Computers in Human Behavior 115 (2021) 106595

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Full length article

Early prediction of undergraduate Student's academic performance in completely online learning: A five-year study

Javier Bravo-Agapito [a,*], Sonia J. Romero [b], Sonia Pamplona [a]

[a] Madrid Open University – UDIMA, Spain
[b] National Distance University of Spain – UNED, Spain

ARTICLE INFO

Keywords:
Analytics
Learning management systems
Online learning
Modeling
Prediction

ABSTRACT

This decade, e-learning systems provide more interactivity to instructors and students than traditional systems and make possible a completely online (CO) education. However, instructors could not warn if a CO student is engaged or not in the course, and they could not predict his or her academic performance in courses. This work provides a collection of models (exploratory factor analysis, multiple linear regressions, cluster analysis, and correlation) to early predict the academic performance of students. These models are constructed using Moodle interaction data, characteristics, and grades of 802 undergraduate students from a CO university. The models result indicated that the major contribution to the prediction of the academic student performance is made by four factors: Access, Questionnaire, Task, and Age. Access factor is composed by variables related to accesses of students in Moodle, including visits to forums and glossaries. Questionnaire factor summarizes variables related to visits and attempts in questionnaires. Task factor is composed of variables related to consulted and submitted tasks. The Age factor contains the student age. Also, it is remarkable that Age was identified as a negative predictor of the performance of students, indicating that the student performance is inversely proportional to age. In addition, cluster analysis found five groups and sustained that number of interactions with Moodle are closely related to performance of students.

Author contribution

Javier Bravo-Agapito, Conceptualization, Software, Investigation, Writing - original draft. Sonia Janeth. Romero, Methodology, Formal analysis, Data curation, Visualization, Writing - original draft. Sonia Pamplona, Writing - original draft, Validation, Resources.

1. Introduction

Nowadays, it is widely accepted that e-learning systems provide more interactivity and flexibility since it is possible to use them anytime in different devices and locations (Chen, Myers, & Yaron, 2000). Especially, these systems are utilized for online and blended universities by instructors to transmit, organize and provide educational contents to students. However, online instructors do not have complete information about students' behavior and academic performance as face-to-face (F2F) instructors do. For example, F2F instructors can adapt their explanations when they feel that students could have failures in their learning process, because these instructors have constant visual contact to students. Then, they can interact continuously with students. However, an online instructor can virtually interact with students since an online course typically contains weekly tasks, which are also continuously read and evaluated by online instructors. The problem is that this online supervision should have a high amount of communication and feedback, but this practice is carried out very rarely. Moreover, descriptions of experience show that a supervision key of 1:10 is sufficient to ensure that the quality of online supervision is indistinguishable from the F2F supervision for the same work intensity of the teacher, but this supervision key is usually larger than 1:10 in online teaching.

Currently, e-learning is implemented in everyday university life, as the advantages are well known: temporal independence, organization of the learning process, illustration, and redundancy. In fact, universities use Learning Management Systems (LMS) to provide blended or completely online (CO) learning. LMS store a vast quantity of data about interactions of students in log files. These files usually contain variables in the data such as the number of logins, number of accesses to elements of an online course, number of assignments completed, number of days in the online course, activities grades, term grade, course grade and so

* Corresponding author.
E-mail addresses: javier.bravo@udima.es (J. Bravo-Agapito), sjromero@psi.uned.es (S.J. Romero), sonia.pamplona@udima.es (S. Pamplona).

https://doi.org/10.1016/j.chb.2020.106595
Received 15 January 2020; Received in revised form 30 June 2020; Accepted 2 October 2020
Available online 12 October 2020

on. These data could be interesting for online instructors since they could contain information about behavior of the students that could influence in their academic performance. In other words, it is useful to know which variables are related to student performance, and then instructors could trigger actions to improve the learning process of students. However, it is important to highlight that CO students are particularly different from F2F students. For example, CO students could be aged from 18 to 70 years old while F2F students usually are aged from 18 to 23 years old. This fact creates a special diversity in CO instruction. This paper provides insights whether this diversity of students' profile could influence in the academic performance.

Early Warning Systems (EWS) utilize data mining methods to detect students at risk of failure of courses in different education levels and contexts (Howard, Meehan, & Parnell, 2018). According to Knowles (2015), Early Warning Indicators (EWI) provide instructors with an advanced warning that students need help in their learning process. These systems contain predictive models with a collection of variables that are related to EWI. These variables usually contain information about demographic and institutional data, student characteristics, term or mid-term grades, and LMS interaction data. The Wisconsin Dropout Early Warning System (DEWS) is a successful development that provides instructors a forward-looking view of student performance (Knowles, 2015). However, a limitation in DEWS is that it is able to utilize demographic and institutional data, student characteristics, term grades, but no able to utilize LMS interaction data and it is only utilized in F2F education. Other successful development of EWS is in Purdue University (Sclater, Peasgood, Mullan, 2016). This system is based on prediction models using the student performance, LMS student interaction, and prior academic history and student characteristics to predict student performance. This system represents each predicted student performance as a traffic light, using red signal (unsuccessful performance), yellow signal (potential problems in the performance), and green (successful performance).

The present research is carried out at the Madrid Open University (UDIMA). This university provides a CO teaching to students. It offers CO undergraduate and graduate courses in different topics such as: Law, Criminology, Computer Science, Civil Engineering, Business Administration, Economy, Journalism, History, Psychology, and Education. It is important to note that every UDIMA course has the same methodology and similar structure. Specifically, each course is composed by 10 learning units and a set of activities. The UDIMA utilizes the LMS Moodle to transmit and organize educational contents of all courses. It means that every instructor can provide students with educational content, tasks, tests, videos, lessons, and so on. Also, the system stores log data about student interaction, student grades and activities submitted, test attempts, etc. It is important to note that the UDIMA students are enrolled in three courses per semester on average.

The present research has three main research goals: to identify the variables that influence students' academic performance in the target course (G1), to identify the variables that influence students' academic performance in subsequent four years to achieve early prediction (G2), and to describe typology of students based on their interactions with the LMS (G3).

2. Theory

Several works have been done recently to predict academic performance based on LMS data. One challenge that it is noted is the difficulty of finding a set of variables that can consistently predict student performance across multiple courses (Conijn, Snijders, Kleingeld, & Matzat, 2017). One of the reasons for this difficulty is that instructional conditions could influence the predictions of academic success based on log files of LMS (Gašević, Dawson, Rogers, & Gasevic, 2016). The students may use different LMS features and there may be differences in how and the extent to which the LMS tools are utilized.

Another relevant research result is that more elaborate theoretical reasoning is needed in learning analytics to achieve generalizable results (Conijn et al., 2017). That is to say, researchers need to use meaningful measures from log files that are congruent with learning and instructional theories. The theory of self-regulation learning has been used as theoretical grounding (Gašević et al., 2016; You, 2016). The use of this theory implies that indicators that reflect regular study and time-management-related behaviors should be given more attention and be further investigated (You, 2016).

According to the theory of self-regulation learning, it could be useful to measure the time that students devote to the accomplishment of the learning activities. In fact, the indicator elapsed time has been considered for modeling the student performance in several studies (Peña-Ayala, 2014). However, time has not proved to be a significant variable to predict student performance in all cases. In a study about the use of discussion forums as an indicator of student performance (Romero, López, Luna, & Ventura, 2013), the number of messages is one of the most important attributes and the total time in the forum is one of the least relevant attributes. The results of a study on the relationship between log data and cognitive activities are even more revealing (Lerche & Kiel, 2018). The time spent editing a wiki did not contribute to explaining the performance because the texts of the wiki are created within a local word processor, so that students do not need to be connected to the LMS to create wiki content, they only connect to it to copy and paste the text created locally. Therefore, these results suggest that modeling the time spent on a task might be a decision that depends on the instructional design and the characteristics of the used LMS.

Another issue to consider when predicting academic performance is the cost of measuring each variable. Variables expensive to measure limit the generalizability of the results (Sandoval, Gonzalez, Alarcon, Pichara, & Montenegro, 2018). An example of a variable expensive to measure is the elaborated time-based measure regular study used in You (2016). In a course where the main materials were instructional videos, this variable was calculated tracking the time point when the student first encountered the video and the length of visualization time. Each student received a score per week. No point was given if the student did not access the instructional video within the scheduled time period, and a half point was given when the students accessed the content but did not finish watching the video. One point was given a student watched the weekly assigned videos from beginning to end within the scheduled week.

Besides, most of the studies have not been conducted in CO degrees. Studies are taken place in campus-based universities with some online courses (You, 2016; Lerche & Kiel, 2018), blended courses (Conijn et al., 2017; Gašević et al., 2016; Sandoval et al., 2018; Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García, 2014), and even within online assignments in a F2F course (Cerezo, Sánchez-Santillán, Paule-Ruiz, & Núñez, 2016). Therefore, the students' interactions with the LMS features in these contexts will be different than in a CO course.

Finally, a key requirement in an online setting is to be able to do an early prediction of the academic performance to appropriately address students' weaknesses (Lu et al., 2018). Two approaches can be founded in the literature. The first one is to do an early prediction of the academic performance at the end of a semester (You, 2016; Lu et al., 2018). In this case, data from LMS are typically gathered from week 1 to week 6. The second one is to do an early prediction at the end of a degree, when a student finishes it (Xu, Moon, & Schaar, 2017). In this case, data from the LMS are gathered during the first semester of the degree.

The research herein was performed in a CO university, which shares a common instructional design between all the offered degrees, which distinguishes this study from previous ones, carried out in blended scenarios with different instructional designs.

The independent variables have been selected for the study in accordance with our instructional design, which is primarily composed of questionnaires, forums, and tasks. Therefore, according to (Lerche & Kiel, 2018) we have not considered measuring the time spent on tasks or Moodle questionnaires since both activities are usually done offline,

outside the LMS. In other words, this variable could not be measured through the logs of an LMS. On the other hand, consistent with the results of Romero et al. (2013), we have not considered the time spent in forums, but we have considered the number of accesses to the forum and the number of messages added to it.

In addition, we use low-cost variables that do not require active effort for data collection or elaborated measures. The common instructional conditions and the use of low-cost variables improve the generalizability of our study. Besides, we use an early prediction approach that allows us to predict academic performance at the end of a complete degree.

3. Material and methods

3.1. Sample

The sample was composed by 802 students: 377 females and 425 males. They were all students of UDIMA in Spain. Data of students' interaction with the LMS were collected from four courses in the academic year 2012–2013. In addition, to perform the early prediction, longitudinal data of academic achievement was gathered during the years 2013–2014, 2014–2015, 2015–2016, and 2016–2017. The courses selected were: Knowledge Management (N = 151, which is 18.8% of the sample), General Sociology (N = 135, 16.8%), Information Technology and Communication (N = 157, 19.6%) and Learning and Information Technologies (N = 359, 44.8%). Criteria of course selection were: first, that were of the first semester to make the early prediction, second that they were transversal and obligatory so that we had a representation of all the degrees. The age of participants ranges from 21 to 70 years old (m = 36.54; s.d. = 10.14). Distribution according the university degree was: 2% Economy, 26% Criminology, 4.5% Civil Engineering, 2% Humanities, 7% Tourism, 10.6% Psychology, 6% Computer Science, 11% Business Administration, 10% Journalism, 3.5% History, 13.7% Law, and 11.5% Labor Science.

3.2. Data collection and variables

Courses on UDIMA has a common structure: all courses consist of 10 didactic units and several activities including 3 to 6 forums or glossaries, 2 to 6 questionnaires (with 2 attempts each one) and 2 to 6 assignments.

Table 1 list the variables included in the present research, all the variables were measured at the end of the academic semester, X_1 to X_{16} are independent variables obtained from Moodle log files, and X_{17} to X_{19} are independent variables obtained from the student profile in the academic system of UDIMA. It is worth pointing out that the academic system of UDIMA did not provide us prior knowledge data and personal data of students such as if a student has a job, family, health issue, etc., since it is not mandatory to collect these data from students in UDIMA. Y_1 to Y_6 are grades of each academic year and are dependent variables and are obtained by the academic system of UDIMA. The extraction of the Moodle log files was carried out using an ad hoc computer application developed by the authors of the present work.

It is important to point out that grade point average (GPA) is calculated as it is showed in (1). In this formula, CG_i represents the grade of course i; W_i is the number of credits of course i; W_T is the number of passed credits; i=1 indicates the first academic year; L indicates the highest academic year. For example, let say that the student S_1 passed 30 credits with the following courses grades and credits: 7.5 (6 credits), 8 (6 credits), 5 (3 credits), 6 (3 credits), 9.5 (6 credits), 6.7 (6 credits). Then, GPA for S_1 is 7.44. Therefore, GPA in our study is ranged from 5 to 10.

$$\sum_{i=1}^{L} \frac{CG_i * W_i}{W_T} \quad (1)$$

Table 1
Description of the variables extracted from the log files (X1-X16), demographics (X17-X19), and academic achievement (Y1–Y6).

Short name	Description	Type
Total_logins	Total number of logins to the Moodle platform	X_1
N_access_forum	Frequency of the student access to all the forums	X_2
N_added_messages_forum	Total number of messages added by the student in the forums	X_3
N_access_didactic_units	Frequency of the student access to teaching materials	X_4
N_access_glossaries	Frequency of the student access to all the glossaries	X_5
Total_assignments	Total number of assignments of the course	X_6
N_assignments_consulted	Frequency of the student consults the assignments	X_7
N_assignments_submitted	Total number of assignments submitted	X_8
N_access_questionnaires	Frequency of the student access to all the questionnaires	X_9
N_attempts_questionnaires	Frequency of the student tries to solve the questionnaires	X_{10}
N_answered_questions	Total of questions answered in all the questionnaires	X_{11}
N_questionnaire_views	Frequency of the student observes the questionnaires	X_{12}
N_questionnaires_submitted	Frequency of the student submit a questionnaire	X_{13}
N_reviews_questionnaire	Frequency of the student revises attempts to questionnaires	X_{14}
Days_first_access	Number of days until first access to the virtual classroom	X_{15}
N_entries course	Total number of entries to the course	X_{16}
Age	Age of the student	X_{17}
Sex	Gender of the student	X_{18}
Degree	Degree in which student is enrolled	X_{19}
Grade_course	Final grade obtained in the course	Y_1
GPA_12_13	GPA in academic year 2012–2013	Y_2
GPA_13_14	GPA in academic year 2013–2014	Y_3
GPA_14_15	GPA in academic year 2014–2015	Y_4
GPA_15_16	GPA in academic year 2015–2016	Y_5
GPA_16_17	GPA in academic year 2016–2017	Y_6

3.3. Design and data analysis

Five-year data were included in order to develop an early prediction of the academic performance. We set significance level of 0.05 for all analyses. Data analysis included several procedures, detailed below:

1) Description of variables: a complete description of the 16 variables extracted from the log files was made in order to characterize the students' behavior in the LMS and also to test the assumptions of the regression models. This analysis includes Pearson correlation matrix to test collinearity.
2) Factor analysis: as most of the variables extracted from the log files are highly correlated an Exploratory Factor Analysis (EFA) was conducted using Principal Component Analysis with varimax rotation to reduce the number of independent variables to factors more interpretable and not correlated. The input for EFA was the Pearson correlations matrix between the 16 variables extracted from the log files. This matrix was examined using the Bartlett test of Sphericity KMO index and Measures of Sample Adequacy. To decide the number of factors to retain a mixed approach was employed (Ruscio & Roche, 2012), combining scree test with optimal coordinates and parallel analysis with 500 simulated resamples (Horn, 1965).
3) Multiple linear regressions: a multiple regression model was estimated by the method of maximum likelihood to meet two of the research goals: (G1) to identify the variables that influence students' academic performance in the target course (Y_1) and (G2) to test the influence of those independent variables on the average final grade of the subsequent academic years from academic course 2012–2013

to 2016–2017 (Y_2 to Y_6). As predictor variables were used the factors identified in the EFA and the only continuous sociodemographic variable (age).

4) Cluster analysis: in order to achieve the third goal (G3) a cluster analysis was performed. This analysis was made following a technique that combines factorial methods and cluster analysis in four steps (Lebart, Morineau, & Piron, 2000). First, a Principal Component Analysis (PCA) was made with the continuous variables (Husson, Lê, & Pagès, 2010). Second, an agglomerative hierarchical classification was made using Ward's method (Pardo & Del Campo, 2007). Third, a classification through mobile centers was made using the K-Means method; in addition, the validation of the identified groups was made using hypothesis testing. Finally, the description of each group was made using the continuous and categorical variables. If the variables are continuous, the average of each group is compared with the general average and, if the variable is categorical, the percentages are compared using *v.test* statistic (Husson, Lê, & Pagès, 2010).

Data analysis was carried out with R v.3.2.4 (The R Development Core Team, 2018). EFA and cluster analysis were made using the FactorMineR v.1.32 (Husson, Josse, & Le, 2016). The remaining analysis was made in Jamovi (The jamovi project, 2019).

3.4. Procedure

The present research was developed in six phases. In the first phase the distribution and correlation of the 16 variables extracted from the log files were analyzed, in the second phase the EFA was made, finding three factors: Access Factor (AF), Questionnaire Factor (QF) and Task Factor (TF) that will be explained in the results section. The third phase consisted in an exploratory analysis of the three factors conformed by the EFA in order to detect outliers and to analyze their distribution. The process of elimination of outliers was as follow: a) the descriptive statistics of the distributions including quartiles, boxplot and skewness/kurtosis were calculated, b) the outliers were eliminated using the boxplot and the quartiles (outliers were mainly values that were above the 75th percentile, that is, extremely high values of the variables), c) specifically, we eliminated cases with values greater than 3000 in AF, greater than 250 in QF and greater than 130 in TF. The boxplots and distribution histograms of the factors after elimination of outliers will be presented in the result section.

The fourth phase was the prediction of the performance in the specific course (Y_1). In the data pre-processing 20 cases were detected as outliers and removed from the sample, so, results for this initial analysis were based on 782 students. In the fifth phase we made predictions of the academic performance on each subsequent academic year (Y_2 to Y_6) with a reduced sample of 525 valid cases: 593 students finished their studies (183 students abandoned) and 68 cases were detected as outliers. The sixth phase was to perform the cluster analysis.

4. Results

4.1. Description of data

4.1.1. Distributions

As can be seen in Table 2, the independent variables exhibit great dispersion, positive skewness and they are leptokurtic (except total of assignments).

4.1.2. Correlations

Table 3 shows almost consistently significant correlations. For that reason, and also due to the skewed and leptokurtic form of the distributions presented in Table 2 we decide to perform an EFA. This EFA checks in advance whether some of the variables extracted from the log files could be better represented in a series of combined factors, more stable against individual outliers and thus against incorrect forecast.

4.2. Exploratory factor analysis

4.2.1. Adequacy of the analysis

Barlett's test ($\chi^2 = 14868$; df = 120; $p < .001$) indicates the acceptance of null hypothesis, then, it is adequate to perform the factor analysis. Table 4 presents the KMO index of sampling adequacy. The measures KMO indicate a high correlation between the variables (greater than 0.7) confirming the adequacy of performing the factor analysis.

4.2.2. Selection of the number of factors to retain

Following the scree plot of the parallel analysis (see Fig. 1), it can be seen that the variables can be summarized in three factors that explain 70.9% of the variance (see Table 5).

4.2.3. Factor loadings and composition of factors

Once the three factors have been selected, we proceed to analyze the factorial loads that are presented in Table 6. It can be observed that the independent variables could be reduced on three factors. The first factor can be called Questionnaire Factor (QF) because it summarizes 6 independent variables: the frequency of accesses and attempts of questionnaires, the number of questions answered, the frequency of questionnaires visualization, the number of questionnaires submitted, and the number of questionnaires reviewed. The second factor is composed by 6 variables that are all related to accesses: total number of logins, the number of course accesses, the frequency of access to forums

Table 2
Description of the variables extracted from the log files.

Variable	Mean	*sd*	Min	Max	Skewness	Kurtosis
Total_logins	766.33	629.33	2	4121	1.47	3.82
N_access_forum	82.62	106.53	0	661	2.40	6.72
N_added_messages_forum	2.04	3.14	0	25	2.95	12.10
N_access_didactic_units	72.22	82.07	0	497	2.03	4.64
N_access_glossaires	3.58	5.12	0	47	2.61	10.78
Total_assignments	4.81	1.42	2	6	−0.95	−0.34
N_assignments_consulted	43.58	31.19	0	216	1.23	2.91
N_assignments_submitted	4.48	2.78	0	20	0.32	0.70
N_access_questionnaires	30.12	28.29	0	236	2.18	8.01
N_ attempts_questionnaires	8.53	7.72	0	50	1.62	3.80
N_answered_questions	17.84	26.38	0	233	3.63	17.80
N_questionnaire_views	9.27	8.29	0	51	1.54	3.35
N_questionnaires_submitted	8.45	7.67	0	50	1.61	3.70
N_reviews_questionnaires	15.91	18.27	0	145	2.55	9.57
N_entries_course	133.38	124.17	0	795	2.13	5.88
Days_first_access	6.06	11.54	1	137	5.38	44.96

Table 3
Pearson correlation matrix.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
X_1	1															
X_2	.376**	1														
X_3	.269**	.464**	1													
X_4	.397**	.879**	.538**	1												
X_5	.133**	.378**	.227**	.391**	1											
X_6	-.124**	.076*	-.143**	.120**	.130**	1										
X_7	.383**	.395**	.138**	.466**	.377**	.443**	1									
X_8	.278**	.297**	.071*	.359**	.239**	.474**	.709**	1								
X_9	.332**	.281**	.229**	.273**	.213**	.151**	.304**	.234**	1							
X_{10}	.303**	.248**	.194**	.242**	.158**	-.15**	.239**	.267**	.801**	1						
X_{11}	.138**	.235**	.091*	.245**	.187**	.225**	.370**	.338**	.458**	.563**	1					
X_{12}	.294**	.246**	.187**	.251**	.181**	-.089*	.280**	.293**	.791**	.975**	.628**	1				
X_{13}	.304**	.249**	.194**	.244**	.157**	-.134**	.241**	.269**	.800**	.999**	.564**	.976**	1			
X_{14}	.323**	.250**	.219**	.246**	.167**	-.145**	.235**	.213**	.908**	.830**	.446**	.819**	.830**	1		
X_{15}	.522**	.849**	.422**	.801**	.357**	0.038	.515**	.374**	.461**	.409**	.310**	.403**	.410**	.422**	1	
X_{16}	-.194**	-.222**	-.164**	-.222**	-.077*	−0.039	-.163**	-.170**	-.181**	-.186**	-.101**	-.179**	-.187**	-.160**	-.258**	1

*p < 0,05 **p < 0,01.

Table 4
KMO Measure of sampling adequacy.

Variable	KMO
Overall	.851
Total_logins	.876
N_access_forum	.794
N_added_messages_forum	.878
N_access_didactic_units	.847
N_access_glossaires	.914
Total_assignments	.668
N_assignments_consulted	.799
N_assignments_submitted	.836
N_access_questionnaires	.859
N_ attempts_questionnaires	.813
N_answered_questions	.901
N_questionnaire_views	.955
N_questionnaires_submitted	.809
N_reviews_questionnaires	.866
N_entries_course	.882
Days_first_access	.953

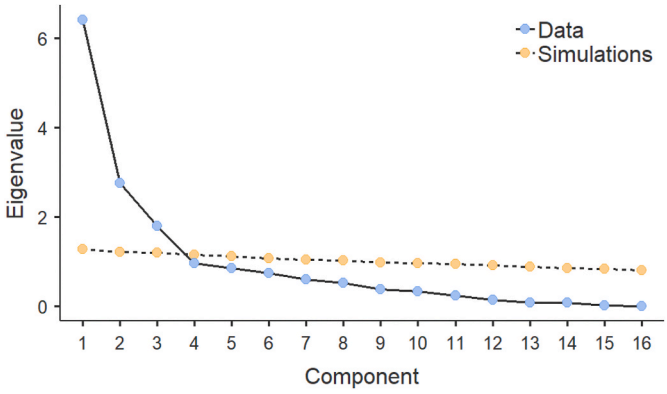

Fig. 1. Scree plot of the parallel analysis.

Table 5
Summary of three factors selected for the EFA.

Component	SS Loadings	% of Variance	Cumulative %
1	4.87	32.5	32.5
2	3.43	22.9	55.4
3	2.33	15.5	70.9

Table 6
Factor loadings and uniqueness.

Variable	F1	F2	F3	Uniqueness
Total_logins		.547		.638
N_access_forum		.888		.181
N_added_messages_forum		.684		.484
N_access_didactic_units		.893		.152
N_access_glossaires		.432		.721
Total_assignments			.835	.264
N_assignments_consulted			.756	.245
N_assignments_submitted			.786	.289
N_access_questionnaires	.860			.193
N_ attempts_questionnaires	.957			.055
N_answered_questions	.600		.425	.454
N_questionnaire_views	.952			.059
N_questionnaires_submitted	.956			.055
N_reviews_questionnaires	.886			.167
N_entries_course		.846		.170
Days_first_access				.883

and glossaries, the number of added messages to forums, and the frequency that students visit didactic units. For this reason, we decided to call this factor Access Factor (AF). Finally, the third factor is made up of three variables related to the assignments (tasks): the total of assignments performed, consulted, and submitted. Then, it has been decided to call it Task Factor (TF).

It is important to note that the variable "number of days until the first access to the classroom" does not load in any factor and also has a uniqueness value greater than 0.8. For these reasons and for its biased distribution, we have decided not to include it in any factor.

4.3. Description of factors

Figs. 2–4 show the distributions of the QF, AF, and TF factors and their descriptive statistics are presented in Table 7. The statistics of asymmetry and kurtosis, being less than 0.6, indicate that the distributions are not significantly biased nor have a high kurtosis. In addition, box plots not exhibit outliers. Although, the TF factor contains outliers, we do not remove these data as this factor already had adequate kurtosis levels and symmetry.

4.4. Prediction of the academic performance in the course

The multiple regression analysis was made with four independent variables: using the three factors identified in the EFA and incorporating age. It is important to note that age variable is the only quantitative demographic variable. Table 8 shows that all variables significantly predict the final grade of the course (Y_1). It is worth noting that the age variable predicts negatively, which indicates that at more age less grade is obtained. In addition, multicollinearity was tested by tolerance values and Variance Inflation Factor (VIF). There is no collinearity among variables since none of tolerance values are less than 0.1. Moreover, none of the VIF values are above 2.5 supporting no collinearity.

Table 9 shows that the model fit is good and the four independent variables explains 45.5% of the variance of the students' academic performance in the target course. On the other hand, independence was tested with the Durbin-Watson statistic, in this case D.W. = 1.88 indicating independence.

4.4.1. Prediction of the academic performance in subsequent academic years

Table 10 exhibits the model fit measures for the 2012/13, 2013/14, 2014/15, 2015/16, and 2016/17 years. The model for the 2012/13 year explains 17,5% of the variance of the students' academic performance measured by the average grade of the academic year. The model for the 2013/14 year explains 13.9% of the variance. The model for the 2014/15 year explains 11.5% of the variance. For the 2015/16 year the model explains 11.2% of the variance, and for the 2016/17 year the model explains 10.04% of the variance. Likewise, the model fit is worse as the year progresses. Thus, the difference in BIC and AIC criteria each year is over 10 indicates a significant decrease in the model fit of the models.

Table 11 exhibits B coefficients, Standard Error (SE), standardized B

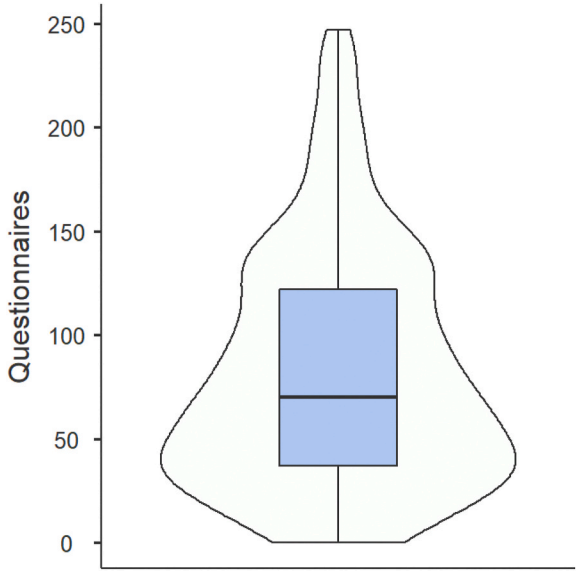

Fig. 2. Distribution of QF factor.

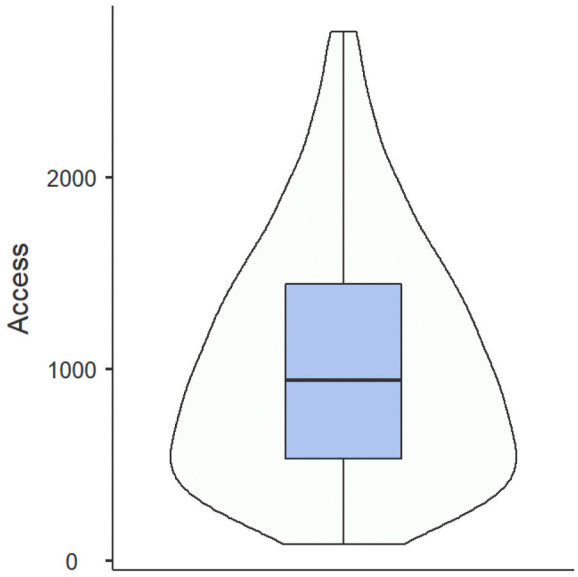

Fig. 3. Distribution of AF factor.

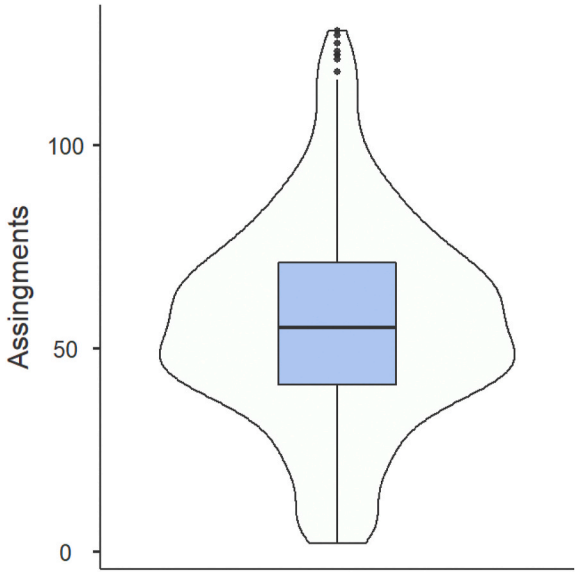

Fig. 4. Distribution of TF factor.

Table 7
Description of the factors.

Factor	Mean	Median	Min	Max	Skewness	Kurtosis
QF	82.4	70	0	247	.568	-.010
TF	56.4	55	2	128	.242	.171
AF	1040	942	84	2759	.452	-.240

Table 8
Model coefficients and collinearity diagnosis.

Variable	*B*	*SE*	t-value	p value	Standard B	VIF	Tolerance
Intercept	5.149	.320	16.05	<.001	17.47		
Task	.038	.002	13.45	<.001	.418	1.36	.734
Access	.000	.000	6.52	<.001	.209	1.45	.688
Questionnaire	.007	.001	7.35	<.001	.217	1.23	.813
Age	-.049	.008	−5.87	<.001	-.163	1.09	.922

Table 9
Model fit measures.

Model	R	R^2	Adjusted R^2	AIC	BIC	RMSEA
1	0.677	0.458	0.455	3448	3480	.126

Note: AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria; RMSE: Root Mean of Square Error.

Table 10
Model fit measures.

Model	R	R^2	Adjusted R^2	BIC	AIC	RMSEA
2012/13	0.433	0.187	0.175	898	925	0.844
2013/14	0.396	0.157	0.139	668	692	0.904
2014/15	0.369	0.136	0.115	579	603	0.924
2015/16	0.356	0.127	0.112	488	510	0.944
2016/17	0.423	0.119	0.104	316	334	0.953

coefficients, t values, p values, and collinearity statistics of the models corresponding to each academic year. From these data it can be seen that prediction accuracy descended each year. For instance, at the end of 2012/13 the Task ceases to be a predictor of the performance, and at the end of 2013/14 the Access ceases.

4.5. Cluster analysis

The dendrogram shows that the students can be classified into five groups: the first is made up of 214 (36.1%), the second 205 (34.6%), the third by 87 (14.6%), the fourth by 27 (4.6%) and the last by 60 (10.2%) students. In Fig. 5, the students and the group to which they belong are projected in the first factorial plane.

With regards to cluster interpretation, it can be seen that there are five groups. These groups are described in the following lines:

- Group 1: According to *v.test* this group is mainly composed by students of Computer Science, Journalism, and Psychology, the continuous variables that characterize them are summarized in Table 12. In this table we can see that it is a group that is characterized by having few interactions in the platform (they write few messages in the forums, enter with low frequency to the platform and the course, rarely access the questionnaires, do not access to course materials, submit few tasks, …) and lower grades both in the course and in subsequent years. Also, they are characterized by having a lower age. In Fig. 5 we can see this group at the left.

- Group 2: Group two is mainly composed by Computer Science, Criminology, and Psychology students. The continuous variables that characterize them are summarized in Table 13. In this table we can see that it is a group that is characterized by consulting and sending many assignments and they have a high frequency of viewing questionnaires. Their course grade is also high. However, their average grade in the academic year 2014/15 is lower than the overall mean. Also, they are characterized by having a lower age, lower frequency of logins and entries in the course, and poor participation in forums. In Fig. 5 we can see this group at the center.

- Group 3: According to *v.test* this group is mainly composed by students of Civil Engineering. The continuous variables that characterize them are summarized in Table 14. In this table we can see that it is a group that is characterized by their high access to didactic material, forums, glossaries, and their high participation in the forums. They are also characterized by having a higher grade in the course and in subsequent years. Also, they are characterized by have a higher age. In Fig. 5 we can see this group at the right.

Table 11
Model coefficients and collinearity diagnosis of subsequent academic years (2012–2017).

Year	Variable	*B*	*SE*	t-value	p value	Standard *B*	VIF	Tolerance
2012/13	Intercept	6.275	.196	31.99	<.001			
	Task	.002	.002	1.33	.183	.071	1.25	.643
	Access	.003	.000	4.48	<.001	.232	1.13	.812
	Questionnaire	.000	.000	3.67	<.001	.190	1.25	.823
	Age	-.016	.004	−3.36	<.001	-.166	1.10	.764
2013/14	Intercept	6.013	.265	22.65	<.001			
	Task	-.001	.002	−0.37	.710	-.023	1.27	.638
	Access	.002	.001	2.73	.007	.168	1.14	.763
	Questionnaire	.000	.000	2.93	.004	.184	1.24	.873
	Age	-.025	.006	−4.00	<.001	-.244	1.09	.645
2014/15	Intercept	6.217	.297	20.89	<.001			
	Task	.000	.003	.08	.932	.005	1.29	.618
	Access	.001	.001	1.29	.198	.086	1.14	.743
	Questionnaire	.000	.000	3.77	.028	.164	1.21	.622
	Age	-.164	.007	−2.28	.023	-.154	1.05	.653
2015/16	Intercept	5.995	.332	18.04	<.001			
	Task	.001	.003	.31	.752	.023	1.29	.658
	Access	.002	.001	2.00	.047	.149	1.13	.797
	Questionnaire	.000	.000	3.21	.002	.242	1.28	.826
	Age	-.019	.008	−2.44	.016	-.178	1.08	.645
2016/17	Intercept	5.763	.042	14.31	<.001			
	Task	-.004	.004	-.97	.334	-.094	1.29	.667
	Access	.004	.001	2.53	.013	.236	1.19	.784
	Questionnaire	.000	.000	1.34	.180	.130	1.22	.821
	Age	-.031	.010	−2.99	.003	-.286	1.06	.792

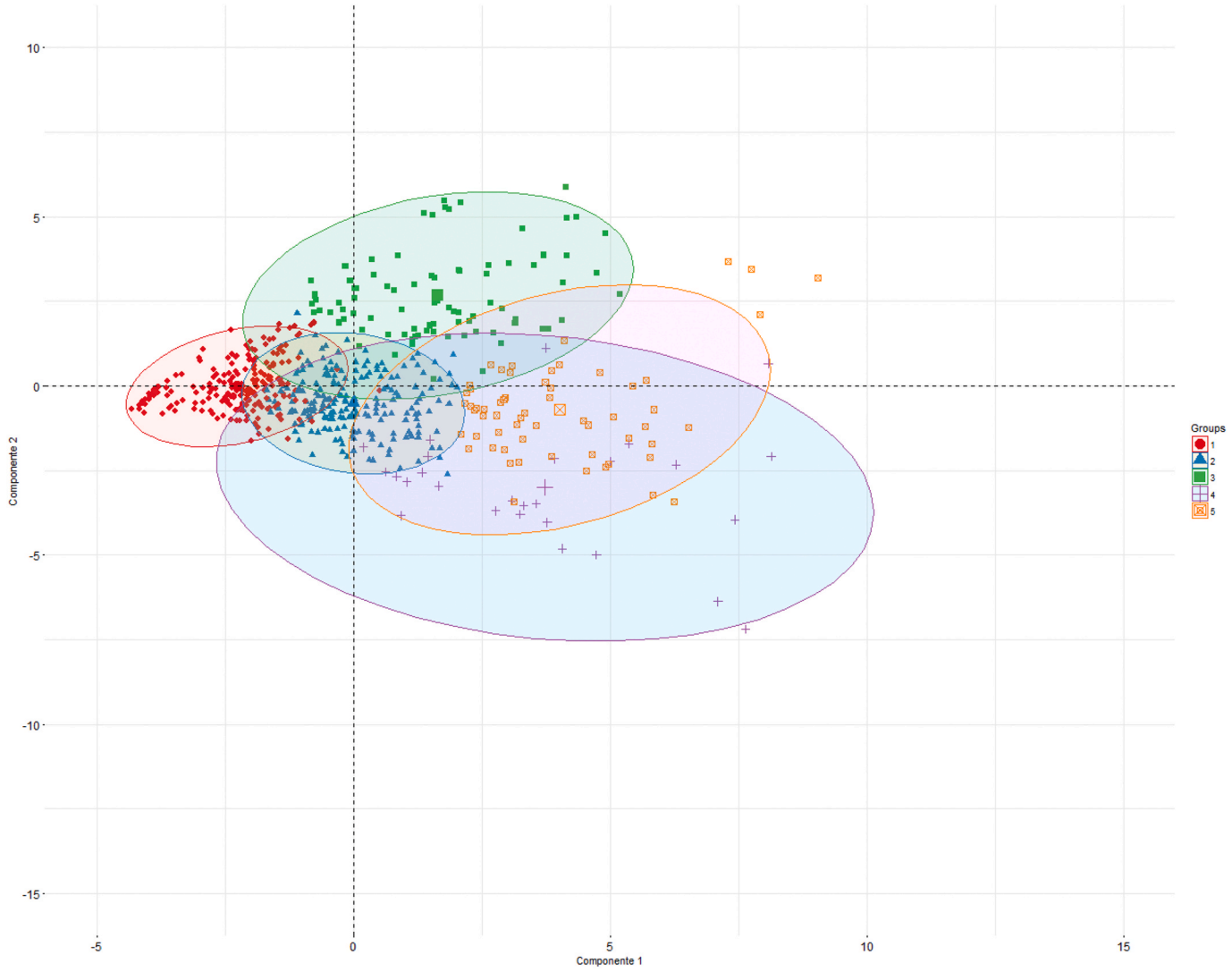

Fig. 5. Students projected in the first factorial plane.

Table 12
Continuous variables that describe the first group.

Variables	v.test	Mean in category	Overall mean	*sd* in category	Overall *sd*
GPA_14_15	−2.39	7.43	7.53	0.76	0.75
Age	−2.87	35.03	36.56	9.26	9.72
GPA_16_17	−3.08	7.21	7.32	0.58	0.62
GPA_15_16	−3.26	7.38	7.51	0.73	0.71
GPA_13_14	−3.72	7.34	7.51	0.77	0.83
N_added_messages_forum	−5.16	1.24	2.20	1.80	3.40
Total_logins	−5.20	667.86	847.08	501.08	628.22
GPA_12_13	−7.74	7.23	7.63	0.97	0.95
N_access_glossaires	−8.00	1.48	3.79	2.91	5.25
Total_assignments	−8.51	4.55	5.09	1.03	1.14
N_access_forum	−8.79	38.96	93.61	39.77	113.29
N_access_didactic_units	−9.04	39.61	82.10	38.69	85.65
N_answered_questions	−10.21	4.70	20.76	4.09	28.69
N_entries_course	−10.79	73.33	147.83	48.92	12.58
N_assignments_submitted	−10.91	3.69	5.22	2.22	2.54
Grade_course	−11.74	6.15	7.67	3.10	2.36
N_assignments_consulted	−12.09	30.59	50.17	16.10	29.53
N_reviews_questionnaire	−12.74	5.02	16.73	3.72	16.75
N_access_questionnaires	−13.69	12.08	31.47	7.73	25.83
N_attempts_questionnaires	−13.85	3.75	9.11	2.46	7.05
N_questionnaires_submitted	−13.86	3.67	9.02	2.46	7.04
N_questionnaire_views	−14.36	3.97	9.98	2.65	7.63

Table 13
Continuous variables that describe the second group.

Variables	v.test	Mean in category	Overall mean	*sd* in category	Overall *sd*
Total_assignments	10.36	5.76	5.09	0.48	1.14
N_assignments_submitted	6.84	6.20	5.22	1.85	2.54
Grade_course	6.72	8.57	7.67	1.02	2.36
N_assignments_consulted	3.64	56.27	50.17	21.07	29.53
N_questionnaire_views	2.03	10.86	9.98	3.63	7.63
GPA_14_15	−2.07	7.44	7.53	0.74	0.75
Age	−2.50	35.18	36.56	9.67	9.72
Total_logins	−3.48	722.96	847.08	470.74	628.22
N_entries_course	−4.58	115.16	147.83	54.62	12.58
N_added_messages_forum	−5.64	1.11	2.20	1.78	3.40
N_access_forum	−5.85	55.99	93.61	49.77	113.29
N_access_didactic_units	−6.04	52.76	82.10	37.07	85.65

Table 14
Continuous variables that describe the third group.

Variables	v.test	Mean in category	Overall mean	*sd* in category	Overall *sd*
N_access_didactic_units	16.46	222.80	82.10	88.77	85.65
N_access_forum	15.77	271.94	93.61	125.68	113.29
N_entries_course	13.33	315.31	147.83	144.72	125.86
N_added_messages_forum	12.37	6.40	2.20	5.59	3.40
N_access_glossaires	8.68	8.34	3.79	8.26	5.25
Total_logins	6.33	1243.94	847.08	862.86	628.22
N_assignments_consulted	5.99	67.83	50.17	25.07	29.53
GPA_12_13	4.84	8.09	7.63	0.84	0.95
GPA_14_15	4.74	7.88	7.53	0.71	0.75
GPA_13_14	4.28	7.86	7.51	0.73	0.83
Grade_course	4.21	8.67	7.67	1.05	2.36
GPA_16_17	3.65	7.55	7.32	0.72	0.62
Age	3.60	40.05	36.56	9.92	9.72
GPA_15_16	3.46	7.75	7.51	0.62	0.71
N_assignments_submitted	3.25	6.04	5.22	1.95	2.54

- Group 4: According to *v.test* this group is mainly composed by students of Criminology, Journalism, and Business Administration. The continuous variables that characterize them are summarized in Table 15. In this table we can see that this group is characterized by their high use of questionnaires (views, attempts, submitted, access, reviews) and their high participation in the forums, logins and entries in the course. They are also characterized by having a low participation in assignments. In Fig. 5 we can see this group at the lower-right.

- Group 5: According to *v.test* this group is mainly composed by students of Law and Criminology. The continuous variables that characterize them are summarized in Table 16. In this table we can see that it is a group that is characterized by their high access to didactic material, forum, and assignments. They are also characterized by having a higher use of questionnaires (views, attempts, submitted, access, question answered). In Fig. 5 we can see this group at the upper-right.

5. Discussion

The results obtained provide information to meet the objectives outlined in the introduction of the present paper. On the one hand, we have found a group of variables that allows predicting the academic performance of a sample of undergraduate students using data collected from an LMS during an academic semester (G1 and G2). These variables may be considered as EWI in order to carry out preventive support measures. On the other hand, we analyzed the relationship between variables and developed a complete description of the five students' profiles that we found (G3).

The factors that have been predictive in the short term (course) are Task (frequency of actions that each student performs with the Moodle tasks), Access (related to the frequency of logins, messages sent in the forums and access to educational material, forums and glossaries), Questionnaire (frequency of actions that each student performs with the Moodle questionnaires), and Age.

Task is the factor that most contributes to the student performance in the course. This result makes sense because tasks are the most important assignments of the course, account for about twenty per cent of the final grade, require generally about 20 h to complete and have a deadline set in the course.

The factor Access indicates that students who use Moodle more intensively and more frequently also receive higher grades. This result can be explained by the university instructional design, as instructors publish new materials in the virtual classroom on a weekly basis and reply to forums daily. Therefore, adequate study of the course requires access to the virtual classroom on a regular basis and examine the new educational material, forums, and glossaries. Furthermore, this result is in line with the works of Nistor and Neubauer (2010) and Romero et al. (2013) in relation to the positive relationship between the number of messages in the forums and academic performance.

The factor Questionnaire is related to regular study. The questionnaires are usually made up of multiple-choice questions and serve to assess concept understanding. They are published at the beginning of the semester and closed at the end. Each student is given two attempts per questionnaire. In this way, learners who study the course on a regular basis take the questionnaires throughout the course and use the two attempts. Therefore, a high number of questionnaires could indicate a regular study.

The composition of our prediction model could be explained by the theory of self-regulated online learning which proposed that online students who manage their time appropriately, are critical in examining the content and persevere in understanding the learning material are more likely to achieve higher academic grades in the online settings (Broadbent & Poon, 2015).

Through the five-year analysis it is observed that the variables that positively influence long-term performance are the same to those that influence the short term (Access, Questionnaire, and Age) except Task. It is important to note that one of the variables that appears constantly in the models is Age, as a negative predictor of performance. This result

Table 15
Continuous variables that describe the fourth group.

Variables	v.test	Mean in category	Overall mean	*sd* in category	Overall *sd*
N_ attempts_questionnaires	12.88	26.22	9.11	9.04	7.05
N_questionnaires_submitted	12.86	26.07	9.02	8.85	7.04
N_access_questionnaires	12.30	91.29	31.47	46.77	25.83
N_ reviews_questionnaires	12.03	54.66	16.73	29.53	16.75
N_questionnaire_views	11.71	26.81	9.98	9.19	7.63
N_added_messages_forum	2.97	4.11	2.20	3.20	3.40
N_entries_course	2.70	211.81	147.83	147.27	125.86
Total_logins	2.57	1150.92	847.08	513.00	628.22
N_assignments_consulted	−5.75	18.22	50.17	10.77	29.53
N_assignments_submitted	−5.88	2.40	5.22	1.59	2.54
Total_assignments	−13.80	2.11	5.09	0.56	1.14

Table 16
Continuous variables that describe the fifth group.

Variables	v.test	Mean in category	Overall mean	*sd* in category	Overall *sd*
N_answered_questions	13.98	69.91	20.76	54.67	28.69
N_questionnaire_views	12.69	21.85	9.98	6.27	7.63
N_questionnaires_submitted	11.79	19.20	9.02	5.61	7.04
N_ attempts_questionnaires	11.76	19.28	9.11	5.58	7.05
N_ reviews_questionnaires	11.21	39.75	16.73	17.15	16.75
N_assignments_consulted	10.46	88.03	50.17	36.95	29.53
N_access_didactic_units	10.29	64.06	31.47	19.03	25.83
N_entries_course	6.92	254.56	147.83	143.49	125.86
N_assignments_submitted	6.84	7.35	5.22	2.65	2.54
Total_assignments	5.24	5.83	5.09	0.37	1.14
N_access_didactic_units	4.99	134.50	82.10	87.33	85.65
Total_logins	4.58	1199.81	847.08	701.68	628.22

might be supported by the fact that cognitive mechanisms, necessary to learn something new, declines with age (Ericsson, 2018; Park, Polk, Mikels, Taylor, & Marshuetz, 2001).

Regarding the limitations of our prediction model, we did not consider personal student data in the initial set of variables (19 independent variables). On one hand, Hattie (2009) indicates that the prior knowledge of learning is an interesting variable that could influence the student learning. On the other hand, Hattie (2009) found that the prior knowledge has less contribution to learning than other variables related to actions of students in classrooms. Then, we did not include the prior knowledge of learning in the initial set of variables since it is not mandatory that UDIMA collects these data.

Following the third objective five clusters of students were clearly identified and validated. Only few previous researchers have employed this technique and it is one of the principal contributions of the present research. Group 1 is characterized by having few interactions in the LMS, lower grades both, in short and large term, and lower age. This group is similar to the cluster 4 found in the study of Cerezo et al. (2016). Group 2 shows higher frequency of consults and submitting of assignments, and a higher view of questionnaires. Their course grade is also high. Being focused on practical tasks, this group would be similar to cluster 3 of Cerezo et al. (2016). Group 3 exhibits high frequency of access to teaching materials, forums, glossaries, and higher participation in forums. They have higher grades in short and long term, and higher age. This group may be similar to the cluster 2 of Cerezo et al. (2016) that also is characterized by their number of actions and frequency of access to teaching materials. Group 4 shows a high use of questionnaires and higher participation in the forums, logins, and entries in the course, but low participation in assignments. It is important to emphasize that a group with these characteristics was not found in the Cerezo et al. (2016). Finally, the group 5 is characterized by higher frequency of access to teaching materials, forums, and assignments, being similar to the cluster 1 reported by Cerezo et al. (2016).

It is important to note the similarities between the results of the multiple regression and the cluster analysis as the groups that are characterized by higher grades (groups 2 and 3) are also the ones with the highest frequency of access to teaching materials, forums, glossaries, and higher participation in forums, and again, they are the ones who are older. These results are matching with previous studies that have also observed that academic performance is strongly determined by patterns of interactions that involve active participation (Agudo-Peregrina et al., 2014; Gómez-Aguilar, Hernández-García, García-Peñalvo, & Therón, 2015). An interesting peculiarity of group 2, focused on practical tasks, is that only exhibits a higher grade in the course, it seems to indicate an influence of tasks in the short term but not in the long term. It may be interesting to deepen this result in future research. Results of correlation analysis are also matching with those found in the cluster and regression analysis because all the variables correlate significantly and positively except the number of days to first access the virtual classroom.

One contribution of the present work is that it is done in a university that is completely online in which all the courses share the same instructional design, which provides a great advantage for the early prediction because the variables can be interpreted in a similar way and we can include participants of different degrees. The results of the present research may be used in other online universities that have similar instructional conditions (activities, didactic units, glossaries, questionnaires, …). Another important contribution is that by obtaining similar results in the short and long term regression models we can identify the common predictor variables that may be considered as EWI in order to carry out preventive support measures as educational programs to promote self-regulation in the groups with detected problems (for example, younger students, who do not participated in forums, that submit few questionnaires and assignments, etc.).

Notwithstanding the contributions of this research, it has an inherent limitation to the cluster since, due to its nature, it depends on the sample being analyzed, and so, the results of the present research are not generalizable to the population.

6. Conclusions

This work proposed a collection of models that could be useful to consistently predict the academic performance of students at the end of a degree. These models utilized variables of two data sources: LMS interaction data of students and institutional data that included information of student enrollment, age and sex of students, and GPA of each academic year from 2012 to 2017. The models presented in this work make an early prediction using LMS students' interaction data of the first semester of the 2012–2013 academic courses.

According to our results we found that the implications for online education are mainly related to intervention strategies. They should be designed to improve the students' academic performance in an early stage of their studies.

Another contribution that it is important to emphasize is that the present research context was a CO university with degrees that share the same instructional approach. This uniformity between courses avoids possible measurement bias errors in our results. In addition, we have utilized variables that are grounded on the self-regulation online learning theory. These characteristics reinforce the generalizability of our models to similar online contexts.

References

Agudo-Peregrina, A. F., Iglesias-Pradas, S., Conde-González, M. A., & Hernández-García, A. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior, 31*, 542–550.

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies academic & achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education, 27*, 1–13.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Compututers & Education, 96*, 42–54.

Chen, F., Myers, B., & Yaron, D. (2000). *Using handheld devices for tests in classes. CMU-CS-00-152*.

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies, 10*(1), 17–29.

Ericsson, K. A. (2018). The differential influence of experience, practice, and deliberate practice on the development of superior individual performance of experts. In *The cambridge handbook of expertise and expert performance* (pp. 745–769). Cambridge, UK: Cambridge University Press.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68–84.

Gómez-Aguilar, D. A., Hernández-García, A., García-Peñalvo, F. J., & Therón, R. (2015). Tap into visual analysis of customization of grouping of activities in eLearning. *Computers in Human Behavior, 47*, 60–67.

Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement*. New York, NY: Routledge Taylor & Francis Group.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/bf02289447

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education, 37*, 66–75.

Husson, F., Josse, J., & Le, S. (2016). *FactoMineR: Factor analysis and data mining with R. R package version 1.32. Paris, France*.

Husson, F., Leˆ, S., & Pagès, J. (2010). *Exploratory multivariate analysis by example using R*. London, UK: Chapman y Hall.

Knowles, J. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining, 7*(3), 18–67.

Lebart, L., Morineau, A., & Piron, M. (2000). *Statistique exploratoire multidimensionnelle*. Francia: Dunod: París.

Lerche, T., & Kiel, E. (2018). Predicting student achievement in learning management systems by log data analysis. *Computers in Human Behavior, 89*, 367–372.

Lu, O. H. T., Huang, A. Y. Q., Huang J C, H., Lin, A. J. Q., Ogata, H., & Yang, S. J. H. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Journal of Educational Technology & Society, 21*(2), 220–232.

Nistor, N., & Neubauer, K. (2010). From participation to dropout: Quantitative participation patterns in online university courses. *Computers & Education, 55*(2), 663–672.

Pardo, C. E., & Del Campo, P. C. (2007). Combinación de métodos factoriales y de análisis de conglomerados en R: El paquete FactoClass. *Revista Colombiana de Estadística, 30*(2), 231–245.

Park, D. C., Polk, T. A., Mikels, J. A., Taylor, S. F., & Marshuetz, C. (2001). Cerebral aging: Integration of brain and behavioral models of cognitive function. *Dialogues in Clinical Neuroscience, 3*(3), 151–165.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432–1462.

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68*, 458–472.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of a known factorial structure. *Psychological Assessment, 24*(2), 282–292. https://doi.org/10.1037/a0025697

Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an institutional context. *The Internet and Higher Education, 37*, 76–89.

Sclater, N., Peasgood, A., & Mullan, J. (2016). *Learning analytics in higher education: A review of UK and international practice full report. Tech. Report*. Bristol, UK: Jisc.

The jamovi project. (2019). *jamovi* (Version 1.1) [Computer Software]. Retrieved from https://www.jamovi.org.

The R Development Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Xu, J., Moon, K. H., & Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing, 11*(5), 742–753.

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education, 29*, 23–30.