**République Algérienne Démocratique et Populaire**

**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

**Université Abbas Laghrour Khenchela**

**Faculté des Sciences et de la Technologie**

**Département de Mathématique et Informatique**

# Mémoire de fin d'étude

Domaine : **Informatiques**

Filière : **Informatiques**

*Spécialité :* ***sécurité et technologie web***

## Thème

# L'extraction des règles d'association entre les attributs d'une base de données éducative

# Cas d'étude : Etudiants de la première année mathématique et informatique

**Encadré par :**

-Dr.Houassi Hichem

**co-encadreur :**

-Ledmi Makhlouf

**Réalisé par:**

- Mordjane Moncef

-Ghoul abd arraouf

***Année Universitaire : 2019/2020***

# Remerciement

*Nos remerciements vont à l'adresse de tous ceux qui, de près ou de loin, ont aidé à la concrétisation de ce travail.*

*Nous remercions Dieu pour nous avoir permis d'aller jusqu'au bout de ce mémoire, nos parents pour leurs honnêtes et infaillibles sacrifices.*

*Nous vaudrions remercier particulièrement notre encadreur, Dr.Houassi Hichem et co-encadeur Ladmi makhlouf, pour le soutien prodigué et lui exprimer notre reconnaissance et toute notre gratitude pour avoir encadré ce travail et pour la confiance qu'il nous a accordée, ses conseils et ses encouragements.*

*Nous remercions également le président du jury et ses membres pour avoir accepté de présider et prendre part à l'évaluation de ce travail.*

*Que tous ceux que nous n'avons pas nommément cités trouvent ici l'expression de notre profonde gratitude et notre salut éternel.*

**Résumé**

La fouille de données éducatives est un processus conçu pour l'analyse de données issues de situations éducatives pour mieux comprendre les étudiants et les situations dans lesquelles ils apprennent. Parmi les travaux de recherche en fouille de données, l'extraction des règles d'association qui est la tâche principale qui a attiré le plus d'attention des chercheurs et pour laquelle beaucoup de travaux ont été effectués, elle permet la découverte de règles intelligibles et exploitables dans un ensemble de données volumineux, règles exprimant des associations entre items ou attributs dans une base de données. Ce mémoire et traite les règles d'association.

Notre travail est lié à cet aspect. Nous avons d'abord collectée les données de 250 étudiants via un questionnaire constitué de 55 questions. Notre application extraire des règles d'association entre les attributs des étudiants de première année du département de sciences et de la technologie utilisant l'algorithme apriori. Ensuite, nous présenterons les résultats de fouille de données en terme de règles d'association fortes (l'algorithme apriori).

Dans ce travail on a effectué des plusieurs scenarios pour extraire et analyser les règles d'association fort

L'extraction des règles d'association fortes se fait automatiquement après l'appel de l'algorithme sur la base de données sélectionnée, ces règles d'association fortes nous l'obtenons après avoir exécuté plusieurs scénarios .avant chaque extraction d'un scénario il faut choisir le nombre de rôle, la valeur du seuil minimal de support et de confiance.

Enfin, les résultats de ces différents scénarios seront présentés et discutés.

**Mot clés :** Extraction de connaissances à partir de données, Fouille de données éducatives, Algorithme apriori, Fouille de données règle d'association.

## Abstract

Educational data mining is a process designed for analyzing data from educational situations to better understand students and the situations in which they learn. Among the research work in data mining, the extraction of association rules is the main task that has attracted the most attention of researchers and for which a lot of work has been done, it allows the discovery of intelligible rules and usable in a large data set, rules expressing associations between items or attributes in a database. This thesis and deals with the rules of association

Our work is linked to this aspect. We first collected data from 250 students via a questionnaire consisting of 56 questions. Our application extracts association rules between the attributes of first year students in the science and technology department using the a priori algorithm. Then, we will present the results of data mining in terms of strong association rules (the prior algorithm).

In this application we performed several scenarios to extract and analyze the strong association rules

The extraction of strong association rules is done automatically after the call of the algorithm on the selected database, these strong association rules we obtain after having executed several scenarios before each extraction of a scenario you have to choose the number of roles, the value of the minimum threshold of support and trust.

Finally, the results of these different scenarios will be presented and discussed.

**Keywords:** Knowledge extraction from data, Educational data mining, Aprior algorithm, Support and trust, Data mining, Association rules.

**ملخص**

التنقيب عن البيانات التعليمية هو عملية مصممة لتحليل البيانات من المواقف التعليمية لفهم الطلاب بشكل أفضل والمواقف التي يتعلمون فيها. من بين الأعمال البحثية في مجال التنقيب عن البيانات ، يعد استخراج قواعد الارتباط المهمة الرئيسية التي جذبت أكبر قدر من اهتمام الباحثين والتي تم القيام بالكثير من العمل من أجلها ، فهي تتيح اكتشاف قواعد واضحة و قابلة للاستغلال في مجموعة بيانات كبيرة ، قواعد تعبر عن الارتباطات بين العناصر أو السمات في قاعدة بيانات. تتناول هذه الأطروحة قواعد التأسيس يرتبط عملنا بهذا الجانب. قمنا أولاً بجمع بيانات من 250 طالبًا من خلال استبيان يتكون من 56 سؤالًا ، ويستخلص تطبيقنا قواعد الارتباط بين سمات طلاب السنة الأولى في قسم العلوم والتكنولوجيا باستخدام خوارزمية مسبقة. بعد ذلك ، سوف نقدم نتائج التنقيب عن البيانات من حيث قواعد الارتباط القوية (الخوارزمية السابقة.

في هذا التطبيق ، أجرينا عدة سيناريوهات لاستخراج قواعد الارتباط القوية وتحليله.

يتم استخراج قواعد الارتباط القوية تلقائيًا بعد استدعاء الخوارزمية في قاعدة البيانات المحددة ، قواعد الارتباط القوية هذه التي نحصل عليها بعد تشغيل عدة سيناريوهات قبل كل استخراج للسيناريو عليك أن تختار عدد الأدوار وقيمة الحد الأدنى للدعم والثقة.

أخيرًا ، سيتم عرض ومناقشة نتائج هذه السيناريوهات المختل.

**الكلمات المفتاحية:** استخراج المعرفة من البيانات ، التنقيب عن البيانات التعليمية ، الخوارزمية السابقة ، الدعم والثقة ، التنقيب في البيانات ، قاعدة الارتباط.

# Table des matières

## Chapitre II : LES REGLES D'ASSOCIATION



## Chapitre III : POPULATION DE L'ETUDE

**Chapitre IV : IMPLEMENTION ET RELISATION**

# Liste des tableaux

l'université et qui détestent l’université.

# Liste des figures

# *Introduction générale*

# Introduction générale :

## Contexte du travail

Durant ces dernières années, avec l'augmentation sans cesse de capacité de stockage des ordinateurs, les quantités de données collectées, dans divers domaines d'application de l'informatique, deviennent de plus en plus importantes, souvent, ces masses de données contiennent des informations cachées qui peuvent être pertinentes. En effet, Il est chaque jour plus facile de collecter des données mais notre capacité à en extraire des connaissances reste limitée. Cela suscite le besoin d'analyser et d'interpréter ces données afin d'en extraire des connaissances utiles. Pour répondre à ces opportunités, l'extraction de connaissances dans les bases de données (ECBD ou « Knowledge Discovery in Data bases ») est le domaine de recherche au sein duquel coopèrent les statisticiens, les spécialistes en bases de données et en intelligence artificielle, ou encore chercheurs en conception d'interfaces homme-machine. Ce domaine connaît une croissance spectaculaire, il a été défini dès 1991 comme le processus non trivial d'extraction d'informations valides, nouvelles, potentiellement utiles, et compréhensibles à partir de données [1].

Parmi les travaux de recherche en fouille de données, l'extraction des règles d'association, elle représente la tâche principale qui a attiré le plus d'attention des chercheurs et pour laquelle beaucoup de travaux ont été effectués, elle permet la découverte de règles intelligibles et exploitables dans un ensemble de données volumineux. Les règles exprimant des associations entre items ou attributs dans une base de données. Ce travail de mémoire traite le problème de règles d'association à partir de données éducatives et plus particulièrement les données des étudiants universitaire en première années afin d'aider les nouveaux étudiants à l'université.

## Objectifs

L'objectif de notre travail est d'extraire les règles d'association fortes qui se fait automatiquement après l'appel de l'algorithme Apriori sur la base de données sélectionnée qui contient des données pour les étudiants de première année Mathématiques et informatique et affiche systématiquement les résultats finaux (Règles Fortes) avec la confiance de chaque règle d'association.

## Organisation du mémoire

A part l'introduction le mémoire se reparti en quatre chapitres :

**Chapitre I : « les fouilles de donnée éducative »**

Ce chapitre présente les fouilles de donnée éducative (éducationnel data mining), processus ECD et les meilleurs outils pour faire de la fouille de données.

**Chapitre II : « les règles d'association »**

Ce chapitre est consacré à la présentation sur les règles d'association et les paramètres de l'algorithme apriori

**Chapitre III : «la population de l'étude »**

Ce chapitre présente la population de l'étude et les données collectées utilisant un questionnaire composé de 55 questions destiné aux étudiants de la première année mathématique et Informatique.

**Chapitre IV : « implémentation et réalisation »**

Le dernier chapitre présente notre méthodologie pour l'extraction de règles d'association et les analyses de résultats.

# *Chapitre I:*

# *Le fouille des données éducative*

# 1. Introduction

Récemment, nos possibilités de produire et de recueillir des données avaient augmenté rapidement. Des millions de bases de données ont été employés dans la gestion d'entreprise, l'administration de gouvernement, la gestion des données scientifique, et beaucoup d'autres applications, le nombre de telles bases de données ne cessent d'augmenter jour après jour.

Cette croissance explosive des données et des bases de données a produit des besoins urgents pour les nouvelles techniques et les outils qui peuvent intelligemment et automatiquement transformer les données traitées en informations utiles et connaissances. [2]

Le Data Mining (Fouille de Données) est un domaine qui consiste à comprendre les données, généralement par le moyen de méthodes statistiques. En d'autres termes, le data mining cherche à identifier des connaissances à partir des bases de données.

Comme ce processus peut être très difficile, il est souvent comparé au minage de l'or dans les rivières: le gravier des alluvions représente l'énorme quantité de données et les pépites d'or représentent la connaissance cachées que l'on veut trouver.

# 2. Le Data Mining

## 2.1.Définition

Le Data Mining ou Knowledge Discovery in Data (KDD) est une technique d'analyse de modèle de données qui agrège de grands ensembles de données, permettant à l'utilisateur de prévoir les tendances futures en extrayant des données utilisables à partir d'un plus grand ensemble de données brutes, ce qui les aide à se rapprocher de son objectif et faire en sorte qu'ils prennent de meilleures décisions. C'est une nouvelle discipline à l'interface de la statistique et des technologies de l'information : bases de données, intelligence artificielle, apprentissage automatique[12].

Le Data Mining est : « le cœur du processus d'ECD, c'est le module dont le rôle est de fouiller dans les grandes masses de données pour extraire les connaissances cachées, son importance est due à la disponibilité d'énormes quantités de données dans un état de croissance permanente ». [3]

## 2.2. Les étapes du processus ECD

ECD est un processus anthropocentré, les connaissances extraites doivent être les plus intelligibles possibles à l'usager. Elles doivent être validées, mises en forme et agencées. Nous allons détailler toutes ces notions et les situer dans le processus général de ECD, se processus et schématisé dans la figure 1.1.

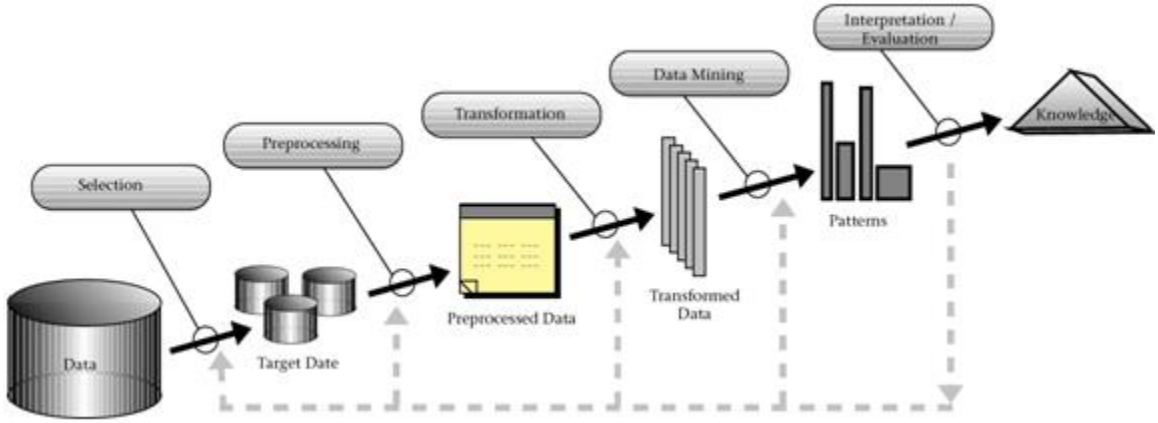


**Figure 1.1:** Processus ECD. [4]

Le processus ECD se décompose en plusieurs étapes, les différentes phases sont les suivant :

***Sélection***

Les données nécessaires au processus d'exploration de données peuvent être obtenues à partir de nombreuses sources de données différentes et hétérogènes. Cette première étape permet d'obtenir les données de diverses bases de données, fichiers et sources non électroniques. **[4]**

Cette phase ne se limite pas à la seule sélection des données qui vont être exploitées par le système ECD. Elle comprend également l'analyse du problème à résoudre**,** ce qui permet d'en déduire le ou les types de données qui sont exploitées, ainsi que les méthodes qui pourraient être utilisées pour résoudre ce problème. **[5]**

Un système ECD idéal est un système qui nécessite l'intervention d'aucune entité, c'est-à-dire un système automatisé qui va extraire de nouvelles connaissances à partir de grandes bases de données mises à sa disposition sans l'intervention de l'utilisateur. Actuellement, ce type de système présente de nombreux inconvénients. Le premier de ceux-ci est la perte de temps et de ressources nécessaires à l'exploitation de l'ensemble des données disponible au système**.[5]**

Parmi les conséquences de ces inconvénients, on peut citer :

– Les recherches lancées par le système peuvent toucher divers domaines ou thèmes qui n'ont

aucun rapport avec l'objectif défini par l'utilisateur.

– Le système peut fournir des connaissances qui ne présentent aucun intérêt ou sont incompréhensibles pour l'utilisateur

– L'utilisateur submergé de nouvelles connaissances, ne peut distinguer des connaissances proposées celles qui lui sont réellement intéressantes.

Ceci implique que l'utilisateur doit avoir la possibilité de communiquer avec le système afin d'orienter la recherche selon ses objectifs. Pour faciliter la communication de l'utilisateur avec le système ECD, un ensemble de primitives (data mining primitives) a été conçu. Ces primitives incluent :

- ✓ Spécification des données
- ✓ Spécification du type de connaissances à extraire
- ✓ Spécification des connaissances préalables
- ✓ Spécification de la mesure
- ✓ Représentation de la connaissance extraite

***Prétraitement***

Les données à utiliser par le processus peuvent contenir des données incorrectes ou manquantes. Il peut y avoir des données anormales provenant de plusieurs sources impliquant différents types de données et mesures. Il peut y avoir aussi beaucoup d'activités différentes effectuées à ce moment. Les données erronées peuvent être corrigées ou supprimées, tandis que les données manquantes doivent être fournies ou prévues (souvent à l'aide d'outils d'exploration de données). **[4]**

Les données à analyser par les méthodes de data mining sont parfois incomplètes, inconsistantes, erronées, incompatibles entre elles, inadaptées ou encombrantes. Ces types de données sont courants et se retrouvent régulièrement dans les bases de données et d'entrepôts de données. **[5]**

Dans cette phase, plusieurs procédures sont nécessaires et chacune d'entre-elles a des tâches bien précises dans le traitement des données.

La procédure de nettoyage des données : elle se compose des tâches de traitement des données manquantes, de traitement des données erronées et inconsistantes.

– Tâche de traitement des données manquantes : Plusieurs méthodes permettent d'accomplir

cette tâche. Le choix de la méthode dépend des données et de l'objectif de l'étude.

1. méthode consiste à ignorer les instances incomplètes
2. méthode consiste à compléter les données manuellement
3. méthodes qui consistent à compléter les données incomplètes à l'aide de constantes globales
4. méthodes qui consistent à remplacer la valeur manquante d'un attribut par la valeur moyenne de cet attribut
5. méthodes qui remplacent la valeur manquante par la valeur la plus probable

– Tâche de traitement des données de type bruit

1. méthode de groupement
2. méthode combinant une solution algorithmique à l'utilisation d'un expert
3. méthode de régression.

***Transformation***

Les données provenant de différentes sources doivent être converties en un format commun pour le traitement. Certaines données peuvent être encodées ou transformées en formats plus utilisables. La réduction des données peut être utilisée pour réduire le nombre de valeurs de données possibles considérées. **[4]**

Permet de modeler les données sous une forme exploitable par les méthodes de data mining.

1. méthode d'agrégation
2. méthode de généralisation des données
3. méthode de normalisation
4. méthode d'ajout d'attributs

La procédure de réduction des données permet de réduire la taille des données tout en gardant leur intégrité.

Les méthodes de réduction les plus connues sont :

1. Agrégation des données cibles
2. Réduction dimensionnelle
3. Compression des données
4. Discrétisation et génération de concept hiérarchique

***Data Mining (Exploration de données)***

En fonction de la tâche d'exploration de données en cours d'exécution, cette étape applique des algorithmes aux données transformées pour générer les résultats souhaités. **[5]**

C'est le cœur du processus d'ECD. Il s'agit à ce niveau de trouver des connaissances à partir des données. Tout le travail consiste à appliquer des méthodes intelligentes dans le but d'extraire cette connaissance. Il est possible de définir la qualité d'un modèle en fonction de critères comme les performances obtenus, la fiabilité, la compréhensibilité, la rapidité de construction et d'utilisation et enfin l'évolutivité.

Tout le problème du data mining réside dans le choix de la méthode adéquate à un problème donné. Il est possible de combiner plusieurs méthodes pour essayer d'obtenir une solution optimale globale.

Les méthodes de fouille de donnée qui sont les plus couramment utilisées dans les systèmes ECD sont les méthodes de type classification, régression, structuration et association

– Méthodes de classification et de structuration (algorithme des k-moyennes (k-means), algorithme du plus proche voisin),

– Méthodes d'explication et de prédiction (arbre de décision, réseaux de neurones, réseaux bayésiens, règles d'associations),

– Méthodes de visualisation et de description.

***Interprétation / évaluation***

La manière dont les résultats de l'exploration de données sont présentés aux utilisateurs est extrêmement importante, car leur utilité en dépend. Diverses stratégies de visualisation et d'interface graphique sont utilisées à cette dernière étape. **[4]**

## 3. Data Mining dans l'éducation :

### 3.1. Définition d'EDM :

« Educationnel Data Mining est une discipline émergente qui se préoccupe de développer des méthodes pour explorer les données uniques et de plus en plus volumineuses obtenues à partir de contextes éducatifs. Elle utilise ces méthodes pour mieux comprendre les étudiants et les contextes dans lesquels ils apprennent. [Society international de données pour éducation, 2011] »

Cette définition de l'EDM est proposée par l'International Educationnel Data Mining Society, qui organise la Conférence internationale sur l'exploration de données éducatives et publie le Journal of Educationnel Data Mining.

## 3.2. Domaines d'application de l'EDM :

Cristobal Romero et Sébastian Ventura ont noté les domaines d'application de l'EDM comme suit : [6]

- ✓ Analyse et visualisation de données
- ✓ Donner de la rétroaction aux instructeurs auxiliaires
- ✓ Recommandations pour les étudiants
- ✓ Prédire le rendement des élèves
- ✓ Modélisation d'étudiant
- ✓ Détecter les comportements indésirables des étudiants
- ✓ Regrouper des étudiants
- ✓ Analyse de réseau social
- ✓ Développer des cartes conceptuelles
- ✓ Construire un didacticiel
- ✓ Planification et ordonnancement

## 3.3. Composantes d'EDM :

La figure suivante présente les trois composantes principales d'EDM : L'informatique, éducation et statistique. Ces composantes ont intersectées et formes d'autres domaines liées à EDM, tels que les systèmes d'apprentissage automatique, les systèmes d'analyses d'apprentissage et les systèmes éducatifs basé sur l'ordinateur.

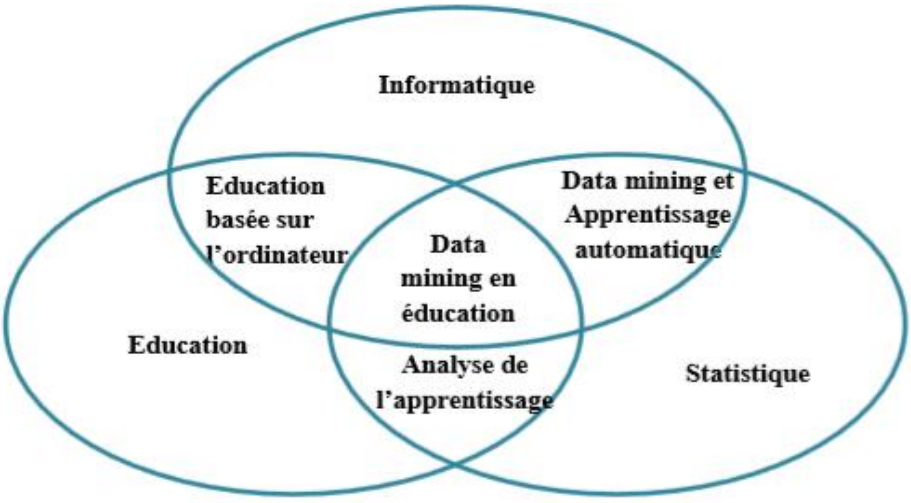


**Figure 1.2:** Educational Data mining (Romero et Ventura, 2013)

## 4. Techniques et algorithmes communs en fouille de données en éducation :

Il existe différents types d'algorithmes pour la fouille de données dans un contexte d'éducation. Ces types varient selon les objectifs de:

- Prédiction :

  Cela concerne par exemple les techniques de classification, régression, ou l'estimation latente de connaissances « Latent Knowledge Estimation »

- Description par la découverte de structures :

  Avec le clustering, l'analyse en facteurs, la découverte de structure de domaine ou de modèles, ou encore avec l'analyse de réseaux.

- Fouille de relations ou de dépendances (qui peut être exploitée dans la description ou dans la prédiction) :

  En utilisant entre autres la fouille de règles d'association, la fouille de corrélations, la fouille de motifs dans les séquences « Sequential Pattern Mining » et la fouille de données causales (causal data mining).

Ces différentes techniques ont pour point commun la distillation de données utiles ou interprétables (selon des degrés variables) pour un jugement humain.

### 4.1. Apprentissage supervisé :

L'apprentissage supervisé se réfère à la capacité d'apprendre à réaliser des tâches à partir de données déjà étiquetées. D'une certaine manière, le but est de refaire la même tâche à partir de données déjà existantes.

Cet apprentissage peut concerner, selon ce qu'on cherche à déterminer :

- soit les méthodes de classification (partant de l'entrée on souhaite déterminer la sortie)
- soit les régressions (partant de la sortie, l'entrée est à déterminer).

Nous décrirons comme exemples les arbres de décision et les réseaux de neurones.

#### *a. Les arbres de décision :*

Un arbre de décision est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteintes en fonction de décisions prises à chaque étape. [6]

Parmi les applications des arbres de décision en éducation nous pouvons citer celle où l'analyse permet de :

- Recommander une séquence optimale d'apprentissage pour les apprenants (et donc de faire de l'adaptation),
- Prédire les performances ou la satisfaction d'un apprenant,ou encore détecter le modèle mental d'un apprenant dans un ITS58

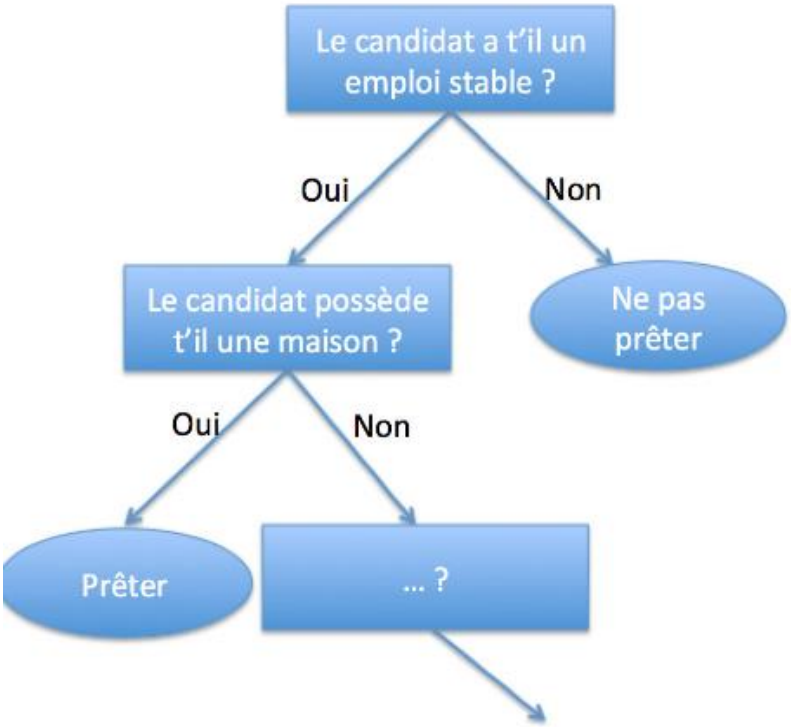


**Figure 1.3:** Accorder ou non un prêt bancaire. Chaque individu est évalué sur un ensemble de variables testées dans les nœuds internes. Les décisions sont prises dans les feuilles.[7]

***a. Les réseaux de neurones***

Un réseau de neurones artificiels est un système dont la conception est à l'origine schématiquement inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques [7].

Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage de type probabiliste, en particulier bayésien. Ils sont placés d'une part dans la famille des applications statistiques, qu'ils enrichissent avec un ensemble de paradigmes permettant de créer des classifications rapides (réseaux de Kohonen en particulier), et d'autre part dans la famille des méthodes de l'intelligence artificielle auxquelles ils fournisse un

mécanisme perceptif indépendant des idées propres de l'implémenter, et fournissant des informations d'entre au raisonnement logique formel. [8]

## 4.2. Apprentissage non supervisé

À l'inverse de l'apprentissage supervisé, dans l'apprentissage non-supervisé, il n'existe pas d'attribut particulier cible, on dit alors que les données ne sont pas étiquetées. Nous illustrons cette approche d'apprentissage machine par le clustering, les règles associatives et le « sequence mining ».[9]

***a. Clustering***

Le « clustering non supervisé » aussi appelé classification non supervisée, est un processus qui permet de rassembler des données similaires. Le fait qu'il ne soit pas supervisé signifie que des techniques d'apprentissage machine vont permettre de trouver certaines similarités pour pouvoir classer les données et ce de manière plus ou moins autonome.

Ce type d'analyse permet d'avoir un profil des différents groupes. Cela permet donc de simplifier l'analyse des données en faisant ressortir les points commun et les différences et en réduisant ainsi le nombre de variable des données. Cette technique n'est pas seulement utilisée dans le domaine génétique, mais permet aussi par exemple de lister de potentiels clients lors d'une action publicitaire.[10]

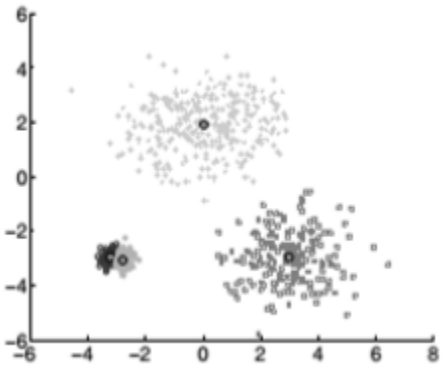


**Figure 1.4:** Résultat de la convergence du clustering

***a. Sequence mining***

Le sequence mining ou fouille de séquences examine la récurrence de certaines occurrences ou motifs, selon un ordre lié à la séquence. Des techniques de fouille de texte (Text-mining) peuvent également être intégrées pour attribuer une syntaxe particulière. L'intérêt en e-

learning réside dans le fait de pouvoir recommander certaines ressources ou parcours en ne connaissant qu'une partie de la séquence. Cela sous-entend bien évidemment le fait d'avoir identifié les séquences les plus fréquentes. Un des algorithmes les plus utilisés est le GSP, la variable de temps peut optionnellement être incluse. Free-span, Prefix-Span sont d'autres algorithmes plus rapides mais se basant sur la même idée que GSP. [9]

# 5. Meilleurs outils pour faire de la fouille de données

## 5.1 Python

Python est un langage de programmation très puissant utilisé en Data Mining pour faire de l'analyse statistique, la classification, le clustering et l'analyse prédictive.

## 5.2 Le langage R

R est un langage de programmation et un logiciel libre destiné aux statistiques et à la science des données soutenu par la R Foundation forStatisticalComputing. Il permet de faire l'analyse statistique, la classification, le clustering et l'analyse prédictive

## 5.3 Tanagra

Tanagra est un logiciel gratuit de Data Mining destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données. C'est un projet ouvert au sens qu'il est possible à tout chercheur d'accéder au code et d'ajouter ses propres algorithmes pour peu qu'il respecte la licence de distribution du logiciel.

## 5.4. RapidMiner

C'est outil Open source à la fois gratuit et commercial. RapidMiner est une plate-forme logicielle de science des données développée par la société du même nom qui fournit un environnement intégré pour la préparation des données, l'apprentissage automatique, l'apprentissage en profondeur, l'exploration de texte et l'analyse prédictive.[10]

## 5.5. WEKA

Weka est une suite populaire de logiciels d'apprentissage automatique. Écrite en Java, développée à l'université de Waikato, Nouvelle-Zélande. Weka est un Logiciel libre disponible sous la Licence publique générale GNU. Il permet de faire l'analyse statistique, la classification, le clustering et l'analyse prédictive.[11]

## 6. Conclusion

Le Data Mining est utilisé par les data scientistes pour tirer une connaissance ou des informations cachées dans des grands volume de données afin de permettre une meilleure prise de décision par les gestionnaires. Les outils comme Python, le Langage R ,Tanagra RapidMiner et WEKA sont mieux utilisés.

# *Chapitre II:*

# *Les règles d'association*

## 1. Introduction

Dans le domaine du data mining la recherche des règles d'association est une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations entre deux ou plusieurs variables stockées dans de très importantes bases de données, ayant un intérêt pour le décideur. Piatetsky-Shapiro [13] présentent des règles d'association extrêmement fortes découvertes dans des bases de données en utilisant différentes mesures d'intérêt. En se basant sur le concept de relations fortes, Rakesh Agrawal et son équipe [14] présente des règles d'association dont le but est de découvrir des similitudes entre des produits dans des données saisies sur une grande échelle dans les systèmes informatiques des points de ventes des chaînes de supermarchés.

Par exemple, une règle découverte dans les données de ventes dans un supermarché pourrait indiquer qu'un client achetant des oignons et des pommes de terre simultanément, serait susceptible d'acheter un hamburger. Une telle information peut être utilisée comme base pour prendre des décisions marketing telles que par exemple des promotions ou des emplacements bien choisis pour les produits associés. En plus des exemples ci-dessus concernant le panier de la ménagère, les règles d'association sont employées aujourd'hui dans plusieurs domaines incluant celui de la fouille du web, de la détection d'intrusion et de la bio-informatique.

## 2. Domaines d'application

Etant un outil efficace de fouille de données, la recherche des règles d'associations est appliquée dans tous les domaines du Data Mining. Vu ses avantages offerts, cette technique est devenue un sujet attractif et actif appliqué à un large champ d'applications dans divers domaines. Nous citons ici une liste non exhaustive des applications dont les résultats ont pu être améliorés par l'analyse des règles d'association extraites.

–**Marketing et Planification commerciale** : placement des articles achetés fréquemment ensemble (étagère ou une page de catalogue), organisation des catalogues, choix des articles en promotion, …etc.

–**Réseaux de télécommunication :** filtrage des alarmes non informatives, identification des causes d'anomalies, prédiction des anomalies, …etc.

–**Recherche médicale :** aide au diagnostic et définition de traitement, identification de population à risque vis-à-vis de certaines maladies, prédiction de résultats d'analyses par combinaison de caractéristiques des patients et de résultats d'autres analyses.

–**Analyse de données spatiales :** détection des relations entre caractéristiques des

données, prédiction d'évènements, etc.

–**Internet :** Amélioration des modes d'accès aux informations, Modification de la structure des pages et des liens, Personnalisation des pages suivant le profil utilisateur, l'intelligence économique dans les sites E-commerce, suggestion aux clients (comme c'est la cas du site Amazon.com), etc.

- **Le domaine industriel :** prévision des ventes, surveillance des unités de production, diagnostic et analyse des pannes, contrôle de qualité, etc. Multimédia: analyse d'imagerie, prévision météorologique, aide aux enquêtes, etc. [14]

## 3. Notions et définitions sur les règles d'association

Dans la section suivante nous allons détailler plusieurs concepts impliqués dans la recherche et l'extraction des règles d'association à savoir: transaction, item, règle d'association

### 3.1. Transaction et ensemble d'items :

Soient T= {I1 I2, I3, •• • , Im} l'ensemble d'attributs binaires appelés items. Soit T une base de données des transactions. Chaque transaction t est représentée comme un vecteur binaire, avec t [k] = 1 si la transaction t achète l'item Ik sinon t[k] = 0 Chaque transaction est définie par un seul numéro dans la base des transactions

### Exemple :

Soient T= {Pain, Lait, Couches, Bière, Œufs, Coca}, l'ensemble de tous les items des paniers et D = {1,2,3,4,S} l'ensemble de toutes les transactions. Le tableau .21

Pourrait se mettre sous forme binaire comme suit:

| TID | Pain | Lait | Couches | Biére | Œufs | Coca |
|---|---|---|---|---|---|---|
| **1** | **1** | **1** | **0** | **0** | **0** | **0** |
| **2** | **1** | **0** | **1** | **1** | **1** | **0** |
| **3** | **0** | **1** | **1** | **1** | **1** | **0** |
| **4** | **1** | **1** | **1** | **1** | **1** | **1** |
| **5** | **0** | **0** | **0** | **0** | **1** | **1** |
| **6** | **1** | **1** | **1** | **0** | **0** | **0** |

**Tableau 2.1** : tableau des transactions présentés en binaire

### 3.2. Item

Un item est tout article, attribut, littéral appartenant à un ensemble fini d'éléments distincts X = {x1, x2, ..., x3}. Par exemple, dans les applications de type analyse du panier de la ménagère, les articles en vente dans un magasin sont des items. L'ensemble X peut contenir les items A, B, C et D correspondant aux articles lait, beurre, pain et confiture par exemple.

| ID ticket | pain | Lait | Jus | beurre | oeufs | Cola |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 1 | 0 | 0 | 0 |

**Tableau 2.2 :** tableau des Item

### 3.3. Item Set :

Un itemset ou motif est tout sous-ensemble d'items de X. Un itemset constitué de k-items sera appelé un k-itemset. Par exemple, l'itemset {A, B, C} est un 3-itemset noté ABC.

| ID ticket | Pain | Lait | Jus | beurre | oeufs | Cola |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 1 | 0 | 0 | 0 |

**Tableau 2.3 :** tableau des Itemset

### 3.4. Item set fréquent :

Un Item set est fréquent si et seulement si son support est supérieur à un support minimum.

### 3.5. Transaction $T_j$ :

Est un item set auquel est associé un identificateur unique.

| ID ticket | pain | Lait | Jus | Beurre | oeufs | Cola |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 1 | 0 | 0 | 0 |

**Tableau 2.4 :** Tableau de transaction

L'ensemble des transactions sont stockés dans une base de transactions T.

| ID ticket | Pain | Lait | Jus | beurre | oeufs | Cola |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 1 | 0 | 0 | 0 |

**Tableau 2.5 :** Base de Transaction

Une règle associative ou règle d'association R, est définie comme une implication de la forme:

R: I1 → I2, tel que I1 ⊂ S, I2 ⊂ S et I1 ⋂ I2 = Ø

I1 est appelé **Condition** et I2 **Conclusion**

## 4. Critères d'évaluation des règles d'association

Pour garder les bonnes associations, on utilise deux critères d'évaluation, chaque règle est évaluée par deux mesures (facteurs) : le support et la confiance.

### 4.1. Le support

Est une mesure d'importance statistique (statistical significance). Pour qu'une règle x → y vérifie un facteur de support S si et seulement si au moins S % des transactions dans la base de données x et y vérifie

Support : probabilité d'acheter le produit X et le produit Y

$$\frac{\text{Nombre de transaction contenant les produit X et Y}}{\text{Nombre total de transaction}}$$

Le support permet de mesurer la fréquence de l'association .[15]

## 4.2. La confiance

Est une mesure de la force de la règle (strength of the rule). Pour qu'une règle x →y vérifie un facteur de confiance C si au moins C % des transactions dans la base de données qui vérifient x vérifie aussi y.

Par exemple, dans une base de données d'enseignement , si on considère la règle suivante : « 75 % des étudiants qui suivent le cours "Linux/ Unix ", suivent également le cours de "Programmation C ", et 30 % de tous les étudiants ont en fait suivis les deux cours ». On peut dire que cette règle est vérifiée avec une certitude supérieure à 75% (confiance de la règle), et que la règle est supportée par au moins 30% des étudiants (support de la règle). Pour être acceptable, il faut que le support de cette règle (30%) soit supérieur à une autre valeur définie à l'avance par l'utilisateur (support minimum), et que la confiance de cette règle (75%) soit supérieure à une autre valeur définie à l'avance (confiance minimale)

**Confiance :** probabilité d'acheter le produit y étant donné que le produit X a été acheté (X=>Y)

$$\frac{\text{Nombre de transaction contenant les produit X et Y}}{\text{Nombre de transaction contenant le produit X}}$$

**La confiance** permet de mesurer la force de l'association [15]

## 5. Processus d'extraction de règles d'association :

Le processus d'extraction de règles d'association est constitué de plusieurs phases allant de la sélection et la préparation des données jusqu'à l'interprétation des résultats, en passant par la phase de recherche des connaissances (extraction des ensembles fréquents d'attributs et génération des règles d'association). Ci-dessous une description de différentes phases de ce processus.

## 5.1. Sélection et préparation des données (nettoyage)

Cette phase consiste à sélectionner les données (attributs et objets) de la base de données utiles à l'extraction des règles d'association et transformer ces données en un contexte d'extraction. L'extraction de règles d'association peut être effectuée à partir des bases de données de divers types, comme des données spatiales, temporelles, orientées objets, multimédia, etc. Cette première phase est très importante car à partir de la qualité des données en entrées dépend la qualité des résultats.

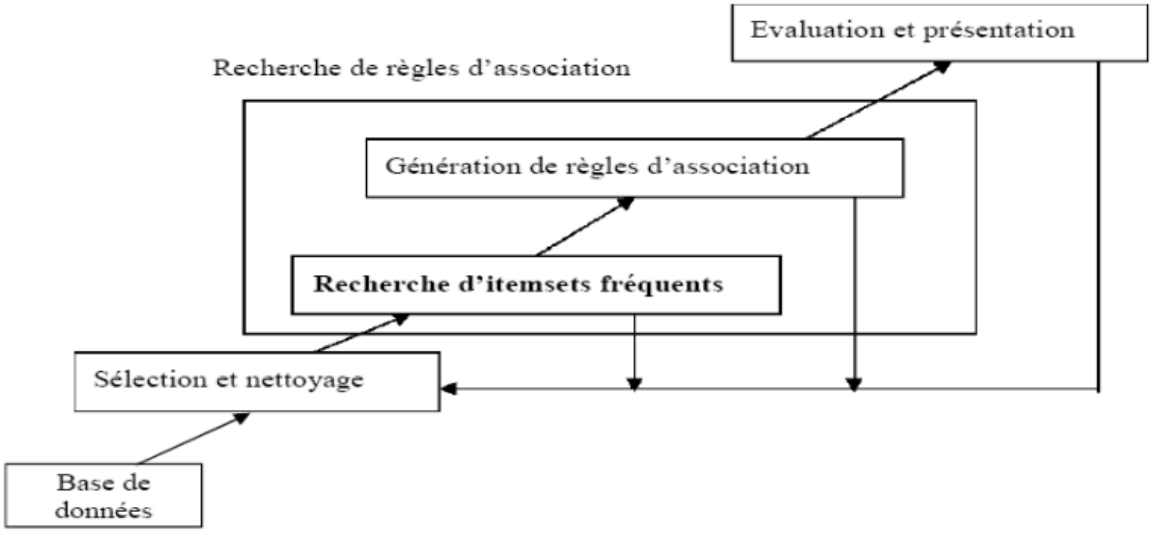


**Figure 2.1:** Processus d'ECD adapté à la recherche de règles d'association

Cette phase est nécessaire pour pouvoir appliquer les algorithmes d'extraction des règles sur des données de natures différentes provenant de sources différentes, de concentrer la recherche sur les données utiles pour l'application et de minimiser le temps d'extraction. A noter que le problème des données incomplètes (valeurs manquantes), et les données erronées ou incertaines et la taille du jeu de données doivent être pris en considération dans cette phase. Par exemple, le tableau ci-dessous 2.5 suivant représente un contexte d'extraction D constitué de 6 objets, chacun représenté par son identifiant et de quatre items. Ce contexte sera utilisé comme exemple dans tout le reste de ce chapitre.

| Item id | Objet |
|---|---|
| 1 | A C D |
| 2 | B C E |
| 3 | A C B E |
| 4 | B E |
| 5 | A B C E |
| 6 | B C E |

**Tableau 2.6 :** Contexte d'extraction de règles d'association D.

## 5.2. Recherche d'item sets fréquents

Cette phase consiste à extraire du contexte D tous les itemsets qui sont fréquents. La recherche des itemsets fréquents est un problème non trivial car le nombre d'itemsets fréquents potentiels est exponentiel en fonction du nombre d'items du contexte D. Dans le cas d'un ensemble d'items I de taille m, le nombre d'itemsets potentiels est de 2m-1. Ces items

forment le treillis des itemsets de I, dont la hauteur est de m+1. Les balayages du contexte doivent être réalisés lors de cette phase et il est donc nécessaire de développer des méthodes efficaces d'exploration de cet espace de recherche exponentiel. La phase découverte des items fréquents constitue la phase la plus coûteuse en temps d'exécution et en espace. L'espace de recherche est de taille exponentielle par rapport au nombre d'items. Plusieurs méthodes ont été proposées dans le but de réduire l'espace de recherche de cette phase ainsi que le nombre de balayages du contexte réalisé. Voici un exemple d'un treillis des itemsets du contexte D donné dans le tableau 2.6 précédent

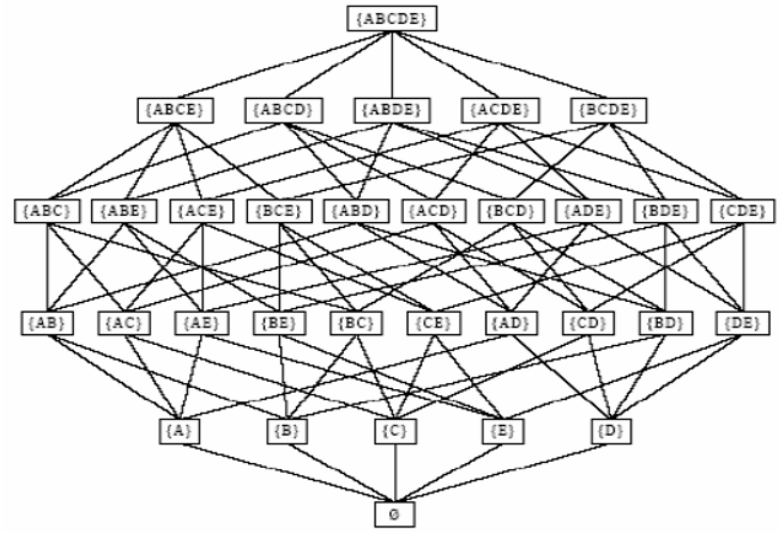


**Figure 2.2 :** Représentation sous forme de treillis d'itemsets fréquents du contexte D

## 5.3.Génération des règles d'association

Dans le domaine de data-mining, La génération des règles d'association est la méthode d'extraction des informations insignifiant et des connaissances utiles qui sont cachés dans une base de données volumineuse. Dans ce contexte, un nombre d'algorithmes s'inspirant de deux approches pour la génération des règles d'association :

- Des algorithmes, s'inspirant de l'approche classique, basés sur l'extraction des itemsets fréquents puis la génération des règles d'association. La particularité de ces algorithmes est qu'ils génèrent un nombre énorme de règles rendant leur exploitation quasiment impossible.
- Algorithme utilisent une nouvelle approche qui est basé sur la génération des règles d'association avec l'expansion des règles.

## 5.4. Visualisation et interprétation

C'est la phase finale du processus d'ECD. Cette phase consiste en la visualisation par l'utilisateur des règles d'association extraites du contexte et leur interprétation afin d'en déduire des connaissances utiles pour l'amélioration de l'activité concernée. Ainsi l'expert du domaine peut juger de leurs pertinences et utilités. Mais le nombre important des règles d'association extraites impose le développement d'outils de classification de règles selon leurs propriétés, de sélection de sous-ensembles de règles selon des critères définis par l'utilisateur, et de visualisation de ces règles sous une forme intelligible. Cette nouvelle problématique est également appelée « Knowledge Mining ».

La forme de présentation de règles peut être textuelle, graphique ou bien une combinaison de ces deux formes intelligibles. Ceci va donner naissance à un nouveau domaine de recherche : la fouille visuelle de données « Visual Data Mining » afin d'améliorer le processus d'extraction de connaissances en proposant des outils de visualisation adaptés à différentes problématiques.

Les connaissances de l'utilisateur concernant le domaine d'application sont nécessaires lors des phases de pré-traitement afin d'assister la sélection et la préparation des données et de post-traitement, pour l'interprétation et l'évaluation des règles extraites. En fonction de l'évaluation des règles extraites, les paramètres utilisés lors des précédentes phases (critères de sélection et préparation des données et seuils minimaux de support et de confiance) peuvent être modifiés avant d'effectuer à nouveau l'extraction des règles d'association, ceci afin d'améliorer la qualité du résultat.

Il ressort de la grande majorité de ces applications qu'au final, beaucoup de règles sont générées par les algorithmes et qu'il est parfois difficile aux experts du domaine de les exploiter dans leur intégralité, car cela engendre un travail cognitif très important. Devant cette tâche, leur premier souhait est souvent de réduire cet ensemble pour ainsi diminuer le temps d'expertise correspondant. En effet, dans le domaine industriel, les experts n'ont pas forcément beaucoup de temps à consacrer à l'analyse des résultats. Dans le chapitre suivant, nous allons décrire cette problématique, et passer en revue les propositions faites dans ce sens.

# 6. Algorithmes de recherche de règles d'association

## 6.1. L'algorithme Apriori :

Proposé par Agrawal et Srikant en1994, l'algorithme Apriori représente la base de tous les algorithmes de recherche des règles d'association. Il extrait les Itemsets fréquents pour les règles d'association [16].

### 6.1.1. Le principe de l'algorithme Apriori :

L'algorithme Apriori utilise une approche itérative, où k - Itemsets sont employés pour explorer les (k + 1) - Itemsets. D'abord, les 1 - Itemsets sont trouvés par balayage de la base de données pour calculer le support de chaque item, et la collecte de ces Itemsets qui ont un support ~ minsup .

L'ensemble résultant est noté L1 puis utilisé pour trouver L2 , les 2-itemsets, qui est utilisé pour trouver L3 , et ainsi de suite jusqu'à ce qu'aucun k-Itemsets puisse être trouvé. L'obtention de chaque Lk nécessite une analyse complète de la base de données.

### 6.1.2. L'algorithme Apriori

REQUIRE : Un support seuil S

ENSURE : La liste des itemsets fréquents

L1 ‹— Liste des items dont le support est supérieur à S; i ‹— i+1;

REPEAT

```
i ‹— 1;
  À partir des Li-1, construire l'ensemble Ci des itemsets fréquents, candidats
comprenant i items ;
  Li ‹— {};
  POUR tout élément e € Ci FAIRE
     SI support(e) > S ALORS
       ajouter e à Li;
     FINSI
  FINPOUR
```

UNTIL L1 == {}

**Exemple :** Recherche des itemsets fréquents

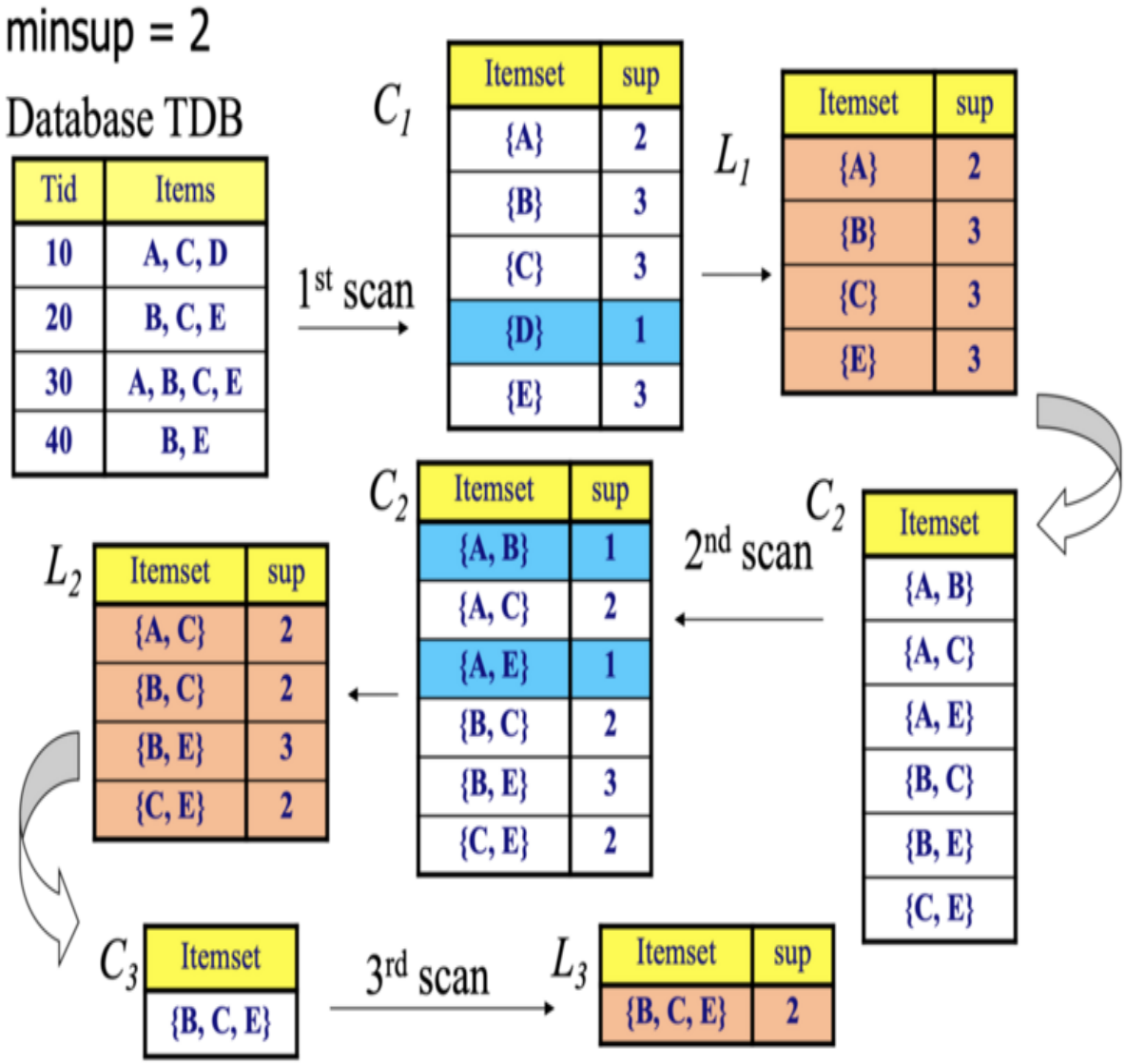


**Figure2.3 :** recherche des itemsets fréquents

### 6.1.3. Générer les règles d'association à partir d'ltemsets fréquents :

Une règle d'association forte satisfait à la fois le minsup et le minconf. Les règles d'association peuvent être générées comme suit [17] :

1. Pour chaque itemset fréquent, générer tous les sous-ensembles non vides de
2. Pour tout sous-ensemble non vide s de l, la règle "S->(l--S) est générée si le support de Cl - s) divisé par le support de s est supérieur ou égale à minsup. Où, (l--S)est l'ensemble des éléments qui appartiennent à l mais pas a s

## Exemple :

Considérons les données de tableau 4.2 Supposons que les données contiennent l'itemset fréquent l = {Pain, Couches, Bière}. Les sous-ensembles non vides de l: {Pain, Couches}, {Pain, Bière}, {Couches, Bière}, {Pain}, {Couches}, et {Bière}. Les règles d'associations résultantes sont indiquées ci-dessous:

1. {Pain, Couches} → {Bière}, confiance = 66 %
2. {Pain, Bière} → {Couches}, confiance = 100 %
3. {Couches, Bière} → {Pain}, confiance = 66 %
4. {Pain} →{Couches, Bière}, confiance = 50 %
5. {Couches}→{Pain, Bière} , confiance = 50 %
6. {Bière} → {Pain, Couches} confiance = 66 %

Si le minsup par exemple est de 60%, alors la première, la deuxième, la troisième, et la dernière règle seront affichées en sortie.

### 6.1.4. Avantages :

On peut résumer les avantages des règles d'association dans:

1. La possibilité de découverte des connaissances utiles, cachées dans les bases de données
2. Leurs facilités de compréhension, efficacité et simplicité.
3. Leur formalisme non supervisé et général.
4. Le forage des règles d'association est un grand succès dans divers domaines que ce soit dans des activités commerciales, sociales ou humaines.

### 6.1.5. Inconvénients :

Quelques inconvénients des règles d'association:

1. La découverte d'un nombre important de règles d'association dont la plupart ne sont pas intéressantes.
2. Le temps de recherche des Itemsets fréquents est énorme.
3. Les algorithmes utilisés ont trop de paramètres, par conséquent l'extraction de données, pour les non experts, devient compliquée.
1. Un problème de sécurité pourrait être posé: des renseignements confidentiels peuvent être facilement divulgués, en utilisant cette technique .

## 7. Conclusion :

Nous avons présenté dans ce chapitre les concepts de base liée à l'extraction des règles d'association, l'algorithme Apriori et son déroulement sur un exemple concret. De plus, nous avons présenté les avantages les inconvénients de cette approche.

Dans le chapitre suivant nous allons détailler la phase de collecte de données utilisant un questionnaire de 55 questions en plus d'une étude statistique des données collectées.

# Chapitre III :

# Population de l'etude

## 1. Introduction

Après avoir collecté et stocké les données de l'étude qui concernent les étudiants de mathématiques et informatique en première année par le biais d'un questionnaire distribué sur les étudiants concerné, nous passons a l'étape suivante dans le processus découverte des connaissances à partir de grandes bases de données, le prétraitement des données est une étape critique dans ce processus, en fait, cela améliore la qualité des données soumises ultérieurement aux algorithmes d'exploration de données. Dans ce chapitre nous avons fait une analyse et présentation des données générés par l'étape précédente afin de donner une vision globale sur les données étudiées.

## 2. Contexte de l'étude : Faculté des sciences et de la technologie à l'université de Khenchela

On a effectué notre étude au sein de la faculté des sciences et de la technologie au niveau de l'université Abbas Laghrour de Khenchela.

## Présentation de la faculté

La faculté des sciences et de la technologie est l'une des plus importantes facultés de l'université de Khenchela. elle regroupe six département et trois domaines qui sont : Mathémathiques et informaique ,Science de la matière et Science et technologies.

La Faculté des Sciences et de la Technologie est organisée selon l'organigramme présenté dans la figure 3.1.

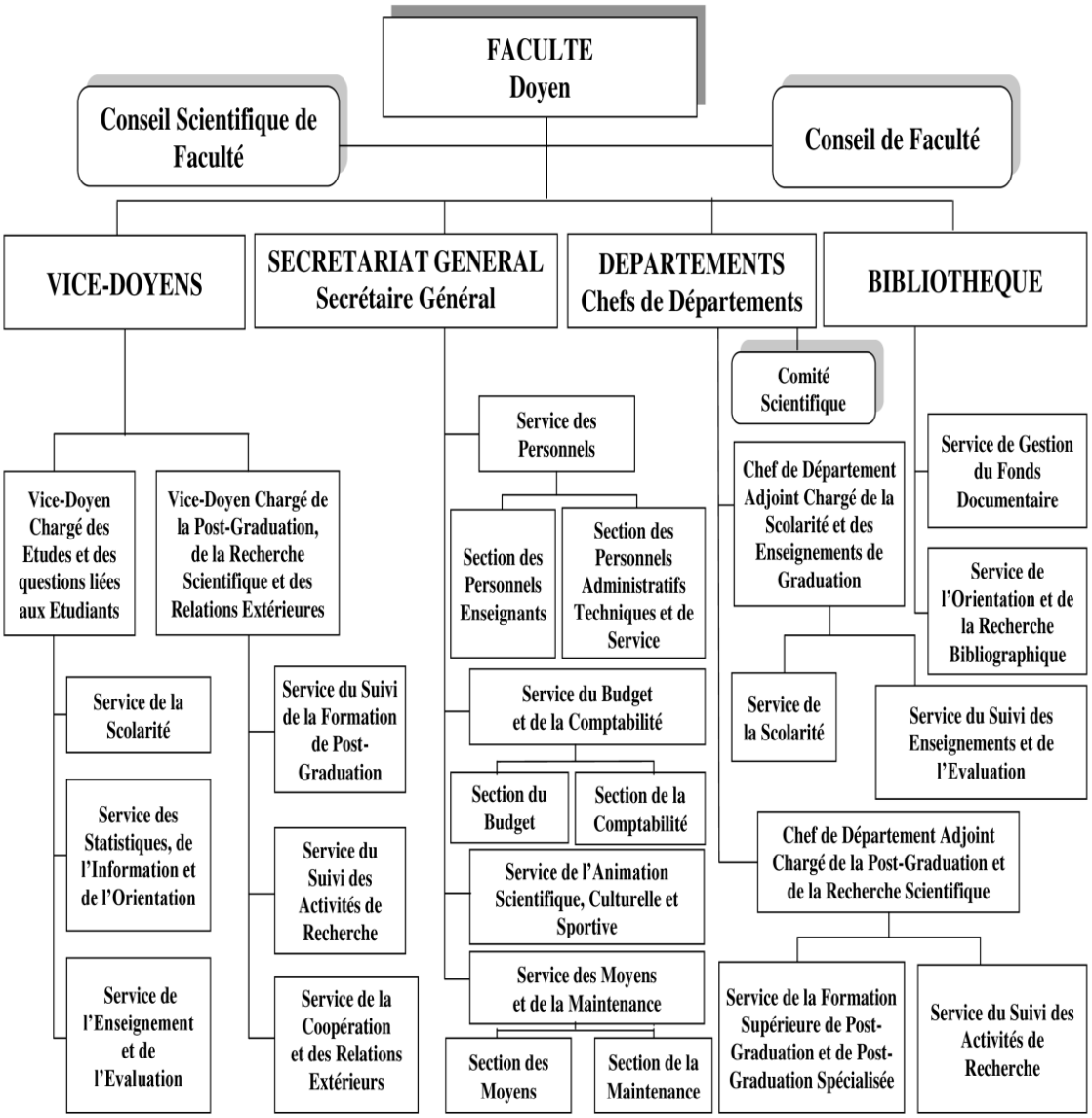


**Figure 3.1 :** Structure de la faculté des Sciences et de la Technologie

## 3. Conception du questionnaire

Le questionnaire utilisé dans notre destiné à une étude pour déterminer les raisons de la réussite en première année universitaire (pour la première fois, c'est-à-dire avant la reprise de l'année), notre questionnaire est préparée au sein du laboratoire d'Informatique ICOSI, Le questionnaire contient un ensemble de questions réparties en six catégories comme suit:

### a. Questions personnelles et sociales en première année :

Cette catégories essaie d'identifier le statut personnel de l'étudiant avec 11 questions, y compris le sexe, l'âge, a-t-il des problèmes familiaux, a-t-il une maladie chronique, le niveau d'éducation de ses parents et la profession de ses parents.

**Exemple** :

- **Quel est votre sexe ?**

✓ Homme

✓ femme

- **Avez-vous des problèmes familiaux en première année ?**

✓ Oui

✓ Non

## b. Questions sur la vie quotidienne :

Cette catégories décrit la situation quotidienne de l'étudiant en posant 11 autres questions différentes pour connaître les facteurs qui ont un impact dans ses études comme le niveau de vie de l'étudiant (élevé, moyen ou bas)

L'étudiant détermine également la distance entre sa résidence et l'université (Moins de 20 km, Entre 20 et 50 km , Plus de 50 km) et le type de logement dans lequel il réside (la résidence des parents. Logement personnel ou universitaire)

Exemple :

- **Votre niveau de vie est :**

o Elevé

o Moyen

o Faible

o Elevé

- **Comment allez-vous habituellement à l'université ?**

o Sur les jambes

o En bus

o En voiture

## c. Questions sur les études secondaires :

Cette catégorie est utilisé pour connaître les capacités de l'étudiant au secondaire et contient 5 questions, connaitre ses capacités en physique et chimie en troisième année secondaire.

Savoir s'il a redoublé ou non l'année au lycée, savoir s'il a effectué des cours spéciales ou non, et connaître son opinion sur ses résultats (très suffisant, suffisant, pas suffisant)

**Exemple :**

- **Quelles sont vos capacités en mathématiques en 3ème secondaire ?**
  - Bon
  - Moyen
  - Sous la moyenne
  - Faible

## d. Questions sur la situation pré-universitaire :

Cette catégories contient 6 questions visant à recueillir des informations relatives à l'obtention de son baccalauréat et à sa moyenne, ainsi qu'à savoir s'il a obtenu ou non d'autres diplôme et à essayer de connaître son opinion sur sa conviction dans son choix du domaine qu'il a choisi.

**Exemple :**

- **Quelle était votre moyenne au baccalauréat ?**
  - Entre 10 et 12
  - Entre 12 et 14
  - Plud de 14

## e. Les études de la première année universitaire à la première fois :

Cette catégorie contient 19 questions et est considéré comme la partie la plus importante car elle contient les facteurs qui affectent la réussite scolaire de l'étudiant, qu'il s'adapte au système universitaire, la difficulté d'étudier en première année, revoir ses cours ou non à la maison, sa relation avec les enseignants, ainsi que sa vie quotidienne en dehors de l'étude, et son comportement à l'université. Toutes ces questions sont essentielles pour déterminer les raisons du succès de la première année.

**Exemple :**

- **Utilisez-vous des livres de bibliothèque ?**
  - Oui toujours
  - Oui, parfois
  - Oui, rarement
  - Jamais
- **Revoyez-vous vos leçons ? Où**
  - Oui à la maison
  - Oui, dans la bibliothèque
  - Oui, ailleurs
  - Non

## f. Résultats de la première année universitaire (sans redoublement)

On considère que la dernière partie du questionnaire contient trois questions qui incluent les résultats du premier et deuxième semestre et le résultat final de la première année.

**Exemple**

- **Quels sont vos résultats dans les six premiers ?**
  - Succès lors de la première session
  - Réussi après le cours de rattrapage
  - Dette réussie
  - Ajourné
- **Quels sont vos résultats finaux de la première année ?**
  - Admis avec dette
  - Admis sans dette
  - Ajournée

## 4. Saisie des données

Après avoir terminé le premier processus et afin de faciliter le processus de collecte de données mentionné dans le questionnaire ci-dessus, nous avons développé un petit logiciel qui permet à l'utilisateur d'entrer des données dans une base de données via une interface graphique.

Chacune des six catégories du questionnaire est présenté sous la forme d'un bouton dans le logiciel, on a ajouté des boutons qui permet la saisie, la suppression ou la modification des données si nécessaire.

Après avoir saisie les données des questionnaires collectés, notre application crée un fichier avec une extension .csv Comme un état de sortie. Il peut être consulté dans Excel et utilisé pour l'analyse et l'apprentissage par d'autres applications.

## 4.1. L'interface graphique du logiciel de saisie des données

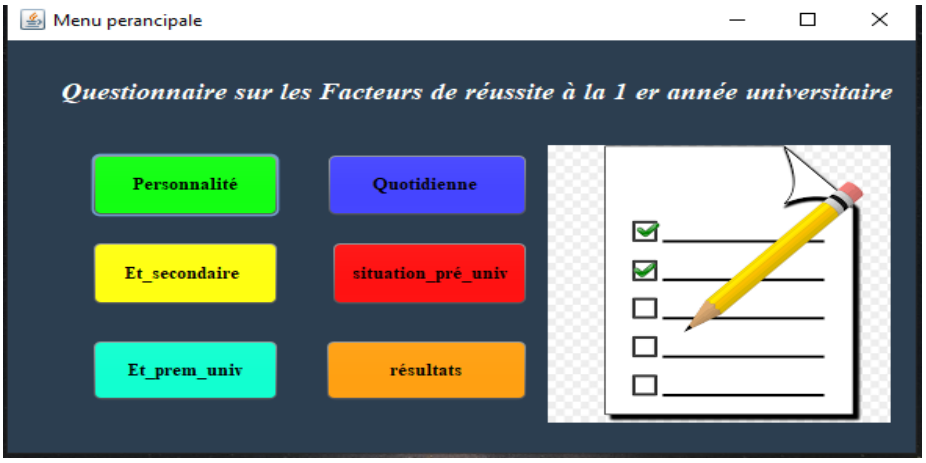


**Figure3.2**: Interface Principale du logiciel de saisie les données

L'interface de notre application de saisie contient les six boutons suivant : Personnalité, Quotidienne, Et secondaire, Situation_pré_univ,Et_prem_univ, résultats. Par exemple, en cliquant simplement sur le bouton "Quotidienne l'interface suivante s'ouvre pour la saisie des données de catégorie concerné dans le questionnaire :

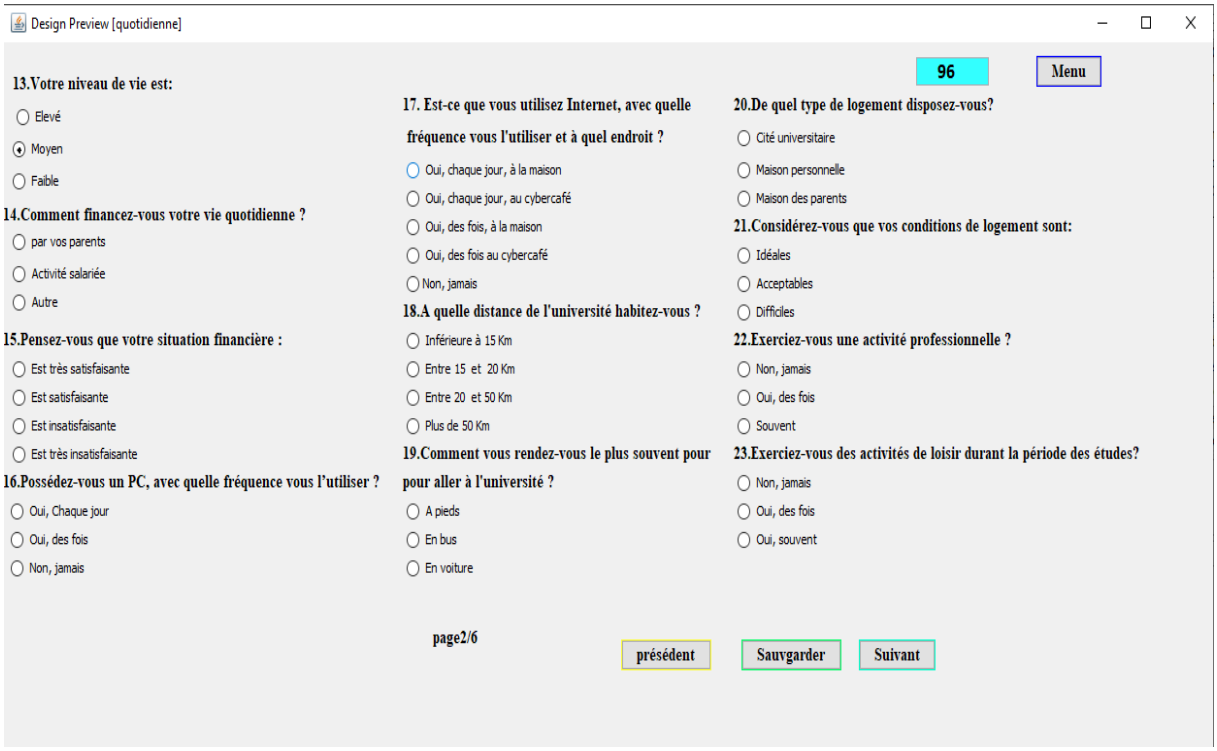


**Figure 3.3:** Interface pour les Questions sur la vie quotidienne

L'interface de la figure 3.2 contient les différentes questions du questionnaire liées à la deuxième catégorie de questions, qui sont des " Questions sur la vie quotidienne ". Chaque question contient au moins deux options où l'utilisateur clique sur l'option spécifique qui correspond au choix de l'étudiant dans le questionnaire.

Comme nous le voyons à la question 13, le choix 2 a été sélectionné.

L'interface contient aussi le numéro du questionnaire comme indiqué dans l'interface précédente, le numéro du questionnaire est 96.

En plus d'un groupe de boutons suivant :

- **Le Bouton sauvegarder :** Ce bouton est utilisé lorsque tous le questionnaire est rempli, ce bouton permet l'enregistrement des données saisies dans la base de données

- **Le bouton supprimer :** Ce bouton supprime toutes les données relatives au numéro de questionnaire à supprimer, en entrant le numéro de questionnaire dans la barre qui lui est affectée et en cliquant sur le bouton supprimer. Nous obtenons une boîte de dialogue indiquant que le questionnaire a été supprimé
- **Le Bouton menu :** Lorsque vous appuyez sur ce bouton, l'interface principale apparaît

## 4.2. Sortie de l'application

Les données collectées et saisies via notre application de saisie sont enregistrées dans un fichier du format CSV (Comma Separated Values), ce fichier est un fichier Excel et peut être utilisé comme une data set pour les étapes suivantes du processus du data mining.

Le format CSV est un format de texte simple qui est utilisé dans de nombreux contextes lorsque de grandes quantités de données doivent être fusionnées sans être directement connectées les unes aux autres.

L'extension de ce type de fichiers est .csv, et ils peuvent être utilisés par plusieurs outils informatiques et bases de données, lorsqu'on souhaite déployer le contenu d'une base de données sur une feuille de calcul.

Des Tableauurs tels qu'Excel (Microsoft) ou Calc (LibreOffice) et des bases de données telles que MySQL et Oracle sont capables d'importer et exporter des fichiers CSV. Toutefois, en raison de sa structure basique, le format de fichier CSV ne convient que pour des données structurées simples.

## 5. Analyse des données collectées

Le processus d'analyse des données de l'étude est de faire une vision globale sur ces données afin de faciliter l'interprétation des résultats finaux.

## Statistiques selon l'attribut « Sexe »

Les résultats de statistiques sur l'attribut sexe sont présentés dans le Tableau3.1.

| Sexe | La fréquence | Pourcentage |
|---|---|---|
| **Male** | 105 | 42% |
| **Femelle** | 145 | 58% |
| **Total** | 250 | 100% |

**Tableau3.1 :** Statistiques sur l'attribut sexe des étudiants

Le nombre total d'étudiants de notre base d'apprentissage est 250, il représente le nombre total des étudiants qui répondent sur notre questionnaire.

Le nombre de femmes est supérieur au nombre d'hommes, le pourcentage de femmes est de 58% avec un nombre de 145 étudiantes, tandis que pour la population masculine, ils sont 105 étudiants et leur pourcentage est de 42%.

## Statistiques selon l'attribut « Age »

Dans le Tableau3.2 on a représenté les statistiques sur l'attribut Age des étudiants concerné par l'étude. Nous avons 3 classes d'étudiants âgés de plus de vingt ans, dix-neuf ans et moins de dix-huit ans, et nous constatons que le pourcentage le plus élevé concerne les étudiants de 19 ans avec un pourcentage de 57,2% et leur nombre est de 143 étudiants, puis en deuxième position les étudiants dont l'âge est plus de 20 ans de 24% et est il y a 60 étudiants, et en dernier lieu les étudiants de moins de 18 ans avec un pourcentage de 18,8% avec 47 étudiants

| Age | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **>=20** | Il représente le pourcentage des étudiants de plus ou égal de 20 ans | **60** | **24%** |
| **19** | Il représente Le pourcentage d'étudiants âgés de 19 ans | **143** | **57.2%** |
| **<=18** | Il représente le pourcentage d'élèves dont l'âge est inférieur ou égal à 18 ans | **47** | **18.8%** |
| **Total** | | **250** | **100%** |

**Tableau 3.2 :** L'âge des étudiants de première année universitaires

## Statistiques selon l'attribut « Résidence »

Dans Tableau3.3, nous notons que le pourcentage d'étudiants résidant dans la wilaya est le plus grand pourcentage, est estimé à 45,2%, et le nombre d'étudiants pour ce ratio est estimé à 113 étudiants. Alors que le reste des étudiants sont répartis dans chacune des communes et des daïras alors qu'un petit nombre d'étudiants résidant dans les villages avec d'environ 3, 2%.

| Résidences | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **DAI** | Le pourcentage d'étudiants résidant dans daïra | **65** | **26%** |
| **WIL** | Le pourcentage d'étudiants résidant dans wilaya | **113** | **45.2%** |
| **COM** | Le pourcentage d'étudiants résidant dans commun | **64** | **25.6%** |
| **VIL** | Le pourcentage d'étudiants résidant dans village | **8** | **3.2%** |
| **Total** | | **250** | **100%** |

**Tableau 3.3:** Statistiques sur la résidence des étudiantes

## Statistiques selon l'attribut « Niveau de vie» des étudiants

Le niveau de vie des étudiants était plus élevé pour ceux ayant un niveau de vie moyen avec un taux de 89, 2%, puis pour les étudiants, dont le niveau de vie était élevé et faible avec un taux faible.

| Niveau de vie | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **MED** | Le pourcentage des étudiants dont le niveau de vie est moyen | **223** | **89.2%** |
| **HIG** | Le pourcentage d'élèves dont le niveau de vie est élevé | **19** | **7.6%** |
| **LOW** | Le pourcentage d'élèves dont le niveau de vie est faible | **8** | **3.2%** |
| **Total** | | **250** | **100%** |

**Tableau 3.4:** Niveau de vie des étudiants

## Statistiques selon l'attribut « Posséder PC»

L'ordinateur est l'un des facteurs qui influent sur les résultats des étudiants. Dans le Tableau3.5, nous trouvons la raison du pourcentage d'étudiants qui utilisent parfois l'ordinateur à un taux de 53, 2%, puis le groupe d'étudiants qui utilisent l'ordinateur quotidiennement arrive à 37, 6% et au dernier rang est la catégorie qui n'utilise pas l'ordinateur par 9 , 2 pour cent.

| Posséder PC | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **YSO** | Il représente le pourcentage des étudiants qui utilisent un ordinateur tous les jours. | **94** | **37.6%** |
| **YID** | Il représente le pourcentage d'élèves qui utilisent parfois l'ordinateur | **133** | **53.2%** |
| **NNE** | Il représente le pourcentage d'élèves qui n'utilisent jamais d'ordinateur | **23** | **9.2%** |
| **Total** | | **250** | **100%** |

**Tableau 3.5:** Le pourcentage des étudiants qui utilisent un ordinateur

## Statistiques selon l'attribut « performance en mathématique »

Le Tableau3.6 montre la compétence des étudiants en mathématiques, le rapport était égal pour les élèves qui ont obtenu des notes très bons et faibles à 4%, et la plus grande pourcentage de ceux qui ont obtenu un bon résultat à un taux élevé de 58, 8%,puis représente le pourcentage des étudiants qui obtiennent un résultat passable à 24%,enfin représente le pourcentage des étudiants qui obtiennent un résultat moyen à 9.2%

| *Performance en mathématique* | *Description* | *La fréquence* | *Pourcentage* |
|---|---|---|---|
| ***PAS*** | Il représente le pourcentage des étudiants qui obtiennent un point passable | **60** | **24%** |
| ***GOO*** | Il représente le pourcentage des étudiants qui obtiennent un bon point | **147** | **58.8%** |
| ***MEY*** | Il représente le pourcentage des étudiants qui obtiennent un point moyen | **23** | **9.2%** |
| ***VGO*** | Il représente le pourcentage des étudiants qui obtiennent un très bon point | **10** | **4%** |
| ***LOW*** | Il représente le pourcentage des étudiants qui obtiennent un point faible | **10** | **4%** |
| ***Total*** | | **250** | **100%** |

**Tableau 3.6:** Performance des étudiants en mathématiques

## Statistiques selon l'attribut « Résultats scolaires »

Dans le Tableau3.7, le pourcentage d'étudiants qui obtiennent des résultats satisfaisants pendant les études secondaires est le plus élevé et est estimé à 62%, et ils sont 155 étudiants. Le pourcentage le plus faible est la catégorie dont les résultats ne sont pas très satisfaisants, avec un très faible pourcentage de 0, 8 % et seulement 2 étédiants alors que le ratio est proche Parmi les résultats étaient très satisfaisants et les résultats étaient insatisfaisants par 27,6 % et 9, 6 %, respectivement

| **Résultats scolaires** | **Description** | La fréquence | Pourcentage |
|---|---|---|---|
| **UNS** | Le pourcentage des étudiants dont les résultats scolaires sont Très satisfaisants | 69 | 27.6% |
| **SAT** | Le pourcentage des étudiants dont les résultats scolaires sont satisfaisants | 155 | 62% |
| **VSA** | Le pourcentage des étudiants dont les résultats scolaires sont insatisfaisants | 24 | 9.6% |
| **VUS** | Le pourcentage des étudiants dont les résultats scolaires sont Très insatisfaisants | 2 | 0.8% |
| **Total** | | 250 | 100% |

**Tableau3.7 :** Résultats scolaires

## Statistiques selon l'attribut « Résultat du baccalauréat »

Dans le Tableau3.8, nous avons trois catégories, la première catégorie des étudiant qui ont une moyenne entre 10 et 12, et c'était le plus grand pourcentage avec 55, 6 % qui représente l'équivalent de 139 étudiants, la deuxième catégorie concerne les étudiants qui ont une moyenne entre 12 et 14, et elle était la deuxième du classement avec 38%, d'étudiants est estimé à 95 étudiants. La troisième catégorie concerne les étudiants qui ont une moyenne supérieure à 14, et c'était le dernier taux du classement à 6,4% est estimé à 16 étudiants.

| *Résultat du baccalauréat* | *Description* | *La fréquence* | *Pourcentage* |
|---|---|---|---|
| ***10_12*** | Le pourcentage des étudiants ayant réussi le baccalauréat se situe entre 10 et 12 | ***139*** | ***55.6%*** |
| ***>14*** | Le pourcentage des étudiants ayant réussi le baccalauréat est supérieur à 14 | ***16*** | ***6.4%*** |
| ***12_14*** | Le pourcentage des étudiants ayant réussi le baccalauréat se situe entre 12 et 14 | ***95*** | ***38%*** |
| ***Total*** | | ***250*** | ***100%*** |

**Tableau 3.8:** Résultat du baccalauréat

## Statistiques selon l'attribut « le choix de domaine »

Le Tableau9 présente les résultats selon l'attribut « est ce que le domaine dans lequel l'étudiant a étudié est de son choix ou non ». Le pourcentage élevé d'étudiants qui ont choisi leur domaine d'études est de 68, 8%. Alors que l'autre groupe des étudiants qui n'ont pas choisi leur domaine d'études le pourcentage était de 31, 5%.

| **le choix de domaine** | **Description** | **La fréquence** | **Pourcentage** |
|---|---|---|---|
| **YES** | Le pourcentage d'étudiants qui ont choisi leur domaine d'études fait partie des spécialisations proposées. | **172** | **68.8%** |
| **NO** | Le pourcentage d'étudiants qui n'ont pas choisi leur domaine d'études fait partie des spécialisations proposées | **78** | **31.5%** |
| **Total** | | **250** | **100%** |

**Tableau 3.9:** Le domaine choisi par l'étudiant pour étudier en première année d'université**.**

## Statistiques selon l'attribut « Difficultés en première année »

Comme le montre le Tableau ci-dessous, nous constatons que la plupart des étudiants ont rencontré des difficultés au cours de la première année de l'enseignement universitaire, et cela est dû à plusieurs facteurs différents de l'enseignement secondaire. Leur pourcentage était élevé, c'est-à-dire plus de la moitié du nombre d'étudiants, le taux atteint 67.2%, tandis que le

reste d'étudiants qui n'ont pas eu de difficultés environ 82 étudiants sur un total de 250 étudiants

.

| ***Difficultés en première année*** | ***Description*** | ***La fréquence*** | ***Pourcentage*** |
|---|---|---|---|
| ***YES*** | Le pourcentage des étudiants ayant rencontré des difficultés en première année | ***168*** | ***67.2%*** |
| ***NO*** | Pourcentage des étudiants n'ayant pas rencontré de difficultés en première année | ***82*** | ***32.8%*** |
| ***Total*** | | ***250*** | ***100%*** |

**Tableau 3.10:** Le pourcentage d'étudiants ayant rencontré des difficultés, en première année universitaire

## Statistiques selon l'attribut « Universite_Like »

Nous notons d'après le Tableau3.11 qu'un grand nombre d'étudiants n'aiment pas l'université et ne viennent pas assister à leurs cours, et qu'un très faible pourcentage d'étudiants qui viennent parfois ou toujours à l'université.

| **Universite_Like** | **Description** | **La fréquence** | **Pourcentage** |
|---|---|---|---|
| **NO** | Le pourcentage d'étudiants qui ne vont pas à l'université | **121** | **48.4%** |
| **YID** | Le pourcentage d'étudiants qui fréquentent toujours l'université | **27** | **10.8%** |
| **JDU** | Représente le pourcentage d'étudiants qui détestent l'université | **38** | **15.2%** |
| **YSO** | Le pourcentage d'étudiants qui fréquentent parfois l'université | **64** | **25.6%** |
| **Total** | | **250** | **100%** |

**Tableau3.11:** Les pourcentages d'étudiants qui fréquentent toujours l'université et qui détestent l'université.

## Statistiques selon l'attribut « Révisez cours »

La révision des cours a toujours été un catalyseur majeur pour la réussite et la supériorité des étudiants dans leur parcours académique à tous les niveaux d'études. Le Tableau12 montre le pourcentage des étudiants revoyant leurs leçons.

Le pourcentage le plus élevé des étudiants concerne les étudiants qui ont révisé leurs cours à la maison et était 48%. Le deuxième pourcentage était pour les étudiants qui avaient révisé leurs cours à la bibliothèque et était estimé à 25,2%. Le pourcentage des étudiants qui n'ont pas révisé leurs cours était de 17,2%, ce qui est un pourcentage significatif.

| ***Révisez cours*** | ***Description*** | ***La fréquence*** | ***Pourcentage*** |
|---|---|---|---|
| ***NO*** | Le pourcentage d'étudiants qui ne révisent pas leurs leçons | ***43*** | ***17.2%*** |
| ***YAH*** | Le pourcentage d'étudiants révisant leurs cours à la maison | ***120*** | ***48%*** |
| ***YAL*** | Le pourcentage d'étudiants révisant leurs cours à la bibliothèque | ***63*** | ***25.2%*** |
| ***YOTH*** | Le pourcentage d'étudiants revoyant leurs cours ailleurs | ***24*** | ***9.6%*** |
| ***Total*** | | ***250*** | ***100%*** |

**Tableau 3.12:** Révisez cours

## Statistiques selon l'attribut « Assister cours »

Le Tableau 3.13 nous montre le pourcentage d'étudiants qui suivent leurs cours. Le pourcentage d'étudiants qui n'assistent jamais à leurs cours était le plus bas de 12%, ce qui équivaut à 30 étudiants, et vice versa pour les étudiants qui assistent à leurs cours ou sont parfois assistent à leurs cours, leur taux est élevé de 32,8% et 55,2% en ordre.

| Assister cours | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **YSO** | Le pourcentage d'étudiants qui assistent parfois à leurs cours. | **138** | **55.2%** |
| **YID** | Le pourcentage d'étudiants qui assistent toujours leurs cours. | **82** | **32.8%** |
| **NNE** | Le pourcentage d'étudiants qui n'assistent jamais à leurs cours. | **30** | **12%** |
| **Total** | | **250** | **100%** |

**Tableau3.13** : Assister cours

## Statistiques selon l'attribut « Assister TD »

Le Tableau3.14 nous montre le pourcentage d'étudiants qui suivent leurs TD. Le pourcentage d'étudiants qui assistent rarement à leurs TD était le plus bas de 1,2% , ce qui équivaut à 3 étudiants, et vice versa pour les étudiants qui assistent à leurs TD leur taux est élevé de 79,2% L'équivalent de198 étudiants , Le pourcentage d'étudiants qui assistent parfois à leurs TDs est égale 14,4 % C'est aussi un pourcentage important , enfin Le pourcentage d'étudiants qui n'assistent jamais à leurs TD est de 5,2 , ce qui équivaut à 13 étudiants .

| AssisterTDs | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **YID** | Le pourcentage d'étudiants qui assistent toujours leurs TDs | **198** | **79.2%** |
| **NNE** | Le pourcentage d'étudiants qui n'assistent jamais à leurs TDs. | **13** | **5.2%** |
| **YSO** | Le pourcentage d'étudiants qui assistent parfois à leurs TDs | **36** | **14.4%** |
| **YON** | Le pourcentage d'étudiants qui assistent rarement à leurs TDs | **3** | **1.2%** |
| **Total** | | **250** | **100%** |

**Tableau3.14:** Assister TDs

## Statistiques selon l'attribut « Révisez avec collègues »

Il existe des types d'étudiants concernant la révision avec leurs collègues, certains soutiennent cette idée et certains ne la soutiennent pas, mais en général la révision avec des collègues est un moyen efficace pour l'étudiant de réussir et de se débarrasser de plusieurs problèmes dans ses leçons, le Tableauci-dessous nous montre le point de vue des étudiants sur ce sujet :

Le pourcentage le plus élevé concerne les étudiants qui révisent parfois leurs leçons avec des collègues, et ce pourcentage est estimé à 46,8%, Autrement dit, près de la moitié des étudiants. Le faible pourcentage est pour les étudiants qui ne revoient jamais leurs leçons avec leurs collègues,leur pourcentage était très proche entre d'étudiants qui revoient rarement leurs leçons avec leurs collègues, et les étudiants qui revoient toujours leurs leçons avec leurs collègues à 20% et 23.2%.

| Révisez avec collègues | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **YSO** | Le pourcentage d'étudiants qui revoient parfois leurs leçons avec leurs collègues | **117** | **46.8%** |
| **YON** | Le pourcentage d'étudiants qui revoient rarement leurs leçons avec leurs collègues | **50** | **20%** |
| **NNE** | Le pourcentage d'étudiants qui ne revoient jamais leurs leçons avec leurs collègues | **25** | **10%** |
| **YID** | Le pourcentage d'étudiants qui revoient toujours leurs leçons avec leurs collègues | **58** | **23.2%** |
| **Total** | | **250** | **100%** |

**Tableau3.15:** Révisez avec collègues

## Statistiques selon l'attribut « Utiliser la bibliothèque »

La bibliothèque a toujours été connue comme un moyen d'aider les étudiants à améliorer leur niveau d'éducation en utilisant des livres, des mémoires et diverses références qui y sont disponibles.

L'utilisation de la bibliothèque est une raison importante de la réussite de l'étudiant dans ses études, grâce au Tableau16, nous notons que la plupart des étudiants utilisent la bibliothèque rarement ou ne l'utilisent jamais d'un taux significatif de 42.8% et 30.4% en

ordre ,tandis qu'un très faible pourcentage d 'étudiants qui utilisent toujours la bibliothèque à un taux n'excédant pas 8,4%, ce pourcentage sont très peu nombreux par rapport au premier

| Utiliser la bibliothèque | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **NNE** | Le pourcentage d'étudiants qui n'utilisent jamais la bibliothèque | **76** | **30.4%** |
| **YON** | Le pourcentage d'étudiants qui utilisent rarement la bibliothèque | **107** | **42.8%** |
| **YSO** | Le pourcentage d'étudiants qui utilisent parfois la bibliothèque | **46** | **18.4%** |
| **YID** | Le pourcentage d'étudiants qui utilisent toujours la bibliothèque | **21** | **8.4%** |
| **Total** | | **250** | **100%** |

**Tableau3.16:** Utilisation de la bibliothèque

## Statistiques selon l'attribut « Absences »

Le Tableau17 Analyse l'absence d'étudiants à l'université, le pourcentage le plus élevé concerne les étudiants qui manquent rarement leurs études , et ce pourcentage est estimé à 48.8%, Autrement dit, près de la moitié des étudiants. Le faible pourcentage est pour les étudiants manquent toujours leurs études, leur taux est de 4%, et le pourcentage était très proche entre les étudiants qui manquent parfois et qui ne manquent jamais leurs études à 22.8% et 24.4% successivement.

| Absences | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **YSO** | Le pourcentage d'étudiants qui manquent parfois leurs études | **57** | **22.8%** |
| **YON** | Le pourcentage d'étudiants qui manquent rarement leurs études | **122** | **48.8%** |
| **NNE** | Le pourcentage d'étudiants qui ne manquent jamais leurs études | **61** | **24.4%** |
| **YID** | Le pourcentage d'étudiants qui manquent toujours leurs études | **10** | **4%** |
| **Total** | | **250** | **100%** |

**Tableau3.17 :** Absences

## Statistiques selon l'attribut « Résultats du premier Semestre »

Le Tableau3.18 montre que la majorité des étudiants ont réussi au premier semestre à un taux élevé de 54%, tandis que les étudiants n'ayant pas réussies au premier semestre de 14,8%, tandis que le reste des étudiants qui ont réussi après la session rattrapage de 22%, Le pourcentage le plus faible est pour les étudiants qui ont réussi le premier semestre avec des dettes était de 9,2%.

| Les résultats du premier Semestre | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **AAR** | Le pourcentage d'étudiants admis après la session rattrapage | **55** | **22%** |
| **ASN** | Le pourcentage d'étudiants admis a la session normale | **135** | **54%** |
| **AAD** | Le pourcentage d'étudiants admis avec dettes | **23** | **9.2%** |
| **AJO** | Le pourcentage d'étudiants ajourné | **37** | **14.8%** |
| **Total** | | **250** | **100%** |

**Tableau3.18 :** Les résultats du premier Semestre

## Statistiques selon l'attribut « *Les résultats du deuxième Semestre* »

On a remarqué dans le Tableau3.19 que les résultats du deuxième Semestre sont presque les mêmes que les résultats du premier Semestre, la majorité des étudiants ont réussi au deuxième semestre à un taux élevé de 51.6%, tandis que les étudiants n'ayant pas réussies au deuxième semestre de 14,8% c'est le même pourcentage dans les résultats du premier semesstre, tandis que le reste des étudiants qui ont réussi après la session rattrapage de 26.8%, Le pourcentage le plus faible est pour les étudiants qui ont réussi le deuxième semestre avec des dettes était de 6.8%.

| Les résultats du deuxième Semestre | Description | La fréquence | Pourcentage |
|---|---|---|---|
| **AAR** | Le pourcentage d'étudiants admis après la session rattrapage | **67** | **26.8%** |
| **ASN** | Le pourcentage d'étudiants admis a la session normale | **129** | **51.6%** |
| **AAD** | Le pourcentage d'étudiants admis avec dettes | **17** | **6.8%** |
| **AJO** | Le pourcentage d'étudiants ajourné | **37** | **14 .8%** |
| **Total** | | **250** | **100%** |

**Tableau3.19 :** Les résultats du deuxième Semestre

## Statistiques selon l'attribut « Les résultats finals »

Le Tableau20 présente les résultats finaux de la première année d'études universitaires pour la première fois, où nous constatons que le pourcentage le plus élevé était pour les étudiants qui ont réussi sans les dettes à 62%, ce qui est un bon pourcentage. Les étudiants qui ont passé la première année avec des dettes étaient de 21,6%, ce qui est un pourcentage significatif aussi. Alors que le pourcentage d'étudiants qui ont échoué en première année de l'université était de 16,4%, soit 41 étudiants.

| Les résultats finals | La fréquence | Pourcentage |
|---|---|---|
| **AAD** | **54** | **21.6%** |
| **ASD** | **155** | **62%** |
| **AJO** | **41** | **16.4%** |
| **Total** | **250** | **100%** |

**Tableau3.20 :** Les résultats finals

## 6. Conclusion

L'objectif de ce chapitre était de collecter le maximum de données dans une base de données relatives aux étudiants de première année en mathématiques et en informatique afin de l'utiliser pour la génération de modèle d'un côté et de générer des dépendances entre ses attributs, Grâce aux techniques de fouille de données (data mining ) et les algorithmes offerts par l'outils de data mining Weka, nous allons faire des études sur ces données comme présente le chapitre suivant.

# *Chapitre IV:*

# *Implémention et Réalisation*

# 1. Introduction

L'objectif principal de l'implémentation de l'application de datamining après un série de plusieurs étapes dans le processus de développement est de développer une application qui sert à extraire des règles d'association entre les attributs des étudiants de première année du département de sciences et de la technologie utilisant l'algorithme apriori.

Ce chapitre est essentiellement axé sur les grandes lignes qui nous ont permis de réaliser et de mettre en œuvre ce projet, et les outils exploités pour le développement du logiciel tels que l'environnement de programmation, le matériel utilisé et les principales fonctions utiliser, ainsi que les technologie de datamining utilisées pour implémenter ce type de système.

Ensuite, nous présenterons les résultats de fouille de données en termes de règles d'association (l'algorithme apriori) générées par notre système et en terminera ce chapitre par une analyser des résultats obtenus.

# 2. Environnement de travaille et outils utilisés

L'environnement de travail est constitué par deux parties nommées environnement matériel et environnement logiciel.

## 2.1 Environnement matériel

L'environnement matériel utilisé pour accomplir ce travail est caractérisé par :

a. Système d'exploitation : Windows 10 professionnel 64-bit
b. CPU : Intel(R) Core (TM) i5-4300U CPU @ 1.90GHz (4 CPUs), ~2.5GHz
c. Mémoire : 4096MBRAM

## 2.2 Environnement logiciel

L'environnement logiciel utilisé composé des outils logiciels suivants :

### 2.2.1 java

Java est un langage de programmation et une plate-forme informatique qui ont été créés par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. Java est rapide, sécurisé et fiable. Des ordinateurs portables aux centres de données, des consoles de jeux aux superordinateurs scientifiques, des téléphones portables à Internet, la technologie Java est présente sur tous les fronts !

Il est fourni avec un ensemble d'outils (le JDK Java Développement Kit) et un ensemble

de packages : ensemble de classes. Ces différentes classes de base couvrent beaucoup de domaine (entrées/sorties, interface graphique, réseau, etc.) Cette richesse en "librairies standards" explique sûrement en partie le succès de Java. Le langage lui-même se trouve dans le package java Lang .

Java est donc :

- Un langage de programmation objets.
- Une architecture de machine virtuelle.
- Un ensemble d'outils.
- Un ensemble de librairies (packages) de base.

### 2.2.3. NetBeans :

NetBeans est un environnement de développement intégré (IDE) pour Java, placé en open source par Sun en juin 2000 sous licence CDDL (Common Développement and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, XML et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages web).

NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X et Open VMS. NetBeans est lui-même développé en Java, ce qui peut le rendre assez lent et gourmand en ressources mémoires.

### 2.2.4. weka :[18]

Weka (Waikato Environment for KnowledgeAnalysis) est un ensemble d'outils permettant de manipuler, d'analyser des fichiers de données et pour implémenter différents algorithmes d'apprentissage artificielle(les arbres de décision, les réseaux de neurones, etc.).Il est écrit en java, développée à l'université de WaikatoenNouvelleZélande 1992.

Parmi tous les outils d'exploration de données disponibles, Weka est le plus couramment utilisé en raison de ses performances et de son support rapide pour l'algorithme apriori et classification.

WEKA supporte plusieurs outils d'exploration de données standards, et en particulier, des préprocesseurs de données, des classificateurs, des analyseurs de régression, des outils de visualisation, et des outils d'analyse discriminante. Le format des données d'entrée par défaut de WEKA est ARFF (Attribute Relation File Format). D'autres formats peuvent être importés comme CSV, Binaire, BDD SQL (avec JDBC) à partir d'une URL, etc. WEKA contient plus de

70 algorithmes de classification / régression supervisés (Tableau II.1), plus de 15 évaluateurs d'attributs et plus de 10 algorithmes de recherche pour la sélection d'attribut, des algorithmes de recherche de règles d'association et plusieurs interfaces graphiques GUI. WEKA s'ouvre avec quatre options (Explorer, expérimenter, KnowledgeFlow et CLI simple). Principalement, Explorer et Expérimentateur sont utilisés pour l'extraction de données. A titre de comparaison de multiples algorithmes, l'Expérimentateur est utilisé, mais pour des résultats spécifiques à l'extraction de données, l'Explorateur est utilisé.


**Figure 4.1:**Interface graphique de WEKA

## Avantage :

- Une interface très complète : WEKA présente quatre modes et implémente beaucoup d'algorithmes pour chaque tâche.
- Possibilité de traiter les données d'une base de données.
- Traitement des données manquantes.
- Une bonne gestion des erreurs (les contrôles logiques)
- Beaucoup de filtres pour faire des transformations sur les données.
- Possibilité de faire des comparaisons entre les différentes méthodes.
- Multi-plate-formes (Windows, Lunix, MAC OS).
- Extensible.

## Inconvenient:

- Nécessite une lecture attentive de la documentation, car la manipulation est difficile.
- Absence de tests statistiques.
- Une limitation technologique (JAVA) sur la taille de la base.
- Une limitation technologique (JAVA) sur la rapidité.

## Format de fichier supporté

WEKA utilise (entre autres) le format de fichier arff pour enregistrer les données. Un fichier arff est composé d'une liste d'exemples définis par leurs valeurs d'attributs.

Un fichier arff comprend toujours trois types d'informations : un nom pour la base de données, des attributs et des données.

La chaîne de caractères @RELATION permet de donner un nom à la base de données. Par exemple, dans le cas du fichier data.arff, le nom donné est data.

@RELATION data

La chaîne de caractères @ATTRIBUTE permet de définir un attribut. Un attribut peut être de 4 types :

– réel (NUMERIC ou REAL)

– Nominal ({valeurs-possible}) : par exemple :

@ATTRIBUTE class {data-setosa,data-versicolor,data-virginica} signifie que l'attribut class peut avoir comme valeur soit data-setosa, soit data-versicolor ou soit data-virginica.

– Chaîne de caractère (STRING)

– Date (date [<date-format>])

## 3. Architecture du système développé

Le système développé a une architecteur modulaire composée de quatre modules principaux a savoir :

- Préparation de la base d'aprentissage
- Paramétrage de l'algorithme de génération des règles d'association
- Génération des règles d'association
- Présentation des règles d'association aux utilisateurs

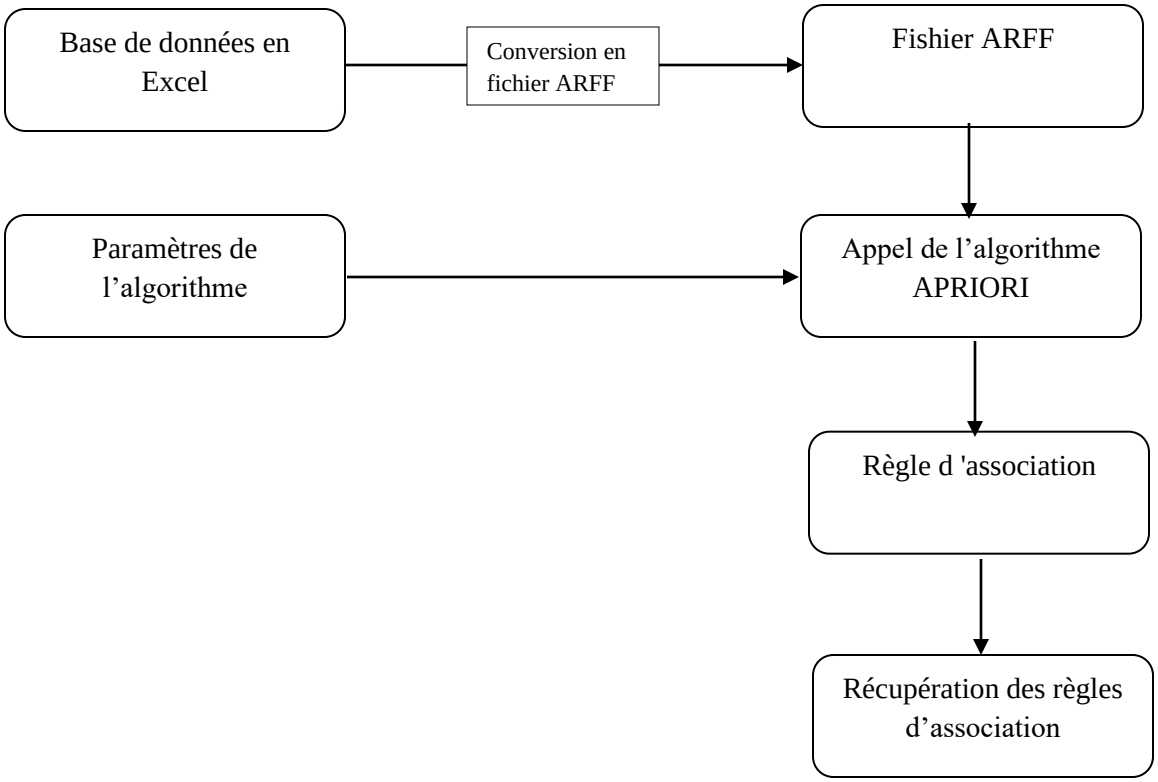


**Figure 4.2:** Architecture du système développé

## 4. Appel des classes Weka dans Java [19]

Pour pouvoir utiliser les classes de Weka dans java on doit suivre les étapes suivantes :

### a) Lecture du fichier d'apprentissage

```
FileReader reader = new FileReader(fichierEntrainement);
```

Cette méthode nous permet de chargement d'un fichier ARFF qui contient les informations d'apprentissage

### b) création des exemples d'apprentissage à partir du fichier

```
Instances instances = new Instances(reader);
```

Pour créer des exemples d'apprentissage à partir du fichier ARFF

### c) choix de l'attribut

laméthode suivant permet de choisir les attribut de la classe apprendre

```
instances.setClassIndex(instances.numAttributes() - 1);
```

**d) Appele de l'algorithm qui génère les règles d'association**

L'Appelle de l'algorithme apriori de weka

Apriori model = new Apriori();

Passage des paramètres de l'algorithme apriori

```
String[] options=new String[6];
options[0]="-N";
options[1]="20";
options[2]="-C";
options[3]="0.8";
options[4]="-M" ;
options[5]="0.2";
model.setOptions(options);
model.getOptions();
```

**e) Récupération et affichage des règles d'association**

L'affichage des règles d'association a partir de l'algorithme apriori

```
System.out.println(model)
```

## 5. Fonctionnement du système développé

Le système développé permet d'extraire les connaissances sous forme de règles d'association à partir des donées des étudiants de première année MI.

Le système conçu appliquera l'algorithme Apriori comme une techniques d'extraction des règles d'association pour en extraire les règles. Notre système permet à l'utilisateur de modifier et afficher les paramètres de l'algorithme de l'extraction des règles d'association. L'utilisateur pourra modifier les paramètres suivant : le nombre de rôle, support et confiance, et ensuite appliquer l'algorithme.

L'application affichera aux utilisateurs une liste de règles d'association qui respectent les paramètres saisie.

## 5.1. Paramètres du système développé

### 5.1.1. Choix des paramètres de l'algorithme de génération des règles d'association

Dans notre système avant chaque extraction il faut choisir le nombre de rôle, la valeur du seuil minimal de support et de confiance.

Le paramétrage se fait par la modification le nombre de rôle, des coefficients support et confiance a travers une simple interface.

**Le nombre de rôle**

Le nombre de répétition de notre algorithme apriori.

**Support**

Le choix du support se fait directement en introduisant le seuil de support désiré dans la case Support.

**Confiance**

Le choix du coefficient de confiance se fait directement en introduisant le seuil de confiance désiré dans la case Confiance.

**Exemple**

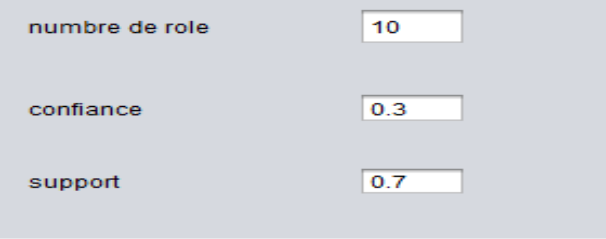


**Figure 4.2:** choix des paramètres

Dans ce cas, l'utilisateur choisit d'appliquer l'algorithme de génération de règles d'associations selon que le nombre de rôle est égale a 10, et avec un support de 70% et une confiance de 30%.

### 5.1.2. Choix de la base de données utilisée

En cliquant sur le bouton « select », le système donne la possibilité de sélectionner n'importe quelle base de données pour l'utilisé comme illustre les figures 4.3 et 4.4.

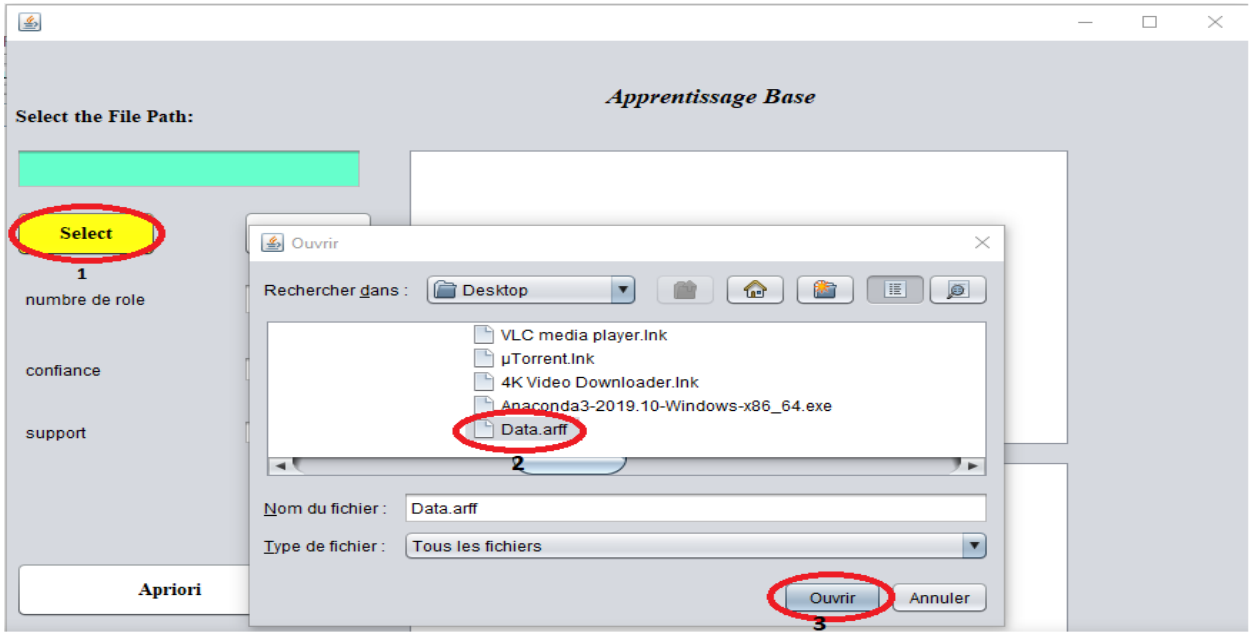


**Figure 4.3:** sélection le fichier ARFF.

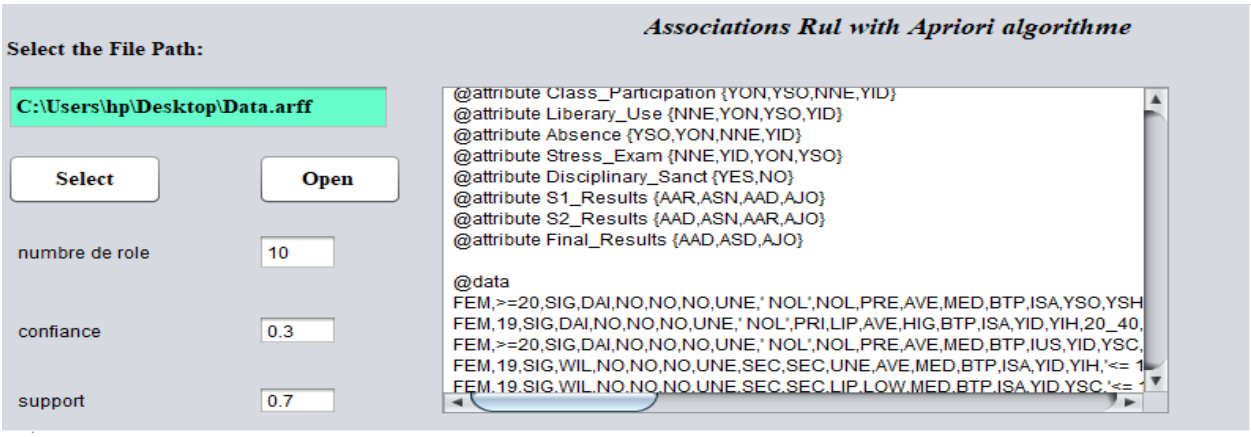


**Figure 4.4:** affichage des attributs et les informations

### 5.1.3. Extraction des règles d'association fortes

L'extraction des règles d'association fortes se fait automatiquement après l'appel de l'algorithme sur la bas de données sélectionnée et affiche systématiquement le résultat final dans la case «Règles Fortes» avec la confiance de chaque règle d'association comme illustre la figure 4.5.

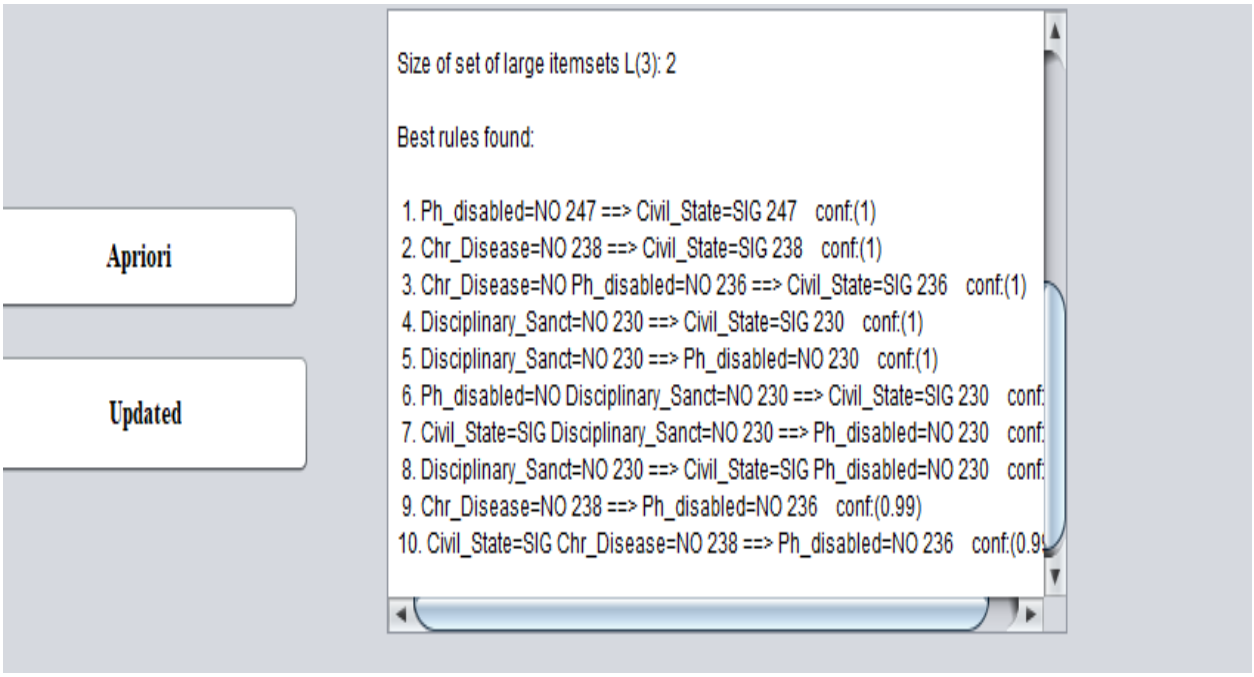


**Figure 4.5:** affichage les règles forts.

## 6. Scénariosde teste

Dans cette partie on va exprimer les scenarios de test de l'algorithme APRIORI utilisant plusieurs valeurs des propriétés confiance et support :

| scenario | Confiance | Support | Nombre de règle fort |
|---|---|---|---|
| 1 | 0.2 | 0.2 | 3289 |
| 2 | 0.2 | 0.5 | 16 |
| 3 | 0.2 | 0.9 | 0 |
| 4 | 0.5 | 0.2 | 1809 |
| 5 | 0.5 | 0.5 | 16 |
| 6 | 0.5 | 0.9 | 0 |
| 7 | 0.9 | 0.2 | 16 |
| 8 | 0.9 | 0.5 | 0 |
| 9 | 0.9 | 0.9 | 0 |

**Table 4.1:** tableau des scenarios de teste

Apres notre observation du tableau 1 il est clair que le cinquième scenarioest le meilleur parmi les autres.

Donc, dans le reste du chapitre on traite les résultats obtenus par les paramètres du scénario 5.

La figure présente l'ensemble des règles d'associations générés par le scénario 5.

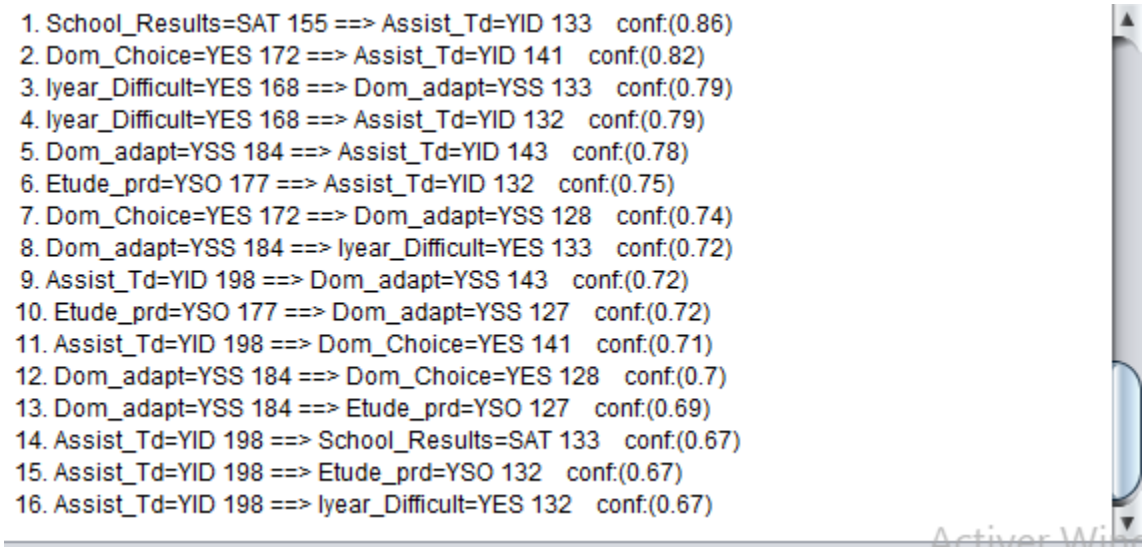


**Figure 4.6:** resultat de scenario n °5

**Règle 1:**School_Results=SAT 155 ==>Assist_Td=YID 133 conf:(0.86)

La règle n °1 indique que les étudiantes qui ont satisfaits par leurs resultats secondaire assistent les TD toujours .Cette regle est tres forte car elle a une confiance de 86%.

**Règle 2:**Dom_Choice=YES 172 ==>Assist_Td=YID 141 conf:(0.82)

Cette règle montre que les étudiants qui ont chosaient leurs domaine d'étude implique que ces étudiants assistent les TD toujours. Avec une confiance de 82%.

**Règle 3:**Iyear_Difficult=YES 168 ==>Dom_adapt=YSS 133 conf:(0.79)

On note dans cette règle que les étudiantsqui ont des difficultés pour adapte aux domaines dans la première année sont adaptent dans a le deuxième semestre. La confiance de cette règle est 79%.

**Règle4:**Iyear_Difficult=YES 168 ==>Assist_Td=YID 132 conf:(0.79)

Pour cette règle on a remarqué que les étudiants qui ont des difficultés dans la première année assistent les TD toujours. La confiance 79%.

**Règle5:**Dom_adapt=YSS 184 ==>Assist_Td=YID 143 conf:(0.78)

Les étudiants qui adaptent avec leur domaine a la deuxième semestre assistent les TD toujours. Avec une confiance de 78%.

**Règle6:**Etude_prd=YSO 177 ==>Assist_Td=YID 132 conf:(0.75)

La révision des étudiants avec leur collègues sa veut dire qu'ils assistent les TD toujours la confiance 75%.

**Règle 7:**Dom_Choice=YES 172 ==>Dom_adapt=YSS 128 conf:(0.74)

Les étudiants qui ont choisirent le domaine vont adapte avec ce domaine dans le deuxième semestre. Avec une confiance de 74%.

**Règle 10 :**Etude_prd=YSO 177 ==>Dom_adapt=YSS 127 conf:(0.72)

Les étudiants qui révisent avec leur collègue adapte avec le domaine a le deuxième semestre. Confiance de 72%.

## 7. Conclusion

Dans ce chapitre nous présenté l'environnement matériel et l'environnement logiciel, nous avons parlé au fonctionnement du système développé et les paramètres de ce système et Nous avons également fourni comment pouvoir utiliser les classes de Weka dans java, Les différents scenarios de test et l'algorithme utilisé pour notre étude(l'algorithme apriori). Enfin On a effectué des scenarios pour extraire et analyser les règles d'association fort.

# Conclusion générale

# Conclusion générale

Avec l'augmentation de la capacité de stockage, nous avons assisté durant ces dernières années à une croissance importante des moyens de génération et de collection des données. C'est ainsi que l'on a commencé à parler de découverte de connaissances à partir de données (KDD) ou encore de Data Mining ou de Fouille de données. Les techniques de Data Mining permettent de découvrir des informations importantes (cachées) dans les données.

Ce travail de mémoire fait partie d'un projet consistant à produire une application de Datamining qui sert à extraire des règles d'association entre les attributs des étudiants de première année universitaire mathématiques et informatique utilisant l'algorithme apriori, pour cela nous avons utiliser l'algorithme apriori pour l'extraction des règles d'association. Nous y avons appliqué plusieurs scénarios pour l'extraction des règles d'association et on a choisit les meilleures relations (les règles d'associations fortes) qui a pour but de découvrir des relations significatives entre les attributs de la base de données collectée.

L'objectif de notre travail est d'extraire les règles d'association fortes qui se fait automatiquement après l'appel de l'algorithme Apriori sur la base de données qui contient des données pour les étudiants de première année Mathématiques et informatique.

# *Annexe*

# ANNEXE

| *N* | *les attributs* | *code attrib* | *valeur attri* | *code et valeur attrib* |
|---|---|---|---|---|
| 1 | The sex of the student | Sex | (Male,Female) | Male, Female |
| 2 | Age | Age | ( 18 or younger, 19, 20 or older) | <=18,19,>=20 |
| 3 | Civil state | Civil_State | Single, Married Without Children, Married With Children) | ( SIG, MWC, MWH) |
| 4 | Residence | Residence | (Wilaya, Daïra, Commune, Village) | (WIL,DAI,COM ,VIL) |
| 5 | Family problems | Fam_Problems | (Yes, No) | (YES,NO) |
| 6 | Chronic disease | Chr_Disease | (Yes, No) | (YES,NO) |
| 7 | Physically disabled | Ph_disabled | (Yes, No) | (YES,NO) |
| 8 | Mother's educational level | Mother_ Educ_Level | (No Level, Primary, Secondary , University) | (NOL, PRI, SEC , UNI) |
| 9 | Mother_ Profession | Mother profession | (Liberal Profession, State Employee, Private Employee , Unemployed) | (LIP,STE,PRE ,UNE) |
| 10 | Father's educational level | Father_ Educ_Level | (No Level, Primary, Secondary, University ) | (NOL,PRI,SEC, UNI) |

| N | les attributs | code attrib | valeur attri | code et valeur attrib |
|---|---|---|---|---|
| 11 | Father profession | Father_ Profession | (Liberal Profession, State Employee, Private Employee, Retirement , Unemployed) | (STE,LIP,PRE, RET,UNE) |
| 12 | confidence in your scientific knowledge | Confidence_ scientific | ( Strong, Low , Averag) | (STR,LOW,AVE) |
| 13 | Life level | Life_Level | (High, Medium, Low) | (HIG,MED,LOW) |
| 14 | Source of funding | Source_ Funding | (By the Parents, Employed ) Activity, Other | (BTP,EMA,OTH) |
| 15 | Family situation | Family_ Situation | (is Very Satisfactory, is Satisfactory, isUnsatisfactory, is Very Unsatisfactory) | (IVS,ISA,IUS, IVU) |
| 16 | The student have a PC, how often it use it | Own_PC | (Yes Every Day, Yes Sometimes , No Never) | (YID,YSO,NNE) |
| 17 | The student uses the Internet , how often he uses it and where | Internet_uses | (Yes Every Day at Home, Yes Every Day at the Cyber cafe, Yes Sometimes at Home, Yes, Sometimes at the Cyber cafe, No Never) | (YIH,YIC,YSH, YSC,NNE) |
| 18 | The distance between housing and university | Dist_Hous _Univ | (Less than 20KM , Between 20 and 50KM, Greater than 20 KM) | (<15, 15_20, 20_40,>40) |

| *N* | *les attributs* | *code attrib* | *valeur attri* | *code et valeur attrib* |
|---|---|---|---|---|
| 19 | Means of transportation to the university | Means_Trasport | (On Foot, By Bus, By Car) | (OFO,BBU,BCA) |
| 20 | Housing type | Housing_Type | (University Campus, Personal House, Parents' House) | (UNC,PAH,PEH) |
| 21 | Housing conditions | Housing_ Conditions | (Idéales, Acceptables , Difficiles) | (IDE,ACC,DIF) |
| 22 | The student exercises a leisure activity | Leisure_ Activity | (No Never, Yes, ) Sometimes,Often | (NNE,YSO,OFT) |
| 23 | exercise leisure activities during the study period | Etude_prd | ( No never, Yes sometimes, Yes, often) | ( NNE, YSO, YFT) |
| 24 | Student's performance in physics and chemistry | Performance _Chi | (Good, Average, Fair, Poor) | (VGO,GOO,FAI ,PAS,LOW) |
| 25 | Student's performance in mathematics | Performance _Mat | (Good, Average, Fair, Poor) | (VGO,GOO,FAI ,PAS,LOW) |
| 26 | The student repeated a year in secondary | Repeat _Year | (Yes, No) | (YES,NO) |
| 27 | the student does special courses in terminale | Special _Courses | (Yes in Math, Yes in Physics , yes in , Science in Other , No) | (YSM,YIO,NO) |
| 28 | School results | School _Results | (Very Satisfactory , Satisfactory, Unsatisfactory) | (VSA,SAT,UNS ,VUS) |

| *N* | *les attributs* | *code attrib* | *valeur attri* | *code et valeur attrib* |
|---|---|---|---|---|
| 29 | Baccalaureate type | Bac_Type | (Mathematics, Experimental Sciences, Technical Math, Other Speciality) | (MAT,ESC,TMA ,OTHS) |
| 30 | Baccalaureate result | Bac_Everage and 12, between 12 and 14, Superior to 14) | ( between 10 | (10_12, 12_14,>14) |
| 31 | another diploma | Auther _Dip | ( No None, Yes, license, Yes, Master, Yes, Classic system , Other) | ( NAUC, YLE, YMA, YCS, OTH) |
| 32 | People who influence your choice of field foryear the first academic | Choice _Field _Infl | (None, My Family, My Loved ones, Other) | ( AUC, MAF, MRE , OTH) |
| 33 | Your first-year domain is among the top three choices | Dom_ Choise | (Yes, No) | (YES,NO) |
| 34 | Did you like the affected domain | Dom_Like | (Yes, No) | (YES,NO) |
| 35 | first year domain | Dom_adapt | (Mathematics and informatics ,science and technology,Material Sciences,Other) | (MI,ST,SM, OTH) |

| *N* | *les attributs* | *code attrib* | *valeur attri* | *code et valeur attrib* |
|---|---|---|---|---|
| 36 | the student is adapted to the university regime | Regime _adapt | (Yes in the First Quarter, Yes in the Second Quarter, Yes in, No) | (YFS,YSs,NO) |
| 37 | The student has difficulties in the first year | 1Year _Difficult | (Yes, No) | (YES,NO) |
| 38 | Parents help | Parents _Help | (Yes Often, Yes Sometimes, Yes Rarely, Never) | (OFT,SOM,ONL ,NEV) |
| 39 | At home, the student does his homework requested by the teachers of the first year of university | Homework | (Yes Often, Yes Sometimes, Yes Rarely, Never) | (YFT,YSO,YON ,NNE) |
| 40 | The student likes going to university | Universite_Like | (Yes always, I hate the university Yes sometimes,No) | ( YID,JDU,YSO ,NO) |
| 41 | the student revise his courses | Courses _Review | (Yes at Home, Yes at the Library, Yes Others, No) | (YAH,YAL, YOTH,NO) |
| 42 | Relations with teachers | Relation _Teacher | (Very Good, Good, Bad, Very Bad) | (GOO,VGO,BAD ,VBA) |
| 43 | Parents' encouragement | Parents _Encourag | (Yes Often, Yes Sometimes, Yes Rarely , Never) | (YFT,YSO,YON ,NNE) |

| | | | | |
|---|---|---|---|---|
| 44 | The student make discussions at home with his family which encourages him | enc_ava | (Yes Often, Yes Sometimes, Yes Rarely, Never) | (YFT,YSO,YON ,NNE) |

| *N* | *les attributs* | *code attrib* | *valeur attri* | *code et valeur attrib* |
|---|---|---|---|---|
| 45 | Do you attend the first years | Assist | (Yes always, Yes sometimes, Yes, only ,No never) | (YID,YSO,YON ,NNE) |
| 46 | Prepare TDs during the first year | Assist_Td | (Yes always, Yes sometimes ,Yes, only ,No never) | (YID,YSO,YON ,NNE) |
| 47 | Review with colleagues | Review _colleag | (Yes Often, Yes Sometimes, Yes Rarely, Never) | (YID,YSO,YON ,NNE) |
| 48 | Frequency of review with colleagues | Freq_ Review_ Colleag | (Every Day, Several Times a Week, About Once a Week, During Exam Period Only, Never) | (ID,PFS,UFS ,PES,NEV) |
| 49 | Class participation | Class_ Participation | (Yes Often, Yes Sometimes, Yes Rarely, Never) | (YID,YSO,YON ,NNE) |
| 50 | Use of library | Library_Use | (Yes Often, Yes Sometimes, Yes Rarely, Never) | (YID,YSO,YON ,NNE) |
| 51 | Absence during lessons | Absence | (Yes Often, Yes Sometimes, Yes Rarely, Never) | (YID,YSO,YON ,NNE) |
| 52 | Stress during exams | Stress _Exam | (Yes Often, Yes Sometimes, Yes Rarely, Never) | (YID,YSO,YON ,NNE) |
| 53 | The student sanctioned by | Disciplinary _Sanctions | (Yes, No) | (YES,NO) |

| *N* | *les attributs* | *code attrib* | *valeur attri* | *code et valeur attrib* |
|---|---|---|---|---|
| 54 | Results in S1 | S1_Results | (Admis a la session normale,Admis aprés la session rattrapage,Admis avecdettes,Ajourné) | (ASN,AAR,AAD ,AJO) |
| 55 | Results in S2 | S2_Results | (Admis a la session normale,Admis aprés la session rattrapage, Admis avec dettes ,Ajourné) | (ASN,AAR,AAD ,AJO) |
| 56 | Final result | Final_Results | (Admis sans dattes,Admis avec dattes,Ajourné) | (ASD,AAD,AJO) |

# Références Bibliographique

## Références bibliographique

[1] Boulicaut, J.-F., & Crémilleux, B. (2004). Extraction de motifs dans des bases de données

[2] Chen, M.-S., Han, J., & Yu, P. S. (1996). Data Mining: An Overview from a Database Perspective.

[3] Edelstein, H., A. (1999). Introduction to Data Mining and Knowledge Discovery (3rd ed).

[4] Margaret H. Dunham «Data Mining Introductory and Advanced Topics», Prentice Hall, 2003.

[5] ZERF Nadjet « *Gestion des connaissances dans le domaine médical* », Mémoire de magister Université Saad Dahlab Blida, 2010.

[6] https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision consulté le 15/04/2020

[7] https://maximilienandile.github.io consulté le 15/04/2020

[8] https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels consulté le 19/04/2020

[9] Mahdi Miled, 2014, RESSOURCES ET PARCOURS POUR L'APPRENTISSAGE DU LANGAGE PYTHON

[10] https://fr.wikipedia.org/wiki/Partitionnement_de_donn%C3%A9es consulté le 15/05/2020

[11] https://fr.wikipedia.org/wiki/Weka_(informatique) consulté le 10/06/2020

[13] Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G.Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', AAAI/MIT Press, Cambridge, MA.

[14] R. Agrawal; T. Imielinski; A. Swami: *Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference 1993: 207-216*

[2] https://fr.wikipedia.org/wiki/R%C3%A8gle_d%27association consulter le 05/06/2020

[15] http://www.solutionstat.ca/pdf/MBA.pdf consulter le 15/06/2020

[16] http://dspace.univ-msila.dz:8080//xmlui/handle/123456789/2756 consulter le 20/06/2020

[17] https://fr.wikipedia.org/wiki/Algorithme_APriori# consulter le 21/06/2020

[18] https://www.cs.waikato.ac.nz/ml/weka/ consulter le 15/08/2020

[19] https://weka.sourceforge.io/doc.dev/weka/associations/Apriori.html#setNumRules-int- consulter le 21/08/2020