

Detection of Academic Patterns in the Course of Leveling with Dropout in the Technical University State of Quevedo Based on Socioeconomic Factors

Eduardo Amable Samaniego Mena¹, Angel Torres Quijijej¹, Juan Carlos Pisco Vanegas¹, Byron Oviedo Bayas¹, Jean Carlos Almeida Murillo²

¹*Universidad Técnica Estatal de Quevedo, Facultad de Posgrado.*

²*Independent Researcher*

Email: esamaniego@uteq.edu.ec

Currently, universities in Ecuador face great challenges due to student dropout, so it is urgent to take measures to identify the factors that affect this problem. To achieve the objectives, a descriptive and exploratory analysis was carried out on a set of data obtained from the Academic Management System department of the State Technical University of Quevedo. Through this project it is proposed to address the problem by implementing Data Mining algorithms with the aim of discovering patterns that cause student dropout. The knowledge discovery in database (KDD) methodology was used, which consists of selection, preprocessing, transformation, data mining and data evaluation phases. For the preparation and purification of the data set, RStudio and Weka are used to apply the J48, Decision Stump, Random Tree, Random Forest, Hoeffding Tree, LMT and Rep Tree algorithms. To choose the optimal algorithm for the study, each of them will be quantitatively evaluated by precision. In the results obtained, it was highlighted that the Random Tree algorithm had better results in the analyses, followed by the Random Forest, this algorithm allowed us to identify patterns directly associated with student dropout.

Keywords: Student Dropout, KDD Methodology, Weka, RStudio.

1. Introduction

Currently, students face numerous challenges to complete their studies, including economic problems, limitations in internet access and family difficulties, among others. Universities aim to generate knowledge and pass it on to students to encourage critical thinking and scientific research. However, the student's purpose is often thwarted when, due to various factors, they drop out of college, which is known as "student dropout" (Vega et al., 2022).

Dropout in higher education represents a complex and multifaceted challenge, without a universally accepted definition. This phenomenon can manifest itself as the interruption of studies within the same institution, transfer to another, or total abandonment of the educational system. Despite the increase in access to higher education, completion rates have not experienced an equivalent increase. This approach focuses on institutional environments as modifiable factors to address and mitigate student attrition. The variety of causes and contexts that influence dropout requires specific approaches in each study, previous research has tended to focus on the individual perception of students, which has limited the consideration of other determining factors that also influence this phenomenon (Lorenzi, 2024).

One of the pillars of a country's economic and social development is education; The population that has the opportunity to study a professional career generally achieves a better job and higher economic income than people who did not, or who only attended basic education. In some countries, certain sectors of the population do not have educational opportunities, so they are limited in their development as individuals and as a society. Access to education is, without a doubt, transcendental, however, the fact that students complete their professional training and do not fail in the attempt to achieve it, is equally important, as it allows them to have the necessary tools and skills to face future challenges and opportunities (Cornejo Sifuentes et al., 2023) .

Globally, statistics show that only 50% of students who start higher education manage to graduate, according to a study conducted in Latin America. Student dropout is a phenomenon present in all countries, including developed ones. In the Latin American context, the figures indicate that only 1 in 10 students between the ages of 25 and 29 has completed their higher education, according to the Regional Office of Education for Latin America and the Caribbean (OREALC)(Michelle & Polytechnic, 2024). The selection process for university admission should not only provide for students' future academic performance, but also include methods that identify the potential for success in those from disadvantaged backgrounds. This would help to avoid discrimination on the basis of ethnicity, race, age or gender, and to reduce the risk of dropping out of education systems. Student dropout, known as a global problem, has effects that go beyond the individual and affect the development of societies in which dropouts are integrated (Tapasco Alzate et al., 2019).

University dropout is a far-reaching problem that affects the whole society, given its impact on the reduction of intellectual capital and the risk of increasing poverty, this phenomenon is affected by various factors, including the educational environment and the socioeconomic conditions of students (Villegas & Núñez Lira, 2024). Student dropout is relevant when observing that most students come mainly from middle and low socioeconomic strata, in addition to belonging to rural areas, which places them at a disadvantage compared to the rest of the student population. This context has significant repercussions on university dropouts, which are perceived as a social problem with effects at the personal, family and institutional levels (Cárdenas Matute et al., 2023).

In the last decade, school dropout has become one of the most researched problems by academic bodies. A low educational level in the population translates into reduced productivity in all economic sectors. Unfortunately, this is a problem with global repercussions, according to recent UNESCO data (Lee et al., 2020). The increase in access to higher education since

the end of the twentieth century has resulted in the implementation of numerous institutional policies aimed at promoting effective democratization. In Ecuador, the Organic Law of Higher Education (LOES) guarantees equal opportunities, recognizing education as a right for all people. In recent years, there has been a significant increase in the massive admission to higher education in all social strata. However, state universities face limitations in their academic offerings, which forces applicants to seek places in co-financed or private universities. This situation generates various problems for both institutions and students. (Cárdenas Matute et al., 2023)

Student dropout refers to the time when a student unsuccessfully attempts to complete their educational project, resulting in the abandonment of their academic activities for two consecutive semesters or classes. This phenomenon represents a challenge for the education system, since it implies the definitive abandonment of academic training by the student, regardless of the circumstances. Educational institutions are forced to implement strategies to integrate students into university life and thus improve their academic success (Madrid Orrego et al., 2023). University dropout, which involves a student dropping out of a program of study or an institution before completing their degree, is a global phenomenon of great relevance.

This phenomenon has significant economic and social consequences, not only for the affected student, but also for society as a whole (Castro et al., 2023). The educational system of Ecuador faces significant difficulties that prevent its adequate development, with school dropout being a prominent concern, particularly in public institutions, the reasons behind this situation are diverse and cover social, cultural, political, economic and social aspects (Fariás Moreira & Angulo Caicedo, 2024). The local effects of school dropout are considerably impactful and difficult to predict. For this reason, investigating the factors that lead to school dropout among students in the Ecuadorian education system has become crucial. This approach seeks to improve understanding of the multiple negative impacts on students, their families, and society at large. Among the most prominent problems are: lack of access to opportunities, health and well-being problems, economic insecurity and an increased risk of social exclusion. In addition, students from vulnerable and marginalized sectors are more likely to drop out of school due to the unfavorable economic situation of their families in rural areas (Loaiza Maldonado. et al., 2023).

School dropout represents a complex educational challenge that profoundly impacts not only the people directly affected, but also the social and economic fabric of a country. In Ecuador, this problem acquires significant importance, given that its causes and effects are linked to the cultural diversity and socioeconomic challenges characteristic of this South American nation (Carrión & Rivera, 2023). Student dropout represents a considerable challenge in the field of higher education, affecting technical, technological and university institutions alike. This phenomenon not only has negative repercussions on the political, economic, social and cultural aspects of national development, but also adversely affects educational processes at these levels (Armas, 2020).

In Ecuador, access to higher education is guaranteed by the Organic Law of Higher Education (LOES), which is based on equal opportunities for all students based on their merits. In addition, it is recognized as a public good through which students can generate relevant and high-quality knowledge. In response to this regulation, higher education institutions (HEIs)

have increased their academic offerings to ensure access to the higher education system. However, admission policies, academic requirements, minimum GPAs, availability of places and scores in the admission process have been significant barriers for many students. As a result, those who are not accepted into their preferred programs often choose to abandon the opportunity or, failing that, enroll in careers that are not of interest to them in order not to be excluded from the system, which has become one of the main causes of academic dropout in the university environment (Enrique et al., 2024).

Faced with this scenario, it seeks to detect the patterns that cause student dropout in leveling courses, focusing on the case of the State Technical University of Quevedo (UTEQ), taking into account socioeconomic and academic factors. To achieve this objective, the most significant variables related to dropout will be identified, and the most appropriate analysis technique to detect dropout patterns will be determined.

2. Materials and Methods

The Academic Management System (EMS) of the UTEQ constitutes the centralized platform for the storage and management of digital academic resources, which are available to students on a continuous basis. This system has facilitated the collection of data necessary for the implementation of this project, incorporating socioeconomic variables in the analysis. The data used in the study were extracted from UTEQ's SGA platform, which acts as the main repository of information related to academic performance, student records, and other pertinent data during the student's academic career at UTEQ. The data analyzed correspond to two academic years, 2022 and 2023. In 2022, UTEQ offered a total of 25 careers and 11 subjects. In contrast, during the year 2023, significant changes were observed in the academic structure: in the first period, 24 careers and 18 subjects were offered, while in the second period of the same year they increased to 30 careers and 18 subjects. These changes in the academic offer have been considered for the analysis of the data and the evaluation of the impact on student performance. For the elaboration of the project, the KDD methodology was used (Menéndez Domínguez et al., 2022) As shown in Figure 1.

Illustration 1: KDD methodology



Within this methodology, the 5 phases provided by the methodology were used, being selection, preprocessing, transformation, data mining and finally interpretation. Phase 1 of selection involved the identification and acquisition of relevant data sources to detect academic patterns and their relationship with dropout in leveling courses at UTEQ, with a focus on socioeconomic variables. Data on students' academic performance and socioeconomic status were collected, including contextual information such as start dates. Initially, we worked with a dataset composed of 99 variables and 23,008 records. Where an exhaustive filtering was

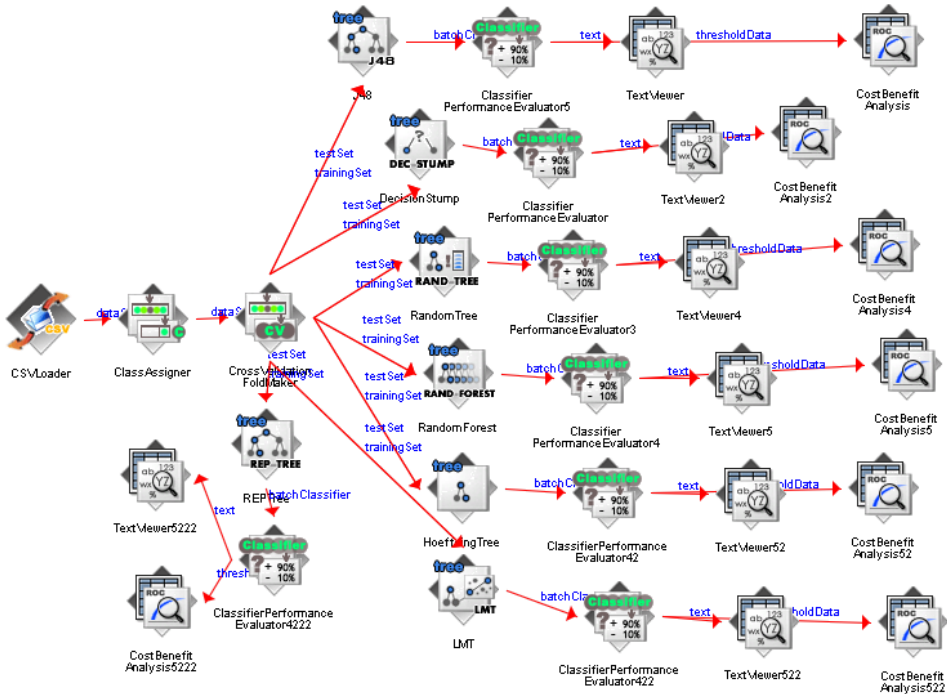
carried out to focus on the pertinent variables related to failed students, thus guaranteeing the relevance and coherence of the data for the analysis.

In phase 2 of Preprocessing, missing values, duplicates, and errors in academic and socioeconomic data were addressed using the RStudio tool. Text normalization was applied with the `iconv` function to remove accentuations and missing values were imputed, assigning 0 to certain columns and "NO PADDING" to others. Unnecessary content in square brackets in the "Subject" variable was eliminated and the word "leveling" was filtered in the "Career" variable to consolidate the information. The variable "Period" was standardized by removing redundant characters to improve the temporal clarity of the data. In addition, the data was filtered to focus on failed students, thus preparing a cleaner and more relevant dataset. Considering that in phase 3 of transformation, numerical variables were treated and categorized into descriptive groups such as "Insufficient", "Sufficient", "Good", "Remarkable" and "Outstanding" to facilitate the identification of patterns and improve the quality of the analysis allowing a more accurate interpretation of academic data, highlighting associations and trends relevant to student performance. In order to identify patterns in the dropout of leveling courses in phase 4 of data mining, several machine learning algorithms were used, including J48, Decision Stump, Random Tree, Random Forest, Hoeffding Tree, LMT and Rep Tree. These algorithms allowed to analyze the relationships between socioeconomic and academic variables and to detect significant patterns to finally move on to phase 5 of evaluation in which it focused on determining the best classification algorithm using performance metrics, including accuracy and the Kappa coefficient. These metrics made it possible to evaluate the accuracy of the predictions and the agreement between the predicted and actual results, ensuring the selection of the most effective model.

3. Results and Discussion

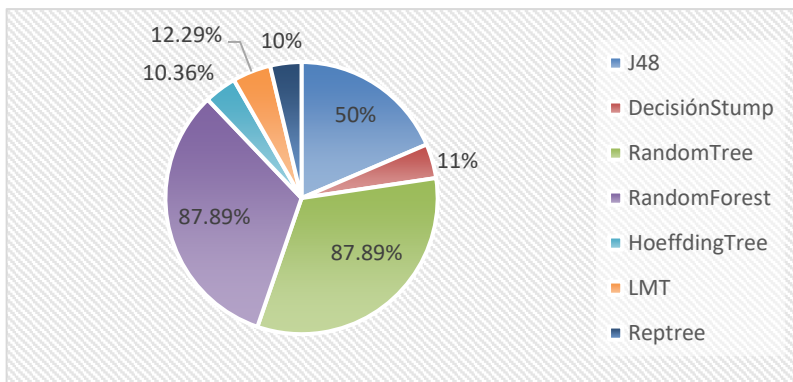
Three analyses were conducted, starting with the analysis of the full dataset. Figure 2 shows a detailed schematic of the process for obtaining variables, which includes several interconnected components. The procedure starts with the `ArffLoader`, which is responsible for loading the data from the original set. The obtained variables are then used to divide the dataset using the `Crossvalidation` component, which prepares the data for training and testing processes. A variety of machine learning algorithms, including J48, Decisión Stump, Random Tree, Random Forest, Hoeffding Tree, LMT, and Rep Tree, are then applied to build predictive models. Once the models are generated, their performance is evaluated using the `Classifier PerformanceEvaluator` component, which performs the evaluation with test data. Finally, the results of the assessment are documented in a text file using the `CostBenefit Analysis` component.

Illustration 2: Modeling Classification Algorithms



In the first analysis, the effectiveness of different algorithms was evaluated using a specific dataset, as illustrated in Figure 3, which shows the variable selection process. The results of this analysis indicated that the Random Forest and Random Tree algorithms stood out remarkably in terms of accuracy. Both algorithms achieved an outstanding accuracy of 87.89%, which reflects a strong agreement between the predictions generated by the models and the actual observations.

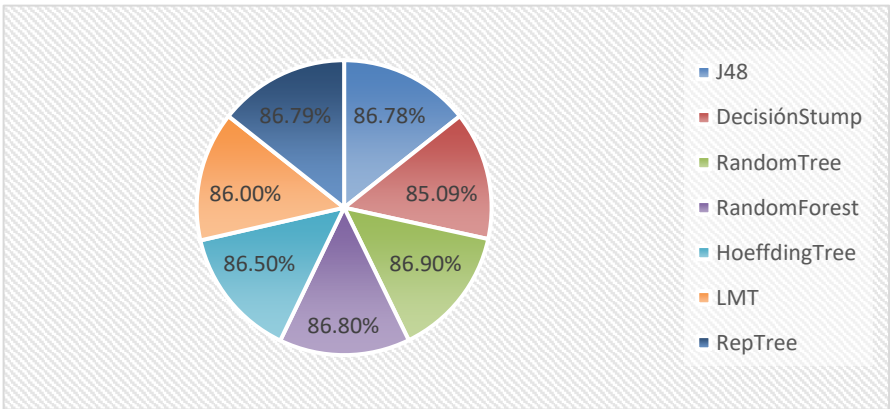
Illustration 3: Result of the first analysis



In the second analysis, the performance of the same algorithms was evaluated, but this time considering exclusively the socioeconomic variables of the dataset. As shown in Figure 4, the

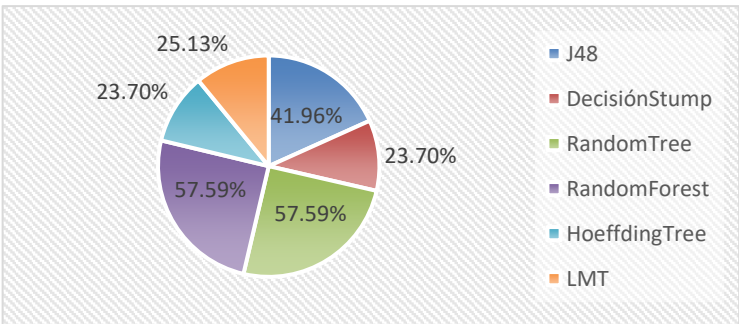
results indicated that the Random Tree algorithm excelled in terms of accuracy in this specific setting. The Random Tree achieved an accuracy of 86.9%, which, although slightly lower than that obtained in the first analysis with the full set of variables, is still highly significant. This result suggests that the Random Tree maintains a strong predictive capacity even when restricted to a subset of variables, in this case, socioeconomic variables. The accuracy of 86.9% indicates that the model has a high degree of agreement between its predictions and actual observations, which is indicative of robust performance in the classification task based on socioeconomic factors

Illustration 4: Result of the second analysis



In the third analysis, the performance of the same algorithms was examined using exclusively academic variables, as illustrated in Figure 5. The results obtained revealed that the Random Tree and Random Forest algorithms stood out for their efficiency in terms of accuracy, reaching a remarkable accuracy of 57.59%. This level of precision, although lower than that recorded in the two previous analyses, indicates that both algorithms maintain a respectable predictive capacity when they focus only on academic variables. The accuracy of 57.59% suggests a moderate correlation between the model's predictions and actual observations, indicating that academic variables have a significant, though not as strong as socioeconomic variables, on the models' performance

Illustration 5: Result of the third analysis



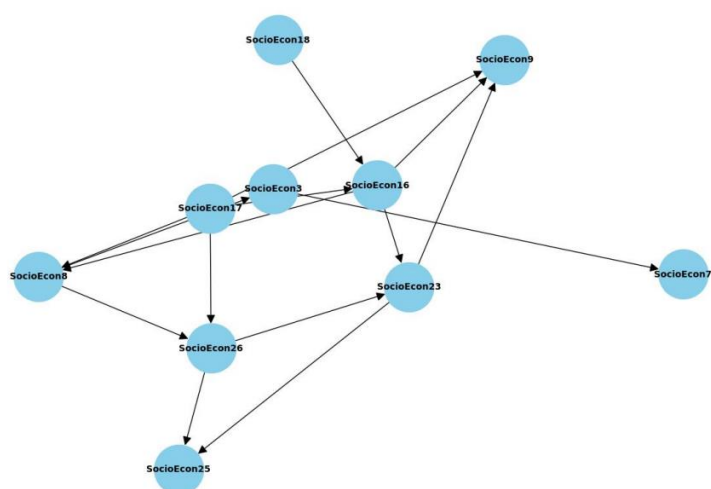
Student dropout is a complex phenomenon that can be influenced by a variety of interrelated factors. Among these, academic performance stands out as one of the most prominent. In

particular, the performance reflected in the final grade of the first cut can be a critical indicator of the difficulties faced by the student. Poor academic performance may indicate problems in adapting to the pace of the course or meeting academic expectations, which often leads to demotivation and, ultimately, dropout from the educational program. In addition to academic performance, the status of head of household emerges as a determining factor in student dropout. Students who take on additional responsibilities, such as caring for family members or managing family income, face a significant burden that can affect their ability to devote time and effort to their studies. This delicate balance between academic obligations and personal responsibilities can be overwhelming, leading them to prioritize family needs over their education, which increases the risk of dropping out.

Another crucial aspect to consider is access to basic resources, such as adequate housing, internet access, and the availability of a computer. The lack of these resources can create significant barriers to effective participation in learning, especially in distance education contexts. Lack of access to technology and digital resources not only hinders a student's ability to keep up with the curriculum, but also widens educational disparities, contributing significantly to school dropouts. The digital divide represents a considerable challenge, exacerbating inequalities and limiting educational opportunities for those who lack the necessary resources. In addition, socioeconomic and contextual factors, such as family income level and parental job stability, also play a role in student dropout. Students from backgrounds with unstable economic or work constraints may face greater difficulty concentrating on their studies due to financial concerns and a lack of a supportive environment.

In addition to the use of classification algorithms, the use of association rules was implemented to identify the factors that were associated as shown in Figure 6

Illustration 6: Variable Association Nodes



As illustrated in Figure 6, the analysis of the selected nodes and edges reveals significant interconnections between various socioeconomic variables. This analysis reveals a notable correlation between access to digital services, such as the internet and computers, and other characteristics of the home environment, such as the type of housing and the presence of *Nanotechnology Perceptions* Vol. 20 No. S9 (2024)

recreational areas. These findings suggest that the level of technological development in a home can have a considerable impact on the lifestyle and physical environment of its residents. The observed relationship between access to educational resources, such as encyclopedias and libraries, and the availability of digital services reinforces the idea that access to technology influences the way information is accessed and used at home. This correlation has important implications for access to education and knowledge, given that the availability of digital services can facilitate participation in educational activities and the use of information resources that, in turn, can influence academic performance and educational success.

Another aspect highlighted in Figure 6 is the connection between financial responsibility, represented by the person covering the student's expenses, and the type of housing. This relationship suggests that the financial resources available in a household can influence its ability to access different types of housing. The quality and type of housing, in turn, can affect the overall well-being of household members, including the student. Adequate and stable housing is crucial to providing an environment conducive to study and personal development. In addition, a relationship is observed between internet use in the last six months and the student's financial responsibility, which underscores how economic resources can be linked to access to digital services. This connection has important implications for participation in online education and other educational opportunities. Access to the internet, facilitated in part by the financial capacity of the household, can be a determining factor in a student's ability to participate in virtual educational programs, access study materials, and participate in academic activities that require a digital connection.

4. Conclusions

A variety of decision tree models were used using the Weka tool, including Stump, Hoeffding Tree, J48, LMT, Random Forest, Random Tree, and REP Tree, and evaluated using three datasets: complete, academic, and socioeconomic. The analysis of the full set allowed for a comprehensive assessment of the predictive performance of each model. In the case of academic data, it was observed that students with low grades in the first cut have a high probability of dropping out of the leveling course, underlining the importance of initial academic performance as a predictor of dropout. On the other hand, the analysis of socioeconomic data revealed that the lack of access to the internet and computers at home, combined with economic dependence, significantly increases the risk of dropping out of the course. These findings highlight the critical influence of academic and socioeconomic variables on student dropout, suggesting the need for comprehensive strategies that address both academic support and socioeconomic barriers to improve student retention and educational success.

References

1. Armas, S. (2020). Analysis of the factors that affect the student dropout of Ecuadorian children and adolescents in the period 2009-2019. National Polytechnic School, 1–64. <http://bibdigital.epn.edu.ec/handle/15000/21111>
2. Cárdenas Matute, J. M., Valle Franco, A., & Tapia Segarra, J. I. (2023). Factors that affect *Nanotechnology Perceptions* Vol. 20 No. S9 (2024)

student dropout in the academic unit of Social Sciences of the Catholic University of Cuenca. *DigitalAwareness*, 6(3), 30–48. <https://doi.org/10.33262/concienciadigital.v6i3.2621>

3. Carrión, A., & Rivera. (2023). Causes and consequences of school dropout in the Ecuadorian context: Causas e consequências do abandono escolar no contexto equatoriano, *Ciencias de la Educación*, Research Article. 85(11), 927–945. <https://doi.org/10.23857/pc.v8i11.6246>
4. Castro, L. F., Espitia, E., & Edwin, R. (2023). Analysis of characteristics influencing student dropout in the context of a Latin American university. *Eia*, 20(40), 1–28.
5. Cornejo Sifuentes, M. C. S. G., Vega Pérez, L. G., Naranjo Cantabrana, M. G., Osúa Acosta, I. I. F., Ávila Santana, F. A., & Sotomayor Fierro, M. de los Á. (2023). Predictive Model of School Dropout in Higher Education: An Approach from Data Mining Using the CRISP-DM Methodology. *Ciencia Latina Revista Científica Multidisciplinar*, 7(5), 7797–7812. https://doi.org/10.37811/cl_rcm.v7i5.8363
6. Enrique, L., Granda, Á., Yautibug, F. C., & Raúl Aucancela Copa. (2024). Dropout in Higher Education in Ecuador, Causes and Consequences. 11474–11490.
7. Farias Moreira, C. R., & Angulo Caicedo, R. A. (2024). Factors that affect student dropout: Case of the Luis Tello Higher Technological Institute. *Ciencia Latina Revista Científica Multidisciplinar*, 8(1), 10518–10533. https://doi.org/10.37811/cl_rcm.v8i1.10357
8. Lee, L. E., Martínez, S. I., Castán Rocha, J. A., Terán Villanueva, J. D., Menchaca, J. L., Treviño Berrones, M. G., & Rocha, E. C. (2020). Evaluation of Prediction Algorithms in the Student Dropout Problem. *Journal of Computer and Communications*, 08(03), 20–27. <https://doi.org/10.4236/jcc.2020.83002>
9. Loaiza Maldonado., D. J., García Neira., I. A., Romero Ambi., J. P., Díaz Triviño., M. de L., & Ronquillo Cabezas, P. S. . (2023). Identification of the factors of academic dropout in the educational system of Ecuador. *Ciencia Latina Revista Científica Multidisciplinar*, 7(2), 11121–11136. https://doi.org/10.37811/cl_rcm.v7i2.6190
10. Lorenzi, G. (2024). Dropout in Argentine higher education. Digital skills as a new challenge for inclusion. 1–23.
11. Madrid Orrego, D. M., Figueroa Royero, L., Berrio Caballero, H., Martínez Choles, A. Y., & Charris Rodríguez, J. A. (2023). Data mining for the identification of Student Dropout in Public Universities. *LATAM Latin American Journal of Social Sciences and Humanities*, 4(2), 1464–1476. <https://doi.org/10.56712/latam.v4i2.697>
12. Menéndez Domínguez, V. H., Guerrero Sosa, J. D. T., Castellanos Bolaños, M. E., & Cervera Pérez, J. W. (2022). Use of data mining for the characterization of researchers and academic bodies. In *RIDE Ibero-American Journal for Educational Research and Development* (Vol. 12, Issue 24). <https://doi.org/10.23913/ride.v12i24.1144>
13. Michelle, D., & Politécnica, U. (2024). Incident factors of probability of student dropout in a higher technological institute : a statistical analysis. 5.
14. Tapasco Alzate, O. A., Ruiz Ortega, F. J., Osorio García, D., & Ramírez Ramírez, D. (2019). Student dropout: incidence of institutional factors related to admission processes. *Education and Educators*, 22(1), 81–100. <https://doi.org/10.5294/edu.2019.22.1.5>
15. Vega, H., Sanez, E., De la Cruz, P., Moquillaza, S., & Pretell, J. (2022). Intelligent System to Predict University Students Dropout. *International Journal of Online and Biomedical Engineering*, 18(7), 27–43. <https://doi.org/10.3991/ijoe.v18i07.30195>
16. Villegas, B. R., & Núñez Lira, L. A. (2024). Factors associated with student dropout in the university environment. A systematic review 2018-2023. *RIDE Ibero-American Journal for Educational Research and Development*, 14(28). <https://doi.org/10.23913/ride.v14i28.1923>