Algerian People's Democratic Republic
Ministry of Higher Education and Scientific Research

University Mohamed Seddik Benyahia Jijel

Faculty of exact science and informatics

Department of informatics

Option

Information System and Decision Support

Detection and analysis of LMD students' profiles using cluster analysis

Supervised by:

Dr. Doulkifli Boukraâ

presented by:

Rami abdelouahab

Salah-eddine chouieb

Academic year 2015-2016

Algerian People's Democratic Republic
Ministry of Higher Education and Scientific Research

University Mohamed Seddik Benyahia Jijel

Faculty of exact science and informatics

Department of informatics

Option

Information System and Decision Support

Detection and analysis of LMD students' profiles using cluster analysis

Supervised by:

Dr. Doulkifli Boukraâ

presented by:

Rami abdelouahab

Salah-eddine chouieb

Academic year 2015-2016

ACKNOWLEDGMENTS

First of all, I would like to give thank to Allah Almighty for inspiring me and guiding me to finish this work. I would like to thank my father and my mother for tolerating me throughout the whole process of my studies, my brothers, sisters, aunts, uncles, cousins and all my family and friends. I shall not forget our supervisor Dr. Doulkifli Boukraâ for his unwavering support, collegiality and mentorship throughout this project.

Rami abdelouahab

First and Foremost, let me express my thankfulness and gratefulness to Allah Almighty my Source of Strength, then to my mother, my father, my brothers, my sisters and all my family and friends; thank you for being with me all the time and for supporting me, without forgetting to thank Dr. Doulkifli Boukraâ for his instructions and advice.

Salah-eddine chouieb

الملخص

يهدف هذا العمل لتسهيل و تحسين مراقبة و تحليل نتائج الطلاب في مختلف العناصر التعليمية. تقنية التكتل تسمح بالكشف عن لمحات الطلاب من خلال نتائجهم المتحصل عليها.

من أجل هذا، قمنا بإدخال نتائج الطلاب في مختلف العناصر التعليمية في قاعدة البيانات وبعدها حولنا هذه المعلومات إلى شكل قابل للاستعمال وطبقنا عليها خوارزمية K-means للكشف عن لمحات الطلاب الموجودة .في الأخير عرضنا تطبيقان لإنشاء و تحليل هذه اللمحات.

Résumé

Ce travail a pour but de faciliter et d'améliorer le suivi et l'analyse des résultats des étudiants dans divers éléments éducatifs. La technique de clustering permet de détecter les profiles des étudiants à travers les résultats obtenus.

Pour cela, nous avons alimenté les schémas sources de données par les résultats des étudiants, puis nous avons transformé et préparé ces données sous une forme utilisable pour appliquer l'algorithme le K-means qui permet de détecter les profils. Enfin nous avons présenté deux applications pour créer et analyser les profiles.

Abstract

This work is intended to facilitate and improve the monitoring and analysis of the students' results in various educational elements. The clustering technique allows detecting students' profiles based on the obtained results.

For that, we have populated the results of students in various educational courses into the data source schema, and then we transformed and prepared the data for applying the K-means algorithm to detect students' profiles. Finally we presented two applications for creating and visualizing the profiles and analyzing their evolution.

Contents table

Lists of figures

List of tables

Chapter I

General introduction

Data mining, or the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools, among others, predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Clustering technique involves placing data into related groups, and that can help businesses manage their data better. For retail businesses, data clustering helps with customer shopping behavior, sales campaigns and customer retention.

In the domain of education, methods and techniques are used for exploring data originating from various educational information systems. This data can be used for analyzing and predicting student's behaviors and the skills obtained during the learning path. Clustering algorithms are applied in *Educational Data Mining* to create groups of students with similar learning style and also provides a fairly ambiguous profile of learning behaviors.

To achieve the last mentioned objectives, we need tools that can explore a large amount of data to comprehend relations between data and to extract models that explain these data. We aim at detecting student's behaviors that explain the obtained skills along the academic path.

Our study environment is the University of Jijel, particularly, the department of Mathematics and Informatics (MI) and the department of Informatics. Student's data concerning marks or personal information is being stored and managed in both departments; this data comes from different sources, for example professors provides data about student's marks, courses progress, absence lists, etc. For students' personal information, the data comes from the central schooling.

In a previous work of a master's project, a data ware housing system was developed to store data about pedagogic following that is courses follow-up, work directed, practical works, and mainly, students' marks in modules, semesters and etc.

Our work consists in exploiting the data with data mining technique. Our principal motivation is to find hidden models relative to students like the explanation of good and bad results obtained by students and to discover student's profiles.

The objective of this work is exploiting the data by data mining techniques. We aim at detecting and analyzing students' profiles based on their results. These profiles correspond to clusters. We exploit the techniques of clustering on students based on the marks obtained in each module. These profiles help the decisions makers better understand the learning process and the skills obtained by students.

To reach our objectives, we contribute with the next elements:

1. An implementation of data mining with the clustering technique. This implementation requires collecting and preparing the necessary data, the data mining itself and the explanation of results.
2. Two applications to be used respectively by (1) a *data miner* user and (2) an *analyst* user. The data miner application is used to create and save models, corresponding to student profiles. The analyst application allows the user to visualize the profiles and to analyze profile evolutions.

Our master thesis is structured according to the following plan:

In chapter 2, we present the LMD system. Next we present the definition, the process and techniques of data mining. Finally we present the different concepts of educational data mining; we also presented some related works concerning clustering in educational data mining.

In chapter 3, we present the source schema and the population of this schema by the collected data.

Chapter 4 is dedicated to data acquisition from the source schema and the preparation of these data for the data mining activity.

Chapter 5 is about the implementation of data mining. We present, the third-party tools that we used to carry out the data mining and to detect the students' profiles, as well an algorithm for labeling these profiles. Finally we present our applications and some results.

Chapter II

State of the art

1. Introduction

Data is growing massively during the last decades; databases in different fields such as finance, economics, science, social media, etc. hold a lot of information. This information can be used to detect and extract hidden knowledge and patterns in fields like economics and use it for strategic decisions or other demands. Data mining is used explore and extract knowledge with different techniques.

In this chapter we will define the educational LMD system that is used at the University of Jijel. Next we present some concepts of data mining. Finally we will present educational data mining and some of its uses.

2. LMD System

LMD system is originally a European higher education system that stands for Licence (Bachelor's degree), Master and Doctorate. The system was designed to create a unified higher education system within the European Union. The LMD was first introduced in Algeria during the academic year 2004/2005 and was initially applied in 10 of higher education institutions from a total of 58 institutions .The LMD system is organized into three study phases (Cycles):

- Licence (Bachelor's degree):Baccalaureate +3
- Master: Baccalaureate +5
- Doctorate: Baccalaureate+8

2.1. Advantages [1] :

- **Step of education**
  - The courses are offered in the field of education and course of study.
  - The courses are grouped into fields of education.
  - One field is containing several sciences.
  - A year of studies is divided into semesters. In each semester, the courses are grouped into teaching units and units. Each course or unit has a corresponding coefficient and a number of credits.
- **Teaching evaluation:**
  - The educational knowledge evaluation is based on the student marks but most importantly, on the number of credits that are obtained by the student.
  - For each semstre, the required number of credits is equal to 30.

    Thus, in order
    - to obtain the Licence degree the student must accumulate 180 credits (6 semesters).
    - to obtain the Master's degree the student must accumulate 120 credits (4 semesters).

  Figure 2.1 demonstrate the description of the LMD system.

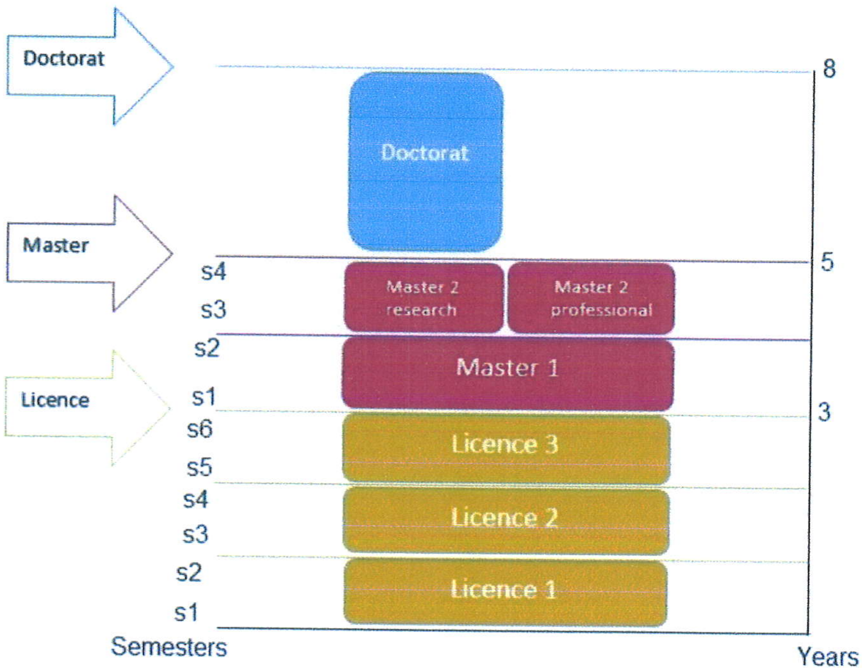


Figure 2.1: LMD system description

2.2. The LMD general characteristics [2]

In this section, we present some characteristics and terminology of the LMD system

- **Study unit**:

  The study units are divided into:

  - **Fundamental units:** they gather the courses that are considered fundamental to the degree and that have a tight relation with the field of studies.
  - **Exploratory units** the courses in these units provide the student with extra knowledge that is useful, although not directly related to the field of studies.
  - **Methodology & general culture units**: these units provide the student with the necessary tools of the scientific research and it acquires his endogenous capacity of work (computing, statistical systems, scientific research methodology, foreign languages…etc).

- **The credit-based system:** a credit represents the time size of the work that the student must accomplish under the received training that is divided into two parts. The first part contains collective and attending work that can be exemplified through lectures, practical and class works. The second part of the work is individual which means that the student does it alone (e.g at home) through the presentation, practicum, reports and dissertations.
- Each module has a credit. The unit credit is determined by the total of the modules credits or its composing modules.
- Once necessary conditions for succeeding a unit are met, the unit is obtained. Besides, since a credit corresponds to the time size of student work the passing mark that the student will get in any module represents the evaluation to this work and thereby allows him/her to obtain the module' s credit.
- The LMD system gives two training paths :(1) the academic path which grants the scientific diplomas at the level of BA, master and doctorate studies and (2) the professional path which provides the BS and the professional master diplomas to prepare the student for professional career.
- When the students obtain the 180 required credit at the first stage, they will culminate with the BA of BS diploma and they will have the right to obtain the master diploma after obtaining 120 credit.

3. Data Mining

3.1. Definition

Data Mining, also commonly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. [3]

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Figure 2.1 depicts data mining at the crossroad of many disciplines.

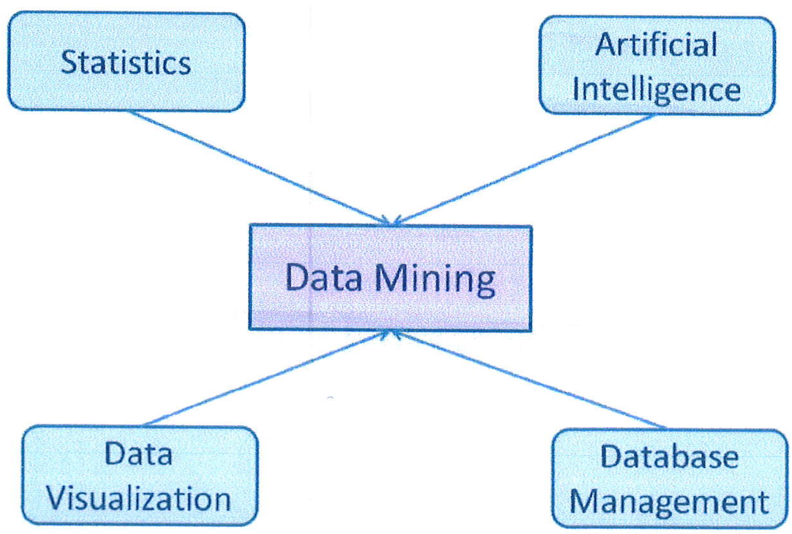


Figure 2.2: Contributing disciplines to data mining

3.2. History

The term "data mining" was introduced the first time back in the 80s within the research community. By the early 1990s, data mining was commonly recognized as a sub-process within a larger process called Knowledge Discovery in Databases or KDD (although in the modern context of data mining Knowledge Discovery in Data would be more apt, as we are no longer preoccupied solely by databases).

The most commonly used definition of KDD is that attributed to Fayyad et al.:«The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data» (Fayyad et al. 1996). [4]

3.3. Mechanisms and techniques [4]

The mechanisms and techniques within the remit of data mining can be described as an amalgamation of approaches to machine learning and statistics; from this perspective data mining can be said to have "grown" out of the disciplines of machine learning and statistics.

Traditionally data mining techniques can very broadly be categorized as being directed as either: (i) pattern extraction/identification, (ii) data clustering or (iii) classification/categorization. Each is briefly considered in more detail in the following subsections. Within the current data mining literature we can also find reference to many other techniques that have been adopted from fields such as statistics and mathematics, for example linear regression and Principal Component Analysis (PCA).

3.3.1. Pattern extraction

Searching for patterns is one of the main goals in data mining these patterns can take many forms like customer purchasing patterns; a pattern is any frequently occurring combination of entities, events, object, etc. The exemplar pattern mining technique is Association Rule Mining (ARM) as first proposed by Agrawal et al [5].

3.3.2. Clustering

Clustering is concerned with the grouping of data into categories. We try to cluster data into a predefined number of clusters such as in k-means or according to some proximity threshold, as in the case of the well-established KNN algorithm based on a similarity measure.

Clustering is usually used to profile individuals that have some common behaviors like customer habits.

3.3.3. Classification

Classification is concerned with the construction of classifiers that can be applied to unseen data so as to categorize that data into groups (classes). As such classification has parallels with clustering, the distinction, however, is that classification requires pre-labelled training data from which the classifiers can be built; such classification is sometimes referred to as supervised learning while clustering is considered to represent unsupervised learning.

3.4. Application of data mining

Data mining is used in different fields and domains such as:

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection
- Customer profiling

3.5. Data mining process within KDD [3]

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, maybe combined in a common source.
- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: at this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: it is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

The following Figure 2.3(fig 2.3) illustrates the KDD process.

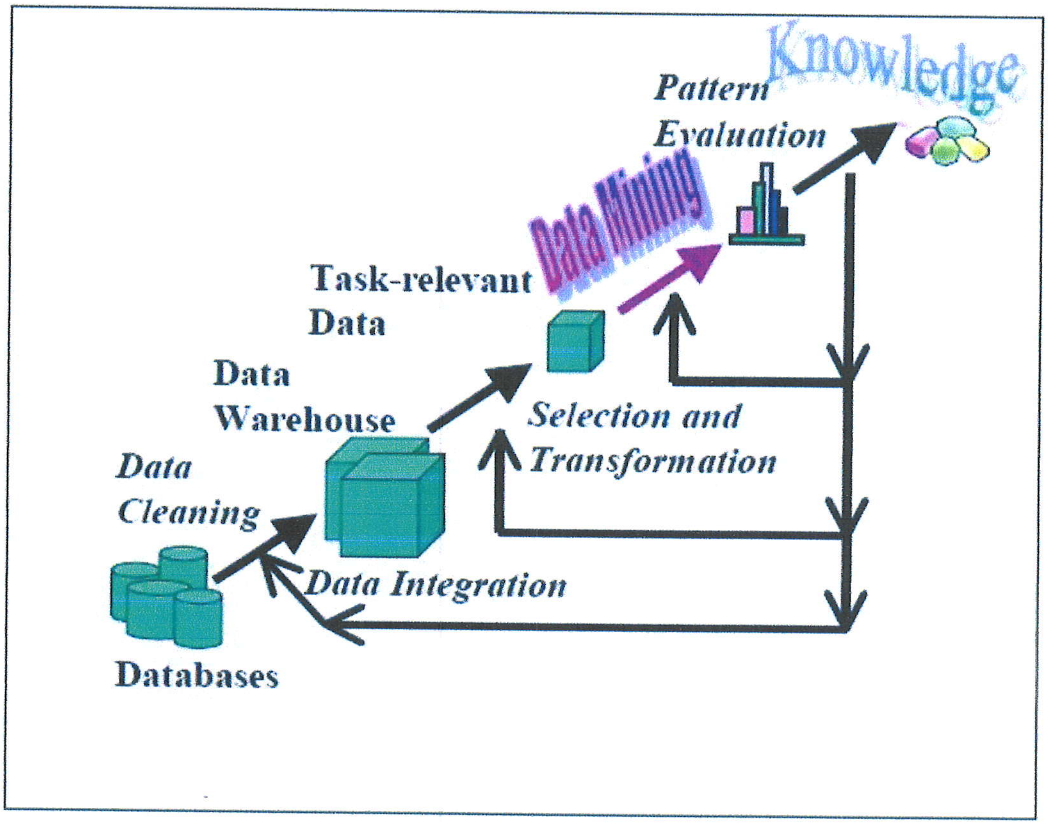


Figure 2.3: KDD process

4. Clustering

4.1. Definition

Clustering is a data mining technique based on grouping similar objects in a same cluster and dissimilar objects in different clusters from a data set. Clustering is an unsupervised learning technique, because the initially clusters are not known. There are various clustering algorithms such as: K-means, KNN (k-nearest neighbors), k-modes, hierarchical clustering etc.

4.2. Distance Measures:

The object or individuals are grouped based on dissimilarities/similarities which may be measured using distances. To form the clusters, different types of distances can be used depending on the characteristics (variables) that describe the objects or the individuals. Examples of distances are,

- Euclidean distance : it represents the geometric distance on multidimensional space, its formula: $D(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
- Squared Euclidean distance : squaring the Euclidean distance adds more weight to the difference between object computed with:
- $D(x,y)=\sum_{i=1}^{n}(x_i-y_i)^2$
- City-block (Manhattan) distance :$D(x,y)=\sum_{i=1}^{n}|y_i-x_i|$

4.3. Techniques [6]

The two main types of clustering techniques are those that create a hierarchy of clusters and those that do not.

- **Hierarchical clustering:** The hierarchical clustering techniques create a hierarchy of clusters from small to big. With a hierarchy of clusters defined it is possible to choose the number of clusters that are desired.
- **Non-Hierarchical clustering:** The non-hierarchical techniques in general are faster to create from the historical database but require that the user makes some decision about the number of clusters desired. The non-hierarchical techniques most of the times are run multiple times starting off with some arbitrary or even random clustering and then iteratively improving the clustering by shuffling some records around.
- Examples of non-hierarchical clustering techniques are k-means, K--NN (nearest neighbors), k-modes, etc.

4.3.1. Basic K-means algorithm:

Algorithm K-means

Select K points as the initial centroids.

Repeat

Form K clusters by assigning all points to the closest centroid.

Recompute the centroid of each cluster.

Until the centroids do not change

- **Example of k-means:**

  As a simple illustration of a k-means algorithm, we consider the following dataset table (2.1) consisting of two variables and seven individuals. We try to group the dataset into two clusters (K=2) choosing the Euclidian distance:

Individuals	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Table 2.1 : Data set

At the first step we define the initial cluster means:

The first centroids can be chosen randomly, but such a random selection is likely to have it effect the algorithm converging time. Therefore, the first centroids that are usually chosen furthest apart table (2.2):

	Individual	Mean Vector (centroid)
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Table 2.2 : Initial centroids

The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean. The mean vector is recalculated each time a new member is added. This leads to the following series of steps table (2.3):

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)

4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

Table 2.3 : Serie of steps

Now the initial partition has changed, and the two clusters at this stage having the following characteristics:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.8, 2.3)
Cluster 2	4, 5, 6, 7	(4.1, 5.4)

Table 2.4 : New clusters

But we cannot yet be sure that each individual has been assigned to the right cluster. So, we compare each individual's distance to its own cluster mean and to that of the opposite cluster. And we find table (2.5):

Individual	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

Table 2.5 : Distances comparison

Only individual 3 are nearer to the mean of the opposite cluster (Cluster 2) than its own (Cluster 1). In other words, each individual's distance to its own cluster mean should be smaller that the distance to the other cluster's mean (which is not the case with individual 3). Thus, individual 3 is relocated to Cluster 2 resulting in the new partition, table (2.6) shows the finals clusters:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)

Table 2.6 : Final clusters

The iterative relocation would now continue from this new partition until no more relocations occur. However, in this example each individual is now nearer its own cluster mean than that of the other cluster and the iteration stops, choosing the latest partitioning as the final cluster solution.

Also, it is possible that the k-means algorithm will not reach a final stable solution. In this case it would be a good idea to consider stopping the algorithm after a pre-chosen maximum of iterations.

4.4. Application of clustering

Clustering is being applied in various fields; some of the applications are:

- Educational data mining
- Pattern recognition
- Image analysis
- Bioinformatics
- Voice mining
- Image processing
- Text mining
- Web cluster engines

5. Educational data mining (EDM)

5.1. Definition [7]

Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. Whether educational data is taken from students' use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities, it often has multiple levels of meaningful hierarchy, which often need to be determined by properties in the data itself, rather than in advance. Issues of time, sequence, and context also play important roles in the study of educational data.

5.2. History

The expression "Educational data mining" appeared in a series of workshops on the theme of the analysis of the uses (log data) by students, the first goes back to the ITS 2000 Conference in Montreal. In 2005, the first workshop titled "Educational Data Mining" (EDM, Educational Data Mining) was held in Pittsburgh in conjunction with the AAAI conference (Association for the Advancement of Artificial Intelligence). Since 2008,EDM refers to an international conference held annually. In addition, a conference on the nearby theme of learning analytics (LAK2011) appeared in 2011. In 2009 was published the first international journal EDM number. [8]

5.3. Application of educational data mining

A list of the primary applications of EDM is provided by Cristobal Romero and Sebastian Ventura.[10] In their taxonomy, the areas of EDM application are:

- Analysis and visualization of data
- Providing feedback for supporting instructors
- Recommendations for students
- Predicting student performance
- Student modeling
- Detecting undesirable student behaviors
- Grouping students
- Social network analysis
- Developing concept maps

- Planning and scheduling

5.4. How EDM methods are applied? [11]

Applications of EDM methods comprise several steps (Figure 2.4). Initially, a design is planned, i.e., the main aim of the study and the required data are identified. Afterwards, the data is extracted from the appropriate educational environment. Frequently, data will need to be pre-processed, since it may come from several sources or have different formats and levels of hierarchy. Models or patterns are obtained from applying EDM methods, which have to be interpreted. If the conclusions suggest applying changes to the teaching/learning process or are not conclusive (because the problem has not been adequately addressed, the raw data are small or not suitable, or the selected methods are not powerful enough), the analysis is performed again after modifying the teaching/learning process or the study design.

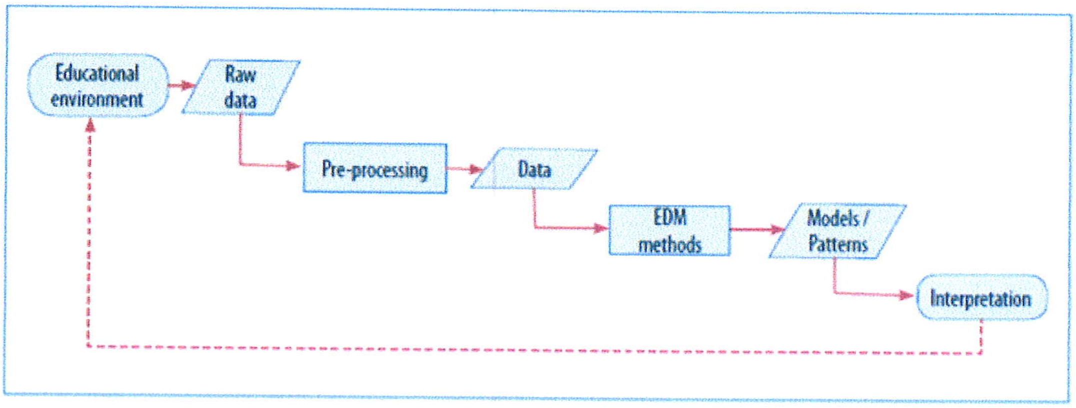


Figure 2.4: EDM process

5.5. EDM using clustering

5.5.1. Goal

The application of clustering in the educational field aims at grouping students into categories according to their behavior in an educational environment based on their grades or participation in different tasks. Another goal is to determine positive and negative skills of students.

5.5.2. Existing work on clustering educational data

The following Table 2.1 Summarizes existing work about the use of clustering, among others, in the educational field.

	Problem/ Objective	Algorithm/Method	Dataset/Data source
[12]	Evaluating undergraduate student academic performance	Using a combination of DM methods like ANN(Artificial Neural Network), Farthest First method based on k -means clustering and Decision Tree as a classification approach	Student data of the Computer Science department at Faculty of Science and Defense technology, National Defense university of Malaysia (NUDM)
[13]	To predict the potentiality of students performance who can fail during an online curriculum in a Learning Management System(LMS)	Expectation Maximization, Hierarchical Clustering, Simple k-Means and X-Means as provided in WEKA software has been used.	Real life dataset provided by Juris campus accessible at http://www.juriscampus.fr/
[14]	Shows the applications of various DM techniques on student academic data.	Apriori Algorithm is applied to academic records of students to obtain the best association rules which help in student profiling-K-means clustering is used to group students categorically.	Student academic record file, no mention of where it's obtained from.

[15]	To identify the significant variables that affects and influences the performance of undergraduate students	C-Means clustering method	Academic dataset from the state University of Santander (IUS). Contains basic student data like faculty, academic program, gender, age, origin, student category and academic achievements data
[16]	To develop student profiles of learner behavior from learner's activity in an online learning environment and also to create click-stream server data	Two clustering methods used, Hierarchical clustering (Ward's clustering) and Non- Hierarchical Clustering method (k-means clustering)	Information not provided

Table 2.7: Existing work on clustering educational data

6. Conclusion

In this chapter we presented the LMD system and different concepts related to data mining. Then we narrowed our interest to educational data mining and the use of clustering. In the next chapter, we will to describe data sources that will feed the data mining process about clustering student data at the University of Jijel.

Chapter III

Database population

1. Introduction

In order to achieve the goal of our work we need to input more data into the source database schemas which were created in a previous work in [17].In the work [17], a data warehousing system was created to input data from two sources schemas: *suivi_pédagogique* , and *résultat*. The database *Suivi_pédagogique* contains data about pedagogic activities and the schema *résultat* contains the results of students.

The two schemas *suivi_pédagogique, résultat* feed data warehousing system with the necessary data for decision making. Our work can be seen as the front-end of the warehousing system by using data mining techniques in order to discover hidden knwoledge. However, the scope of our work in only limited to schema *résultat,* thus we ignore the other schema.

In this chapter we present the data warehousing system. Next we will describe the collected data. Later, we will give details on how we input data into the source schema and the tools that we used.

2. Presentation of the data source schema

In the following section we present the data source schema *résultat*

2.1. Data source *résultat*

The schema *résultat* contains data about:

- Students assignments to groups
- Results of students per study element, study unit and per semester.

Figure 3.1 depicts the conceptual model of source *résultat.*

2.1.1. Conceptuel Data source *résultat*

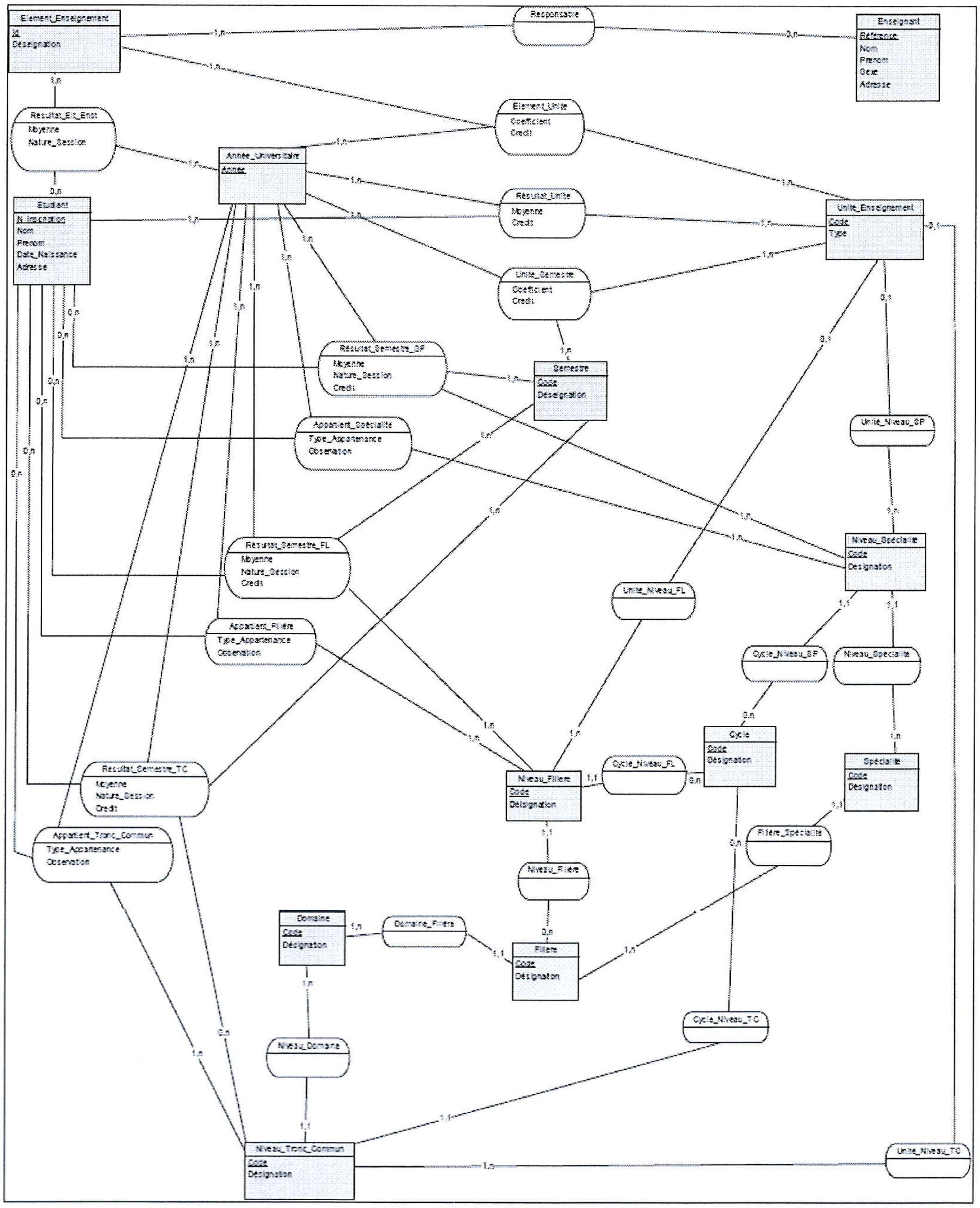


Figure 3.1: Conceptual schema of source *résultat*

2.1.2. Description of data source *résultat*:

In table 3.1 we describe each entity of the data source *résultat*.

Entities		
Name	Description	Example
Etudiant	Contains informations about Students.	**N_inscription :** 09/3032842 **Date_Naissance :** 30/05/90
Enseignant	Contains informations about Professors.	**Référence :** E1
Element_Enseignement	Contains education elements	**Id :** ED **Désignation :** Entrepôt de Données
Unité_Enseignement	Contains education units	**Id :** UEF1 **Type :** Fondamentale
Année_Universitaire	Contains years	**Année :** 2009-2010
Semestre	Contains semesters of each year	**Code :** S1L **Désignation :** Premier semestre licence
Cycle	Study cycle :license, master, doctorate	**Code :** L **Désignation :** Licence
Domaine	Contains the Domains	**Code:** MI **Désignation :** Mathématique et informatique
Filiere	Contains the study courses	**Code :** INF **Désignation :** Informatique

Spécialité	Contains the specialties	**Code:** SIAD **Désignation :** Systèmes d'information et aide à la décision
Niveau_Tronc_C	Contains levels of each specialty	**Code :** 1.L.MI **Désignation :** Première année licence mathématique et informatique
Niveau_Filière	Contains levels of each Study course	**Code :** 2.L.INF **Désignation :** Deuxième Année Licence Informatique
Niveau_Spécialité	Contains levels of each Specialties	**Code :** 1.M.SIAD **Désignation :** Première Année Master Systèmes d'information et aide à la décision

Table 3.1: Entities of the source schema *résultat*

In table 3.2, we describe each association of the data source *résultat.*

Associations	
Responsible	Represents the responsible of the education element
Element_Unité	Represents the membership of a element to a unit
Resultat_Elt_Enst	Represents the results obtained by students for the education elements
Resultat_Unité	Represents the results obtained by students for the education units
Unité_Semestre	Represents the membership of the units to semesters
Résultat_Semestre_TC	Represents the results of semesters obtained by the students Who are members of a level of common core
Résultat_Semestre_FL	Represents the results of semesters obtained by students Who are members of a level of course
Résultat_Semestre_SP	Represents the results of semesters obtained by students Who are members of a level of specialty
Appartient_Domaine	Represents the membership of levels to a domain
Appartient_Filière	Represents the membership of level s to a course
Appartient_Spécialité	Represents the membership of level s to a specialty
Niveau_Domaine	Represents the membership of level s to a domain
Niveau_Filière	Represents the membership of levels to a course
Niveau_Spécialité	Represents the membership of level s to a specialty
Unité_Niveau_TC	Represents the membership of education units to domains levels

Unité_Niveau_FL	Represents the membership of education units to levels That belong to courses
Unité_Niveau_SP	Represents the membership of education units to levels That belong to specialties
Domaine_Filière	Domains are divided to many courses ,and the course has Only one domain
Filière_Spécialité	Courses are divided to many specialties ,and the specialty has only one course
Cycle_Niveau_TC	Represents the membership of domain levels to cycles

Table 3.2: Association schema source *résultat*

2.1.3. Relationnel model of data source *résultat*

The relational model of data source is described as follows:

- Etudiant (N_Inscription, Nom, Prénom, Date_Naissance, Adresse).
- Enseignant (Référence,Nom,Prénom,Sexe,Adresse).
- Element_Enseignement (Id,Désignation).
- Unité_Enseignement (Code,*#Code_Niv_Spécialite*,*#Code_Niv_Filiere*,*#Code_Niv_Tronc_C*, Type).
- Année_Universitaire (Année).
- Semestre (Code,Semestre,*#Année*).
- Cycle (Code,Désignation).
- Domaine (Code,Désignation).
- Filiere (Code,Désignation,*#Code_Domaine*).
- Spécialité (Code, Désignation,*#Code_Filiere*).
- Niveau_Tronc_C (Code, Désignation,*#Code_Domaine*,*# Code_Cycle*).
- Niveau_Filiere (Code,Désignation, *Code_Filiere*,*# Code_Cycle*).
- Niveau_Spécialite (Code,#Code_Spécialite,# Code_Cycle,Désignation).
- Responsable (*#Référence*,*#Id_Elt_enst*,*#Année*).

- Element_Unité*(#Id_elt_enst,#Code_Unité,#Année*,Coefficient,Credit).
- Résultat_Elt_Enst(*#N_Inscription,#Id_Elt_Enst,#Année*,Moyenne,Nature_Session).
- Résultat_Unité (*#N_Inscription,#Code_Unité,#Année*,Moyenne,Credit).
- Unité_Semestre(*#Code_Unité,#Code_Semestre,#Année*,Coefficient,Cedit).
- Résultat_Semestre_TC(*#N_Inscription,#,#Code_Semestre*,#Année,*#code_Niv_TC*,Moyenne_Semestre,Nature_Session).
- Résultat_Semestre_FL(*#N_Inscription,#Code_Semestre,#Année,#Code_Niv_Filière*, Moyenne_Semestre,Nature_Session).
- Résultat_Semestre_SP(*#N_Inscription,#Code_Semestre,#Année,#Code_Niv_Spécialie*,Moyenne_Semestre,Nature_Session).
- Appartient_Domaine(*#N_Inscription,#Année,#Code_Niv_TC*,Type_Appartenance,
- Observation).
- Appartient_Filiere(*#N_Inscription,#Année,#Code_Niv_Filiere*,Type_Appartenance, Observation).
- Appartient_Spécialité(*#N_Inscription,#Année*,Type_Appartenance,Observation).

3. Description of the collected data:

3.1. Collection of data

In the work of [17], the data that was collected corresponds to the licence and master's cycles of one promotion only, namely that which started in 2009 and post-graduated in 2013. For the purpose of our work, this amount of data is not sufficient. Therefore, we had to collect more data and insert it into the source schema *résultat.*

For the sake of clarity, we refer to the existing data from the work [17] as: *promotion 1.*This promotion corresponds to the students who are had registered for the first study year in 2009-2010 with the following academic path:

- First year bachelor (License) 2009-2010.
- Second year bachelor (License) 2010-2011.
- Third year bachelor (License) 2011-2012.
- First year Master 2012-2013.
- Second year Master 2013-2014.

We asked for more data about subsequent promotion from the department of MI (Math and Informatics) and from the department of informatics. These data correspond to the results of students for two more promotions named *promotion 2* and *promotion 3,* such as

Promotion 2: correspond to the students who registered for the first study year in 2010-2011 with the following academic path:

- First year bachelor (License) 2010-2011.
- Second year bachelor (License) 2011-2012.
- Third year bachelor (License) 2012-2013.
- First year Master 2013-2014.
- Second year Master 2014-2015.

Promotion 3: correspond to the students who registered for the first study year in 2010-2011 with the following academic path:

- First year bachelor (License) 2011-2012.
- Second year bachelor (License) 2012-2013.
- Third year bachelor (License) 2013-2014.
- First year Master 2014-2015.
- Second year Master 2015-2016.

For the *promotion 3*, we have just collected the data for the bachelor cycle since their master cycle is not completed yet.

Table 3.3 summarizes the academic paths of the promotions

	2009 / 2010	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015
L1	P1	P2	P3			
L2		P1	P2	P3		
L3			P1	P2	P3	
M1 SIAD				P11	P21	P31
M2 SIAD					P11	P21
M1 R&S				P12	P22	P32
M2 R&S					P12	P22

Table 3.3: Promotions description in details

- P1: represents the promotion that started in the study year 2009/2010 and finished the license cycle in 2011/2012.
- P2: represents the promotion that started in the study year 2010/2011 and finished the license cycle in 2012/2013.
- P3: represents the promotion that started in the study year 2011/2012 and finished the license cycle in 2013/2014.
- P11: represents the promotion that started in the study year 2012/2013 and finished the master cycle in 2013/2014.
- P12: represents the promotion that started in the study year 2012/2013 and finished the master cycle in 2013/2014.
- P21: represents the promotion that started in the study year 2013/2014 and finished the master cycle in 2014/2015
- P22: represents the promotion that started in the study year 2013/2014 and finished the master cycle in 2014/2015
- P31: represents the promotion that started in the study year 2014/2015 and still not finished.
- P3: represents the promotion that started in the study year 2014/2015 and still not finished.

Let us note that we needed to collect these data so that we can detect profile evolution for the same level through different study years. Thus:

- P2, P3 are used to detect profile evolution in the license cycle.
- P21, P31 are used to detect profile evolution in the master cycle specialty SIAD (system information et aide a la decision) first year.
- P22, P32 are used to detect profile evolution in the master cycle specialty RES (Réseaux et sécurite) first year.

We obtained the data from the department of Mathematics and informatics and the department of Informatics in the form of paper sheets. The collected data deal with results of students per study element, study unit and per semester. Finally, we manually inserted the data into excel files.

3.2. Data organization

The following tables contain details about the collected data:

- The first table describes the domain, the study courses and the specialties used in the input.
- The others tables contain statistics about the number of modules, number of the study units, and number of students per domain, courses and specialties for the previously described promotions.

The organization of the domain, study courses, specialties are presented in following tables:

Domain	*Study course*	*Specialty*
Math and informatics (MI)	Informatics (INF)	Système d'information et aide à la décision (SIAD)
		Réseaux et sécurité (RES)

Table 3.4: Presentation of the domain, study course and specialties

Study year	Study level	domain	Number of Study elements	Number of Study units	Number of Students
2009-2010	First year	MI	18	6	392
2010-2011					626
2011-2012					832

Table 3.5: Data statistics about the domain level

Study year	Study level	Course	Number of Study elements	Number of Study units	Number of Students
2010-2011	Second year	INF	15	6	105
2011-2012					121
2012-2013					232

Table 3.6: Data statistics about study course level, second year

Study year	Study level	course	Number of Study elements	Number of Study units	Number of Students
2011-2012	Third year	INF	09	04	123
2012-2013					120
2013-2014					176

Table 2.7: Data statistics about study course level, third year

Study year	Study level	Specialty	Number of Study elements	Number of Study units	Number of Students
2012-2013	First year	SIAD	16	06	33
2013-2014					29
2014-2015					-
2012-2013		RES	16	06	24
2013-2014					24

2014-2015					-
2013-2014	Second year	SIAD	09	04	22
2014-2015					29
2015-2016					-
2013-2014		RES	09	04	15
2014-2015					24
2015-2016					-

Table 3.8: Data statistics about specialties levels

4. Database population

4.1. Tools

4.1.1. Oracle database

Oracle database is an object-relational database management system produced and marketed by Oracle Corporation. For the relational database, the data are stored in bi-dimensional tables that have columns and rows. Oracle database is a powerful database management system that offers a high performance for managing relational databases.

4.1.1.1. Available features in Oracle [18]

Scalability and Performance

- Concurrency
- Read Consistency
- Locking Mechanisms
- Quiesce Database
- RAC
- Portability

Manageability

- Self managing database
- OEM

- SQL*Plus
- ASM
- Scheduler
- Resource ManagerBackup and Recovery

High availability

Business Intelligence

- Data Warehousing
- ETL
- Materialized views
- Bitmap indexes
- Table compression
- Parallel Execution
- Analytic SQL
- OLAP
- Data mining
- Partitioning

Content Management

- XML
- LOB
- Oracle Text
- Oracle Ultra Search
- Oracle interMedia
- Oracle Spatial

Security

Data integrity/Triggers

- Integrity constraints
- Triggers

We used oracle database for creating the data source schema and for storing data.

4.1.2. Oracle SQL Developer

4.1.2.1. Overview

Oracle SQL Developer is a free integrated development environment that simplifies the development and management of Oracle Database in both traditional and Cloud deployments. SQL Developer offers complete end-to-end development of PL/SQL applications, a worksheet for running queries and scripts, a DBA console for managing the database, a reports interface, a complete data modeling solution, and a migration platform for moving your 3rd party databases to Oracle.[19]

4.1.2.2. Tasks

- Creating connections.
- Creating and exploring objects.
- Interrogating and updating data.
- Executing query.
- Creating and debugging PL/SQL blocks, procedures and functions.
- Executing and defining database reports.

4.1.2.3. Characteristics

- Open source.
- Runs on different platforms.
- Easy to navigate and manage data base users throw the graphic user interface (GUI).
- Uses driver JDBC Thin.
- Requires JRE 1.5 or later versions.

4.1.3. Our choice of versions

We worked with the following versions:

- Oracle Database Entreprise Edition 11g Release 2.
- Sqldeveloper 4.1.3.

4.2. Setting up of the work space

We created the source schemas described earlier in SQL Developer.

4.3. Input in the source schema

As we mentioned earlier, we obtained the students' results from the departments of Math and Informatics (MI) and the department of Informatics as paper sheets. We then had to digitize them into excels data, then we imported them using SQL Developer to the source database.

Figure 3.2 shows screenshots of the import activity into the table *RESULTAT_ELT_ENS*

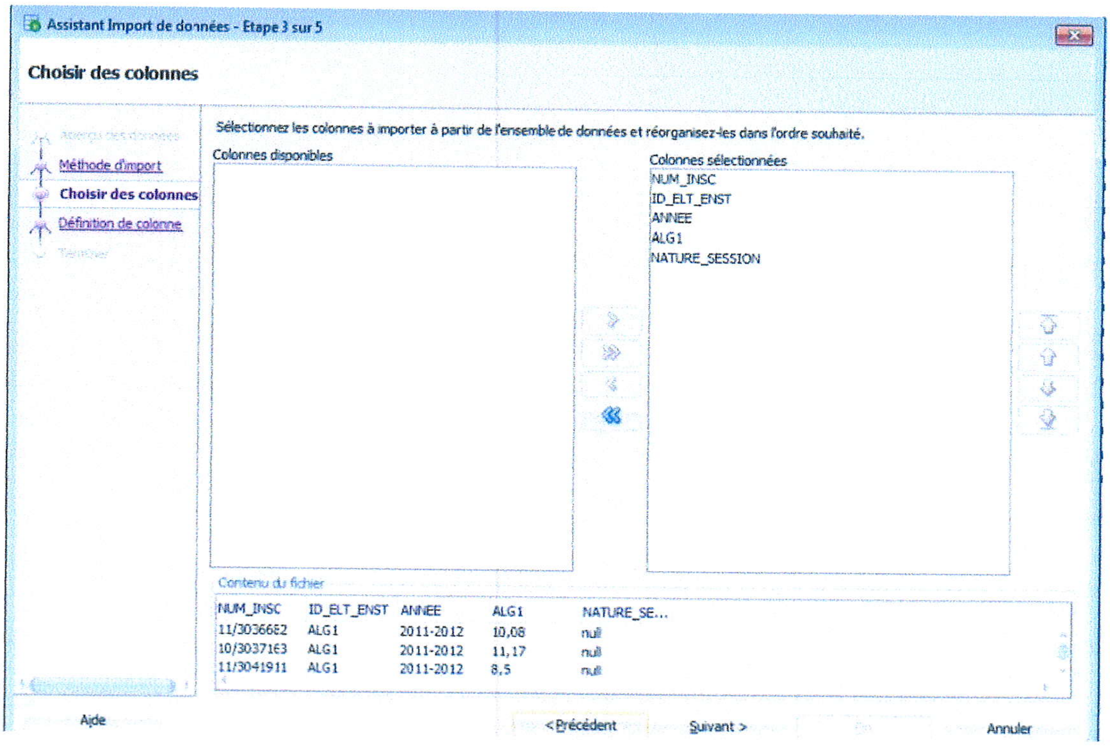


Figure 3.2: Data import screenshot

5. Conclusion

In this chapter we presented the source schema .We also described the existing data and the collected data that are needed for carrying out the data mining activity. Finally we presented the tools that we used, namely Oracle database, SQL Developer and Excel. The next chapter deals with data preparation for the data mining activity, and more precisely for clustering technique.

Chapter IV

Data Acquisition and Preparation

1. Introduction

In the previous chapter we presented the data source for data mining. In this chapter we present the activity of preparing the data so we can apply our data mining (clustering) on this data using educational data mining process (EDM).

2. Data treatment with the educational data mining process (EDM)

In this section we present data treatment according to our needs for clustering the students. This activity corresponds to the steps: raw data and pre-processing in the educational data mining process (EDM).

2.1. Problem understanding

The environment of our study is the University of Jijel, specifically the department of Math and informatics (MI) and the department of informatics.

Our objective is to detect students' skills and to shape them as profiles based on their grades in different modules at different levels during two cycles (Bachelor and Master). Besides, we aim at analyzing students' profile evolution for whole promotions or for individuals. To reach these goals, we will use cluster analysis, also known as clustering.

At the end, we expect to discover a list of profiles of students per level and per promotion. Each profile reflects the students' skill i.e. how good they are in some computer science fields, e.g. algorithms.

2.2. Data understanding

2.2.1. Collecting initial data

Data are obtained from the data source described previously containing grades for study modules and units for students.

2.2.2. Choosing and explaining the attributes

In any data mining activity, the variables describing the objects or individuals are not all needed. In our case, we are interested only in student marks in modules at different levels. Thus, we discard all the unnecessary attributes of the data sources like students semester credits, units credits study year credit, semester average and study year average and unit's average.

2.2.3. Checking the quality of Data

Before using the data, we have to check their quality (errors, null values, etc), so that it does not affect our analysis.

We completed a thorough analysis of the data, after which we encountered a problem with missing values. These missing values correspond to the missing marks of some students whom academic path is not regular. In other words, a regular academic path is one where a student steps normally from one level at a year to the next in the next later. This condition of a regular academic year is necessary in order to have the students' marks for a sequence of levels, starting from one academic year. Thus, students who repeat a level are just discarded from the data.

2.3. Data preparation

Data will need to be pre-processed, since it may come from several sources or have different formats and levels of hierarchy[8].the preparation includes the creation of new attributes, new tables ,etc.

2.3.1. Creation of tables

The collected data refers to different cycles, levels and different study years. Since we want modules obtained by students in different study levels we had to create partitions from the collected data according to the cycles and levels existed in the department of Math and Informatics and the department of Informatics figure 4.1 presents these partitions.

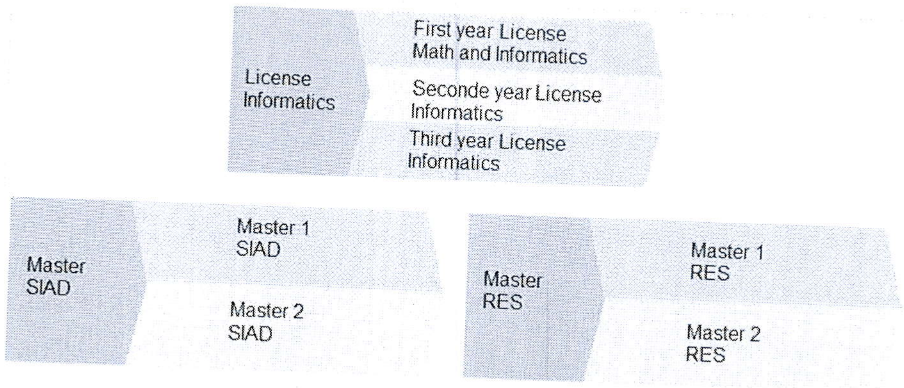


Figure 4.1: Informatics partitions

For each partition we create new tables with the attributes that are: the inscription number of the student, the study year and the names of modules in that partition.

Example for the first year License students Math and Informatics:

- Num_ins: inscription number of student.
- Year: study year.
- *Analyse1*: grade obtained in *analyse1*.
- *Analyse2*: grade obtained in *analyse2*.
- *Algebre1*: grade obtained in *algebre1*.
- *Algebre2*: grade obtained in *algebre2*.
- *Informatique1*: grade obtained in *informatique1*.
- *Infromatique2*: grade obtained in *informatique2*.
- *Mecanique*: grade obtained in *mecanique*.
- *Eléctricite*:grade obtained in *eléctricite*.
- *Theorie information*: grade obtained in *theorie information*.
- *Statistique descriptive*: grade obtained in *statistique descriptive*.
- *Structure machine*: grade obtained in *structure machine*.
- *Calcule formel*: grade obtained in *calcule formel*.
- *Englais1*: grade obtained in *anglais1*.
- *Englais2*: grade obtained in *angliais2*.
- *TP bureautique*: grade obtained in *TP bureautique*.
- *Technologie WEB*: grade obtained in *Technologie WEB*.

- *Technique expression*: grade obtained in *technique expression.*
- *Histoire de science*: grade obtained in *histoire de science*.

For the remaining levels we created the table as shown previously.

2.3.2. Data cleaning

In this activity we are forced to handle missing values corresponding to the grades obtained by the students as they are likely to affect our analysis; so we had to use only the data which refers to regular students in that study year and ignore students with missing values for grades.

2.3.3. Data integration

Since we apply clustering in different levels and for different promotions and the last presented tables contain all students' marks from different promotions we had to create for each level and promotion an integration table which holds only the student that are in the same promotion.

- For first year License promotion *2009-2010* we created table L1_2009-2010 by the following script:

  CREATE TABLE L1_2009-2010 AS SELECT (NUM_INS, ANALYSE1, ANALYSE2, ALGEBRE1, ALGEBRE2, INFORMATIQUE1, INFORMATIQUE2, MECANIQUE, ELECTRICITE, TP_BUREAUTIQUE, STRUCTURE_MACHINE, CALCULE_FORMEL, TECHNIQUE_EXPRESSION, HISTOIRE_SCIENCE, ANGLAIS1, ANGLAIS2, TECHNOLOGIE_WEB, THEORIE_INFORMATION, STATISTIQUE_DISCRIPTIFE)
  FROM L1 WHERE STUDY_YEAR='2009-2010';
- For first year License promotion *2010-2011* we created table L1_2010-2011 by the following script:

  CREATE TABLE L1_2010-2011 AS SELECT (NUM_INS, ANALYSE1, ANALYSE2, ALGEBRE1, ALGEBRE2, INFORMATIQUE1, INFORMATIQUE2, MECANIQUE, ELECTRICITE, TP_BUREAUTIQUE, STRUCTURE_MACHINE, CALCULE_FORMEL, TECHNIQUE_EXPRESSION, HISTOIRE_SCIENCE, ANGLAIS1, ANGLAIS2, TECHNOLOGIE_WEB, THEORIE_INFORMATION, STATISTIQUE_DISCRIPTIFE)

FROM L1 WHERE STUDY_YEAR='2010-2011';

- For first year License promotion *2011-2012* we created table L1_2011-2012 by the following script:

  CREATE TABLE L1_2011-2012 AS SELECT (NUM_INS, ANALYSE1, ANALYSE2, ALGEBRE1, ALGEBRE2, INFORMATIQUE1, INFORMATIQUE2, MECANIQUE, ELECTRICITE, TP_BUREAUTIQUE, STRUCTURE_MACHINE, CALCULE_FORMEL, TECHNIQUE_EXPRESSION, HISTOIRE_SCIENCE, ANGLAIS1, ANGLAIS2, TECHNOLOGIE_WEB, THEORIE_INFORMATION, STATISTIQUE_DISCRIPTIFE)

  FROM L1 WHERE STUDY_YEAR='2011-2012';

- For second year License promotion *2010-2011* we created table L2_2010-2011 by the following script:

  CREATE TABLE L2_2010-2011 AS SELECT (NUM_INS, SI, LOG_M, TL, AR_OR, AN, ANG3, GL, SE1, ANG4, ASD2, COGN, ASD1, BD, PL, PROB_S)

  FROM L2 WHERE STUDY_YEAR='2010-2011';

- For second year License promotion *2011-2012* we created table L2_2011-2012 by the following script:

  CREATE TABLE L2_2011-2012 AS SELECT (NUM_INS, SI, LOG_M, TL, AR_OR, AN, ANG3, GL, SE1, ANG4, ASD2, COGN, ASD1, BD, PL, PROB_S)

  FROM L2 WHERE STUDY_YEAR='2011-2012';

- For second year License promotion *2012-2013* we created table L2_2012-2013 by the following script:

  CREATE TABLE L2_2012-2013 AS SELECT (NUM_INS, SI, LOG_M, TL, AR_OR, AN, ANG3, GL, SE1, ANG4, ASD2, COGN, ASD1, BD, PL, PROB_S)

  FROM L2 WHERE STUDY_YEAR='2012-2013';

- For third year License promotion *2011-2012* we created table L3_2011-2012 by the following script:

  CREATE TABLE L3_2011-2012 AS SELECT (

  NUM_INS, PRGT, PRL, TG, DIG, RES, SE2, COGP, COMP, MASI)

  FROM L3 WHERE STUDY_YEAR='2011-2012';

- For third year License promotion *2012-2013* we created table L3_2011-2013 by the following script:

  CREATE TABLE L3_2012-2013 AS SELECT (

NUM_INS, PRGT, PRL, TG, DIG, RES, SE2, COGP, COMP, MASI)
FROM L3 WHERE STUDY_YEAR='2012-2013';

- For third year License promotion *2013-2014* we created table L3_2013-2014 by the following script:
  CREATE TABLE L3_2013-2014 AS SELECT (
  NUM_INS, PRGT, PRL, TG, DIG, RES, SE2, COGP, COMP, MASI)
  FROM L3 WHERE STUDY_YEAR='2013-2014';
- For first year Master SIAD promotion *2012-2013* we created table SIAD_M1_2012-2013 by the following script:
  CREATE TABLE SIAD_M1_2012-2013 AS SELECT (NUM_INS, ANGI_S, SED, AD, ED, MH, BDI, GOE, E_BUS, MAD, ANGII_S, SIA, IA, MMR, AOTE, OS, G_LA)
  FROM SIAD_M1 WHERE STUDY_YEAR='2012-2013';
- For first year Master SIAD promotion *2013-2014* we created table SIAD_M1_2013-2014 by the following script:
  CREATE TABLE SIAD_M1_2013-2014 AS SELECT (
  NUM_INS, ANGI_S, SED, AD, ED, MH, BDI, GOE, E_BUS, MAD, ANGII_S, SIA, IA, MMR, AOTE, OS, G_LA)
  FROM SIAD_M1 WHERE STUDY_YEAR='2013-2014';
- For second year Master SIAD *2013-2014* we created table SIAD_M2_2013-2014 by the following script:
  CREATE TABLE SIAD_M2_2013-2014 AS SELECT (
  NUM_INS, RT, AMD, SP, SSI, BDDA, SMNR_S, FD, ANGIII_S, PFE_S)
  FROM SIAD_M2 WHERE STUDY_YEAR='2013-2014';
- For second year Master SIAD *2014-2015* we created table SIAD_M2_2014-2015 by the following script:
  CREATE TABLE SIAD_M2_2014-2015 AS SELECT (
  NUM_INS, RT, AMD, SP, SSI, BDDA, SMNR_S, FD, ANGIII_S, PFE_S)
  FROM SIAD_M2 WHERE STUDY_YEAR='2014-2015';
- For first year Master RES *2012-2013* we created table RES_M1_2012-2013 by the following script:
  CREATE TABLE RES_M1_2012-2013 AS SELECT (

NUM_INS, R_PR, SR, ANGII_R, MCP, ANGI_R, A_F_ED, A_SII, DI, TI, PR_S, CRYPT, APG, BDA, R_AV, V_S, EV_P)
FROM RES_M1 WHERE STUDY_YEAR='2012-2013';

- For first year Master RES *2013-2014* we created table RES_M1_2013-2014 by the following script:
  CREATE TABLE RES_M1_2013-2014 AS SELECT (
  NUM_INS, R_PR, SR, ANGII_R, MCP, ANGI_R, A_F_ED, A_SII, DI, TI, PR_S, CRYPT, APG, BDA, R_AV, V_S, EV_P)
  FROM RES_M1 WHERE STUDY_YEAR='2013-2014';

- For second year Master RES *2013-2014* we created table RES_M2_2013-2014 by the following script:
  CREATE TABLE RES_M2_2013-2014 AS SELECT (
  NUM_INS, SMSI, TSC_S, S_WEB, SMNR_R, RFM, ASSR, SAP_P, PFE_R, ANGIII_R)
  FROM RES_M2 WHERE STUDY_YEAR='2013-2014';
- For second year Master RES *2014-2015* we created table RES_M2_2014-2015 by the following script:
  CREATE TABLE RES_M2_2014-2015 AS SELECT (
  NUM_INS, SMSI, TSC_S, S_WEB, SMNR_R, RFM, ASSR, SAP_P, PFE_R, ANGIII_R)
  FROM RES_M2 WHERE STUDY_YEAR='2014-2015';

3. Conclusion

In this chapter we set the objective of students data mining as that of *profiling bachelor and master students of Informatics at the University of Jijel*. As a data mining technique, we chose clustering since it naturally fits our goals. Later we defined the attributes that we need for clustering. Finally we created our own tables with the cleaned data described previously.

Chapter V

Implementation of Data Mining

1. Introduction

In the previous chapter we prepared the data to use it in data mining. This chapter is the next step in the educational data mining (EDM) process which is the EDM methods and models/patterns. We will present the two applications that we developed: the first one will be used by the data miner and the second one will be used by the analyst in order to visualize the results of the data mining activity.

2. Building models

In this section, we describe the tool that we used for data mining, namely WEKA. Next, we present the development environment for the application that is dedicated to the final users: the data miner user and the analyst user.

2.1. Presentation of the used tool

2.1.1. WEKA [20]

Waikato Environment for Knowledge Analysis (WEKA) is a free popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.

2.1.2. Advantages of WEKA

- Free availability under the GNU General Public License.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces (GUI).

2.1.3. The used version

We used version WEKA 3.6.9. Besides, since we are using WEKA which is an external tool to the database and as it requires a specific kind of data format, we had to extract the data from the tables that we prepared at the previous step in a CSV format and to feed WEKA with this format before carrying out data minin.

We extract the data from the tables in a CSV format using SQL Developer as shown below in figures 5.1. (step 1) and 5.2. (step 2).

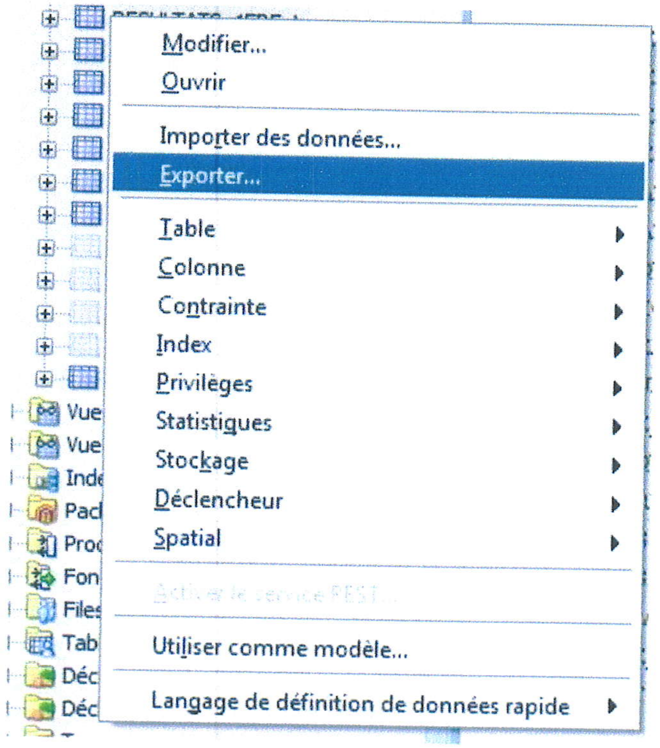


Figure 5.1 : CSV export 1

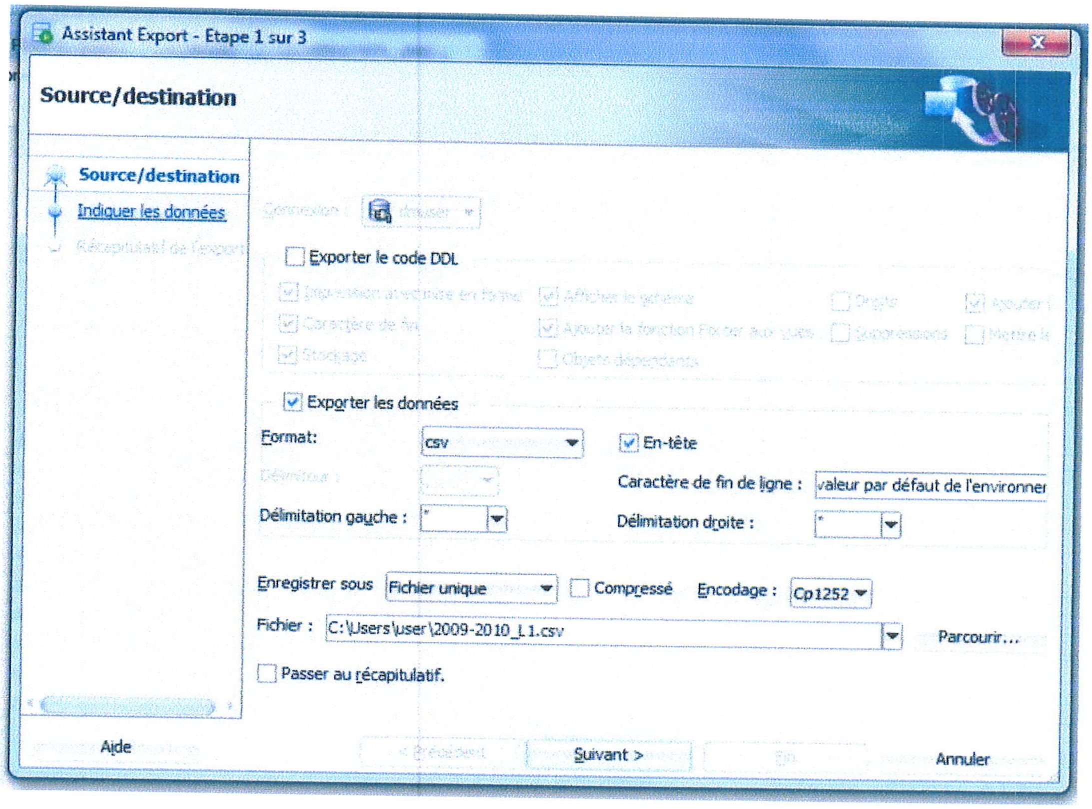


Figure 5.2: CSV export 2

Moreover, for the data to be self explained, we proposed a dataset labeling standard for the exported file: we start the study year like: 2009-2010 we follow it by a "_" then the study cycle and level, example: 2009-2010_L1.

2.2. Preparation of the work space

We installed WEKA 3.9.6 then we launched WEKA Explorer.

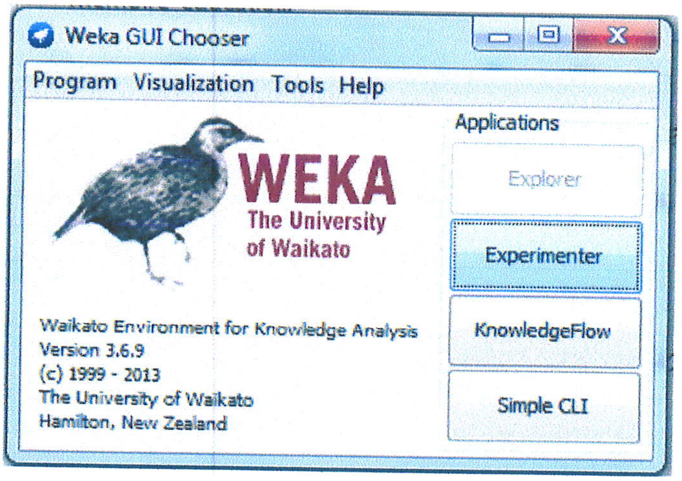


Figure 5.1 : WEKA GUI Chooser

2.3. Selecting the dataset

After having exported the needed data for data mining and opened the dataset in the CSV format; we selected the attributes by default, i.e. we used all the attributes. Let us recall that in our case, a data mining attribute corresponds to a course (module). The dataset selection step is shown in figure 5.4.

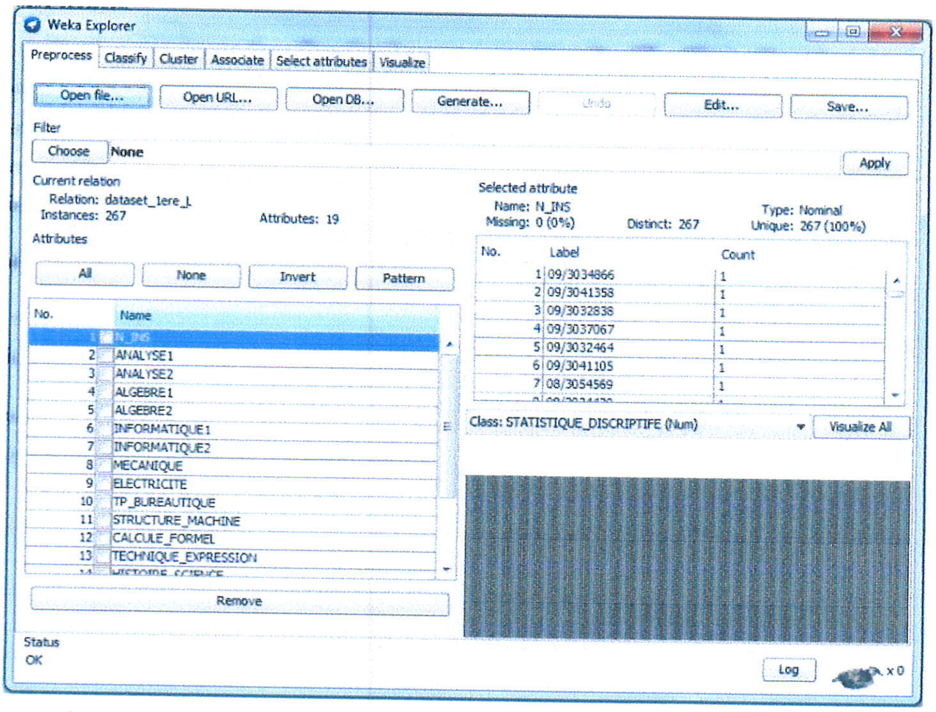


Figure 5.4 : WEKA Explorer

2.4. Algorithm and parameters selection

The next step is the choice of the clustering algorithm. In our case we chose the k-means algorithm with different K values. For the K values we used rule of thumb it is a simple method. This method can by apply to any type of data set:

$K=\sqrt{N/2}$ where n is the number of individuals. [21]

As an example, promotion of 2009-2010_L1, we have 268 students. Thus, the number of cluster is given by $K=\sqrt{268/2} \approx 11$ clusters

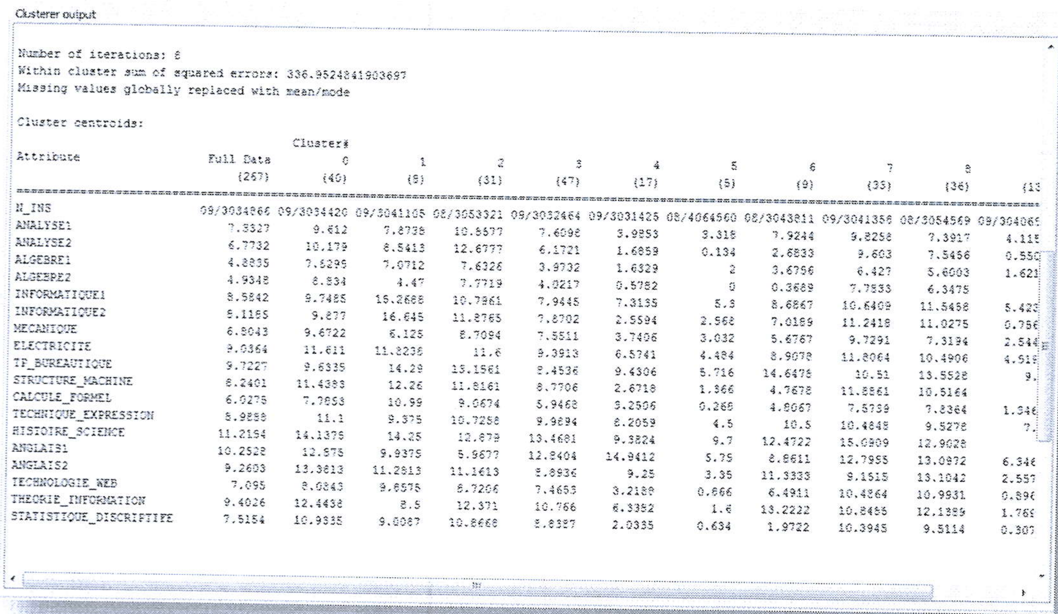
The other parameters of the clustering activity are set as follows:

- Distance: Euclidean distance.
- Max number of iterations: By default 500.
- Seed: by default 10.

2.5. Execution of the model

After having selected the dataset and the algorithm parameters, we started the data mining process which corresponds to building the clusters.

Figure 5.5 shows the results obtained after applying the clustering algorithm on promotion *2009-2010 first year bachelor* students.



Clusterer output

Number of iterations: 6
Within cluster sum of squared errors: 336.9524841903697
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (267)	Cluster# 0 (40)	1 (8)	2 (31)	3 (47)	4 (17)	5 (5)	6 (9)	7 (33)	8 (36)	(13
N_INS	09/3034866	09/3034420	09/3041105	08/3053321	09/3032464	09/3031425	08/4064560	08/3043811	09/3041356	08/3054569	09/304065
ANALYSE1	7.3327	9.612	7.8738	10.8677	7.6098	3.9853	3.318	7.9244	9.8258	7.3917	4.115
ANALYSE2	6.7732	10.179	8.5413	12.6777	6.1721	1.6859	0.134	2.6833	9.603	7.5456	0.550
ALGEBRE1	4.8835	7.6299	7.0712	7.6326	3.9732	1.6329	2	3.6796	6.427	5.6003	1.621
ALGEBRE2	4.9348	8.834	4.47	7.7719	4.0217	0.5782	0	0.3689	7.7833	6.3475	
INFORMATIQUE1	8.5842	9.7485	15.2688	10.7961	7.9445	7.3135	5.3	8.6867	10.6409	11.5458	5.423
INFORMATIQUE2	8.1185	9.877	16.645	11.8765	7.8702	2.5594	2.568	7.0189	11.2418	11.0275	0.756
MECANIQUE	6.8043	9.6722	6.125	8.7094	7.5511	3.7406	3.032	5.6767	9.7291	7.3194	2.544
ELECTRICITE	9.0364	11.611	11.8238	11.6	9.3913	6.5741	4.484	8.9078	11.8064	10.4906	4.919
TP_BUREAUTIQUE	9.7227	9.6339	14.29	13.1561	8.4536	9.4306	5.716	14.6478	10.51	13.5528	9.
STRUCTURE_MACHINE	8.2401	11.4383	12.26	11.8161	8.7706	2.6718	1.366	4.7678	11.8861	10.5164	
CALCULE_FORMEL	6.0278	7.7853	10.99	9.0674	5.9468	3.2506	0.268	4.8067	7.5739	7.8364	1.346
TECHNIQUE_EXPRESSION	8.9888	11.1	9.375	10.7258	9.9894	8.2059	4.5	10.5	10.4848	9.5278	7.
HISTOIRE_SCIENCE	11.2194	14.1376	14.25	12.879	13.4681	9.3824	9.7	12.4722	15.0909	12.9028	
ANGLAIS1	10.2528	12.875	9.9379	5.9677	12.8404	14.9412	5.79	8.8611	12.7955	13.0972	6.346
ANGLAIS2	9.2603	13.3813	11.2813	11.1613	8.8936	9.25	3.35	11.3333	9.1515	13.1042	2.557
TECHNOLOGIE_WEB	7.095	8.0843	9.8578	8.7206	7.4653	3.2189	0.866	6.4911	10.4864	10.9931	0.896
THEORIE_INFORMATION	9.4026	12.4438	8.5	12.371	10.766	8.3352	1.6	13.2222	10.8455	12.1389	1.769
STATISTIQUE_DISCRIPTIFE	7.5154	10.9335	9.0087	10.8668	8.8327	2.0335	0.634	1.9722	10.3945	9.5114	0.307

Figure 5.5: Clustering results

After executing the clustering we obtained 11 clusters where each cluster is defined by a centroid that is the mean values of the modules (attributes); we call a cluster a profile. Since a profile contains the marks of student in the modules, it is difficult to give an interpretation of a profile or what does this profile describes.

Therefore we proposed a simple algorithm to detect and label the profiles. Our algorithm ranks the profiles according to each module then detect the best ranked modules in the cluster and label the cluster based on these module ranks.

5.2.1. Pseudo algorithm for profile labeling

The algorithm below describes the labeling of profiles.

***Algorithm** labeling*

```
RanksMatrix: matrix of integers
CL: Vector of Clusters
D: Vector of Dimentions
Labels: Vector of Strings
   For each D_i ∈ D do
       For each CL_j ∈ CL do
          RanksMatrix(i,j):=rank of the cluster j in the dimension i
       End For
   End For
   Labels ← ϕ
   For i: =1 to RanksMatrix.columns do
          -- Search the minimum rank for the column i
          MinRank: = the minimum rank for the column i
        string: =""
          For j: =1 to RanksMatrix.raws do
              If (RanksMatrix(j,i)=minRank) then
                String+= MinRank +" in"+ D_j.name
              End If
       End For
       If (Labels not contains string) then
           Add string to Labels
       Else
           While (Labels contains string) do
             --Search the next minimum rank for column i
             NextMinRank ← the next minimum rank value
             If (RanksMatrix(j,i)=NextMinRank) then
                 String+=NextMinRank +" in"+ D_j.name
            End while
            Add string to Labels
       End If

 End For

End Algorithm.
```

2.5.2. Example of profile labeling

Let us suppose that we obtained three profiles based on three modules, as shown in table 5.1

modules	Profile 1	Profile 2	Profile 3
Module 1	10	14	7
Module 2	16	11	9
Module 3	13	13	6

Table 5.1 : Modules values

After ranking the profiles, we obtain the following ranks table 5.2:

modules	Profile 1	Profile 2	Profile 3
Module 1	2	1	3
Module 2	1	2	3
Module 3	1	1	2

Table 5.2: Ranks table

- Now for profile 1 the Min Rank is 1 and the corresponding Modules are Module 2, Module 3. Thus, the profile name would be" *1 in Module 2 and 1 in Module 3*".
- For profile 2 the Min Rank is 1 and the corresponding Modules are Module 1, Module 3. Thus, the profile name would be"*1 in Module 1 and 1 in Module 3*".
- For profile 1 the Min Rank is 2 and the corresponding Module is Module 3.Thus, the profile name would be" *2 in Module 3*".

We apply the labeling algorithm to the results obtained previously on promotion *2009-2010 first year bachelor* students. We obtained the following profiles labels:

profile	Label	Number of students
Profile 1	1 in algebre1 and 1 in algebre2 and 1 in technique expression and 1 in anglais2 and 1 in statistique descriptif	40
Profile 2	1 in informatique1 and 1 in informatique2 and 1 in electricite and 1 in structure machine and 1 in calcule formel	8
Profile 3	1 in analyse1 and 1 in analyse2	31
Profile 4	4 in mecanique and 4 in histoire de science and 4 in anglais1	47
Profile 5	1 in anglais1	17
Profile 6	8 in algebre1 and 8 in informatique2 and 8 in histoire de science	5
Profile 7	1 in TP Bureautique and 1 in theorie information	9
Profile 8	1 in mecanique and 1 in histoire de science	33
Profile 9	1 in technologie web	36
Profile 10	7 in TP Bureautique	13
Profile 11	11 in analyse1 and 11 in analyse2 and 11 in algebre1 and 11 in algebre2 and 11 in informatique1 and 11 in informatique2 and 11 in mecanique and 11 in electricite and 11 in TP Bureautique and 11 in structure machine and 11 in calculi formel and 11 in technologie web	28

Table 5.3: Profiling results

3. Evaluation of the results

After the implementation of data mining, we had to evaluate the quality of the obtained results before jumping to the next step in the EDM process which is the interpretation.

The profiles obtained previously do not precisely meet our goal which is profiling informatics students. This is due the existence of some courses that do not describe an informatics student, such as: anglais1,2, mecanique, TP Bureautique,etc. Therefore we had to execute the model again without these modules and we obtained the following results shown in the table 5.4:

profile	Label	Number of students
Profile 1	1 in analyse2 and 1 in algebre2 and 1 in theorie information and 1 in statistique discriptif	19
Profile 2	1 in algebre1 and 1 in informatique1 and 1 in informatique2 and 1 in structure machine and 1 in calcule formel	9
Profile 3	2 in algebre2	53
Profile 4	5 in analyse1 and 5 in theorie information	47
Profile 5	6 in informatique2	22
Profile 6	5 in informatique1	11
Profile 7	3 in theorie information	12
Profile 8	1 in analyse1	28
Profile 9	2 in informatique1 and 2 in infromatique2 and 2 in	22

	structure machine and 2 in calcule formel and 2 in theorie information	
Profile 10	10 in analyse1 and 10 in analyse2 and 10 in algebre1 and 10 in algebre2 and 10 in informatique1 and 10 in informatique2 and 10 in and 10 in structure machine and 10 in calcule formel and 10 in theorie information and 10 in statistique descriptif	18
Profile 11	11 in analyse1 and 11 in analyse2 and 11 in algebre1 and 11 in algebre2 and 11 in informatique1 and 11 in informatique2 and 11 in and 11 in structure machine and 11 in calcule formel and 11 in theorie information and 11 in statistique descriptif	26

Table 5.4: Profiling results after evaluation

3. Presentation of the applications

3.1. The need of applications

The creation of the model is not the end of our work: we needed to let the users manipulate and create new models for more flexibility and to better understand the created models.

We have developed two applications. The first one is dedicated to the creation of the models and the second one is meant for visualizing and manipulating the models. We named the first one *Data Miner* and the second *Analyst*.

3.2. The development environment

3.2.1. Eclipse

3.2.1.1. Presentation

Eclipse is an open source integrated development environment (IDE) developed by the eclipse foundation that supports many languages besides JAVA such as: C, C++, PYTHON, PHP, HTML, etc. Eclipse supports plug-ins that allows developer to develop with other languages. Eclipse is available under Linux, Mac and Windows.

3.2.1.2. Characteristics

- Open source.
- Plug-ins for different languages.
- Improved code completion and quick fix support.
- Static analysis of Java code.
- Easy window building with Windows Builder.

3.2.1.3. The used version

We used the version KEPLER 2.0.0.2 of eclipse.

3.2.2. Functions of the applications

a. Application "Data miner"

The Data Miner application is used for creating and saving models. First, the user selects a pre-prepared CSV dataset created by a data mining engineer.

The following figure shows the Data miner GUI.

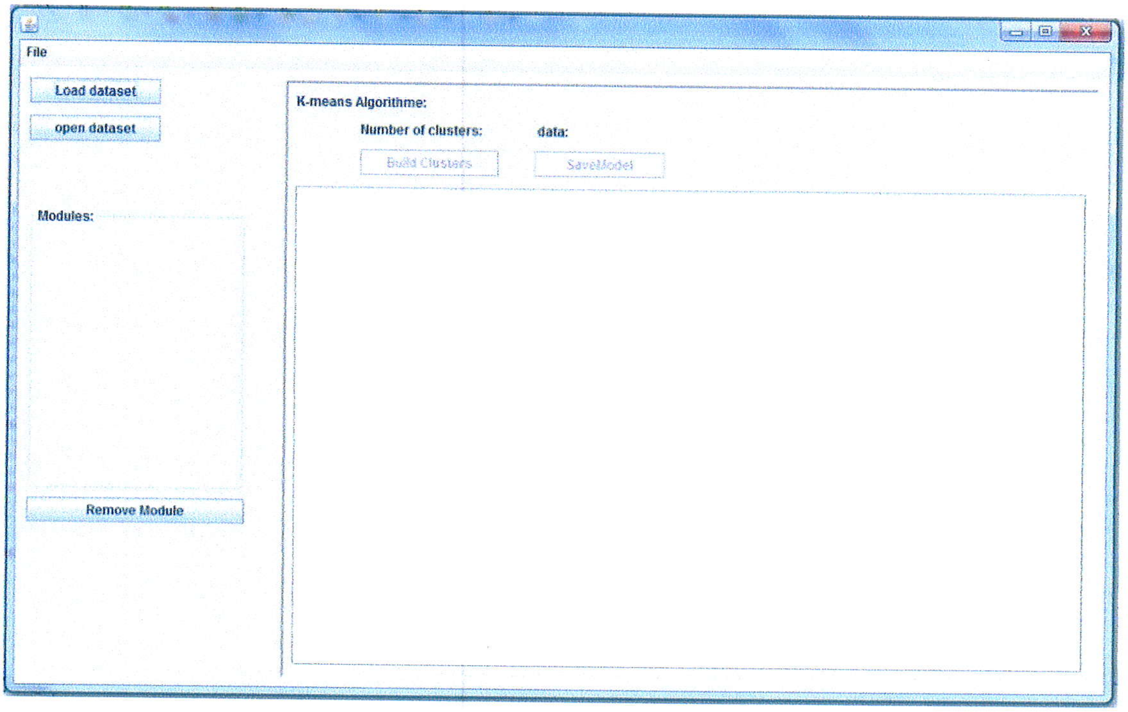


Figure 5.6: Application data miner

The main functions of the application are carried out using the following buttons

- **Load dataset Button** → when clicked a dialog frame appears and the user select the pre-prepared CSV datasets and loads it into the application.
- **Open dataset Button** →when clicked another dialog frame and the user selects a loaded dataset to apply the clustering to.
- **Remove Module Button** → to discard some unwanted modules for the profiling.
- **Build Clusters Button** → to create the clusters.
- **Save Model Button** →to save the clustering model.

The next figure illustrates an example:

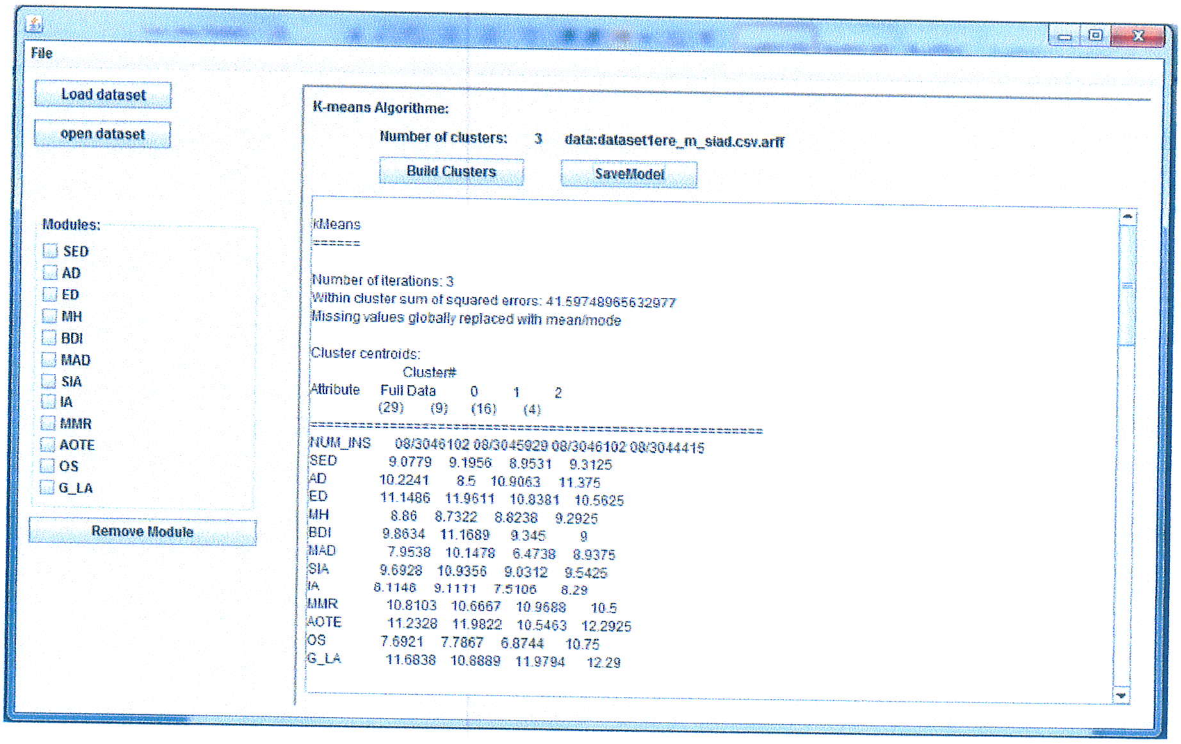


Figure 5.7: Data miner application example

b. Application "*Analyst*"

Application "*Analyst*" is used to visualize the saved models created by the Data Miner application. Also it can trace the evolution of profiles through different study years and visualize the global profile of any given student composed of the profiles of his academic path, for the bachelor or master cycle or for both.

Figure 5.8 show the graphical user interface for application "*Analyst*".

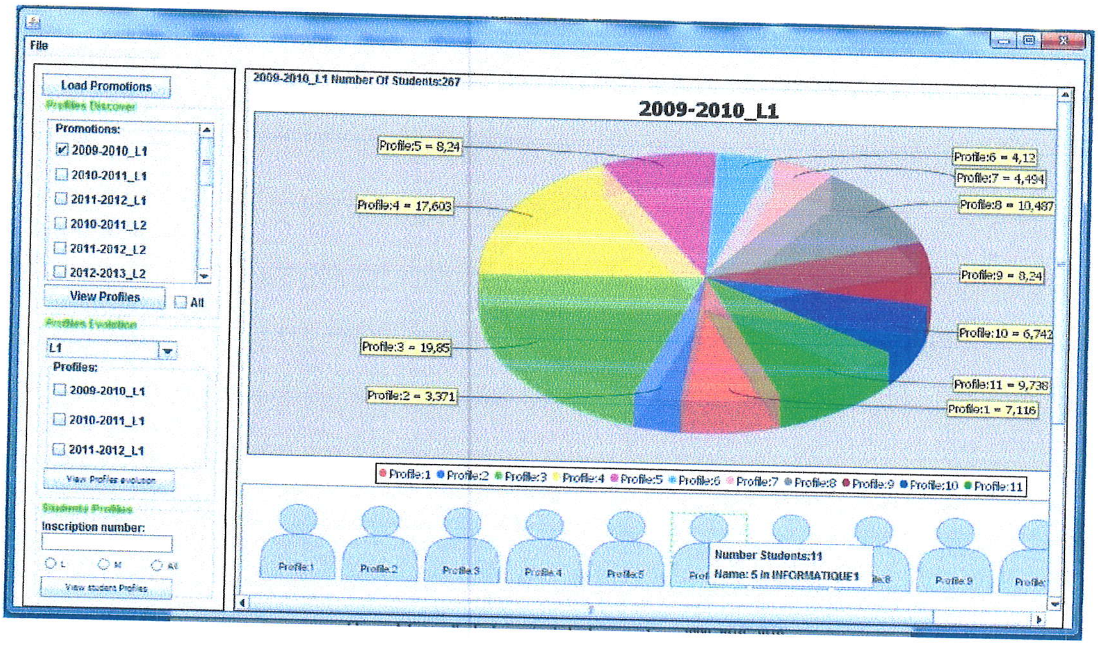


Figure 5.8: Profiles Discovery

The main functions of the application are carried out using the following buttons

- **View Profiles Button →** to view existing profiles from the chosen promotions.
- **View Profiles evolution Button →** to detect and visualize profile evolution through different years. The next figure (5.9) shows the result of analyzing profiles evolution in *the first year bachelor* for promotions *2009-2010*; *2010-2011*and *2012-2012*.

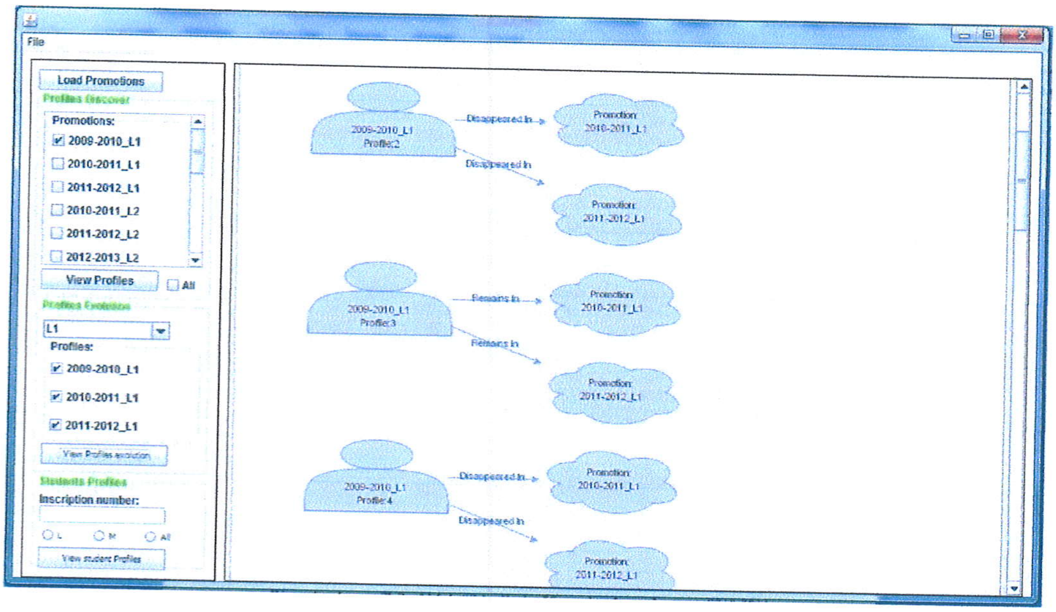


Figure 5.9: Profiles evolution

We notice that profile3 which is "*2 in Algebre2*" remains the study years: *2010-2011* and *2011-2012*.But profile 2 in promotion *2009-2010_L1* disappeared in both *2010-2011* and *2011-2012* promotions. Also the profile "*3 in Informatics 1 and 3 Informatics2*" remains in the promotion *2011-2012*.

- **View Student Profiles Button** → for showing the profiles obtained by a given student in both study cycles. Figure 5.10 and Figure 5.11 shows the profiles of a student in the bachelor and master cycles.

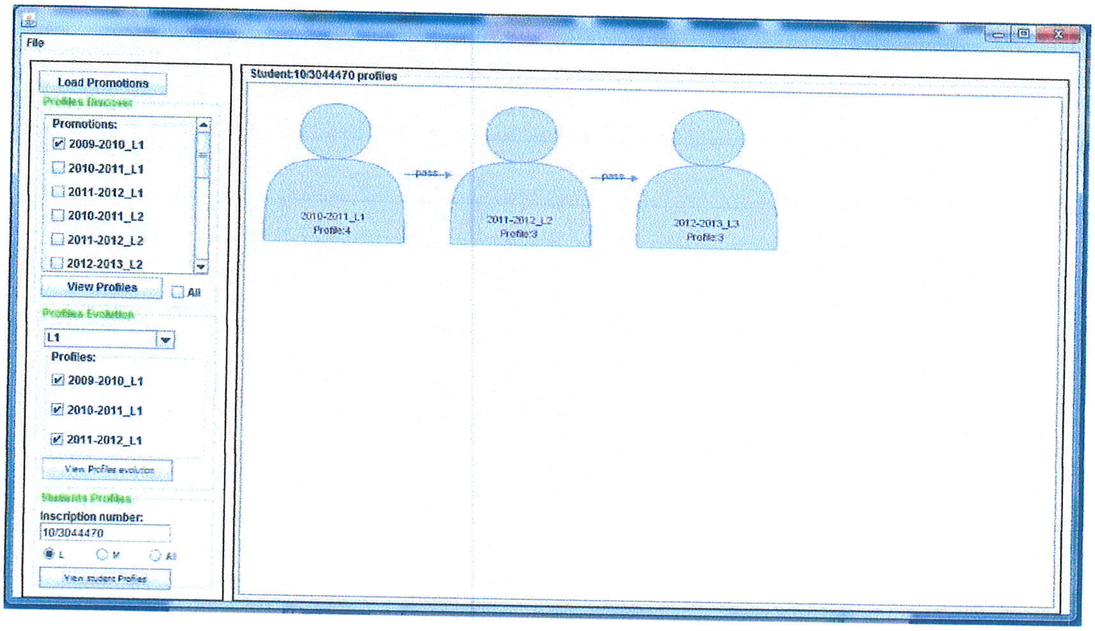


Figure 5.10: Student bachelor profiles

The profiles corresponds to " *1 in theorie information*" , "*1 in analyse*", "*1 in TG and 1 in RES and 1 in COGP and 1 in MASI*"

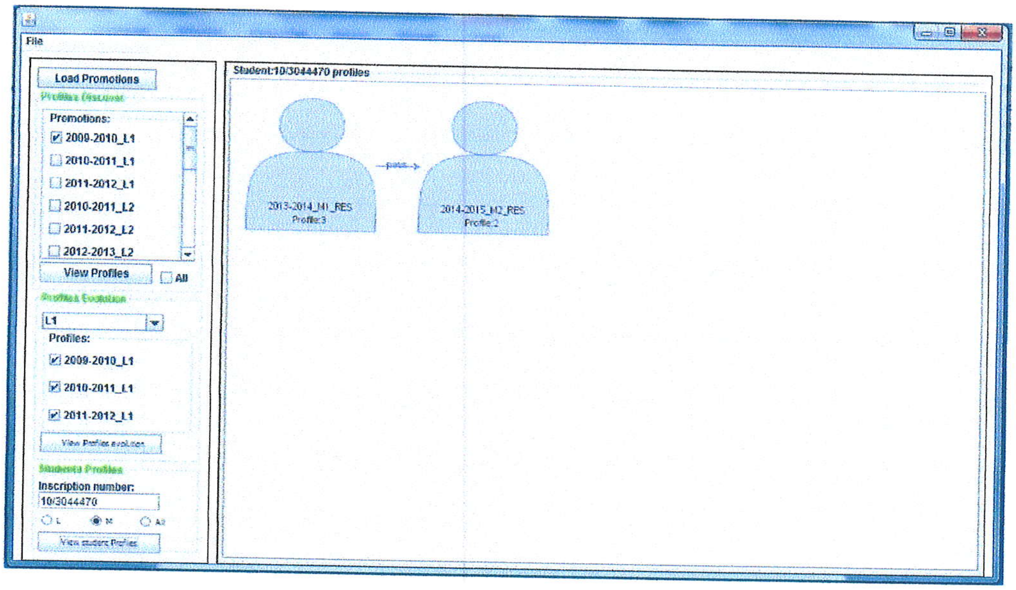


Figure 5.11 : Student master profiles

The profiles correspond to *"1 in Méthodologies de conduite de projets and 1 in Analyse, fouille et extraction de donnéess and 1 in Cryptographie and 1 in Bases de données avancées", " 1 in projet fin etude".*

4. Conclusion

In this chapter we presented the implementation of data mining, the used tools such as WEKA for clustering. We chose K-means algorithm with the number of K=$\sqrt{N/2}$ which is the rule of thumb for clustering. We also illustrated *"Data miner"* application and *"Analyst"* application and some results.

General conclusion

In this master thesis, we presented the use of a data mining technique, namely clustering in the context of higher education. The objective of the project was the detection and analysis of students' profiles in the department of Informatics.

First, we presented the state of art of data mining in general and the educational data mining in particular. Later, we presented the population of the source database with the collected data from the departments. After the population, we prepared the data so that we could apply the clustering technique. On these data we applied the K-means algorithm.

The obtained results group the students into students' profiles based on their marks in modules at different levels and cycles.

This work is the result of long months of effort. Starting off from the step of bibliography search, we noticed that data mining is a wide domain that allows the extraction of knowledge from different domains and areas.

We believe having reached the goals of this project, which is the detection and analysis of students' profiles from the students data obtained from the departments of Math and Informatics and the department of Informatics.

In addition, data mining is reach of technique applicable in domain of higher education with other perspectives.

Regarding the perspectives, we can use decision trees to predict the profile of current students based on existing data or we can combine clustering and association rules to detect certain links between profiles.

Moreover, since our work focuses on student profiles in both bachelor and master cycles at the department of Math and Informatics and the Department of Informatics, we can think of expanding our work to other departments and add more criteria such as the baccalaureate marks to detect student profiles at the baccalaureate stage and see if they affect students profiles' along the academic path at the University.

Bibliography

[1] http://www.usthb.dz/en/spip.php?rubrique41

[2] http://eco.univ-setif.dz/en/article.php?id=575

[3] Osmar R. Zaïane, 1999 CMPUT690 Principles of Knowledge Discovery in Databases

[4] Data Mining: Past, Present and Future FRANS COENEN

[5] Agrawal, R.; Imieliński, T.; Swami, A. "Mining association rules between sets of items in large databases», 1993

[6] Institute of Innovation in Technology &Management IITM Journal of Information Technology

[7] www.educationaldatamining.org

[8] Educational data mining/fr Édition : Michel C. Desmarais, Polytechnique Montréal Contribution : Ryan S.J.d. Baker, Worcester Polytechnic Institute Adaptation : Nicolas Balacheff, LIG, Grenoble

[9] C. Romero, S. Ventura. Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 40(6), 601-618, 2010

[10] Huebner, Richard A. "A survey of educational data-mining research"

[11] Educational Data Mining and Learning Analytics: differences, similarities, and time evolution Laura Calvet linane and angel Alejandro Juan Pérez

[12] S. Chen and X. Liu, "An integrated approach for modeling learning patterns of students in web-based instruction: A cognitive style perspective," ACM Trans. Comput. Interact., vol. 15, no. 1, 2008.

[13] A. Bovo, S. Sanchez, O. Heguy, and Y. Duthen, "Clustering moodle data as a tool for profiling students," in Proc. 2013 Second Int. Conf. E-Learning E-Technologies Educ., Sep. 2013, pp. 121-126.

[14] H. Grob, F. Bensberg, and F. Kaderali, "Controlling open source intermediaries-a web log mining approach," IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews, vol. 1, pp. 233-242, 2004,

[15] M. Pechenizkiy, T. Calders, E. Vasilyeva, and P. De Bra, "Mining the student assessment data: Lessons drawn from a small scale case study," EDM, 2008.

[16] Y. Psaromiligkos, M. Orfanidou, C. Kytagias, and E. Zafiri, "Mining log data for the analysis of learners" behaviour in web-based learning management systems," Oper. Res.,vol. 11, no. 2, pp. 187-200, Jan. 2009.

[17] Hadjer Meriai Amel Meledjem. Fouille de données dans un entrepôt de données dédié à l'évaluation des activités pédagogiques

[18] http://www.adp-gmbh.ch/ora/misc/features.html

[19]http://www.oracle.com/technetwork/developer-tools/sql-developer/overview/index-097090.html

[20] https://en.wikipedia.org/wiki/Weka_(machine_learning)

[21] Review on determining number of Cluster in K-Means Clustering.Trupi M.Kodinariva, Dr. Prashant R. Makwana.