2013

Datamining Report

Building a hypercube

AMBOUNDA ZEKPA Apoté - HURIER Médéric – PY- PATINEC Armand – THEPSOUVANH Phongphet

MIAGE NANCY
01/01/2013

Table of Contents

I. Classifications: K-Means Method

1. Classifications

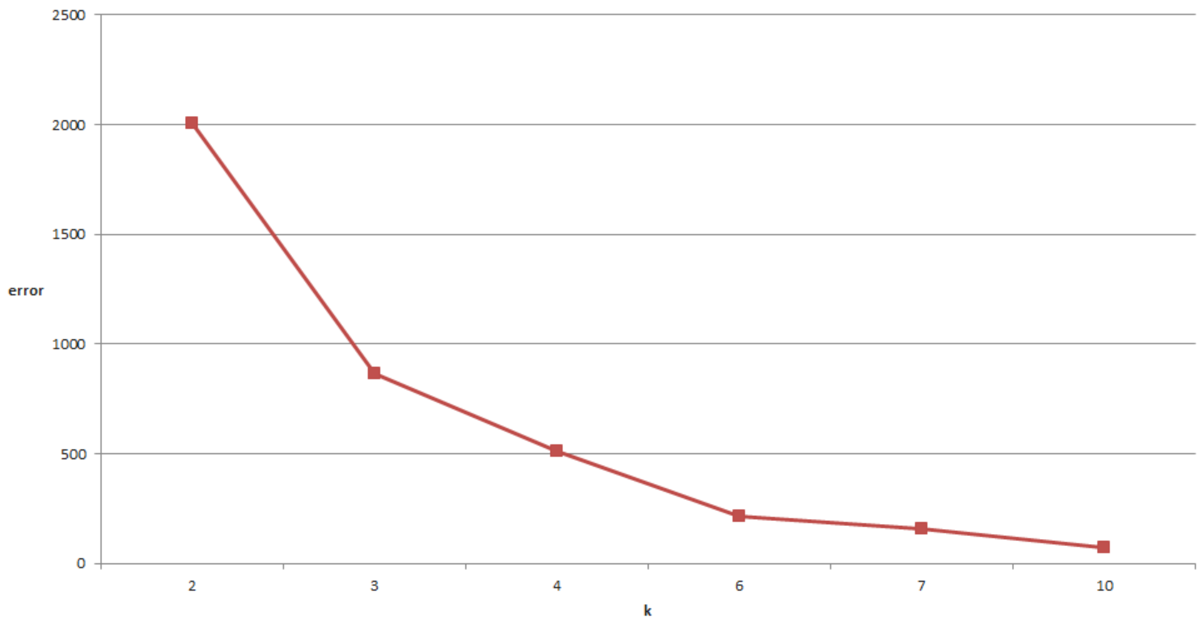
For each classification, we use the method we explain in part 3.3 of this document. Each part show:

- the optimal number of group
- for a fixed number of group, the size of clusters
- when we obtain relevant cases, the characteristics of a cluster

A. Students by final note

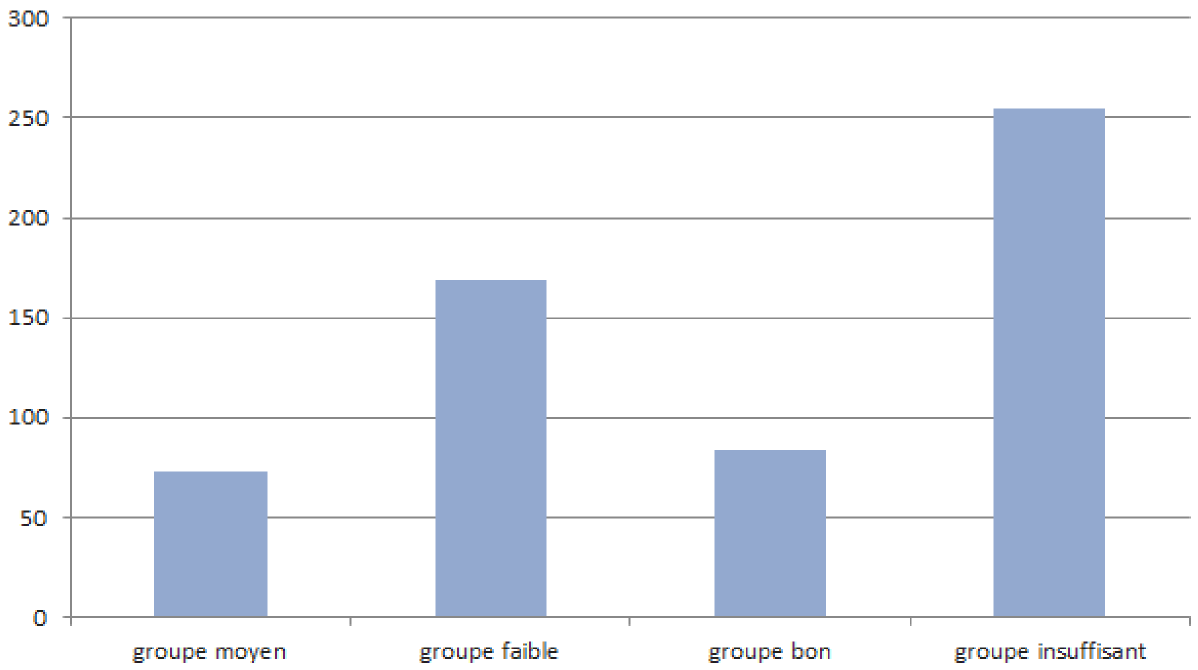
1. Optimal number of group

In this figure, we work on 581 students that we classify by their final note:



There is no step to help us determine the optimal number of group. We fixe it to 4 because it’s a layer that determine if a student if admin and his mention.

2. Clusters



We see that the cluster 1 and 3 are twice the size of the cluster 0 and 2.

3. Characteristics

Showing some representative individual show us the following properties:

- Cluster 0 contains average students (marks between 9 and 12)
- Cluster 1 contains weak students (marks above 6)
- Cluster 2 contains good student (marks above 13)
- Cluster 3 contains below average students (between 6 and 9)

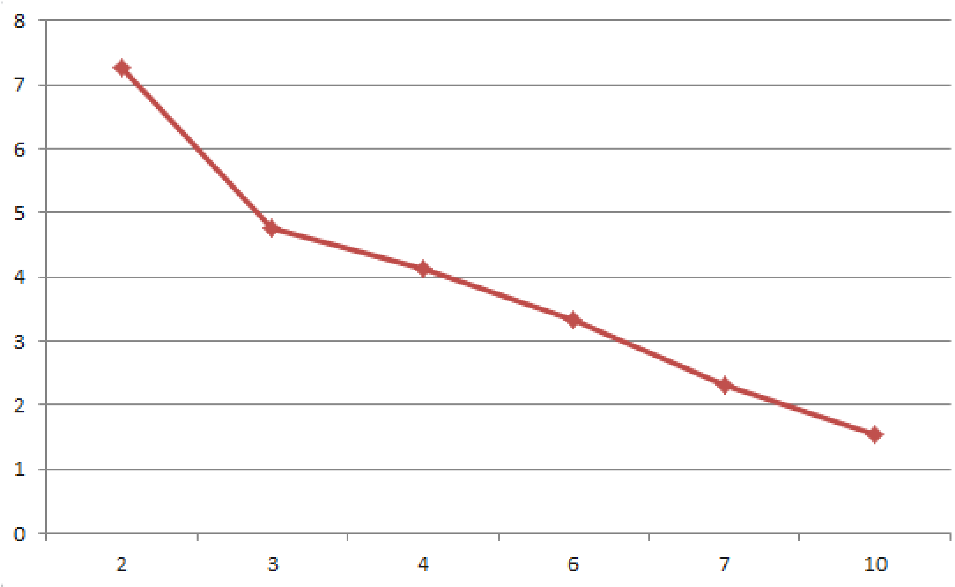
We think it’s a good classification of the students; we have a majority of the student average or below average. **Il follow a normal distribution**.

B. Students by modules marks

1. Optimal number of group

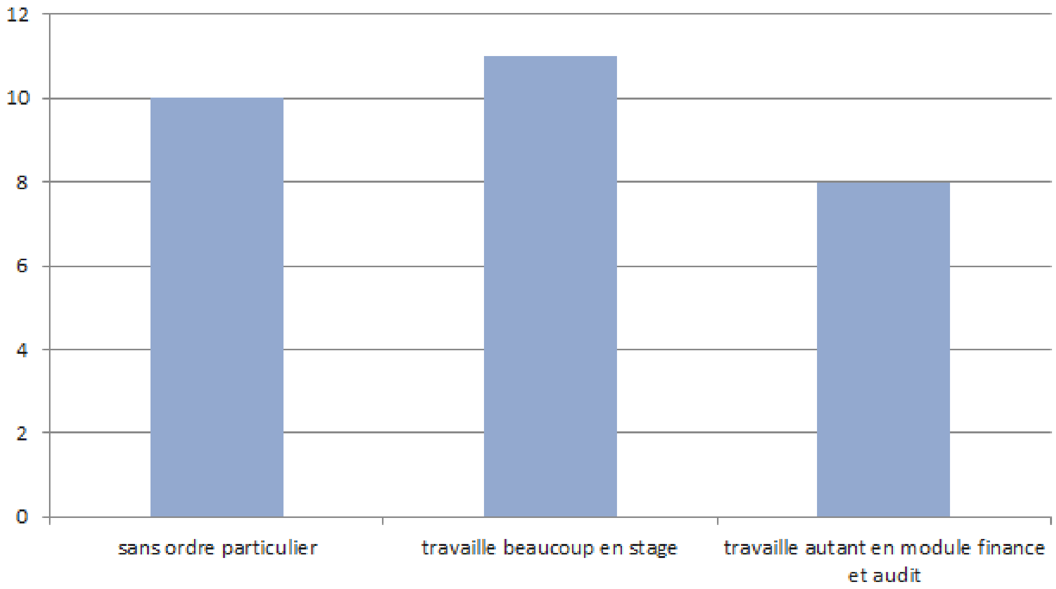
There are many categories of modules, so we decide to work only of our 29 students from Morocco.

Here is the curve of optimal clustering.



We find that the error step is between 2 and 3, so we decice a fixed value of 3 clusters.

2. Clusters



Clusters show that there is an equal repartition. We find interresting caracteristics.

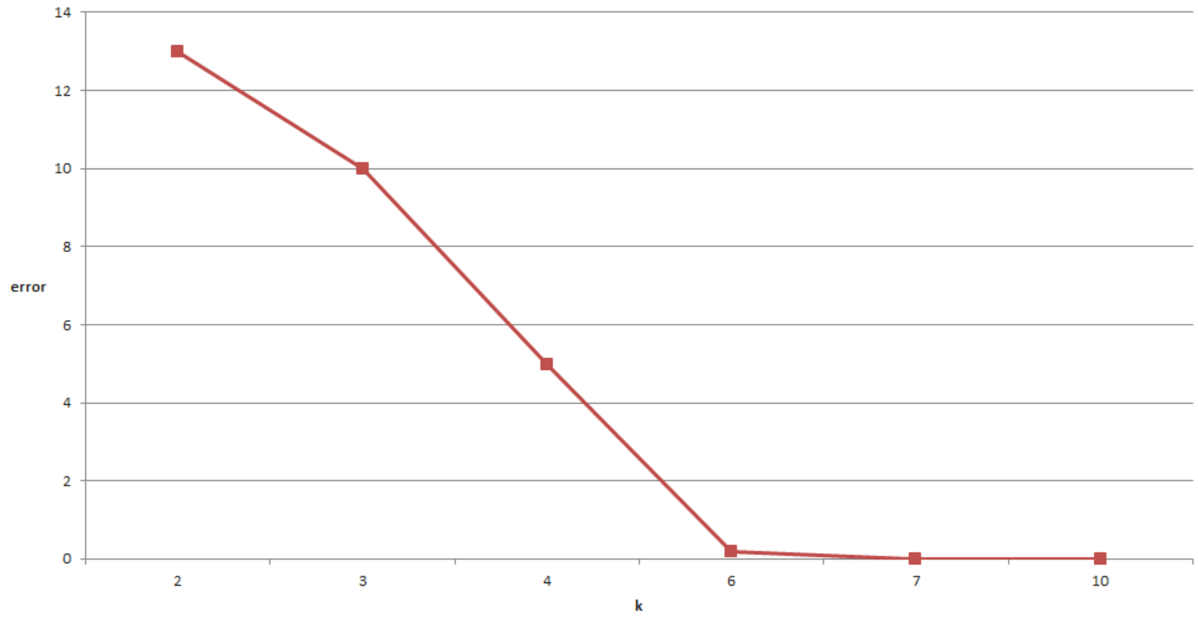
3. Characteristics

- Cluster 0 shows students which have done a good stage (mark > 12)
- Cluster 1 shows students with good grade in management and finance.
- Cluster 2 shows no particular order, we think it's the student that we could not sort

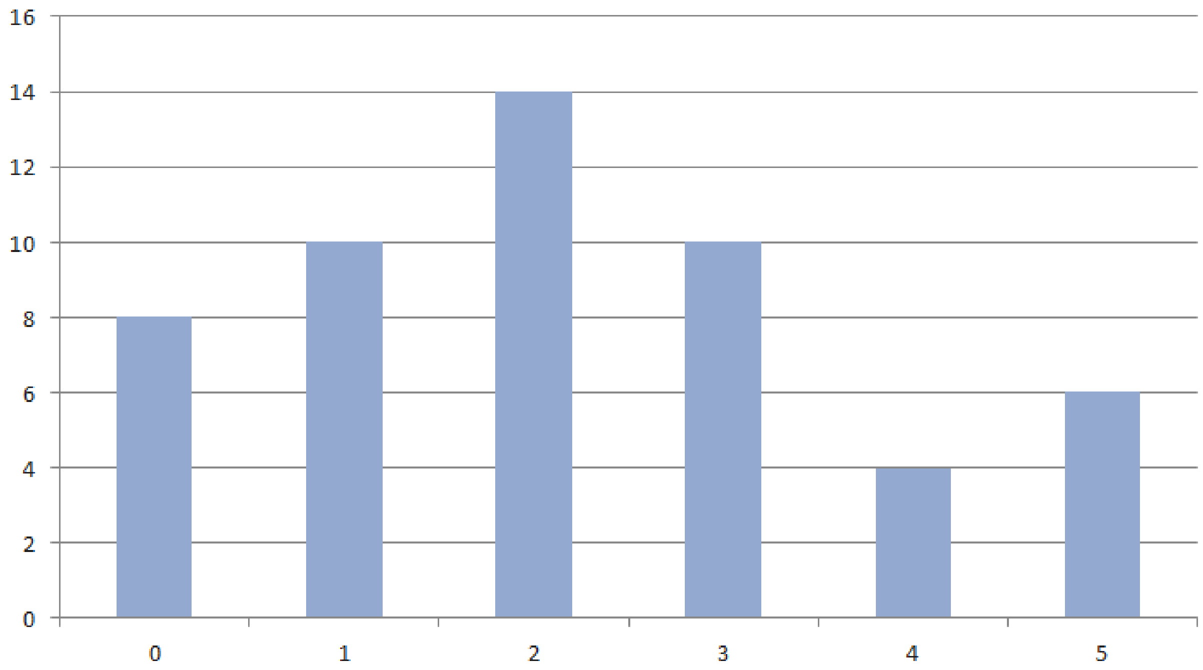
B. Subjects

1. Optimal number of group

For our classification by subject, we find an interesting step for 6 clusters. We decide to keep for value for our analysis.



2. Clusters



Clusters have a good repartition of individuals, although there is no perfect equality in size.

3. Caracteristics

- Cluster 0 contains abstract subjects (outils conceptuels, probabilié, droit and algorithme)
- Cluster 1 contains technical subject (langage C, approche systèmique, base de données, CSI and gestion financière)
- Cluster 2 contains mathematical subjects (jeu d'entreprise, compilation, mathématique général)
- Cluster 3 contains difficult subjects (analyse de données, assembleur, comptabilité, allemand)
- Cluster 4 contains beginner subject (droit, outils de base, structure des ordinateurs)
- Cluster 5 contains financial subjects (économie général, recherche opérationnelle, contrôle de gestion)

II. Hypercube browsing

1. Dimensions

Here are the dimensions in the cube:

- Summary of score
- Effective of student
- Effective of student of achieve their study

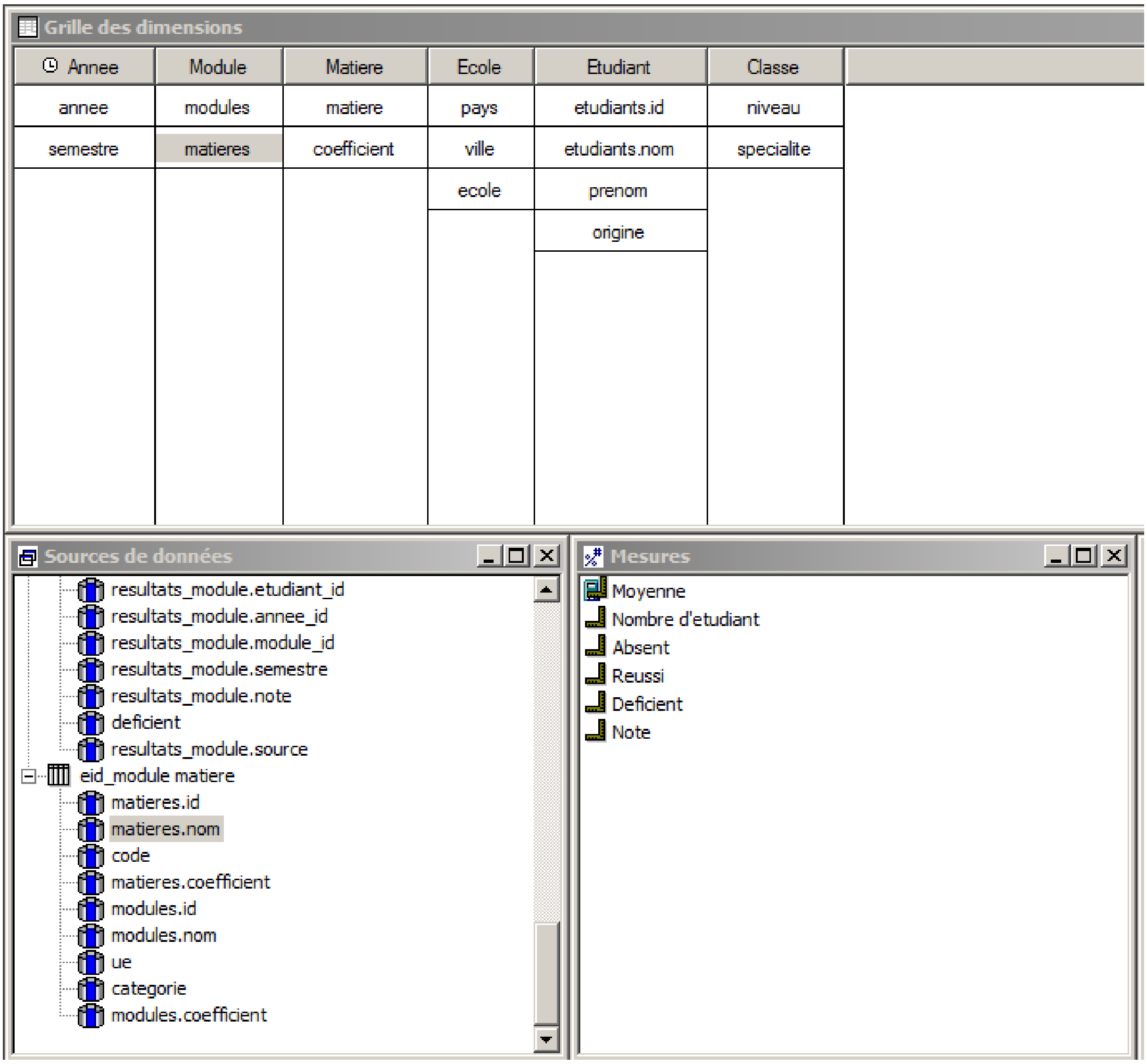
- Effective of deficient student

2. Measures

Here are the dimensions in the cube:

- Summary of score
- Effective of student
- Effective of student of achieve their study
- Effective of deficient student

3. Screen Shots



Grille des dimensions

Annee	Module	Matiere	Ecole	Etudiant	Classe
annee	modules	matiere	pays	etudiants.id	niveau
semestre	matieres	coefficient	ville	etudiants.nom	specialite
			ecole	prenom	
				origine	

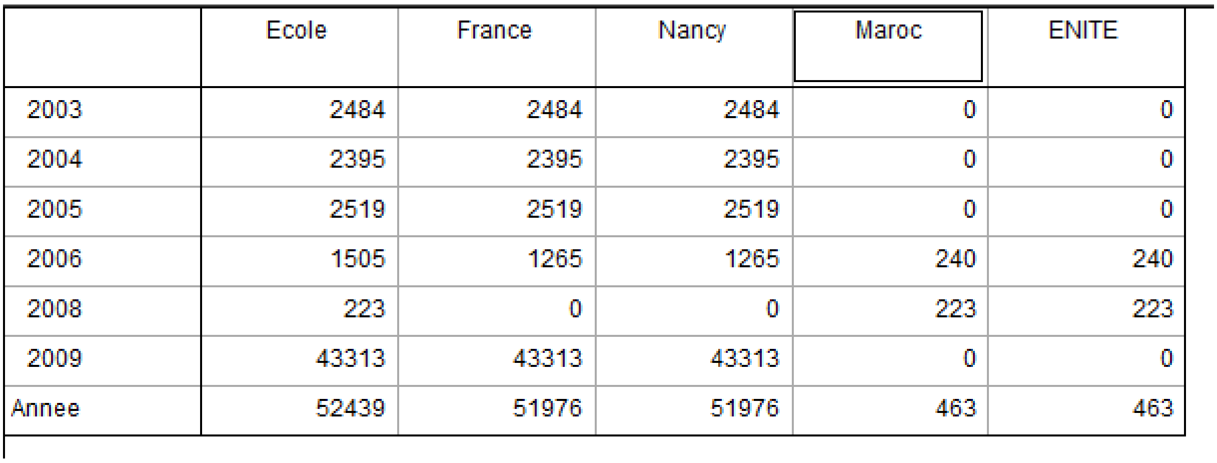
Sources de données

- resultats_module.etudiant_id
- resultats_module.annee_id
- resultats_module.module_id
- resultats_module.semestre
- resultats_module.note
- deficient
- resultats_module.source
- eid_module matiere
  - matieres.id
  - matieres.nom
  - code
  - matieres.coefficient
  - modules.id
  - modules.nom
  - ue
  - categorie
  - modules.coefficient

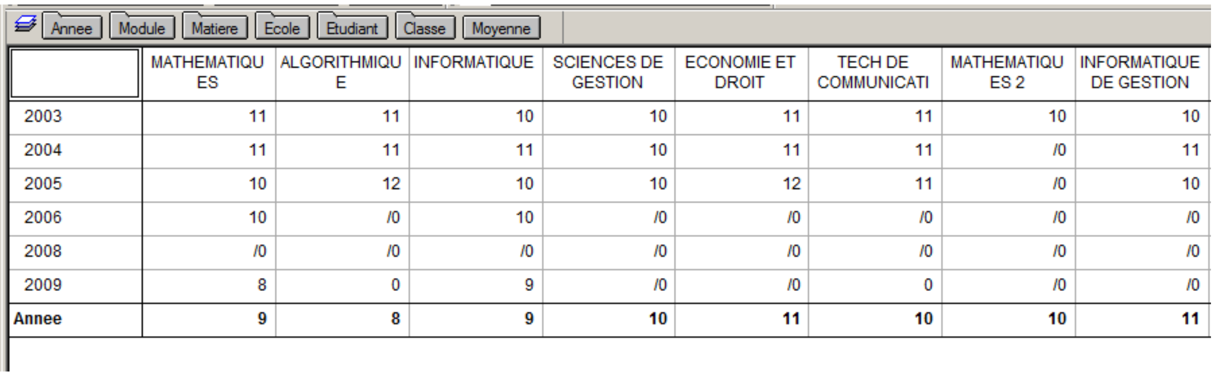
Mesures

- Moyenne
- Nombre d'etudiant
- Absent
- Reussi
- Deficient
- Note

Number of students



	Ecole	France	Nancy	Maroc	ENITE
2003	2484	2484	2484	0	0
2004	2395	2395	2395	0	0
2005	2519	2519	2519	0	0
2006	1505	1265	1265	240	240
2008	223	0	0	223	223
2009	43313	43313	43313	0	0
Annee	52439	51976	51976	463	463



	MATHEMATIQUES	ALGORITHMIQUE	INFORMATIQUE	SCIENCES DE GESTION	ECONOMIE ET DROIT	TECH DE COMMUNICATI	MATHEMATIQUES 2	INFORMATIQUE DE GESTION
2003	11	11	10	10	11	11	10	10
2004	11	11	11	10	11	11	/0	11
2005	10	12	10	10	12	11	/0	10
2006	10	/0	10	/0	/0	/0	/0	/0
2008	/0	/0	/0	/0	/0	/0	/0	/0
2009	8	0	9	/0	/0	0	/0	/0
Annee	**9**	**8**	**9**	**10**	**11**	**10**	**10**	**11**

III. Analysis and Conception

1. Tools and Softwares

Following are the tools and software used in the project:

- Microsoft Excel : files treatment
- Microsoft Visual Basic for Applications : helped to create and edit macros for data extraction into the Excel files, instead of using Talend
- Cognos. software: to build the hypercube
- python : create and edit script in order to determine the K-means using the PyCluster library

- MySQL : database management

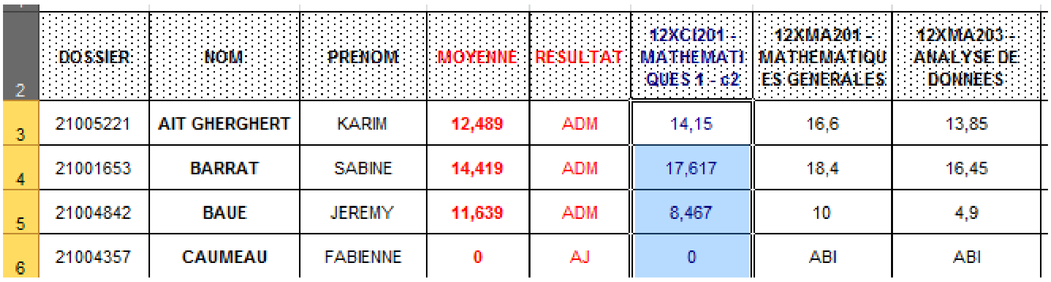
- Microsoft Access : edit input files before exporting toTransformer

2. Data extraction

We had at our disposal 40 Excel files containing results of MIAGE from academic year 2002-2003 to 2010-2011. After a quick exploration, it appears that the sheets structure of files was different based on years. Hence, we needed a particular treatment for each one since there are different levels (DEUG , L2, L3, L3 IGA, M1, M2, …). We also had to deal with the change of policy due to LMD system reform.

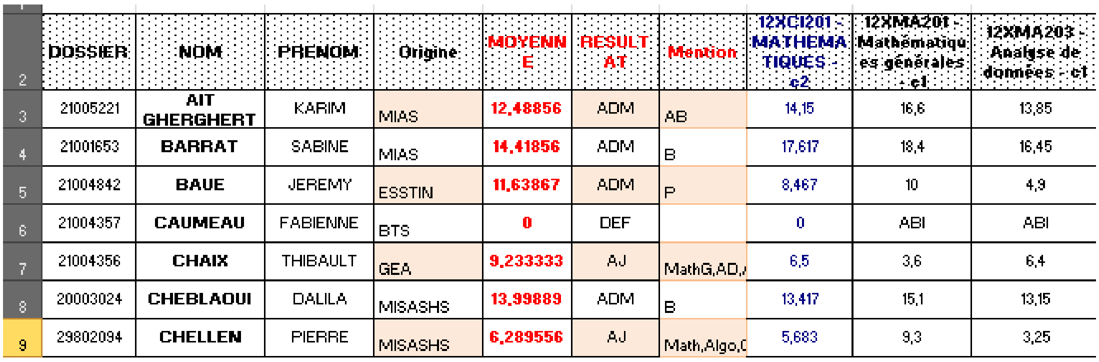
A single convention format was defined in order to standardize the structure of a main output file for a further database export. Non-significant columns have been deleted during the analysis process. That means the columns with information missing or with no interest regarding our priorities. Below are screenshots of Excel files before formatting and after.

Before:



2	DOSSIER	NOM	PRENOM	MOYENNE	RESULTAT	12XCI201 - MATHEMATIQUES 1 - c2	12XMA201 - MATHEMATIQUES GENERALES	12XMA203 - ANALYSE DE DONNEES
3	21005221	AIT GHERGHERT	KARIM	12,489	ADM	14,15	16,6	13,85
4	21001653	BARRAT	SABINE	14,419	ADM	17,617	18,4	16,45
5	21004842	BAUE	JEREMY	11,639	ADM	8,467	10	4,9
6	21004357	CAUMEAU	FABIENNE	0	AJ	0	ABI	ABI

After:



2	DOSSIER	NOM	PRENOM	Origine	MOYENNE	RESULTAT	Mention	12XCI201 - MATHEMATIQUES - c2	12XMA201 - Mathématiques générales - c1	12XMA203 - Analyse de données - c1
3	21005221	AIT GHERGHERT	KARIM	MIAS	12,48856	ADM	AB	14,15	16,6	13,85
4	21001653	BARRAT	SABINE	MIAS	14,41856	ADM	B	17,617	18,4	16,45
5	21004842	BAUE	JEREMY	ESSTIN	11,63867	ADM	P	8,467	10	4,9
6	21004357	CAUMEAU	FABIENNE	BTS	0	DEF		0	ABI	ABI
7	21004356	CHAIX	THIBAULT	GEA	9,233333	AJ	MathG,AD,/	6,5	3,6	6,4
8	20003024	CHEBLAOUI	DALILA	MISASHS	13,99889	ADM	B	13,417	15,1	13,15
9	29802094	CHELLEN	PIERRE	MISASHS	6,289556	AJ	Math,Algo,(	5,683	9,3	3,25

1. Before LMD reform:

This concerns the academic years 2002-2003, 2003-2004, 2004-2005. The idea is to collect the maximum information contained into the sheets and to put it in a final sheet that is called “export”.

The “export” sheet has the format:

Dossier	Nom	Prenom	Origine	Moyenne	Resultat	Mention	Module + coefficient	Matière + coefficient	Matière + coefficient
data	data	data	data	data	data	data	data	data	data

Before LMD reform, there was a single final result referenced in the “Moyenne” column. This fianl format stands for all the excel files with no exception. Unless the file is useless. In this case it is not exported in the database.

2. After LMD reform:

This concerns the academic year 2005-2006 and further. At this time, we faced a new situation, for the academic year is divided into 2 semesters. The “PV” sheet compile for each student, their annual, semester 1, semester 2, modules and subjects’ marks.

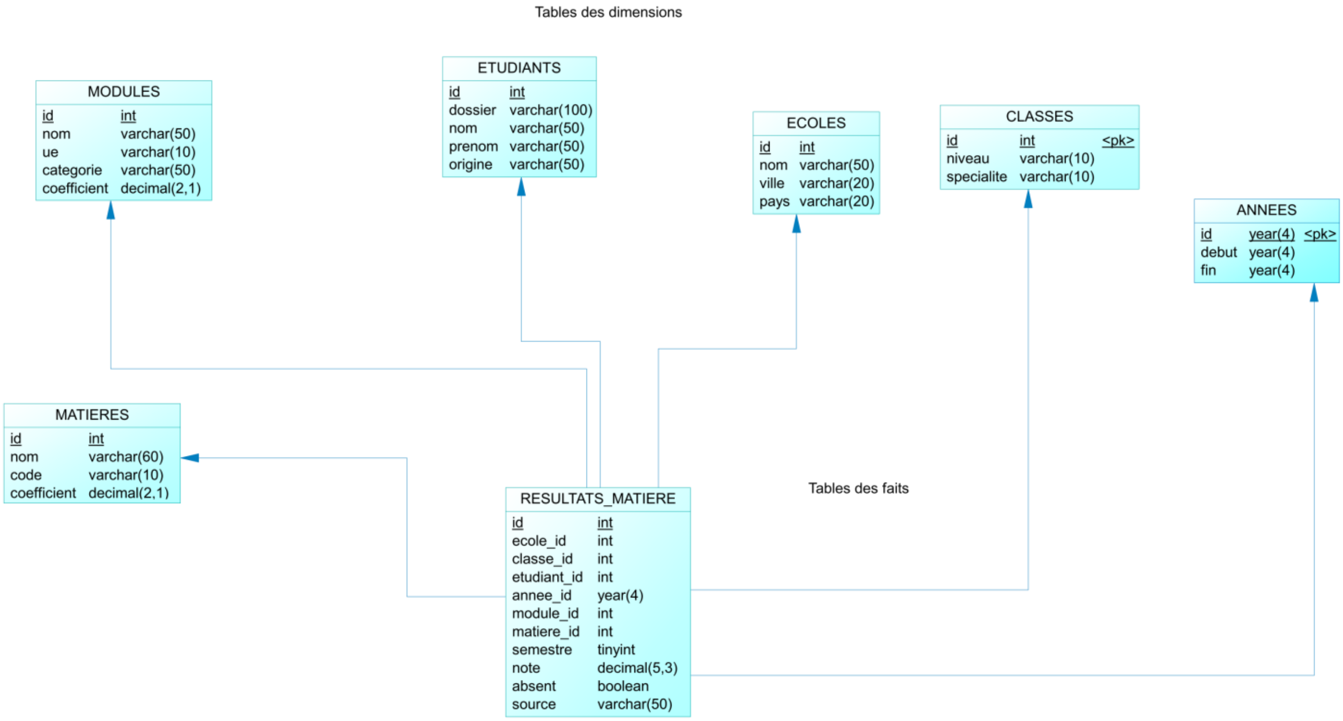
Dossier	Nom	Prenom	Origine	Moyenne (Semester 1 or Semester 2 or Annual)	Resultat	Mention	Module + coefficient	Matière + coefficient	Matière + coefficient
data	data	data	data	data	data	data	data	data	data

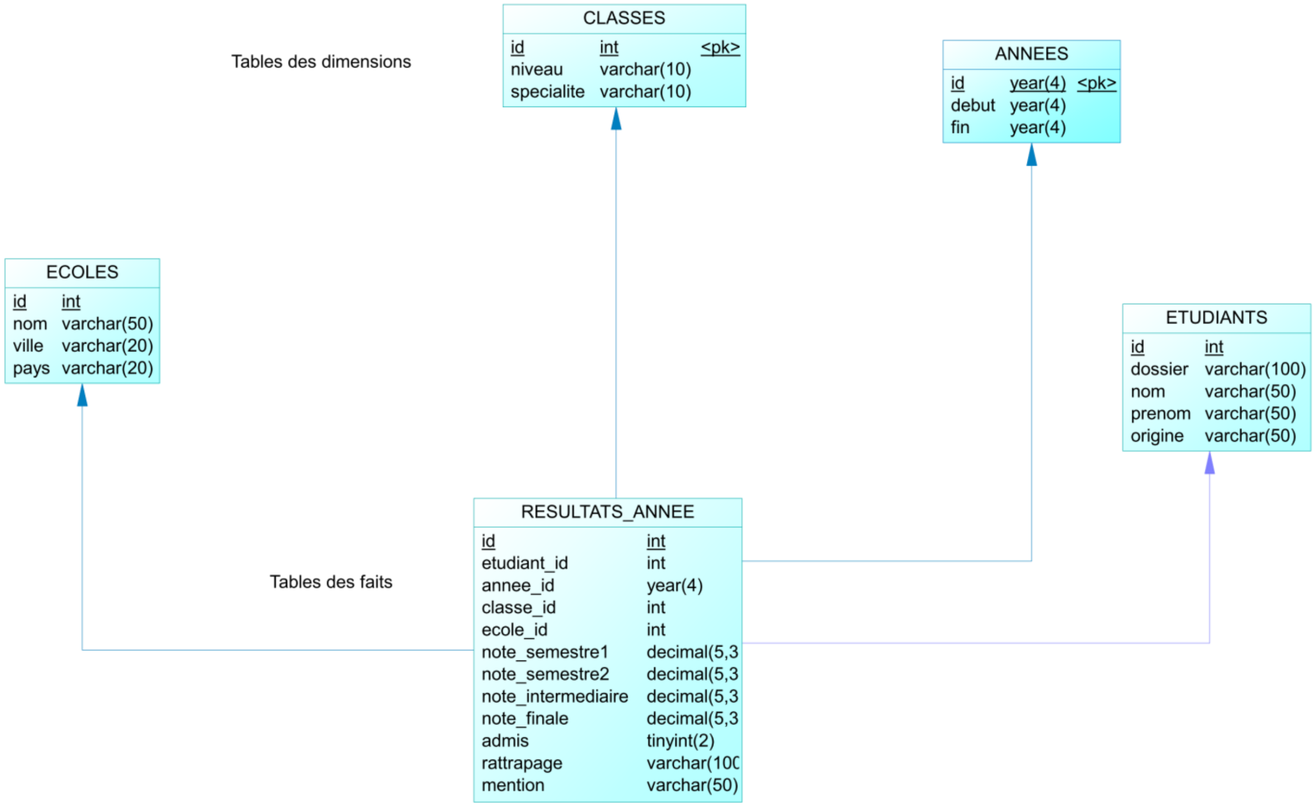
3. Setting up the Database

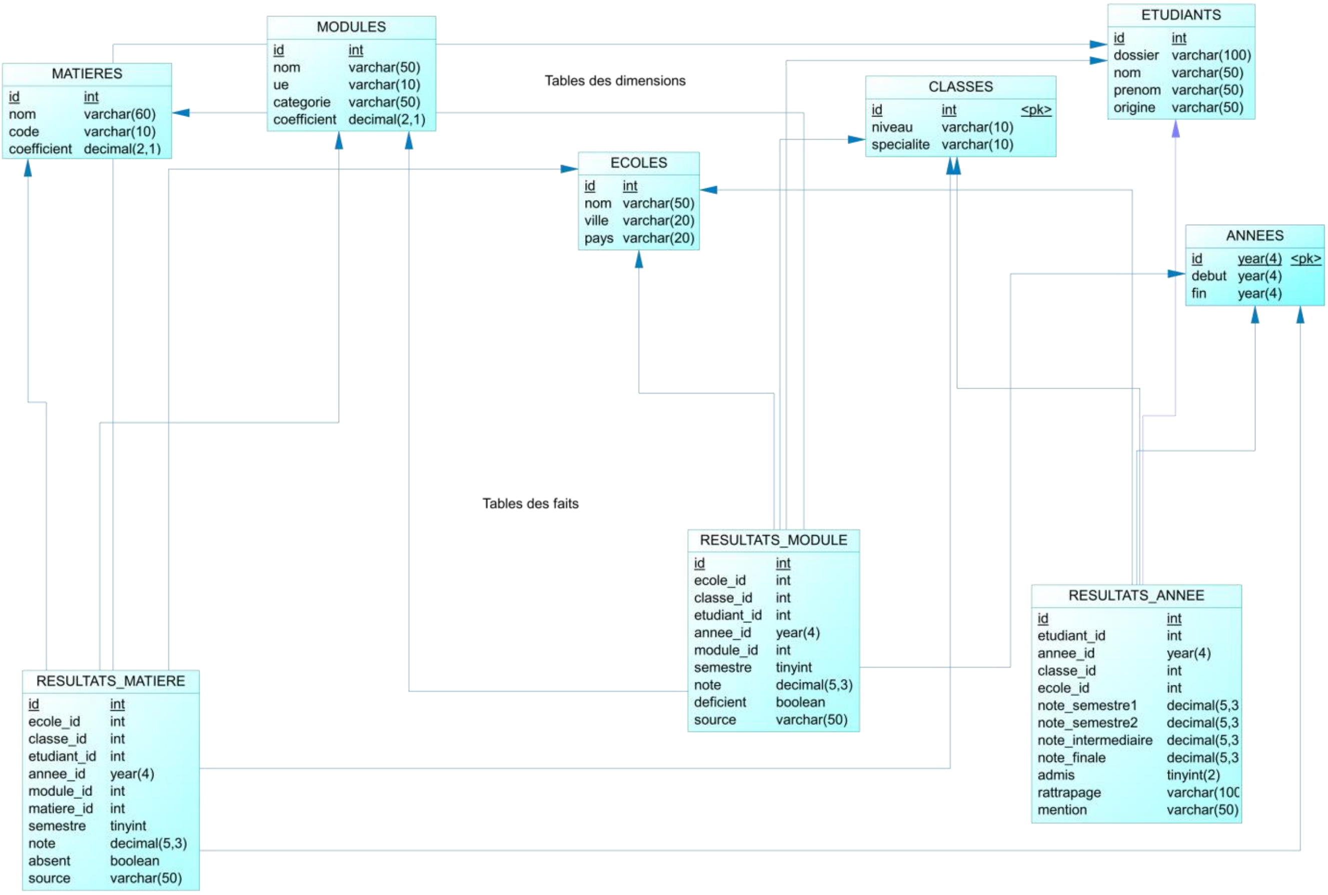
We chose to implement our database based on a star schema. The benefit is the denormalized form of the database. It is composed by:

- 6 dimensions tables: MATIERES, MODULES, CLASSES, ECOLES, ETUDIANTS, ANNEES
- 3 facts tables: RESULTATS_MATIERES, RESULTATS_MODULES, RESULTATS_ANNEE

Below is the Physical Data Model representation:







MODULES	
id	int
nom	varchar(50)
ue	varchar(10)
categorie	varchar(50)
coefficient	decimal(2,1)

MATIERES	
id	int
nom	varchar(60)
code	varchar(10)
coefficient	decimal(2,1)

ETUDIANTS	
id	int
dossier	varchar(100)
nom	varchar(50)
prenom	varchar(50)
origine	varchar(50)

CLASSES		
id	int	<pk>
niveau	varchar(10)	
specialite	varchar(10)	

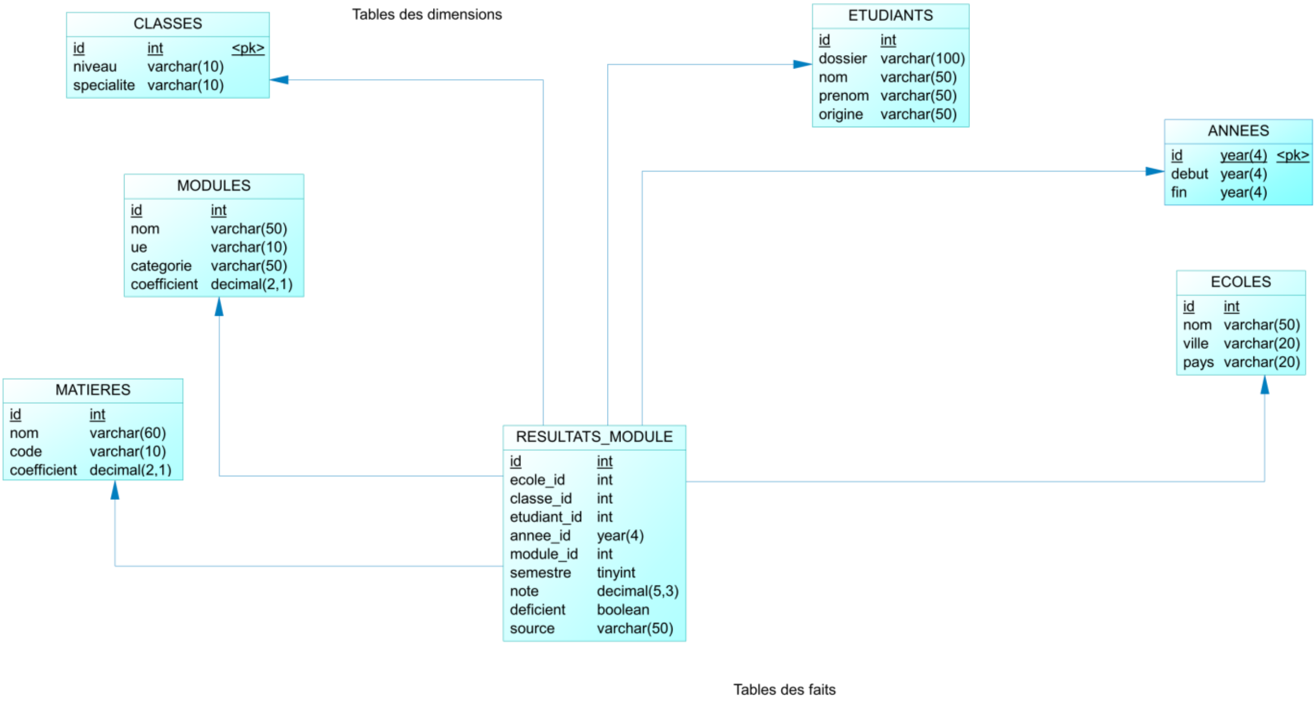
ECOLES	
id	int
nom	varchar(50)
ville	varchar(20)
pays	varchar(20)

ANNEES		
id	year(4)	<pk>
debut	year(4)	
fin	year(4)	

RESULTATS_MODULE	
id	int
ecole_id	int
classe_id	int
etudiant_id	int
annee_id	year(4)
module_id	int
semestre	tinyint
note	decimal(5,3)
deficient	boolean
source	varchar(50)

RESULTATS_ANNEE	
id	int
etudiant_id	int
annee_id	year(4)
classe_id	int
ecole_id	int
note_semestre1	decimal(5,3
note_semestre2	decimal(5,3
note_intermediaire	decimal(5,3
note_finale	decimal(5,3
admis	tinyint(2)
rattrapage	varchar(10C
mention	varchar(50)

RESULTATS_MATIERE	
id	int
ecole_id	int
classe_id	int
etudiant_id	int
annee_id	year(4)
module_id	int
matiere_id	int
semestre	tinyint
note	decimal(5,3)
absent	boolean
source	varchar(50)



CLASSES		
id	int	<pk>
niveau	varchar(10)	
specialite	varchar(10)	

ETUDIANTS	
id	int
dossier	varchar(100)
nom	varchar(50)
prenom	varchar(50)
origine	varchar(50)

ANNEES		
id	year(4)	<pk>
debut	year(4)	
fin	year(4)	

MODULES	
id	int
nom	varchar(50)
ue	varchar(10)
categorie	varchar(50)
coefficient	decimal(2,1)

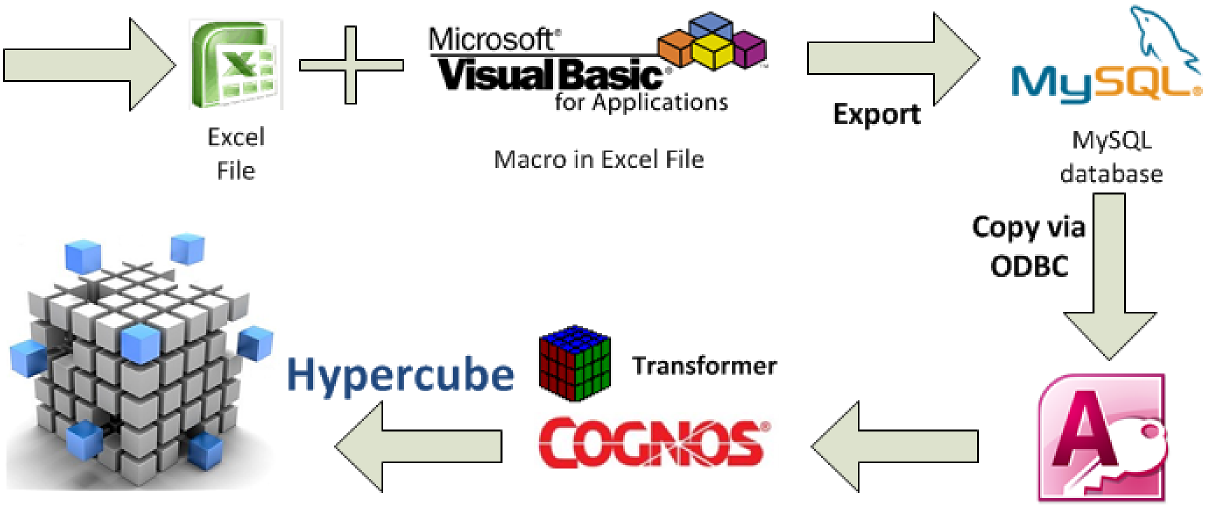
ECOLES	
id	int
nom	varchar(50)
ville	varchar(20)
pays	varchar(20)

MATIERES	
id	int
nom	varchar(60)
code	varchar(10)
coefficient	decimal(2,1)

RESULTATS_MODULE	
id	int
ecole_id	int
classe_id	int
etudiant_id	int
annee_id	year(4)
module_id	int
semestre	tinyint
note	decimal(5,3)
deficient	boolean
source	varchar(50)

4. Data integration

Instead of using Talend, we chose another solution. It appears that depending on the academic years, modules and subjects are not named identically. This is why we had to format all files before import as explained above. Several hours were spent in order to obtain corrects files. It was decided not to spend more time on data integration regarding the heaviness of Talend. It was then decided to define a procedure with the purpose to facilitate and control the integration data process. Since we worked on a Microsoft Windows environment, it was natural to choose a specific tool as Visual Basic for Application. VBA is useful to develop solid macros adapted to each “export” sheet contained in the Microsoft Excel files. The export files then feed a Mysql database.



3. Conception

1. Algorithm

We use **k-means clustering** to classify students and subjects. This method is adapted for big data model because the algorithm is **simple, fast and we can customize the notion of distance** that the function uses (the similarities between profiles). It uses an **input matrix** with **individuals** (student or subject) on rows and **dimensions** on columns (module mark, subject mark, student mark).

Each cell is a numeric value, each missing value is ignored.

After choosing a value for the k parameter, the algorithm tags individuals with a label (a number between 0 and k). **Clusters have different sizes**, and show properties that used to characterize groups.

For each value of k, we have an error value that indicates the optimal k value. We chose to report only the optimal clusters, but we can use the algorithm for any value of k.

2. Notion of distance

A distance between two individuals X and Y is a function that asserts:

- d(x, y) >= 0
- d(x, y) = d(y, x)
- d(x,y) = 0 only if x=y
- d(x, y) <= d(x,z)+d(z,y) for all x,y,z (triangle inequality)

There is two kind of distance function:

- semi-metric that does not satisfy triangle inequality (based on correlation **r** where **d = 1 - r**)
- metric that satisfies triangle inequality (based on Euclidean)

In our cube, **our main measure are marks between 0 (bad) and 20 (good)**. So we can easily obtain a distance between profiles because all marks are numerical values. We use the **Euclidean distance** for simple cases like only one mark, but we will lose information for multiple marks using this solution.

An example with 3 students and 3 modules with the same coefficient:

	Student A	Student B	Student C
Mathematics (1)	20	0	10
Computer Science (1)	10	10	10
Communication (1)	0	20	10
Final note	**10**	**10**	**10**

The final note is the same for every student, but they didn't obtain it by the same way. For example, Student A is good in mathematics, average in computer science but didn't pass his communication module and Student B obtain the opposite marks.

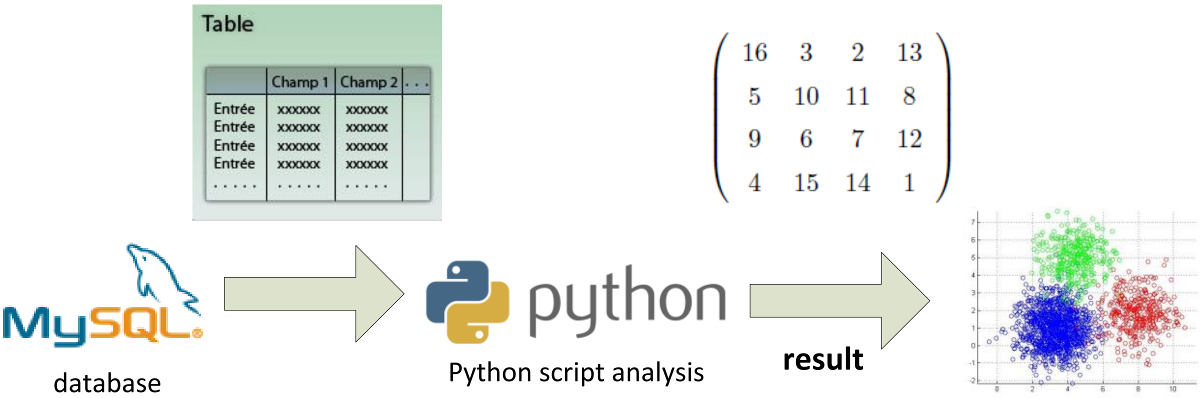
A teacher or an academy chief is not only interested by the final note of his students, but also by the modules: all or just some of them. It's a more precise approach to use the ***Pearson correlation* coefficient**. Using it, we can group students not only by grade, but also by the similarity between the modules they attent.

We choose to report only the correlation distance function on normalized data.

We also explored these functions:

- [a] Absolute value of the Pearson correlation coefficient (semi-metric)
- [u] Uncentered pearson correlation (semi-metric)
- [x] Absolute uncentered pearson (semi-metric)
- [s] Spearman rank correlation (semi-metric)
- [k] Kendall's t (metric)
- [e] Euclidean distance (metric)
- [b] City-block distance (metric)

3. Process



$$\begin{pmatrix} 16 & 3 & 2 & 13 \\ 5 & 10 & 11 & 8 \\ 9 & 6 & 7 & 12 \\ 4 & 15 & 14 & 1 \end{pmatrix}$$

- From the MySQL database, we extract the input data from the fact tables using a specific query.
- our Python script uses these datas to build an N x M matrix and a mask for all the missing values
- we use an extern library : PyCluster, to compute the labels (k-means method)
- the script ouput is an array with the matrix data, headers, labels, and the error value

We chose this technique instead of Tanagra and other commercial product.

Advantages:

- Python and MySQL are fast data extraction and computation
- Scripts can be modified and adapted quickly and on demand
- No need to compile our program because Python is an interpreted language
- We can customize the ouput (CSV, HTML, PDF , XLS, or TXT)
- It is multiplaform and require few dependencies

Disadvantages:

- Scripts are not an easy solution for end users (command line)
- We cannot change the algorithm easily, it requires a new development

IV. Conclusion

1. Review

In the end of the projet, we managed to produce:

- a fixed **Standard of Sheet** for input source
- a reliable **SQL Database** implementing a star schema
- an efficient **VB Macro** to extract datas from the sheet to the database
- a partial **Cognos Hypercube** with multiple measure and dimension
- an extensible **Python Script** to compute k-means

Our assembly line is complete, although some informations are missing from the sheet. To improve our solution, **we could organize an audit** to implement more dimension, and ways to handle missing data.

Also, **we could implement an end user interface for our k-means program**. For the time being, the decision maker needs to refer to the technical staff to make more classifications.

2. Personal opinion

We have mixed review feelings on this project, but we admit that it’s a real showcase of Business Intelligence needs and solutions.

The main con of the project is **the renormalization of the sheets** that required a huge amount of time to achieve. We chose to use a standard input and automate the rest rather than treat each case individually. **This project taught us that standard source of information are crucial to build new kind of applications.**

Even if this project has required a lot of manual work, we are proud of the knowledge we acquired on the tools we used. We had seen numerous aspects of datamining, data exploration and classification on a real case. It will definitely be an useful experience regarding what we will encounter in our career.