Analysis of Student Performance Applying Data Mining Techniques in a Virtual Learning Environment

https://doi.org/10.3991/ijet.v18i11.37309

Leonardo Aguagallo[1], Fausto Salazar-Fierro[1,2], Janneth García-Santillán[3], Miguel Posso-Yépez[1], Pablo Landeta-López[1], Iván García-Santillán[1](✉)
[1]Universidad Técnica del Norte, Ibarra, Ecuador
[2]Universidad Nacional Mayor de San Marcos, Lima, Perú
[3]Unidad Educativa Juan Pablo II, Ibarra, Ecuador
idgarcia@utn.edu.ec

Abstract—Students' academic performance is a key factor for educational institutions and society, which is an important indicator of the quality of the teaching-learning process and the appropriation of knowledge. Its analysis allows an understanding of the behavior of students and teachers, generating valuable knowledge for making timely academic decisions. In this study, the following phases were carried out: (i) identification of factors associated with the academic performance of engineering university students, (ii) early prediction of academic success (student performance), and (iii) identification of use patterns in a virtual learning environment (VLE). The Knowledge Discovery in Databases (KDD) methodology was applied based on predictive and descriptive data mining techniques, using academic and socioeconomic data and interactions (resources and activities) with the VLE. The tools and programming languages used were Pentaho Data Integration for data integration and processing; Jupyter Notebook, Python, and Scikit-Learn for correlation analysis and prediction modeling; and R Studio for the clustering task. The results show that VLE resources such as files, links, and activities such as participation in forums are factors related to good academic performance. On the other hand, it was possible to make predictions of academic success (pass or fail) with an accuracy greater than 95% and to identify the main patterns of use of the VLE. The group with excellent academic performance (grades 9 to 10) is recognized for using file-type resources and high participation in class and forum activities.

Keywords—student performance, educational data mining, virtual learning environment, learning management systems, Knowledge Discovery in Databases, machine learning in education, learning analytics

1 Introduction

1.1 Problem statement

The low academic performance of students is an indicator of the educational reality of an educational institution [1], as well as a source of concern and interest for

parents, institutions, and governments [2]. Educational institutions have made an effort to offer quality education to achieve good academic performance in their students [3], reflected in the acquisition of learning results. These and other factors have produced interest in studies on academic performance since they are an instrument for creating indicators that guide decision-making [4] in the teaching-learning process. Low student performance and desertion affect higher education and academic quality standards. This phenomenon significantly increases the waste of public resources assigned to education [5].

The way of learning and teaching has changed with the growing technological advance and the implementation of information and communication technologies (ICT) in education [6]. Educational institutions rely more and more on virtual learning environments (VLE) [7], both in-person and online modality, especially in recent years (2020–2022) due to the Covid-19 pandemic. VLEs (such as Moodle, Google Classroom, Chamilo, among others), also known as learning management systems (LMS), complement learning, interaction, and collaboration in the face-to-face classroom, managing student activities and resources. In addition, VLEs store interactions such as system access, delivery times, use of resources and activities, task delivery, etc. The interactions are a product of student behavior under an institutional framework [8] and generate a large amount of data to be analyzed to improve the learning and teaching process.

Currently, the benefits and disadvantages of virtual and face-to-face education models are well-known. Some pros of face-to-face learning are learning from other students, real-time interaction, and improving social skills. Certain cons are no flexibility in place and time, use of traditional teaching, and dependent on the teacher. On the other hand, various pros of virtual learning are accessibility of time and place and cost affordability. Some cons are technology issues, a sense of isolation, and a long time in front of a screen [9]. There are several standards and best practices for virtual education [10], which consider, among others, the following topics: (a) Instructional design, indicating the learning objectives and appropriate online teaching methods; (b) Platforms and technologies, designed with accessibility and usability standards to facilitate interaction and communication between students and teachers; (c) Evaluation methods, adapted to the virtual format and that provide effective feedback to students; (d) Student support, including additional resources, tutorials, and advice on other services available online; (e) Teacher training, for the management of virtual teaching that allows creating meaningful learning experiences in students.

In this context, Educational Data Mining (EDM) arises, which deals with discovering knowledge from academic data sources [8]. The EDM comprises predictive and descriptive techniques used in data mining to understand and improve the use and exploitation of VLEs [11]. EDM has focused on the application of tasks of classification, regression, grouping, association, visualization, modeling, and monitoring of learning activities [12]. Some research approaches are the prediction of student dropout [5][13], prediction of academic success (student performance) [14][15], identification of behavior patterns [15], and factors associated with academic performance [16], etc.

The objectives of this study applying EDM are the following: (i) the identification of factors associated with school performance; (ii) the creation of a model for predicting academic success (pass or not); and (iii) obtaining use patterns of VLE in engineering

students. In this work, academic data (grades), socioeconomic and interactions generated in the institution's VLE (university careers) are used using KDD methodology [17] and the tools and programming languages Pentaho Data Integration (called Kettle) for data integration and data processing; Jupyter Notebook, Python, and Scikit-Learn for correlation analysis and prediction modeling; and R Studio for clustering. Prediction models are trained using attributes (variables) available from the beginning of the academic period to obtain an early prediction that helps make timely decisions and not at the end of the period where any preventive/corrective action may be late for the benefit of the students. This aspect represents an important contribution to this research. In addition, the analysis was carried out with data from the institution's VLE (called SIIU), obtaining usage patterns based on clustering algorithms as the well-known K-Means and others rarely mentioned in the literature, such as K-Prototype and K-Modes. It denotes another significant contribution. These algorithms are iterative and try to partition the dataset into *K* pre-defined clusters.

1.2 Related work

The following works were the starting point for the development of this study:

On the one hand, using mainly descriptive data mining techniques, in [16] they identified the factors related to student performance by applying the SNS, DSSD, NMEEF-SD, BSD, SD-Map, and APRIORI-SD algorithms, which are methods by subgroup discovery. Here it was determined that the economic situation, parents' educational level, participation in activities, forums, and visualization of resources are key factors in academic performance. In [15], they identified that the behavior patterns related to academic success are variables of the delivery time of tasks and activities, such as participation in forums using clustering algorithms such as K-Means and Expectation-Maximization (EM). Similarly, in [18], using K-Means, they searched for characteristics that affect school performance, showing that participation in classes, debates, and review of resources play an important role. In [11], they identified patterns of resource use by teachers. Resources like files and URLs and activities such as homework and quizzes are mostly used. Thus, pattern detection using mining data techniques is useful in supporting decision-making processes [19], including higher education, as in our case.

On the other hand, using predictive data mining techniques, in [6], they used both multi-level and standard regression (linear and logistic) to predict student performance. They obtained 69% accuracy in predicting pass-fail probabilities applying binary Logistic regression on in-between assessment grades and predictor variables from the VLE (Moodle) such as number of online sessions, total time online (min), number of resources and activities viewed, number of clicks, number of quizzes passed, average time per session (min), average assessment grade, etc. In [14], they analyzed the students' academic success using a data set with educational information and VLE activities. The Random Forest, Naive Bayes, and SMO algorithms showed great efficiency with an accuracy greater than 90.9%. In [20], they found variables that make it possible to predict school performance, pointing out that homework, access to the VAS, questionnaires, and age are the main predictor variables. In [21], they looked at the SVM, C4.5, and KNN algorithms for predicting academic success. The results showed

relevant factors: final grades, economic status, parental level of education, the distance between home and institution, student interest, and access to the VLE.

The rest of the document is as follows: Section 2 details the methodology used in this study, including predictive and descriptive techniques for analyzing academic performance. Section 3 shows the main results of the different tasks and algorithms used, as well as the comparison with other studies. Finally, the conclusions and future work are presented in Section 4.

2 Materials and methods

2.1 Data set

This study was carried out at the Faculty of Engineering in Applied Sciences. The data set was extracted from the University Institutional Integrated System (SIIU), a VLE developed in a Web environment with the Oracle 11g database. The selected data corresponds to the three years 2018–2021 containing six academic periods (Sep2018–Feb2019, Mar2019–Aug2019, Sep2019–Feb2020, Mar2020–Aug2020, Sep2020–Feb2021, Mar2021–Aug2021), belonging to 5 careers (Software, Industrial, Mechanics, Electronics, and Mechatronics), as shown in Table 1.

The data set is composed of the following variables:

- **Academic data:** career to which the student belongs, level (1st-10th), enrollment number (times enrolled in the current course: 1, 2 or 3), partial grade 1 and 2 (each semester has two in-between assessment grades out of 10 each), a final grade (out of 10), passed the course (yes or no) and percentage of absence (1–100).
- **Socioeconomic data:** age (years), gender (male, female), marital status (single, married, divorced, widowed, free union), disability (1–100), ethnic self-definition (white, mestizo, indigenous, afro-descendant), availability of internet (Yes or No) and monthly income (high, medium, low).
- **Interactions with the institutional EVA:** use of resources used: documents, links, files (yes or no); use of activities utilized: exams, forums, debates, participation, projects, tests, tasks, and papers (yes or no); the number of resources and activities used.

Table 1. Summary of the data set used in this study, which contains six academic periods and five careers each during 2018–2021

Academic Period	Career	Levels	N° Subjects	N° Activities	N° Resources	N° Records	N° Students
Sep2018–Feb2019	Mechanics	1 to 3	18	9697	4715	650	179
	Electronics	4 to 9	38	11993	8785	836	229
	Mechatronics	4 to 10	38	19494	15975	1171	234
	Industrial	4 to 10	33	19582	42433	1710	243
	Software	1 to 10	60	30835	33991	1912	373
Subtotal			**187**	**91601**	**105899**	**6279**	**1258**
Mar2019–Aug2019	Mechanics	1 to 4	24	11788	7186	718	216
	Electronics	5 to 10	34	8682	6300	585	201
	Mechatronics	4 to 10	36	15502	12370	846	189
	Industrial	5 to 9	34	10607	25194	881	207
	Software	1 to 10	58	27473	34234	1736	372
Subtotal			**186**	**74052**	**85284**	**4766**	**1185**
Sep2019–Feb2020	Mechanics	1 to 5	30	8485	10205	587	206
	Electronics	6 to 10	28	7455	10553	554	175
	Mechatronics	6 to 10	26	8696	7583	537	137
	Industrial	6 to 9	27	8141	11611	693	153
	Software	1 to 10	56	24410	39519	1521	361
Subtotal			**167**	**57187**	**79471**	**3892**	**1032**
Mar2020–Aug2020	Mechanics	1 to 6	20	27992	64088	3909	318
	Electronics	6 to 10	21	8213	18009	1398	124
	Mechatronics	6 to 10	23	23947	99586	3023	135
	Industrial	6 to 10	21	33071	480753	4635	143
	Software	1 to 10	47	47926	185775	6568	543
Subtotal			**132**	**141149**	**848211**	**19533**	**1263**
Sep2020–Feb2021	Mechanics	1 to 7	35	26989	76193	3351	202
	Electronics	8 to 10	15	6144	14814	984	113
	Mechatronics	8 to 10	13	12419	31014	1602	92
	Industrial	7 to 10	14	22138	90418	2716	95
	Software	1 to 10	55	36278	162676	5163	313
Subtotal			**132**	**103968**	**375115**	**13816**	**815**
Mar2021–Aug2021	Mechanics	1 to 8	43	25823	59500	3424	244
	Electronics	8 to 10	13	2562	2595	361	106
	Mechatronics	9 to 10	8	2691	4248	384	77
	Industrial	7 to 10	7	10212	31631	971	68
	Software	1 to 10	50	27161	89477	3688	302
Subtotal			**121**	**68449**	**187451**	**8828**	**797**
TOTAL			**925**	**536406**	**1681431**	**57114**	**6350**

Table 1 shows several records exist for each student, according to each academic level and subject. In total, there are 57114 raw records in the dataset.

2.2 Methodology

The widely used KDD methodology [17] was selected in this study for data analysis and consists of the following phases: (i) data collection and integration, (ii) processing, (iii) data mining, and (iv) validation and interpretation [22]. The result of each phase is the input for the next. Figure 1 shows the KDD methodology and its stages.

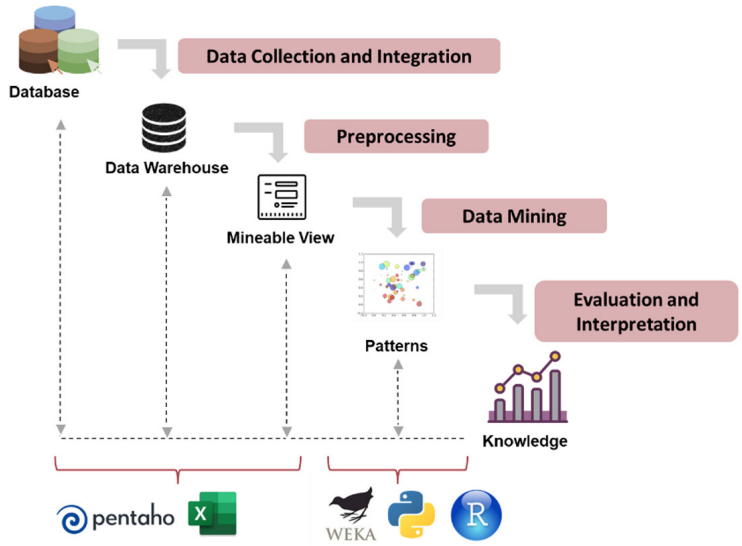

Fig. 1. KDD methodology [17] used in this study, together with the tools applied in each phase

2.3 Data collection and integration

Regarding the collection, the raw data was extracted from the Oracle database of the SIIU system in CSV format. For the data integration, the Kettle tool of Pentaho Data Integration Suite [23] was used, linearly obtaining the student's data, that is, a record that contains all the academic, socioeconomic, and interaction information with the VLE for each academic period. (semester). In this phase, a data warehouse ready to be processed was obtained.

2.4 Data processing

The processing phase was carried out in three parts: selection, elimination, and data transformation. The selection and elimination stages allowed a cleaner data set, free of inconsistencies and empty records. In the transformation stage, a discretization method [24] was applied to convert the continuous variables to ordinals. Socioeconomic data (age, income, disability) and academic data (percentages absent and grades) were

discretized in a personalized way using interval ranges established by the institution; for example, with a percentage of non-attendance greater than 30%, the student loses the course [5]. Data on the number of resource and activity interactions were discretized into categories or ranges using the equal width method [25]. The transformed data are shown in Table 2.

Table 2. Categories used to transform quantitative attributes into qualitative belonging to engineering students

Attribute	Category	Interval
Age (years)	Low	17–25
	Medium	26–30
	High	≥ 31
Grades (points)	Insufficient	< 7
	Sufficient	7–7.99
	Good	8–8.99
	Excellent	9–10
Non-attendance (%)	Very low	0–7
	Low	8–15
	Medium	16–23
	High	24–29
	Very high	30–100
Disability level (%)	None	< 30
	Slight	30–49
	Moderate	50–74
	Severe	75–84
	Very severe	85–100
Monthly income (USD)	Very low	< 400
	Low	400–713
	Medium	714–1427
	High	1428–2000
	Very high	> 2000
N° of activities and resources	Very Low	< 12
	Low	12–23
	Medium	24–34
	High	35–45
	Very high	> 45

The result of this phase is a general mineable view, with the 26 qualitative attributes (academic, socioeconomic, and interactions with the EVA) and a total of 57114 raw records mentioned above. The variables were selected based on previous studies indicated in the related works section and following the objectives of this research.

2.5 Data mining

The data mining process was developed in three stages: (i) correlation analysis, (ii) prediction model (classification), and (iii) cluster model (grouping). Jupyter Notebook with Python and the Scikit-Learn library were used for the correlation analysis and the prediction model, while R Studio was used for clustering. Each of the stages is detailed below.

Correlation analysis. This first stage determines the factors (variables) associated with student academic performance. A correlation analysis was conducted between each independent variable X (academic-socioeconomic data and interactions with the EVA), and the dependent variable Y (final grade). The correlation method was selected based on the framework proposed in [26], considering the variables' data type to be correlated. For this reason, the Spearman correlation was used, an adequate method for categorical (nominal) variables. Goodman-Kruskal gamma and Somers' coefficients are other methods to measure the strength of the association between ordinal variables [27] [28], hence, these were not used in this study. First, the significance level was calculated, which is used to determine if there is a statistically significant relationship between the two variables. If so, we calculate the correlation coefficient, which indicates the strength [–1; 1] and direction (positive or negative) of the relationship between the two variables studied. Two variables are associated when one variable provides information about the other; that is, the increase or decrease of one variable gives some indications of the behavior of the other variable.

Prediction model (classification). The second stage consisted of building a binary classification model to predict a student's academic success (pass or fail) and was divided into two steps:

a) The model's most relevant predictor variables were selected using the XGBoost model [29]. This predictive algorithm is part of the family of trees, which orders the importance of the characteristics of the prediction model, placing attributes: academic, socioeconomic, and interaction with the VLE as independent variables and the attribute Passed the course (yes or no) as the dependent variable.
b) Due to the good results reported in previous studies, two prediction models were built using the well-known Random Forest [14] and Support Vector Machine [21] classifiers. Random Forest (RF) comprises a set of decision trees, where each tree predicts a category, and then the most voted category is chosen. Support Vector Machine (SVM) is a classifier based on the search for an optimal hyperplane to determine to which class the input data belongs. The independent variables used to train each model are the grade (partial 1) for the first half of the semester (an early stage) and the most relevant variables selected in the previous step. The independent variables are qualitative (ordinal and nominal), using the previously mentioned discretization conversion. Additionally, the GridSearch technique was applied to identify the best parameters and hyperparameters of each model. For RF, 100 trees (n_estimators) and a random_state=1 were instantiated, and for SVM the model was configured with a Gaussian kernel (rbf), C=1 and Gamma=1 [30].

Cluster analysis (clustering). In this third stage, a cluster analysis was carried out to discover patterns of use of the EVA. The variables significantly correlated with the final grade (Correlation analysis section) were taken. Then, groups that reflected characteristics of VLE use related to academic performance were obtained. The K-Means, K-Prototype [31], and K-Modes [32] algorithms were used for grouping. K-means is appropriate for numerical data, K-Modes is suitable for categorical data only, and K-Prototype properly handles mixed data types (numerical and categorical variables). The three algorithms were configured to obtain four groups (value k=4), considering the categories of the final grade (insufficient, sufficient, good, and excellent) according to the university's regulations.

2.6 Validation and interpretation

The correlation analysis of variables and the academic success prediction model were evaluated with quantitative metrics. For the correlation analysis, the significance level of the statistical test (p-value) determines the existence of a statistically significant correlation and is validated when the p-value ≤ 0.05 [33]. The correlation coefficient measures the degree to which two variables are associated [34]. This value ranges from −1 to 1, these limits being a perfect correlation, while the value of zero indicates a null correlation. The prediction models were validated through a confusion matrix and metrics such as accuracy, recall, precision, F1-score, ROC curve, the area under the curve (AUC), and the Kappa coefficient. The definitions and formulas of each metric can be seen in detail in [35] [36]. The most relevant for this study are defined in the next section. The validation for clustering tasks consists of choosing an optimal value for the number of clusters (k). As indicated before, the value of k=4 was selected, considering the categories of the final grade (insufficient, sufficient, good, and excellent).

3 Results and discussion

3.1 Correlation analysis

The level of significance of the statistical correlation test (p-value) showed a significant linear relationship ($p \leq 0.05$) for most of the variables, except for the career to which the student belongs and the availability of the Internet. These two variables are not associated with the student's academic performance. Therefore, they were discarded from the correlation analysis, leaving the dataset with 24 attributes for this task. The statistically significant correlation coefficients (Spearman) are shown in Table 3.

Table 3. Correlation coefficients for the 24 categorical variables using Spearman's correlation

	Level	N° Enrollment	Grade 1	Grade 2	Passed	% of Absence	N° Activities	N° Resources	Documents	Link	File	Exams	Forums	Projects	Tests	Tasks	Papers	Age	Gender	Marital Status	Disability	Ethnic	Income	Final Grade
Level	1																							
N° Enrollment	0.06	1																						
Partial grade 1	0.11	−0.11	1																					
Partial grade 2	0.14	−0.09	0.52	1																				
Passed	0.03	−0.09	0.36	0.42	1																			
% of absence	0.00	0.08	−0.12	−0.15	−0.16	1																		
N° Activities	−0.03	0.02	−0.12	−0.12	−0.06	0.13	1																	
N° Resources	0.06	−0.02	0.05	0.08	0.02	−0.02	0.00	1																
Documents	−0.11	0.00	−0.06	−0.02	−0.03	0.01	0.00	0.03	1															
Link	0.02	−0.02	0.03	0.03	0.03	−0.09	−0.04	−0.04	0.09	1														
File	0.10	0.01	0.03	0.01	0.01	0.05	−0.01	0.08	−0.20	−0.10	1													
Exams	−0.02	0.00	−0.04	−0.05	0.00	−0.02	0.02	0.01	−0.01	0.07	0.03	1												
Forums	0.05	−0.06	0.20	0.21	0.10	−0.36	−0.31	0.04	0.05	0.25	−0.15	−0.02	1											
Projects	0.03	0.06	−0.14	−0.15	−0.07	0.25	0.26	−0.02	−0.02	−0.18	0.14	0.00	−0.70	1										
Tests	0.06	−0.06	0.20	0.21	0.09	−0.35	−0.31	0.04	0.03	0.25	−0.15	0.02	0.94	−0.69	1									
Tasks	0.00	0.05	−0.14	−0.13	−0.08	0.29	0.27	−0.03	−0.06	−0.16	0.08	−0.01	−0.67	0.54	−0.66	1								
Papers	0.06	−0.06	0.20	0.21	0.09	−0.35	−0.32	0.04	0.03	0.24	−0.16	0.00	0.95	−0.70	0.94	−0.67	1							
Age	0.31	0.09	−0.06	−0.06	−0.06	0.06	0.01	0.03	−0.07	−0.06	0.07	−0.01	−0.07	0.08	−0.06	0.07	−0.06	1						
Gender	−0.19	0.01	−0.05	−0.07	−0.03	−0.02	−0.01	−0.02	0.00	0.00	−0.02	−0.01	0.05	−0.06	0.05	−0.05	0.05	−0.01	1					
Marital status	−0.03	−0.02	0.02	0.02	−0.01	−0.04	0.00	0.01	0.00	−0.01	−0.01	0.00	0.01	−0.02	0.01	−0.02	0.01	−0.12	0.07	1				
Disability	0.03	0.01	−0.03	−0.03	−0.04	0.00	0.01	0.00	0.00	−0.01	−0.01	0.01	−0.02	0.02	−0.02	0.02	−0.02	−0.03	−0.03	0.00	1			
Ethnic	−0.07	0.00	−0.04	−0.03	−0.01	−0.02	0.00	−0.02	0.01	−0.01	0.00	−0.01	0.02	−0.02	0.02	−0.02	0.02	0.07	0.03	−0.04	−0.02	1		
Income	−0.02	−0.01	0.02	0.02	0.00	0.00	−0.01	0.00	0.02	0.00	0.00	0.00	0.02	−0.02	0.02	−0.01	0.02	−0.08	0.00	−0.02	0.01	−0.10	1	
Final grade	0.13	−0.12	**0.79**	**0.81**	**0.50**	−0.16	−0.13	0.08	−0.06	0.04	0.01	−0.05	**0.23**	−0.16	**0.23**	−0.15	**0.23**	−0.06	−0.07	0.02	−0.04	−0.04	0.02	1

Note: Values in bold indicate the most significant correlation with the variable Final grade (student performance).

The interpretation of the correlation was carried out according to the scale proposed in Table 4.

Table 4. Meaning of the correlation coefficient ranging [−1, 1] [26]

R Coefficient	Interpretation of Linear Relationship
0.8	Strong positive
0.5	Moderate positive
0.2	Weak positive
0.0	No relationship
−0.2	Weak negative
−0.5	Moderate negative
−0.8	Strong negative

According to the correlation analysis (values in bold in Table 3), the attributes most associated with academic performance (final grade) are partial grades 1 and 2, showing a strong positive correlation. Furthermore, other related attributes are forums, debates, and participation in classes, tests, and papers, indicating a weak positive correlation. On the other hand, a low negative correlation was obtained with the student's absences. It may be because many teachers do not usually register their students' attendance in the SIIU system.

3.2 Academic success prediction models

Table 5 shows the metrics of the predictions of academic success (pass=1 or fail=0) obtained with the two models tested (RF and SVM), where good results are evident in both models. The model accuracy is the ratio between correctly classified predictions and the total number of estimates. The Kappa coefficient reveals whether the outcomes found in a confusion matrix are significantly better than those produced in a random classification. This value is in the range [0, 1], with 1 being the best score.

Table 5. Validation metrics of academic success prediction models

Classifier	Accuracy	Precision	Recall	F1-Score	Kappa	AUC
RF	96.7%	97.51%	98.92%	0.98	0.77	0.86
SVM	95.1%	95.40%	99.45%	0.97	0.59	0.73

It is noted that the RF classifier obtained better metrics than SVM in almost all of them, except recall. Both classifiers perform appropriately, fulfilling the requirements of usually accepted values higher than 85% accuracy and Kappa coefficient value greater than 0.70 [37], except with SVM. Figures 2 and 3 show both models' confusion matrix and ROC curve on the test set. The overall dataset (57114 records) was randomly split into 70% for training (39979 records) and 30% for testing (17135 records).

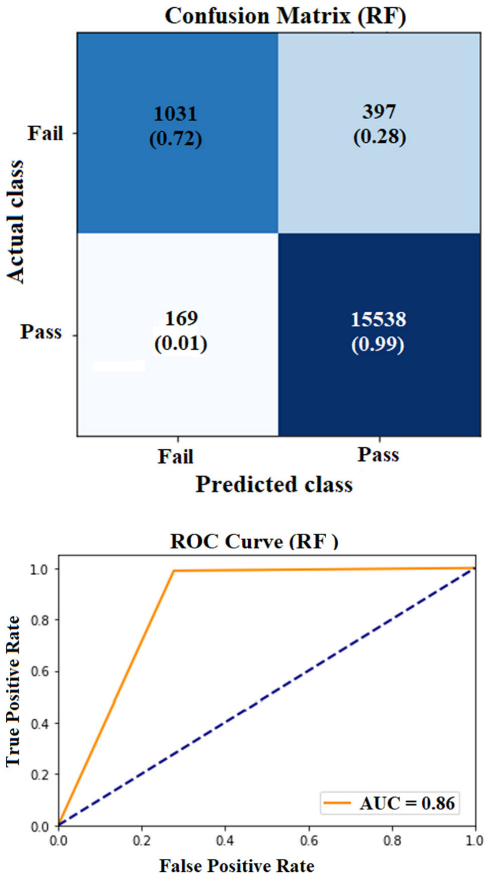

Fig. 2. Confusion matrix and ROC curve from the Random Forest (RF) model for predicting academic success on the test set

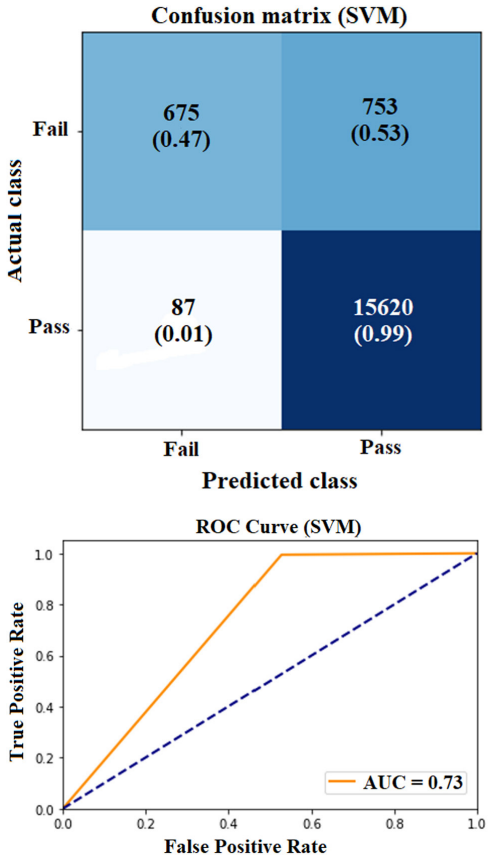

Fig. 3. Confusion matrix and ROC curve from the SVM model for predicting academic success on the test set

The confusion matrix allows us to see, through a contingency table (cross-tabulation), the distribution of type I (false positive, FP) and II (false negative, FN) errors committed by the classifier throughout all the categories of the problem (Pass and Fail). A false positive is an outcome where the model incorrectly predicts the positive class (Pass). In contrast, a false negative is an outcome where the model incorrectly predicts the negative class (Fail).

In this study, we prefer to decrease the false-positive rate more than the false-negative rate. The FP is the worst because a potential student to fail would be left without timely tutoring or academic follow-up. Thus, the RF algorithm (with FP=397) also works better than SVM (FP=753).

Besides, it is possible to notice an imbalanced dataset problem in the confusion matrices (Figures 2 and 3), i.e., a disparity in the frequencies of the observed classes (Pass=91.7% vs. Fail=8.3%). In this case, a classifier will skew the predicted

probabilities, tending to predict the abundant class more often. One way to solve the class imbalance problem is using better accuracy metrics like the F1 score, a weighted average of the precision and recall values [5]. This metric ranges from [0 to 1], with 1 being the best score. Both classifiers (Table 5) keep a good performance, with RF (0.98) being slightly better than SVM (0.97).

A ROC (Receiver Operating Characteristic) curve is a graph showing the performance of a classification model at different threshold levels (Figures 2 and 3). The AUC (Area Under the Curve) measures the ability of a binary classifier to distinguish between the positive and negative classes. It is used as a summary of the ROC curve. The higher the AUC, the better the model's performance. An excellent model has an AUC close to 1. The RF and SVM algorithms perform well, although RF has an AUC=0.86 greater than SVM=0.73 (Table 5).

On the other hand, the XGBoost algorithm used in selecting the most relevant predictor variables of the models (RF and SVM) showed that the attributes of monthly income and exams are variables with little relevance for predicting academic success. Thus, they were not considered in both predictive models.

3.3 Cluster analysis (clustering)

The K-means, K-prototype, and K-modes algorithms were applied to form clusters with a k=4 considering the excellent, good, sufficient, and insufficient student performance groups. The clusters and their centroids are shown in Table 6, considering the 24 attributes explained.

The relevant patterns of use of the VLE of each cluster (0–3) are interpreted, based on the centroids, as follows:

- **Excellent group**: refer to students with grades (final, partial 1 and 2) between 9.00 and 10.00. They are characterized by using file-type resources and highly participating in forums and classes. However, they register a low use of the rest of the resources and activities of the EVA; probably because these students easily understand the contents in class, it is unnecessary to use additional resources/activities.
- **Good group**: Your grades are between 8.00 and 8.99. They register the use of resources such as files and links but low participation in forums, debates, and class participation.
- **Sufficient group**: they register grades between 7.00 and 7.99. This group minimally passes the course and is characterized by not using the links. In addition, they present a low use of files, participation in forums and discussions, and a low use of tasks and projects.
- **Insufficient group**: their grades are between 0.00 and 6.99, so they fail the semester. In general, they are characterized because they register a low use of VLE resources and activities; specifically, they do not use the resources like links and files. Also, they do not deliver homework or projects.

Table 6. Clusters and their centroids formed with the k-means, k-modes, and k-prototype clustering algorithms

Attributes	K-Means				K-Modes				K-Prototype			
	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Level	6	8	9	8	9	6	8	5	8	1	6	4
N° Enrollment	1	1	1	1	1	1	1	1	1	1	1	1
Partial grade 1	Good	Sufficient	Good	Excellent	Excellent	Sufficient	Good	Sufficient	Good	Good	Sufficient	Good
Partial grade 2	Sufficient	Insufficient	Good	Excellent	Excellent	Insufficient	Good	Sufficient	Excellent	Good	Sufficient	Good
Final grade	Sufficient	Insufficient	Good	Excellent	Excellent	Sufficient	Good	Sufficient	Good	Sufficient	Sufficient	Sufficient
Passed the course	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
% of absence	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low
N° Activities	Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low
N° Resources	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low
Documents	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Link	No	No	Yes	No	No	No	No	Yes	No	No	No	No
File	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes	No	Yes	No
Exams	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Forums	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes
Projects	Yes	Yes	No	No	No	Yes	No	No	No	No	Yes	No
Tests	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes
Tasks	No	Yes	No	No	No	Yes	No	No	No	No	Yes	No

(Continued)

Table 6. Clusters and their centroids formed with the k-means, k-modes, and k-prototype clustering algorithms *(Continued)*

Attributes	K-Means				K-Modes				K-Prototype			
	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Papers	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes
Age	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
Gender	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male
Marital status	Single	Single	Single	Single	Single	Single	Single	Single	Single	Single	Single	Single
Disability	None	None	None	None	None	None	None	None	None	None	None	None
Ethnic	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo	Mestizo
Monthly income	Low	Low	Low	Low	Low	Low	Low	Medium	Low	Low	Low	Low
Number of records	7993	6870	27640	14611	15399	13385	15914	12416	22109	15191	12026	7788
% of records	14.00	12.00	48.00	26.00	26.96	23.44	27.86	21.74	38.71	26.60	21.06	13.64

4 Discussion

This study shows that the factors most associated with academic performance are partial grades (1 and 2), followed by forums, debates, class participation, tests, and papers (Table 3). The first components (grades) are required to approve the subjects, while the others strengthen retention of what is learned in class. On the other hand, the career to which a student belongs, and the availability of the Internet are not factors related to academic success.

The results of this study agree with some previous works, where it is pointed out that intermediate grades [21], forums and participation in classes [15][16][18], links, and tasks [11] are key factors in academic performance. Regarding the discrepancies, it was obtained that the monthly income had a low relationship with academic performance, unlike the work done in [16].

Regarding the prediction models (RF and SVM), these yield results with an accuracy greater than 95%, specifically, Random Forest with 96.7% and SVM with 95.1% using data from EVA interactions, socioeconomics, and partial grades.

Our results overcome those obtained in [6] with 69% accuracy using binary logistic regression where predictor variables such as number of clicks, number of sessions, online time, number of resources viewed, pages, links, and intermediate grades from the Moodle platform are used.

Besides, our scores are better than [14] with 90.9% accuracy using Naive Bayes but less than 100% using Random Forest. The data used is extracted from Moodle, such as cumulative grade point average (CGPA), risk of failure rate, assignments, plagiarism count, final exam, number of accesses on and off campus, and connection time.

The feasibility of making predictions of academic success or not at an early stage, that is, using the data from the first half of the period and thus being able to take corrective actions for the benefit of the student, is an important contribution to the present study.

Regarding the patterns of use of the EVA, based on the cluster analysis, it shows that the use of resources such as files and links and participation activities in forums and classes are characteristics related to high academic performance (excellent and good). The low use of the rest of the resources and activities is evident in all students, even those with good academic performance. This aspect may be because they easily understand the content in class, and there is no need to review additional VLE resources. Besides, they can take advantage of the activities with the highest score, ignoring the less weighted ones. Additionally, there is evidence of a low percentage of absences in each cluster (insufficient, sufficient, good, and excellent), reinforcing the idea that teachers are not regularly registering attendance.

Some limitations of this study are related to the delivery time of homework [15] and access time (total and per session) to the VLE [6], where they were not considered. Neither were counted the variables concerning the academic information of the parents, information on the educational environment (duration of the module, modality of study, and classification of the institution), and psychological conditions of the students [21], whose data are related to academic performance in the cited studies. These data are not currently recorded in the VLE of our institution (SIIU). Other machine learning

algorithms, such as artificial neural networks, have not been tested, which have reported good results in prediction tasks, even in different application domains [38]. Besides, the best analytical model generated in this study is not deployed in a cloud education Big Data platform [39].

5 Conclusions and future work

In the present study, the factors most associated with academic performance are identified: partial grades (1 and 2), followed by participation in classes, forums, debates, tests, and papers (Table 3). In addition, the variable passed the course (pass or fail) is predicted early with an accuracy greater than 95% (Table 5), using only the data from the first half of the semester, which allows preventive/corrective actions to be taken on time. VLE usage patterns show that the use of resources such as files and links; activities such as participation in forums and classes, and delivery of tasks and projects are related to high academic performance (Section 3.3). In future work, it is suggested to consider, in the predictive models, predictor variables of data about the times of access to the EVA, academic information of the parents and the educational environment, and some psychological data of the student. Finally, it is suggested to carry out either the deployment in the cloud of the analytical models generated or integrated into the institutional SIIU system.

6 References

[1] M. De Miguel Díaz, P. Apocada Urquijo, J. M. Arias Blanco, T. Escudero Escorza, S. Rodriguez Espinar, and J. Vidal García, "Performance assessment in higher education. Comparison of results between LOGSE and COU students," *Rev. Investig. Educ.*, vol. 20, no. 2, pp. 357–383, 2002, [Online]. Available: https://revistas.um.es/rie/article/view/98971

[2] M. Guacales-Gualavisi, F. Salazar-Fierro, J. García-Santillán, S. Arciniega-Hidrobo, and I. García-Santillán, "Computer System Based on Robotic Process Automation for Detecting Low Student Performance," in *Information Technology and Systems*, 2021, pp. 141–150. https://doi.org/10.1007/978-3-030-68285-9_15

[3] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," *Mater. Today Proc.*, vol. 47, pp. 5260–5267, 2021. https://doi.org/10.1016/j.matpr.2021.05.646

[4] G. M. Garbanzo Vargas, "Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública," *Rev. Educ.*, vol. 31, no. 1, pp. 43–63, 2007, [Online]. Available: https://www.redalyc.org/articulo.oa?id=44031103. https://doi.org/10.15517/revedu.v31i1.1252

[5] D. Vila, S. Cisneros, P. Granda, C. Ortega, M. Posso-Yépez, and I. García-Santillán, "Detection of desertion patterns in university students using data mining techniques: A case study," in *Communications in Computer and Information Science*, 2019, vol. 895, pp. 420–429. https://doi.org/10.1007/978-3-030-05532-5_31

[6] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, 2017. https://doi.org/10.1109/TLT.2016.2616312

[7] A. Juma, J. Rodríguez, J. Caraguay, M. Naranjo, A. Quiña-Mera, and I. García-Santillán, "Integration and evaluation of social networks in virtual learning environments: A case study," in *Technology Trends*, 2019, pp. 245–258. https://doi.org/10.1007/978-3-030-05532-5_18

[8] D. J. Lemay, C. Baek, and T. Doleck, "Comparison of learning analytics and educational data mining: A topic modeling approach," *Comput. Educ. Artif. Intell.*, vol. 2, p. 100016, 2021. https://doi.org/10.1016/j.caeai.2021.100016

[9] V. Gherheș, C. E. Stoian, M. A. Fărcașiu, and M. Stanici, "E-learning vs. face-to-face learning: Analyzing students' preferences and behaviors," *Sustainability*, vol. 13, no. 8. 2021. https://doi.org/10.3390/su13084381

[10] FAO, *E-learning methodologies and good practices: A guide for designing and delivering e-learning solutions from the FAO elearning Academy*, 2nd ed. Rome, 2021.

[11] J. Calderon-Valenzuela, K. Payihuanca-Mamani, and N. Bedregal-Alpaca, "Educational data mining to identify the patterns of use made by the university professors of the Moodle platform," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 1, pp. 321–328, 2022. https://doi.org/10.14569/IJACSA.2022.0130140

[12] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telemat. Informatics*, vol. 37, pp. 13–49, 2019. https://doi.org/10.1016/j.tele.2019.01.007

[13] S.-U. Hassan, H. Waheed, N. Aljohani, M. Ali, S. Ventura, and F. Herrera, "Virtual learning environment to predict withdrawal by leveraging deep learning," *Int. J. Intell. Syst.*, vol. 34, no. 8, pp. 1935–1952, 2019. https://doi.org/10.1002/int.22129

[14] R. Hasan, S. Palaniappan, A. Raziff, S. Mahmood, and K. U. Sarker, "Student Academic Performance Prediction by using Decision Tree Algorithm," in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, 2018, pp. 1–5. https://doi.org/10.1109/ICCOINS.2018.8510600

[15] R. Cerezo, M. Sánchez-Santillán, M. P. Paule-Ruiz, and J. C. Núñez, "Students' LMS interaction patterns and their relationship with achievement: A case study in higher education," *Comput. Educ.*, vol. 96, pp. 42–54, 2016. https://doi.org/10.1016/j.compedu.2016.02.006

[16] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, and D. J. Murray, "Identifying key factors of student academic performance by subgroup discovery," *Int. J. Data Sci. Anal.*, vol. 7, no. 3, pp. 227–245, 2019. https://doi.org/10.1007/s41060-018-0141-y

[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996. https://doi.org/10.1145/240455.240464

[18] S. Bharara, S. Sabitha, and A. Bansal, "Application of learning analytics using clustering data mining for students' disposition analysis," *Educ. Inf. Technol.*, vol. 23, no. 2, pp. 957–984, 2018. https://doi.org/10.1007/s10639-017-9645-7

[19] G. Chacón-Encalada, L. Jaramillo-Mediavilla, W. Rivera-Montesdeoca, L. Suárez-Zambrano, and I. García-Santillán, "Detection of Space and Time Patterns in the ECU 911 Integrated Security System Using Data Mining Techniques," in *Information and Communication Technologies*, 2021, pp. 61–74. https://doi.org/10.1007/978-3-030-89941-7_5

[20] J. Bravo-Agapito, S. Romero, and S. Pamplona, "Early prediction of undergraduate student's academic performance in completely online learning: A five-year study," *Comput. Human Behav.*, vol. 115, p. 106595, 2021. https://doi.org/10.1016/j.chb.2020.106595

[21] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Predicting student success using big data and machine learning algorithms," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 12, pp. 236–251, 2022. https://doi.org/10.3991/ijet.v17i12.30259

[22] A. M. Wilson, L. Thabane, and A. Holbrook, "Application of data mining techniques in pharmacovigilance," *Br. J. Clin. Pharmacol.*, vol. 57, no. 2, pp. 127–134, 2004. https://doi.org/10.1046/j.1365-2125.2003.01968.x

[23] Hitachi Vantara Corporation, "Download Pentaho," *Hitachi Vantara Corporation*, 2022. https://www.hitachivantara.com/es-latam/products/data-management-analytics/pentaho/download-pentaho.html

[24] M. HacibeyoĞLu and M. Ibrahim, "Comparison of the effect of unsupervised and supervised discretization methods on classification process," *Int. J. Intell. Syst. Appl. Eng.*, pp. 105–108, 2016. https://doi.org/10.18201/ijisae.267490

[25] D. Lind, W. Marchal, and S. Wathen, *Statistics applied to business and economics*, Sixteenth Edition. McGraw-Hill, 2015.

[26] H. Khamis, "Measures of association: How to choose?," *J. Diagnostic Med. Sonogr.*, vol. 24, no. 3, pp. 155–162, 2008. https://doi.org/10.1177/8756479308317006

[27] K. Lavidas et al., "Factors affecting response rates of the web survey with teachers," *Computers*, vol. 11, no. 9, 2022. https://doi.org/10.3390/computers11090127

[28] K. Lavidas, S. Papadakis, D. Manesis, A. S. Grigoriadou, and V. Gialamas, "The effects of social desirability on students' self-reports in two social contexts: Lectures vs. Lectures and lab classes," *Information*, vol. 13, no. 10, 2022. https://doi.org/10.3390/info13100491

[29] R. Mitchell and E. Frank, "Accelerating the XGBoost algorithm using GPU computing," *PeerJ Comput. Sci.*, vol. 3, p. e127, 2017. https://doi.org/10.7717/peerj-cs.127

[30] A. Müller and S. Guido, *Introduction to Machine Learning with Python—A Guide for Data Scientists*, First Edition. United States of America: O'Reilly Media, Inc., 2016.

[31] G. Szepannek, "ClustMixType: User-friendly clustering of mixed-type data in R," *R J.*, vol. 10, p. 200, 2018. https://doi.org/10.32614/RJ-2018-048

[32] M. K. Ng, M. J. Li, J. Z. Huang, and Z. He, "On the impact of dissimilarity measure in k-modes clustering algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 503–507, 2007. https://doi.org/10.1109/TPAMI.2007.53

[33] J. H. McDonald, *Handbook of Biological Statistics*, Third. Baltimore, Maryland: Sparky House, 2014.

[34] N. S. Chok, "Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data," 2010.

[35] S. Manimurugan, S. Al-Mutairi, M. M. Aborokbah, N. Chilamkurti, S. Ganesan, and R. Patan, "Effective attack detection in Internet of medical things smart environment using a deep belief neural network," *IEEE Access*, vol. 8, pp. 77396–77404, 2020. https://doi.org/10.1109/ACCESS.2020.2986013

[36] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Real-time twitter sentiment analysis for Moroccan Universities using machine learning and big data technologies," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 05 SE-Papers, pp. 42–61, 2023. https://doi.org/10.3991/ijet.v18i05.35959

[37] I. D. García-Santillán and G. Pajares, "On-line crop/weed discrimination through the Mahalanobis distance from images in maize fields," *Biosyst. Eng.*, vol. 166, pp. 28–43, 2018. https://doi.org/10.1016/j.biosystemseng.2017.11.003

[38] I. D. Herrera-Granda et al., "Artificial Neural Networks for Bottled Water Demand Forecasting: A Small Business Case Study," in *Advances in Computational Intelligence*, 2019, pp. 362–373. https://doi.org/10.1007/978-3-030-20518-8_31

[39] J. Wang, "Comprehensive test and evaluation path of college teachers' professional development based on a cloud education big data platform," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 05 SE-Papers, pp. 79–94, 2023. https://doi.org/10.3991/ijet.v18i05.38497

7 Authors

Leonardo Aguagallo is a computer engineer from the Universidad Técnica del Norte in Ibarra-Ecuador (email: lmaguagalloa@utn.edu.ec).

Fausto Salazar-Fierro is a Ph.D. (c) from the Universidad Nacional Mayor de San Marcos in Lima-Perú and an assistant professor at the Universidad Técnica del Norte in Ibarra-Ecuador (email: fasalazar@utn.edu.ec).

Janneth García-Santillán is an MSc. from the Universidad Europea de Madrid in Spain and a teacher at the Unidad Educativa Juan Pablo II in Ibarra-Ecuador (email: janneth.garcia@educacion.gob.ec).

Miguel Posso-Yépez is a Ph.D. Universidad de las Palmas de Gran Canaria in Spain and an associate professor at the Universidad Técnica del Norte in Ibarra-Ecuador (email: maposso@utn.edu.ec).

Pablo Landeta-López is an MSc. from the Universidad de las Fuerzas Armadas in Sangolquí-Ecuador and an assistant professor at the Universidad Técnica del Norte in Ibarra-Ecuador (email: palandeta@utn.edu.ec).

Iván García-Santillán is a Ph.D. from the Universidad Complutense de Madrid in Spain and an associate professor at the Universidad Técnica del Norte in Ibarra-Ecuador (email: idgarcia@utn.edu.ec).

Article submitted 2022-12-09. Resubmitted 2023-03-07. Final acceptance 2023-03-24. Final version published as submitted by the authors.