REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOULOUD MAMMERI DE TIZI OUZOU
FACULTE DE GENIE ELECTRIQUE ET D'INFORMATIQUE
DEPARTEMENT D'INFORMATIQUE

Mémoire

De Fin d'études

En vue de l'obtention du diplôme de master en informatique

Option : Conduite De projet Informatique

Thème

Application des outils du datamining pour l'analyse des erreurs des apprenants en Learning by doing

Proposé et Dirigé par :

M^eme Bouarab-Dahmani.F

Réalisé par :

M^elle Medjnoun Chabha
M^elle Messad Djamila

Promotion 2011/2012

Remerciements

Au terme de ce travail,

Nous tenons en premier lieu à remercier le Bon Dieu pour le courage et la patience qu'il nous a donné afin de mener ce projet à terme.

Nous remercions vivement notre promotrice Mme. Bouarab farida d'avoir accepté de nous encadrer et de nous avoir proposé ce thème, ainsi que pour son aide précieuse, son travail encourageant, et ses conseils judicieux.

Nous remercions également nos parents et tous les enseignants qui ont contribué à notre formation.

Sans oublier de remercier les membres de jury d'avoir accepter de juger notre travail.

Enfin, nos remerciements vont à tous ceux qui, de loin ou de prés ont aidé à l'élaboration de ce mémoire.

Dédicaces

Je dédie ce modeste travail :

- A ceux qui m'ont tout donné sans rien attendre en retour mis à part ma réussite, à ceux qui m'ont appris à aller au bout de Mes ambitions, à ceux qui ont toujours cru en moi : à mes très

Chers parents.

--A mes adorables chères frères : Salim et Jugurtha,

-A mes adorables chères soeurs : Nabila et son époux Belaid, Malika et son époux Makhlouf, Sara et Katia.

- A mes très chers amis : Kahina, Djaffar , kamilia ,Belaid Ouissem,Nawel…

-A tout mes amis(es).

-A ma collègue Djamila et toutes sa famille.

Chabha

Dédicaces

Je dédie ce modeste travail

- A la mémoire de mon grand père, que dieu bénisse son âme.
- A mes chers parents

En témoignage de ma profonde gratitude et de mon incontestable reconnaissance, pour tous les sacrifices qu'ils me contentent, toute la confiance qu'ils m'accordent et tout l'amour dont ils m'entourent.

- A mes chers frères et sœurs

Zahia, Sadia, ahmed, brahim et que je ne trouverais jamais assez de mots pour leurs exprimer mon amour, En leurs espérant le plein succès.

- A la famille

Ma grand-mère maternel, mes chères tantes et oncles à qui je souhaite le parfait bonheur.

- A mon binôme chabha et à toute sa famille.
- Aux amis

Chères amies ouissem et nawel que je remercie pour leur soutien et leurs encouragements, à tous les collègues du Master II, à mes amis et à tous ceux qui me sont chers.

Djamila

TABLE DES FIGURES

Figure1* : *Data mining, union de discipline varies* *04

***Figure2 : Le processus de data mining* 04**

***Figure 3* : *Interventions des types d'évaluation pendant une année scolaire* 39**

***Figure 4* : *Diagramme entités/associations* 42**

***Figure 5*: *Processus de transformation de données d'apprentissage* 49**

***Figure 6*: *La fenêtre weka GUI chooser*..... 55**

***Figure 7 : Fenêtre Explorer*..... 58**

***Figure 8: Le fichier arff associé à la table input_base* 59**

***Figure 9 : L'onglet classify de l'explorer weka*..... 61**

***Figure 10 : L'arbre de décision associé aux résultats de la classification*..... 66**

***Figure 11 : Page d'accueil Websiela3* 76**

***Figure 12 : Espace datamining*..... 76**

***Figure 13 : Choix des résultats de la session à consulter* 77**

***Figure 14 : résultat du test de la session 1* 77**

***Figure 15 : résultat de test de la session 1 pour les règles d'association* 78**

SOMMAIRE

Introduction générale

CHAPITRE I : Le data mining

I.1 Introduction **01**

I.2 Qu'est-ce que le Datamining **01**

I.3 Pourquoi le Datamining **01**

I.4 Les raisons du développement du Datamining **02**

I.5 Quelles sortes de données **02**

I.6 Pratique du datamining **04**

I.7 Le processus du Data Mining **04**

I.7.1 Poser le problème **05**

I.7.2 La recherche des données **05**

I.7.3 La sélection des données pertinentes **05**

I.7.4 Le nettoyage des données **05**

I.7.5 Les actions sur les variables **06**

I.7.6 La recherche du modèle **06**

I.7.7 L'évaluation du résultat **07**

I.7.8 L'intégration de la connaissance **07**

I.8 Techniques de Data Mining **08**

I.8.1 ***Apprentissage supervisé*** **08**

I.8.1.1 Arbres de décision **08**

I.8.1.2 Les réseaux de neurones **10**

I.8.2 ***Apprentissage non supervisé*** **10**

I.8.2.1 Clustering **11**

I.8.2.2 Les règles d'associations **11**

I.8.3 ***Apprentissage incrémental*** **12**

I.9 Exploitation du Data Mining 13

I.9.1 Logiciels de data mining **..... 13**

I.9.2 Domaines d'application du data mining **..... 13**

I.9.2.1 Le data mining dans le secteur bancaire 13

I.9.2.2 Le data mining dans la bio-informatique et la biotechnologie 14

I.9.2.3 Le data mining dans la détection de fraude 14

I.9.2.4 Le data mining dans le secteur des assurances 14

I.9.2.5 Le data mining dans la télécommunication 14

I.9.2.6 Le data mining dans la médecine et la pharmacie 15

I.9.2.7 Le data mining dans le e-commerce et le world wide web 15

I.9.3 ***Avantages du data mining*** **..... 15**

I.9.4 ***Inconvénient du data mining*** **..... 16**

I.10 Conclusion 16

CHAPITRE II: Data mining en e-learning

II.1 Introduction 17

II.2 E-learning 17

II.3 Etude du data mining en e-learning du point de vue data mining 18

II.3.1 Le problème de classification en e-learning 20

II.3.2 Méthode de la logique floue 20

II.3.3 Réseaux de neurones artificiels et calcul évolutionniste 21

II.3.4 Graphes et arbres 21

II.3.5 Règles d'association 22

II.3.6 Le problème de clustering en e-Learning 23

II.3.7 Autres problèmes de data mining dans l'e-Learning 25

II.4 Etude du data mining en e-learning du point de vue e-Learning 27

II.5 Le processus de data mining en e-learning 28

II.6 Discussion et opportunité pour l'utilisation du data mining dans les systèmes d'e-learning 30

II.6.1 Optimisation des logiciels en E-learning 31

II.6.2 Améliorations d'expérience des étudiants d'e-learning 31

II.6.3 Outils de support pour les tuteurs d'e-learning 32

II.7 Data Mining en e-learning au-delà des publications universitaires 33

II.8 Conclusion 35

CHAPITRE III : Data mining pour le suivi des apprenants

III.1 Introduction 36

III. 2 Evaluation, diagnostic et erreurs 36

III.2.1 Les erreurs dans l'évaluation 36

III.2.2 Diagnostic 37

III.2.3 Evaluation des apprenants 38

III.3 Evaluation des apprenants en utilisant les techniques du data mining 40

III.4 Le cas de l'analyse de résultats d'évaluations de type ODALA 41

III.4.1 Modélisation des résultats de l'évaluation 41

III.4.2 Classification des apprenants 44

III.4.3 Déduction de règles d'associations entre les erreurs 47

III.5 Le problème de formats de données 49

III.5.1 Le format CSV 49

III.5.2 Le format arff 50

III.6 Conclusion 53

CHAPITRE IV : Implémentation

IV.1 Introduction 54

IV.2 Utilisation de l'outil Weka 54

IV.2.1 Présentation de l'outil weka 54

IV.2.1.1 Description globale 54

IV.2.1.2 Format de l'ensemble de données d'apprentissage 56

IV.2.1.3 Le traitement Explorer proposés par Weka 58

IV.2.2 Application des algorithmes de datamining via l'outil weka 59

IV.2.2.1 *Classification des apprenants* 59

IV.2.2.2 *Recherche de règle d'associations entre les erreurs* 71

IV.2.3 Intégration de l'outil weka dans Web Siela 75

IV.3 Utilisation de l'environnement java 79

IV.3.1 Outils de programmations 79

IV.3.2 Application de l'applet 79

IV.4 Conclusion 86

Conclusion générale

Introduction générale

L'un des enjeux de la recherche sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH) est la personnalisation de l'apprentissage, cette personnalisation débute par la conception de système d'enseignement, appliquant des stratégies définies pour suivre, aider l'apprenant durant son apprentissage en évaluant ses connaissances, ses aptitudes et en détectant et analysant ses erreurs afin de suivre la progression de son niveau d'apprentissage et ses capacités d'acquisition, l'évaluation peut être effectuée en se basant sur le modèle de l'apprenant et par historisation des sessions d'apprentissages. L'évaluation doit se concentrer sur les stratégies d'enseignements afin de faire évoluer l'apprentissage des apprenants.

C'est à ce niveau que s'inscrit notre travail, qui vise à intégrer les techniques de data mining pour le suivi des apprenants sous la plate-forme « websiela3 ».

L'objectif de notre travail est de mettre en place les techniques pour gérer, au niveau de la base de données des apprenants et contenus pédagogiques, les corrélations, dépendances, associations, modèles, classes, pour former des groupes homogènes et proposer de nouvelles décisions comme l'adaptation de contenu par rapport aux résultats obtenus.

L'évaluation dans le contexte de notre travail consiste à analyser les réponses et résultats des apprenants, ainsi que leurs classement par historisation de sessions d'apprentissage, le suivi des apprenants sur lequel nous avons fait nos tests se base principalement sur le nombre d'erreurs commises par un apprenant ou sur son niveau d'acquisition et de compréhension.

Pour réaliser ce travail, nous avons organisé notre mémoire de la manière suivante :

- Le premier chapitre présente le data mining, son processus, ces différentes techniques, ainsi que les domaines concernés par ce concept.
- Le deuxième chapitre est la phase intermédiaire entre le data mining et l'e-learning, ou les techniques d'exploration de données ont été intégrés aux problèmes e-learning.
- Le troisième chapitre c'est la partie conception qui présente l'évaluation au service d'apprentissage, l'évaluation avec quelques techniques de data mining notamment les entités concernés par cette évaluation.
- Le quatrième chapitre c'est réalisation qui débute par le prototypage sous l'outil *weka* qui est relatif à la réalisation de quelques algorithmes de data mining sur certaines données, et se finaliser par l'utilisation de l'environnement java pour monter la portabilité et l'avantage de ces algorithmes.

Chapitre 1 : Le data mining

I.1 Introduction

La notion de **data mining** est aussi connu sous le nom« d'exploration de données ». Le data mining est souvent utilisé pour définir le comportement type d'un consommateur en supermarché, par exemple. Mais qu'est-ce que c'est exactement ?

Afin de présenter au mieux la notion de data mining, nous avons divisé ce dossier en quatre grandes parties, la première partie approfondira la notion de data mining. S'en suivra le processus du data mining, puis une liste explicative des différentes méthodes exploitées en data mining, avec l'utilité de chacune. Ensuite, nous présenterons concrètement quelles sont les utilisations possibles du data mining.

I.2 Qu'est-ce que le data mining (Fouille de données)?

Le data mining est un nouveau champ situé au croisement de la statistique et des technologies de l'information (bases de données, intelligence artificielle, apprentissage etc.) dont le but est de découvrir des structures, des connaissances « intéressante » et des modèles dans de vastes ensembles de données.

On peut le définir aussi comme étant l'ensemble des algorithmes et méthodes destinés à l'exploration et l'analyse de (souvent) grandes bases de données informatiques en vue de détecter dans ces données des règles, des associations, des tendances inconnues (non fixées apriori), des structures particulières restituant de façon concise l'essentiel de l'information utile pour l'aide à la déc ision[1].

I.3 Pourquoi le data mining?

Problème : L'explosion des données

- Les outils de collecte automatique des données et les bases de données conduisent à d'énormes masses de données stockées dans des entrepôts [2].
- La submergé des données, donc manque de connaissance!

Solution: Entrepôts de données et fouille de données

- Entrepôts de données et analyse on-line
- Extraction de la connaissance intéressante (règles, contraintes) à partir de grandes bases de données.

I.4 Les raisons du développement du data mining

Pour quoi ça s'est développé ?

- Intérêt économique : du produit aux clients.
- Technologie de l'information : faible coût de stockage de données, saisie automatique de transaction (code bar, click, données de localisation GPS, internet)
- Augmentation de la puissance de calculs des ordinateurs.

Donc extraire de la connaissance à partir de grande bases de données devient possible.

Objectifs du data mining

- Développer des techniques et systèmes efficaces et extensibles pour l'exploration de BD larges et multidimensionnelles.
- faciliter l'utilisation des systèmes de data mining.
  - Limiter l'intervention de l'utilisateur
  - Représentation simple de la connaissance
  - Visualisation sous forme exploitable.

I.5 Quelles sortes de données

Le data mining n'est pas spécifique à un type de médias ou de données. Il est applicable à n'importe quel type d'information.

- ***Fichiers plats*** : Sont des fichiers de données simples dans le format texte ou binaire avec une structure connue par l'algorithme du data mining qui va t être y appliqué.
- ***Bases de données relationnelles*** : Les algorithmes du data mining appliqués sur des bases de données relationnelles sont plus polyvalents que les algorithmes 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 des données.
- ***Entrepôts de données*** (data warehouses) : Ensemble de données historiées variant dans le temps, organisé par sujets, consolidé dans une base de données unique, géré dans un environnement de stockage unique particulier, aidant à la prise de décision dans l'entreprise.
- ***Datamart*** (Magasin de données) sous-ensemble de données extrait du data warehouse et ciblé sur un sujet unique particulier, aidant à la prise de décision dans l'entreprise

- ***Bases de données transactionnelles*** : En général, une base de données transactionnelle est un fichier où chaque enregistrement représente une transaction. Une transaction contient un identifiant unique de transaction (*transactionID*) 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'identifiant du consommateur, l'identifiant de la personne qui a vendu, et ainsi de suite. **[3]**
- **Les bases de données multimédia**

  Les bases de données multimédia comportent des documents sonores, des vidéos, des images et des médias en textes et audio. Elles peuvent être stockées sur des bases de données orientées objets ou objets relationnelles ou simplement sur un fichier système. Le multimédia est caractérisé par sa haute dimension ce qui rend le data mining sur ce type de données très difficile. Le data mining sur les supports des multimédias exige la vision par ordinateur, l'infographie, l'interprétation des images et les méthodologies de traitement de langages naturels
- ***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 **[4].**
- ***Le World Wide Web***

  Le World Wide Web est le support de données le plus hétérogène et le plus dynamique disponible, 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 garantie 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 **[4]**

I.6 Pratique du data mining

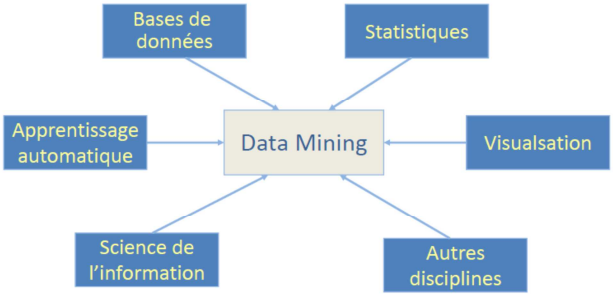


Figure1: Data mining, union de discipline varies

Le data mining utilise plusieurs autres disciplines. A la fin de tout le processus, le but est de prendre des décisions, après une analyse rationnelle. Plusieurs flux de données sont utilisées, et diverses techniques sont appliquées dans le but d'apporter au décideur, en partant de simples données pas très parlantes, à des connaissances suffisantes pour effectuer des choix techniques, commerciaux, de gestions, …

I.7 Le processus du data mining

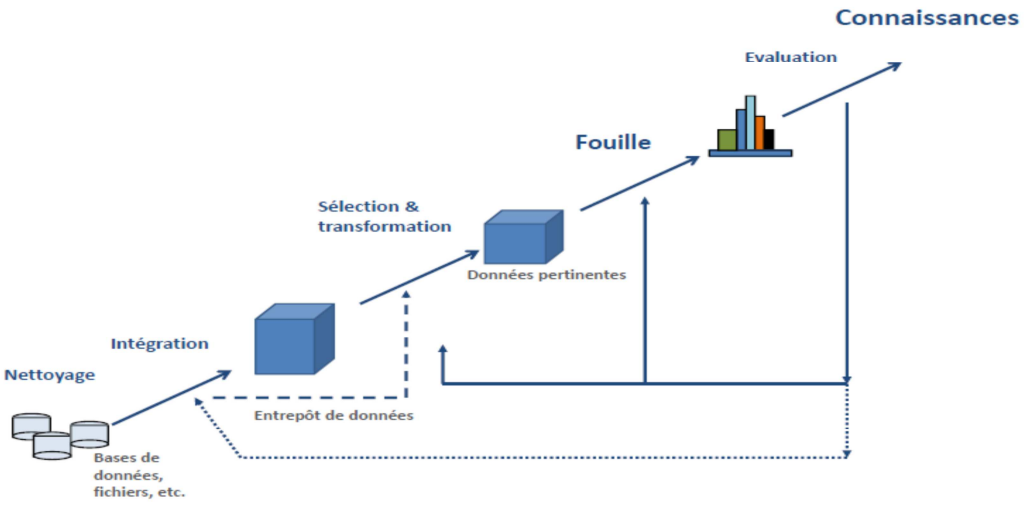


Figure2: Le processus de data mining

I.7.1 Poser le problème:

- Exposer le problème et définir les objectifs. Comme il faut connaître la typologie du problème (affectation ou structuration).
- Ayant défini le type de problème, on doit bien savoir ce qu'on attend comme résultat et ce l'exploitation qu'on en fera.

Ces dernières connaissances faciliteront les choix à effectuer dans les étapes suivantes.

I.7.2 La recherche des données :

- Déterminer la structure générale des données ainsi que les règles utilisées pour les constituer.
- La sélection des données doit être optimale et peut nécessiter la consultation d'experts, afin de déterminer les attributs les plus aptes à décrire la problématique.

En suite, grâce à des taxonomies, il faudra réduire le nombre des variables obtenues pour faciliter une généralisation du problème. Cette étape peut fortement conditionner la qualité des résultats du processus de data mining.

I.7.3 La sélection des données pertinentes :

- Pour effectuer une collecte et une sélection de données. Il faut constituer une base d'informations qui permet de construire l'apprentissage, c'est à dire la construction de modèles en recherchant dans le passé des évènements similaires **[2].**
- Ce travail peut nécessiter l'intervention de toute une équipe et sera plus ou moins facilité selon les technologies en place dans l'entreprise (base de données ouverte, entrepôt de données exhaustif, …).
- La sélection des données peut aboutir sur un échantillon ou une exhaustivité de données qui seront ensuite nettoyées.

I.7.4 Le nettoyage des données :

- La phase de nettoyage des données permet d'améliorer la qualité des données afin de minimiser l'effet d'anomalies telles que des erreurs de saisie, des champs nuls.
- Les modalités de contrôle de l'origine des données dépendent de la taille de la base d'exemples (importante ou restreinte) et de son type d'alimentation (automatique ou manuelle).

- La recherche des valeurs aberrantes peut être effectuée en isolant les pics de certaines valeurs dans une distribution statistique, ou en utilisant d'autres méthodes comme la détermination de score.
- Les valeurs manquantes sont gérées, soit en excluant les enregistrements incomplets, en remplaçant les données manquantes, ou en les gérant via des algorithmes précis.
- Une analyse est effectuée pour déceler l'existence d'enregistrements totalement nuls. Elle permet d'en identifier les causes externes possibles panne de capteurs, saut de lignes par l'agent de saisie, …)

Afin d'avoir un modèle performant et faciliter l'apprentissage, il faut améliorer la qualité des données par l'utilisation de bruits ou des processus flous. Ces opérations, tous comme les précédentes permettent d'obtenir des données fiables.

I.7.5 Les actions sur les variables :

Une fois les variables sont pertinentes, et que les données sont fiables, il faut les transformer pour préparer le travail d'analyse. Il s'agit d'intervenir sur des variables pour qu'elles soient mieux exploitables par des outils de modélisation. Ces transformations peuvent être de plusieurs types :

La transformation mono variable :

Lorsqu'on veut améliorer une seule variable, par normalisation ou transformation logarithmique. Il est aussi important de changer les dates en durées pour faciliter le travail de modélisation. Si on a affaire à des coordonnées géographiques, l'utilisation de géocodage ou de logiciels d'information géographique peut être nécessaire afin de rendre des coordonnées plus significatives.

La transformation multi variable :

Concerne la combinaison de plusieurs variables élémentaires en une nouvelle variable agrégée. En effet, les données brutes sont parfois insuffisantes pour apporter un pouvoir prédictif à un modèle. Les types de transformation sont multiples. On peut utiliser la fréquence, les combinaisons linéaires, les combinaisons non linéaires, …

I.7.6 La recherche du modèle :

Après avoir obtenu des variables réduites, on passe à la phase de modélisation. Elle consiste à extraire des données à partir d'un volume de données bruitées et à la présenter sous une forme synthétique. Elle est parfois décrie sous le terme de data mining.

La recherche du modèle se fait dans la phase d'apprentissage sur une base d'apprentissage qui doit être distincte de la base de tests .La construction de ce modèle peut se faire de manière automatique et interactive. Sa performance dépend du choix d'algorithmes de calculs.

Parmi les techniques de modélisation utilisables, citons trois groupes :

- La recherche des modèles à base d'équations.
- L'analyse logique.
- Les techniques de projection.

Quelque soit la précision du modèle, sa précision devra être vérifié par une évaluation.

I.7.7 L'évaluation du résultat :

L'évaluation du résultat permet d'estimer la qualité du modèle, c'est-à-dire sa capacité à déterminer correctement les valeurs qu'il est censé avoir apprises sur des cas nouveaux. Cette évaluation prend généralement une forme qualitative et une forme quantitative.

- **L'évaluation qualitative :** permet d'illustrer le poids ou l'influence d'un facteur. Elle peut se faire sous une forme graphique. Dans ce cas, elle améliore la compréhension des résultats.
- **L'évaluation quantitative :** utilise des techniques et notions telles que l'intervalle de confiance, pour fiabiliser les conclusions apportées sur des données futures.
- **La validation par des tests :** Après avoir construit un modèle, il est possible d'en tester la pertinence sur la base d'apprentissage. Cela étant, il faut éviter d' « apprendre » les données plutôt que le modèle. Il vaudra mieux constituer une base de test ne servant qu'au test. De cette manière on vérifiera que le modèle est capable classer convenablement les données qu'il n'a jamais rencontrée. La stabilité entre les résultats observés sur le fichier d'apprentissage et le fichier test constitue la capacité d'apprentissage.

I.7.8 L'intégration de la connaissance :

La connaissance n'est rien tant qu'elle n'est pas convertie en décision, puis en action. Il est essentiel d'implanter le modèle et ses résultats dans les systèmes informatiques ou dans les processus de l'entreprise. Cette intégration peut se faire soit sous la forme de données (résultat du modèle) ou sous la forme d'un traitement (algorithme du modèle).

C'est dans cette dernière phase qu'il faut dresser un bilan du déroulement des étapes précédentes. Ce bilan sert à améliorer l'existant en termes de données et de collecte de données.

Les étapes précédentes illustrent, les étapes suivies pour faire du data mining. Ce pendant, comme nous l'avons vu à l'étape 6, pour construire un modèle, des techniques propres à la discipline de data mining sont utilisés.

I.8 Techniques de data mining

Le data mining est un ensemble de techniques et de méthodes du domaine des statistiques, des mathématiques et de l'informatique permettant l'extraction, à partir d'un important volume de données brutes, des connaissances originales auparavant inconnues. Il s'agit de fouilles visant à découvrir l'information cachée que les données renferment et que l'on découvre à la recherche d'associations, de tendances, de relations ou de régularités.

Ces techniques sont classées de la manière suivante :

I.8.1 Apprentissage supervisé

En sciences cognitives, l'apprentissage supervisé est une technique d'apprentissage automatique plus connu sous le terme anglais de machine learning qui permet à une machine d'apprendre à réaliser des tâches à partir d'une base d'apprentissage contenant des exemples déjà traités. Chaque élément (item) de l'ensemble d'apprentissage (training set) étant un couple entrée-sortie.

De part sa nature, l'apprentissage supervisé concerne essentiellement les méthodes de classification de données (on connaît l'entrée et l'on veut déterminer la sortie) et de régression (on connaît la sortie et l'on veut retrouver l'entrée). **[5]**

I.8.1.1 Arbres de décision **:** Comme son nom le suggère, un outil d'aide à la décision. C'est une technique de classification automatisée, l'analyse d'un ensemble d'enregistrements préalablement classifiés permet de générer une structure arborescente optimale pour la classification des autres enregistrements disponibles [6].

Un arbre de décision permet de répartir une population d'individus en groupes homogènes selon des attributs discriminants en fonction d'un objectif fixé et connu. Un arbre de décision est composé d'un nœud racine, un ensemble de nœuds internes et un ensemble de feuilles. Chaque nœud interne de l'arbre représente un attribut, les nœuds

externes appelés feuilles, représentent les classes d'affectation, les arcs entre les nœuds représentent quant à eux les tests sur ces derniers. Les nœuds internes d'un arbre de décision possèdent un seul nœud parent, et deux ou plusieurs nœuds descendants **[4].**

Il est très utilisé en marketing, ces méthodes sont devenues un des outils les plus populaires du data mining. On peut les utiliser pour prédire une variable quantitative (on parle alors d'arbre de régression) ou qualitative (il s'agit d'arbre de décision, de classification ou de segmentation).

Pour parcourir un arbre de décision et trouver une solution il faut partir de la racine. Chaque nœud est une décision atomique, chaque réponse possible est prise en compte et permet de se diriger vers un des fils du nœud, de proche en proche, on descend dans l'arbre jusqu'a tomber sur une feuille. La feuille représente la réponse qu'apporte l'arbre au cas que l'on vient de tester.

- **Principe**

Pour chaque nœud, on choisit la variable qui, par ses catégories, sépare le mieux les individus en fonction des catégories de la variable cible. Le choix du nœud est donc en fonction du choix des branches du nœud. On réitère l'opération jusqu'à ce qu'elle ne soit plus possible ou plus souhaitable (selon des critères dépendant du type d'arbre).

Les feuilles ainsi construites sont toutes majoritairement constitués d'individus d'une seule classe. La branche allant de la racine à une feuille constitue une règle de classement. L'ensemble des règles constitue le modèle.

- **Les algorithmes existants : *ID3*, C4.5, C5.0**

Il existe plusieurs algorithmes pour fabriquer des arbres de décision.

Citons particulièrement **[7]:**

- *ID3 (Inductive Decision Tree, Quinlan 1979)* Construit un arbre de décision de façon récursive en choisissant l'attribut qui maxime le gain d'information selon l'entropie. Cet algorithme traite uniquement les variables qualitatives, un nœud est créé pour chaque valeur des attributs sélectionnés. ID3 est un algorithme basique facile à implémenter dont la première fonction est de remplacer les experts dans la construction d'un arbre de décision.
- *C4.5 (Quinlan 1993)* est une amélioration d'ID3 qui permet de travailler à la fois avec des données discrètes et des données continues. Il permet également de travailler avec des valeurs d'attribut absentes (valeurs manquantes). Un dernier élément de performance de

C4.5 réside dans l'élagage de l'arbre construit afin de supprimer les règles inutiles et de rendre l'arbre plus compact.

- *C5.0 (Quinlan 1998)* L'arbre C5.0 est un perfectionnement par le chercheur australien J.RossQuinlan de ses précédents arbres ID3 (1986) et C4.5(1993), expliqués précédemment. L'algorithme C5.0 fonctionne en cherchant à maximiser le gain d'information réalisé en affectant chaque individu à une branche de l'arbre.

L'originalité de C5.0 comme le C4.5 est qu'il est accompagné d'une procédure transformant les arbres en ensemble de règles. Les règles redondantes sont supprimées, ce qui réduit la complexité de l'ensemble des règles.

C5.0 généralise les règles ne menant pas à une diminution du taux d'erreur. C5.0 n'est pas binaire du fait de son traitement des variables qualitatives

Remarque : Les arbres de décision sont capables de produire des règles compréhensibles et effectuent la classification sans exiger beaucoup de calcul, mais manque de performance dans le cas de plusieurs classes; les arbres deviennent très complexes et ne sont pas nécessairement optimaux.

I.8.1.2 Les réseaux de neurones (issue de l'intelligence artificielle) c'est un ensemble de traitements qui peuvent être des unités soit d'entrée soit de sortie ou bien cachées. Un réseau de neurones est composé de plusieurs neurones interconnectés. Un poids est associé à chaque arc. A chaque neurone on associe une valeur. C'est une technique de classification automatisée. De même que pour les arbres de décision, le principe consiste à apprendre à correctement classifier des données à partir d'un jeu d'exemples déjà classifiés. Présentement, les champs d'information des enregistrements forment les entrées d'un réseau dont la sortie correspond à la classe de l'enregistrement. L'apprentissage consiste alors à faire passer les enregistrements classés en entrée du réseau, et à corriger un petit peu l'erreur fatalement obtenue en sortie en modifiant les nœuds internes du réseau. Au fur et à mesure, celui-ci s'adapte, et finit par classer correctement les enregistrements.
Généralement les réseaux de neurones ne sont pas souvent utilisées dans les tâches du data mining parce qu'ils produisent des modèles souvent incompréhensibles et demande un longtemps d'apprentissage **[3]**

I.8.2 Apprentissage non supervisé

On parle d'apprentissage non supervisé lorsque l'on cherche à extraire des informations nouvelles et originales d'un ensemble de données dont aucun attribut n'est plus important

qu'un autre. Le résultat des algorithmes de data mining non supervisé doit être analysé afin d'être retenu pour un usage ou tout simplement rejeté.

I.8.2.1 Clustering (segmentation)

Le clustering est une méthode statistique d'analyse de données qui a pour but de regrouper un ensemble de données en différents groupes homogènes.

Chaque sous-ensemble regroupe des éléments ayant des caractéristiques communes qui correspondent à des critères de proximité.

Le but des algorithmes de clustering est donc de minimiser la distance intra-classe (grappes d'éléments homogènes) et de maximiser la distance interclasse afin d'obtenir des sous-ensembles le plus distincts possible.

La mesure des distances est un élément prépondérant pour la qualité de l'algorithme de clustering.

- *L'algorithme des k-moyennes (K-Means)*

L'algorithme des K-moyennes est dédié aux tâches de clustering, il permet de diviser une population donnée en K groupes homogènes appelés clusters. Le nombre de clusters K est déterminé par l'utilisateur selon ses attentes.

I.8.2.2 Les règles d'associations : Provenant de bases de données est à l'origine utilisée pour faire l'analyse de séquences et qui décrivent des associations entre certains éléments. Cette technique permet de faire ressortir les associations entre les produits de base (les produits essentiels, ceux pour lesquels le client se déplace) et les produits complémentaires, ce qui permet de mettre en place des stratégies commerciales visant à accroître les profits en favorisant, par exemple, les ventes complémentaires. Elle permet, dans ce cas, de savoir quels sont les produits achetés simultanément, dans un supermarché par exemple, par un très grand nombre de clients ; elle est aussi appliquée pour résoudre des problèmes d'analyse de parcours de navigation de sites web. La recherche de règles d'association peut être utilisée de manière supervisée.

La méthode peut être appliquée à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services : services bancaires, services de télécommunications, par exemple. Elle peut être également utilisée dans le secteur médical pour la recherche de complications dues à des associations de médicaments ou à la recherche de fraudes en recherchant des associations inhabituelles

Une *règle d'association* est de la forme : Si **condition** alors **résultat**. Dans la pratique, nous nous limitons généralement à des règles où la condition se présente sous la forme d'une conjonction d'apparition d'articles et le résultat se constitue d'un seul article. Par

exemple, une règle à trois articles sera de la forme : Si X et Y alors Z ; règle dont la sémantique peu être énoncée : Si les articles X et Y apparaissent simultanément dans un achat alors l'article Z apparaît [3].

Les résultats de cette méthode sont clairs :

- règles faciles à interpréter.
- Simplicité de la méthode et des calculs (calculs élémentaires des fréquences d'apparition).
- Aucune hypothèse préalable (Apprentissage non supervisé).

Mais cette méthode est coûteuse en temps de calcul, ainsi la production d'un nombre important de règles qui peuvent être par fois évidentes, mais, n'apportent pas d'information (ou inutiles), elle est non efficace pour les articles rares

Ils existent plusieurs algorithmes pour cette technique le plus important est :

- *L'algorithme apriori*

Apriori est un algorithme classique de recherche de règles d'association. Comme tous les algorithmes de découvertes d'associations, il travaille sur des bases de données transactionnelles (des enregistrements de transactions).

Il repose sur le fait qu'un ensemble fréquent d'éléments a pour sous-ensembles des ensembles fréquents d'éléments. En effet si l'ensemble {A, B} est fréquent dans la base, les ensembles {A} et {B} sont eux-mêmes fréquents dans la base.

I.8.3 Apprentissage incrémental

L'apprentissage incrémental permet à une machine d'apprendre par ajout successif d'informations. Pour être considéré comme tel, un système d'apprentissage doit :

- Etre capable d'apprendre de nouvelles informations à partir de nouvelles données.
- Etre capable de se passer des données d'origine pour entraîner le nouveau classifieur.
- Préserver le savoir précédemment acquis.
- Etre capable de reconnaître de nouvelles classes introduites dans les nouvelles données.

I.9 Exploitation du data mining

I.9.1 Logiciels de data mining

La fouille de données n'existerait pas sans outils. Ces outils sont les logiciels commerciaux, et les logiciels libres, tout dépend de ce qu'on veut en faire. Les logiciels commerciaux sont destinés aux entreprises, ou aux organismes ayant de gros volumes de données à explorer, tandis que les logiciels libres sont destinés plus particulièrement aux étudiants, à ceux qui veulent expérimenter des techniques nouvelles, et aux PME.

les outils les plus utilisés sont, dans l'ordre, **SPSS**, **RapidMiner**, **SAS**, **Excel**, **R**, **KXEN**, **Weka**, **Matlab**, **Knime**, **Microsoft SQL Server**, **Oracle DM** et **STATISTICA**.

STATISTICA, IBM SPSS Modeler, et R ont reçu les taux de satisfaction les plus élevés à la fois en 2010 et 2009.

I.9.2 Domaines d'application du data mining

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 :

I.9.2.1 Le data mining dans le secteur bancaire

Le secteur bancaire est à la tête de tous les autres domaines industriels grâce 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".
- 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.

I.9.2.2 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.

I.9.2.3 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.
- 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.

I.9.2.4 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'assurances.

I.9.2.5 Le data mining dans la télécommunication

À nos jours, toute activité de télécommunication a utilisé une 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.

I.9.2.6 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.
- 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.

I.9.2.7 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 de 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 de données, le Web mining peut être fait par les façons susdites.

I.9.3 Avantages du data mining

Pour mieux cerner l'utilité du data mining citons ces principaux avantages :

- Le data mining aide à la prise de décision. Par l'analyse des données, la méthode peut synthétiser la situation et alors accélérer la prise de décision des dirigeants à un problème donné, mais le data mining ne remplace pas ces dirigeants.
- Le data mining permet de faire des liens pertinents entre des données qui, auparavant, n'ont aucune corrélation.
- Améliorer la satisfaction des clients en analysant leur besoins et en proposant des améliorations en fonction des évènements passés.
- Facilite le développement de nouveaux produits.
- Accélère la gestion des stocks, des inventaires, de la logistique.
- Peut augmenter les revenus tout en diminuant les coûts.

Le data mining a été étudié pour augmenter et optimiser la productivité et l'intérêt d'une entreprise ou l'amélioration d'un critère. Cependant cette méthode souffre de quelques défauts qui sont présentés ci-après.

I.9.4 Inconvénient du data mining

- Taille de la base : le data mining est totalement dépendant de la base de données qu'il analyse et donc des méthodes et technologies qui permettent d'y accéder
  - le stockage des données requiert de très grands espaces. Il doit, souvent se faire sur une machine spécifique.
  - le temps de transfert des données entre la base de données et le poste de travail augmente la durée des traitements.
- Sujets d'analyse

Dans l'exemple d'analyse d'un site web, la structure de celui-ci est régulièrement modifiée et rend donc une analyse précise difficile.

- L'excès de confiance

Le data mining permettait de prévoir l'avenir, d'anticiper la réaction des clients par rapport à une modification, donc il est prudent de recouper les informations obtenues avec d'autres études, statistiques et autres, et de disposer d'un statisticien et d'un commercial pour l'analyse des résultats

- La nécessité de disposer de personnel qualifié

Le data mining reste un processus complexe qui nécessite du personnel habitué à l'utiliser. Dans le cas minimum, il faudrait disposer d'un informaticien, d'un commercial connaissant bien la clientèle ainsi que d'un statisticien.

I.10 Conclusion

Nous avons pu comprendre à travers ce rapport le concept du data mining, et de préciser les utilisations possibles. Nous avons remarqué que cette discipline est pratiquement présente dans toutes les applications quotidiennes. Les grandes distributions sont les cas les plus concernés par cette méthode, Les outils du data mining peuvent prédire les futurs tendances et actions, permettant de prendre les bonnes décisions. C'est ce qui rend le data mining la technologie la plus importante nous allons détaillées plus dans la prochaine section ce concept mais cette fois seulement en e-learning.

Chapitre 2 : Le data mining en e-learning

II.1 Introduction

L'utilisation d'internet à des fins d'enseignements de multiples formes se développe actuellement dans tous les secteurs de l'éducation et particulièrement dans le monde universitaire. Cette nouvelle modalité d'enseignement offre des avantages dont celui notamment de faciliter l'enseignement à distance **[9]**. Dans un rapport statistique du National Center for Education, il apparaît que les formations à distance aux Etats-Unis connaissent une explosion fulgurante pour les institutions d'enseignements postsecondaires. Le nombre d'institutions offrant des cursus à distance dans l'enseignement public est passé de 58% en 1997-1998 à plus de 90% en 2000-2001 avec une multiplication par deux du nombre d'étudiants en parallèle. Un autre rapport (U.S. Departement of Education, National Center for 1999-2000) utilisait Internet pour suivre leurs cours et ce chiffre atteint près de 80% pour les étudiants en master. Cette utilisation massive d'internet dans l'enseignement à distance n'est pas étrangère à sa progression dans les universités américaines. En France, même si cette impression repose sur des données plus subjectives, de nombreuses universités et écoles supérieures mettent en place des formations utilisant le réseau internet à des fins pédagogiques. Dans ce rapport nous faisons un petit aperçu montrant l'application des outils du data mining dans l'e-learning du point de vue e-learning et aussi du point de vue data mining.

II.2 E-learning

Comme il existe de nombreuses définitions disponibles, nous ne présentons ici que quelques-unes d'entre elles afin d'acquérir une certaine compréhension **[3].**

Définition 1

Tom Kelly, Cisco: "E-learning est un environ d'information, de communication, d'éducation et de formation. Indépendamment de la façon dont les formateurs classent la formation et l'éducation, l'apprenant ne veut que les compétences et les connaissances pour faire un meilleur emploi ou pour répondre à la question reçue à partir d'un client. "

Définition 2

Apprendre Circuits1: "E-learning couvre un large éventail d'applications et des processus tels que l'apprentissage basé sur le Web, apprentissage assisté par ordinateur, les classes virtuelles et la collaboration numérique. Il comprend la livraison de contenus via l'internet, intranet / extranet, cassettes audio et vidéo, de diffusion par satellite, TV interactive et CD-ROM. "

Introduire le data mining en e-learning?

L'offre de cours e-learning est désormais abondante, et de nombreuses nouvelles plates-formes e-learning et des systèmes ont été développés et implémentés avec des degrés de succès variables. Ces systèmes génèrent une quantité exponentielle croissante de données, et une grande partie de cette information peux avoir le potentiel de devenir une nouvelle connaissance pour améliorer le système e-learning. Le processus du data mining devrait permettre l'extraction de cette connaissance. Il n'existe pas de nombreuses implémentations réelles et pleinement opérationnelles pour l'intégration de data mining dans les systèmes E-learning. Du point de vue de problème E-learning la plupart d'offre de travail concerne l'évaluation de l'apprentissage des étudiants, le matériel d'apprentissage, l'évaluation des cours et l'adaptation de cours reposant sur le comportement des étudiants de l'apprentissage.

II.3 Etude du datamining en E-learning du point de vue datamining :

Notre intention est d'organiser les résultats de l'étude de différentes façons qui pourraient correspondre à la diversité des lecteurs universitaires aux milieux professionnels. Dans cette section, nous présentons les recherches interrogées en fonction des problèmes de data mining (classification, clustering, etc), Par conséquent, le data mining peut être utilisé pour extraire des connaissances à partir de systèmes E-learning à travers l'analyse des informations disponibles dans le formulaire, reconnaissez que ce n'est pas la première fois qu'une initiative similaire des techniques et méthodes (par exemple, réseaux de neurones, arbres de décision (DT)).

En fait, la plupart des recherches existantes porte sur les problèmes de classification et le clustering. Pour cette raison, des sections spécifiques seront consacrées à ceci. Mais d'abord, essayant de trouver une place pour le data mining dans le monde d'e-learning.

Où va s'adapter le Data Mining en e-learning?

Certains chercheurs ont souligné la relation étroite entre les domaines de l'Intelligence Artificielle (IA) et l'apprentissage de la machine ML (Learning Machine), les principales sources de techniques de data mining, les méthodes et les processus d'éducation.

Dans **[10]**, l'auteur établit les occasions de la recherche dans AI et l'éducation sur la base de trois modèles de processus pédagogiques:

- ***Modèles utilisés comme outil scientifique :*** qui est un moyen pour comprendre et prévoir quelques aspects d'une situation pédagogique.
- ***Modèles utilisés comme composant:*** correspond à quelques caractéristiques de processus d'enseignement ou de l'apprenant et utilisés comme un composant d'un objet éducatif.
- ***Modèles utilisés comme base pour dessin d'objets pédagogiques:*** associent avec l'outil de dessin de l'ordinateur pour l'éducation en fournissant des méthodologies du dessin et des composants du système, ou contraindre des outils qui peuvent être disponible aux apprentis.

Une étude sur la manière dont les techniques du data mining pourraient être incorporées avec succès aux environnements e_learning et comment ils peuvent améliorer les taches d'apprentissage générés. Autre étude, un cluster de donnée a été suggéré comme un moyen d'encourager le groupe-basée sur l'apprentissage collaboratif et de fournir le diagnostic sur l'étudiant.

Une révision des possibilités de l'application des techniques du Web mining (chapitre 1) pour avoir quelques défis courants dans l'e-learning a été présentée dans **[11]**. L'approche proposée pourrait améliorer l'efficacité et l'éducation de la distance dans deux chemins:

- d'un côté, la découverte des chemins individuels total pour les étudiants pourrait aider dans le développement d'une efficace éducation personnalisée, fournir une indication la mieux organisée du didacticiel (Logiciel à vocation pédagogique assisté par ordinateur) de l'organisation de l'éducateur.
- En revanche, la structure virtuelle de la connaissance pourrait être identifiée à travers les méthodes du Web mining: La découverte de règles d'association pourrait le rendre possible aux tuteurs afin d'identifier le modèle de la connaissance et de réorganiser le cours virtuel basé sur les modèles découverts.

Une analyse des techniques ML encore, une source commune pour l'exploitation de techniques de data mining a été utilisée pour automatiser la construction et l'énumération des modèles des apprenants. Aussi bien que la connaissance de l'origine nécessaire pour le modèle de l'apprenant, a été faite.

II.3.1 Le problème de classification en e-learning :

Dans les problèmes de classification, nous avons l'habitude de modéliser les relations existantes (le cas échéant) entre un ensemble d'éléments de données à plusieurs variables et un certain ensemble de résultats pour chacun d'eux sous la forme de classe d'appartenance. Bien que beaucoup de méthodes de classification qui s'inscriraient dans un processus d'exploration de données existent, dans ce qui suit, nous verrons que quelques techniques seulement (ou familles de techniques) ont été appliquées en E-learning.

II.3.2 Méthodes de la logique floue :

La logique floue est basée sur des méthodes qui ont récemment fait leurs premiers pas dans le champ d'E-learning.

Dans **[12]**, un modèle neuro flou pour l'évaluation des élèves dans un système intelligent de tutorat (ITS) a été présenté. La théorie floue a été utilisée pour mesurer et transformer l'interaction entre l'étudiant et son ITS en un terme linguistique. Puis, les réseaux artificiels de neurones ont été formés pour réaliser les relations floues opérées avec la composition max-min. Ces relations floues représentent l'estimation du degré d'association entre une réponse observée et une caractéristique des étudiants faite par les tuteurs.

Une approche floue d'un groupe de décision pour aider les utilisateurs et les experts du domaine dans l'évaluation des sites Web éducatifs qui ont été réalisé dans le système SIPA. Dans la suite des travaux de Hwang et ses collègues **[13],** une approche fondée sur des règles de méthodes floues pour intégrer la gestion du système de connaissance a été proposée et a servi de base pour la conception d'un système de gestion intelligent de surveillance de serveurs Web éducatif. Ce système est capable de manipuler les défaillances possibles dans les serveurs Web éducatif, l'amélioration de leur stabilité et leur fiabilité. Il assiste l'auto-évaluation des étudiants et leur fournit des suggestions basées sur les techniques de raisonnement flou.

Deux phases d'algorithme de data mining flou et de l'apprentissage ont été décrites.

1. Dans une première phase, Il intègre un algorithme de règle d'association, appelé *Apriori*, avec la théorie des ensembles flous pour trouver des informations intégrées qui pourraient être renvoyées à des enseignants pour le raffinage ou la réorganisation de matériel d'enseignement et de test.

2. Dans une deuxième phase, il utilise un algorithme d'apprentissage inductif de la famille AQ : AQR, pour trouver les descriptions des notions indiquant les concepts manquants lors de l'apprentissage des élèves. Les résultats de cette phase pourraient également être réinjectés aux enseignants pour le raffinage ou la réorganisation des parcours d'apprentissage.

II.3.3 Réseaux de neurones artificiels et calcul évolutionniste :

Quelques recherches sur l'utilisation des réseaux neuronaux artificiels et modèles évolutifs de calcul pour traiter des sujets e-learning peuvent être trouvées dans **[14], [15].**Un système de soutien de navigation basé sur un réseau de neurones artificiel a été mis en avant dans **[14].** Afin de décider sur les stratégies de navigations appropriées. Le réseau de neurone a été utilisé comme une stratégie de navigation dans le système.

L'évaluation a validé les connaissances acquises par le réseau neuronal et le niveau de l'efficacité de la stratégie de navigation.

Dans **[14]**, « les algorithmes évolutionnaires ont été utilisés pour évaluer les apprenants ». Une combinaison de classificateurs multiples (MCC), pour le classement des élèves et la prédiction de leurs notes finales, basé sur des caractéristiques extraites à partir des données enregistrées dans une éducation basée sur le système Web. La précision de la classification et de prévision sont améliorées grâce à la pondération des vecteurs de données en utilisant une fonction d'algorithme génétique. Dans **[16],** nous trouvons une génération de code aléatoire et le processus de mutation suggéré comme méthode pour examiner la capacité de compréhension des élèves.

II.3.4 Graphes et arbres :

La théorie des graphes et / ou des arbres a été appliquée à l'apprentissage électronique.

Dans **[17]**, un outil automatique, basé sur la performance d'apprentissage des apprenants et les préférences de communication, pour la génération et découverte d'apprenant simple, les modèles ont été décrits, avec le but ultime de créer un environnement éducatif personnalisé.

L'approche a été basée sur la partie algorithme qui produit des règles partielles élaguées de DTs. Dans **[18]**, un outil qui peut aider, comme trace des insuffisances dans la compréhension des apprenants a été présenté. Il a recours à un type d'arbre de données abstraites (ADT), construit à travers les concepts couvert dans un laboratoire, conférence, ou

cours. Une fois l'arbre qu'ADT est créé, chaque, Nœud peut être associé avec des différentes entités telles que la performance de l'apprenant, classe, performance, ou développement de laboratoire. En utilisant cet outil, un professeur pourrait aider des étudiants en découvrant les concepts qui ont besoin de couverture supplémentaire, pendant que les étudiants peuvent découvrir les concepts pour lesquels ils auraient besoin de passer le temps actif supplémentaire.

Un outil exécute une analyse quantitative basé sur la performance des étudiants d'apprentissage été introduit dans **[19],** Il propose le nouveau didacticiel (support pédagogique), décrit par un diagramme, en combinant des outils fournis par la théorie de cartes conceptuelles et diagrammes d'influence **[28]**. Dans **[28]**, les systèmes d'e-learning basés sur le Web personnalisés ont été définis, en appliquant l'usage des techniques de Web mining aux services de la recommandation personnalisée. L'approche est basée sur une méthode de classification de la page Web qui utilise l'énumération attribut-orientée.

II.3.5 Règles d'association :

Les règles d'association pour la classification, appliquée en e-learning, ont été étudiées dans les domaines des systèmes de recommandation de l'apprentissage, l'organisation matérielle d'apprentissage, évaluations de l'apprentissage des étudiants et l'adaptation de cours selon les comportements des étudiants **[21].**

Les techniques de data mining tels que les règles d'association, et l'extraction de motifs fréquents intersession (entre différentes sessions) et intra-session (dans une même session), ont été appliquées afin d'extraire des connaissances utiles qui pourraient aider les éducateurs, les gestionnaires de l'éducation, et les tuteurs web pour évaluer et interpréter les activités de cours en ligne. Une approche similaire peut être trouvée dans **[21]**, où les règles de contraste, définis comme des ensembles de règles conjonctives décrivant les modèles de la disparité de performance entre des groupes d'étudiants, ont été utilisés. Une approche assistée par ordinateur pour diagnostiquer les problèmes d'apprentissages des étudiants dans les cours de science (science naturelle) et d'offrir aux étudiants des conseils ont été présentés dans **[28]**, basée sur le modèle de la relation concept d'effet (CER) (une spécification de la technique de règle d'association).

Un environnement d'apprentissage hypermédia avec un composant tutoriel a été décrit. Il est appelé *Logiocando* et s'adresse aux enfants du quatrième niveau de l'école primaire (9-10 ans). Il comprend un module de tuteur, basée sur les règles si-alors, qui émule l'enseignant en fournissant des suggestions sur la façon et quoi étudier. Dans **[20]**, nous trouvons la description d'une méthode d'évaluation du processus d'apprentissage qui a recourt à des règles d'association, et bien connue méthode d'apprentissage (ID3 de DT). Un cadre pour l'utilisation du Web mining appuyé sur la validation des conceptions de sites d'apprentissage a été étudié, en appliquant l'association et les techniques de séquence.

Dans **[28],** un cadre d'e-learning personnalisé basé sur le total des profils d'utilisation et une ontologie de domaine ont été présentés, et une combinaison de méthodes d'extraction a été utilisée dans le web sémantique. L'algorithme de règle d'association *Apriori* a été appliqué pour capturer les relations entre les références URL basées sur les modèles de navigation des étudiants. Un modèle de rétroaction des résultats du test (TRF), modèle qui analyse les relations entre l'apprentissage des étudiants, le temps et les résultats du test correspondant a été introduit. L'objectif était double: **d'une part, de développer un outil pour soutenir le tuteur à réorganiser le matériel de cours**, de l'autre, **une personnalisation adaptée aux besoins individuels des étudiants**. L'approche était basée sur l'exploration des règles d'association.

Un mécanisme fondé sur des règles pour la génération adaptative de problèmes dans le contexte de la programmation des tuteurs basés sur la recommandation Web a été proposé dans **[22].**Un système basé sûr la recommandation web, utilisé pour fournir aux étudiants des suggestions lors de la difficulté dans le choix des cours, a été décrit. L'approche intègre l'algorithme apriori.

II.3.6 Le problème de clustering en e-learning :

Contrairement à des problèmes de classification, dans les données de clustering ou regroupement qui a comme objectif ***la modélisation d'une relation entre un ensemble d'éléments de données multi variées et un certain ensemble de résultats pour chacun d'eux*** (être présent sous la forme d'étiquettes d'appartenance à une classe). Au lieu de cela, nous avons pour but ***de découvrir les modèles des groupes dans lesquels les éléments de données sont souvent regroupés*** :

- ***Selon certain élément mesurant la similarité***

Nous trouvons une première application de méthodes de groupement, basé sur le réseau d'essai et de diagnostic du système a été mis en œuvre. Elle implique un ou plusieurs critères de test feuilles génératrices de problème et une approche de programmation dynamique, pour générer des feuilles de test. L'approche proposée emploie la théorie de la logique floue pour déterminer ***les niveaux de difficulté*** des items du test en fonction du ***statut d'apprentissage*** et les ***caractéristiques personnelles*** de chaque élève, puis appliquer un modèle de réseau neuronal artificiel: *Fuzzy Adaptive Resonance Theory* , et de regrouper les items du test en groupes, ainsi que la programmation dynamique **[23]** pour la construction de feuille de test. Une autre étude approfondie décrivant l'utilisabilité de réseaux de neurones artificiels pour l'évaluation des élèves dans un superviseur tutoriel (TS) du système, ainsi que la capacité d'un TS floue à s'adapter à la difficulté de question dans le processus d'évaluation, a été réalisée. Une enquête sur la façon dont les données et les techniques de data mining pourraient être intégrées avec succès à l'environnement E-learning, et comment cela pourrait améliorer le processus d'apprentissage qui a été présenté en **[23].** Ici, le regroupement des données est suggéré comme un moyen de promouvoir le groupe basé sur l'apprentissage collaboratif fournit un diagnostic différentiel des étudiants.

- ***Regroupement des utilisateurs en fonction de leurs comportements***

Des actions, associées à l'utilisation du Web par les étudiants ont été recueillis et prétraitées dans le cadre d'un processus de data mining. L'algorithme attente-maximisation (EM) a ensuite été utilisé pour regrouper les utilisateurs en grappes en fonction de leurs comportements. Ces résultats pourraient être utilisés par les enseignants et fournir des conseils spécialisés aux élèves appartenant à chaque groupe. L'hypothèse simplificatrice est que les élèves appartenant à chaque groupe devraient être partagés .Les comportements des stratégies d'utilisation du Web personnalisé devient plus évolutive. Les administrateurs système peuvent également bénéficier de ce savoir acquis en ajustant l'environnement e-Learning qu'ils gèrent. L'algorithme EM a été également la méthode de choix, où le regroupement a été utilisé pour découvrir des comportements d'utilisateur dans les activités de collaboration en applications E-learning .Certains chercheurs ,proposent l'utilisation de techniques de clustering au groupes de matériaux de cours similaires: un outil basé sur l'ontologie, au sein du cadre du Web sémantique, a été mis en œuvre **[24]** dans le but d'aider les utilisateurs d' e-Learning à trouver et organiser les ressources didacticiels distribués.

Un élément de cet outil a été l'application de l'algorithme bissection K-Means, utilisé pour le regroupement des matériaux similaires d'apprentissage. *Kohonen* bien connue algorithme SOM a été utilisé dans **[25]** pour concevoir un système intelligent d'outil de recherche au matériel d'apprentissage dans les classes du cluster similaires, basés sur ses similitudes sémantiques. Le Clustering a été proposé dans **[24]** à un groupe similaire des documents d'apprentissage en fonction de leurs sujets et les similitudes. Un document d'indice graphique (DIG) pour la représentation de document a été présenté, et certains algorithmes de classification classiques (Clustering et K-NN) ont été mises en œuvre.

Différentes variantes de modèle de la cartographie topographique générative (GTM), une alternative probabiliste pour SOM, ont été utilisées dans **[24]** pour le regroupement et la visualisation de données multi variées concernant le comportement des étudiants dans un cours virtuel. Plus précisément, une variante de GTM connu pour avoir un robuste en présence de données atypiques ou aberrantes a été utilisée avec succès pour identifier les grappes de comportements de l'apprentissage des étudiants atypiques. Une autre variante du GTM pour la détermination de la pertinence a été utilisée pour classer les caractéristiques des données disponibles en fonction de leur pertinence pour la définition des pôles des étudiants.

II.3.7 Autres problèmes de data mining en e-Learning :

Comme indiqué précédemment, la plupart des recherches actuelles portent sur les problèmes de classification (apprentissage supervisé) et de clustering (apprentissage non supervisé) dans l'environnement e-learning. Cependant, il existe plusieurs applications qui aborde d'autres problèmes tels que, la prévision et visualisation, ce que nous allons voir dans ce qui suit.

- **Techniques de prédiction :**

Le but de l'exploration de données est la description et la prédiction. Il existe de nombreuses stratégies dans l'exploration de données qui peut être conduit à la prédiction. L'un d'eux est la classification. *Donc on ne peut jamais prédire sans classifier*.

La prédiction est intéressante en E-learning :

- La prévision du comportement des élèves et des performances lors de l'utilisation des systèmes d'e-learning porte le potentiel pour faciliter l'amélioration des cours virtuels ainsi que l'e-learning dans les environnements général.

- Une méthodologie pour améliorer la performance des cours élaborés à travers l'adaptation présentée dans **[26].**
- Cours et fichiers stockés dans une base de données pourraient être exploités par des enseignants en utilisant des *algorithmes évolutionnaires* (sont une famille d'algorithmes s'inspirant de la théorie de l'évolution pour résoudre des problèmes divers. Ils font ainsi évoluer un ensemble de solutions à un problème donné, dans l'optique de trouver les meilleurs résultats), l'application des techniques de data mining en problèmes e-learning, la découverte des relations importantes (exemple : déduction de règles d'associations entre les erreurs dans le prochain chapitre), avec l'objectif de découvrir les relations entre les niveaux de connaissances des élèves, temps d'utilisation et les notes des élèves.
- trouver les sources d'erreurs dans la prédiction de comportements des connaissances des apprenants a été réalisé.
- La régression a été appliquée dans **[27]** pour prédire si la prochaine réponse de l'étudiant serait correct, et combien de temps il ou elle aurait pris pour générer cette réponse.
- Des expériences sur la prédiction des notes finale des élèves dans un cours de prise à un campus virtuel ont été réalisées. La pertinence relative aux caractéristiques spécifiques qui décrivent le comportement des apprenants sur les cours en ligne a également été évaluée. Ce travail a été étendu en utilisant des réseaux de neurones pour la prédiction des notes finale des élèves.
- la découverte de règles de prédictions intéressantes dans les informations utilisées par l'étudiant, afin de les utiliser pour améliorer l'adaptabilité des Cours sur le Web.

Un outil pour la détection automatique des comportements atypiques sur l'utilisation du système d'e-learning par les apprenants a été défini. Il a recourt à un modèle prédictive pour la détection de processus d'apprentissage irrégulier sur la base de temps de réponse des élèves.

- **Techniques de visualisation :**

C'est les phases la plus importante d'un processus de data mining (et qui est généralement négligée) est celui de l'exploration de données grâce à la méthode de visualisation.

- ***La Visualisation*** était entendue dans le contexte des réseaux d'analyse sociaux adaptée à *l'enseignement collaborative* à distance, où la cohésion de petits groupes d'apprentissage a été mesurée. La cohésion est calculée de plusieurs façons afin de mettre en évidence :
  - les personnes isolées,
  - sous-groupes actifs et divers rôles des membres de la structure du groupe de communication.

La méthode permet l'affichage des propriétés globales à la fois au niveau individuel et au niveau du groupe, ainsi que pour assister efficacement le tuteur virtuel en suivant les modes de *collaboration* (mode de sous groupes d'apprentissage) au sein du groupe.

Un outil éducatif de data mining est présenté, qui montre, de façon hiérarchique et partiellement ordonnée, l'interaction des étudiants avec l'environnement e-learning et de leurs tuteurs virtuels. L'outil fournit une analyse de cas et visualise les résultats dans un arbre d'événements, exploitant les bases de données MySQL pour obtenir des événements tutoriel.

II.4 Etude du data mining en e-learning du point de vue e-Learning

Dans ce point, nous présentons la recherche inspectée selon les problèmes d'apprentissage dans lesquels les méthodes de data mining sont appliquées.

Pour éviter des redondances inutiles, une étude est faite sur la littérature disponible afin de préciser les approches de data mining, selon les différents sujets e-learning abordés. Cette étude inclut les références sur l'un des sujets d'e-learning suivants :

- Applications portant sur *l'évaluation de la performance d'apprentissage des étudiants*.
- Les applications qui fournissent l'adaptation des cours et des recommandations d'apprentissage basées sur le comportement d'apprentissage des étudiants.
- Approches relatives à *l'évaluation du matériel d'apprentissage* et les cours d'enseignement basés sur le Web.
- Les applications qui impliquent des commentaires à la fois des enseignants et des étudiants concernant les cours d'E-learning, basées sur le comportement d'apprentissage des étudiants.
- Développements pour la *détection des comportements d'apprentissage* atypiques des étudiants.

***Adaptation des sujets E-learning** :*

On s'est inspiré du classement précédent des sujets E-learning, pour aboutir au résultat suivant :

Cours/apprenants : adaptation de cours en se basant sur le comportement des apprenants, les commentaires des apprenants sur les cours seront pris en considération.

Tuteurs/apprenants : améliorer l'expérience des étudiants d'e-learning, encore plus a porté des tâches aidant en ligne des tuteurs, y compris l'analyse et l'évaluation de la performance des étudiants.

Cours/tuteurs : évaluation du matériel didactique, commentaires des tuteurs concernant les cours.

L'un des sujets d'e-learning avec les résultats les moins obtenues dans cette étude est l'analyse du comportement d'apprentissage des étudiants atypiques.

II.5 Le processus de data mining en e-learning

Le développement traditionnel des cours e-learning est une activité laborieuse **[28]** dans lequel le développeur (habituellement l'enseignant bien sûr) doit :

- choisir le contenu qui sera affiché,
- prendre des décisions sur la structure du contenu,
- et déterminer les éléments de contenu les plus appropriés pour chaque type d'utilisateur potentiel du cours.

En raison de la complexité de ces décisions, il sera nécessaire dans la plupart des cas d'évaluer et éventuellement modifier le contenu du cours, la structure et la navigation basée sur l'utilisation d'informations des étudiants, de préférence même à la suite d'une approche d'évaluation empirique continue .

Pour faciliter cela, nous avons besoin de méthodes d'analyse des données et des outils pour :

- observer le comportement des étudiants
- aider les enseignants à détecter les éventuelles erreurs, les insuffisances et les améliorations possibles.

L'analyse traditionnelle des données en e-learning est l'hypothèse ou une supposition conduite dans le sens que l'utilisateur commence à partir d'une question et explore les données pour confirmer l'intuition. Même si cela peut être utile quand un nombre modéré de facteurs et de données sont impliqués, il peut être très difficile pour l'utilisateur de trouver des modèles plus complexes qui se rapportent aux différents aspects des données.

Une alternative à l'analyse des données traditionnelles consiste à utiliser le data mining comme une approche inductive pour découvrir automatiquement des informations cachées dans les données.

L'exploration de données construit des modèles analytiques qui découvrent des modèles intéressants et des tendances à partir des informations utilisés par l'étudiant qui peuvent être utilisés par l'enseignant pour améliorer l'apprentissage de l'étudiant et l'entretien du parcours.

- ***Application du data mining :***

L'application du data mining en systèmes E-learning est un cycle itératif dans lequel les connaissances extraites doivent inclure la boucle et un guide du système, afin de faciliter et d'améliorer l'apprentissage dans son ensemble, non seulement transformer des données en connaissances, mais une aide à la prise de décision par le filtrage des connaissances extraites. Pour le processus data mining en E-learning se compose des quatre mêmes étapes dans le processus de data mining générale comme suit:

- ***Recueillir des données***. Le système E-learning (LMS: Learning Management System) est utilisé par des étudiants et l'information d'interaction est stockée dans la base de données. (Exemple l'utilisation du système Moodle par les étudiants).
- ***Prétraitement des données***. Les données sont nettoyées et transformés en un format approprié pour être exploitées. Afin de prétraiter les données du système (exemple Moodle : système de gestion de cours d'apprentissage open source), nous pouvons utiliser un outil d'administration de base de données ou un outil de prétraitement spécifique.
- ***Appliquer l'exploration de données*** : Les algorithmes de data mining sont appliquées pour construire et exécuter le modèle qui découvre et résume les connaissances de l'intérêt de l'utilisateur (enseignant, étudiant, administrateur, etc.) Pour ce faire, nous pouvons utiliser un outil data mining.
- ***Interpréter, évaluer et déployer les résultats***. Les résultats ou les modèles obtenus sont interprétés et utilisés par l'enseignant pour d'autres actions. L'enseignant peut utiliser les informations découvertes pour prendre des décisions concernant les étudiants et les activités (de Moodle) du cours afin d'améliorer l'apprentissage des étudiants.

Nous pouvons voir que l'application du data mining dans l'E-learning n'est pas très différente que celle d'autres domaines d'applications. Cependant, il ya quelques questions importantes qui rendent l'exploration de données en E-learning différente que dans les autres :

- ***Données :*** Dans d'autres systèmes basés sur le web les données utilisées est normalement un journal d'accès au serveur web simple, mais en *E-learning* il ya beaucoup plus d'informations sur l'interaction de l'étudiant. Systèmes LMS pouvez enregistrer quel que soit les activités des étudiantes qui sont impliquées, comme la lecture, l'écriture, exécutions de plusieurs tâches, ils fournissent une base de données qui stocke toutes les informations système avec les informations personnelles des utilisateurs (profil) et les données d'interaction de l'utilisateur.

- ***Objectif*** : L'objectif de l'exploration de données dans chaque domaine d'application est différent et peut être plus ou moins objective.
  Par exemple, dans l'e-commerce l'objectif est d'accroître leurs profits, ce qui est tangible et peut être mesurée en termes de mentant d'argent, le nombre de clients et fidélisation de la clientèle. Mais *l'objectif de l'exploration de données en E-learning est d'améliorer le processus d'apprentissage et de guider les étudiants dans leur apprentissage par exemple : le taux de réussite, taux d'assimilation*… Cet objectif est plus subjectif et plus difficile à mesurer.

- ***Techniques :*** Les systèmes éducatifs ont des caractéristiques qui nécessitent un traitement différent de problème. Quelques techniques traditionnelles peuvent être appliquées directement, mais d'autres ne peuvent pas et peuvent être adaptées aux problèmes éducatifs spécifiques. De plus, quelques techniques du data mining spécifiques peuvent être utilisées spécifiquement pour diriger le processus d'apprentissage.

II.6 Discussion et opportunité pour l'utilisation du Data Mining dans les systèmes d'E-Learning

Dans ce qui suit, nous analysons en détail un peu plus l'état actuel de la recherche en data mining appliquées en E-Learning, en soulignant ses perspectives d'avenir et les opportunités, ainsi que ses limites qu'on peut caractériser comme suit :

II.6.1 Optimisation des logiciels en e-learning

La possibilité de suivre le comportement des utilisateurs dans les environnements virtuels E-learning rend possible l'extraction des bases de données qui en résultent. Cela ouvre de nouvelles possibilités pour les concepteurs pédagogiques et didactiques qui créent et organisent les contenus d'apprentissages.

Afin d'améliorer le contenu et l'organisation des ressources de cours virtuels, les méthodes d'exploration de données concernées par l'évaluation du matériel d'apprentissage, pourraient être utilisées. Les problèmes de classification sont dominants dans ce domaine, même si la prédiction et le clustering sont également présents.

Certaines publications (journal, conférence…etc.) pourraient aussi indirectement servir à améliorer les ressources de cours, dans le cas où on peut conclure (mais pas toujours) que *Si l'évaluation des étudiants n'était pas satisfaisante, elle pourrait laisser entendre sur le fait que les ressources de cours et du matériel d'apprentissage sont insuffisantes*.

Les méthodes de data mining appliquées pour évaluer le matériel d'apprentissage dans un cours e-learning, comprennent: Techniques de règles d'association, et les techniques de clustering.

II.6.2 Améliorations d'expérience des étudiants d'e-learning

L'un des objectifs les plus importants dans l'e-learning, et l'un de ses majeurs défis est l'amélioration de l'expérience e-learning des étudiants inscrits à un cours virtuel.

Plusieurs publications [**28**] ont porté sur l'auto-évaluation, recommandation des stratégies d'apprentissage, l'adaptation de cours utilisateurs en fonction du profil et des nécessités de l'étudiant. Divers modèles d'exploration de données ont été appliquées à ces problèmes, y compris les règles d'associations, réseaux neuronaux, arbres de décision...

L'application des techniques de data mining (exploration de texte ou de l'exploration du web) pour analyser les journaux Web, afin de découvrir des modèles de navigation utiles, ou déduire des hypothèses qui peuvent être utilisés pour améliorer les applications web, est l'idée principale derrière l'utilisation du Web mining.

Les méthodes de clustering et de visualisation pourraient améliorer l'expérience d'e-learning :

- la segmentation permet de regrouper les acteurs similaires en fonction de leurs similitudes
- la visualisation permet de décrire et explorer ces groupes de manière intuitive.

Les Combinaisons de méthodes d'exploration de données ont démontré leur potentiel dans des environnements basés sur le Web, tels que la combinaison de plusieurs classificateurs et algorithmes génétiques (*un algorithme génétique appartient à la famille des algorithmes évolutionnistes. Leur but est d'obtenir une solution approchée à un problème d'optimisation, lorsqu'il n'existe pas de méthode exacte ou que la solution est inconnue pour le résoudre en un temps raisonnable*).

II.6.3 Les outils de support pour les tuteurs d'e-learning

La fourniture d'un ensemble d'outils automatiques, ou semi-automatique, pour les tuteurs virtuels qui leur ont permis d'obtenir une rétroaction objective du comportement d'apprentissage des étudiants afin de suivre leur processus d'apprentissage, a été une ligne importante de recherche sur l'exploration de données pour l'e-learning,

Sur la base de publications interrogées, les outils expérimentaux mis au point avec cet objectif à l'esprit pourrait être regroupés en:

- Outils pour évaluer les performances d'apprentissage des étudiants.
- Les outils qui permettent d'effectuer une évaluation des matériels d'apprentissage
- Les outils qui fournissent une rétroaction aux tuteurs basée sur le comportement d'apprentissage des étudiants

Diverses méthodes d'exploration de données ont été appliquées pour évaluer la performance d'apprentissage des étudiants, y compris: le clustering, arbres de décision, réseaux de neurones, et les règles d'association.

Une des activités les plus difficiles et les plus longues pour les enseignants dans les cours d'enseignement à distance est le *processus d'évaluation*, en raison du fait que, dans ce type de cours, le processus d'examen devrait plutôt se faire au moyen de ressources de collaboration tels que l'e-mail, forums de discussion, etc. Par conséquent, cette évaluation est généralement effectuée conformément à un grand nombre de paramètres, dont l'influence dans la note finale n'est pas toujours bien définie et / ou compris. Donc, il serait utile pour découvrir des caractéristiques qui sont très pertinents pour l'évaluation des étudiants. De cette façon, il serait possible pour les enseignants à fournir une rétroaction aux étudiants concernant leurs activités d'apprentissage en ligne et en temps réel.

II.7 Data mining en e-learning au-delà des publications universitaires

(Les systèmes et les projets de recherche)

Au-delà des publications universitaires, des méthodes d'exploration de données ont été intégrées dans des plates-formes mises en œuvre dans des véritables systèmes d'apprentissage. Une étude générale de ces types de systèmes: **[29],** couramment utilisé dans les universités et l'enseignement supérieur, a montré deux types principaux de plates-formes:

Le premier type (par exemple, TopClass) n'a pas fait une distinction entre l'enseignant et l'auteur, tandis que le second est clairement défini et a marqué les rôles

Nom du projet	Les techniques DM appliquées	Sujet e-Learning	Université ou une institution	URL du projet
ATutor	analyse statistique	Système d'évaluation et suivi du comportement des étudiants	Université de Toronto, Canada	www.atutor.ca/
aLFanet	agents logiciel, machine d'apprentissage, règles association	Adaptation de cours aux comportements de navigation des étudiants	Université de Netherlands. Espagne Portugal, Germanie et Netherlands	http://alfanet.ia.uned.es/alfanet
AHA!	Règles de prédiction	Adaptation de cours aux comportements de navigation des étudiants	University of Technology and Cordoba University. Netherlands and Spain	http://aha.win.tue.nl
Blackboard	Analyse statistique	Système d'évaluation et suivi du comportement des étudiants	Blackboard	www.blackboard.com/us/index.aspx

Tableau 1 : Quelques projets E-Learning dans lesquels les techniques du data mining sont utilisées [28]

Tous ces systèmes prétendent être innovants et soulignent l'importance du contenu, mais, malheureusement, ils fournissent difficilement des informations sur les méthodes didactiques et des modèles qu'ils implémentent, il est donc difficile de les évaluer.
En ce qui concerne l'*adaptation* est une partie intégrante des systèmes, il exige une personnalisation étendue. La plupart des systèmes étudiés font des tâches d'apprentissage de soutien collaboratives, mais ils ne permettent pas l'utilisation d'un scénario spécifique. Ils permettent une collaboration, en fournissant simplement les outils de base pour son implémentation **[28].**

Plusieurs grands projets de recherche ont porté sur l'intégration des méthodes de Data Mining en e-Learning (voir tableau 1). Le projet Alfanet consiste en une plate-forme e-learning qui fournit aux personnes un apprentissage interactif, adaptatif et personnalisé grâce à l'internet.

- ***Alfanet*** comprend une composante pour fournir un support à l'interprétation et à la présentation des questionnaires adaptatives dynamiques et de leur évaluation au moment de l'exécution, sur la base des préférences des étudiants et leur profil. La composante d'adaptation applique les techniques de règles d'association, et les architectures multi-agents pour fournir des recommandations en ligne en temps réel et des conseils aux apprenants basées sur les interactions précédentes des utilisateurs, la structure du cours, la caractérisation du contenu et les résultats des questionnaires.

- ***Le projet AHA***! été initialement développé pour soutenir un cours en ligne pour ajouter l'adaptation au cours hypermédia à l'Université de Technologie d'Eindhoven. AHA! est actuellement dans sa version 3.0. Un de ses caractéristiques les plus importantes est l'adaptation de la présentation de système de navigation d'un cours sur la base du niveau de la connaissance d'un étudiant particulier. AHA applique des règles de prédiction spécifiques pour atteindre les objectifs d'adaptation.

- ***Un ATutor*** est un Open Source conçu avec des fonctionnalités d'accessibilité et d'adaptabilité. *ATutor* a également adopté des spécifications permettant aux développeurs de créer des contenus réutilisables qui peuvent être échangés entre les différents systèmes d'apprentissage. En *ATutor*, les tuteurs peuvent affecter un crédit partiel pour certaines réponses et peut voir les notes, par étudiant, et pour tous les étudiants sur tous les tests.

- ***Blackboard*** est une autre suite commerciale e-Learning qui permet aux tuteurs de créer des cours d'e-learning et de développer des parcours d'apprentissage personnalisés pour le groupe ou pour chaque étudiant, en fournissant des outils qui facilitent l'interaction, la communication et la collaboration entre tous les acteurs. Le système fournit une analyse des données aux études et le point de test, et les résultats peuvent être exportés pour une autre analyse. Le rapport inclut le nombre de fois et les dates dans les quels chaque étudiant a accédé au contenu des cours, forums de discussion et les affectations.

II.8 Conclusion

Du point de vue des problèmes de data mining négocié dans les travaux inspectés, nous avons vu que ceux-ci sont dominés par la recherche sur la classification et le clustering. La plupart des travaux traitent l'estimation des étudiants d'apprentissage, matériels d'apprentissage et évaluation du cours, et l'adaptation du cours basé sur les comportements des étudiants d'apprentissage.

Notre travail vise à intégrer le data mining dans les systèmes e-learning, de nombreuses applications de data mining ont été développés pour extraire les informations sur les apprenants et les classer dans un environnement e-learning, afin de les aider à prédire leurs résultats.

Les techniques de règles d'association ont été utilisées pour construire des profils des apprenants, les techniques de classification ont été utilisées pour classer et analyser les informations de l'apprenant à travers l'arbre de décision avec un résultat plus précis et la déduction des règles d'association qui peuvent être utilisées pour des améliorations dans un environnement e-learning.

Chapitre 3 : Le data mining pour le suivi des apprenants

III.1 Introduction

L'évaluation est un aspect primordial et inévitable dans une démarche d'apprentissage. Pour pouvoir évaluer, l'enseignant construit et utilise des outils qui lui permettent de récolter les informations indiquant l'état du savoir de l'apprenant.

L'étude des erreurs occupe une place importante dans le processus d'évaluation. Afin de remédier aux erreurs, l'enseignant dispose d'une évaluation d'un procédé de diagnostic mis en place pour la détection des erreurs. D'une manière générale, l'évaluation permet la régulation et la vérification des apprentissages.

III.2 Evaluation, diagnostic et erreurs

III.2.1 Les erreurs dans l'évaluation

Il est difficile de cerner le terme erreur par des mots dans une définition unique, certains auteurs expliquent que la notion d'erreur s'appuie sur l'existence d'une démarche particulière suivie par l'apprenant qui répond à une question ou qui tente de résoudre un problème**.**

- **Définition**

Le Robert Langue Française propose cette définition : « Erreur : Chose fausse, erronée par rapport à une norme (différence par rapport au modèle ou au réel). Faute, inexactitude ».

Une erreur est un indice de la non maîtrise de l'item de connaissance du domaine auquel elle est rattachée elle est détectée par une utilisation erronée donc non conforme à la définition et règles régissant l'élément de connaissance en question.

Une erreur ne doit plus être sanctionnée mais analysée pour pouvoir trouver les causes qui l'ont engendrée.

- ***Erreur ou faute :***

Trop souvent, l'erreur est associée à la faute et c'est à ce titre qu'elle est sanctionnée généralement par une mauvaise note (trop de fautes assurent un zéro) la faute a une connotation négative tant pour le maître puisqu'elle traduit une absence de savoir, que pour les élèves puisqu'elle véhicule une notion de jugement moral. A contrario et c'est ce qui fait l'intérêt du concept d'erreur, c'est qu'il constitue un outil pédagogique pour aider l'élève dans son apprentissage.

- ***Difficulté et erreur* :**

Les relations qui existent entre ces deux concepts sont des relations de causes à effet :

Il est commun de dire que l'enfant en difficulté commet des erreurs. Cependant, l'absence d'erreur ne traduit pas toujours l'absence de difficultés. En effet, l'élève peut éprouver des

difficultés même s'il arrive au terme de son travail (difficultés au niveau de la rapidité du travail, d'une défaillance de la mémoire, etc…). Ainsi, travailler à partir des erreurs, les exploiter, les analyser ne permettrait-il pas aux élèves et au maître de comprendre le domaine enseignée ?

- ***erreur ou et échec :***

L'échec se présente comme la conséquence directe des erreurs. Cependant une distinction s'impose :

- L'erreur correspond à un moment dynamique de la phase d'apprentissage durant lequel il convient de la recenser grâce aux évaluations et de proposer des remédiations.
- L'échec correspond, lui, à un constat final. On peut parler d'échec lorsque, dans une phase d'apprentissage, toutes les remédiations aux erreurs n'ont pas permis aux élèves d'obtenir le résultat escompté

- **Analyse des erreurs**

L'analyse des erreurs précède le processus de diagnostic. Elle s'agit en effet d'identifier l'erreur, en décrire la nature, définir ses caractéristiques et décrire le contexte où elle se produit. Cette analyse ne se substitue pas aux méthodes de diagnostic, mais, par contre, elle nous permet de décrypter le processus de résolution avancé par l'apprenant et de fournir des informations sur sa compréhension et sur ses habiletés.

III.2.2 Diagnostic

Le terme diagnosis vient du grec, mot composé de « dia » (qui signifie la séparation) et de « gnosis » qui signifie apprendre à connaître. Ainsi, l'origine étymologique de ce mot signifie apprendre à séparer les connaissances : séparer le faux du vrai, le bien du mal, ce qui est sain de ce qui est malade, ce qui est normal de ce qui est en panne, autrement dit le sens étymologique du mot diagnostic, c'est savoir discerner. Il s'agit en effet d'arriver à la connaissance des problèmes et difficultés, connaissance suffisante pour éventuellement conduire à une solution **[32].** De plus, il a pour but de connaître les sources réelles de ces problèmes et difficultés.

L'importance du diagnostic : Il est très utile pour l'enseignant de se servir d'erreurs comme moyen pour avoir plus d'informations relativement à ce qui se passe dans la tête des apprenants afin d'analyser les processus de leurs pensées. Entre autres, il est très profitable de découvrir les mécanismes mentaux qui ont conduit les apprenants à telle ou telle erreur **[32].**

Il ne suffit pas de considérer le produit final qu'est l'erreur. Il ne suffit pas non plus de la catégoriser pour déterminer leur nature ou de faire un bilan en cours d'apprentissage, mais, il est important d'utiliser le diagnostic pour connaître les sources réelles des difficultés rencontrées par les apprenants et ainsi pouvoir aider ces apprenants à les résoudre.

III.2.3 Evaluation des apprenants

Pour Charles Hadji **[30],** Evaluer signifie interpréter, vérifier ce qui a été appris, compris, retenu, vérifier les acquis dans le cadre d'une progression, juger un travail en fonction des critères donnés, estimer le niveau de compétence d'un apprenant, situer l'apprenant par rapport à ses compétences, déterminer le niveau d'une production donnée par l'apprenant.

L'évaluation joue un rôle primordial dans l'activité pédagogique, non seulement comme le moyen de vérifier les acquisitions mais aussi comme le moyen de motiver ces apprentissages.

La définition du dictionnaire (***Petit Robert 2006***) indique que l'évaluation est l'action d'évaluer, qui permet de déterminer la valeur ou l'importance d'une chose.

En pédagogie, l'évaluation désigne le processus destiné à apprécier, qualitativement et quantitativement, le niveau atteint par l'apprenant, en fonction des méthodes éducatives utilisées et en rapport avec les objectifs fixés.

L'évaluation dans notre contexte consiste à quantifier et qualifier le degré de maîtrise des éléments de connaissances enseignés sous forme de notes, dans le but d'aider les apprenants à améliorer leurs compétences leurs degrés de maîtrise des éléments enseignés pour chaque exercice résolu durant la session d'apprentissage, cette évaluation permet de suivre la profession par historisation en effet l'apprenant sera facilement suivi par l'enseignant.

L'intérêt des évaluations dans le contexte de la formation ouverte et à distance peut se résumer aux points suivants :

- Pour mesurer la différence entre les objectifs formulés au début du cours et les résultats obtenus à la fin de la formation.
- Pour mesurer l'efficacité du dispositif d'apprentissage.
- Pour explorer et comparer différents moyens pour atteindre un même objectif d'apprentissage.
- Pour prendre des décisions afin de faire évoluer ou améliorer la formation.

- ***L'évaluation au service de l'apprentissage***

Le but premier de l'évaluation est d'améliorer la qualité des apprentissages.

Elle permet de vérifier les acquis des élèves et de mesurer leurs progrès par rapport à l'atteinte souhaitée des résultats d'apprentissage d'un programme.

L'évaluation permet d'observer plus en profondeur le processus d'apprentissage et peut ainsi contribuer à l'adaptation d'approches et de stratégies pédagogiques.

On peut situer les types d'évaluations suivants selon le moment et l'intention :

- *L'évaluation diagnostique* : a comme fonction essentielle d'identifier les difficultés d'apprentissage des élèves. Souvent effectuée au début d'une période de formation
- *L'évaluation formative* : a comme fonction exclusive la régulation des apprentissages pendant un cours ou une séquence d'apprentissage, elle permet à l'élève de savoir où il en est dans son apprentissage
- *L'évaluation sommative* : Elle sert à sanctionner ou certifier le degré de maîtrise des résultats d'apprentissage des programmes d'études, elle se situe en général en fin d'une séquence d'apprentissage.
- *L'évaluation pronostique :* Qui intervient avant la phase d'apprentissage, a pour rôle d'évaluer les prés requis de l'apprenant.
- *L'auto-évaluation* : l'évaluateur est l'apprenant lui-même, comme il se peut que deux autres apprenants évaluent le travail d'un autre apprenant sous la supervision de l'enseignant.

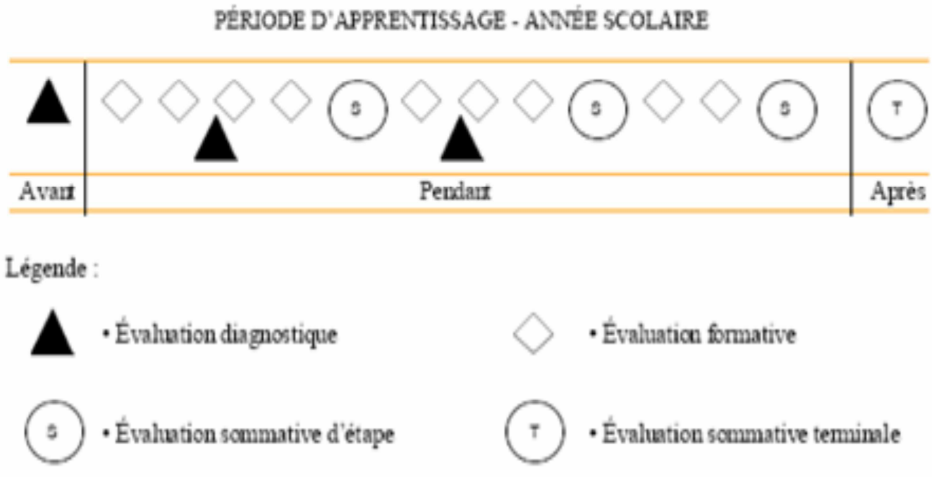


Figure 3 : Interventions des types d'évaluation pendant une année scolaire [31]

Remarque :

Tous ces types d'évaluation peuvent être appliquées dans toutes les situations d'apprentissage, y compris dans les environnements qui fonctionnent en présentiel ou à distance **[32].**

Dans *l'évaluation formative*, la justification des résultats permet d'expliquer les fautes de l'apprenant qui est un élément important. Dans *l'évaluation sommative*, l'élément essentiel est la note obtenu par l'apprenant. Dans *l'évaluation pronostique* ne sera ni la note ni les erreurs faites par l'apprenant, c'est le jugement qu'émettra l'enseignant sur l'accession de l'apprenant à une nouvelle formation, enfin dans *l'évaluation diagnostique*, l'accent est mis sur les compétences de l'apprenant, donc l'élément central dans cette évaluation est le niveau de compétence que reflète le résultat de l'évaluation sachant que le mot compétence est considéré au sens large, qui englobe des savoir-faire (capacités), des savoirs-être (comportements), des savoirs-agir (méthodes de production) ainsi que des connaissances.

III.3 Evaluation des apprenants en utilisant les techniques du data mining:

Le data mining ou découverte de la connaissance dans les bases de données (KDD) est l'extraction automatique de modèles implicites et intéressants de grandes collections de données.

L'exploitation de règle d'association a été appliquée aux systèmes e-learning pour l'analyse de l'association qui consiste à trouver des corrélations entre les items dans un dataset, y compris, les tâches suivantes:

- Guider automatiquement les activités des apprenants, recommander des matières d'apprentissage,
- Identifier des attributs qui caractérisent des modèles de disparité de la performance entre plusieurs groupes d'apprenant,
- Découvrir les corrélations intéressantes à travers l'usage de l'information pour fournir la réaction aux auteurs du cours,
- Trouver les relations entre chaque modèle du comportement de l'apprenant,
- Trouver les erreurs qui se produisent souvent ensemble,
- Guider la recherche du modèle de transfert le mieux adapté.

- Extraction de modèles utiles pour aider des éducateurs à évaluer et interpréter les activités du cours en ligne, et personnaliser l'e-learning basé sur les profiles d'usage global et une ontologie du domaine.
- Evaluer les activités des apprenants et donner une indication de la manière d'organiser au mieux l'espace web et être capable de faire des suggestions aux apprenants qui partagent des caractéristiques semblables; générer des activités personnalisées aux groupes différents d'apprenants;

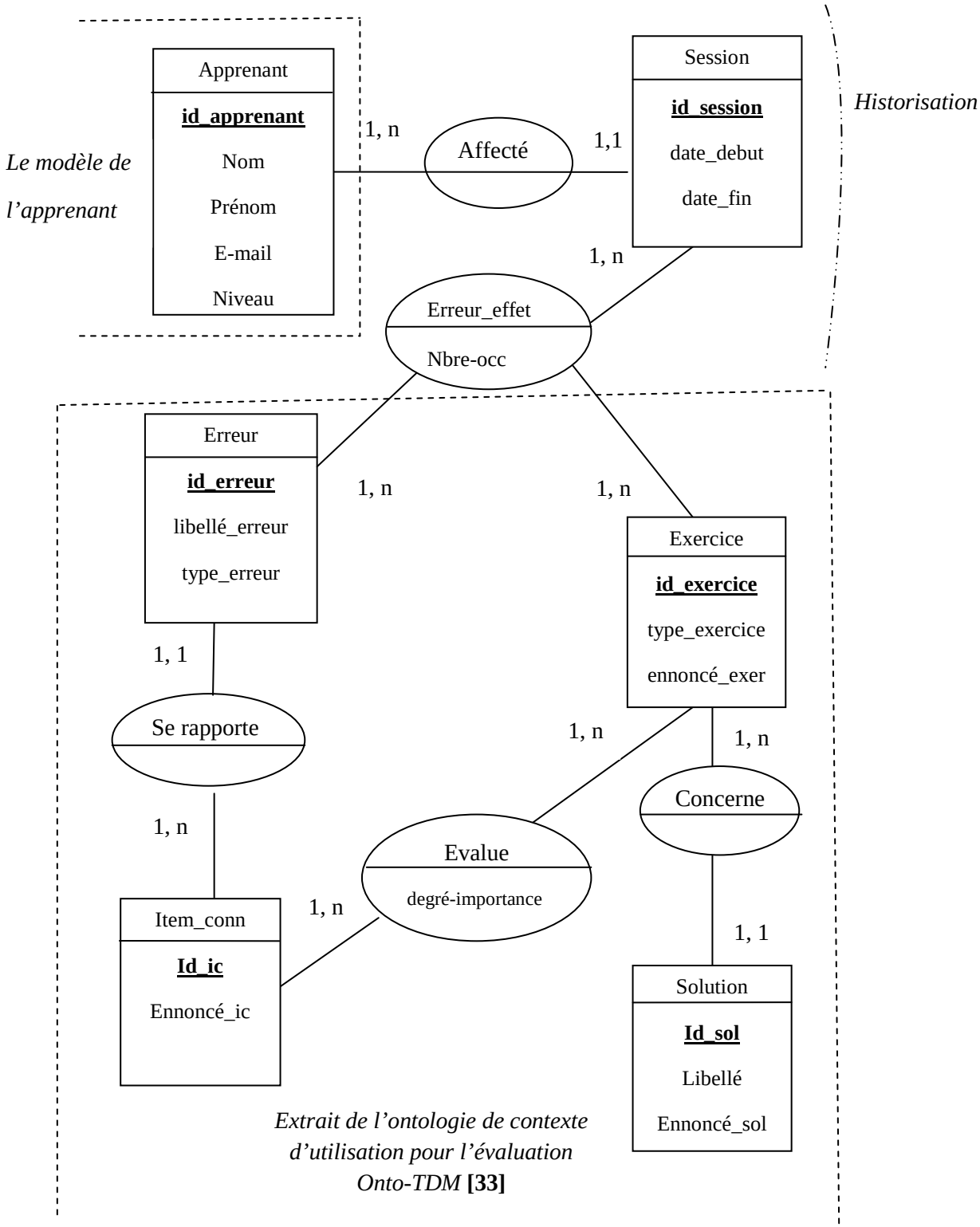
III.4 Le cas de l'analyse des résultats d'évaluation de type ODALA

L'approche d'évaluation ODALA (Ontology Driven Auto-evaluation for Learning Approch), propose une méthodologie et des techniques pour le développement d'un système d'évaluation basé sur l'ontologie de domaine d'enseignement Onto-TDM proposée dans **[33].**l'approche ODALA est fondé sur la notion d'erreurs commises et/ou successibles d'être commises classées en différents types (sémantique, lexicaux-syntaxique…) en sessions d'apprentissages. Dans le cadre de ce mémoire l'analyse (classification, règle d'association) est faite en utilisant les techniques de data mining.

III.4.1 Modélisation des résultats de l'évaluation

En plus du modèle du domaine représenté dans Onto-TDM, nous avons besoin d'apporter d'autres structures pour capter les données d'une session d'apprentissage .Le modèle entités association de la figure 4 donne une idée sur cette structure.

***Le modèle entité/association** :*



Extrait de l'ontologie de contexte d'utilisation pour l'évaluation Onto-TDM **[33]**

Figure 4 : Diagramme entités/associations

Selon la réalité étudiée on peut établir les règles suivantes :

- Un apprenant peut être inscrit dans une ou plusieurs sessions,
- Chaque exercice peut avoir plus d'une solution,
- Un exercice peut traiter plusieurs items de connaissances,
- Une erreur se rapporte à un ou plusieurs items de connaissances,
- Un apprenant peut effectuer plus d'une erreur dans un exercice d'une ou plusieurs sessions.

- ***Le modèle relationnelle :***

table	**commentaire**
Apprenant (**Id_apprenant,** Nom**,** Prénom**,** E-mail, niveau)	Appartient au modèle de l'apprenant
Session (**id_session,** date_debut**,** date_fin, note, *id_apprenant****)**	C'est l'intersection de l'apprenant avec le domaine
Exercice (**id_exercice,** type_exercice**,** ennoncé_exer)	Table contenant les exercices correspondants à une session d'apprentissage
Erreur (**id_erreur,** libellé_erreur, type_erreur,*id_ic*)*	Contient toutes les erreurs connus sémantiques ou lexicaux-syntaxiques
Solution (**Id_sol,** Libellé, Ennoncé_so, *id_exercice**)	toutes les solutions possibles pour chaque exercice
Item_conn (**Id_ic,** Ennoncé_ic)	Contient tous les items de connaissances qu'on peut attribuer à des notions d'apprentissage
Erreur_effect (**id_session, id_erreur, id_exercice,** Nbre-occ)	Table contenant toutes les erreurs effectuées par un apprenant lors d'une session donnée en résolvant un exercice donné.
Evalue (**id_exercice, id_ic,** degré-importance)	Tale qui contient les items de connaissances évalués dans un exercice donné.

Tableau 2 : Table relationnelle pour la session d'apprentissage.

III.4.2 Classification des apprenants

L'arbre de décision

Un arbre de décision permet à partir des données connues sur le problème de donner des prédictions par réduction, niveau par niveau, du domaine des solutions.

Un arbre de décision se fait à partir de données déjà classifiées.

- ***L'algorithme C4.5***

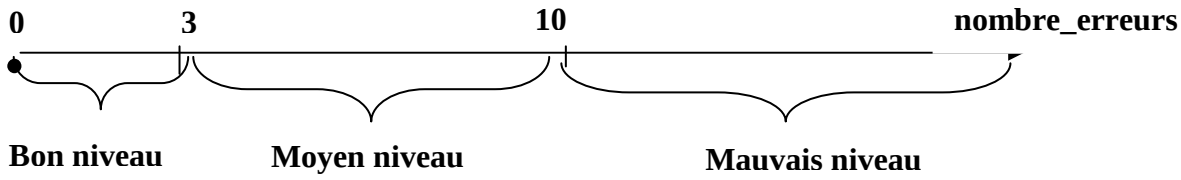
Le C4.5 élague l'arbre construit afin de supprimer les règles inutiles et de rendre l'arbre plus compact.

L'application de cet algorithme permet de regrouper les individus qui se ressemblent dans une même classe, dans le cas de session d'apprentissage, il s'agit par exemple de regrouper les apprenants ayant effectués les même erreurs, regrouper les apprenants habitant dans la même ville, ou encore les apprenants ayant un bon niveau ou non …

L'algorithme C4.5 nous permet de classifier les apprenants d'une session d'exercice, par exemple s'ils sont des apprenants d'un bon, moyen ou mauvais niveau. Cette classification est faite à partir d'autres attributs qu'on appelle des attributs prédicateurs (un attribut prédicateur est l'attribue sur le quel on se base pour réaliser une classification). Dans notre exemple l'attribut prédicateur est le nombre d'erreur effectué dans une session en résolvant un exercice donné.

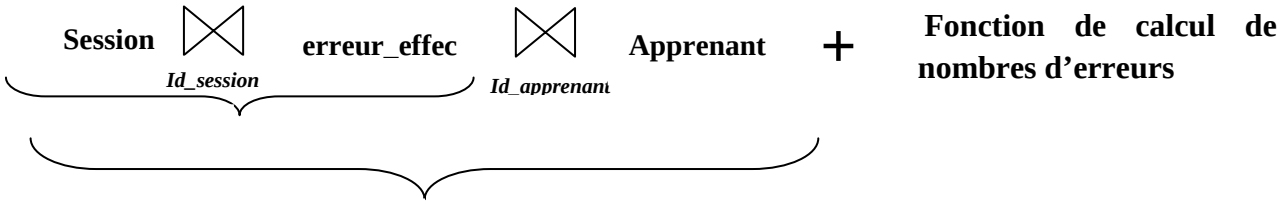
- ***Hypothèse :***

On suppose par exemple qu'un apprenant de bon niveau est un apprenant qui a au moins un nombre d'erreur effectué <= 3, concernant le niveau moyen (3<nombre d'erreur<10) et le mauvais (nombre d'erreur >10).



- ***Les données en entrées :***

Trois tables de la base de données nous intéressent pour avoir les données en entrées de l'algorithme C4.5 relativement à l'objectif (classification des apprenants par niveau selon le nombre d'erreurs effectués) qui sont : **apprenant**, **erreur_effec** et **session**. On va déduire une nouvelle table: **input_base**. En faisant une jointure entre les trois tables**,** comme suit :



- nombre_err : est un attribut calculé par la fonction ***SUM,*** *qui calcule la somme des occurrences de toutes les erreurs pour chaque apprenant.* La requête SQL qui permet d'avoir ce résultat est la suivante :

SELECT apprenant.id_app, apprenant.niveau,SUM(nbre-occ) as " nombre_err "

FROM apprenant,erreur_effec,session

WHERE session.id_sess=erreur_effec.id_sess and apprenant.id_app=session.id_app

GROUP BY apprenant.id_app // ou GROUP BY session.id_sess (puisque une session est affectée à un seul apprenant)

Le tableau suivant montre la structure de la base input:

id_apprenant	nombre_err	niveau
1	2	bon
2	2	bon
3	4	moyen
4	4	moyen
8	0	bon
9	11	mauvais
5	11	mauvais
6	3	bon
7	12	mauvais

Tableau 3 : input_base

Apres avoir récupérer cette table On peut directement classifier les apprenants selon leur niveau, c.-à-d. en fonction du nombre d'erreurs effectuées pendant la session d'apprentissage.

- ***Étapes de l'algorithme* :**

Etapes C4.5	**Application en cas d'évaluation**
- Ensembles d'apprentissage et un ensemble de test, - Décider si un nœud est terminal, si tous les éléments associés à ce nœud sont dans une même classe ou si on ne peut sélectionner aucun test, - Si un nœud n'est pas terminal, lui associer un test - Affecter une classe à une feuille	- La table des données input_base - si tous les éléments associés à l'attribut *bon* sont dans la même classe alors le nœud correspondant est un nœud terminal - le cas du nombre_err >3, ce nœud n'est terminal, on peut lui associer deux tests : cas de nombre_err ≤ 4, ou non. - Affecter les trois feuilles de l'arbre résultat aux trois classes d'apprenants.

- ***Résultat pour le cas du tableau 3***

Le résultat de cette étape sera donc trois classes différentes d'apprenants. L'algorithme retourne pour chaque classe le nombre d'apprenants associés. Dans ce cas nous avons

Classe à ***bon niveau*** : 4

Classe b ***moyen niveau*** : 2

Classe c ***mauvais niveau*** : 3

- ***Interprétation des résultats***

D'après Les résultats générés par l'algorithme C4.5 on sait bien qu'il s'agit de classer les apprenants comme suit

Classe 1 : *bon niveau*
- Id_apprenant=1, nombre_err=2
- Id_apprenant=2, nombre_err=2
- Id_apprenant=8, nombre_err=0
- Id_apprenant=6, nombre_err=3

Classe 2 : *moyen niveau*
- Id_apprenant=3, nombre_err=4
- Id_apprenant=4, nombre_err=4

Classe 3 : *mauvais niveau*
- Id_apprenant=9, nombre_err=11
- Id_apprenant=5, nombre_err=11
- Id_apprenant=7, nombre_err=12

III.4.3 Déduction de règles d'associations entre les erreurs :

Les règles associatives sont des règles qui décrivent des associations entre certains éléments. Cette technique permet de faire ressortir les associations entre les erreurs de base (les erreurs essentielles, ceux qui sont répétées souvent par plusieurs apprenants).

***Objectif* :** dans ce cas on cherche a généré toutes les corrélations possible entre les erreurs.

- ***L'algorithme apriori*** *(rappel)*

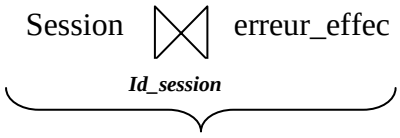
Il repose sur le fait qu'un ensemble fréquent d'éléments a pour sous-ensembles des ensembles fréquents d'éléments. En effet si l'ensemble {A, B} est fréquent dans la base, les ensembles {A} et {B} sont eux-mêmes fréquents dans la base.

- ***Hypothèse***

Comme cet algorithme ne nécessite pas des classes à priori, c'est à lui de générer des règles à travers les données donc on a qu'à lui donner les attributs dont on cherche de connaitre les corrélations ; (il ne nécessite pas de classe).

- ***Les entrées***

Les données à intégrer à ce niveau ***input_base2*** sont obtenues par une jointure mais cette fois de deux tables seulement table *session* et table *erreur_effec* sur l'attribut id_apprenant



La requête associée :

SELECT id_apprenant, id_exercice, id_erreur FROM session, erreur_effec WHERE session.id_session=erreur_effec.id_session

Le tableau suivant montre la structure de la table input_base 2 :

id_apprenant	id_exercice	id_erreur
1	1	1
1	1	2
1	2	4
1	3	NULL
3	2	3
3	2	3
3	2	4
3	3	NULL
2	1	1
2	1	2
2	3	NULL
4	2	4
4	3	NULL

Tableau 4 : input_base2

Après avoir récupérer cette table, on peut exécuter directement l'algorithme apriori sur les trois attributs id_apprenant, id_exercice, et id_erreur afin d'extraire toutes les règles possibles (règle de type si alors).

Pour avoir un ensemble bien précis de règles, on peut spécifier les attributs sur lesquelles on applique l'algorithme, par exemple ici on peut l'appliquer seulement sur les deux attributs : id_apprenant et id_erreur pour avoir les règles d'associations entre les erreurs par apprenants (quelles sont les erreurs effectuées souvent par les apprenants,), comme on peut sélectionner les attributs : id_exercice et id_erreur pour avoir la relation de type : dans quelles exercices les erreurs se répètent ; ou les quelles traités correctement (sans erreurs)

- ***Résultats pour le cas du tableau 4***

1. *Id_apprenat=1, id_exercie=1 ==>id_erreur=1*
2. *Id_apprenat=1, id_exercie=1 ==>id_erreur=2*
3. *Id_exercice=1==>id_erreur=1, id_erreur=2*
4. *Id_apprenat=3, id_exercie=2 ==>id_erreur=3*
5. *id_exercie=3==>id_erreur=NULL*
6. *id_exercie=2 ==>id_erreur=4*

.
.

Les résultats cités au dessus sont un extrait sur ce qu'on peut avoir en appliquant l'algorithme apriori, comme on peut avoir plus de résultats avec toutes les combinaisons possibles d'attributs.

- ***Interprétation :***

D'après nos connaissances on peut interpréter ces résultats comme suit :

- Si id_exercice=3 alors pas d'erreurs (règle 5)

C'est-à-dire que l'exercice 3 est effectué avec succès par tous les apprenants de l'exemple précédent.

- Si id exercice=1 alors id_erreur=1 et id_erreur=2 (régle3), c.-à-d. que ces deux erreurs reviennent toujours dans cette exercice (seulement dans la table donnée en entrées).
- Si id_apprenant=3 et id_exercice=2 alors id_erreur=3 (régle4); on peut conclure ici qu'on répétant le même exercice, l'apprenant 3 refait la même erreur.
- Etc.…

Ce type de règle peut aider les enseignants dans le processus d'évaluation des apprenants et surtout de suivre leur niveau de compréhension à travers la maitrise de résolution d'exercices cela permet de prendre des décisions concernant la session d'apprentissage.

III.5 Le problème de format de données (Les inputs) :

Pour réaliser les algorithmes précédents, il faut avoir un format de données spécifique exigés par les outils de data mining, le format accepté par l'outil utilisé dans notre étude est le format arff. Pour l'obtenir, nous proposons un processus de transformation (Figure 5).

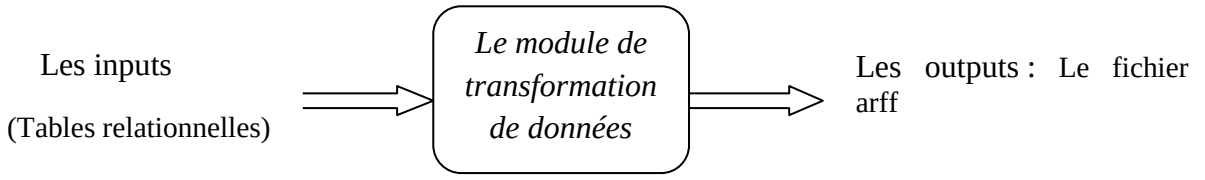


Figure 5 : Processus de transformation de données d'apprentissage.

Le module de transfert de données contient deux étapes :

- Transformer la table relationnelle en fichier CSV (*Comma-separated values),* en faisant un export de la table concernée et on choisissant le format CSV.
- Puis transformer le fichier CSV en format ARFF approprié en utilisant l'outil en ligne.

L'implémentation de ce module restera une perspective de notre travail.

III.5.1 Le format CSV

Est un format informatique ouvert pour l'échange des données tabulaires pour (sous forme de tableaux). En français, « valeurs séparées par des virgules » est un format dans lequel les données sont enregistrées **ligne par ligne**, et où les valeurs sont **séparées par des virgules**.

Un fichier CSV est un fichier texte, par opposition aux formats dits « binaires ». Chaque ligne correspond à une rangée du tableau et les cellules d'une même rangée sont séparées par une virgule. Le fait que les fichiers CSV soient essentiellement utilisés autour du logiciel Microsoft Excel, et que les séparateurs ne soient pas standardisés rend ce format peu pratique pour une utilisation autre que des échanges de données ponctuels. Ce format est toutefois assez populaire parce qu'il semble relativement facile à générer.

Les champs texte peuvent également être délimités par des guillemets. Lorsqu'un champ contient lui-même des guillemets, ils sont doublés afin de ne pas être considérés comme début ou fin du champ.

Si un champ contient un signe utilisé comme séparateur (virgule, point-virgule, etc.), les guillemets sont obligatoires afin que ce signe ne soit pas confondu avec un séparateur.

Exemple, avec des champs séparés par des points-virgules et des guillemets :

```
Robert;Dupont;rue du Verger, 12;…
"Michel";"Durand";" av. de la Ferme, 89 ";…
"Michel ""Michele""";"Durand";" av. de la Ferme, 89";…
"Michel;Michele";"Durand";"av. de la Ferme, 89";…
```

Ce texte représente les données suivantes :

Robert	Dupont	rue du Verger, 12	…
Michel	Durand	av. de la Ferme, 89	…
Michel "Michele"	Durand	av. de la Ferme, 89	…
Michel;Michele	Durand	av. de la Ferme, 89	…

III.5.2 Le format ARFF

Un fichier ARFF (Attribute-Relation File Format) est un fichier de texte ASCII qui décrit une liste d'exemples ou d'instances partageant un ensemble d'attributs, contenant deux sections : La première section l'En-tête qui est suivie par la section des données.

L'en-tête d'un fichier arff contient le nom de la relation, la liste d'attributs et leurs types

Les lignes qui commencent par le symboles % sont des commentaires.

- **La section de l'en-tête**

L'en-tête d'un fichier arff contient la déclaration de la relation ainsi que les déclarations des attributs

- **La déclaration de @relation**

Le nom de la relation est défini à la première ligne du fichier arff. Le format est le suivant :

```
@relation <nom-relation>
```

Où <nom-relation>est une chaine caractère. Cette chaine doit être mise entre quottes si le nom inclut des espaces.

- **Les déclarations de @attribute**

Les déclarations des attributs prend une forme de séquence ordonnées d'énoncés de @attribute. Chaque attribut de l'ensemble des données à son propre énoncé **@attribute** qui définie uniquement son nom et son type de données.

L'ordre dans lequel les attributs sont déclarés indique la position de la colonne dan la section de données du fichier. Par exemple, si un attribut est le troisième déclaré, l'outil weka (voir quatrième chapitre) s'attend à trouver les valeurs des attributs dans la colonne délimité par la troisième virgule.

Le format de l'énoncé @attribute est le suivant :

@attribute <nom attribut> <type donnée>

Où le < nom attribut> doit commencer avec un caractère alphabétique. Si les espaces seront inclus dans le nom alors le nom de l'attribut entier doit être mis entre quottes.

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

- numérique
- un entier est traité comme numérique
- le réal est traité comme numérique
- <spécification nominale>
- Chaine de caractère
- date [<date format>]

Où <spécification nominale> et <date format> sont défini après.

- **Les attributs Numériques**

Peuvent être des nombres réels or entiers.

- **Les attributs Nominales**

Les valeurs nominales sont définies en fournissant une <spécification nominale> listing les valeurs possibles {<nom-nominal1>, <nom-nominal2>, <nom-nominal3>, ...}

par exemple, la valeur de clase pour le l'ensemble iris(les données existant sur weka) peut être définie comme suit :

```
@ATTRIBUTE class     {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Les valeurs contenants des espaces doivent être mises entre quottes.

- **Les attributs String**

Les attributs String nous permettent de créer des attributs contenant des valeurs textuels arbitraires. Les attributs string sont déclarés comme suit :

```
@ATTRIBUTE LCC    string
```

- **Les attributs de date :**

Les déclarations de L'attribut date prend la forme :

```
@attribute <nom> date [<format-date>]
```

Où <nom> est le nom pour l'attribut et <format-date> est une chaine de caractère facultative (qui spécifie comment les valeurs de la date devraient être analysées).

- **La section de données ARFF**

La section de données du fichier arff contient la ligne de déclarations de données et les lignes d'instances actuelles.

- **La déclaration de @data**

La déclaration de @data est une seule ligne indiquant le début du segment des données du fichier. Le format est le suivant :

```
@data
```

- **L'instance de données**

Chaque instance est représentée sur une seule ligne.

Les valeurs de l'attribut pour chaque instance sont délimitées par les virgules. Une virgule peut être suivie par zéro ou plusieurs espaces. Les valeurs de l'attribut doivent apparaître dans l'ordre dans lequel ils ont été déclarés dans la section de l'en-tête (c.-à-d., les données qui correspondent à l'énième déclaration de *@attribute* est toujours le champ énième de l'attribut).

Une valeur manquante est représentée par un point d'interrogation seul, comme dans:

```
@data
   4.4,?,1.5,?,Iris-setosa
```

Remarque :

Fautes de temps on n'a pas pu avoir un autre moyen pour le transfert des données d'apprentissage en fichier arff à part l'outil en ligne et les modifications manuelles mais, ce module de transfert de données peut être un programme java, prenant deux arguments : le premier argument est le fichier CSV pris par le programme en entrées afin de générer le deuxième argument le fichier arff associé en sortie.

III.6 Conclusion :

Les techniques de data mining continuent à guider et aider des étudiants à acquérir des capacités cognitives de haut niveau de compréhension, d'application, et à l'évaluation, le résultat de cette évaluation est souvent non compris et n'est pas définie , mais avec le data mining l'évaluation est faite en se basant sur certains critères, nous avons choisi comme critère de base dans cette étude le nombre d'erreurs effectués, d'après nos connaissances il n'existait pas d'études qui s'intéressait à ce point ; cette évaluation permet donc de mieux guider le processus d'apprentissage à travers une session d'exercices, en déduisant les corrélations entre les erreurs afin de suivre la progression du processus d'apprentissage et d'améliorer les performances des apprenants, en plus d'évaluation on s'est intéresser aussi au classement des apprenants en utilisant l'arbre de décision comme algorithme de classification pour pouvoir prédire les résultats des apprenants par la suite, on peut aussi classer les apprenants non seulement par niveau mais aussi par leurs capacités, par comportement…

Pour avoir des résultats concrets sur tous les tests précédents on aura besoin d'utiliser un outil de data mining pour appliquer les algorithmes cités au-dessus qui est l'outil weka décrit dans le chapitre suivant.

Chapitre 4 :

Implémentation

IV.1 Introduction

L'évaluation de l'apprenant est l'étape la plus importante dans le processus d'apprentissage à travers une session d'exercices dont les enseignants ont besoin d'aide à la prise de décision.
Afin d'extraire les informations pertinentes de la session d'apprentissage au sein des bases de données e-learning, plusieurs techniques de data mining ont été proposées et implémentées sur des plates formes différentes, une de ces plateforme est weka[35] dont nous allons implémenter deux algorithme l'un concernant les arbres de décisions le C4.5 pour la classification des apprenants et l'autre concerne les règles d'associations entre les erreurs commises par un apprenant dans un exercice (apriori) , que nous avons appliqués de deux manière :

- Sous l'outil de datamining weka.
- Dans l'environnement java.

Que nous avons détaillé dans les deux parties suivantes :

IV.2 L'utilisation de l'outil Weka

Dans un premier temps nous vous proposons de découvrir comment utiliser Weka, puis nous aurons à exploiter ces nouvelles connaissances pour analyser nos propres données.

IV.2.1 Présentation de l'outil weka

Ce point est structuré principalement en quatre grandes parties qui représentent les quatres traitements fournis par l'environnement Weka à savoir l'Explorer, l'Experimenter, Knowledge Flow et le simple CLI.

IV.2.1.1 Description globale

Le logiciel Weka est l'interface d'une librairie Java pour la fouille de donnée. Une bonne part des algorithmes de fouille de donnée sont disponibles au travers de cette interface. La version actuelle est la 3.5.8 Cette librairie est développée par le groupe de recherche "machine learning" du département d'informatique de l'université de Waikato en Nouvelle-Zélande. Le Weka est un oiseau Zelandais en voie de disparition [34].
L'espace de travail Weka contient une collection d'outils de visualisation et d'algorithmes pour l'analyse des données et la modélisation prédictive. Les algorithmes peuvent être appliqués directement à un ensemble de données ou appelés via un programme Java.
Les principaux points forts de Weka sont qu'il:

- est librement disponible (en particulier gratuitement) sous la licence publique générale GNU,

- est très portable car il est entièrement implémenté en Java et donc fonctionne sur quasiment toutes les plateformes modernes, et en particulier sur quasiment tous les systèmes d'exploitations actuels.
- contient une collection complète de préprocesseurs de données et de techniques de modélisation,
- est facile à utiliser en raison de l'interface graphique qu'il contient.

Weka contient les outils pour le prétraitement de données, la classification, la régression, la segmentation (*clustering*), les règles d'association et la visualisation. En effet, Weka permet d'effectuer un prétraitement sur un ensemble de données, d'appliquer un algorithme d'apprentissage, et d'analyser les résultats et les performances d'un classificateur. Il est aussi bien adapté pour intégrer de nouveaux algorithmes d'apprentissage.

Toutes les techniques de Weka reposent sur la supposition que les données sont disponibles dans un unique fichier plat ou une *Relation binaire*, ou chaque type de donnée est décrit par un nombre fixe d'attributs (les attributs ordinaires, numériques ou symboliques).

Aperçu du logiciel

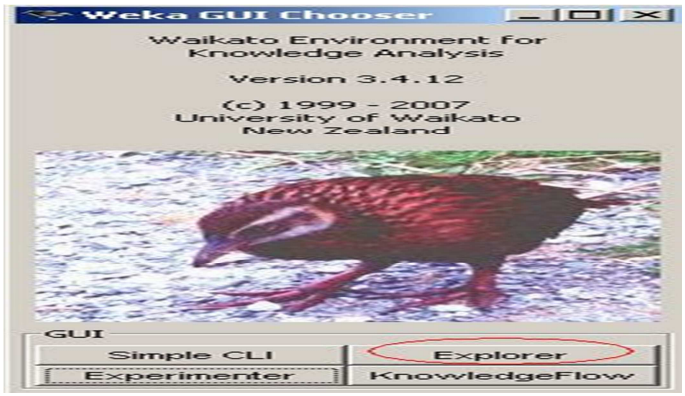


Figure 6 : la fenêtre Weka GUI chooser

L'interface graphique de Weka propose quatre traitements séparés des données :

• « *Simple CLI* » : interface en ligne de commande simple qui permet l'exécution directe des commandes de weka pour les systèmes d'exploitation qui ne fournissent pas leur propre interface ligne de commande

• « *Explorer* » : l'interface principale de weka, ce module regroupe tous les packages importants de Weka à savoir le prétraitement, les algorithmes d'apprentissage, le groupement (*clustering*), les associations, la sélection des attributs et la visualisation. Cette interface possède plusieurs onglets qui donnent accès aux principaux composants de l'espace de travail

• « *Experimenter* » : permet d'exécuter plusieurs algorithmes d'apprentissage en mode lot (batch) et de comparer leurs résultats.

« *Knowledge Flow* » : fournit les mêmes fonctionnalités que le composant "*Explorer*". Ces fonctionnalités sont représentées sous forme graphique et sont utilisées pour construire un schéma de flux de connaissances via une interface drag-and-drop.

IV.2.1.2 Format de l'ensemble de données d'apprentissage

Avant d'utiliser cet environnement dans différentes expérimentations, les données doivent être conformes aux formats de données disponibles au niveau de l'environnement Weka pour l'utilisation de tout outil de cet environnement.

Le format des fichiers de données dans Weka est le format arff (Attribute-Relation File Format). Vous pouvez visualiser ce format en ouvrant avec un éditeur de textes un des fichiers de données proposé par Weka. Ces fichiers se situent dans le répertoire data du répertoire d'installation de Weka. Le format csv (*comma separed values*) est également reconnu par Weka.

Généralement, les données servant à l'expérimentation dans n'importe quel domaine sont stockées dans des fichiers Excel et le format des données utilisées par les tâches de fouille de données est "arff". Avant d'appliquer donc les algorithmes d'apprentissage, ces données doivent être transformées en un fichier de données avec l'extension "arff" pour être lu soit par *Explorer*, *Experimenter* ou *KnowledgeFlow*. Les transformations consistent à :

- Sauvegarder le fichier de données d'apprentissage et de test au niveau de l'outil Excel sous la forme .csv (comma separated value);
- Lire le fichier avec l'extension "csv" dans un éditeur;
- Ajouter en entête du fichier les informations nécessaires au module "Explorer";

➢ Définition des données d'apprentissage par une relation sous forme de :

@relation nom des données d'apprentissage (Dataset).e.g @relation input_base, définition des attributs du dataset avec leur type, e.g, @attribute id_apprenant NUMERIC.

➢ Fermeture de ces ajouts par le mot réservé @data.

- Transformer toutes les virgules par le point, et tous les points virgule par la virgule. Ces deux petites tâches doivent être faites dans l'ordre.
- Sauvegarder le fichier transformé des données sous l'extension "arff"(@ relation file format.

Note : Pour convertir un fichier arff en ce format il faut passer par des outils ou un programme java qui fait cette conversion.

Généralement les données sont stockées dans une base de données relationnelle. Plusieurs outils gratuits existent qui permettent d'exporter le résultat d'une requête SQL dans un fichier csv. Par exemple le **SQL-Viewer**.

- **Format ARFF, un exemple**

```
% 1 . Title: base
@RELATION imput_base
@ATTRIBUTE id_apprennant NUMERIC
@ATTRIBUTE id_exercice NUMERIC
        @ATTRIBUTE nombre_err  NUMERIC
@ATTRIBUTE class {bon, moyen, mauvais}
```

Entête

```
@DATA
@data
1, 2, bon
7, 3, bon
2,4, moyen
6,4, moyen
3,0, bon
4, 11, mauvais
5, 11, mauvais
```

Données

- **Format ARFF, l'entête**

Le nom du tableau de données

@RELATION imput_base

@ATTRIBUTE id_app NUMERIC

@ATTRIBUTE id_exe NUMERIC

Variables continues

@ATTRIBUTE nombre_err NUMERIC

@ATTRIBUTE class {bon, moyen, mauvais} ← **Variable discrète**

- **Format ARFF, Les données**

@DATA

1, 2,	bon
1, 2,	bon
2, 4,	moyen
2, 4,	moyen
3, 0,	bon
4, 11,	mauvais
5, 11,	mauvais

Numérique

Nominal

Le séparateur est la virgule. Si le nom d'une modalité contient un blanc mettre entre guillemets. Une fois les transformations effectuées, le processus de fouille de données est déclenché.

IV.2.1.3 Le traitement Explorer proposés par Weka

- Le Weka Explorer permet de lancer une méthode à partir d'un fichier ARFF.
- Les résultats sont mis sous la forme d'un fichier texte normalisé.
- Permet de sélectionner la méthode la mieux adaptée ou la plus efficace.

Cet outil est activé en cliquant sur l'option Explorer de la figure 5.

Au sommet de la fenêtre, juste au-dessous de la barre du titre, on trouve une ligne d'onglets. Quand l'Explorateur est commencé en premier seulement le premier onglet est actif; les autres sont grisés. C'est parce que c'est nécessaire d'ouvrir (et potentiellement pré-processus) un ensemble de données avant de commencer à explorer les données.

Les onglets sont comme suit:

- *Preprocess* : pour choisir un fichier, inspecter et préparer les données.
- *Classify* : pour choisir, appliquer et tester différents algorithmes de classification.
- *Cluster* : pour choisir, appliquer et tester les algorithmes de segmentation
- *Associate* : pour appliquer l'algorithme de génération de règles d'association..
- *Select Attributes* : permet de faire de la sélection d'attributs à partir de différentes méthodes.
- *Visualize* : pour afficher (en deux dimensions) certains attributs en fonctions d'autres.

La sélection de cette option fait apparaître la fenêtre représentée par la figure 6.

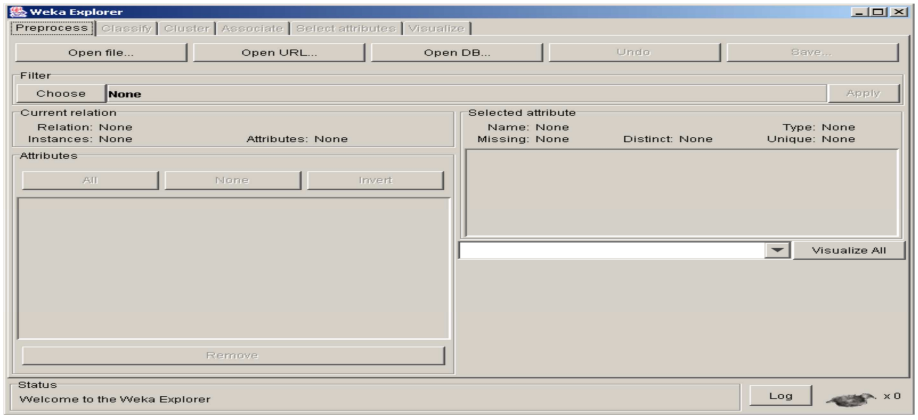


Figure 7 : Fenêtre Explorer

La zone inférieure de la fenêtre (Y compris le status box, le bouton Log, et l'oiseau Weka) reste visible quelle que soit la section.

Boîte statut (Status Box): La boîte d'état apparaît au bas de la fenêtre, il affiche des messages que vous tenir informés de ce qui se passe.

IV.2.2 Application des algorithmes de datamining via l'outil weka

IV.2.2.1 Classification des apprenants

- **La préparation des données**

Après avoir appliqué les transformations nécessaires, jointure entre les trois tables apprenant, session, erreur_effec, nous avons obtenus la table ***input_base*** (voir rapport 3) qu'il faut transformer au format *arff* acceptable par Weka

Préparation du fichier de données sous format arff

Plusieurs outils existent qui permettent d'exporter le résultat d'un fichier csv (SQL Viewer). Ce fichier obtenu nous pouvons le transformer en fichier arff de différentes manières :

- Ouvrir le fichier csv avec l'éditeur de weka
- Le sauvegardé sous format *arff*

Effectué certains changement comme mentionnées précédemment nous obtenons le fichier arff acceptable par weka ou on en utilisant un :

- Outil de conversion disponible sur internet au lien suivant : http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php; Voici le fichier arff généré à partir de la table *input_base*

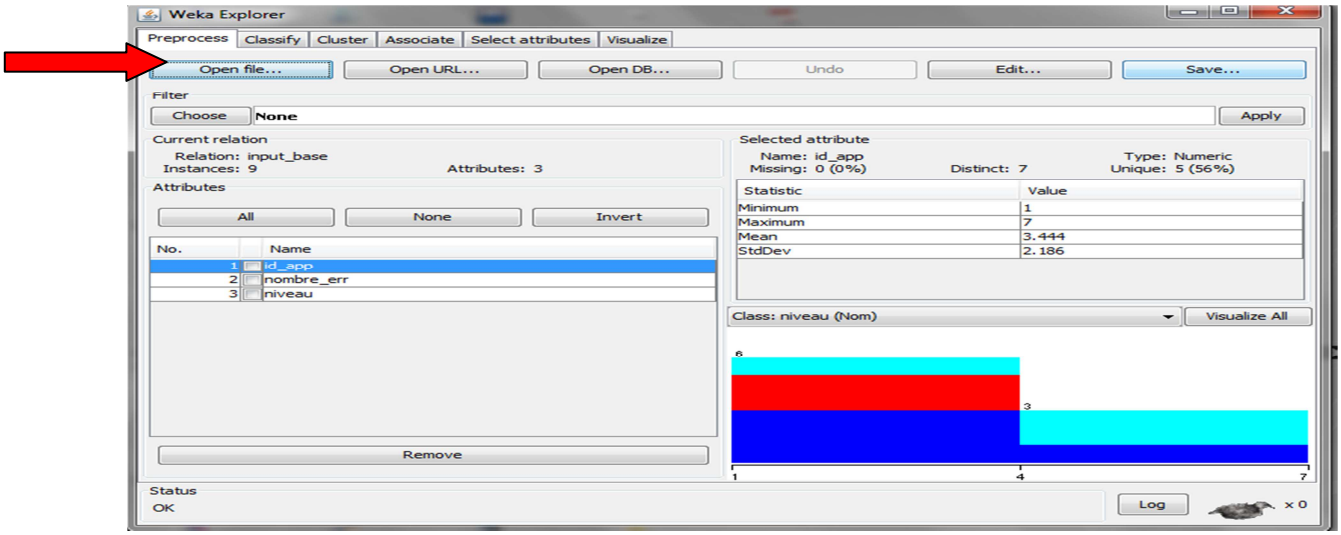
```
@relation input_base
@attribute id_app numeric
@attribute nombre_err numeric
@attribute 'niveau' {'bon','moyen','mauvais'}

@data
1, 2,  bon
2, 2,  moyen
3, 4,  moyen
4, 4,  moyen
8, 0,  bon
9, 11, mauvais
5, 11, mauvais
6, 3,  bon
7, 12, mauvais
```

Figure 8 : le fichier arff associé à la table input_base

- **Etablissement de l'arbre de décision**

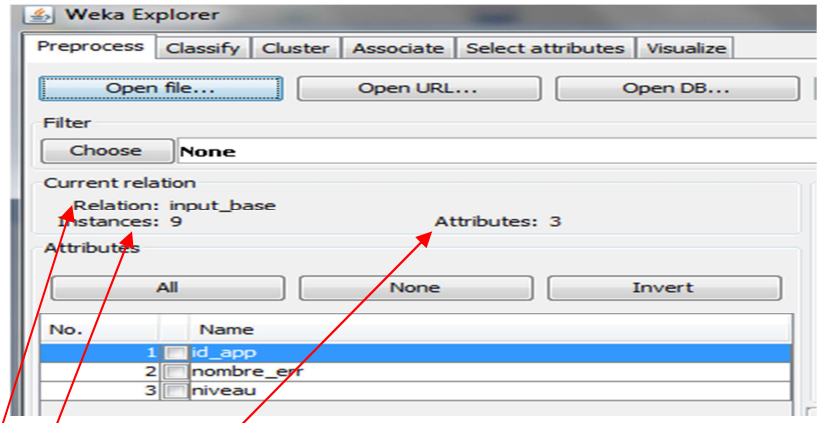
Pour se connecter au fichier précédemment préparé, cliquer sur ***Explorer*** sur la première fenêtre apparue lors du lancement de WEKA. Ensuite, cliquer sur ***Open file...*** de l'onglet ***Preprocess***.

Dans la fenêtre qui s'ouvre choisir le type de fichier arff et naviguer jusqu'à trouver votre fichier puis cliquer sur ouvrir Vous aurez un a perçu du genre :



Cet aperçu comporte plusieurs parties. On y trouve :

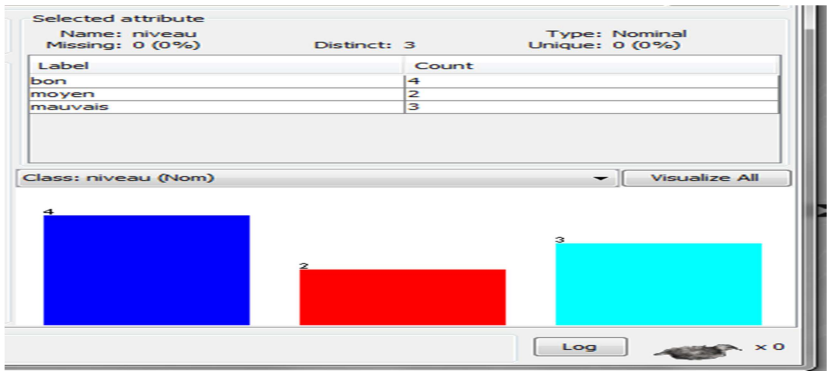
La partie ***Current relation*** :



Qui montre :

- **Relation** : Le nom du fichier arff utilisé, dans ce cas c'est ***input_base.***
- **Instances** : Le nombre *d'instances* du fichier, dans ce cas c'est 9.
- **Attributes** : Le nombre *d'attributs traités*, dans ce cas c'est 3, à savoir *l'id_app, le nombre_err* et *niveau?*

La partie ***Selected attribute*** :



Qui donne des statistiques sur l'attribut sélectionné dans la partie **Attributes** (le nom de l'attribut, son type : numérique ou texte, nombre d'occurrences distinctes, le minimum et le maximum si c'est un attribut numérique, le nombre d'occurrences si c'est un attribut de type texte...). Cette partie comporte aussi un graphe composé de plusieurs couleurs. Chaque couleur représente l'effectif d'une valeur précise de la variable qualité. Pour savoir la signification de chaque couleur, il suffit de sélectionner la variable qualité dans la partie Attributs. Dans notre cas Cette partie comporte un graphe en rectangles composés de 3 couleurs : le bleu (bon niveau), le rouge (moyen niveau) et le vert (mauvais niveau). On peut ainsi savoir le nombre d'occurrences de chaque valeur, ainsi que la couleur associée.

- **Application de l'arbre de décision**

Pour établir l'arbre de décision cliquer sur l'onglet ***Classify***, choisir l'option *Use training set* de Test options comme suit :

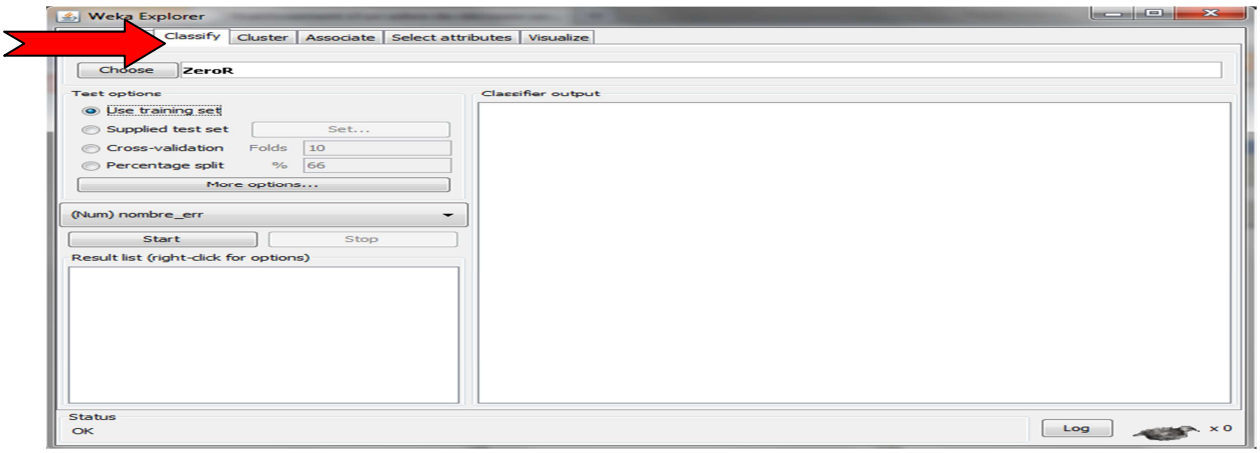
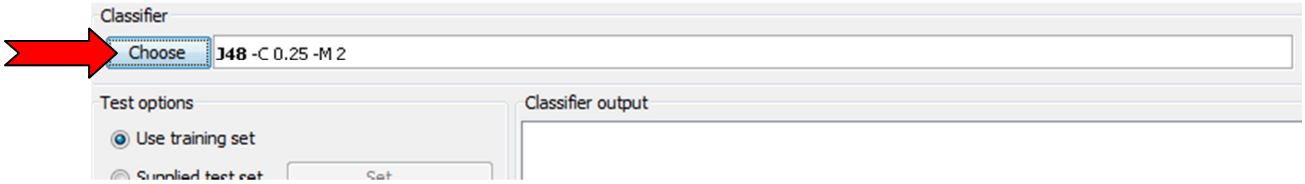


Figure 9: l'onglet classify de l'explorer weka

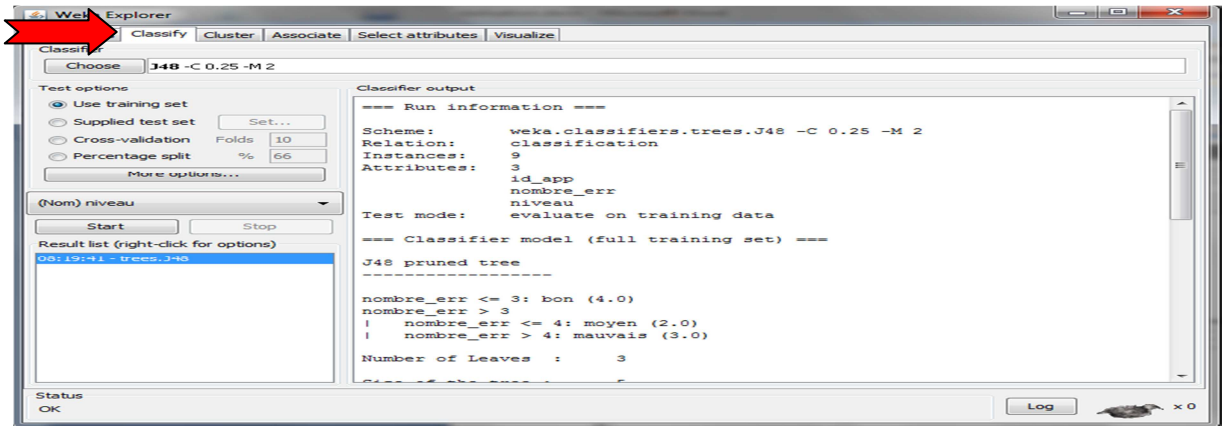
La zone Test options permet de choisir de quelle façon l'évaluation des performances du modèle appris se fera.

L'option *Use training* : utiliser l'ensemble d'apprentissage comme ensemble de test,

Ensuite, cliquer sur le bouton ***Choose*** de Classifier pour choisir un algorithme parmi ceux proposés par WEKA. Dans notre cas on va utiliser l'algorithme **J48** pour l'élaboration de l'arbre de décisions.

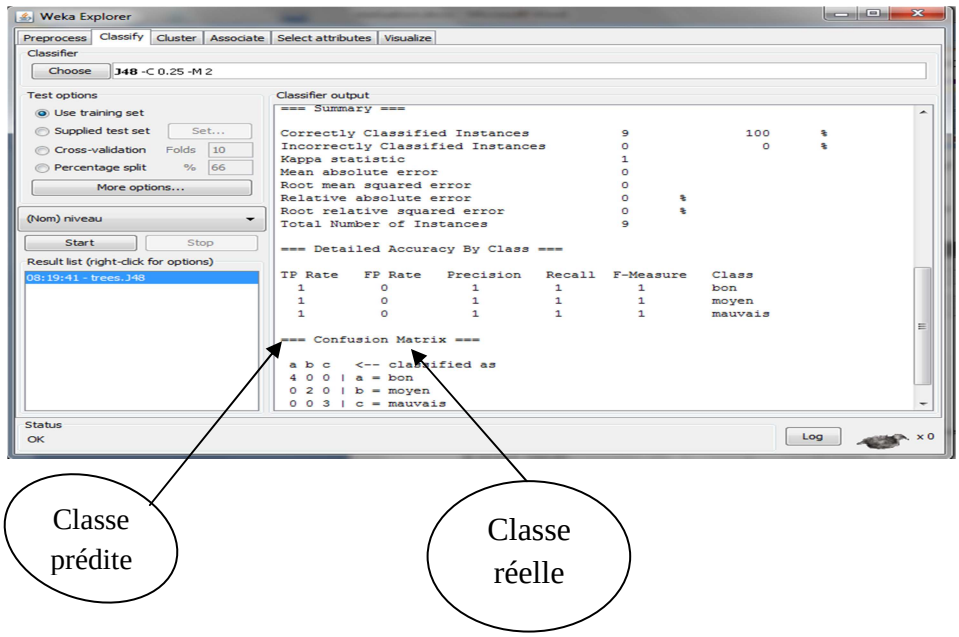


Cliquer sur ***Start*** pour effectuer l'analyse.



Dans la partie ***Classifier output*** vous avez des statistiques sur le fichier exploité, à savoir le ***nombre d'instances*** Total Number of Instances de votre fichier, le nombre d'instances correctement classifiées ***Correctly Classified Instances*** et incorrectement classifiées ***Incorrectly Classified Instances*** et autres statistiques.

Sur le même écran vous avez aussi la matrice de confusion : ***Confusion Matrix*** de cette analyse.



Les résultats du *sommaire* donnent le nombre total d'instances classifiées correctement et incorrectement, la valeur de kappa, l'erreur absolue moyenne, l'erreur racine carrée moyenne, l'erreur relative absolue et l'erreur racine carrée relative. Le choix d'un ou de plusieurs de ces critères dépend de l'intérêt que leur porte l'utilisateur final.

Les résultats par classes nous fournissent le taux d'instances classifiées correctement et incorrectement via les taux TP (true positif) et FP (false positif), la précision et d'autres informations statistiques. Quant à la matrice de confusion, cette table nous donne plus de précision concernant le nombre d'instances correctement et incorrectement classifiées correspondants aux taux TP et FP pour chaque classe. Les éléments de la diagonale de la matrice représentent les instances correctement classifiées. La figure précédente montre les résultats obtenus par l'application de l'algorithme J48 (une implantation de C4.5). Les résultats sont structurés en trois volets : un volet résultats sommaires, un volet résultats par classes et un volet matrice de confusion.

- **Les mesures générales :**
  - ***Correctly Classifiéd Instances***: Le nombre d'apprenants bien classifiés, en valeur absolue, puis en pourcentage du nombre total d'apprenant.
  - ***Incorrectly Classied Instances*** : Sous le même format, le nombre d'apprenant mal classifiés.
  - ***Kappa statistic :*** Le coefficient Kappa est censé mesuré le degré de concordance de deux ou de plusieurs juges. Dans Weka, on est toujours dans le cas de deux juges.

Dans Weka, le jugement, c'est la classe d'un exemple, et les deux juges sont le classifier et la classe réelle de l'exemple. L'accord/désaccord entre les deux juges se lit directement dans la matrice de confusion.

Le coefficient Kappa prend ses valeurs entre -1 et 1.

- Il est maximal quand les deux jugements sont les mêmes : tous les exemples sont sur la diagonale.
- Il vaut 0 lorsque les deux jugements sont indépendants.
- Il vaut -1 lorsque les juges sont en total désaccord.

- **Les mesures d'exactitude par classe :**

Ces valeurs se trouvent dans la partie « ***Detailed Accuracy By Class*** ». Pour chaque classe, Weka fournit cinq mesures **[36].**

Pour une classe donnée, un classifieur, et un exemple, quatre cas peuvent se présenter :

- L'exemple est de cette classe, et le classifieur ne se trompe pas : c'est un ***vrai positif.***
- L'exemple est de cette classe, mais le classifier se trompe : c'est un ***faux négatif.***
- L'exemple n'est pas de cette classe, mais le classifieur la lui attribue quand même : c'est ***faux positif***.
- L'exemple n'est pas de cette classe, et le classifieur ne le range pas non plus dans cette classe : c'est un ***vrai négatif***. Le tableau suivant résume les différentes situations.

	Classe prédite	
	prédite	Non prédite
Exemple de cette classe	Vrai positif	Faux négatif
Exemple n'appartenant pas a cette classe	Faux positif	Vrai négatif

- ***TP Rate*** : Rapport des vrais positifs. Il correspond à :

$$\frac{\textit{Nombre de vrais positifs}}{(\textit{Nombre de vrais positifs} + \textit{Nombre de faux négatifs})}$$

C'est donc le rapport entre le nombre de bien classé et le nombre total d'éléments qui devraient être bien classés. Il se calcule en utilisant la première ligne du tableau de précédent.

- ***FP Rate*** : Rapport des faux positifs. Il correspond, symétriquement à la définition Précédente :

$$\frac{\textit{Nombre de faux positifs}}{(\textit{Nombre de faux positifs} + \textit{Nombre de vrais négatifs})}$$

Il se calcule en utilisant la deuxième ligne du tableau précédent.

La donnée des taux TP Rate et FP Rate permet de reconstruire la matrice de confusion pour une classe donnée. Symétriquement, la matrice de confusion permet de calculer TP Rate et FP.

- ***Précision*** : La précision indique le degré de vérité des résultats obtenus, elle se calcule comme suit : Précision=TP / (TP+FP).

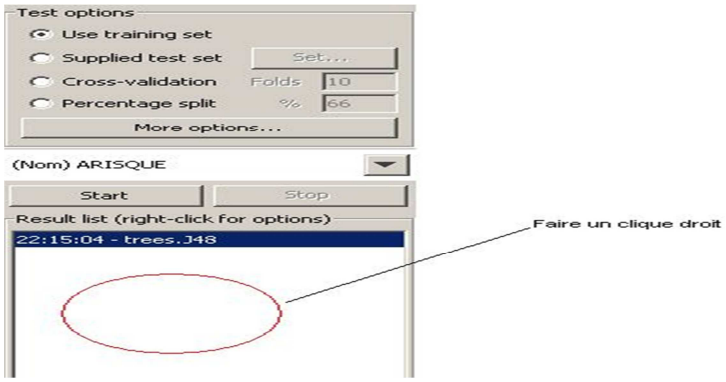
Une valeur de 1 exprime le fait que tous les exemples classés positifs l'étaient vraiment.

- ***Recall*** : le rappel indique le degré de pertinence du résultat. Un Recall de 1 signifie que tous les exemples positifs ont été trouvés.il se calcule comme suit :

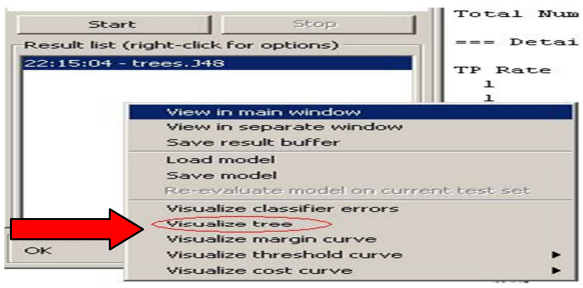
Rappel = TP / (TP + FN).

- ***F-Measure*** :[36] Cette quantité permet de regrouper en un seul nombre les performances du classifieur (pour une classe donnée) pour ce qui concerne le Recall et la Précision : f_mesure = (2 * Precision * Reppel) / (precision + rappel).

Confusion Matrix : est la matrice de confusion, c'est-à-dire un outil servant à mesurer la qualité d'un système de classification. Chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe réelle. Pour afficher l'arbre de décision, cliquer droit dans la partie ***Result list*** (*right-click for options*)



Lors de ce clic droit un menu d'options s'affiche comme suit : Choisir l'option **Visualize tree**.



L'arbre de décision s'affiche ressemblant à ceci :

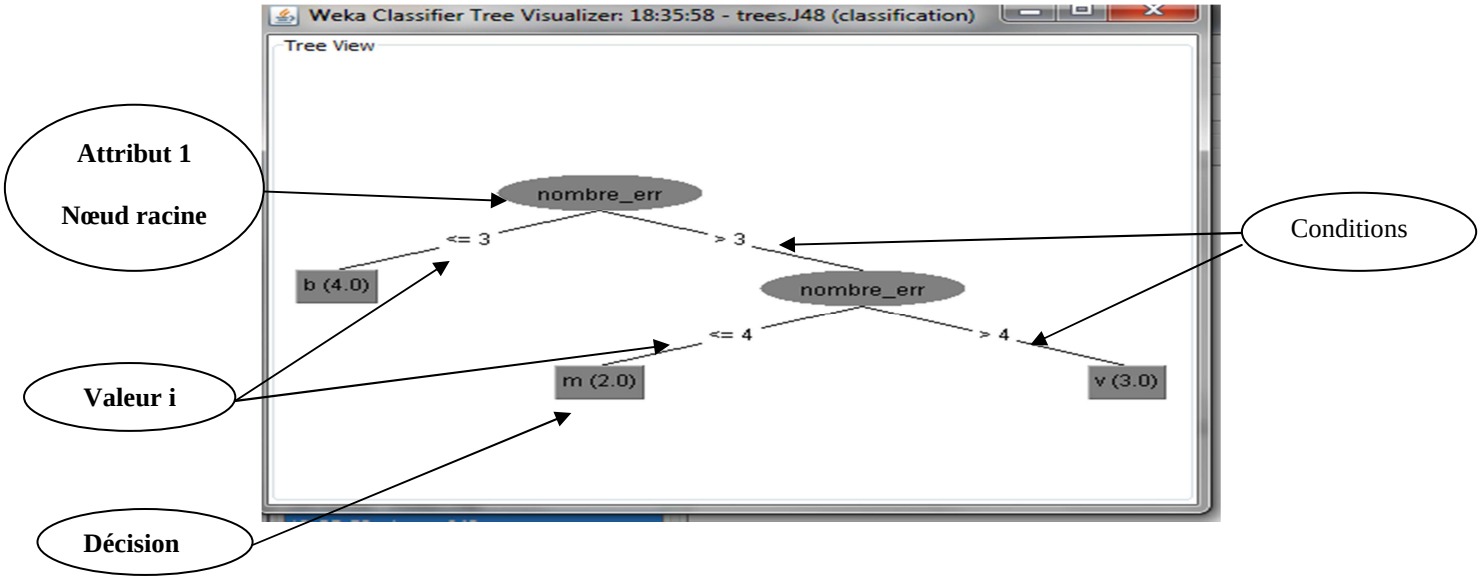


Figure 10 : l'arbre de décision associé aux résultats de la classification

- **Interprétation des résultats de l'arbre**

Dans cet exemple on peut analyser le résultat de la manière suivante, Premièrement, j'ai deux catégories principales d'apprenant :

- Des apprenants ayant un nombre d'erreur inférieur ou égal à 3.
- Des apprenants ayant un nombre_err supérieur strictement à 3.

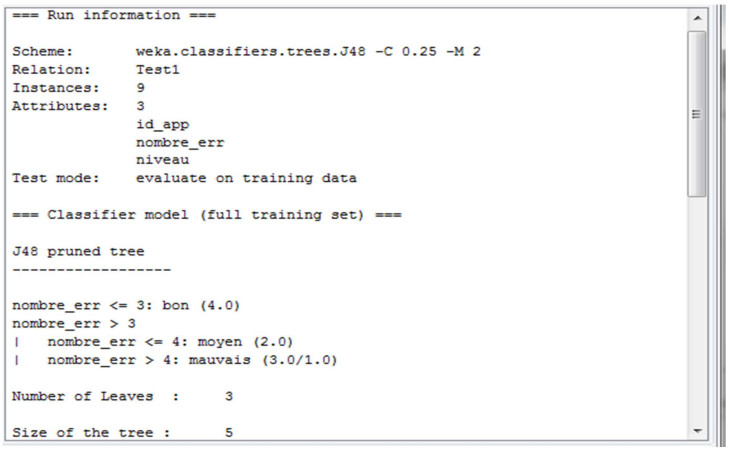
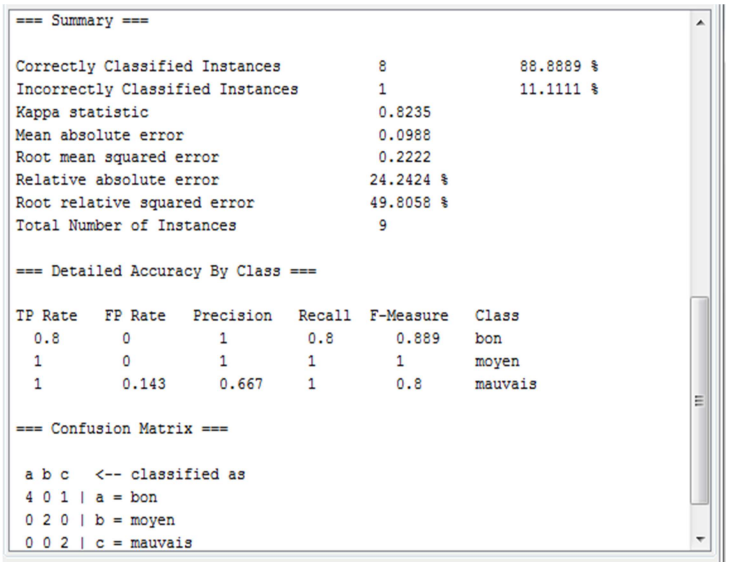
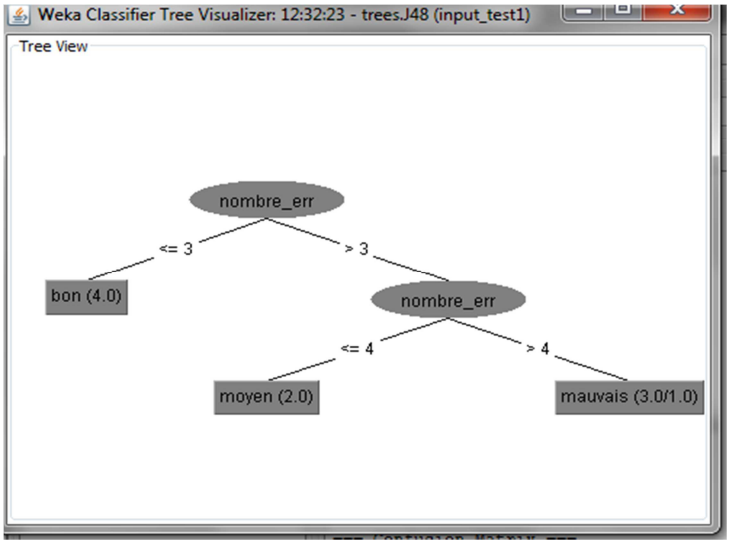
L'arbre montre que les apprenants ayant un nombre_err inférieur ou égal à 3 sont tous des apprenants bons niveau. Les apprenants ayant un nombre_err supérieur strictement à 3 sont subdivisés en deux parties :

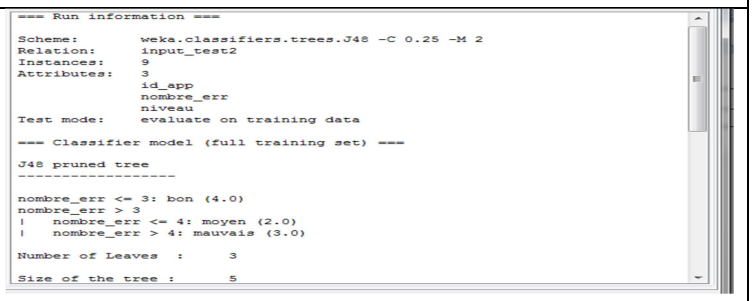
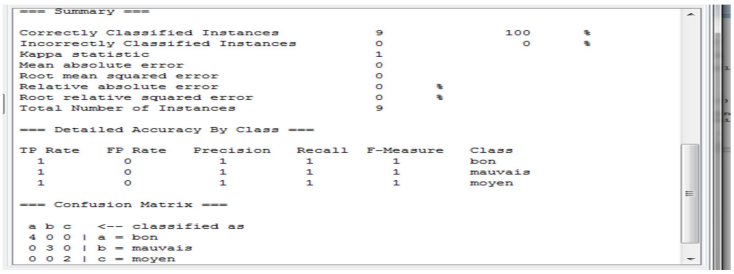
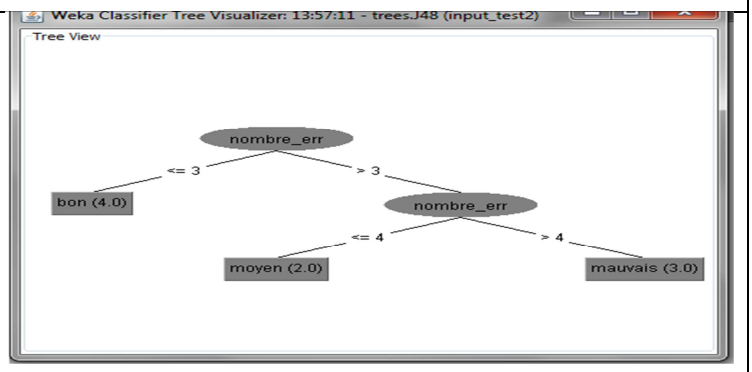
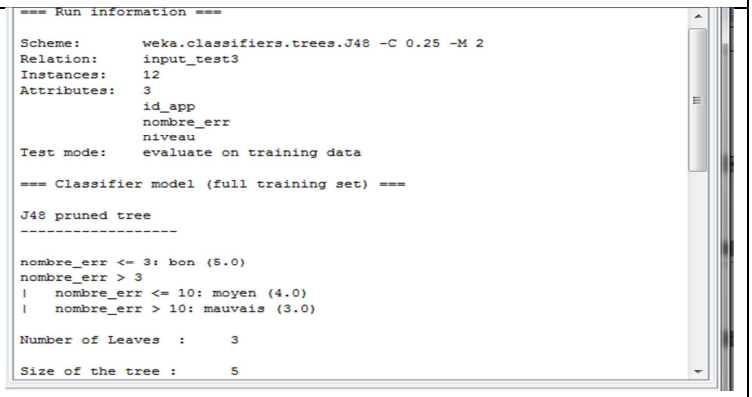
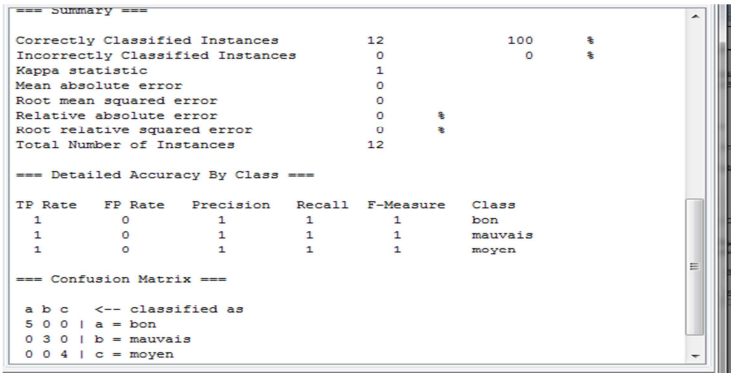
- Des apprenants ayant un nombre d'erreur inférieur ou égal à 4.
- Des apprenants ayant un nombre d'erreur supérieur strictement à 4.

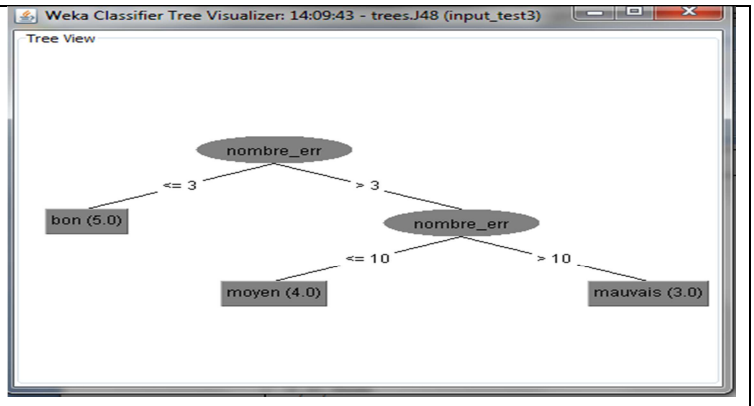
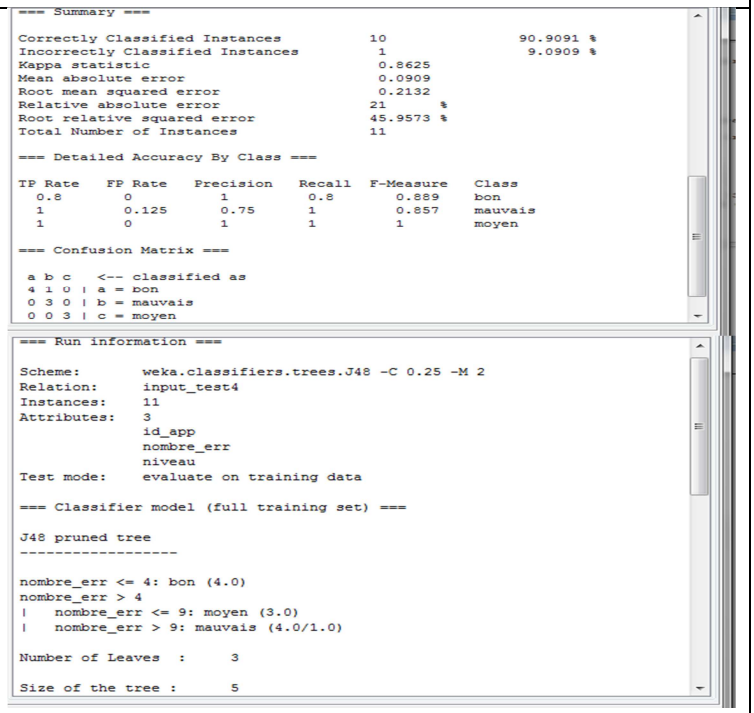
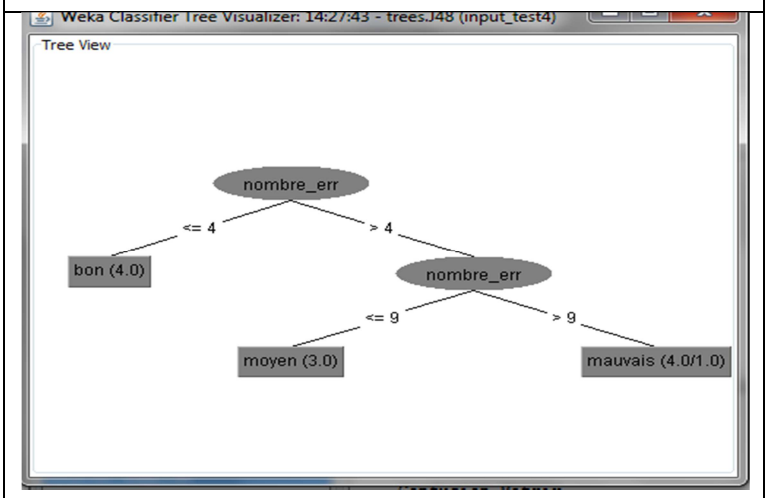
L'arbre montre que les apprenant ayant un nombre d'erreur supérieur strictement à 3 et inférieur ou égal à 4 sont tous des apprenant de niveau moyen. Les apprenants ayant un nombre d'erreur supérieur strictement à 4 sont des apprenants du mauvais niveau.

- **Intérêt de l'algorithme J48 pour l'analyse des erreurs :**

Afin de mettre en lumière l'intérêt de l'algorithme C4.5, nous avons effectués plusieurs tests sur des base_input contenant des valeurs différentes nous avons obtenues les résultats montrés dans le tableau qui suit, donc afin de comprendre l'intérêt de cet algorithme, analysons les résultats effectué sur l'outil weka et essayons d'en tirer son intérêt.

Fichier input	**Arbre de décision C4.5 généré par Weka**
Test1	**Resultat1**
@relation input_test1 @attribute id_app numeric @attribute nombre_err numeric @attribute niveau {bon, moyen, mauvais} @data 1, 2, bon 2, 4, moyen 3, 0, bon 4, 11, mauvais 5, 11, mauvais 6, 3, bon 7, 12, bon 8, 2, bon 9,4, moyen	 === Run information === Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2 Relation: Test1 Instances: 9 Attributes: 3 id_app nombre_err niveau Test mode: evaluate on training data === Classifier model (full training set) === J48 pruned tree ------------------ nombre_err <= 3: bon (4.0) nombre_err > 3 | nombre_err <= 4: moyen (2.0) | nombre_err > 4: mauvais (3.0/1.0) Number of Leaves : 3 Size of the tree : 5  === Summary === Correctly Classified Instances 8 88.8889 % Incorrectly Classified Instances 1 11.1111 % Kappa statistic 0.8235 Mean absolute error 0.0988 Root mean squared error 0.2222 Relative absolute error 24.2424 % Root relative squared error 49.8058 % Total Number of Instances 9 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure Class 0.8 0 1 0.8 0.889 bon 1 0 1 1 1 moyen 1 0.143 0.667 1 0.8 mauvais === Confusion Matrix === a b c <-- classified as 4 0 1 | a = bon 0 2 0 | b = moyen 0 0 2 | c = mauvais 

Test2	**Résultat 2**
@relation input_test2 @attribute id_app numeric @attribute nombre_err numeric @attribute niveau {bon, moyen, mauvais} @data 1, 2, bon 2, 4, moyen 3, 0, bon 4, 11, mauvais 5, 11, mauvais 6, 3, bon 7, 12, mauvais 8, 2, bon 9, 4, moyen	  
Test3	**Resultat3**
@relation input_test3 @attribute id_app numeric @attribute nombre_err numeric @attribute niveau {bon, mauvais, moyen} @data 1, 2, bon 11, 2, bon 12, 4, moyen 2, 4, moyen 3, 0, bon 4, 11, mauvais 5, 11, mauvais	 

6, 3, bon 7, 12, mauvais 8, 3, bon 9, 4, moyen 10, 10, moyen	
Test 4	**Resultat4**
@relation input_test4 @attribute id_app numeric @attribute nombre_err numeric @attribute niveau { bon , mauvais , moyen } @data 1, 2, bon 2, 12, bon 2, 6, moyen 3, 0, bon 4, 13, mauvais 5, 13, mauvais 6, 3, bon 7, 12, mauvais 8, 4, bon 9, 7, moyen 10, 9, moyen	 

Nous remarquons que dans :

Test1 : Un apprenant a été mal classé, il été classé de bon niveau alors qu'il est de mauvais niveau (nombre_err =12) ce qui est montré dans l'arbre dans la feuille de décision mauvais (3.0/1.0).

Donc l'algorithme nous a permet de détecté les anomalies de notre classification, ce qui est un point fort dans le C4.5 et qui aide le mieux le processus de suivi des apprenant dans des sessions d'apprentissage.

Pour la condition nous remarquons que : L'algorithme a choisit les tests qui aide au mieux la classification, ici deux tests (seuil=3, seuil=4)

Bon niveau =<3, 3<Moyen niveau=<4 , Mauvais niveau >4.

Test 2 : bonne classification du moment que tous les attributs sont bien classées .les conditions restent les même puisque les valeurs sont très proche aux valeurs du test 1.La différence réside dans les valeurs des mesure de qualité (pour test2 nombre d'instance correctement classifiés égal aux nombre d'instance total qui est 9 et le nombre d'instance mal classifies = 0 par contre dans le test 1 nombre d'instance correctement classifiés = 8, et le nombre d'instance mal classifiés égal a 1).

Test 3 : bonne classification, mais la condition a changé du moment qu'on a ajouté de nouvelles valeurs l'algorithme a pris un nouveau seuil de moyen niveau (si 10>= nombre_err >3 alors niveau moyen, et si nombre_err>10 alors mauvais niveau).Les valeurs de mesure de qualité et les même que le test 1 puisque c'est une bonne classification.

Test 4 : bonne classification. On remarque en enlevant l'apprenant ayant un nombre_err=10 et en le remplaçant par un nombre_err= 9 la condition change le seuil devient9, même chose pour le premier seuil qui =4 au lieu de 3.on a classifié un apprenant (apprenant 2) comme étant du bon niveau or qu'il ne l'est pas ,mais l'algorithme a constaté ça ,il la classifié dans la classe mauvais niveau, donc l'outil est capable d'apprendre des règle qui lui permettent de généré le meilleur arbre de décision correspondant à une bonne classification.

D'après les différents résultats des tests précédent nous déduisant que l'algorithme C4.5 est un algorithme intéressant pour la classification des erreurs car il permet de détecté les instances mal classés, il permet aussi de changer la condition du test selon les variables données en entrés ce qui nous donne des bons résultats, et qui utilise ces règle apprise durant la classification.

IV.2.2.2 Recherche de règle d'associations entre les erreurs :

- **Préparation des données :**

Les règles d'association à établir nous permettront de détecter les corrélations entre les erreurs effectuées par des apprenants dans des exercices d'une session d'apprentissage. Nous utiliserons dans ce cas la table ***Input_base2***, qui est la jointure entre les deux tables session, erreur_effec,

C'est ce résultat qu'on a exporté vers un fichier CSV (expliquer précédemment).

- **Application de l'algorithme de règle d'association apriori:**

Après transformation du fichier CSV en un fichier ARFF acceptable par Weka on va lui appliquer l'algorithme apriori comme suit :

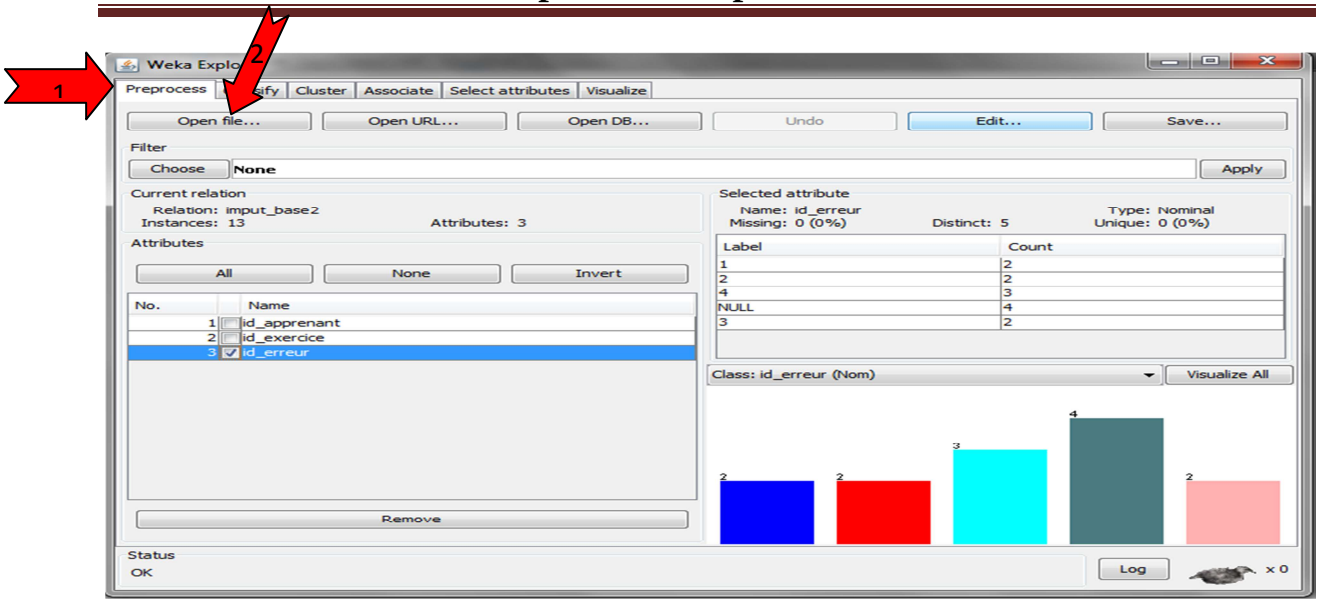
- Après ouverture de fenêtre ***Explorer*** montré précédemment : Le ***open fil*** va nous permettre d'importer la base correspondante.
- La structure de cette base peut être affichée via le bouton ***edit*** comme suit :

Viewer

Relation: apriori

No.	id_apprenant Nominal	id_exercice Nominal	id_erreur Nominal
1	1	1	1
2	1	1	2
3	1	2	4
4	1	3	0
5	3	2	3
6	3	2	3
7	3	2	4
8	3	3	0
9	2	1	1
10	2	1	2
11	2	3	0
12	4	2	4
13	4	3	0

L'ouverture du fichier de notre base arff fait apparaitre la fenêtre suivante :



Pour savoir la signification de chaque couleur il suffit de choisir l'attribut qualité dans notre cas, on choisit id_err donc on voie bien sur la figure que l'erreur id_err=1 correspond a la couleur bleu et le nombre d'occurrence de cette erreur est 2, et que id_err=4 correspond a la couleur vert et son nombre d'occurrence est 3.

➢ Afin d'avoir les règle d'associations entre les erreurs effectués, il suffit de choisir l'onglet ***Associates*** de l'outil Explorer, puis choisir l'algorithme désiré, dans notre cas c'est l'algorithme Apriori, via le bouton ***choose.***

➢ Une fois l'algorithme choisi, on peut l'appliquer directement en cliquant sur le bouton ***Start***, ce qui permet de générer la fenêtre suivante :

La confiance d'une règle : si condition alors conclusion est le rapport : Nombre de données ou les items de la condition et de la conclusion apparaissent simultanément / nombre de données ou les items de la condition apparaissent simultanément

- **Interprétation des résultats :**

L'algorithme apriori génère les meilleures règles qu'on peut avoir sur la base, (donc on peut interpréter Les règle résultantes comme suit :

- Pour toutes les occurrences de l'exercice 3 dans la base (occurrence est égal à 4).

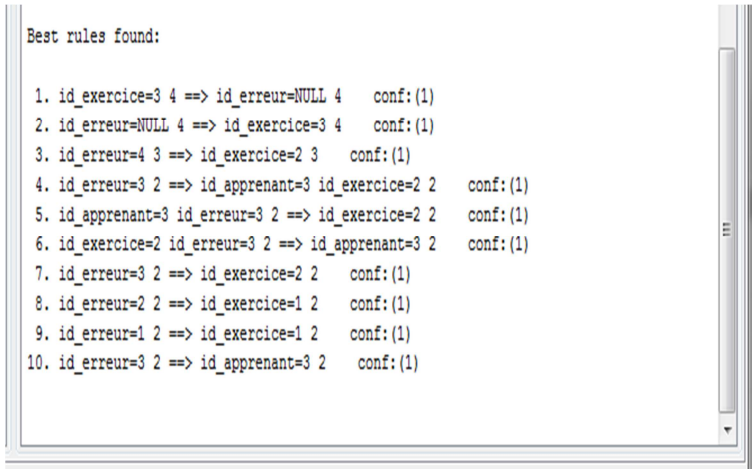
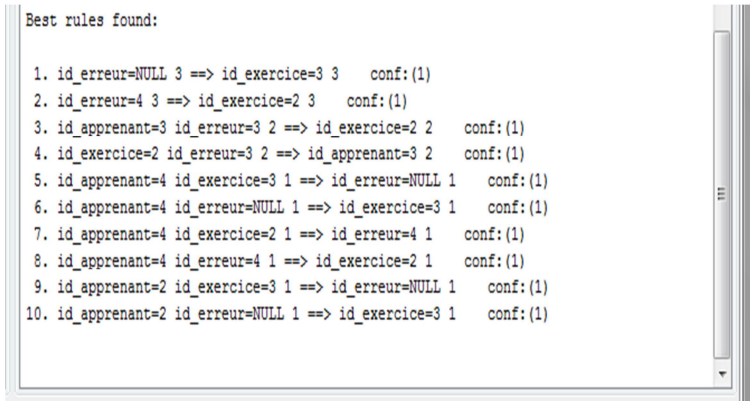
Correspond une occurrence Null de l'id_erreur dans la base (nombre d'occurrence= 4 aussi), Null c.à.d. pas d'erreurs.

Si l'id_exercice est 3 alors l'id_erreur =Null. C.-à-d. qu'a chaque fois on effectue l'exercice 3 ; pas d'erreurs.

- La règle 2 est la réciproque de la 1.
- Pour toute occurrence de l'erreur 4 dans la base correspond une occurrence de l'exercice 2, Le nombre d'occurrence de l'erreur 4 est 3, Le nombre d'occurrence de l'exercice 2 est 3.
- Si l'id_erreur est 4 alors l'exercice effectué est l'exercice 2, on peut dire que l'erreur 4 est toujours faite dans l'exercice2.
- A chaque fois que l'apprenant 3 effectue l'exercice 2, il refait toujours la même erreur.
- Idem avec règle 4.
- Idem avec règle 4.
- L'erreur 3 s'effectue toujours dans l'exercice 2.
- L'erreur 2 s'effectue toujours dans l'exercice 1.
- L'erreur 1 s'effectue toujours dans l'exercice 1.
- L'erreur 3 est effectuée toujours par l'apprenant 3.

- **Intérêt de l'algorithme apriori**

Afin de mettre en lumière l'intérêt d'un algorithme de règle d'association (Apriori), nous avons effectués plusieurs tests sur des base_input contenant des valeurs différentes nous avons obtenues les résultats montrés dans le tableau qui suit, afin de comprendre l'intérêt de cette algorithme, analysons les résultats de l'algorithme apriori effectué sur l'outil weka et essayons d'en tirer son intérêt.

Fichier input	**Résultat de l'algorithme apriori généré par Weka**
Test 1	**Résultat 1**
@relation imput_test1 @attribute id_apprenant {1, 3, 2,4} @attribute id_exercice {1 ,2 ,3} @attribute id_erreur {1, 2, 4, NULL, 3} @data 1, 1, 1 1, 1, 2 1, 2, 4 1, 3, NULL 3, 2, 3 3, 2, 3 3, 2, 4 3, 3, NULL 2, 1, 1 2, 1, 2 2, 3, NULL 4, 2, 4 4, 3, NULL	
Test 2	**Résultat 2**
@relation imput_test2 @attribute id_apprenant {1, 3, 2, 4} @attribute id_exercice {1, 2, 3} @attribute id_erreur {1, 2, 4, NULL, 3,5} @data 1, 1, 3 1, 1, 5 1, 2, 4 1, 3, 1 3, 2, 3 3, 2, 4 3, 3, NULL 2, 1, 1 2, 1, 2 2, 3, NULL 4, 2, 4 4, 3, NULL	

Nous remarquons en comparons :

Test 1 et Test 2: dans le test 1 nous avons la règle suivante :

A chaque fois qu'un apprenant résout l'exercice 3, il le résout sans erreur, dés que nous ajoutons un apprenant qui résout l'exercice trois avec une erreur cette règle disparait et d'autre règle apparaissent, mentionnant aussi que le nombre d'occurrence des items fréquents change (de même pour les autre règles, a chaque modification des données, les règles générées changent) .

D'après les résultats des tests précédents nous pouvons conclure que les algorithmes de datamining en générale et ceux de la classification et règle d'association en particulier sont des moyen qui permettent d'accélérer le processus d'évaluation grâce aux résultat qu'il fournissent dans un temps réduit ,il peut aussi satisfaire aux mieux l'enseignant ,en l'aidant à la prise de décision ,et d'autre part l'apprenant qui est l'intéressé des décisions prise par le pédagogue.

IV.2.3 Intégration de l'outil weka dans Web Siela3 :

Une plate forme d'e-learning ou de formation à distance est un logiciel qui assiste la conduite des enseignements à distance, elle regroupe les outils nécessaires aux trois principaux utilisateurs : enseignant, apprenant et administrateur. Mais peu de plates formes intègrent les outils du datamining pour le suivi du processus d'apprentissage des apprenants, et s'ils existent, ils sont de nombre limité. Ce qui a été l'objectif de notre travail et de mettre en lumière l'intérêt de l'utilisation des outils du datamining dans l'e-learning .

Pour cela après avoir étudié un outil de datamining Weka vu précédemment et vu l'importance que porte l'application de ces algorithme en e-learning , nous avons intégré un espace datamining dans une plate forme intitulée Websiela **(s**ystème **i**nformatique pour l'**e**nseignement du **l**angage **a**lgorithmique sur le **Web)** ,qui est un système d'enseignement permettant aux auteurs d'introduire le domaine d'enseignement (cours, exercices, solutions, exemples et erreurs) et de les mettre sur un serveur web à la disposition des apprenants, qui peuvent s'inscrire et de bénéficier des enseignements disponibles.

- Dans le système web siela déjà existant dans un environnement PHP (Easy PHP) nous avons ajouté un espace de ***data ming***, auquel on peut accéder en cliquant sur le bouton Datamining de la page d'accueil de Web Siela , voir la figure suivante :

Figure 11 : ***page d'accueil Websiela3***

Il suffit de cliquer sur le bouton ***Datamining*** pour avoir l'espace de data mining, ou on a le menu qui apparait dans la fenêtre suivante :

- Pour utiliser Weka cliquer sur ***utiliser Weka***.
- Pour voir les résultats de classification pour les différentes sessions, cliquer sur ***Résultat Classification***.
- Pour voir les résultats de règle d'association, cliquer sur ***Résultats Règles d'associations***

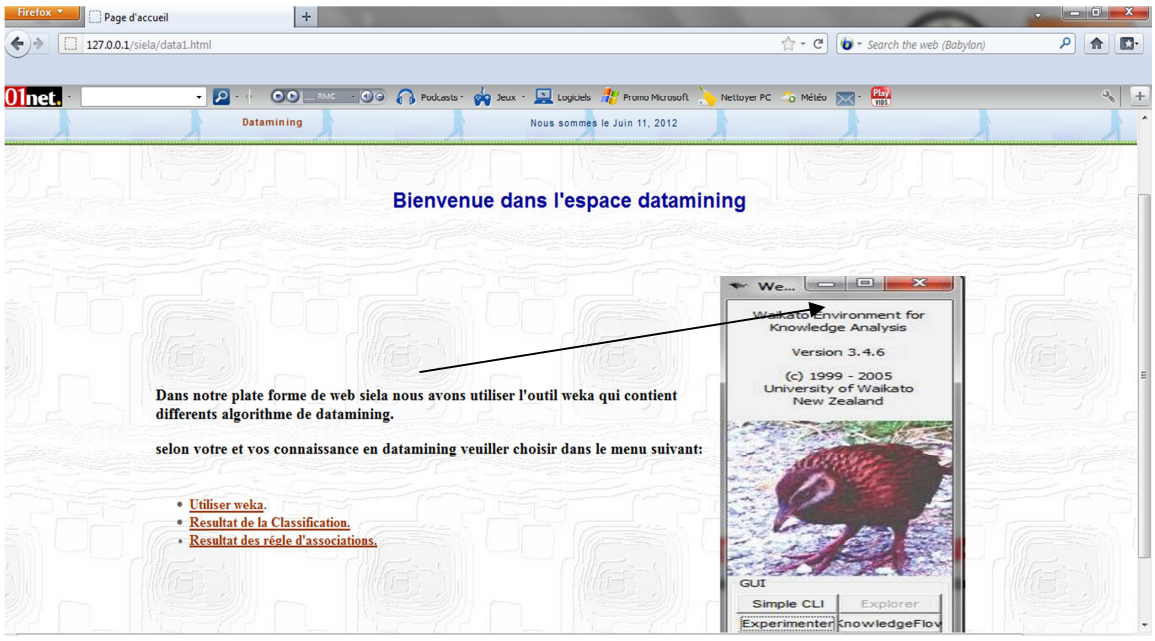


Figure 12*: espace datamining*

Une fois cliqué sur résultat de classification, la fenêtre suivante apparait, permettant de choisir le résultat du test de la session que l'on veut consulter.

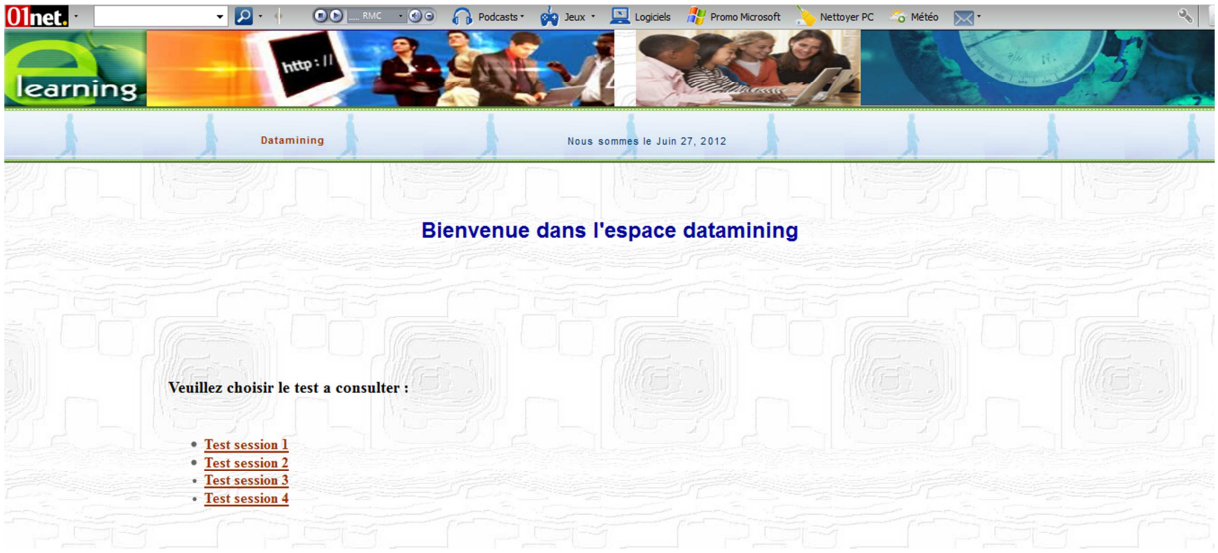


Figure 13 : choix des résultats de la session à consulter

En cliquant sur l'un des résultats (**Test session1**), il sera affiché dans la fenêtre suivante :

```
J48 pruned tree
------------------

nombre_err <= 3: bon (4.0)
nombre_err > 3
|   nombre_err <= 4: moyen (2.0)
|   nombre_err > 4: mauvais (3.0/1.0)

Number of Leaves  :      3

Size of the tree :       5


Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances           8               88.8889 %
Incorrectly Classified Instances         1               11.1111 %
Kappa statistic                          0.8235
Mean absolute error                      0.0988
Root mean squared error                  0.2222
Relative absolute error                 24.2424 %
Root relative squared error             49.8058 %
Total Number of Instances                9

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
   0.8          0           1          0.8          0.889    bon
   1            0           1          1            1        moyen
   1            0.143       0.667      1            0.8      mauvais

=== Confusion Matrix ===

 a b c   <-- classified as
 4 0 1 | a = bon
 0 2 0 | b = moyen
 0 0 2 | c = mauvais
```

Figure 14 : résultat du test de la session 1

Pour consulter les résultats des règles d'associations, il suffit de choisir le test à consulter, cliqué, la fenêtre suivante s'affiche :

```
Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:     whatever
Instances:    13
Attributes:   3
              id_apprenant
              id_exercice
              id_erreur
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 29

Size of set of large itemsets L(3): 12

Best rules found:

 1. id_erreur=NULL 3 ==> id_exercice=3 3    conf:(1)
 2. id_erreur=4 3 ==> id_exercice=2 3    conf:(1)
 3. id_apprenant=3 id_erreur=3 2 ==> id_exercice=2 2    conf:(1)
 4. id_exercice=2 id_erreur=3 2 ==> id_apprenant=3 2    conf:(1)
 5. id_apprenant=4 id_exercice=3 1 ==> id_erreur=NULL 1    conf:(1)
 6. id_apprenant=4 id_erreur=NULL 1 ==> id_exercice=3 1    conf:(1)
 7. id_apprenant=4 id_exercice=2 1 ==> id_erreur=4 1    conf:(1)
 8. id_apprenant=4 id_erreur=4 1 ==> id_exercice=2 1    conf:(1)
 9. id_apprenant=2 id_exercice=3 1 ==> id_erreur=NULL 1    conf:(1)
10. id_apprenant=2 id_erreur=NULL 1 ==> id_exercice=3 1    conf:(1)
```

Figure 15 : résultat de test de la session 1 pour les règles d'association

IV.3 L'utilisation de l'environnement java

Afin de testé l'utilisabilitè des algorithmes de datamining, nous avons appliqué l'algorithme de classification J48, dans un environnement Java.

IV.3.1 Outils de programmation

Pour implémenter un tel système nous avons besoin d'un langage tel que JAVA.

Pourquoi JAVA?

Le langage java, né en 1995, est un langage orienté objets, il permet d'écrire de façon simple et claire des programmes portables sur la majorité des plates-formes. De plus, il bénéficie d'une très grande bibliothèque de classes avec lesquelles l'utilisateur pourra composer des interfaces graphiques, créer des applications multithreads, animer une page HTML par des applets ou encore communiquer en réseau.

Dans notre cas nous avons besoin d'implémenter une applet. Pour cela nous devons utiliser une librairie Weka qui est un ensemble d'outils nécessaire à la construction et à la mise en service des algorithmes du datamining. Ces outils peuvent servir également à l'analyse et au test. Parmi les plateformes libres, il y en a quelques-unes très connues pour être utilisées dans le développement de plusieurs applications. Mais celui qu'on a utilisé est Weka.

IV.3.2 Application de l'applet

Nous désignons par l'étape du data mining l'application de notre applet sur le jeu de données précédemment préparé et dont le résultat sera la classification de ces données en un nombre de classes homogènes, selon un arbre de décision (J48 implantation de C4.5). via une applet java en utilisant la librairie Weka.

Premièrement, nous présentons la façon de création de notre projet dans éclipse (Eclipes GALILIO V1.0 2000,2009).qui s'effectue par les étapes suivante :

1. Création d'un projet java. Incluant les différentes classes du projet.
2. Créer un répertoire de bibliothèque qui s'appelle "lib" de même niveau avec "src", puis mettre le fichier "external library Weka – weka-lib.jar" dans le répertoire "lib".
3. Dans Eclipse choisir le projet -> clique droit -> Properties -> Java Build Path -> Libraries -> Add Externals Jars... -> choisir : lib/weka-lib.jar -> ok.
4. Maintenant on peut lancer ***DecisionTreeApplet.java*** comme une applet à partir du code source.

Détail de la première étape : nous avons implémenté 6 classes correspondant aux 6 fichiers java mais 3 classes principales du projet sont : Database, DecisionTree et DecisionTreeApplet.

- **Database** : contient la base de donnée à exécuter et fournit quelques fonctions de traitement sur la base de données comme :
  - Récupérer une base de données sauvegarder dans un fichier ***arff*** ou télécharger une base de données sur Internet et l'enregistrer
  - Prendre les attributs nominaux pour choisir l'attribut à classer, prendre tous les attributs de la base de données comme les candidats qu'on peut utiliser pour la construction de l'arbre de décision.
  - Calculer le « gain d'information » des attributs choisis à utiliser dans la construction de l'arbre de décision.
  - Dans le cas d'une grande base de données séparer une partie pour l'apprentissage et une pour le test de la base de données. Partie d'apprentissage : il y a 3 manières d'extraire : Extraire aléatoirement un certain nombre des instances de la base, extraire un certain nombre de premières instances, extraire un certain nombre de dernières instances, Partie de test : il y a 5 façons de créer l'échantillon de test : L'échantillon de test ≡ ceux d'apprentissage, l'échantillon de test est le complément de ceux d'apprentissage, extraire aléatoirement un certain nombre des instances de la base, extraire un certain nombre de premières instances, extraire un certain nombre de dernières instances.

- **DecisionTree** : fournit des fonctions de manipulation avec l'arbre de décision ci-dessous :
  - Construire un arbre de décision à partir de l'échantillon d'apprentissage contenant des attributs choisis par l'utilisateur.
  - Calculer le taux d'erreur de l'arbre de décision construit sur la partie d'apprentissage.
  - Prédire ou bien classer les instances de la base de données par l'évaluation de l'arbre de décision.
  - Visualiser graphiquement l'arbre de décision : nous prenons la description de l'arbre de décision en Weka pour l'afficher graphiquement sur l'écran.

- **DecisionTreeApplet** : c'est une interface graphique d'interaction entre le programme et l'utilisateur. Cette classe fournit aux utilisateurs les fonctions suivantes :
  - Télécharger une base de données, afficher tous ses attributs.

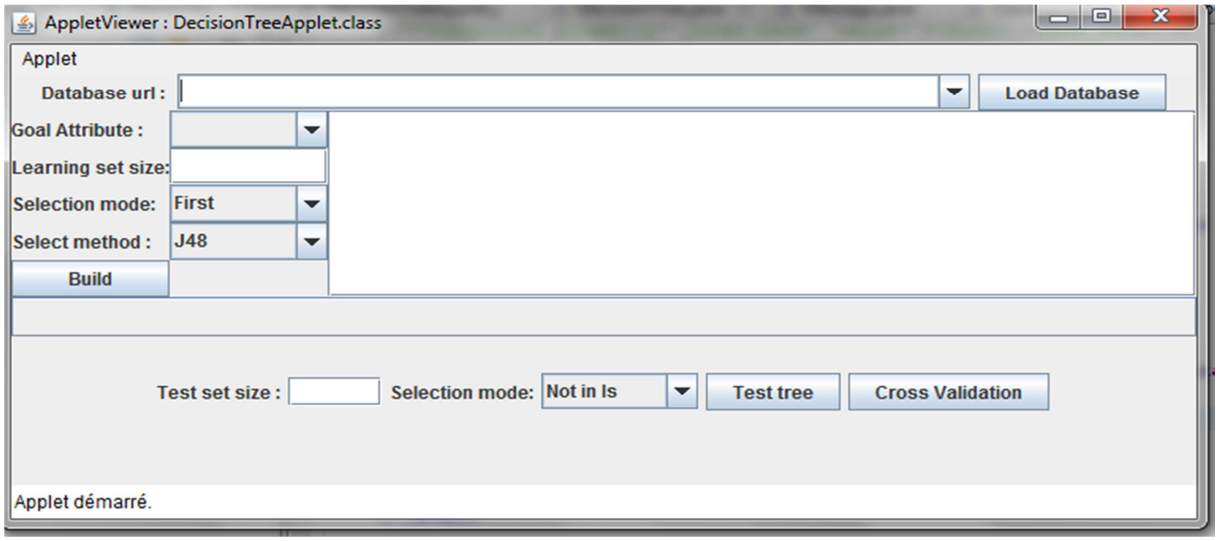
- Établir le type de l'arbre de décision : J48, ID3,... Diviser la base de données en partie d'apprentissage et en partie de test, construire l'arbre de décision.
- Calculer le taux d'erreur de l'arbre sur l'échantillon d'apprentissage, tester l'arbre sur l'échantillon de test visualiser le gain-info des attributs utilisés.
- Visualiser graphiquement l'arbre de décision

• **DataInput** : Enregistre quelques sources des bases de données au format arff. Ces sources sont affichées dans le champ d'adresse pour que l'utilisateur choisisse une source à travailler ou a testé.

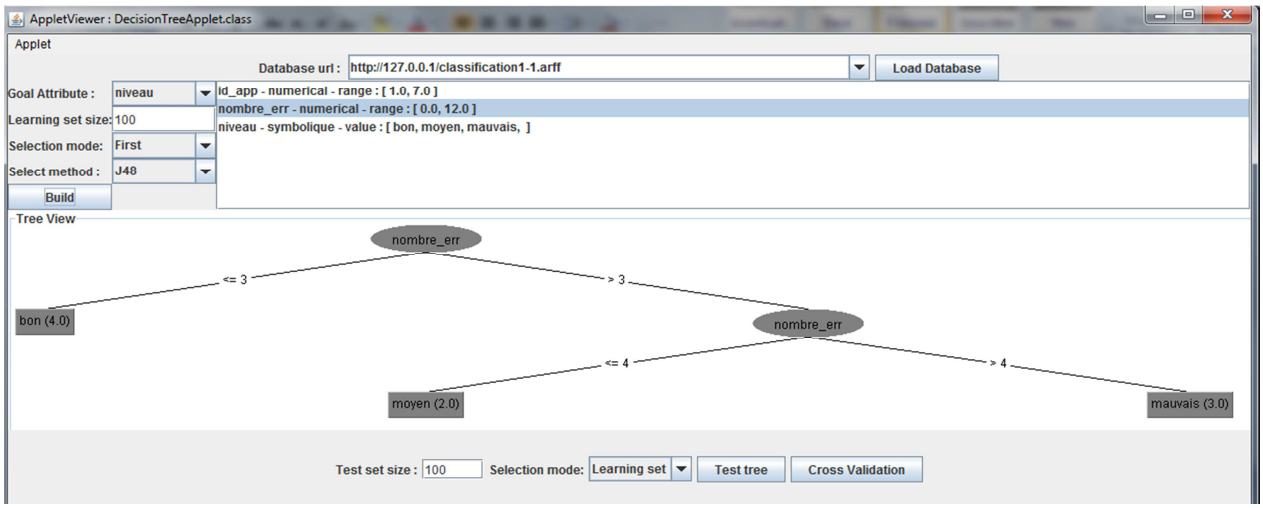
• **Constant** : Enregistre les codes des différents arbres de décision.

• **Message** : C'est un dialogue pour informer l'utilisateur de quelques informations, par exemple : l'échec du téléchargement de la base de données.

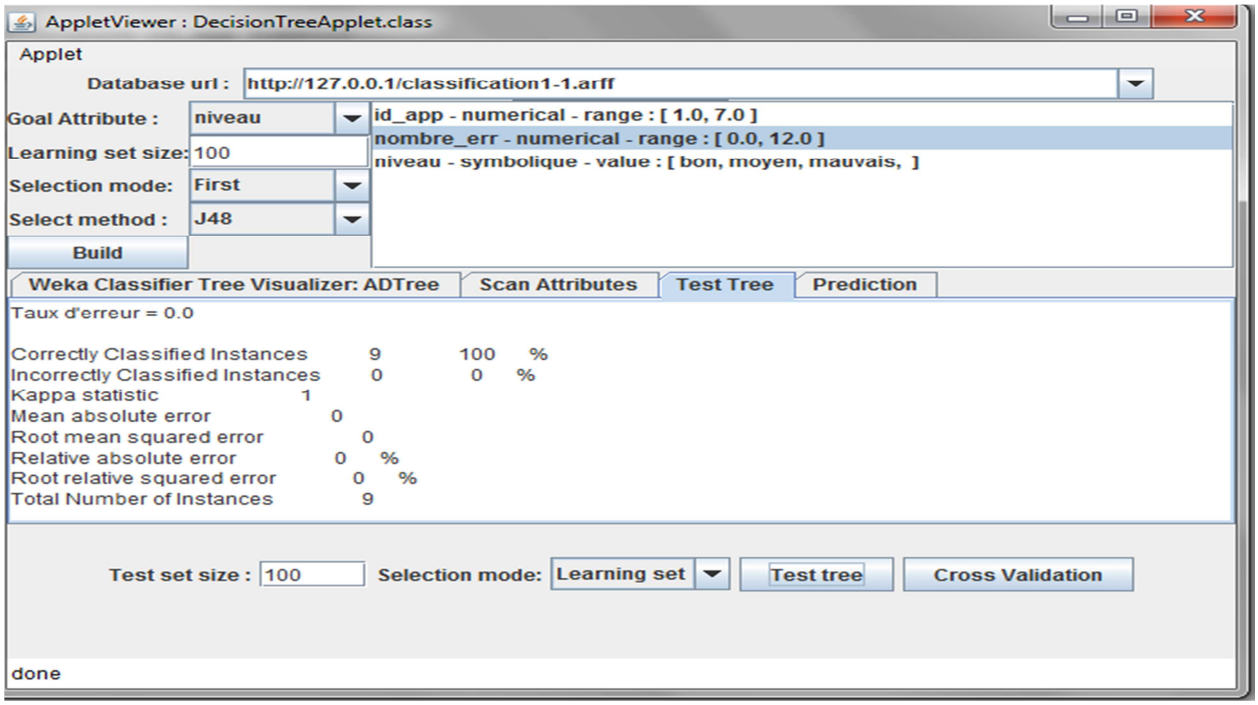
La figure suivante montre la première fenêtre qui apparait lors du lancement du système et qui est la fenêtre principale de l'Applet, elle désigne que l'Applet est démarrée.



Pour télécharger les données sur lesquelles on effectue le test de classification, il suffit de parcourir et choisir un fichier et un clic sur le bouton ***Load Dtabase***, ensuite choisir l'algorithme de classification via ***Selection method*** (j48 dans notre cas), enfin il suffit d'un clic sur le bouton ***Build pour construire l'arbre de décision,*** l'interface suivante apparait :

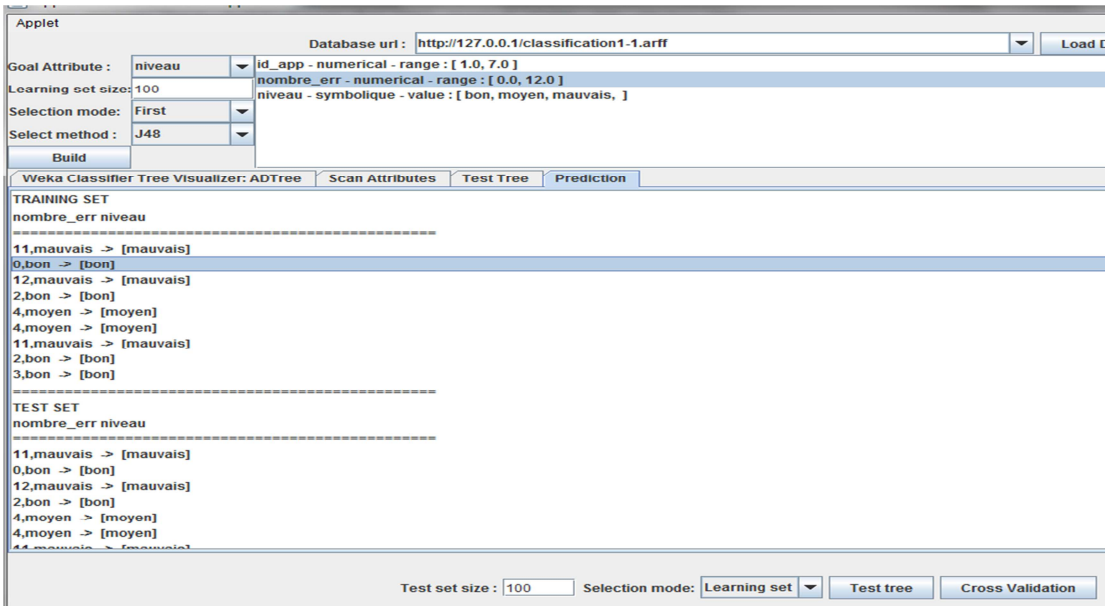


Afin de tester la qualité des résultats de la classification, cliquer sur Test tree l'interface suivante apparait :



Qui contient les mesures générales de qualité (les instances correctement classifiées, incorrectement classifiées…)

Par ailleurs, vous pouvez voir l'évaluation de l'arbre de décision sur l'attribut à classer, ou bien la prédiction de l'arbre, dans la table « Prédiction ». Nous présentons l'évaluation de l'arbre pour la partie d'apprentissage et aussi pour celle de test.



Extrait des points intéressant du code java :

```
Import weka.core.*;
Import weka.core.converters.*;
Import weka.core.converters.ConverterUtils.DataSource;
Import weka.classifiers.Classifier;
Import weka.classifiers.trees.ADTree;
Import weka.classifiers.trees.adtree.*;
Import weka.attributeSelection.*;
Import weka.experiment.InstanceQuery;
Import weka.filters.unsupervised.attribute.Remove;
Import weka.gui.treevisualizer.Node;
Import weka.gui.treevisualizer.PlaceNode2;
Import weka.gui.treevisualizer.TreeBuild;
Import weka.gui.treevisualizer.TreeVisualizer;
Import weka.classifiers.Evaluation;
```

Entête de la classe appel des méthodes de Weka

```
      // telecharcher la base de données
      Int loadDatabase(String file)
      {       try
            {
      DataSource source = new Data Source (file);
      data = source.getDataSet();
      data.setClassIndex(data.numAttributes() );
                  filtered_data = data;
            }
```

Méthode pour le téléchargement de la base de données

Classe DataInput.java

```
Public class DataInput {

      Public static String dataInputs[] = {
"","file:///C:/Users/chabha/Desktop/test1.arff",

"file:///D:/NOTRE%20BASE/test%20classification/test1.arff”,

"file:///D:/NOTRE%20BASE/test%20classification/test2.arff",

"file:///D:/NOTRE%20BASE/test%20classification/test3.arff",

"file:///D:/NOTRE%20BASE/test%20classification/test4.arff"

            };

}
```

La classe DataInput contenant les chemins vers les fichiers arff (base d'apprentissage)

Classe constant.java

```
import java.util.Vector;

public class Constant {

      public final static int J48 = 0;

      public final static int AD_TREE = 1;

      public static String methods [] = {

            "J48",

           "AD_TREE"

} ;

}
```

Classe effectuant l'enregistrement de différentes méthodes des arbres de décisions

IV.4 Conclusion

Au niveau de ce chapitre nous avons présenté l'environnement de développement Weka et l'utilisation des ces algorithmes sur une base d'apprentissage.

L'objectif de cette étude réside dans un premier lieu dans l'utilisation de l'arbre de décision pour la classification de l'apprenant.

Pour répondre à un autre objectif de la fouille de données il est important de disposer de moyen pour sélectionner des règles utiles intéressantes, surprenantes, exprimant des connaissances insoupçonnées. Nous avons intéressé à l'extraction des règles d'association entre les erreurs (avec l'algorithme apriori) .une étude comparatif des différents tests nous a permet d'en extraire ses avantage.

La réalisation du projet sous java nous a permet d'utiliser un algorithme de data mining existant dans l'environnement weka appelé via un programme java, qui retourne en faite le même résultat avec l'outil.

Conclusion générale

L'objectif de notre travail était de réaliser un suivi des apprenants dans les systèmes e-learning en utilisant les outils du data mining. Nous avons conduit nos tests l'analyse des erreurs sur deux paramètres : le nombre d'erreurs, les règles d'associations entre erreurs.

Au terme de notre travail nous avons pu :

- Enrichir la version Websiela3 par l'option *data mining*
- intégrant un nouveau mode d'évaluation qui est l'évaluation par utilisation de techniques de data mining :
  - Enrichir l'ontologie ONTO-TDM.
  - Proposer quelques types de suivi d'apprenants en utilisant la classification et les règles d'associations.
- Nous familiariser avec l'outil de data mining weka que nous avons utilisé pour la réalisation de notre travail.
- Améliorer la réalisation des algorithmes de data mining en utilisant l'environnement java.

Comme perspectives nous proposons :

- L'intégration des résultats des algorithmes de data mining appliquées sur les données d'apprentissage dans l'ontologie de domaine ONTO-TDM afin d'avoir une historisation de session d'apprentissage (avoir les résultats de chaque algorithme pour chaque apprenant dans toutes les sessions).
- Un outil de transformation du format bases de données vers le format ARFF

Références bibliographiques

[2]:M.Fieschi. Datamining .Simon Fraser University, Canada**. 2006**

[3]: O. ZAÏANE. Principles of Knowledge Discovery in Databases. University of Alberta.**1999**.

[4] : *D.* CHAMI .Une plate forme orientée agent pour le data mining. **2009-2010**.

[5]: S . RAHMOUN.Méthodes d'apprentissage pour améliorer la QoS d'une flotte de logiciels embarqués(Mémoire).Université Abou Bakr Belkaid– Tlemcen. **2010-2011**.

[6]: J.Baltie. Data Mining: ID3 and C4.5. **2002**.

[7]: B.LIAUDET. Cours de data mining modélisation supervisée, les arbres de décision, Option Ingénierie d'Affaires et de Projets – Finance. **2004**.

[8]: D.Moïse, F.Fenouillet. Le « e-Learning » est-il efficace ? Une analyse de la littérature anglo-saxonne(conference).**2004**.

[10]: M.Baker. The Roles of Models in Artificial Intelligence and Education Research: A Prospective View. International Journal of Artificial Intelligence in Education. **2000**.

[11:M. Park.Web Mining for Distance Education, In: IEEE International Conference on Management of Innovation and Technology. Department of Computer Science University of Saskatchewan 57 Campus Drive, CANADA.**2002.**

[12]: L.Gaudioso, E.Mining. Student Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces. In: Workshop in Artificial Intelligence in Computer Supported Collaborative Learning in conjuntion with 16th European Conference on Artificial Intelligence, ECAI. Valencia, Spain.**2003**.

[13]: G.Hwang. Development of an Intelligent Management System for Monitoring Educational Web Servers. In: 10th Pacific Asia Conference on Information Systems, PACIS.**2004**.

[14]: D. Traynor, J.P.Gibson. Synthesis and Analysis of Automatic Assessment Methods in CS1. In: The 36th SIGCSE Technical Symposium on Computer Science .USA.**2005**.

[15]: K,Mizue .O,Toshio. Neural Network Navigation Support-Knowledge-Navigation in Hyperspace: The Sub-symbolic Approach. Journal of Educational Multimedia and Hypermedia.**2005**.

[16]:B,Minaei-Bidgoli .Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System. In: Cantu, P.Genetic and Evolutionary Computation Conference.**2003**.

[17]: V.Carchiolo.A,M Malgeri,G .Mangioni.Courses Personalization in an e- Learning Environment. In: The 3rd IEEE International Conference on Advanced Learning Technologies.**July 2003**.

[18]: C.Lance, J.Hankins.Student Progress Monitoring Tool Using Treeview. In: The 37th Technical Symposium on Computer Science Education.USA. **2006**.

[19]: F,Chang. New Courseware for Quantitative Measurement of Distance Learning Courses. Journal of Information Science and Engineering.**2003**.

[20]: k, Becker. Distance Education: a Web Usage Mining Case Study for the Evaluation of Learning Sites. In: The 3rd IEEE International Conference on Advanced Learning Technologies, ICALT'03. IEEE Computer Society.**2003**.

[21]: G,Hwang. A Computer-Assisted Approach to Diagnosing Student Learning Problems in Science Courses. Journal of Information Science and Engineering .**2003**.

[22]: P.Markellou, I.Mousourouli,S.Spiros.A.Tsakalidis. Using Semantic Web Mining Technologies for Personalized e-Learning Experiences. In: Uskov, V. (ed.): The Fourth IASTED Conference on Web-based Education**.2005**.

[23]: G.Carpenter. S.Grossberg. Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. Neural Networks 4.**1991**.

[24]: J.Tane, C.Schmitz., G.Stumme. Semantic Resource Management for the Web: An eLearning Application. In: Fieldman, S.: The 13th World Wide Web Conference , WWW. ACM Press. New York .**2004.**

[25]: A.Drigas,J.Vrettaros. An Intelligent Tool for Building e-Learning Contend-Material Using Natural Language in Digital Libraries. WSEAS Transactions on Information Science and Applications.**2004**.

[26]:C.Romero,S.Ventura.P,De Bra. Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware. User Modeling and User-Adapted Interaction.**2004**.

[27] : M.Feng,N.Heffernan,K. Koedinger. Looking for Sources of Error in Predicting Student's Knowledge. In: The Twentieth National Conference on Artificial Intelligence by the American Association for Artificial Intelligence, AAAI'05, Workshop on Educational Data Mining, Pittsburgh, Pennsylvania. **July 9-13-2005**.

[28]: C.Lakhmi, A. aymond. Tedman and Debra K. Tedman (Eds.). Evolution of Teaching and Learning Paradigms in Intelligent Environment.**2007.**

[29]: P.Van Rosmale, F.Brouns,C.Tattersall, H.Vogten, J.van Bruggen. Towards an Open Framework for Adaptive, Agent-Supported e-Learning. International Journal Continuing Engineering Education and Lifelong Learning.**2005**.

[30]: H. Charle. « L'évaluation démystifiée », ESF Editeur. **1977**.

[31] : G.rard .Scallon. « L'évaluation formative ». **2000**.

[32] : T.SERIR. « Notation d'apprenant et scénarisation du processus d'évaluation Cas : Enseignement de l'algorithmique à des débutants ».**2009**.

[33] : F.Bouarab.« modélisation basé ontologies pour l'apprentissage interactif-application à l'évaluation des connaissances de l'apprenant ».**2010**.

Webo graphie :

[1]: http://chirouble.univlyon2.fr/~ricco/cours

[9] : http://www.namahn.com/resources/documents/note-e-learning

[34] : www.info2.uqam.ca/…h/…/documents_weka/environnement_weka

[35] : http://www.cs.waikato.ac.nz/ml/weka/

[36] : http://www.lgi2p.ema.fr/~dray/EMN05/docs/lireWeka.pdf