Bulletin of the *Transilvania* University of Braşov
Series V: Economic Sciences • Vol. 8 (57) No. 1 - 2015

Data mining techniques used to analyze students' opinions about computization in the educational system

Nicoleta PETCU[1]

Abstract: *Both the educational process and the research one, together with institutional management are unthinkable without the information technologies. Through them, one can harness the work capacity and creativity of both students and professors. The aim of this paper is to present the results of a quantitative research regarding: the scope of using computers, the importance of using them, faculty activities that involve computer usage, number of hours students work with them at university, Internet and web-sites usage, e-learning platforms, investments in technology in the faculty and access to computers and other IT resources. The major conclusions of this research allow us to propose strategies for increasing the quality, efficiency and transparency of didactic, scientific, administrative and communication processes.*

Key-words: *computerization of educational process, quantitative, statistical tests, decision tree.*

1. Introduction

Increasing the usage of information technologies would play a major role in connecting the education system to the requirements of a society based on knowledge. In the current situation, where the fast evolution of information technology makes way in all areas of work, the education system needs state of the art informatics systems. They, in its turn, have to adapt on an ongoing basis to the dynamic of nowadays technology. In order to achieve the computerization of the educational process some new organizational structures, new management methods and new ways to use information technology are needed (Strajeri, 2009). Acquiring new knowledge and using it in highly qualified training relies on recent discoveries from the area of information technologies and communication. For the European countries *the knowledge society* has required a computerized educational system. A report of the European Commission from 2001 mentioned that: „*the embodiment of information and communication technologies in the educational systems is a process*

[1] Transilvania University of Braşov, nicoleta.petcu@unitbv.ro

that will have major implications for organizing education and teaching methods, on the long run." (European Commission, Annual Report 2000/01). This has moved forward the development and validation of strategies and instruments that can improve the efficiency of the educational system for a larger number of beneficiaries (Cucos, 2006,30). "All EU states are in the process of integrating the digital in the educational system, e-books and digital pedagogy are the new challenges for the ministers of education. The computerization of the education system is a pedagogic strategy adapted/adaptable at the policy level. This happens in the context of a postindustrial, computerized cultural model. From a technological perspective one has to address the managerial issue, at a system level (monitoring the resources, designing the curriculum for the educational plan, global evaluation, the institutionalization of a databases etc) and the process (the curriculum design of the programs and school books, stimulating the intensive learning/ self-learning, development of assessment tests)' (Cucoş, 2006, 28).

2. Quantitative research – general data

The current paper aims to identify students' opinions regarding: the importance of using computers at the university, the activities for which they use them, the number of hours at which students use computers at the university, the usage of Internets and different websites, the e-learning platform, investments in technology in different faculties.

Data collection was made through an online, structured survey made out of 26 questions. This took the form of a link that was then sent to all the faculties from the Transilvania University. It resulted in 192 filled-in surveys. The data from these surveys was then analyzed using SPS and SPSS Clementine.

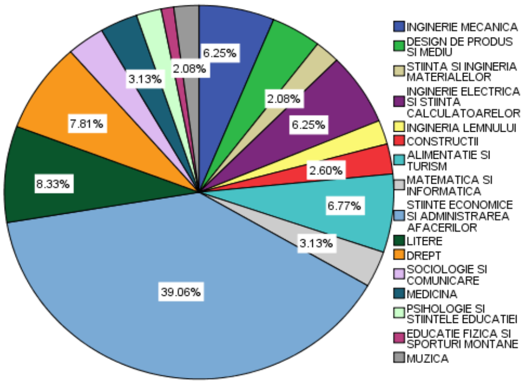

Fig.1. *Respondents distribution on faculties*

Most of the respondents came from the Faculty of Economical Sciences and Business Administration (39,13%) and from Law and Letters (16,1%)

2.1. General hypothesis of the research

1. The majority of students use computers at university for: solving some exercises, writing documents (Word), Power Point presentations, and data collection from the Internet.
2. Generally, students use information from the Internet to write their projects because they have access to a larger volume of information and they save time.
3. A considerably high percentage of the students consider that the following websites are the best for project writing: scribd.com and regielive.ro
4. Most of the students have access to the Internet and use the e-learning platform.
5. A relatively low percentage of students have noticed that the university has made investments in IT technology.

3. Results

Among the students that have either courses or seminars where they use computers: 38,40% use them to solve assignments or to make PowerPoint presentations, 18% use them for writing documents, 15,10% use them to create databases, 10,9% for Internet searches, 10% use them for programming, 7,5% for printing, 48% have said that they use computers for at least two subjects, 90,22 % use the Internet to write an essay or a project and the most accessed site is www.google.com. The new technologies and instruments help students to learn faster and easier according to 75% of the respondents. This means that ¾ of them know and use the E-learning platform. Students' opinions regarding this platform have revealed the following: 81% have been satisfied and very satisfied of being able to see their grades, 42,5% have been satisfied and very satisfied of being able to see the courses, 19,2% have been satisfied and very satisfied of being able to communicate with their professors, 64,4% have been satisfied and very satisfied of being able to see the payment situation.

Considering students' recommendation for using computers at the university, they have marked as important or very important the following issues (cumulated percentages): one should give students more basic knowledge about computers 25,5%, one should have enough computers/projectors at the faculty level 51%, one should acquire licensed software 40,1%, students should be allowed to use computers even when not during classes 45,8%, the faculty should stimulate students' interest for developing practical abilities related to computers and IT 37,5%.

Hypotheses testing

Hypothesis no. 1

H_0: At the university level at most 60% of the students take part at courses/seminars where computers are used.

H_1: At the university level more than 60% of the students take part at courses/seminars where computers are used.

	N	Mean	Std. Deviation
Do you have courses/seminars where you use computers?	192	.81	.395

One-Sample Test

	Test Value = 0.6					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Do you have courses/seminars where you use computers?	7.263	191	.000	.207	.15	.26

We will accept the alternative hypothesis H1 (Sig. (2-tailed) < 0,05): with a probability of 95% one can say that, at the university level, more than 60% of the students take part in courses and seminars where they use a computer.

Hypothesis no.2

H_0: At most 80% of the students use the e-learning platform.

H_1: More than 80% of the students use the e-learning platform.

	N	Mean	Std. Deviation
Do you use the e-learning platform: http://portal.unitbv.ro ?	192	.76	.428

One-Sample Test

	Test Value = 0.8					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Do you use the e-learning platform: http://portal.unitbv.ro ?	-1.282	191	.202	-.040	-.10	.02

We will accept the null hypothesis (Sig. (2-tailed) > 0,05): with a probability of 95% one can say that, at the university level, at most 80% of the students use the e-learning platform.

Hypothesis no.3

H_0: The access to the e-learning platform is independent from having access to the university's IT resources.

H_1: There is a dependency between accessing the e-learning platform and having access to the university's IT resources.

			Could you use the IT resources/computers at the faculty when you needed to?		Total
			Nu	Da	
Do you use the e-learning platform: http://portal.unitbv.ro?	No	Count	0	22	22
		Expected Count	.8	21.2	22.0
	Yes	Count	4	81	85
		Expected Count	3.2	81.8	85.0
Total		Count	4	103	107
		Expected Count	4.0	103.0	107.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.075[a]	1	.300		
N of Valid Cases	107				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is .82.

b. Computed only for a 2x2 table

We will accept the null hypothesis H0 (Sig. (2-tailed) > 0,05), the two variables are independent.

Hypothesis no. 4

H_0: At most 70% of the students from *Transilvania* University use the Internet to write their projects.

H_1: More than 70% of the students from *Transilvania* University use the Internet to write their projects.

	N	Mean	Std. Deviation
Do you use the Internet when writing a project?	192	.91	.292

One-Sample Test

	Test Value = 0.7					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Do you use the Internet when writing a project??	9.779	191	.000	.206	.16	.25

We will accept the alternative hypothesis H1 (Sig. (2-tailed) < 0,05),), with a probability of 95% one can say that, at the university level, more than 70% of the students use the Internet when writing a project.

Decision trees used in multidimensional classifications

The decision and classification trees are one of the main Data Mining techniques. Analyzing decision trees allows forecasting the affiliation of some objects/instances to different categories, starting with their rates depending on one or more predictors[2].

Objective – forming homogeneous subgroups from the point of view of the Y dependent variable.

The decision and classification trees can be:

- Classification trees, when the prediction result is the data affiliation class;
- Regression trees, when the forecast result can be considered a real number (oil price, the price of a house);
- CART (C&RT) Classification And Regression Tree (Breiman, 1984) – combines the two cases mentioned above.

Growing a decision tree - The root node divides the collectivity into groups depending on the Y dependent variable. The first segmentation variable is chosen and it results in son-nodes. If a son-node is homogenous, it becomes a pure node (a leaf node); if it is not, a new segmentation variable is chosen and the process continues until the tree ends in leaf nodes.

Segmentation (partition) evaluation measures – *Information theory*

One of the evaluation measures for segmentation (partition) could be ***Informational Gain.***

Shannon Entropy – the quantity (amount) of information for knowing Y values

$$E(Y) = -\sum_{i=1}^{p} \frac{n_{i.}}{n} \cdot \log_2\left(\frac{n_{i.}}{n}\right)$$

Conditional Entropy – the quantity (amount) of information for knowing Y values that are conditioned by the values of X

$$E(Y/X) = -\sum_{j=1}^{q} \frac{n_{.j}}{n} \sum_{i=1}^{p} \frac{n_{ij}}{n_{.j}} \cdot \log_2\left(\frac{n_{ij}}{n_{.j}}\right)$$

Informational Gain - $G(Y/X) = E(Y) - E(Y/X)$

[2] Gorunescu F., *DATA MINING ,* Ed. Albastra, Cluj-Napoca, 2006, pg. 142

Normalized Informational Gain (Gain Ratio) - which considers the marginal distribution of X.

$$GR(Y/X) = \frac{G(Y/X)}{E(X)}$$

where

$$E(X) = -\sum_{j=1}^{q} \frac{n_{.j}}{n} \cdot \log_2\left(\frac{n_{.j}}{n}\right)$$

One can differentiate among the methods by considering the variable type: if the dependent variable is quantitative (range) a regression tree will be generated; if the depended variable is qualitative (categorical) a classification tree will result.

Classification and Regression Tree – a node that generates a decision tree with which one can predict or classify future values. The method implies using a recursive partitioning by splitting the recordings from an exercise cluster in segments, diminishing impurities at each step. The dependent and independent variables can be either quantitative or categorical, the separation is a binary one (one two subgroups are created).

C5.0 –the optimal splitting is made using the informational gain method. The dependent variable must be a categorical one (Petcu, 2010, 93).

For the first segmentation model we used the ***C5.0*** technique and the following question was considered: ”*Please rank the following recommendations regarding computer usage at the faculty, considering their importance to you (assign 1 for the first place, 2 for the second as importance etc)*

...Offer students more basic training for using computers
...To have more computers/projectors at the faculty
...Acquire licensed software
...Allow access to computers even when not at class
...Stimulate students' interest for developing practical abilities related computers and IT.

Variables included in the model:

☞ Target variable Y – a new variable was generated that divided the students in two groups, Economic Sciences students and students from other faculties.

☞ The entry variables were the response options to the question presented above.

The root node stands for the following distribution of students: 75 from Economic Sciences and 117 from “Other faculties”.

The first and most important segmentation variable, at this stage, was selected: “***Stimulate students’ interest for developing practical abilities related computers and IT "***. Two groups resulted: one that consisted of 107 students (55,73%), out of which 51 from Economic Sciences and 56 from “Other faculties”, that rated this recommendation as very important and important; a second group that consists of 85 students (44,27%), out of which 24 at Economic Sciences and 61 at other faculties for which other response variables were important.

The tree analysis has lead to the framing of the following rules, which characterize the groups of students that form the nodes:

- The first group also considered as important the following response options: ***Allow access to computers even when not at class, Acquire licensed software, offer students more basic training for using computers***
- The second group has considered as important the following response options: ***have more computers/projectors at the faculty, offer students more basic training for using computers***

The second segmentation model was build using the question from the survey related to the e-learning platform. The **C&RT** technique was used. Only the students that use the e-learning platform were selected.

	1Totaly unsatisfied	2	3	4	5 Very satisfied
Grade visualization	4	2	21	48	71
Courses visualization	26	23	35	41	21
Communication with professors	31	33	54	20	8
Payment visualization	9	16	27	45	49
ERASMUS	27	32	50	31	6

Table 1. *How satisfied are you with the information offered by the e-learning platform?*

The first and most important segmentation variable, at this stage, was: “***payment visualization"***. Two groups of students have resulted: one that consists of 94 students (64,38%), out of which 51 from Economic Sciences and 43 at Other faculties, that consider this service as important and very important; one group that consists of 52 students (35,62%), out of which 14 from Economic Sciences and 38 from other faculties for whom other response options are more important.

The tree analysis has lead to the framing of the following rules:

- From the first group 72 students (49,31%), out of which 37 from Economic Sciences and 35 from other faculties, have stated that they are not satisfied with ***communicating with professors*** through the e-learning platform but they are satisfied with ***grades' visualization***. Part of these (40) was unsatisfied with ***course visualization***;

- From the second group 52 (35,61%), out of which 14 from Economic Sciences and 38 from other faculties have stated that they are not satisfied with ***payment visualization***, ***ERASMUS***. Part of these (46) was also not satisfied about ***communicating with professors***.

4. Conclusions

This research has analyzed the attitudes and opinions of the students from Transilvania University regarding the computerization of the education system. After analyzing the data, one can draw the following conclusions: the students are aware of the fact that investments have been made at the university level in IT technologies; most of the students use computers for: solving exercises, writing documents (Word), PowerPoint presentations, data collection from the Internet; most of them have at least 2 subjects where they use computers at the university. Students use the e-learning platform especially for grade and payment visualization. Some of the respondents are discontented with matters such as communicating with professors and courses visualization thru the platform

After using data mining techniques to analyze the data, we draw the following conclusions: students have ranked on the first three places the recommendations- ***Acquire licensed software, Allow students to use computers at the university even when not at class, Stimulate students' interest for developing practical abilities related computers and IT.***

These results can inform decision making that will improve the efficiency of the informational learning process: increasing the efficiency of teaching time; storing information on an electronic device; working on projects at a higher level of specialization; the teaching system is easier; developing communication competencies and individual study.

Connecting this study to the computerization of teaching at Transilvania University, several support activities for the educational system can be coined for *"the Internal Evaluation- Annual Report of quality for the year 2013-2014"*: carrying on with the process of maintaining the material basis for students, that was developed in the period 2006-2008; consistent informational interconnectivity, at current standards in the fields, and ensuring a high performance of the Internet network, Point-To-Point Motorola Canopy 600 wireless equipments, fulfilling UTBv objectives regarding the informational infrastructure, which are: implementing and running the metropolitan educational network, with direct reference to having access to high speed Internet connection, long distance learning and e-learning, voice intranet services (Voice over IP – VoIP) and video (videoconferences) as well as administrative apps distributed in the network.

6. References

Cucoş, Constantin. 2006. *Informatizarea în educaţie. Aspecte ale virtualizării formării*. Iaşi: Polirom

Gorunescu, F. 2006. *DATA MINING concepte, modele şi tehnici.* Cluj-Napoca: Ed. Albastra.

Griffith, Arthur. 2010. *SPSS for Dummies* - 2nd Edition*,* Indianapolis: Editura Wiley Publishing Inc.

Noveanu Eugen., and Dan Potolea. 2008. *Informatizarea sistemului de învățământ Programul S.E.I.- raport de cercetare evaluative EVAL S.E.I 2008*, Bucuresti : Ed. Agata.

Petcu, Nicoleta. 2010. *Tehnici de data mining rezolvate în SPSS Clementin.* Cluj-Napoca. Editura Albastră.

Petcu, Nicoleta. 2003. *Statistica teorie si aplicatii in SPSS*. Brasov: Ed. Infomarket.

Radu, Ion T. 2000. *Evaluarea în procesul didactic.* Bucuresti: Editura Didactica si Pedagogica.

Stoetzel, Jean, and Girard Alain. 1975. *Sondajele de opinie publică.* Bucureşti: Ed. Ştiinţifică şi Enciclopedică.

Strajeri, Mihaela Luminiţa. 2009. Invățământul superior românesc şi necesitatea schimbării”. *Curentul National.*

http://ec.europa.eu/dgs/education_culture/index_en.htm, European Commission, Directorate General for Education and Culture. *Basic Indicators on the Incorporation of ICT into European Education Systems. Facts and Figures*. 2000/01 Annual Report

http://www.unitbv.ro/Portals/0/Hotarari/HS%20nr.%2033%20din%2013.02.2015.pdf Raportul anual de evaluare internă a calității anul universitar 2013-2014