appelle la console, c'est-à-dire à R proprement dit,

– le quadrant supérieur droit permet de connaître :
○ La liste des objets en mémoire ou environnement de travail (onglet Environment),
○ l'historique des commandes saisies dans la console (onglet History).

– le quadrant inférieur droit affiche :
○ La liste des fichiers du répertoire de travail (onglet Files),
○ les graphiques réalisés (onglet Plots),
○ la liste des extensions disponibles (onglet Packages),
○ l'aide en ligne (onglet Help),
○ un Viewer utilisé pour visualiser certains types de graphiques au format web. [16]

4.4 Installation des packages

L'installation des packages est une étape préalable essentielle à la mise en place de l'analyse. Deux packages principales ont été utilisés dans ce travail : RCHIC et Tidyverse. Le package RCHIC permet de mobiliser des outils d'analyse statistique implicative directement dans l'environnement R. Le package Tidyverse, quant à lui, regroupe un ensemble d'outils facilitant la manipulation, le nettoyage et la visualisation des données. Ces paquets constituent ainsi la base technique nécessaire à la préparation et à la conduite rigoureuse de l'analyse.

4.4.1 RCHIC

L'installation du package RCHIC (présenté dans le chapitre 1) nécessite l'ajout préalable de plusieurs packages indispensables à son bon fonctionnement. Chacun de ces packages remplit une fonction spécifique :

– stringr : fournit des fonctions simples et cohérentes pour la manipulation des chaînes de caractères,

– tcltk2 : permet l'utilisation d'interfaces graphiques via Tcl/Tk, notamment pour la création de fenêtres interactives dans R,

– Rcpp : facilite l'intégration de code C++ dans R, ce qui permet d'optimiser les performances des calculs complexes,

– BiocManager : gestionnaire de paquets pour la plateforme Bioconductor, utilisé ici pour installer le paquet Rgraphviz,

– Rgraphviz : permet la visualisation de graphes, étape essentielle pour représenter les structures issues de l'analyse implicative.[17]

La structure d'installation, à effectuer uniquement lors de la première utilisation sur un système Windows, est la suivante :

Installation des dépendances de base

install.packages(c("stringr", "tcltk2", "Rcpp"))

Installation du gestionnaire Bioconductor (si nécessaire)

if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")

Installation du l'extension Rgraphviz via Bioconductor

BiocManager : :install("Rgraphviz")

Installation manuelle du l'extension RCHIC depuis l'archive ZIP

install.packages("https ://members.femto-st.fr/raphael-couturier/sites/femto-st.fr.raphael-couturier/fil
repos = NULL, type = "win.binary") [17]

Après l'exécution de ces instructions, le package RCHIC est correctement installé. Pour l'utiliser, il convient de le charger dans l'environnement R à l'aide de la commande suivante [17] :

library(rchic)

rchic()

Lors de ce chargement, une fenêtre spécifique s'ouvre automatiquement, signalant que le paquet fonctionne correctement. Cette interface graphique permet d'accéder aux différentes options d'analyse implicative proposées par RCHIC. La figure (3.3) illustre cette fenêtre telle qu'elle apparaît à l'ouverture du RCHIC.

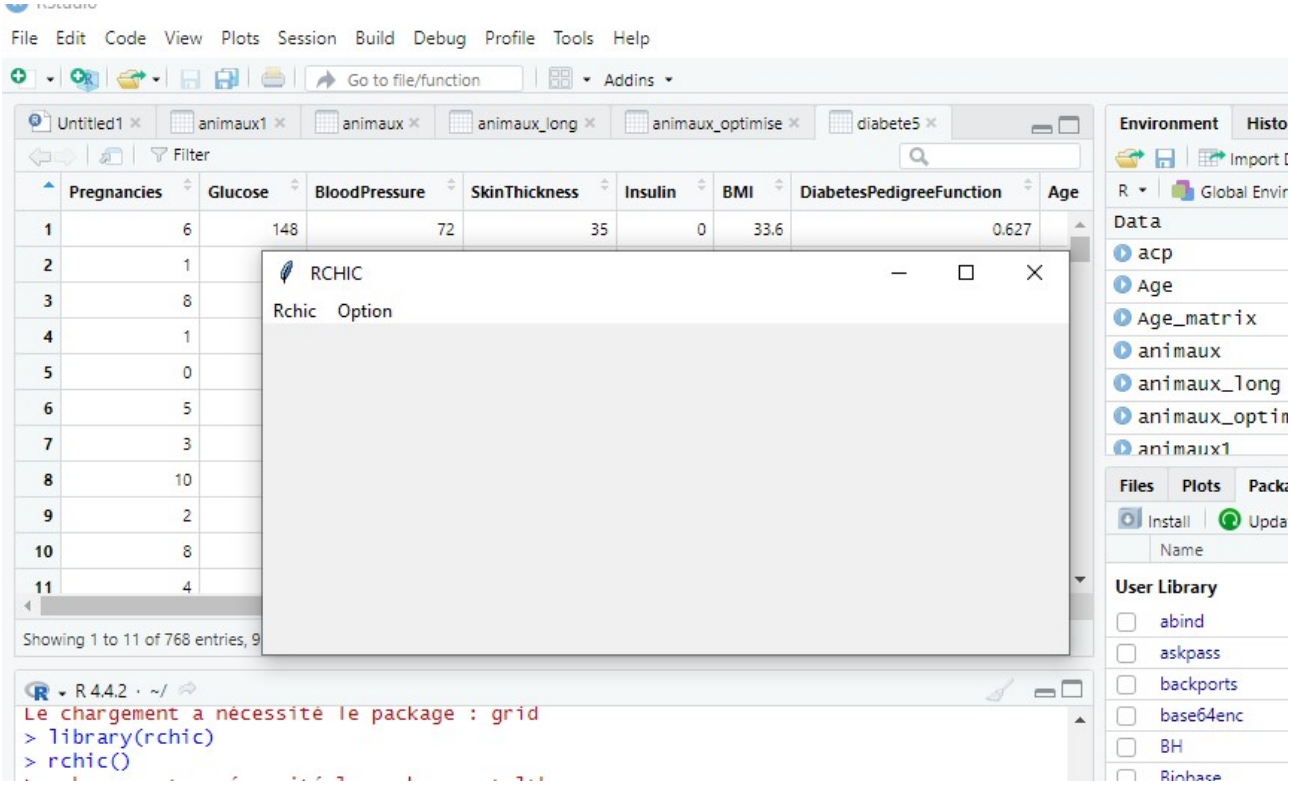

Figure (3.3) : Fenêtre RCHIC.

4.4.2 Tidyverse

Le terme Tidyverse est une contraction de tidy (qu'on pourrait traduire par "bien rangé") et de universe. Il s'agit en fait d'une collection de packages conçues pour travailler ensemble et basées sur une philosophie commune.[16]

Un des objectifs de ces packages est de fournir des fonctions avec une syntaxe cohérente, qui fonctionnent bien ensemble, et qui retournent des résultats prévisibles. Elles sont en grande partie issues du travail d'Hadley Wickham, qui travaille désormais pour RStudio.[16]

Pour installer le package Tidyverse, il est possible d'utiliser l'une des deux méthodes classiques : soit en cliquant sur le bouton Install dans l'onglet Packages de RStudio, soit en saisissant la commande suivante dans la console R [16] :

install.packages(tidyverse)

Cette commande permet d'installer plusieurs packages qui constituent le ≪cœur≫ du Tidyverse, à savoir :
– ggplot2 (visualisation),
– dplyr (manipulation des données),
– tidyr (remise en forme des données),
– purrr (programmation),
– readr (importation de données),
– tibble (tableaux de données),
– forcats (variables qualitatives),
– stringr (chaînes de caractères). [16]

De la même manière, charger le package comme suit [16] :

library(tidyverse)

4.5 Justification du choix du logiciel R et des outils utilisés

Après avoir présenté les choix techniques opérés, il convient à présent d'en expliciter les raisons. Cette section vise ainsi à justifier le recours au langage R ainsi qu'aux outils mobilisés dans le cadre de cette étude, en mettant en évidence leur pertinence au regard des objectifs méthodologiques poursuivis.

4.5.1 Choix du logiciel R

Le choix du logiciel R dans le cadre de ce travail s'est imposé naturellement en raison de sa spécialisation dans le traitement et l'analyse statistique des données.

De plus, R présente une grande compatibilité avec différents formats de données (CSV, Excel, SQL, etc.) et permet une reproductibilité des analyses grâce à sa structuration par scripts. Ces caractéristiques facilitent la traçabilité des étapes d'analyse, un critère fondamental dans un travail académique et scientifique.

4.5.2 Choix du RStudio

RStudio a été utilisé comme interface de développement en raison de sa convivialité et de son intégration complète avec le logiciel R. Cet environnement de développement intégré (IDE) facilite la rédaction, l'exécution et l'organisation du code, tout en offrant des outils puissants pour la visualisation des résultats, la gestion des fichiers et la surveillance des variables en mémoire.

4.5.3 Justification des packages utilisées

Pour mener à bien l'Analyse Statistique Implicative, plusieurs packages ont été installées et utilisées.

Le package RCHIC constitue l'élément central de ce travail, car il implémente les outils nécessaires à l'application de la méthode ASI. Il permet de générer des chaînes implicatives, de visualiser les relations conditionnelles entre variables, et de produire des graphiques explicites facilitant l'interprétation.

Par ailleurs, Le package Tidyverse a été mobilisé pour réaliser les opérations de nettoyage, de filtrage, de regroupement et de mise en forme des données en amont de l'analyse implicative. Son intégration a grandement facilité la préparation des données, en garantissant une organisation rigoureuse et une meilleure lisibilité des traitements effectués.

4.6 Présentation des données à traiter

Dans cette section, nous présentons les données que nous avons utilisées pour notre analyse. Il s'agit d'un jeu de données médical portant sur le diabète. Avant de passer à l'analyse, nous trouvons utile de préciser le contexte d'origine des données, leur structure, ainsi que les objectifs poursuivis à travers leur traitement par la méthode d'Analyse Statistique Implicative.

4.6.1 Origine et contexte des données

Le jeu de données utilisé dans cette étude provient de l'Institut national du diabète et des maladies digestives et rénales (National Institute of Diabetes and Digestive and Kidney Diseases - NIDDK) aux États-Unis. Il a été constitué dans l'objectif de développer des outils de diagnostic permettant de prédire la probabilité qu'une patiente soit atteinte de diabète de type 2, à partir de mesures cliniques simples.[18]

Les cas sélectionnés proviennent d'un sous-ensemble spécifique d'une base de données médicale plus large. Toutes les personnes incluses sont des femmes âgées de 21 ans ou plus, d'origine amérindienne Pima, vivant en Arizona. Ce groupe a été choisi parce que le diabète y est relativement fréquent, ce qui en fait une population intéressante pour mieux comprendre les facteurs liés à cette maladie. [18]

Dans cette recherche, ce jeu de données est mobilisé pour illustrer et appliquer une méthode d'analyse originale : l'Analyse Statistique Implicative, à l'aide du RCHIC. L'objectif est d'explorer les relations implicatives entre les variables cliniques, en mettant en évidence des configurations fréquentes et stables associées à la présence de diabète.

4.6.2 Description de la structure du jeu de données

Le jeu de données utilisé est fourni sous la forme d'un fichier au format CSV (Comma-Separated Values), facilement exploitable avec le logiciel R. Il contient 768 lignes, représentant chacune une patiente, et 9 colonnes, correspondant aux variables observées. Parmi ces 9 variables, on distingue 8 variables explicatives, qui décrivent les caractéristiques médicales et personnelles des patientes, et 1 variable cible, qui indique la présence ou l'absence du diabète.[18]

Les variables incluses dans ce jeu de données, accompagnées de leurs abréviations respectives, sont les suivantes :

– GRO : nombre de grossesses,

– GLU : concentration de glucose dans le plasma,

– PA : pression artérielle diastolique (en mm Hg),

– EPC : épaisseur du pli cutané tricipital (en mm),

– INS : taux d'insuline sérique (en U/ml),

– IMC : indice de masse corporelle (poids en kg / taille² en m²),

– FHD : indicateur de prédisposition héréditaire au diabète,

– AGE : âge de la patiente (en années),

– RES : variable cible indiquant si la patiente est atteinte de diabète (1) ou non (0). [18]

La figure (3.4) suivante illustre la structure générale du jeu de données ainsi que les variables analysées.

	GRO	GLU	PA	EPC	INS	IMC	FHD	AGE	RES
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0

Figure (3.4) : Extrait du jeu de données Pima Indian Diabetes au format .csv.

4.7 Préparation des données à traiter

Avant de pouvoir appliquer l'ASI à l'aide de RCHIC, il est indispensable de préparer les données de manière rigoureuse. En effet, RCHIC repose sur des exigences spécifiques en matière de format et de structuration, rendant cette phase préparatoire essentielle à la validité de l'analyse.

L'ensemble du processus de préparation a été réalisé à l'aide du logiciel R, via l'interface RStudio, qui offre un environnement souple et puissant pour le traitement, la transformation et l'organisation des données.

Cette phase de préparation comprend trois sous-étapes principales :

– le nettoyage des données brutes,

– la discrétisation des variables continues en catégories,

– la transformation binaire des variables catégorielles.

Ces opérations permettent de structurer les données dans un format exploitable par Rchic, garantissant ainsi la validité et la pertinence des analyses implicatives à venir.

4.7.1 Nettoyage des données brutes

La première étape consiste en un nettoyage des données brutes. Elle comprend notamment la suppression des lignes en double ainsi que l'élimination des observations présentant des valeurs nulles (zéros) dans certaines variables du jeu de données. Cette sélection rigoureuse vise à garantir une première qualité des données, nécessaire pour pouvoir appliquer ultérieurement l'Analyse Statistique Implicative.

Initialement, le jeu de données comprenait les informations de 392 patientes. Après nettoyage et filtrage, le nombre d'observations retenues a été réduit, constituant une base fiable, prête à être utilisée pour l'étape suivante, à savoir la discrétisation des variables.

La figure (3.5) présente la structure du jeu de données après le nettoyage, constituant ainsi une base saine et exploitable pour les étapes ultérieures de préparation.

	GRO	GLU	PA	EPC	INS	IMC	FHD	AGE	RES
1	1	89	66	23	94	28.1	0.167	21	0
2	0	137	40	35	168	43.1	2.288	33	1
3	3	78	50	32	88	31.0	0.248	26	1
4	2	197	70	45	543	30.5	0.158	53	1
5	1	189	60	23	846	30.1	0.398	59	1
6	5	166	72	19	175	25.8	0.587	51	1
7	0	118	84	47	230	45.8	0.551	31	1
8	1	103	30	38	83	43.3	0.183	33	0
9	1	115	70	30	96	34.6	0.529	32	1
10	3	126	88	41	235	39.3	0.704	27	0
11	11	143	94	33	146	36.6	0.254	51	1
12	10	125	70	26	115	31.1	0.205	41	1
13	1	97	66	15	140	23.2	0.487	22	0

Figure (3.5) : Extrait du jeu de données après nettoyage.

4.7.2 La discrétisation des variables continues en catégories

Après le nettoyage initiale des données, nous avons procédé à une étape clé : la discrétisation des variables continues. Cette opération consiste à convertir les variables quantitatives (telles que la glycémie, l'indice de masse corporelle, le taux d'insuline, ainsi que d'autres variables mesurées) en variables qualitatives ordinales.

Nous avons choisi d'appliquer la discrétisation par défaut, qui divise automatiquement chaque variable numérique en trois intervalles fixes. Par exemple, la variable INS (Insuline) a été répartie en trois catégories : faible (valeurs comprises entre 1 et 282), moyenne (283 à 564) et élevée (565 à 846). Cette méthode facilite un traitement uniforme des données tout en respectant les caractéristiques propres à chaque variable. La seule variable exclue de cette opération a été RES, qui renseigne sur la présence ou l'absence de diabète chez les patientes. Elle a été reformulée en deux modalités explicites : Diabétique (pour la valeur 1) et Non Diabétique (pour la valeur 0), puis convertie en facteur (c'est-à-dire en variable catégorielle).

La figure (3.6) illustre la nouvelle structure du jeu de données après discrétisation, où chaque variable continue a été convertie en modalité qualitative, prête à être utilisée pour l'étape suivante : la transformation en binaire.

	GRO	GLU	PA	EPC	INS	IMC	FHD	AGE	RES
1	faible	faible	moyenne	faible	faible	faible	faible	faible	Non_Diabetique
2	faible	moyenne	faible	moyenne	faible	moyenne	elevee	faible	Diabetique
3	faible	faible	faible	moyenne	faible	faible	faible	faible	Diabetique
4	faible	elevee	moyenne	elevee	moyenne	faible	faible	moyenne	Diabetique
5	faible	elevee	moyenne	faible	elevee	faible	faible	moyenne	Diabetique
6	faible	elevee	moyenne	faible	faible	faible	faible	moyenne	Diabetique
7	faible	moyenne	elevee	elevee	faible	moyenne	faible	faible	Diabetique
8	faible	faible	faible	moyenne	faible	moyenne	faible	faible	Non_Diabetique
9	faible	moyenne	moyenne	moyenne	faible	moyenne	faible	faible	Diabetique
10	faible	moyenne	elevee	moyenne	faible	moyenne	faible	faible	Non_Diabetique
11	moyenne	moyenne	elevee	moyenne	faible	moyenne	faible	moyenne	Diabetique
12	moyenne	moyenne	moyenne	moyenne	faible	faible	faible	faible	Diabetique
13	faible	faible	moyenne	faible	faible	faible	faible	faible	Non_Diabetique

Figure (3.6) : Extrait du jeu de données après discrétisation.

4.7.3 Transformation binaire des variables catégorielles

La dernière étape de notre préparation consiste à transformer les variables catégorielles en variables binaires. Cette opération, appelée binarisation, est nécessaire pour rendre les données compatibles avec le logiciel RCHIC, qui nécessite un format booléen (1 ou 0) indiquant la présence ou l'absence d'une modalité.

Concrètement, nous avons transformé chaque modalité d'une variable qualitative en une variable binaire prenant la valeur 1 si l'observation appartient à cette modalité, et 0 sinon. Par exemple, la variable INS (Insuline), discrétisée en trois modalités (faible, moyenne, élevée), a été remplacée par trois variables : INS_faible, INS_moyenne, INS_élevée. Chacune indiquant la présence (1) ou l'absence (0) de la modalité correspondante.

Nous avons également appliqué cette transformation à la variable RES, qui renseigne sur la présence ou l'absence de diabète. Dans notre étude, nous nous sommes concentrés sur les cas de diabète. Ainsi, seule la modalité « Diabétique » a été retenue, sous la forme d'une variable binaire nommée RES_Diabétique, valant 1 pour les patientes diabétiques et 0 sinon.

La figure (3.7) présente la structure finale du jeu de données après transformation binaire, prête à être utilisée dans l'ASI à l'aide du logiciel Rchic.

	GRO_faible	GRO_moyenne	GRO_elevee	GLU_faible	GLU_moyenne	GLU_elevee	PA_faible
1	1	0	0	1	0	0	0
2	1	0	0	0	1	0	1
3	1	0	0	1	0	0	1
4	1	0	0	0	0	1	0
5	1	0	0	0	0	1	0
6	1	0	0	0	0	1	0
7	1	0	0	0	1	0	0
8	1	0	0	1	0	0	1
9	1	0	0	0	1	0	0
10	1	0	0	0	1	0	0
11	0	1	0	0	1	0	0
12	0	1	0	0	1	0	0
13	1	0	0	1	0	0	0

Figure (3.7) : Extrait du jeu de données après transformation en variables binaires.

4.8 Application de l'Analyse Statistique Implicative (ASI)

Une fois les données préparées selon les exigences de format et de structure, nous avons procédé à l'application de l'ASI à l'aide de RCHIC, outil spécifiquement conçu pour ce type d'analyse. À cette étape, les données ont été entièrement converties en format binaire, pour garantir la compatibilité avec le traitement implicatif.

Une fois l'exécution lancée, RCHIC procède au calcul de l'indice d'implication et de la valeur de confiance pour chaque règle implicative extraite à partir des données binarisées. Ces résultats sont automatiquement enregistrés dans un fichier nommé transaction.out. Ce fichier rassemble l'ensemble des relations implicatives identifiées, accompagnées de leurs indicateurs statistiques respectifs. La figure (3.8) illustre la structure du fichier de sortie produit par Rchic.

hyp -> con	occurrence(hyp)	occurrence(con)	support(rule)	confidence	classical index
INS_faible -> FHD_faible	352.0000000000000000	340.0000000000000000	89.7959183673469425	86.9318181818181728	50.1564980935995663
FHD_faible -> INS_faible	340.0000000000000000	352.0000000000000000	86.7346938775510239	90.0000000000000000	50.1806054406201767
INS_faible -> AGE_faible	352.0000000000000000	327.0000000000000000	89.7959183673469425	84.6590909090909065	71.6221600770950317
AGE_faible -> INS_faible	327.0000000000000000	352.0000000000000000	83.4183673469387656	91.1314984709480171	74.3297110437219288
INS_faible -> IMC_faible	352.0000000000000000	239.0000000000000000	89.7959183673469425	63.0681818181818201	73.5746711492538452
IMC_faible -> INS_faible	239.0000000000000000	352.0000000000000000	60.9693877551020478	92.8870292887029336	92.4181310721936740
INS_faible -> EPC_faible	352.0000000000000000	147.0000000000000000	89.7959183673469425	38.9204545454545467	63.1979137659072876
EPC_faible -> INS_faible	147.0000000000000000	352.0000000000000000	37.5000000000000000	93.1972789115646236	88.1535588470984948
INS_faible -> GLU_faible	352.0000000000000000	131.0000000000000000	89.7959183673469425	37.2159090909090864	80.8713659644126892
GLU_faible -> INS_faible	131.0000000000000000	352.0000000000000000	33.4183673469387799	100.0000000000000000	99.9998434565190166
INS_faible -> GLU_elevee	352.0000000000000000	80.0000000000000000	89.7959183673469425	15.9090909090909083	17.2035992145538330
GLU_elevee -> INS_faible	80.0000000000000000	352.0000000000000000	20.4081632653061220	70.0000000000000000	0.0001664290318670
INS_faible -> PA_elevee	352.0000000000000000	78.0000000000000000	89.7959183673469425	18.4659090909090899	38.2013142108917236
PA_elevee -> INS_faible	78.0000000000000000	352.0000000000000000	19.8979591836734677	83.3333333333333428	3.2987220873037848
INS_faible -> GRO_moyenne	352.0000000000000000	75.0000000000000000	89.7959183673469425	17.8977272727272734	39.8339629173278809
GRO_moyenne -> INS_faible	75.0000000000000000	352.0000000000000000	19.1326530612244881	84.0000000000000000	4.8524324473477431
INS_faible -> AGE_moyenne	352.0000000000000000	63.0000000000000000	89.7959183673469425	14.7727272727272734	39.5132839679718018

Figure (3.8) : Extrait du fichier transaction.out issu des données Pima Indian Diabetes.

Afin d'interpréter les résultats fournis par RCHIC, nous avons tout d'abord fixé un code couleur permettant de représenter visuellement l'intensité d'implication, c'est-à-dire la force du lien entre les antécédents et les conséquents dans chaque règle. Les implications dont l'intensité est supérieure ou égale à 90% sont représentées en rouge,celles comprises entre 80% et 89% en vert, entre 70% et 79% en bleu et entre 65% et 69% en bleu clair.

Ensuite, afin d'affiner l'analyse, nous avons défini différents seuils de confiance, notamment 70% et 60%, permettant de filtrer les règles selon leur degré de fiabilité. Cette double démarche nous a permis d'identifier les relations implicatives les plus significatives.

4.8.1 Résultats obtenus en utilisant un seuil de confiance égale à 70

La figure (3.9) ci-dessous illustre le graphe obtenu avec un seuil de confiance de 70 %, ce qui nous permet de retenir uniquement les règles les plus fiables. Ce choix vise à identifier en priorité les facteurs les plus importants associés à un état diabétique. Néanmoins, certaines implications révèlent des informations complémentaires, parfois indirectes, que nous prendrons également en compte lorsqu'elles contribuent à une meilleure compréhension de la logique des relations présentes dans les données.

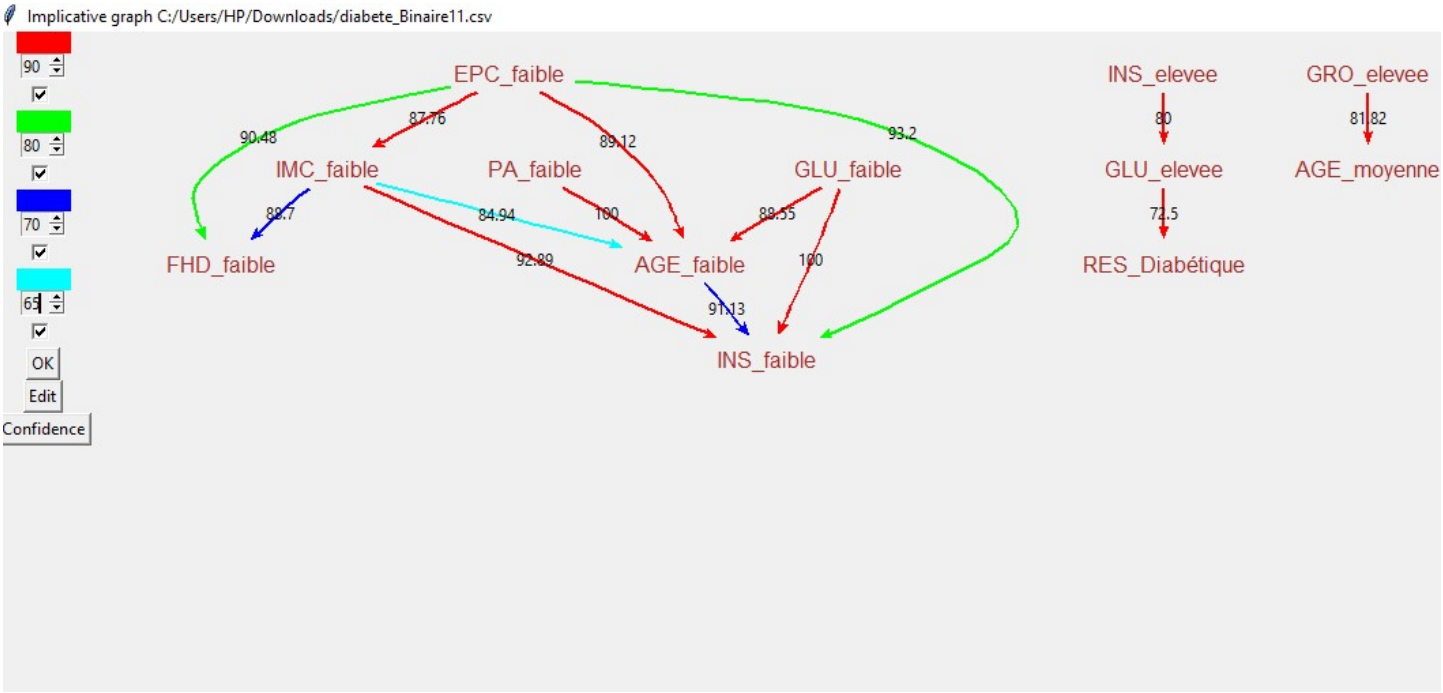

Figure (3.9) : Graphe implicatif avec un seuil de confiance égal à 70.

Avec un seuil de confiance de 70 %, nous constatons que le graphe est plus lisible et que les résultats sont cohérents. Les variables avec des valeurs faibles impliquent d'autres variables également faibles, et les valeurs élevées en impliquent d'autres élevées. Cette organisation permet d'extraire plus facilement les principales relations, en commençant par celles qui sont liées à un état de diabète.

• Implication principale vers un état de diabète

Dans cette première partie, le graphe d'implication met en évidence une relation directe et significative entre certains marqueurs biologiques et un état de diabète. Parmi les implications extraites, une seule établit un lien direct vers un état de diabète. Elle est définie comme suit :

– (INS_elevée → GLU_elevée) → RES_Diabétique

Cette implication indique que, lorsque le taux d'insuline est élevé et que celui de glucose l'est également, la personne présente un risque accru de se retrouver en état de diabète.

Analyse et interprétation

Cette implication met en évidence l'un des mécanismes caractéristiques du diabète de type 2, qui correspond à la résistance à l'insuline. Il s'agit d'une situation dans laquelle les cellules de l'organisme ne réagissent plus efficacement à l'action de l'insuline. Normalement, cette hormone, produite par le pancréas, permet au glucose d'entrer dans les cellules pour y être utilisé comme source d'énergie. En cas de résistance, les cellules deviennent moins sensibles, ce qui pousse le pancréas à produire davantage d'insuline pour tenter de compenser. Malgré cette surproduction, le glucose reste en circulation dans le sang, entraînant une hyperglycémie persistante. À plus long terme, cette surcharge peut conduire à un épuisement du pancréas, jusqu'à l'arrêt progressif de la sécrétion de l'hormone, favorisant ainsi l'installation durable du diabète de type 2. [19]

Ce profil clinique, bien documenté, confirme la validité et la pertinence des résultats obtenus par l'ASI.

• Implications significatives chez des patientes non diabétiques

Dans cette seconde partie, nous nous intéressons aux implications qui ne mènent pas directement à un état de diabète, mais qui présentent néanmoins des anomalies biologiques notables. Ces associations sont particulièrement observées chez des patientes jeunes, avec des profils physiologiques présentant des valeurs faibles de plusieurs indicateurs métaboliques. Ces configurations, bien qu'absentes de diagnostic de diabète, soulèvent des interrogations cliniques importantes. Les règles implicatives suivantes illustrent ces relations :

– GLU_faible → AGE_faible

Ce qui signifie que les patientes ayant un faible taux de glucose sont généralement jeunes,

– PA_faible → AGE_faible
Ce qui signifie qu'une pression artérielle basse est également observée chez des patientes jeunes,

– EPC_faible → AGE_faible
Ce qui indique qu'une faible épaisseur du pli cutané est typiquement associée à un jeune âge,

– IMC_faible → INS_faible
Ce qui signifie que les patientes maigres présentent souvent un taux d'insuline réduit,

– GLU_faible → INS_faible
Ce qui montre qu'un faible taux de glucose s'accompagne généralement d'un faible taux d'insuline.

Analyse et interprétation

Ces implications montrent qu'un certain nombre de patientes jeunes présentent des profils biologiques anormalement bas pour plusieurs variables comme la glycémie (glucose), l'insuline, l'indice de masse corporelle (IMC) ou encore la pression artérielle. Ces résultats suggèrent des anomalies métaboliques, même en l'absence de diagnostic de diabète.

L'hypoglycémie, c'est-à-dire un taux de glucose sanguin anormalement bas, constitue une situation à ne pas négliger, même chez des sujets non diabétiques. Le glucose est en effet la principale source d'énergie pour les cellules, en particulier pour le cerveau. Un déficit chronique ou brutal peut engendrer des troubles importants (fatigue, confusion, perte de connaissance). Lorsqu'elle est associée à une insuline basse, cela peut refléter soit une régulation excessive de la glycémie, soit un dysfonctionnement de la sécrétion pancréatique. [20]

Selon **le Manuel MSD (2023)**, l'hypoglycémie chez les patients non diabétiques est rare, et peut résulter de plusieurs causes :

– Troubles hormonaux (ex : insuffisance surrénalienne),

– Tumeurs pancréatiques (insulinome),

– Mauvaise alimentation ou jeûne prolongé,

– Maladies chroniques (foie, reins, cœur).[20]

En conséquence, un suivi médical régulier est indispensable pour identifier la cause exacte de cette hypoglycémie et mettre en place une prise en charge adaptée. La vigilance est d'autant plus importante que ces patientes ne présentent pas encore de diabète, mais possèdent déjà un terrain métabolique déséquilibré.

4.8.2 Résultats obtenus en utilisant un seuil de confiance égale à 60

La figure (3.10) ci-dessous montre le graphe d'implication que nous avons obtenu en utilisant un seuil de confiance fixé à 60%.

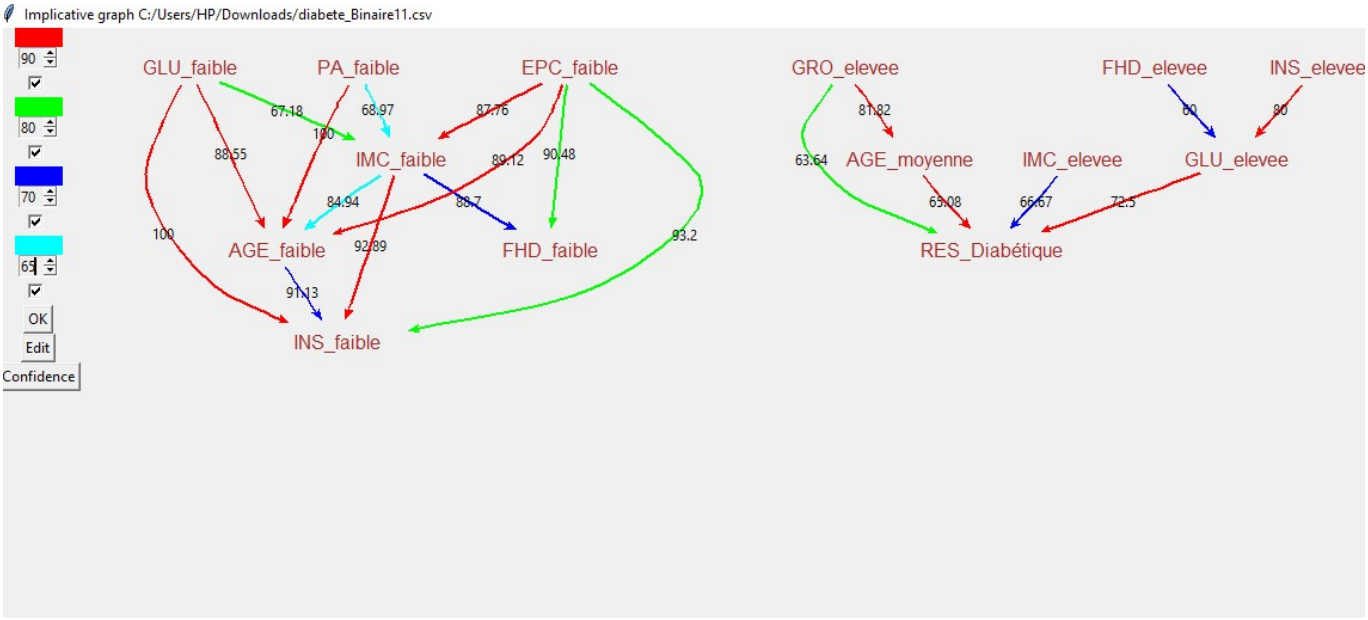

Figure (3.10) : Graphe implicatif avec un seuil de confiance égal à 60.

Avec ce seuil de confiance, nous avons remarqué que le graphe est plus complet que le précédent (Figure (3.9)). Il contient plus de règles, ce qui nous permet d'explorer un plus grand nombre de relations entre les variables. Même si ces implications proviennent d'un seuil de confiance moins élevé (60%), elles restent pertinentes et apportent des informations intéressantes sur les liens entre les variables. Les nouvelles implications mises en évidence sont les suivantes :

– (FHD_elevée → GLU_elevée) → RES_Diabétique

Cette implication suggère qu'une combinaison d'antécédents familiaux de diabète et d'un taux de glucose élevé est fortement liée à un état diabétique.

– (GRO_elevée → AGE_moyenne) → RES_Diabétique

Cela indique que chez les patientes ayant eu de nombreuses grossesses et présentant un âge moyen, le risque de diabète est également accru.

– IMC_elevée → RES_Diabétique
Cette implication met en évidence que l'obésité, caractérisée par un IMC élevé, est fortement associée à l'état diabétique.

Analyse et interprétation

Les implications obtenues à partir du graphe d'analyse implicative avec un seuil de confiance de 60% permettent de mieux comprendre certains profils à risque de développer un diabète de type 2, même avant l'apparition clinique de la maladie. Voici une analyse détaillée :

– (FHD_elevée → GLU_elevée) → RES_Diabétique

Cette implication illustre l'importance du facteur héréditaire dans la survenue du diabète. Avoir un ou plusieurs membres de la famille atteints de diabète augmente fortement le risque personnel, surtout si ce facteur est combiné à une hyperglycémie. Selon **l'Inserm**, le risque est multiplié si les deux parents sont atteints. Ce lien s'explique à la fois par une prédisposition génétique et par des habitudes de vie souvent partagées au sein de la famille.[21]

– (GRO_elevée → AGE_moyenne) → RES_Diabétique

Cette implication peut faire référence au diabète gestationnel ou à l'épuisement progressif du métabolisme lié à plusieurs grossesses. Chaque grossesse sollicite fortement la régulation du glucose. Chez certaines femmes, cela peut entraîner une insulinorésistance temporaire, voire permanente si le nombre de grossesses est élevé. De plus, un âge moyen (environ 30-40 ans) augmente ce risque. [22]

– IMC_elevée → RES_Diabétique.

Cette implication met en évidence le rôle déterminant de l'obésité, reconnue comme l'un des principaux facteurs de risque du diabète de type 2. En particulier, la graisse abdominale contribue à l'apparition d'une résistance à l'insuline, qui empêche le glucose d'être correctement utilisé par les cellules. Le pancréas compense en produisant davantage d'insuline, ce qui peut aboutir à un déséquilibre métabolique durable. Selon **l'Organisation Mondiale de la Santé (OMS)**, 80% des personnes atteintes de diabète de type 2 sont en surpoids ou obèses.[23]

D'après les deux graphes étudiés, nous avons pu identifier plusieurs facteurs associés au développement du diabète de type 2, parmi lesquels :

– La résistance à l'insuline,

– Un nombre élevé de grossesses,

– L'obésité,

– Un terrain héréditaire favorable à la maladie.

Ces éléments confirment l'importance d'une surveillance précoce et d'une prévention adaptée chez les personnes présentant ces caractéristiques.

4.9 L'intérêt de l'ASI dans notre étude

L'Analyse Statistique Implicative s'est révélée être un outil précieux dans notre étude. Elle permet de faire ressortir des relations importantes entre des variables, même lorsque celles-ci ne sont pas facilement visibles avec des méthodes statistiques classiques. Grâce à cette méthode, nous avons pu identifier des profils à risque et des facteurs associés au diabète de type 2 de manière claire et organisée.

Ce qui rend l'ASI particulièrement intéressante, c'est la fiabilité des résultats qu'elle fournit. En effet, plusieurs études dans le domaine médical ont confirmé que les liens mis en évidence par cette méthode correspondent souvent à des faits cliniques bien établis. Dans notre cas, les résultats obtenus sont en accord avec les connaissances actuelles sur les causes et les facteurs de risque du diabète.

De plus, cette méthode nous a permis de repérer des informations utiles chez des personnes non diabétiques, mais présentant des caractéristiques ou des déséquilibres biologiques qui peuvent indiquer un risque de développer la maladie plus tard. Cela montre que l'ASI peut aider à repérer certains signes à un stade très tôt.

Nous soulignons également que cette méthode peut être facilement utilisée dans d'autres situations, que ce soit pour étudier d'autres maladies chroniques ou pour identifier des personnes à risque. Sa simplicité d'utilisation et la solidité de ses résultats en font un outil très intéressant, autant pour la recherche que pour la prévention en santé publique.

4.10 Conclusion

Dans ce dernier chapitre, nous avons présenté la mise en œuvre pratique de notre étude en mobilisant l'Analyse Statistique Implicative. Nous avons d'abord introduit les outils utilisés, notamment le logiciel R, l'environnement RStudio et le package RCHIC, avant de détailler le jeu de données retenu ainsi que les traitements préparatoires effectués.

Nous avons ensuite appliqué l'ASI pour générer et interpréter les graphes d'implication, ce qui nous a permis de faire ressortir des relations significatives entre les variables. Ce travail a mis en lumière l'intérêt de cette méthode pour mieux comprendre les facteurs liés au diabète de type 2, en apportant un regard structuré sur les profils à risque.

Conclusion générale

5 Conclusion générale

Ce mémoire s'inscrit dans une démarche visant à appliquer l'Analyse Statistique Implicative à l'étude de données médicales, dans le but d'identifier les profils à risque de diabète de type 2. Dans un premier temps, nous avons présenté les fondements du Data Mining, une discipline essentielle pour extraire des connaissances à partir de grands ensembles de données. Cependant, ces approches classiques présentent certaines limites, notamment lorsqu'il s'agit de mettre en évidence des relations directionnelles ou implicites entre variables.

Afin de pallier ces limites, nous avons introduit l'ASI, une méthode d'analyse non symétrique permettant de dégager des règles logiques fondées sur des relations stables et signifiantes. Elle constitue une alternative pertinente pour explorer les structures cachées au sein des données, en particulier dans des contextes médicaux complexes tels que celui du diabète.

• Contribution

À travers les différentes étapes de ce mémoire, nous avons :

– Démontré l'intérêt théorique et pratique de l'ASI.

– Présenté l'environnement technique de l'analyse, en mobilisant le logiciel R et le package RCHIC.

– Appliqué l'ASI à un jeu de données médicales afin d'extraire des règles d'implication entre des variables comme le glucose, l'insuline, la pression artérielle, etc.

– Identifié des profils de patientes à risque et mis en évidence des déséquilibres chez certaines patientes non diabétiques, suggérant un potentiel de prévention.

– Vérifié que certains résultats obtenus sont en adéquation avec les connaissances médicales actuelles, renforçant ainsi la crédibilité de cette approche.

Ce projet nous a permis de mettre en pratique nos connaissances en statistiques, en analyse de données et en santé publique. Il a constitué notre première application concrète d'une méthode statistique implicative dans un domaine essentiel : la prévention du diabète. Ce travail nous a offert une expérience enrichissante, tant sur le plan technique, grâce à l'utilisation d'outils comme R et RCHIC, que sur le plan méthodologique, en nous initiant à une nouvelle approche d'analyse exploratoire des données médicales.

- **Perspectives et travaux futurs**

Cette étude ouvre la voie à plusieurs perspectives intéressantes, tant dans le domaine médical que dans d'autres secteurs, parmi lesquelles figurent notamment :

– Étendre l'analyse à un échantillon plus large, pour valider les résultats sur des données plus diversifiées et renforcer leur généralisation.

– Appliquer l'ASI à d'autres maladies, y compris les maladies rares, qui sont souvent difficiles à étudier en raison du faible nombre de cas. Grâce à sa capacité à détecter des relations cachées, l'ASI pourrait aider à mieux comprendre ces pathologies peu connues.

– Développer des outils de visualisation interactifs permettant aux professionnels de santé ou aux utilisateurs finaux de comprendre facilement les résultats, même sans formation statistique.

– Explorer l'utilisation de l'ASI dans d'autres domaines, notamment dans le e-commerce, où cette méthode pourrait aider à identifier des profils d'acheteurs, détecter des habitudes d'achat, ou recommander des produits de manière plus ciblée, en se basant sur des relations implicites entre les comportements des clients.

Bibliographie

[1] SAMET, Ahmed. Théorie des fonctions de croyance : application des outils de data mining pour le traitement des données imparfaites. 2014. Thèse de doctorat. Artois.

[2] Sellami, LYNDA. Approche DATA mining pour le détection d'intrusions. 2009. Thèse de doctorat. Université de Béjaia-Abderrahmane Mira.

[3] Ghanem,souhila. Des Contributions autour de l'Analyse Statistique Implicative. 2022. Thèse de doctorat. Université de Béjaia-Abderrahmane Mira.

[4] NEMICHE, Mohamed. Master MASI.

[5] Khaled,Hayette. 2020. Comparaison de l'Analyse Statistique Implicative et des Outils de Fouille de Données. Thèse de doctorat. Université de Béjaia-Abderrahmane Mira.

[6] GRAS, Régis, RÉGNIER, Jean-Claude, MARINICA, Claudia, et al. L'analyse statistique implicative Méthode exploratoire et confirmatoire à la recherche de causalités. Cépaduès Editions, 2013.

[7] RÉGNIER, Jean-Claude, SLIMANI, Yahia, GRAS, Régis, et al. Analyse statistique implicative. Des sciences dures aux sciences humaines et sociales. 2015.

[8] GRAS, Régis, KUNTZ, Pascale, et BRIAND, Henri. Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. Mathématiques et sciences humaines. Mathematics and social sciences, 2001, no 154.

[9] GHANEM, Souhila et COUTURIER, Raphaël. Ajout de la confiance au graphe implicatif. In : Actes du 8ème Colloque International sur Analyse Statistique Implicative. 2015. p. 117-129.

[10] COUTURIER, Raphaël et GRAS, Régis. CHIC : traitement de données avec l'analyse implicative. In : EGC. 2005. p. 679-684.

[11] COUTURIER, Raphaël. Traitement de l'analyse statistique dans chic. Caen : Actes des Journees : LA FOUILLE DANS LES DONNEES PAR LA METHODE D'ANALYSE STATISTIQUE IMPLICATIVE Applications et traitament par CHIC, 2000.

[12] COUTURIER, Raphaël et ALMOULOUD, Saddo Ag. Historique et fonctionnalités de CHIC. 2009.

[13] VERSCHEURE, Ingrid, AMADE-ESCOT, Chantal, et CHIOCCA, Catherine-Marie. Représentations du volley-ball scolaire et genre des élèves : pertinence de l'inventaire des rôles de sexe de Bem. Revue française de pédagogie. Recherches en éducation, 2006, no 154, p. 125-144.

[14] DE AQUINO, Rafael Santos, ACIOLY-REGNIER, Nadja Maria, DOS ANJOS CARNEIRO-LEÃO, Ana Maria, et al. Multiculturalité et santé mentale à l'école en temps de pandémie. In : congrès international d'Actualité de la Recherche en Éducation et en Formation (AREF). 2022.

[15] KHALED, Hayette, GHANEM, Souhila, et COUTURIER, Raphael. Analysis of Bejaia University Computer Science students' marks through the CHIC software and Statistical Implicative Analysis. In : 2014 4th International Symposium ISKO-Maghreb : Concepts and Tools for knowledge Management (ISKO-Maghreb). IEEE, 2014. p. 1-8.

[16] BARNIER, Julien, BIAUDET, Julien, BRIATTE, François, BOUCHET-VALAT, Milan, GALLIC, Ewen, GIRAUD, Frédérique, GOMBIN, Joël, KAUFFMANN, Mayeul, LALANNE, Christophe, LARMARANGE, Joseph, ROBETTE, Nicolas. Introduction à l'analyse d'enquêtes avec R et RStudio. Mars 2018.

[17] COUTURIER, Raphaël. RCHIC [en ligne]. Disponible sur : https://members.femto-st.fr/raphael-couturier/en/rchic [consulté le 28 février 2025].

[18] NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES. Pima Indians Diabetes Dataset [en ligne]. Disponible sur : https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database[consulté le 15 avril 2025].

[19] MSD MANUAL. Diabète sucré (DS) [en ligne]. Disponible sur : https://www.msdmanuals.com/fr/accueil/troubles-hormonaux-et-m%C3%A9taboliques/diab%C3%A8te-sucr%C3%A9-ds-et-troubles-du-m%C3%A9tabolisme-de-la-glyc%C3%A9mie/diab%C3%A8te-sucr%C3%A9-ds [consulté le 15 mai 2025].

[20] MSD MANUAL. Hypoglycémie [en ligne]. Disponible sur : https://www.msdmanuals.com/fr/accueil/troubles-hormonaux-et-m%C3%A9taboliques/diab%C3%A8te-sucr%C3%A9-ds-et-troubles-du-m%C3%A9tabolisme-de-la-glyc%C3%A9mie/hypoglyc%C3%A9mie [consulté le 16 mai 2025].

[21] INSERM. Diabète de type 2 [en ligne]. Disponible sur : https://www.inserm.fr/dossier/diabete-type-2/[consulté le 17 mai 2025].

[22] MSD MANUAL. Diabète pendant la grossesse [en ligne]. Disponible sur : https://www.msdmanuals.com/fr/accueil/probl%C3%A8mes-de-sant%C3%A9-de-la-femme/grossesse-compliqu%C3%A9e-par-la-maladie/diab%C3%A8te-pendant-la-grossesse[consulté le 17 mai 2025].

[23] ORGANISATION MONDIALE DE LA SANTÉ. Diabète [en ligne]. Disponible sur : https://www.who.int/fr/news-room/fact-sheets/detail/diabetes [consulté le 17 mai 2025].

Résumé

Ce mémoire porte sur l'application de l'Analyse Statistique Implicative (ASI) à l'étude du diabète de type 2, à partir du jeu de données médicales Pima Indian Diabetes. L'objectif principal est de détecter des profils à risque en identifiant des liens logiques entre différentes variables cliniques, souvent invisibles avec les méthodes statistiques classiques. Après avoir présenté les fondements théoriques du Data Mining et les limites des approches traditionnelles, l'ASI est introduite comme une méthode complémentaire, capable de faire émerger des connaissances nouvelles.

À l'aide du logiciel R et le package RCHIC, nous avons mis en œuvre une analyse complète, allant du prétraitement des données à la visualisation des résultats. L'étude a permis de révéler des profils à haut risque de diabète, ainsi que des déséquilibres biologiques chez des patientes non diabétiques, suggérant un terrain métabolique fragile. Ces résultats soulignent l'intérêt de l'ASI pour la prévention, l'interprétation fine des données, et l'aide à la décision médicale.

Ce travail met en valeur l'intérêt de l'Analyse Statistique Implicative pour mieux comprendre les facteurs du diabète de type 2. Il souligne son utilité pour la prévention et l'analyse approfondie des données médicales.

Mots-clés : ASI, Data Mining, R, RCHIC, Pima Indian Diabetes.

Abstract

This thesis focuses on the application of Statistical Implicative Analysis (SIA) to the study of type 2 diabetes, using the widely recognized Pima Indian Diabetes dataset. The main objective is to identify at-risk profiles by uncovering logical relationships between clinical variables that may not be detected by traditional statistical methods. After presenting the foundations of Data Mining and the limitations of conventional approaches, SIA is introduced as an alternative method capable of generating new and meaningful insights.

Using R software and the RCHIC package, we carried out a full analysis—from data preprocessing to result visualization. The study revealed strong implicative links associated with diabetic patients, as well as biological imbalances among non-diabetic women, suggesting early metabolic vulnerability. These findings demonstrate the relevance of SIA for prevention, early detection, and medical decision support.

This work highlights the relevance of Statistical Implicative Analysis in better understanding the factors of type 2 diabetes. It emphasizes its usefulness for prevention and the in-depth analysis of medical data.

Keywords : SIA, Data Mining, R, RCHIC, Pima Indian Diabetes.