MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE ABDELHAMID IBN BADIS DE MOSTAGANEM

Faculté des Sciences Exactes et d'Informatique
Département de Mathématiques et d'Informatique
Filière Informatique

MEMOIRE DE FIN D'ETUDES
Pour l'Obtention du Diplôme de Master en Informatique
Option : Ingénierie des Systèmes d'Information

Etude Comparative Des Principaux Outils Open Source De Datamining

Etudiants :
CHAOUCH SADEK
DEBBAB ABBOU BAKR

Encadrant(e) :
BENAMMEUR ABDELKADER

Année Universitaire 2014/2015

SOMMAIRE

« Listes des figures »

Introduction générale

Aujourd'hui, une énorme quantité de données et d'informations est disponible pour tout le monde. En fait, le développement des calculateurs et des capacités de stockage a conduit à la croissance exponentielle des quantités d'information collectées et qui peuvent être stockées dans de nombreux types de bases de données et des référentiels d'information, en plus d'être disponibles sur Internet ou sous forme imprimée. Avec une telle quantité de données, il existe un réel besoin pour des techniques puissantes permettant l'interprétation de ces données qui dépasse la capacité de l'humain lors de la compréhension et la prise de décision. Afin de révéler les meilleurs outils pour faire face à cette tâche qui aide à la prise de décision, le présent document mène une étude comparative entre certains outils et logiciels de découverte de connaissances et de fouille de données disponibles.

En fait, une multitude d'outils et de logiciels existe dans la littérature couvrant plus ou moins divers domaines d'application. C'est outils et logiciels diffèrent dans plusieurs points dont la catégorie des méthodes qu'ils présentent, qu'elles soient celles des mathématiques ou de l'apprentissage automatique, ou dans les phases de fouille de données qu'ils implémentent, tel que le prétraitement, la fouille ou la validation. Cependant, ceci peut constituer un inconvénient pour un utilisateur possédant un jeu de données et désirant mener un processus d'extraction de connaissances. En fait, si les méthodes de prétraitement et de fouille que cet utilisateur veut lancer sur son jeu de données se trouvent dans des outils différents, et même dans des versions différentes de ces même outils, il sera donc confronté au problème d'importation et d'exportation des données et des résultats intermédiaires et finaux entre ces outils, du fait que chacun d'eux propose son propre format pour les données et les résultats. Aussi, un problème de variation de performances se pose entre ces différents outils.

Cette problématique nous a poussés à mener une étude comparative entre les principaux outils et logiciels de fouille de données. Nous nous sommes limités au logiciels open source pour la disponibilité de leur code source, et pour ne pas se confronter au problème de licence, bien que les logiciels payants offre généralement de meilleurs produits. Le but après cette étude est d'implémenter une plateforme de fouille de données intégrant plusieurs outils et bibliothèques, et ainsi tirant avantage de tous ces outils.

Le présent mémoire est organisé comme suit : le premier chapitre introduit le domaine de la fouille de données en proposant quelques définitions et en présentant le processus d'extraction de connaissances avec ses tâches et méthodes. Dans le deuxième chapitre, nous exposons les résultats de notre étude, en commençant par limiter les critères selon lesquels nous nous sommes basés lors de la comparaison, et en présentant ensuite les différents outils à la lumière de ces critères. Le troisième chapitre traite de la conception de notre application et son implémentation. Enfin, nous terminons avec une conclusion et quelques perspectives.

«Liste des tableaux »

Chapitre I

Généralités sur la fouille de données et le processus d'extraction de connaissances

I. Introduction

L'homme a besoin d'aide dans sa capacité d'analyse et d'exploration de données qu'il manipule. Cette exigence a généré un besoin urgent d'outils automatisés qui peuvent aider à transformer de vastes quantités de données en informations et en connaissances utiles. Cette tâche, connue sous plusieurs noms, dont celui d'exploration de données ou de fouille de données (Data mining), correspond au processus de découverte de connaissances intéressantes dans de grandes quantités de données stockées dans des bases de données, d'entrepôts de données, ou d'autres référentiels d'information. L'exploration de données implique une intégration des techniques provenant de plusieurs disciplines telles que la technologie des bases de données et d'entreposage de données, les statistiques, l'apprentissage automatique, l'informatique haute performance, la reconnaissance des formes, les réseaux neuronaux, la visualisation de données, la recherche d'information, l'image et le traitement du signal ...etc.

Actuellement, de nombreux outils et logiciels de découverte de connaissances à partir de données sont disponibles pour tout le monde et à usage différent tels que Weka, RapidMiner, R Statistics, Orange, Tanagra ...etc. Ces outils fournissent un ensemble de méthodes et d'algorithmes qui aident à une meilleure utilisation des données et informations disponibles pour les utilisateurs; y compris les méthodes et algorithmes pour l'analyse des données, l'analyse des clusters comme les algorithmes génétiques, les plus proches voisins, l'analyse de régression, les arbres de décision, l'analyse prédictive, l'extraction de texte, ...etc.

II. Historique

L'exploration de données est apparue dans les années 1980 au MIT « Massachusetts Institute of Technology» à Cambridge. Elle n'a pas connu un grand intérêt, car les autres sciences n'étaient pas assez avancées. Les bases du Data mining n'ont vraiment été fondé qu'en 1990 grâce à la combinaison de plusieurs facteurs, à la fois technologiques, économiques et même sociopolitiques. Les volumes gigantesques de données constituent, dès lors, des mines d'informations stratégiques aussi bien pour les décideurs que pour les utilisateurs.

III. Définition

Le terme Data mining est souvent utilisé pour définir l'ensemble des outils permettant à l'utilisateur d'accéder aux données d'un établissement et de les analyser. Cependant, il est limité ici aux outils ayant pour objet de générer des informations riches à partir des données de l'entreprise, notamment des données historiques, en vue de découvrir des modèles implicites dans les données. De ce fait, il permet par exemple à un magasin de dégager des profils de clients et des différents achats et de prévoir ainsi les ventes futures. Il permet ainsi d'augmenter la valeur des données contenues dans les bases ou les entrepôts de données.

C'est un processus d'extraction d'informations utiles à partir de données pré-traitées. Il existe de nombreuses techniques de fouille de données proprement dites dont le choix dépend du type de connaissances souhaitées

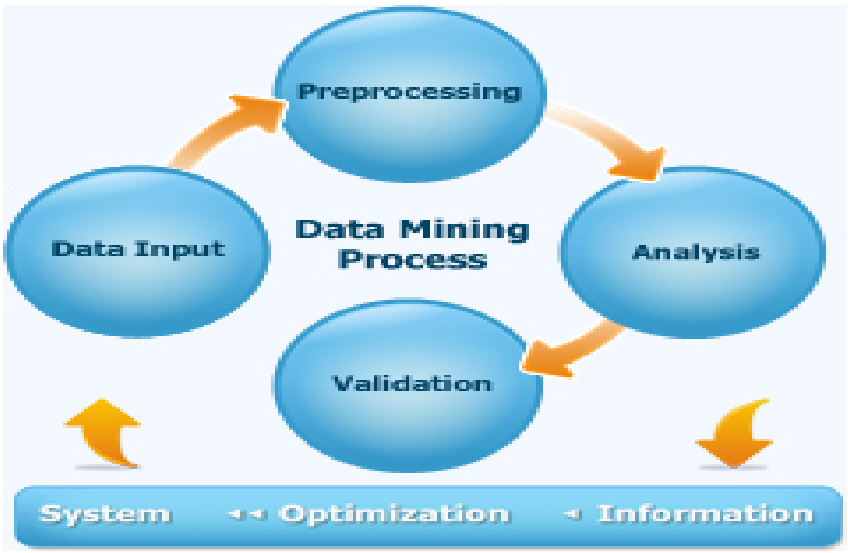


Figure I.1 : processus de data mining

IV. Domaines d'application:

Le Data mining regroupe un ensemble de théories et d'algorithmes qui restent ouverts aux différents domaines d'applications, dont nous citons :[34]

IV.1. Médecine et Pharmaceutique :

. Diagnostique assisté par ordinateur (CAD) à travers des systèmes experts.

. Explication ou prédiction de la réponse d'un patient à un traitement.

. Identification des thérapies à succès (combinaison de prescriptions).

. Étude des corrélations entre le dosage dans un traitement et l'apparition d'effets secondaires.

IV.2. Assurances et santé :

. Découverte d'associations des demandes de remboursements.

. Identification de clients potentiels à de nouvelles polices d'assurances.

. Détection d'association de comportements pour la découverte de clients à risque.

. Détection de comportement frauduleux.

IV.3. Banques et Finances :

. Détection d'usage frauduleux de cartes bancaires.

. Gestion du risque lié à l'attribution de prêts par le scoring.

. Découverte de relations cachées entre les indicateurs financiers.

. Détection de règles de comportement boursier par l'analyse des données du marché.

IV.4. Vente, Distribution et Marketing :

. Détection d'associations de comportements d'achat.

. Découverte de caractéristiques de clientèle.

. Prédiction de probabilité de réponse aux campagnes de mailing.

IV.5. Bourse :

. Analyse du cours de la bourse pour pouvoir passer des ordres automatiques de transactions boursières.

V. Processus d'ECD (Extraction de Connaissance à partir de Données)

Proposé par Oussama Fayyad pour répondre aux besoins des entreprises, le processus d'ECD est un processus itératif et interactif qui exige l'intervention de l'utilisateur dans chacune de ces étapes, et nécessite parfois de le relancer pour obtenir un meilleur résultat .[35]

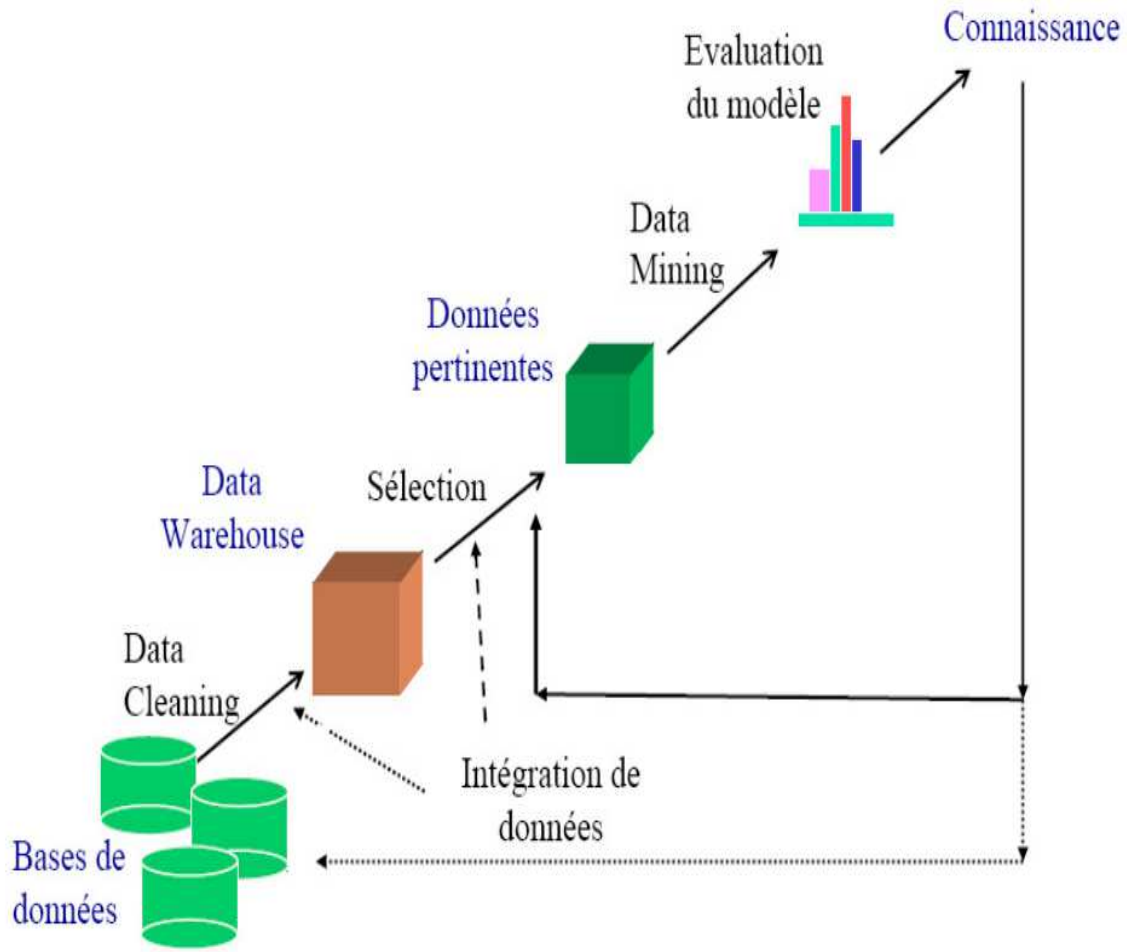


Figure I.2 : Le processus ECD

Ce processus est composé de neuf étapes :

Consolidation : cette étape consiste à regrouper les données, relatives au système cible de l'étude, en une unique source : l'entrepôt de données. Ces données sont le plus souvent hétérogènes. Par exemple, dans le cas de la surveillance des machines tournantes, elles peuvent être quantitatives comme les mesures physiques, les données de fiabilité ou qualitatives (textuelles) telles les informations des rapports d'intervention. Le lien entre ces données est le plus souvent temporel, mais on trouve aussi des applications se référant au nombre de cycles/tâches effectuées par le système cible.

Sélection et création d'un ensemble de données sur lequel va être appliqué le processus : dans cette étape nous devons sélectionner, dans l'entrepôt, les données qui seront retenues pour construire le modèle. Ainsi, dans le cadre de l'exemple cité dessus, on peut travailler sur les signaux vibratoires des machines asynchrones plutôt que sur leurs mesures de puissance ou de courant et ceci uniquement sur une période de temps avant une défaillance.

Prétraitement et nettoyage des données : cette étape inclut des opérations comme l'enlèvement du bruit et des valeurs aberrantes, si nécessaire, ou des décisions sur les stratégies qui vont être utilisées pour traiter les valeurs manquantes, etc.

Transformation de données : dans cette étape nous cherchons les méthodes correctes pour représenter les données. Ces méthodes incluent la réduction des dimensions et la transformation des attributs.

Choix de la meilleure tâche de Data mining : nous devons choisir quel type de Data Mining utilisé, en décidant le but du modèle (classification, régression, regroupement, ...etc.).

Application d'un algorithme de Data mining : nous devons choisir la méthode spécifique pour faire la recherche des motifs, en décidant quels modèles et paramétrés sont appropriés.

Évaluation : inclut l'évaluation et l'interprétation des motifs découverts. Cette étape donne la possibilité de retourner à une des étapes précédentes, mais aussi d'avoir une représentation visuelle des motifs, d'enlever les motifs redondants ou non-représentatifs et de les transformer dans des termes compréhensibles pour l'utilisateur.

Utilisation des connaissances découvertes : inclut l'incorporation de ces connaissances dans d'autres systèmes pour d'autres actions en mesurant l'effet de ces connaissances sur le système, vérifier et résoudre les conflits possibles avec les connaissances antérieures.

VI. Les tâches du Data Mining

Contrairement aux idées reçues, le Data mining n'est pas le remède miracle capable de résoudre toutes les difficultés ou besoins de l'entreprise. Cependant, une multitude de problèmes d'ordre intellectuel, économique ou commercial peuvent être regroupés, dans leur formalisation, dans l'une des tâches suivantes :

. Classification

. Estimation

. Prédiction

. Regroupement par similitudes

. Segmentation (ou clusterisation)

. Description

. Optimisation.

Afin de lever toute ambiguïté sur des termes qui peuvent paraître similaires, il semble raisonnable de les définir.

VI.1. La classification

Étant donné un ensemble prédéfini de classes d'objets, affecter un objet à une classe, selon une certaine mesure de proximité est le rôle de la classification. Les techniques de classification commencent par définir un plan d'expérience ou un ensemble de donnée d'apprentissage sur lequel on applique les méthodes de classification. Puis, pour mesurer leur pouvoir de classement correct, on applique les mêmes méthodes sur un jeu d'essai.

La classification se fait naturellement depuis déjà bien longtemps pour comprendre et communiquer notre vision du monde (par exemple les espèces animales, minérales ou végétales)[2]. Elle relie (range) les données dans des groupes prédéfinis (les catégories ou les classes). Souvent appelée apprentissage supervisé parce que les classes sont déterminées avant qu'on examine les données, la classification consiste à « examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini. »

Dans le cadre informatique, les éléments sont représentés par des enregistrements et le résultat de la classification viendra alimenter un champ supplémentaire. La classification permet de créer des classes d'individus (terme à prendre dans son acception statistique). Celles-ci sont discrètes : homme / femme, oui / non, rouge / vert / bleu, ...etc. Par exemple, un décideur veut classer ses employés par tranches de revenu, ou n'importe quelle autre caractéristique associée à cette personne, comme l'âge, le sexe et la profession.

Les techniques les plus appropriées à la classification sont :

. Les arbres de décision,

. Le raisonnement à base de cas.

. Les Réseaux de neurones.

. La machine à vecteurs de support (SVM). [2]

. L'analyse des liens.

VI.2. L'estimation

L'estimation est similaire à la classification, sauf que la variable cible est numérique plutôt que catégorique. Les modèles sont construits en utilisant des données, qui fournissent la

valeur de la variable cible, ainsi que les « prédicteurs ». Par exemple l'estimation de la pression artérielle d'un patient d'hôpital, basée sur son âge, son sexe, son indice de masse corporelle, et le taux de sodium. La relation entre la pression artérielle et le prédicteur variable de l'ensemble de formation nous donnerait un modèle d'estimation. Nous pouvons alors appliquer ce modèle à de nouveaux cas. Le résultat d'une estimation permet d'obtenir une variable continue. Celle-ci est obtenue par une ou plusieurs fonctions combinant les données en entrée. Le résultat d'une estimation permet de procéder aux classifications grâce à un barème. Par exemple, on peut estimer le revenu d'un ménage selon divers critères (type et nombre de véhicules, profession ou catégorie socioprofessionnelle, type d'habitation, ...etc.). Il sera ensuite possible de définir des tranches de revenus pour classifier les individus.

Un des intérêts de l'estimation est de pouvoir ordonner les résultats pour ne retenir si on le désire que les n meilleures valeurs. Cette technique sera souvent utilisée en marketing, combinée à d'autres, pour proposer des offres aux meilleurs clients potentiels. Enfin, il est facile de mesurer la position d'un élément dans sa classe si celui-ci a été estimé, ce qui peut être particulièrement important pour les cas limitrophes.

Les techniques les plus appropriées à l'estimation sont :

. Les réseaux de neurones.

. L'analyse statistique classique : régression linière simple.[2]

VI.3. La prédiction

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

Exemples de tâches de prédiction appliquée au marketing : « Prédire le prix d'un stock de trois mois dans le futur »

Les techniques les plus appropriées à la prédiction sont : Les réseaux de neurones [2].

VI.4. Le regroupement par similitudes

Le regroupement par similitudes consiste à grouper les éléments qui vont naturellement ensembles. La technique la plus appropriée au regroupement par similitudes est l'extraction des règles d'association [2].

VI.5. L'analyse des clusters (la segmentation)

Le Clustering désigne le regroupement des données, des observations ou des cas dans des classes d'objets similaires. Un cluster maximise la similarité des objets de du même cluster et minimise la similarité des objets de cluster différents. En effet, il n'y a pas de variable cible pour le clustering. La tâche de clustering ne cherche pas à classer, estimer, ou prédire la valeur

d'une variable cible, mais plutôt à segmenter l'ensemble des données en sous-groupes relativement homogènes à l'aide de mesures de distances.

L'analyse des clusters consiste à segmenter une population hétérogène en sous populations homogènes. Contrairement à la classification, les sous populations ne sont pas préétablis.

Les technique les plus appropriées à la clusterisation sont :

. Les réseaux de neurones

. Les machines à vecteurs support

VI.6. La description

Parfois, les chercheurs et les analystes essaient simplement de trouver des façons de décrire des tendances cachées dans les données. Les descriptions des modèles et des tendances servent à expliquer ou vérifier un fait. Par exemple : « ceux qui ont le plus de diplômes sont les plus susceptibles d'avoir un poste à responsabilité. »

C'est souvent l'une des premières tâches demandées à un outil de Data mining. On lui demande de décrire les données d'une base complexe. Cela engendre souvent une exploitation supplémentaire en vue de fournir des explications. La technique la plus appropriée à la description est aussi l'extraction des règles d'association [2].

VI.7. L'optimisation

Pour résoudre de nombreux problèmes, il est courant pour chaque solution potentielle d'y associer une fonction d'évaluation. Le but de l'optimisation est de maximiser ou minimiser cette fonction. Quelques spécialistes considèrent que ce type de problème ne relève pas du Data mining. Les techniques les plus appropriées à l'optimisation sont :

. Les réseau de neurones [2].

. Les algorithmes génétiques

VII. Le cercle vertueux

Enfin, on ne met pas en œuvre une technique de Data mining pour faire une simple exploration. Il faut l'inscrire dans un contexte plus global, appelé le cercle vertueux. Celui-ci est composé de quatre étapes :

. Identifier le domaine d'étude

. Préparer les données

. Agir sur la base de données

. Évaluer les actions

La première étape consiste à identifier le domaine d'étude. Il faut répondre aux questions : de quoi parlons-nous et que voulons-nous faire ? A ce stade, on définit un objectif général. Lorsque le domaine est délimité, il faut recenser les données relatives au domaine, puis les

regrouper pour en faciliter l'exploration. Nous parlons de regroupement logique, ce qui inclus le client / serveur, même si ce n'est pas recommandé. La troisième étape consiste à mettre en œuvre une ou plusieurs techniques de Data mining pour une première analyse.

Après évaluation et étude des résultats, des actions sont mises en œuvre. La dernière étape consistera à évaluer ces actions, et par-là même la performance du Data mining, voir le retour sur investissements. L'achèvement du premier cycle débouche souvent sur l'expression de nouveaux objectifs affinés, ce qui nous ramène à la première étape.[2]

VIII. Conclusion

Dans ce premier chapitre, nous avons introduit le domaine de l'extraction de connaissances à partir des données, communément connu sous le nom anglais de Data mining, ou son équivalent français de fouille de données. Nous avons essayé dans un premier temps de définir le domaine, pour ensuite citer quelques domaines d'application où ce champs a connu de grands succès où nous pouvons constater la grande émergence de ce champs d'application dans ces diverses domaines. Ensuite, nous avons présenté le processus complet d'ECD, où nous avons révélé qu'il s'agit de tout un processus et nous avons exposé les différentes étapes qu'il comporte. Enfin, nous avons survolé les différentes tâches que peut réaliser le processus d'extraction de connaissances et nous avons présenté chacune d'elles.

Dans le deuxième chapitre, nous allons entamer la présentation de notre travail, en exposant tout d'abord les critères de comparaison que nous avons retenu pour notre étude, et sur lesquels nous nous sommes basés lors de la comparaison des différents outils. Ensuite, nous citons de brèves aperçues des différents outils et logiciels open source que nous avons étudié. Nous nous sommes limités dans notre travail aux outils open source, car l'objectif final de l'étude est d'intégrer les principaux d'entre eux dans une plate-forme de Data mining.

Chapitre II

Étude des outils open source de fouille de données

I. Introduction

La variété des logiciels et algorithmes d'apprentissage et l'exploration de données donne un avantage aux utilisateurs dans les différents domaines et cela suit aux bouleversements des données en termes de quantité et du temps d'exécution ainsi qu'aux résultats désirer. Cela donne naissance au terme de plate-forme. Une plate-forme est en informatique une base de travail à partir de laquelle on peut écrire, lire, développer et utiliser un ensemble de logiciels. Elle peut être composée de matériel, de système d'exploitation et d'outils logiciels. Les plates-formes informatiques sont généralement conçues, développées, construites, mises en service et maintenues par des constructeurs informatiques, ou des prestataires de services.

Dans ce chapitre nous allons faire une étude comparative des différentes plate-formes open sources de Data mining existantes dans la littérature. Pour chaque plate-forme, nous citons les différents aspects qu'elle touche parmi un ensemble de critères que nous avons retenu.

II. Critères de comparaison des outils de Data mining

Nous commençons notre étude tout d'abord par la précision des critère selon lesquels nous avons comparé les outils de fouille de données pour mettre notre étude dans son vrai contexte[36].

II.1. Groupes d'utilisateurs

Business application : ce groupe utilise la fouille de données comme un outil pour résoudre les applications d'entreprise commercialement pertinents tels que la gestion de la relation client, la détection des fraudes, et ainsi de suite. Ce champ est principalement couvert par une variété d'outils commerciaux fournissant un soutien pour les bases de données avec de grands ensembles de données, et intégration profonde dans le flux de travail de l'entreprise.

Recherche appliquée : les utilisateurs sont principalement intéressés par les outils avec des méthodes bien éprouvées, une interface utilisateur graphique (GUI), et des interfaces avec les bases de données ou des formats de données relatives aux domaines. Cela permet d'appliquer la fouille de données à des problèmes de recherche, de la technologie et des sciences de la vie.

Développement algorithmique : développe de nouveaux algorithmes d'exploration de données, et les compare avec les méthodes existantes.[36]

Éducation : pour l'enseignement universitaire. Les outils de fouille de données devraient être très intuitive, avec une interface interactive confortable pour l'utilisateur, et peu coûteux.

II.2. Structures de données :

Certains jeux de données structurés sont caractérisés par la même dimension. Le format le plus important et ayant une dimension supérieure contient des séries chronologiques comme des éléments. Les tâches typiques prévoient des valeurs futures, en trouvant des modèles typiques dans une série temporelle ou en trouvant des séries similaires par clustering. L'analyse des séries chronologiques joue un rôle important dans de nombreuses applications

différentes, y compris la prévision des marchés boursiers, la prévision de la consommation d'énergie et d'autres marchés, et la supervision de la qualité dans la production.

Une tendance plus récente est l'application de méthodes de fouille de données sur les images et les vidéos. Le principal défi est la manipulation de très grands volumes de données, des Giga et Téra Octets, causés par la forte dimensionnalité des exemples. Les applications typiques sont des images microscopiques de la biologie et de la médecine, ou des captures caméras dans le contrôle de la qualité et de la robotique, de la biométrie et de la sécurité.

Un autre format de structure de données aux dimensions de l'image comprend des graphiques qui peuvent être représentés en tant que matrices de contiguïté, décrivant le lien entre différents nœuds d'un graphe.

II.3. Interaction et Visualisation

Il existe trois principaux types d'interaction entre un utilisateur et un outil d'exploration de données :

• Interface textuelle pure en utilisant un langage de programmation : difficile à gérer, mais facilement automatisé.

• Interface graphique avec structure : un menu facile à manipuler, mais pas si facilement automatisé.

• Interface utilisateur graphique : l'utilisateur sélectionne des blocs fonctionnels ou des algorithmes à partir d'une palette de choix, définit les paramètres, les place dans un espace de travail, et les relie pour créer des flux de travail complèts : un bon compromis, mais difficile à gérer pour les grands flux de travail.[36]

Presque tous les outils fournissent des techniques de visualisation puissants pour présenter leurs données explorées et les résultats, en particulier les outils pour « Business application » et « la recherche appliquée », qui sont en mesure de générer des rapports complets contenant les résultats les plus importants dans une forme lisible pour les utilisateurs. Les méthodes interactives peuvent supporter une analyse exploratoire de données. Par exemple, la méthode « brossage » qui permet à l'utilisateur de sélectionner des points de données spécifiques à un chiffre ou sous-ensembles de données (par exemple, les nœuds d'un arbre de décision) et mettre en évidence ces points de données dans d'autres parcelles (plot).

II.4. Tâches et méthodes

Les méthodes d'exploration de données ne sont pas disponibles dans tous les outils. La liste suivante contient la fréquence d'apparition de certaines méthodes spécifiques :

Fréquentes : la classification en utilisant des probabilité estimée, fonctions de densité, analyse de corrélation statistique, sélection de fonction, et la pertinence des tests.

Dans de nombreux outils : arbres de décision, regroupement, régression, nettoyage et filtrage des données, extraction de caractéristiques, analyse en composantes principales,

analyse factorielle, évaluation de fonctionnalité avancée et de sélection, calcul de similitudes, réseaux de neurones artificiels, validation croisé, et les tests de pertinence statistique.

Dans certains outils : classification floue, analyse en composantes indépendantes, amorçage, mesures de complexité, le modèle de fusion, machines à vecteurs supports, méthodes k-plus proches voisins, réseaux bayésiens.

Rare: Random forests (contenue dans Random forests, WEKA, et toutes ses dérivées), l'apprentissage des systèmes flous (contenues dans KnowledgeMiner, See5 et Gait-CAD), Rough sets(dans ROSETTA, et Nurseslabs), et l'optimisation du modèle par algorithmes évolutionnaires (en QUILLE, Adam, et D2K).

II.5. Import / Export des données

Les données sont normalement générées et hébergée par différentes sources telles que les bases de données. Dans les applications d'affaires, les interfaces avec les bases de données comme Oracle ou de toute base de données supportant le langage de requêtes SQL sont les moyens les plus courants de l'importation de données.

Afin d'importer et d'exporter des modèles mis au point en tant que composants dans d'autres processus et systèmes, l'PMML32 standard basé sur XML a été développée par le Groupe Data Mining et est soutenu par de nombreuses entreprises comme IBM et SAS. Une autre initiative est la norme Object Linking and Embedding Database (OLEDB ou OLE-DB) pour l'extraction de données, est une API conçue par Microsoft pour accéder à différents types de données stockées d'une manière uniforme. OLEDB est un ensemble d'interfaces mises en œuvre en utilisant le Component Object Model (COM). Pour l'échange de données entre les différents outils. Une autre initiative traite de « Java Specification Requests for data mining », définit une API Java extensible pour les systèmes d'exploration de données. Le consortium comprend de nombreuses sociétés liées, comme Oracle, SAS, SPSS (aujourd'hui IBM), SAP

II.6. Plate-forme

Les outils d'exploration de données peuvent être subdivisés en différents solutions autonomes. La solutions Client / Serveur est dominante, en particulier dans les produits conçus pour les utilisateurs professionnels qui sont disponibles pour les différentes plates-formes, y compris Windows, Mac OS, Linux, ou superordinateurs spéciaux. Il y a un nombre croissant de systèmes basé sur Java pour les utilisateurs dans la recherche et la recherche appliquée. Un nombre croissant d'interfaces Web fournissant l'extraction de données en mode SaaS (logiciel en tant que service, avec des outils comme données appliqué) et un soutien plus fort des solutions de Data mining Client / Serveur basée sur des grilles (outil Adam).

Ces deux tendances peuvent avoir le risque potentiel de blesser les politiques de confidentialité parce que la protection des données est difficile et de nombreuses entreprises sont très prudents avec les données sensibles.

II.7. Licences

Le type le plus populaire des licences open-source est la Licence Publique Générale GNU de la Free Software Foundation (GNU-GPL). Il permet la libre redistribution, l'intégration dans d'autres paquets, et la modification du logiciel tant que tous les utilisateurs ultérieurs reçoivent le même niveau de liberté « copie gauche ». Cette restriction garantit que tous les logiciels contenant des composants GNU-GPL doivent être distribués sous licence GNU-GPL[36].

Maintenant nous présentons notre étude des outils de fouille de données.

III. La plate-forme WEKA

WEKA (Waikato Environment for Knowledge Analysis) est un environnement pour l'analyse de connaissances développé à l'université de Waikato, Nouvelle-Zélande. C'est un logiciel libre disponible sous la licence publique générale GNU (GPL) qui permet de réaliser des analyses en Data mining. WEKA est écrite en Java et fonctionne sur quasiment tous les systèmes d'exploitation actuels. Il offre une panoplie d'algorithmes d'apprentissage [8] et permet d'appliquer toute la chaîne du processus d'extraction de connaissance à partir de données (prétraitement, classification supervisée, règles d'associations, visualisation, etc...). La disponibilité du code source permet d'implémenter et de tester des nouveaux algorithmes tout en s'appuyant sur une plate-forme éprouvée et un code objet. Le logiciel comprend plusieurs outils dont un API (Aplication Programming Interface) qui permet d'utiliser les outils WEKA dans d'autres programmes (Figure II.1).

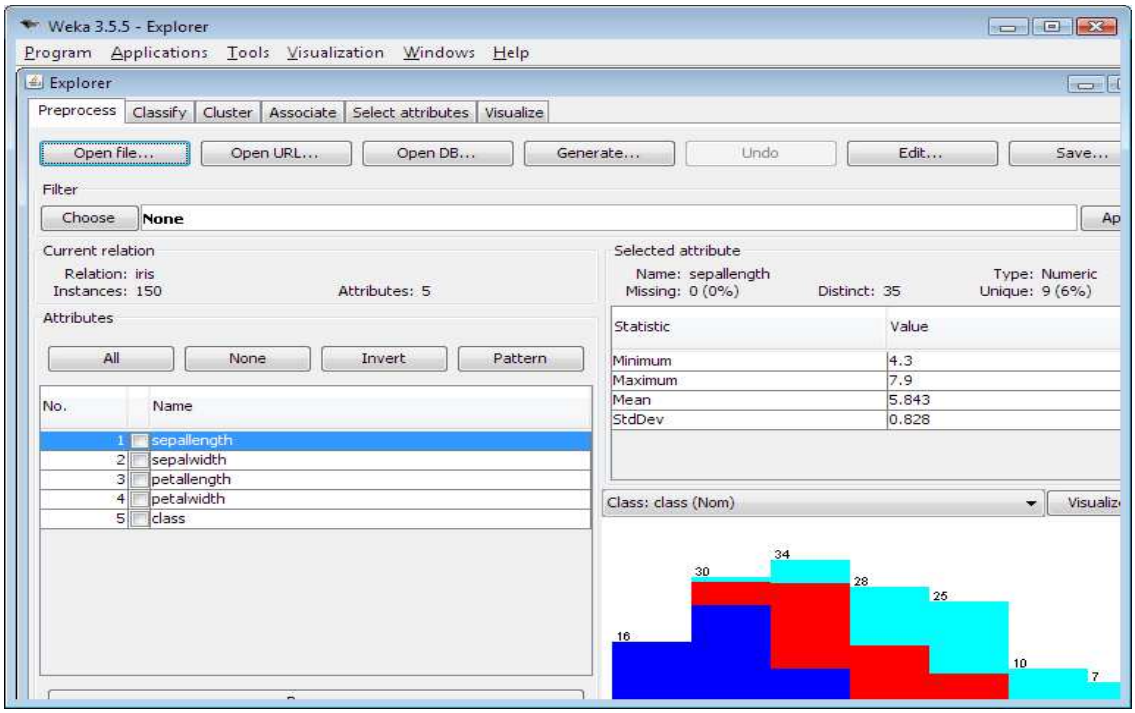


Figure II.1 : La plate-forme WEKA

III.1. Caractéristiques générales

WEKA supporte plusieurs outils d'exploration de données standards, et en particulier, des préprocesseurs de données, des classificateurs, des analyseurs de régression, des outils de visualisation, et des outils d'analyse discriminante [10]. Le format des données d'entrée par défaut de WEKA est ARFF (Attribute Relation File Format). D'autres formats peuvent être importés comme CSV, Binaire, BDD SQL (avec JDBC) à partir d'une URL, etc. WEKA contient plus de 70 algorithmes de classification / régression supervisés (Tableau II.1), plus de 15 évaluateurs d'attributs et plus de 10 algorithmes de recherche pour la sélection d'attribut, des algorithmes de recherche de règles d'association et plusieurs interfaces graphiques GUI.

WEKA s'ouvre avec quatre options (Explorer, expérimenter, KnowledgeFlow et CLI simple). Principalement, Explorer et Expérimentateur sont utilisés pour l'extraction de données. A titre de comparaison de multiples algorithmes, l'Expérimentateur est utilisé, mais pour des résultats spécifiques à l'extraction de données, l'Explorateur est utilisé.

III.2. Les différents onglets de WEKA

Classify : WEKA donne accès à plusieurs méthodes supervisées comme les réseaux bayésiens, les arbres de décision, les règles de décision,...etc. Pour chaque méthode on trouve beaucoup d'algorithmes, ce qui permet de toucher à tous les types de données.

Cluster : WEKA propose plusieurs algorithmes de clustering qui traitent les différents types de données comme : K-Means, l'algorithme hiérarchique, OPTICS, des algorithmes basés sur la densité, COBWEB, DBSCAN, EM,...etc ;

Associate : les algorithmes de génération des règles d'association disponibles sous WEKA sont : Apriori, FP-GROWTH, Tertus, Filtred-Associator,...etc, qui traitent les différents types de données.

Select attributs : présente plusieurs algorithmes qui aident à optimiser la sélection des attributs pour réduire le jeu de données.

Visualise : Affiche les objets en représentant leurs coordonnées par rapport à chaque deux variables.

Le tableau ci-dessous présente quelque méthodes d'apprentissage avec ces algorithmes :

Méthode	Algorithmes
Les réseaux de Bayésiens	Naive Bayes Simple, NaiveBayes
Les arbres de décision	Id3, J48, DecisionStump
Les réseaux de neurones	RBF Network, Multilayer Perceptron, Voted Perceptron
Régression avec méthode de moins carrées	Linear Regression
Les séparateurs à vaste marge	SMO

K plus proches voisins	IB
Les méthodes ensemblistes	Bagging , AdaBoostM1

Tableau II.1 : Quelques exemples des algorithmes supervisés utilisés dans WEKA

III.3. Avantage :

. Une interface très complète : WEKA présente quatre modes et implémente beaucoup d'algorithmes pour chaque tâche.

. Possibilité de traiter les données d'une base de données.

. Traitement des données manquantes.

. Une bonne gestion des erreurs (les contrôles logiques) .

. Beaucoup de filtres pour faire des transformations sur les données.

. Possibilité de faire des comparaisons entre les différentes méthodes.

. Multi-plate-formes (Windows, Lunix, MAC OS).

. Extensible.

III.4. Inconvénients

. Nécessite une lecture attentive de la documentation, car la manipulation est difficile.

. Absence de tests statistiques.

. Une limitation technologique (JAVA) sur la taille de la base.

. Une limitation technologique (JAVA) sur la rapidité.

IV. La plate-forme TANAGRA

TANAGRA est un logiciel open source de fouille de données qui dérive de SIPINA [11]. Cet environnement est une plate-forme libre d'expérimentation destinée à l'enseignement et à la recherche développée par Ricco RAKOTOMALALA à l'université de Lumière de Lyon [12]. Elle présente une interface graphique conviviale, et permet l'enchaînement de plusieurs traitements visualisés par un graphe. Les données y sont introduites par un simple fichier texte, où les variables sont séparées par des tabulations. La sortie des différentes opérations quant à elle est réalisée en utilisant HTML [13].

Développé avec le langage C++, ses différents composants peuvent être assemblés à la manière d'une structure en arbre afin d'implémenter l'ordre de réalisation des tâches de fouille de données (Figure II.2).

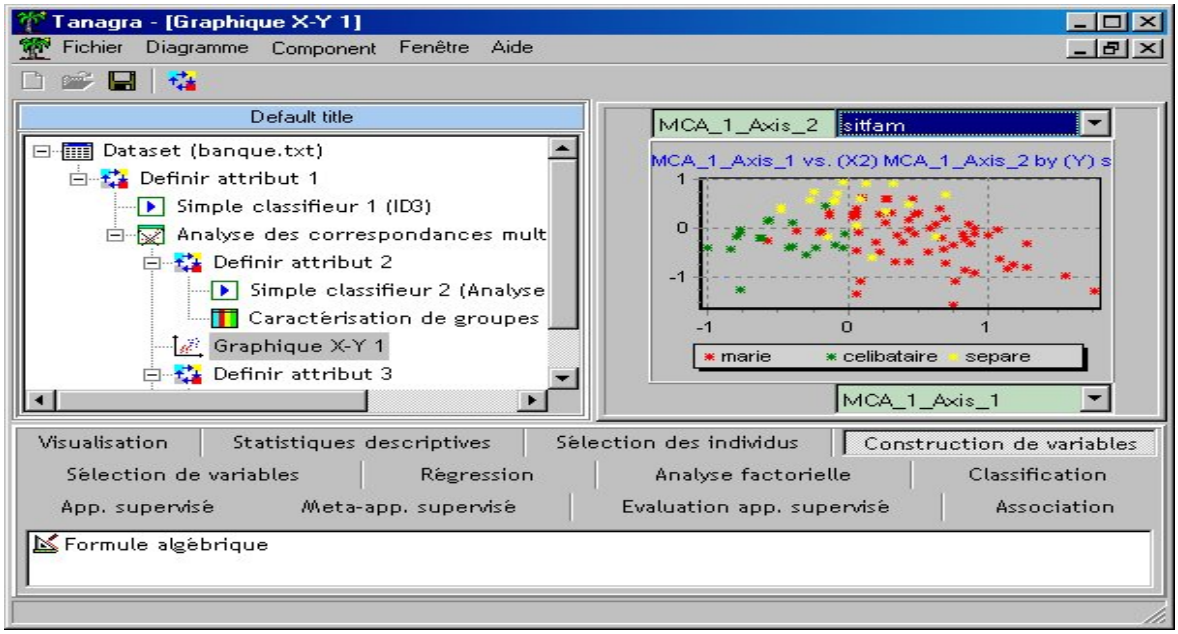


Figure II.2 : La plateforme TANAGRA

Par ailleurs, TANAGRA regroupe un ensemble important de méthodes d'apprentissage supervisé telles que :

. Binary logistic regression : Régression logistique binaire, méthode du maximum de vraisemblance (14) (15)
. k-Nearest Neighbor (k-NN) : Méthode des k-plus proches voisins, s'appuyant sur des distances pouvant appréhender tous types de variables (Heterogenous Value Difference Metric) (16) (17)
. Multi-layer perceptron : Perceptron multicouches, algorithme du rétro-propagation du gradient. (18) (19) (20)
. Prototype-NN : Des noyaux sont préalablement construits. A chacun est affectée une modalité d'appartenance de la variable à prédire, généralement celle qui est majoritaire. On affecte alors à l'individu à classer la modalité du noyau dont le centre de gravité lui est le plus proche.
. ID3 : Algorithme de base des arbres de décision (21).
. Linear Discriminant Analysis : Analyse discriminante linéaire prédictive (modèle bayesien) (22).
. Naive Bayes : Modèle bayesien naïf, modèle d'indépendance conditionnelle (23).
. Radial basis function : Réseau de neurones RBF (Radial basis function). Il s'agit d'un perceptron simple où la couche d'entrée est constituée de noyaux. L'implémentation choisie est off-line, c-à-d les noyaux sont construits ex-ante, avant le processus d'apprentissage du réseau, par un clustering par exemple (24).
. Multiple linear regression : Régression multiple linéaire méthode des moindres carrés (25).

L'objectif principal du projet TANAGRA est d'offrir aux chercheurs et aux étudiants une plate-forme de Data mining facile d'accès, respectant les standards des logiciels du domaine, notamment en matière d'interface et de mode de fonctionnement, et permettant de mener des

études sur des données réelles et/ou synthétiques (7). Le second objectif est de proposer aux chercheurs une architecture leur permettant d'implémenter aisément les techniques qu'ils veulent étudier et de comparer les performances des algorithmes. TANAGRA se comporte plus comme une plate-forme d'expérimentation qui leur permettrait d'aller à l'essentiel en leur épargnant toute la partie ingrate de la programmation de ce type d'outil : la gestion des données (7). Le dernier objectif, vise à diffuser une méthodologie possible d'élaboration de ce type de logiciel. L'accès au code leur permettra de voir comment se construit ce type de logiciel, quels sont les écueils à éviter, quelles sont les principales étapes d'un tel projet, et quels sont les outils et les bibliothèques qu'il faut préparer pour le mener à bien. En ce sens, TANAGRA est plus un outil d'apprentissage des techniques de programmation.(7)

IV.1. Avantages :

. L'application des algorithmes présentés donne de bons résultats ;

. Disponibilité de plusieurs transformations possibles d'un type d'attribut vers un autre ;

. Riche en fonctionnalités statistiques et des composantes de l'analyse des données ;

IV.2. Inconvénients :

. Nécessite un émulateur pour lui permettre de fonctionner sur d'autre système d'exploitation

. Pas de traitement des bases de données ;

. Pas de traitement des valeurs manquantes ;

. Manque des contrôles logiques (exemple : deux attributs avec le même nom) ;

. La gestion de la mémoire n'est pas très performante (règles d'association, calcul matricielle,...etc.) ; [6]

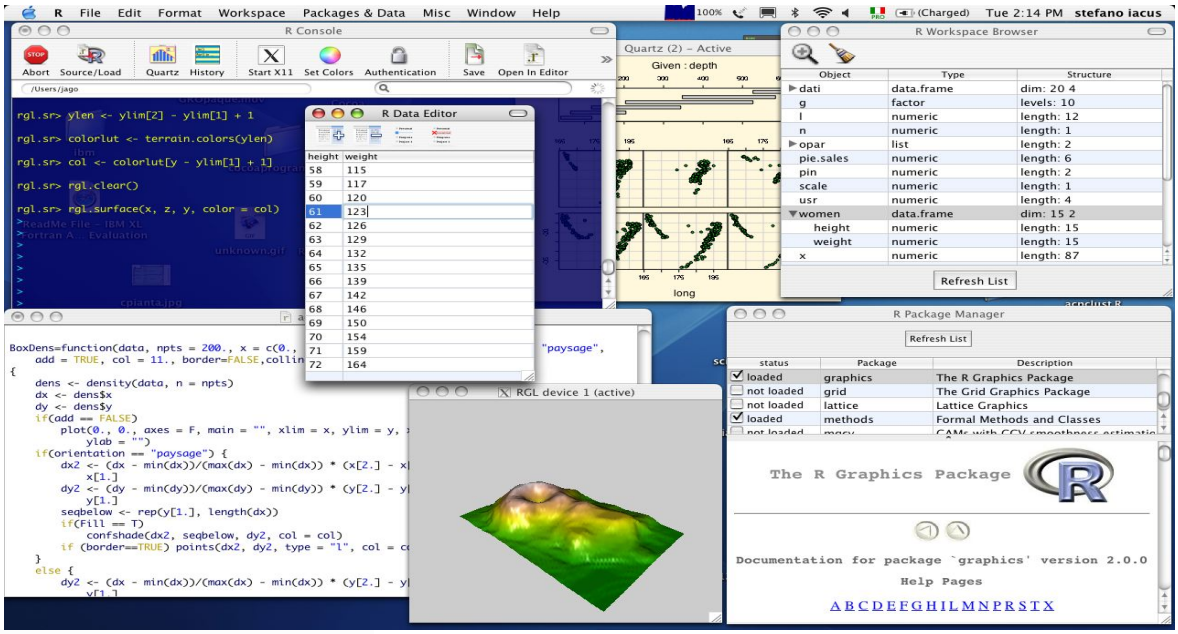
. La composante Data visualization offre la possibilité d'afficher les données sous forme de tableau ou en nuages de points seulement (6)

. Non extensible(6)

V. La plate-forme R

R est un logiciel libre, gratuit et multiplate-forme (Linux, Mac OS X et Windows) de traitement des données, d'analyse statistiques et de Data mining mettant en œuvre le langage de programmation S (Scheme) (26) (27). Il est codé dans les langages C, C++, Fortran et Java (28), piloté en ligne de commande, et distribué par GNU (26) (27). R est également un langage de programmation basé sur le calcul matriciel. La manipulation d'objets de type vecteur, liste, matrice permet une flexibilité de programmation d'algorithmes plus ou moins évolués répondant aux attentes de chacun.

Composé d'un ensemble des paquet ou packages, la logiciel R permet à ses utilisateurs de créer de nouveaux paquets ou installer, désinstaller, charger, décharger et mettre à jour des paquets existant (29). Ce système permet d'augmenter considérablement la puissance de R (29) en rendant les possibilités d'utilisation immenses dans des domaines d'études très différents (écologie, psychologie, économie...) et faisant intervenir des techniques très diverses (modélisation linéaire et non linéaire, statistique spatiale, classification, tests statistiques...). Le partage grandissant de nouveaux paquets rend ce logiciel très dynamique et qui s'enrichit jour après jour (26). Parmi les paquets, il en existe quelques uns permettant d'interfacer avec d'autres outils tels que PostgreSQL et MySQL pour les bases de données, le logiciel libre GRASS pour les SIG, RExcel pour Excel ou encore Latex et OpenDocument



pour l'exportation de résultats.

Figure II.3 : La plate-forme R

V.1. Caractéristiques Générales

R est un outils open source basé sur les packages, piloté en ligne de commande. Il y a centaines paquets supplémentaires librement disponibles, qui offrent toutes sortes de techniques de Data mining, d'apprentissage automatique et de statistiques.

. Il permet aux statisticiens de faire des analyses très complexes et compliquées sans connaître ou être fondés dans le système informatique

. Il dispose d'un grand nombre d'utilisateurs, en particulier dans les domaines de la bio-informatique et les sciences sociales.

. Adapté pour le calcul statistique.[3]

Parmi les packages des algorithmes d'apprentissage sous R, on cite : [29]

. Arbres de décision : dans le package « rpart »

. Réseaux Bayésiens : dans le package « bnlearn »

. Réseaux de neurones : dans le package « mlbench »

. Régression avec la méthode des moindres carrées : dans les packages « stats », « tseries », « systemfit », « Rcmdr » et « FactoMineR »

. Séparateurs à vaste marge : dans le package « e1071 » basée également sur la bibliothèque LIBSVM [30].

. k-plus-proches-voisins : dans le package « KernSmooth » [31]

V.2. Avantages

. R est un langage de tableau puissant dans la tradition de Mathematica et MATLAB
. Riche en bibliothèque statistique.[3]
. Possibilité de faire un programme d'apprentissage automatique en 40 lignes de code seulement
. Fort en programmation numérique
. Riche en visualisation graphique.
. Plus facile à combiner avec d'autres calculs statistiques.[3]

V.3. Inconvénients

. Moins spécialisée dans le domaine d'exploration de données.[3]

. Il y a une courbe d'apprentissage abrupte pour les gens qui ne sont pas familiarisé avec les langages de tableau.

VI. La plate-forme ORANGE

Orange est un outil open source dédié à la fouille de données et à l'apprentissage automatique. Il propose à la fois une bibliothèque de fonctions qui permet le développement de scripts et une interface utilisateur très intuitive qui permet de construire graphiquement l'enchaînement des traitements souhaités. Le logiciel Orange est actuellement développé principalement par le Laboratoire de Bioinformatique de la Faculté d'Informatique et des Sciences de l'Information à l'Université de Lubiana en Slovénie. Il a été créé en 1997 par Janez Demšar et Blaž Zupan, aujourd'hui membres du Laboratoire de Bioinformatique. Pour l'anecdote, mais elle donne également une idée des performances du logiciel, Orange est arrivé premier sur 126 participants lors d'une compétition internationale dans le domaine de la fouille de données organisée en 2012 .

Orange est distribué sous licence GNU General Public License version 3 et est réalisé en langages C++ et Python avec l'utilisation de la libraire Qt pour la partie interface graphique utilisateur. [4]

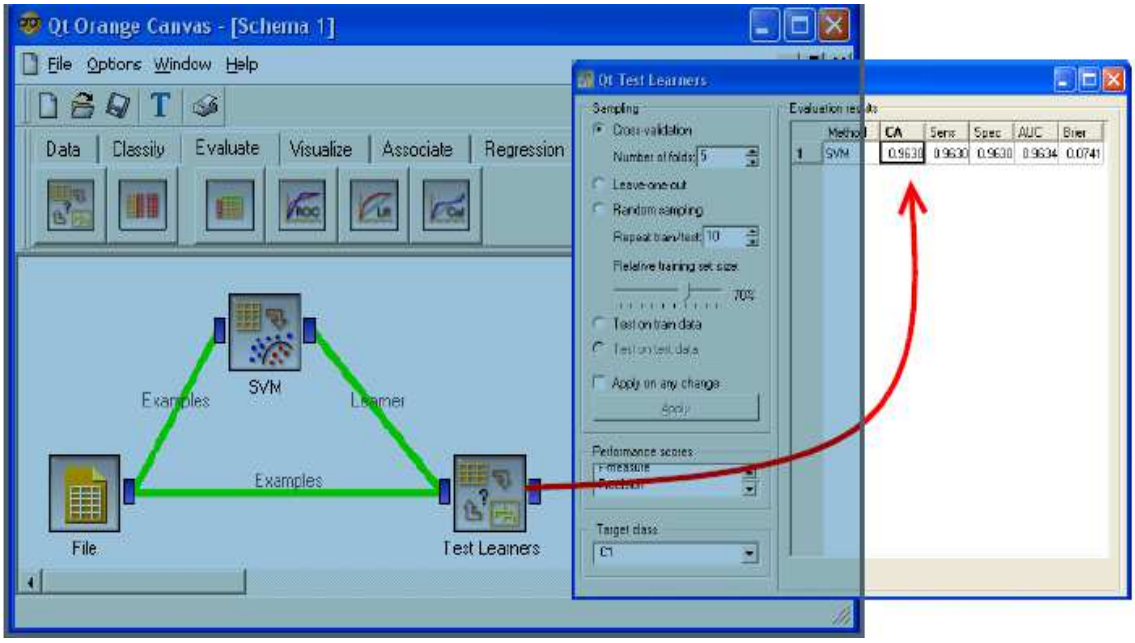


Figure II.4 : La plate-forme Orange

Le widget File de la composante Data permet de définir les donnée sous divers formats (csv, tab, txt, data, arff, svm,...etc). Orange peut se connecter avec un SGBD (MySQL,PostgreSQL SQLite) via une variable de chemin de classe du pilote JDBC[6].

VI.1. Les différentes composantes d'Orange

. La composante Data offre également d'autres opérations sur les données comme la sélection des attributs, l'échantillonnage, la discrétisation, la numérisation,...etc ;

. La composante Visualize propose plusieurs méthodes de visualisation des données introduites (l'histogramme, les nuages de points, la projection linéaire, les coordonnées en parallèles et les statistiques des attributs) ;

. Les algorithmes de clustering disponibles sont : K-Means Clustering pour les données mixtes, l'algorithme hiérarchique qui travaille avec une matrice de distances et l'algorithme SOM pour les données numériques ;

. Pour les algorithmes de classification, on trouve : un arbre de classification, K-NN, C4.5, les réseaux bayésiens,...etc. qui traitent des données mixtes ;

. Un seul algorithme de génération des itemsets fréquents (Apriori).[6]

VI.2. Avantages :

. Interface agréable qui permet de manipuler les données graphiquement ;

. Possibilité d'effectuer plusieurs opérations sur les données (Union, jointure, échantillonnage,...etc.) ;

. Possibilité de réaliser une combinaison de traitements et de faire plusieurs traitements en parallèle ;

. Visualisation graphique riche et interactive ;

. Disponibilité sous Windows et MAC OS ;

. Extensibilité.[6]

VI.3. Inconvénients :

. Pas de traitement des bases de données ;

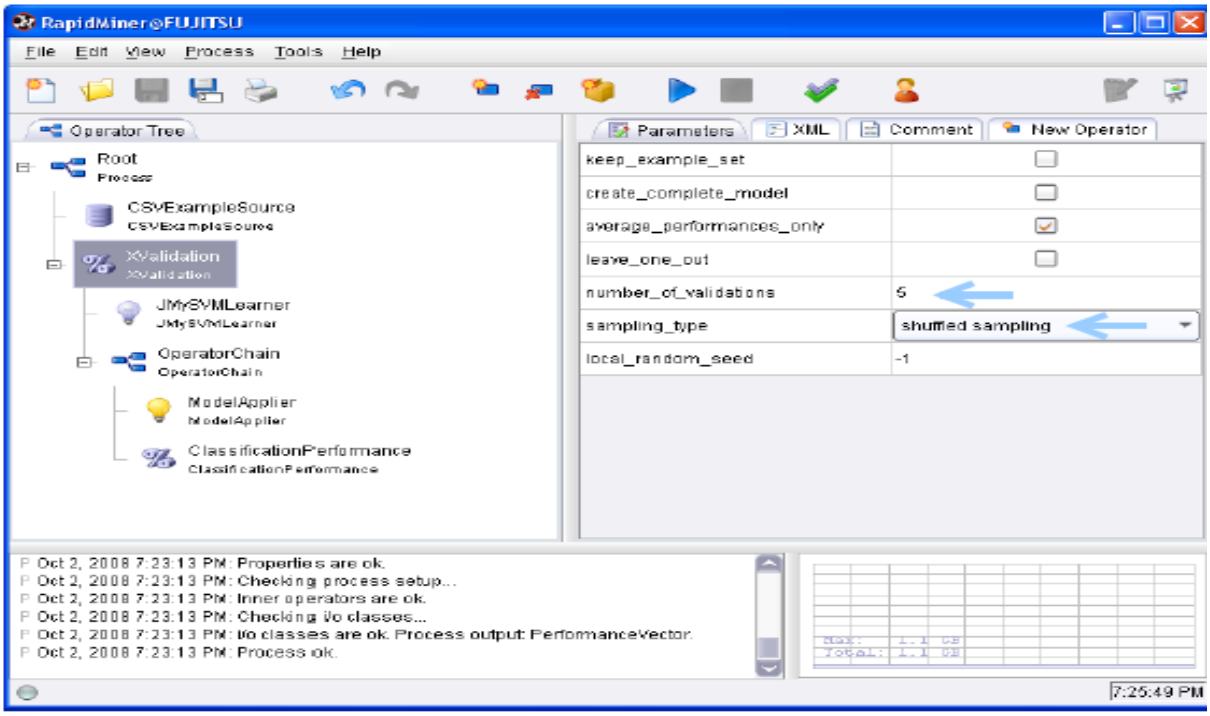
. Manque des contrôles logiques (exemple : deux attributs avec le même nom) ;

. L'absence de tests statistiques.

. Orange est riche dans la visualisation et la manipulation des données, mais il est moins adapté pour les opérations statistiques.[6]

VII. La plate-forme RapidMiner

RapidMiner est un logiciel open source et gratuit dédié au Data mining. Il contient de nombreux outils pour traiter des données : lecture de différents formats d'entrée, préparation et nettoyage des données, statistiques, tous les algorithmes de data mining, évaluation des



performances et visualisations diverses.[2]

Figure II.5 : La plate-forme RapidMiner

VII.1. Caractéristiques générales

RapidMiner supporte les formats de fichier suivants : Aml, arff, att, bib, CLM, cms, cri, csv, dat, ioc, log, matte, mode ObF, a bar, one pair, res, sim, Thr, WGT, WLS, xrff .[1]
On peut le connecter avec Oracle, Microsoft SQL Server, PostgreSQL, ou bases de données MySQL. Si le système de gestion de base de données n'est pas supporté, il peut être corrigé en ajoutant la variable de chemin de classe du pilote JDBC.
. L'Ensemble de données est exprimé en XML. [5]
. RapidMiner comprend WEKA.
. La Visualisation en 3D est très utiles pour l'utilisateur .
. Près de 22 formats de fichiers pris en charge.
RapidMiner est un outil d'exploration de données qui a une simulation ARENA parce que chaque processus est décrit d'une manière similaire. RapidMiner peut utiliser tous les algorithmes de WEKA ainsi que ses propres algorithmes [5]. Il représente une nouvelle approche de conception même des problèmes très complexes en utilisant un concept d'opérateurs modulaires qui permet la conception de chaînes d'opérateurs imbriqués complexes pour grand nombre de problèmes d'apprentissage [3]. Rapidminer utilise le langage XML pour décrire le processus de découverte de connaissances de la modélisation des arbres de l'opérateur [3]. Il a des opérateurs flexibles pour les formats d'entrée de données et de fichiers de sortie.

VII.2. Avantages

. Plus 1,500 méthodes d'intégration de données, transformation de données, l'analyse et la modélisation ainsi que la visualisation.
. Aucune autre solution sur le marché n'offre plusieurs procédures et donc plus de possibilités de définir les processus d'analyse optimales[3]
. RapidMiner offre de nombreuses procédures, en particulier dans la zone de sélection d'attribut et de détection des valeurs aberrantes, qui aucune autre solution offre.[3]

VII.3. Inconvénients

. Nécessite la capacité de manipuler les instructions SQL et des fichiers.[3]
. Résultat uniquement basé sur la matrice de confusion.

Plateforme	Algorithmes	Open source, Libre	Site officiel
TANAGRA	Binary logistic	+,+	http://www.eric.univ-

	regression, K Plus proches voisins, ID3 ...		lyon2.fr/~ricco/tanagra.fr/
RapidMiner	Naive Bayes, la décision Stump, Hoeffding Arbre	+,+	https://rapidminer.com
WEKA	Id3, J48,RBFNetwork, MultilayerPerceptron, VotedPerceptron	+,+	http://www.cs.waikato.ac.nz/ml/weka/
R	package « rpart », package « bnlearn», package « mlbench »	+,+	http://www.r-project.org/
Orange	K-Means Clustering -l'algorithme SOM ,- K-NN,C4.5 -les réseaux bayésiens -Apriori	+,+	http://orange.biolab.si/

Tableau II.2 : Tableau récapitulatif des différentes plates-formes existantes (1)

Plate-formes	**Type de Plate-forme**			**Type d'apprentissage**		**Langage**
	exécution	**développement**	**bibliothèque**	**ISSH**	**ISIE**	
TANAGRA	**-**	**-**	**+**	**+**	-	**C++**
RapidMiner	**+**	**-**	**+**	**+**	**-**	Java
WEKA	**+**	**+**	**+**	**+**	**-**	
R	**+**	**+**		**+**	**+**	C++,
Orange	**+**	**-**	**+**	**+**	**-**	C++,python ,Qt

Tableau II.3 : Tableau récapitulatif des différentes plates-formes existantes (2)

Avec :

. ISSH :Inductif, Supervisé, Statique, Hors ligne

. ISIE : Inductif, Supervisé, Incrémental, En ligne

Nous complétons notre étude par quelques sondages effectuées sur l'utilisation de ces outils.

Selon le site « Kdnuggets » (année 2012), la première fois que le nombre d'utilisateurs de logiciels libres / open source a dépassé le nombre d'utilisateurs de logiciels commerciaux. Parmi les électeurs, 28% ont utilisé un logiciel commercial mais pas un logiciel libre, 30% logiciels libres mais pas commercial, et 41% les deux à la fois.

Quelle sont les outils d'analyse et d'exploration de données que vous avez utilisé au cours des 12 derniers mois pour un projet réel (pas seulement d'évaluation) [798 électeurs]	
R (245)	30.7% 23.3%
Rapid-I RapidMiner (213)	26.7% 27.7%
Weka (118)	14.8% 11.8%
Orange (42)	5.3% 1.3%

VIII. Conclusion

Dans ce deuxième chapitre, nous avons exposé diverses caractéristiques des principaux outils open source de fouille de données. A travers cette étude, nous avons constaté la diversité des algorithmes d'apprentissage automatique et des méthodes statistiques disponibles dans chaque outil. Le choix d'un algorithme ou d'une méthode appropriée dépend fortement du contexte de son application, de la nature des données et des ressources disponibles. Dans le chapitre suivant, nous allons passé à la description de l'application développée.

Chapitre III

Conception et implémentation

I. Introduction

Les développeurs des outils de fouille de données cherchent toujours à évaluer les performances de leurs outils, ce qui crée de la concurrence entre les différents entreprises. Pour qu'une entreprise améliore ces outils il faut qu'elle intègre de nouvelles méthodes et algorithme au sein de ces produits. Dans ce chapitre, on va présenter notre outil de fouille de données. Cet outil est l'extrait de l'intégration de deux outils de fouille de données (WEKA et R statistics) regroupé sous une seule interface pour l'utilisateur.

II. Java NetBeans

NetBeans est à l'origine un IDE Java développé par une équipe d'étudiants à Prague, racheté ensuite par Sun Microsystems quelque part en 2002. C'est une plate-forme, qui permet d'écrire des applications Swing ce qui fait de NetBeans une boite à outils facilement améliorable ou modifiable. La licence de NetBeans permet de l'utiliser gratuitement à des fins commerciales ou non. Elle permet de développer tous types d'applications basées sur la plate-forme NetBeans. Les modules qu'on peut écrire peuvent être open-source comme ils peuvent être closed-source, Ils peuvent être gratuits, comme ils peuvent être payants. Il existe d'autres systèmes de développement rapide mais Netbeans est particulièrement très bien placé.

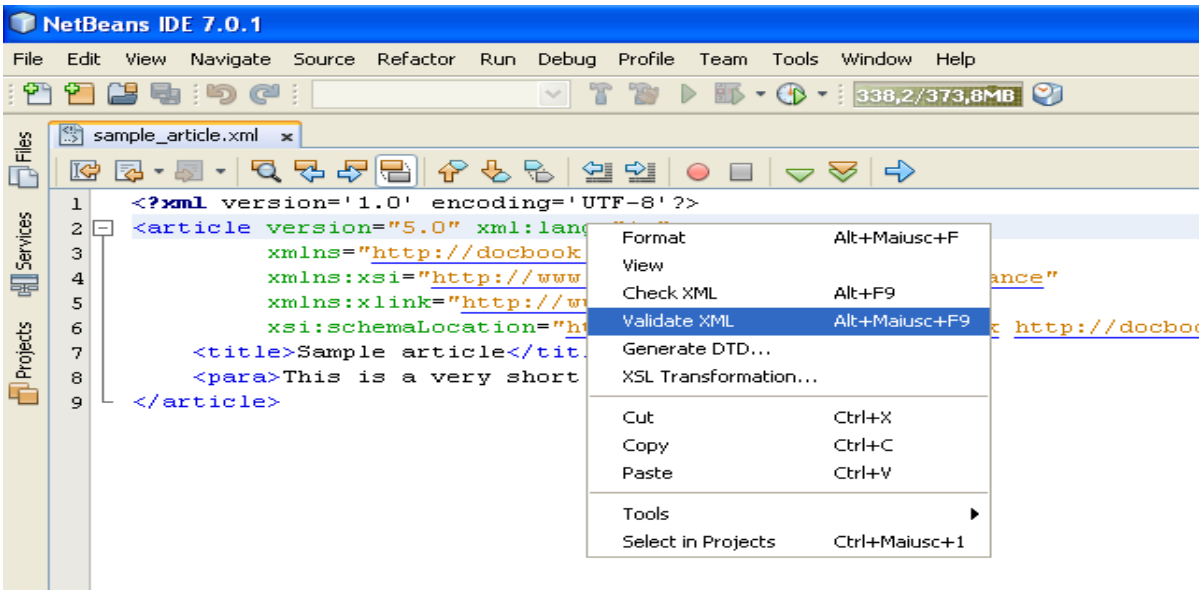


Figure III.1 : La plateforme NetBeans 7.0.1

III. WEKA

C'est uniquement le fichier JAR que nous avons utilisé pour exposer l'API WEKA pour Netbeans de sorte que nous vous pouvons programmer en toute souplesse qu'avec WEKA. On ajoute le fichier « weka.jar » à notre projet et on va avoir l'accès à toutes les classes de l'API WEKA. Dans la figure suivante, nous présentons les principaux imports de l'API.

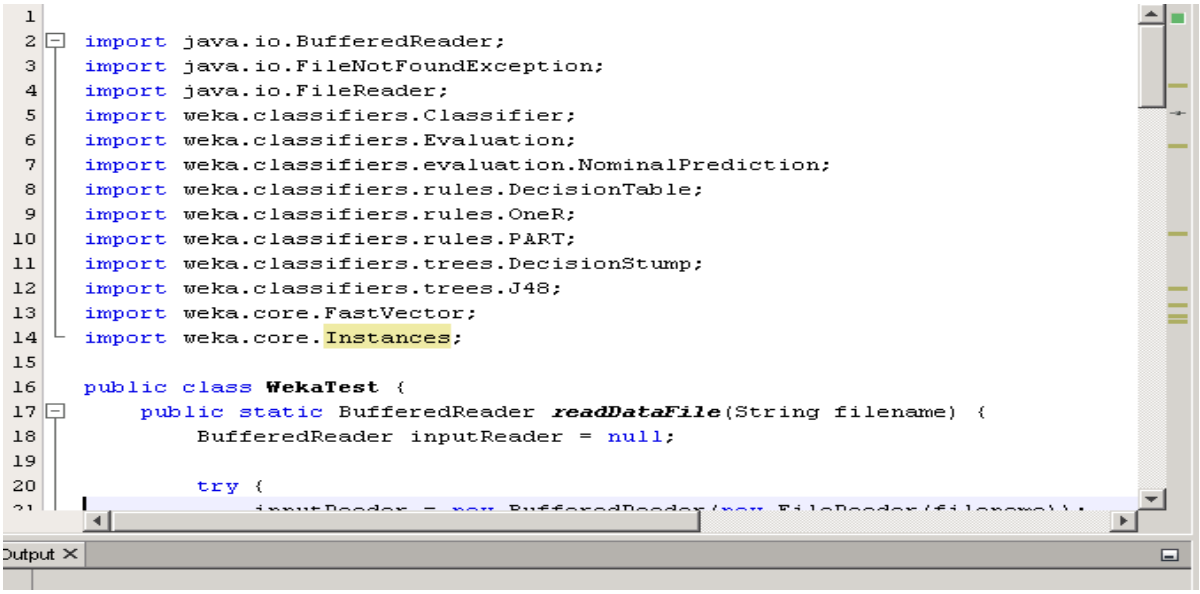


Figure III.2 : Les principales classes du fichier weka.jar

IV. R

L'utilisation de l'outil R sous Java a nécessité l'utilisation de la bibliothèque « Rcaller.jar ». C'est est une bibliothèque logicielle qui entre en scène avec sa simplicité et peut être utilisée dans des projets relativement petits. « Rcaller » convertit les structures de données en un R code, les envoie à un processus de R créé à l'extérieur, renvoie les résultats générés au format XML qui est le moyen universel de stocker des données. La structure XML est ensuite analysé et les valeurs retournées sont accessibles directement en Java. On peut créer un processus externe pour chaque opération, Cependant, cela peut provoquer un inconvénient pour la performance. « Rcaller » prend également en charge l'exécution séquentielle des commandes dans un processus de R . Comme il ne partage pas la même zone de mémoire lors de l'appel code externe, elle permet l'exécution de plusieurs processus simultanément et divise les environnements de fonctionnement.

De la même façon que WEKA, on ajoute le fichier « Rcaller.jar » à notre projet sauf que « Rcaller » nécessite le package « Runiversal.r » pour qu'il soit installé.

Classes & Fonction : Rôle
Rcode : Création d'un code source
SetRcode : Exécuter le code
StartPlot : Dessiner une parcelle
R_require : Chargement de package

Run only : Permettre de visualiser les données graphiquement
Run And Returnresult : Envoi les données à R et visualiser les résultats
AddRcode : Passant objets Java à R

Tableau III.1 : Description des principales fonctions de la bibliothèque « Rcaller »

V. Description de l'outil développé

L'outil permet l'exploration de données et le calcul statistique. Il englobe le domaine mathématique et le domaine informatique dans une seule interface. Il permet aussi de visualiser les données sous forme textuelle et graphique.

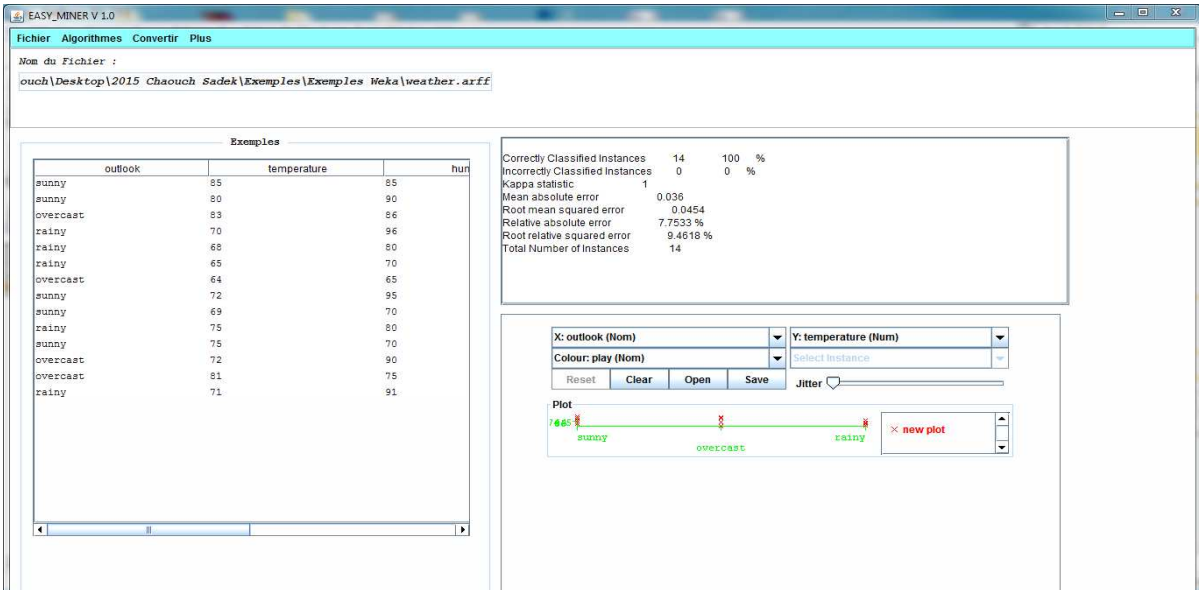


Figure III.3 : L'interface de l'outil

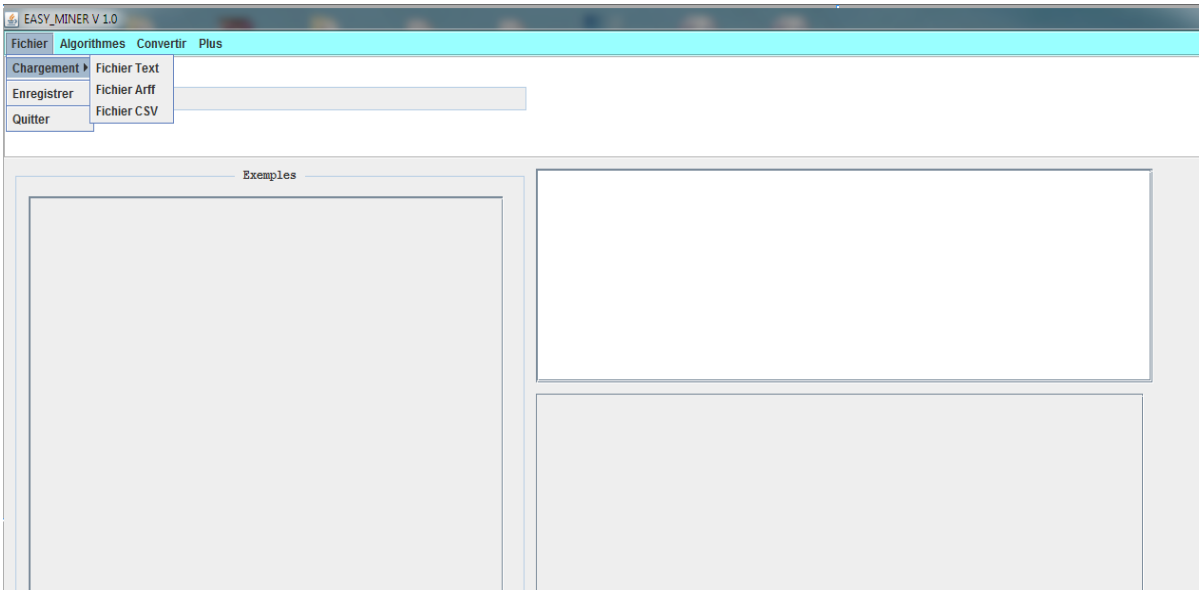


Figure III.4 : Le menu « Fichier »

A partir du menu « Fichier », on peut :

. Définir la forme du fichier qu'on souhaite le charger (txt, arff, csv).

. Enregistrer le fichier en cas de modification.

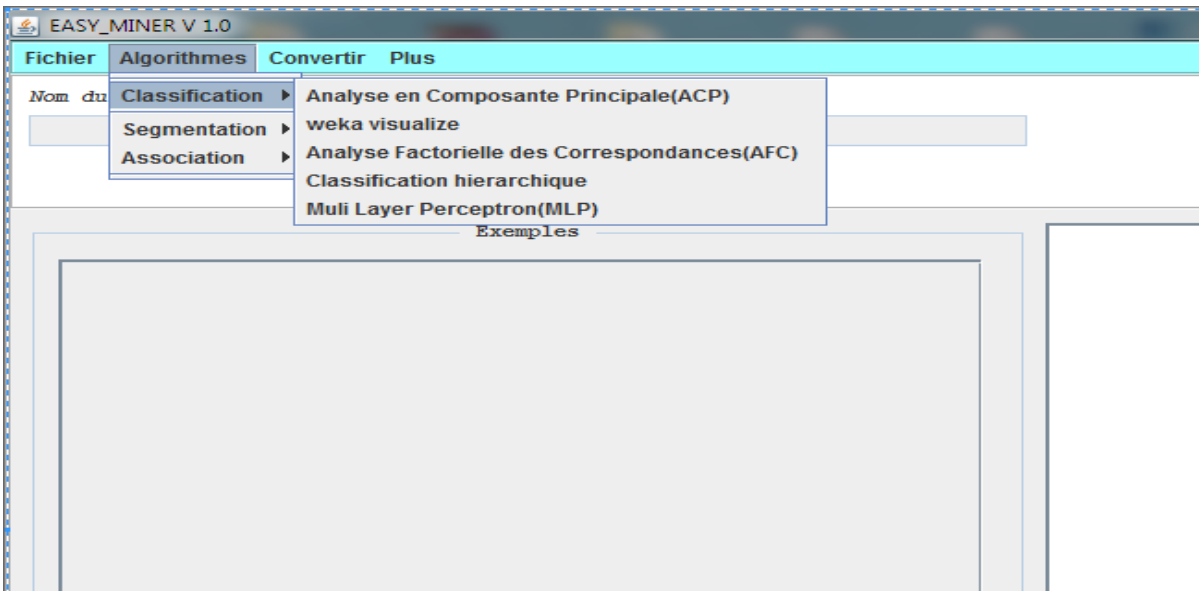


Figure III.5 : Le menu « Algorithmes »

Dans le menu « Algorithme » on trouve :

. Les Algorithmes de classification comme :

Analyse en composante principale (ACP)

Analyse factorielle des correspondances(AFC)

Classification hiérarchique

Les réseaux de neurones (Multi layer Perceptron)

. Les Algorithmes de segmentation (clustering) comme :

Kmeans(centre mobile)

. les Algorithmes d'association

Apriori

FP_Growth

K-n-n(k plus proche voisins)

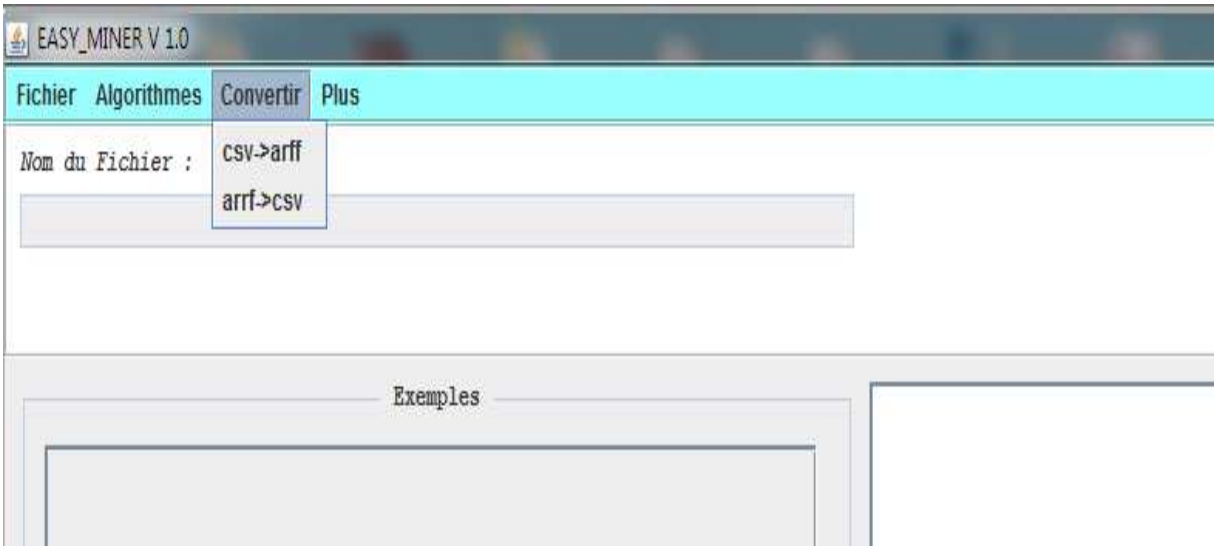


Figure III.6 : Menu « Convertir »

A partir du menu « Convertir » on peut :

. Convertir un fichier « arff » en un fichier « csv »
. Convertir un fichier « csv » en un fichier « arff »

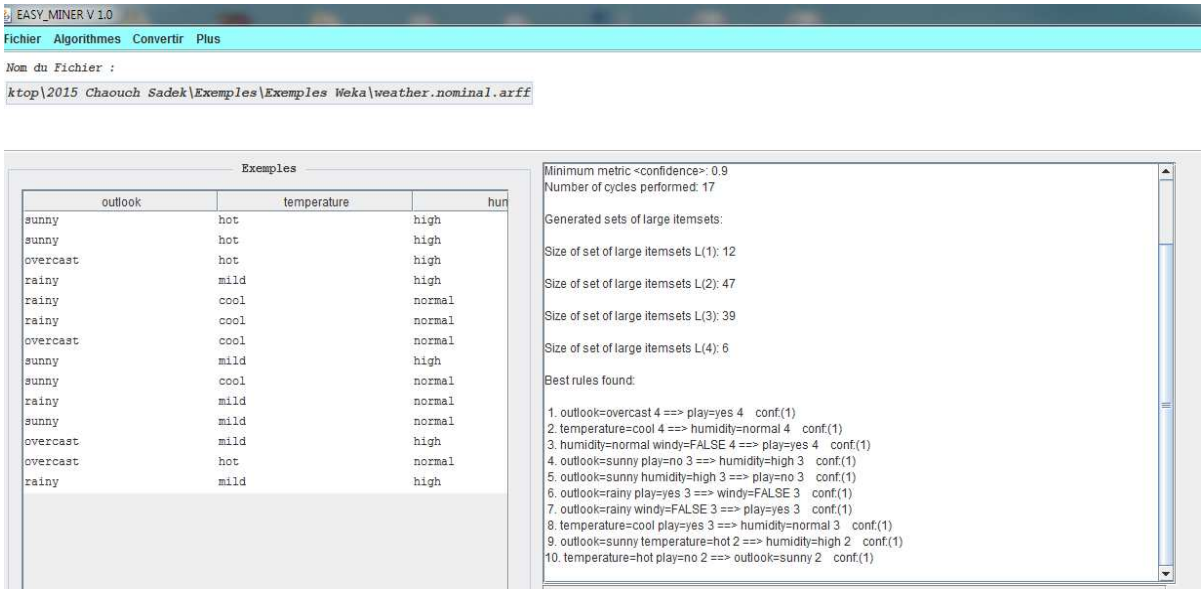


Figure III.7 : Résultat d'application de l'algorithme Apriori sur un ensemble de données

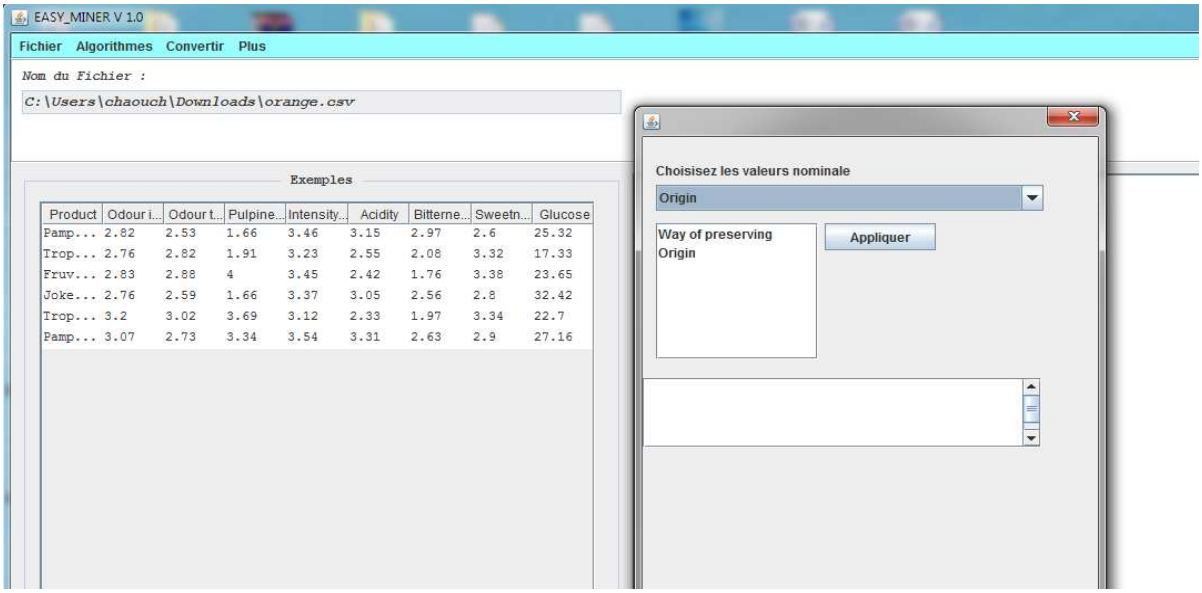
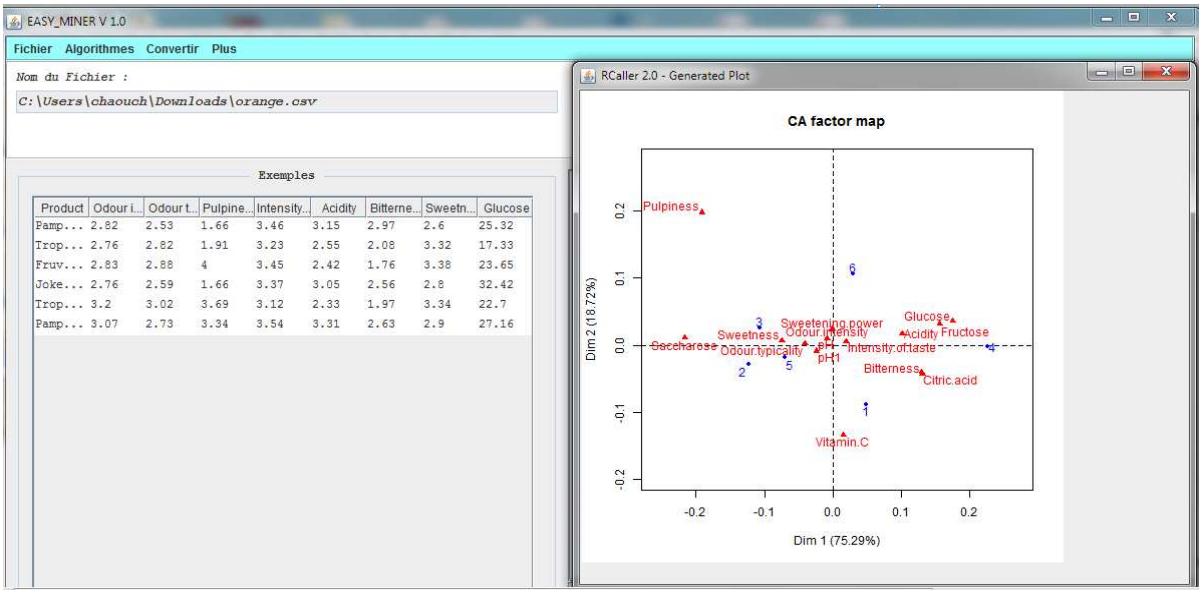


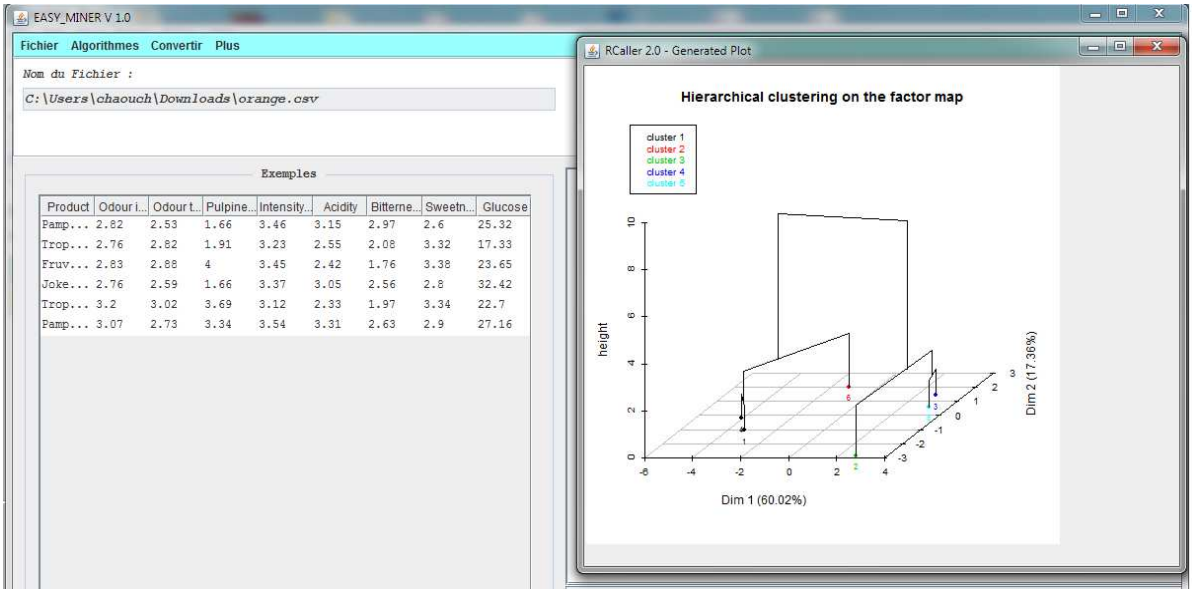
Figure III.8 : Choix des valeurs nominales

Lors d'application de l'ACP, la classification hiérarchique ou l'AFC sur un ensemble de données, on doit préciser les valeurs nominales.



Product	Odour i...	Odour t...	Pulpine...	Intensity...	Acidity	Bitterne...	Sweetn...	Glucose
Pamp...	2.82	2.53	1.66	3.46	3.15	2.97	2.6	25.32
Trop...	2.76	2.82	1.91	3.23	2.55	2.08	3.32	17.33
Fruv...	2.83	2.88	4	3.45	2.42	1.76	3.38	23.65
Joke...	2.76	2.59	1.66	3.37	3.05	2.56	2.8	32.42
Trop...	3.2	3.02	3.69	3.12	2.33	1.97	3.34	22.7
Pamp...	3.07	2.73	3.34	3.54	3.31	2.63	2.9	27.16

Figure III.9 : Résultat d'application de l'AFC sur un ensemble de données



Product	Odour i...	Odour t...	Pulpine...	Intensity...	Acidity	Bitterne...	Sweetn...	Glucose
Pamp...	2.82	2.53	1.66	3.46	3.15	2.97	2.6	25.32
Trop...	2.76	2.82	1.91	3.23	2.55	2.08	3.32	17.33
Fruv...	2.83	2.88	4	3.45	2.42	1.76	3.38	23.65
Joke...	2.76	2.59	1.66	3.37	3.05	2.56	2.8	32.42
Trop...	3.2	3.02	3.69	3.12	2.33	1.97	3.34	22.7
Pamp...	3.07	2.73	3.34	3.54	3.31	2.63	2.9	27.16

Figure III.10 : Résultat d'application de la classification hiérarchique

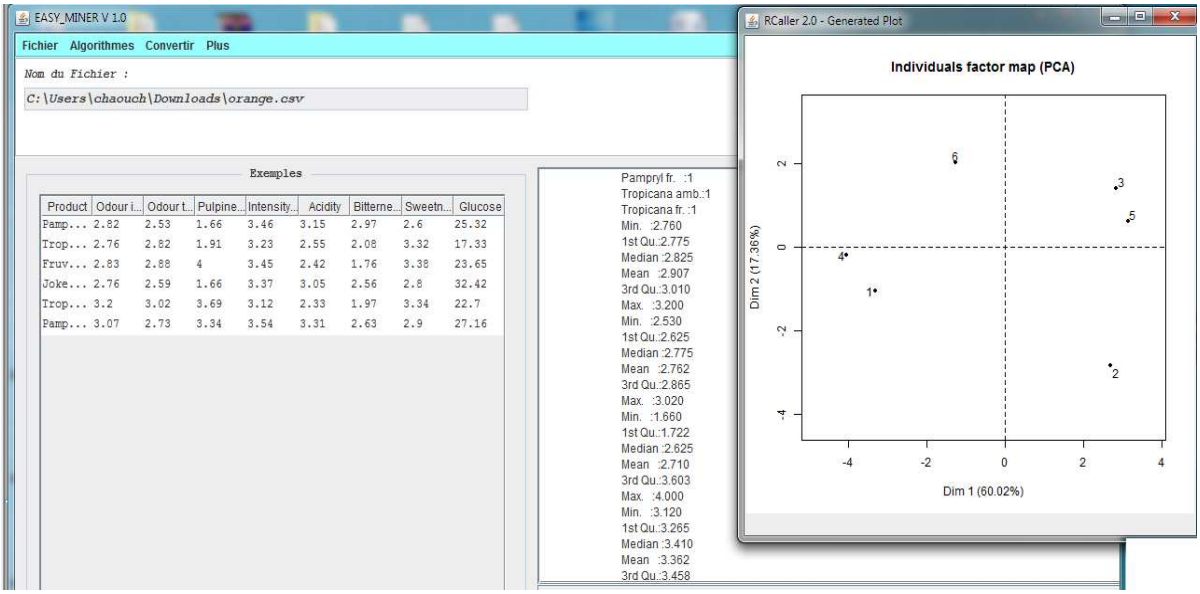


Figure III.11 : Résultat d'application de la ACP sur un ensemble de données

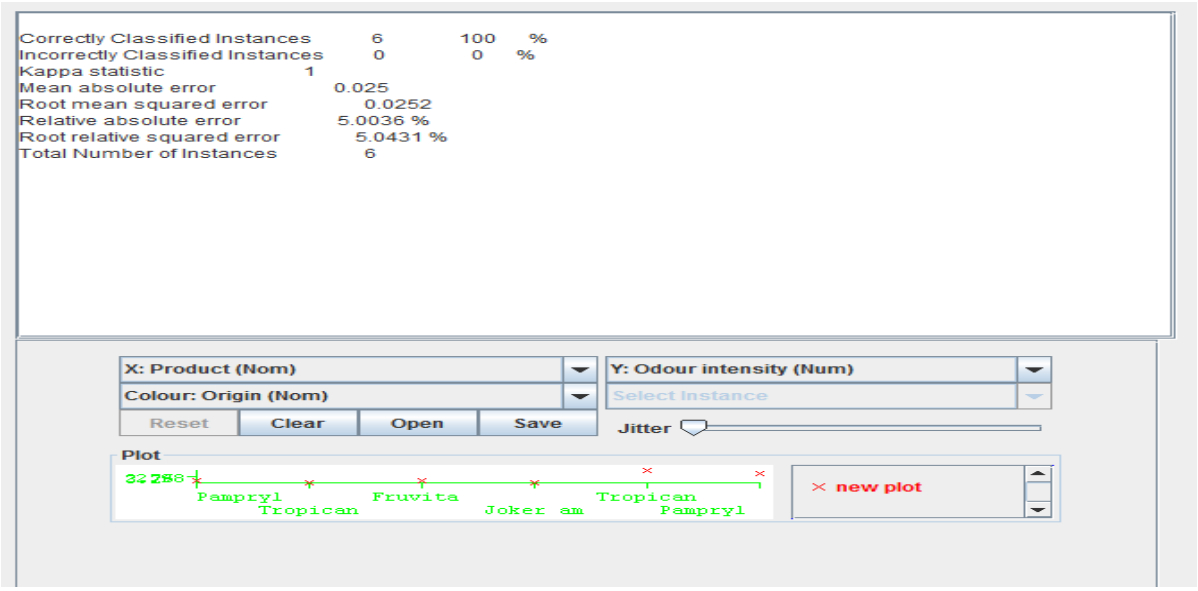


Figure III.12 : Résultat d'application des réseaux de neurones sur un ensemble de données

Enfin, l'outil enregistre les parcelles dans un fichier PDF dans le répertoire du projet.

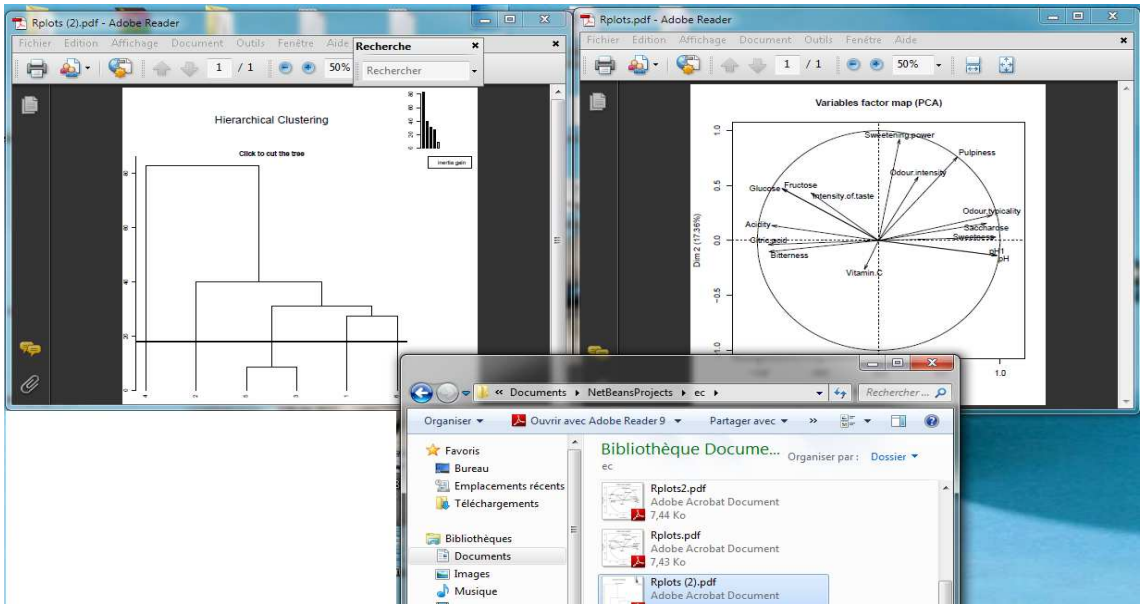


Figure III.13 : Liste des fichiers PDF enregistrés

VI. Conclusion

Nous avons présenté, dans ce chapitre, l'architecture générale de notre application qui se caractérise d'un aspect extensible. L'utilisation d'une telle architecture, donne plus de pouvoir à l'utilisateur en lui permettant d'exécuter d'autres algorithmes. Idéalement, cette indépendance inclut la possibilité pour le logiciel principal d'évoluer tout en restant compatible avec les API existants.

Bibliographie

[2]. Master Maths Finances _ 2010/2011_Data Mining février 2011 « Prise en main de RapidMiner » http://www.fil.univ-lille1.fr/~decomite/ue/MFFDD/tp1/rapidminer.pdf

[3]. **Kalpana Rangra et Dr. K. L. Bansal.** *" Comparative Study of Data Mining Tools",* **International Journal of Advanced Research in Computer Science and Software Engineering " Volume 4, Issue 6, June 2014"**

[4]. Stéphane Legrand." Test du logiciel open source de data mining « Orange »". CNAM - NFE211 - 2012/2013 17 octobre 2013

[5] .Yrd.doc.dr.AYKA CAKMAK PEHLIVANLI ,The comparison of datamining tools,data warehouse and datamining ,department of computer engineering Istanbul university *16.11.2011*

[6]**. KELLOU Kenza,MOKHTARI Abdeldjalil**. Réalisation d'une plateforme d'expérimentations et de tests d'algorithmes de data mining. **Mémoire de fin d'études, Option : Systèmes informatiques**

[7].http://chirouble.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html

[8]. **Ian H. Witten, Eibe Frank, et Mark A. Hall.** *Data Mining: Practical machine learning tools and techniques.* s.l. : 3e édition Morgan Kaufmann, 2011

[9]. **Remco, Remco Bouckaert.** *Bayesian network classi_ers in weka.12 mai 2008,university of waikato*

[10]. **Wang, Wenjia.** *Tutorial for Weka a data mining tool.* 2010.

[12]. **R, Rakotomalala.** *TANAGRA : un logiciel gratuit pour l'enseignement et la recherche.Université lyon 2 laboratoire IRIC*.

[13]. **NAOUI, Slimane OULAD.** Prétraitement & Extraction de Connaissances en Web Usage Mining S2WC2 : un WUM Framework Centré Utilisateur. Algérie : s.n., 2009.

[14]. **L. Lebart, A. Morineau et M. Piron.** *Statistique exploratoire multidimensionnelle.* s.l. : Dunod, 2000.

[15]. **Giraud, R.** *L'économétrie.* France : Collection QSJ - Presses Universitaires de France, 2000.

[16]. **D. Aha, D. Kibler, M. Albert.** *"Instance-based learning algorithms", Machine Learning.* 1991.

[17]. **D. Randall, T. Martinez.** *Improved heterogenous distance functions.* : Journal of Artificial Intelligence Research (JAIR), 1997.

[18]. **Mitchell, T.** *Machine learning.* lill : Mc Graw-Hill International Editions, 1997.

[19]. **K. Mehrotra, C. Mohan, S. Ranka.** *Elements of artificial neural network.* s.l. : MIT Press, 1997.

[20]. **T. Hastie, R. Tibshirani, J. Friedman.** *The elements of statistical learning. Data Mining, inference and predictions.* s.l. : Springe, 2001.

[21]. **Quinlan, J.R.** *Induction of Decision Trees- Machine Learning-.* 1986.

[22]. **T. Hastie, R. Tibshirani, J. Friedman.** *The elements of statistical learning. Data Mining, inference and predictions.* s.l. : Springer, 2001.

[23]. **P. Domingos, M. Pazzani.** *On the optimality of the simple bayesian classifier under zero-one loss, Machine Learning.* 1997.

[24]. **K. Mehrotra, C. Mohan, S. Ranka.** *Elements of artificial neural network.* s.l. : MIT Press, 1997.

[25]. **Johnston, J.** *Econometric methods.* s.l. : McGraw-Hill, 1972.

[26]. **Guyader, Vincent.** Abcd R. [En ligne] [Citation : 17 mai 2014.] http://Abcd'R - astuces et Scripts R.htm.

[27]. **Smith, David.** *R users: Be counted in Rexer's 2013 Data Miner Survey.* s.l. : Revolution Analytics Blog, 30 janvier 2013.

[28]. **Gentleman, Ross Ihaka et Robert.** [En ligne] [Citation : 17 MAI 2014.] http://The R Project for Statistical Computing.htm.

[29]. **Rakotomalala, Ricco.** Introduction à R Arbre de décision. [En ligne] [Citation : 17 MAI 2013.] http://eric.univ-lyon2.fr/~ricco/cours/cours_programmation_R.html.

[30]. **Regression, Support Vector.** [En ligne] 15 Avril 2009. [Citation : 17 Mai 2013.] http://tutoriels-data-mining.blogspot.com/2009/04/support-vector-regression.html.

[31]. **Monbet, V.** Analyse Discriminante Décisionnelle. *Analyse de données.* 2012.

[32] http://www.petite-entreprise.net/P-2595-83-G1-principales-taches-du-data-mining.html

[33] Georges El Helou , Charbel Abou khalil « Data Mining ,*Techniques d'extraction des connaissances »* Module 4.1 - Management et NTIC, Professeur : Mélissa Saadoun ,Projet soutenu le 16 février 2004

[34] http://www.rithme.eu/?m=resources&p=dmdomains&lang=fr

[35] http://kdacy-consulting.com/main/le-processus-ecd

[36] Ralf Mikut,and Markus Reischl "Data mining tools**", _ 2011 John Wiley & Sons,**