Research paper

Prediction of University-Level Academic Performance through Machine Learning Mechanisms and Supervised Methods

Predicción del rendimiento académico universitario mediante mecanismos de aprendizaje automático y métodos supervisados

Leonardo Emiro Contreas-Bravo[1*], Nayive Nieves-Pimiento[2], Karolina González Guerrero[3*]

[1]Universidad Distrital Francisco José de Caldas (Bogotá, Colombia), [2]Universidad ECCI (Bogotá, Colombia), [3]Universidad Militar Nueva Granada (Bogotá, Colombia).

* Correspondence E-mail: lecontrerasb@udistrital.edu.co

Abstract

Context: In the education sector, variables have been identified which considerably affect students' academic performance. In the last decade, research has been carried out from various fields such as psychology, statistics, and data analytics in order to predict academic performance.

Method: Data analytics, especially through Machine Learning tools, allows predicting academic performance using supervised learning algorithms based on academic, demographic, and sociodemographic variables. In this work, the most influential variables in the course of students' academic life are selected through wrapping, embedded, filter, and assembler methods, as well as the most important characteristics semester by semester using Machine Learning algorithms (Decision Trees, KNN, SVC, Naive Bayes, LDA), which were implemented using the Python language.

Results: The results of the study show that the KNN is the model that best predicts academic performance for each of the semesters, followed by Decision Trees, with precision values that oscillate around 80 and 78,5% in some semesters.

Conclusions: Regarding the variables, it cannot be said that a student's per-semester academic average necessarily influences the prediction of academic performance for the next semester. The analysis of these results indicates that the prediction of academic performance using Machine Learning tools is a promising approach that can help improve students' academic life allow institutions and teachers to take actions that contribute to the teaching-learning process.

Keywords: educational data analysis, Machine Learning, higher education

Language: English

Resumen

Contexto: En el sector educativo se han identificado variables que inciden considerablemente en el rendimiento académico de los estudiantes. En la última década se han llevado a cabo investigaciones desde diversos campos como la psicología, la estadística y el análisis de datos con el fin de predecir el rendimiento académico.

Método: La analítica de datos, especialmente a través de herramientas de Machine Learning, permite predecir el rendimiento académico utilizando algoritmos de aprendizaje supervisado basados en variables académicas, demográficas y sociodemográficas. En este trabajo se seleccionan las variables más influyentes en el transcurso de la vida académica de los estudiantes mediante métodos de filtro, embebidos, y de ensamble, así como las características más importantes semestre a semestre utilizando algoritmos de Machine Learning (árbol de decisión, KNN, SVC, Naive Bayes, LDA), implementados en el lenguaje Python.

Resultados: Los resultados del estudio muestran que el KNN es el modelo que mejor predice el rendimiento académico para cada uno de los semestres, seguido de los árboles de decisión, con valores de precisión que oscilan alrededor del 80 y 78,5 % en algunos semestres.

Conclusiones: Con respecto a las variables, no se puede decir que el promedio académico semestral de un estudiante influya necesariamente en la predicción del rendimiento académico del siguiente semestre. El análisis de estos resultados indica que la predicción del rendimiento académico utilizando herramientas de Machine Learning es un enfoque promisorio que puede ayudar a mejorar la vida académica de los estudiantes y permitir a las instituciones y a los docentes adoptar acciones que ayuden al proceso de enseñanza-aprendizaje.

Palabras clave: análisis de datos educativos, *Machine Learning*, educación superior

Idioma: Inglés

1. Introduction

One of the areas that significantly impacts society is education, as it has a great influence on reducing poverty and unemployment, as well as on improving the life conditions of the community [1] .In the education sector, metrics have been identified such as the annual dropout rate, the dropout rate per cohort, the graduation rate, and the inter-monthly absence rate [2], which allow measuring students' academic performance [3]. Academic performance is a multidimensional concept that depends on multiple aspects such as the objectives of the teacher, the institution, and the student, *etc.* It also requires an integration of different techniques and methodologies for its prediction [4].

Academic performance involves each of the actors in the teaching-learning process, which has been approached from different fields of knowledge (psychology, education, medicine,

statistics, among others), issuing various definitions [5], [6]. This concept is considered to represent a level of knowledge demonstrated in an area or subject while considering age and academic level [7]. In other words, academic performance is measurable from an assessment of the student; it is the sum of different and complex factors that generate an impact on him/her [8]. Similarly, for [9], there are a series of factors that revolve around effort and indicate the success or failure of the student [10]. Currently, with the incursion of the web and ICTs applied to education, this has undergone a series of changes, among which a large volume of data has emerged given the interaction between students, teachers, and institutions [11], [12]. These data are stored, and little of them is used to improve the academic performance and orientation of the student [13]. Therefore, it is necessary to investigate a decision-making model that contributes to the improvement of academic performance.

Decision-making models in the education sector have undergone a certain evolution in terms of the type of data analytics used, as suggested by [14]: descriptive analytics (performance of all the activities studied) carried out with spreadsheets; diagnostic analytics (past performance to analyze information) conducted by means of computer science; and predictive analytics (anticipating behaviors based on historical relationships between variables) performed using data mining and machine learning techniques.

1.1   Related works

Machine learning is a subdiscipline of artificial intelligence that is based on addressing and solving problems from numerical disciplines such as probabilistic reasoning, research based on statistics, information retrieval, and pattern recognition. In this way, machines, through the execution of algorithms, become capable of performing tasks commonly performed by humans [15]. This field is subdivided into several branches, as shown in Fig. 1. Supervised learning takes place when each of the observations of the data set has a related variable or information that indicates what happened (*i.e.*, when entries are labeled). Machine learning (ML) has begun to permeate the educational field, allowing for the collection, cleaning, analysis, and visualization of data on educational actors, in order to optimize related aspects of the teaching-learning process [15], which is why it is currently regarded as one of the techniques that will help decision-making in these contexts [16].

In the last decade, multiple studies have been carried out which seek to establish the variables that specifically affect academic performance. Research has been carried out in areas such as psychology, where, apart from demographic data, the influence of variables related to interest, motivation, attendance, integration, self-regulation, commitment, participation, anxiety, and communication on academic performance have been considered [17]-[21]. From the field of statistics, contributions have been made such as those reflected in [21]-[23], which applied statistical models that seek to examine the variables involved in university admission (admission and pre-university exams), proposing a model that involves various interrelated variables in an attempt to predict academic performance. Some early research have grouped the variables into economic, demographic, and psychological factors [24] [25]. Others have expanded the number of factors, grouping them into demographic, socioeconomic, institutional, sociocultural, socioeconomic, pedagogical, academic, psychological, intellectual, and technological factors, and, due to the rise of ICTs, they have included the learning analytics factor (online interactions) [10].
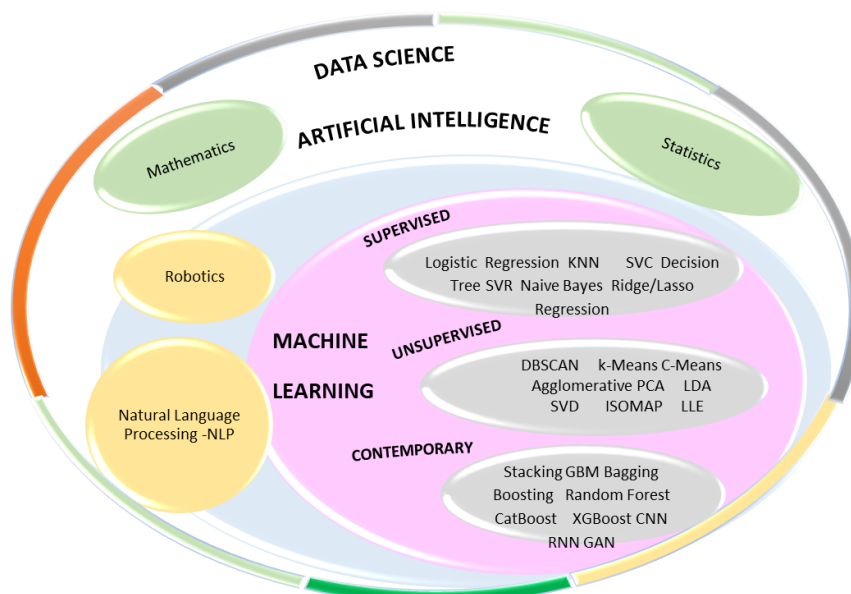
Fig. 1. Overview of machine learning

Recent works have made it possible to group the variables into fewer factors, such as previous academic performance, demographics, e-learning activity, and psychological and environmental factors [26], considering their influence on the variable under study. Table I shows some previous works that have used supervised algorithms as prediction models of academic performance. The variables associated with these studies were grouped into the factors of the classification proposed in [27]. This classification is obtained considering previous research and our reference research [27]-[29], grouping the variables that are easy to identify, of a controllable nature, that are supported by theory, and that can be grouped into previously defined factors. It can be seen that most variables are grouped mainly within the academic and sociodemographic factors (place of residence, number of family members, level of education of the parents, distance traveled to the educational center), followed by psychosocial factors and academic management.

Table I.

Previous work on predicting academic performance using supervised algorithms

Factor	Variables	Previous work with supervised machine
Academic	Government test score, grade point average from the last year of high school, admission test result, academic average or GPA (Grade Point Average), grades by subject, behavior in seminars, conferences and extracurricular activities	[30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]
Socio - demographic	Age, gender, language, marital status, nationality, socioeconomic variables such as stratum, family income, place of residence, parental education level, occupation, number of family members, distance traveled per journey to school	[32], [33], [34], [35], [36], [41], [42], [43], [44], [45]
Online learning	Number that the student has accessed the platform, number of tasks assigned by the teacher, number of exams taken, participation in discussion forums, amount of material viewed, hours online, number of attendances or absences	[2], [46], [47], [48]

Academic management	Year of admission to the university, number of credits, scholarships obtained, credits taken, credits approved, credits lost, final grade for each subject, number of subjects taken, number of subjects passed, number of subjects missed, number of subjects repeated, number of times that the student has missed a subject	[27], [37], [39], [40], [45], [46], [47], [49]
Psychosocial	Interest, motivation, assistance, integration, teamwork, self-regulation, commitment, participation, stress, anxiety	[30], [34], [38], [46], [47], [50], [51], [52], [53]
Academic environment	Type of class/course, duration of the semester, type of program, duration of classes, faculty, course preparation, material, assignments, available resources	[46], [48], [49], [53] [54];

1.2 Contributions and organization

This work explores three concepts that converge in the models: academic performance and its possible ways of evaluating it; the factors that affect it; and supervised machine learning algorithms. In the literature review in [27], [28], [29], which was previously published by the authors, there are related investigations that propose models with various variables that influence performance, but these are usually applied to studying academic performance in an exam, in a specific course, in a year, or to obtain an academic degree. In this sense, this research addresses the problem of determining it throughout the student's academic life (ten academic semesters) by using data transformation tools, feature selection methods, and supervised ML algorithms.

The fields or areas of knowledge that have studied the multidimensional variable of academic performance are diverse. This has been approached from the field of psychology [17], [18], [19], [55], [56], [57], which has applied tools related to questionnaires on students' perceptions regarding academic performance, followed mainly by statistical tools that have a much more marked focus on demographic data and their influence on the variable of interest [21], [22], [58], [59]. Likewise, research related to data science is important, especially studies that use data mining algorithms and ML applied to the field of education. Therefore, a significant contribution is to propose a methodology and a model to establish university academic performance. Approximately 324 variables are analyzed in this work (50 variables analyzed for each academic semester). The authors provide the essential steps to be followed in order to correctly apply ML algorithms to the field of education (in this case, for a 10-semester engineering program). The results show that, with a good dataset, it is possible to analyze situations of academic life or indicators of educational quality that lead to an improvement of the educational process at the university and secondary and primary education levels. This is an interesting contribution for teachers and researchers in the field of education and engineering who wish to investigate issues of education and ML, since engineering articles generally do not provide a clear and easy-to-learn methodology.

Using ML algorithms (Decision Trees, KNN, SVC, Naive Bayes, LDA), various models have proposed in order to predict the academic performance of engineering students in each of their 10 academic semesters. The number of records used to analyze the 50 variables on average in each of the 10 semesters ranges between 2.300 and 2.100 for the first four semesters studied, as well as between 2.100 and 1.800 for the other semesters. These proposed models and their relevant variables allow for decision-making regarding both students and teachers. This, despite the fact that all of the variables present in the consulted literature are not used.

The rest of the article is organized as follows: Section 2 describes the research methodology; Section 3 details the tests and their results; Section 4 presents a discussion of the results obtained; and Section 5 outlines the conclusions.

2. Materials and methods

The methodology employed in this research is presented in the following eight steps: 1) referential information; 2) data source; 3) data cleaning and conditioning; 4) statistics; 5) data transformation, 6) selection of characteristics; 7) prediction algorithms; and 8) performance metrics.

2.1 Reference information

Initially, a review was carried out in databases such as Springer Links, Proquest, IEEE Explorer, and Science Direct, using combinations of keywords, *i.e.*, "academic performance + machine learning, supervised learning + academic performance, academic performance + EDM, data mining + academic performance, improving educational + Machine Learning". The aim was to identify the supervised learning ML algorithms for evaluating academic performance in higher education along with its relevant variables. This referential research was carried out for a period of five years using the method for systematic literature reviews (SRL) proposed by [60], whose initial phase has already been published [61].

2.2 Data source

Universidad Distrital Francisco José de Caldas (Bogotá DC, Colombia) provided a database with a total of 1.614.472 data from 4.738 students of the Industrial and Electrical Engineering programs between 2008 and 2018. These data from both teachers and students are summarized in 324 variables and grouped into five factors defined in Table I: pre-university academic, socio-demographic, socio-economic, academic management, and academic environment. Based on this information, a methodology was proposed, as well as supervised algorithms that allow predicting university academic performance.

2.3 Data cleaning and conditioning

This process initially consisted of eliminating unwanted observations, correcting structural errors, managing values, and handling missing data, as this would probably be reflected as abnormal data and cause poor prediction in the final models. Likewise, information from students who had inconsistent records was discarded, and new variables were created from the information provided (*e.g.*, distance traveled per journey to school, per-semester average, number of subjects taken). Thus, the information was organized, considering the aforementioned factors and the vast majority of variables that group each factor, which resulted in 4.500 records of undergraduate students.

2.4 Data statistics

The supplied datasets (.CSV files) were merged, thus obtaining input data. Descriptive statistics were carried out through Python libraries in order to learn more about the data framework [62].

2.5 Data transformation

As it is possible that an independent variable exerts a greater influence on the dependent variable (in this case, academic performance) due to the fact that its numerical scale is greater than that of the other variables, it was necessary to carry out different types of transformations in order to obtain a better quasi-Gaussian curve for the variables of the dataset (Rescale, Standardize, Normalize, Yeo-Johnson, Box-cox). These transformations

sought to eliminate influence effects, since they are mainly syntactic modifications carried out on data without changing the algorithm [63].

2.6 Feature selection

In order to take advantage of the information provided, a good selection must be made of the most inclusive or relevant characteristics of the output variable [64]. The literature presents two options: the use of feature selection methods (which include and exclude the most relevant features for the development of the problem without changing them and which are generally divided into filter, wrapping, embedded, and assembler methods); and dimensional reduction methods (which create new combinations of attributes from base ones).

2.7 Prediction algorithms

The supervised machine learning algorithms implemented in the dataset were KNN, Decision Trees, SVC, Naive Bayes, and LDA. It is worth mentioning that it was necessary to calculate the dependent variable of study (academic performance) semester by semester in accordance with the norms established by the University and the Colombian government, since its wide range of numerical values generated inconsistencies in the execution. The scale generated to define the variable is shown in Table II, which is based on the ranges established by the Colombian Ministry of National Education.

Table II

Performance variable conventions

Performance	Average	Number
Superior performance	50-45	4
High performance	44-40	3
Basic performance	39-30	2
Low performance	29-0	1

K-Nearest Neighbors (KNN) is one of the classification algorithms whose performance depends on the selection of the hyper parameter K and the distance measure used between two data points (Euclidean, Manhattan, or Minkowski) [65]. Decision Trees are a kind of diagram that consists of internal nodes corresponding to a logical test on an attribute and connection branches used to illustrate the whole process and show the result [66]. The top node in a tree is the root node and represents the entire dataset [67]. In order to establish which is the best partition of the node, different metaheuristics have been suggested which seek to minimize entropy, *i.e.*, information gain and the Gini index. SVM (Support Vector Machines) allow searching for a hyperplane in a high dimensional space that separates the classes in a dataset. It is implemented using a kernel (linear or nonlinear) [68]. Naive Bayes is a classifier supported by Bayes' theorem with good classification precision. It is implemented by estimating a posterior probability [69]. Finally, LDA makes predictions by estimating the probability that a new set of entries belongs to each class. The class that gets the highest probability is the output class, and a prediction is thus made [70].

2.8 Performance metrics

There are several ways to evaluate the results of a ML algorithm. According to [71], the quality of the classification should be evaluated by one of the four different performance metrics: accuracy, precision (specificity), recall (sensitivity) and the F1 score. Values that are determined from the confusion matrix (Table III).

Table III

Confusion matrix

		Predicted values	
		Positive	*Negative*
Actual values	*Positive*	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Accuracy is defined as the number of correctly predicted instances over the total number of records, precision is the ratio of correctly predicted positive instances to the total predicted positive instances, sensitivity is calculated as the ratio of the number of correctly predicted instances to the total number of positives, and the F1 score is the weighted average of precision and sensitivity.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

3. Results

By applying the methodology described above, various results were obtained for steps 4, 5, 6, 7, and 8.

3.1 Regarding statistics

The base dataset consists of 324 variables on average which influence students' academic performance and were grouped by semester. It was necessary to create other variables mentioned in the literature that could influence performance, *e.g.*, the number of subjects taken, missed, and repeated. Universidad Distrital Francisco José de Caldas constantly measures the variables of interest and commitment of the students during their time at the university, applying measurement mechanisms per semester (known as academic tests). Another variable created was distance. This variable is considered, since the time it takes for the student to go from his residence to the university can his/her influence academic performance. The distance between the student's residence and the university was determined by means of approximations using the Google Maps tool, drawing a radial perimeter, taking the centroid of each location on the map of Bogotá as a reference.

3.2 Data transformation

Data transformations are used to change the type or distribution of data variables towards a standard range, so that they can be compared and subjected to different correlation and/or prediction models [72]. From the 4.500 student records, different types of data transformations were carried out, since it is often possible to improve the performance of a range of ML algorithms when the input characteristics are close to a normal distribution [73] or are quasi-Gaussian (Fig. 2a). As an example, Fig. 2b epicts the curves of how the performance of the models (KNN, Decision Trees, NB, and SVC) varies with and without a transformation method in the context under study. The performance metric used to compare the results of each model is accuracy. Fig. 2b compares the model performance improvement when using data without transformation (NO_TRANS) *vs.* using data

transformation methods (Rescale, Standardize, Normalization, Robust Standardization, Box-Cox, or Yeo-Johnson) on the implemented supervised algorithms. The improvement in accuracy typically ranges from 3 to 7% when a transformation method is applied to the data. This is not the case for the SVC algorithm with non-linear kernel. The best results are obtained when the Yeo-Johnson transformation is used.
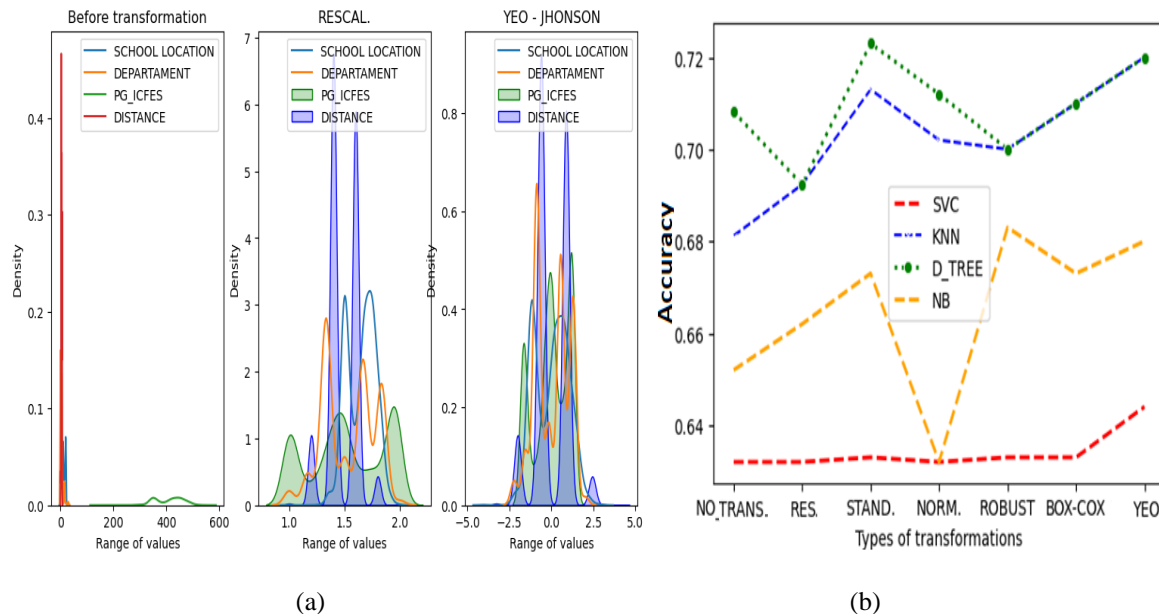

Fig. 2. a) Distribution for some variables before and after transformation methods (Rescale, Yeo-Johnson); b) metrics before and after the use of transformations

3.3 Feature selection

It was interesting to determine, for each academic semester, which would be the most influential variables in a student's performance as he/she advances through his/her university life. To this effect, filter methods were used (Pearson correlation, ANOVA, Chi Square, and mutual information); as well as envelope methods (recursive feature elimination RFE with logistic regression, RFE with logistic regression, RFE-SVC, RFE-Linear regression, RFE-Decision Trees, Backward Selection, Forward Selection, Bi-Directional Elimination), embedded methods (linear regression, Lasso regularization), assembly methods (CART, Random Forest, ExtraTreesClassifier, XGBoost, CatBoost, LightGBM). The number of characteristics that yielded the best per-semester value of the performance metric in the models for Industrial Engineering is shown in Fig. 3a. It is worth mentioning that the results obtained by each method in each of the semesters were tabulated and, in general, the characteristics produced by the assembly methods are the ones that provide the best results when the supervised learning algorithms (KNN and Decision Trees) are applied. This step is considered fundamental for the models, as it is necessary for those that provide information to the model to be the relevant variables, not those that introduce noise. As an example, Fig. 3b shows the results regarding the precision of the models involving Decision Trees and KNN when trying to predict academic performance in the sixth semester of Industrial Engineering with different amounts of characteristics while using 10-fold cross-validation. In a previous work, the authors had presented a first attempt to predict the

academic performance of students in only the first semester, with a model precision of 66,6%, in which they established the pre-university variables that influenced the academic performance of students (10 out of 25 were selected) [74].
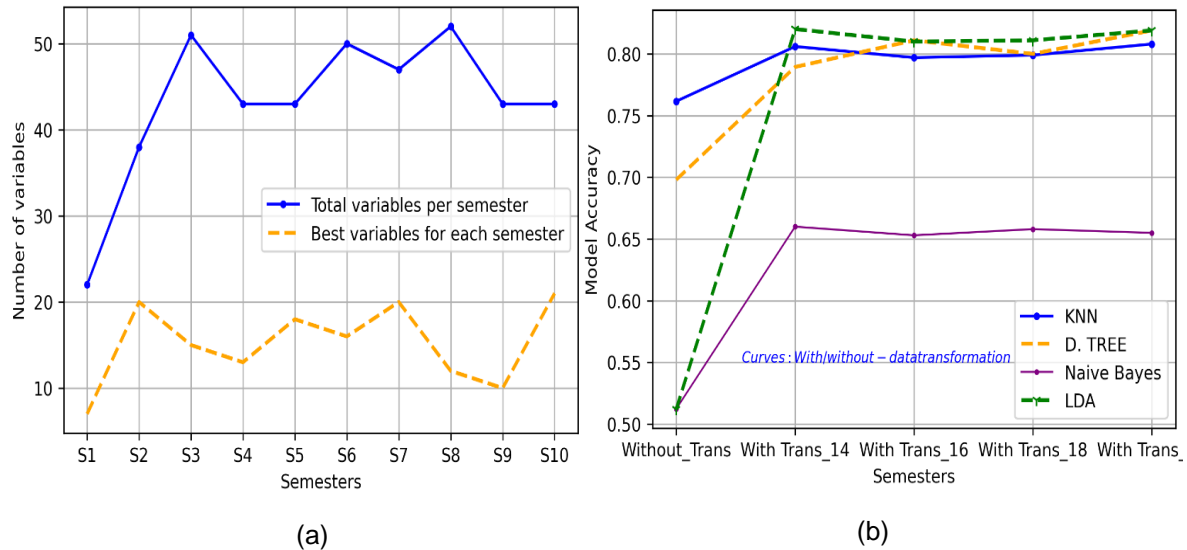

(a)   (b)

Fig. 3. a) Number of best characteristics to predict academic performance in each semester; b) accuracy according to the number of best variables selected by the methods

Another interesting aspect was the fact that engineering courses usually have a common component (basic engineering subjects). Therefore, this study sought to identify which would be the subjects of this basic component that most influence the determination of university academic performance when estimated consecutively for the first three semesters (Table IV).

Table IV

Common variables in the determination of academic performance within the basic cycle of Industrial and Electrical Engineering

Semester	Common variables	
1	ICFES Global Score ICFES Area of Biology ICFES Math Area	ICFES Area of Biology School Location Residence Location
2	ICFES Global Score ICFES Math Area ICFES Area of Biology Residence Location	Student Average (1 Semester) Grade_Lecture FJC Number of Subjects Repeated (1 Semester) Grade_Text
3	Student Average (1 Semester) Student Average (2 Semester) Grade Differential Calculation Number of Credits Studied (2 Semester)	Grade Algebra Number of Subjects Studied (2 Semester) Number of Subjects Approved (1 Semester) Grade_Integral Calculus

3.4 Prediction algorithms

As previously mentioned, models with supervised learning algorithms were implemented: SVC, KNN, Decision Trees, Naive Bayes, and LDA for the dataset corresponding to each of the 10 academic semesters, which had to be divided into training and test data. The literature presents options in order to avoid subsampling or oversamplings such as cross-validation

(it works in the search for less variance), which works by dividing the data set into k parts (k = 10), which are called folds (*i.e.*,10-fold), where the first fold will act as a validation set and the model is trained with the k-1 (fold). Each time the model is validated with a different fold, it will be trained with the remaining k-1. In addition, the random method was used (70% for training data, 30% for the test data) in order to estimate the performance of the algorithms [73]. Some of the best results of the performance metrics of the algorithms are shown in Table V for Industrial Engineering.

Table V

Results of the models with cross-validation (CV) and random method (for different semesters)

Accuracy	CV Method – Training data	CV Method – Test data	Random method – Training data	Random method – Test data	Accuracy	CV method – Training data	CV method – Test data	Random method – Training data	Random method – Test data
		SEMESTER 1					*SEMESTER 6*		
KNN	0,615	0,616	0,660	0,602	KNN	0.836	0.817	0.811	0.806
SVC	0,534	0,513	0,626	0,637	SVC	0.773	0.770	0.705	0.235
D_TREE	0,639	0,637	0,634	0,621	ARBOL	0.805	0.765	0.821	0.789
NAIVE	0,582	0,623	0,609	0,607	NAIVE	0.685	0.730	0.692	0.660
LDA	0,628	0,638	0,627	0,642	LDA	0.812	0.806	0.810	0.820
		SEMESTER 2					*SEMESTER 7*		
KNN	0,815	0,779	0,836	0,810	KNN	0,804	0,7339	0,782	0,761
SVC	0,590	0,425	0,710	0,670	SVC	0,703	0,634	0,645	0,687
D_TREE	0,817	0,736	0,847	0,766	ARBOL	0,820	0,710	0,777	0,681
NAIVE	0,670	0,676	0,674	0,679	NAIVE	0,669	0,608	0,690	0,650
LDA	0,805	0,771	0,810	0,764	LDA	0,776	0,743	0,787	0,757
		SEMESTER 3					*SEMESTER 8*		
KNN	0,782	0,785	0,766	0,792	KNN	0,766	0,7075	0,776	0,746
SVC	0,607	0,709	0,597	0,660	SVC	0,664	0,6126	0,637	0,603
D_TREE	0,782	0,762	0,808	0,782	ARBOL	0,791	0,669	0,757	0,692
NAIVE	0,606	0,661	0,615	0,644	NAIVE	0,615	0,5039	0,598	0,510
LDA	0,783	0,785	0,7838	0,776	LDA	0,763	0,7108	0,769	0,7301

There are different libraries that are used to optimize the hyperparameters of the classification algorithms, such as Scikit-learn (GridSearch, Random Search) and Scikit-Optimize. In this work, the optimization of parameters was carried out by means of Grid Search, an approach that is in charge of constructing and exhaustively checking all the combinations in the parameter space (specified in advance) of an algorithm. To determine the best value for the hyperparameters, the cross-validation method was used to avoid over-fitting the model. Some hyperparameters that should have been optimized for each of the models are shown in Table VI.

Table VI

Example of hyper parameters for the algorithms

Algorithm	Hyperparameters to optimize with some value options	
KNN	n_neighbors: [i for i in range(1,17)] metric: [Euclidean, Manhattan, Minkowski]	p: [i for i in range(1,6)] weights:[uniform] algorithm: [auto, kd_tree, ball_tree, brute]
SVC	kernel: [rbf, poly, sigmoid,linear] random_state: [1, 10, None]	C: [0.01,0.1, 1, 10, 100, 1000] max_iter: [i for i in range(1,30,2)] gamma : [1, 0.1, 0.01, 0.001, 0.0001]
DECISION TREES	max_depth: [2, 4, 6, None] criterion: [gini, entropy] splitter: [best, random]	max_features :[10, 20, 30, 40, 50, None] min_samples_split : [i for i in range(2,20)] min_samples_leaf': [i for i in range(2,10)]

		random_state: [1, 5, 10]
NAIVE BAYES	(It has no optimization. It has no hyperparameters)	
LDA	solver: [svd, lsqr,eigen]	n_components:[1,2,3,4,5,None]
	shrinkage: [auto, 0, 0.001, 0.01, 0.1, 0.5,1]	

3.5 Performance metrics

There are different metrics to determine if a model performs well. Fig. 4a shows the value of the accuracy metric of each model for each of the academic semesters after implementing the Yeo-Johnson transformation (it was the one that yielded the best, quasi-Gaussian data) and selecting the most influential characteristics in the response variable (academic performance). An accuracy value (ratio between the correctly predicted observation and the total number of observations) close to 1 indicates that all the predictions are correct. A value close to 0 suggests a very bad prediction model. The KNN algorithm yielded the best results in the vast majority of academic semesters (greater than 77,5%), closely followed by the Decision Trees (greater than 76,9%) and the LDA method (greater than 76,5%). The KNN algorithm was not only the best at predicting academic performance in each semester; it also showed precision values between 76 and 82% when evaluating the students' year of study (every other semester). This is a contribution of this work, in the sense that previous works show predictions for a group of data in particular that correspond to a subject, several subjects of a semester, or, in the best of cases, to a sum of semesters in particular. Instead, this research aimed to predict academic performance throughout students' academic life, *i.e.*, semester after semester and year after year in different engineering curricular programs (Fig. 4b).
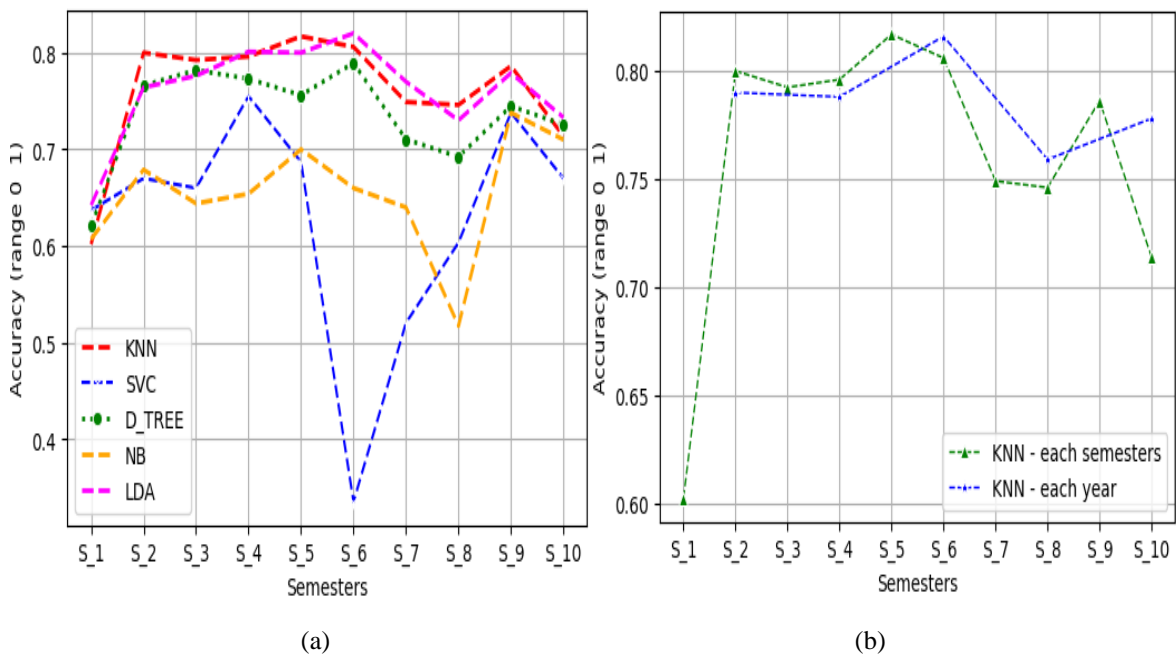

(a)   (b)

Fig. 4. a) Summary of evaluation metrics of classic ML models with supervised learning (industrial engineering); b) KNN precision in predicting academic performance (per semester and per year for industrial engineering)

The best variables considered to predict academic performance in each of the semesters are shown in Table VII. As expected, there are differences between the predictions made for each semester and those for each year.

Table VII

Relevant variables by academic semester to determine academic performance (four semesters)

Semester	Relevant Variables	
1	ICFES Global Score	ICFES Area of Biology
	ICFES Math Area	ICFES Area of Biology
	School Location	Residence Location
2	ICFES Global Score	ICFES Math Area ICFES
	ICFES Area of Biology	Physics Area
	Residence Location	ICFES Language Area
	Grade_Text	School Location
	ChemistryGrade	Student Average (1 Semester)
	Number of Subjects Repeated (1 Semester)	Number of Subjects Approved (1 Semester)
3	Student Average (1 Semester)	Grade_Drawing
	Grade Differential Calculus	Number of Credits Studied (2 Semester)
	Student Average (2 Semester)	Number of Subjects Approved (2 Semester)
	Grade_Materials	Grade_Integral Calculus
	Grade_Algebra	Number of Subjects Studied (2 Semester)
4	Student Average (1 Semester)	Number of Subjects Approved (1 Semester)
	Grade_Algebra	Ethics_Grade
	Grade_Oriented Programming	Student Average (2 Semester)
	Grade_Multivariate Calculus	Grade_Statistics_One
	Note Thermodynamics	Grade_General Theory

There are different metrics to evaluate the algorithms. As an example, it is shown in Table VIII that the KNN algorithm not only obtained high accuracy results, but it also surpassed the others in precision, recall, and the F1 score. Precision values (which measure the ability of a classifier not to label an observation as positive when it should be considered as negative) has values close to those of accuracy. The recall and F1 scores are also good values, as they are close to 1 (100%)

Table VIII

Metrics of the algorithms implemented for the first four academic semesters (industrial engineering)

Semester	Algorithm	accuracy_score	precision_score	recall_score	f1_score
1	KNN	0,606	0,603	0,603	0,603
	SVC	0,636	0,632	0,632	0,632
	DECISION_TREE	0,634	0,639	0,639	0,620
	NAIVE BAYES	0,639	0,639	0,639	0,639
	LDA	0,642	0,642	0,642	0,642
2	KNN	0,836	0,803	0,803	0,792
	SVC	0,771	0,675	0,674	0,675
	DECISION_TREE	0,767	0,761	0,761	0,762
	NAIVE BAYES	0,679	0,674	0,674	0,673
	LDA	0,810	0,764	0,763	0,764
3	KNN	0,777	0,752	0,736	0,722
	SVC	0,597	0,582	0,597	0,582
	DECISION_TREE	0,808	0,793	0,778	0,763
	NAIVE BAYES	0,615	0,615	0,615	0,615
	LDA	0,784	0,769	0,754	0,738
4	KNN	0,823	0,809	0,793	0,778
	SVC	0,757	0,742	0,715	0,700
	DECISION_TREE	0,826	0,812	0,796	0,781
	NAIVE BAYES	0,710	0,695	0,680	0,665

LDA	0,810	0,795	0,780	0,765

Then, the supervised algorithms were implemented. Note that the best features were provided and that the hyperparameters were optimized by implementing the Grid Search method. Fig. 5 shows the average results regarding the precision metric for predicting academic performance in Industrial and Electrical Engineering programs.
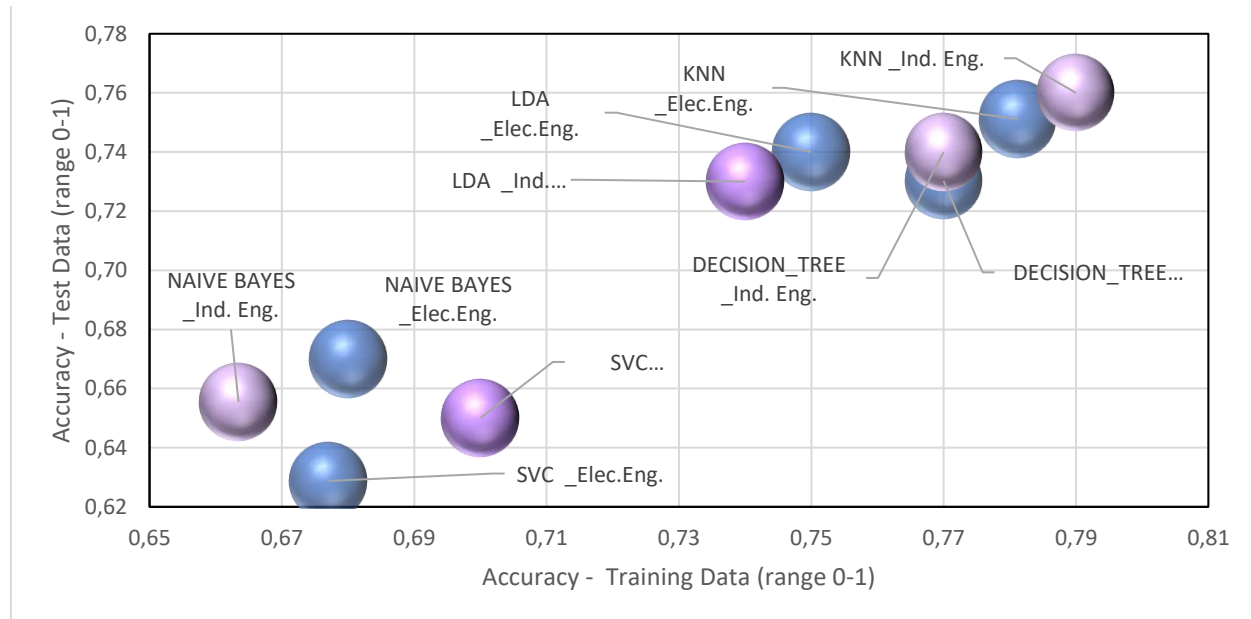
Fig. 5. Precision with supervised algorithms for Industrial and Electrical Engineering programs

4. Discussion

Based on the results obtained, it can be stated that, in the last decade, the development of tools and techniques in the field of computer science has allowed data analytics to penetrate many fields, such as the education sector, where it is used not only in the prediction of students' academic performance but also of other indicators of educational quality such as dropout and graduation rates. Thus, this project lays the foundations to continue with the exploration of how to estimate, predict, or group students in order to take appropriate actions that guide their academic course. The characteristics selection methods employed show that the gender variable was not relevant when determining academic performance, which is somewhat similar to the findings of studies such as [75] and [75]. However, when reviewing the literature, works from the field of psychology were found, such as [76], [77], and [78], which state the opposite. Something similar occurs with the place of origin, as studies such as [78] and [79] argue that students from diverse geographic locations have specific knowledge, prior experiences, and different ways of life that are guided in various ways by teachers to meet educational needs, which affects the way they learn. However, there are studies such as [80] and [81] which suggest that this has no significant effect. Thus, it seems that, depending on the analyzed group, the results may or may not be similar in terms of the influence of the independent variables on academic performance, which happened in this work.

The academic average or GPA (Grade point average) is one of the variables that exerts the most influence on the determination of performance, albeit not in all 10 semesters, as

there are semesters where performance is influenced by the number of subjects taken and their corresponding grades. Socioeconomic variables do not show a high influence on the precision of the prediction for first semester students, unlike the results obtained by [82]. Also for first semester students, [40] achieved 49,078% accuracy with their best algorithm; in this work, this value was between 60 and 65%. The performance of the algorithms yielded better values, despite being similar to those implemented by [35], [52], [83], and [84]. Nevertheless, the results of this work are very different to those of [38] with regard to the SVC algorithm; the latter obtained values close to 90%, whereas our study was below 80%. It is worth mentioning that [38] considered psychological parameters, learning strategies, and learning approaches that were not taken into account in this work. This suggests that the determination of academic performance is a complex process that varies from institution to institution, and that it may not be possible to generalize with regard to the influencing variables and the best algorithms. It is instead possible to estimate according to particular conditions while considering general variables and factors. According to the results, the KNN algorithm allows predicting, with metrics such as accuracy, the per-semester or yearly academic performance of a student while only considering some academic variables (subjects attended, averages, subjects failed and approved, and overall approved credits) and some pre-university demographic variables.

However, this study has some methodological limitations, such as the lack of available and/or reliable data. The date used in this research is the product of the information provided by the systems office of Universidad Distrital Francisco José de Caldas (data warehouse), to which a generous cleaning process should have been applied in light of the errors found. Another important limitation is access to information, since some information related to economic and psychosocial variables was not provided by the University (it was regarded as private information). Therefore, some variables such as commitment, participation, stress, anxiety, assertive communication, and family income could not be analyzed in order to determine whether or not they influenced the academic performance of the students. These limitations could be corrected by means of a database that allows access to as many variables as possible, as well as investigating variables of the students' environment which may influence their performance.

5. Conclusions

According to the work presented and the results obtained, the following main conclusions can be drawn:

Not applying transformation or feature selection methods on the data generates models with low performance metrics, even when the hyperparameters of the supervised learning algorithms are optimized. This is reflected in the good results obtained with the Yeo-Johnson transformation method *vs.* those yielded by Rescale, Standardize, Normalize, and Box-Cox. These transformations seek to eliminate influence effects, and they are mainly syntactic modifications carried out on data without involving a change to the algorithm.

The pre-university variables pertaining to demographic and socio-demographic factors are not conclusive when trying to predict students' academic performance, as their accuracy is around 65%.

Although not always, the best result is provided by the same algorithm regarding per-semester academic performance. It can be stated that the KNN algorithm (accuracy greater than 77,5%) provides good results, especially in even semesters, closely followed by Decision Trees (greater than 76,9%) and the LDA method (greater than 76,5%).

The results indicate that the prediction of academic performance using machine learning tools is a promising approach that can help improve students' academic life and can allow institutions and teachers to take actions that contribute to the teaching-learning process.

Machine learning tools have been increasingly used in education in the last decade. This aspect, added to the detection of the variables that most influence academic performance, will allow to continue implementing other algorithms belonging to other branches within this field, such as assembly and deep learning methods.

In order to continue with the process of searching for models and algorithms that better predict academic performance, it is necessary to implement contemporary assembly methods (Bagging, Boosting, Voting), which belong to another branch of ML and are based on establishing different methods that work together in order to reduce errors.

References

[1] M. Ferreyra, J. Botero, P. Haimovich, and S. Urzúa, "Momento decisivo La educación superior en América Latina y el Caribe," Washington, 2017. [Online]. Available: https://openknowledge.worldbank.org/bitstream/handle/10986/26489/211014ovSP.pdf

[2] E. J. de La Hoz, E. J. de La Hoz, and T. J. Fontalvo, "Methodology of Machine Learning for the classification and prediction of users in virtual education environments," *Inf. Tecnol.*, vol. 30, no. 1, pp. 247-254, Feb. 2019. https://doi.org/10.4067/S0718-07642019000100247

[3] Ministerio de Educación, "Sistema nacional de información de la educación superior," 2019. [Online]. Available: https://snies.mineducacion.gov.co/portal/

[4] I. A. Khan and J. T. Choi, "An application of educational data mining (EDM) technique for scholarship prediction," *Int. J. Softw. Eng. Its Appl.*, vol. 8, no. 12, pp. 31-42, 2014. https://doi.org/10.14257/ijseia.2014.8.12.03

[5] H. Lamas, "Sobre el rendimiento escolar," *Prósitos y Represent. Rev. Psicol. Educ.*, vol. 3, no. 1, pp. 313-386, 2015. https://doi.org/10.20511/pyr2015.v3n1.74

[6] J. Espinosa, J. Hernández, J. Rodríguez, M. Chacín, and V. Bermúdez, "Influencia del estrés sobre el rendimiento académico," *AVFT-Archivos Venez. Farmacol. y Ter.*, vol. 39, no. 1, 2020. https://doi.org/10.5281/zenodo.4065032

[7] M. G. Jiménez, J. A. I.- Psicothema, and 2000, "La predicción del rendimiento académico: regresión lineal versus regresión logística," *Psicothema*, vol. 12, pp. 222-248, 2000. https://www.psicothema.com/pdf/558.pdf

[8] Garbanzo and G. María, "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. https://www.redalyc.org/articulo.oa?id=44031103

[9] L. Rojas, *"Validez predictiva de los componentes del promedio de Admisión a la universidad de costa rica utilizando el Género y el tipo de colegio como variables control,"* Rev. Elec. Actual. Investig. en Educ., vol. 13, no. 1, pp. 17-25, Jan. 2013. https://revistas.ucr.ac.cr/index.php/aie/article/view/11707/18183

[10] D. García, J. Manuel, and M. Pichardo, "Learning analytics as an analysis factor of university academic performance," in *CEUR Workshop Proceedings,* 2019, pp. 42-50. http://ceur-ws.org/Vol-2231/LALA_2018_paper_14.pdf

[11] J. Huamán, "Evaluación del rendimiento académico estudiantil de la cohorte 2011-2015, según áreas de la carrera de estomatología Universidad Peruana Cayetano Heredia". Título de Cirujano Dentista, Departamento Académico de Odontología Social, Universidad Peruana Cayetano Heredia, 2018. [Online]. Available: https://repositorio.upch.edu.pe/handle/20.500.12866/1429

[12] D. A. Montoya-Arenas, E. M. Bustamante-Zapata, C. M. Díaz-Soto, and D. Pineda, "Factores de la capacidad intelectual y de la función ejecutiva relacionados con el rendimiento académico en estudiantes universitarios," *Rev. la Esc. Cienc. Salud Univ. Pontif. Boliv.*, vol. 40, no. 1, pp. 10-18, 2021. https://doi.org/10.18566/medupb.v40n1.a03

[13] L. Contreras, J. Rodríguez, and H. Fuentes, "Analítica académica: nuevas herramientas aplicadas a la educación," *Rev. Boletín Redipe*, vol. 10, no. 3, pp. 137-158, 2021.

[14] P. Murnion and M. Helfert, "Academic analytics in quality assurance using organisational analytical capabilities," in *Annual Conf. UK Acad. Info. Sys. (UKAIS)*, 2013. [Online]. Availavle: https://doi.org/10.13140/2.1.3368.1600

[15] G. Hackeling, *Mastering machine learning with scikit-learn: Learn to implement and evaluate machine learning solutions with scikit-learn*, 2nd ed., vol. 1., Bigmingham, UK: Packt Publishing Ltd., 2014.

[16] L. Contreras, H. Fuentes, and J. Rodríguez, "Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático," *Form. Univ.*, vol. 13, no. 5, pp.

233-246, 2020. https://doi.org/10.4067/S0718-50062020000500233

[17] T. C. Hakyemez and S. Mardikyan, "The interplay between institutional integration and self-efficacy in the academic performance of first-year university students: A multigroup approach," *Int. J. Manag. Educ.*, vol. 19, no. 1, 2021. https://doi.org/10.1016/j.ijme.2020.100430

[18] G. Guizado, M. Valenzuela, and P. Vallejo, "Desempeño docente y el rendimiento académico de los estudiantes de la Facultad de Tecnología en la Universidad Nacional de Educación de Perú," *Rev. Conrado*, vol. 16, no. 72, 200-203, 2020. https://conrado.ucf.edu.cu/index.php/conrado/article/view/1231

[19] E. Zárate, B. Lavado, and W. Pomahuacre, "Competecia comunicativa intercultural y rendimiento académico en lenguas extranjeras," *Rev. Conrado*, vol. 16, no. 74, 30-37, 2020. https://conrado.ucf.edu.cu/index.php/conrado/article/view/1330

[20] T. Icekson, O. Kaplan, and O. Slobodin, "Does optimism predict academic performance? Exploring the moderating roles of conscientiousness and gender," *Stud. High. Educ.*, vol. 45, no. 3, pp. 635-647, 2020. https://doi.org/10.1080/03075079.2018.1564257

[21] A. M. Pavelea and O. Moldovan, "Why some fail and others succeed: Explaining the academic performance of PA undergraduate students," *NISPAcee J. Public Adm. Policy*, vol. 13, no. 1, pp. 109-132, 2020. https://doi.org/10.2478/nispa-2020-0005

[22] H. Vargas, L. Solórzano, and W. Chanini, "Modelo matemático entre el puntaje de examen de ingreso y el rendimiento académico de los estudiantes ingresantes a la Universidad Nacional Jorge Basadre Grohmann, año académico 2018," *Ciencias*, vol. 3, no. 3, 45-51, 2019. https://doi.org/10.33326/27066320.2019.3.949

[23] A. Lenskiy, R. Shariat, and S. Seol, "The effect of academic breaks on undergraduate academic performance," *Int. J. Electr. Eng. Educ.*, 2020. [Online]. Available: https://doi.org/10.1177/0020720920922518

[24] M. Oladejo, "A path-analytic study of socio-psychological variables and academic performance of distance learners in nigerian universities," Doctoral thesis, Univ. Lagos, 2010. [Online]. Available: https://doi.org/10.13140/RG.2.2.19443.73762

[25] M. Kotzé; Niemann, "Psychological resources as predictors of academic performance of first-year students in higher education," Acta *académica.*, vol. 45, no. 2, pp. 85-121, 2013. https://journals.ufs.ac.za/index.php/aa/article/view/1399

[26] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, pp. 1-21, Dec. 2020. https://doi.org/10.1186/S41239-020-0177-7/TABLES/15

[27] G. Tarazona, L. Contreras, and H. Fuentes, "Machine Learning variables and algorithms that influence academic performance: A review," *Int. J. Mech. Prod. Eng. Res. Dev.*, vol. 10, no. 3, pp. 16011-16028, 2020. http://www.tjprc.org/view_paper.php?id=14467

[28] L. Contreras, H. Fuentes, and J. Rodríguez, "Academic Interruption Model using Automatic Learning Algorithms" Sylw*an J.*, vol. 10, no. 3, pp 16075-16086 ,2020. http://www.tjprc.org/view_paper.php?id=14480

[29] L. Contreras, H. Fuentes, and J. Molano, "Analítica académica: nuevas herramientas aplicadas a la educación," *Rev. Bol. Redipe*, vol. 10, no. 3, pp. 137-158, 2021. https://doi.org/10.36260/rbr.v10i3.1225

[30] A. Rico, N. Gaytán, and D. Sánchez, "Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes," *Diálogos sobre Educ.*, vol. 19, art. 509, 2019. https://doi.org/10.32870/dse.v0i19.509

[31] Y. Widyaningsih, N. Fitriani, and D. Sarwinda, "A semi-supervised learning approach for predicting student's performance: First-year," 2019 12th International Conference on Information & Communication Technology and System (ICTS), pp. 291–295, 2019. https://doi.org/10.1109/ICTS.2019.8850950

[32] F. Otálora, "Modelo para la identificación de patrones de desempeño académico estudiantil para fortalecer el acompañamiento académico en la Universidad Nacional de Colombia," MSc. dissertation, Dept. Elect. Eng., Universidad Nacional de Colombia, 2019. [Online]. Available: https://repositorio.unal.edu.co/handle/unal/77758.

[33] R. Istvan and V. Lasagna, "Sistema informático para la detección temprana de deserción estudiantil universitaria," *Innovación y Desarro. Tecnológico y Soc.*, vol. 1, no. 2, pp. 1-15, 2019. https://doi.org/10.24215/26838559e006

[34] S. S. M. Ajibade, N. Bahiah Binti Ahmad, and S. Mariyam Shamsuddin, "Educational data mining: Enhancement of student performance model using ensemble methods," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 551, no. 1, art. 012061, 2019. https://doi.org/10.1088/1757-899X/551/1/012061

[35] C. Jalota and R. Agrawal, "Analysis of educational data mining using classification," in *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, 2019, pp. 243-247. https://doi.org/10.1109/COMITCon.2019.8862214

[36] O. Castrillón, W. Sarache, and S. Ruiz, "Predicción del rendimiento académico por medio de técnicas de inteligencia artificial," *Rev. Form. Univ.*, vol. 13, no. 1, pp. 93-102, 2020. https://doi.org/10.4067/S0718-50062020000100093

[37] A. Das and E. Rodríguez, "A predictive analytics system for forecasting student academic performance: Insights

from a pilot project at eastern Washington university," *2019 Jt. 8th Int. Conf. Informatics, Electron. Vision, ICIEV*, 2019, pp. 255-262. https://doi.org/10.1109/ICIEV.2019.8858523

[38] I. Burman and S. Som, "Predicting Students Academic Performance Using Support Vector Machine," in *Proc. 2019 Amity Int. Conf. Artif. Int., AICAI 2019*, Apr. 2019, pp. 756-759. https://doi.org/10.1109/AICAI.2019.8701260

[39] M. V. Amazona and A. A. Hernández, "Modelling student performance using data mining techniques," in *Proc. 2019 5th Int. Conf. Comp. Data Eng., ICCDE' 19*, May 2019, pp. 36-40. https://doi.org/10.1145/3330530.3330544

[40] A. I. Adekitan and E. Noma-Osaghae, "Data mining approach to predicting the performance of first year student in a university using the admission requirements," *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1527-1543, 2019. https://doi.org/10.1007/s10639-018-9839-7

[41] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 381-407, 2019. https://doi.org/10.1007/s10462-018-9620-8

[42] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Comput. Human Behav.*, vol. 98, pp. 166-173, Apr. 2019. https://doi.org/10.1016/j.chb.2019.04.015

[43] Bendangnuksung, "Students' performance prediction using deep neural network," *Int. J. Appl. Eng. Res.*, vol. 13, no. 02, pp. 1171-1176, 2018. https://www.ripublication.com/ijaer18/ijaerv13n2_46.pdf

[44] Y. Nieto, V. García-Díaz, C. Montenegro, and R. G. Crespo, "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft Comput.*, vol. 23, no. 12, pp. 4145-4153, 2018. https://doi.org/10.1007/s00500-018-3064-6

[45] L. Wang and Y. Yuan, "A prediction strategy for academic records based on classification algorithm in online learning environment," *Proc. - IEEE 19th Int. Conf. Adv. Learn. Technol. ICALT 2019*, vol. 2161-377X, pp. 1-5, 2019. https://doi.org/10.1109/ICALT.2019.00007

[46] Y. K. Salal, S. M. Abdullaev, and M. Kumar, "Educational data mining: Student performance prediction in academic," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 4C, pp. 54-59, 2019. https://www.semanticscholar.org/paper/Educational-Data-Mining-%3A-Student-Performance-in-Salal-Abdullaev/b21fa7245581c3baad2d468cb9d706940de7e010

[47] S. Hirokawa, "Key attribute for predicting student academic performance," in *ICETC '18: 10th Int. Conf. Ed. Tech. Comp*, 2018, pp. 308-313. https://doi.org/10.1145/3290511.3290576

[48] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143-19165, 2019. https://doi.org/10.1109/ACCESS.2019.2896880

[49] J. Sotomonte, C. Rodríguez, C. Montenegro, P. Gaona, and J. Castellanos, "Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos," *Rev. Científica*, vol. 3, no. 26, p. 35, 2016. https://doi.org/10.14483/23448350.11089

[50] M. Alloghani, D. Al-Jumeily, A. Hussain, A. J. Aljaaf, J. Mustafina, and E. Petrov, "Application of machine learning on student data for the appraisal of academic performance," *Proc. - Int. Conf. Dev. eSystems Eng. DeSE*, vol. 2018, pp. 157-162, Sep. 2019. https://doi.org/10.1109/DeSE.2018.00038

[51] M. Mohammadi, M. Dawodi, W. Tomohisa, and N. Ahmadi, "Comparative study of supervised learning algorithms for student performance prediction," in *1st Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2019*, 2019, pp. 124-127. https://doi.org/ 10.1109/ICAIIC.2019.8669085

[52] H. Anderson, B. Afshan, and R. Baker, "Predicting Graduation at a Public R1 University," 2019. [Online]. Available: https://learninganalytics.upenn.edu/ryanbaker/paper323.pdf

[53] J. Hou and Y. Wen, "Prediction of learners' academic performance using factorization machine and decision tree," in *2019 IEEE Int. Congr. Cybermatics*, 2019, pp. 1-8. https://doi.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00024

[54] Y. S. Alsalman, N. Khamees Abu Halemah, E. S. Alnagi, and W. Salameh, "Using decision tree and artificial neural network to predict students academic performance," in *2019 10th Int. Conf. Inf. Commun. Syst. ICICS 2019*, 2019, pp. 104-109. https://doi.org/ 10.1109/IACS.2019.8809106

[55] T. Icekson, O. Kaplan, and O. Slobodin, "Does optimism predict academic performance? Exploring the moderating roles of conscientiousness and gender," *Stud. High. Educ.*, vol. 45, no. 3, pp. 635-647, Mar. 2020. https://doi.org/10.1080/03075079.2018.1564257

[56] R. C. Céspedes, A. Vara-Horna, D. López-Odar, I. Santi-Huaranca, A. Díaz-Rosillo, and Z. Asencios-González, "Ausentismo, presentismo y rendimiento académico en estudiantes de universidades peruanas," *Rev. Psicol. Educ.*, vol. 6, no. 1, pp. 83-133, Jan. 2018. https://doi.org/10.20511/PYR2018.V6N1.177

[57] P. Luján, L. Trelles, and M. Mogollón, "Asertividad y rendimiento académico en estudiantes de la facultad de ciencias administrativas de la Universidad Nacional de Piura," *UCV - Sci.*, vol. 11, no. 1, 13-20, 2019. https://revistas.ucv.edu.pe/index.php/ucv-scientia/article/view/1170

[58] Y.-W. Liang, D. Jones, and R. A. Robles-Pina, "Ethnic and gender stereotypes on college students' academic performance," *Res. High. Educ. J.*, vol. 35, 2018. https://www.aabri.com/manuscripts/182858.pdf

[59]   C. Durán and A. Rosado, "La comprensión lectora y el rendimiento académico en estudiantes de ingeniería," *Rev. Colomb. Tecnol. Av.*, vol. 1, no. 33, pp. 9-15, Mar. 2019, https://doi.org/10.24054/16927257.V33.N33.2019.3317

[60]   B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7-15, Jan. 2009. https://doi.org/ 10.1016/j.infsof.2008.09.009.

[61]   K. Gonzalez, J. Rodríguez, and L. Contreras, "Academic performance and alternatives with prediction- oriented machine learning: A review of the state of the art," *Int. J. Mech. Prod. Eng. Res. Dev.*, vol. 10, no. 3, pp. 16329-16340, 2020. http://www.tjprc.org/view_paper.php?id=14520

[62]   K. C. Santosh, "AI-driven tools for coronavirus outbreak: Need of active learning and cross-population train/test models on multitudinal/multimodal data," *J. Med. Syst.*, vol. 44, no. 5, pp. 1-5, May 2020. https://doi.org/10.1007/s10916-020-01562-1

[63]   J. García, P. Sánchez, M. Orozco, and S. Obredor, "Extracción de conocimiento para la predicción y análisis de los resultados de la prueba de calidad de la educación superior en Colombia," *Rev. Form. Univ.*, vol. 12, no. 4, pp. 55-62, 2019. https://doi.org/ 10.4067/S0718-50062019000400055

[64]   M. Zaffar, M. A. Hashmani, K. S. Savita, and S. S. H. Rizvi, "A study of feature selection algorithms for predicting students' academic performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 5, pp. 541-549, 2018. https://doi.org/10.14569/IJACSA.2018.090569

[65]   A. K. Das and E. Rodriguez-Marek, "A Predictive Analytics System for Forecasting Student Academic Performance: Insights from a Pilot Project at Eastern Washington University," in *2019 Joint 8th Int. Conf. Informatics Elec. Vision (ICIEV) and 2019 3rd Int. Conf. Imaging,* 2019, pp. 255-262. https://doi.org/10.1109/ICIEV.2019.8858523

[66]   V. L. Uskov, J. P. Bakken, A. Byerly, and A. Shah, "Machine Learning-based predictive analytics of student academic performance in STEM education," in *2019 IEEE Global Eng. Educ. Conf. (EDUCON)*, 2019, pp. 1370-1376. https://doi.org/10.1109/EDUCON.2019.8725237

[67]   R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177-194, 2017. https://doi.org/10.1016/j.compedu.2017.05.007

[68]   J. Horak, J. Vrbka, and P. Suler, "Support vector machine methods and artificial neural networks used for the development of bankruptcy prediction models and their comparison," *J. Risk Financ. Manag.*, vol. 13, no. 3, p. 80, Mar. 2020. https://doi.org/10.3390/JRFM13030060

[69]   F. Ofori, E. Maina, and R. Gitonga, "Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review," *J. Inf. Technol.*, vol. 4, no. 1, pp. 33-55, 2020. https://ir-library.ku.ac.ke/handle/123456789/20243?show=full

[70]   J. Brownlee, "Machine Learning Mastery," 2020. https://machinelearningmastery.com/ (accessed Dec. 21, 2020).

[71]   F. J. Kaunang and R. Rotikan, "Students' academic performance prediction using data mining," in *3rd Int. Conf. Informatics Comput. ICIC 2018*, 2018, pp. 1-5. https://doi.org/10.1109/IAC.2018.8780547

[72]   Pandas.org, "pandas.DataFrame.transform," 2021. https://pandas.pydata.org/

[73]   R. M. Aguilar, J. M. Torres, and C. A. Martín, "Automatic learning for the system identification. A case study in the prediction of power generation in a wind farm," *RIAI - Rev. Iberoam. Autom. e Inform. Ind.*, vol. 16, no. 1, pp. 114-127, 2019. https://doi.org/10.4995/riai.2018.9421

[74]   L. E. Contreras, H. J. Fuentes, and J. I. Rodríguez, "Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático," *Form. Univ.*, vol. 13, no. 5, pp. 233-246, 2020. http://dx.doi.org/10.4067/S0718-50062020000500233.

[75]   H. Almarabeh, "Analysis of students' performance by using different data mining classifiers," *Int. J. Mod. Educ. Comput. Sci.*, vol. 8, pp. 9-15, 2017. https://doi.org/10.5815/ijmecs.2017.08.02

[76]   X. J. Lin *et al.*, "Stress and its association with academic performance among dental undergraduate students in Fujian, China: A cross-sectional online questionnair survey," *BMC Med. Educ.*, vol. 20, art. 181, 2020. https://doi.org/10.1186/s12909-020-02095-4

[77]   T. Deliens, P. Clarys, I. de Bourdeaudhuij, and B. Deforche, "Weight, socio-demographics, and health behaviour related correlates of academic performance in first year university students," *Nutr. J.*, vol. 12, art. 162, 2013. https://doi.org/10.1186/1475-2891-12-162

[78]   E. T. Ortlieb and E. H. Cheek, "How geographic location plays a role within instruction: Venturing into both rural and urban elementary schools," *Educ. Res. Q.*, vol. 31, no. 2, pp. 48-64, 2008. https://www.proquest.com/docview/215932925

[79]   J. Cresswell and C. Underwood, "Location, location, location: Implications of geographic situation on australian student performance in PISA 2000," 2004. https://research.acer.edu.au/acer_monographs/2

[80]   A. Porto and L. Di Gresia, "Performance of University students and their determinants," 2005. [Online]. Available: http://sedici.unlp.edu.ar/bitstream/handle/10915/54674/Documento_completo__.pdf-PDFA.pdf?sequence=1

[81]   R. Garzón, M. O. Rojas, L. Del Riesgo, M. Pinzón, and A. L. Salamanca, "Factores que pueden influir en el rendimiento académico de estudiantes de bioquímica que ingresan en el programa de medicina de la Universidad

del Rosario-Colombia," *Educ. Médica*, vol. 13, no. 2, pp. 85-96, 2010. https://scielo.isciii.es/scielo.php?script=sci_abstract&pid=S1575-18132010000200005

[82] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Bus. Res.*, vol. 94, no. 2018, pp. 335-343, Feb. 2019. https://doi.org/10.1016/j.jbusres.2018.02.012

[83] A. Rico and D. Sánchez, "Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN/Design of a model to automate the prediction of academic performance in students of IPN," *RIDE Rev. Iberoam. para la Investig. y el Desarro. Educ.*, vol. 8, no. 16, pp. 246-266, 2018. https://doi.org/10.23913/ride.v8i16.340

[84] S. Bhutto, I. F. Siddiqui, Q. A. Arain, and M. Anwar, "Predicting students' academic performance through supervised Machine Learning," in *ICISCT 2020 - 2nd Int. Conf. Inf. Sci. Commun. Technol.*, Feb. 2020. [Online]. Available: https://doi.org/10.1109/ICISCT49550.2020.9080033

Leonardo Emiro Contreras-Bravo
Born in Sincelejo, Sucre, Colombia. He obtained a degree as mechanical engineer at Universidad Francisco de Paula Santander. He obtained a degree as Master of Engineering at Universidad Nacional de Colombia. PhD candidate in Engineering at Universidad Distrital Francisco José de Caldas. Member of the Dimsi research group (Diseño, modelamiento y simulación; Design, modeling, and simulation). He currently works as a professor in the Industrial Engineering program at the Department of Engineering of Universidad Distrital Francisco José de Caldas. His research interests are design, engineering education, and data analytics.
Email: lecontrerasb@udistrital.edu.co

Nayive Nieves-Pimiento
Born in Barrancabermeja, Santander, Colombia. She obtained a degree as mechanical engineer at Universidad Pontificia Bolivariana. She obtained a degree as Master of Environmental Sciences at Universidad Jorge Tadeo Lozano. PhD student in Natural Resources Engineering at Universidad de Oviedo (Spain). Member of the Environmental Management and Sustainable Development research group [Gestión ambiental y desarrollo sostenible]. She currently works as a professor in the Environmental Engineering program at the Department of Engineering of Universidad ECCI. Her research interests are design, sustainable engineering, and analytics.
Email: nnievesp@ecci.edu.co

Karolina González-Guerrero
Born in Colón, Putumayo, Colombia. She obtained an Education degree at Universidad Pedagógica Nacional. She obtained a Master's degree in Education at Pontificia Universidad Javeriana. She obtained a PhD degree in Education at Universidad Santo Tomás. She is currently a member of the PYDE research group and a full professor at Universidad Militar Nueva Granada. Her research interests are social sciences, education sciences and, general education (including training and pedagogy).
Email: karolina.gonzalez@unimilitar.edu.co