Exploring the influence of ICT in online students through data mining tools

Javier Bravo
javier.bravo@udima.es

Sonia Janeth Romero
soniajaneth.romero@udima.es

María Luna
maria.luna@udima.es

Sonia Pamplona
sonia.pamplona@udima.es

Madrid Open University, UDIMA
Carretera de la Coruña 38.500. Vía de servicio 15.
28400, Collado Villalba, Madrid (Spain)

ABSTRACT

The aim of the present work is to evaluate differences according to age in digital competence, usages, and attitude towards ICT in a sample of 1231 online students of a distance university. To fulfill this goal, hypothesis testing, correlation analysis, and data mining techniques were performed on the basis of a 72-item survey. Results showed no strong differences between extreme groups of age. Besides, some interesting correlations between variables and additional information through association rules were found. This study led to better knowledge of online students in order to improve teaching and learning processes.

Keywords

Association rules, ICT attitude, ICT usages, distance education, online education, correlation, Mann-Whitney test, digital competence.

1. INTRODUCTION

During the last decades the proportion of higher education students taking at least one online course has outstandingly increased [1]. A research line developed in the field of e-learning in higher education focuses on the students' access, competences, actions and attitudes towards digital tools and devices and on how those variables are related to learning and well-being. As it is known, distance education, fostered by ICT, increases the variety of learners attending higher studies, creating new challenges for educators and institutions [2]. Specifically there are recent studies on whether the students' age is an important variable. In contrast to the concept of "digital natives" [10] , several studies find no evidences for strong discontinuity of young people on the use and attitudes about digital technology [6] [8]. Nevertheless differences related to age have been found, such us a deeper approach to studying of older students and less time spent using ICT (although this last difference is more noticeable at face-to-face universities than at the distance ones) [5,6].

In this paper we try to address this problem by developing the following objectives: (1) Compare digital competence, uses, and attitude towards ICT between young and students over 50; (2) Analyze relationships between variables in young and students over 50; (3) Obtain additional information about relationships between variables and group of ages by using data mining tools.

This paper is organized as follows. The next section presents a brief selection of related works. The method is described in section three. In section four the results are exhibited. Finally, the section five concludes the paper with discussion and plans for future work.

2. RELATED WORKS

Recently, some research studies were proposed to address the usage of data mining techniques in education especially in association rule mining.

Fattah et al. presented an association rule discovery model to investigate and analyze a university admission system database [3]. The model discovered the relation between students' data and their application status in the university system. The information discovered was very important to the admissions office in the analyzed university because it showed how to filter the applicants with respect to their record in high school.

García et al. described a data mining tool that uses association rule mining and collaborative filtering in order to make recommendations to instructors about how to improve e-learning courses [4]. This tool enables the sharing and scoring of rules discovered by other teachers in similar courses. The work showed and explained some examples of rules discovered in an adaptive web-based course.

Romero et al. explored the extraction of rare association rules when gathering student usage data from a Moodle system [11]. They showed how some specific algorithms, such as Apriori-Inverse and Apriori-Rare, are better at discovering rare-association rules than other non-specific algorithms, such as Apriori-Frequent and Apriori-Infrequent. Finally, they showed how the rules discovered by rare association rule mining algorithms can help the instructor to detect infrequent student behavior/activities in an e-learning environment such as Moodle.

Merceron and Yacef gave an interpretation of two measures of interest through association rules: cosine and added value [9]. In addition, they presented a case study that depicts a standard situation: a LMS that provides additional resources for students as a complement to the face-to-face teaching context. An important conclusion of this work is that common LMS are far from being data mining friendly. Thus, LMS should be enhanced with a special module with good facilities for exploring data.

Kumar and Chadha [7] used association rules mining in discovering the factors that affect assessment in Haryana University (India). They analyzed data for some courses taught in order to measure the students' performance based on factors such as instructor behavior, curriculum design, time schedule and students' interests.

3. METHOD

3.1 Participants, variables and instruments

A total of 1231 students participated voluntarily (with informed consent) in this study, 600 females and 631 males. They were all students recruited from Madrid Open University in Spain. 63.44% of the sample were studying a Bachelor's degree and 36.56% were Master's students . 40.76% of the students worked in ICT related areas and 57.66% had completed undergraduate studies previously. All the participants were between 18 and 69 years old (mean= 36.01, SD=9.59) and 110 are older than 50 (age 50+ group).

A survey of 72 items was designed to test students' self-reported ICT abilities, uses and attitudes. This survey is divided into four parts: demographical data and academic performance, actions with digital devices (computers, Smartphones, and other digital devices), frequencies of use of ICT tools (digital devices, communications, Moodle, file managing, and other tools) and attitude towards ICT in the process of learning.

(1) Demographical data and academic performance. Students were asked about: age (integer), gender (1=female, 2=male), grade (from 0 to 10), study area (58 values), first enrollment in UDIMA (from 2009-10 to 2014-15), previous degrees, if they work in fields related to ICT and average grade on academic record (retrieved using student identity).

(2) Actions with digital devices. It is composed of 20 items distributed on three scales: Actions with Other Digital Devices (AODD-4 items), Actions with Computers (AC-8 items) and Actions with Smartphones and Tablets (AST-8 items). The format used is 4-point Likert type, from 1 (I cannot do it) to 4 (I can do it and explain it to others). Descriptive results on this block of the instrument are: AODD (min=4; max=16; mean=15.02; SD=1.75); AC (min=10; max=32; mean=28.65; SD=3.96); AST (min=11; max=32; mean=29.04; SD=3.76).

(3) Frequencies of use of ICT tools. It is composed of 25 items distributed on five parts: a) Other Digital Devices (FODD-5 items), b) Communications (FC-5 items), c) Moodle (FM-7 items) d) File Management (FFM-3 items), and e) Other Tools (FOT-4 items). The format used is 4-point Likert type, measuring frequency of use, from 1 (I do not use/do not know) to 4 (Very often). Descriptive results on this block of the instrument are: FDD (min=5; max=20; mean=14.97; SD=3.01); FC (min=7; max=24; mean=17; SD=3.69); FM (min=7; max=28; mean=19.48; SD=4.23), FFM (min=3; max=12; mean=6.32; SD=2.33); FOT (min=4; max=16; mean=6.69; SD=2.46).

(4) Attitude towards ICT in the learning process. It is composed of 24 items distributed on three scales: affective, cognitive and behavioral. The format used is 5-point Likert type from 1 (totally disagree) to 5 (totally agree). Two items on each dimension are inversely rated. The higher test score indicates greater favorable attitude towards the incorporation of ICT in the learning process. Descriptive results shows: min=29; max=120; mean=97.79; SD=13.49. Cronbach Alpha (considering all 24 items of attitude) was 0.89 indicating the high reliability of the test.

3.2 Data analysis

Data analysis included hypothesis testing, correlation and data mining analysis. These are detailed below.

3.2.1 Hypothesis testing and correlation

a) Hypothesis testing. Wilcoxon and Mann-Whitney tests were made to test the hypothesis about differences between extreme ages (24- and 50+).

b) Correlation. Pearson correlation matrix between continuous variables was made in order to evaluate possible associations.

3.2.2 Data Mining Techniques

To complement and provide additional information we used four data mining techniques: OneR, Decision trees (J48), Naïve Bayes and association rules. In this stage it is important to perform the preprocessing phase [12].

a) Preprocessing phase. It is important to note that in this analysis we utilized the whole data. In addition, the items of "Frequencies of use ICT tools" were grouped in four nominal variables: FODD (it contains the sum of five items of "Other Digital Devices"), FC (sum of five items of "Communications"), FM (sum of seven items of "Moodle"), FFM (sum of three items of "File Management"), FOT (sum of 4 items of Other tools). In the same way, the items of "Attitude towards ICT in the learning process" were grouped in the ictAttitude variable. Classification techniques works better with nominal variables. Therefore, age and ictAttitude were discretized to ictAttitude3groups and age4groups respectively. The ictAttitude variable is a continuous variable ranged from 29 to 120. We discretized this variable in three nominal values according to its 33 percentile, 66 percentile, and 99 percentile. Regarding to age variable, it ranged from 18 to 69. This variable was discretized in four values according to its four quartiles. This new variable is called ictAttitude3groups.

b) Selection of variables. In this phase we only use 14 variables: codDegree, gender, firstEnrollment, gradeRound, AODD, AC, AST, FODD, FC, FM, FFM, FOT, age4groups, and ictAttitude3groups. In addition, we utilized the WraperSubsetEval method provided by Weka [13]. This metaselection method selects the most appropriate variables for a data mining technique. This method receives two variables: the selection method and the search method. Since OneR, J48, and Naïve Bayes are classification techniques we indicated to this method to use J48 for selection method (selection mode: 10-fold cross-validation). Also, BestFirst forward method was used for searching method. As a result 8 variables were selected: gender, AODD, AC, FM, FFM, FOT, age4groups, and ictAttitude3groups.

c) Application of techniques. In this phase we utilized three classification techniques and association rules. The classification techniques utilized were: OneR, J48, and Naïve Bayes. We utilized for these techniques the 8 variables listed above (class variable: ictAttitude3groups). In order to select the most appropriate technique we calculated the accuracy (number of correctly classify instances) of each technique. As the size of data was enough to apply the split method, which divides the sample in two parts: training and testing data, we utilized it instead of the cross-validation method. Moreover, we utilized 80% of data for training and 20% for testing. It is well known that if the data size is large enough both methods of dividing the data should give similar accuracies. Concerning association rules we utilized the Apriori algorithm with confidence=0.9 and minimum support=0.1.

4. RESULTS

4.1 Hypothesis testing and correlations

a) Hypothesis testing. First, assumption evaluation was made in order to decide statistical techniques to compare students of extreme ages: above 50 years (50+), and below 24 years (24-). Levene's test shows a lack of homoscedasticity between groups in the variables related with actions: L=20.068, (1, 125), p<.001 for

AODD; L=19.177, (1, 125), p<.001 for AC and L=58.028, (1, 125), p<.001for FDD. The Shapiro-Wilk test shows a lack of normality in all the variables except FC. Due to failure to meet the assumptions we decided to use nonparametric statistic (Mann-Whitney test) to compare both groups. As it can be seen in Table 1, significant differences between both groups (50+ and 24-) were found on actions with other digital devices, actions with Smartphones and tablets, and frequency of use of communication tools. The sum of ranks (see Table 1) indicated high scores on AODD, AST and FC in the group of age 24-.

b) Correlation. The Pearson correlation matrix between continuous variables was calculated for both groups. In the group of age 50+ all the action variables correlated significantly and positively with frequency variables and attitude towards ICT: the lowest positive and significantly correlation was r=.111(AC-FC) and the highest r=.724 (AC-AODD), this result was expected because the more digital competence self-perceived, the more frequent use of ICT tools and the more positive attitude towards ICT use in the learning process. On the other hand, AODD, AST and FC correlated negatively with age (r=-.148, p<.01; r=-.208, p<.01 and r=-.069, p<.05 respectively) indicating as the age increased, the self-perceived competence in actions with digital devices and Smartphones decreased and the use of communication tools was less frequent. The average grade correlated positively with self-perceived competence in actions with computers (r=.136, p<.01). And finally, the number of years studying through distance learning correlated with AC (r=.087, p<.01), FDD (r=.068, p<.05), FM (r=.083, p<.01), attitude (r=.128, p<.01) and age (r=.278, p<.01) indicating that the more years of experience studying online, the higher self-perceived competence of actions with computers and higher frequency of use of digital devices, Moodle, and better attitude toward ICT.

In the group of young students we found less significant correlations, as the number of years studying online did not correlate with any variable, the average grade only correlate positively with AC (r=.197, p<.01), the age correlated inversely with AST (r=-.226, p<.05) and FDD (r=-.307, p<.01). The attitude toward ICT correlated positively with all the actions (r=.205, p<.01 for AODD; r=.221, p<.05 for AC and r=.305, p<.01 for AST) and also with FM (r=.300, p<.01). Finally, the scales of frequencies correlated positively with each other but not with the actions scales.

Table 1. Mann-Whitney U, Wilcoxon W, Z and significance

	ADD	AST	FC
U	5098.500	3429.500	5528.500
W	10454.500	8785.500	10781.500
Z	-4.082	-6.895	-2.438
Sig.	.000	.000	.015
Age	Rank sum	Rank sum	Rank sum
24-	18225.5	17170.5	19894.5
50+	10454.5	11509.5	8785.5

4.2 Data mining techniques

4.2.1 Classification techniques

The accuracies of applying the OneR, J48 and Naïve Bayes technique are as follows: 41.9%, 43.8%, and 42.9%, respectively. For example, J48 classifies correctly in 43.8% of the instances. It is clear that neither of the three techniques has an accuracy greater than 44%. As none of these techniques obtained reliable results, we applied the association rules technique.

4.2.2 Association rules

This trial consisted of using 14 variables with the Apriori algorithm. It is important to highlight that the Apriori algorithm only works with nominal variables. Therefore, the grade variable was removed from the data. The parameters for this algorithm were: minimum support=0.1; confidence=0.9; number of rules=20; instances=1231; attributes=(codDegree, gender, firstEnrollment, AODD, AC, AST, FODD, FC, FM, FFM, FOT, age4groups, ictAttitude3groups).

The result of applying this algorithm is presented in Table 2. It is shown that the Apriori algorithm selected 16 rules. Thus, it is shown that AR6 indicates that students with the best score in actions with computers and best ICT attitude will have the best score in actions with other digital devices. The AR7 and AR14 rules indicate that both genders will have the best score in "actions with other digital devices" if they have the best score in "actions with computers". Both association rules are redundant, since this fact is indicated in AR11. The AR9 contains the variable age. It indicates that the students between second and third quartile of age with the best score in "actions with computers" are experts managing other digital devices.

Regarding the other association rules interesting relations between several variables were found. The first rule relates actions with computers, actions with Smartphones, and actions with other digital devices. Thus, it means that students with the best ICT attitude who demonstrate a high level of actions with computers and Smartphones, will have a high level of actions with other digital devices. The AR2 rule shows a similar relation, but only for male students. The AR3 is informed about a general relation between actions with computers, actions with Smartphones, and actions with other digital devices. A value of 32 indicates a high level of actions with computers and Smartphones, and a value of 16 is also a high level of actions with other digital devices. Thus, a student with a high level of actions with computers and Smartphones, he/she will have a high level of actions with other digital devices. Interesting information is revealed in the AR4 rule, since it relates the first enrollment, action with computers, Smartphones, and with other digital devices. Thus, the students of the 2014-15 year show a high level of action with computers and Smartphones, and also with other digital devices. This association rule is similar to the AR10 and AR15, but with less information. The rules AR5 and AR6 show that students with a high ICT attitude, and a high value in actions with computers or Smartphones, will have a high level in actions with other digital devices. Finally, the AR8 and AR16 rules relate the gender, action with computers and with other digital devices. Consequently, female and male students report the same abilities in actions with Smartphones and other digital devices.

5. CONCLUSIONS

The present study gathered a large sample composed of 1231 online students in a distance university with a range of age from 18 to 69 years. Our results agree to a great extent with other related studies [5][6]. In fact, we did not find enough evidence of strong differences among extreme groups of age, although results showed slight differences in variables related with the frequency of use and perceived competence with Smartphones and communication tools.

Another interesting conclusion is that attitude towards ICT did not correlate inversely with age, on the contrary, students aged 50+ exhibited positive attitudes towards the implementation of ICT for the learning process. These conclusions lead to better knowledge about students attending online higher education. Therefore, these results should provide improvements in the methodology of the e-Learning courses and foster the utilization of communication tools (less utilized by 50+ students).

This work also showed that data mining techniques can provide complementary information to traditional analysis methods. Although classification techniques did not provide reliable results, since its accuracy was less than 44%, the association rules technique provided deeper information. In fact, the Apriori algorithm obtained 16 association rules. These association rules showed relationships between the following variables: actions with computers, Smartphones and other digital devices, gender, ITC attitude, and first enrollment in UDIMA. This information was not provided by the hypothesis testing, therefore, we have demonstrated that association rules are appropriate to analyze these data.

For future work it will be appropriate to analyze other parameters of the Apriori algorithm that could provide rules with more information. For instance, to test and evaluate other selection methods based on Lift or Leverage is an interesting future line of research [9].

Table 2. Best rules of the Apriori algorithm

Rule		Cov.	Conf.
AR1	AC=32 AST=32 ictAttitude3groups=3 ==> AODD=16	137	1
AR2	gender=2 AC=32 AST=32 ==> AODD=16	221	0.99
AR3	AC=32 AST=32 ==> AODD=16	307	0.99
AR4	firstEnrollment=2014-15 AC=32 AST=32 ==> AODD=16	134	0.99
AR5	AST=32 ictAttitude3groups=3 ==> AODD=16	167	0.98
AR6	AC=32 ictAttitude3groups=3 ==> AODD=16	187	0.97
AR7	gender=1 AC=32 ==> AODD=16	143	0.97
AR8	gender=2 AST=32 ==> AODD=16	264	0.97
AR9	AC=32 age4groups=3 ==> AODD=16	123	0.96
AR10	firstEnrollment=2014-15 AC=32 ==> AODD=16	191	0.96
AR11	AC=32 ==> AODD=16	426	0.96
AR12	AST=32 ==> AODD=16	390	0.95
AR13	gender=2 firstEnrollment=2014-15 AC=32 ==> AODD=16	126	0.95
AR14	gender=2 AC=32 ==> AODD=16	283	0.95
AR15	firstEnrollment=2014-15 AST=32 ==> AODD=16	182	0.93
AR16	gender=1 AST=32 ==> AODD=16	126	0.91

6. ACKNOWLEDGMENTS

This work has been funded by European Commission Lifelong Learning Programme through the AGE50+ project (hash code: 1B8E6E156B13780A).

7. REFERENCES

[1] Allen, I. E., & Seaman, J. 2013. *Changing Course: Ten Years of Tracking Online Education in the United States*. Sloan Consortium. PO Box 1238, Newburyport, MA 01950, USA.

[2] Altbach, P. G., Reisberg, L., & Rumbley, L. E. 2009. Trends in global higher education: Tracking an academic revolution. UNESCO.

[3] Fattah Mashat, A., M.Fouad, M., S. Yu, P. and F. Gharib, T. 2013. Discovery of Association Rules from University Admission System Data. *International Journal of Modern Education and Computer Science* 5,4 (May 2013), 1–7.

[4] García, E., Romero, C., Ventura, S. and de Castro, C. 2011. A collaborative educational association rule mining tool. The Internet and Higher Education 14, 2 (March 2011), 77–88.

[5] Hosein, A., Ramanau, R., & Jones, C. 2010. Learning and living technologies: a longitudinal study of first year students' frequency and competence in the use of ICT. *Learning, media and technology* 35, 4 (2010), 403-418.

[6] Jelfs, A., & Richardson, J. T. 2013. The use of digital technologies across the adult life span in distance education. *British Journal of Educational Technology* 44, 2 (2013), 338-351.

[7] Kumar, V. and Chadha, A. 2012. Mining association rules in student's assessment data. *International Journal of Computer Science Issues* 9, 5 (September 2012), 211–216.

[8] Lai, K. W., & Hong, K. S. 2014. Technology use and learning characteristics of students in higher education: Do generational differences exist?. *British Journal of Educational Technology.* doi:10.1111/bjet.12161

[9] Merceron, A. and Yacef, K. 2008. Interestingness measures for association rules in educational data. *Proceedings of the 1st International Conference on Educational Data Mining* (Montreal, Quebec, Canada, June 20-21, 2008). 57–66.

[10] Prensky, M. 2001. Digital natives, digital immigrants part 1. *On the horizon*, *9*, 5 (2001), 1-6.

[11] Romero, C. and Romero, J. 2010. Mining rare association rules from e-learning data. *Proceedings of the 3rd International Conference on Educational Data Mining* (Pittsburgh, PA, USA, June 11-13, 2010). 171–180.

[12] Vialardi, C., Bravo, J., & Ortigosa, A. 2008. Improving AEH Courses through Log Analysis. *Journal of Universal Computer Science* 14,17 (Sept. 2008), 2777-2798.

[13] Witten, I.H., Frank, E., & Hall, M.A. 2011. Data Mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers. Burlington, MA 01803, USA.