How to Cite:
Bayas, B. O., Morán, E. R. M., & Castro, L. M. U. (2022). Student dropout in times of COVID-19: A case study Universidad Técnica Estatal de Quevedo. *International Journal of Health Sciences*, *6*(S2), 13869–13879. https://doi.org/10.53730/ijhs.v6nS2.8634

Student dropout in times of COVID-19: A case study Universidad Técnica Estatal de Quevedo

Byron Oviedo Bayas
PhD. en Tecnologías de la Información y Comunicación, Director of Scientific Journals, Research Professor at the School of Engineering Sciences of the Universidad Técnica Estatal de Quevedo – Quevedo, Ecuador
Corresponding author email: boviedo@uteq.edu.ec

Evelym Ruth Morán Morán
Student of the Research Master's Degree in Applied Statistics at the Postgraduate Institute of the Technical University of Manabí - Portoviejo, Instituto de Investigación de la Universidad Técnica Estatal de Quevedo - Quevedo, Ecuador
emoran3145@utm.edu.ec

Lelly María Useche Castro
PhD. in Statistics, Director of the Multivariate and Stochastic Analysis Group (G.A.M.E) Department of Mathematics and Statistics, Instituto de Ciencias Básicas, Universidad Técnica de Manabí, Portoviejo, Ecuador
lelly.useche@utm.edu.ec

***Abstract*---**This article presents a study on student dropout in COVID19 time. As a case study, the socioeconomic and academic performance data of students of the State Technical University of Quevedo, Ecuador in the periods from 2019 to 2022 where educational activities were developed in virtual modality are analyzed. With the requested information, the objective variable (Permanence) was constructed and a process of exploration and analysis of the information was carried out. From this process, 58 variables were chosen and used for the construction of two classification models using Decision Trees and Logistic Regression. Of the algorithms studied, Decision Trees was the best model for identifying students who dropped out, with a value higher than 95% correct classification.

***Keywords*---**student dropout, ranking models, logistic regression, decision trees.

Introduction

Higher education centers around the world contribute to the development and growth of people's skills, providing them with the necessary competencies to better prepare them for the challenges of today's society. The educational process, although full of good intentions, does not always have positive results for a large part of the students. Demotivation, economic problems, and family environment, among others, lead them not to continue with their professional training process, causing them to drop out of their career. The retention rate, as is known to the dynamic process developed by higher education institutions to meet the needs of their students and ensure that they remain until the end of their academic goals, is one of the most important indicators for which authorities and teachers work tirelessly together in general (Munizaga et al., 2018). Some researchers also know this indicator as student attrition which, in common sense, represents the opposite of the retention rate and define it as: "the premature abandonment of a study program before reaching the degree or degree and which considers a sufficiently long time to rule out the possibility of the student rejoining" (Saldaña *et al.*, 2010).

Although Latin America and the Caribbean are among the regions with the greatest problems related to student dropout at the university level (Olarte, 2020), developed countries such as the United States, Spain, France, and Australia, also face difficulties with this indicator, with a retention rate below 70%. The best results worldwide are found in Germany, Switzerland, Finland, and the Netherlands whose value is between 80% and 90%. The Universidad Técnica Estatal de Quevedo (UTEQ) is a public university in Ecuador, that, like the others, focuses its educational processes on increasing the retention rate and providing Ecuadorian families and society with competent professionals ready to take on the challenges of the new times. At UTEQ, there has been a notable increase in student dropout in recent years, which has become a pandemic. The lack of technological equipment for online classes, the lack of work in families, among other economic factors have taken many students out of education.

Studies in the educational sciences and the contributions of other sciences to education have made interdisciplinarity very common to seek better results in all social spheres. This justifies the fact that sciences such as Statistics and more recently Learning Machines focus their techniques on finding answers to many educational problems, including student dropout. In this sense, Torres (2021) proposes in his doctoral dissertation analysis of student dropouts in the professional school of accounting at the Universidad Nacional de San Agustín de Arequipa, using multivariate techniques to group the subjects with a high degree of complexity and thus explore student dropout.

Another study to highlight is that of Martelo *et al.* (2018) who determined factors that influence university dropout through the multivariate analysis technique. The research was quantitative with a non-experimental transectional correlational design. The sample consisted of 59 dropouts from the Systems Engineering program at the University of Cartagena. A 32-item questionnaire was designed for data collection. The analysis used the multivariate technique of factor analysis, which yielded the 10 most relevant factors for the study of student dropout. For

his part, López (2021) also estimates student dropout using multivariate analysis techniques in a technological higher education institution. In this study, to evaluate the dependent variable, the author requested data from the institution's secretary's office, from the enrollment lists of each semester, identifying which students dropped out. A multivariate analysis model with logarithmic regression was applied to obtain a predictive model, identifying that the significant variables are repetition and the career to which the student belongs (Nyandra & Suryasa, 2019). In Oviedo (2016), a study of student dropout in Ecuador was conducted using Bayesian classifiers with metaheuristic learning. The authors reported that economic variables have a high incidence of dropout. In general, many studies address this problem and, unfortunately, there is no general model that characterizes dropout in all educational centers due to the specificities of each center and study program. Because of this and because of the situation faced by the educational processes due to the pandemic, this paper presents a study of the behavior of dropout in the State Technical University of Quevedo through classification models.

Materials and Methods

To conduct the study, a quasi-experimental approach was defined with a population of 359802 student records of the Quevedo State Technical University during the period from 2019 to 2021. The socioeconomic data were obtained from an online survey conducted by the Universidad Técnica Estatal de Quevedo to its students at the beginning of each term. This survey aims to identify unfavorable economic situations for the knowledge of Student Welfare. The information on academic performance was also provided by the university management system (Miguel Medina Romero, 2021).

Figure 1 shows the scheme that responds to the methodological strategy followed for the study. To execute each of the stages, the free software R (Rizzo, 2019) was used due to the facilities it provides for the manipulation and graphic analysis of the data. Its efficiency in the Statistical area and Data Science in general, makes it one of the most used options, together with Python (Tattar, 2017), currently used by scientists around the world.

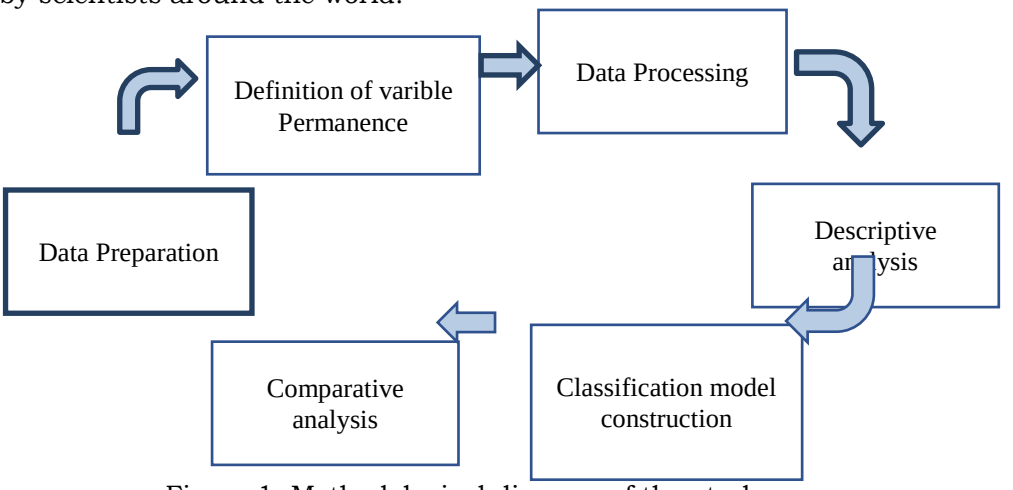

Figure 1. Methodological diagram of the study

Each of the stages of the methodological process is discussed below:

- Data preparation: In this step, criteria are applied to choose the records that are part of the population, taking into account that students in the leveling stage and students in the last semester in each period will not be taken into account. This is because the identifier of the leveling student changes when he/she enters the first semester of the course and the students of the last period is not important for the study.
- Definition of the variable Permanence: This variable takes the value of 1 if the student is still enrolled in the next school year regardless of whether he/she was promoted or not, and 0 if the student did not continue his/her studies.
- Scanning analysis: In this stage, missing value analysis, recoding, and variable selection are performed. This process is important to ensure that the applied classification algorithms use the most refined information possible.
- Descriptive analysis: Although this step is quite extensive due to the variety of analyses that can be developed, in this paper we will only analyze the relationship of the predictor variables with the permanence variables, as a way to identify possible dependencies in the process.
- Construction of classification models: For this process, the knowledge base is divided into a training group and a test group. Then two algorithms, Decision Trees (Landgrebe, 1991) and Logistic Regression (Ungar, 2007) are applied to the training data.
- Comparative analysis: To establish which of the algorithms obtained the model more adjusted to the reality of the test data, some indicators of the classification models are used, such as Accuracy, Precision, Specificity, and AUC (Kulis, 2013).

Results

This section presents the results of each of the stages defined in the methodological process. Regarding the data, information was obtained from 6 academic periods (from 2019-2020-I to 2021-2022-II) for a total of 359802 records with 224 variables. In the filtering process, some cases were eliminated leaving 299694 cases in the knowledge base. The results of the definition of the variable Permanence are shown in the graph, where it can be seen that about 20% of the records represent cases of student dropout.

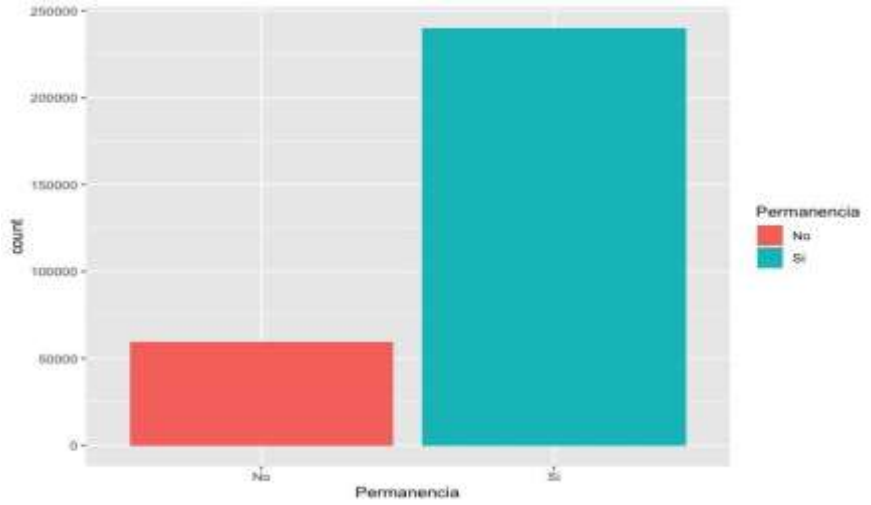

Figure 2. Frequency of decision variables Permanence

Concerning the preprocessing of the data, analyzing the missing values by variables. It was identified that the great majority of these variables have a percentage of missing values that exceeds 25%, variables such as Type of work of the father, with whom he lives, If the father is deceased, Level of degree of the father and mother, among many others, do not have enough records to carry out an imputation process and are no longer important for the study. In this process 162 variables were eliminated, most of the socioeconomic variables. This left 58 variables for the study.

After analyzing the missing values by records, it was identified that a total of 119728 records had some missing value and due to the characteristics of the study, these records were eliminated, commonly known as the Listwise Technique (Julian, 1995). Coding problems were also identified in some variables as shown in Table 1.

Table 1. Variable coding

Variable	Type	Domain	Solution
Age	Discreet	[0,79]	Delete all records x < 18 and x > 40
Average time spent doing homework	Discreet	{0, 1, 2, 3, 4, 5, 6, 9, 11, 12}	It was coded into 3 categories x∈{0,1,2,3} = Little, x∈{4, 5, 6} =Normal and x > 6 = Much.
Average time spent doing chores at home	Discreet	{0, 1, 2, 3, 4, 5, 6, 9, 11, 12}	It was coded into 3 categories x∈{0,1,2,3} = Little, x∈{4, 5, 6} =Normal and x > 6 = Much.

After this process, there remained the study of 174215 cases and 58 variables including the objective variable, 8 of which were coded as numerical (final grade, first cut grade, second cut...) and the others as categorical. Regarding the descriptive analysis of the variables, the behavior of the categorical variables concerning permanence follows the same pattern as described in Figure 4. It can be observed that for each category of the variables Supple (1: students with Supple, 0: not) the value "Yes" is more frequent than "No" for permanence. The same happens with the numerical variables as can be seen in graph 5. In this case, the final grade presents a greater dispersion for the students who drop out.

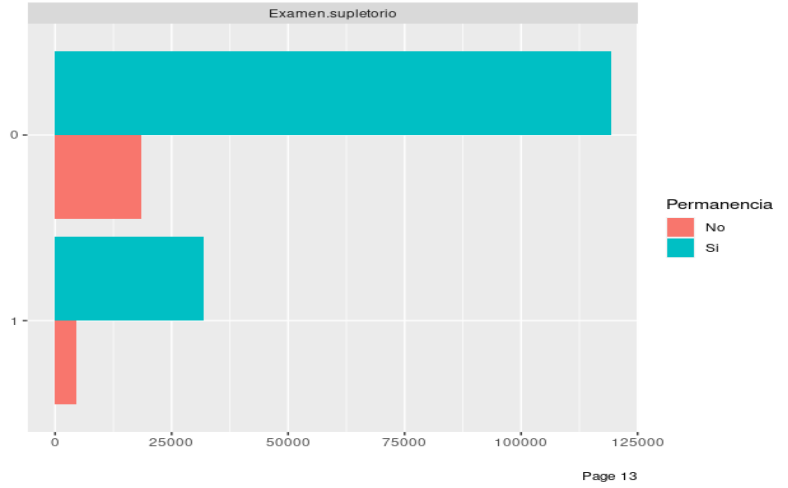

Figure 3. Frequency of the variable Supplementary Examination for Permanence

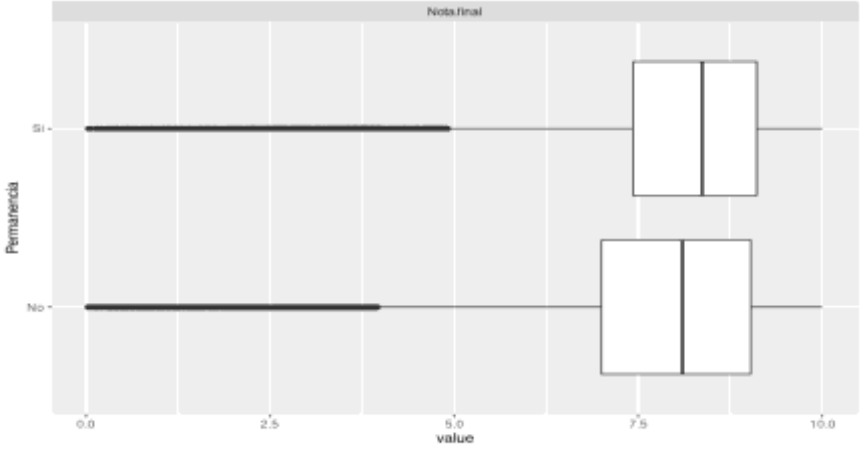

Figure 4. Boxplot of the final grade as a function of permanence

For the learning process, the knowledge base is divided into a training group (75%) and a test group (25%). The training base should contain the greatest diversity of examples to prepare the model for the testing stage. The data used in this work are unbalanced (Figure 2), a situation that is not beneficial for learning (Bonaccorso, 2017). For such reason, a controlled selection was performed by

randomly choosing 15352 cases from each category of the variable Permanence, as shown in Figure 6.

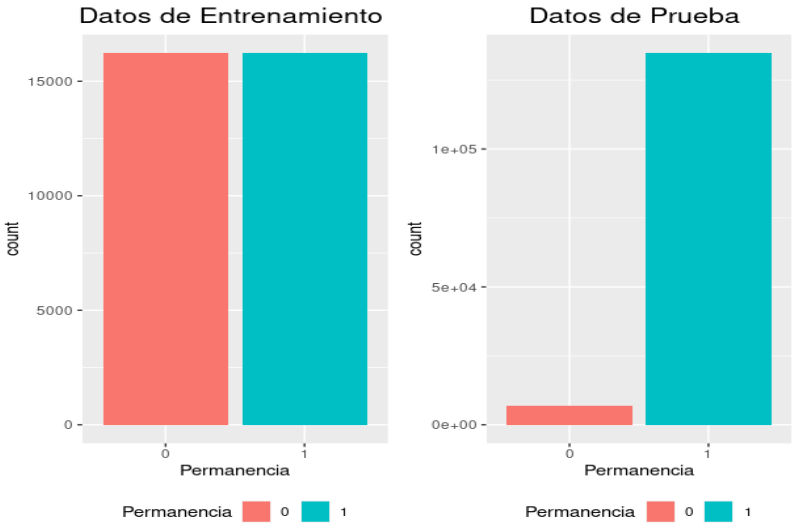

Figure 6. Distribution of training and test data.

After applying the Logistic Regression algorithm and the Decision Trees to the case study, the results shown in Table 2 were obtained: These results are quite interesting if the indicators Accuracy and Specificity are analyzed. The former determines the ability of the models to estimate the students who remain at the university and the latter the students who drop out. In this sense, the decision trees can identify more than 98% of the students who drop out, while the Logistic Regression only identifies 77%. As for the students who do not drop out, the best results are achieved by Logistic Regression. On the other hand, the Accuracy (General ability of the classifier) and the AUC (Ability to differentiate between the two classes) results are very similar.

Table 2. Comparative results

Indicator	Logistic Regression	Decision Trees
Exactitud	76.8%	75%
Precisión	98.4%	78%
Especificidad	77.51	98.54%
AUC	0.83	0.80

Table 3 shows the most important variables (sig < 0.1) for the Logistic Regression model. It can be seen that only 4 faculties are significant out of the 9 that the University has. All the academic performance variables, such as cut-off grades, final exam grades and attendance, are significant. Among the other socioeconomic

variables, residing in the province of El Oro and with whom he/she lives with are also important for the model, as well as the use of technology

Table 3. Significant variables for the Logistic Regression Model

Variable	Values
Faculty	Environmental Sciences, Engineering Sciences, Business Sciences and Distance Studies
Level	2nd, 3rd, 4th, 5th, 6th, 7th, 8th, and 9th
Cut-off mark 1	
Cut-off mark 2	
Final exam grade	
Supplementary Exam	1: Yes
Attendance	
State	Failed
Province	Del Oro
Who you live with	Son(A), Cousin(A)
Lives with	Yes
Who covers expenses	Other
Education level of the head of household	Not studied
Occupation of head of household	Operators and craftsmen, Plant and machine operations, Agricultural and fishing work, Service work and tradesmen, Unskilled workers.
Has anyone in the household used the Internet	Yes
Uses non-work email	Yes
Has a washing machine in the household	Yes
How many color televisions do you have in the household?	1

Has internet at home	Yes
Have a desktop computer	Yes
Have laptop computer	Yes
Where you do your homework	Yard
Use the library	Yes
Use Cyber	Yes

In the case of the Decision Trees, Figure 7 shows the structure obtained, where it can be seen that this model uses fewer variables than the Logistic Regression. For this model, the socioeconomic variables are not important and in the case of the variables Level and Faculty, it uses fewer categories.

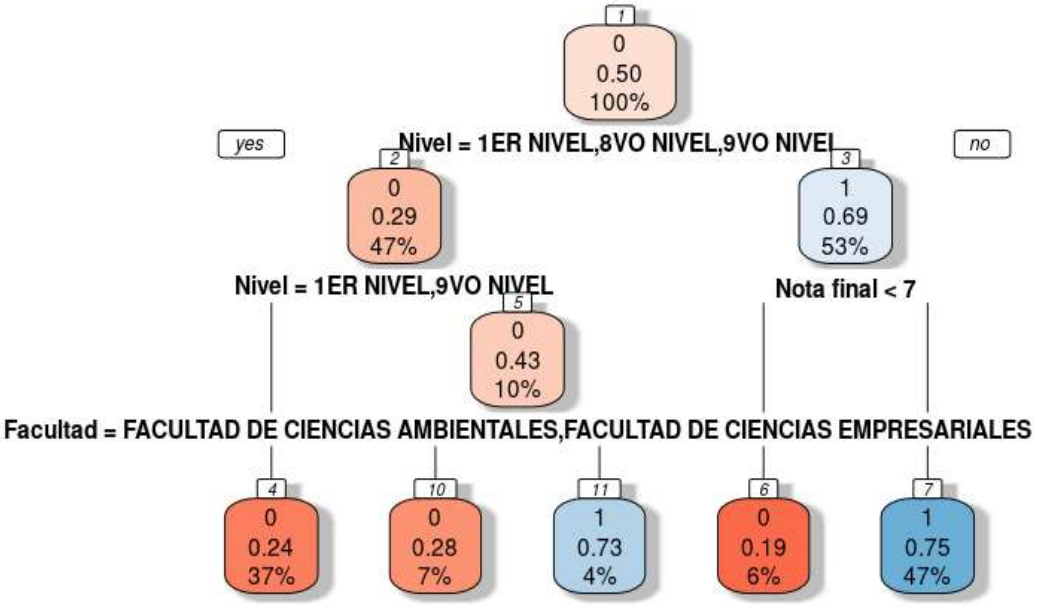

Figure 7. Decision Tree Model

In general, and considering that the objective of the research was to study student desertion in times of pandemic, using real data from the Technical State University of Quevedo, it is considered that the model obtained by the Decision Trees is the one that best fits the conditions of the study.

Conclusions

This paper presented a study of student attrition using supervised learning methods. The information was provided by the Universidad Técnica Estatal de Quevedo, Ecuador, and responds to 6 periods of study in COVID-19 time. From the results achieved in this work, it is concluded that:

- The quality of the data was not good, many variables had to be eliminated due to lack of information, and others such as age and average time spent on chores and at home had to be recorded because they had many imprecise values.
- To deal with the imbalance of the knowledge base, a random selection method by class was applied, obtaining the same amount of information for each value of the variable Permanence.
- Two learning models were applied for classification problems Logistic Regression and Decision Trees. Taking into account the indicators that were analyzed, it was determined that the Decision Trees were able to identify more than 95% of the cases of students who dropped out, unlike the Logistic Regression which was only able to identify 77% of the cases.
- Regarding the characteristics of the models, the Decision Trees used fewer variables than the Logistic Regression and the resulting model did not take into consideration any socio-economic variables.

In general, it was possible to carry out the study, providing results that may be important to help the decision-making process of the educational center authorities and reduce the impact of dropout in the classrooms. It is recommended, in the first place, to carry out a process of control and analysis of the online survey that the university conducts with the students at the beginning of each semester, to guarantee a better quality of the information.

References

Bonaccorso, G. (2017). *Machine Learninh Algorithms, A reference guide to popular algothms for data science and machine learning.* Packt Publiching.

Byron Oviedo, D. M. (2016, Junio 15). A hierarchical clustering method: Applications to educational data. doi:https://doi.org/10.3233/IDA-160839

Julian, C. J. (1995). *An Ad Hoc Analysis Strategy With Missing Data. The Journal of Experimental Education, 63(4), 333–342.* (Vol. 63). doi:https://doi.org/10.1080/00220973.1995.9943468

Kulis, B. (2013, Julio 31). Metric Learning: A Survey. *Foundations and Trends® in Machine Learning, 5*(4). doi:http://dx.doi.org/10.1561/2200000019

Landgrebe, S. R. (1991, Junio Mayo). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3). doi:https://doi.org/10.1109/21.97458

López, C. N. (2021). Diseño de un modelo matemático para estimar la deserción estudiantil mediante técnicas de análisis multivariado en una institución de educación superior tecnológica. *Universidad de Ambato*. doi:https://repositorio.uta.edu.ec/jspui/handle/123456789/32219

Martelo, R., Acevedo, D., & Martelo, P. (2018, Octubre 27). Análisis Multivariado aplicado a determinar factores clave de la deserción universitaria. *Espacios,*

39(10). Retrieved from https://www.revistaespacios.com/a18v39n10/a18v39n10p13.pdf

Miguel Ángel Medina Romero , Darwin Eliecer Solano Bent, Edwar Benjamín Bahoque Flórez, Rubén Jaime Huancapaza Cora , Pablo Ignacio Manrique Oroza , Edgar Salas Luzuriaga. A key to the quality of the educational institution: improvement plans. Journal of Positive Psychology & Wellbeing. 2021, Vol. 5, No. 4, 2390 – 2401.

Munizaga, F., Cifuentes, M. B., & Beltrán, A. (2018, Julio 5). Retención y Abandono Estudiantil en la Educación Superior Universitaria en América Latina y el Caribe: Una Revisión Sistemática. *Revista Académicas Universidad Tecnológica de Panamá, 26*.

Rizzo, M. L. (2019). *Statistical Computing with R.* New York, Estados Unidos: Tayloe & Francis Group. doi:https://doi.org/10.1201/9780429192760

Saldaña Villa, M., & Barriga, O. A. (2010, Diciembre). Adaptación del modelo de deserción universitaria de Tinto de la Universidad Católica de la Santísima Concepción, Chile. *Revista Ciencias Sociales, 16*(4), Universidad de Zulia Venezuela.

Tattar, P. N. (2017, Agosto 31). *Statistical Application Development with R and Python - Second Edition: Develop applications using data processing, statistical models, and CART.* USA: PACKT.

Torres, E. M. (2021). Análisis de la deserción estudiantil en la escuele profesional de contabilidad de la Universidad Nacional de San Agustín de Arequipa, usando técnicas multivariantes. *Universidad Nacional de San Agustín de Arequipa Escuala de Posgrado*. Retrieved from http://hdl.handle.net/20.500.12773/13772

Ungar, A. I. (2007, Agosto 4). Active learning for logistic regression: an evaluation. doi:https://doi.org/10.1007/s10994-007-5019-5

Nyandra, M., & Suryasa, W. (2019). Lifestyle for stress buffer and reverse cell aging. *International Journal of Health Sciences*, *3*(1), 17–23. https://doi.org/10.29332/ijhs.v3n1.276