International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 6, December 2021, pp. 5402~5410
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i6.pp5402-5410

Forecasting model with machine learning in higher education ICFES exams

Daniel Esteban Martínez Cervera[1], Octavio José Salcedo Parra[2], Marco Antonio Aguilera Prado[3]
[1,2]Faculty of Engineering, Universidad Distrital Franciso José de Caldas, Bogotá, Colombia
[2]Faculty of Engineering, Universidad Nacional de Colombia, Bogotá, Colombia
[3]Vice-presidency forr Research, Universitaria Agustiniana, Bogotá, Colombia

Article Info

Article history:

Received Sep 8, 2020
Revised Apr 9, 2021
Accepted May 11, 2021

Keywords:

Forecast models
K-closest neighbor
K-means
Machine learning
Naïve bayes
Neural network
Prediction

ABSTRACT

In this paper, we proposed to make different forecasting models in the University education through the algorithms K-means, K-closest neighbor, neural network, and naïve Bayes, which apply to specific exams of engineering, licensed and scientific mathematical thinking in Saber Pro of Colombia. ICFES Saber Pro is an exam required for the degree of all students who carry out undergraduate programs in higher education. The Colombian government regulated this exam in 2009 in the decree 3963 intending to verify the development of competencies, knowledge level, and quality of the programs and institutions. The objective is to use data to convert into information, search patterns, and select the best variables and harness the potential of data (average 650.000 data per semester). The study has found that the combination of features was: women have greater participation (68%) in Mathematics, Engineering, and Teaching careers, the urban area continues to be the preferred place to apply for higher studies (94%), Internet use increased by 50% in the last year, the support of the family nucleus is still relevant for the support in the formation of the children.

Corresponding Author:

Daniel Martínez
Faculty of Engineering
Universidad Distrital Franciso José de Caldas
Cra. 7 #40b-53, Bogotá, Colombia
Email: demartinezc@correo.udistrital.edu.co

1. INTRODUCTION

Education is one of the factors that has manifested exceptional changes in our country. It has had a notorious impact on society because it has helped to decrease poverty, improve quality of life, support innovation, and industrialization, among others. Many authors like Hanushek [1], Coleman [2], Barrera [3], and in Finland [4], have studied this topic. They have demonstrated that knowledge can be affected by different variables that allow improving the cognitive abilities of students: their way of being, their level of emotions, and economic aspects, among others [5].

Colombian Ministry of National Education approved decree 3963 in 2009. The purpose of the decree is to make an instrument for the external evaluation of quality in higher education. Likewise, with the tests the respective indicators are obtained, both for the student and for the institution of Higher Education, seeking to improve the policies, studies and regulations proposed by the government [6].

The amount of information stored about the students who take the Saber Pro Tests constitutes a great challenge for identifying the factors that have the greatest influence when obtaining good results.

However, details such as behavior patterns, academic background, family, or economic group are not presented. The above mentioned may lead to a bias in the results for improving effectiveness in higher education [6]. There are even studies conducted in Bogotá on both the process of getting to university and the careers and the strong relationship between external variables and student grades [7]-[9], however, of all the information that ICFES presents, the most important characteristic that affects education over time, will be the efficacy with which higher education can improve and how much information bias in databases affects the data acquisition.

With the above, the article tries to verify the achievement of university education quality through the use of Scrum methodologies, the knowledge discovery in databases (KDD), process of evaluation of a large amount of data to obtain intrinsic or extrinsic information) [10], supervised algorithms (K-neighbors plus nearby, neural networks, naive Bayes) and unsupervised (K-Means) in machine learning, to obtain patterns, verify which model obtains better evaluations and determine which variables stand out.

The main data source is a Colombian Institute for the Evaluation of Education (ICFES) FTP server that supports the Colombian Ministry of Education to present the tests, it stores all the information about each student every semester, with average information of 650,000 thousand students. The models used obtain data from 2017 to 2019 (Results and information external to the University) and each algorithm is designed to analyze each data and store it in Data frames, validate biases, data crossing, transformation, training, test, results, and selection. This information seeks to support the institutions, ICFES, learning processes.

2. RESEARCH METHOD

2.1. Research activities

Due to the new challenges and problems presented in engineering, the methodology, the follow-up to carry out the prediction model [11], and the search for information was taken from ICFES database. This information included the students' background, family characteristics, school, and test results. Following the above, the modules to take into account for the development of the model to be evaluated are economic module, software design, teaching, engineering projects, and mathematical and scientific thinking. Planning is one of the highest costs in the development of Engineering processes, due to the construction of the project, the development of the team, related areas, and the process of extracting and transforming data until reaching the prediction model [12]. As shown in Table 1, the dataset corresponds to the classifying the results ICFES network, along with the student's history and the scope of data from 2017 to 2019 [13]:

Table 1. Variables of the ICFES database

Consecutive Number - allows traceability	Economic situation
Gender	Kind of job (Mother and Father)
Department	Daily time spent reading
Education (Father, Mother)	Daily time spent using the internet
Social stratum	Name of school
Family members	Bilingual school
Home appliances (computer, washing machine, microwave, tv, car, motorcycle, videogames)	Type of school
Quantity of books owned	School area (Urban, rural)
School hours	Percentile (economic, software design, teaching, education, engineering projects and Mathematical and Scientific Thinking.

The selection range that is analyzed corresponds to the period between February 15, 2019, and may 30, 2020 (data have been combined annually since 2017). This gives better coverage of information, separating the sprints in:

a. Information gathering
b. Check data concordance
c. Anomalous data
d. Attribute selection to perform the Prediction Model and data transformation
e. Use of each algorithm’s mathematical models to separate independent -dependent variables
f. Application and results [14].

All of the above is based on the knowledge discovery in databases (KDD) methodology, which allows to extract knowledge from large amounts of information in a database [10]. Four algorithms are applied to the procedure, three supervised type [15], and one of them unsupervised type, because the prediction models have been proposed to improve the quality of the software and one of the popular ways is through machine learning [16]. This was built through Python with the Spyder application, considering the following:

a. Saber Pro's general and Saber Pro's specific traceability key, to identify performance, unknown values and change over time.
b. Separation of training variables by 70% and 30% for tests.
c. The variable Y stores the class attribute and the variable X takes the attributes that will be analyzed.

The variable selection was made through the documents with the needed information acquired from the application of the use of information and the FTP server (for transferring files), based on the studies carried out by several authors such as Hanushek [1], Colemann [2], Barrera [3] and in Finland [4] and the articles [8], [9], [17]-[19] that have shown that the results of standardized tests go beyond knowledge, characterizing variables such as student's financial capacity, identity, learning, objects they have in their home and expressions of their personality in addition to determining other variables that may be affecting the student, either at family or educational level, among others. The variables are shown as follows: Socioeconomic level, parent's education, family, parents' educational level, average overall score, academic results in critical reading, mathematics, natural and social sciences, English, and two sub-tests for quantitative reasoning.

2.2. Algorithms

2.2.1. K-Closest neighbors

It is one of the top 10 in data mining techniques, it is known for assigning similar labels to the class. based on the observations, and the collected training, the relationship between them is distinguished [20], along with unknown examples in "neighboring" classifications. Through what has been said, it looks for an optimal training set, so that the prediction can identify data from the smallest ones [21]. One of the methods for calculating the distance is the KNN () function, which uses the Euclidean distance:

$$D(p,q) = \sqrt{(\rho_1 - q_1)^2(\rho_2 - q_2)^2 + \cdots + (\rho_n - q_n)^2} \quad (1)$$

where p and q are the distance between two points that are related by n characteristics, until obtaining a strong amount of close data. The characteristics were divided by annual data (more than a million records) and the selected variables (24 variables) according to the contributions made by the investigated authors, thereby defining 12 neighbors for the analysis.

$$n\_neigbors = 12 \quad (2)$$

$$knn = KNeighborsClassifier(n\_neighbors) \quad (3)$$

The K-neighbors classifier function has:

a. Distance calculated with the Euclidean function
b. A uniform weight, that is, all the points in each neighborhood that weigh the same

$$knn.fit(X\_train, y\_train.values.ravel()) \quad (4)$$

The "fit" estimator adjusts the data to be able to predict the classes to which they are related. Once adjusted, the prediction with the predict () function is applied for the training data with the test, and the additional use of the confusion matrix, in order to verify the accuracy of the model and thus evaluate the effectiveness.

$$print(confusion\_matrix(y\_test, prediction)) \quad (5)$$

Use of the confusion matrix next to the variable that has the prediction of the data stored.

2.2.2. K-means

It is one of the algorithms that dates back to the middle of the last century, which is based on the grouping of values with a minimum distance in a group, taking into account the following points:

a. Definition of clusters.
b. K proposed groupings, represented by centroids (the mean location of the points in the group) $(x_1, x_2, x_3, \ldots x_n \in \mathbb{R})$.
c. Random initialization of values for each cluster.
d. Repeat steps a, b and c until the data convergence [22], [23].
e. Based on the above, the process behind K-means is applied

Based on what was mentioned above, the process that is behind K-means is applied.

$$\min_{\{m_k\},1\le k\le K} \sum_{k=1}^{k} \sum_{x\in c_k} \pi_x dist(x, m_k) \quad (6)$$

Where π_x is the weight of the objects of x, m_k which is the data assigned to each cluster. The cluster is equivalent to the variable k and c_k is the centroid of each cluster, until the convergence of the clusters [24].

In the development process in Python, a sequence of numbers is taken randomly to apply the transformation of each cluster

$$Nc = range(1, len(X)) \quad (7)$$

Taking into account that x, m_k are the data, the variable X is taken because it stores the variables to be evaluated and the variable Y is taken as a class to identify the relationship with each cluster;

$$kmeans = [KMeans(n\_clusters = i)\ for\ i\ in\ Nc] \quad (8)$$

The amount of Cluster in the variable n_clusters start together with the cycle of the set of selected random variables until the conversion is carried out, and while this process is being carried out, the data transformation is applied to determine the average fit of the data.

$$score = [kmeans[i].fit(X).score(X)\ for\ i\ in\ range(len(kmeans))] \quad (9)$$

The process is stored in the score variable, where the fit function fits the data. While the score (X) function calculates the sub-precision of the data subset. At the end, the K-Means execution, together with the data transformations, is applied to the Centroids to be able to visualize the relationship of the data, thanks to the cluster_centers_ function to generate the centroids, which is the prediction of the model verify its accuracy.

$$kmeans = KMeans(n\_clusters = i).fit(X) \quad (10)$$

$$centroids = kmeans.cluster\_centers\_ \quad (11)$$

$$labels = kmeans.predict(X) \quad (12)$$

2.2.3. Naive Bayes

It is a learning algorithm based on Bayesian rules, with a strong assumption of the attributes that are conditionally independent of the class, offering a high precision in the classification algorithms, to estimate the probability P (X | Y) in which each class is Y given the objects X, the classification is carried out [25], [26].

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (13)$$

One of the ways to apply naive Bayes is through the Gaussian model being the main source, where X_i is taken as the amount of data, Y the class, the parameters σ_y and μ_y (mean of the variable X which is associated with the class Y) are estimated with the maximum probability, and together with a normal distribution [27].

In the prediction model, the previous formula is applied through the function gnb=GaussianNB (), which takes the data and the class (X | Y) taking into account factors in the formula such as the classification, the maximum likelihood and the amount of information. However, in order to reach this process, the best characteristics are chosen best = SelectKBest (k=14); In this case, 60% of the variables are taken, so that the characteristics that are selected can classify the data in a uniform way.

$$X\_new = best.fit\_transform(X, y.values.ravel()) \quad (14)$$

$$selected = best.get\_support(indices = True) \quad (15)$$

The fit_transform () function fits the data and transforms it into an array to be able to use get_support (), which allows obtaining an index of the characteristics and thus verifying the degree of correlation of the variables through the columns [] function, shown as follows and where they are labeled according to the dataset (annual ICFES data).

$$used\_features = X.columns[selected] \quad (16)$$

In the end, the process undertaken is:

$$gnb.fit(X\_train[used\_features].values, y\_train.values.ravel()) \quad (17)$$

$$y\_pred = gnb.predict(X\_test[used\_features]) \quad (18)$$

a. For the 70% of training data with 30% of test data, the Gaussian naive Bayes model gnb=GaussianNB () is applied, which is stored in the variable gnb.
b. The fit with the fit fuction is made.
c. The prediction of the characteristics related to the $predict()$ function is made.

2.2.4. Neural networks

Inspired by biological neural networks, where the input values received from other units connected to it are added, comparing the amount with the threshold value and if it equals or exceeds it, it sends the exit [28]. The interest is centered on the possibility of learning through a set of observations, taking into account the cost function C: F → R, where C represents how far an optimal solution can be according to the problem to be solved. This is the reason why we try to minimize the processing cost through the distribution of the data and the mean square error [29]. It is exposed to pairs of high amount of data, which is related to y=f(x), where ot takes X as input for the generation of a prediction, the Y error is computed and fed back the network, whose internal parameters are adjusted in sequence until verifying the accuracy of the neural network [30]. The prediction model was made through the perceptron model:

$$model = Sequential() \quad (19)$$

Create the set of nodes through the Sequential () function that groups layers in a linear way (one in front of the other), which then is divided into eight nodes and an output, so that, when performing tests, the more of hidden layer nodes are used, the lower the accuracy.

$$model.add(Dense(24, input\_shape = (X.shape[1],), activation = 'relu')) \quad (20)$$

$$model.add(Dense(1, activation = 'sigmoid')) \quad (21)$$

We adhere the 24 variables selected previously to the *model* variable together with the Dense () function, and applying the activation function = 'relu', which is the one that returns an output from an input value, and for this case, the function that was used was the Sigmoide function.

$$model.compile(loss = 'mean\_squared\_error', \quad (22)$$

$$optimizer = 'adam', \quad (23)$$

$$metrics = ['binary\_accuracy']) \quad (24)$$

It is calculated as the average of the squared differences and the predicted and actual values through the 'mean_squared_error'; with the 'adam' optimization, which is based on the adaptive estimation of first and second order moments, and the output metrics that are adjusted in binary form.

$$model.fit(X\_test, y\_test, epochs = 100) \quad (25)$$

$$scores = model.evaluate(X\_test, y\_test) \quad (26)$$

$$model.predict(X\_train).round() \quad (27)$$

At the end, the model is adjusted through epochs to perform the iterations between nodes, the evaluation on each iteration of both training and test data (evaluate ()) and the prediction of the model.

3. RESULTS AND DISCUSSION

The prediction models allowed us to select the variables that stand out from the student's information and the scores obtained through percent tages.

3.1. Models

3.1.1. Naive Bayes

Figure 1 shows how in Colombia the number of people who want to study and obtain a university degree has gradually increased, however, the choice of university major is being affected by variables such as: gender, residential areas, social stratum and information technologies. This happens because some students don't have the tools and opportunities to be able to study, it is the case of rural areas where the percentage remained stable during these three years. Even so, people are seeking to study a professional major in large cities, due to the accessibility of information, the program they want to study and the time they plan to invest during their studies along with a high percentage of daily reading and Internet use.

When comparing the Bogota's 2016 Saber Pro Tests article [8], social variables are added to verify how it affects their academic performance. The average numbers tend to have results in the median. However, each one suggests that when the education level of the family is higher as well as when the amount of money they make is higher and when the region they live is better, they have better chances to access and stay in any university, something that happens in the Naïve Bayes algorithm because the variables are similar and the data are between 40% and 80%, both use a correlation matrix to select a higher precision, and thus, the demand for students will continue to increase.

3.1.2. Neuronal networks

In Figure 2 the use of Tensor Flow and Keras, that are neural networks libraries, allowed the model to evaluate characteristics of the data, both numerical and categorical, as well as all the columns of the data. It helped determine which of them affected in greater proportion the separation of the information into nodes, until a single output is obtained. When comparing multilayer perceptron algorithm, which is a variant of neural networks [18], the results were as follows:

a. The multilayer perceptron had a percentage that does not exceed 1%, however, the same occurs with the margin of error, where the number of nodes created did not generate a significant response when compared to the prediction related to students learning.
b. In the results obtained with the neural networks, most of the variables presented null percentages or percentages below 50%, due to the amount of information analyzed. Gender continues to greatly affect the selection of majors related to Mathematics, Engineering and Teaching, since the percentages of students who study this type of majors require greater motivation (the government is an important supporter that provides alternatives and incentives for the students' population, as well as the economic resources that owns each family group). Additionally, if compared with secondary sources, the authors mention that all people regardless of gender, family and area of residence, can have access to education, having motivational support factors that drive their desire to learn such as the love towards reading books.

3.1.3. K nearest neighbors

In Figure 3. The development of this prediction model takes into account the variables that are closest to each other, compared to all the others. It relates an average of 1 million of individual data which allows to generate more details around the important variables.

The analysis on each variable shows relevant results, the majority above the average, which allowed obtaining clear trends on Higher Education in Colombia:

a. Women have a higher participation (68%) in majors related to Mathematics, Engineering and Teaching.
b. The urban area continues to be the place of preference for applying for higher education (94%).
c. The use of internet increased by 50% in the past year.
d. The support of the family continues to be important in the academic learning of the children.
e. Motivation to learn increased by 50%, taking into account the variable "Daily Reading Dedication".
f. The mother's education exceeds that of the father by 5%.

In the case of the "Evolution of the inequality of educational opportunities from secondary education to university" [17], it is displayed as follows:

The education of the parents and especially the higher percentage contribution of the mother, which is increasing. However, after some time, the mean has increased by 5%, and, according to secondary sources, it is one of the most important topics that the student needs to retain knowledge. This determines the good or bad results that students will have on a Saber 11 or a Saber pro test for when students to have better results in the Saber 11 and Saber Pro tests, especially in the areas of mathematics and reading.

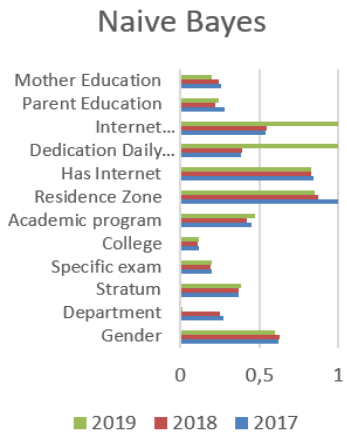


Figure 1. Results naive Bayes

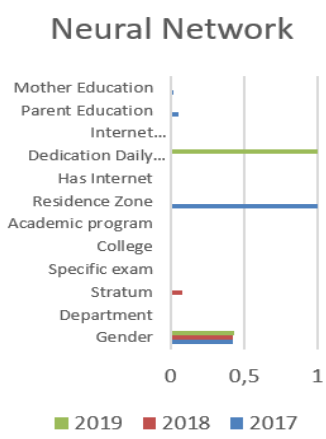


Figure 2. Results neural network

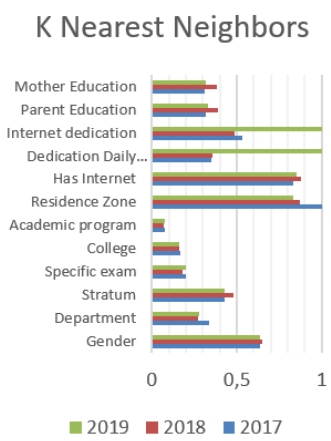


Figure 3. Results KNN

3.1.4. K-means

The K means prediction model algorithm has an unsupervised learning approach, this means that it is used to explore data that does not have a specific objective or it is not possible to determine what information is stored, considering null data, not relevant data, and data with low percentages. This situation is considered. Biased information in ICFES' database, so it is very important to verify what relationship each one has, the dispersion of the data and which are related to each other. The largest amount of data is grouped in an amount of approximately 100 thousand, and in which the approximation is found in the mean (55%) taking into consideration the variables that are related to the study: gender, stratum, scores, education, academic program and autonomous learning, highlighting other variables of a specific nature of each major). There's an important impact on the following variables that were analyzed (Mathematics, Engineering and a general degree). Besides, the Model implies that entrepreneurship predominates in the next generations, since students want to know everything in order to achieve their goals.

When comparing the factors associated with academic performance in mathematics in the Saber 11 tests from 2015 to 2016 [9], 709,421 data sets were analyzed with a percentage of accuracy of 67%. In the case of the mathematics test this has a precision above the average, with a total of 313,231 students. The same occurs with behavior patterns such as: stratum, working day, age, gender and having technological devices, which identifies that social and individual variables are key for both Saber 11 and Saber Pro tests.

3.2. Best forecasting model

The results obtained while applying the prediction model, it was possible to see that the model with greater precision was the one in which the K closest neighbors algorithm was applied, since the studied variables presented percentages above the average considering gender, place of residence, daily reading and internet. Likewise, the model with K closest neighbors had results that were closest to the data when applying the algorithm together with the confusion matrix, which generated that the degree of correlation was efficient. However, a characteristic that should be highlighted is that it was one of the algorithms that had "neighboring" classifications with 12 variables. This means that it had fewer variables compared to the other Models, for example, in the case of naive Bayes and K-means, models that had 14 variables and the iterations in neural networks were used with all the proposed variables, something that allowed the degree of training with the Model applied with the closest neighbors K algorithm to be optimal.

4. CONCLUSION

The prediction models seen in this process allowed us to estimate the different alternatives, trends and needs of the student population in Colombia. In this case, the statistics provided by ICFES compared to the models used with supervised and unsupervised algorithms have been very useful to establish the progress of each student according to their results' history and their basic information at the end of their university degree. With all the information mentioned and analyzed above, it was possible to determine that the alternatives provided by the algorithms focused on the prediction for the development of a Model allowed us to analyze the statistical data and to show the different behaviors of the variables that are associated with the standardized Saber Pro tests.

During the development of the prediction models, the algorithms applied presented a considerable amount of results, however the presence of null data in the ICFES information altered the process in certain variables with the use of supervised algorithms, which did not occur with the applied model with an unsupervised algorithm, since it takes into account unknown variables or non-relevant information, so that an improvement to the process is the union of a supervised algorithm with another unsupervised or semi-supervised, so that although it does not count with the data, it allows that the accuracy of the model is higher and even add data both at the University and College level.

Likewise in the Model were used regression algorithms, neural networks, grouping, bayesians, however, it can be added deep learning algorithms, dimension reduction and decision trees, with the purpose of be able to perform functions such as: reduction of dimension in the number variables to use, data learning through layers of algorithms, in case of data such as images are added, further that there are some columns that work in binary form (economic situation, diseases, property, among others) which a tree-like structure allow representing nodes information, where each branch represents a result.

ACKNOWLEDGEMENTS

We thank the academic institution Universidad Distrital Francisco José de Caldas for supporting this research process.

REFERENCES

[1] E. A. Hanushek, "Addressing cross-national generalizability in educational impact evaluation," *International Journal of Educational Development*, vol. 80, 2021, Art. no. 102318, doi: 10.1016/j.ijedudev.2020.102318.

[2] M. D. Martín-Lagos López, "Education and inequality: A meta-synthesis after the 50th anniversary of coleman's report," *Revista de Educación*, vol. 2018, no. 380, pp. 186-209, 2018, doi: 10.4438/1988-592X-RE-2017-380-377.

[3] F. Barrera-Osorio and H. Bayona-Rodríguez, "Signaling or better human capital: Evidence from Colombia," *Economics of Education Review*, vol. 70, pp. 20-34, 2019, doi: 10.1016/j.econedurev.2019.02.006.

[4] A. Federick, "Finland Education System*," International Journal of Science and Society (IJSOC),* vol. 2, no. 2, pp. 21–32, 2020, doi: 10.200609/ijsoc.v2i2.88.

[5] M. Castro and J. V. Ruíz, "La educación secundaria y superior en Colombia vista desde las pruebas Saber," *Prax. Saber, Praxis & Saber*, vol. 10, no. 24, pp. 341-366, 2019, doi: 10.19053/22160159.v10.n25.2019.9465.

[6] Ministerio de Educación Nacional, C. Velez, "Decreto No. 3963 del 14 de octubre de 2009," *Decreto*, no. 3963, p. 4, 2009.

[7] E. Delahoz-Dominguez, R. Zuluaga, and T. Fontalvo-Herrera, "Dataset of academic performance evolution for engineering students," *Data in Brief*, vol. 30, 2020, Art. no. 105537, doi: 10.1016/j.dib.2020.105537.

[8] D. E. R. Ospina, "Class relations in the higher educational system and its effects in academic performance: The case of Bogotá," *Multidisciplinary Journal of Educational Research*, vol. 9, no. 1, pp. 1-24, 2019.

[9] R. T. Pereira, J. C. Zambrano, and A. H. Troya, "Identification of factors associated with academic performance in mathematics in the saber 11th tests applying educational data mining," *The 17th LACCEI International Multi-Conference for Engineering, Education, and Technology*, 2019, pp. 24-26, doi: 10.18687/LACCEI2019.1.1.297.

[10] M. Guarascio, G. Manco, and E. Ritacco, "Knowledge discovery in databases," *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatic*, vol. 1-3, pp. 336-341, 2018.

[11] S. Chaouch, A. Mejri, and S. A. Ghannouchi, "A framework for risk management in Scrum development process," *Procedia Computer Science*, vol. 164, pp. 187-192, 2019, doi: 10.1016/j.procs.2019.12.171.

[12] K. Vaid and U. Ghose, "Predictive Analysis of Manpower Requirements in Scrum Projects Using Regression Techniques," *Procedia Computer Science*, vol. 173, pp. 335-344, 2020, doi: 10.1016/j.procs.2020.06.039.

[13] Mineducación and Icfes, "Diccionario de variables Saber-Pro Periodo 2012-2019," Instituto Colombiano para la Evaliacion de la Educacion*,* pp. 1-20, 2019.

[14] I. Kumar, K. Dogra, C. Utreja, and P. Yadav, "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 1003-1007, doi: 10.1109/ICICCT.2018.8473214.

[15] D. Bzdok, M. Krzywinski, and N. Altman, "Points of significance: Machine learning: Supervised methods," *Nature Methods*, vol. 15, no. 1, pp. 5-6, 2018.

[16] N. Li, M. Shepperd, and Y. Guo, "A systematic review of unsupervised learning techniques for software defect prediction," *Information and Software Technology*, vol. 122, 2020, Art. no. 106287, doi: 10.1016/j.infsof.2020.106287.

[17] J. A. Sarmiento Espinel, A. C. Silva Arias, and E. van Gameren, "Evolution of the inequality of educational opportunities from secondary education to university," *International Journal of Educational Development*, vol. 66, pp. 193-202, 2019, doi: 10.1016/j.ijedudev.2018.09.006.

[18] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers & Education*, vol. 131, pp. 22-32, 2019, doi: 10.1016/j.compedu.2018.12.006.

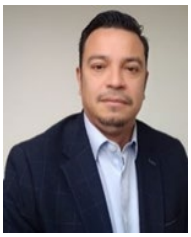
[19] S. Wiyono and T. Abidin, "Comparative Study of Machine Learning Knn, Svm, and Decision Tree Algorithm To Predict Student'S Performance," *International Journal of Research-GRANTHAALAYAH*, vol. 7, no. 1, pp. 190–196, Jan. 2019, doi: 10.29121/granthaalayah.v7.i1.2019.1048.

[20] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, "Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 1, pp. 1-24, 2019, Art. no. e1289, doi: 10.1002/widm.1289.
[21] D. A. Anggoro and N. D. Kurnia, "Comparison of accuracy level of support vector machine (SVM) and K-nearest neighbors (KNN) algorithms in predicting heart disease," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 1689-1694, 2020, doi: 10.30534/ijeter/2020/32852020.
[22] D. S. Maylawati, T. Priatna, H. Sugilar, and M. A. Ramdhani, "Data science for digital culture improvement in higher education using K-means clustering and text analytics," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 10, no. 5, pp. 4569-4580, 2020, doi: 10.11591/ijece.v10i5.pp4569-4580.
[23] A. D. Rachid, A. Abdellah, B. Belaid, and L. Rachid, "Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 8, no. 4, pp. 2367–2383, 2018, doi: 10.11591/ijece.v8i4.pp2367-2383.
[24] J. Wu, "Advances in K-means Clustering A Data Mining Thinking," *Springer*, vol. 44, no. 8, 1st ed., China, 2012. doi: 10.1007/978-3-642-29807-3.
[25] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Systems*, vol. 192, 2020, Art. no. 105361, doi: 10.1016/j.knosys.2019.105361.
[26] M. A. Burhanuddin, R. Ismail, N. Izzaimah, A. A.-J. Mohammed, and N. Zainol, "Analysis of Mobile Service Providers Performance Using Naive Bayes Data Mining Technique," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, pp. 5153-5161, 2018, doi: 10.11591/ijece.v8i6.pp5153-5161.
[27] H. C. Kim, J. H. Park, D. W. Kim, and J. Lee, "Multilabel naïve Bayes classification considering label dependence," *Pattern Recognition Letters*, vol. 136, pp. 279-285, 2020, doi: 10.1016/j.patrec.2020.06.021.
[28] V. H. Medina, J. Rodriguez, and M. A. Ospina, "A Comparative Study Between Feature Selection Algorithms," *International Conference on Data Mining and Big Data-DMBD 2018*, vol. 10943, 2018, pp. 65-76.
[29] T. B. Lopez-Garcia, A. Coronado-Mendoza, and J. A. Domínguez-Navarro, "Artificial neural networks in microgrids: A review," *Engineering Applications of Artificial Intelligence*, vol. 95, 2020, Art. no. 103894, doi: 10.1016/j.engappai.2020.103894.
[30] J. Viquerat and E. Hachem, "A supervised neural network for drag prediction of arbitrary 2D shapes in laminar flows at low Reynolds number," *Computers & Fluids*, vol. 210, 2020, Art. no. 104645, doi: 10.1016/j.compfluid.2020.104645.

BIOGRAPHIES OF AUTHORS

Daniel Esteban Martínez Cervera was born in Bogotá, Colombia, on October 22, 1995. He received his BS degree in System Engineer from the ean University in 2017, Colombia, Specialist in Software Engineering from Francisco Jose de Caldas District University in 2018 and his MSc. In Information and Communication Sciences from Francisco Jose de Caldas District University in 2021. He has publications about Efficient algorithm for the calculation of molecular masses, Data mining and Machine Learning in journals. He has participed in programming marathons from 2013. His current research interests are in Machine Learning, Data Mining, Software Engineering and web development.

Octavio Salcedo was born in Morroa, Sucre, Colombia, on September 28, 1969. He received his BS degree in System Engineer from the Universidad Autonoma de Colombia, Colombia, and his MS degree in Teleeinformatics from the Universidad Distrital Francisco José de Caldas, Colombia, in 1998, MS degree in Economy from the Universidad de los Andes, Colombia, in 2004. A diploma of advanced Studies (DEA) from the Pontificia Universidad de Salamanca, Campus of Madrid, Spain in 2010. A PhD in Informatics Engineering from the Pontificia Universidad de Salamanca, Campus of Madrid, Spain in 2013. A PhD in political studies from the Universidad Externado de Colombia, Colombia in 2013. He has several publications in telecommunications area in international journals, books, and conferences. His current research interests are in data networks and Networkings and in political studies and Economy PhD Salcedo is Titular Professor at the Universidad Distrital Francisco José de Caldas and He is Associate Professor Universidad Nacional de Colombia and director of Internet Inteligente research group.



Marco Aguilera Prado Economist from Universidad Autónoma de Occidente, M.A. Planning and Administration of Regional Development, PhD(c) Engineering from Universidad Distrital Francisco José de Caldas. Professor at Universitaria Agustiniana Vice-Rectory for Research. He has wide experience in quantitative modelling for social fenomena. He has been professor in different universities in Colombia. His recent works includes research management and research in Economics of Education and Bibliometrics of information on social phenomena for modeling and forecasting. Leadership, teamwork, managing for results. Experience in different IES Colombian for 15 years: research, teaching and project management.