Machine Learning for the identification of students at risk of academic desertion

Leidy Daniela Forero Zea, Yudy Fernanda Piñeros Reina and José Ignacio Rodríguez Molano

Universidad Distrital Francisco José de Caldas, Bogotá, Colombia
{ldforeroz, yfpinerosr}@correo.udistrital.edu.co,
jirodriguezm@udistrital.edu.co

Abstract. In Latin America, desertion rates in higher education range between 40% and 75%. There are many reasons for a student to deserted of their studies. However, the importance of identifying the level of risk related to such desertion is reflected in the socio-economic impact for the institutions as well as for the country. Technological advancements in database management and artificial intelligence have led to the development of techniques such as Machine Learning, which supports decision-making when facing a problem and adapts accordingly to the required conditions.

The following article shows a case study of the identification of students in Industrial Engineering at risk of dropping out in the Universidad Distrital Francisco José de Caldas from the 2003-1 to 2018-1 academic semesters. The algorithm is selected based on which is more suitable to the nature of data, through the comparison of automated learning techniques in Azure Machine Learning Studio.

Keywords: Academic Desertion, Algorithm, Artificial Intelligence, Machine Learning.

1 Introduction

The approach for student desertion and retention involves variables on individual, institutional and family-related levels, which on top of that are classified in terms of psychological, economic, sociological, organizational and interactional aspects [1] [4]. This implies handling large amounts of information that require the use of technologies with high, medium or low complexity depending on the nature of the study. From this standpoint, Artificial Intelligence offers a variety of techniques for database management and analysis including Machine Learning.

Machine Learning, emerges as an artificial intelligence method derived from computing programs that access data and uses it to learn and predict results. These decision-making methods are assessed and receive feedback so that they can determine the algorithm that is more suitable for the type of data and response related to the research.

2 State of the art

2.1 Academic desertion in Colombia and context of the Universidad Distrital

Academic desertion can be defined as the condition of those people enrolled in higher education that abandon the institution during two or more consecutive periods at the time of their studies [5].

According to data gathered in 2015 by the Colombian System for the Prevention of Desertion in Higher Education Institutions (by its acronym in Spanish SPADIES) in the statistical report of desertion and graduation of the Colombian Ministry of National Education and considering the desertion rate by cohort and by training level, the undergraduate students that deserted represent 41.60% out of the total number of students that enroll in an academic period on a national scale and 45.25% on a district scale [6].

On another note, the last report "Statistics of permanence, graduation and desertion of students in the Engineering Faculty in undergraduate programs from 2009 to 2017" carried out by the Consultant System Office in the Universidad Distrital Francisco José de Caldas (UDFJC), 53.8% of the faculty students between 2009 and 2017, either deserted or lost their student status [11]. Furthermore, specific information on desertion for the curricular program of Industrial Engineering indicates that the desertion percentage in the assessed periods is equivalent to 48% [11].

2.2 Machine Learning and classification of its algorithms

Machine Learning is a form of artificial intelligence where large volumes of data are accessed and interpreted, the system is trained and new information is predicted through learning algorithms. These techniques are classified according to various criteria: the type of learning, the tasks dictated to the algorithm and the types of models used.

Classification is a task in which an individual of the system intends to determine to which class it belongs according to the process of learning characteristics, patterns and behaviors that other individuals have previously adopted and their documented records [10]. Thus, these algorithms are divided into two categories: binary classification and multiclass classification. Binary classification refers to those algorithms whose observation result must be catalogued as either positive or negative. Therefore, classification depends on a threshold used to compare the generated score for each iteration of the algorithm. The multiclass classification differs in the sense that it predicts the classes with the highest scores [3].

Binary classification algorithms. This article will use the binary classification learning algorithms, since the objective of the development of the case study is to predict a desertion or no desertion label for a student whose values of the variables allow their classification; these algorithms are Averaged Perceptron, Bayes Point Machine, Boosted Decision Tree, Decision Forest, Decision Jungle, Logistic Regression and Two-Class Neural Network[2][9].

3 Methodology for the application of classification models using ML

3.1 Definition of the computational tool to use.

There are many options for commercial Machine Learning software. However, in this case, it is suggested that it is directed towards the predictive model as well as understandable and easy to use, which allows the manipulation of data and the types of results that it can generate.

Next, a comparative table is presented including the advantages, functionality and generalities of each tool keeping in mind the previously mentioned criteria (See Table 1).

Table 1. Characteristics of Machine Learning software. Source: Author.

Software	Description	Functions
Azure Machine Learning Studio	Collaborative solution with drag-and-drop interface developed to create and implement predictive analytics solutions in minutes. Designed for applied machine learning.	-Predictive Modeling
Google Cloud ML Engine	Engine Managed service that provides a balanced, scalable and automatic predictive training compiled in mathematical models that allow to understand the information extracted from the data set.	-Deep learning -Model formation -Predictive modeling
AWS	It allows to create, train technically and implement deep learning models quickly and easily, with high performance automatic learning algorithms; includes data storage, business intelligence, batch processing, transmission processing and organization of data workflows.	-Self-learning -Machine Learning algorithm library -Model formation

It is determined that the Azure Machine Learning Studio from Microsoft is used whose predictive model function is required based on the expected response and handled database. Additionally, it is code-free which facilitates the understanding of applied automated learning and includes a free trial version so any user can access it.

3.2 Used database and determination of variables

Considering the concepts and trends mentioned in Machine Learning and the goal of identifying students at risk of Academic desertion of the Industrial engineering program of the District University Francisco José de Caldas, Bogotá, Colombia. The following sample was taken:

Sample Description. The software training process is done with a student status database of the program of Industrial Engineering between 2003 and 2018 (Active student, graduate, suspended, sanctioned, retired); which was provided by the UDFJC for

academic purposes, and They do not contain personally identifiable information about people, as recognized by Habeas data. A database treatment was required, from which unnecessary, incomplete and inconsistent data were eliminated; Bearing in mind that the required information is specifically that of the students who left the university, either as a graduate student who is classified as not abandoned or by desertion or abandonment, leaving a total sample of 3201 data, each one 24 academic and social variables, which are presented below. (see Table 2).

Table 2. Description of the database variables. Source: Author.

Variable	Id.	Definition
Student code	CE	Identification number within the institution. Numeric character string
State	ES	Characterizes the student in the state Abandoned, which covers sus pended students, deserted or who did not pass academic test, or No Abandoned that refers to graduate students.
Sex	SX	Gender of the student: Male (M), Female (F)
Stratum	ST	Socioeconomic classification. Integer value between 0 and 6
Age of Entry	EI	The age at which the student enters the first semester
Type of Inscription	TI	Classification of the inscription: Normal, Displaced, Indigenous, External transfer.
Average	PR	Accumulated average of the career until the last semester taken. Value between 0 and 5.
Number of Academic Tests	PN	State at risk of losing student quality. Value between 0 and 4
Approved Subjects	EA	Number of subjects taken and passed. Whole value between 0 and the number of subjects in the academic program
Failed Subjects	ER	Number of subjects studied, but not approved
Score ICFES	PI	Total score obtained in the Saber test 11. Weighted average of the scores in the five (5) Areas. Value between 0 and 100

Variable	Id.	Definition
Biology	B	Score obtained in the Saber 11 test for each subject. Value between 0 and 100
Chemistry	Q	
Physics	F	
Social	S	
Verbal Aptitude	AV	
Spanish and Literature	EL	
Mathematics Abilities	AM	
Mathematical Knowledge	CM	
Philosophy	F	
History	H	
Geography	G	
Foreign language	IE	
Interdisciplinary	I	

The sample was submitted to a statistical analysis of independence and homogeneity on the average variable of students who deserted to verify the validity of the data, which were approved with a correlation index of 0.14 and 0,496 for test Kruskall Wallis with a confidence level of 95%, respectively.

Based on the analysis of variables, their level of correlation is established through the Pearson coefficient assessed in the software. The highest correlation indexes delivered by the Pearson correlation can be seen in Table 3:

Table 3. Pearson correlation coefficients. Source: Author.

CI Pearson	Average	Stratum	Age of Entry	N. of Ac. Tests	Approved Subjects	Failed Subjects	Score ICFES	Math. Knowledge	Math. Abilities
Average	1.000	0.113	0.176	0.035	0.757	0.003	0.290	0.193	0.180
Stratum	0.124	1.000	0.058	0.057	0.096	0.056	0.087	0.021	0.004
Age of Entry	0.176	0.057	1.000	0.080	0.179	0.076	0.174	0.042	0.036
N. of Ac. Tests	0.035	0.056	0.080	1.000	0.055	0.881	0.055	0.137	0.015
Approved Subjects	0.757	0.087	0.179	0.055	1.000	0.033	0.285	0.148	0.136
Failed Subjects	0.003	0.049	0.076	0.881	0.033	1.000	0.002	0.131	0.022
Score ICFES	0.290	0.060	0.174	0.055	0.285	0.002	1.000	0.041	0.467
Math. Knowledge	0.193	0.021	0.042	0.137	0.148	0.131	0.041	1.000	0.384
Math. Abilities	0.180	0.003	0.036	0.015	0.136	0.022	0.467	0.384	1.000

However, the algorithm that is more suitable to the needs of the case study is chosen with the purpose of predicting when a student is at risk of dropping out of the Industrial Engineering program of the UDFJC given that the term desertion includes those students that entered the program as such yet did not culminate their careers.

3.3 Determination of performance metrics of the selected algorithms

In order to validate the performance of the implemented algorithms, the crossed validation tool provided by the software used which generates assessment metrics (accuracy, precision, recall and F1 score) for binary classification algorithms. Additionally, the duration of the experiment is also considered as well as the factors generated by the confusion matrix which includes the percentage of true positives, true negatives, false positives and false negatives. These variables are understood as a whole to determine the assertiveness of the algorithms and then choose the best conditions to predict the classification of the status of students. Then (Table 4), the definition of these measures is presented.

Table 4. Definition of the assessment metrics. Source: Author.

Measure	Definition	Formula
Accuracy (A)	Is the ratio of true results to total cases. Measure the goodness of a classification model	$A = \frac{TP + TN}{\text{Total of cases}}$
Precision (P)	It is the proportion of true results on all positive results.	$P = \frac{TP}{TP + FP}$

Recall (R)	It is the relation of all the correct results returned by the model.	$R = \frac{TP}{TP + FN}$
F1 Score (F1S)	It is the weighted average of Precision and Recall. The summary of the evaluation is considered.	$F1S = \frac{2 * (R * P)}{R + P}$
Duration of the experiment (T)	It is the time it takes the model in the training of the database. It is measured in seconds.	
True positives (TP)	Number of cases of No Abandonment, whose prediction was "No abandonment".	
Negative positives (TN)	Number of cases of No Abandonment, whose prediction was "Abandonment".	
False positives (FP)	Number of cases of Abandonment, whose prediction was "Abandonment".	
False negatives (FN)	Number of cases of Abandonment, whose prediction was "No Abandonment".	

3.4 Application of the different algorithms to the case study

Based on the objective of the case study, it is stated that according to the desired response the algorithm must be a binary (two-class) classification (supervised learning) since the program needs to choose between two response options: Desertion or Non-desertion. To determine the appropriated algorithm within the Binary Classification, a software test is performed as shown in Figures 1 and 2.

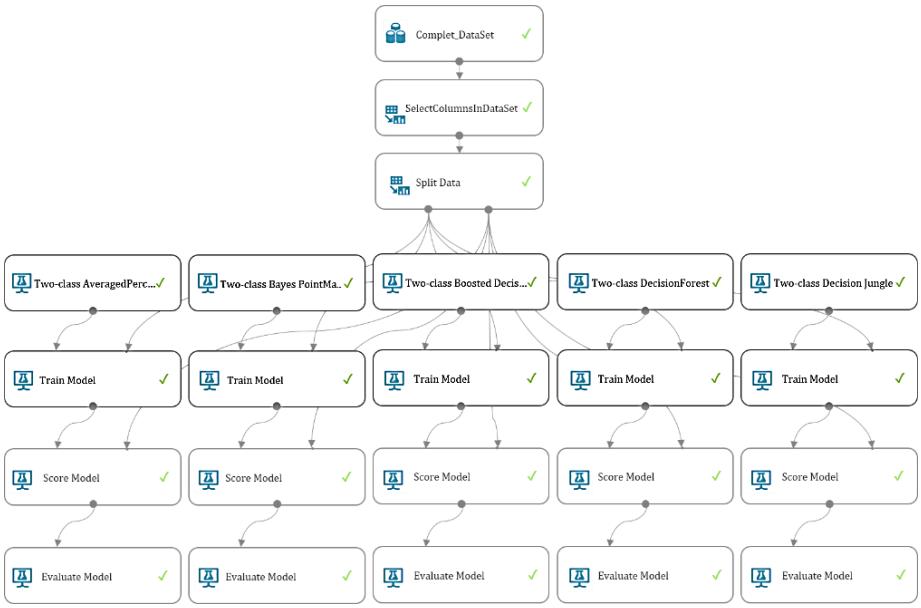

Figure 1. Assessment structure of the algorithms. Part a) Source: Author.

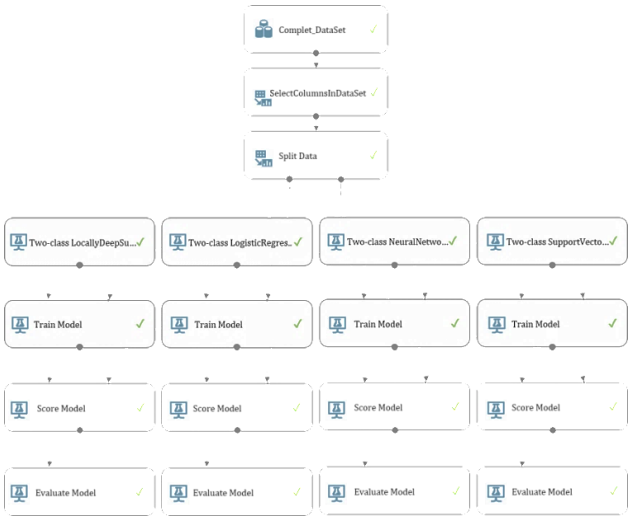

Figure 2. Assessment structure of the algorithms. Part b) Source: Author.

The assessment scheme for the algorithms shown in Figure 1 and Figure 2 consists of a software tool that can find the value of the Assessment Criteria. Afterwards, the Select Columns module is used to choose the variables that will be a part of the software training. In Split Data, the proportion of data used for training is determined. The subsequent modules refer to the learning, training and result assessment algorithms. For the testing process, each algorithm is trained for the Two-Class Classification based on the status variable with a significant sample of 80% of the data and the remaining 20% is assessed later. The obtained results are shown in Table 5.

Table 5. Comparison of the measurements for the assessment of algorithms. Source: Author.

Algorithm	TP	TN	FP	FN	A	P	R
Averaged Perceptron	0.879	0.955	0.121	0.045	0.919	0.879	0.947
Bayes point machine	0.852	0.963	0.148	0.037	0.908	0.852	0.958
Boosted decision tree	0.903	0.965	0.097	0.035	0.936	0.903	0.958
Decision forest	0.899	0.956	0.101	0.044	0.93	0.899	0.947
Decision jungle	0.891	0.982	0.109	0.018	0.938	0.891	0.979
Locally deep support vector machine	0.890	0.979	0.110	0.021	0.936	0.89	0.975
Logistic Regression	0.850	0.963	0.150	0.037	0.906	0.85	0.958
Neural network	0.870	0.975	0.130	0.025	0.923	0.87	0.972

Using the previous results, it is determined that the learning algorithm that is more suitable for the nature of data and the type of response is a Two Class Boosted Decision

Tree. Establishing an equitable weighing strategy for each criteria, this represents a larger set of characteristics that enable proper adjustment and minimization of prediction errors.

3.5 Performance assessment of the chosen algorithm

Aiming to improve the performance of the selected algorithm, five tests are carried out (See Table 6) with different groups of variables (Abbreviations Table 2), where the X represents the variables chosen from the database that are included in the algorithm. Since the status is the main variable used to train all algorithms, it is not included in the tests.

Table 6. Comparison matrix of attribute addition. Source: Author.

TEST	SX	ST	—	TI	PR	NP	EA	ER	PI	B	Q	F	S	AV	EL	AM	CM	F	H	G	IE	I
1		X			x	x	x	x	x							x	x					
2	x	x	x		x	x	x	x	x							x	x					
3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	X	x	x
4					x	x			x													
5	x		x	x	x	x		x	x							x	x					

Table 7 shows a summary of the Assessment Criteria per test, which is useful to identify the variables that have a significant impact on the results of Accuracy, Precision, Recall and F1Score.

Table 7. Test comparison of Assessment Criteria.

Test	Accuracy	Precision	Recall	F1Score
1	0.994	0.989	0.996	0.993
2	0.995	0.99	1	0.995
3	0.994	0.986	1	0.993
4	0.911	0.887	0.915	0.901
5	0.938	0.909	0.954	0.931

It was confirmed that the variables with the highest correlation such as Average (PR), Approved Status (EA) and Number of Tests (NP), Failed Subjects (ER) must necessarily be included in the test and generates the highest values in the assessment indexes. In contrast, other variables such as the areas of knowledge do not have a significant impact in the improvement of results.

4 Results of the predictive experiment.

The chosen learning algorithm is used to create the Training Experiment (Figure 3) which comes from the selection of the previous variables based on the determination of

the status and the use of the Two-Class Boosted Decision Tree as an ordered set of systematic operations that the algorithm uses to determine the result.

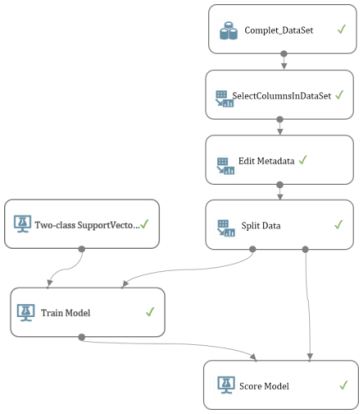

Figure 3. Training Experiment. Source: Author.

The Predictive Experiment (Figure 4) shows the final structure of the software modelling trained with 80% of the data as well as the algorithm that is more suitable for the characteristics and the type of response. Additionally, the Web Service Input module requests information to the user to deliver the corresponding prediction. These requirements are the group of variables in test (Table 6) that are more appropriate according to the assessment criteria (Table 7).

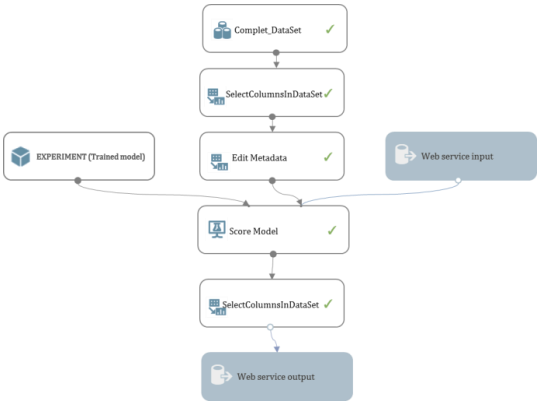

Figure 4. Predictive experiment. Source: Author.

With the prediction module created, an example can be tested with the data in Table 8.

Table 8. Data for the practical example. Source: Author

Variable	Val.	Variable	Val.	Variable	Val.
Sex	M	Score ICFES	303	Mathematical Knowledge	66
Stratum	1	Biology	51	Philosophy	0
Age of Entry	20	Chemistry	61	History	0
Type of Inscription	Normal	Physics	56	Geography	0
Average	3.0	Social	50	Foreign language	0
Number of Academic Tests	1	Verbal Aptitude	57	Interdisciplinary	0
Approved Spaces	42	Spanish and Literature	56		
Failed Spaces	15	Mathematics Abilities	68		

Figure 5 presents the result of the practical example where the student deserts of the university under the given characteristics.

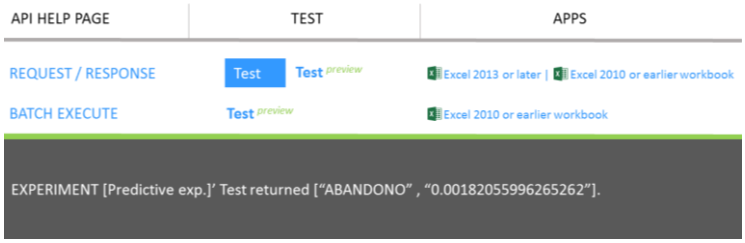

Figure 5. Results of the practical example. Source: Author.

5 Conclusions.

The data of the practical example is used to determine that the student deserts of university without specifying the academic or social variable that the student quits from. In this case, the student has passed 26% of the subjects in his syllabus and a GPA of 3.0, which classifies him as a student in academic probation. Since these variables have stronger weights within the prediction model, it is valid to state that the desertion is due to low academic performance of the student who decides to deserted. This analysis would lead to promote strategies that reduce the academic hardships of students by offering subjects during mid-year breaks, more scholarships to students with better GPAs, among others. It is noteworthy to highlight that there are many academic or social reasons that lead a student to deserted. However, the previous analysis of predictions establishes more effective retention mechanisms within higher education institutions. Learning techniques such as Machine Learning consider both the main and secondary variables, as well as their normal and abnormal values through previous training. In this scenario, the responses obtained are reliable with a precision of 90.3% and an accuracy of 93.6%.

At last after developing this case study, we identify Machine Learning tools that can be used in business, commercial, financial, industrial, academic, scientific and other fields. However, from an educational perspective and in order to predict whether a student will desert or not, with the use of Azure Machine Learning Studio; from the identification of the relevant variables; the treatment of the database is done by filtering the useful information for the study; the program is trained with the historical behavior of the variable (s) it seeks to predict; the current information is entered and the results are analyzed in order to make decisions and action against the possible scenarios. This algorithm can be replicated with different objectives and applied in multiple areas that require the prediction of random variable behaviors.

References

1. Chung, J.Y., Lee, S.: *Desertion early warning systems for high school students using machine learning*. Child. Youth Serv. Rev. (2018).
2. Copeland, M. et al.: *Microsoft Azure Machine Learning.* In: Microsoft Azure. (2015).
3. Guides, A.M.D.: *Amazon Machine Learning*.
4. Himmel, E.: *Modelo de análisis de la deserción estudiantil en la educación superior. Calidad. en la Educación. (Himmel, E.: Model for the analysis of student desertion in higher education.)* 0, 17, 91 (2018).
5. Melo-Becerra, L.A. et al.: *La educación superior en Colombia: situación actual y análisis de eficiencia. Desarro. Soc*. (Melo Becerra, et al.:*Higher education in Colombia: current situation and efficiency analysis.*) 59–111 (2017).
6. Ministerio de Educación: *Estádisticas de deserción y graduación*. (Colombian Ministry of National Education: *Desertion and graduation statistics*) (2016).
7. Ministerio de Educación Nacional: *Deserción estudiantil en la educación superior colombiana.* (Colombian Ministry of National Education: *Student desertion in Colombian higher education.*). (2009).
8. Ministerio de Educación Nacional.: *Estadisticas de Educacion Superior*, (Colombian Ministry of National Education: S*tatistics of higher education*) (2016).
9. Sisodia, D., Sisodia, D.S.: *Prediction of Diabetes using Classification Algorithms.* Procedia Comput. Sci. 132, 1578–1585 (2018).
10. Tan, M., Shao, P.: *Prediction of Student Desertion in E-Learning Program Through the Use of Machine Learning Method.* Int. J. Emerg. Technol. Learn. 10, 1, 11–17 (2015).
11. Universidad Distrital: *Estadística de la permanencia, graduación y deserción de los estudiantes en la Facultad de Ingeniería en programas de pregrado 2009-2017.* (Universidad Distrital: S*tatistics on the permanence, graduation and desertion of students in the Faculty of Engineering in undergraduate programs 2009-2017*.)(2018).
12. Navarro, C. A. T., & Neira, J. A. C. (n.d.). *Design of Expert System for Decision Making in Materials Purchasing*. Retrieved from http://www.scielo.org.co/pdf/cuadm/v30n52/v30n52a03.pdf
13. Hidalgo, L. A. (n.d.). *Artificial Intelligence*. https://doi.org/M-26913-2004.

14. Alberto Ruiz Marta Susana Basualdo, C., & Jorge Matich, D. (n.d.). *Cátedra: Informática Aplicada a la Ingeniería de Procesos-Orientación I Redes Neuronales: Conceptos Básicos y Aplicaciones*. (Alberto Ruiz Marta Susana Basualdo, C., & Jorge Matich, D. *Chair: Computer Science Applied to Process Engineering-Orientation I Neural Networks: Basic Concepts and Applications*.)Retrieved from https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf
15. Gómez, J., Sánchez, J., & William Restrepo, J. (2012). *Aplicación de Redes Neuronales en la Clasificación de Arcillas*. *Revista EIA* (Vol. 17). (Gómez, J., Sánchez, J., & William Restrepo, *Application of Neural Networks in the Classification of Clays*)Retrieved from http://www.scielo.org.co/pdf/eia/n17/n17a14.pdf
16. Garcia, M. R., & Rodríguez, J. E. R. (2004). Sistemas Basados En El Conocimiento. *Revista Vínculos*, (*Knowledge-Based Systems*. Vínculos Magazine) *1*(1), 37–44. https://doi.org/10.14483/2322939X.4070
17. Gorges, C., Öztürk, K., & Liebich, R. (2019). *Impact detection using a machine learning approach and experimental road roughness classification*. https://doi.org/10.1016/j.ymssp.2018.07.043
18. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). *Feature selection in machine learning: A new perspective. Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077
19. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). *Machine learning: a review of classification and combining techniques. Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3
20. Badaró, S., Javier Ibañez, L., & Agüero, M. J. (n.d.). *Sistemas Expertos: Fundamentos, Metodologías y Aplicaciones*. (*Expert Systems: Fundamentals, Methodologies and Applications*) Retrieved from https://www.palermo.edu/ingenieria/pdf2014/13/CyT_13_24.pdf
21. Goldberg, D. E. (David E., & E., D. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Longman Publishing Co., Inc. Retrieved from https://dl.acm.org/citation.cfm?id=534133
22. Moreno -Eva, A., Béjar, A.-J., Belanche, L., Cortés, U., Gavaldà, R., Manuel, J., … Sànchez, M. (n.d.). *Aprendizaje automático*. Retrieved from www.edicionsupc.es
23. Kim, I., Choi, H. J., Ryu, M., Lee, K., Yu, J. H., Kim, W., … Lee, J. E. (2018). A predictive model for high/low risk group according to oncotype DX recurrence score using machine learning. https://doi.org/10.1016/j.ejso.2018.09.011
24. Patricia, S., Moreno, B., & Támara, L. G. (2017). Acercamiento a la deserción estudiantil desde la integración social y académica (Approach to student desertions from the perspective of social and academic integration). *Revista de La Educación Superior*, *46*(183), 63–86. https://doi.org/10.1016/j.resu.2017.05.004