Kasetsart Journal of Social Sciences 46 (2025) 460227

Kasetsart Journal of Social Sciences

journal homepage: http://kjss.kasetsart.org

Predicting student dropout risk using machine learning: A case study at the Technical University of Manabí

María Gabriela Cuzme Romero*,†, Jaime Meza†, Rodolfo García†

Information System Department Computational System, Faculty of Informatics, Technical University of Manabí, Portoviejo, Manabí 130101, Ecuador

Article Info

Article history:
Received 26 June 2023
Revised 26 May 2024
Accepted 7 June 2024
Available online 30 June 2025

Keywords:
algorithms,
crisp dm,
data mining,
random forest,
student dropout

Abstract

Student dropout is a problem that occurs in every Higher Education institution, which is why this research proposes the application of machine learning methods and algorithms allowing the risk of dropout in students at the Technical University of Manabí to be estimated. For this, the data collection process was developed, taking as reference the demographic and academic information of 10,002 students from different majors, information that was extracted through the Academic Management System. With the CRISP DM methodology, phases and processes were specified, taking as a sample a degree program from each of the faculties during the academic period from May 2014 to February 2019. Subsequently, the inclusion criteria were established to verify the regularity of the students who attended classes, through the Pearson correlation coefficient, the relationship between dropouts and non-regular students. Through the process carried out, three scenarios were obtained and the logistic regression algorithms, KNN, neural networks, support vector machines and the random forest algorithm were executed. As a result, the academic and demographic information of the students allowed us to verify the correlation between dropout students and non-regular students, so irregularity was an estimator of the risk of dropout that occurred in the institution, where the best scenario was during the second, third and fourth level of studies, with a margin of error of 0.05 evaluated by the metric evaluation system and with a high correlation in each of them, in the area under the AUC curve of 0.95 and an F1 score of 0.95.

* Corresponding author.
E-mail address: maria.cuzme@utm.edu.ec (Cuzme Romero, M. G.).
† Co-first author.

https://doi.org/10.34044/j.kjss.2025.46.2.27
2452–3151/© 2025 Kasetsart University.

Introduction

The Secretary of Higher Education, Science, Technology and Innovation (SENESCYT), is the regulatory body of the National Leveling and Admission System (SNNA), which currently governs all Higher Education institutions nationwide. Its main objective is to guarantee access to free higher education based on equal opportunities, meritocracy and transparency, through the use of new technologies, which is why the Technical University of Manabí is governed by the provisions issued by this regulatory body, (Canales & Rios, 2018). The objective of this study is to develop a predictive model to estimate the risk of dropout in students of the Technical University of Manabí, through the use of machine learning techniques and the selection of the most relevant characteristics in the academic database of Manabí, the institution.

Over time, public and private higher education universities negatively impact the social, economic, political and cultural processes of the nation's sustainable development. Due to this, some strategies that promote knowledge management and the quality of student training are implemented; however, university dropout is presented as a problem that limits the vision and mission of training competent professionals to improve the development of the country, due to different family, personal and pedagogical factors, where students abandon the university, impacting the country's progress in different social and scientific fields (Parra-Sánchez et al., 2023).

In addition, through different investigations, the results of a large number of students who do not complete their university studies are shown, related to the cost to the state. In this sense, education is characterized as a primary mechanism for countries to achieve higher levels of development. Therefore, through this research, the main drawbacks that may arise in the students of the Technical University of Manabí are described, as well as the analysis of the characteristics to define the best scenario and the irregularity of the students through the automatic learning model (Brownlee, 2023).

Literature Review

Through the bibliographic review, the research questions were defined according to the specific objectives that addressed the objective of the proposed research, for which a search of prior information was carried out to verify the existence of studies related to the research questions. Proposals. In addition, a cross-validation of studies was carried out to ensure that the inclusion and exclusion criteria were met. Due to this, according to the authors of the articles found, they allowed verifying the viability of the information search, as shown in Table 1.

Inclusion Criteria

- Articles published last in the last 5 years.
- Articles related to the subject of university dropouts and appearing in publications cataloged in specialized databases such as the area of computer science engineering.

Table 1 Research questions

Specific Objectives	Research Questions
SO1: Carry out a review of the state of the art to determine the most appropriate machine learning model creation techniques for predicting the risk of student dropout.	SO1 – AQ1.1: What kind of studies contribute to the definition of automatic learning techniques? SO 1 – AQ1.2: What are the characteristics of automatic learning techniques?
SO 2: Determine the current situation of student desertion at the Universidad Técnica de Manabí based on the analysis of data from the Academic Management System and the application of interviews with students and teachers.	SO 2 – AQ 2.1: What are the main dropout indicators found in the Academic Management System? SO 2 – AQ 2.2: What is the current status of school dropouts at UTM? SO 2 – AQ 2.3: What kind of techniques will be used for this type of study?
SO3: Implement the automatic learning model that, based on the determination of the most relevant characteristics, allows all of us predicting the risk of desertion in students of the University.	SO 3 – AQ 3.1: What is the automatic learning model to implement? SO 3 – AQ 3.2: How can the automated model predict the risk of attrition?
SO4: Perform the model validation using classifier performance evaluation metrics in prediction tasks, such as sensitivity, specificity, F1 score, and AUC ROC curve.	SO 4 – AQ 4.1: How can the proposed model be validated with automatic learning techniques? SO 4 – AQ 4.2: Which method is the most optimal to guarantee the evaluation metrics in the performance of the prediction tasks?

- Articles on aspects related to automatic learning predictive models.
- Articles indexed on the Scopus Database.
- Articles that included automatic learning techniques such as decision trees, Bayesian, neural networks, among others; these are used in the methodological process.
- Articles that mentioned definitions and study processes that have used the CRISP – DM methodology.

Exclusion Criteria

- Articles that dealt with the topic of university dropout and appearing in publications cataloged in specialized databases such as the area of pedagogy or social sciences in general.
- Articles that did not include automatic learning terms and those related to distance education in the abstract.
- Articles that did not take into account the dropout process in schools, and the comparison of various machine learning techniques.

After the conformation of the inclusion and exclusion criteria, the conformation of the control group was made, which was related to the studies that met the characteristics of the research analyzed by the researchers, considering the title of the study, abstract and keywords.

Through the search of the Scopus database, the inclusion and exclusion criteria were applied in order to determine the seven primary studies related to the problem, that was posed with the use of automatic learning techniques, resulting in relevant information on the above-mentioned topic.

It is concluded that according to the literary review, it was possible to see that in Latin America there is a high rate of student dropout compared to Europe due to the inconveniences that students sometimes have in continuing with their university studies. Some authors mention that the main characteristics are related to academic, social, demographic and economic aspects; Likewise, they showed that the most used algorithms are decision trees, logistic regression, Bayes and Random Forest, allowing reliable levels of precision to be achieved to identify predictors of dropout in public and private universities.

Consequently, this study will contribute to the generation of knowledge about the problem raised; through a more comprehensive perspective from the process of obtaining information, systematization and data modeling using machine learning techniques and finding the main indicators of student dropout in this Higher Education Institution.

Methodology

Data mining is a discipline oriented to the development of new methods and techniques to explore data, for which the present work is a type of predictive research and is linked to the information of the students of the Technical University of Manabí as a study of case. Due to this, the author (Rodríguez et al., 2016) mentions that to handle a large amount of stored data and to make the procedures less complex, methodologies have been designed to guide these processes, for which a comparative study highlights the methodology. CRISP-DM (Cross-Industry Standard Process for Data Mining) as the most used methodology in this type of procedures, managing to reach a deeper level of specification in each of its six phases, (Castro et al., 2018), as shown in Figure 1.

Figure 1 Phases of the CRISP – DM methodology
Resource: Cortina (2016)

Description of the Dataset

The dataset was collected form the Technical University of Manabí (UTM), the data of the students who were enrolled between the academic periods May 2014 to February 2019. Data collected were from the major with the highest number of students for each college and who were in the extensions of the institution.

Identification of Features

In order to identify the most important characteristics that control the results of these dataset, it was divided in three data groups: (1) identification data, (2) demographics, and (3) academic data[1].

[1] Dataset available https://n9.cl/exnsk

Data Processing

Phase 1: Understanding the Business. - The Universidad Técnica de Manabí (UTM), is a public university located in Portoviejo city, Manabí and has 10 colleges that are detailed below: Humanistic and Social Sciences, Health Sciences, Computer Sciences, Veterinary Sciences, Sciences Zootechnics, Administrative and Economic Sciences, Mathematical Sciences, Physics and Chemistry Science, Agronomic Study, Agricultural Study, Philosophy, Letters and Education Sciences; in addition, this school has 32 face-to-face courses, 8 virtual courses; and, four Institutes.

Therefore, the problem that arose about student desertion in the different majors was analyzed, where previous studies had been carried out by personnel of the Department of Student Welfare through surveys and manual processes in each academic department; where the process of student desertion was detailed in relation to the irregularity that they presented during their university studies. There was also a database with the academic and social information of the students. The main objectives for this process are detailed below:

- Identify student dropouts for each level.
- Classify regular students and those who are not.
- Establish the correlation of students between desertion and academic irregularity.
- Discover patterns and characteristics of students that affect academic irregularity.
- Generate training and evaluation data.
- Test various classification algorithms used in self - learning.

Phase 2: Data Understanding. - In this phase, the initial collection of the data was carried out to analyze and become more familiar with the information, verifying the quality itself. For this reason, the data of the students who were enrolled between the academic periods between May 2014 to February 2019 were necessary, taking as a sample the major with the highest number of students for each college and who were in the extensions of the institution, as shown in Table 2.

After the description of the data, the exploration was carried out, applying statistical tests through the Pandas Profiling tool that details the features and allowed the creation of distribution graphs of the data of the students of the colleges. In addition, several SQL requests were made to the University Database Management System for the data mining process. Due to the number of records and the crossing of tables, the information was obtained in .xls files where there were 286,798 records that belonged to the students in each subject. After exploring the data, some observations related to the duplication of records were verified, elimination of repeated fields, null fields, normalized values and fields.

Phase 3: Data Preparation. – After the information was selected through data mining techniques, a subset of data was selected, calculating as an estimator of the risk of desertion in a period of time, the condition of the academic regularity of the students, having to complete at least 60 percent of the subjects enrolled in each level of study. Due to the foregoing, it was based on the provisions of article 14 of the regulations of the Academic Regime of the Higher Education Council and article 10 of the current regulations of the Academic Regime of the Universidad Técnica de Manabí. In addition, the inclusion criteria were taken as:

- It is included if he was enrolled in at least 60 percent of the school subjects of the current level of studies, and the academic information was calculated in relation to the previous level, that is, only during one academic period.
- Social information for each student must be included.
- The label to take into account is: Are you a regular student of the next level enrolling in 60 percent of subjects?

Through SQL requests to the university database and through the tools presented by Python, each of the data were cleaned, where the total number of students by the colleges that are part of the sample was 9,994 and 286,728 records for each subject, having as analysis a correlation between the final grade and the attendance variable of 0.77, which defined that, in general, class attendance was a determining factor in the results of the teachers and therefore in retention.

In addition, the Python statsmodels.api tool was used, which includes a mathematical process within a function in which it verifies the correlation of the variables, categorizing them and seeing their importance.

Table 2 Sample of the population

Major	Number of students
Civil ingineering	1,941
Medicine	1,878
Business administration	1,608
Computer science	1,314
Social work study	1,037
Agronomic engineering	584
Agricultural engineering	569
Zootechnics	445
Aquaculture and fisheries	344
Physical education Pedagogy	282
Total of students	10,002

Results

The modeling phase applied the Pearson correlation coefficient with all the data that were obtained for each level. The correlation was verified with a value corresponding to 0.94, between non-regular students and dropout students as an estimate of the risk of desertion, as shown in the Table 3.

After the analysis was carried out and checking the correlation between the dropouts and the non-regular students, the selection of the data mining techniques that best fit the present investigation was made. However, due to the imbalance of the information on the regularity of the students in the colleges during the selected academic periods, it was necessary to carry out other processes before applying these techniques and defining each of the scenarios, as shown in the Figure 2.

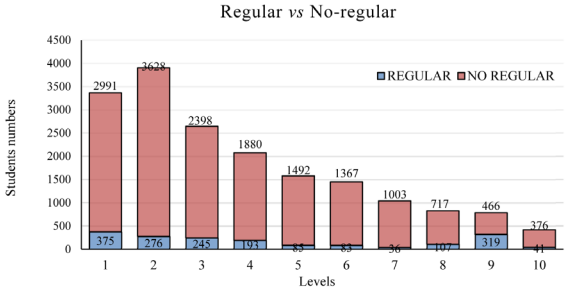

Figure 2 Analysis of the regularity of the students in each of the levels

The regularity of students on the first levels of study is detailed, which has greater emphasis on the second level because many students change their university majors and therefore the number of these results is greater. In addition, it was anticipated that the percentage of irregularity of the students from the eighth level was higher because for the students from these levels of study, depending on their curricular map in which they are, and the major they followed, the corresponding graduation process begins, and as many of them skip subjects, there is a minimum percentage of subjects that they take, which do not meet the inclusion criteria defined in the previous phases of the methodology. Therefore, after this analysis, the imbalance of the majority and minority classes was verified, creating three scenarios according to the educational levels.

For the data imbalance process, it was necessary to verify the correlation of the variables through the stats model tool, which is a Python module that provides classes and functions for the estimation of many statistical models. Due to this, the sampling methods that were used allowed eliminating instances of the majority class through under-sampling and those that generate new instances of the minority class using over-sampling.

After the results were obtained according to the inclusion criteria, the following techniques were chosen to apply: Logistic Regression, Random Forest, K-nearest neighbors (KNN), Neural Network, Support Vector Machines (SVM - RBF). Each of these mentioned techniques were part of the Python scikit learn tool that provided a few facilities for performing these types of supervised classification algorithms.

To carry out the tests in each of the scenarios, the data were divided into two parts, one of which is that the training data will be set to create the different prediction models depending on the method that will be used. On the other hand, to verify the reliability of the model, the set will be the one that tests the process, which allows evaluating, and also it is capable of predicting this model from new data, applying sampling methods using over-sampling and under-sampling that is useful for data imbalance.

Three scenarios were established to apply each of the algorithms in relation to the different levels of study of the students and the percentage of regularity of the same, according to the points established in the inclusion criteria. In addition, the previously detailed learning techniques, which were evaluated by the evaluation metric system of the area under the AUC ROC curve and the F1 score, were applied. Likewise, according to the results obtained, one of the best methods for data imbalance applied in this research work was the undersampling and oversampling of the sampling method used in the data mining processes, as shown in the Figure 3.

Table 3 Pearson correlation matrix between student dropout and irregularity

	Regular	Non Regular	Dropout students	Non Dropout students
Regular	1	0.241326	0.307019	0.929028
Non-Regular	0.241326	1	0.940415	0.267564
Dropout students	0.307019	0.940415	1	0.271540
Non Dropout students	0.929028	0.267564	0.271540	1

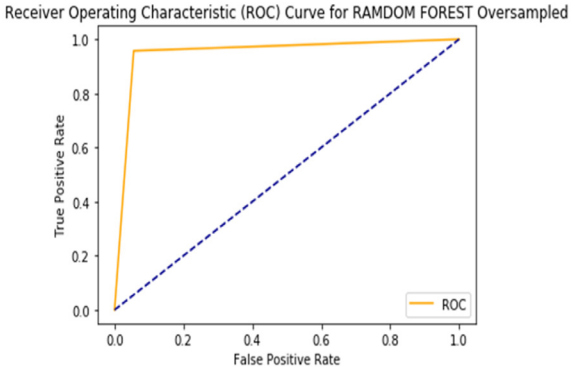

Figure 3 AUC ROC curve of the random forest learning algorithm

In the previous figure, the oversampled sampling method was used, which details an area under the AUC ROC curve of 0.95, with a margin of error of 0.05, that is, it allows interpreting a lower percentage of error in the regularity of the students during the study and second, third and fourth level of their studies. Likewise, the F1 score is 0.95 in relation to the weighted average of precision and sensitivity of the used model for which there is some similarity between each of the metrics, obtaining a better evaluation of the model in relation to the obtained result.

Phase 5: Evaluation. - In the development of this research, five models were applied in three different scenarios, where the percentage of success was different in each of them. Each of the results is shown in Table 4.

The iteration of the values in each of the algorithms is visualized, in which the scenario with the best results is the second one that refers to the second, third and fourth level of study, where the Random Forest classification machine learning algorithm performs a better prediction on the regularity and irregularity of the students in the different majors of the students.

In addition, according to the study of the art carried out, it was possible to realize that in Latin America there is a high rate of student desertion compared to Europe, due to some problems that students sometimes present to continue with their university studies, according to the results obtained in the present investigation. Likewise, some authors mentioned that the main characteristics are referring to the academic, social, demographic and economic aspects, where they pointed out that the most used algorithms were decision trees, logistic regression, Bayes and Random Forest, allowing to reach reliable levels of precision by identifying dropout predictors in public and private universities.

Consequently, this study will contribute to the generation of knowledge about the problem raised through a more comprehensive perspective from the process of obtaining information, systematization and data modeling.

Discussion

As time goes by, public and private universities in Ecuador record data on dropout of undergraduate students. The personal, academic and institutional factors that affect the university stage influence students to abandon their studies. In addition, it is necessary to obtain data to recognize the causes where students at the first levels leave their careers and withdraw, allowing university authorities to make decisions to reduce this rate, and thus be able to take corrective actions that are focused on formulating policies and strategies.

Therefore, according to what (Garrido Silva & Pajuelo Diaz, 2023) mentions, student dropout is a problem that directly affects the educational training of young people, which is why it is an issue of great concern within university institutions. Given the impact of this situation on the educational system, it is important to be able to monitor it, as well as analyze its causes and consequences in order to propose corrective measures that mitigate its impact on students.

Table 4 Comparative results of the learning algorithms in each of the scenarios

N	Algorithm	Scenario 1 Level 1,2,3			Scenario 2 Level 2,3,4			Scenario 3 Level 3,4,5		
		Score F1	Area Under Auc Roc Curve	Test Error	Score F1	Area Under Auc Roc Curve	Test Error	Score F1	Area Under Auc Roc Curve	Test Error
1	Knn	0.89	0.88	0.12	0.93	0.93	0.07	0.86	0.85	0.15
2	Red Neuronal		0.82			0.91			0.93	
3	Svm – Rbf	0.88	0.88	0.12	0.90	0.91	0.09	0.85	0.82	0.18
4	Random Forest	0.91	0.91	0.09	0.95	0.95	0.05	0.92	0.93	0.07
5	Logistic Regression	0.88	0.87	0.13	0.89	0.90	0.10	0.83	0.82	0.18

In general terms, combating this problem requires an improvement in both individual and institutional conditions, establishing a reciprocal interrelation between these two factors that allows students to complete their higher studies according to what Otero Escobar (2021) mentions, which does not deviate from the causes that generally affect the majority of young people in the world.

Through the search strings of the Scopus databases used in the state of the art, the inclusion and exclusion criteria were applied with the objective of determining the primary studies related to the problem raised with the use of learning techniques in an automatic way allowing the creation of a predictive model of student dropout at the first levels of each of the careers, which determines the probability of a student abandoning his studies, (Roldan Robles, 2019).

Therefore, according to the proposed research, it was determined that the academic and demographic information of the students through the data mining process showed the correlation between dropout students and non-regular students, so irregularity was an estimator of the risk of dropout that occurred in the institution.

Likewise, through the analysis of the characteristics to define the scenarios and execute the supervised machine learning model, the irregularity of the students in the first academic levels was verified, showing that many students abandon these levels during their university career. Where their academic performance and variables from their personal environment were taken as a reference, such helped obtain results through supervised algorithms and applied evaluation metrics.

Therefore, based on the evaluation metrics, the expected results were calculated with the Random Forest classification algorithm that made a better prediction on student dropout with a margin of error of 0.05, assessed by the metrics of the area under the AUC ROC curve of 0.95 and the F1 Score of 0.95, with a high correlation in the values of each of the evaluation metrics, allowing us to determine that student irregularity is the main drawback of dropout, where attendance affects academic performance.

Recommendation

Include the personal information of the students in the system, multiple selection fields, in such a way as to avoid certain errors that occur at the time of the data manipulation. Periodically update the demographic and personal information of the students and keep all those records independently.

Carry out a mid-term follow-up through the Academic Management System of the students during the ordinary academic period to verify the development of each one of them and thus avoid student desertion. Carry out a data mining system, which contains observer and report generators that provide consolidated and quality data mainly to the academic units of each of the majors, and make it easy to get them, in such a way that it helps in decision-making with the other authorities.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

Brownlee, J. (2023). A Gentle Introduction to k-fold Cross-Validation. *Machine Learning Mastery*. https://machinelearningmastery.com/k-fold-cross-validation/

Canales, A., & Ríos, D. (2018). Factores explicativos de la deserción universitaria. *Revista Calidad en la Educación: Educación Superior: Diversidad y Acceso,* (26), 173–201. https://doi.org/10.31619/caledu.n26.239

Castro R., L.F., Espitia P., E., & Montilla, A.F. (2018). Applying CRISP-DM in a KDD Process for the Analysis of Student Attrition. In C. J. Serrano & J. Martínez-Santos (Eds.) *Advances in Computing, CCC 2018*. Springer. https://doi.org/10.1007/978-3-319-98998-3_30

Cortina, V. G. (2016). Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario. https://hdl.handle.net/10016/22198

Garrido Silva, C. A., & Pajuelo Díaz, J. (2023). Deserción en estudiantes de educación superior: un estudio de caso [Dropout among students in higher education: a case study]. *Universidad, Ciencia y Tecnología*, *27*(119), 18–28. https://doi.org/10.47460/uct.v27i119.703 [in Spanish]

Otero Escobar, A. D. (2021). Deserción escolar en estudiantes universitarios: estudio de caso del área económico-administrativa. [School drop-off in university students: case study of the economic-administrative area], *RIDE Revista Iberoamericana Para La Investigación Y El Desarrollo Educativo*, *12*(23). https://doi.org/10.23913/ride.v12i23.1084 [in Spanish]

Parra Sánchez, J. S., Torres Pardo, I. D., & Martínez de Merino, C. Y. (2023). Factores explicativos de la deserción universitaria abordados mediante inteligencia artificial. [Explanatory Factors of University Dropout Explored Through Artificial Intelligence], *Revista Electrónica de Investigación Educativa*, 25 (e18), 1–17. https://doi.org/10.24320/redie.2023.25.e18.4455 [in Spanish]

Roldan Robles, P. R. (2019). Desarrollo de una arquitectura conceptual para el análisis de contenidos en redes sociales sobre el tema del aborto usando Python [Bachelor's thesis, Repositorio Digital Universidad Técnica del Norte]. https://repositorio.utn.edu.ec/handle/123456789/9026

Rodríguez León, L. C., & García Lorenzo, D. C. M. M. (2016). Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor [Adaptation to a methodology of data mining for applying to un-supervised problems type attribute-value]. *Universidad y Sociedad*, *8*(4), 43–53. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202016000400005 [in Spanish].