OASP-RITA-09-2021-0077.R1

A machine learning model to predict standardized tests in engineering programs in Colombia

Misorly Soto-Acevedo, Alfredo M. Abuchar-Curi, Rohemi A. Zuluaga-Ortiz, Enrique J. Delahoz-Domínguez*

Forecasting of Standardized Test Results for engineering students through Machine Learning

***Abstract*— This research develops a model to predict the results of Colombia's national standardized test for Engineering programs. The research made it possible to forecast each student's results and thus make decisions on reinforcement strategies to improve student performance. Therefore, a Learning Analytics approach based on three stages was developed: first, analysis and debugging of the database; second, multivariate analysis; and third, machine learning techniques. The results show an association between the performance levels in the Highschool test and the university test results. In addition, the machine learning algorithm that adequately fits the research problem is the Generalized Linear Network Model. For the training stage, the results of the model in Accuracy, AUC, Sensitivity, and Specificity were 0.810, 0.820, 0.813, and 0.827, respectively; in the evaluation stage, the results of the model in Accuracy, AUC, Sensitivity, and Specificity were 0.820, 0.820, 0.827 and 0.813 respectively.**

***Index Terms*— learning Analytics, Machine Learning, Predictive Evaluation, standardized tests.**

1. INTRODUCTION

Quality, when viewed as a process, can be objectively gauged through performance indicators. For instance, longitudinal analysis of standardized tests [1], or the interplay between economic variables, infrastructure, and academic outcomes [2] can offer valuable insights. Hence, to achieve educational quality, implementing ongoing self-assessment policies is necessary, aiming towards a continuous improvement process [3]. Thus, quality should be objectively evaluated. Internationally, one method of estimating quality in education is through the Accreditation Board of Engineering and Technology (ABET), a non-governmental, non-profit entity comprising technical and technological societies. These societies establish the policies of the process and accredit programs in applied sciences, computing, engineering, and engineering technologies both within and outside the United States [4]. Currently, numerous universities are undergoing not only national accreditation processes but also international ones, which requires measurements at different stages of the learning process and compliance with the demanded quality standards.

Standardized tests are the principal means used in Colombia and globally to measure academic achievement [5]. In Colombia, standardized tests measure the academic achievement of students in secondary education (Saber 11) and higher education (Saber Pro). The Saber Pro tests are not only applied as an indicator of excellence and quality for the academic programs of universities, but the national government, through the National Accreditation Council (CNA), also offers universities and their academic programs the possibility to undergo an accreditation process. This process is voluntary and seeks to promote quality improvement. It is also a way for these institutions to be accountable to society and the state regarding their educational service [6]. In the high-quality accreditation processes of the university or the program, a significant aspect is the outcome of the Saber Pro tests. It is crucial to achieve good results in these tests for those universities aspiring to achieve high-quality accreditation or maintain it.

The academic performance of students is a very complex issue and of great interest in the educational and investigative field. It is one of the biggest challenges facing educational institutions in basic and university education in Colombia and worldwide [8], [9]. Despite the various strategies of the National Ministry of Education for improving the quality of education in Colombia, the results in standardized evaluations at the middle and primary levels (Saber 11) indicate that there is still much to improve [7].

From a social perspective, standardized tests are an opportunity for students to demonstrate their knowledge level and access scholarships and school fee discounts with a good performance. Therefore, it would be important for a student to have objective information to effectively identify the competencies on which they should focus their study process to maximize their performance on the standardized test. Thus, the SaberPro test, in addition to being a tool for evaluating the quality of higher education in Colombia, becomes a tool of social mobility for students, associating high performance on the test with recognition in the form of economic benefits, reputation, and self-confidence.

For example, Timarán, Hidalgo, and Caicedo [10] in their research analyzed the variables of gender, age, monthly family income, type of school, score obtained on the Saber 11 tests and geographical area and their impact on the academic performance of Colombian students who took the

Manuscrito recibido el día de mes de año; revisado día de mes de año; aceptado día de mes de año.

English versión received Month, day-th, year. Revised Month, day-th, year. Accepted Month, day-th, year.

Misorly Soto-Acevedo, Facultad de Ciencia Básicas, Universidad Tecnológica de Bolívar, Cartagena, Colombia (e-mail: msoto@utb.edu.co)

Rohemi A. Zuluaga-Ortiz, Facultad Ingeniería, Universidad del Sinú, Cartagena, Colombia (e-mail: rohemi.zuluaga@unisinu.edu.co)

Alfredo M. Abuchar-Curi, Facultad de Ingeniería, Universidad Tecnológica de Bolívar, Cartagena, Colombia (e-mail: aabuchar@utb.edu.co)

*Enrique J. Delahoz-Dominguez, Department of Productivity and Innovation, Universidad de la Costa, Barranquilla, Colombia (e-mail: edelahoz13@cuc.edu.co)

Saber 11 tests in 2015 and 2016. They concluded that the models found, based on the data found in the ICFES databases, are consistent with observed reality. Therefore, it is pertinent to emphasize the importance of implementing objective models for decision-making in the educational field.

On the other hand, Pentel and Kaiva [11] conducted research in Estonia. The authors sought to predict the outcomes of student exams based on previous grades and demographic data, and to identify the most relevant subjects and features contributing to the outcome of the state examination. They considered variables such as gender, native language, and grades in some subjects and developed continuous and discrete models. For continuous models, they used linear regression, K-nearest neighbors, and random forests, while for discrete models they used logistic regression, K-nearest neighbors, C4.5, and random forests. The authors found several subjects that influenced the outcome of the examination. As expected, the most significant predictors included the subjects in which the examination was taken. Still, this was not always the case; some subjects had a surprising effect on some results of the state examinations, and some subjects had a strong negative impact on these.

Furthermore, Yang and Li [12] proposed a model to track student progress, forming a central component of e-Learning systems. They incorporated variables such as student grades in subjects and scores in learning skills using back-propagation neural networks. They concluded that the experimental results showed that the potential of the predicted progress can intuitively express the students' potential to progress in terms of their skills and performance. In addition, the experimental results also showed that the estimated student characteristics, the students' expected performance, and the causal relationship based on these attributes are correct. Significantly, the estimated performance based on grouping yielded more accurate results and took less time using a smaller amount of training data, which also approved that the results of the student classification are correct and meaningful.

For the current study, a database of the Saber 11 and Saber Pro test results from different students supplied by the Colombian Institute for the Evaluation of the Quality of Education (ICFES) is available. Consequently, the study's objective is to create a prediction model that allows identifying and classifying, based on the results of the Saber 11 tests, the socioeconomic variables, the university and the program selected by the students, the results of the Saber Pro tests in some Engineering programs in Colombia, using machine learning techniques.

Finally, to accomplish this objective, the work is divided into four sections: theoretical framework, methodology, research results, conclusions, and discussion.

I. Theoretical framework

This section presents the machine learning models used in the research. The machine learning models are presented from subsections A to G, and subsection H displays the evaluation metrics used to compare the models. It is important to note that the model selection follows the literature section. Lastly, this chapter contains eight subsections from A to H.

A. K Nearest Neighbours

The k-nearest neighbours algorithm (KNN) is a supervised machine learning model for classification or regression. This algorithm classifies by creating behaviour patterns to identify the belonging of an observation to a specific category. This is accomplished by calculating the distances between observations [13]. This analysis of distances is known as a proximity, similarity, or nearest point algorithm. In other words, given a new example, KNN finds its most similar examples, called nearest neighbours, based on a distance metric such as Euclidean distance and predicts its value by aggregating the target values associated with its nearest neighbours [14].

B. Generalized linear network model

A generalized linear model (GLMNET) is a flexible modification of linear regression because it allows the output variable to be a non-linear behaviour function of the input through an activation function [15]. Typically, GLMNET models use modifications from logistic regression, Poisson regression, and linear regression. Generally, GLMs are combined with other techniques for their high performance in various scenarios, such as object recognition, human action recognition, syntactic analysis, and automatic translation [16].

C. Random Forest

Random Forests (RF) is a supervised machine learning model widely used for classification [17]. This algorithm utilizes decision trees to create multiple responses and then classifies according to the most frequent response [18]. The Random Forests model has two fundamental parameters: the number of trees (k) and the number of variables used to split the nodes (m).

D. Support Vector Machine

Support Vector Machine models (SVM) are binary classifiers that utilize the kernel function and are based on machine learning theory [19]. This algorithm aims to identify the optimal separating hyperplane that can classify different observations. This method is widely used for data that is correlated and non-linear.

E. Naïve Bayes

The Naïve Bayes (NB) model is based on Bayes' theorem with assumptions of independence among the predictors. This theory allows for the calculation of the probability of P(a|b), from P(a), P(b), and P(b|a). This classifier assumes that the effect of the dependent variable b's value on a given class a is independent of the values of other dependent variables [20]. This assumption is called class conditional independence. Due to the simplicity of the model, it is widely used and often delivers surprisingly better results than those of more sophisticated algorithms.

F. Decision Trees

The Decision Tree model (DT) is one of the most widely used supervised algorithms in machine learning today. It is a non-parametric method used for classification and regression [21]. This algorithm aims to create a model that seeks to estimate the value or category of a response

ISSN 2255-5706 © IEEE-ES (Capítulo Español)

variable through simple decision rules derived from the characteristics of the data.

G. Boosting

Boosting algorithms focus on powerful and sophisticated predictions that are made from a single model. These algorithms seek to enhance predictive power by training a sequence of weak models, with the final model drawing upon each lesson learned by the individual models [22]. This algorithm is also known as a generic, non-specific algorithm, making it crucial to define the base model (such as DT, GLMNET, NB, among others), which is then improved upon. This study will apply Boosting to the generalized linear model (GLMBOOSTING).

H. Performance metrics

Machine learning models for classification are evaluated using performance metrics derived from a confusion matrix. This matrix compares predicted values against known actual values (see Table I). Hence, from the confusion matrix, we extract the following [23]:

- True Positive (TP) occurs when an observation is predicted as positive and is indeed positive.
- True Negative (TN) occurs when an observation is predicted as negative and is indeed negative.
- False Positive (FP) occurs when an observation is predicted as positive but is actually negative.
- False Negative (FN) occurs when an observation is predicted as negative but is actually positive.

TABLE I
MATRIZ DE CONFUSIÓN.

Predicción \ Real	1	0
1	VP	FP
0	FN	VN

From the confusion matrix, the Accuracy metric is derived. This performance metric is the percentage of correct predictions for the test data. It is calculated as the ratio of correct predictions to the total number of predictions (see Equation (1)).

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FV} \quad (1)$$

Another performance indicator of the model is the area under the curve (AUC) of the receiver operating characteristics (ROC). The ROC curve presents the true positive rate (sensitivity) as a function of the false positive rate (specificity). The ROC value maximizes the sensitivity and specificity values simultaneously. The area under the curve metric is used for binary classification problems and is one of the most widely used. The AUC value of a model will be approximately equal to the probability that the model classifies a randomly chosen positive example better than a randomly chosen negative example.

On the other hand, the true positive rate (sensitivity) is defined as the ratio between True Positives over the sum of False Negatives and True Positives. And finally, the true negative rate (specificity) is defined as the ratio between True Negatives over the sum of False Positives and True Negatives.

II. METHODOLOGY

The present research is framed under an approach of quantitative data analysis, in the area known as Learning Analytics, which promotes the use of data generated at different levels of the process, in this case educational, to create tools that support objective decision-making for different groups interested in the process: students, teachers, educational managers, governmental entities, accreditation entities, among others. The research is divided into three stages: Principal Component Analysis (PCA), model training, and finally, model evaluation.

Therefore, it can be concluded that this research is of an applied type considering that it seeks to infer, through machine learning, the results of students in the Saber Pro test based on the results obtained in the Saber 11 tests and some specific aspects of the selected university.

It is essential to point out that the software used for the data analysis, construction, and evaluation of the models is R [26].

A. Descripción de la base de datos

For the research, there is a database provided by the Colombian Institute for the Evaluation of the Quality of Education (ICFES) corresponding to the Saber Pro 2018 tests and the corresponding previous results of Saber 11 [27], which includes the results of 12410 students from different engineering programs at the national level. The database includes students from different universities nationwide. Table II shows the selected programs, the number of students registered in the database, and information on whether the program is accredited.

Additionally, it is essential to note that in the database designed by Delahoz-Dominguez et al. [27], the information used corresponds to the variables from the Saber 11 tests, the socioeconomic variables, the university, and the program selected by the students (see Table IV).

Finally, before the development of the research stages, the database is prepared while maintaining the fidelity of the information to improve the performance of the models. This conditioning included the creation of categories for each variable to reduce the variability of the information presented by variables.

TABLE II
ENGINEERING PROGRAMS IN THE STUDY

Programa	Number of total students	Accredited
Civil engineering	3320	No
Electrical engineering	278	No
Chemical engineering	1001	No
Electronic Engineering	849	Si
Industrial Engineering	5318	Si
Mechanical engineering	1136	Si
Mechatronics Engineering	78	Si

III. RESULTS

This chapter is divided into three stages according to what is presented in the methodology chapter: Principal

Component Analysis, model training, and model evaluation.

A. *Stage 1: Principal Component Analysis (PCA)*

The Principal Component Analysis (PCA) is carried out with an orthogonal rotation (maximum variance rotation). In Fig. 1, the biplot graph is presented, and levels of the global score in the Saber Pro tests observe a clear differentiation. Also, a direct relationship between the results of the different areas evaluated in the Saber Pro tests and the performance levels defined in the global score is clearly observed. Similarly, in the fi

rst quadrant, a strong relationship can be seen between students who had good results in the Saber 11 tests and the performance levels of the Saber Pro test.

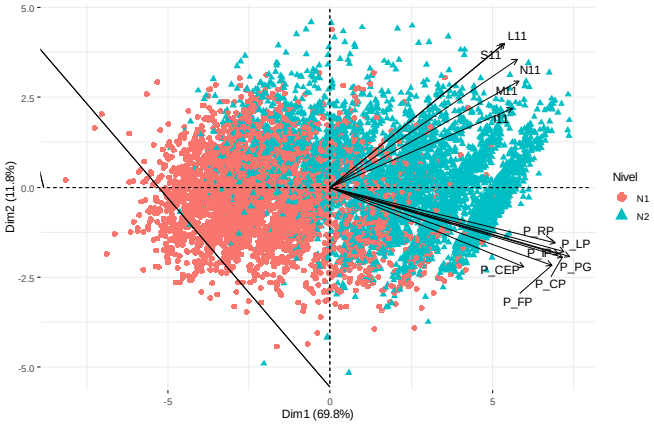

Fig. 1: PCA Analysis. Dimensions 1-2

Consequently, Table III presents the distribution of observations from the 1-2 plane of the PCA analysis. It is observed that students with higher academic performance are in greater proportion in quadrants one and four, with approximately 74.4% of students who are at level two. On the other hand, in quadrants two and three, approximately 83.1% of students are at level one.

TABLE III
DISTRIBUTION OF OBSERVATIONS IN THE PCA PLANE 1-2

Nivel	Quadrant			
	I	II	III	IV
1	5.1%	36.6%	46.5%	11.9%
2	37.5%	20.2%	5.4%	36.9%

Finally, in the first exploratory stage of the database, Figure 2 presents information on the importance of the variables.

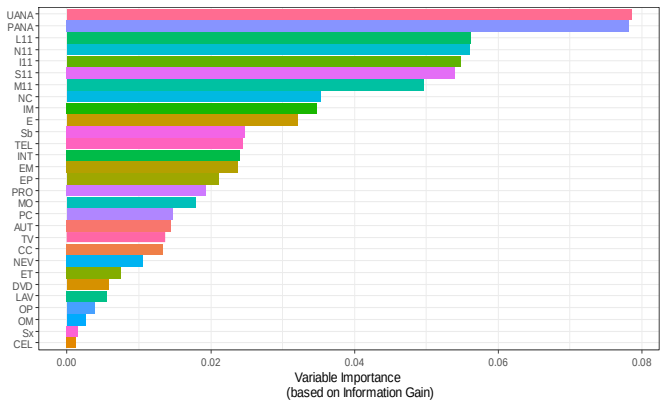

Fig. 2: Importance of variables.

Dentro de los resultados presentados en la Fig. 2 se puede evidenciar que para la construcción de un modelo que sea capaz de predecir el resultado de las pruebas Saber PRO es importante tener en cuenta la acreditación de la universidad, del programa académico y los resultados del estudiante en su evaluación de Saber 11.

B. *Stage 2: Model Training*

For model training, the cross-validation method was used and we worked with a training base corresponding to 70% of the data, leaving 3300 in N1 and 3576 in N2, and a base for model evaluation corresponding to 30% of the data, leaving 1453 in N1 and 1494 in N2. Since a very even distribution is observed, the data are not balanced. Additionally, the configuration of the different models is presented in Table V.

Thus, for the parameterization of the KNN model, the k value indicates the appropriate number of groups for the dataset. For the parameterization of the GLMBOOST model, the Mstop value indicates the number of iterations the model will perform.

For the parameterization of the GLMNET model, the value of the elastic net penalty is controlled by Alpha, while the lambda value controls the overall strength of the model's penalty; in Table V, this model presents an Alpha value equal to 0.55 and lambda equal to 0.0057, this indicates the configuration of a model with lasso penalty.

Thus, for the parameterization of the RF model, the Mtry value indicates the number of variables to split at each node; the splitrule value indicates the mode for estimating probability in classification, and finally, min.node.size indicates the minimum node size.

For the parameterization of the SVM model, the Sigma value acts as a smoothing parameter and the cost value (C) controls the complexity of the boundary between the support vectors. For the parameterization of the NB model, the Laplace value indicates a type of additive smoothing for the model (if it is zero, there is no smoothing) and, in addition, the use of the kernel (whether adjusted or not). Finally, for the parameterization of the DT model, the complexity factor is required, and this decreases the overall lack of adjustment by a factor according to the determined value.

ISSN 2255-5706 © IEEE-ES (Capítulo Español)

TABLE IV
VARIABLES

Variables	Código	Clase	Categorías
Gender	Sx	Cat	Femenino (F); Masculino (M)
Parent Education	EP	Cat	Complete Professional Education (EPC); Incomplete professional education (EPI); None (n); Don't know (ns); Postgraduate (post); Complete primary (pc); Incomplete primary (pi); Secondary complete (bc); Secondary incomplete (bi); Complete technical or technological (TC); Incomplete technical or technological (it)
Mother Education	EM	Cat	
Parent Occupation	OP	Cat	Employee with position as director or general manager (G); Auxiliary or administrative level employee (A); Managerial level employee (D); Employee of technical or professional level (T); Employee worker or operator (EO); Entrepreneur (E); Household (H); Other activity or occupation (OA); Pensioner (P); Small Business Owner (PE); Independent professional (I); Self-employed (CP)
Mother Occupation	OM	Cat	
Stratum	E	Cat	Stratum 1 (E1); Stratum 2 (E2); Stratum 3 (E3); Stratum 4 (E4); Stratum 5 (E5); Stratum 6 (E6)
Sisben	Sb	Cat	It is classified at another level of SISBEN (SO); Level 1 (S1); Level 2 (S2); Level 3 (S3); It is not classified by SISBEN (SN)
People home	Per	Num	One; Two; Three; Four; Five; Six; Seven; Eight; Nine; Ten; Eleven; Twelve or more
Type of floor	TP	Cat	Cement, gravel, brick; Rough wood, board, plank; Polished wood, tile, tablet, marble, carpet; Earth, sand
Family has internet	INT	Cat	Yes; No
Family has TV service	TV	Cat	Yes; No
Family has a computer	PC	Cat	Yes; No
Family has washing machine	LAV	Cat	Yes; No
Family has microwave	MO	Cat	Yes; No
Family has a car	AUT	Cat	Yes; No
Familia tiene DVD	DVD	Cat	Yes; No
Family has fridge	NEV	Cat	Yes; No
Family has cell phone	CEL	Cat	Yes; No
Family has a phone	TEL	Cat	Yes; No
Monthly household income	IM	Cat	Less than 1 SMLV; Between 1 and less than 2 SIDL; Between 2 and less than 3 SMLV; Between 3 and less than 5 SMLV; Between 5 and less than 7 SMLV; Between 7 and less than 10 SMLV; 10 or more SMLV
Currently studying or working	ET	Cat	No; Yes, 2 hours or more a week; Yes, less than 2 hours a week
School name	N.C	Cat	
Funding of the school	NC	Cat	Unofficial (NO); Official (OF)
Character of the school	CC	Cat	Academic (ACÁ); Not applicable (NA); Technician (TEC); Technical/Academic(T/A)
Saber11 Math Score	M11	Num	Range: 0-100
Saber11 Critical Reading Score	L11	Num	Range: 0-100
Saber 11 Social and citizenships	S11	Num	Range: 0-100
Puntaje Ciencias Naturales Saber11	N11	Num	Range: 0-100
English Score Saber11	I11	Num	Range: 0-100
University Name	UN	Cat	
Saber Pro Quantitative Reasoning Score	RP	Num	Range: 0-100
Puntaje Lectura Crítica Saber Pro	LP	Num	Range: 0-100
Saber Pro Citizenship Skills Score	CP	Num	Range: 0-100
English SaberPro Score	IP	Num	Range: 0-100
SaberPro Written Communication Score	CEP	Num	Range: 0-100
SaberPro Global Score	PG	Num	Range: 0-300
SaberPro Formulation of engineering programs	FP	Num	Range: 0-300
Academic Program	PRO	Cat	Civil Constructions (CCI); Aeronautical Engineering (AER); Cadastral Engineering and Geodesy (CYG); Civil Engineering (CIV); Control Engineering (NOC); Production Engineering (PRO); Productivity and Quality Engineering (PYC); Transport and Roads Engineering (TYV); Electrical Engineering (ELE); Electromechanical Engineering (ELM); Electronic Engineering (ETR); Electronic and Telecommunications Engineering (ETRT); Industrial Automation Engineering (AUTI); Automation Engineering (IAU); Control Engineering (NOC); Industrial Control and Automation Engineering (CYA); Industrial Engineering (IND); Mechanical Engineering (MEC); Mechatronics Engineering (MTR); Chemical Engineering (QUI); Topographic Engineering (TOP)
Accredited university	UANA	Cat	UA (Accredited University); UNA (Unaccredited University)
Accredited Program	PANA	Cat	PA (Accredited Program); PNA (Non-accredited Program)

Note. The categories for the Class column are: Numeric and Cat.

TABLE V
MODEL'S PARAMETRIZATION

Model	Parameters
KNN	k = 43
GMLBOOST	Mstop = 250
GLMNET	Alpha = 0.55; lambda = 0.0057
RF	Mtry = 7; splitrule = extratrees; min.node.size = 1
SVM	Sigma = 0.11; c = 0.25
NB	Laplace = 0; Kernel ajustado
DT	Cp = 0.0018

Figure 3 shows the comparative results of the models used in the research, and here, the Generalized Linear Model in Net (GLMNET) and the GLMBOOST model deliver the best results.

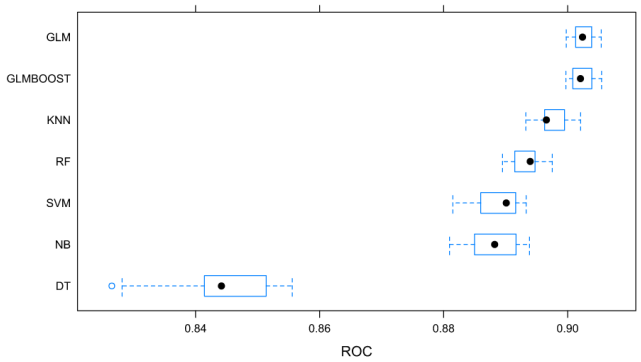

Fig. 3: Comparison between applied models

As observed in Figure 4, the GLMNET model exhibits lower variability in the box and whisker plot of cross-validation results compared to the GLMBOOST model. Additionally, the GLMNET model has lower computational cost due to its lower complexity than the GLMBOOST model. For these reasons, the GLMNET model is considered the best model for predicting the performance level in the Saber Pro exams.

Table VI presents the results of the model training stage. The Generalized Linear Model in Net (GLMNET) algorithm achieves the highest performance in terms of AUC, while the Decision Tree (DT) algorithm performs the worst. The GLMNET model achieves an Accuracy, AUC, Sensitivity, and Specificity of 0.810, 0.820, 0.813, and 0.827, respectively, while the DT model achieves 0.770, 0.803, 0.837, and 0.770 in these metrics.

TABLE VI
PERFORMANCE METRICS FOR TRAINING DATA

	Accuracy		AUC		Sensitivity		Specificity	
Model	Mean	sd	Mean	sd	Mean	sd	Mean	sd
KNN	0.810	0.015	0.810	0.014	0.813	0.015	0.807	0.016
GLMNET	0.810	0.015	0.820	0.014	0.813	0.016	0.827	0.017
RF	0.810	0.015	0.813	0.014	0.815	0.020	0.812	0.019
SVM	0.810	0.014	0.816	0.014	0.825	0.017	0.807	0.017
NB	0.790	0.015	0.790	0.015	0.742	0.014	0.837	0.014
DT	0.770	0.015	0.803	0.014	0.837	0.031	0.770	0.028
GLMBOOST	0.810	0.014	0.818	0.014	0.808	0.014	0.829	0.016

C. *Stage 2: Evaluation phase*

Finally, the last phase of the methodology involves evaluating the trained models to assess their performance in predicting new observations. Accordingly, Table VII presents the performance of the models in the evaluation stage. Once again, the Generalized Linear Model in Net (GLMNET) outperforms the other models regarding predictive capacity, while the Naïve Bayes (NB) model shows lower performance in this stage. The GLMNET model achieves an Accuracy, AUC, Sensitivity, and Specificity of 0.820, 0.820, 0.827, and 0.813, respectively, while the NB model achieves 0.790, 0.792, 0.937, and 0.742 in these metrics.

TABLE VII
PERFORMANCE METRICS FOR TEST DATA

	Accuracy		AUC		Sensitivity		Specificity	
Model	Mean	sd	Mean	sd	Mean	sd	Mean	sd
KNN	0.810	0.015	0.810	0.014	0.807	0.013	0.813	0.014
GLMNET	0.820	0.014	0.820	0.014	0.827	0.012	0.813	0.013
RF	0.813	0.015	0.813	0.014	0.812	0.013	0.815	0.170
SVM	0.816	0.014	0.817	0.014	0.807	0.012	0.825	0.013
NB	0.790	0.015	0.792	0.015	0.837	0.014	0.742	0.014
DT	0.803	0.015	0.805	0.014	0.770	0.013	0.837	0.013
GLMBOOST	0.818	0.014	0.818	0.014	0.829	0.012	0.801	0.013

Indeed, as observed in Figure 4 and Figure 5, the performance of the models is illustrated through the ROC curve. In these figures, it is evident that the Generalized Linear Model in Net (GLMNET) has a larger area under the curve, indicating a more significant predictive capacity than the other models.

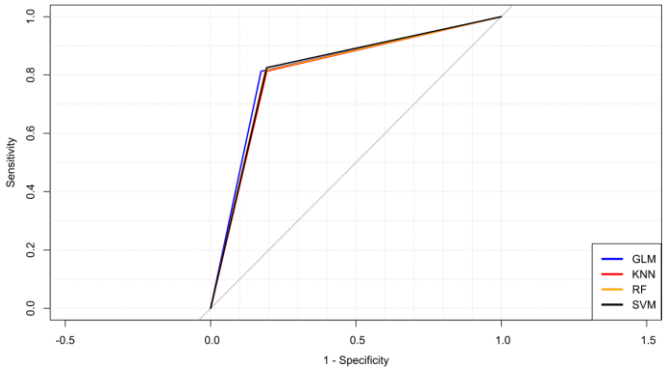

Fig. 4: ROC curve for GLMNET, KNN, RF and SVM models

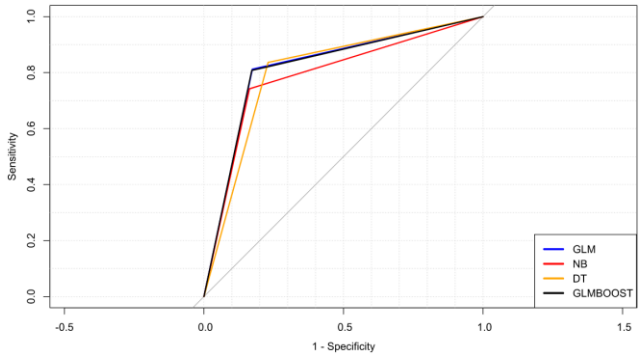

Fig. 5: ROC curve for GLMNET, NB, DT and GLMBOOST models

IV. DISCUSSION

This section presents a comparative interpretation of the main findings associated with developing a predictive model for the standardized SaberPro tests in Colombia. The results have implications beyond fitting the data to a supervised learning model and would serve as a tool for managing university resources. Early identification of student's future academic performance would allow for efficient management of educational resources.

ISSN 2255-5706 © IEEE-ES (Capítulo Español)

Firstly, the multivariable descriptive analysis through PCA reveals differences in mathematics results between students with excellent and poor performance in the SaberPro tests. Various reasons can explain these differences. i) Differences in geographical location: Previous studies have shown that regional heterogeneity is a significant source of regional inequality in emerging countries [28]. ii) Differences between private and public universities [29].

Secondly, the analysis of variable importance in generating the classification of the machine learning model indicates that university accreditation is the most critical factor. These results are consistent with the research by [30], where a logistic regression model was used to classify engineering programs in Colombia using a linear regression model.

Thirdly, regarding the ability of machine learning models to predict performance results used in our research, the accuracy and AUC results were 82%. Compared with the work by Kaur et al. [31], who built a Neural Network model to estimate students' academic performance using economic, family, and educational variables, achieving an accuracy of 75%. Similarly, Jishan et al. [32] developed a model using Naïve Bayes and Neural Networks algorithms with purely academic variables, obtaining an AUC value of 81%. Furthermore, Lau et al. [33] developed a Neural Network model using 11 input variables, two hidden layers, and one output layer, employing the Levenberg-Marquardt algorithm as the back-propagation training rule, achieving an accuracy of 84.8%. Therefore, the results of our research related to the predictive capacity of the implemented machine learning models are competitive and consistent with studies with similar approaches in education. However, comparing the models goes beyond the figures of performance metrics. From a methodological perspective, it is essential to note that the mentioned research only used variables from the academic context, generating a bias by associating student performance with a single context. On the contrary, socioeconomic variables were considered in our research, assimilating the learning process as the interaction of multiple variables in the student's environment.

Finally, from an applied perspective, our research contributes to the spectrum of knowledge of predictive models that are useful for managing higher education institutions. However, this does not limit the possibility of extrapolating the methodology to other academic levels and areas of knowledge. Considering the above, it is vital to note that education, as the main driver of development for societies, needs tools that can identify deficiencies in students' learning process. The proposed tool will not only support predicting students' academic performance levels but also identify the critical variables that provide the most information to explain academic performance. This will enable interventions in various aspects that are opportunities for improvement (socioeconomic conditions, academic weaknesses, among other variables considered in the study).

V. CONCLUSION

In the present research, a database of the results of the Saber 11 and Saber Pro tests from different students was used, provided by ICFES, to create a machine learning model for predicting the academic performance of students entering a university. The research allows for forecasting the results of each student and, thus, making decisions on reinforcement strategies to improve the results in the Saber Pro tests. Ultimately, the proposed model enables the identification of groups of students who may have low performance in the Saber Pro tests in order to create strategies throughout the teaching-learning process for these students and thereby improve their academic training, enhance the results of the Saber Pro tests, and boost their professional and career performance.

Finally, the study's main limitation is the potential bias due to only including engineering programs. However, this methodology can be applied to other areas of study. In this regard, future research can consider its application to other fields of study and an analysis of fuzzy inputs of the variables to observe changes in the response variable in a broader and more dynamic spectrum of inputs.

ACKNOWLEDGEMENT

Special thanks to the Colombian Institute for the Evaluation of Education Quality for providing the data used in this research.

REFERENCES

[1]J. Aparicio, S. Perelman, y D. Santín, «Comparing the evolution of productivity and performance gaps in education systems through DEA: an application to Latin American countries», *Oper. Res.*, jun. 2020, doi: 10.1007/s12351-020-00578-2.

[2]D. Visbal-Cadavid, M. Martínez-Gómez, y F. Guijarro, «Assessing the efficiency of public universities through DEA. A case study», *Sustainability*, vol. 9, n.º 8, p. 1416, 2017.

[3]M. Campo, «Capital humano para el avance colombiano, Editorial en Educación superior 20». p. 1, 2012.

[4]L. Valencia, H. Trefftz, y I. Delgado-González, «Acreditación Internacional de Carreras de Ingeniería», *Educ. En Ing.*, vol. 15, n.º 29, pp. 28-33, 2020.

[5]R. Hoyos Martínez, M. Borja Maturana, R. Gómez Lorduy, y G. Casadiegos Aponte, «Calidad en la escuela vs. prácticas pedagógicas: los relatos como medio para la reflexión y la emancipación de los maestros en tiempos de la eficiencia», *Esfera*, vol. 5, n.º 2, p. 16, 2015.

[6]J. Guerrero, «La acreditación de alta calidad en Colombia», 2018.

[7]L. A. Sanabria James, M. C. Pérez Almagro, y L. E. Riascos Hinestroza, «Pruebas de evaluación Saber y PISA en la Educación Obligatoria de Colombia», *Educ. Siglo XXI*, vol. 38, n.º 3 Nov-Feb, pp. 231-254, 2020, doi: 10.6018/educatio.452891.

[8]L. A. Melo-Becerra, J. E. Ramos-Forero, y P. O. Hernández-Santamaría, «La educación superior en Colombia: situación actual y análisis de eficiencia», *Desarro. Soc.*, vol. 2017, n.º 78, pp. 59-111, 2017, doi: 10.13043/DYS.78.2.

[9]Y. Bernal y C. Rodríguez, «Factores que Inciden en el Rendimiento Escolar de los Estudiantes de la Educación Básica Secundaria», Universidad Cooperativa de Colombia, 2017.

[10] R. Timarán-Pereira, J. Caicedo-Zambrano, y A. Hidalgo-Troya, «Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°», *Rev. Investig. Desarro. E Innov.*, vol. 9, n.º 2, pp. 363-378, 2019, doi: 10.19053/20278306.v9.n2.2019.9184.

[11] A. Pentel y L. L. Kaiva, «Predicting Students' State Examination Results based on Previous Grades and Demographics», *11th Int. Conf. Inf. Intell. Syst. Appl. IISA 2020*, 2020, doi: 10.1109/IISA50023.2020.9284401.

[12] F. Yang y F. W. B. Li, «Study on student performance estimation, student progress analysis, and student potential prediction based on data mining», *Comput. Educ.*, vol. 123, n.º April, pp. 97-108, 2018, doi: 10.1016/j.compedu.2018.04.006.

[13] S. Zhang, X. Li, M. Zong, X. Zhu, y R. Wang, «Efficient kNN Classification With Different Numbers of Nearest Neighbors», *IEEE Trans.*

Neural Netw. Learn. Syst., vol. 29, n.º 5, pp. 1774-1785, may 2018, doi: 10.1109/TNNLS.2017.2673241.
[14] A. Moldagulova y R. Bte. Sulaiman, «Using KNN algorithm for classification of textual documents», en *2017 8th International Conference on Information Technology (ICIT)*, Amman, Jordan, may 2017, pp. 665-671. doi: 10.1109/ICITECH.2017.8079924.
[15] P. K. Dunn y G. K. Smyth, «Chapter 5: Generalized Linear Models: Structure», en *Generalized Linear Models With Examples in R*, P. K. Dunn y G. K. Smyth, Eds. New York, NY: Springer, 2018, pp. 211-241. doi: 10.1007/978-1-4419-0118-7_5.
[16] D. Zhang, «A Coefficient of Determination for Generalized Linear Models», *Am. Stat.*, vol. 71, n.º 4, pp. 310-316, oct. 2017, doi: 10.1080/00031305.2016.1256839.
[17] E. De La Hoz, R. Zuluaga, y A. Mendoza, «Assessing and Classification of Academic Efficiency in Engineering Teaching Programs», *J. Effic. Responsib. Educ. Sci.*, vol. 14, n.º 1, Art. n.º 1, mar. 2021, doi: 10.7160/eriesj.2021.140104.
[18] G. Louppe, «Understanding Random Forests: From Theory to Practice», *ArXiv14077502 Stat*, jul. 2014, Accedido: 23 de julio de 2019. [En línea]. Disponible en: http://arxiv.org/abs/1407.7502
[19] S. Suthaharan, «Support Vector Machine», en *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, S. Suthaharan, Ed. Boston, MA: Springer US, 2016, pp. 207-235. doi: 10.1007/978-1-4899-7641-3_9.
[20] D. Buzic y J. Dobsa, «Lyrics classification using Naive Bayes», en *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, may 2018, pp. 1011-1015. doi: 10.23919/MIPRO.2018.8400185.
[21] K. David Kolo, S. Adepoju, y J. Kolo Alhassan, «A Decision Tree Approach for Predicting Students Academic Performance», *Int. J. Educ. Manag. Eng.*, vol. 5, n.º 5, pp. 12-19, oct. 2015, doi: 10.5815/ijeme.2015.05.02.
[22] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, San Francisco, California, USA, 2016, pp. 785-794. doi: 10.1145/2939672.2939785.
[23] A. J. Hung, J. Chen, y I. S. Gill, «Automated Performance Metrics and Machine Learning Algorithms to Measure Surgeon Performance and Anticipate Clinical Outcomes in Robotic Surgery», *JAMA Surg.*, vol. 153, n.º 8, p. 770, ago. 2018, doi: 10.1001/jamasurg.2018.1512.
[24] Z. H. Hoo, J. Candlish, y D. Teare, «What is an ROC curve?», *Emerg. Med. J.*, vol. 34, n.º 6, pp. 357-359, jun. 2017, doi: 10.1136/emermed-2017-206735.
[25] D. Gašević, V. Kovanović, y S. Joksimović, «Piecing the learning analytics puzzle: a consolidated model of a field of research and practice», *Learn. Res. Pract.*, vol. 3, n.º 1, pp. 63-78, 2017, doi: 10.1080/23735082.2017.1286142.
[26] R Core Team, *R: A language and environment for statistical computing, R Foundation for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013. Accedido: 15 de abril de 2020. [En línea]. Disponible en: http://www.polsci.wvu.edu/duval/PS603/Notes/R/fullrefman.pdf
[27] E. Delahoz-Dominguez, R. Zuluaga, y T. Fontalvo-Herrera, «Dataset of academic performance evolution for engineering students», *Data Brief*, vol. 30, p. 105537, jun. 2020, doi: 10.1016/j.dib.2020.105537.
[28] P. Herrera-Idárraga, E. López-Bazo, y E. Motellón, «Regional Wage Gaps, Education and Informality in an Emerging Country: The Case of Colombia», *Spat. Econ. Anal.*, vol. 11, n.º 4, pp. 432-456, oct. 2016, doi: 10.1080/17421772.2016.1190462.
[29] J. Moreno-Gómez, J. Calleja-Blanco, y G. Moreno-Gómez, «Measuring the efficiency of the Colombian higher education system: a two-stage approach», *Int. J. Educ. Manag.*, vol. 34, n.º 4, pp. 794-804, ene. 2020, doi: 10.1108/IJEM-07-2019-0236.
[30] E. J. Delahoz-Dominguez, S. Guillen-Ibarra, T. Fontalvo-Herrera, E. J. Delahoz-Dominguez, S. Guillen-Ibarra, y T. Fontalvo-Herrera, «Análisis de la acreditación de calidad en programas de ingeniería industrial y los resultados en las pruebas nacionales estandarizadas, en Colombia», *Form. Univ.*, vol. 13, n.º 1, pp. 127-134, feb. 2020, doi: 10.4067/S0718-50062020000100127.
[31] P. Kaur, M. Singh, y G. S. Josan, «Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector», *Procedia Comput. Sci.*, vol. 57, pp. 500-508, 2015, doi: 10.1016/j.procs.2015.07.372.
[32] S. T. Jishan, R. I. Rashu, N. Haque, y R. M. Rahman, «Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique», *Decis. Anal.*, vol. 2, n.º 1, p. 1, dic. 2015, doi: 10.1186/s40165-014-0010-2.
[33] E. T. Lau, L. Sun, y Q. Yang, «Modelling, prediction and classification of student academic performance using artificial neural networks», *SN Appl. Sci.*, vol. 1, n.º 9, p. 982, ago. 2019, doi: 10.1007/s42452-019-0884-7.

Misorly Soto Acevedo es ingeniera de Sistemas, Politécnico Gran Colombiano. Especialista en Estadística Aplicada, Universidad del Norte. Magister en Estadística Aplicada, Universidad Tecnológica de Bolívar. Profesora catedrática en la Facultad de Ciencias Básicas de la Universidad Tecnológica de Bolívar.

Alfredo Miguel Abuchar Curi es Ingeniero mecánico, Universidad Tecnológica de Bolívar (UTB). Magister en Ingeniería mecánica, universidad de los Andes. Magister en Estadística Aplicada, Universidad Tecnológica de Bolívar. Profesor de Tiempo Completo de la UTB con más de 25 años de experiencia. Actualmente secretario de la Facultad de Ingeniería de la UTB. Su área de investigación es mecánica de fluidos.

Rohemi Alfredo Zuluaga Ortiz es magister en ingeniería de la Universidad Tecnológica de Bolívar (UTB). Actualmente es profesor de la Universidad del Sinú. Sus áreas de investigación son eficiencia y Learning Analytics.

Enrique De La Hoz máster en Investigación de Operaciones por la Universitat de Barcelona y la UPC. Actualmente es profesor de la Universidad de la Costa en el departamento de productividad e Innovación. Sus áreas de investigación son Learning Analytics, sistemas de recomendación y minería de datos a gran escala.