EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

Discovering the anatomy of school dropouts with data science: case study, Technical Professional Education in Mexico

Descubriendo la anatomía de la deserción escolar con ciencia de datos: caso de estudio, Educación Profesional Técnica en México

Juan Gabriel Villagrán-Castañeda[1]; Rosa María Valdovinos-Rosas[2]*; Angélica Guzmán-Ponce[3]; Marcelo Romero Huertas[2]

[1]Colegio Nacional de Educación Profesional Técnica CONALEP, México.
[2]Faculty of Engineering, Universidad Autónoma del Estado de México, México.
[3]Institute of New Imaging Technologies, Department of Computer Languages and Systems, Universitat Jaume I, Spain.

Email de correspondencia: *rvaldovinosr@uaemex.mx*

Cronograma editorial: *Artículo recibido 10/05/2024 Aceptado: 17/04/2025 Publicado: 01/07/2025*

Para citar este artículo utilice la siguiente referencia:
Valdovinos Rosas, R. M. (n.d.). Descubriendo la anatomía de la deserción escolar con ciencia de datos: caso de estudio, Educación Profesional Técnica en México. EDUCA. Revista Internacional Para La Calidad Educativa, 5(2), 1-21. https://doi.org/10.55040/ydfc9j29

Contribución específica de los autores: Los autores han participado conjuntamente en todas las fases de la investigación.

Financiación: No existió financiación para este proyecto.

Consentimiento informado participantes del estudio: Se han solicitado los consentimientos informados de los participantes.

Conflicto de interés: Los autores no señalan ningún conflicto de interés.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

Abstract. School dropouts are a major educational problem that influences the economic development of a country, social well-being, and individual growth. This article uses data science techniques to identify the determinants of dropout in career technical education in Mexico. To do this, we consider students' academic achievements and information related to socioeconomic and psychological aspects. The results indicate the importance of exploring factors beyond academic performance to understand the causes of school dropout in CONALEP.

Keywords: school dropout, data science, decision trees, classification

Resumen. El abandono escolar es un problema educativo importante que influye en el desarrollo económico de un país, el bienestar social y el crecimiento individual. Este artículo utiliza técnicas de ciencia de datos para identificar los determinantes de la deserción escolar en la educación técnica profesional en México. Para ello, se consideran los logros académicos de los estudiantes y la información relacionada con aspectos socioeconómicos y psicológicos. Los resultados indican la importancia de explorar factores más allá del rendimiento académico para comprender las causas de la deserción escolar en el CONALEP.

Palabras clave: deserción escolar, ciencia de datos, árboles de decisión, clasificación.

1 Foundation

School dropouts constitute a major challenge in the education sector. This phenomenon is characterized by students gradually disengaging from the academic environment and eventually leaving the educational system without obtaining a school certificate. The latter prevents them to access to the university, as well as practicing a profession, stopping the professional opportunities that a school certificate would offer them. An example of the latter is the case of Latin America (Alvarado-Uribe et.al, 2022; Fernandez-Haddad and Gonzalez, 2023; Rodríguez et.al, 2023) where 50% of students discontinue their studies, and only half of the remaining complete their programs on time (Quinn, 2013).

School high school dropout in Mexico is one of the main challenges, since it reports the highest dropout rate, far above primary and secondary education. According to the Ministry of Public Education, for the 2021-2022 cycle, school dropouts for high school

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

reached 9.2%, while for primary and secondary education was 0.4% and 2.5% respectively (IMCO, 2022) (Figure 1).

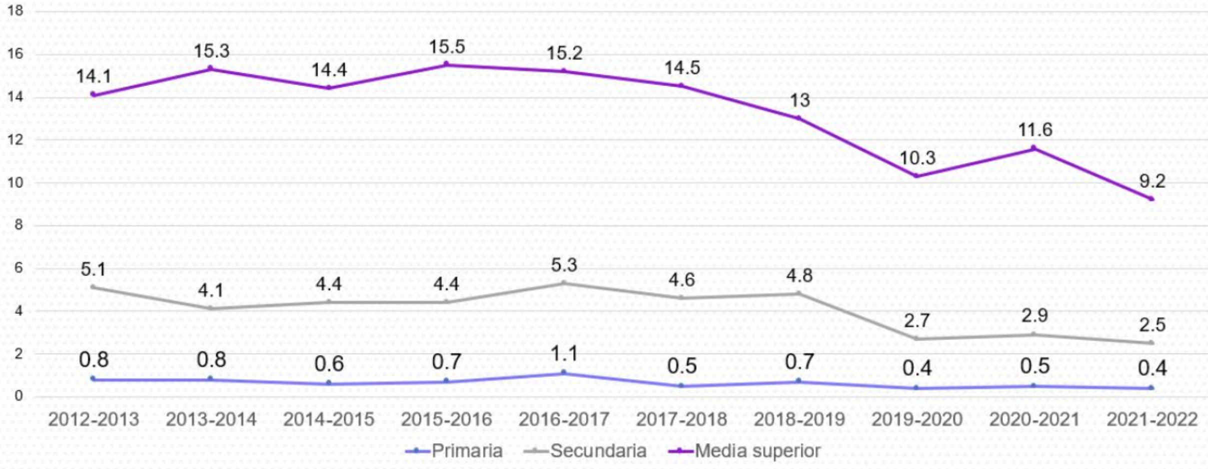


Figure 1. School dropout rate from 2012 to 2021 in Mexico[1].

In the field of educational analytics, the prediction and prevention of school dropout is one of the most pressing challenges. The latter because the problem transcends personal situations, significantly impacting the social and economic progress of countries.

From a comprehensive perspective, school dropout is recognized as a critical and complex issue within the educational landscape. It arises not only due to poor academic performance or school attendance problems, but as a symptom of an extensive collection of situations that encompass several dimensions between the individual, family, school and community levels (De Witte et.al, 2013). In this sense, school dropout is not solely the result of individual academic failure or personal inadequacies, but rather a manifestation of a broader social dynamic that requires a holistic and multifaceted response.

With the advent of Data Science (DS) and Machine Learning (ML), educators and decision makers are increasingly turning to these technologies for practical solutions (Rodríguez et.al, 2023; Nai et.al, 2023; Salinas-Chipana et.al, 2024). These models provide valuable information by analyzing patterns within vast data sets that include academic, demographic, economic, and health-related factors (Sarker, 2021; Martínez-Plumed et.al, 2021). This analysis reveals hidden patterns, predicts academic

[1] https://imco.org.mx/bachillerato-el-escalon-fragil-de-la-educacion/

outcomes (such as at-risk students), improves learning experiences, and formulates effective educational strategies to improve overall educational quality (Bošnjaković and Durdevíc, 2023; Yagci, 2022).

Some state-of-the-art works analyze student behavior and the reasons for dropping out. Santoso (2019) used the *K*-means algorithm, decision trees, and Naïve Bayes to predict student behavior. The study used academic and non-academic data derived from the admissions processes, focusing on socioeconomic information. The results demonstrated that the socioeconomic environment significantly influences academic performance.

In the same line, Lottering et.al (2020) studied the increasing dropout rates in South African higher education. The authors use Data Mining techniques to identify students likely to dropout school, specifically in the Faculty of Information and Communication Technologies. Using the Support Vector Machine and Random Forest classifiers, they were able to identify at-risk students with an *F*-measure score of 99.32%.

The study by Ribeiro and Canedo (2020) was focused on the University of Brasilia (UnB) to identify factors that affect student dropout rates. To do this, the authors consider the course workload, duration, academic performance, and student fee status. The Gradient Boosting Machine model was used, which predicted a dropout rate of 54% among UnB undergraduate students, underscoring the need to improve monitoring of students at risk of dropping out.

A study by Pérez-Gutiérrez (2020) compared decision trees, logistic regression and Naive Bayes to detect student dropout in a Systems Engineering degree. The results revealed correlations between performance in systems engineering, physics, and mathematics courses, and variability in semester grades, with the probability of student dropout. In the end, the authors emphasized the significant impact of systems engineering courses in predicting school dropout.

The study by Urbina-Nájera et.al (2020) used feature selection algorithms to identify critical factors that influence university dropout decisions. They considered a sample of 300 students from public Higher Education Institutions (HEIs) and 200 from private HEIs. They identified 27 relevant factors, with lack of tutoring, an inadequate

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

student environment, and insufficient academic monitoring as the three main factors. Additionally, using decision trees, they identified seven dropout patterns, including variables such as student environment, financial support, and situational discomfort. The study provided comprehensive information, allowing the development of precise and personalized strategies to the characteristics of the students to reduce university dropout rates.

On the other hand, Nahar et.al (2021) predicted the performance of engineering students. Data from 80 computer science and engineering students were analyzed, considering academic records and surveys. Two data sets were created: one to categorize and predict student performance against prerequisites in the course; and another to predict final grades based on partial results. A decision tree was used for the first data set and the Naive Bayes algorithm for the second. The research aimed to guide students proactively based on predictions of their academic results.

Finally, Nájera and Ortega (2022) developed a model to anticipate university dropouts in Engineering, Social and Administrative Sciences. With data from 2010 to 2019, using artificial neural networks and the Random Forest algorithm. The study found that using the majority of attributes and balancing classes the performance of the models can be improved. Future research suggests incorporating more attributes and exploring various classification algorithms to refine the predictive model.

Recognizing the urgent need to improve the quality of higher education and reduce school dropout rates in Mexico, this study offers a valuable contribution through the collection and analysis of a comprehensive data set. As case study, we consider the National College of Professional Technical Education (CONALEP), which is aligned with the National System of Technological Education, emerged in 1978, established as a decentralized institution of the Mexican federal government. The CONALEP goal is the comprehensive training of technical professionals, providing them with skills to continue studying at the university level or join to the productive sector upon graduation. The institution has more than 290,000 students distributed across 308 campuses within the 32 states of Mexico (Bernal-Reyes, 2020).

CONALEP evaluates educational performance using indicators which are determined by the Secretariat of Planning and Institutional Development. These indicators

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

measure student failure rates, graduation efficiency, and dropout rates. The indicators serve not only to reflect the state of the educational system, but they also help to identify how close the institution is to its goals, as well as discover areas of opportunity. The main indicators for evaluating school dropouts include total enrollment, failure rate, dropout rate, and graduation efficiency rate, all reviewed at the end of each school year. This system works as a control panel, facilitating the identification of problems and their analysis to solve.

In this way, CONALEP is committed to the student community and seeks to implement strategies aimed at reducing school dropout rates. Therefore, the academic, socioeconomic and psychological circumstances of students are crucial for their permanence in school.

CONALEP uses the business intelligence tool known as Tableau[2] to help senior management make informed decisions. However, Tableu does not consider the correlation between non-academic indicators and school dropout, only provides predominantly numerical data, such as percentages, without delving into the underlying factors that influence attrition rates.

This research is focusing on the campuses of the State of Mexico that offer the Professional Technician in Computer Science degree. These campuses were selected for their high enrollment and dropout rates reported. Using data science strategies, the complex phenomenon of school dropout rates will be analyzed on data from students of the Computer Science major at the CONALEP. This career was chosen because it is notably affected by a dropout rate close to 50%.

2 Methodology

To develop the research, we use the CRoss-Industry Standard Process for Data Mining (CRISP-DM) model, which is a methodology widely recognized for its structured project approach (Martínez-Plumed et.al, 2021). CRISP-MD breaks down the project into a six-stage life cycle, which are described in the following sections.

[2] https://www.tableau.com/es-mx

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

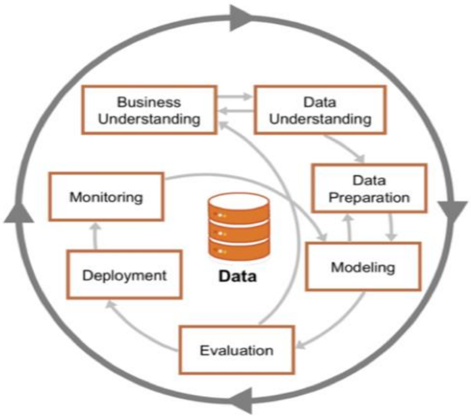


Figure 2. CRISP-MD methodology[3].

2.1 Business Understanding

This phase is essential to fully understand the educational objectives and requirements of the problem to be solved. It involves a thorough analysis of the educational landscape from an institutional perspective, with the goal of translating these requirements into technical objectives for a data-driven educational project.

In CONALEP, the School Administration System (SAE) represents an essential component in the case study. This system manages critical administrative tasks and plays an important role in the overall educational framework. The SAE includes fundamental operations such as managing student enrollment, orchestrating the creation of academic groups, assigning teachers, recording evaluations and generating official documentation.

The approximately 1.5 terabyte database contains complete information on more than 2 million students who have been part of CONALEP. The research presented in this paper is focused on the State of Mexico, which has 41 of the 308 campuses, offers 54 carriers and attend approximately 290 thousand students. The volume and variety of data housed in this system presents both an opportunity and a challenge for educational data mining. Analysis this data can allow us to show patterns and trends associated with school dropouts. Additionally, it allows informed decision-making, explicitly focusing on interventions and improvements in educational practices within CONALEP. The goal is

---

[3] https://medium.com/analytics-vidhya/understanding-crisp-dm-and-its-importance-in-data-science-projects-91c8742c9f9b

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

to identify early warning signs and factors that contribute to school dropouts, facilitating proactive measures to improve student retention and success.

2.2 Data Comprehension

This phase focuses on data collection and exploration, involves a detailed description of the data, as well as identifying their various sources in the educational context. The CONALEP database is a comprehensive source of student information, covering personal and academic aspects, such as group assignments, grades, study plans, and graduate data, among others. This data mine facilitates in-depth analysis of student performance and trends, crucial for identifying patterns in academic performance and dropout rates.

The data used in this research correspond to the State of Mexico, which has 41 of the 308 campuses, offers 54 carriers and attend approximately 290 thousand students. Specifically, the study is focused on students who began their studies in August 2016 and graduated in August 2019, with a total of 2,390 instances. The data set has been divided into two classes: graduate students (G) and dropouts (D).

In addition to the data mine mentioned above, information collected with the Educational Context Survey, created by the Mexican Ministry of Public Education, is available. This survey offers detailed information on demographic and psychosocial aspects of students, including age, gender, employment, social-emotional skills, and extracurricular activities. In this way, the analysis considers academic, socioeconomic and psychological aspects that can influence student dropout. This comprehensive approach is crucial to plan more robust and practical strategies designed to avoid high dropout rates and substantially raise the level of higher education throughout the region. Specifically, three research questions are addressed in this work:

1. How do socioeconomic status, psychological well-being, health, and academic performance influence student dropout rates?
2. What are the main non-academic factors (socioeconomic, psychological aspects and health conditions) that contribute to student dropout?
3. How do psychological well-being along with health status and academic performance affect school dropout rates?

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

These questions are critical to understanding the complex interplay between different factors that can affect student retention and success in higher education.

2.3 Data Preparation

This step includes selecting features relevant to the problem studied, such as student engagement metrics or course completion rates. In addition, it involves the creation of new variables, cleaning data to ensure accuracy and reliability of the model to be trained and modifying the format of some data to make them compatible with machine learning algorithms.

2.4 Modelling

In this phase, one or more machine learning techniques are strategically applied to deduce meaningful insights from the data. The experiments were carried out considering the Weka J48 classifier with its default settings. This classifier was selected due to its effectiveness in handling data sets that include mixed data types. Furthermore, the cross-validation method was applied to ensure the robustness of the model against overfitting and validate its predictive performance on different data subsets.

2.5 Evaluation

The evaluation ensures that the model accurately addresses the defined questions, considering the results related to the educational objectives. To evaluate the model obtained, several metrics derived from the well-known confusion matrix were obtained. This matrix provides valuable information about the overall performance and effectiveness of the classifier. In a two-class problem, the confusion matrix is shown in Table 1 (Saini and Chhabra, 2024).

Table 1. Confusion matrix in a binary problem.

Real	**Predicted**	
	C_1	**C_2**
C_1	True Positive (*TP*)	False Negative (*FN*)
C_2	False Positive (*FP*)	True Negative (*TN*)

From the confusion matrix, the next metrics can be obtained:

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

- *Sensitivity or Recall* measures the number of positive instances thatwere correctly classified, and is given by: $\frac{TP}{TP+FN}$
- *Specificity* establishes the effectiveness rate at which a classifier identifies instances from a class as negative: $\frac{TN}{TN+FP}$
- *Overall Accuracy* measures the probability that the classifier can correctly recognize instances of both classes, and is given by: $\frac{TP+TV}{TP+FN+TN+FP}$.

3. Results and Discussion

In this section are analyzing and interpreting the results obtained considering the following three scenarios:

a. **Scenario 1:** This is the most complete scenario. With this, we analyze the impact that socioeconomic, psychological, health and academic factors (results of disciplinary evaluations) have on student dropout rates.

Scenario 2: In this scenario, academic attributes are excluded, and the analysis is focused on socioeconomic, psychological and health data only. The goal is to discover non-academic factors that influence the student dropout.

b. **Scenario 3:** In this scenario, socioeconomic aspects are excluded and only psychological, health and academic aspects are used. The goal is to understand how these factors interact and their impact on students' educational trajectories.

3.1 Holistic analysis of factors that influence student dropout

The data set used considered all aspects of the study: socioeconomic, academic, psychological and health. With the data used, a decision tree was built comprising 64 nodes and 48 leaves. Graduate students are represented by G, and dropouts are represented by D (Figure 3).

From the tree in Figure 3, students who interrupted their studies are identified by following the branches that end in leaves labeled "D" (Dropout). By carrying out this monitoring it is possible to obtain detailed information about the factors that contributed to student dropout.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

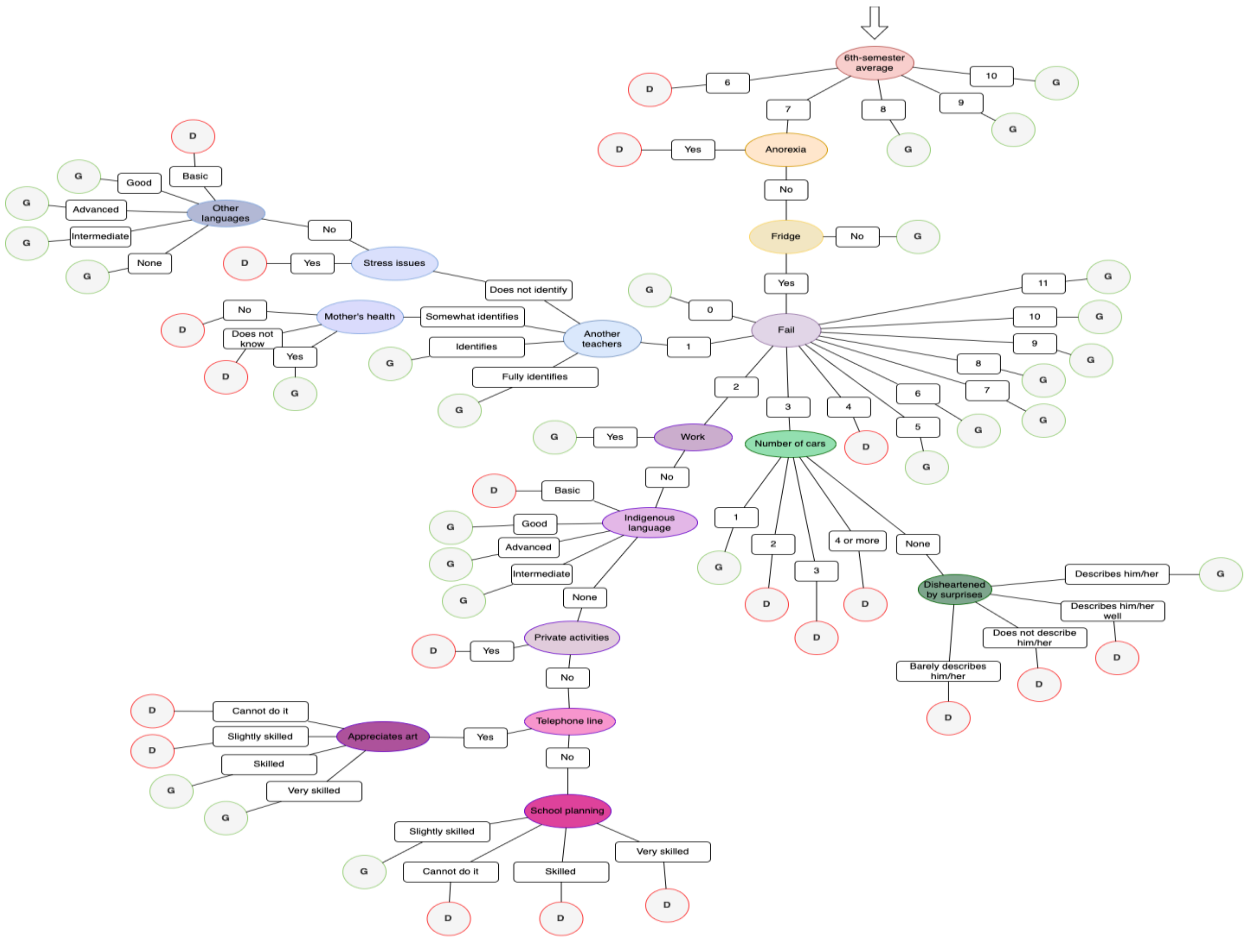
1. Something notable is to observe that when the students reached an average grade of 6 (not passing) in the sixth semester, 1,092 students dropped out due to not reaching the minimum passing grade of 7.
2. For students with an average grade of 7, other determinants were observed:

- Several groups of two students decided to drop out the school due to anorexia problem, stress, or lack of basic ability in a foreign language (English), regardless of the presence of other variables such as failing subjects, not having anorexia, not working, or speaking an indigenous language.
- In contrast, several groups of five students dropped out despite not having anorexia, failing two subjects, not working, not speaking indigenous language, lacking various skills in planning school activities, art appreciation skills or lack thereof.The factors that were observed to have influence in a larger group of students (12 students) were the mother's health and reaction to unexpected problems. Likewise, the number of failed subjects significantly influenced student dropout; most of the students who dropped out were between 2 and 4.

Figure 3 Decision tree from the first scenario.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

– Students with an average grade of 8, 9, or 10 in the sixth semester were classified as graduates by the algorithm. This classification suggests that a high academic level could be an indicator of success, preventing deeper analyzes of dropout factors. However, because not all students earn these grades, aside from academic



performance, there are additional factors that significantly influence student dropout rates.

The Area Under Curve (AUC) is a global indicator of the accuracy of a 2-class model. It is interpreted as the probability of correctly classifying a pair of patterns, this area is always greater than or equal to 0.5 and the range of values moves from 0 to 1 (perfect discrimination). For this scenario, the model obtained an accuracy rate of 95.94%, with a sensitivity rate of 97% (true positive 1,184) for students categorized as graduates and a specificity rate of 94% (true negative 1,110) (Figure 4).

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

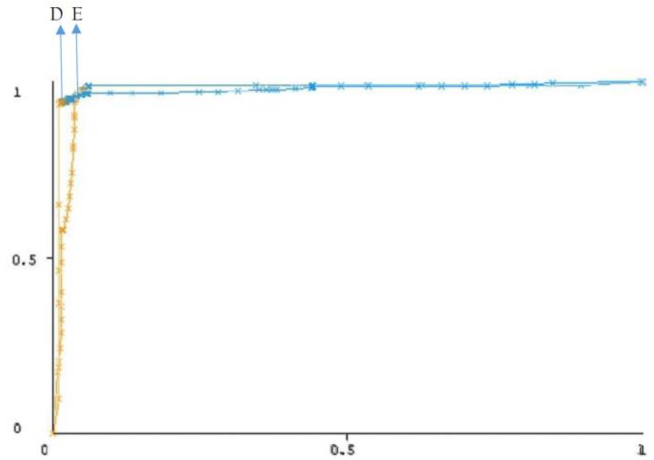


Figure 4. ROC curve for the first scenario.

2.2Exploring beyond academics: what non-academic factors influence student dropout?

For the second scenario, in the training process we do not consider information related to failed subjects per semester and semester averages. The obtained decision tree includes 844 nodes with 638 leaves. With this model, it was possible to accurately identify 1,369 of 2,391 samples, resulting in an accuracy rate of 57.25% (Figure 5).

This result shows the impact that academic information has. Not considering these academic indicators notoriously difficult by the classifiers to identify the underlying causes of school dropout.

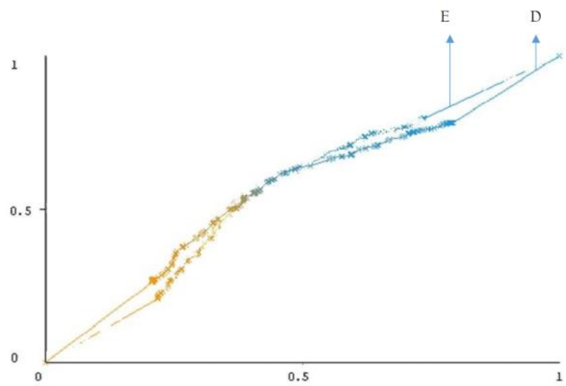


Figure 5. ROC curve for the second scenario.

Due to the large size of the tree, we focus on the most significant results.

1. Students with an average grade of 6 in the sixth semester of the degree: Two cases related to poor nutrition or size of the house abandoned their studies. Groups of seven

students dropped out due to the amount of time they spent studying at home, others because they had limited communication with their friends or because of stress problems and occasional feelings of loss of control. Finally, fourteen students dropped out despite participating well in class, along with three others who did not.

2. Students with an average of 7 in secondary school revealed the following: Four dropouts were linked to families with monthly income between 1,000 and 2,000 Mexican pesos, others attributed to low self-esteem and three to poor communication with teachers. Five students needed further communication with a psychologist or counselor or had inadequate sexuality education. Six students dropped out due to occasional feelings of loss of control and others due to lack of group presentation skills and behavioral problems.

   Twelve dropouts occurred despite having a computer at home, potentially influenced by their mothers' health. Twenty students who dropped out lacked proficiency in reading English texts, thirty-four due to intensive use of the Internet for social media, four more for entertainment, and ninety-five because they had a job.
3. For students with an average of 8 in secondary school, the observations include: Two cases due to incompetence in laboratory or field experiments, other two students were linked to problems solving the EXANI-I guide and lack of ability to express opinions. Five men dropped out due to not receiving support during situations of emotional distress and another five due to vision or behavioral problems.

   Two dropouts occurred among students where the mother's education level was primary school, five due to vision problems, and seven linked to insufficient sexual education or related to stress. Finally, fifteen dropouts associated with depression and attendance in evening classes.
4. Students with an average of 9, the dropout causes included: Not submitting homework or assignments, behavioral problems, and excessive time spent on video games.
5. Students with perfect average of 10, the results identified: Twelve students dropped out despite do not have behavioral or homework-related problems, and two others because they did not lead a healthy lifestyle.

2.3 The triad effect: psychology, health, and education on dropout rates

In Scenario 3, we explore the psychological, health and academic aspects. A decision tree with 51 nodes and 68 leaves was created (Figure 6). Of 2,391 cases, 2,301 were classified correctly, giving an accuracy rate of 96.23%. Of the students who obtained an average below 7 in the 6th semester (i.e., failed), 1,092 students discontinued their education. While students who achieved an average equal or upper 7, the following confusion matrix and its corresponding ROC curve are observed (Figure 7).

1. Students who achieved an average of 7 in the 6th semester, to fail 2 - 4 subjects from the third semester, onwards showed school dropout in several students.
2. Some aspects that, together with failing subjects, result in school dropout are: Three students with an average of 7 in the first semester, who presented depression, dropped out after failing three subjects in the sixth semester. Other three students dropped out despite not being inhibited by unexpected challenges, but after failing one class in the fourth semester and two in the sixth semester.
3. Four students with an average of 8 in high school and did not face nutritional, visual, or hearing disabilities dropped out after failing two subjects in the sixth semester. Other four students, who overcame unexpected challenges and were not affected by Internet-related issues, dropped out after failing one course in the third semester and another in the sixth semester.
4. Five students, who did not suffer from depression and had good self-esteem without hearing difficulties, abandoned their studies after failing three subjects in the sixth semester.
5. Most students who achieved averages of 8, 9 and 10 in the sixth semester finished their studies without notable problems.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

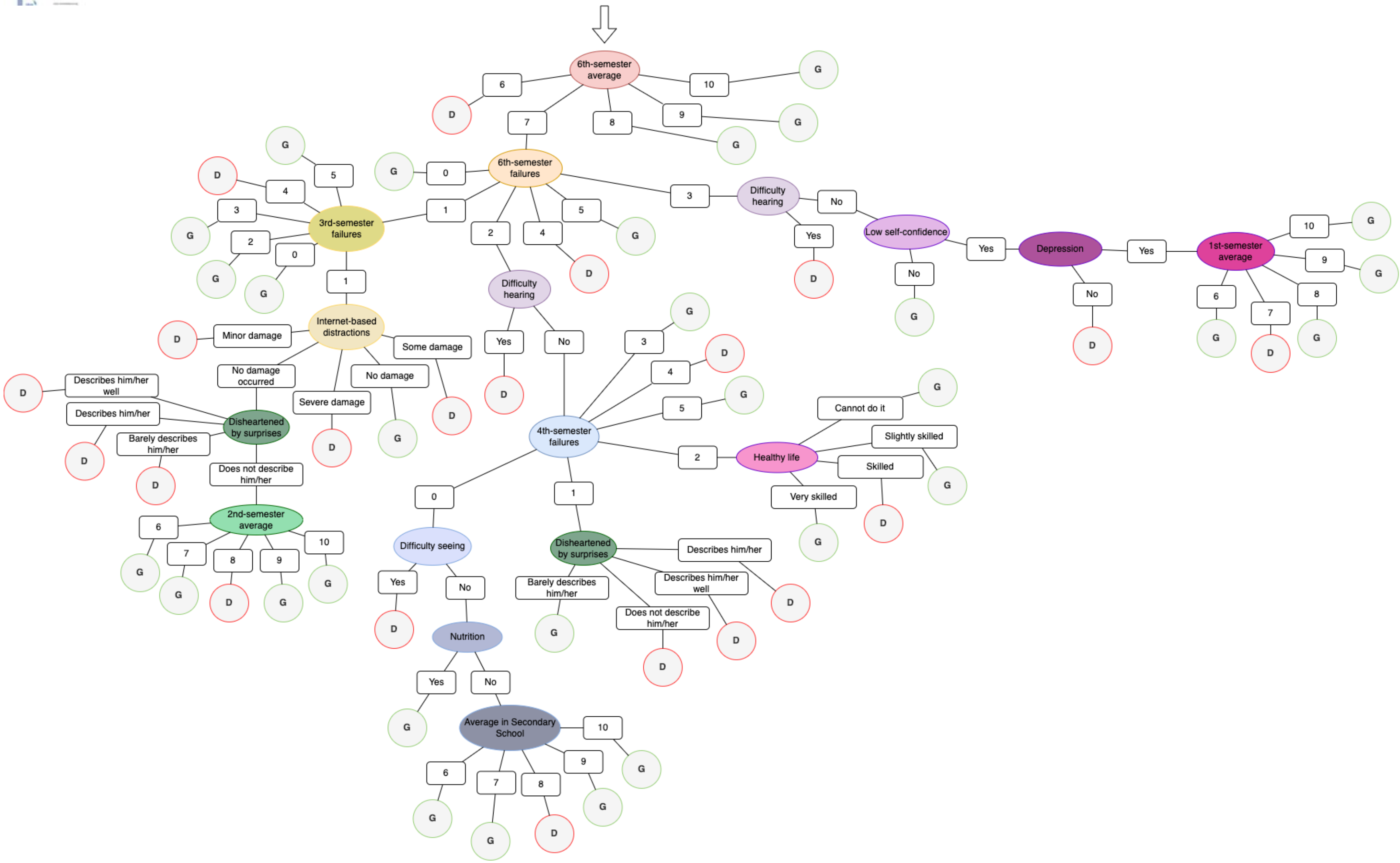


Figure 6. Decision tree from the third scenario

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
ISSN 2792-7660
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org

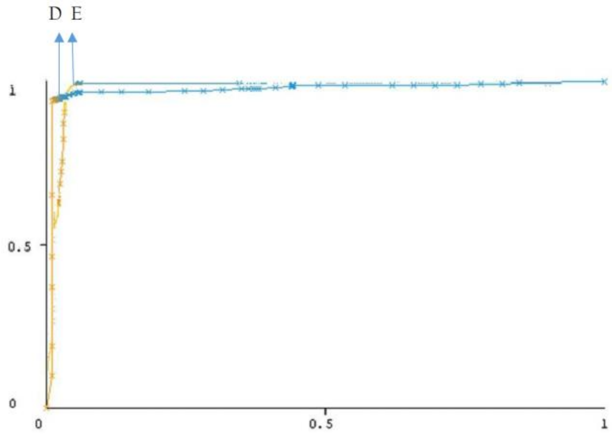


Figure 7. ROC curve for the third scenario.

4. Conclusions

Scholar dropping significantly affects economic and social progress, due to the risk that it implies in both, for the professional development of students and their successful incorporation into the productive sector. This research aims to identify the relationship between academic performance, socio-economic and psychological aspects, with school dropouts.

The main limitations and challenges faced in the research are related to the great variety of responses obtained for each item, their noise and ambiguity, as well as the width of the trees generated. However, was possible to identify three scenarios of study, from which the following concluding remarks are derived:

1. Poor academic performance has been consistently recognized as a major factor in school dropout.
2. Physical and psychological health problems, such as anorexia, stress, vision and hearing problems, nutritional, and maternal health problems, were presented by a large part of the students who dropped out.
3. Aspects such as English proficiency, problems for completing homework, few participations in extracurricular activities or home study, limited interaction with peers and educators and behavioral problems were also identified as determinants of school dropout.

4. Although a lesser extent, basic services such as telephone services, having family vehicles or poor family income are aspects that were also revealed by the model.

As we can see, school dropout does not depend only on academic performance, the social determinants, and psychological aspects to face stress and school or family difficulties must be addressed at same time.

Derived from the above, it is imperative to offer complementary tutoring, provide financial scholarships that allow more time to be spent studying than working, and detect possible physical and psychological health problems from the beginning. These measures are intended to ensure that students can complete their technical education, increasing the number of graduates ready to enter the workforce or pursue higher education.

For that reason, results obtained in the research were disseminated to the psychology and guidance departments of CONALEP schools throughout the Mexico State. The latter, with the intention that students at risk of dropping out could be identified and personalize support provided in a timely manner.

Applying the machine learning techniques described in this research could lead us to discover other causes of dropout depending on the school and the geographical area in which it is located. For example, in the northern states of the country, desertion could occur due to factors other than those in the center of the country due to the proximity to the northern border or due to high levels of insecurity. As future work, is to expanding research by analyzing the school dropouts in all careers offered by CONALEP. In addition, using other machine learning algorithms such as Association Rules, Latent Topics, NaiveBayes, Logistic Functions or another type of tree such as Random Forest or random tree is also considered.

References

Alvarado-Uribe, J., Mejía-Almada, P., Masetto Herrera, A. L., Molontay, R., Hilliger, I., Hegde, V., Montemayor Gallegos, J. E., Ramírez Díaz, R. A., & Ceballos, H. G. (2022). Student dataset from Tecnológico de Monterrey in Mexico to predict dropout in higher education. *Data*, *7*(9), 119. https://doi.org/10.3390/data7090119.

Bernal-Reyes, L. (2020). El CONALEP. desarrollo de una estrategia de formación técnica para el trabajo. *Revista Mexicana de Investigación Educativa*, *25*(84), 121–152.

Bošnjaković, N., & Đurđević Babić, I. (2023). Systematic review on educational data mining in educational gamification. *Technology Knowledge and Learning*. https://doi.org/10.1007/s10758-023-09686-2

De Witte, K., Cabus, S., Thyssen, G., Groot, W., & van den Brink, H. M. (2013). A critical review of the literature on school dropout. *Educational Research Review*, *10*, 13–28. https://doi.org/10.1016/j.edurev.2013.05.002

Fernandez-Haddad, M., & Lara Gonzalez, R. C. (2023). Key factors that influence school dropouts amongst adolescents in marginalized urban areas of Mexico. *Journal of Nonprofit & Public Sector Marketing*, *35*(5), 494–520. https://doi.org/10.1080/10495142.2021.1982111

Instituto Mexicano para la Competitividad IMCO (2022). Bachillerato, el escalón más frágil de la educación, consultado el 30 de julio de 2024 https://imco.org.mx/bachillerato-el-escalon-fragil-de-la-educacion/

Lottering, R., Hans, R., & Lall, M. (2020). A model for the identification of students at risk of dropout at a university of technology. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. DOI: 10.1109/icABCD49160.2020.9183874

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/tkde.2019.2962680

Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., & Islam, A. H. M. S. (2021). Mining educational data to predict students performance: A comparative study of data mining techniques. *Education and Information Technologies*, *26*(5), 6051–6067. https://doi.org/10.1007/s10639-021-10575-3

Nai, R., Meo, R., Morina, G., & Pasteris, P. (2023). Public tenders, complaints, machine learning and recommender systems: a case study in public administration. *Computer Law and Security Report*, *51*(105887), 105887. https://doi.org/10.1016/j.clsr.2023.105887

Nájera, A., & Ortega, L. (2022). Predictive model for taking decision to prevent university dropout. *International Journal of Interactive Multimedia and Artificial Intelligence*, *7*(4), 205–213. https://doi.org/10.9781/ijimai.2022.01.006

Pérez-Gutiérrez, B. R. (2020). Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico. *Revista UIS ingenierías*, *19*(1), 193–204. https://doi.org/10.18273/revuin.v19n1-2020018

Quinn, J. (2013). *Drop-out and Completion in Higher Education in Europe Among Students from Under-represented Groups. An independent report authored for the European Commission*.

Ribeiro, R. C., & Canedo, E. D. (2020). Using data mining techniques to perform school dropout prediction: A case study. In *Advances in Intelligent Systems and Computing* (pp. 211–217). Springer International Publishing. https://doi.org/10.1007/978-3-030-43020-7_28

Rodríguez, P., Villanueva, A., Dombrovskaia, L., & Valenzuela, J. P. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. *Education and Information Technologies*, *28*(8), 1–47. https://doi.org/10.1007/s10639-022-11515-5

Saini, T., & Chhabra, A. (2024). *Performance analysis of different machine learning classifiers for prediction of lung cancer. Artificial Intelligence of Things (págs)*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48774-3_18

Salinas-Chipana, J., Obregon-Palomino, L., Iparraguirre-Villanueva, O., & Cabanillas-Carbonell, M. (2024). Machine learning models for predicting student dropout a review. In *Proceedings of Eighth International Congress on Information and Communication Technology* (pp. 1003–1014). Springer. https://doi.org/10.1007/978-981-99-3043-2_83

Santoso, L. W., & Yulia. (2019). The analysis of student performance using data mining. In *Advances in Intelligent Systems and Computing* (pp. 559–573). Springer Singapore. https://repository.petra.ac.id/18101/1/Publikasi1_03023_4745.pdf

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Urbina-Nájera, A. B., Camino-Hampshire, J. C., & Cruz Barbosa, R. (2020). Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa. *RELIEVE* -

Revista Electrónica de Investigación y Evaluación Educativa, *26*(1). https://doi.org/10.7203/relieve.26.1.16061

Yagcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z