# Using Students' Data to Improve the Quality of the Education in Moroccan Institution

M. Aitdaoud, K. Namir, M. Bentaib, R. Ihya, S. Bouiti, M. Talbi

***Abstract*—The goal of every company and public sector organization is to provide quality service to their customers and make them satisfied. However, as we move to-wards a more connected world where technology has been integrated into the business process, handling data has become more complicated. Today, businesses and High School Institutions (HSI) face one of the biggest challenge, which is characterized by the exponential growth of data storage in various formats such as plain text, relational database, etc…**

**This massive data can be used to improve decision making and management, which requires proper extracting and cleaning methods. For that reason, data warehousing has become a major step in the knowledge discovery in databases (KDD) process which can guarantee a solid description of concepts and methods for transforming transactional data into analytical data formats.**

**The aim of this paper is to provide a way to support and understand the educational processes of a HSI by offering a new description to data and making it more venerable using visualization techniques. We used four different datasets for this study throughout the years (from 2012 to 2016), which was collected from a HSI Enterprise resource planning (ERP) database.**

***Keywords*— Educational Data Mining, KDD, Data warehouse, Data provisioning, Data Visualization.**

## I. INTRODUCTION

The purpose of School Institutions as knowledge centers and human resource developers is to provide an education of quality to its pupils by improving the quality of educational processes and identify the means by which it can be validated and improved. Education quality is one of the key responsibilities of any Secondary School Institutions (SSI) or High School Institutions (HSI) to its stakeholders denoting not only the requirement for production of high level knowledge, but also the need to provide efficient education so that pupils achieve their learning objectives without any problem [1].

One way to enhance the quality of educational processes in HSI is to extract insights from educational data to study the main attributes that may affect the pupil's performance. This will reveal more understanding to teachers, school administrators and parents who have always wanted to know how the pupils are doing in their education [2].

This knowledge can be discovered from data that reside in different databases of the organization, or in surveys that collect data used to evaluate quality criteria (course assessment, lecturer assessment, pupil assessment, etc.) and can be extracted through KDD process. This new research field concerned with methods for exploring the unique types of data that come from educational settings and their use to better understand learners and their characteristic, is called Educational Data Mining (also referred to as "EDM") or Learning Analytics (also referred to as "LA"). EDM is defined as the area of scientific inquiry centered around the development of methods to extract insights within the data that come from educational settings, and using those methods to better understand pupils and the settings which they learn in [3].

The supervision of the performance of high school institutions pupils is vital during an early stage of their curricula. Indeed, their grades in specific major courses (mathematic, French….) as well as their cumulative General Point Average (GPA) are decisive when pertaining to their ability to pursue their academic studies in higher education institutes. Furthermore, these compelling strict requirements not only significantly affect the attrition rates in mathematics and French studies (on top of probation and suspension) but also decide of grant management, developing courseware, and scheduling of programs.

In this paper, we present a study that has a twofold objective. First, it attempts at correlating the aforementioned issues with the pupils' performance in some key courses taken at early stages of their curricula, then, by using statistics and visualizations techniques, some primary useful knowledge is presented and refined in order to demonstrate the capability of data mining to improve the quality of the educational processes by supporting the administration of educational institutions in the decision-making process and in identifying more enhanced policies for educational practices.

In this paper our aim is to investigate the educational domain of data mining using a case study from the pupils' data collected from a HSI Enterprise resource planning (ERP) database. The data include four years' period from 2012 to 2016. It showed the data that we could collect, how we could preprocess and visualize it, and finally how can we benefited from the discovered knowledge.

This paper takes into consideration the ethical and privacy issues. Official approval from the HSI was obtained to have an access to the related databases for the sole use of analysis and knowledge discovery purposes. To achieve privacy, all individual and personal data are extracted from the database before applying any processing.

The rest of the paper contains the following sections: section 2, presents related work in the field of educational data mining and describing the capabilities it has as well as the elements that can be considered as the success factors of its application. Section 3, contains a description of the process of data provisioning. Section 4, describe the purpose of data visualization. Section 5, present the results of the experiments conducted. Finally, in Section 6, we conclude this paper and give an outlook of future work.

## II. Related works

Although, using KDD in higher education is a recent research field, there are many works in this area. That is because of its potentials to educational institutes. Remero and Ventura [4], established a survey on educational data mining between 1995 and 2005, they concluded that educational data mining is a promising area of research and it has a specific requirements not presented in other domains. Thus, work should be oriented towards educational domain of data mining.

An educational institution often has many diverse and varied sources of information. A real life application of KDD was presented by [5], to find weak pupils. They visualized that the education domain offers many interesting and challenging applications for KDD. These applications not only enhance an educational institute in delivering a better quality of education experience, but also aid the institution in running its administrative tasks effectively. With so much information and so many diverse needs, it is foreseeable that an integrated data mining system that is able to cater to the special needs of an educational institution will be in great demand. KDD can be used in educational field to enhance our understanding of educational process to focus on identifying, extracting and evaluating attributes related to the educational process of pupils as described by [6].

Data warehousing is a powerful repository of integrated information, available for querying and analysis [7], which enables educational institutions to better allocate resources and staff, and proactively manage pupil outcomes. The management system can improve their policy, enhance their strategies and thereby improve the quality of that management system [8].

A model was developed by [9], to find similar patterns from the data gathered (such as assignment marks, attendance, test marks etc…) and to make predication about pupils' performance. The goal of the paper was to make the description and predictions about the pupil performance as well as to find similar pattern from the collected data.

## III. Data provisioning

Data provisioning constitutes the prerequisite for any Business Intelligence (BI) project. Clearly, without any data basis, there will be no analysis at all, and without a database of good quality, the quality of the analysis can be expected to be low. However, data collection, extraction, and integration are often the most complex and expensive tasks in a BI project. As already stated by [8]: "What at first appears to be nothing more than the movement of data from one place to another quickly turns into a large and complex task far larger and more complex than the programmer thought." According to [10], companies state that "information integration is thought to consume about 40 % of their budget." Kimball states that the design and development of the underlying "systems consumes the lion's share of effort during a DW/BI project" [11]. In addition, due to current developments, such as big data, more and more data is available holding potential for valuable analysis [12].

### *A. Data Collection and Description*

Data collection is the process of gathering and measuring information of interest that can provide answers to stated research questions and evaluate outcomes. In practice, an often-occurring realistic outcome is a compromise between the desired analysis goals and the available data [13]. In our case study we collected the pupils' data from a private school ERP database, which contains personal records and academic records of pupils (i.e. the tow period grades, number of school absences, Address etc…).

### *B. Data Extraction*

After selecting the relevant data sources and describing them, the next step is to extract relevant data from various possibly heterogeneous data sources and create a data-warehouse (DWH) area for analysis purposes [14]. Typically, data extraction is part of the so-called extraction-transformation-load (ETL) process [15].

A data warehouse is a collection of integrated and thematic databases (Fig. 1), which is designed to support the decision-making function, where each unit of data is relevant to events at a given time [7].

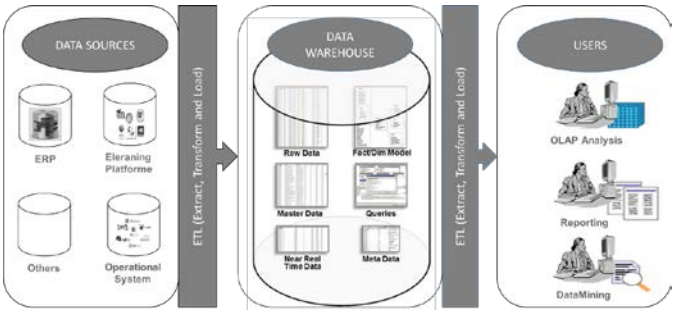


**Fig. 1.** Data Warehousing Architecture

Data warehouses, in contrast, are targeted for decision support. Historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, they tend to be magnitude larger and complex than operational databases.

To facilitate complex analyses and visualization, the data in a warehouse is typically modeled multi-dimensionally and might be implemented on standard or extended relational

DBMSs, called Relational On-Line Analytical Processing (ROLAP) servers. These servers almost always query over a database which is structured as a star schema. As discussed in [16], the star schema provides, obviously, a more compact representation of the multidimensional data. In our cases, the dimensions of our star schema are shown in Fig. 2 (classes_dim, courses_dim, dates_d, lesson_d, student_dim).

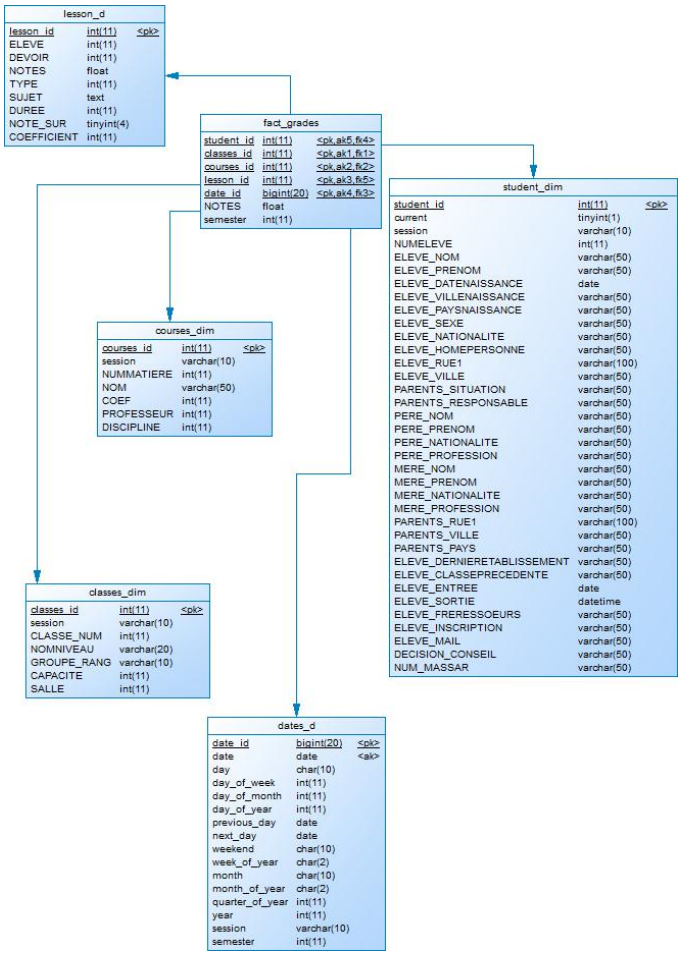


**Fig. 2.** A Star Schema

## IV. Data Visualization

The main challenge in the description and visualization of a collection of process instances is the definition of appropriate summaries and displaying these multivariate summaries in such a way that the complexity of the educational process is captured in an understandable way for the observer. For readers, a graphic is usually more instructive than a table of numbers, and it is of utmost importance to give a correct impression about reality.

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software [17]. Visual Data Mining is the process of searching and analyzing databases to find implicit but potentially useful information [18].

The starting point for all kinds of description and visualization is the selection of an appropriate data frame containing the variables of interest. This means that we have to decide what process instances we want to analyze and what attributes should be used. The selection of the instances is usually accomplished by defining a certain time interval for which data in the cross-sectional view or in the state view are analyzed.

The data structures mentioned (see Fig.2) are those occurring most frequently in applications. Multidimensional tables, often called fact tables, are defined by the values of qualitative variables called dimensions and by a summary attribute for the cells of the fact table. This summary attribute is usually defined by the counts or means for a quantitative variable describing process instances. Simple data structures are defined by a matrix with rows representing the process instances and columns representing the values of the variables for the instances. Although the data have a simple matrix structure, there may be some internal structure, for example, groups of instances according to attributes like sex or age.

## V. Proposed Work

In Morocco, the high school education consists of 3 years of schooling, preceding by 9 years (3 of secondary education and 6 of basic education) and followed by higher education. There are several divisions (e.g. Sciences, Economy, Literature) that share core subjects such as the French Language and Mathematics. Like several other countries (e.g. France or Turkey), a 20 point grading scale is used, where 0 is the lowest grade and 20 is the perfect score. During the school year, pupils are evaluated in tow periods and the last evaluation (GPA of Table 1) corresponds to the average and final grade.

This study was based on data collected throughout the school years (from 2011-2012 to 2015-2016) that belongs to a private school from the grand Casablanca region of Morocco. Although there has been a trend for an increase of Information Technology investment from the Government, by using a management system information to save and control the pupils' grades. Hence the majority of the Moroccan private school uses an ERP software, which provides a set of tools that enhance school administrators to smoothly run the institution and do so in a way that demonstrates efficiency, cost-savings and ingenuity.

As part of the data provisioning for transforming transactional data into analytical data formats, the following steps are performed as part of the preparation and preprocessing of the dataset: Tools, Data collection, Data selection and preparation.

### *A. Tools*

Today, in computer era, different data extraction tools are available in the market and each tool has its own merits and demerits. For the analysis, we have concentrated on HSI dataset and will study the result analysis to look into some

parameter for pupil's performance using data mining clustering, classification and association techniques. For this we are going to use the Talend Open Studio (open source) version 6.3.1 for data integration, then we will be using our own application for students' grades description and visualization. That we name HSIgrading.

Talend Open Studio (TOS) is an open source data integration platform, based on Java language. TOS allows to answer all the problems related to the processing of the data in the decision chain: ETL, EAI: Inter-Application Data Exchange, Synchronization of data. One of the great strengths of TOS is the ability to connect virtually to any data source (DBMSs, Plain text…).

HSIgrading is a web application designed for data description and visualizations (students' grades) of High School Institutions. It allows an easy connection to our data warehouse (Fig. 2) to access all students' data for use. It also allows data visualization with various types of chart for dashboards, to help the different stockholder understanding the students' performance to make better decision and support the student at risk.

### *B. Data Collection*

Data collection is the process of gathering and measuring information of interest that enables to answer stated research questions and evaluate outcomes. In our case study we collected the pupils' data from a private school ERP database, which contains personal records and academic records of pupils (i.e. the tow period grades and number of school absences, Address etc…).

### *C. Data selection and preparation*

Data selection is the first stage of mining process. In our work data is selected from the school institutes management database, throughout the school years (from 2011-2012 to 2015-2016) from the grand Casablanca region of Morocco.

During the preprocessing stage, some features were discarded due to large variances or duplicates information which make them irrelevant for data mining. The remaining attributes are shown in Table 1

Finally, the data was integrated into two datasets related to Mathematics (with 1454 examples) and the French language (1262 records) classes.

**Table 1.** The preprocessed student related attributes

| **Attributes** | **Description (Domain)** |
|---|---|
| student_id | Student's id |
| numeleve | Student's number |
| age | Student's age (numeric: from 15 to 22) |
| eleve_sexe | Student's sex [binary: male (MASCULINE) or female (FEMININE)] |
| address | Student's home address type (string) |
| parents _situation | Parent's status [binary: married(MARIÉS), divorced(DIVORCÉS), father deceased (PÈRE DÉCEDÉ), mother deceased (MÉRE DÉCEDÉ)] |
| Parents _responsable | Student's guardian [nominal: mother(MÉRE) , father(PÈRE) , both(PARENTS) or other(AUTRE)] |
| père _profession | Father's job (nominal) |
| mere _profession | Mother's job (nominal) |
| nomniveau | The class level of the student |
| groupe_rang | The class group of the student |
| absences | Number of school absences (numeric: from 0 to 93) |
| session | School year |
| note_s2 | first period grade (numeric: from 0 to 20) |
| note_s2 | Second period grade (numeric: from 0 to 20) |
| notes | General Point Average : final grade (numeric: from 0 to 20) |
| mention | Student's performance [nominal : excellent(TB), very good(B), good(AB), poor(P) and failure (D)] |
| valider | Student's class graduation [binary: true or false] |

To get better input data for data mining techniques, we did some preprocessing for the collected data. After we integrated it into one file for each dataset (Mathematics and French), to increase interpretation and comprehensibility, we discretized the GPAs "NOTES" attribute to categorical one "MENTION". For example, we grouped all grades into five groups excellent "TB" (GPA >= 16), very good "B" (16 > GPA >= 14), good "AB" (14 > GPA >= 12), poor "P" (12 > GPA >= 10) and failure "D" (GPA < 10).

For our work we choose MySQL database to design the datasets. All the information collected from data collection phase will then transform into data table. The choice of the database was justified by several elements (simplicity, efficiency of data processing and excellent portability).

### *D. Data description and visualization*

After cleaning and loading the data, we can easily compute different statistics using our data visualization application HSIgrading (Fig. 3).

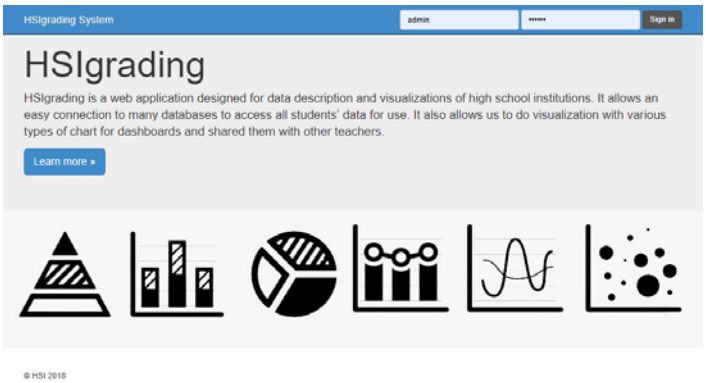


**Fig. 3.** Home page of HSIgrading

HSIgrading is designed to provide incentive for achievement and assist in identifying problem areas of students in a Moroccan HIS. Students' grades are vital information needed in

advancing to the next grade/year level and its accuracy is very important.

In the following, we will present the qualitative and quantitative variables using some basic visualization techniques. We will also comment on the structure and description of the data.

*1) Description and visualization of qualitative variables.*

For qualitative variables, we structured the data as s pivot table with counts for the different attributes using tableau software. The pivot table gives the frequencies of the different combination of values, either as absolute values or as percentages.

We visualize our data using charts; the purpose of a chart is to package information in a way that makes it quickly understandable. The thing that makes charts so useful is that they provide a quickly recognizable shape for our data. For the visualization of more than one variable, we used a bar chart.

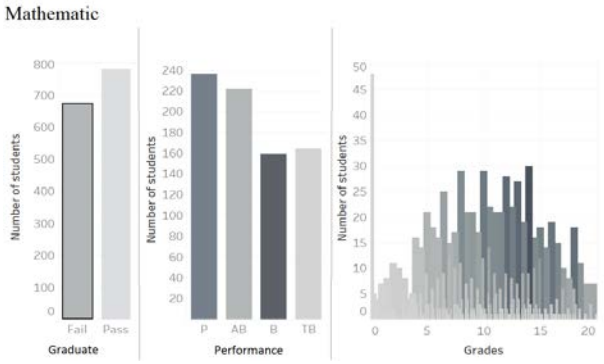


**Fig. 4.** Dashboard of bar chart for pupils' characteristics in Mathematic

Using Tableau software, we created dashboard for the Mathematic courses. On the left side of Fig. 4, there is a bar chart for pupils' graduation status which shows that the number of pupil that passes is not high enough from the number of pupil that fails. In the middle, we note that the majority of pupils scored poor and good results. By looking at the graphic on the right side we confirm our previous statement, which display the grades of pupils and shows that a lot of pupils scored grades between 10 and 14.

Fig. 5 present a dashboard of graphs for the French courses. On the left side, we notice that there are a satisfactory number of the graduate pupils. In the middle, the bar chart of performance shows that the majority of pupils had very good and excellent results. This is confirmed by the last graphic in the right, which display the grades of pupils and shows that a lot of pupils scored grades between 15 and 20.

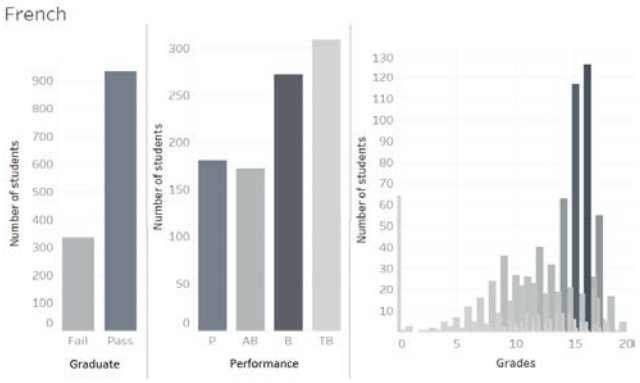


**Fig. 5.** Dashboard of bar chart for pupils' characteristics in French courses.

Many a time, a mosaic plot is more instructive than a bar chart for displaying two or more variables. In a mosaic plot, all data are represented as a square. The horizontal edge of the square is split according to the proportions of the first variable, and we obtain a number of rectangles with areas corresponding to the relative frequencies.

A mosaic plot corresponding to the math clustered bar chart is shown on the left side in Fig. 6. One can learn, for example, that female pupils scored more frequently excellent "TB" and very good "B" than male pupils, but less frequently good "AB" and poor "P" than male pupils. A mosaic plot also incorporating French courses is shown on the right side in Fig.4. We notice the same results obtained in the mathematic courses.

*2) Description and visualization of quantitative variables.*

A standard description of quantitative variables can be obtained by calculating the summary statistics such as mean, variance, standard deviation, minimum, maximum, and quartiles.

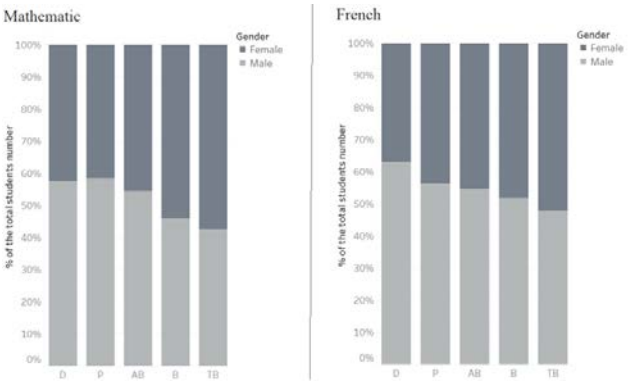


**Fig. 6.** Dashboard of mosaic plot for pupils' grades grouped by gender.

We visualize our data using box plots, also known as a box-and-whisker plot are a good way to summarize large amount of data. It displays the range and distribution of data along a number line. Box plots are descriptive statistics used when

there is no prior information about the distribution of the underlying population.

In the left side of Fig. 7 we compare performance of female and male pupils in the mathematic courses. And on the right side we compare performance of classes. The black lines show the median of the scores in each categories where the boxes denote the quartile of the grades (25%-70% of data points are placed inside the box, we define it as "interquartile range").

Judging by the median of the tow boxes in the left side graph of Fig. 7, we can see that female pupils have slightly better performance than male. However, the difference is not significant. With this plots, we can also compare the overall shape of the two genders, not only their average performance. For example, we can see that the female pupils' score has slightly better performance than male in the lower and upper quartile.

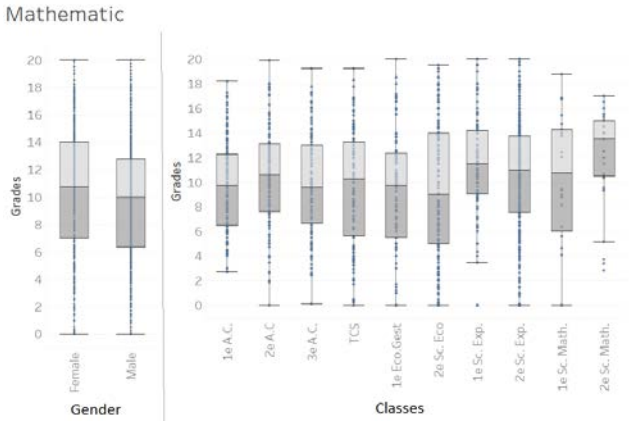


**Fig. 7.** Dashboard of boxplots for pupils' performance in Mathematic courses.

We compare the performance in similar manner for pupils' classes. In the right side graph of Fig. 7, we note that the pupils in second mathematic class scored the best grades. Using the median to compare the results the second mathematic class has the upper hand, however the difference is not significant between all the classes. As the right graph shows the lowest median is scored by second economy class and also it has the most distribution of grades (0 at the min and 20 at the max).

## VI. Conclusion

To conclude, using the data generated from different systems and by using the techniques mentioned before, we can give a better understanding and provide insights that can help solve various problems and even predict them before happening. This can be useful in our case to improve the quality of education system and help to identify the problem easier.

For further work, we will work on using the Principal component analysis (PCA) procedure to identify correlated variables and to define the factors that affect student achievement, also we will apply data mining techniques to predict student grades and performance.

## References

[1] M. Chalaris, S. Gritzalis, M. Maragoudakis, C. Sgouropoulou, and A. Tsolakidis, "Improving quality of educational processes providing new knowledge using data mining techniques," *Procedia-Soc. Behav. Sci.*, vol. 147, pp. 390–397, 2014.

[2] R. Baker, "Data mining for education," *Int. Encycl. Educ.*, vol. 7, no. 3, pp. 112–118, 2010.

[3] S. L. Prabha and A. M. Shanavas, "Educational data mining applications," *Oper. Res. Appl. Int. J.*, vol. 1, no. 1, pp. 23–29, 2014.

[4] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.

[5] Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M. Lee, "Targeting the right students using data mining," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 457–464.

[6] A. El-Halees, "Mining students data to analyze e-Learning behavior: A Case Study," *Dep. Comput. Sci. Islam. Univ. Gaza PO Box*, vol. 108, 2009.

[7] W. H. Inmon, *Building the data warehouse*. John wiley & sons, 2005.

[8] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational web-based system," in *Frontiers in education, 2003. FIE 2003 33rd annual*, 2003, vol. 1, p. T2A–13.

[9] K. Shyamala and S. Rajagopalan, "Data mining model for a better higher educational system," *Inf. Technol. J.*, vol. 5, no. 3, pp. 560–564, 2006.

[10] P. A. Bernstein and L. M. Haas, "Information integration in the enterprise," *Commun. ACM*, vol. 51, no. 9, pp. 72–79, 2008.

[11] R. Kimball and M. Ross, *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence Remastered Collection*. John Wiley & Sons, 2015.

[12] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[13] M. Binder *et al.*, "On analyzing process compliance in skin cancer treatment: An experience report from the evidence-based medical compliance cluster (ebmc2)," in *International Conference on Advanced Information Systems Engineering*, 2012, pp. 398–413.

[14] J. Mundy and W. Thornthwaite, *The Microsoft data warehouse toolkit: with SQL Server 2008 R2 and the Microsoft Business Intelligence toolset*. John Wiley & Sons, 2011.

[15] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Conceptual modeling for ETL processes," in *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, 2002, pp. 14–21.

[16] M. Levene and G. Loizou, “Why is the snowflake schema a good data warehouse design?,” *Inf. Syst.*, vol. 28, no. 3, pp. 225–240, 2003.
[17] P. Stephenson, *Official (ISC) 2® Guide to the CCFP CBK*. CRC Press, 2014.
[18] S. Simoff, M. H. Böhlen, and A. Mazeika, *Visual data mining: theory, techniques and tools for visual analytics*, vol. 4404. Springer Science & Business Media, 2008.

Mohammed AIT DAOUD,
Phd in Computer Sciences – Faculty of Sciences Ben M'Sik, UH2C, Casablanca, Morocco (http://www.univh2c.ma/),
email: aitdaoud.mohammed@gmail.com,
Scopus Author ID: 57191254859
ResearcherID: V-6550-2018
ORCID: orcidID= https://orcid.org/0000-0002-8627-4429.

Khalil NAMIR,
Phd student in Computer Sciences – Faculty of Sciences Ben M'Sik, UH2C, Morocco, (http://www.univh2c.ma/),
email : namirkhalil.95@gmail.com,

Mohssine BENTAIB,
Associated Professor – Faculty of Sciences Ben M'Sik, UH2C, Casablanca, Morocco (http://www.univh2c.ma/),
email: m.bentaib@gmail.com,

Rachida IHYA,
Phd in Computer Sciences – Faculty of Sciences Ben M'Sik, UH2C, Casablanca, Morocco (http://www.univh2c.ma/),
email: rachida.ihya@gmail.com,
Soufiane. BOUITI,
Data Scientist – Association Maarif Centre, FSBM, Casablanca, Morocco (http://maarifcentre.org),
email: soufianebouiti@gmail.com,

Mohammed TALBI,
Dean – Faculty of Sciences Ben M'Sik, UH2C, Casablanca, Morocco (http://www.univh2c.ma/),
email: talbi.ordipu@gmail.com,