EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Classification model for student dropouts using machine learning: A case study

Henry Villarreal-Torres [1], Julio Ángeles-Morales [1], William Marín-Rodriguez [2,3,*], Daniel Andrade-Girón [2], Jenny Cano-Mejía [1], Carmen Mejía-Murillo [1], Gumercindo Flores-Reyes [1], Manuel Palomino-Márquez [1,4]

[1] Universidad San Pedro. Chimbote, Perú.
[2] Universidad Nacional José Faustino Sánchez Carrión. Huacho, Perú.
[3] Universidad Cesar Vallejo. Los Olivos, Lima, Perú.
[4] Seguro Social de Salud - EsSalud. Lima, Perú.

Abstract

Information and communication technologies have been fulfilling a highly relevant role in the different fields of knowledge, addressing problems in various disciplines; there is an increased capacity to identify patterns and anomalies in an organization's data using data mining; In this context, the study aimed to develop a classification model for student dropout, applying machine learning with the autoML method of the H2O.ai framework; the dimensionality of the socioeconomic and academic characteristics has been taken into account, with the purpose that the directors make reasonable decisions to counteract the abandonment of the students in the study programs. The methodology used was of a technological type, purposeful level, incremental innovation, temporal scope, and synchronous; data collection was prospective. For this, a 20-item questionnaire was applied to 237 students enrolled in the master's degree programs in the education of the Graduate School. The research resulted in a supervised machine learning model, Gradient Reinforcement Machine (GBM), to classify student dropout, thus identifying the main associated factors that influence dropout, obtaining a Gini coefficient of 92.20%, AUC of 96.10% and a LogLoss of 24.24% representing a model with efficient performance.

Keywords: autoML; machine learning; Student dropout; higher education; H2O.ai; data mining.

Received on 21 November 2022, accepted on 6 June 2023, published on 15 June 2023

doi: 10.4108/eetsis.vi.3455

*Corresponding author. Email: wmarin@unjfsc.edu.pe

1. Introduction

Education is essential for the development and well-being of a society; therefore, students are the raison d'être for any educational institution. A country's social and economic growth is directly related to the academic performance of its students (Mushtaq and Khan, 2012). In the last decade, the Peruvian state has implemented various measures for quality assurance in university higher education to guarantee that the country's youth have access to a comprehensive and continuous educational service that promotes development through research.

In 2014, with the publication of University Law No. 30220, the National Superintendence of Higher University Education (SUNEDU, in Spanish) was created, an organization that has been playing a leading role in compliance with primary quality conditions by educational institutions during the institutional licensing process. Faced with the requirement to implement primary quality conditions in the higher education system, universities must manage their resources efficiently. In this sense, it is an excellent option to manage information technologies in the

university higher education system, according to the proposal of Villarreal et al. (2021), to have the information available at the right time.

One of the problems that arise in public universities is the insufficient allocation of financial resources; However, the Graduate School of the José Faustino Sánchez Carrión National University carries out its academic and administrative management with autonomy since it has two sources of income; the first, by ordinary budget allocation; and the second, for resources directly collected. University dropout is a problem related to the student as the directly responsible, which generates concern in its directors to know the probabilities of dropout because only a small number of students manage to complete their studies; Student dropout negatively influences the academic and economic development of the organic unit, which is why it is intended, through data mining, to identify behavior patterns in students, analyzing socioeconomic and academic factors that allow the implementation of specific strategies that contribute to maintaining a sustainable economy over time, seeking to reduce the dropout rate.

Based on the report prepared by OECD (2022) with actual data from cohorts in 25 countries, it describes that students admitted to a full-time undergraduate program graduate on average within the theoretical duration by 39%; likewise, the completion rate on average after three years increases by 68% of the complement of students who have not obtained their degree within the theoretical duration; It is highlighted that on average 12% drop out of tertiary education before the start of the second year of studies. In the case of master's students, on average, 51% complete it within the theoretical duration of the study program. Of the complement, an average of 77% of students complete it after three years of the academic course. Other interesting statistics presented by OECD (2021) is estimated in 2019, 38% of students, on average, graduate for the first time before turning 30, excluding international students; In addition, 8%, on average, the proportion of first-time graduates at the master's level of education or it's equivalent compared to OECD information (2020) considering 2018, the majority of first-time graduates 78% obtained a bachelor's degree on average and 10% a master's degree, they also maintain that the three areas or field of study with the highest average proportion is given by business, administration, and law with 25%, followed by health and well-being with 15 %, and finally with engineering, manufacturing, and construction with an average of 14%. The state of the students at the end of their first year of studies can be very significant to understand what happens with the effectiveness of the orientation or preparation. There is an average of 12% of students not enrolled, more than 2% of students completed transfer to another program, and 85% had enrolled in the same or another degree program; In addition, there is an average of 64% of students who have graduated from a bachelor's program, and only 1% from a master's program (OECD, 2019).

In Peru, the figures on the evolution of enrollment according to SUNEDU (2021) at the undergraduate level during 2018 was 1.59 million, a figure that has been reduced by 1.34 million students in 2020, representing a 15.7% difference between periods; the case of postgraduate there is a reduction of 27.7%; during 2018 there were 131.9 thousand and in the 2020 period there were 95.4 thousand students enrolled; The official newspaper El Peruano (2021) details that the universities licensed at the national level indicate that the percentage of interruption of studies has decreased by 4.7%; that is, from 16.2% it has decreased to 11.5% between the semesters 2020-II and 2021-I. Likewise, the regions with the greatest impact were Loreto (16.7%), Callao (14.2%), Áncash (13.9%), Ayacucho (12.8%), and Lima (12.4%), in contrast to Amazonas (4.2%), Huancavelica (6.3 %), Tacna (7.9%), where there was less interruption; and among the causes were connectivity problems, student welfare services, and economic conditions, among others. To reverse this situation to some extent, an investment of 61 million soles has been made to contract internet for students and teachers.

The research was framed in the production of new knowledge through the proposal of the classification model, in addition, the theory of student dropout supported by Díaz (2008) was corroborated. The objective of the research was to develop a dropout classification model in students of education study programs through machine learning and data mining techniques applying H2O.ai's autoML.

2. Literature review

Data mining

Data mining is the process of discovering useful information from immense data structures. It is based on mathematical and statistical analysis aimed at deducing the patterns and trends in the data. Typically, these patterns cannot be detected through traditional exploration since the relationships are too complex or due to the existence of too many volumes of data (Microsoft, 2019; Takaki & Dutra, 2022; Zaina et al., 2022). Likewise, for their implementation, they use statistical techniques and artificial intelligence algorithms to discover patterns or behaviors in large volumes of data (Camborda, 2014; Carrión Ramírez et al., 2023). They use different techniques, such as classification, grouping, and prediction, among others; For this reason, they are effective, for example, in predicting the academic performance of students (Zárate-Valderrama et al., 2021). In turn, Dole and Rajurkar (2014) apply the Naive Bayes algorithm and decision tree to predict graduation and the final condition of students: pass and fail.

Machine learning

Kodelja (2019) argues that some experts in machine learning, a subset of artificial intelligence, claim that machine learning is learning and not something else, while others—including philosophers—reject the claim that machine learning is real learning. For them, real human learning is the highest form of learning. For their part, Xu & Li (2014) argue that machine learning is becoming an essential method for dealing with knowledge acquisition problems; It is defined as a branch of artificial intelligence and refers to the construction and study of systems that can

learn from data. Machine learning is typically concerned with how to build computer programs that automatically improve through the behavior of data; Samuel (1959) defined machine learning as a field of study that allows computers to learn without being explicitly programmed. Dwi et al. (2019) specify Machine Learning as a part of artificial intelligence that focuses on developing a system capable of learning from its own patterns based on a training data set without human intervention. It is applied in various fields, such as education.

Types of machine learning

Jung (2022) describes the types of machine learning as supervised learning —the approach that uses a labeled data set for its prediction, divided into regression and classification (Chatterjee et al., 2023)—; unsupervised learning—a data set that does not need labels; it allows analysts to discover behavior patterns or similarities between functions, it is not intended to detect or predict anything, it is only based on subdivision or grouping (Chatterjee et al., 2023)—; reinforcement learning—is similar to unsupervised learning, learning from an unlabeled data set. The difference with previous tutorials is that you can evaluate the loss function; in these cases, it learns from trial and error experiences depending on the feedback and its factor or agent to perform efficiently (Andrade-Girón et al., 2023; Junco Luna, 2023; Sharmeela et al., 2022).

AUTO ML

He et al. (2020) state that deep learning algorithms (Deep Learning, DL) have achieved extraordinary results in various tasks, such as language modeling, object detection, and image recognition; however, creating a world-class deep learning system for a particular activity is highly dependent on human expertise, which limits its widespread application; this drawback can be solved by introducing the AutoML approach. In recent years it has attracted the attention of various sectors. Many information and communication technology service providers have chosen to implement their respective platforms, such as H2O.ai, DataRobot, DarwinAI, and OneClick.ai. Existing AutoML libraries such as AutoWeka, MLBox, AutoKeras, Google Cloud AutoML, Amazon AutoGluon, IBM Watson AutoAI, and Microsoft Azure AutoML have provided solutions that automatically generate ML-based models (Olusegun Oyetola et al., 2023; Vakhrushev et al., 2021). AutoML, automatic machine learning, Nagarajah & Poravi (2019) describe it as a process that can develop custom models, considerably reducing human intervention; In addition to performing data preprocessing, variable engineering, model building, hyperparameter optimization, and analysis of prediction results and their evaluation, the development of automatic machine learning has made it possible, to a certain extent, to streamline time-consuming machine learning development operations, aiming to reduce the demand on data scientists and can build well-performing machine learning applications, without requiring extensive knowledge of statistics and machine learning (Zöller et al. Huber, 2021). Thus, lately, there has been significant growth in developed libraries. The best known are AutoWEKA, AutoSklearn, AutoPytorch, AutoGluon, H2O.AutoML, MLBox.AutoML and TransmogrifAI.AutoML, among others (AutoML, 2022; Prakash et al, 2023; Rincon Soto & Sanchez Leon, 2022).

H2O.ai platform

LeDell & Poirier (2020) state that H2O is an open-source distributed machine learning platform built to scale to huge data sets. Its application programming interfaces (APIs) are written in R, Python, Java, and Scala. The steps to carry out the automation process using H2O.autoML are data collection, exploration, data preparation, data transformation, model selection, model training, hyperparameter tuning, and prediction (Ajgaonkar, 2022).

Feature Selection

To develop a machine learning model for the prediction of the objective variable, it is necessary to carry out the feature selection process, which aims to identify the interaction of the variables to have the best predictive performance. This process is relevant because it allows knowing the variables that contribute significantly to the predictive model, reducing the number of variables, time, speed, and deployment, making the model less complex and easier to explain (Haque, 2022; Simhan & Basupi, 2023; Zambrano Verdesoto et al., 2023).

There are three kinds of methodologies for feature selection. According to Khun & Jhonson (2022), we have the intrinsic methods —the models based on trees and rules—; multivariate adaptive regression models; and regularization models; The advantage is that they are relatively fast since they are integrated into the model fit; In the case of filter methods, utilizing a supervised analysis it is simple and quick to determine the essential characteristics in the model, they are prone to over select predictors in the model. Finally, the wrapper methods use iterative search procedures, providing subsets of predictors for the model, achieving greater efficiency in prediction performance (do Carmo & da Silva Lemos, 2022; Samuel & Garcia-Constantino, 2022; Santos Amaral et al., 2022).

Student dropout

Tinto (1982) defines dropout as a situation in which a student fails to finish their education; therefore, a dropout would be one who is enrolled in a higher education institution but has no academic activity for three consecutive academic semesters. Gonzales (2005) differentiates two types of dropout in university higher education: the first concerning time (initial, early, and late), and the second concerning space (institutional, internal and the educational system). Likewise, Tinto (1989) sustains the existence of several critical periods that influence student dropout; the first, is during the admission process, where the interested parties form their first impressions or social and intellectual ideas, which generates the expectation of the applicant. The second period is contemplated in the transition between secondary education and the institution, after entering the institution (Driss Hanafi et al., 2023; González Vallejo, 2023; Montes, 2022; Rincón Soto et al., 2023) due to the assembly between college life towards the new way of university life, influencing their mental situation. Tinto (1989) states that dropouts occur during the transition period, with voluntary dropouts being more frequent.

Díaz (2008) presented the analysis models of student dropout to analyze the phenomenon of dropout inherent to university student life and describe the theories from different points of view:

a) Psychological Model, indicates the personality traits that establish the differences between students who complete and drop out of their university studies; it is based on the proposals of Fhisbein and Ajzen (1975), who support the theory of Reasoned Action; Ethington (1990), who is based on the Academic Choice model supported by Eccles et al. (1984) to insert theories about achievement behaviors, such as academic performance that affects the student. Finally, Bean and Eaton (2001) base psychological processes with academic and social integration supported by four psychological theories: Attitude and Behavior theory; Copy Behavior theory—the ability to enter and adapt to a new environment; Self-efficacy theory; and the theory of Attribution.
b) The sociological Model emphasizes the external factors of the students which influence student dropout; Spady (1970) states that one of the causes of dropout is affected by social integration in the university, generated by the influences, expectations, and demands given in the family environment. Likewise, he proposes six predictors for student dropout: academic integration, social integration, socioeconomic status, gender, career quality, and average for each semester.
c) The economic Model is based on two models: the first, Cost/Benefit, related to the social and economic benefits that students perceive to remaining in the university; the second, Subsidy Targeting, aimed at students with low resources or limitations to pay for their studies (Bayona Arévalo & Bolaño García, 2023; Cabrera et al., 1992; Cabrera et al., 1993; Bernal et al., 2000; Jiménez-Pitre et al., 2023; St. John et al., 2000).
d) Organizational Model is based on how the organization integrates students (Berger and Milem, 2000; Berger, 2002; Kuh, 2002; Martínez Sánchez, 2023).
e) Interaction Model, Tinto (1975), maintains that permanence in the institution is a function of the degree of student engagement with the institution and is complemented by Spady's (1970) model, which incorporates the theory of exchange of Nye (1976).

Dimensions of Student Dropout

The variables most frequently considered in the theoretical models related to student dropout were consolidated in the study carried out by Díaz (2008), where four categories are considered: individual (age, gender, family group and integration, social); the academic ones (professional orientation, intellectual development, academic performance, study methods, admission processes, degrees of career satisfaction and academic load); the institutional ones (academic regulations, student financing, university resources, quality of the program or career and relationship with professors and peers); and the socioeconomic ones (socioeconomic stratum, employment situation of the student, employment situation of the parents and educational level of the parents).

3. Methodology

The methodology focused on applying data mining and supervised machine learning techniques, using a set of pre-classified elements to develop the model. The data set was obtained from two sources of information: first, by applying a questionnaire as an instrument, containing 20 items grouped into four dimensions (academic, individual, environmental, and institutional) used to 237 participants. Second, data from the evaluation record was collected through observation. The process involves splitting two data sets; the first to carry out the training, allowing the construction of the classification model, and the second used for the tests, thus obtaining the adjustment parameters. Below, Table 1 presents the items contained.

Table 1. Data collection instrument for participants.

Question	Typo
Academic performance in high school	Ordinal
Failed subjects at the high school level	Ordinal
High school year repetition	Dichotomic
Academic peformeance at the undergraduate level	Ordinal
Failed subjects at the undergraduate level	Ordinal
Sex	Dichotomic
Age range	Ordinal
Marital status	Ordinal
Employment	Ordinal
Number of children	Ordinal
Family income	Ordinal
Motivation towards studying	Dichotomic
Financial situation	Ordinal
Study funding	Dichotomic
Time availability for studying	Ordinal
Stress level	Ordinal
Proper infraestructure	Ordinal
Proper equipment	Ordinal
Proper subjects	Ordinal
Teacher level	Dichotomic

Based on the review of the literature that supports student dropout, the theory of Díaz (2008) has been considered, who adapted the proposed theories to the context of Peruvian reality elaborated by Spady (1970) and Tinto (1989), framed in four factors, as detailed in Table 2.

Table 2. Description of items according to factors proposed by Díaz (2008)

Factors	Items

	Beginning	Ending
Academics	01	05
Individuals	06	12
Environmental	13	16
Institucionals	17	20

For the development of the student dropout model, the R Statistical Software language (v4.2.2; R Core Team 2022) was used, and with the R Studio development environment (v2022.12.0 Build 353; RStudio Team 2022) executed from the system Windows 11 desktop operating (x64 build 22621); Likewise, the H20.ai platform was used to generate the classification model through the package, H2O (v 3.38.0.1; Castellanos & Figueroa, 2023; LeDell et al. 2022; Obregón Espinoza et al., 2023). For dimensionality reduction through feature selection, the following packages were used: familiar (v1.4.1; Zwanenburg & Löck 2021), Information (v0.0.9; Kim 2016), Boruta (v8.0.0; Kursa & Rudnicki 2010), Regularized Random Forest, RRF (v1.9.4; Deng 2013) and FSinR (v2.0.5; Mejías et al., 2022; Aragón-Royón et al. 2020). Additionally, feature selection packages were used to reduce dimensionality and save time and processing capacity to elaborate machine learning models. In addition, the existence of null values, outliers, and cardinality in the variables was verified, which impacts the machine learning models.

4. Results

The descriptive analysis of the scores issued by the participants was carried out, as evidenced in Table 3.

Table 3. Descriptive analysis of the data set of the participants.

#	Tag.	Description	Min	Max	Mean	DE
01	P01	Academic performance in high school	1	5	3.633	0.977
02	P02	Failed subjects at the high school level	1	4	1.578	0.786
03	P03	High school year repetition	1	2	1.932	0.251
04	P04	Academic peformeance at the undergraduate level	2	5	3.443	0.879
05	P05	Failed subjects at the undergraduate level	1	3	1.266	0.530
06	P06	Sex	1	2	1.624	0.485
07	P07	Age range	1	3	2.004	0.805
08	P08	Marital status	1	5	1.975	0.786
09	P09	Employment	1	2	1.831	0.375
10	P10	Number of children	1	3	1.916	0.714
11	P11	Family income	2	5	3.013	0.773
12	P12	Motivation towards studying	1	2	1.038	0.192
13	P13	Financial situation	2	5	3.194	0.773
14	P14	Study funding	1	2	1.068	0.251
15	P15	Time availability for studying	1	5	3.118	1.477
16	P16	Stress level	1	5	2.970	1.418
17	P17	Proper infraestructure	1	5	3.084	1.369
18	P18	Proper equipment	1	5	2.924	1.376
19	P19	Proper subjects	1	5	2.911	1.419
20	P20	Teacher level	1	5	3.650	1.012

To develop these models, independent variables were defined that correspond to 20 items of the instrument and as a dependent variable, student dropout; In addition, two aspects of vital importance have been considered: the selection of characteristics and the percentage for the partition of the data set for training, validation, and testing for each of the models. For the selection of the characteristics, different algorithms were used, obtaining two sets of variables based on the coincidences or similarities in common; the first set, made up of 11 variables (P01, P02, P03, P04, P09, P10, P12, P13, P14, P16, P20); and the second set made up of the variables (P07, P11, P17,

P18, P19), plus the variables of the first, making a total of 16 variables.

Subsequently, the parameters for the invocation of the AutoML method of the H2O object were established, considering the set of independent variables as data parameters and then the objective or destination variable, defined as the dependent variable; the stop or termination parameter, max_models = 100 was considered; in addition, of the option balance_classes = TRUE; With this configuration, the results are presented in Table 4.

Table 4. Machine learning models based on the size of data sets for training, testing, and validation.

#	Model denomination	Items	Data set		
			Training	Test	Validation
01	DeepLearning Grid	16	70	30	0
02	DeepLearning Grid	11	70	30	0
03	GBM Grid	16	70	15	15
04	DeepLearning Grid	11	70	15	15
05	GBM Grid	16	80	20	0
06	GBM Grid	11	80	20	0
07	GBM Grid	16	60	40	0
08	GBM Grid	11	60	40	0
09	GBM Grid	16	75	25	0
10	GBM Grid	11	75	25	0

Table 4 shows the results of the ten executions or iterations carried out according to the predefined configuration; In summary, the main machine learning models with better training metrics are shown in comparison with other models located in lower positions; for example, Extremely Randomized Trees (XRT) and Distributed Random Forest (DRF), Generalized Linear Model (GLM). Table 5 below shows the metrics of the training process for each of the automatically generated models.

Table 5. Model performance metrics with the training and validation data set.

#	Model denomination	Items	AUC	LOGLOS	AUCPR
01	DeepLearning Grid	16	0.981685	0.389653	0.956428
02	DeepLearning Grid	11	0.981136	0.214359	0.951164
03	GBM Grid	16	0.980220	0.183851	0.943741
04	DeepLearning Grid	11	0.982784	0.196832	0.954476
05	GBM Grid	16	0.972311	0.258593	0.923799
06	GBM Grid	11	0.972603	0.204378	0.932085
07	GBM Grid	16	0.974163	0.246842	0.915569
08	GBM Grid	11	0.972010	0.207276	0.920860
09	GBM Grid	16	0.977618	0.218077	0.925325
10	GBM Grid	11	0.972982	0.201235	0.923862

As seen in Table 5, the scores obtained in each metric are very similar and significant during the training process, subsequently carrying out the tests to get the performance metrics of each of the indicated models.

Ranking models have a variety of performance metrics. Among the most relevant, we have the Gini coefficient, used to measure the quality of the prediction model, in whose interpretation a value of zero means perfect equality; that is, there is a deficient model; Whenever it has a value close to unity, it is presented as a maximum inequality, it is considered a perfect classifier. The area under the curve is a metric to evaluate the capacity of the classification model, allowing it to differentiate between true positives and false positives; a value close to unity is considered a perfect model. Unlike the metric, the area under the precision-recall curve does not feel true negatives, something widely used in unbalanced data sets. The log loss metric looks at the approximation of a model's predicted values and actual target ratings, where an assignment close to zero means the model provides the probability correctly.

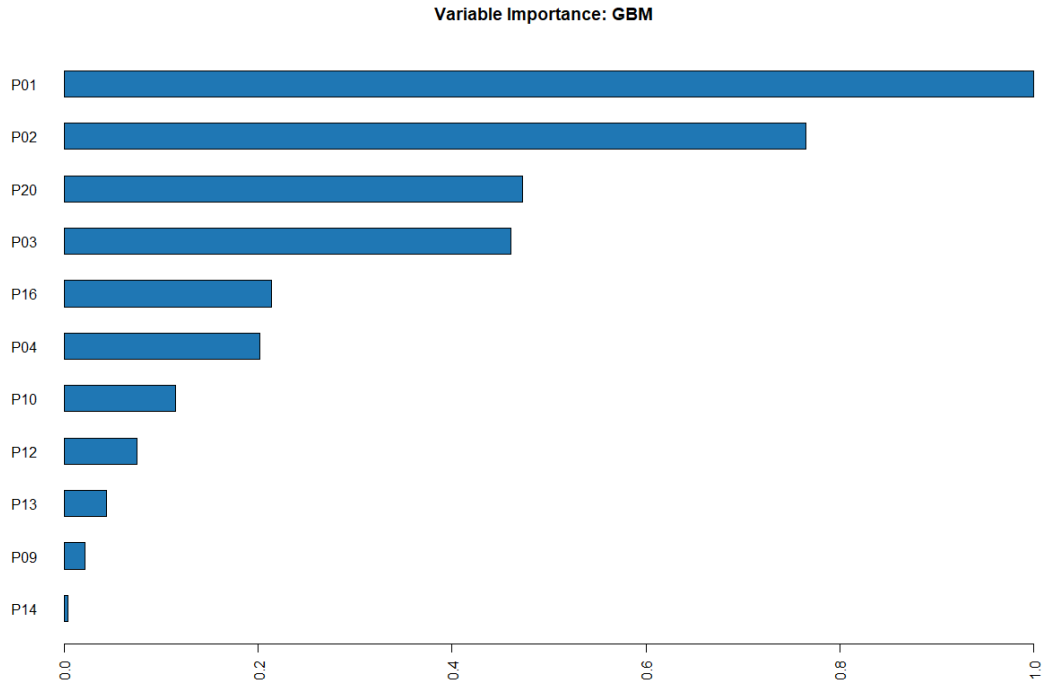
Table 6. Model performance metrics using the test data set.

#	Model denomination	Items	GINI	AUC	AUCPR	LOGLOSS
01	DeepLearning Grid	16	0.895981	0.947991	0.913763	0.850491
02	DeepLearning Grid	11	0.865248	0.932624	0.905851	0.546854
03	GBM Grid	16	1.000000	1.000000	1.000000	0.025920
04	DeepLearning Grid	11	1.000000	1.000000	1.000000	0.044860
05	GBM Grid	16	0.915633	0.957816	0.911510	0.312979
06	GBM Grid	11	0.935484	0.967742	0.937704	0.259712
07	GBM Grid	16	0.943012	0.971506	0.919590	0.293444
08	GBM Grid	11	0.932157	0.966079	0.925879	0.217350
09	GBM Grid	16	0.912281	0.956140	0.922686	0.270146
10	GBM Grid	11	0.898246	0.949123	0.911629	0.295948

Table 6 contains the metrics of each execution and tests carried out with the automatically generated models. The metrics are similar, except for the cases of the third and fourth models, which are overfitted due to the number of observations partitioned into three data sets. Likewise, most models demonstrate better performance in the metrics of the models with fewer items. In this sense, due to the principle of parsimony, the models with 11 items are chosen according to the algorithms used to select characteristics, allowing benefits for their future implementation. Thus, slightly better performance is observed in the tenth Gradient Boosting Machine model, followed by the second DeepLearning model.

Figure 1 contemplates the variables ordered from highest to lowest, according to the importance in the model prediction, based on the percentage values that are scaled to 100%. A strong influence is evident in the experience of the participants at the secondary level: academic performance (29.65%), failed subjects (22.67%), repetition of the year (13.65%), teacher performance (14.03%), with less relevance are the aspects related to the stress of the person (6.35%); performance in undergraduate (5.99%), the number of children (3.40%), motivation (2.23%), economic situation (1.28%), work related to his career (0.62%) and finally the financing of his studies (0.10%).

Figure 1. Importance of variables in the classification model.



Additionally, model metrics were obtained from the confusion matrix. They are detailed in Table 7.

Table 7. Confusion matrix of the generated GBM model.

Prediction values	**Reals**		**Error**	**Index**
	Positive	**Negative**		
Positive	38	0	0.000	= 0 / 38
Negative	4	11	0.267	= 4 / 15
Total	42	11	0.075	= 4 / 53

Accuracy is a metric for determining correct predictions as a proportion of the total number of predictions made. A score close to unity represents optimal performance. From Table 7 we can obtain a precision equivalent to 92%, that is to say, that the model has a successful prediction capacity of 92 cases among 100 observations; for sensitivity, 90% is indicated, indicating a successful prediction of 90 cases out of 100 for the positive class; finally, for specificity, we identified 100% of the cases to predict the negative class.

The ROC curve is a graph that represents the relationship between true positives (sensitivity) and false positives (specificity). Figure 2 demonstrates a curve near the upper left corner, thus indicating optimum performance. It should be specified when the curve approaches the 45° diagonal or baseline. It will be less precise, corresponding to poor performance. Likewise, the lower left side of the graph represents a lower tolerance for false positives, while the upper right side represents a higher tolerance for false positives.

Figure 2. ROC chart of the GBM classification model.

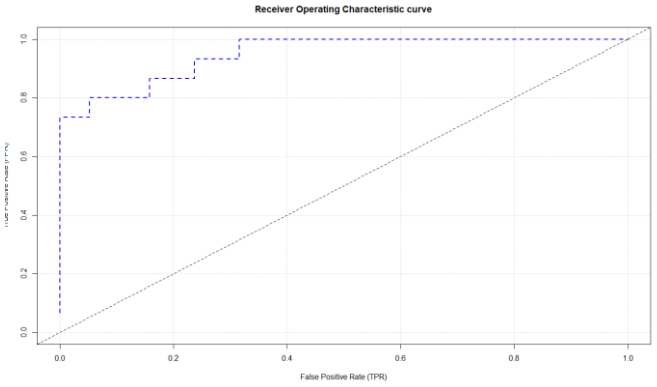
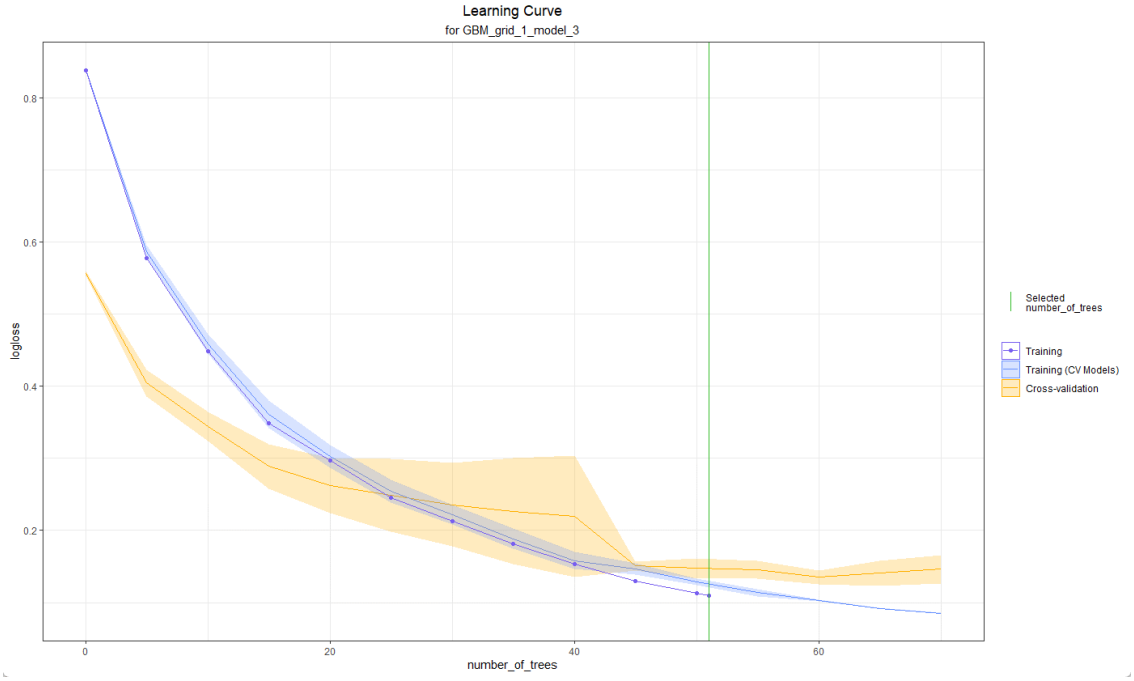


Figure 3 shows the behavior of the GBM classification model through the learning curve and presents a logarithmic loss in the training and validation data set; In addition, it is seen that the curves are stable when having a number greater than 50 trees, that is, adding more instances to the model would not improve its performance much.

In short, the GBM (Gradient Boosting Machine) model is a supervised machine learning method used to classify machine learning problems. It is built using decision trees. The generated GBM model consists of 51 internal trees, with a size corresponding to 8,910 bytes. The tree has a minimum depth of 4 and a maximum depth of 6, with an average depth of 5.29. The minimum number of sheets is 7, and the maximum is 13, with an average of 9.24. This configuration of the GBM model indicates that the internal decision trees have a reasonable depth and a moderate number of leaves. This means that the GBM

model has a good fit and can provide an appropriate classification for the data, as evidenced by performance metrics.

Figure 3. GBM Classification Model Learning Curve.



5. Conclusions

Once the GBM model of student dropout classification has been generated, it can be concluded that it is adequate for the task since it offers adequate precision, sensitivity, and specificity for the prediction of student dropout cases, since it presents a high-performance capacity, depth, and several blades suitable for training. Therefore, the use of this model for the analysis of student dropout is appropriate. It offers several advantages, such as the ability to work with unbalanced data, improve results by tuning model parameters, use a cross-validation method to assess model accuracy, and make forecasts in real-time, which allows managers to make quick and effective decisions to combat student dropout, providing a useful tool for the detection and prevention of dropout. The use of H2O.ai platform, which has a method called H2O.AutoML, used for the automatic generation of learning models, allows the user to select a data set, partition them and generate the model. H2O.ai selects the best model according to the specified parameters. This tool saves the user time and resources since he does not need to choose the model parameters manually. Therefore, using the H2O.ai platform and the AutoML method is a good option for model building.

A relevant aspect of the research was transversality. In the first instance, machine learning could use algorithms to extrapolate insights into a data set; In the case of data mining, this technique has made it possible to identify patterns in the data within the context of university higher education, allowing users to share and reuse acquired knowledge and best practices in other knowledge areas.

Implementing the generated student dropout classification model is recommended since it has great practical utility for those responsible for education and the university community because this tool allows predicting the risk of student dropout early, allowing measures to be taken preventive measures to reduce it. These measures may include offering financial aid, academic advising, tutoring programs, remedial classes, and other forms of student support that can help students stay in college and achieve their educational goals. Academically, taking the model into account allows researchers to save time and resources when evaluating different classification and prediction models automatically, offering the ability to perform a sensitivity analysis to understand better the factors that influence attrition, being a good choice for research. Likewise, it improves the teaching approach and provides a greater understanding of the needs of students to provide them with appropriate support.

References

[1] Ajgaonkar, S. (2022). *Practical Automated Machine Learning Using H2O.ai: Discover the power of automated machine learning, from experimentation through to deployment to production*. Packt Publishing.

[2] Andrade-Girón, D., Carreño-Cisneros, E., Mejía-Dominguez, C., Marín-Rodriguez, W., & Villarreal-Torres, H. (2023). Comparación de Algoritmos Machine Learning para la Predicción de Pacientes con Sospecha de COVID-19. *Salud, Ciencia Y Tecnología*, *3*, 336. https://doi.org/10.56294/saludcyt2023336

[3] Anzanello, M. J., & Fogliatto, F. S. (2011). Learning curve models and applications: Literature review and research directions. *International Journal of Industrial Ergonomics*, 41(5), 573–583. https://doi.org/10.1016/j.ergon.2011.05.001

[4] Aragón-Royón, F., Jiménez-Vílchez, A., Arauzo-Azofra, A. & Benitez, J. (2020). "FSinR: an exhaustive package for feature selection." *arXiv e-prints, arXiv:* 2002. 10330. 2002. 10330, https://arxiv.org/abs/2002.10330.

[5] AutoML. (2022, 15 de diciembre). *AutoML* | Home. https://www.automl.org/automl/

[6] Bean, J. P. & Eaton, S. (2001). The psychology underlying successful retention practices. *Journal of College Student Retention Research, Theory & Practice* Vol. 3, N° 1: 73-89.

[7] Berger, J. & Milem, J. (2000). Organizational Behavior in Higher Education and Student Outcomes. In: J. Smart (Ed.), *Higher Education: Handbook of theory and research*. Vol. 15: 268-338.

[8] Berger, J. (2002). Understanding the Organizational Nature of Student Persistence: Empirically based Recommendations for Practice. *Journal of College Student Retention: Research, Theory and Practice.* Vol. 3, N° 1: 3-21.

[9] Bayona Arévalo, Y., & Bolaño García, M. (2023). Scientific production on dialogical pedagogy: a bibliometric analysis. *Data & Metadata*, *2*, 7. https://doi.org/10.56294/dm20237

[10] Cabrera, A., Nora, A. & Castañeda, M. (1992). The role of finances in the persistence process: a structural model. *Research in Higher Education.* Vol 33, N° 5: 303-336.

[11] Cabrera, A., Nora, A. & Castañeda, M. (1993). College Persistence: structural Equations modelling test of Integrated model of student retention. *Journal of Higher Education.* Vol. 64, N° 2: 123-320.

[12] Carrión Ramírez, B. M., Córdova Medina, H. M., Murillo Párraga, M. V., & Del Campo Saltos, G. S. (2023). Health and Inclusive Higher Education: Evaluation of the Impact of Policies and Programs for People with Disabilities in Ecuador. *Salud, Ciencia Y Tecnología*, *3*, 361. https://doi.org/10.56294/saludcyt2023361

[13] Castellanos, S., & Figueroa, C. (2023). Cognitive accessibility in health care institutions. Pilot study and instrument proposal. *Data & Metadata*, *2*, 22. https://doi.org/10.56294/dm202322

[14] Chatterjee, P., Yazdani, M., Fernández-Navarro, F., & Pérez-Rodríguez, J. (2023). *Machine Learning Algorithms and Applications in Engineering*. CRC Press. https://doi.org/10.1201/9781003104858

[15] Deng, H. (2013). Guided Random Forest in the RRF Package. *ArXiv:* 1306.0237 (9 de noviembre de 2021). Tasa de deserción en educación universitaria. *Diario oficial El Peruano* https://elperuano.pe/noticia/132960-tasa-de-desercion-en-educacion-universitaria-se-redujo-a-115

[16] Díaz, C. (2008). Modelo Conceptual para la Deserción Estudiantil Universitaria Chilena. *Estudios Pedagógicos (Valdivia)*, 34(2), 65-86. https://dx.doi.org/10.4067/S0718-07052008000200004

[17] Do Carmo, D., & da Silva Lemos, D. L. (2022). Quality standards for data and metadata addressed to data science applications. *Advanced Notes in Information Science*, *2*, 161–170. https://doi.org/10.47909/anis.978-9916-9760-3-6.116

[18] Driss Hanafi, M., Lali, K., Kably, H., & Chakor, A. (2023). The English Proficiency and the Inevitable Resort to Digitalization: A Direction to Follow and Adopt to Guarantee the Success of Women Entrepreneurs in the World of Business and Enterprises. *Data & Metadata*, *2*, 42. https://doi.org/10.56294/dm202342

[19] Dwi, M., Prasetya, A., & Pujianto, U. (2018). Technology acceptance model of student ability and tendency classification system. *Bulletin of Social Informatics Theory and Application*, 2(2), 47–57. https://doi.org/10.31763/businta.v2i2.113

[20] Eccles, J. P., Adler, T. & Meece, J. (1984). Sex differences in achievement: a test of alternate theories. *Journal of Personality and Social Psychology.* Vol. 46, N° 1: 26-43.

[21] Ethington, C. (1990). A psychological model of student persistence. *Research in Higher Education.* N° 31, Vol. 31: 279-293.

[22] Fishbein, M. & Ajzen, I. (1975). Attitudes toward objects as predictors of simple and multiple behavioural criteria. *Psycological Review.* N° 81: 59-74.

[23] González, L. E. (2005). *Estudio sobre la repitencia y deserción en la educación superior chilena*. Digital Observatory for higher education in Latin America and The Caribbean. IESALC – UNESCO.

[24] González Vallejo, R. (2023). Metaverse, Society & Education. *Metaverse Basic and Applied Research*, *2*, 49. https://doi.org/10.56294/mr202349

[25] Haque, A. (2022). *Feature Engineering & Selection for Explainable Models: A second course for data scientists*. LULU Internacional.

[26] He, X., Zhao, K., & Chu, X. (2020). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems, 106622*. https://doi:10.1016/j.knosys.2020.106622

[27] Jiménez-Pitre, I., Molina-Bolívar, G., & Gámez Pitre, R. (2023). Visión sistémica del contexto educativo tecnológico en Latinoamérica. *Región Científica*, *2*(1), 202358. https://doi.org/10.58763/rc202358

[28] Junco Luna, G. J. (2023). Study on the impact of artificial intelligence tools in the development of university classes at the school of communication of the Universidad Nacional José Faustino Sánchez Carrión. *Metaverse Basic and Applied Research*, *2*, 51. https://doi.org/10.56294/mr202351

[29] Jung, A. (2022). *Machine Learning. Springer Singapore*. https://doi.org/10.1007/978-981-16-8193-6

[30] Kim, L. (2016). *_Information: Data Exploration with Information Theory (Weight-of-Evidence and Information Value).* R package version 0.0.9, https://CRAN.R-project.org/package=Information.

[31] Kodelja, Z. (2019). Is Machine Learning Real Learning? Robotisation, Automatisation, the End of Work and the Future of Education. *CEPS Journal* Vol 9 No 3. Educational Research Institute, Ljubljana, Slovenia. https://doi.org/10.26529/cepsj.709

[32] Kuh, G. (2002). Organizational culture and student persistence: prospects and puzzles. *Journal of College Student Retention*. Vol. 3, N° 1: 23-39.

[33] Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11.

[34] LeDell, E. & Poirier, S. (2020). *H2O AutoML: Scalable Automatic Machine Learning. 7th ICML Workshop on Automated Machine Learning (AutoML)*, July 2020. URL https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf.

[35] LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M. & Malohlava, M. (2022). *_h2o: R Interface for the 'H2O'*

*Scalable Machine Learning Platform_*. R package version 3.38.0.1, https://github.com/h2oai/h2o-3

[36] Martínez Sánchez, R. (2023). Transforming online education: the impact of gamification on teacher training in a university environment. *Metaverse Basic and Applied Research*, *2*, 47. https://doi.org/10.56294/mr202347

[37] Mejías, M., Guarate Coronado, Y. C., & Jiménez Peralta, A. L. (2022). Artificial intelligence in the field of nursing. Attendance, administration and education implications. *Salud, Ciencia Y Tecnología*, *2*, 88. https://doi.org/10.56294/saludcyt202288

[38] Melgar, A. S., Garay-Argandoña, R., Aranda, E. A. E., & Hernández, R. M. (2020). Management risk factors in educational institutions and their impact on peruvian student dropout. *Elementary Education Online*, *19*(4), 226–233. https://doi.org/10.17051/ILKONLINE.2020.04.124

[39] Montes, H. (2002). *La transición de la educación media a la educación superior, Retención y movilidad estudiantil en la educación superior: calidad en la educación*, pp. 269-276. Publicación del Consejo Superior de Educación. Santiago.

[40] Mushtaq, I., & Khan, S. (2012). Factors Affecting Students' Academic Performance. Global Journal of Management and Business Redearch, 12(9), 17-22. ISSN: 2249-4588

[41] Nagarajah, T., & Poravi, G. (2019). *A Review on Automated Machine Learning (AutoML) Systems*. 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). https://doi:10.1109/i2ct45611.2019.9033810

[42] Nye, J. (1976). Independence and Interdependence. *Foreign Policy*. Spring, Nº 22: 130-161.

[43] Obregón Espinoza, E. L., Neri Ayala, A. C., Ramos y Yovera, S. E., Caro Soto, F. G., & Muñoz Vilela, A. J. (2023). Design Thinking as a tool for fostering innovation and entrepreneurship. *Salud, Ciencia Y Tecnología*, *3*, 368. https://doi.org/10.56294/saludcyt2023368

[44] OECD (2022), Education at a Glance 2022: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/3197152b-en

[45] OECD (2021), Education at a Glance 2021: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/b35a14e5-en.

[46] OECD (2020), Education at a Glance 2020: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/69096873-en.

[47] OECD (2019), *Education at a Glance 2019: OECD Indicators*, OECD Publishing, Paris, https://doi.org/10.1787/f8d7880d-en.

[48] Olusegun Oyetola, S., Oladokun, B. D., Ezinne Maxwell, C., & Obotu Akor, S. (2023). Artificial intelligence in the library: Gauging the potential application and implications for contemporary library services in Nigeria. *Data & Metadata*, *2*, 36. https://doi.org/10.56294/dm202336

[49] Prakash, A., Haque, A., Islam, F., & Sonal, D. (2023). Exploring the Potential of Metaverse for Higher Education: Opportunities, Challenges, and Implications. *Metaverse Basic and Applied Research*, *2*, 40. https://doi.org/10.56294/mr202340

[50] R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[51] Rincon Soto, I. B., & Sanchez Leon, N. S. (2022). How artificial intelligence will shape the future of metaverse. A qualitative perspective. *Metaverse Basic and Applied Research*, *1*, 12. https://doi.org/10.56294/mr202212

[52] Rincón Soto, I. B., Soledispa-Cañarte, B. J., Soledispa-Cañarte, P. A., Cañarte-Rodríguez, T. C., & Sarmiento-Tomalá, G. M. (2023). Neurociencia y educación en la era de la sociedad del tecno-conocimiento. *Salud, Ciencia Y Tecnología - Serie De Conferencias*, *2*(2), 176. https://doi.org/10.56294/sctconf2023176

[53] RStudio Team (2022). *RStudio: Integrated Development for R. RStudio*, PBC, Boston, MA URL http://www.rstudio.com/.

[54] Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44(1), 211-229. https://doi:10.1147/rd.441.0206

[55] Samuel, A. M., & Garcia-Constantino, M. (2022). User-centred prototype to support wellbeing and isolation of software developers using smartwatches. *Advanced Notes in Information Science*, *1*, 140–151. https://doi.org/10.47909/anis.978-9916-9760-0-5.125

[56] Santos Amaral, L., Medeiros de Araújo, G., & Reinaldo de Moraes, R. A. (2022). Analysis of the factors that influence the performance of an energy demand forecasting model. *Advanced Notes in Information Science*, *2*, 92–102. https://doi.org/10.47909/anis.978-9916-9760-3-6.111

[57] Sharmeela, C., Sanjeevikumar, P., Sivaraman, P, & Meera, J. (2022). *IoT, Machine Learning and Blockchain Technologies for Renewable Energy and Modern Hybrid Power Systems*. River Publishers.

[58] Simhan, L., & Basupi, G. (2023). None Deep Learning Based Analysis of Student Aptitude for Programming at College Freshman Level. *Data & Metadata*, *2*, 38. https://doi.org/10.56294/dm202338

[59] Spady, W. (1970). Dropouts from higher education: an interdisciplinary review and synthesis. *Interchange*. Vol. 19, Nº 1: 109-121.

[60] St. John, E., Cabrera, A., Nora, A. & Asker, E. (2000). *Economic influences on persistence. In: J. M. Braxton. Reworking the student departure puzzle: New theory and research on college student retention*. Nashville: Vanderbilt University Press. pp. 29-47.

[61] Superintendencia Nacional de Educación Superior Universitaria [SUNEDU]. (2020). *II Informe bienal sobre la realidad universitaria en el Perú*. https://cdn.www.gob.pe/uploads/document/file/1230044/Informe%20Bienal.pdf

[62] Takaki, P., & Dutra, M. (2022). Data science in education: interdisciplinary contributions. *Advanced Notes in Information Science*, *2*, 149–160. https://doi.org/10.47909/anis.978-9916-9760-3-6.94

[63] Tinto, V. (1982). Limits of theory and practice of student attrition. *Journal of Higher Education*. Vol. 3, Nº 6: 687-700.

[64] Tinto, V. (1989). Definir la deserción: una cuestión de perspectiva. *Revista de Educación Superior* Nº 71, ANUIES, México.

[65] Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., & Farivar, R. (2019). Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. https://doi:10.1109/ictai.2019.00209

[66] Vakhrushev, A., Ryzhkov, A., Savchenko, M., Simakov, D., Damdinov, R. and Tuzhilin, A. (2021). LightAutoML: AutoML Solution for a Large Financial Services

Ecosystem. *Choice Reviews Online*, 45(02), 45–0602—45–0602. https://doi.org/10.5860/choice.45-0602

[67] Villarreal-Torres, H., Marín-Rodriguez, W., Ángeles-Morales, J. & Cano-Mejía, J. (2021). Gestión de Tecnología de Información para universidades peruanas aplicando computación en la nube. *Revista Venezolana de Gerencia*, 26 (Especial 6), 665-679. https://doi.org/10.52080/rvgluz.26. e6.40

[68] Xu, W., & Li, W. (2014). *Granular Computing Approach to Two-Way Learning Based on Formal Concept Analysis in Fuzzy Datasets*. IEEE Transactions on Cybernetics, 46(2), 366–379. https://doi:10.1109/tcyb.2014.2361772

[69] Zaina, R. Z., Culmant Ramos, V. F., & Medeiros de Araujo, G. (2022). Automated triage of financial intelligence reports. *Advanced Notes in Information Science*, *2*, 24–33. https://doi.org/10.47909/anis.978-9916-9760-3-6.115

[70] Zambrano Verdesoto, G. J., Rincon Soto, I. B., & Castro Alfaro, A. (2023). Contributions of neurosciences, neuromarketing and learning processes in innovation. *Salud, Ciencia Y Tecnología*, *3*, 396. https://doi.org/10.56294/saludcyt2023396

[71] Zöller, M. y Huber, M. (2021). Benchmark and Survey of Automated Machine Learning Frameworks. *Journal of Artificial Intelligence Research*, 70, 409–472. https://doi.org/10.1613/jair.1.11854

[72] Zwanenburg, A. & Löck, S. (2021). *Familiar: End-to-End Automated Machine Learning and Model Evaluation*. https://github.com/alexzwanenburg/familiar.

[73] Zwanenburg, A. (2021). *Familiar: Vignettes and Documentation*. https://github.com/alexzwanenburg/familiar.