Validation Machine Learning Models To Predict Score on Graduate Tests based on High School Test and other Factors, Case Study: Colombia.

Maryori Sabalza Mejia, Electronic Engineer[1], Carolina Campillo Jimenez, MSc[2], and Juan Carlos Martinez Santos, PhD[1]

[1]*Universidad Tecnologia de Bolívar, Colombia, msabalzam@utb.edu.co, jcmartinezs@utb.edu.co*

[2]*Universidad Nacional Abierta y a Distancia, Colombia, carolina.campillo@unad.edu.co*

***Abstract*– In Colombia, the state usually administers tests to evaluate the knowledge learned during high school and university. This test is the Saber 11 Test, and it applies at the end of high school. These tests are an indispensable requirement in admissions for the university. Students must take the Saber Pro Test as a grade requirement at the end of said studies, which assesses university quality.**

However, many of the students who performed well on the Saber 11 tests may fail or even never take the Saber Pro Test because many drop out before finishing their degree. Many conditions may affect, but the student of the socio-economic conditions is one of them.

This research shows the validation of machine learning models to predict the Saber Pro Test results based on the results of the Saber 11 test according to a range. This range was a maximum period of five years, considering socio-economic variables that remained constant during this time. Two models were verified that comply with a 100% prediction with the real value, and by the stacking model, the prediction values are correct up to 80.41%.

Keywords-- State tests, Education, High School, College, Machine Learning, Prediction, Colombia.

I. Introduction

The Colombian government for some years implemented the Saber 11 tests as a mandatory measure to enter the university. To measure the quality of education and how prepared the student is. However, as it could determine from these results within four to five years, high grades in the Saber Pro are carried out at the end of undergraduate studies. This research takes the socio-economic details of the students who have presented it, such as socio-economic status, residential area, a computer at home, and the mother's educational level, because these are fixed variables within these four to five years.

The possibility of predicting the Saber Pro Test results will allow universities to implement an early warning system so that from the entrance of the student, the university can take a program of reinforcements and an intense study to improve the probability of good results.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2021.1.1.343
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

The databases with the Saber 11 and Saber Pro results are freely available on the Colombian Institute for the Promotion of Higher Education-ICFES (for its Spanish acronym) [1] website, so some important ones have trained and applied models for different studies using these databases.

This paper is organized as follows. Section II shows the works before this research are mentioned. Section III describes the methodology of the tests in Colombia. Section IV presents the approach of this research. Section V is about the methods and artifacts. Section VI explains each detail of the phases of development. Section VII shows the results obtained. Finally, section IX contains the conclusions of this research.

II. Related Work

The most useful technique to apply before any more complicated method is linear regression. In the case of [2] used this technique to affirm the premise that there are no significant gender differences in the test results Saber Pro and Saber 11.

However, in modules linked more to the sciences, the male genre is favored. In the verbal ability modules, the female genre predominates, mentioning that socio-economic factors
strongly influence the Saber Pro.

As mentioned in [3], there may be some performance patterns in tests such as Saber 11 and Saber Pro that can be evidenced by studying the students' results in a specific period. One of the most used techniques for the database of the Saber Pro is K-Nearest Neighbors (KNN) technique looks for observations that are closer to the predicted value, [4] use KNN to find the "K" closest records in the prediction of students studying engineering. Also, [5] used to predict efficiency in higher education institutions. This research also applied Random Forest, principal component analysis (PCA), vector support machines (SVM), clustering, and K-Means.

Those techniques helped determine the type of students according to the study area, for the test base on correlation, the Euclidean distance between the data. The selection of the factors can influence the results. Decision Tree (DTs) is a technique that makes a series of decisions in the form of a tree, being the final nodes of the prediction. In [6], the sought for the case was the method allowed to predict associated factors such as gender, age, socio-economic status, salary, educational level of the mother and father, and everything conditions the students' lives. That is, what are the socio-economic, academic, and institutional factors that can influence a low or high score in the Saber 11 test. For [7], it is also essential to study the socio-economic variables with the results of the Saber 11 test with a Decision tree to predict which student will have the highest score in the Saber Pro.

The Decision Tree technique can use the cross-validation technique to estimate how accurate the model is. In the case of [8] and [9], they used cross-validation with ten partitions (Kfolds = 10) for the discovery of performance patterns that can influence the English tests of the Saber Pro. The confidence was 60%, and for percentages that can affect the critical reading module results of the Saber Pro with a maximum confidence value of 25%. [10] uses decision tree and random forest techniques to predict the success in the Saber Pro, taking into account variables such as reading comprehension, English, technological deficiencies, and results of previous exams. The decision tree used was CART, and they obtained a precision in the model no greater than 75%.

Another technique used for the Saber Pro was Random forest regression [11], taking into account socio-economic variables such as obtained average root mean square error (RMSE) values of 27,58. The random forest helped to analyze the quality accreditation of the industrial engineering program in Colombia by [12]. Comparing the Saber Pro Test results of each student shows that accredited universities as better results than not accredited ones.

On the other hand, [13] shows the correlation between Saber 11 and Saber Pro for students of systems engineering at the San Gil University (UNISANGIL) Colombia for a specific period of six years. However, the study did not consider socio-economic variables. The data analysis opted to give reinforcement in specific areas and improve the preparation of the Saber Pro.

According to this research mentioned before, we can conclude that there is little prediction information with the database of Saber 11 and Saber Pro's results in a period and evaluating the socio-economic conditions. So we a study of the Saber Pro test prediction system taking the socioeconomic conditions into account. It is vital to mention that the jobs, as mentioned above, do not use the stacking model for comparison as opposed to this research.

III. STATE TESTS

In Colombia, students have to complete the Saber 11 test to register to the university. Since 1960 this test has changed of name, a topic covered, application protocol, but the objective remains; evaluate and compare secondary academic institutions' quality. However, until the 80s, the exam was formalized, and with the [14]. It was declared a mandatory requirement. ICFES applies the ICFES and recognizes a list of foreign exams that allow validation abroad [15]. Today, this test is known as Saber 11. Its objective is to measure the students' skills at the end of elementary, High School precisely in the last year and is an indispensable requirement to enter the university.

This test is similar to the exam taken in China: Gaokao [16] and in the United States: The Suite of Assessments (SAT) developed by the College Board [17].

The test Saber 11 contains 278 questions and covers Mathematics, Critical Reading, Natural, Social and Citizen Sciences, and English, which require two sessions. On the other hand, the Saber Pro is applied to students of the last academic semester of undergraduate with compulsory character to evaluate the competencies obtained in their institution of higher education.

This test is similar to the Graduate Record Examination (GRE) [18] applied in the United States. This test is a requirement for people who wish to do postgraduate studies, whether it master's or a doctorate. The first session evaluates Critical reading, Quantitative reasoning, Citizen competencies, Written communication, and English. The second session assesses the specific competencies of the training.

According to the Colombian National Minister of Education, the curriculum must contain almost 12 academic subjects. Then Saber 11 Test evaluates five categories of knowledge and their possible relation with Saber Pro categories, such as mathematics. We can relate them to quantitative reasoning since the latter implies everything: to calculate, reason, issue, and make decisions. The test also evaluates social and citizen competencies. Critical reading and English are part of both Saber 11 and Saber Pro. On the other hand, the natural sciences could be one of the topics of deepening of the Saber Pro, such as biology, but this depends on its career selection.

IV. OUR APPROACH

The National higher education system database in public and private universities are 2'050.616 young people but only graduated 482.122 [19].

From this relationship of these variables, our research looked to design a Saber Pro Test prediction system based on the results obtained in the Saber 11 tests almost presented almost five years before. To verify whether they influence the results considering the economic variables that do not vary.

For this, our primary metric will be the "root mean square error" (RMSE). This measure is the distance between the prediction and the actual value, which the smaller will guarantee a predictive accuracy of the model. A perfect prediction model will have an RMSE equal to zero; therefore, it must not exceed one's value.

Besides, we will also take the value of the standard deviation in each model into account to know how dispersed the data is, knowing that the smaller the standard deviation is, the lower its dispersion will be and vice versa.

R squared value is also considered for each model because it is the coefficient of determination that will tell us how good the model’s fit is. This value should range between zero and one.

V. METHODS AND ARTIFACTS

For this research, we obtain the database from the official ICFES website. We took the results obtained in 2018 from the Saber Pro. We related them using each participant’s official code, with the results obtained in the Saber 11 test carried out in the years 2013-2014, according to the estimated time range of four to five years.

It is important to note that the results of both tests belong to the same students, and it did not consider that people worked or not while they were carrying out their university studies.

A. Preparation Data

The variables to study from the database are the results of the Saber Pro as output variables. As input the scores of the Saber 11 of each student, a variable selection method was applied to determine which socioeconomic variables factors could be maintained in the time between both tests and influence the results of the Saber Pro, taking into account only these factors.

The variable selection method applied was Chi-Square,which is used for categorical variables in data-sets [20] and follows the equation (1).

$$X^2 = \frac{\left(\text{observed}+\text{expected}\right)2}{expected} \quad (1)$$

Table I shows the top ten (0-10) variables obtained according to the Chi-Square method, the input and output variables of the tests, and description according to the dictionary of variables of the saber tests [21].

B. Models

The analysis of the regressor models is a technique applied to find the relationship between a dependent and independent variable according to the machine learning library, Scikit-Learn [22]. The regression is a prediction between a continuous value associated with an object. Due to the importance of this technique and data type, the development of this research, we used the Python [23], with the help of the Scikit-Learn library [24] we tested the following models.

1) K-Nearest Neighbors: K-nearest neighbors (KNN) [25], An algorithm based on regression of the closest neighbors with the objective that the close points contribute to the regression of the most distant ones.

This output ŷ is the predict values of k-nearest-neighbors can be interpreted in regression [26] with the equation (2).

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i\left(x\right) \quad (2)$$

TABLE I
VARIABLES AND DESCRIPTION

Variable	Description
SCORE-MATH-11	Score obtained in Math Saber 11 test
SCORE-CRITICAL-READING-11	Score obtained in Critical Reading Saber 11 test
SCORE-SOCIAL-AND-CITIZEN-11	Score obtained in Social and citizen Saber 11 test
SCORE-NATURAL-SCIENCE-11	Score obtained in Natural Science Saber 11 test
SCORE-QUANTITATIVE-REASONING-PRO	Score obtained in Quantitative Reasoning Saber Pro test
SCORE-CRITICAL-READING-PRO	Score obtained in Critical Reading Saber Pro test
SCORE-CITIZEN-COMPETENCIES-11	Score obtained in Citizen Competencies Saber Pro test
SCORE-ENGLISH-11	Score obtained in English Saber Pro test
SCORE-GLOBAL-PRO	Score obtained in Global Score Saber Pro test
STUDENT-GENDER	Gender: Female and Male
FAMILY-NUMBER-OF-PERSONS-CHARGE	Number of people who depend financially of the student
FAMILY-FATHER-EDUCATION	Highest educational level achieved by father
FAMILY-MOTHER-EDUCATION	Highest educational level achieved by mother
FAMILY-FATHER-JOB	Father’s occupation or job did for most of the last year
FAMILY-MOTHER-JOB	Mother’s occupation or job did for most of the last year
FAMILY-HOUSING-STRATUM	Socioeconomic status according electric bill
FAMILY-HOME-ROOMS	Number of bedrooms where people in their household sleep
FAMILY-MOTORCYCLE	Have a motorcycle in their home
FAMILY-QUOTES-SHARE-BATHROOM	How many people share the bathroom of the home
FAMILY-NUMBER-BOOKS	Number of books there have at home

2) Decision Tree: Decision Tree (DTs) [27], An algorithm that creates a predictive model by learning simple decision rules inferred from the characteristics of the data. For regression, it is important to determine locations for future splits are Mean Squared Error, Poisson deviance, and Mean Absolute Error. The equations according to this model are: (3), (4), (5), (6), (7).

Mean Squared Error:

$$y_m = \frac{1}{N_m}\sum_{i \in N_m} y_i \quad (3)$$

$$H\left(X_m\right) = \frac{1}{N_m}\sum_{i \in N_m}\left(y_i - y_m\right) \quad (4)$$

Half Poisson deviance:

$$H\left(Q_m\right) = \frac{1}{N_m}\sum_{y \in Q_m} ¿¿¿ \quad (5)$$

Mean Absolute Error:

$$median\left(y\right)_m = median_{i \in N_m}\left(y_i\right) \quad (6)$$

$$H\left(X_m\right) = \frac{1}{N_m}\sum_{i \in N_m}\left|y_i - \left(y\right)_m\right| \quad (7)$$

3) Random Forest: Random Forest (RF)[28], Algorithm meta estimator based on a decision tree, each tree depends, and the values of an independently tested random vector is the same distribution for each one.

The tree predictor h(X) take numerical values, and the output too [29], The mean-squared generalization error is equation (8).

$$E_{X.Y}\left(Y+h\left(X\right)\right)^2 \quad (8)$$

As random forest reaches infinity, it will be true that [29]:

$$E_{X.Y}\left(Y-av_k h\left(X,\theta_k\right)\right)^2 \Rightarrow E_{X.Y}\left(Y-E_\theta h\left(X,\theta_k\right)\right)^2$$
(9)

4) XGBoost: The Xgboost library provides a gradient boosting, based on a generalization of boosting to arbitrary differentiable loss functions model [30]. Each stage fits a regression tree to the gradient of the given loss function.

The equation (10) is based on the following main equation that will be minimized to Euclidean domain [31]:

$$\Im^{(t)}=\sum_{i=1}^{n} l\left(y_i, y_i^{(t-1)}+f_t\left(X_i\right)\right)+\Omega\left(f_t\right) \quad (10)$$

The final minimized equation, eliminating the constant parts is equation (11).

$$\Im^{(t)}=\sum_{i=1}^{n}\left[g_i f_t\left(X_i\right)+\frac{1}{2}h_i f_t^2\left(X_i\right)\right]+\Omega\left(f_t\right) \quad (11)$$

5) ElasticNet: ElasticNet [32], the linear regression model with combined L1 and L2 priors as regularizer. See equation (12).

$$\min_{w} \frac{1}{2\,n_{samples}}\left\|X_w-y\right\|_2^2+\alpha\rho\left\|w\right\|1+\frac{\alpha(1-\rho)}{2}\left\|w\right\|_2^2$$
(12)

6) Stacking: The stacking method to ensemble consists of stacking an individual estimator's output and using a regressor to compute the final prediction [33].

This method has the advantage of two or more models incorporated in a test, which is known as an assembly of models, adds cross validation, and shows the result for this project root-mean-square parameter or R2.

A partition defines the stacked, and this is known as "Stacked generalization" [34] (See equation (13)).

$$R^{n+1} \rightarrow \theta\left\{\theta_{i1},\theta_{i2}\right\} \quad (13)$$

B. Interquartile Range

Interquartile range (IQR) is the distance between the third and first quartile relative to the median value, according to [35] the middle half of the data-set is specifies by IQR and the variability among the data.

Fig. 1 shows a box-plot with whiskers to global score variable to determine the interquartile range using the established formula (14).

$$IQR=Q_3-Q_1 \quad (14)$$

The median inside the box suggests that the data are skewed, and the point outside represents a separate peak in the upper quartile.

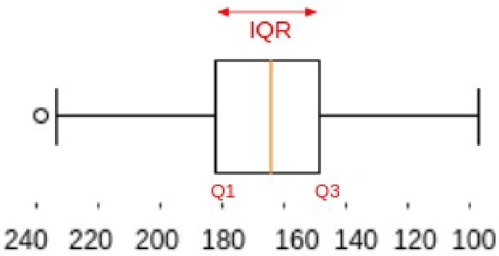

Fig. 1 Box Plot Global Score

VI. IMPLEMENTATION

The research paper was a review that has used the Saber Pro database and determines what models they used and if they took into account the socioeconomic variables. The approach of this research was divided into three phases, as shown in Fig. 2.

A. Phase 1: Exploration and Preparation Data

The first phase consisted of data exploring, analyzing all the variables that could help us in our model, and we created a new database with the variables to study. Our input variables will always be the Saber 11 test results: Mathematics, Critical Reading, Natural Sciences, Social and citizen, and English.

B. Phase 2: Train Saber 11 and Saber Pro

This phase included the Saber Pro of each student variable as output: Global score, Quantitative Reasoning, Critical Reading, Citizen Competencies, and English. Machine learning models were applied to predict each of these output variables. For the train/test, we divide the data according to the Pareto 80-20 principle [36].

C. Phase 3: Saber 11, Saber Pro and Socioeconomic Variables

At this stage, we use four socioeconomic variables from the Saber Pro test database for the model to provide us with information that could influence the Saber 11 test result so

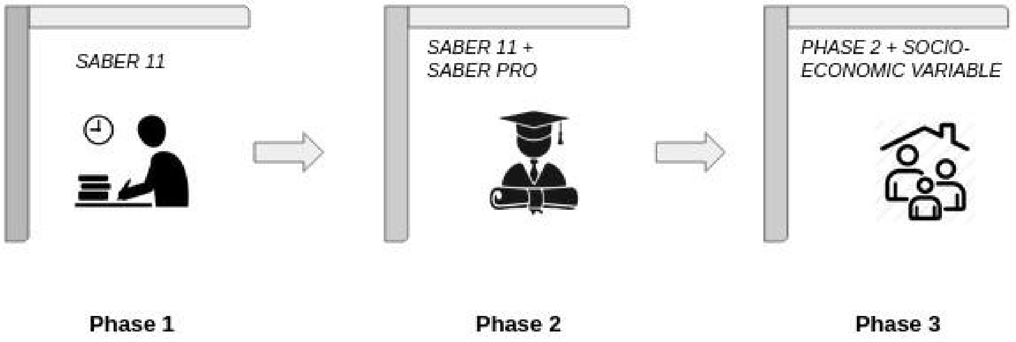

Fig. 2 Exploration and Data Preparation

much. We took these variables as output, and we made the model analysis with each one.

After performing these three phases, we validated the results, and we established a range to determine which models were within them and could predict the test results per student.
We obtained this value from the interquartile range.

VII. RESULTS

Machine learning regression models as KNN, RF, Dts, ElasticNet, and XGBoost are applied to predict the Saber Pro test according to the Saber 11 test results.

The models trained according to the Pareto principle [36], so train: 80% and test: 20%, However, for validation data, we made a new division data performed according to Table II.

TABLE II

Pareto

Data	Number	Percentage
Train	7816	70%
Test	2236	20%
Validate	1117	10%

After dividing the data we established a prediction range. For this, we take the results of the global score and plot them (See Fig. 1).

The following step was to establish a range of values from the real value of the variables obtained and compare with the value predicted by each of the implemented models. This range was obtained from the interquartile range according to the global score values per student, as it shows in Fig. 1.

The next step was to obtain the interquartile range: [146.,162., 180.].

The distance between the mean (162) and the first quartile-Q1: 180 is 18.

The distance between the mean (162) and the third quartile-Q3: 146 is 16.

We average the distances obtained the result for the range is +/- 17.

However, this value is still high for the prediction of real value. We must emphasize that the Saber Pro tests have a maximum score of 300, so we decided to reduce the range by almost half and use +/- 10 points.

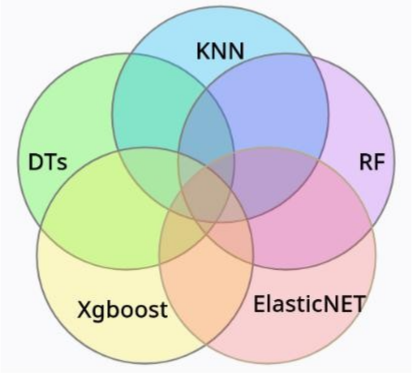

Fig 3. Voting System Based on Venn Diagram

The next step to check a voting system, based on the Venn diagram to five sets (See Fig 3). We seek to validate how many models match within the established range.

Our voting system works by validating which models meet the established range for each variable. If all the models coincide, a score of five points was assigned, in the case of four, four points, in the case of three, three points, and the case of two, finally two points. In case of not coinciding with any model, no points were assigned.

Table III shows the results of our voting system.

We verify that for each variable, the five models coincide in the prediction, specifically an average of 495,2.

However, there are two models that mostly match the range, on average for more than half of the total validated data. Specifically an average of 568,6.

TABLE III

Voting System

Variable	5 Points	4 Points	3 Points	2 Points
Global Score	510	62	204	341
Quantitative Reasoning	633	0	0	484
Critical Reading	401	0	0	716
Citizen Competencies	313	0	0	804
English	619	0	0	498
Total Average Points	**495.2**	**62**	**204**	**568.6**

TABLE IV

Average Per Model

Variable	KNN	DTs	RF	ElasticNet	Xgboost	Stacking	Average per Variable
Global Score	100%	100%	51.21%	51.21%	63.92%	92.93%	76.5%
Quantitative Reasoning	100%	100%	56.67%	56.67%	56.67%	85.41%	75.9%
Critical Reading	100%	100%	35.90%	35.90%	35.90%	73.95%	53.2%
Citizen Competencies	100%	100%	28.02%	28.02%	28.02%	68.83%	58.0%
English	100%	100%	55.42%	55.42%	55.42%	80.93%	74.5%
Total Average Models	**100%**	**100%**	**45.44%**	**45.44%**	**47.98%**	**80.41%**	

Due to these scores in the voting system, we opted to implement a stacking model with the models used. Also, we verify with each of the models to know which are the two models that always match within the range.

After implementing our stacking model, we verify one by one the models (now six) that were within the range established for each output variable.

Table IV shows the results average obtained score Saber Pro according to each model and total average.

Our maximum prediction value was in the global score, where we obtained 76.5 % compared to the real values and the lowest for critical reading with 53.2%.

We got 100% prediction hits in KNN and DTs models. For stacking model we got 80.41% and for the other models that were validated in this research (RF, ElasticNet, Xgboost), we obtained a percentage no greater than 50%.

Regarding the metrics we obtained for KNN: RMSE: 20.15, Standard Deviation: 23,542 and R-square: 0.27.

For Dts: RMSE: 20.81, Standard Deviation: 23.796 and R-square: 0.22.

Regarding the metrics that we obtain in our stacking model: RMSE: 14.76, Standard Deviation: 21,135 and R-square: 0.61.

We show how the R-square improves using the stacking model.

VIII. DISCUSSION

For this research, we did not take the type of study (Full time-Part time-Job+Study) of the students. We not considered the modality of a job in the day and study at night or viceversa. We matched the information data of Saber 11 and Saber Pro, per student, and addition the information about the top ten socioeconomic variables.

IX. CONCLUSIONS AND FUTURE WORK

This paper shows a validation machine learning models to predict score on graduate test based on high school test and socio-economic variables: Gender, educational level of the father and mother, mother’s occupation, father’s occupation, socioeconomic status, number of rooms at home, number of books, number of people with their share bathroom and motorcycle.

The first validation was our voting system, which worked with the assignment of points for the five proposed models (KNN, DTs, RF, Xgboost, ElasticNet).

Then, we proposed a model basing on the stacking model for validation of models by one.

We verify that the stacking model can predict up to 80.41% of the results of the Saber Pro. However, simplemodels such as KNN and DTs can be almost 100% accurate. We verify all this through a voting system.

The trained model will propose that areas should deepen students' performance and guarantee good results in the Saber Pro test's time frame for future work.

Acknowledgment

We would like to thank the anonymous reviewers of LACCEI for their comments and feedback on the ideas in this paper and the Universidad Tecnológica de Bolívar for their support.

References

[1] I. C. para el Fomento de la Educación Superior, Educación superior y desarrollo. El Instituto, 1983, vol. 2.

[2] [2] V. Cantillo and L. García, "Gender and other factors influencing the outcome of a test to assess quality of education in civil engineering in colombia," Journal of Professional Issues in Engineering Education and Practice, vol. 140, no. 2, p. 04013012, 2014.

[3] L. M. G. BELTRÁN, "Patrones de desempeño en pruebas estandarizadas y de calidad en instituciones de educación superior: Evidencia basada en datos a partir de resultados individuales en pruebas saber 11 y pruebas saber pro."

[4] A. I. O. Carrascal and J. J. Giraldo, "Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas saber-pro," Revista Politécnica, vol. 15, no. 29, pp. 128–140, 2019.

[5] D. Visbal Cadavid, A. Mendoza Mendoza, and S. J. Orjuela Pedraza, "Predicción de la eficiencia de las instituciones de educación superior colombianas con análisis envolvente de datos y minería de datos," Pensamiento & Gestión, no. 42, pp. 140–161, 2017.

[6] R. Timarán-Pereira, J. Caicedo-Zambrano, and A. Hidalgo-Troya, "Decision trees for predicting factors associated with academic performance of high school students in saber 11 tests," Revista de Investigación, Desarrollo e Innovación, vol. 9, no. 2, pp. 363–378, 2019.

[7] G. P. Bernal, L. T. Villegas, and M. Toro, "Saber pro success prediction model using decision tree based learning," arXiv preprint arXiv:2006.01322, 2020.

[8] R. T. PEREIRA, N. San Juan de Pasto, C. A. H. TROYA, and C. J. C. ZAMBRANO, "Proceso de descubrimiento de patrones de desempeño académico en la competencia de inglés con crisp-dm," 2016.

[9] I. H. Arteaga and J. C. Alvarado, "Descubrimiento de patrones de desempeño académico en la competencia de lectura crítica."

[10] E. Berdugo, D. Bustos, J. González, A. Palacio, J. Pérez, and E. Rocha, "Identification and comparison of factors associated with performance in the saber pro and saber tyt exams for the period 2016-2019," 2020.

[11] E. Berdugo, D. Bustos, J. González, A. Palacio, J. Pérez, and E. Rocha, "Identification and comparison of factors associated with performance in the saber pro and saber tyt exams for the period 2016-2019," 2020.

[12] E. J. Delahoz-Dominguez, S. Guillen-Ibarra, and T. Fontalvo-Herrera, "Análisis de la acreditación de calidad en programas de ingeniería industrial y los resultados en las pruebas nacionales estandarizadas, en colombia," Formación universitaria, vol. 13, no. 1, pp. 127–134, 2020.

[13] A. R. P. Blanco and L. Y. C. Chacón, "Correlación de los resultados de las pruebas icfes–saber 11 y saber pro de los estudiantes del programa de ingeniería de sistemas, sede san gil unisangil, periodo 2012-2017," Revista Matices Tecnológicos, vol. 12, pp. 34–39, 2020.

[14] Colombia, Ley general de la educación: Ley 30 de 1992: diciembre 28, Ley 107 de 1994: enero 7, Ley 115 de 1994: febrero 8, Decreto 1860 de 1994: agosto 3, Decreto 1857 de 1994: agosto 3. Ministerio de Educación Nacional, 1994.

[15] U. P. Javeriana, "Exámenes extranjeros reconocidos por el icfes," 1993.

[16]A. Muthanna and G. Sang, "Undergraduate chinese students' perspectives on gaokao examination: Strengths, weaknesses, and implications," International Journal of Research Studies in Education, vol. 5, no. 2, pp. 3–12, 2015.

[17] C. M. Fuess, "The college board; its first fifty years." 1950.

[18] G. A. Schaeffer, C. M. Reese, M. Steffen, R. L. McKinley, and C. N. Mills, "Field test of a computer-based gre general test," ETS Res Rep Ser, vol. 1993, pp. I–47, 1993.

[19] S. N. de Informacion de la Educacion Superior, "Estadisticas," 2018.

[20] G. for Geeks, "Ml — chi-square test for feature selection," 2020.

[21] ICFES, "Documentación y diccionarios saber pro," 2019.

[22] Scikit-Learn, "sklearn regression," 2020.

[23] Python.org, "Python," 2020.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python," the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[25] Z. Lateef, "Knn algorithm: A practical implementation of knn algorithm in r," 2019.

[26] Y. Song, J. Liang, J. Lu, and X. Zhao, "An efficient instance selection algorithm for k nearest neighbor regression," Neurocomputing, vol. 251, pp. 26–34, 2017.

[27] T. Point, "R-decision tree," 2019.

[28] P. Analytics, "How to implement random forests in r," 2018.

[29] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[30] xgboost.readthedocs, "Xgboost r tutorial," 2019.

[31] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[32] Scikit-Learn, "sklearn linear model elasticnet," 2020.

[33] "Sklearn ensemble stackingregressor," 2020.

[34] D. H. Wolpert, "Stacked generalization," Neural networks, vol. 5, no. 2, pp. 241–259, 1992.

[35] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, A Modern Introduction to Probability and Statistics: Understanding why and how. Springer Science & Business Media, 2005.

[36] B. C. Arnold, "Pareto distribution," Wiley StatsRef: Statistics Reference Online, pp. 1–10, 2014.