

Learning Analytics or Educational Data Mining? This is the Question...

Daniela Marcu

Ștefan cel Mare University of Suceava
Str. Universității 13, Suceava 720229
Phone: 0230 216 147
mdaniela.marcu@yahoo.ro

Mirela Danubianu

Ștefan cel Mare University of Suceava
Str. Universității 13, Suceava 720229
Phone: 0230 216 147
mdanub@eed.usv.ro

Abstract

In full expansion, a vital area such as education could not remain indifferent to the use of information and communication technology. Over the past two decades we have witnessed the emergence and development of e-learning systems, the proliferation of MOOCs, and generally the rise of Technology Enhanced Education. All of these contributed to generation and storage of unprecedented volumes of data concerning all areas of learning.

At the same time, domains such as data mining and big data analytics have emerged and developed. Their applications in education have spawned new areas of research such as educational data mining or learning analytics.

As an interdisciplinary research area Educational Data Mining (EDM) aims to explore data from educational environment to build models based on which students' behavior and results are better understood. In fact, EDM is a complex process that consists of a few steps grouped in three stages: data preprocessing, modelling and postprocessing. It transforms raw data from educational environments in useful information that could influence in a positive way the educational process.

According to Society for Learning Analytics Research (SoLAR) which took over the wording of the first International Conference on Learning Analytics and Knowledge, learning analytics is "the measurement, collection, analysis and reporting of data about learners and their contexts for purposes of understanding and optimizing learning and the environments in which it occurs" (Siemens, 2011).

This paper proposes a comparative study of the two concepts: EDM and learning analytics.

Due to certain voices in the scientific environment that claim that the two terms refer to the same thing, we want to emphasize the similarities and differences between them, and how each one can serve to raise the quality in educational processes.

Keywords : EDM; LA; Data Mining; Education.

1. Introduction

The educational community has an interest in the great potential of education. Why are researchers so enthusiastic about this? The answer is simple. Seeing the impact of applying data mining to exploiting large data volumes and analyzing data from areas such as the business environment, social media, and other scientific areas, we can think of the benefits for the education system. If we could adapt the methods of finding models in the data, used for analyzing the online activity of clients and social media users for the educational environment, we could get closer evidence of reality on the activities of the training system.

The widespread use of computer-based pre-university learning, the development of Web-based courses, are additional reasons for EDM and LA research.

Designing educational policies based on practical evidence provided by researchers can bring benefits to the educational system.

The exploitation of large volumes of data from different domains is done using specific techniques and methods. It helps to develop tools to facilitate progress in these areas.

The science of extracting useful information from large volumes of data is called Data Mining (DM) (Hand, Mannila & Smyth, 2001).

The concept is based on three key areas: *statistics*, *artificial intelligence* and *machine learning* (Figure 1).

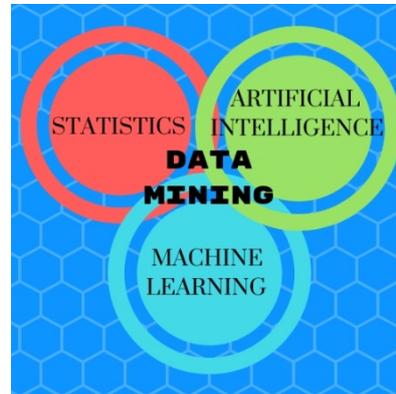


Figure 1. Data Mining

Initially, DM used statistical algorithms. Specific techniques such as decision trees, association rules, clustering, artificial neural networks, and others have been developed (Şuşnea, 2012).

Applying exploitation methods for educational system data to build models to better understand students' behavior and outcomes is named Educational Data Mining (EDM). Since data and education issues are different from those in other areas, classical DM methods have been improved and supplemented with EDM specific methods (Romero & Ventura, 2007). According to some authors, there are four areas of application of EDM aimed at: improving student modeling and domain modeling, e-learning and scientific research (Baker, 2012).

In order to better understand learning, data from pupils and from the educational environment is measured, collected and analyzed. This is the learning analysis and is a related field of EDM. Among the Learning Analytics (LA) methods we can list:

- content analysis
- discourse analysis
- analyzing the social dimension of learning (Ferguson & Buckingham Shum, 2012).

In the following sections we propose to detail relevant aspects about EDM and LA in order to provide viable arguments in a comparative study of the two concepts.

2. Educational Data Mining

Over the past 10 years, the field of research aimed to exploit the unique types of data from education has developed quite internationally. In 2011, in Massachusetts USA, the International EDM Working Group (established in 2007) created the International Society for EDM (online: <http://educationaldatamining.org/about/>). Romania is, however, at a pioneering stage in EDM. There is currently a growing interest in using computers in learning and Web-based training. With the rapid increase in the volume of learning software resources, the Romanian educational system also accumulates huge amounts of data from students, teachers, parents, libraries, secretariats, etc. Getting the information needed to build models to improve the quality of managerial decisions becomes one of the greatest challenges of the present.

Traditional research in the field of education is time-consuming and often non-ecological through the waste of material resources. Developing an experimental study, such as combating school absenteeism, involves firstly the selection of schools, teachers and pupils. It follows the definition of strategies that lead to the identification of sources of school stress, increasing the

motivation of students to attend classes, trust in school, family, and so on. However, the studies depend on context, class, geography, economic development, teacher-student relationships. Changing any parameter can lead to very different conclusions. Soon there may be new factors that could not be taken into consideration earlier in the demotivation of students towards school. Making traditional new studies for this topic involves the use of important temporal resources.

By comparison, EDM proves to be more efficient. The analysis of existing data in the educational system through the use of specific EDM methods allows the identification of new models for new contexts. An enormous advantage is that the same methods can be applied to different data generating specific results without the need for new analysis strategies.

More specifically, let's take the example of a course designed for web-based training (Romero, Ventura, De Bra, 2004). Traditionally, evaluating the effectiveness of a course is done by analyzing the results obtained by the student upon completion of the course, which does not necessarily lead to the improvement of the material or methods and teaching tools used for the future course versions. In fact, in the Romanian pre-university system, the updating of educational programs and educational resources does not present the periodicity expected by the society.

What would it be like the knowledge of EDM data exploitation? EDM methods aim at discovering correlation rules between course components (content, questions, various activities) and student activities. In the *Knowledge Discovery with Genetic Programming for providing feedback to the courseware author*, C. Romero, S. Ventura and P. Bra describe the four main steps in building a software based on EDM (Romero, Ventura, De Bra, 2004): development, use, discovering knowledge, improving

Other classification has three stages: preprocessing, data exploitation and post processing [3]. The cycle of these steps is illustrated in Figure 2.

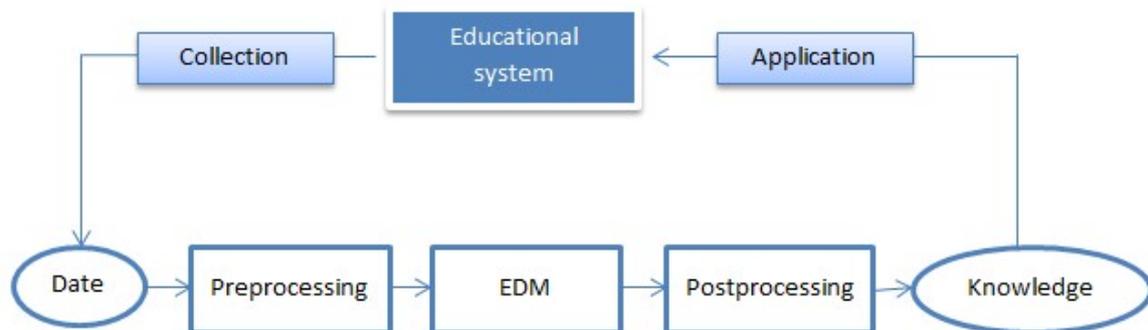


Figure 2. Stages of the process of converting data into information

If we refer again to the analysis of the efficiency of a course, in the first stage, the preprocessing is performed various operations such as:

- the teacher creates the content and provides information on pedagogical and methodological aspects
- the teacher creates course support
- the student uses the course
- the EDM software records information about: the student's time spent in the course, the sections visited, the scores obtained and other interactions
- the information collected is converted into data with a format appropriate for processing.

In the next step, EDM-specific algorithms are applied to obtain different correlation rules. The models will provide information in different formats for analysis: numerical results of the coefficients, tables, diagrams, correlation matrices (an example is illustrated in *Appendix 1 - Correlation matrix obtained with the DataLab application based on the results of the Olympiad of computer science*).

One of the most important rules for discovering knowledge is *if-else*. Several such rules can be defined in EDM: Association, Classification and Prediction (Klosgen & Zytkow, 2002).

The teacher will analyze the results of the analyzes and study the degree of achievement of the initial goals.

Depending on the conclusions, it may take the decision to improve the course and resume its evaluation process. This may prove to be a difficult process because opinions can differ significantly from one teacher to another in relation to the material and the way of interaction with the student the course offers.

3. Methods of data exploitation

There are currently a wide variety of methods of exploiting data in the education system. These can be categorized into two broad categories according to the ways to achieve the objectives:

- predictive: Prediction, Classification, Regression, Outlier Detecting
- descriptive: Clustering, Determination of association rules, Discovery of data for human judgment (Sasu, 2014).

Many of these are general DM methods: prediction, classification, grouping, exploitation of texts and others. But there are also specific EDM methods such as nonnegative matrix factorization and Knowledge tracing (KT) (Romero & Ventura, 2012). Here are some of these:

Prediction

The method can be used in education to predict students' behavior and outcomes. It is based on the creation of predictive models. In the training phase, they learn to make predictions about a set of variables called predictors by analyzing them in combination with other variables. Once the enrollment phase is completed, the patterns can be applied to the data sets for which the prediction is to be applied. It is known the study by Baker, Gowda, Corbett - *Automatically detecting the student's preparation for future learning: help use is key* (Baker, Gowda & Corbett, 2011). The authors create a tool for automatically predicting a student's future performance on the basis of establishing positive or negative correlations between various features such as: student test results, time spent in response, time elapsed between receiving a clue and typing the answer, and others. It is experienced on a group of students, and then applied to another group. The results are then compared to those obtained using the Bayesian Knowledge Tracing (BKT) model.

Classification

The method involves building a predictive model. The data in the training set is characterized by certain attributes. The model must identify belonging to a class based on the set of attributes. Suppose we built an educational software as an interactive game for a given theme. Based on user attributes such as age, gender, geographic area, duration until the game is completed, number of attempts we can build a classifier, and determine the user's belonging to a specific class. The model will learn to identify students. The analyzes can provide information on the need to use this educational method for certain age groups, interests and education.

Methods that use the classification are: decision trees, neural networks, bayesian classifications, and others.

Clustering

The method involves building patterns that identify data clustering after certain similarities. For the model to provide quality predictions, the similarities inside class must be maximized and similarities between classes minimized.

The use of this method in Romanian high school education could aim at grouping pupils according to the pupil's learning style (auditory, visual, practical - kinesthesia) based on the analysis of behavior in relation to certain educational products and pupils' characteristics. The prediction of such a model could lead to an effective recommendation of how to learn educational content. Thus, the instructional process could be carried out efficiently in relation to the learning particularities of each student. At present, there is an attempt to unfold the lessons in a way appropriate to the

students' learning styles, but the reality is that identifying learning styles is superficial. The results of the questionnaires are attached to the class catalog, but this does not lead, in most cases, to the improve teaching methods and techniques used in the lesson. In the absence of clear alternatives, the teacher has to improvise.

The method is successfully used in the detection of plagiarism (Text Mining) and is also applied in the educational sphere.

Outlier Detection

The method involves creating patterns that detect data that have different features than others. In Romanian education, this method could be used to detect students with content assimilation problems, or those with aberrant behavior.

In general, not only one EDM method is used in case studies. Outlier Detection methods can be used, for example, with data clustering techniques and decision tree classification as presented in the study by Ajith, Sai and Tejaswi (2013) - *Evaluation of student performance: an outlier detection perspective* (Ajith, Sai & Tejaswi, 2013). The study aims to identify learners with special learning needs to reduce the school failure rate. Input data are collected from: participation in student lessons, tests, notes on initial tests. In order to achieve the proposed objective, they try to find models for classifying students who will be helpful in setting up study groups.

At present, in Romania, students in the high school education of state do not have the opportunity to trace the course matter in other groups than the classes they belong to. Moreover, pupils diagnosed as having special educational needs participate in classes with other colleagues. The teachers create for them specially programs. Then the courses are held by under the guidance of a single teacher who does not have any pedagogical and methodical experience related to the learning situation! There are special requirements for conducting the educational process. This based on grouping students within the same educational space within the same timeframe to go through different course materials. In the absence of a proper classification, alternative methods and means, and teachers with such experience, things happen more or less in a manner that leads to the best results.

Discovery with Models

Discovery with Models is the fifth category presented in Baker's Taxonomy (Baker, 2012). It is also one of the most widely used methods of data exploitation in the field of education. It is based on the use of a previously validated model as a component in analyzes that use prediction or exploitation of relationships in new contexts (Baker & Yacef, 2009). In this way information on educational materials that contribute most to educational progress can be obtained. A study carried out by Beck and Mostow in 2008 - *How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students* (Beck & Mostow, 2008) - on the analysis of different types of learners demonstrates that the method supports identifying relationships between student behavior and characteristics of variables used.

Nonnegative Matrix Factorization (or Decomposition)

There are several algorithms used for factoring the nonnegative matrix. This transforms (decomposes, factorizes) a matrix V into two W and H matrices with the property that they all have non-negative elements. This is very useful in applications such as determining the effectiveness of an evaluation system in which matrices contain elements related to: exams, abilities, and items. Matrix V is obtained from the product of the two smaller matrices as can be seen in Figure 3. ("Non-negative matrix factorization", 2019).

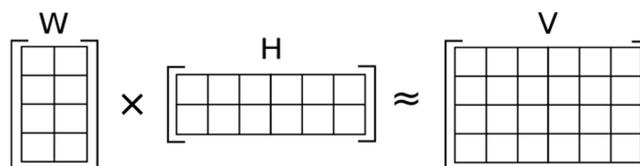


Figure 3. Illustration of approximate non-negative matrix factorization. Source: wikipedia.org

We propose to study the evaluation of two specific abilities defined on the columns of the matrix W for 4 work requirements (items), defined in the W matrix on the four lines.

Matrix H will contain two lines representing the two abilities and 6 columns representing the assessed students.

The result will be recorded in Matrix V that has 4 lines for each of the 4 items and 6 columns for each of the 6 students.

A value of 1 in the W matrix indicates the need for a certain skill (Figure 4) (Desmarais, 2012).

$$\begin{array}{c}
 \mathbf{W} \\
 \text{skills} \\
 \begin{array}{|c|c|}
 \hline
 0 & 1 \\
 \hline
 1 & 0 \\
 \hline
 1 & 0 \\
 \hline
 1 & 1 \\
 \hline
 \end{array} \\
 \text{Items}
 \end{array}
 \times
 \begin{array}{c}
 \mathbf{H} \\
 \text{students} \\
 \begin{array}{|c|c|c|c|c|c|}
 \hline
 1 & 1 & 1 & 0 & 1 & 1 \\
 \hline
 0 & 0 & 1 & 1 & 0 & 0 \\
 \hline
 \end{array} \\
 \text{skills}
 \end{array}
 \approx
 \begin{array}{c}
 \mathbf{V} \\
 \text{students} \\
 \begin{array}{|c|c|c|c|c|c|}
 \hline
 0 & 0 & 1 & 1 & 0 & 0 \\
 \hline
 1 & 1 & 1 & 0 & 1 & 1 \\
 \hline
 1 & 1 & 1 & 0 & 1 & 1 \\
 \hline
 1 & 1 & 2 & 1 & 1 & 1 \\
 \hline
 \end{array} \\
 \text{items}
 \end{array}$$

Figure 4. Non-negative matrix factorization - example

The first item requires the ability 2, $W [1] [2] = 1$. Only the 2 and 3 students have the ability 2, so item 1 will not be promoted by students 1, 2, 4 and 5.

To promote Item 4 both skills are required. Only one of the candidates will promote this item with the maximum score.

Using computerized analysis methods, interpretations can be obtained in a much shorter time and with great accuracy because machines are faster and more accurate than humans.

4. Learning Analysis (LA)

Learning is the product of an interaction between learners and the learning environment, between among students / educators / teachers and others (Elias & Lias, 2011).

The evaluation of learning, in the traditional sense, is based on the evaluation of student / pupil outcomes. This involves assessing knowledge but also trying to answer questions such as: how well this student needs, how can be improved, how to change the course interface to make it more accessible. At present, especially in the pre-university system, learning evaluation is based on questionnaires. Obtaining feed-back is lasting because the non-automatic data processing takes time and the analysis possibilities are quite limited.

The desire to improve the quality of learning and assessment in the educational system is increasing at the international level, but also in our country. Traditional systems are confronted by huge amounts of data and their diversity. Learning Analytics (LA) attempts to answer questions about how this data can be used and how it can be transformed and analyzed to provide useful information that can give value to the learning process (Liu & Fan, 2014).

In 2011, at the first International Conference on Learning Analysis (LAK 2011), the definition of the new research area, LA, was adopted as: "*learning analysis is the measurement, collection, analysis and reporting of pupils and students and about the context of learning, in order to understand and optimize learning and its environments*" (Siemens, 2011).

Data analytics was first used in sales, also called Business Intelligence. This branch of research uses computer techniques to synthesize huge amounts of data and turn them into powerful tools for making the best marketing decisions.

With the development of Web technologies, a branch of data analysis research, Web Analytics, has been developed. Web Analytics tools collect data about users of a site and report on their behavior. This leads to a better understanding of customers and making the best decisions to improve your browsing experience and to keep visitors to the site.

Learning Analytics borrows tools and methods used in Business Intelligence and Web Analytics to analyze educational data.

At present, many universities, companies, and organizations are developing learning platforms for both students and lifelong learning. An enormous advantage of these is to personalize the learning experience and adapt it to the physical deficiencies of the learners.

In a research conducted by the New Media Consortium and the EDUCAUSE Learning Initiative in 2016, areas that will have a particular impact on university education globally by 2020 are identified. One of these is Learning Analytics. In the research report LA is defined as an application in the educational field of Web Analytics. It focuses on the collection and detailed analysis of student interactions with online learning platforms (Johnson, Adams Becker & Cummins, 2016).

A free example of a Web Analytics tool is provided by Google and is called Google Analytics. It provides sophisticated user behavior on a website and provides its administrators with reports about:

- time spent on the site;
- most visited pages;
- the number of users in a given period and how many of them are new customers;
- demographics about users
- devices used for navigation;
- traffic sources and more (Thushara & Ramesh, 2016).

With these reports, can create additional features, add more interesting content, enhance interactivity, customize the interface of the application based on the devices used for viewing.

In the following figures (5,6,7) there are illustrated sections of various reports provided by this tool for the site <https://www.modinfo.ro> - a site dedicated to the preparation of the students from the Romanian high schools at the course of computer science.

Figure 5 provides a diagram representation of the number of visitors per page of the site. We note that students are looking for baccalaureate content (*bac.php*), admission to faculty (*admission.php*) and additional training for performance (*cex.php*).

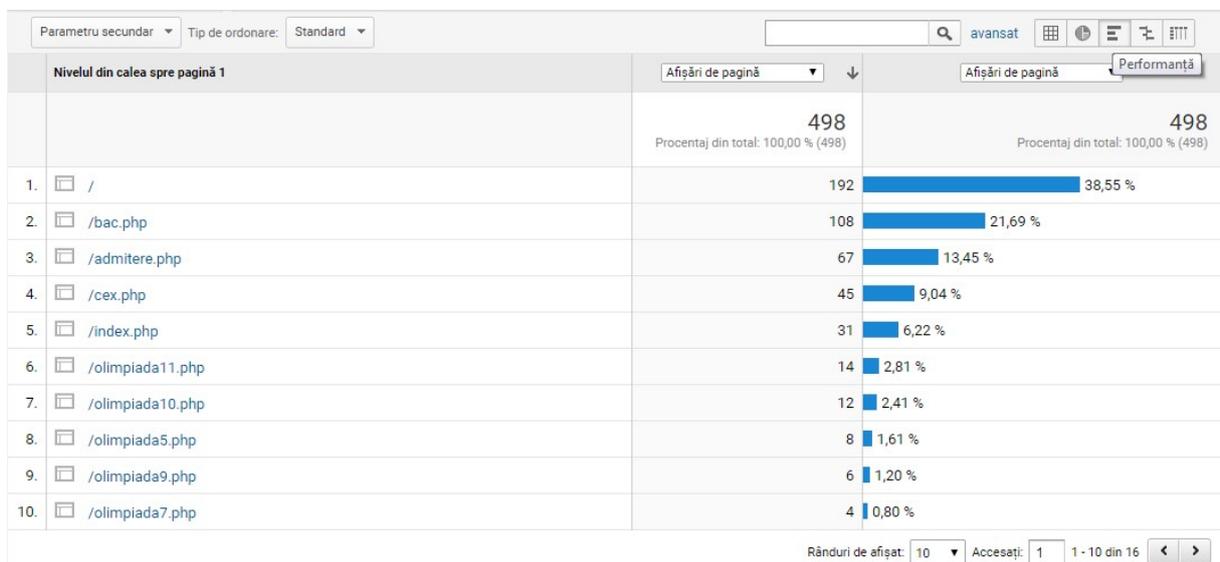


Figure 5. User preferred content

Figure 6 represents the percentage of visitors to the site over a fixed period, by age category. It can be seen that most users are aged between 25 and 34 years. For administrators, given the period under review, this reveals their student's preoccupation for to prepare for the Computer Programming Exam.

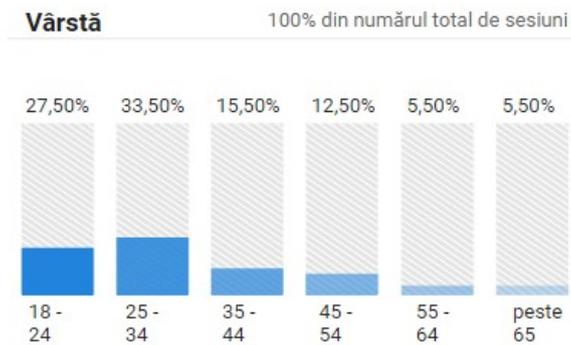


Figure 6. Demographics and interest categories - Age of users

Figure 7 provides information on analyzing the active presence of a specific user on a site within a selected time interval.

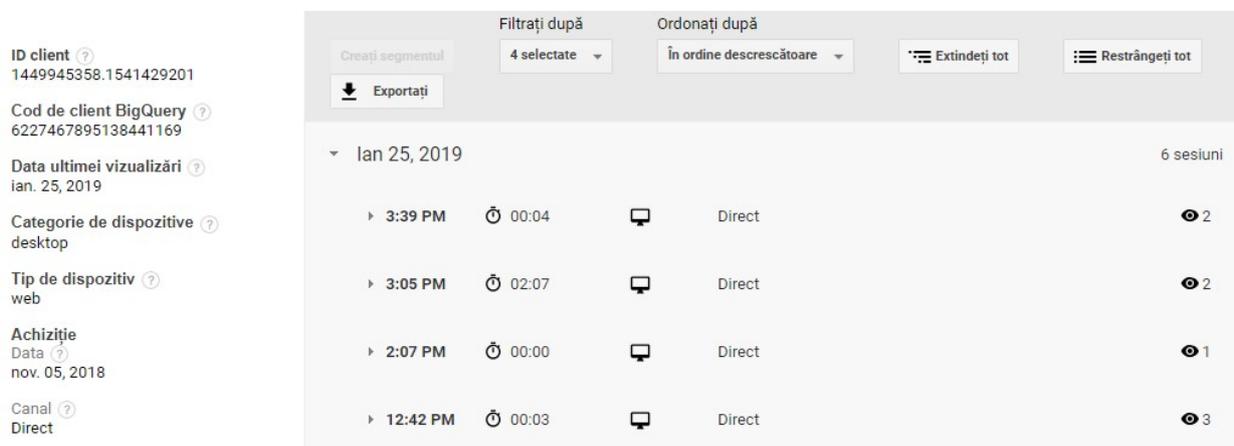


Figure 7. Behavior of a user on the site within a selected time range

Choosing how to use and constructing analytics tools starts from the choice of quantifiable indicators that have to be defined according to the proposed objectives. Examples of such indicators for the educational environment:

- the time required to successfully complete an exercise
- the number of attempts until successful completion of a task
- the number of assessments given through a communication tool within the course and others.

4.1. Learning Analytics methods

Methods used for learning analysis include:

- content analysis: pupils' essays
- discourse analysis: student exchange of messages. The quality of the expression is analyzed.
- student availability in relation to learning: Students interested in the topic will ask questions, access links to supplementary resources
- social learning analysis: exploits student interactions and motivational learning.

LA uses some methods of data mining as EDM. They can be classified in: Prediction, Clustering, Relationship mining, Discovery with models, Distillation of Data for Human Judgment (Nunn, Avella, Kanai, & Kebritchi, 2016).

We will briefly describe the methods that have not already been presented in the previous section.

Relationship mining

It's a method that uses algorithms to find association rules to detect, for example, mistakes made by students when solving a set of exercises. Based on the associations made, one can predict a certain behavior of the student depending on the hypothesis of solving the problem from which he starts. Thus, the teacher or course manager can intervene in order for the pupil / student not to be mistaken. There can be found, for example, relationships between other activities of the student (playing on the computer, talking to a chat room colleague) while solving his or her work tasks and erroneous answers (Baker, Corbett, Koedinger & Wagner, 2004).

Distillation of Data for Human Judgment

This method includes statistics and visualization techniques that help people understand data analytics. The method is the basis for the creation of many useful tools that provide clear analysis that can be quickly understood by unrelated users.

An example is the formation of a map to group learners by the amount of heat emanating from their bodies during learning the instructional material. This can be done with sensors mounted on the body. The analysis provides real-time learning about learning performance indicators (Merceron, 2015).

5. Learning Analytics or Educational Data Mining?

Educational Data Mining is a new field of research. It is based on the models, methods and algorithms built for DM. However, there are also specific methods of applying DM in education. The main purpose of EDM is to explore large sets of data from the educational system to create knowledge-extraction models from the data. The main objective is to provide useful information to education decision makers about existing correlations between sets of data that provide a deeper understanding of the educational needs of students and the system as a whole (de Almeida Neto & Castro, 2017).

Learning Analytics is a newer field of research. It is based on data analysis techniques in Business Intelligence. LA uses highly sophisticated analysis tools and predictive models to improve learning. Most applications using LA have been created for the university system and are dedicated to early detection of concrete problems such as the risk of abandoning a course by certain students. LA also uses the expertise of other research areas, such as EDM and Web Analytics, with the same objectives of predicting learning outcomes and providing useful information for improving the quality of the learning process (Elias & Lias, 2011).

EDM is at the intersection of areas such as artificial intelligence, machine learning, education, and statistics.

Figure 8 shows the LA as an interdisciplinary subdomain of Business Intelligence, Statistics and Education.

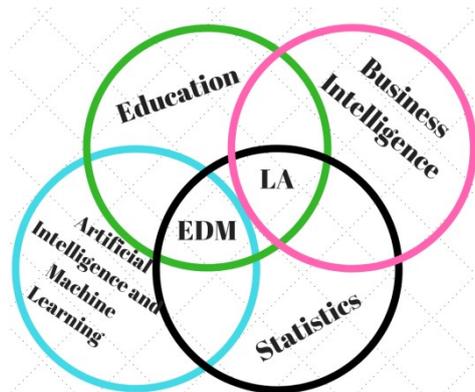


Figure 8. Educational Data Mining and Learning Analytics

The two new areas of research are quite similar in terms of the aims pursued and methods used, but there are also some significant differences between them. Some of the most important resemblances and differences between EDM and LA are shown in Tables 1 and 2.

Table 1. Similarities between EDM and LA

EDM	LA
Both areas contribute to improving the quality of education and alternative education systems as well.	education policies in schools and universities, but in
It is a new field of research. In 2011, in Massachusetts USA, the International Working Group on EDM (established in 2007) created the International Society for EDM.	The definition of this new field of research was adopted in 2011 at the first International Conference on Learning Analytics (LAK 2011).
It is based on the exploitation of large data collections.	It is based on analysis of large data collections.
It is based on the formulation of specific research objectives and the establishment of useful indicators in relation to them.	
It requires the development of specific tools to apply the methods in solving the proposed goals.	
They use methods of data mining.	

Table 2. Differences between EDM and LA

EDM	LA
The main objective is to find models for extracting knowledge from data sets.	Apply Business Intelligence tools and predictive models to analyze data from educational applications.
It focuses on the development of models using data mining techniques to understand students' behavior, their relationship with other factors involved in the education system for to improve the educational system.	Analyzes data from students to improve issues such as the risk of abandoning the course by them or improving the quality of instructional material.
The most used techniques are: classification, prediction, Clustering, Bayesian model, Relationship mining, Discovery with models.	The most commonly used techniques are: Statistics, Visualization, Sentiment analysis, Discourse analysis and so on.
The origin of EDM is in the DM domain.	The origin of LA is in Business Intelligence and Semantic Web.
EDM focuses on the development of exploitation techniques to answer questions such as: " <i>How we can valorize the exploitation of education data sets?</i> " (Liu & Fan, 2014).	LA is focused on data analysis and tends to answer questions like: " <i>how can learning be optimized?</i> " (Liu & Fan, 2014).
EDM uses tools such as: RapidMiner, DataShop, DataLab (see Appendix 1), WEKA, Orange, KNIME, NLTK, TANAGRA, SPSS and others.	LA uses tools such as: GEPHI, EgoNet, SoNIA, SocNetV, SNAPP, Clever, PASS (Slater, Joksimović, Kovanovic, Baker & Gasevic, 2016).

6. Conclusions

The first successful DM application was achieved in sales. Customer data analysis models have been developed (where to buy online, what sites they visit, when they buy, how much they buy, etc.). The results obtained from the analysis proved to be accurate and useful in adopting the most appropriate marketing strategies, problem identification and solution finding. Another area in which DM offers very good results is social media. There is currently an enthusiasm among scholars in the educational community in applying DM techniques to education data.

In Romania, all the actors involved in the educational system (researchers, managers, pupils, parents, teachers, representatives of various organizations) have for some time discussed the opportunity of adapting the curricula to the specific cultural and economic situations of the present so that education becomes more effective and really come to support pupils' training. But the process proves to be slow and almost impossible in the absence of a realistic, evidence-based design. We are convinced that EDM and LA can make a significant contribution to finding the best solutions to enhance the quality of training.

In recent years, universities in the US in particular but also in other parts of the world, including Romania, have begun to explore and develop applications that use EDM and LA to improve the quality of the educational process. Research projects aim, on the one hand, to find

models for extracting knowledge from huge data sets and, on the other, to apply analysis methods to solve problems for which certain predictions have been made. One of these refers to students with risk of abandonment, researched by the University of Alabama based on the analysis of demographic sets of data provide of students (Nunn, Avella, Kanai, & Kebritchi, 2016).

Romania is in a pioneering phase in the development of research projects in the field of data mining and analyzes of learning in the university system. We do not currently have data on the existence of such projects, in the state pre-university system in our country, which encourages us to make the following statements:

- it is necessary to unify the results of the research and the studies carried out in these fields
- EDM and LA are areas of real interest for the educational future research.

Appendix 1. DataLab Application

DataLab is a useful tool in manipulating and interpreting data (University, 2019).

In this appaendix is presents the correlation matrix. In order to calculate it, it is necessary to observe several steps: selecting correlation statistics, coding variables, treating missing data and presentation.

Data were taken from the table with structure shown in Figure 9.

Row ▲	Student	cartier	medalion	numar	TOTAL
1	601	100.0	50.0	30.0	180.0
2	607	100.0	60.0	80.0	240.0
3	617	100.0	0.0	35.0	135.0
4	620	100.0	75.0	100.0	275.0
5	634	100.0	0.0	0.0	100.0
6	635	100.0	80.0	80.0	260.0
7	638	100.0	60.0	75.0	235.0
8	640	100.0	45.0	55.0	200.0
9	700	100.0	60.0	80.0	240.0
10	602	90.0	100.0	100.0	290.0
11	613	70.0	45.0	55.0	170.0
12	639	60.0	65.0	85.0	210.0
13	630	60.0	90.0	50.0	200.0
14	615	60.0	70.0	60.0	190.0
15	611	50.0	60.0	0.0	110.0
16	645	46.0	60.0	75.0	181.0
17	631	40.0	60.0	0.0	100.0
Column Average		42.9828	41.5517	50.9828	135.5172
Standard Deviation		37.4023	30.5925	31.9745	64.3538

Figure 9. Data set for Correlation Matrix

The correlation matrix is an array that displays correlations calculated between different variables (Figure 10).

The matrix is symmetrical.

The data in the original table is fictitious. The name of the problems (*cartier*, *medallion*, *numar*) is real. The table contains the fields:

- *student* - ID with which a student is participating in the National Computer Science Olympiad
- *cartier*, *medalion*, *numar* are the names of the three issues offered to competitors and will be completed with the scores of the candidates that are natural numbers in the [0..100] range.
- *TOTAL* is the total score of each candidate for all three issues.

The variables between which correlations are calculated are: *cartier*, *medalion*, *numar*, *TOTAL*.

For the calculation of the correlation coefficient the minimum and maximum thresholds of 0.2 and 0.7 were used respectively.

Everything outside these thresholds is marked in the matrix. The value of the correlation between a variable x and itself is 1.

The negative correlation coefficient between the *medalion* and the *cartier* variable expresses the probability that a student who better solves the *medalion* problem, solve the *cartier* problem less effectively.

Correlation Matrix 

	cartier	medalion	numar	TOTAL
cartier	1			
medalion	-0.0685	1		
numar	0.2883	0.1285	1	
TOTAL	0.6919	0.4994	0.7255	1
Total Score 	0.6919	0.4994	0.7255	1

These values are a measure of the correlation between each item and Total Score

Figure 10. Correlation Matrix

The calculation of the correlation coefficient was done with the Pearson algorithm ¹.

References

- Ajith, P., Sai, M. S. S., & Tejaswi, B. (2013). *Evaluation of student performance: an outlier detection perspective*. Int. J. Innov. Technol. Explor. Eng, 2(2), 40-44.
- de Almeida Neto, F., & Castro, A. (2017). A reference architecture for educational data mining. *2017 IEEE Frontiers In Education Conference (FIE)*.
- Baker RSJd, Gowda SM, Corbett AT. (2011). *Automatically detecting a student's preparation for future learning: help use is key*. In: Fourth International Conference on Educational Data Mining. Eindhoven, The Netherlands; 179–188.
- Baker RSJd, Yacef K. (2009). *The state of educational data mining in 2009: a review and future visions*. J Edu Data Min, 3–17.
- Baker, R., Corbett, A., Koedinger, K., & Wagner, A. (2004). Off-task behavior in the cognitive tutor classroom. *Proceedings Of The 2004 Conference On Human Factors In Computing Systems - CHI '04*.
- Baker, R.S.J.d. (in press) Data Mining for Education. To appear in McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.
- Beck, J.E. and Mostow, J. (2008). *How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students*. In Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 353-362.
- Desmarais, M. (2012). Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter*, 13(2), 30.
- Elias, T. and Lias, T. E. (2011). *Learning Analytics: Definition, Processes and Potential*.
- Ferguson, R. and Buckingham Shum, S. (2012). Social Learning Analytics: Five Approaches. Proc. 2nd International Conference on Learning Analytics & Knowledge, (29 Apr-2 May, Vancouver, BC). ACM Press: New York
- Hand, D., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. *The MIT Press*, 19(2), 183-184.
- Johnson, L., Adams Becker, S. & Cummins, M. (2016). *NMC Horizon Report: 2016 Higher Education Edition*. The New Media Consortium. Texas, Austin, USA: 38.
- Klosgen, W. and Zytkow, J. M. (2002). *Handbook of Data Mining and Knowledge Discovery*. New York, NY: Oxford University Press.

¹ Pearson's R is a measure of the linear correlation between two variables X and Y. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s. It has a value between +1 and -1.

- Liu, Q., & Fan, G. (2014). Using Learning Analytics Technologies to Find Learning Structures from Online Examination System. *2014 International Conference Of Educational Innovation Through Technology*.
- Merceron, Agathe. (2015). *Educational data mining/learning analytics: Methods, tasks and current trends*. 1443. 101-109.
- Non-negative matrix factorization. (2019). Retrieved 25 January 2019, from https://en.wikipedia.org/wiki/Non-negative_matrix_factorization
- Nunn, S., Avella, J., Kanai, T., & Kebritchi, M. (2016). Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review. *Online Learning*, 20(2).
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems With Applications*, 33(1), 135-146.
- Romero, C., & Ventura, S. (2012). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery*, 3(1), 12-27.
- Romero, C., Ventura, S., De Bra, P. (2004). *Knowledge discovery with genetic programming for providing feedback to courseware author*. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research* 14(5), 425–464.
- Sasu, L. (2014). Introducere în Data Mining Curs 1: Prezentare generală. Retrieved 9 September 2019, from <https://www.slideshare.net/lmsasu/curs1-data-mining>
- Siemens, G. (2011). About: *Learning analytics and knowledge*. In 1st International Conference on Learning Analytics and Knowledge.
- Slater, S., Joksimović, S., Kovanovic, V., Baker, R., & Gasevic, D. (2016). Tools for Educational Data Mining. *Journal Of Educational And Behavioral Statistics*, 42(1), 85-106.
- Șuşnea, E. (2012). Utilizarea tehnicilor data mining într-un sistem educațional de tip e-Learning. București: Pro Universitaria.
- Thushara, Y., & Ramesh, V. (2016). A Study of Web Mining Application on E-Commerce using Google Analytics Tool. *International Journal Of Computer Applications*, 149(11), 21-26.
- University, C. (2019). DataLab Tools + Resources - DataLab - Carnegie Mellon University. Retrieved 9 September 2019, from <https://www.cmu.edu/datalab/tools/index.html>



Daniela Marcu (b. February 14, 1973) received her BSc in Physics (1997), postgraduate studies in Computer Science (2001), PhD student in Computers and Information Technology. Now she is professor of computer science at the "Dimitrie Cantemir" Economic College in Suceava and assistant (2018-2019) at the Faculty of Electrical Engineering and Computer Science at "Stefan cel Mare" University in Suceava, Romania. Her current research interests include different aspects of Big Data applied in the educational field. She has (co-) authored 3 books that have been approved by the Romanian Ministry of Education to be used in high schools in Romania in the 12th grade, real and human profile, and several (co-) books. She is the coordinator of the Training Center for Excellence in Computer Science. She was a member of the National Committee of the Informatics Olympiad. She is currently a member of several national committees for the organization of IT competitions. Over the past 8 years, she has trained many students that have been awarded with a bronze medal at Balkan Olympiad in Informatics (2018) and one at Infomatrix International Computer Project Competition (2009) and also more than 60 medals at international and national programming and software competitions.



Mirela Danubianu (b. July 13, 1961) has obtained the B.S. and M.S. degree in Computer Science from University of Craiova in 1985, and the PhD. degree in Computer Science in 2006 from “Stefan cel Mare“ University of Suceava. She has also obtained the B.E. degree in Economics from University of Craiova in 2001. Currently, she is Associate Professor and Head of the Computers Department at “Stefan cel Mare University” of Suceava. She is the author/co-author of 5 books, 7 chapters and more than 100 papers which have been published in journals and presented at different conferences. Her current research interests include databases theory and implementation, modern data architectures, data analytics, application of Data Science in economics, education and healthcare.