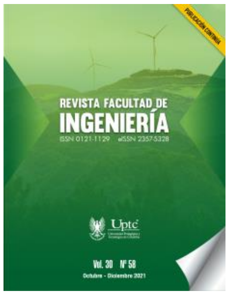
Revista Facultad de Ingeniería
Journal Homepage: https://revistas.uptc.edu.co/index.php/ingenieria



Using Decision Trees to Predict Critical Reading Performance

Andrea Timarán-Buchely[1]
Silvio-Ricardo Timarán-Pereira[2]
Arsenio Hidalgo-Troya[3]

Received: October 11, 2021 **Accepted:** November 28, 2021 **Published:** December 01, 2021

Citation: A. Timarán-Buchely, S.-R. Timarán-Pereira, A. Hidalgo-Troya, "Using Decision Trees to Predict Critical Reading Performance," *Revista Facultad de Ingeniería*, vol. 30 (58), e13495, 2021. https://doi.org/10.19053/01211129.v30.n58.2021.13792

Abstract

In Colombia, all undergraduate students, regardless of the professional training program they take, must complete the general competencies sections of the Saber Pro exam that include Critical Reading, Quantitative Reasoning, Citizen Competencies, Written Communication, and English. This paper presents the application of the classification technique based on decision trees in the prediction of the performance in the Critical Reading section presented by the students of the

[1] Universidad Javeriana Cali (Cali-Valle del Cauca, Colombia). ORCID: 0000-0003-4041-5115
[2] Ph. D. Universidad de Nariño (Pasto-Nariño, Colombia). ritimar@udenar.edu.co. ORCID: 0000-0002-0006-6654
[3] M. Sc. Universidad de Nariño (Pasto-Nariño, Colombia). arsenio.hidalgo@udenar.edu.co. ORCID: 0000-0003-4080-118X

Pontificia Universidad Javeriana Cali in the years 2017 and 2018. The CRISP methodology was used. From the socioeconomic, academic and institutional data stored in the ICFES databases, a data repository was built, cleaned and transformed. A mineable view composed of 2052 records and 17 attributes was obtained. The J48 algorithm of the Weka tool was used to build the decision tree. The score obtained in the Critical Reading section of the Saber Pro exam was taken as a class. According to the results obtained, the Philosophy, Applied Mathematics, and Medicine programs stood out for having the best performance in this test. Among the predictive variables associated with performance in the Critical Reading skill are the faculty, the age group and the student's transportation index, as three important variables related to the good or low academic performance of the students of the Universidad Javeriana Cali. The knowledge generated in this research is constituted in quality information to support the decision-making process of the university directives in order to improve the quality of the higher education offered in this institution.

Keywords: academic performance; critical reading; decision trees; J48 algorithm; Saber Pro.

Aplicación de árboles de decisión para predecir el desempeño en lectura crítica

Resumen

En Colombia, todos los estudiantes de pregrado, sin importar el programa de formación profesional que cursen, deben presentar las pruebas de competencias genéricas del examen Saber Pro que incluyen: Lectura Crítica, Razonamiento Cuantitativo, Competencias Ciudadanas, Comunicación Escrita e inglés. En este artículo se presenta la aplicación de la técnica de clasificación basada en árboles de decisión para predecir el desempeño en la prueba de Lectura Crítica del examen Saber Pro que presentaron los estudiantes de la Pontificia Universidad Javeriana Cali en los años 2017 y 2018. Se utilizó la metodología CRISP-DM. A partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES, se construyó, limpio y transformó un repositorio de datos. Se

DOI: https://doi.org/10.19053/01211129.v30.n58.2021.13792

obtuvo una vista minable compuesta por 2052 registros y 17 atributos. Se utilizó el algoritmo J48 de la herramienta Weka para construir el árbol de decisión. De acuerdo con los resultados obtenidos, se destacaron los programas de Filosofía, Matemáticas Aplicadas y Medicina por tener el mejor desempeño en esta prueba. Entre las variables predictoras asociadas al desempeño en la competencia de Lectura Crítica, están la facultad, el grupo etario y el índice de transporte del estudiante, como tres variables importantes relacionadas al buen o bajo desempeño académico de los estudiantes de la Universidad Javeriana Cali. El conocimiento generado en esta investigación, se constituye en información de calidad para soportar la toma de decisiones de las directivas universitarias en vía del mejoramiento de la calidad de la educación superior que se brinda en esta institución.

Palabras clave: algoritmo J48; árboles de decisión; desempeño académico; lectura crítica; Saber Pro.

Aplicação de árvores de decisão para prever o desempenho de leitura crítica

Resumo

Na Colômbia, todos os alunos de graduação, independentemente do programa de treinamento profissional que estejam cursando, devem apresentar os testes de habilidades genéricas do exame Saber Pro que incluem: Leitura Crítica, Raciocínio Quantitativo, Competências Cidadãs, Comunicação Escrita e Inglês. Este artigo apresenta a aplicação da técnica de classificação baseada em árvores de decisão para predizer o desempenho na prova de Leitura Crítica do exame Saber Pro apresentada pelos alunos da Pontificia Universidad Javeriana Cali nos anos de 2017 e 2018. Foi utilizado o CRISP- Metodologia DM. A partir dos dados socioeconômicos, acadêmicos e institucionais armazenados nas bases de dados do ICFES, um repositório de dados foi construído, limpo e transformado. Foi obtida uma vista lavrável composta por 2.052 registros e 17 atributos. O algoritmo J48 da ferramenta Weka foi usado para construir a árvore de decisão. De acordo com os resultados obtidos, os programas de Filosofia, Matemática Aplicada e Medicina destacaram-se por ter o melhor desempenho nesta prova. Entre as variáveis

DOI: https://doi.org/10.19053/01211129.v30.n58.2021.13792

preditivas associadas ao desempenho na competência Leitura Crítica, encontram-se o corpo docente, a faixa etária e o índice de transporte do aluno, três importantes variáveis relacionadas ao bom ou baixo desempenho acadêmico dos alunos da Universidad Javeriana Cali. O conhecimento gerado nesta pesquisa constitui-se em informação de qualidade para subsidiar a tomada de decisão das diretrizes da universidade no sentido de melhorar a qualidade do ensino superior que é oferecido nesta instituição.

Palavras-chave: algoritmo J48; árvores de decisão; leitura crítica; performance acadêmica; Saber Pro.

I. INTRODUCTION

Saber Pro is part of a toolset used by the Colombian State to assess the quality of higher education. One of its objectives is to assess the development level of specific competencies in students about to complete undergraduate degree programs offered by higher education institutions [1]. Saber Pro is a compulsory requirement for graduation according to Act 1324 of 2009, and it is applied once a year. This test evaluates general and specific competencies. The general competencies assessment is divided into sections: critical reading, written communication, quantitative reasoning, English, and citizenship competencies [1]. Specifically, the critical reading competency assesses the performance associated with reading, critical thinking, and interpersonal understanding [2].

According to Timarán et al. [3], the studies carried out so far at the national level [4,5,6] regarding the Saber Pro exam are based on information processed through statistic analysis, where mainly variables and primary relationships are considered. They do not consider the actual interrelations, which are usually hidden and can only be described employing more complex data analyses, such as data mining.

The use of data mining in education is not a new topic. Its study and implementation have been very relevant in recent years, and its techniques can be used to explain or predict any phenomenon within the educational field [7]. For example, it is possible to predict the dropout probability of any student with a very high-reliability rate through data mining techniques [7,8,9]. Furthermore, educational institutions can use data mining to analyze their student's characteristics or evaluation methods comprehensively and thus discover successful methodologies, frauds or inconsistencies [9].

This article presents the application of the decision tree-based classification technique in predicting the performance in the critical reading section of the Saber Pro exams presented by students from the Pontificia Universidad Javeriana Cali in 2017 and 2018.

II. Materials and Methods

Descriptive studies are designed to describe the distribution of variables without considering the causal or other nature hypotheses. That is why this research was descriptive with a quantitative approach applied to a non-experimental design. The results of the Colombian students that presented the Saber Pro exams in 2017 and 2018, available at the databases of the Colombian Institute for Higher Education (ICFES), were used as the information source. Given that this research involves data mining, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was used. This model is used mainly in the academic and industrial fields, and it is the reference guide most widely used in developing this type of project [10, 11, 12]. It comprises six stages: problem analysis, data analysis, data preparation, modeling, evaluation, and deployment [12].

In the problem analysis stage, the activities that allow to deepen and fully understand the Saber Pro exam and the Critical Reading general competency were carried out, making it possible to collect the correct data to interpret the results adequately.

In the data analysis stage, the socioeconomic, academic, and institutional information available at the ICFES databases, at the time of conducting this research, which corresponded to the results obtained by the students of Universidad Javeriana who took the Critical Reading section of the Saber Pro exams in 2017 and 2018 was identified, compiled, and studied. After integrating the repositories of each year, the result was an initial data set denominated *sbpro_lec_2052A101,* which included 2052 records and 101 attributes

Considering that high dimensionality is an issue for the discovery of patterns in data mining [12], the *sbpro_lec_2052A101* set was cleaned and transformed during the data preparation stage to delete noisy, null, and atypical data, transform some attributes to obtain greater information gain and delete the irrelevant attributes that would not help in the pattern detection process. The result was the data set called *sbpro_lec_2052A23* conformed by 2052 records and 23 attributes, which was the base for the modeling stage.

The decision tree classification model was selected during the modeling stage as the data mining technique more suitable for solving the research problem. This model is probably the most used and popular because it is simple and easy to understand [13,14,15]. The importance of decision trees comes from their capability to build interpretable models, being this a decisive factor in its use. The decision tree classification considers disjoint classes so that the tree will result in only one leaf, assigning a unique class to the prediction [16]. This technique has several advantages. First, the reasoning process behind the model is evident when the tree is examined. This is in contrast to other black-box modeling techniques where the internal logic may be difficult to determine. Second, the process automatically includes in its rule only the attributes that indeed count for the decision making. The attributes that do not contribute to the tree's accuracy are omitted [14].

The process to test the model's quality and validity was established before building the model. Considering that to train and test a classification model the data are divided into two sets, training and test [17], the cross-validation method was used since it reduces the dependence of the experiment's results on the way the division is made [12]. For this particular case, the n-fold cross-validation evaluation method was used. In this method, the training set is randomly divided into *n* disjoint subsets of similar size called folds. The number of subsets can be entered into the field Folds. Subsequently, *n* iterations (equal to the number of subsets) are made, where a different subset is reserved for the testing set for each iteration and the remaining *n-1* (merging all the data) for the construction of the model (training). The partial sampling error of the model is calculated in each iteration. Finally, the model is constructed with all the data, and its error is obtained by averaging those calculated previously in each iteration. Another advantage of cross validation is that the variance of the *n* partial sampling errors allows estimating the learning method's variability regarding the data set. This research used the 10-fold cross validation considering the recommendation by Hernández et al. [12].

The classifier cost for the *sbpro_lec_2052A23* repository was estimated through a confusion matrix during the evaluation stage. The confusion matrix represents in detail the number of instances predicted by class. The sum of the records presented in each row i, i = 1...n constitutes the number of instances that genuinely belong to class i. Similarly, the summation of the examples or records in each column j, j =

1...n is the instances predicted by the algorithm for the j value of the class. The values on the diagonal are the correct matches, and the rest are the classification errors (examples that belonged to the class i of the row i and were classified incorrectly in another) [12].

Furthermore, the discovered patterns were evaluated to determine their validity, remove the redundant or irrelevant patterns, and interpret the patterns useful in terms of being understandable for the user.

In the deployment stage, the discovered patterns were documented, and these can be incorporated into the existing knowledge on academic performance in the Critical reading competency of the students of professional programs in Colombia. The directors of the Universidad Javeriana Cali are responsible for integrating this knowledge into their decision-making processes to improve the education quality of this institution.

III. Results

The purpose of selecting the decision tree classification technique is to obtain a model that can predict the socioeconomic, academic, and institutional factors associated with good (above the mean) or low (below the mean) academic performance in the critical reading section of the Saber Pro exam for new students of the Universidad Javeriana Cali, considering as class attribute the score obtained in this test. To achieve this, several decision tree algorithms were evaluated with the tool Weka, which allowed to select the technique that classified with greater accuracy the *sbpro_lec_2052A23* data set. Results are presented in Table 1.

Table 1. Evaluation of different decision tree techniques.

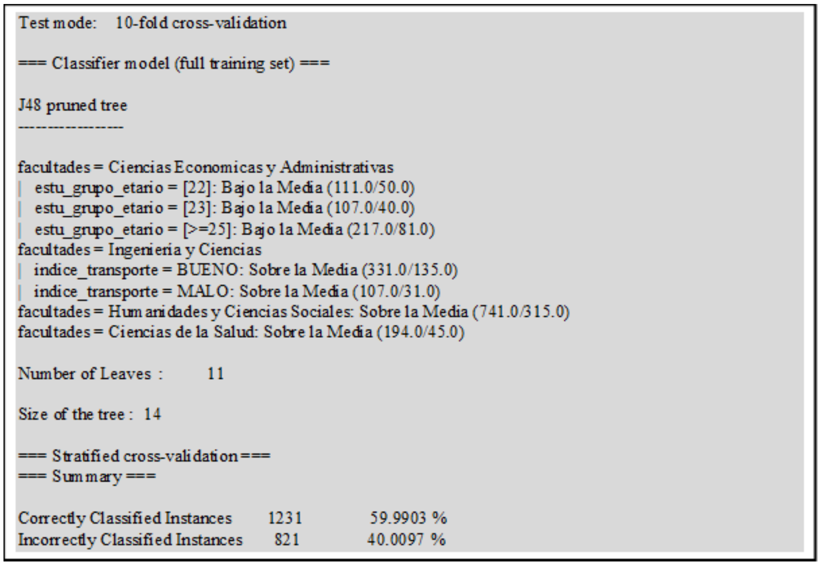
Algorithm	Accuracy
Decision Stump (one-level decision tree)	53.02%
J48	68.85%
LMT (Logistic Model Tree)	62.37%
Random Forest	57.89%
Random Tree	53.89%
RepTree	55.50%

According to Table 1, the most accurate algorithm was J48. That is why this algorithm was selected for the construction of the classification models with a decision tree. Once the algorithm and method for the testing and training of the models were selected, the decision trees and the J48 algorithm, which implements algorithm C. 45 [18], were built with Weka, see 3.9.4 [17]. The J48 algorithm is based on the usage of the information gain criteria. In this way, it is possible to ensure that the variables with a higher number of possible values are not benefitted in the selection. Additionally, the algorithm includes a classification tree pruning once it has been inducted. The most crucial parameter considered in the pruning was the confidence level C, which affects the size and prediction capability of the tree built. The lower this probability, a more significant difference in the prediction errors before and after pruning is required not to prune. The default value for this factor is 25%, and, as this value decreases, more pruning operations are allowed and thus, smaller trees are obtained [19]. Another parameter used to vary the size of the tree was the factor M, which specifies the minimum number of instances or records per tree node. The global score obtained by the students in the Saber Pro exams was selected as a class, which was discretized in the values "above national mean" and "below national mean".

Different decision tree models were generated to choose the decision tree that best classified the students and with the highest interpretability level of the patterns associated with academic performance in Critical reading. This is why two values were set for the confidence factor C, 25% and 50%, combined with two values for factor M, 2.5% (52 examples) and 5% (104 examples). Furthermore, a post-pruning process was implemented to maintain the most representative branches, and therefore the rules, which are those that exceed a minimum support of 5% and a confidence of 60%.

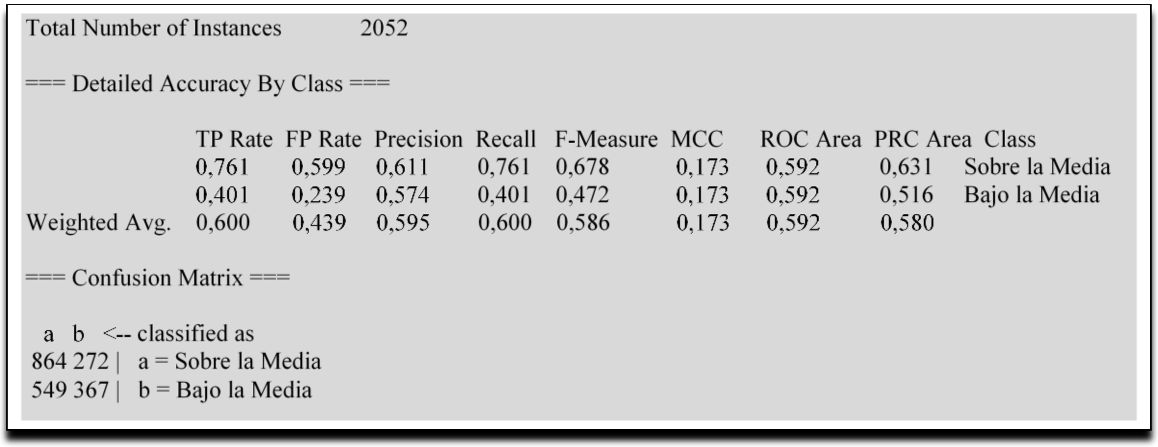
The best tree was built with the parameters C=0.25 and M=104 for pre-pruning and a support greater than or equal to 5% for post-pruning. Figure 1 presents the classification tree obtained.

The confusion matrix (also called contingency table) was used to evaluate or estimate the cost of the classification model built. A confusion matrix is a tool that allows visualizing the performance of a supervised learning algorithm. This is shown in Figure 2.



```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

facultades = Ciencias Economicas y Administrativas
|   estu_grupo_etario = [22]: Bajo la Media (111.0/50.0)
|   estu_grupo_etario = [23]: Bajo la Media (107.0/40.0)
|   estu_grupo_etario = [>=25]: Bajo la Media (217.0/81.0)
facultades = Ingenieria y Ciencias
|   indice_transporte = BUENO: Sobre la Media (331.0/135.0)
|   indice_transporte = MALO: Sobre la Media (107.0/31.0)
facultades = Humanidades y Ciencias Sociales: Sobre la Media (741.0/315.0)
facultades = Ciencias de la Salud: Sobre la Media (194.0/45.0)

Number of Leaves  :       11

Size of the tree :   14

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1231               59.9903 %
Incorrectly Classified Instances       821               40.0097 %
```

Fig. 1. Classification model for critical reading obtained with Weka.



```
Total Number of Instances            2052

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC     ROC Area  PRC Area  Class
                 0,761    0,599    0,611      0,761   0,678      0,173   0,592     0,631     Sobre la Media
                 0,401    0,239    0,574      0,401   0,472      0,173   0,592     0,516     Bajo la Media
Weighted Avg.    0,600    0,439    0,595      0,600   0,586      0,173   0,592     0,580

=== Confusion Matrix ===

   a   b   <-- classified as
 864 272 |   a = Sobre la Media
 549 367 |   b = Bajo la Media
```

Fig. 2. Confusion matrix of the classification model for critical reading.

IV. Discussion and Conclusions

After analyzing the decision tree results of the performance in the critical reading competency of the students from the Universidad Javeriana Cali in the Saber Pro exams in 2017 and 2018 presented in Figure 1, it can be observed that the tree classified 1231 instances correctly, which corresponds to an accuracy of 60%, and 821 instances were classified incorrectly, corresponding to a 40%.

When evaluating the model with the confusion matrix in Figure 2, obtained with the tool Weka, it predicts correctly 864 cases of students whose performance in critical reading is above the mean (VP) and 367 cases below the mean (VN). On the other hand, the model classifies incorrectly as below the mean (FN) 272 cases whose performance is above the mean, and 549 cases as above the mean (FP) whose performance is below the mean.

For the case where students are above the mean in critical reading performance, the model has a prediction accuracy of 0.611, which means that, of the total cases predicted above the mean, 61% are correct. The model's sensitivity (TPR) and recall is 0.761, indicating that the model correctly classifies 76.1% of the students that indeed are above the mean. On the other hand, the model's false positive rate is 0.599, meaning that 59.9% of the students below the mean were classified as above the mean. The F-measure is 0.678, which means that the harmonic mean between the accuracy and recall of those above the mean is 67.8%. When combining these measures, a better performance of the model is appreciated for those above the mean.

For the case where students are below the mean in critical reading performance, the model has a prediction accuracy of 0.574; that is, of the total cases predicted below the mean, 57.4% are correct. The model's specificity (TNR) and recall is 0.401, indicating that the model correctly classifies 40.1% of the students who genuinely are below the mean. Furthermore, the false-negative rate of the model is 0.239, meaning that 23.9% of the students above the mean were classified as below the mean. The F-measure is 0.472, which means that the harmonic mean between the accuracy and recall of those below the mean is 47.2%. When combining these measures, a poorer performance of the model is appreciated for those below the mean.

The model built to detect performance patterns in the critical reading competency of the Saber Pro exam of students from Universidad Javeriana Cali is not highly unbalanced. There is a 220 (10%) cases difference between those that were above the mean (55%) and those below the mean (45%). For this reason, within the evaluation metrics calculated previously, it can be said that this model has an accuracy of 60%, and it is better at predicting students above the mean than those below. This is also noticeable in the relation between the recall and accuracy given in the PRC area, where it is 0.631 for the students above the mean and 0.516 for those below the mean. Furthermore, the Mathews correlation coefficient of the model is 0.173, indicating a weak relationship between the prediction and the observed, that is, a low quality in prediction. Finally, regarding the areas, since the ROC area of the model is above 0.5, the model has a good performance in the classification of the students from Universidad Javeriana Cali regarding the performance in the critical reading competency of the Saber Pro exams of 2017 and 2018.

AUTHORS' CONTRIBUTION

Andrea Timarán-Buchely: Data curation, Formal Analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Silvio-Ricardo Timarán-Pereira: Formal Analysis, Investigation, Writing – review & editing.

Arsenio Hidalgo Troya: Formal Analysis, Investigation, Writing – review & editing.

REFERENCES

[1] Icfes, Saber Pro: Módulos de Competencias Genéricas 2017. Instituto Colombiano para la Evaluación de la Educación Superior, Bogotá D.C., Colombia, 2017. https://www.icfes.gov.co/documents/20143/495161/Guia%20de%20orientacion%20modulos%20de%20competencias%20genericas-saber-pro-2017.pdf

[2] Icfes, Guía de orientación Saber Pro: Módulos de competencias genéricas, Bogotá D.C., Colombia, 2018. https://www.icfes.gov.co/documents/20143/496194/Guia%20de%20orientacion%20modulos%20de%20competencias%20genericas-saber-pro-2018.pdf

[3] R. Timarán, I. Hernández, J. Caicedo, A. Hidalgo, J. Alvarado, *Descubrimiento de patrones de desempeño académico*, Bogotá, Colombia: Ediciones Universidad Cooperativa de Colombia, 2016. DOI: https://doi.org/10.16925/9789587600490

[4] Icfes, Informe nacional de resultados Saber Pro 2012-2015, Bogotá D.C., Colombia, 2016. https://www.icfes.gov.co/documents/20143/194324/Informe%20nacional%20de%20resultados%20saber%20pro%202012%20-%202015.pdf

[5] L. Zapata, *Factores académicos asociados al bajo rendimiento en inglés en las pruebas ECAES presentadas por los estudiantes de la Facultad de Educación en el año 2009*, Grade Thesis, Fundación Universitaria Luis Amigó, Medellín, Colombia, 2011

[6] UNAL, *Análisis de los resultados obtenidos por la Universidad Nacional de Colombia sede Bogotá en las pruebas Saber Pro 2011–2*, Bogotá D.C., Colombia, 2012. https://www.unal.edu.co/diracad/evaluacion/SaberPro_2012/analisis_de_resultados.pdf

[7] R. Timarán, A. Calderón, J. Jiménez, *Detección de Patrones de Deserción Estudiantil con Minería de Datos*, San Juan de Pasto, Colombia: Editorial Universidad de Nariño, 2017

[8] S. Valero, A. Vargas, M. Alonso, *Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos*, 2005. http://fcaenlinea.unam.mx/anexos/1566/1566_u6_act1b.pdf.

[9] 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

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

[11] 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

[12] J. Hernández, M. Ramírez, C. Ferri, *Introducción a la Minería de Datos*. Madrid, España: Editorial Pearson Educación S.A., 2005

[13] J. Han, M. Kamber, *Data Mining: Concepts and Techniques. San Francisco*, USA: Morgan Kaufmann Publishers, 2001

[14] K. Sattler, O. Dunemann, "*SQL Database Primitives for Decision Tree Classifiers*," in *10th ACM International Conference on Information and Knowledge Management,* pp. 379-386, 2001

[15] R. Timarán, J. Caicedo, A. Hidalgo, *Aplicación de la minería de datos en la detección de patrones de desempeño académico en las pruebas Saber Pro,* San Juan de Pasto, Colombia: Editorial Universidad de Nariño, 2021

[16] E. Hernández, R. Lorente, Minería de datos aplicada a la detección de Cáncer de Mama. Universidad Carlos III, Madrid, Spain, 2009. http://www.it.uc3m.es/jvillena/irc/practicas/08-09/14.pdf

[17] I. Witten, E. Frank, M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition).* New York, USA: Morgan Kaufmann, 2011. DOI: https://doi.org/10.1016/C2009-0-19715-5

[18] J. R. Quinlan, *Programs for Machine Learning*. San Francisco, USA: Morgan Kaufmann Publishers, 1993

[19] 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