Table des matières :

Chapitre 3 : conception

Chapitre 4 : implémentation et tests

Table des figures :

Les moyens informatiques modernes ont permis de produire et d'archiver d'énormes masses de données numériques depuis maintenant au moins deux décennies. Ces données sont généralement collectées pour rendre un service donné ou répondre à une question précise, Le problème de l'accès à cette connaissance dépasse largement les capacités humaines d'analyse car ces connaissances sont diffusées dans une quantité importante de données souvent complexes. La mise à disposition de grandes quantités de données d'une part et l'impossibilité de les exploiter pleinement d'autre part, ont favorisé dès le début des années 90 l'essor d'une nouvelle discipline scientifique appelée l'ECD ou l'extraction de connaissance à partir des données.

Dans notre mémoire on s'est intéressé à l'extraction de connaissances dans le domaine de l'éducation en se basant sur plusieurs algorithmes du data mining tel que le clustering, les règles d'associations etc.

L'EDM (le data mining en éducation) offre un ensemble d'algorithmes et de techniques qui permettent d'extraire des connaissances à partir des données qui sont utiles pour l'enseignant afin de prendre les bonnes décisions concernant les apprenants.

Notre mémoire est organisé de la manière suivante :

- dans le premier chapitre on a expliqué la notion d'ECD ,son processus et ces différentes étapes ,comme on a parlé du data mining en générale, ses domaines d'utilisation ainsi que les différents outils utilisés .

- dans le deuxième chapitre on s'est basé sur le data mining en éducation ; en parlant sur les différentes méthodes utilisées ainsi que Principale application du modèle EDM

- Le troisième chapitre décrit les différentes étapes de la conception de notre application englobant le module du preprocessing et les différents modules d'utilisation des algorithmes de data mining.

- le quatrième chapitre est consacré à l'implémentation de notre application en utilisant l'API weka suivi de tests et d'interprétation de résultats, on a aussi opté pour l'implémentation de notre travail avec l'outil R et on a enrichi notre mémoire par une étude comparatif entre les deux environnements.

- Enfin nous concluant notre mémoire par une conclusion générale et des perspectives ouvertes pour notre travail.

Introduction

L'extraction de Connaissances (E.C.D.) est une discipline récente, à l'intersection des domaines des bases de données, de l'intelligence artificielle, de la statistique, des interfaces homme / machine et de la visualisation. A partir de données collectées par des experts, il s'agit de proposer des connaissances nouvelles qui enrichissent les interprétations du champ d'application, tout en fournissant des méthodes automatiques qui exploitent cette information.

Dans cette partie, nous allons donner un aperçu général sur le processus ECD (définition, étapes...) notamment sur l'étape fouille de données .., les techniques utilisées (Règle d'association, classification ...).

I. Extraction de connaissance

1. Définition

L'extraction de connaissances est le processus de création de connaissances à partir d'informations structurées comme les bases de données relationnelles, XML ou non structurées comme les textes, documents, images. A pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes mathématiques, automatiques ou semi-automatiques. Dans le but d'extraire un résultat dans un format lisible par les ordinateurs [19]. Il peut aussi se définit comme « l'acquisition de connaissances nouvelles, intelligibles et potentiellement utiles à partir de faits cachés au sein de grandes quantités de données »

L'utilisation industrielle ou opérationnelle de ce savoir dans le monde professionnel permet de résoudre des problèmes très divers, allant de la gestion des risques à la maintenance préventive, en passant par la détection de fraudes ou encore l'optimisation de processus.

Donc il s'agit de proposer des connaissances nouvelles qui enrichissent les interprétations du champ d'application, tout en fournissant des méthodes automatiques qui exploitent cette information.

L'extraction de connaissances se fait à l'aide d'algorithmes de calcul, de visualisation et d'interprétation des résultats, lors d'interactions avec l'expert. Les méthodes d'exploration Proposent des solutions aux problèmes de recherche d'associations, de classification supervisée et non supervisée, composé de plusieurs phases.

2. Processus

le processus d'Extraction de Connaissances à partir de Données est un processus itératif et interactif complexe ayant pour objectif « d'identifier, dans des données, des patterns ultimement compréhensibles, valides, nouveaux et potentiellement utiles » (Fayyad et al., 1996).

Il possède certaines caractéristiques importantes. Tout d'abord, il traite de grands volumes de données. Dans certains domaines d'application, il est d'usage d'avoir à traiter des relations possédant plusieurs millions d'enregistrements. Ensuite, l'information extraite provient des données disponibles et l'utilisateur doit pouvoir affiner ses recherches au fur et à mesure qu'il avance dans le processus. Il est important d'extraire de la connaissance inconnue auparavant.

Enfin, la connaissance extraite doit être exprimée sous une forme compréhensible par l'utilisateur.

La méthodologie générale d'un projet d'extraction de connaissances est illustrée dans la figure suivante adaptée de (Gardarin, 1999) et (Zaïane, 1999).

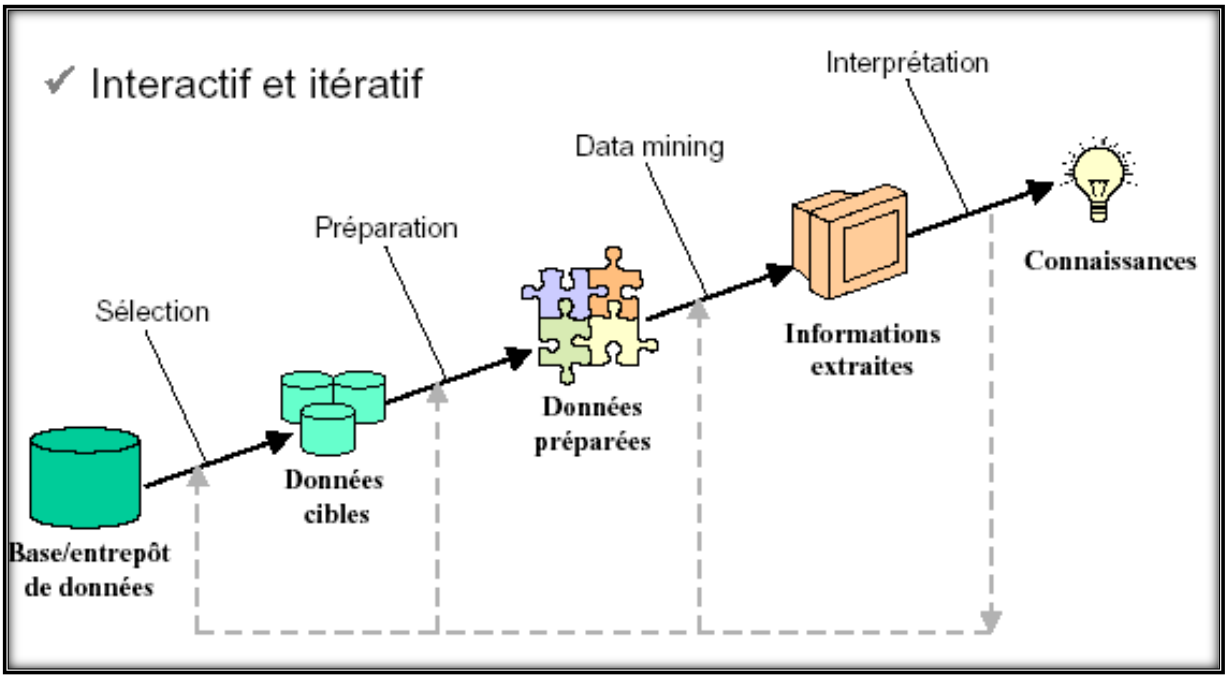


Figure 01: *Processus d'extraction de connaissances [18]*

3. Etapes du processus d'ECD

D'une manière plus concrète, un processus d'extraction de connaissances dans les bases de données peut se décomposer en cinq grandes étapes : [17]

3.1 La consolidation (Définition et compréhension du problème)

Cette première étape permet de rassembler et d'unifier dans un cadre conceptuel commun les différentes données sur le domaine que l'on souhaite étudier. Cette étape inclue également certains traitements nécessaires au nettoyage des données. En effet, un utilisateur du processus est confronté à plusieurs défis. D'une part il doit se familiariser avec les données à analyser, traduire son objectif d'étude en tâche d'ECD et sélectionner les attributs pertinents pour l'étude. Ce qui fait appel aux connaissances sur le domaine analysé. D'autre part il doit être capable de bien choisir, paramétrer, composer et exécuter des outils et méthodes provenant de divers domaines (statistique, intelligence artificielle, bases de données...) afin de réussir son objectif d'analyse. Ce qui fait appel aux connaissances du domaine de l'analyste. A la fin de cette étape, on obtient généralement un entrepôt de données, qui servira de référentiel de source de données pour la suite du processus. [2]

3.2 La sélection (collecte de données)

Dans cette étape, on s'intéresse à la manière dont les données sont générées et collectées. D'après la définition du problème et des objectifs du data mining, on peut avoir une idée sur les données qui doivent être utilisées. Ces données n'ont pas toujours le même format et la même structure. On peut avoir des textes, des bases de données, des pages web, ...etc. Parfois, on est amené à prendre une copie d'un système d'information en cours d'exécution, puis ramasser les données de sources éventuellement hétérogènes (fichiers, bases de données relationnelles, temporelles, ...) [20]. Quelques traitements ne nécessitent qu'une partie des données, on doit alors sélectionner les données adéquates. Généralement les données sont subdivisées en deux parties : une utilisée pour construire un modèle et l'autre pour le tester. On prend par exemple une partie importante (suffisante pour l'analyse) des données (80 %) à partir de laquelle on construit un modèle qui prédit les données futures.
Pour valider ce modèle, on le teste sur la partie restante (20 %) dont on connaît le comportement.

3.3 Prétraitement

Les données collectées doivent être "préparées". Avant tout, elles doivent être nettoyées puisqu'elles peuvent contenir plusieurs types d'anomalies : des données peuvent être omises à cause des erreurs de frappe ou à causes des erreurs dues au système lui-même, dans ce cas il faut remplacer ces données ou éliminer complètement leurs enregistrements. Des données peuvent être incohérentes c.-à-d qui sortent des intervalles permis, on doit les écarter où les normaliser. Parfois on est obligé à faire des transformations sur les données pour unifier leur poids. Un exemple de ces transformations est la normalisation des données qui consiste à la projection des données dans un intervalle bien précis [0,1] ou [0,100] par exemple. Un autre exemple est le lissage des données qui considère les échantillons très proches comme étant le même échantillon. Le prétraitement comporte aussi la réduction des données qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration. Une méthode largement utilisée dans ce contexte, est l'analyse en composantes principales (ACP). [21]

Une autre méthode de réduction est celle de la sélection et suppression des attributs dont l'importance dans la caractérisation des données est faible, en mesurant leurs variances. On peut même réduire le nombre de données utilisées par le data mining en écartant les moins importantes. Dans la majorité des cas, le prétraitement doit préparer des informations globales sur les données pour les étapes qui suivent tel que la tendance centrale des données (moyenne, médiane, mode), le maximum et le minimum, le rang, les quartiles, la variance, ... etc. Plusieurs techniques de visualisation des données telles que les courbes, les diagrammes, les graphes,... etc, peuvent aider à la sélection et le nettoyage des données. Une fois les données collectées, nettoyées et prétraitées on les appelle entrepôt de données (data warehouse).

3.4 La fouille de données (ou Data Mining)

C'est l'étape qui constitue véritablement le cœur du processus d'Extraction de Connaissances dans les données. C'est ici que l'on va chercher à extraire automatiquement des modèles à partir des données. Cette étape est la plus critique du point de vue

algorithmique. En effet, l'espace de recherche est en général très vaste, et c'est ici qu'il faudra utiliser, lorsque cela est possible, les contraintes spécifiées par l'utilisateur lors de l'extraction. Cela permet de réduire l'espace de recherche à parcourir et donc de faire en sorte que l'algorithme s'exécute dans des temps raisonnables. De plus, un autre aspect important à prendre en compte est la manière dont l'algorithme va lire les données en entrée. Si le jeu de données est volumineux, faire plusieurs passes sur les données peut être coûteux en temps. [21]

Des techniques telles que les réseaux de neurones, les arbres de décision, les réseaux bayésiens et clustering sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat.

3.5 L'interprétation et l'évaluation

À cette étape, il faut regarder si les modèles et motifs retournés par la phase de fouille de données correspondent à une connaissance nouvelle pertinente pour le domaine considéré. Dans le cas de l'extraction d'ensembles d'items ou de motifs séquentiels, il est courant de récupérer plusieurs milliers de motifs dont seuls quelques-uns se révéler ont véritablement intéressants, ce qui implique une phase d'analyse assez longue. Dans le cas des clusters ou des arbres de décisions, les modèles renvoyés étant de taille plus réduite, leur analyse peut en sembler simplifiée, mais leur interprétation et leur validation restent très difficiles. [2]

4. Avantages de l'exploration de données

- ✓ Prédiction automatisée des tendances et des comportements
- ✓ Il peut être implémenté sur de nouveaux systèmes ainsi que sur des plates-formes existantes
- ✓ Il peut analyser une énorme base de données en quelques minutes
- ✓ Découverte automatisée des modèles cachés
- ✓ Il y a beaucoup de modèles disponibles pour comprendre facilement des données complexes
- ✓ Il est très rapide et permet aux utilisateurs d'analyser d'énormes quantités de données en moins de temps.
- ✓ Il produit des prévisions améliorées

II. Data mining

Après avoir expliqué la notion d'extraction de connaissance et son processus d'exécution, nous allons passer à présenter la définition du data mining et ces techniques utilisées dans différents domaines.

1. Définition

Plusieurs définitions ont été proposées pour le data mining notamment :

Le data mining est l'analyse de grandes ensembles de données observationnelles pour découvrir des nouvelles relations entre elles et de les reformuler afin de les rendre plus utilisables de la part des propriétaires. [1]

Le data mining est un domaine interdisciplinaire utilisant dans le même temps des techniques d'apprentissage automatiques, de reconnaissance des formes, des statistiques, des bases de données, et de visualisation pour déterminer les manières d'extraction des informations de très grandes bases de données.

Data Mining (fouille de données) se définit comme un processus analytique destiné a explorer de large quantité de données dans différents domaines, afin de dégager une certaine structure et/ou des relations systématiques entre variables, puis en validant les conclusions et appliquant les structures trouvées à de nouveaux groupes de données.

Généralement on s'accorde à définir le data mining comme étant l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de bases de données informatiques, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données »,

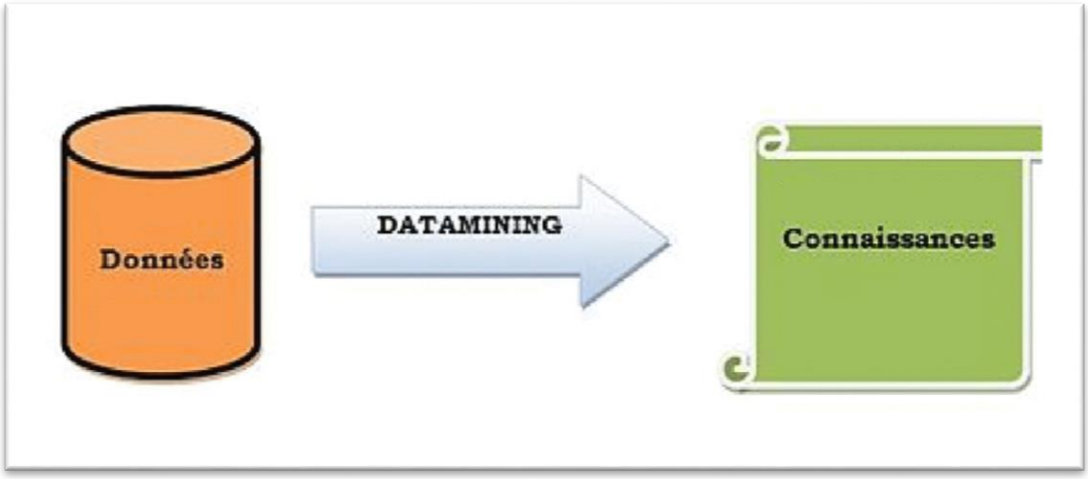


Figure 02 : *Data mining vers la connaissance*

2. Data Mining versus KDD (Knowledge Discovery in Databses)

Habituellement les deux termes sont inter changés.

KDD (Knowledge Discovery in Databases) : C'est le processus de trouver information et/ou partons utiles à partir de données.

Data Mining : C'est l'utilisation des algorithmes pour extraire information et/ou partons comme partie du processus KDD,c'est le cœur du processus d'extraction de connaissances.

3. Domaines d'application

Le Data mining est une approche d'analyse de données, adaptée et utilisée dans un large nombre de domaines d'activités.

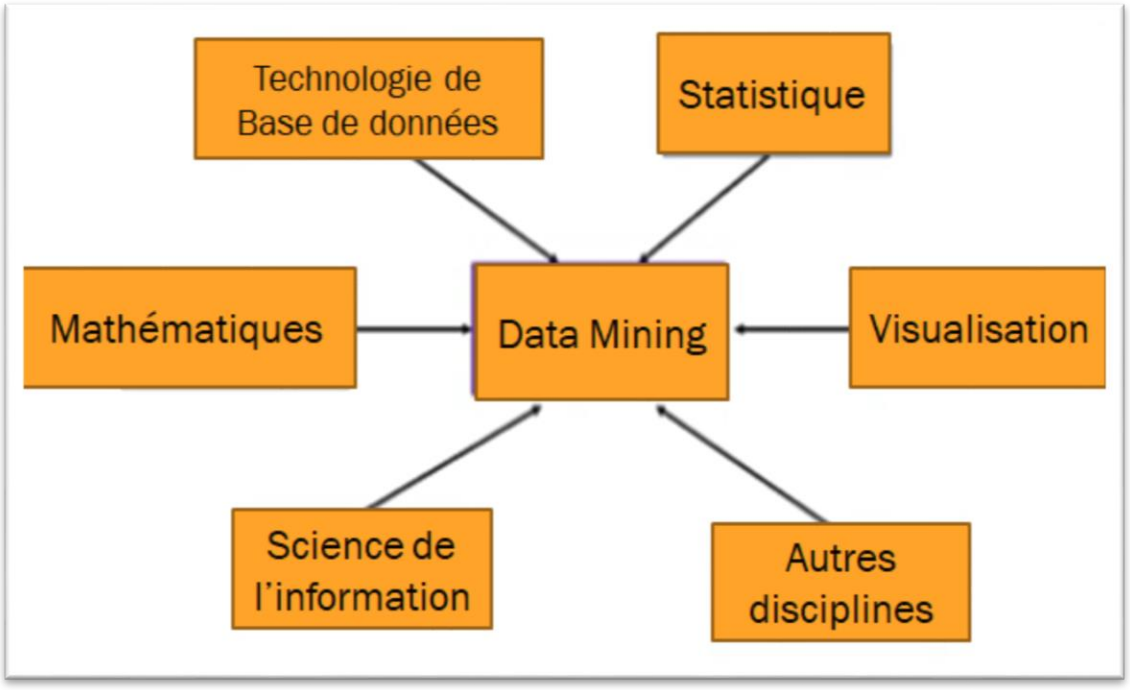


Figure 03: *Domaines d'application de data mining [18]*

La technologie de data mining a une grande importance économique grâce aux possibilités qu'elle offre pour optimiser la gestion des ressources (humaines et matérielles).

Les domaines d'application actuels du data mining sont les suivants :

- **Le data mining dans le secteur bancaire :** Le secteur bancaire est à la tête de tous les autres domaines industriels grâce à son utilisation des techniques du Data Mining dans ses grandes bases de données clients. Bien que les banques ont employées des outils d'analyse statistiques avec un peu de succès pendant plusieurs années, les modèles précédemment invisibles des comportements des clients deviennent maintenant plus clair à l'aide des nouveaux outils du data mining.

Quelques applications du data mining dans ce domaine sont :

- Prédire la réaction des clients aux changements des taux d'intérêt.
- Identifier les clients qui seront les plus réceptifs aux nouvelles offres de produits.
- Identifier les clients "fidèles".
- Déterminer les clients qui posent le risque le plus élevé de manquer à leurs engagements aux prêts.
- Détecter les activités frauduleuses dans les transactions par cartes de crédit.

- Prédire les clients qui sont susceptibles de changer leurs cartes d'affiliation au cours du prochain trimestre.

- Déterminez les préférences des clients pour les différents modes de transaction à savoir par le biais de guichets ou par l'intermédiaire de cartes de crédit, etc.

- **Le data mining dans la bio-informatique et la biotechnologie :** La Bio-informatique est un domaine de recherche en développement rapide, qui a des racines aussi bien dans la biologie que dans la technologie d'informations.

Quelques applications du data mining dans ce domaine sont :

- la prédiction des structures de différentes protéines.
- la détermination de la complexité des structures de plusieurs médicaments.

- **Le data mining dans le marketing direct et le collecte de fonds :** Le marketing direct est un ensemble de techniques permettant d'identifier les consommateurs (particuliers et entreprises) d'un produit stockés dans une base de données, de leur adresser directement et individuellement une proposition commerciale, afin d'obtenir une réponse directe, à laquelle l'entreprise répondra tout aussi directement. Le Marketing Direct est le premier domaine qui a employé les outils du data mining pour son profit.
- ***Le data mining dans la détection de fraude :*** La fouille de données est largement appliquée dans des processus de détection de fraude divers tel que :

- Détection de fraude de cartes de crédits.

- Détection de fraude dans les listes des électeurs en utilisant les réseaux de neurones en combinaison avec le data mining.

- La détection des fraudes dans les demandes de passeport par la conception d'un système de diagnostic par apprentissage en ligne.

- Détection de fausses demandes de remboursement médicale.

- **Le data mining dans la gestion de données scientifiques :** Quelques exemples de la fouille de données dans l'environnement scientifique sont :

- Les études sur les changements climatiques du globe : Il s'agit d'un domaine de recherche *chaud* et essentiellement un exercice de vérification axée sur l'exploitation. À travers les données climatiques qui ont été recueillies au fil des siècles et qui sont en train d'être étendues dans le passé lointain et, en même temps, à travers des activités telles que l'analyse des échantillons de carottes de glace de l'Antarctique, des différents modèles de prédiction ont été proposées pour les futures conditions climatiques.

- Les études sur les bases de données de géophysique à l'Université d'Helsinki : Ils ont publié une analyse scientifique des données sur l'agriculture et l'environnement. En conséquence, ils ont optimisé le rendement des cultures, tout en réduisant au minimum les ressources fournies. Afin de réduire au minimum les ressources, ils ont identifié les facteurs qui influent sur le rendement des cultures, comme les engrais chimiques et les additifs (phosphate), Le contenu d'humidité et le type du sol.

- **Le data mining dans le secteur des assurances :** Les compagnies d'assurance peuvent bénéficier des méthodes du data mining, qui aident les entreprises à réduire les coûts, augmenter les profits, de conserver les clients actuels, d'acquérir de nouveaux clients, et développer de nouveaux produits.

Cela peut être fait par le biais de :

- Evaluation du risque d'un bien assuré prenant en compte les caractéristiques du bien et de son propriétaire.
- Formulation des modèles statistiques des risques d'assurance.
- Utilisation du modèle de l'exploitation de Poisson / Log-normale afin d'optimiser les polices d'assurance.

- **Le data mining dans la télécommunication :** À nos jours, toute activité de télécommunication a utilisé une la technique de data mining.
- Analyse des achats de services de télécommunications.
- Prédiction de modèles d'appels téléphoniques.
- Gestion des ressources et de trafic réseau.
- Automatisation de la gestion du réseau et de la maintenance en utilisant

l'intelligence artificielle pour diagnostiquer et réparer les problèmes de transmission du réseau, etc.

- **Le data mining dans la médecine et la pharmacie :** Quelques exemples de l'usage médicaux et pharmaceutiques des techniques de Data Mining pour l'analyse de bases de données médicales.
- Prédiction de présence de maladies et/ou de complications.
- Le choix d'un traitement pour le cancer.
- Choix des antibiotiques pour des infections.
- Le choix d'une technique particulière (de sutures, matériel de suture, etc.) dans une des procédures chirurgicales.
- Approvisionnement des médicaments les plus fréquemment prescrits

- **Le data mining dans le commerce au détail :** Les techniques du data mining ont été très utiles pour le CRM (Customer Relation ship Marketing) en développant des modèles pour :
- La prédiction de la propension du client à acheter.
- L'évaluation des risques pour chaque transaction.
- Connaitre la distribution et l'emplacement géographique des clients.
- L'analyse de la fidélité des clients dans les opérations à base de crédit.
- Évaluation de la menace concurrentielle dans une région.

- **Le data mining dans le e-commerce et le World Wide Web :** Quelques façons d'utilisation des outils du data mining dans le e-commerce sont :

- En formulant des tactiques du marché dans les opérations de business.
- En automatisant des interactions d'affaires avec des clients, pour que les clients puissent traiter avec tous les acteurs dans la chaîne d'approvisionnement.

La détermination de la taille d world Wide Web est extrêmement difficile. En 1999,elle a été estimée à 350 milliards de pages avec un taux de croissance de 1 million de pages /jour. En considérant le World Wide Web comme la plus grande collection de bases dedonnées, le Web mining peut être fait par les façons susdites

- ***Le data mining dans le marché boursier et l'investissement :*** L'évolution rapide de la technologie informatique au cours des dernières décennies a facilité l'investissement par des professionnels (et des amateurs), avec la possibilité d'accéder et d'analyser d'énormes quantités de données financières. Les outils du data mining sont utilisés pour :

- Aider les spécialistes du marché boursier à prédire les mouvements du prix des actions.
- Le data mining des anciens des prix et des variables liées aide à découvrir des anomalies de marché boursier comme le scandale hawala.

- ***Le data mining dans l'analyse de chaîne d'approvisionnement***

Les techniques du data mining ont trouvé une large application dans l'analyse des chaînes d'approvisionnements. Des exemples d'utilisation du data mining par les fournisseurs sont :

- Analyser le processus des données pour gérer l'évaluation d'acheteur.
- L'extraction des données de paiement avec l'utilité de mettre à jour la politique tarifaire.

4. Méthodes du data mining

Pour arriver à exploiter les quantités importantes de données, le data mining utilise des méthodes d'apprentissages automatiques. Une amalgame est faite à tord entre toutes ces méthodes. Ces méthodes sont de deux types : les méthodes descriptives et les méthodes prédictives, selon qu'il existe ou non une variable "cible" que l'on cherche à expliquer.

4.1 Méthodes descriptive

Le principe de ces méthodes est de pouvoir mettre en évidence les informations présentes dans le data warehouse mais qui sont masquées par la masse de donnée.

Parmi les techniques et algorithmes utilisés dans l'analyse descriptive, on cite :

- Analyse factorielle (ACP et ACM)
- Méthode des centres mobiles
- Classification hiérarchique
- Classification neuronale (réseau de Kohonen)
- Recherche d'association [22]

4.2 Méthodes prédictive

Contrairement à l'analyse descriptive, cette technique fait appels à de l'intelligence artificielle. L'analyse prédictive, est comme son nom l'indique une technique qui va essayer de prévoir une évolution des événements en se basant sur l'exploitation de données stockés dans le data warehouse. [22]

En effet, l'observation et l'historisation des événements peuvent permettre de prédire une suite logique. Le meilleur exemple est celui des prévisions météorologiques qui se base sur des études des évolutions météorologiques passées. En marketing, l'objectif est par exemple de déterminer les profils d'individus présentant une probabilité importante d'achat ou encore de prévoir à partir de quel moment un client deviendra infidèle.

- Parmi les techniques et algorithmes utilisés dans l'analyse prédictive, on cite : arbre de décision, réseaux de neurones, régression linéaire, analyse discriminante de Fishere et Analyse probabiliste

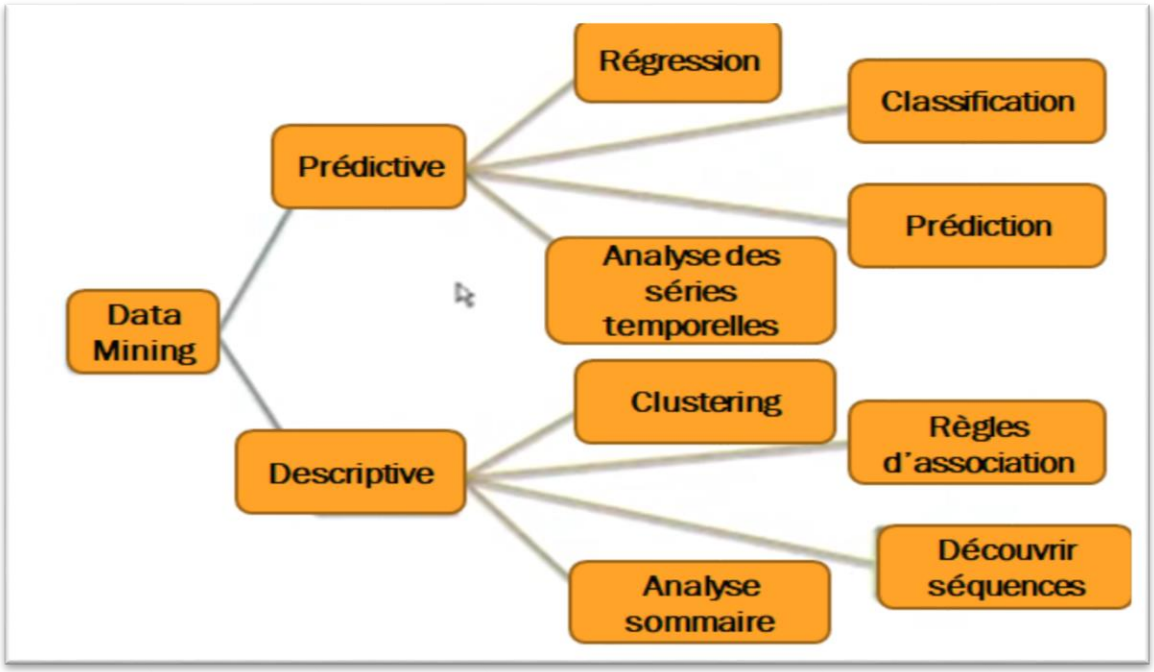


figure 04: *les méthodes de data mining [23]*

5. Tâches réalisées en data mining

Les tâches les plus courantes que le data mining est amené à accomplir sont : la description, l'estimation, la prévision, la classification, le clustering, l'association.

- **La description :** C'est souvent l'une des premières tâches demandées à un outil de data mining. On lui demande de décrire les données d'une base complexe. Cela engendre souvent une exploitation supplémentaire en vue de fournir des explications

Parfois, les chercheurs et les analystes essaient simplement de trouver des façons de décrire des tendances cachées dans les données. Les descriptions des modèles et des tendances servent

à expliquer ou vérifier un fait. Par exemple : « ceux qui ont le plus de diplômes sont les plus susceptibles d'avoir un poste à responsabilité. ». [22]

- **L'estimation :** L'estimation est similaire à la classification, sauf que la variable cible est numérique plutôt que catégorique. Les modèles sont construits en utilisant des données, qui fournissent la valeur de la variable cible, ainsi que les « prédicteurs ». Par exemple : « l'estimation de la pression artérielle d'un patient d'hôpital, basée sur son âge, son sexe, son indice de masse corporelle, et le taux de sodium. La relation entre la pression artérielle et le prédicteur variable de l'ensemble de formation nous donnerait un modèle d'estimation. Nous pouvons alors appliquer ce modèle à de nouveaux cas. [22]

- **La prédiction :** La prédiction est semblable à la classification et l'estimation, sauf que pour la prévision, les résultats se situent dans l'avenir. Exemples de tâches de prévision appliquée au marketing : « Prédire le prix d'un stock de trois mois dans le futur » [22]

- **La classification :** La classification consiste à examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini. Supposons qu'un décideur veuille classer ses employés par tranches de revenu, ou n'importe quelle autre caractéristique associée à cette personne, comme l'âge, le sexe et la profession. Cette tâche est une tâche de classification. [22]

- **Le clustering :** Le Clustering désigne le regroupement des données, des observations ou des cas dans des classes d'objets similaires. Un cluster maximise la similarité des objets de du même cluster et minimise la similarité des objets de cluster différents. En effet, il n'y a pas de variable cible pour le clustering. La tâche de clustering ne cherche pas à classer, estimer, ou prédire la valeur d'une variable cible. Mais plutôt à segmenter l'ensembledes données en sous-groupes relativement homogènes à l'aide de mesures de distances. [23]

- **L'association :** La recherche de règles d'association est la tâche la plus intéressante du data mining. C'est également celle qui est la plus répandue dans le monde des affaires, notamment en marketing pour l'analyse du panier de consommation. La recherche de règles d'association cherche à découvrir les règles de quantification ou de relation entre deux ou plusieurs attributs. Les règles d'association sont de la forme «Si antécédent, puis conséquente », avec une mesure confiance associée à la règle. La recherche de règles d'associations dans une grande base de données permet de découvrir des règles cachées utiles pour la prise de décision.

Exemple de règle célèbre : lorsqu'un homme achète des couches pour bébés, il achète 2 packs de bières dans 65% des cas. Il serait alors intéressant pour le gestionnaire d'adapter ses promotions à ces nouvelles règles. [23]

6. Techniques de data mining

Les techniques du Data Mining représentent une partie très importante dans la tâche de l'apprentissage machine (machine learning).

Différentes techniques sont proposées. Elles sont à choisir en fonction de la nature des données et du type d'étude que l'on souhaite entreprendre.
Chacune des taches regroupe une multitude d'algorithmes pour construire le modèle auquel elle est associée. Ces techniques se divisent en connaissances dirigées ou Techniques supervisé et connaissances non dirigées ou Techniques non supervisé.

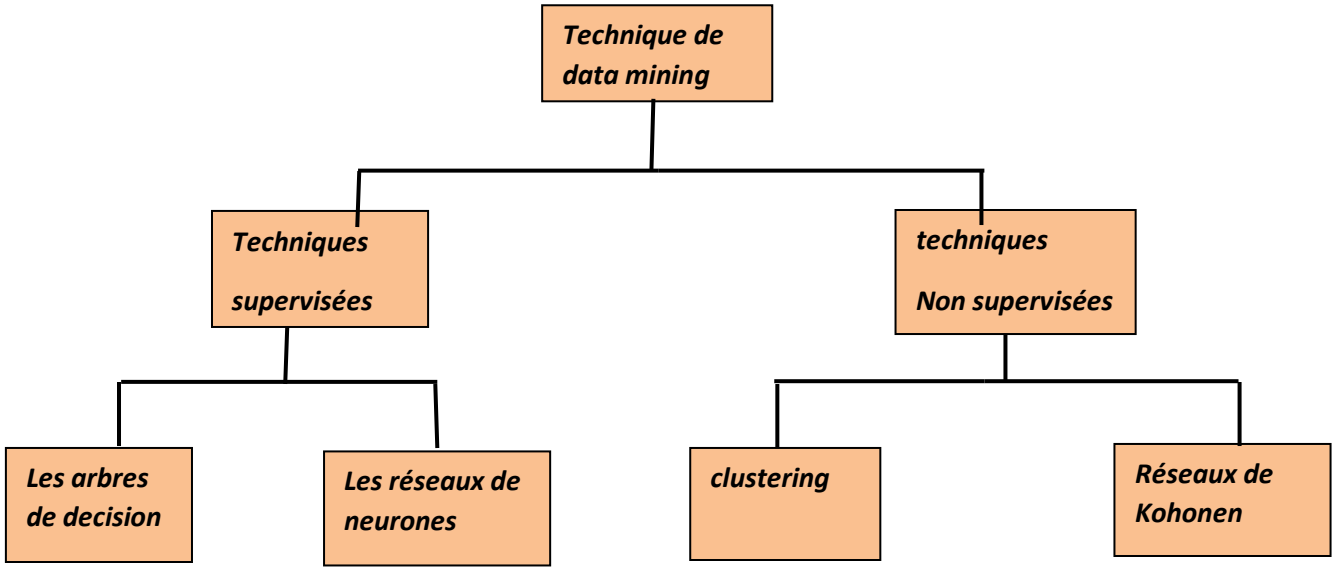


Figure 05: *Les techniques de Data mining[24]*

6.1 Techniques supervisées

Dans la modélisation supervisée, ou prédictive, l'objectif est de prédire un événement ou d'estimer les valeurs d'un attribut numérique continue. Dans ces modèles, il existe des champs où les attributs d'entrée et une zone de sortie ou de la cible. Les champs d'entrée sont également appelés prédicteurs, car ils sont utilisés par le modèle pour identifier une fonction de prédiction de champ de sortie. Nous pouvons penser à des prédicteurs que la partie X de la fonction et le domaine cible que la partie Y, le résultat. Le modèle utilise les champs de saisie qui sont analysées en ce qui concerne leur effet sur le champ cible. La reconnaissance de formes est "surveillé" par le domaine cible. Des relations sont établies entre les champs d'entrée et de sortie. Une cartographie " fonction d'entrée-sortie " est généré par le modèle, qui associe des prédicteurs et à la sortie permet la prédiction des valeurs de sortie, étant donné les valeurs des champs d'entrée. [23] Les modèles prédictifs sont subdivisés en modèles de classification et d'estimation :

- **Classification**

Dans ces modèles les groupes ou classes cibles sont connus dès le départ. Le but est de classer les cas dans ces groupes prédéfinis ; en d'autres termes, à prévoir un événement. Le modèle généré peut être utilisé comme un moteur de marquage pour l'affectation de nouveaux cas pour les classes prédéfinies. Il estime aussi un score de propension pour chaque cas. Le score de propension dénote la probabilité d'occurrence du groupe cible ou d'un événement.

Les techniques les plus appropriées à la classification sont :

- **les arbres de décision** : Les arbres de décision fonctionnent en séparant de façon récursive la population initiale. Pour chaque groupe, ils sélectionnent automatiquement l'indicateur le plus significatif, le prédicteur qui donne la meilleure séparation par rapport au champ cible. À travers des cloisons successives, leur objectif est de produire sous-segments pures, avec un comportement homogène en termes de production. Ils sont peut-être la technique la plus populaire de classification. Une partie de leur popularité, c'est parce qu'ils produisent des résultats transparents qui sont facilement interprétables, offrant un aperçu de l'événement à l'étude. Les résultats obtenus peuvent avoir deux formats équivalents. Dans un format de règle, les résultats sont représentés dans un langage simple que les règles ordinaires : SI (VALEURS PREDICTIVES) ALORS (RESULTAT CIBLE ET SCORE DE CONFIANCE). Dans une forme d'arborescence, les règles sont représentées graphiquement sous forme d'arbre dans laquelle la population initiale (nœud racine) est successivement divisée en des nœuds terminaux ou feuilles de sous-segments ayant un comportement similaire en ce qui concerne le champ. [7]
Les algorithmes d'arbres de décision constituent selon la vitesse et l'évolutivité. Algorithmes disponibles sont : - C5.0 - CHAID - Classification et arbres de régression - QUEST.

- **Estimation**

Ces modèles sont similaires à des modèles de classification, mais avec une différence majeure. Ils sont utilisés pour prédire la valeur d'un champ continu en fonction des valeurs observées des attributs d'entrée. [7]
La technique la plus appropriée à l'estimation est :

- **les réseaux de neurones** : Les réseaux de neurones sont des puissants algorithmes d'apprentissage automatique qui utilisent des fonctions de cartographie complexe, non linéaire pour l'estimation et classification. Ils sont constitués de neurones organisés en couches. La couche d'entrée contient les prédicteurs ou neurones d'entrée. La couche de sortie comprend dans le champ cible. Ces modèles permettent d'estimer des poids qui relient les prédicteurs (couche d'entrée à la sortie). Modèles avec des topologies plus complexes peuvent également inclure, couches cachées intermédiaires, et les neurones. La procédure de formation est un processus itératif.

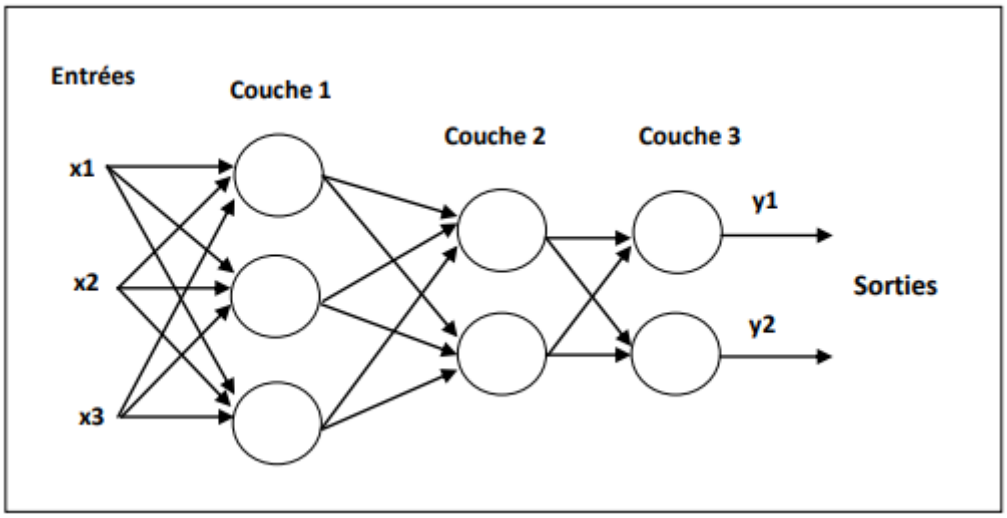


Figure 6 : *réseau de neurones artificiel*

- **La prédiction**

Ressemble à la classification et à l'estimation mais dans une échelle temporelle différente [7]. Tout comme les tâches précédentes, elle s'appuie sur le passé et le présent mais son résultat se situe dans un futur généralement précisé. La seule méthode pour mesurer la qualité de la prédiction est d'attendre !

Les techniques les plus appropriées à la prédiction sont :

- Les arbres de décision
- les réseaux de neurones

6.2 Techniques non supervisées

Dans les modèles non supervisés ou non orientés, il n'y a pas de champ de sortie, il n'y a que des entrées [5]. La reconnaissance de formes est non orienté ; elle n'est pas guidée par un attribut cible spécifique. Le but de ces modèles est de découvrir des motifs de données dans l'ensemble des champs d'entrée. Les modèles non supervisés comprennent :

• Les modèles de dispersion : Dans ces modèles les groupes ne sont pas connus à l'avance. Au contraire, nous voulons que les algorithmes pour analyser les schémas de données d'entrée et d'identifier les regroupements naturels de données ou de cas. Lorsque de nouveaux cas sont marqués par le modèle de cluster généré ils sont affectés à l'un des groupes révélés.

• Les modèles d'association de séquences (Le regroupement par similitude) :Ces modèles font également partie de la classe de la modélisation non supervisé. Ils ne comportent pas de prédiction directe d'un seul champ. En fait, tous les champs concernés ont un double rôle, car ils agissent comme des entrées et des sorties en même temps. Des modèles d'association de détecter des associations entre des événements discrets, des produits ou des attributs. Les modèles de séquence détectent des associations au fil du temps.

➢ Clustring

Clustring hiérarchique il est considéré comme la "mère" de tous les modèles de clustering[18]. Il est appelé hiérarchique ou d'agglomération, car il commence avec une solution où chaque enregistrement comprend un groupe et peu à peu les groupes se former jusqu'au point où tous tomber dans un super-cluster. À chaque étape, il calcule les distances entre toutes les paires d'enregistrements et les groupes les plus similaires. Une table (horaire d'agglomération) ou un graphique (dendrogramme) résume les étapes de regroupement et les distances respectives. L'analyste doit consulter ces informations, identifier le point où l'algorithme commence à cas disjoints de groupe, et de décider ensuite sur le nombre de grappes à conserver. Cet algorithme ne peut pas traiter efficacement plus de quelques milliers de cas. Ainsi, il ne peut pas être directement appliqué dans la plupart des tâches de regroupement d'entreprise. Une solution habituelle consiste à une utilisation sur un échantillon de la population de clustering. Cependant, de nombreux autres algorithmes efficaces qui peuvent facilement gérer des millions d'enregistrements, le regroupement par échantillonnage n'est pas considéré comme une approche idéale.

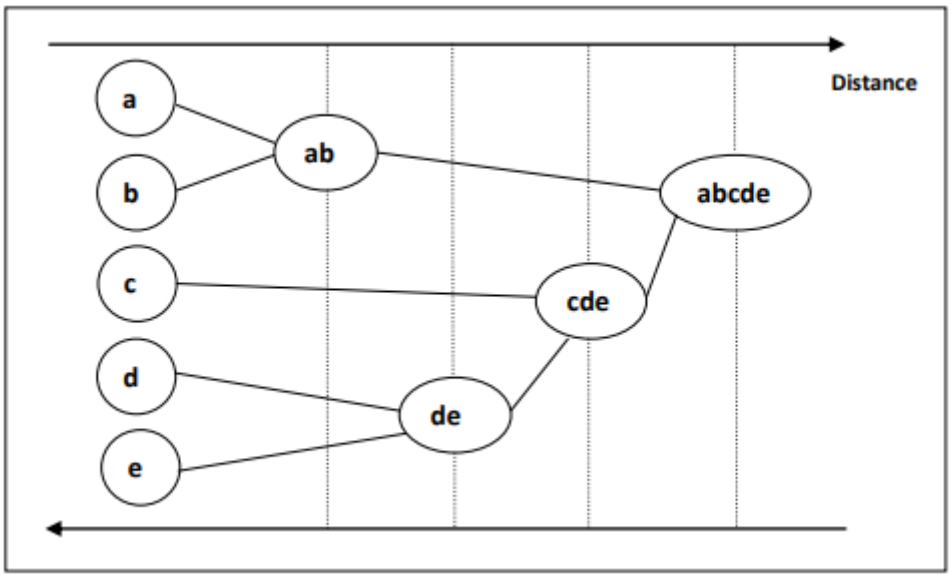


Figure 06 : *Clustering hiérarchiques*

- K-means : C'est un moyen efficace et peut-être l'algorithme de segmentation le plus rapide qui peut gérer deux longues (plusieurs enregistrements) et des ensembles de données larges (de nombreuses dimensions de données et des champs d'entrée) [9]. Il s'agit d'une technique de segmentation basé sur la distance et, à la différence de l'algorithme hiérarchique, il n'a pas besoin de calculer les distances entre toutes les paires d'enregistrements. Le nombre de grappes d'être formés et est prédéterminée spécifiée par l'utilisateur à l'avance. Habituellement, un certain nombre de solutions différentes doit être jugé et évalué avant d'approuver le plus approprié.

- Carte auto-organisatrice de Kohonen : [9] Réseaux de Kohonen sont basés sur des réseaux de neuronaux et produisent typiquement une grille à deux dimensions ou une

carte des grappes, où les cartes d'auto-organisation. Réseaux de Kohonen prennent généralement plus de temps à former que les K-means, mais ils fournissent un point de vue différent sur le regroupement qui est la peine d'essayer.

7. Outils de data mining

Figure 7: *Outils de data mining*

Les systèmes d'information actuels regorgent de données de tout type et la difficulté réside maintenant dans l'exploitation et l'interprétation de ces données pour en extraire des informations, puis des connaissances [11]. De nombreux outils de fouille de données ont été développés pour atteindre cet objectif.

Nous présentons ci- dessus une liste des outils Data mining open source :

- **Sipina :** C'est est un logiciel gratuit de Data Mining spécialisé dans l'induction des arbres de décision. C'est un des très rares outils en libre accès intégrant des fonctionnalités interactives lors de la construction d'un arbre de décision. La version actuelle n'a guère évolué depuis 2000. Elle est néanmoins distribuée car il y a très peu d'équivalents gratuits au monde. Configuré judicieusement, SIPINA peut traiter de très gros volumes (plusieurs millions d'observations et Traitement des très grands fichiers) tout en conservant ses fonctionnalités interactives. Le logiciel reste en anglais, mais les mots clés sont relativement simples à appréhender.

- **R-project :** R est un langage et une infrastructure spécialisés pour les traitements statistiques. R est l'un des nombreux projet GNU distribué sous licence GPL (logiciel libre). R est écrit en langage compilé (principalement en C), ce qui autorise de bonnes performances. La qualité de cet environnement et son ouverture ont permis à une myriade de théoriciens, statisticiens et informaticiens de compléter cette plate-forme d'un nombre impressionnant de fonctionnalités. Des dizaines de packages offrant des milliers de fonctions en font probablement la plate-forme la plus complète. Ce n'est cependant pas l'outil le plus simple d'abord.

- **Scilab et Mixmod :** Scilab est un langage et une infrastructure spécialisés pour les traitements mathématiques numériques et la modélisation.. Sa licence autorise une utilisation gratuite ainsi que la modification des sources. Scilab supporte un spectre très large d'applications, et de nombreuses contributions sont opérationnelles sur Cette plate-forme.
  Mixmod propose des fonctionnalités de clustering (analyse discriminante et maximum de vraisemblance). Mixmod est relativement simple d'utilisation et s'avère adapté pour un volume raisonnable de données.

- **Autoclass-c :** Logiciel spécialisé dans le clustering (analyse discriminante et maximum de vraisemblance). Développé par un laboratoire de la NASA et disponible dans le domaine public. Outil performant écrit en C qui n'a plus évolué depuis le milieu 2002.

- **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. L'objectif principal du projet tanagra est d'offrir aux chercheurs et aux étudiants une plate-forme de data mining facile d'accès, respectant les standards des logiciels du domaine, notamment en matière d'interface et de mode de fonctionnement, et permettant de mener des études sur des données réelles et/ou synthétiques.

- **Weka :** Weka (Waikato Environment for KnowledgeAnalysis) est un ensemble d'outils permettant de manipuler et d'analyser des fichiers de données, implémentant la plupart des algorithmes d'intelligence artificielle, dont les arbres de décision et les réseaux de neurones. Il est écrit en java, disponible sur le web. Weka contient des outils pour les données de prétraitement, la classification, la régression, le clustering, règles d'association, et la visualisation. Il est aussi bien adapté pour le développement de nouveaux programmes d'apprentissage de la machine.

- **Alceste :** Il s'agit d'un logiciel d'analyse de données textuelles, ou statistique textuelle. Alceste, à partir d'un corpus, effectue une première analyse détaillée de son vocabulaire, et constitue le dictionnaire des mots ainsi que de leur racine, avec leur fréquence. Ensuite, par fractionnements successifs, il découpe le texte en segments homogènes contenant un nombre suffisant de mots, et procède alors à une classification de ces segments en repérant les oppositions les plus fortes. Cette méthode permet d'extraire des classes de sens, constituées par les mots et les phrases les plus significatifs, les classes obtenues représentent les idées et les thèmes dominants du corpus. L'ensemble des résultats triés selon leur pertinence, accompagnés de nombreuses représentations graphiques et de différents rapports d'analyse, permet à l'utilisateur une interprétation aisée et efficace.

- **RapidMiner :** RapidMinerest l'un des outils de data mining les plus populaires. Il est accessible gratuitement et facile à utiliser, même sans connaissances particulières en programmation. RapidMiner a été écrit en Java et contient plus de 500opérateurs avec des approches différentes pour démontrer les connexions dans les données - entre autres, il y a des options pour l'exploration de données, l'exploration de texte et l'exploration Web, mais aussi pour l'analyse d'humeur (Analyse du sentiment, Opinion Mining). Le programme importe également des tableaux Excel, des fichiers SPSS et des ensembles de données à partir de nombreuses bases de données et intègre également les outils d'exploration de données WEKA et R. La force particulière de RapidMiner réside dans l'analyse prédictive, c'est-à-dire la prédiction des développements futurs à partir des données collectées. En comparant les logiciels de data mining, RapidMiner est l'un des outils les plus puissants.

- **Knime :** Bien que KNIME fût à l'origine destiné à un usage commercial, il est toujours disponible en tant que logiciel open source. Il a été écrit en Java et édité avec Eclipse. Si l'on regarde ce logiciel de data mining en comparaison avec d'autres, on remarque tout d'abord son périmètre fonctionnel : avec plus de 1 000 modules et des applications prêtes à l'emploi, cet outil permet de découvrir les structures de données cachées. Les modules peuvent être complétés par d'autres fonctions commerciales. Parmi les fonctions, l'analyse intégrative des données est particulièrement convaincante : KNIME est l'un des outils les plus puissants dans ce domaine et permet l'intégration de nombreuses méthodes d'apprentissage machine et de data mining. Il est également particulièrement efficace dans le prétraitement des données, c'est-à-dire l'extraction, la transformation et le chargement des données. Son pipeline modulaire en fait un outil d'exploration de données orienté flux de données.

- **Sas :** (Statistical Analysis System) est le principal outil de data mining pour l'analyse d'entreprise - et aussi le plus cher des programmes listés ici. Cependant, c'est celui qui convient le mieux aux grandes entreprises. SAS se distingue particulièrement bien dans le domaine du pronostic et de la visualisation interactive des données, ce qui est idéal pour les grandes présentations. Cependant, ce logiciel ne peut être utilisé gratuitement que si vous obtenez une licence correspondante d'un établissement public. En principe, SAS est toujours soumis à une redevance. Les coûts sont réglés sur demande, des conditions spéciales, par exemple pour les autorités ou les établissements d'enseignement sont possibles. SAS est principalement utilisé dans les entreprises pharmaceutiques où il s'est imposé comme le standard. Il est également fréquemment utilisé dans le secteur bancaire et offre des solutions optimales pour la BI et le web mining. L'outil dispose notamment de son propre logiciel de Business Intelligence. Cela en fait l'un des outils de data mining les plus puissants du marché.

- **Orange :** Orange est un logiciel complet d'exploration de données qui montre tout ce que vous pouvez faire avec Python : il offre des applications utiles pour l'analyse de données et de textes ainsi que des fonctionnalités pour l'apprentissage machine et dans le domaine du data mining. Il travaille avec des opérateurs pour la classification, la

régression, le clustering et bien plus encore. Cet outil de data mining intègre également la programmation visuelle. Un autre avantage pour les nouveaux arrivants est qu'il y a de nombreux tutoriels en ligne disponibles pour l'outil. Une autre particularité d'Orange est de connaître les préférences de ses utilisateurs dans le temps et de se comporter en conséquence. Cela peut rendre l'utilisation de l'outil de data mining encore plus pratique

.

8. Type de données fouillées par Data mining

Le Data Mining n'est pas spécifique à un type de médias ou de données [22]. Il est applicable à n'importe quel type d'information. Le Data Mining est utilisé et étudié pour les Bases de Données incluant les Bases de Données relationnelles et les Bases de Données Orientées-Objets, les data warehouses, les Bases de Données transactionnelles, les supports de données non structurés et semi-structurés comme le World Wide Web, les Bases de Données avancés comme les Bases de Données spatiales, les Bases de Données multimédia, les Bases de données de séries temporelles et les Bases de Données textuelles et même fichiers plats.

- **Les fichiers plats :** Les fichiers plats sont actuellement la source de données la plus commune pour les algorithmes du Data Mining et particulièrement dans le niveau de recherches [12]. Les fichiers plats sont des fichiers de données simples dans le format texte ou binaire avec une structure connue par l'algorithme du Data Mining.

- **Les bases de données relationnelles :** Les algorithmes du Data Mining appliqués sur des Bases de Données relationnelles sont plus polyvalents que les algorithmes spécifiquement faits pour les fichiers plats puisqu'ils peuvent profiter de la structure inhérente aux bases de données relationnelles. Le Data Mining peut profiter du SQL pour la sélection, la transformation et la consolidation, il passe au-delà de ce que le SQL pourrait fournir, comme la prévision, la comparaison, la détection des déviations, etc.

- **Les data warehouses** : Un Data Warehouse est un support de données rassemblées de multiples sources de données (souvent hétérogènes) et est destinée à être utilisé dans l'ensemble sous le même schéma unifié. Supposons une entreprise Our Video Store qui a des contrats d'exclusivité dans toute l'Amérique du Nord. La plupart des magasins de vidéos appartenant à l'entreprise Our Video Store peuvent avoir des bases de données et des structures différentes. Si le directeur de l'entreprise veut accéder aux données de tous les magasins pour prendre des décisions stratégiques, il serait plus approprié si toutes les données étaient stockées dans un seul emplacement avec une structure homogène qui permet l'analyse interactive des données. Autrement dit, les données de différents magasins peuvent être chargées, nettoyées, transformées et intégrées ensemble.

- **Les bases de données transactionnelles :** En général, une Base de Données transactionnelle est un fichier où chaque enregistrement représente une transaction.

Une transaction contient un identifient unique de transaction (transaction ID) et une liste d'items composant la transaction (les achats d'un client lors d'une visite). Les bases de données transactionnelles peuvent contenir d'autres informations tels que la date de la transaction, l'identifient du consommateur, l'identifient de la personne qui a vendu, et ainsi de suite. [6] Par exemple, dans le cas du magasin de vidéos, la table de locations qui est illustrée par la figure 5 représente la base de données transactionnelle où chaque enregistrement est un contrat de location contenant un identifiant du consommateur, une date et une liste des items loués (cassettes vidéo, jeux, VCR, etc.). Pour le Data Mining sur ce type de données nous utilisons souvent les règles d'association (appelées aussi Analyse du panier de la ménagère) dans lesquelles les associations des items arrivant ensemble ou séquentiellement soient étudiées.

- **Les bases de données multimédia :** Les bases de données multimédia4 comportent des documents sonores, des vidéos, des images et des médias e4n textes et audio. Elles peuvent être stockées sur des bases de données o4rientées objets ou objets relationnelles ou simplement sur un fichier système. Le multimédia est caractérisé par sa haute dimension ce qui rend le datamining sur ce type de données très difficile. Le data mining sur les supports des multimédias requiert exige la vision par ordinateur, l'infographie, l'interprétation des images et les méthodologies de traitement de langages naturels.

- **Les bases de données spatiales :** Ce sont des bases de données, qu'en plus de leurs données usuelles, elles contiennent des informations géographiques comme les cartes et les positionnements mondiaux ou régionaux. De telles bases de données présentent de nouveaux défis aux algorithmes de data mining.

- **Les bases de données de séries temporelles :** Les bases de données de séries temporelles contiennent des données relatives au temps, comme les données du marché boursier ou les activités enregistrées. Ces bases de données ont couramment un flux continu de nouvelles données entrantes, qui parfois rend l'analyse en temps réel un besoin exigeant. Le data mining pour ce genre de bases de données est généralement l'étude des tendances et des corrélations entre les évolutions des différentes variables, aussi bien que la prédiction des tendances et des mouvements des variables par rapport au temps. Par exemple, une base de données du trafic automobile qui stocke une description symbolique de séries temporelles de ce dernier, il sera possible de répondre à la requête : « définir les grand axes où le commerce est fluide le week-end ». La figure suivante montre quelques exemples de données en séries temporelles :

Figure 08: *Données en série temporelle*

- **Le World Wide Web :** Le World Wide Web est le support de données le plus hétérogène et le plus dynamique disponible. Un grand nombre d'auteurs et d'éditeurs contribuent sans arrêt à son accroissement et évolution, et chaque jour un énorme nombre d'utilisateurs accède à ses ressources. Les données dans le World Wide Web sont organisées dans des documents interconnectés. Ces documents peuvent être des textes, audio, vidéos, données brutes et même des applications. Conceptuellement, le World Wide Web est composé de trois grands composants : le contenu du Web, qui englobe les documents disponibles ; la structure du Web, qui garantit les hyperliens et les relations entre documents ; et l'usage du Web, en décrivant quand et comment les ressources seront accédées. Une quatrième dimension peut être ajoutée concernant la nature dynamique ou l'évolution des documents. Le data mining pour le World Wide Web, ou le web mining, essaie d'aborder toutes ces questions et il est souvent divisé en contenu Web mining, la structure Web mining et l'usage Web mining.

9. Avantages et inconvénients de Data mining

Avantage :

- ✓ Ils utilisent l'évaluation de la fonction objective sans prendre en compte sa nature ce qui lui donne plus de souplesse et un large domaine d'application.
- ✓ Ils sont dotés de parallélisme car ils travaillent sur plusieurs points en même temps il s'agit des individus de la population.
- ✓ L'utilisation de règles de transition probabilistes de croisement et de mutation permet dans certains cas d'éviter des optimums locaux et d'aller vers un optimum global

Inconvénients :

- ✓ Temps de calcul très élever car ils nécessitent de nombreux calculs particulièrement au niveau de la fonction d'évaluation.
- ✓ Difficiles à mettre en œuvre à cause des paramètres parfois difficiles à déterminer comme la taille de la population ou le taux de mutation. Ce qui implique la nécessité de plusieurs essais car le succès de l'évolution en dépend, ce qui limite encore l'efficacité de l'algorithme.
- ✓ du choix de la fonction d'évaluation qui est critique, elle doit prendre en compte les bons paramètres du problème. Elle doit donc être choisie avec soin.
- ✓ Il est impossible d'être sure que la solution obtenue après un nombre fini de générations soit la meilleure, on peut seulement être sûr que l'on s'est approché de la solution optimale.
- ✓ Problème des optimums locaux : lorsqu'une population évolue, il se peut que certains individus deviennent majoritaires. À ce moment, il se peut que la population converge vers cet individu et s'écarte ainsi d'individus plus intéressants mais trop éloignés de l'individu vers lequel la population converge.

Conclusion :

Data Mining et Anlaytics sont les technologies de base pour le nouveau monde basé sur la connaissance où nous construisons des modèles à partir de données et de bases de données pour comprendre et explorer notre monde. L'extraction de données peut améliorer nos activités, notre gouvernement et notre vie. Avec les bons outils, tout le monde peut commencer à explorer cette nouvelle technologie, sur la voie de devenir un professionnel de l'extraction de données.

Introduction

Les techniques d'exploration de données sont utilisées pour extraire des connaissances utiles à partir de données brutes. Les connaissances extraites sont précieuses et affectent considérablement le décideur .l'exploration des données éducatives (EDM) est une méthode d'extraction utile pouvant potentiellement affecter une organisation.

Dans ce chapitre nous allons aborder la notion d'Edm, ses objectifs ainsi que les différentes méthodes du data mining utilisées dans l'éducation.

1. Définition

L'exploration de données (DM) dans l'éducation est un domaine de recherche interdisciplinaire émergent, également appelé exploration de données pédagogiques (EDM) [14]. Il s'agit de développer des méthodes pour explorer les types uniques de données provenant d'environnements éducatifs. Son objectif est de mieux comprendre comment les élèves apprennent et d'identifier les contextes dans lesquels ils apprennent pour améliorer les résultats scolaires et pour mieux comprendre et expliquer les phénomènes éducatifs. Les systèmes d'information éducatifs peuvent stocker une quantité énorme de données potentielles provenant de sources multiples, dans différents formats et à différents niveaux de granularité. Chaque problème éducatif particulier a un objectif spécifique avec des caractéristiques particulières qui nécessitent un traitement différent du problème minier. Les problèmes signifient que les techniques traditionnelles de gestion du contenu ne peuvent pas être appliquées directement à ces types de données et de problèmes. En conséquence, le processus de découverte des connaissances doit être adapté et certaines techniques de MS spécifiques sont nécessaires.

2. Objectifs

Une politique d'analyse de données en temps réel amène normalement comme résultats un taux de rétention et de réussite de 5 à 20 % supérieur [14] et augmente l'efficacité de l'ensemble des services des institutions en éliminant les goulots, les redondances et autres entraves à l'efficacité. En fait l'analyse des données permet l'identification précise et rapide des problèmes bien avant qu'ils deviennent irrémédiables ou ingérables.

Un bon système permet de :

- ✓ Suivre ce qui se passe au niveau individuel ou collectif
- ✓ Détecter les problèmes et anomalies
- ✓ Être alerté et intervenir
- ✓ Comprendre ce qui se passe
- ✓ Orienter
- ✓ Prédire, monter des scénarios, faire des simulations : «si on modifie ce paramètre, qu'arrive-t-il ?»
- ✓ Optimiser

Les données utilisées peuvent provenir de plusieurs sources, aussi bien du registraire que de l'environnement numérique de travail utilisé par les étudiants, du système d'évaluation que de la bibliothèque ou de d'autres services aux étudiants, d'autant plus s'ils sont numériques et associés à l'identifiant de l'étudiant.

- Les cibles des analyses sont :

La gestion des inscriptions

Le financement et le budget

Les progrès des étudiants

Les apprentissages des étudiants

La gestion des enseignements

L'infrastructure informatique

L'évolution de la planification stratégique

Le suivi des finissants

L'administration de la recherche

On se sert de l'analyse des données pour :

- ✓ améliorer la prise de décision administrative et mieux répartir les ressources ;
- ✓ identifier les apprenants à risque, cibler les interventions pour les aider à réussir ; à partir d'un partage transparent des données, obtenir une meilleure compréhension des défis de l'institution.
- ✓ transformer positivement le modèle académique et les approches pédagogiques, par exemple en retirant des prérequis inadéquats, des activités superflues, etc.
- ✓ fournir aux apprenants des informations sur leurs propres habitudes d'apprentissage et fournir des recommandations ou des comparaisons à la moyenne par exemple ;
- ✓ augmenter la productivité et la réactivité en fournissant des données à jour et les défis qui peuvent leur être associées ; etc.

3. Processus [26]

• Analyses des données d'apprentissage (Learning Analytics)

• Modélisation de l'apprenant

•Production de services pédagogiques

•Acquisition automatique des connaissances du domaine (modélisation du domaine)

• Diagnostic comportemental

• Diagnostic épistémique

• Rétroactions pédagogiques et didactiques.

Comme le montre la figure suivante :

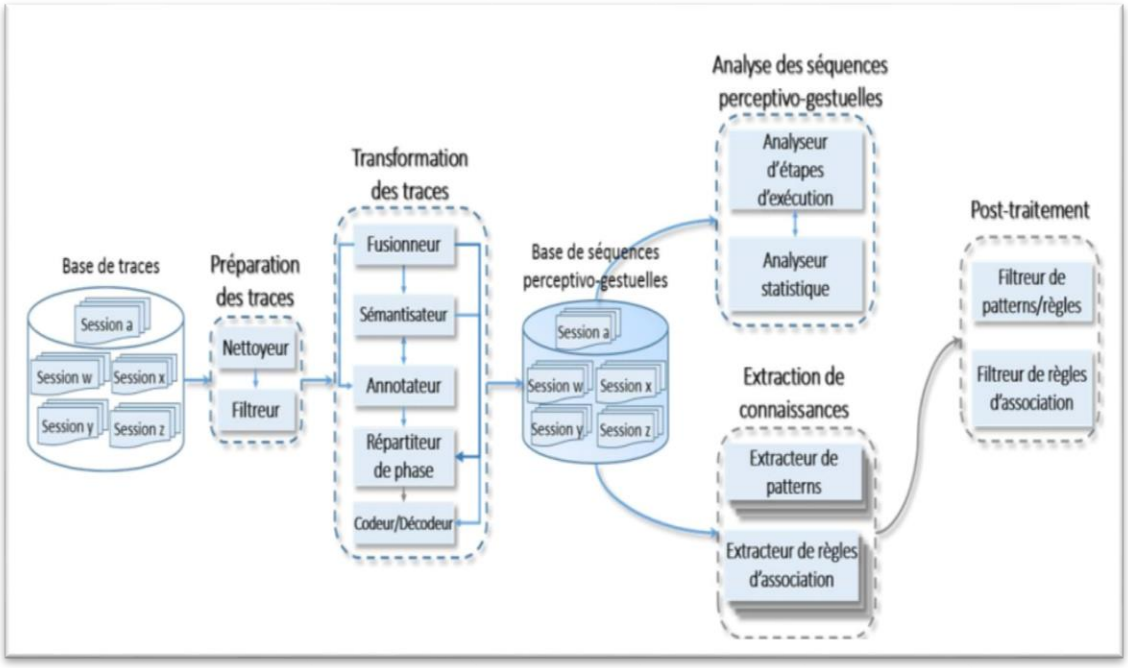


Figure 9: *Processus d'extraction de connaissance en éducation [26]*

Le data mining pour l'éducation est une discipline scientifique émergente ; elle fait se rencontrer plusieurs discipline, en particulier : Le data mining, le machine learning, la statistique, les mathématiques et l'informatique.

Les sciences de l'apprendre (en anglais Learning Science)C'est un processus conçu pour l'analyse de données issues uniquement de milieux éducatifs, pour mieux comprendre les apprenants et les situations dans lesquelles ils apprennet.il a pour but de construire des modèles et produire des résultats qui puissent aider à la conception et la réalisation d'applications et d'environnements innovants pour l'apprentissage ainsi que d'apporter une contribution théorique a la psychologie de l'éducation ou encore observer les taux d'échecs et de réussite afin de faciliter leurs orientations ou d'autres domaines en éducation.

4. Méthodes data mining pour l'éducation

Ryan Baker classifie les méthodes du data mining pour l'éducation comme suit [26] :

- **Prédiction :** C'est la modélise des données numériques pour prédire des valeurs inconnues ou manquantes. La prévision dans l'exploration de données consiste à identifier les points de données uniquement sur la description d'une autre valeur de données associée. Il n'est pas nécessairement lié à des événements futurs, mais les variables utilisées sont inconnues. La prédiction détermine la relation entre une chose que vous connaissez et une chose que vous devez prévoir pour référence future.

- **Clustering :** L'analyse typologique ou le groupement est la tache de regrouper un ensemble d'objet (cluster) tel que chaque un d'entre eux est séparé des autres et regroupe des données différentes. Plus la distance entre deux clusters est importante, plus les données qu'ils referment sont différentes. Il s'agit d'une tâche principale d'exploration minière de données, et une technique courante pour les statistiques d'analyse de données utilisés dans de nombreux domaine y compris l'apprentissage automatique, la reconnaissance de forme, analyse d'image, la recherche d'information et la bio-informatique.

- **L'exploration minière de la relation :** Implique de découvrir des rapports entre les variables dans un ensemble de données et de les coder comme des règles pour l'usage postérieur. Par exemple, l'exploitation de rapport peut identifier les rapports parmi des produits achetés dans des achats en ligne.

- **Règle d'association :** 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 ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans de très importantes bases de données. Peut être employée pour trouver les erreurs d'étudiants qui se produisent, associant le contenu aux types d'utilisateurs pour établir des recommandations pour le contenu qui est susceptible d'être intéressant, ou pour faire des changements aux approches d'enseignements.

- **Découverte avec des modèles :** En découverte , les modèles sont généralement basés sur le clustering, prévision, ou ingénierie des connaissances utilisant le raisonnement humain plutôt que des méthodes automatisées. Le modèle développé est alors utilisés dans le cadre d'autres modèles complets tels que les l'exploitation minière des navires. Les applications clés de cette méthode incluent la découverte de relations entre les comportements, les caractéristiques et les variables contextuelles des élèves dans l'environnement d'apprentissage. Cette méthode permet également de découvrir d'autres questions de recherche vastes et spécifiques dans un large éventail de contextes.

- **Distillation de données pour le jugement humain :** La distillation des données pour le jugement humain vise à produire des données compréhensibles. Présenter les données de différentes manières aide le cerveau humain découvre de nouvelles connaissances wledge. Différentes sortes de les données nécessitent des méthodes spécifiques pour les visualiser. Toutefois , les méthodes de visualisation utilisées dans l'exploration de données éducatives sont différentes de celles utilisées dans différents ensembles de données dans qu'ils considèrent la structure des données sur l'éducation et la sens caché à l'intérieur. La distillation des données pour le jugement humain est appliquée dans données nationales à deux fins : classification et / ou identification. Ce secteur de l'exploration éducative de données améliore les modèles,

ou les dispositifs d'étudiant, les comportements d'étudiant, ou les données d'étude comportant la collaboration parmi des étudiants.

5. Principales application du modèle EDM

Un domaine clé de la GED est la performance des étudiants en mines [25]. Un autre domaine clé est l'extraction des données d'inscription. Principales utilisations de la GED prévoir la performance des élèves et l'apprentissage des apprentissages afin de recommander des améliorations aux pratique pédagogique. La GED peut être considérée comme l'une des sciences de l'apprentissage, ainsi que comme un domaine de l'exploration de données.

Les principales applications de l'EDM sont listées comme suit :

- **Analyse et visualisation des données :** Il est utilisé pour mettre en évidence des informations utiles et faciliter la prise de décision. Dans le milieu éducatif, par exemple, il peut aider les éducateurs et les administrateurs de cours à analyser les activités de cours des étudiants et les informations d'utilisation pour avoir une vue générale de l'apprentissage d'un élève. Les statistiques et les informations de visualisation sont les deux techniques principale sont été les plus largement utilisés pour cette tâche. La statistique est une science mathématique concernant la collecte, l'analyse, l'interprétation ou explication et présentation des données. Il est relativement facile d'obtenir des statistiques descriptives de base à partir de logiciel statistique, tel que SPSS. Analyse statistique des données éducatives (journaux /fichiers / bases de données) peuvent nous indiquer des informations telles que les endroits où les étudiants entrent et sortent, les pages les plus populaires consultées par les étudiants, nombre de téléchargements de ressources d'apprentissage en ligne, nombre de pages différentes consultées et durée totale de navigation.

Il fournit également des connaissances sur les résumés d'utilisation et des rapports sur les tendances hebdomadaires et mensuelles des utilisateurs, quantité de matériel que les élèves peuvent passer et l'ordre dans lequel ils étudient les sujets, les modes d'étude l'activité, le calendrier et l'enchaînement des événements, ainsi que l'analyse du contenu des notes et des résumés des élèves.

La visualisation utilise des techniques graphiques pour aider les gens à comprendre et à analyser des données. Il y a plusieurs études orienté vers la visualisation de différentes données éducatives telles que des modèles d'utilisateurs annuels, saisonniers, quotidiens et horaires comportement sur les forums en ligne. Certaines de ces enquêtes sont des graphiques statistiques permettant d'analyser les affectations complémentaires, questions admises, pointage de l'examen, données de suivi des étudiants pour analyser l'assiduité des étudiants, résultats des travaux et des quiz, des informations hebdomadaires sur les étudiants et les activités du groupe

- **Prédire le rendement des élèves :** Dans ce cas, nous estimons la valeur inconnue d'une variable qui décrit l'élève. En éducation, les valeurs normalement prédits sont la performance de l'élève, ses connaissances, ses résultats ou ses notes. Cette valeur peut être numérique / continu (tâche de régression) ou catégorique / discret (tâche de classification). L'analyse de régression est utilisée pour trouver la relation entre une

variable dépendante et une ou plusieurs variables indépendantes. La classification est utilisée pour grouper articles individuels en fonction de caractéristiques quantitatives inhérentes aux articles ou sur un ensemble de formation de l'article. La prévision de la performance d'un élève est l'une des applications les plus populaires du DM en éducation. Différentes techniques et modèles sont appliqués comme les réseaux de neurones, les réseaux bayésiens, les systèmes de règles, la régression et l'analyse de corrélation pour analyser des données éducatives. Cette analyse nous aide à prédire la performance des élèves, à savoir prédire sur son succès dans un cours et sur sa note finale en fonction des caractéristiques extraites des données. Différents types de systèmes à base de règles ont été appliqués pour prédire la performance des élèves (prédiction de marques) dans un environnement d'apprentissage en ligne (en utilisant des règles d'association floue). Plusieurs techniques de régression sont utilisées pour prédire des marques comme la régression linéaire pour prédire le rendement scolaire des élèves, la régression linéaire par étapes pour prédire temps à consacrer à une page d'apprentissage, régression linéaire multiple pour identifier les variables permettant de prédire le succès collèges et pour prédire les résultats des examens dans les cours à distance.

- **Regrouper des étudiants :** Dans ce cas, des groupes d'élèves sont créés en fonction de leurs caractéristiques personnalisées, caractéristiques personnelles, etc.

L'instructeur / développeur peut utiliser ces grappes / groupes d'élèves pour créer un système d'apprentissage personnalisé qui peut promouvoir un apprentissage en groupe efficace. Les techniques DM utilisées dans cette tâche sont la classification et la mise en cluster.

Différents algorithmes de classification utilisés pour regrouper les étudiants sont la classification hiérarchique par agglomération, *K* –means et clustering basé sur un modèle. Un algorithme de clustering est basé sur de grandes séquences généralisées qui aident à trouver des groupes d'élèves ayant des caractéristiques d'apprentissage similaires, comme l'algorithme de classification hiérarchique qui sont utilisés dans les systèmes d'apprentissage électronique intelligents pour regrouper les étudiants en fonction de leurs préférences en matière de style d'apprentissage

- **Gestion des inscriptions :** Ce terme est fréquemment utilisé dans l'enseignement supérieur pour décrire des stratégies et des tactiques bien planifiées pour inscription d'un établissement et atteindre les objectifs fixés. La gestion des inscriptions est un concept organisationnel et un ensemble systématique d'activités conçues pour permettre aux établissements d'enseignement d'exercer davantage d'influence sur leurs élèves. Ces pratiques incluent souvent le marketing, les politiques d'admission, les programmes de rétention et l'aide financière attribution. Les stratégies et les tactiques reposent sur la collecte, l'analyse et l'utilisation de données permettant de projeter des résultats positifs.

Les activités qui produisent des améliorations mesurables des rendements sont poursuivies et / ou étendues, Tandis que les activités qui ne le sont pas sont interrompues ou restructurées. Les efforts compétitifs pour recruter des étudiants sont un accent commun des responsables des inscriptions.

Conclusion

Le data mining pour l'éducation est un puissant outil d'analyse et d'exploration de grandes bases de données en vue d'extraire des connaissances importantes pour les personnes concernées.

Dans ce chapitre nous avons fait un survol sur les différentes méthodes de data mining utilisés en éducation afin d'avoir un aperçu complet sur eux, ainsi son processus et ce pour répondre aux différents objectifs.

Introduction

La réalisation d'une application ou bien d'un outil informatique doit être impérativement précédée d'une méthodologie de conception qui a pour objectif la formalisation des étapes préliminaires du développement de cet outil afin de rendre ce développement plus fidèle aux besoins de son utilisateur.

La phase de conception nous permet de lister et décrire les solutions et les résultats attendues, le fonctionnement futur de système afin d'en faciliter la réalisation à l'aide du langage UML (Unified Modeling Language) qui permet de bien représenter les aspects statiques et dynamiques de notre projet.

1. Objectifs de notre application

L'objectif principal de notre travail est de mette en œuvre une application web qui sert d'intermédiaire entre les besoins des institutions d'éducation et les outils existants pour exécuter les algorithmes de data mining et ce afin de faciliter la tâche aux administrateurs des établissements, de simplifier leur activité et de faire de data analytique et extraction des données pour des données en éducation.

Notre application doit répondre aux besoins suivants :

- Utilisation des différents algorithmes de datamining.
- Classer les élèves selon un caractère prédéfini.
- Faire l'association entre les différentes matières.
- Faire une orientation pour les élèves selon la filière.
- La régression des élèves selon les différentes matières
- Faires des clusters selon la moyenne générale.

2. Architecture globale de notre application

Le schéma suivant représente l'architecture globale de notre application qui est divisée en plusieurs parties principales qu'on va détailler par la suite,

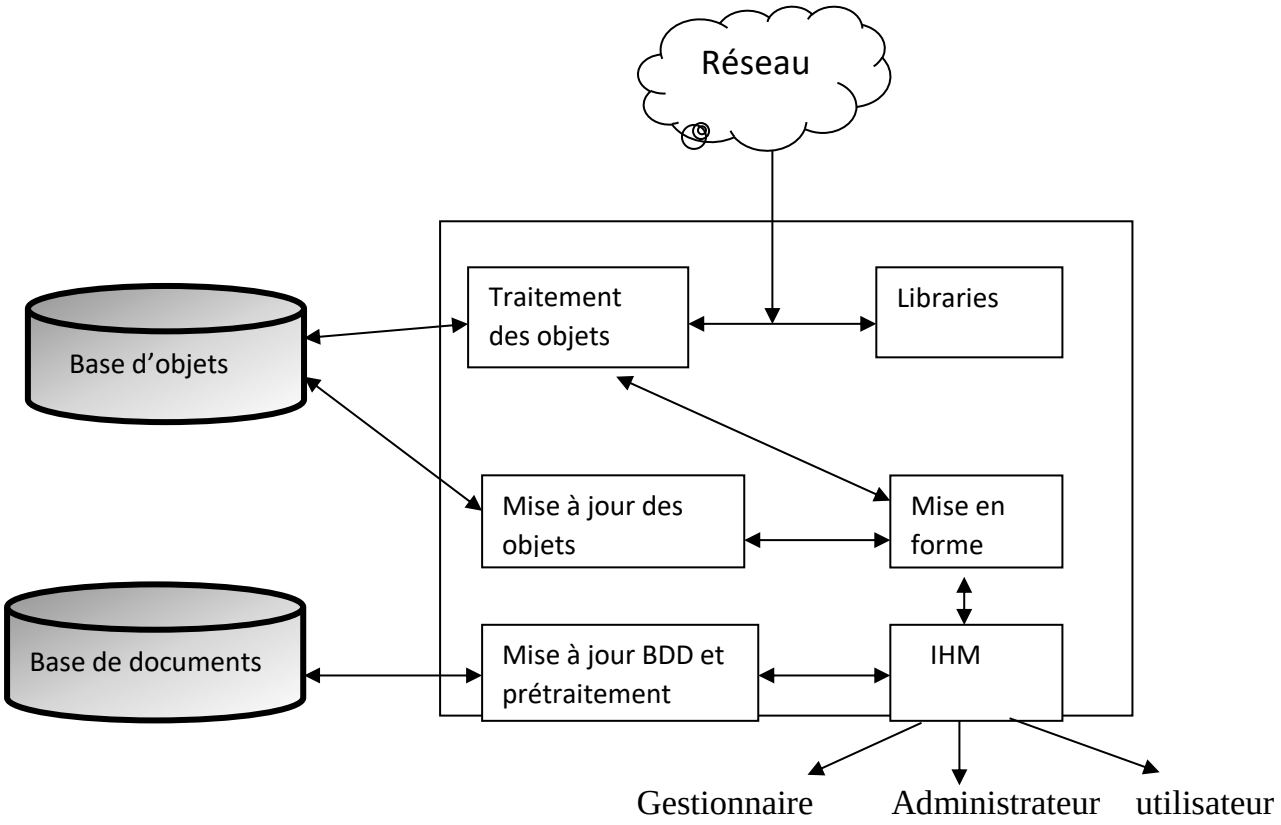


Figure 10: *Architecture globale de l'application*

- **Modules de l'application :**

On a cinq modules principaux :

✓ Traitements des objets :

C'est la partie la plus importante de notre application elle consiste à prendre un objet en entré puis écrire les programmes qui vont concrétiser les objectifs en activant un ou plusieurs algorithmes de data mining .

✓ Module de librairie (macro instruction) :

Plusieurs librairies telles que weka,R,RapidMiner peuvent être utilisées , une fois téléchargées on leur fera appel pour les activer .

On peut faire une activation à distance via le réseau ou une communication directe si elle se trouve sur le serveur.

✓ Module mise en forme et édition :

C'est un module très important en datamining il est en communication avec l'IHM pour le choix de l'ergonomie.

✓ IHM (Interface homme machine) :

Communique avec tous les acteurs : administrateur d'institution d'éducation, utilisateur final, gestionnaire de scolarité, programmeur (informaticien).

- ✓ Mise à jour des données et prétraitement :

C'est un module qui est en contact avec la base de donnée et avec prétraitement des objets pour le mettre à jour à l'arrivé de chaque nouvel objet.

3. Spécification des cas d'utilisation

3.1 Définition de cas d'utilisation

Un cas d'utilisation est une manière spécifique d'utiliser un système. Les acteurs sont à l'extérieur du système ; ils modélisent tout ce qui interagit avec lui. Un cas d'utilisation réalise un service de bout en bout, avec un déclenchement, un déroulement et une fin, pour l'acteur qui l'initie.

3.2 Le langage de modélisation UML

- ✓ Présentation d'UML:

UML « Unified Modeling Language» : se définit comme un langage de modélisation graphique et textuel destiné à comprendre et décrire des besoins d'une entreprise ou d'une société, spécifier et documenter des systèmes, concevoir des solutions et communiquer des points de vue. [27]

Ce langage est né de la fusion des trois méthodes qui ont influencé la modélisation objet en 1997 : OMT, Booch et OOSE. Principalement issus des travaux d'une équipe d'experts : GradyBooch, James Rumbaugh et Ivar Jacobson. UML est à présent un standard adopté par l'Object Management Group (OMG).Ses deux principaux objectifs sont la modélisation de systèmes utilisant les techniques orientées objet, depuis la conception jusqu'à la maintenance, et la création d'un langage abstrait compréhensible par l'homme et interprétable par les machines.

La version actuelle est UML 2 qui propose 13 diagrammes, dont nous allons utiliser : le cas d'utilisation.

- ✓ Les outils de modélisation UML :

Plusieurs logiciels, citons
– Together, Borland
– Win Design
– Plugin Omondo, Eclipse
– ArgoUML
Et tous les autres….
Dans notre cas, on a opter pour ArgoUML.

3.3 Identification des acteurs

- ✓ Définition d'un acteur : [28] Un acteur représente un rôle joué par une entité externe (utilisateur humain, dispositif matériel ou autre système) qui interagit directement avec le système, il peut consulter et/ou modifier directement l'état du système, en émettant et/ou recevant des messages susceptibles d'être porteur des données.

✓ Les acteurs de notre système :

Les acteurs de notre système sont :

- **le gestionnaire** : une interface « gestionnaire » est mise à sa disposition en cas de nécessité pour ajouter un nouveau besoin.
- **le programmeur** (informaticien) : Après avoir accéder au nouveau besoin, il mettra à jour le code adéquat.
- **l'utilisateur final** : Dans l'espace « utilisateur » il va choisir l'algorithme Data mining qui répond à ses besoins.

3.4 Les cas d'utilisation

Des schémas UML pour les cas d'utilisation (use case) :

- **Cas d'utilisation pour l'administrateur :**

✓ Consulter les nouveaux besoins
✓ Arranger le code selon les nouveaux besoins.
✓ adapter les outils de data mining en besoin.

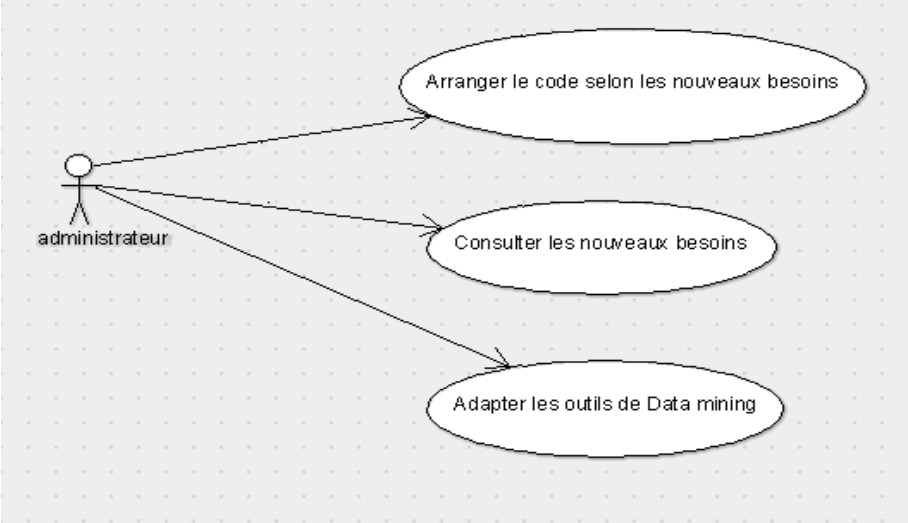


Figure 11: *Diagramme de Cas d'utilisation de l'administrateur*

- **Cas d'utilisation pour le gestionnaire :**

✓ Ajouter de nouveaux besoins : le nom de besoin et petite explication

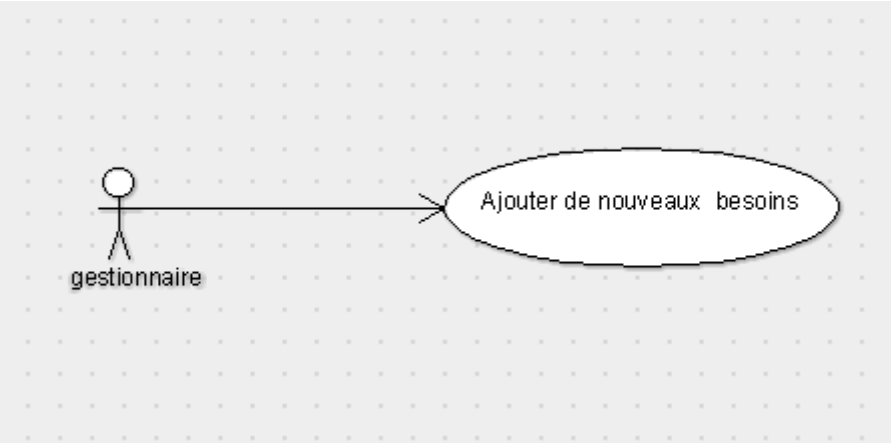


Figure 12: *Diagramme de cas d'utilisation de gestionnaire*

- **Cas d'utilisation pour l'utilisateur final :**

✓ faire les prétraitements.
✓ Activation des algorithmes de data mining.

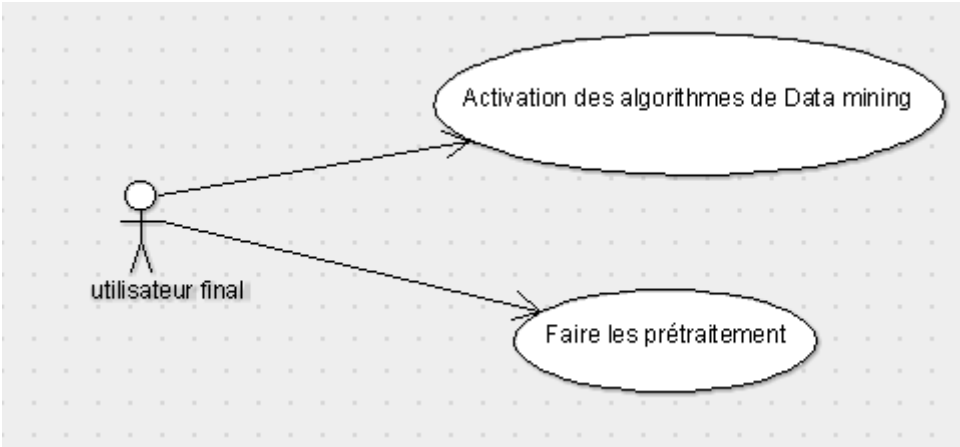


Figure 13: *Diagramme de cas d'utilisation de l'utilisateur final*

4. Le schéma de fonctionnement de traitement d'un objet

Le schémas ci-après représente le fonctionnement de traitement d'un objet (et ce dans l'environnement Weka):

- le preprocessing : qui consiste la préparation et la transformation des fichiers de données à utiliser pour l'extraction des données
- choisir un algorithme à exécuter de data mining puis l'affichage et la sauvegarde des résultats.

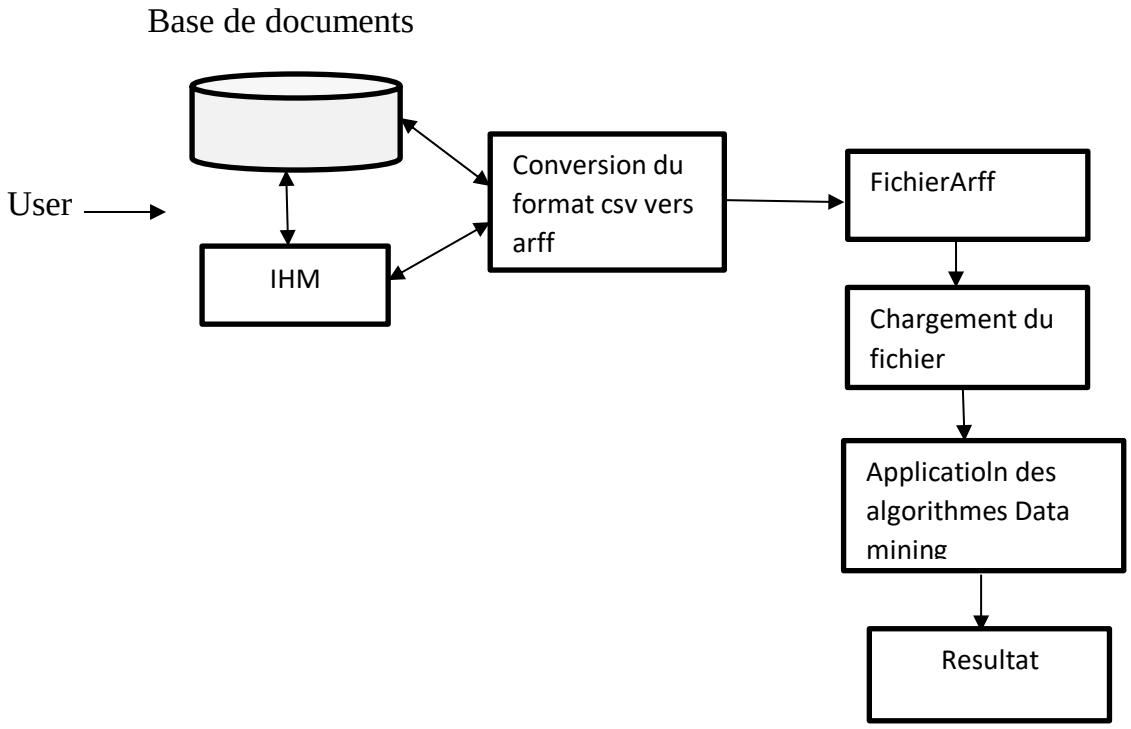


Figure 14: *processus d'exécution d'un objet*

- processing :(fichiers csv vers fichier arff)

Les outils de data mining exigent un format de données spécifique pour implémenter les algorithmes qu'ils proposent. Dans notre cas le format accepté par Weka est le format ARFF, pour cela on a pu développer un module sous java qui prend en entrée un fichier en format csv pour produire en sortie un fichier en format ARFF.

- Format CSV :[15]

Comma-separated values, connu sous le sigle CSV, est un format texte ouvert représentant des données tabulaires sous forme de valeurs séparées par des virgules.

Un fichier CSV est un fichier texte, par opposition aux formats dits « binaires ». Chaque ligne du texte correspond à une ligne du tableau et les virgules correspondent aux séparations entre les colonnes. Les portions de texte séparées par une virgule correspondent ainsi aux contenus des cellules du tableau.

L'extension de ce type de fichiers est .csv, et ils peuvent être utilisés entre différents outils informatiques et bases de données et facile à convertir à un fichier ARFF qui est le format le plus utilisé par Weka.

- **Exemple**

Fichier au format.csv	**Représentation tabulaire**
Sexe,Prénom,Année de naissance M,fateh,2003 F,Sara,2003 F,nasrine,2004	<table><tr><td>Sexe</td><td>Prénom</td><td>Année de naissance</td></tr><tr><td>M</td><td>fateh</td><td>2003</td></tr><tr><td>F</td><td>Sara</td><td>2003</td></tr><tr><td>F</td><td>nasrine</td><td>2004</td></tr></table>

Figure 15 : *Exemple de format csv*

- **Format ARFF :**

ARFF(Attribute-Relation File Format) signifie Format de fichier relation-attribut. [15] C'est un fichier texte ASCII qui décrit une liste d'instances partageant un ensemble d'attributs.

Les fichiers ARFF comportent deux sections distinctes. La première section contient les informations d'en-tête, suivies des informations de données .L'en- tête du fichier ARFF contient le nom de la relation, une liste des attributs (les colonnes dans les données) et leurs types. Voici sa structure :

@RELATION <nom-relation>
@ATTRIBUTTE <nom-attribut1><type donneé1>
@ATTRIBUTTE <nom-attribut2><type donneé2>
@ATTRIBUTTE <nom-attributN><type donnéeN>

l'entête

@DATA Les données

ARFF val-attribut1, val-attribut2,….. val-attributN

Le <nom-attribut> doit commencer avec un caractère alphabétique, et s'il contient des espaces alors le nom de l'attribut entier doit être mis entre quottes.

Le <type donnée> peut être un parmi les quatre types supportés par Weka :

- Numeric : peut-être un nombre réel ou entier.

- String.

- Date<date-format>.

- <spécification nominale > : définit ainsi {<nom nominal1> ,<nom nominal2> , <nom nominal3> ,…..}

Dans la section de données ARFF :

- Chaque instance est représentée en une seule ligne.

- Les valeurs de l'attribut pour chaque instance sont délimitées par des virgules et doivent apparaitre dans l'ordre dans lequel ils ont été déclarés dans la section de l'entête.

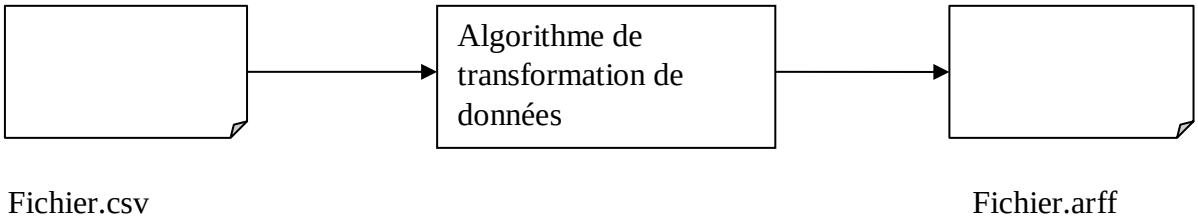


Figure 16: *Processus de transformation de fichier*

Exemple d'objet :

❖ Edition des étudiants admis meilleurs en math :

- On classe les élèves selon la moyenne générale de sorte que ceux qui ont une note supérieur ou égale à 10 sont admis et les autres sont non admis
- parmi les admis qu'elles sont ceux meilleur en mathématique (la note de math supérieur à 15)

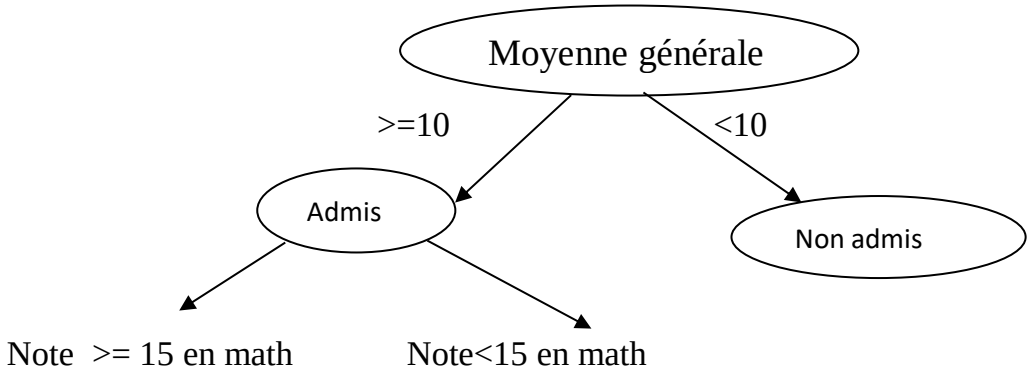


Figure 17: *exemple de classification des élèves*

❖ **algorithmes de l'objet :**

```
Début
    En entrée : un pv_note
        Appel fonction C4.5(pv_note,c1)
            Return C1
                Appel fonction C4.5(C1, C2)
                    Return   C2

        /* C1 := les élèves admis
        /* C2 := les élèves avec une note de mathématique >= 15
fin
```

Figure 18: *Algorithme de l'objet*

Pour le déroulement de cet exemple on doit essentiellement passer par les étapes suivantes :

1. cliquer sur bouton «processing»
2. Charger un fichier.CSV et le transformer en un fichier.ARFF.
3. Importer le fichier.ARFF.
4. Choix du type de l'algorithme de data mining :

Cette étape permet de choisir un algorithme de data mining parmi les 3 catégories d'algorithmes suivant :

- classification
- Clustering
- règles d'association

Dans notre cas c'est l'algorithme de classification

- **Choix de l'algorithme de classification et de paramètre :**

Il existe plusieurs modèles d'algorithmes de classification
(J48, SVM ,Naive-Bayes,OneR….) dans notre cas on a opter pour l'algorithmes J48

5. Exécution :

Consistera l'exécution de l'algorithme avec le mode choisit et les paramètres sélectionnés.

6. **Affichage :**

visualisation de l'arbre de décision pour l'algorithme J48.

7. sauvegarde de résultats.

Conclusion

Dans ce chapitre nous avons montré l'architecture globale de notre système d'extraction des données pour le domaine d'éducation suivi par une explication détaillée de cas d'utilisation de chaque auteur.

Les résultats de ce chapitre seront enrichis par des détails d'implémentation dans le chapitre suivant.

Introduction

Après avoir vu l'extraction de connaissances en générale, nous allons faire l'extraction de connaissances des élèves de Cem, ce qui nous pousse tout d'abord à présenter les langages de programmation et les outils utilisés ainsi que l'environnement d'exécution. Puis nous allons voir quelques interfaces illustrant notre application et on va finir par présenter quelques tests.

I. Implémentation avec java et weka

1. L'environnement de développement

1.1 Langage de programmation

Pour la réalisation de notre application nous avons utilisés le langage JAVA et l'implémentation de l'API Weka ainsi que l'utilisation de Swing et les JFrame pour la création de notre application.

- **Langage Java**

Java est un langage de programmation orienté objet qui produit des logiciels pour plusieurs plates-formes.[19] Lorsqu'un programmeur écrit une application Java, le code compilé (appelé byte code) s'exécute sur la plupart des systèmes d'exploitation (OS), notamment Windows, Linux et Mac OS. Java tire une grande partie de sa syntaxe des langages de programmation C et C ++.

Java a été développée au milieu des années 90 par James A. Gosling, un ancien informaticien de Sun Microsystems.

- Java produit des applets (programmes exécutés par un navigateur) facilitant l'interface utilisateur graphique et l'interaction des objets par les utilisateurs Internet. Avant les applets Java, les pages Web étaient généralement statiques et non interactives. Les applets Java ont perdu de leur popularité avec la sortie de produits concurrents, tels qu'Adobe Flash et Microsoft Silverlight.

1.2 Les outils

- **IDE Netbeans**

NetBeans est un environnement de développement intégré (IDE) pour Java, placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License). [19] 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.

NetBeans utilise des composants, également appelés modules, pour permettre le développement de logiciels. NetBeans installe les modules de manière dynamique et permet aux utilisateurs de télécharger des fonctionnalités mises à jour et des mises à niveau authentifiées numériquement.

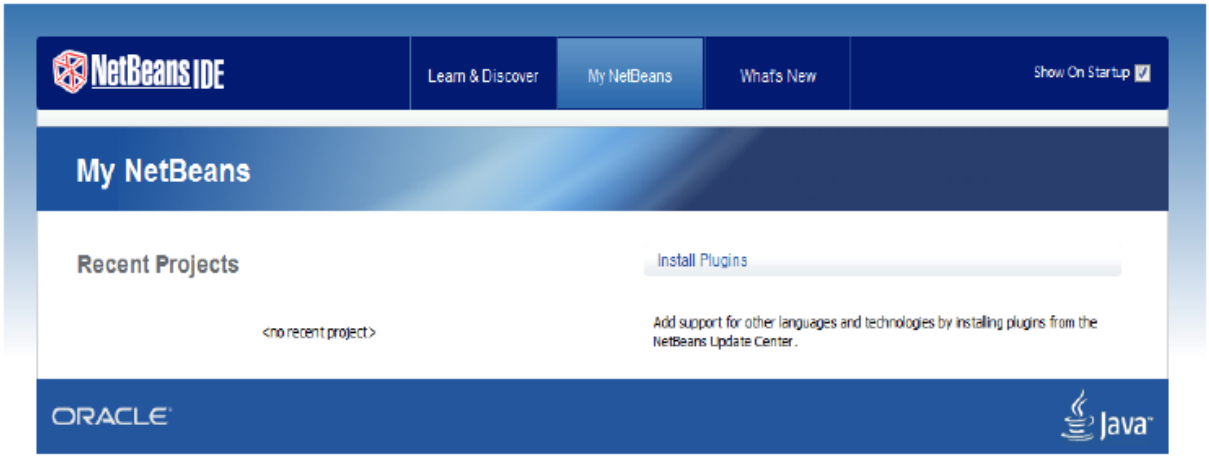


Figure 19 : *Page démarrage NetBeans IDE*

➢ Excel

Excel est un programme informatique développé et distribué par Microsoft Corp. Il s'agit d'un logiciel qui permet de réaliser des tâches comptables et financières grâce à ses applications pour créer et travailler avec des feuilles de calcul. [15]
Excel permet de créer facilement des tableaux ou des fichiers de toutes sortes, et d'y intégrer des calculs. Les valeurs du tableau se mettront donc à jour automatiquement en fonction de vos saisies et calculs. Les comptables et les administrateurs l'utilisent beaucoup. Excel permet également de générer de jolis graphiques (à bâtons, en camembert…) pour mieux visualiser les valeurs et les interpréter. C'est un puissant outil de visualisation mathématique.

On dit aussi qu'Excel est un logiciel d'analyse de données, autrement dit il fait subir à des données brutes des transformations de toutes sortes (mise en forme, calculs, gestion …etc.). En vue d'une utilisation spécifique, analyser des données ce n'est pas de les rendre jolies mais c'est leur créer une association pour les rendre utilisables.

1.3 Utilisation de l'API weka sous netbeans

➢ Chargement de l'API

Pour le chargement de l'API weka sous netbeans on a procédé ainsi :

- Installer le logiciel Weka 3.7
- Copier le fichier weka.jar qui se trouve dans le répertoire « C:\Program Files (x86)\Weka-3-7 » et le mettre dans le répertoire de notre projet

- A partir de netbeans et dans l'arborescence du notre projet, au niveau de librairies on clique avec le bouton droit et on choisit « Add JAR/Folder… ».
- Une fenêtre apparut. On sélectionne notre fichier weka.jar

➢ Utilisation de l'API :

Une fois le fichier weka.jar est ajouté à la librairie, on peut utiliser l'API weka pour implémenter les algorithmes de Data mining en ajoutant tout d'abord les packages nécessaires pour chaque algorithme, puis ajouter les instructions pour la programmation des différentes taches.

Dans ce qui suit on va présenter quelques packages utilisés dans notre code :

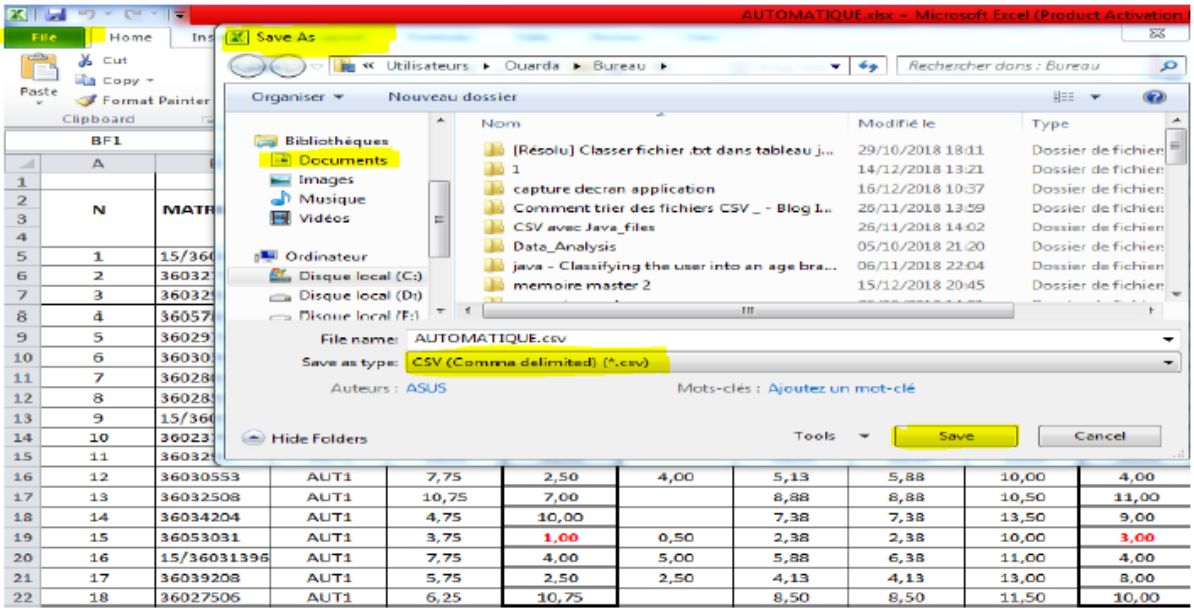
```
import weka.associations.Apriori;
import weka.associations.FPGrowth;
import weka.attributeSelection.AttributeSelection;
import weka.attributeSelection.CfsSubsetEval;
import weka.attributeSelection.GreedyStepwise;
import weka.classifiers.Evaluation;
import weka.classifiers.functions.LinearRegression;
import weka.classifiers.functions.SMOreg;
import weka.classifiers.meta.FilteredClassifier;
import weka.classifiers.trees.J48;
import weka.clusterers.ClusterEvaluation;
import weka.clusterers.SimpleKMeans;
import weka.core.Instance;
import weka.core.Instances;
import weka.core.Utils;
import weka.core.converters.ArffSaver;
import weka.core.converters.ConverterUtils.DataSource;
import weka.filters.Filter;
import weka.filters.unsupervised.attribute.NumericToNominal;
import weka.filters.unsupervised.attribute.Remove;
import weka.filters.unsupervised.instance.NonSparseToSparse;
import weka.gui.treevisualizer.PlaceNode2;
import weka.gui.treevisualizer.TreeVisualizer;
```

Figure 20: *Les packages utilises pour l'implémentation les algorithmes Weka*

➢ L'environnement de l'exécution :

Pour l'exécution de notre application nous devons cliquer sur le button run et de générer un fichier csv.

:On génère le fichier csv à partir de logiciel Excel en enregistrant le fichier sous l'extension « .csv »comme le montre la figure qui suit :



	N	MATR								
1	1	15/360								
2	2	360327								
3	3	360329								
4	4	360576								
5	5	360297								
6	6	360303								
7	7	360286								
8	8	360283								
9	9	15/360								
10	10	360231								
11	11	360321								
12	12	36030553	AUT1	7,75	2,50	4,00	5,13	5,88	10,00	4,00
13	13	36032508	AUT1	10,75	7,00		8,88	8,88	10,50	11,00
14	14	36034204	AUT1	4,75	10,00		7,38	7,38	13,50	9,00
15	15	36053031	AUT1	3,75	1,00	0,50	2,38	2,38	10,00	3,00
16	16	15/36031396	AUT1	7,75	4,00	5,00	5,88	6,38	11,00	4,00
17	17	36039208	AUT1	5,75	2,50	2,50	4,13	4,13	13,00	8,00
18	18	36027506	AUT1	6,25	10,75		8,50	8,50	11,50	10,00

Figure 21 : *Génération de fichier CSV sous EXCEL*

1.4 Présentation des interfaces de l'application

Dans ce qui suit nous présenterons quelques interfaces essentielles qui illustrent les étapes de l'exécution de notre application.

A partir de la page principale on peut accéder à l'espace data mining, en cliquant sur « Data mining » : Preprocessing et utilisation des algorithmes data mining.

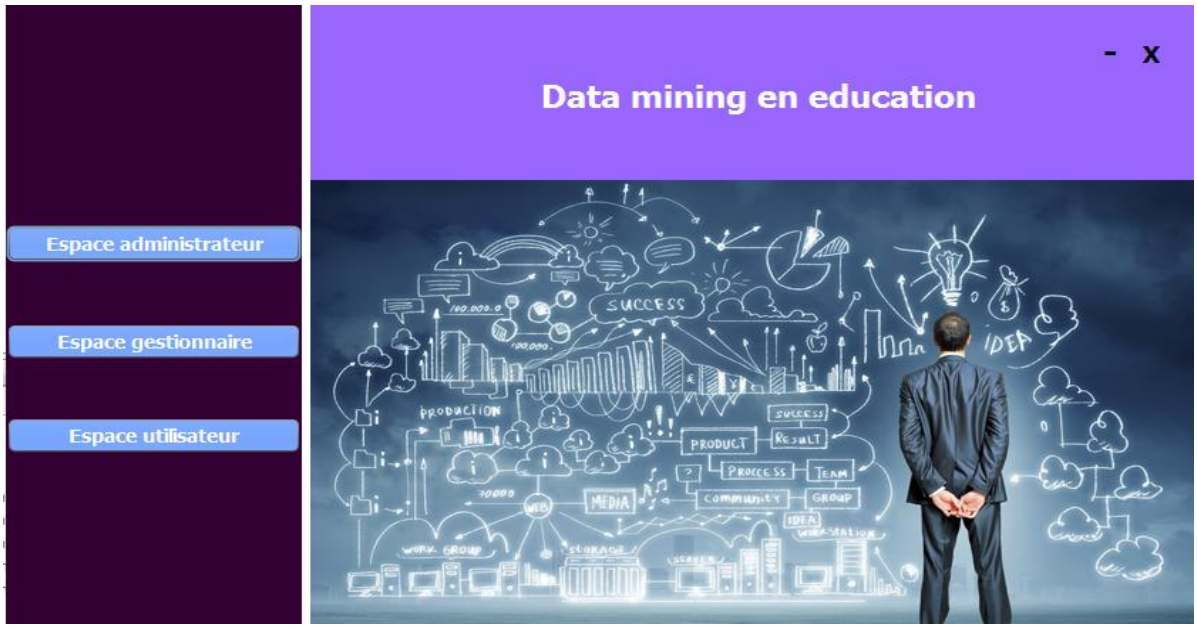


Figure 22: *Page principale*

❖ Page administrateur

Le lien « administrateur » envoie à l'espace prive de l'administrateur ou il a la possibilité d'arranger le code selon les nouveaux besoins de l'application. Une fois authentifié, l'administrateur accédera à son espace.

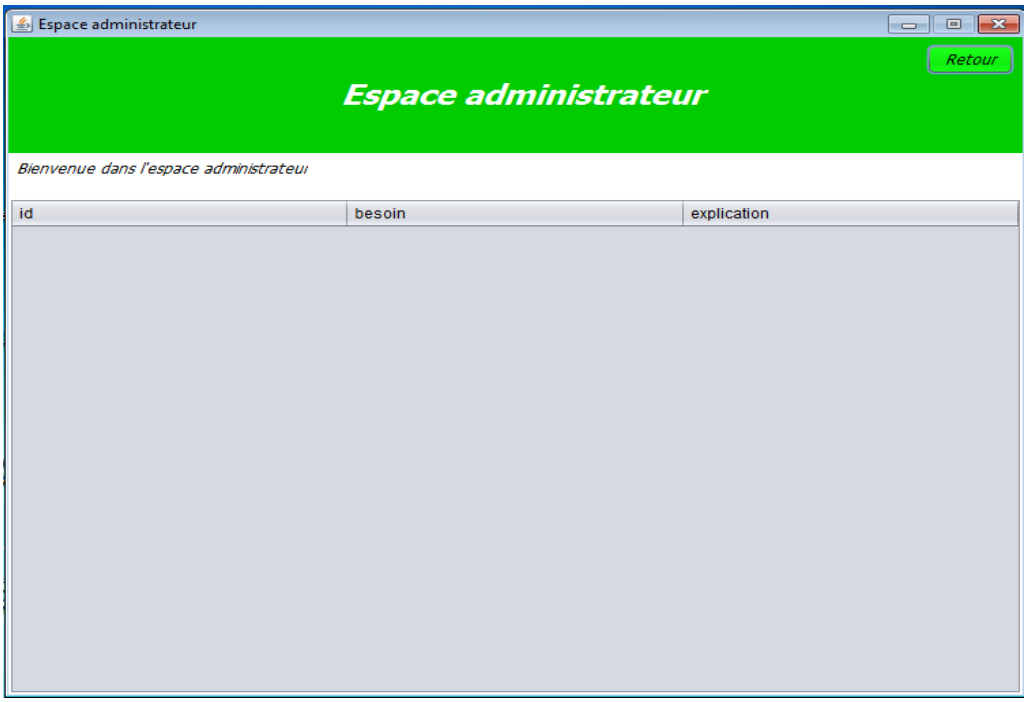


Figure 23 : *Espace administrateur*

❖ Page utilisateur

Figure 24: *Activation des algorithmes Data mining*

La page utilisateur dispose d'un ensemble de boutons qu'on va expliquer par la suite :

- **Preprocessing :** L'espace preprocessing est utilisé afin de préparer le fichier d'entrée.

Il nous permet de convertir le fichier CSV en un fichier ARFF comme on le voit sur la figure qui suit :

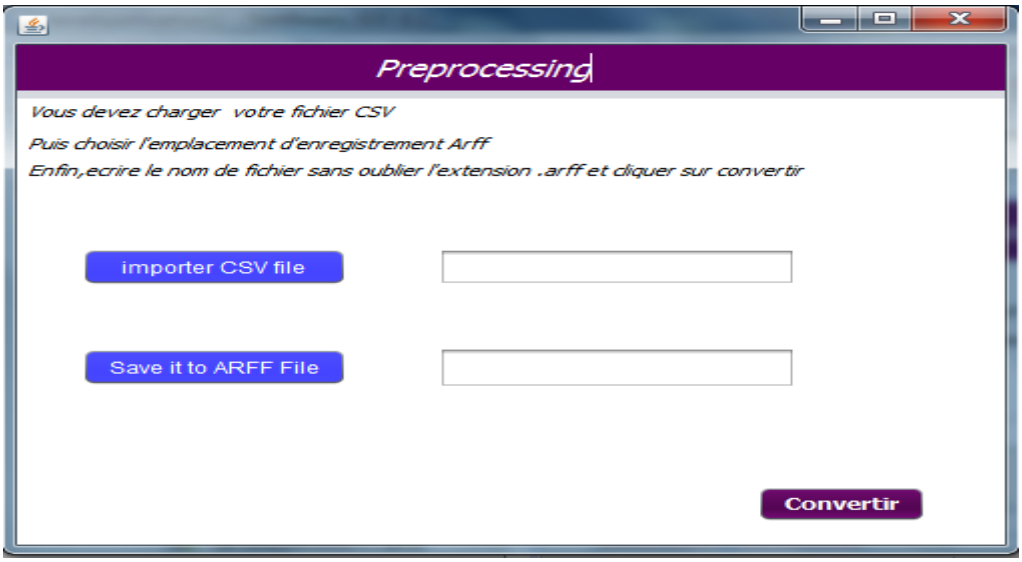


Figure 25: *Espace Preprocessing*

- **Classification :** Si on veut faire une classification des notes des élèves, on clique sur le bouton « classification » on aura l'interface suivante qui va nous permettre d'importer le fichier de données, de choisir le paramètre de classification ; 'orientation ' ,'admis' ou bien 'moyenne générale' ;comme on dispose du bouton « visualisation » pour tracer l'arbre de décision adéquat .

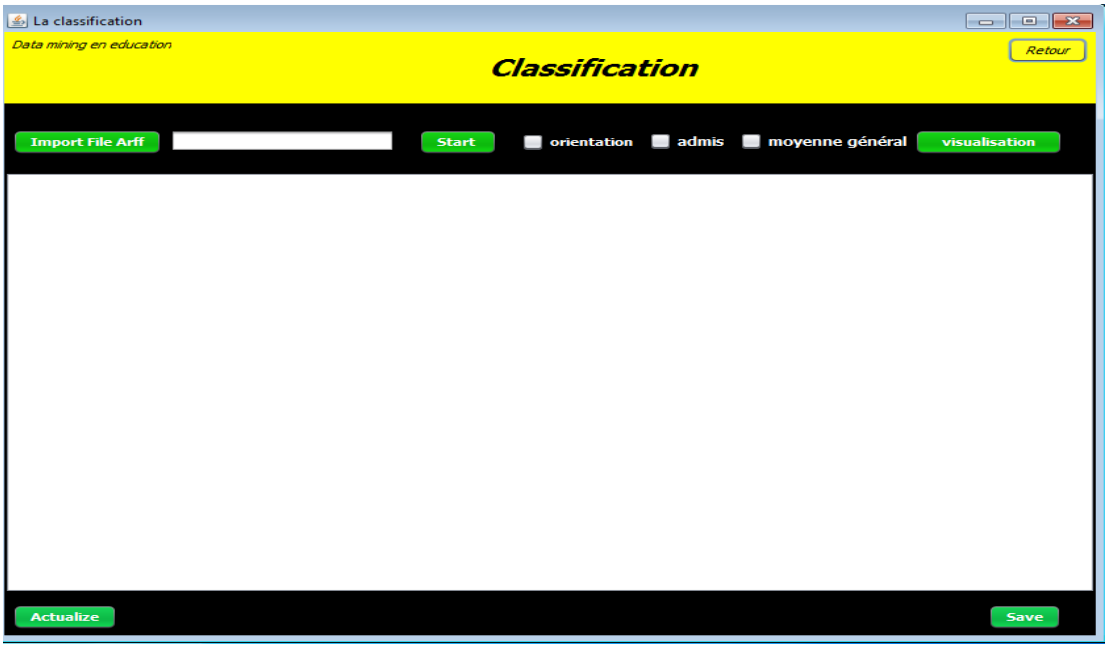


Figure 26: *Interface classification*

- **Règles d'association :** Pour découvrir des relations entre deux ou plusieurs attributs stockés dans des bases de données, on cliquant sur le bouton « association » et on aura l'interface suivante :

Figure 27: *Interface règle association*

- **Prédiction :** Pour faire des prédictions sur les nouvelles instances.

Une fois le modèle est généré on peut commencer à faire les prédictions en entrant un autre fichier arff contenant les nouvelles instances et importer notre modèle déjà enregistrer.
La figure suivante montre l'interface de prédiction :

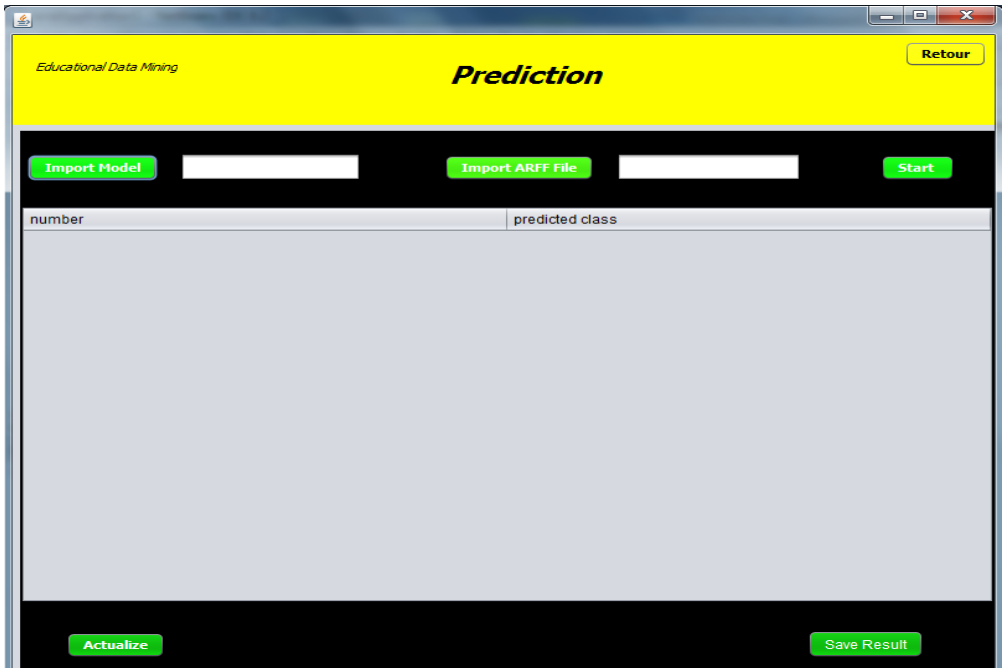


Figure 28: *Interface Prédiction*

- **Les objets traités :** Dans cet espace l'utilisateur il peut jeter un coup d'œil sur la liste des besoins traités auparavant.

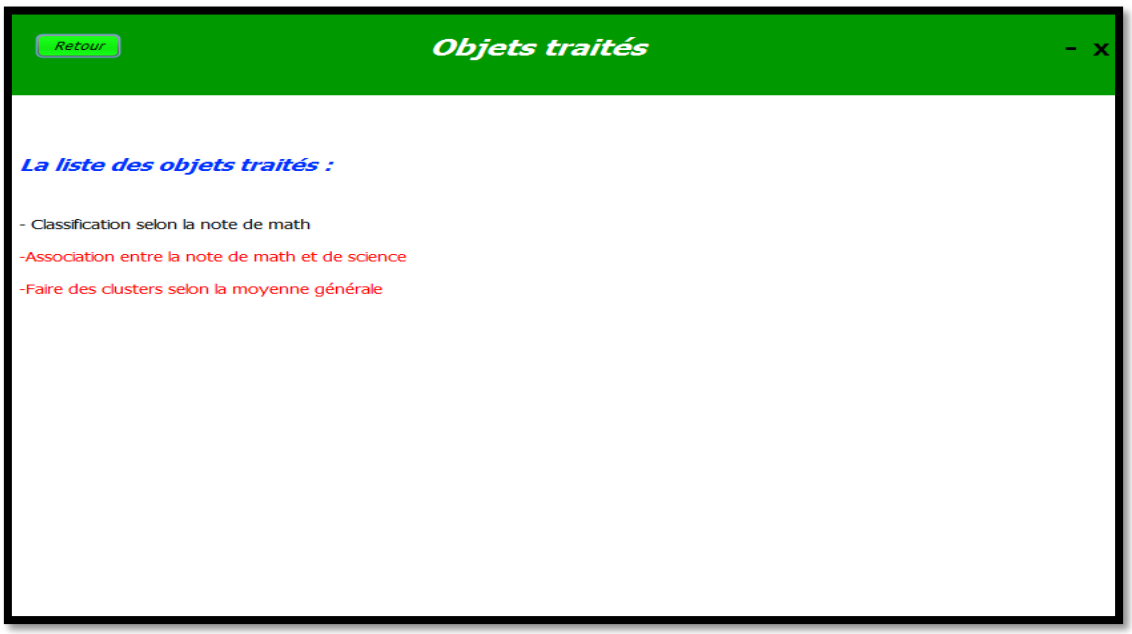


Figure 29: *Objets traités*

❖ **Page gestionnaire**

Le lien « gestionnaire » permet au gestionnaire d'accéder à son espace où il choisit son besoin ; sois par la liste qu'on lui a proposé ou bien en ajoutant un nouveau besoin.

Figure 30: *Espace gestionnaire*

Si le besoin désiré par le gestionnaire ne figure pas sur la liste proposée, il clique sur le lien « Ajouter un nouveau besoin » ou il remplit la fenêtre suivante :

Figure 31: *Espace gestionnaire*

2. Tests :

➢ Algorithme J48(classification) :

Pour ce cas on effectuera le test suivant : évaluation des apprenants selon leur moyenne, pour celui qui a une moyenne supérieur ou égale à 10 il est admis sinon il est non admis (ajourné), une deuxième classification sera effectuée sur la liste des élèves admis pour leur faire une orientation sois 'scientifique' ou bien 'littéraire'.

La fenêtre ci-dessous montre un exemple d'exécution de l'algorithme J48 :

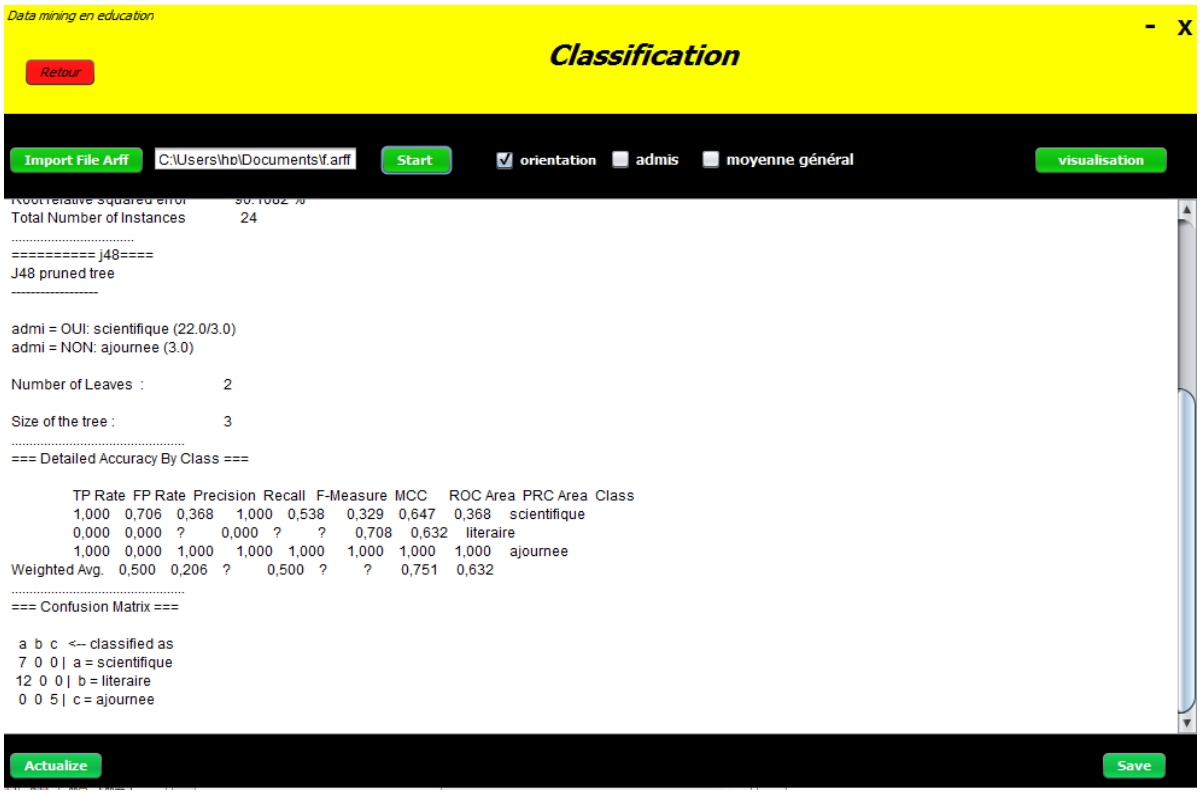


Figure 32: Exemple d'exécution de l'algorithme J48

Remarque

L'arbre de décision qui affiche la classification des élèves selon la moyenne est le suivant :

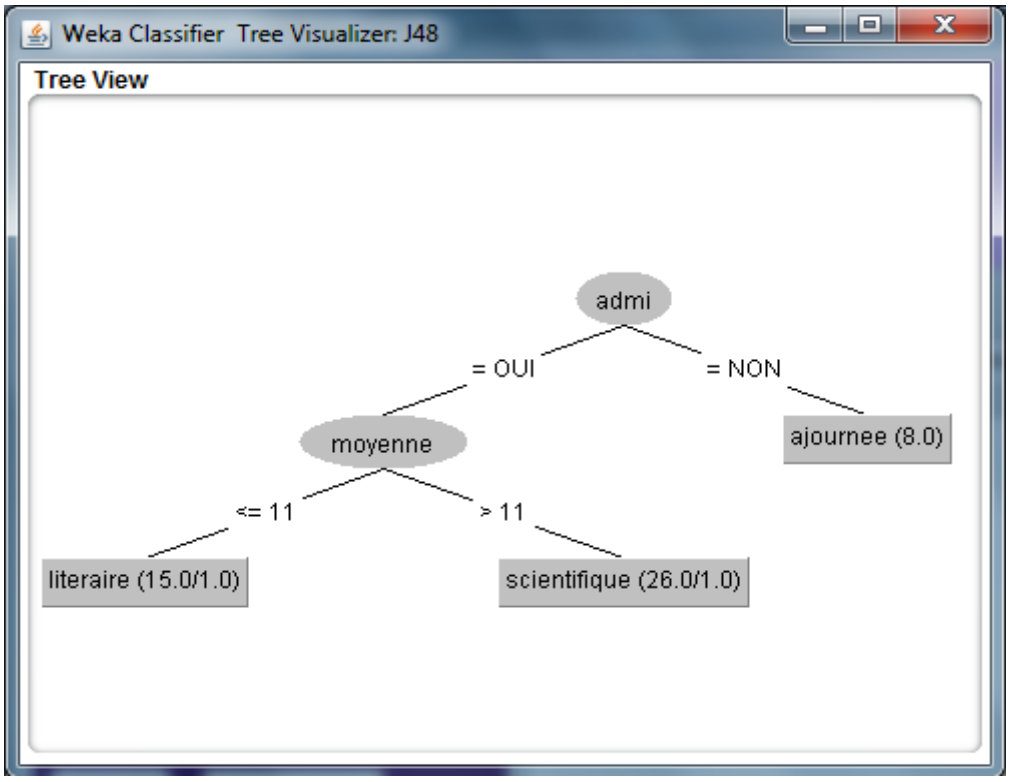


***Figure 33:** Arbre de décision*

- **Interprétation des résultats :**

On remarque que le nombre d'instances correctement classifiées est : 24 instances, le pourcentage est 100%.
Le nombre total d'instances est 24.

D'après la matrice de confusion, on trouve que :

3 apprenants sont **admis**, 22 apprenants pour le **non admis**.

L'arbre de décision quant à elle, nous permet de déduire les informations suivantes :

Si l'apprenant n'est pas admis : ajournée(8).

Si l'apprenant est admis et la moyenne >11 on est dans la classe scientifique.

Si l'apprenant est admis et la moyenne <=11 on est dans la classe littéraire.

- **Prédiction :** Après avoir entré le model de classification J48 et le fichier ARFF et cliquer sur le bouton start on aura les résultats de prédiction dans un tableau comme on le voit dans cette figure :

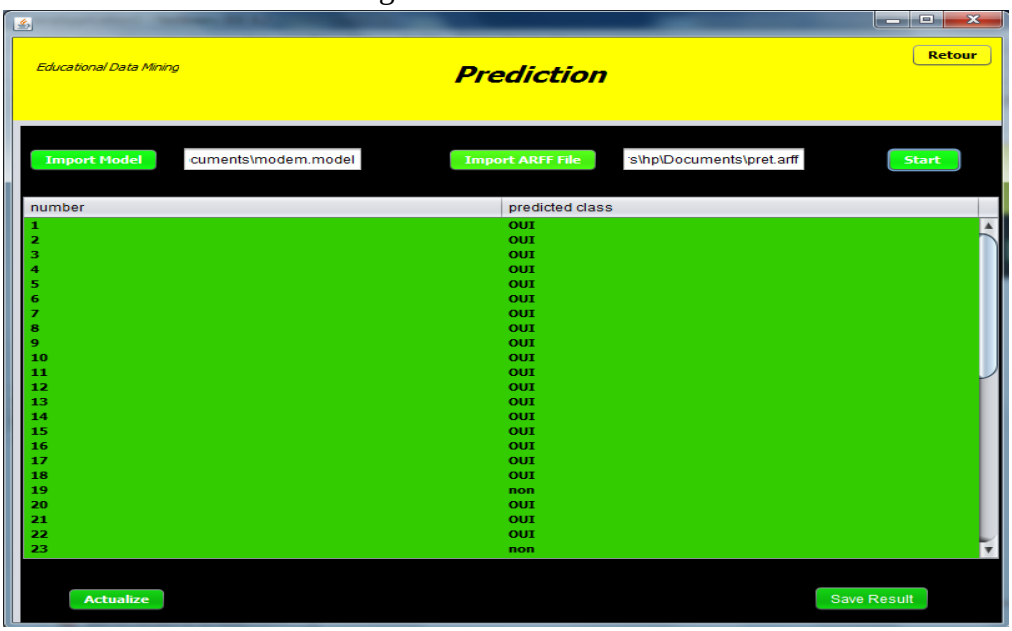


Figure 34: *Résultat de prédiction*

- **Algorithme apriori (règles d'association) :** Le test qu'on a effectué pour l'algorithme Apriori est la découverte de relation entre la note de mathématique et l'orientation des élèves ; le résultat est illustré dans la figure suivante :

Figure 35: *Exécution de l'algorithme Apriori*

- **Interprétation des résultats :**

On remarque que si la note de math=non(<10) alors l'orientation est littéraire par contre si la note de math=oui(>=10) donc orientation est scientifique.

II. implémentation avec r Project

1. Environment de development

1.1 Rstudio

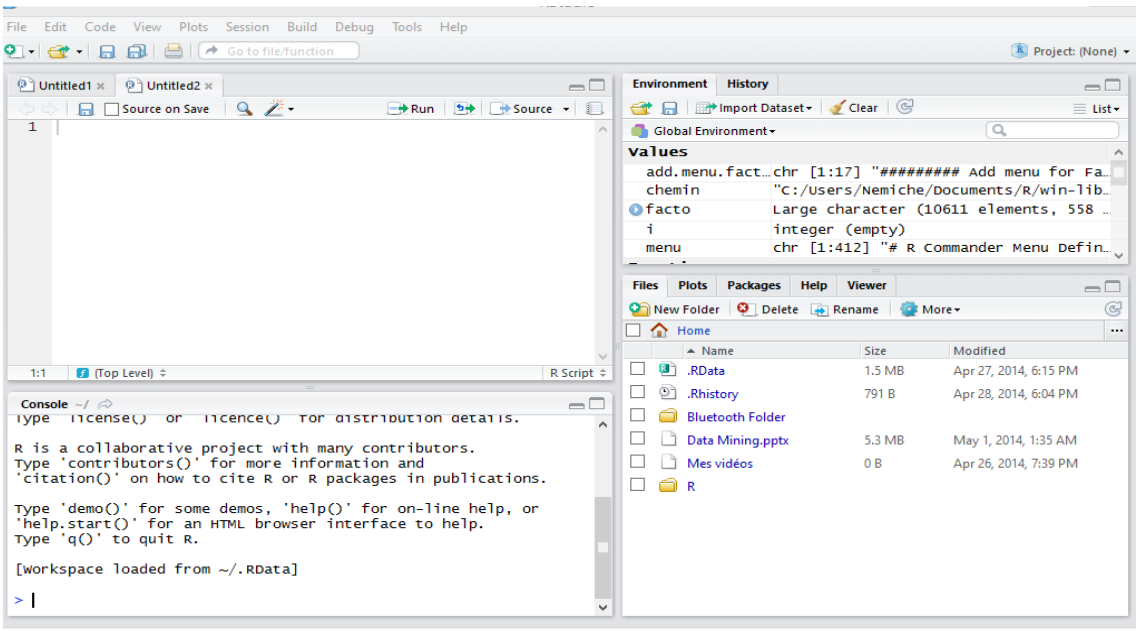


Figure 36: *Interface de RStudio*

RStudio est un environnement de développement gratuit, libre et multiplateforme pour R, un langage de programmation utilisé pour le traitement de données et l'analyse statistique. Il est disponible sous la licence libre AGPLv3, ou bien sous une licence commerciale, soumise à un abonnement annuel. [29]

RStudio est un environnement de développement intégré (IDE) qui fournit une interface en ajoutant de nombreuses fonctionnalités et outils pratiques. Ainsi, l'accès à un compteur de vitesse, à des rétroviseurs et à un système de navigation facilite grandement la conduite. L'utilisation de l'interface de RStudio facilite également l'utilisation d`e R il a été écrit en langage C++, et son interface graphique utilise l'interface de programmation Qt.

RStudio est disponible en deux versions : RStudio Desktop, pour une exécution locale du logiciel comme tout autre application, et RStudio Server qui, lancé sur un serveur Linux, permet d'accéder à RStudio par un navigateur web. Des distributions de RStudio Desktop sont disponibles pour Microsoft Windows, OS X et GNU/Linux3 ;Il comprend une console, rédacteur en soulignant la syntaxe qui prend en charge l'exécution directe du code, ainsi que des outils pour le traçage, l'histoire, le débogage et la gestion de l'espace de travail.

1.2 R projet

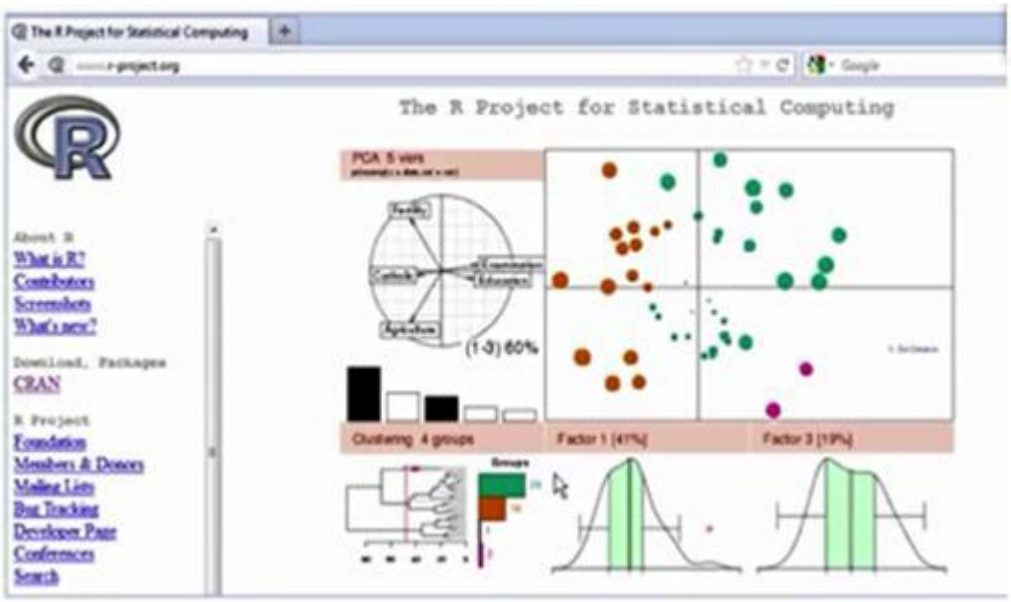


Figure 37: *Interface de Rproject*

R un langage de programmation et un logiciel libre dédié aux statistiques et à la science des données soutenu par la R Foundation for Statistical Computing. R fait partie de la liste des paquets GNU2 et est écrit en C (langage), Fortran et R. [28]

GNU R est un logiciel libre distribué selon les termes de la licence GNU GPL et disponible sous GNU/Linux3, FreeBSD4, NetBSD5, OpenBSD6, Mac OS X7 et Windows8

Le langage R est largement utilisé par les statisticiens, les data miner, data scientiste pour le développement de logiciels statistiques et l'analyse des données.

Le logiciel R est un logiciel de statistique créé par Ross Ihaka Robert Gentleman. Il est à la fois un langage informatique et un environnement de travail : les commandes sont exécutées grâce à des instructions codées dans un langage relativement simple, les résultats sont affichés sous forme de texte et les graphiques sont visualisés directement dans une fenêtre qui leur est propre. C'est un clone du logiciel S-plus qui est fondé sur le langage de programmation orienté objet S, développé par AT&T Bell Laboratoires en 1988. Ce logiciel sert `à manipuler des données, `à tracer des graphiques et `à faire des analyses statistiques sur ces données.

Le logiciel R est particulièrement performant pour la manipulation de données, le calcul et l'affichage de graphiques. Il possède, entre autres choses :

– un système de documentation intégré très bien conçu (en anglais) ;

– des procédures efficaces de traitement des données et des capacités de

stockage de ces données ;

– une suite d'opérateurs pour des calculs sur des tableaux et en particulier

sur des matrices ;

– une vaste et cohérente collection de procédures statistiques pour l'analyse

de données ;

– des capacités graphiques évoluées ;

– un langage de programmation simple et efficace intégrant les conditions,

les boucles, la récursivité, et des possibilités d'entrée-sortie.

➢ Syntaxe du langage R

R est à la fois un langage de programmation et un outil d'analyse de données. Un utilisateur n'est pas obligé de savoir programmer, mais il doit être capable d'utiliser des fonctions. En gros, toutes les commandes à taper font appel à des fonctions. [29]

La syntaxe de R ressemble superficiellement à celle des langages de type "C" et il s'agit d'une implémentation du langage S avec des extensions sémantiques dans la direction du langage Schème.

1.3 Le package rattle

➢ Définition

Rattle est une interface graphique populaire pour l'exploration de données en utilisant R. Il présente des résumés statistiques et visuels de données, transforme les données pour qu'elles puissent être facilement modélisées, construit des modèles d'apprentissage automatique supervisés et non supervisés à partir des données, présente les performances des modèles de manière graphique et attribue de nouveaux jeux de données à déployer en production. L'une des principales caractéristiques est que toutes vos interactions via l'interface utilisateur graphique sont capturées sous la forme d'un script R pouvant être facilement exécuté en mode R indépendamment de l'interface Rattle. [29]

➢ Installation de rattle

Rattle est un paquet R et s'installe (en principe) facilement. Pour chaque opération, Rattle risque de demander l'installation d'autres paquets .la plupart s'installer facilement
Install.packages(« rattle »)

➢ Execution de rattle

Library(rattle)
Rattle()

➢ Principaux packages

Beaucoup de packages supplémentaires à installer, parmi les principaux :

- **GTK+** : permet la construction d'interfaces graphiques pour l'utilisateur

- **RGtk2**: installations supplémentaires pour les interfaces graphiques de programmation
- **Stringr**: facilite la manipulation de chaîne de caractères
- **XML**: utile dans la lecture et la création de fichiers XML
- **CairoDevice**: permet d'afficher des graphiques à l'écran et de les enregistrer dans un fichier ou en mémoire

2. Processus de l'application [30]

✓ importation des données

- Dans l'onglet « Data », nous cliquons sur le bouton « Nom du fichier » pour spécifier le fichier de données.

- Nous spécifions le séparateur : « SEPARATOR = . ». Puis nous cliquons sur le bouton EXECUTER.
  Les données sont chargées, les variables sont typées (numérique ou catégorielle) automatiquement.
- Rattle dénombre le nombre de valeurs distinctes pour chaque variable (« Comment »). Cela devrait nous aider à identifier leur type réel. Nous spécifions leur statut dans l'analyse. Nous distinguons principalement les variables prédictives (INPUT) et la variable cible (TARGET).
- Enfin, nous partitionnons les données en échantillons d'apprentissage (70%) et de test (30%) avec l'option PARTITION. Notons qu'il est possible de subdiviser les données en 3 parties :

➢ apprentissage,
➢ validation
➢ test.

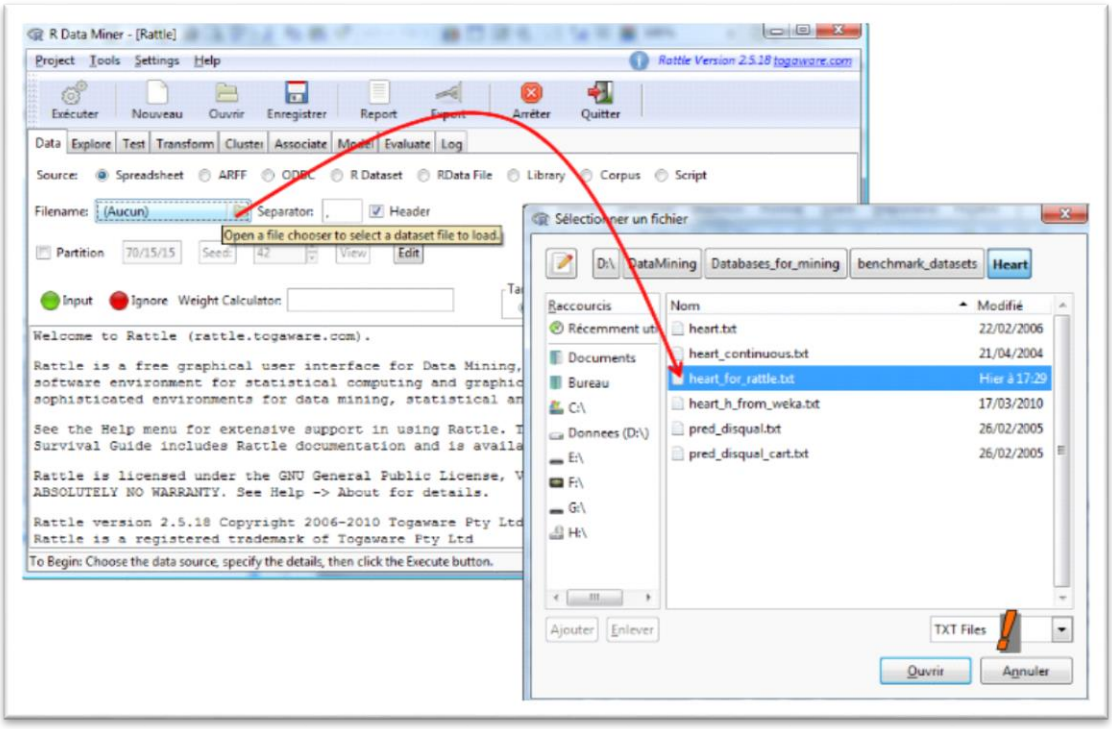


Figure 38: importation des données *dans* Rattle

✓ Description des données

Données Explorer Test Transformer Cluster Associer Model Evaluer Journal

Type : Résumé Distributions Corrélation Composantes principales Interactif

Résumé Décrire De base Coefficient d'aplatissement Coefficient d'asymétrie Afficher les valeurs manquantes Tableaux croisés

```
Le jeu de données est résumé ci-dessous.

Les données sont limitées au jeu de données de formation.
 Notez que les données contiennent 3 observations avec des données manquantes.
Cochez la case 'Afficher les données manquantes' pour plus de détails.

Data frame:crs$dataset[crs$train, c(crs$input, crs$risk, crs$target)]   34 observations and 9 variables    Maximum # NAs:3


             Levels Storage NAs
sexe              2 integer   3
note_math           integer   0
note_physique       integer   0
note_arabe          integer   0
moyenne              double   0
orientation       3 integer   0
situation         2 integer   0
nom              44 integer   3
admi              2 integer   0

+-----------+-----------------------------------------------------------------------+
|Variable   |Levels                                                                 |
+-----------+-----------------------------------------------------------------------+
|sexe       |femme,homme                                                            |
+-----------+-----------------------------------------------------------------------+
|orientation|ajournee,literaire,scientifique                                        |
+-----------+-----------------------------------------------------------------------+
|situation  |BIEN,PROB                                                              |
+-----------+-----------------------------------------------------------------------+
```

Figure 39: *Description des données*

L'onglet « Explore » est dédié à la description des données. Nous obtenons l'énumération des valeurs des variables catégorielles, ainsi que leur distribution de fréquences. Pour les variables quantitatives, nous avons les quartiles et la moyenne. Tous les indicateurs sont calculés sur l'échantillon d'apprentissage.

Avec l'option SUMMARY / DESCRIBE, la description est plus détaillée. Nous obtenons le nombre de valeurs distinctes, les déciles ainsi que les 5 plus grandes et les 5 plus petites valeurs pour les variables quantitatives. Ces informations sont précieuses pour détecter rapidement la présence de valeurs aberrantes dans notre échantillon.

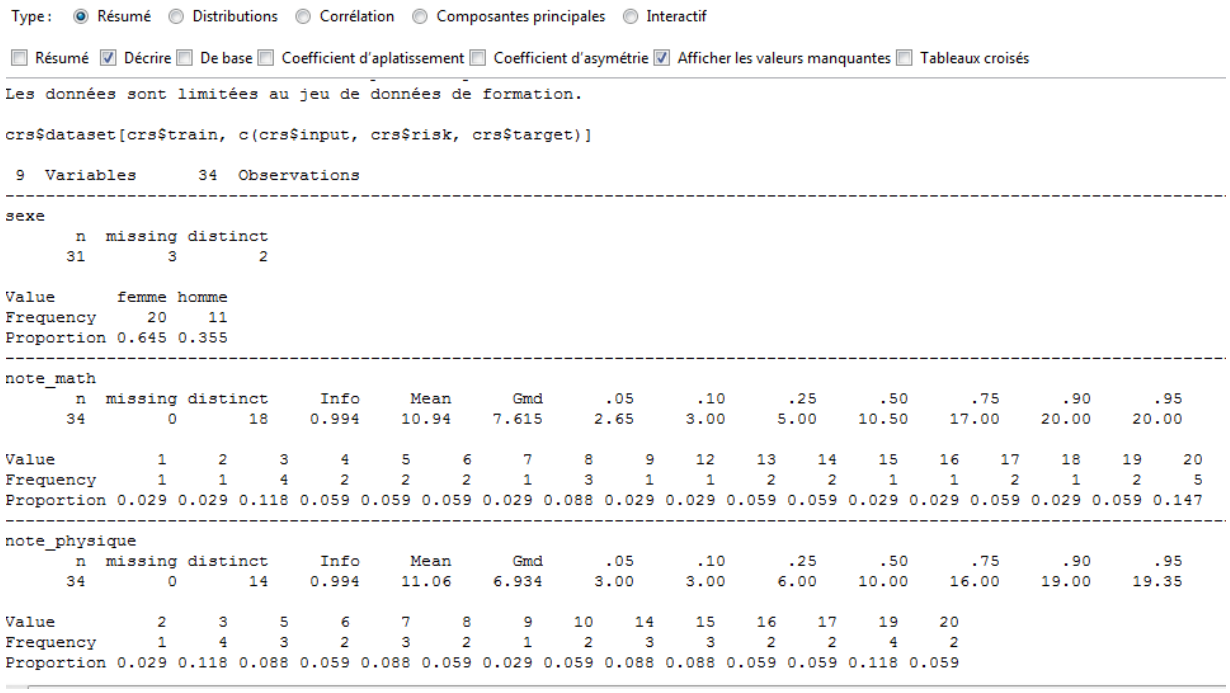


Figure 40 : *Résultat de l'onglet explore*

Toujours dans l'onglet « Explore », lorsque nous passons à l'outil DISTRIBUTIONS, nous avons accès aux outils graphiques. Nous souhaitons par exemple obtenir les boîtes à moustaches de la variable 'moyenne' globalement et conditionnellement aux valeurs de la variable à prédire 'situation' :

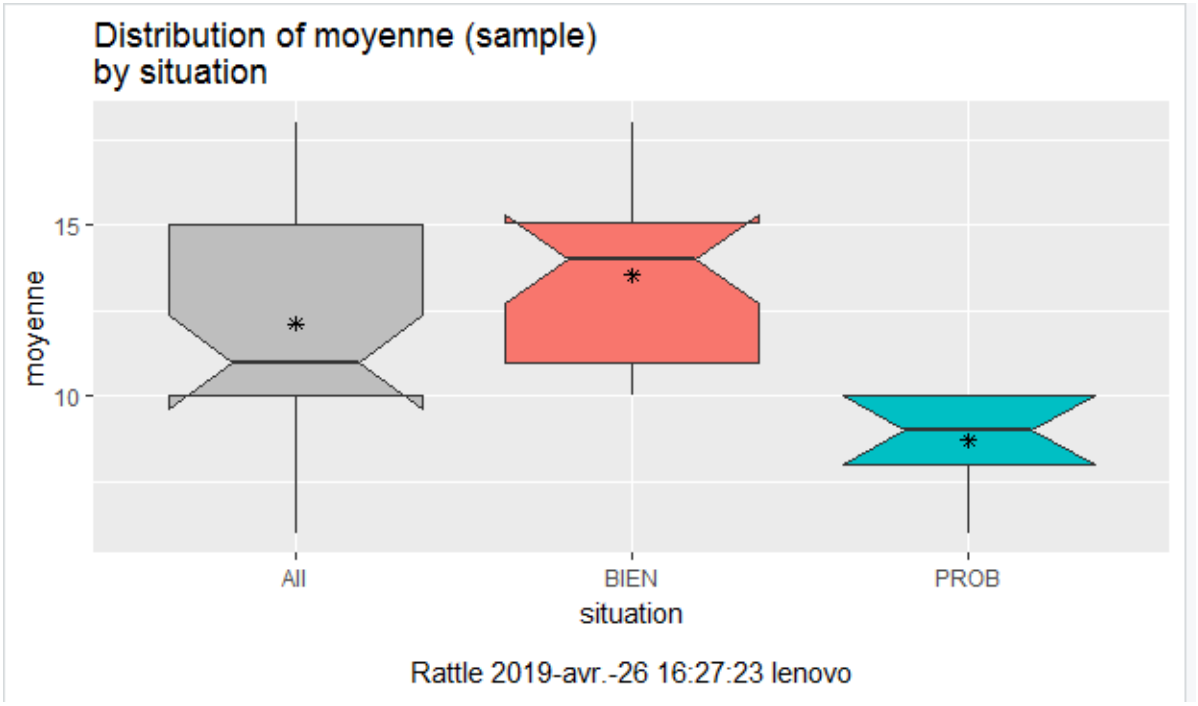


Figure 41 : *Résultat de l'onglet distribution 'graphique en boite'*

Comme nous pouvons aussi avoir une autre représentation graphique 'en barre' comme le montre la figure suivante.

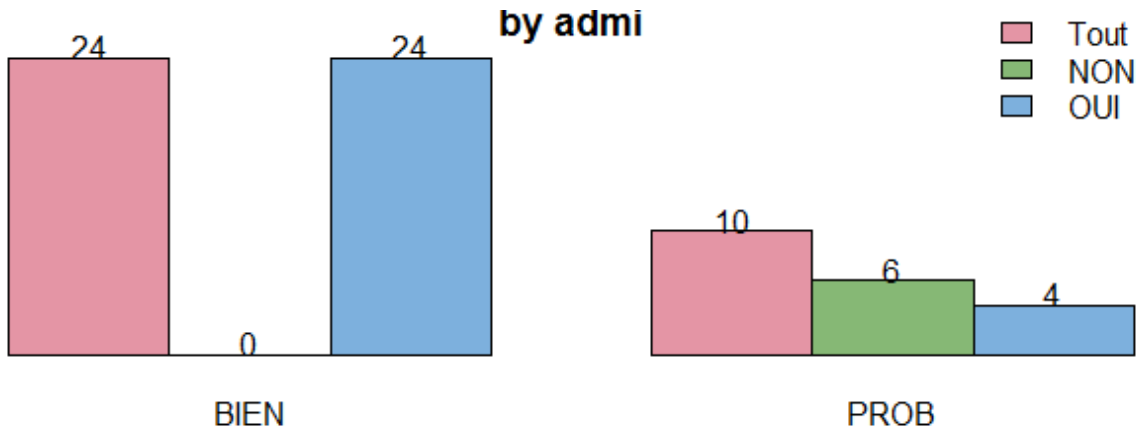


Figure 42: *Résultat de l'onglet distribution 'graphique en barre'*

En un coup d'œil, nous obtenons une série d'informations très instructives par exemple l'ensemble des élèves qui ont une bonne situation a la maison sont admis ; leur moyenne varient entre (10 et 15) contrairement aux élèves qui n'ont pas une bonne situation a la maison ou qui souffrent de problèmes (pauvreté, parents divorcés …) se sont eux les élèves qui échouent à l'école.

- ✓ Nous pouvons obtenir une représentation graphique de l'arbre de décision en actionnant le bouton DRAW.

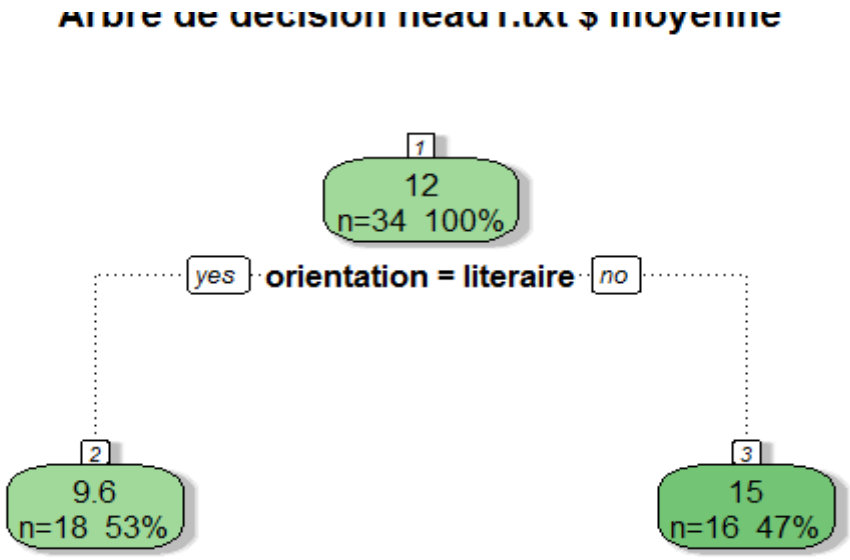


Figure 43: *Arbre de décision sous R selon l'orientation*

Avec cet arbre de décision c'est facile de faire l'évaluation et apprendre que le nombre d'élèves orientés littéraire et de 18éléves ce qui fais 53% du nombre totale.

- La figure qui suit montre un autre type d'arbre de décision avec plus d'informations car on peut conclure le nombre et le pourcentages d'élèves qui ont une moyenne inférieur à 15 et dans notre cas 26éléves ou 76% des élèves ; on peut aussi associer cette information a leur note de mathématique .comme on peut voir le nombre d'élèves qui ont une note de mathématique inférieur à 11 c'est 14éléves parmi 26 de là on comprends que c'est la note de mathématique qui a rabaisser la moyenne des élèves.

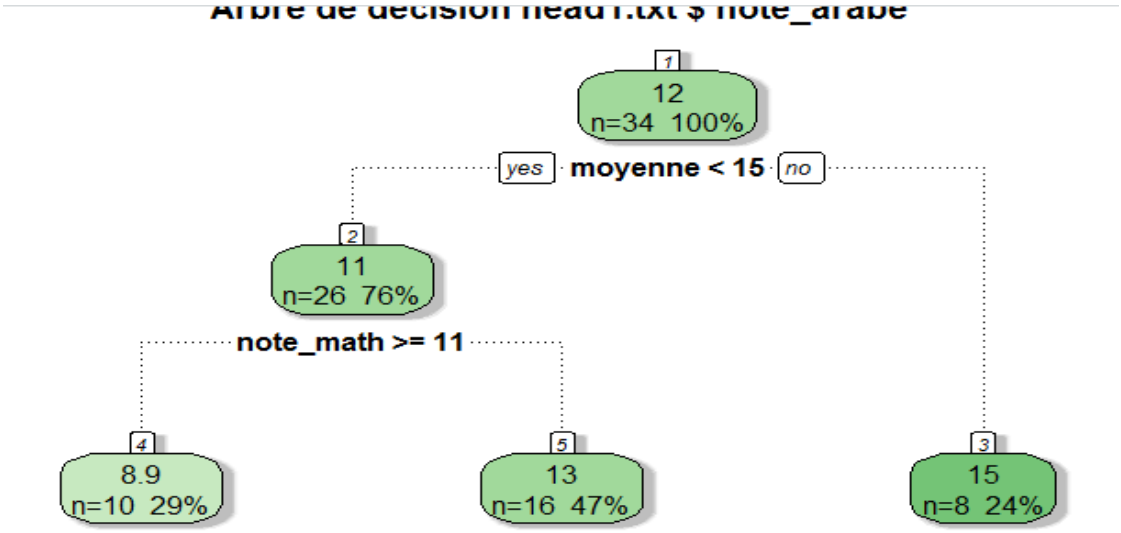


Figure 44: *Arbre de décision sous Rselon la moyenne et la note des mathématiques*

III. Discussion

L'intérêt croissant pour l'extraction de connaissances utiles à partir de données dans le but d'être bénéfique pour le propriétaire des données donne lieu à de multiples outils d'exploration de données ; dans notre projet on a eu l'occasion de travailler avec les deux outils weka et rattle gui, dans ce qui suit on a opter pour faire une étude comparatifs entre les deux logiciels en se basant sur les points forts et faibles de chaqu'un d'eux .

Remarque

Les points forts et faibles décrits ci-dessous sont limités à notre compréhension et nos constations issues de nos travaux pour ce projets.

Critères	**Weka**	**Rproject**
Gratuité	Gratuit et open source	Gratuit et open source
Documentation	Une variété de documentation est disponible.	Peu de documentation sur internet.
Ergonomie	Ergonomie et lisibilité limitée et pas évidente pour une prise en main par les débutants.	Résultats facile à interpréter, graphiques de qualité de publication.
Types de données traitées	Gestion de données CSV calamiteux.	Uniquement quantitative et qualitatives et binaires (pas de traitement, de données textuelle)
Analyse de données	Pas très puissant dans quelques techniques telles que l'analyse des clusters	Analyse descriptive assez complète (illustration graphiques, modalités statistiques descriptives …)
Disponibilité	Disponible sur Windows, macintosh et linux	Disponible sur Windows, macintosh et linux
Capacité de données	Peut rencontrer des problèmes de traitement en cas de grande quantité de données	Grande capacité de gérer les bases de données

Conclusion

Le développement de notre application nous a nécessité un ensemble d'outils et une bibliothèque puissante (weka) qui nous a servi à sa création.

Dans ce chapitre nous avons présenté ces outils, cette bibliothèque et les langages de développement utilisés pour la réalisation de notre application et présenter quelques-unes de ses principales interfaces et l'implémentation de quelque algorithme de Data mining et effectuer quelque test. Puis faire une comparaison avec l'outil rattle de R.

Notre travail a été consacré pour réaliser un outil informatique dans le but de l'extraction de connaissances en éducation, ça nous permet l'analyse des résultats des élèves du collège par extraction des connaissances nouvelles et utiles pour les utilisateurs.

Nous avons débuté notre projet par une recherche bibliographique, qui étais une phase très importante .Elle nous a permis d'avoir une idée sur le concept d'extraction de connaissances et notamment dans l'éducation, les différents algorithmes du data mining existant ainsi que leur domaine d'utilisation.

Nous avons ensuite fais une étude complète et détaillée sur les techniques et outils du data mning les plus récents en ayant une idée sur les avantages et inconvénients de chaqu'un ; on a aussi cité les différents besoin des algorithmes du data mining en éducation et on a répondu à ces besoin par l'utilisation de l'outil weka après avoir tenté de le faire avec l'outil R et affronter plusieurs obstacles comme expliqué dans notre mémoire.

Enfin nous avons présenté les différentes étapes de la réalisation de notre projet ainsi que les différentes technologies utilisées.

Au terme de notre projet nous avons pu :

- approfondir nos connaissances acquises durant notre cursus notamment sur le data mining.
- D'apprendre beaucoup de notions et techniques sur le développement sous Netbeans avec java.
- Utiliser l'API weka via l'environnement java.
- Présenter les différents algorithmes du data mining et proposer quelques tests en utilisant (J48, kmeans, Apriori)

En guise de perspectifs nous proposons :

- ✓ Utilisation des nouveaux outils de data mining pour une meilleure exploitation des données.
- ✓ améliorer notre application pour que les algorithmes proposés répondent à tous les besoins des utilisateurs.

Webographie

[1]http://www.daviddlewis.com/resources/testcollections/reuters21578/

[2]https://www.ibm.com/support/knowledgecenter/fr/SSCVKV_10.0.0/Campaign/CampaignProcesses/Extract_process

[3] https://www.commentcamarche.net/faq/43812-le-big-data-les-avantages-pour-l-entreprise

[4] https://www.memoireonline.com/08/13/7308/m_Analyse-et-detection-de-lattrition-dans-une-entreprise-de-telecommunication20.html

[5] Wikipédia -apprentissage non supervisé

[6]https://fr.wikiversity.org/wiki/Datamining/Logiciels

[7]https://fr.slideshare.net/doniahammami/techniques-du-data-mining

[8] https://jafwin.com/2019/01/14/top-5-des-outils-les-plus-utilises-en-data-mining/

[9] kmeans clustering- Wikipédia

[10] https://hufr.wordpress.com/2015/10/23/logiciels-open-source-de-fouille-de-donnees-data-mining/

[11] https://www.ionos.fr/digitalguide/web-marketing/analyse-web/outils-de-data-mining-comparaison/

[12] http://www.ordinateur.cc/systèmes/Compétences-informatiques-de-base/

[13] http://cybertim.timone.univ-mrs.fr/enseignement/doc enseignement/informatique/introdatawarehouse/docpeda_fichier

[14] Wikipédia-extraction de connaissance en éducation

[15] WWW. Wikipédia.com

[16]https://cran.r-project.org/web/packages/rattle/vignettes/rattle.pdf

Bibliographie

[17] Dr. Abdelhamid DJEFFAL, Cours Fouille de données avancée, Master 2 IDM, Année Universitaire 2012/2013

[18] Data mining-Mohamed NEMICHE0 Faculté des Sciences d'Agadir(2014/2015) Master MASI

[19]ATMANI Saliha et BELAIDI Lynda : « Extraction de connaissances à partir de résultat d'évaluation d'apprenants en utilisant le data mining ».Mémoire de Master de l'université Mouloud Mammeri TiziOuzou. 2012/2013.

[20] MENOUAR tarek et DERMOUCHE Mohamed:« Application de techniques de Data mining pour la classification automatique des données et la recherche d'associations » Mémoires de fin d'études ESi, 2009/2010

[21] BRAHIMI belkacem : « Extraction de connaissances à partir de données

incomplètes et imprécises » mémoire de magister 2011.

[22] M. J. BERRY, G. S. LINOFF, Data Mining Techniques For Marketing, Sales, and Customer Relationship, Management, Second Edition, 2004.

[23] S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, New Age International (P) Ltd., Publishers, New Delhi, 2007.

[24] G. CALAS, Études des principaux algorithmes de data mining, Spécialisation Sciences Cognitives et Informatique Avancée, France.

[25] Data mining-Mohamed NEMICHE0 Faculté des Sciences d'Agadir(2014/2015) Master MASI

[26] LE DATA MINING (APPLIQUÉ AUX DONNÉES D'APPRENTISSAGE): L'ÉQUILIBRE ENTRE ÉTHIQUE ET EFFICACITÉ

[27] cours 1 ère année M1 LMD –S2 UH BC-faculté des sciences et science de l'ingénieur – département TCT

[28] Pascal Roques, « UML2 par la pratique, Etude de cas et exercice corriges », 5 éme Edition, Edition EYROLLES

[29] R and Data Mining: Examples and Case Studies- Yanchang Zhao-octobre 2015

[30] DataMining sous R – Le package « rattle ».