Computer Science and Information Systems 20(1):175–194 https://doi.org/10.2298/CSIS211110050J

Analyzing Feature Importance for a Predictive Undergraduate Student Dropout Model

Alberto Jiménez-Macias[1], Pedro Manuel Moreno-Marcos[1], Pedro J. Muñoz-Merino[1], Margarita Ortiz-Rojas[2, 3], Carlos Delgado Kloos[1]

[1] Universidad Carlos III de Madrid, Avda de la Universidad, 30
E-28911, Leganes, Spain
{albjimen, pemoreno, pedmume, cdk}@it.uc3m.es
[2] Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Km. 30.5
Vía Perimetral EC-090112, Guayaquil, Ecuador
maelorti@espol.edu.ec
[3] Ghent University, Henri Dunantlaan 2,
B-9000, Ghent, Belgium
margaritaelizabeth.ortizrojas@ugent.be

Abstract. Worldwide, one of the main concerns of universities is to reduce the dropout rate. Several initiatives have been taken to avoid this problem; however, it is essential to recognize at-risk students as early as possible. This article is an extension of a previous study that proposed a predictive model to identify students at risk of dropout from the beginning of their university degree. The new contribution is the analysis of the feature importance for dropout segmented by faculty, degree program, and semester in the different predictive models. In addition, we propose a dropout model based on faculty characteristics to try to infer the dropout based on faculty features. We used data of 30,576 students enrolled in a Higher Education Institution ranging from years 2000 to 2020. The findings indicate that the variables related to Grade Point Average(GPA), socioeconomic factor, and a pass rate of courses taken have a more significant impact on the model, regardless of the semester, faculty, or program. Additionally, we found a significant difference in the predictive power between Science, Technology, Engineering, and Mathematics (STEM) and humanistic programs.

Keywords: dropout model, features importance, data mining, learning analytics.

1. Introduction

One of the main issues Higher Education Institutions (HEIs) often face is the high rates of students' dropout [41]. For example, Scheneider [37] reported that about 30% of students in the USA drop out in the first year, and the World Bank reported a dropout rate of about 22% in Latin America [27]. In addition, Schnepf [38] reported dropout rates between 16% and 33% in Europe, although lower rates (11% and 4%) were reported in Asian countries, like Korea and Japan. As high dropout rates can be found in most parts of the world, it is very relevant to analyze how this issue can be detected and how these rates can be decreased.

In order to solve this issue of the dropout rates, it is possible to collect data from students and analyze it using learning analytics. As universities store dozens of records about students, such as students' grades throughout their degree programs, and know whether students completed their studies or dropped out, this information can help anticipate dropout cases and reduce this problem. Mainly, it is possible to detect students at risk and develop predictive models to forecast possible dropouts [8, 39]. Early detection of dropout can be beneficial since this may allow carrying out interventions to address this issue. Some possible interventions can be offering an orientation to students (e.g., counseling sessions to guide students in the courses they should take) [42], offering financial aids or scholarships to students with economic issues [12], and so on. Moreover, apart from specific interventions designed for students, analyses can also provide insights about possible difficulties in how the degree program is organized (e.g., if the workload of a specific semester is unbalanced, that could cause dropout).

In addition, to provide proper and timely support to the students, it is also essential to detect the main factors behind the predictive models. For example, Del Bonifro et al. [15] identified that the number of credits acquired by the students was a significant factor for dropout. However, that number was not available at application time, and thus variables related to performance in high school had to be used. In addition, Abu-Oda and El-Halees [1] discovered that some courses might have a more significant influence on dropout (e.g., students who got a high grade in a specific course had a lower probability of dropout).

In this line, this work aims to conduct a study using several degree programs in an HEI in Ecuador to discover more important variables in students' dropout. This paper aims to address the following main research question: What variables significantly influence students' dropout? To analyze this question, the objectives of this paper are as follows:

Analyze the feature importance for a student dropout predictive model considering all degrees

1. Analyze the feature importance for a student dropout predictive model considering the semesters throughout a degree
2. Analyze the feature importance for a student dropout predictive model considering the faculty level
3. Propose a predictive model based on a model of the University's faculties characteristics to estimate the dropout

This paper is an extension of [20]. In [20], we presented an algorithm for early dropout prediction and resulted in the algorithm's prediction power in different degrees. This paper extends the analysis using feature importance for the different variables involved in the prediction. We aimed to determine the most critical variables and if the differences depend on time, faculty, or degree.

The article is structured as follows. Section 2 presents an overview of the relevant literature. Section 3 describes the dataset and how the predictive model is designed. Section 4 shows the results and discussion, and Section 5 draws conclusions from this analysis and suggests possible directions for future work in this area.

2. Related Work

There have been many contributions focused on dropout prediction. These contributions have been made at two different levels. Some have been done at the course level (i.e., predict who will drop out of a course), while others have been done at the degree program level (i.e., predict who will drop out a degree). Among the former group, there has been research on Massive Open Online Courses (MOOCs), where Moreno-Marcos et al. [29] made a review on prediction in MOOCs and found many different variables were relevant to predict students' performance, including variables related to self-regulated learning, interactions with videos and exercises in the platform. In addition, research has been done in university courses different from MOOCs. For example, Burgos et al. [10] predicted dropout in university courses through Moodle data, and they claimed their models helped reduce dropout by 14%. Moreover, Pereira et al. [34] predicted dropout as a lack of attendance in programming courses.

On the other hand, several works have focused on dropout detection at the degree program level. For example, Luo and Pardos [24] used data from 8 years of course enrollments to predict whether the students would graduate on time. Furthermore, Chen et al. [13] predicted dropout in nine different majors (mostly STEM majors) and found that survival analysis approaches could achieve promising results. When developing these models, one crucial aspect is the anticipation because if models are only accurate at the end of the course, they may not be effective. In this line, Márquez-Vera et al. [28] predicted whether students would continue studying in the following academic year using data from 419 Mexican students and found accurate results within the first 4-6 weeks of the course. Furthermore, Jiménez et al. [22] emphasized the importance of early predictions, and they optimized models to obtain reasonable accuracies of dropout prediction in the university programs after two semesters.

Apart from the anticipation, one of the key aspects to make these predictions are the predictor's variables used to generate the models. In this case, variables are often retrieved from the academic record (e.g., Student Information System). One typical variable is the GPA, combined with other grades. For example, Kang and Wang [23] included both the overall GPA and the term GPA to predict dropout, combined with other variables such as gender, ethnicity, time status (e.g., full-time, half-time), classification (e.g., freshman, sophomore.), and age. They found that while GPA is strongly associated with dropout, other variables could also achieve strong results.

Moreover, Ameri et al. [5] combined several features, including demographics, family background (e.g., parents' educational level), pre-enrollment attributes (e.g., high-school GPA and grades from the admission test), financial attributes, enrollment attributes, and academic attributes. They used GPA, the percentage of passed, dropped, and failed credits, and the credit hours attempts among those academic grades. They concluded that the variables with the highest impact were the high school GPA, the GPA, the percentage of failed credits, and the financial attributes.

Another relevant issue when designing a predictive model is to select the prediction algorithm. For example, Aulck et al. [6] used logistic regression, Random Forest (RF), and k-NN to predict dropout and found better predictors with logistic regression. In addition, Barbosa Manhães et al. [7] used several machine learning algorithms to predict dropout in several undergraduate STEM degree programs in a Brazilian university. They found that multilayered perceptron, logistic regression, Support Vector Machines

(SVMs), and RF were more accurate. Furthermore, Ortigosa et al. [32] applied an early student-dropout prevention system and used it in production. They suggested that tree-based methods, such as RF, can outperform other models (as in [21-31]), but they pointed out that the explicability of the models was essential when they were put in production. Because of that, they initially used RF in a lab environment, but they preferred using the C5.0 decision tree model in the production stage. For that stage, another relevant issue is the generalizability of the models, i.e., ensuring that a model trained with some students is valid for other students.

While some authors have tried to mitigate this issue of generalizability with machine learning techniques, such as assembling [14], it is not feasible to get a one-size-fits-all model, and the instructional conditions should be considered [18]. For example, significant differences can be found when predicting using models trained with different courses or students [30]. Because of that, it is essential to develop separate models for each degree program to keep the context as similar as possible and consider that models may not generalize over time (e.g., when the study plan changes). Moreover, it is vital to analyze the differences of the predictive models depending on the context (e.g., degree program or faculty).

In this context, this work aims to analyze the importance of variables when predicting dropout in different degree programs of a Higher Education Institution. This will better understand how the prominent variables in the models generalize or differ when changing the context. Mainly, the analysis contributes to understanding the importance of variables over time and across degree programs and faculties. Moreover, it provides insights into the essential variables when designing predictive models so that other researchers can adapt predictive systems in their institutions more efficiently.

3. Early Dropout Prediction Model

This section describes the dropout prediction model. The dataset, preprocessing process, and used algorithms are described. Figure 1 from our previous work [20] describes the predictive model; in the transformation phase, we cleaned the provided data. Then, we calculated additional indicators to be used as input in the model. After that, we removed the degree programs that had no graduating students or dropped out because these are new degrees at the university. Next, we calculated the indicators related to the academic history of the students in the selected degrees. Finally, we split the data set, leaving 80% for training and 20% for testing, ran the predictive model on the data, and evaluated the results obtained with the selected metrics.

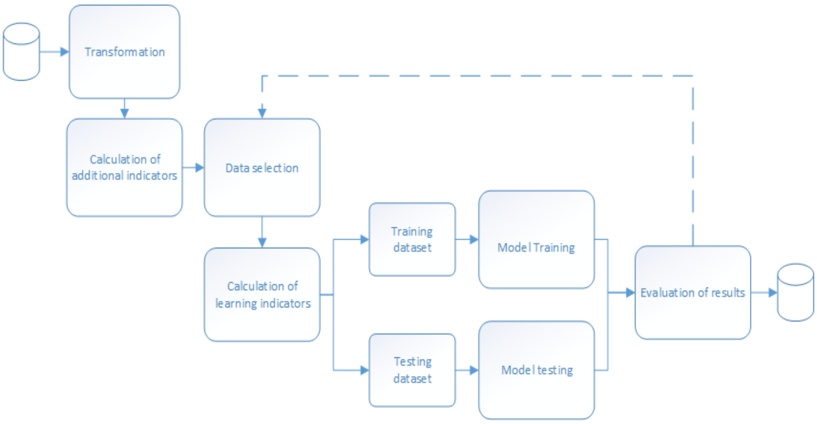

Fig. 1. Predictive model phases [20]

3.1. Dataset and pre-processing

The dataset used for this analysis contains demographic and academic data provided by the university's Information Technology department. This dataset contains academic records of 30,576 students enrolled from 2000 to the first semester of 2020 in 25-degree programs in 8 faculties. Table 1 shows the number of students in the degree programs with the largest number of students, and Table 2 shows the five faculties with the highest number of active students.

The data used as predictors included the following categories: (1) socio-demographic information (for example, employment status, city of residence, marital status, school of origin), (2) financial information (socioeconomic factor), (3) information on study program (e.g., number of credits to complete, course code) and (4) academic performance data (e.g., courses taken, courses ratio, course credits, among others).

During the pre-processing stage, cleaning, changing, and unifying techniques were performed due to the data structure. For the socioeconomic factor variable, we scaled the different student values to a single scale between 0 and 1, where 0 indicates a low socioeconomic level and 1 indicates a high socioeconomic level.

For this study, the academic semesters are identified according to the academic calendar. Particularly, the first and second semesters were called "ordinary semesters" and they normally started in May and October respectively. In addition, there is a third semester, called the "extraordinary semester", which comprises a short period of two months, generally in March and April of each year and this semester is usually used by students for two reasons: to pass courses previously failed or pending registration because the courses do not give flow to others. Students can take a maximum of two courses during the extraordinary semester, although they can decide whether to take them or not. Only ordinary semesters were considered for the predictive models because they had a similar duration of in-class weeks. The students' semesters were sorted chronologically to calculate the variables from each student's first semester in each course.

Table 1. Degree programs with the most active students

Degree Program code	Degree Program name	Faculty	Number of students
CI007	Mechanics	Faculty of Engineering in Mechanics and Production Sciences	792
CI005	Civil Engineering	Faculty of Engineering in Earth Sciences	743
CI013	Computing	Faculty of Engineering in Electricity and Computing	716
LI002	Economy	Faculty of Social and Humanistic Sciences	656
LI007	Business Administration	Faculty of Social and Humanistic Sciences	546
CI001	Industrial Engineering	Faculty of Engineering in Mechanics and Production Sciences	532
CI002	Chemical Engineering	Faculty of Natural Sciences and Mathematics	470
LI006	Production for Media	Faculty of Art, Design and Audiovisual Communication	469
LI005	Graphic Design	Faculty of Art, Design and Audiovisual Communication	457

Table 2. Faculty with most active students

Faculty name	Faculty	Number of students
FIEC	Faculty of Engineering in Electricity and Computing	2582
FCSH	Faculty of Social and Humanistic Sciences	2421
FIMCP	Faculty of Engineering in Mechanics and Production Sciences	2142
FCNM	Faculty of Natural Sciences and Mathematics	1184
FICT	Faculty of Engineering in Earth Sciences	1132
FADCOM	Faculty of Art, Design and Audiovisual Communication	995
FCV	Faculty of Life Sciences	623
FIMCM	Faculty of Maritime Engineering and Marine Sciences	536

After pre-processing, variables related to academic performance were precalculated to be used in the models. The complete list of variables is presented in Table 3. The variable V1 represents the socioeconomic level of the student and family of the student, V2 indicates the total number of times the student enrolled for the second time in a course after having failed it on a previous occasion in his or her current program, V3 indicates the total number of times the student enrolled for the third time in a course after having failed it twice on a previous occasion in his or her current program, V4 indicates the total number of years in which the student has not taken courses, V5 indicates the average of the grades in courses passed and failed registered in the current academic year, excluding cancelled courses, V6 indicates the weighted average of the grades in courses taken considering a penalty according to the number of times taken the same course, V7 indicates the proportion of courses passed by the student in the current undergraduate program, V8 indicates the proportion of courses failed by the student in the current undergraduate program, V9 indicates the proportion of courses with canceled enrollment by the student in the current undergraduate program. Some available sociodemographic variables, such as residence, school of origin, marital status, and employment status, were discarded due to the low correlation between the model output variable. University GPA is calculated by obtaining the average of the final grades among all the courses taken by the student. The final grade for each course is calculated by averaging the two highest grades of the three midterm grades.

Table 3. Learning indicators for the model

Variable ID	Variable	Category	Description	Values
V1	Economic factor	Financial information	Socio-economic factor	0 to 1
V2	Seg mat	Study program	Number of times of second enrollment in a course after failing it the first time	0 to a max number of courses
V3	Ter mat	Study program	Number of times of third enrollment in a course after failing it the second time	0 to a max number of courses
V4	Gap year	Study program	The period in years that the student takes to return to study	0 to maximum not defined (increase in intervals of 0.5 each semester without enrollment)
V5	Average APRP	Academic performance data	GPA of taken courses in the current undergraduate program	0 to maximum possible score (taken courses in other programs are excluded)
V6	Average weighted	Academic performance data	GPA of courses with a number of credits greater than zero, considering a penalty depending on the number of times the student takes the same course	0 to maximum possible score (Penalty: 0,9 for a second time and 0,8 for a third time)
V7	Ratio pass	Academic performance data	The ratio of passed courses by the student	0 to 1
V8	Ratio fail	Academic performance data	The ratio of failed courses by the student	0 to 1
V9	Ratio cancel	Academic performance data	The ratio of the canceled courses by the student	0 to 1

The predicting variable is whether or not a student will drop out in a degree, i.e., it is a categorical variable with only two possible values (0 for dropout and 1 for completion). We establish the dropout criterion when an undergraduate student has not taken any course for quite some time since their last enrollment. Considering the internal rules and guidelines of the university, this period is five years. Therefore, students who had not enrolled for more than five years were detected as dropouts.

In addition, undergraduate students who completed 90% of their college degrees were considered non-dropouts. This rate has also appeared in previous works [3], which showed that individuals with this high level of completion dropout of their college degrees for reasons unrelated to their academic performance.

3.2. Dropout algorithm for each degree

As for the machine learning algorithm, Random Forest classifier (RFC) computation was used [9], and using the RandomForestClassifier method of the sklearn library within the ensemble methods the GridSearchCV method to find the best parameters using Python as the programming language. We evaluated by purchasing the performance using the Area Under the Curve (AUC) metric, similar to [19,40 ,17]. RFC is a remarkable tree-based learning computation known for its low overprediction bias and high accuracy [9].

After pre-processing, the dropout model was trained for each degree program segmented by each semester and faculty meaning that the model used different input data for training depending on the semester or faculty as appropriate. In addition, specific data of students who had already graduated or dropped out were used to test the model. Subsequently, the model was run using active students in each of the first five semesters; for example, the first-semester degree program model was used for all students who had completed a semester in their degrees. The second model was used with students who had completed two semesters, and so on. After the fifth semester, all students who had completed more than five semesters in their undergraduate programs were grouped into a single model, including their interactions. As indicated in [16], regular school dropout occurs between the second and third years of Ecuador studies, and that is why the model focus on the first semesters.

3.3. Dropout algorithm for each Faculty

Using the dataset described in section 3.1, we propose a characterization model for the faculties using student interactions. The purpose of the model is to obtain the following characteristics: *average_weighted, ratio_pass, type_faculty* as described in Table 4. Table 2 describes the university faculties that are part of the study; the model learned based on the students who have studied in the eight faculties using as inputs the three variables described in Table 4.

Table 4. Learning indicators of faculty for the model

Variable ID	Variable	Description	Values	Type
V1	Average weighted	GPA of courses with a number of credits greater than zero, considering a penalty depending on the number of times the student takes the same course taken in the faculty's programs	0 to max (maximum possible is 10)	Output
V2	Ratio pass	The proportion of courses passed in the faculty	0 to 1	Output
V3	Type faculty	Indicates the type of faculty between STEM and NO_STEM	0 = No_STEM or 1 = STEM	Output

Along with the characteristics model described in the previous paragraph, we propose a global dropout model using as a focus the faculty that is the most global level within the university, a similar perspective to the model proposed in [25] where the author uses global and local data for each degree to calculate dropout. We use the output of the faculty characteristics model described in Table 4 as input to the dropout model and obtain as model output the probability of dropout in the faculties using Random Forest Classifier (RFC) as the algorithm. The model learns based on the faculties described in Table 2, obtaining eight rows to learn what is dropout and eight rows to learn what is not dropout.

4. Results and Discussion

4.1. Prediction Accuracy of the Model

In the first analysis, we averaged the values of each model's metrics using data from ten-degree programs. The results indicated that the different models could accurately predict dropout across all degree programs. Table 5 shows the results for ten randomly selected degree programs out of the 25-degree programs where the model was run, confirming the results as mentioned earlier. Using the AUC metric, the lowest results were obtained for CI004 with 0.80 and CI005 with 0.76. Although the results are not as good compared to the others of this study, there are semesters in both degree programs where the AUC metric is above average. In general, the model obtains good results for the metrics analyzed based on [29], indicating that obtaining an AUC is suitable between 0.8 and 0.9 inclusive. Our model can predict both a student who has completed a semester and a mid-career, similar to the levels obtained in [25,36] above 80%. student analyzed. Therefore, the design of proposed predictive model could be replicated using the same algorithm and variables in other higher education institutions with similar conditions to the one used in the present research.

Table 5 Prediction accuracy for the undergraduate programs between 2000 and first semester 2020

Code of Degree Programs	Degree	AUC average
CI003	Mining Engineering	0.87
CI004	Petroleum Engineering	0.80
CI005	Civil Engineering	0.76
CI008	Statistical Engineering	0.99
CI009	Logistics and Transportation Engineering	0.99
CI013	Computer Engineering	0.99
CI018	Telematics Engineering	0.98
ECCBA	Economy	0.98
INACP	Auditing and Certified Public Accountancy Engineering	0.99
INALL	Food Engineering	0.98

4.2. Feature Importance

This section shows the results obtained after analyzing the feature importance for the different degree programs in the trained model in the different semesters and faculties. The weight value of each variable was obtained through the Mean Decrease metric in Gini coefficient using the Python Shap library. Table 6 presents the frequency of use and the average weight of the variables. All values are used in at least one model within the degree programs. The variables that are used in all models are *average_aprp, rate_pass, economic_factor, average_weighted*.

The results show that the two most important variables in all models are: *average_aprp* and rate_pass. *Economic_factor* is the third most important model, confirming studies considering that this variable influences student dropout [11]; this result can be inferred from the country's economic indicators. The variable *average_weighted* has an average weight of 0.16 due to the relationship with the variable *average_aprp*. Both use the student grades in the different courses taken. The variables *seg_mat* and *ter_mat* have an average weight of 0.05 and 0.07, respectively. The *ratio_cancel* variable has an average weight of 0.02 in the model because students prefer to continue in the course until the end despite failing or dropping out after the first evaluation but without taking the corresponding administrative steps to cancel it. The *rate_fail* variable is only used in 1.1% of all the models (in two models) due to its high relationship with the *rate_pass* variable; it has high importance in the few models used. In conclusion, the ratio variables are correlated among the three, and consequently the low frequency of use of the ratio_cancel and ratio_fail variables. For future research, we recommend using the first four variables whenever possible because of their high predictive power, which was also supported in [2,11].

Table 6. General results of features importance

Variable ID	Variable	Frequency of use	Average weight
V5	Average APRP	100%	0.28
V7	Ratio pass	100%	0.27
V1	Economic factor	100%	0.16
V6	Average weighted	100%	0.16
V2	Seg mat	87.91%	0.07
V4	Gap year	80.22%	0.02
V3	Ter mat	75.82%	0.05
V9	Ratio cancel	39.56%	0.02
V8	Ratio fail	1.1%	0.19

Feature Importance per semester of the models for all the degree programs

To understand the behavior of the models, we averaged the values of the importance of the features in the models for each semester. Table 7 shows the average weights of the models for all the degree programs segmented by semester, the sum of the importance of all the characteristics is one for each semester. The four variables with the highest weights in the predictive model are: *average_aprp, rate_pass, economic_factor, average_weighted* in the first year. The variables *seg_mat, gap_year and ter_mat*, have no value initially because they only have information from one semester, and it is not possible to repeat a course or have years without studying. The *economic_factor* variable loses importance as the semesters of study increase; this could be due to the student's effort to finish his studies despite the economic problems that may occur as he progresses. The variables *average_aprp and ratio_pass* have a constant behavior during the models in the different semesters, due to the significant difference in these variables between graduates and dropouts. On the other hand, the variable *term_mat* has slightly increased weight from the third semester because a student may be taking his third enrollment in some course. The variables *gap_year and ratio_cancel* have little significant weight in the models, and the variable *ratio_fail* is used only in two models in the model of 4 and 6 or more semesters but had low weight.

As the student progresses in his degree, other variables are included as necessary in the model. For example, when the student is in six semesters or more, the *rate_fail* variable has a significant weight in the model, while the *socioeconomic_factor* variable has a low importance weight. The findings show a correlation between grades and pass rate variables, students in their first semester try to obtain a good grade point average in the courses registered. However, as they advance in their degree program, it is more important to pass the course without having so much weight on the grade obtained in the courses.

Table 7. Results of the average weight of features importance per semester

Semester	Average APRP	Ratio pass	Economic factor	Average weighted	Seg mat	Gap year	Ter mat	Ratio cancel	Ratio fail
1	0.33	0.23	0.24	0.20	0.01	-	-	-	
2	0.29	0.27	0.17	0.17	0.08	0.01	0.01	0.02	-
3	0.27	0.26	0.16	0.16	0.07	0.02	0.05	0.02	-
4	0.26	0.26	0.15	0.15	0.07	0.03	0.06	0.01	0.01
5	0.27	0.27	0.14	0.15	0.08	0.03	0.06	0.01	-
6 or more	0.27	0.34	0.08	0.12	0.08	0.03	0.05	0.02	0.02

Feature Importance per faculty

The behavior between the faculties could be different between the variables in the models. To clarify this hypothesis using data from all semesters, Table 8 shows the average weights of the importance of faculty characteristics. In the faculties shown, the four most important variables for the models are *average_aprp, ratio_pass, economic_factor,* and *average_weighted*. The variables *average_aprp, ratio_pass,* and *average_weighted* have the same trend between each faculty model. Therefore, student performance is an essential variable in the dropout model regardless of the student's faculty. In addition, the *economic_factor* variable presents two groups in its behavior. One is associated with the STEM degree programs (FIEC, FIMCP, FICT) with high values.

The other is related to the humanistic degree programs (FICSH, FCNM, FCV, etc. FIMCM, FADCOM) with low values because students of humanistic degrees usually have a higher economic level. The seg_mat variable is more important in the FCV and FIMCM faculties since they are relatively new and students leave semesters without studying. However, it is not a significant value concerning the rest, but shows a difference with the other faculties. The variable *ter_mat* has a lower weight in the FIMCP and FCV faculties, while the variable *gap_year* has a lower weight in FIEC, FICMP and FIMCM. In the FICSH, FICT, and FADCOM faculties, the variable *ratio_cancel* has a very low weight, almost zero concerning the other variables; we could infer that the students rarely canceled the courses in these faculties. Finally, in all the models, the variable ratio_fail does not have a significant weight. The models trained by faculty showed the same results based on the weight of the four variables compared to the models segmented by semesters. For students of all faculties, STEM or no-STEM is essential to pass the courses with a good grade. A gender is a variable that does not affect student dropout for this university, however Pilotti et al. [35] found that gender does affect student performance for STEM and non-STEM.

Table 8. Results of the average weight of features importance per faculty.

Faculty name	Average APRP	Ratio pass	Economic factor	Average weighted	Seg mat	Gap year	Ter mat	Ratio cancel	Ratio fail
FIEC	0.245	0.250	0.242	0.137	0.066	0.019	0.057	0.015	-
FICSH	0.261	0.280	0.113	0.165	0.058	0.032	0.073	0.06	-
FIMCP	0.278	0.224	0.216	0.156	0.069	0.017	0.039	0.023	-
FCNM	0.250	0.260	0.180	0.159	0.062	0.032	0.066	0.022	-
FICT	0.285	0.260	0.230	0.127	0.050	0.020	0.045	0.008	-
FADCOM	0.299	0.308	0.135	0.160	0.057	0.026	0.029	0.009	-
FCV	0.252	0.224	0.173	0.162	0.093	0.035	0.102	0.025	-
FIMCM	0.301	0.253	0.172	0.141	0.081	0.016	0.053	0.017	-

We selected four faculties from Table 2 with more students, segmented into two groups: STEM degrees and humanistic degrees. Figure 2 shows a comparison between the feature importance of the FIEC, FICMP faculties corresponding to STEM (FIEC and FIMCMP) and the faculties FCSH, FADCOM corresponding to humanistic degrees (FCSH, FADCOM) faculties. The results showed differences in behavior in some variables; for example, *economic_factor* in STEM schools' degrees had a greater weight than in humanities because the socioeconomic level in humanities faculties was usually higher than in STEM. The variable *ratio_pass* had a greater weight in humanistic faculties to STEM; this indicated that students could pass more courses in humanistic degrees. It should be clarified that the educational contents of the courses are different for the STEM and humanities faculties during the basic formation in degrees. For example, the courses: Calculus, Statistics, Linear Algebra, Programming has different content and evaluation forms for each of the faculties. In the year 2020, changes were made in the other degrees' curricular to unify these courses. All students can see the same content regardless of the faculty during their fundamental training.

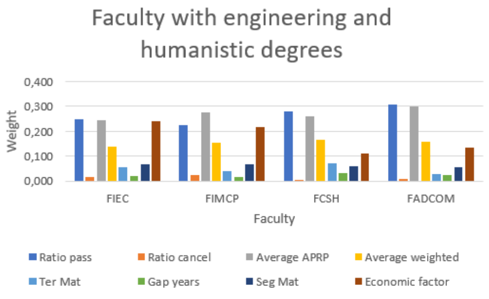

Fig.2. Feature importance between faculty STEM and humanistic degrees**.**

Among the findings, after analyzing the features importance for the different faculties, we found similar behavior in most of them. Thus, we decided to explore feature importance in some degree programs of the faculties to understand the behavior of these variables and identify if the same pattern of the faculties is found.

Feature Importance per degree program

Models predicting better results using characteristics for each degree than global characteristics were similar to [25,36]. We have used this local approach with attributes of each degree, obtaining similar results. In addition, we have performed an analysis in the ten-degree programs with the highest number of active students presented in Table 1, trying to understand the variables' behavior in the different degree programs. Table 9 shows the average weight of the variables in each of the degree programs. The four most important variables for these degree programs were *average_aprp*, *ratio_pass*, *economic_factor*, and *average_weighted*; This finding confirms the faculties' analysis results. The variable *average_aprp* had the lowest weight in the CI002 degree program compared to other programs where the behavior is similar. The variable *ratio_pass* had the lowest weight in the CI001. The *economic_factor* variable had a similar behavior for degrees: CI007 (STEM degrees), LI002, LI007, LI006, LI005 (humanistic degrees), and other behavior for degrees: CI005, CI013, CI001, CI017 (STEM degrees), LI006 (humanistic degree). The LI007 degree program presented a high value in weight for the *average_weighted* variable, significantly different from the other degree programs.

The variable *seg_mat h*ad the lowest value in the degree program CI005 and LI006. The variables *gap_year* and *ter_mat* presented the highest in the CI007 degree program; despite being small, they presented a more significant difference. The variable *ratio_fail* showed little significant weight for degree programs CI007 and LI005. In the faculties analysis, we found that the faculties with a low weight for the *ratio_fail* variable were FICSH, FICT, and FADCOM; the CI007 degree program belongs to the FIMCP faculty; therefore, if we only carried out an analysis by faculties, we could not find this type of findings. Finally, the variable *ratio_fail* had no importance for any model of the ten-degree programs shown. The results allow us to understand better the behavior of the variables in each degree program. These findings cannot be found in a general way if we only analyze the faculties due to the context of each degree program with different behavior on the part of the students.

Table 9. The average weight of feature importance per degree.

Degree	Average APRP	Ratio pass	Economic factor	Average weighted	Seg mat	Gap year	Ter mat	Ratio cancel	Ratio fail
CI007	0.285	0.268	0.177	0.140	0.073	0.064	0.151	0.008	-
CI005	0.258	0.224	0.259	0.148	0.051	0.015	0.055	0.016	-
CI013	0.221	0.265	0.259	0.122	0.073	0.015	0.066	0.020	-
LI002	0.287	0.284	0.134	0.133	0.065	0.024	0.085	0.010	-
LI007	0.253	0.219	0.159	0.217	0.087	0.037	0.052	0.011	-
CI001	0.226	0.177	0.318	0.166	0.076	0.018	0.035	0.014	-
CI002	0.196	0.206	0.320	0.150	0.060	0.012	0.041	0.025	-
LI006	0.278	0.240	0.204	0.163	0.054	0.026	0.040	0.027	-
LI005	0.265	0.299	0.188	0.149	0.063	0.027	0.028	0.005	-
CI017	0.238	0.213	0.287	0.127	0.070	0.014	0.053	0.024	-

We chose the most active students to identify patterns or differences between a STEM degree and a humanistic degree. Figure 3 shows the behavior of the variables in the CI007 degree models as it is the degree with the most active STEM students. The obtained grades and the passing rate were essential variables in all the semesters. The *socioeconomic_factor* was important in the model at the beginning of the degree. Still, when the student passed the middle of the degree, it had little relevance in the model due to the students' effort to finish the degree. The variable *seg_mat* was important in the second-semester model; we could infer that in the second semester of the degree, students repeated courses, while the weight of the variable *ter_mat* from the third semester maintained a significantly low value indicating that few students reached the third enrollment.

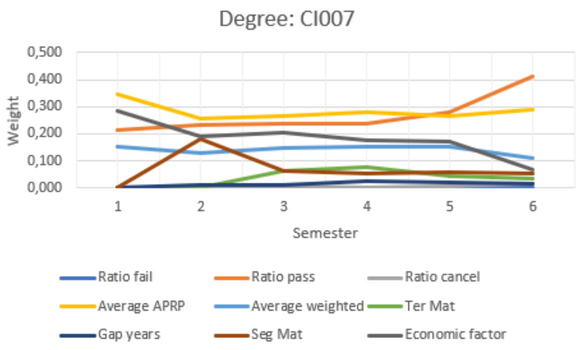

Fig.3. Weight of the variables segmented by the semester of the CI007

Figure 4 shows the behavior of the variables in the models for the LI002 degree because it is the degree with the most active students in the humanistic degree. The two most important variables in the degree model were *average_aprp* and *rate_pass*, exchanging importance after the second year (four semesters) of studies. These variables have a large gap concerning the other variables, with an average weight of less than 0.15 during each semester's different models. In the first semester, the economic level of the student influences the degree of dropout; as the student advances in the levels of the study program, the weight of this variable tends to decrease. The variables *seg_mat*, *gap_year*, and *ter_mat* were important in the first semester because there were only data from one semester studied. The *seg_mat* variable was important from the second semester. There is already data with students studying a course for the second time. Also, with the variable *ter_mat* from the third semester, students were already taking courses for the third time. Between the third and fifth semesters, more students took courses for the third time than for the second time. Finally, the variables *ratio_fail*, *gap_years*, and *ratio_cancel* were barely significant for the degree.

The first difference that we can observe in the analysis of degrees concerning faculties is the *average_arprp* variable. This variable is the most important in the first two years of both degrees, but it is the fourth important variable in the faculties analysis models. The variables *ter_mat* and *seg_mat* have similar behavior to the STEM degree. In contrast, to the humanistic degree, the variable *ter_mat* has a higher weight than *seg_mat*. In this case, humanistic degree students take more courses for the third time than STEM students. While during the middle stage of their degree, the variables have almost similar behavior in both cases.

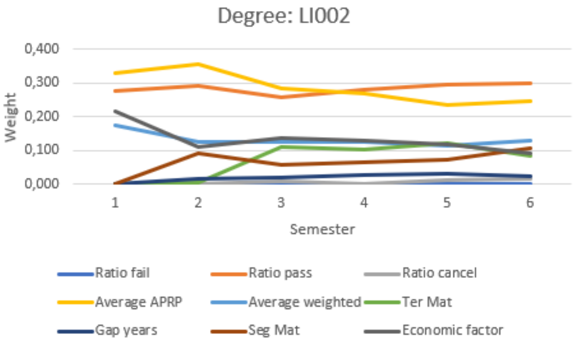

Fig.4. Weight of the variables segmented by the semester of the LI002

4.3. Model of Faculty

Using the model of faculty characteristics described in Section 3.3, we calculated the three variables listed as results in Table 4 for the faculties using the students currently associated with the respective faculties as input. Table 10 shows the results obtained using the proposed model of characteristics. According to the results obtained, the eight faculties can be grouped into 3 clusters as they have similar results for the characteristics. The proposed model has a limitation, so some metrics could not evaluate it [33] because the data used to learn the model are only from 8 faculties. Therefore, we would need data from more faculties from other universities to measure the algorithm's robustness or perform data simulations.

Table 10. Results of model faculty

Faculty ID	Average weighted	Rate pass	Faculty type
F1	7.13	0.6	NO STEM
F2	7.7	0.89	STEM
F3	7.13	0.6	NO STEM
F4	7.13	0.6	NO STEM
F5	7.84	0.92	NO STEM
F6	7.7	0.89	STEM
F7	7.7	0.89	STEM
F8	7.7	0.89	STEM

Using the data described in Table 10, we run the proposed dropout model. Table 11 shows the result obtained using the model for the faculties; we calculated the dropout probability using two different data sets: academic data for the last 20 years and academic data for the last ten years. To find a difference between the behavior of graduates and dropouts in the faculties, we shorten the range of years by dividing the input data used in the model into two groups.

The results show that there is a significant difference between the dropout measurement using as a data set the last 20 and 10 years in Faculties 1,3,4,5 because the students who are currently students have a weighted average and a pass rate closer to the students who graduated in the last ten years concerning those who graduated in the last 20 years. On the other hand, in Faculties 2,6,7,8, there is a contracting effect in which the increase in dropouts is because the data used as input are nearer to the group of students who have dropped out in the last ten years.

The proposed dropout model cannot be evaluated using some metric [33] because the data used with which the model was trained was very few: dropouts with eight rows and non-dropouts with eight rows. We would need more data from faculties of other universities to evaluate the proposed model and observe the behavior of the proposed variables that characterize the faculties.

Table 11. Results of dropout model faculty

Faculty ID	Dropout (using data from the last 20 years)	Dropout (using data from the last 10 years)
F1	0.83	0.74
F2	0.09	0.21
F3	0.83	0.74
F4	0.83	0.74
F5	0.17	0.04
F6	0.09	0.21
F7	0.09	0.21
F8	0.09	0.21

5. Conclusions

This study analyzed which variables affect the most dropout risk in a predictive model for higher education students. Unlike previous works that limited the analysis to just accuracy or results by a specific degree, our work analyzed the variable's behavior segmented by faculty, degree program, and semester to see whether or not there were any differences. The results indicated that the variables related to GPA, socioeconomic factor, and the pass rate of courses taken have a more significant impact on the model than the first semester of studies. This in-depth analysis identified that STEM programs present a different behavior than humanities programs; for example, socioeconomic status has a more significant influence on STEM models than humanities careers, while the pass rate is more important in humanities careers than in STEM. Additionally, in all models, we found a relationship between variables related to academic performance such as GPA and pass rate of courses taken with student dropout similar to [4].

Generalizing the analysis of variable importance for dropout may induce biases due to the data variability in the different careers. For this reason, we analyze from the most general at the faculty level to each semester in each career, using data from a single higher education institution (HEI) that could not be generalized to the level of all Ecuador in all HEIs due to the context of each one. To complement the analysis, we propose a model for predicting dropout based on teacher characteristics by the students of the respective degrees. Data from one HEI was used; future work will require other HEIs to measure the model's effectiveness and draw conclusions about the variables that influence dropout.

These findings' implications lead to in-depth analysis to avoid generalizations when working with predictive dropout models [25,36]. Even within the same institutions, there can be individual differences that can affect the models.

Regarding the limitations of this study, it should be noted that the recent modification of the study plan for all the degree programs had a significant impact. The students who start their studies with the new student plan will finish their degree program in the next four years and not in five. Therefore, the input data of the dropout predictive model will change significantly concerning the one proposed in the present study. Another limitation is the "extraordinary semester"; optional semesters of study during the vacation semester were not included in this study because they are not mandatory.

As future work, we want further to explore the field of academic analytics with this data to help decision-makers, such as program coordinators, make proper adjustments in the university's programs. For instance, we could perform a more in-depth analysis, exploring the results per cohort and different time frames. Moreover, it would be relevant to compare the results obtained in the present study with models of universities in several countries, analyzing the similarities and differences in the multiple contexts.

Acknowledgment. This work was supported in part by the FEDER/Ministerio de Ciencia, Innovación y Universidades–Agencia Estatal de Investigación, through the Smartlet Project (grant TIN2017-85179-C3-1-R) and the H2O Learn Project (grant PID2020-112584RB-C31); the Madrid Regional Government through the e-Madrid-CM Project (Grant S2018/TCS-4307) which is co-funded by the European Structural Funds (FSE and FEDER); and the European Commission through the LALA project (grant 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP).

References

1. Abu-Oda, G.S., El-Halees, A.M.: Data mining in higher education: university student dropout case study. International Journal of Data Mining & Knowledge Management Process5(1), 15 (2015)
2. Al-Noshan, A. A., Al-Hagery, M. A., Al-Hodathi, H. A., & Al-Quraishi, M. S. Performance evaluation and comparison of classification algorithms for students at Qassim University. Int. J. Sci. Res, 8(11), 1277-1282 (2018).
3. Albarracín, P., Daniel, J.: Identificación del perfil de egreso correspondiente a la licenciatura de la carrera de laboratorio clınico e histotecnologico de la Universidad central del ecuador periodo 2017-2022 (2016)
4. Ameen, A. O., Alarape, M. A., & Adewole, K. S. STUDENTS'ACADEMIC PERFORMANCE AND DROPOUT PREDICTION. Malaysian Journal Of Computing, 4(2), 278-303 (2019).
5. Ameri, S., Fard, M.J., Chinnam, R.B., Reddy, C.K.: Survival analysis based framework for early prediction of student dropouts. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 903–912 (2016)
6. Aulck, L., Velagapudi, N., Blumenstock, J., West, J.: Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364 (2016)
7. Barbosa Manhaes, L.M., da Cruz, S.M.S., Zimbrao, G.: Towards automatic prediction of student performance in stem undergraduate degree programs. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing. pp. 247–253(2015)
8. Barbu, M., Vilanova, R., Lopez Vicario, J., Pereira, M.J., Alves, P., Podpora, M.,´Angel Prada, M., Moran, A., Torreburno, A., Marin, S., et al.: Data mining tool for academic data

exploitation: literature review and first architecture proposal.Projecto SPEET-Student Profile for Enhancing Engineering Tutoring (2017)

9. Breiman, L.: Random forests. Machine learning45(1), 5–32 (2001)
10. Burgos, C., Campanario, M.L., de la Pena, D., Lara, J.A., Lizcano, D., Martinez, M.A.: Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers & Electrical Engineering 66, 541–556(2018)
11. Crawford, C.: Socioeconomic differences in university outcomes in the uk: dropout, degree completion and degree class. Tech. rep., IFS Working Papers (2014)
12. Chen, R.: Financial aid and student dropout in higher education: A heterogeneous research approach. In: Higher education, pp. 209–239. Springer (2008)
13. Chen, Y., Johri, A., Rangwala, H.: Running out of stem: a comparative study across tem majors of college students at-risk of dropping out early. In: Proceedings ofthe 8th international conference on learning analytics and knowledge. pp. 270–279(2018)
14. Chung, J.Y., Lee, S.: Dropout early warning systems for high school students using machine learning. Children and Youth Services Review96, 346–353 (2019)
15. Del Bonifro, F., Gabbrielli, M., Lisanti, G., Zingaro, S.P.: Student dropout prediction. In: International Conference on Artificial Intelligence in Education. pp.129–140. Springer (2020)
16. Fabara, E.: Cuadernos del contrato social por la educacion. Cuaderno8, 97–98(2013)
17. Fei, M., & Yeung, D. Y. (2015, November). Temporal models for predicting student dropout in massive open online courses. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 256-263). IEEE
18. Gasevi ́c, D., Dawson, S., Rogers, T., Gasevic, D.: Learning analytics should notpromote one size fits all: The effects of instructional conditions in predicting academic success. The Internet and Higher Education 28, 68–84 (2016)
19. Gitinabard, N., Khoshnevisan, F., Lynch, C. F., & Wang, E. Y. Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. arXiv preprint arXiv:1809.00052. (2018).
20. Heredia-Jimenez, V., Jimenez, A., Ortiz-Rojas, M., Marın, J.I., Moreno-Marcos,P.M., Munoz-Merino, P.J., Kloos, C.D.: An early warning dropout model in higher education degree programs: A case study in ecuador (2020)
21. Howard, E., Meehan, M., Parnell, A.: Contrasting prediction methods for earlywarning systems at undergraduate level. The Internet and Higher Education37,66–75 (2018)
22. Jimenez, F., Paoletti, A., Sanchez, G., Sciavicco, G.: Predicting the risk of academic dropout with temporal multiobjective optimization. IEEE Transactions on Learning Technologies12(2), 225–236 (2019)
23. Kang, K., Wang, S.: Analyze and predict student dropout from online programs. In:Proceedings of the 2nd International Conference on Compute and Data Analysis.pp. 6–12 (2018)
24. Luo, Y., Pardos, Z.: Diagnosing university student subject proficiency and predicting degree completion in vector space. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
25. Manrique, R., Nunes, B. P., Marino, O., Casanova, M. A., & Nurmikko-Fuller, T. An analysis of student representation, representative features and classification algorithms to predict degree dropout. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (pp. 401-410) (2019).
26. Marcılio, W.E., Eler, D.M.: From explanations to feature selection: assessing shap values as feature selection mechanism. In: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). pp. 340–347. IEEE (2020)
27. Marta Ferreyra, M., Avitabile, C., Botero ́Alvarez, J., Haimovich Paz, F., Urz ́ua,S.: At a crossroads: higher education in Latin America and the Caribbean. TheWorld Bank (2017)

28. Marquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Mousa Fardoun, H.,Ventura, S.: Early dropout prediction using data mining: a case study with high school students. Expert Systems33(1), 107–124 (2016)
29. Moreno-Marcos, P.M., Alario-Hoyos, C., Muñoz-Merino, P.J., Kloos, C.D.: Pre-diction in moocs: A review and future research directions. IEEE Transactions on Learning Technologies12(3), 384–401 (2018)
30. Moreno-Marcos, P.M., De Laet, T., Munoz-Merino, P.J., Van Soom, C., Broos, T.,Verbert, K., Delgado Kloos, C.: Generalizing predictive models of admission test success based on online interactions. Sustainability11(18), 4940 (2019)
31. Najdi, L., Er-Raha, B.: A novel predictive modeling system to analyze students a trisk of academic failure. International Journal of Computer Applications156(6),25–30 (2016)
32. Ortigosa, A., Carro, R.M., Bravo-Agapito, J., Lizcano, D., Alcolea, J.J., Blanco,O.: From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system. IEEE transactions on learning technologies12(2), 264–277 (2019)
33. Pelanek, R.: Metrics for evaluation of student models. Journal of Educational DataMining7(2), 1–19 (2015)
34. Pereira, F.D., Oliveira, E., Cristea, A., Fernandes, D., Silva, L., Aguiar, G., Alamri,A., Alshehri, M.: Early dropout prediction for programming courses supported by online judges. In: International Conference on Artificial Intelligence in Education.pp. 67–72. Springer (2019)
35. Pilotti, M. A., Abdelsalam, H. M., Anjum, F., Daqqa, I., Muhi, I., Latif, R. M., ... & Al-Ameen, T. A. Predicting Math Performance of Middle Eastern Students: The Role of Dispositions. Education Sciences, 12(5), 314. (2022).
36. Rovira, S., Puertas, E., & Igual, L. Data-driven system to predict academic grades and dropout. PLoS one, 12(2), e0171207 (2017).
37. Schneider, M.: Finishing the first lap: The cost of first year student attrition inamerica’s four year colleges and universities. American Institutes for Research (2010)
38. Schnepf, S.V.: Do tertiary dropout students really not succeed in european labour markets? (2014)
39. Suganya, S., Narayani, V.: Analysis of students dropout forecasting using data mining,”. In: 3rd Internaational Conference on Lastest Trends in Engineering, Science,Humanities and Management (2017)
40. Tang, C., Ouyang, Y., Rong, W., Zhang, J., & Xiong, Z. Time series model for predicting dropout in massive open online courses. In International Conference on Artificial Intelligence in Education (pp. 353-357). Springer, Cham . (2018).
41. Tinto, V.: Dropout from higher education: A theoretical synthesis of recent research. Review of educational research45(1), 89–125 (1975)
42. Vossensteyn, J.J., Kottmann, A., Jongbloed, B.W., Kaiser, F., Cremonini, L., Sten-saker, B., Hovdhaugen, E., Wollscheid, S.: Dropout and completion in higher education in europe: Main report (2015)

Alberto Jiménez-Macías is a PhD student at Universidad Carlos III de Madrid. He obtained a bachelor’s degree in Telematics Engineering and a master’s degree in Computer Science at the Escuela Superior Politécnica del Litoral (ESPOL) (Ecuador). He carried out development and research work at the Information Technology Center (CTI-ESPOL) for 8 years. His areas of interest are Learning Analytics, Educational Data Mining and Educational Technology.

Pedro Manuel Moreno Marcos is Visiting Professor in the Department of Telematics Engineering at Universidad Carlos III de Madrid (UC3M). He received his Bachelor in

Telecommunications Technologies Engineering in 2015 as well as his Master Degrees in Telecommunication Engineering and Telematic Engineering, which were both obtained in 2017. All of them (bachelor and masters) were obtained at Universidad Carlos III de Madrid. Moreover, he has obtained several awards, including the Extraordinary Awards in the Bachelor Degree and both Master Degrees, and other awards which recognize his academic achievements and Master Thesis. In 2017, he obtained a FPU fellowship to carry out his PhD, which was finished in July 2020. He also obtained the Outstanding Thesis Award. In January 2021, he obtained the positive evaluations from ANECA for the academic positions of Assistant Professor, Private University Professor, and Associate Professor (as non-civil servant). He worked as Specific Teaching Assistant in the academic year 2020/2021, as Assistant Professor in the academic year 2021/2022, and he has been Visiting Professor since September 2022. Currently, he has made 14 publications in JCR-indexed journals and multiple contributions in other journals and conferences. His areas of research interest include learning analytics, Educational Data Mining and MOOCs (Massive Open Online Courses).

Pedro J. Muñoz-Merino is Full Professor at the Department of Telematics Engineering at Universidad Carlos III de Madrid. His main topic of research is on learning analytics. In 2003, he received his Telecommunication Engineering degree from the Polytechnic University of Valencia, and in 2009 his PhD in Telematics Engineering from the Universidad Carlos III de Madrid. Pedro has published more than 150 papers, including more than 50 in journals indexed in the JCR. Pedro has participated in many research projects at the international and national level as well as with companies, being the Principal Investigator in several of them. Pedro has had different Chair positions in different international conferences related to educational technologies such as General Chair at EC-TEL 2023, Program Chair at EC-TEL 2022, Workshop Chair at LAK 2020, Publication Chair at EDM 2020, Program Chair at II LALA conference 2019, Poster chair at EDM 2017, Demo & poster Chair at EC-TEL 2014.

Margarita Ortiz-Rojas holds a PhD in Educational Sciences from Ghent University. Currently, she is the director of the Center of Educational Services at ESPOL in Ecuador. Her research interests include pedagogical innovations, technology in education, learning analytics, gamification and e-learning".

Carlos Delgado Kloos received the Ph.D. degree in Computer Science from the Technische Universität München and in Telecommunications Engineering from the Universidad Politécnica de Madrid. He is Full Professor of Telematics Engineering at the Universidad Carlos III de Madrid, where he is the Director of the GAST research group, Director of the UNESCO Chair on “Scalable Digital Education for All”, and Vice President for Strategy and Digital Education. He is also the Coordinator of the eMadrid research network on Educational Technology in the Region of Madrid. He is Senior Member of IEEE. He has been the Manager of ICT research projects at the Spanish Ministry and has carried out research stays at several universities such as Harvard, MIT, Munich, and Passau.

Received: November 10, 2021; Accepted: November 25, 2022.