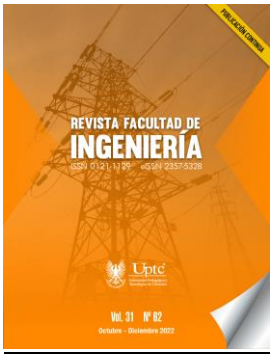
Revista Facultad de Ingeniería
Journal Homepage: https://revistas.uptc.edu.co/index.php/ingenieria



Applying Predictive Data Mining to Discover Factors Associated to the Language Skill Performance from Elementary School Students

Ricardo Timarán-Pereira[1]
Javier Caicedo-Zambrano[2]
Andrea Timarán-Buchely[3]

Received: August 31, 2022 **Accepted:** December 22, 2022 **Published:** December 31, 2022

Citation: R. Timarán-Pereira, J. Caicedo-Zambrano, A. Timarán-Buchely, "Applying Predictive Data Mining to Discover Factors Associated to the Language Skill Performance from Elementary School Students," *Revista Facultad de Ingeniería*, vol. 31 (62), e14814, 2022. https://doi.org/10.19053/01211129.v31.n62.2022.14814

Abstract

In this paper, predictive data mining techniques are applied to determine the academic performance from fifth grade students in the Saber 5° tests Language skill

[1] Ph. D. Universidad de Nariño (Pasto-Nariño, Colombia). ritimar@udenar.edu.co. ORCID: 0000-0002-0006-6654
[2] Ph. D. Universidad de Nariño (Pasto-Nariño, Colombia). jacaza1@udenar.edu.co. ORCID: 0000-0002-5399-0410
[3] Universidad Javeriana (Cali-Valle, Colombia). ORCID: 0000-0003-4041-5115

at Colombian elementary schools in 2017. We employed the CRISP-DM methodology. Socioeconomic, academic, and institutional information was available at the ICFES databases. A minable dataset was obtained using data cleaning and transformation techniques. A decision tree was built with the Weka tool J48 algorithm. Some of the predictors of the discovered patterns are the nature and location of the school, whether or not students failed a school year, the age group, the mother's educational attainment, and the rates of ICTs and household appliances. The findings of this research serve as quality information for the decision-making at the Ministry of National Education (MEN) and the secretaries of education, and for the directors of elementary educational institutions to define improvement plans that result in the quality of elementary school education in Colombia.

Keywords: classification; data mining; decision trees; performance patterns; predictive model; Saber 5° tests.

Minería predictiva aplicada al descubrimiento de factores asociados al desempeño en la competencia de lenguaje de los estudiantes de básica primaria

Resumen

En este artículo se aplican técnicas predictivas de minería de datos para descubrir patrones de desempeño académico en la competencia de Lenguaje de las pruebas Saber 5° que presentaron los estudiantes de las instituciones educativas colombianas de básica primaria en el año 2017. Para tal fin, se utilizó la metodología CRISP-DM y se tuvo en cuenta la información socioeconómica, académica e institucional de las bases de datos del ICFES. Se obtuvo un conjunto de datos minable utilizando técnicas de limpieza y transformación de datos y se construyó un árbol de decisión con el algoritmo J48 de la herramienta Weka. Entre los factores predictores de los patrones descubiertos están la naturaleza y la ubicación del colegio, si los estudiantes reprobaron o no algún grado, el grupo etario, la educación de la madre y los índices de TICs y electrodomésticos en los hogares. El conocimiento producido en esta investigación es información de calidad para la

toma de decisiones en el MEN y las secretarías de educación y para que las directivas de las instituciones educativas de básica primaria definan planes de mejoramiento que redunden en la calidad de la educación en Colombia.

Palabras clave: árboles de decisión; clasificación; minería de datos; modelo predictivo; patrones de desempeño; pruebas Saber 5°.

Mineração preditiva aplicada à descoberta de fatores associados ao desempenho na competência linguística de alunos do ensino fundamental

Retomar

Neste artigo, técnicas preditivas de mineração de dados são aplicadas para descobrir padrões de desempenho acadêmico na competência linguística dos testes Saber 5 apresentados por alunos de escolas primárias colombianas em 2017. Para isso, foi usada a metodologia CRISP-DM e o socioeconômico, foram consideradas as informações acadêmicas e institucionais das bases de dados do ICFES. Um conjunto de dados mineráveis foi obtido usando técnicas de limpeza e transformação de dados e uma árvore de decisão foi construída com o algoritmo J48 da ferramenta Weka. Entre os preditores dos padrões descobertos estão a natureza e a localização da escola, se os alunos foram ou não reprovados, a faixa etária, a escolaridade da mãe e as taxas de TIC e eletrodomésticos nas residências. O conhecimento produzido nesta pesquisa é informação de qualidade para a tomada de decisões nos MEN e nas secretarias de educação e para os diretores de instituições de educação primária definirem planos de melhoria que resultem na qualidade da educação na Colômbia.

Palavras-chave: árvores de decisão; classificação; mineração de dados; modelo preditivo; padrões de desempenho; testes Saber 5º.

I. INTRODUCTION

According to the Orientation Guide of the Saber 5º tests [1], they “evaluate elementary school fifth-grade students aiming to contribute to the improvement of the quality of Colombian education through periodic tests that evaluate the students' competences in Language, Mathematics, Natural Sciences, and Citizenship, and analyze the factors that affect their achievements.” Its nature enables assessing their progress in a given period and establish the impact of specific programs and improvement actions. According to the ICFES [2] guidelines for sample and census applications, the design of the Saber 5° tests “aligns with the Basic Skills Standards established by the Ministry of National Education (MEN), understood as common a reference that enables to determine how well students and the education system as a whole meet the quality expectations in terms of what they know and are able to do.”

In Colombia, several studies on the Saber 5° tests have been conducted. Torres et al. [3], Martín [4], and Gutiérrez [5] seek to identify the variables associated with academic performance in the Saber 5° Tests based on only one fundamental area —Natural Sciences, Mathematics or Language. In another study [6], the associated factors of the 5° and 9° grade tests were analyzed; one of the conclusions was that the higher the socioeconomic level of the students and their families, the higher the expected performance in both areas and grades. In addition, private school students tend to obtain higher test scores, and the differences —compared to those who attend official schools— increase as socioeconomic conditions improve. The report on Associated Factors of the Saber 5° and 9° Tests published by the ICFES [7] identifies variables related to performance in the Saber tests. Statistical techniques that enabled visualizing elements that affect academic performance were applied. To extend evaluation processes, the ICFES studied factors associated with school performance using theoretical models. The aim was to explain the relationship between the elements that determine learning at three levels: educational institutions, classrooms, and students [2].

According to Timaran et al. [8-9] “the studies carried out to date to determine academic performance in the Saber tests are based on information processed

through statistical analysis, considering variables and primary relationships, but ignoring the true interrelationships, which are usually hidden and can only be discovered giving a more complex treatment to the data, which is possible with data mining."

Educational institutions can use data mining to make comprehensive analysis of the characteristics of their students and evaluation methods, to reveal successful processes or, on the contrary, detect fraud or inconsistencies, and determine the probability of dropout of any student [10-13].

In this paper, predictive data mining techniques are applied to discover academic performance patterns in the Language skill of the Saber 5° test taken by fifth grade students at Colombian elementary schools in 2017.

II. Materials and Methods

The data used in this research was obtained from the ICFES databases, where the socioeconomic, academic, and institutional information of the students of the elementary educational institutions who presented the Saber 5° tests are stored.

This research was descriptive, under the quantitative approach, and applied a non-experimental design. The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was used. According to [14], [15], [16], it is "one of the models used, mainly, in academic and industrial environments and the most widely used reference guide in the development of data mining projects". It comprises six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

In the business understanding phase, the necessary information was collected and selected. It allowed the researchers to know and appropriate the knowledge about the Saber 5° tests and the evaluated skills, with emphasis in Language. This process made it possible to collect the correct data to obtain adequate results.

In the data understanding phase, the socioeconomic, academic, and institutional information of Colombian students who presented the Saber 5° tests in 2017 was identified, collected, and familiarized. As a result, an initial data set called *sbr11_776436A56* was obtained, it had 776436 records and 56 attributes.

In the data preparation phase, data cleaning and transformation techniques were applied to the set sbr11_776436A56 to eliminate noisy, null, atypical data, and transform some attributes so that they obtain more information, and to eliminate those irrelevant attributes that did not contribute to the pattern detection process. As a result, the data set called *sbr11_776436A15* was obtained. It consisted of 776436 records and 15 attributes, which served as the basis for the modeling phase.

In the modeling phase, the classification model with decision trees was selected as the most appropriate data mining technique to solve the research problem due to the ease and simplicity to interpret the patterns obtained [17-19]. This technique has several advantages: first, the reasoning process behind the model is clearly evident when examining the tree, contrary to other black box modeling techniques, where the internal logic can be difficult to figure out; second, the process automatically includes only the attributes that really matter in decision-making and omit the ones that do not contribute to the accuracy of the tree [20-21].

In the evaluation phase, the cost of the classifier for the sbr11_776436A15 repository was estimated through the confusion matrix [18]. Following the recommendation of Hernández et al. [20], we used the cross-validation method with 10 partitions (10-fold cross validation) to test the quality and validity of the model and to reduce the dependency of the result on the way in which the partition is performed [20]. Likewise, the discovered patterns were evaluated to determine their validity, remove redundant or irrelevant patterns, and interpret useful patterns so they are understandable to the user, considering their support and confidence.

In the deployment phase, the discovered patterns were documented. These are quality information to help the decision-making of organizations such as the MEN and elementary educational institutions to create plans to improve the quality of elementary education in Colombia.

III. RESULTS

A. Exploratory Data Analysis

To understand the data, we analyzed the socioeconomic variables of the students who took the Saber 5° tests in 2017. The results show that by gender most of them are men, with a 52.3%. By age, most are between 10 years old (39.9%) and 11 years old (35.2%). Almost all students (89.3%) do not live in overcrowded households. Moreover, most of their parents have a high school degree (48.2% of the mothers and 47% of the fathers). The majority of students have good rates of ICTs and home appliances (43.6% and 47.7%, respectively). The largest number of students are from the Atlantic region (25.3%), followed by the Eastern (16.9%), and Pacific (15.2%) regions. Finally, 53% of students had a Language performance “below the national mean” and 47% were “above the national mean.”

B. Results

By selecting decision trees as classification technique, we aim to obtain a model that enables forecasting the socioeconomic, academic, and institutional factors associated with good (above mean) or poor (below mean) new cases of academic performance in the Saber 5° test. The target attribute was the score obtained in the Language skill. Different decision tree algorithms were evaluated with the Weka tool to select the one that best classifies the sbr11_776436A15 dataset. Results are shown in Table 1.

Table 1. Evaluation of different decision tree algorithms

Algorithm	Accuracy
Decision Stump (One-level decision tree)	61.95%
J48	64,44%
LMT (Logistic Model Tree)	65.48%
Random Forest	73.28%
Random Tree	63.65%
RepTree	55.50%

According to Table 2, the algorithms with the highest accuracy were Random Forest and LMT, but we did not choose them due to their interpretability. For this reason, we chose the J48 algorithm for the construction of decision tree classification models since it facilitates understanding the patterns.

Once the algorithm, training, and testing methods for the models were selected, we built decision trees with the J48 algorithm of the WEKA tool V. 3.9.4 [22] that implements algorithm C4.5 [23]. For pruning the tree, the confidence level C, which influences the size and predictability of the tree was considered. The default value of this factor is 25% and as it decreases more pruning operations are allowed, thus getting smaller trees [24]. The parameter M was also used; it determines the minimum number of records per tree node. The Language score obtained by the students was chosen as class attribute and it was discretized as “above the mean” and “below the mean.”

Different tree models were generated to choose the decision tree that best classifies students and has the highest level of interpretability of the patterns associated with academic performance. Therefore, two values for the confidence factor C were set: 25% and 5%, combined with two values for the factor M: 1% (7764 examples) and 0.05% (3882 examples). The Cross Validation test with 10 folds was used. Table 2 shows the different trees constructed and their percentage of accuracy.

Table 2. Selection of the best tree.

Tree	C	M%	% Accuracy
leng_c25m7764	25	1	63.94
leng_c25m3882	25	0.05	64.38
leng_c005m7764	5	1	63.97
leng_c005m3882	5	0.05	64.37

According to Table 2, the tree built with the parameters C=0.25 and M=3882 was the most accurate. Figure 1 shows the obtained classification tree. To evaluate or estimate the cost of the constructed classification model, the confusion matrix was used. This tool enables visualizing the performance of a supervised learning algorithm, as shown in Figure 2.

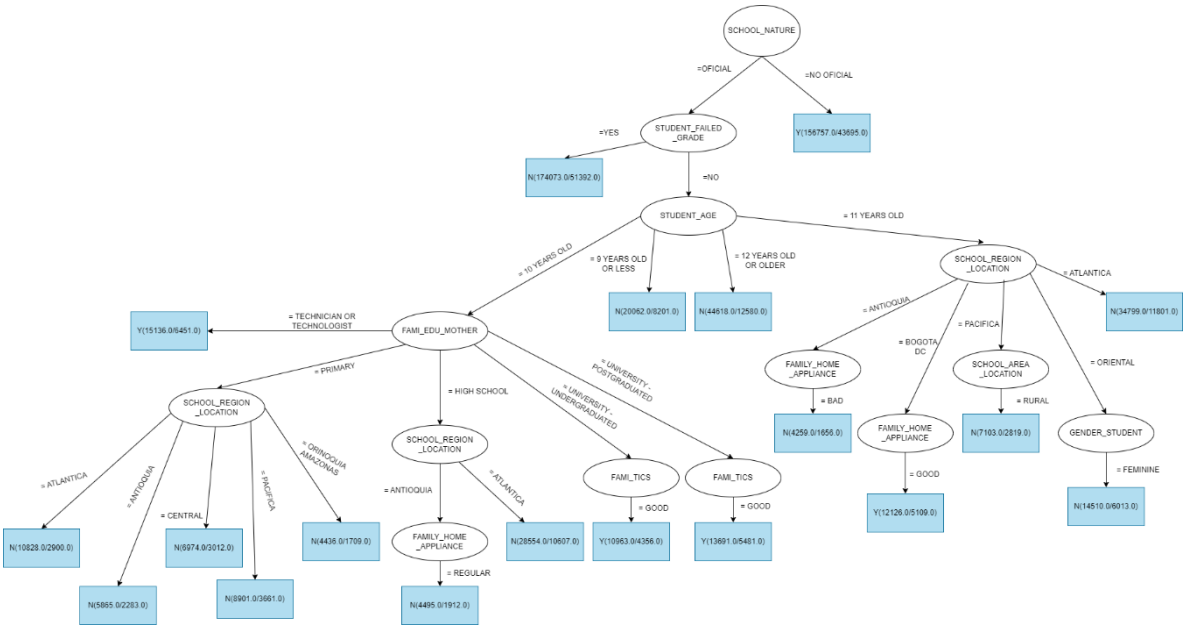
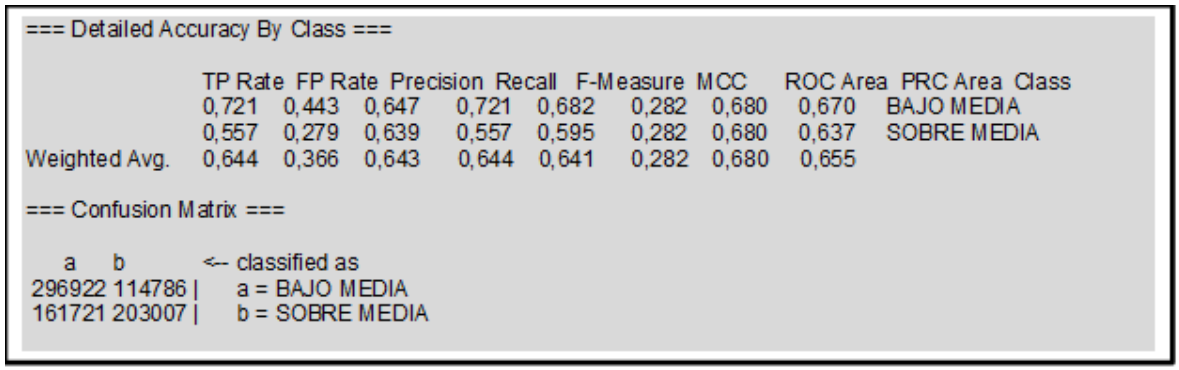


Fig. 1. Most accurate tree model.

C. Discussion

Analyzing the Saber 5º tests Language performance according to the model shown in Figure 1, 499,929 instances were classified correctly —i.e., an accuracy of 64.4%— and 276,507 instances were classified incorrectly —i.e., 35.6%.

Evaluating the model with the confusion matrix, obtained with the Weka tool as shown in Figure 2, it correctly predicted 296,922 cases of students whose performance in Language was below the mean (True Negatives-TN) and 203,007 cases above the mean (True Positives-TP). Likewise, 161,721 cases whose performance is above the mean were incorrectly classified as below the mean (False Negatives-FN) and 114,786 cases whose performance was below the mean were incorrectly classified as above the mean (False Positives-FP).



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,721    0,443    0,647      0,721   0,682      0,282    0,680     0,670     BAJO MEDIA
                 0,557    0,279    0,639      0,557   0,595      0,282    0,680     0,637     SOBRE MEDIA
Weighted Avg.    0,644    0,366    0,643      0,644   0,641      0,282    0,680     0,655

=== Confusion Matrix ===

   a      b        <-- classified as
296922 114786 |     a = BAJO MEDIA
161721 203007 |     b = SOBRE MEDIA
```

Fig. 2. Confusion Matrix generated with the Weka Tool.

For the case of students with a score below the mean, the model has a prediction accuracy of 0.65, i.e., 65% of the predicted cases are correct. The sensitivity (TPR) and Recall of the model is 0.72, which indicates that the model correctly classifies 72% of the students who are actually in this category. On the other hand, the False Negative Rate (FP Rate) of the model is 0.44, which means that 44% of students who were above the mean were classified as below the mean. The F-measure is 0.68, which means that the harmonic mean between the precision and the recall of those below the mean is 68%. The combination of these measures indicates a better performance of the model for them.

For the case of students who are above the mean in the language score, the model has a prediction accuracy of 0.64, i.e., 64% of the predicted cases are correct. The sensitivity (TPR) and Recall of the model is 0.56, which indicates that the model correctly classifies 56% of the students who are actually in this category. On the other hand, the False Positive Rate (FP Rate) of the model is 0.28, which means that 28% of students who were below the mean were classified as above the mean. The F-measure is 0.60, which means that the harmonic mean between the precision and the recall of those above the mean is 60%. In the combination of these measures, a worse performance of the model can be seen for the latter.

The model built to detect performance patterns in the Saber 5° tests Language skill is not unbalanced. There is no significant difference between cases below the mean (53%) and those above the mean (47%). Therefore, Cross Validation was used to build the tree.

Within the evaluation metrics calculated above, the model has an accuracy of 64.4% and it predicts better the students *below the mean* than those above the mean. This is also observed in the relationship between Recall and Accuracy given in the PRC area, which is 0.67 for students below the mean and 0.64 for those above the mean. The Mathews-MCC correlation coefficient of the model is 0.28, which indicates a medium agreement between what is predicted and what is observed. That is, a regular quality in the prediction. Regarding the areas, the ROC area of the model (0.68) indicates that it has a good performance in the classification of Colombian students with respect to the Language score obtained in the Saber 5° tests because it is greater than 0.5.

To choose the most representative performance patterns in Language, as seen in Figure 1, students who pass a minimum support of 1% and a minimum confidence of 60% were considered. The following rules are the interpretation of these patterns (in descending order by support):

Rule 1. If the student comes from an official school and failed school years, then his/her performance in Language is likely to be below the national mean, with a support of 22.4% and a confidence of 70.5%. 42.3% of the analyzed students below the mean meet this pattern**.**

Rule 2. If the student comes from an unofficial school, then his/her performance in Language is likely to be above the national mean, with a support of 20.2% and a confidence of 72.1%. 43% of the analyzed students above the mean meet this pattern.

Rule 3. If the student comes from an official school, did not fail school years, and is 12 years old or older, then his/her performance in Language is likely to be below the national mean, with a support of 5.7% and a confidence of 71.8%. 10.8% of the analyzed students below the mean meet this pattern.

Rule 4. If the student comes from an official school, did not fail school years, belongs to the 10 years old age group, the educational attainment of their mother is high school, and the student comes from the Atlantic region, then their performance in Language is likely to be below the national mean, with a support of 3.7% and a

confidence of 62.9%. 6.9% of the analyzed students below the mean meet this pattern.

Rule 5. If the student comes from an official school, did not fail school years, belongs to the 10 years old age group, the educational attainment of their mother is university and the ICT index at home is good, then their performance in Language is likely to be above the national mean, with a support of 1.4% and a confidence of 60.3%. 3% of the analyzed students above the mean meet this pattern.

Rule 6. If the student comes from an official school, did not fail school years, belongs to the 10 years old age group, the educational attainment of their mother is elementary school, and the student comes from the Atlantic region, then their performance in Language is likely to be below the national mean, with a support of 1.4% and a confidence of 73.2%. 2.6% of the analyzed students below the mean meet this pattern.

IV. Conclusions

In this research, the decision tree classification model was selected to detect performance patterns in the Saber 5° tests Language skill presented by Colombian students of elementary schools in 2017 because it is easier to interpret the patterns. To prepare the data, the construction and evaluation of the model, we followed the stages of the CRISP-DM methodology.

Among the most important factors discovered —associated with good or bad academic performance in Language— we found the nature and location of the school, whether or not students failed a school year, the age group, the educational attainment of the mother, and ICT rates and home appliances. The main characteristic of the students' good performance in Language is that they belong to unofficial (private) schools with a support of 20.2% —in relation to all the students who presented the Saber 5° tests— and a confidence of pattern of 72.1%. Likewise, students from official schools who failed school years had low performance with a support of 22.4% and a pattern confidence of 70.5%.

According to the analyzed quality metrics, the model predicts better students below the mean (negative) than those above the mean (positive). This means that the model is more sensitive than specific; thus, it tries to avoid false negatives.

Future work includes building decision tree models to predict the performance of Colombian students in the Saber 5° tests Mathematics skill and analyzing the results. Apply descriptive data mining techniques to find the association and similarity relationships between the socioeconomic, academic, and institutional attributes of the students who took the Saber 5° tests.

AUTHORS' CONTRIBUTION

Ricardo Timarán-Pereira: Research, application of data mining techniques, writing, editing and review.

Javier Caicedo-Zambrano: Research, preparation of the theoretical framework, writing, editing and review.

Andrea Timaran-Buchely: Research, data cleaning and transformation, writing, editing and review.

REFERENCES

[1] Icfes, *Saber 5°: Guía de orientación. Instituto Colombiano para la Evaluación de la Educación (ICFES),* Colombia, Mineducación, 2017.

[2] Icfes, *Pruebas Saber 3°, 5° y 9°: Lineamientos para las aplicaciones muestral y censal (ICFES)*, Colombia, Mineducación, 2014.

[3] J. Torres, L. Pachajoa, R. Pantoja, "Resultados de las Pruebas Saber en el grado quinto del área de las ciencias naturales en tres instituciones educativas oficiales del municipio de Pasto," *Revista Fedumar Pedagogía y Educación*, vol. 1, no. 1, pp. 55-69, 2014.

[4] S. Martín, *Pruebas Saber de lenguaje 3° y 5°: Posibilidades y retos desde la perspectiva de la evaluación formativa*, Colombia, Universidad Pedagógica Nacional, 2015.

[5] Y. Gutiérrez, *Relación entre la estructura familiar y el rendimiento académico en el área de matemáticas*, Colombia, Editorial Milla, 2015.

[6] Icfes, *Lineamientos generales Saber 5° y 9°*, Colombia, Instituto Colombiano para la Evaluación de la Educación (ICFES), 2009.

[7] Icfes, *Informe técnico Saber 5° y 9*, Colombia, Instituto Colombiano para la Evaluación de la Educación (ICFES), 2011.

[8] R. Timarán, J. Caicedo, A. Hidalgo, *Aplicación de la Minería de datos en la Detección de Patrones de Desempeño Académico e las Pruebas Saber Pro*, Colombia, Editorial Universidad de Nariño, 2021.

[9] R. Timarán, J. Caicedo, A. Hidalgo, *Minería de datos educativa para el descubrimiento de factores asociados al desempeño académico en las Pruebas Saber 11º*. Colombia, Editorial Universidad de Nariño, 2021.

[10] S. Valero, A. S. Vargas, M. García, "Minería de datos: Predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos," *En Línea*, vol. 779, no. 73, pp. 33-38, 2005.

[11] H. Escobar, M. Alcívar, C. Márquez, C. Escobar, "Implementación de Minería de Datos en la Gestión Académica de las Instituciones de Educación Superior," *Didasc@lia: Didáctica y educación*, vol. 8, no. 3, pp. 203-212, 2017.

[12] R. Timarán, A. Calderón, J. Jiménez, "Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil," *Ventana Informática*, vol. 28, pp. 31-47, 2013. https://doi.org/10.30554/ventanainform.28.181.2013

[13] S.R. Timarán, J. Jiménez, A. Calderón, *Detección de patrones de deserción estudiantil con minería de datos*, Colombia, Editorial Universidad de Nariño, 2017.

[14] A. Azevedo, M. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," in *Proceedings of IADIS European Conference on Data Mining*, Netherlands, 2008, pp. 182-185.

[15] J. Hernández, M. Ramírez, C. Ferri*, Introducción a la Minería de Datos*. Spain, Editorial Pearson Prentice Hall, 2005.

[16] J. Villena, *CRISP-DM: La metodología para poner orden en los proyectos de Data Science*, 2016, https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science.

[17] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, Third Edition (3 edition.). Burlington, MA: Morgan Kaufmann, 2011.

[18] K. Sattler, O. Dunemann, "SQL database primitives for decision tree classifiers," in *Proceedings of the tenth international conference on Information and knowledge management*, USA, 2001, pp. 379–386.

[19] R. Timarán, M. Millán, "New algebraic operators and SQL primitives for mining classification rules," in *Computational Intelligence*, USA, 2006, pp. 61–65.

[20] J. Hernández, M. Ramírez, M, C. Ferri, *Introducción a la Minería de Datos*, Spain, Editorial Pearson Prentice Hall, 2005.

[21] I. Witten, E. Frank, M, Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Third Edition, Morgan Kaufmann, 2011.

[22] M. Hall, E. Frank, I. Witten, *Practical Data Mining: Tutorials*, University of Waikato, 2011.

[23] J. Quinlan, *Programs for machine learning*, Morgan Kaufmann, 1993.

[24] M. García, A. Álvarez, *Análisis de datos en WEKA–pruebas de selectividad*, 2010. http://www.it.uc3m.es/~jvillena/irc/practicas/06-07/28.pdf