

LITERATURE ANALYSIS ON THE APPLICATION OF MACHINE LEARNING ALGORITHMS FOR THE PREVENTION OF UNIVERSITY DROP OUT

José Guadalupe Landaverde-Vaca^{1,2}, Ana María Pacheco-Martínez²,
& Carlos Fresneda-Portillo³

¹Engineering, Universidad Popular Autónoma del Estado de Puebla (Mexico)

²Quantitative Methods, Universidad Loyola Andalucía (Spain)

³Mathematics and Computing, Universidad de Burgos (Spain)

Abstract

Education is a fundamental human right, established by international agreements such as the Sustainable Development Goal 4, which promotes an inclusive and quality education. In Mexico, upper secondary and tertiary education exhibit the highest student dropout rates, adversely impacting social equity, employment opportunities, economic development, and the sustainability of educational institutions. In response to this challenge, machine learning, artificial intelligence, and deep learning techniques have emerged as relevant tools for the early prediction of student dropout and the development of more effective retention strategies. The aim of this study was to systematically analyze recent literature on predictive models applied to dropout in higher education, identifying the most effective algorithms and the key factors influencing the phenomenon. The methodology followed the PRISMA framework for systematic reviews. Searches were conducted in the Scopus and Web of Science databases covering publications from January 2020 and April 2025. After removing duplicate records, applying JCR indexing criteria, and assessing thematic relevance, 45 high-impact studies were selected. The data were subsequently organized and analyzed using tabular and graphical tools in RStudio. The results indicate that the most effective algorithm is Random Forest, particularly when combined with balancing techniques such as SMOTE Tomek Links, achieving performance metrics above 99%. The most influential factors were found to be current academic performance variables, followed by socioeconomic and admission factors. The analysis confirms the feasibility of implementing predictive models in higher education to reduce student dropout, providing to academic administrators valuable information to target support measures and enhance the efficiency of the educational system. In the specific case of a private institution such as the Universidad Popular Autónoma del Estado de Puebla (UPAEP), implementing the most effective dropout prediction algorithm would have a positive impact by enabling early and personalized interventions for at-risk students, helping them overcome academic, financial, or emotional challenges, and promoting their retention and academic success. Faculty members would have better tools to identify and support at-risk students, strengthening academic support. These actions are expected to significantly reduce dropout rates, optimize institutional resources, and reinforce the university's academic reputation.

Keywords: *University dropout, predictive factors, machine learning, prediction models.*

1. Introduction

Education is a fundamental human right recognized worldwide through Article 26 of the Universal Declaration of Human Rights, which states that everyone has the right to education (Organización de las Naciones Unidas, 2024a). This principle is reinforced by the Sustainable Development Goal 4 (SDG 4), which aims to ensure inclusive and equitable quality education and promote lifelong learning opportunities for everyone (Organización de las Naciones Unidas, 2024b). Within this framework, the United Nations Educational, Scientific and Cultural Organization (UNESCO) plays a central role in promoting the right to education by strengthening education policies, teacher training, infrastructure improvements, and the strategic integration of emerging technologies, including artificial intelligence, with the goal of eliminating structural barriers that limit access to and retention in school (UNESCO, 2024).

In Mexico, the national education system has aligned itself with these international commitments through reforms aimed at improving educational coverage and quality. The 2020–2024 Sectoral Education

Program establishes, among its priority strategies, the consolidation of support and school coexistence programs to reduce school dropout rates and improving graduation rates, strengthening the transition between levels and modalities of the education system (Secretaría de Educación Pública, 2020). However, despite the reported progress, official statistics indicate that dropout rates remain higher at the upper secondary and higher education levels, maintaining a downward trend in recent decades (Instituto Nacional de Estadística y Geografía (INEGI), 2025). This phenomenon has profound implications by contributing to a reduction in employment opportunities, being associated with lower wages, and contributing to higher levels of informality (Bentaouet Kattan & Székely, 2015).

University dropout is a complex and multidimensional phenomenon that does not stem from a single cause, but rather from the interaction of academic, socioeconomic, institutional, and personal factors that influence students' educational trajectories (Barragán Moreno & González Támara, 2024). Abandoning higher education represents a crucial challenge for higher education institutions, due to its negative impact on both students and society (Vaarma & Li, 2024). In this sense, university dropout has been defined as the premature abandonment of an academic program before obtaining the corresponding degree, once a sufficient period has elapsed to rule out the student's readmission, and it can be analyzed from psychological, economic, sociological, organizational, and interactional perspectives (Jiménez-Gutiérrez et al., 2024).

Over the past two decades, the increasing availability of digitized educational data has fostered research aimed at the early prediction of dropout rates using machine learning, artificial intelligence, and deep learning techniques. These approaches have proven useful in both online and face-to-face educational contexts. Recent studies have identified algorithms such as neural networks, boosting models, and ensemble methods as effective tools for predicting dropout rates, as well as the relevance of academic and interaction variables on LMS platforms to explain this phenomenon (Urteaga et al., 2020; Fernandez-Garcia et al., 2021; Rebelo Marcolino et al., 2025; Vaarma & Li, 2024). However, literature reviews warn of fragmented approaches, heterogeneous definitions of dropout, and limitations in the explanatory power of purely automated models, underscoring the need for systematic and integrative analyses (Elibol & Bozkurt, 2023; Parra-Sánchez et al., 2023). In this context, given that student dropout is a core issue for any educational institution and considering that the highest dropout rates in Mexico occur in upper secondary and higher education, this research focuses on analyzing the literature on the application of machine learning, artificial intelligence, and deep learning methods for the early prediction of university dropout. The proposed literature review seeks to answer the following questions: 1) what algorithms have been used for predicting and preventing university dropout, and which are the best based on validation metrics? And 2) what factors significantly determine university dropout?

2. Methodology

The study was based on a systematic literature review focused on analyzing research that uses machine learning, artificial intelligence, and deep learning algorithms to predict student dropout rates in higher education. The review encompassed publications from January 2020 to April 2025 and was structured according to the PRISMA guidelines, following three main stages: article search, study selection and registration, and data extraction and analysis (Page et al., 2021). Publications were identified in the Scopus and Web of Science databases using a search strategy applied to document titles that combined terms related to student dropout and predictive approaches based on artificial intelligence. An initial 261 articles were identified, of which, after removing duplicates, 174 unique publications were obtained.

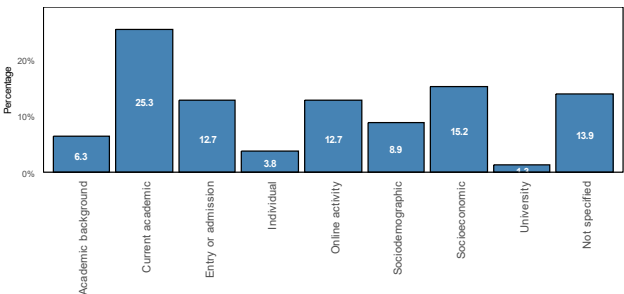
In a second stage, the studies were subjected to more restrictive selection criteria, considering only articles indexed in journals included in the Journal Citation Reports (JCR) and with open access to the full text. Subsequently, an additional filter was applied, prioritizing publications ranked in the 50th percentile or higher of the JCR and focused on higher education student dropout rates. This resulted in the final selection of 45 articles for detailed analysis. Data extraction was carried out by consolidating the records into a unified database, followed by a thorough reading of each study to identify key information related to samples, variables, predictive algorithms, and evaluation metrics. The analysis was supported by tabular and graphical tools implemented in RStudio.

Finally, due to the heterogeneity and large number of variables reported in the analyzed studies, a categorization strategy based on previous proposals in the literature was adopted, complemented by an additional dimension of online activity, in order to distinguish between face-to-face and virtual modalities (de Oliveira et al., 2021). This categorization allowed the variables to be organized into eleven analytical dimensions: academic background, entry or admission, current academic, final academic, online activity, individual, socioeconomic, sociodemographic, external support, environment and university.

3. Results

The study's findings show a clear predominance of supervised learning algorithms, particularly those based on decision trees and ensemble methods. Among these, Random Forest (RF) stands out as the most widely used and best-performing algorithm overall, accounting for 13.3% of individual applications and reaching nearly 19.9% when combined with other algorithms or balancing techniques; other relevant models include Categorical Boosting, Decision Tree, k-Nearest Neighbors, and Logistic Regression. Regarding the factors that significantly determine university dropout rates, the results show a clear consensus on the centrality of current academic variables, which represent 25.3% of the identified influencing factors. These include indicators such as overall GPA, number of credits earned, courses not passed, percentage of courses successfully completed, and attendance. Second are socioeconomic factors, with an incidence of 15.2%, followed by entry or admission variables and those associated with activity on online learning platforms (figure 1).

Figure 1. Important dimensions.



The comparative analysis by educational modality reveals that, although current academic variables maintain a similar relevance in both modalities, there are significant differences in other dimensions. For example, “In person” modality, entry or admission factors and socioeconomic factors carry more weight, while in the online modality, variables associated with activity on LMS platforms are clearly predominant (Figure 2). Likewise, when differentiating by educational sector, it is identified that current academic factors are more relevant in the private sector, while in the public sector, entry factors and online activity variables carry more weight (Figure 3).

Figure 2. Important dimensions by modality: online (blue) face-to-face (red).

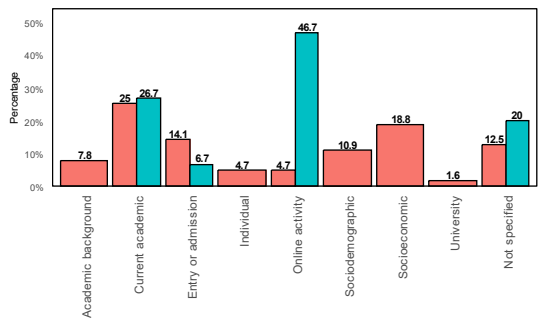
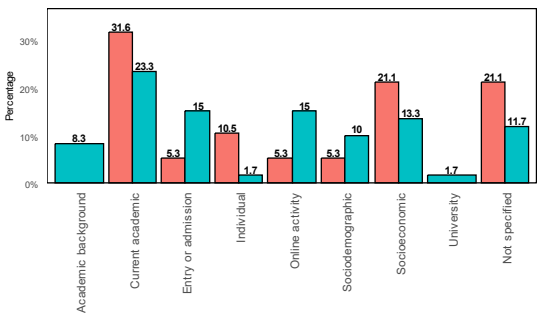


Figure 3. Important dimensions by sector: private (red) and public (blue).



Finally, the supplementary analysis reveals that most studies are conducted in public institutions and in person educational settings. Research with institutional coverage predominates, based primarily on internal data generated by the educational institutions themselves. Furthermore, a trend toward multidisciplinary approaches is observed, as well as a wide geographical diversity in the analyzed studies. In terms of computational tools, Python emerges as the most widely used programming environment, followed by R, reflecting current preferences in the development of predictive models applied to university dropout rates.

4. Discussion

The analysis of the algorithms used reveals a clear predominance of methods based on decision trees and ensemble techniques, particularly Random Forest, which is consistent with previous studies in educational data mining (Andrade-Girón et al., 2023; Dekker et al., 2009). Several studies report that these models achieve high levels of performance, especially when combined with class balancing techniques

such as SMOTE-Tomek Links (Fernández-Delgado et al., 2014; Kumar et al., 2024). However, the discussion cautions that extremely high metrics should be interpreted with care, as they may reflect overfitting, biases derived from preprocessing, or particular characteristics of the datasets, rather than generalizable performance. Furthermore, the observed methodological heterogeneity—stemming from the diverse use of balancing techniques (Chen et al., 2020; Fernandez-Garcia et al., 2021) and explainable artificial intelligence approaches (Delen et al., 2024; Rebelo Marcolino et al., 2025)—hinders direct comparison between studies. This situation is reinforced by the variability in the validation metrics employed, where accuracy continues to predominate despite its limitations in contexts of unbalanced classes (Herzog, 2006). Regarding predictor variables, there is broad consensus on the centrality of current academic indicators, such as grade point average, accumulated credits, and failed courses, as the most consistent predictors of university dropout (Kocsis & Molnár, 2025). This finding is consistent with the classic literature on student dropout (Bean, 1979; Tinto, 1993). However, the recurring exclusion of socioeconomic, personal, and contextual variables limits the explanatory power of the models and reinforces the need to understand university dropout as a socially situated and structural phenomenon (Cabrera et al., 1992; Parra-Sánchez et al., 2023). In this sense, the importance of moving towards more transparent, explainable, and socially responsible approaches to the use of artificial intelligence in education is emphasized (Holstein et al., 2019; Williamson & Eynon, 2020).

5. Conclusions and limitations

The systematic review successfully answered the research questions and achieved the stated objectives by providing a structured synthesis of the methodological, algorithmic, and explanatory approaches used to predict university dropout. The findings reveal a clear predominance of ensemble algorithms, particularly Random Forest, which emerges as the most frequently employed method and the one with the best reported performance in the analyzed studies, both as a standalone classifier (13.3%) and in combination with other models or class-balancing techniques (19.9%). The analysis of factors associated with university dropout, based on eleven dimensions, indicates that current academic characteristics constitute the most influential dimension (25.3%), followed by socioeconomic characteristics (15.2%). However, the predominant use of ensemble models calls for caution in interpreting the high-performance levels reported, as these may reflect overfitting or limited capacity for generalization, while the recurrent exclusion of socioeconomic, personal, and contextual variables constrains the explanatory power of the models and their applicability for designing comprehensive interventions. Moreover, the substantial methodological heterogeneity of the studies, the frequent use of secondary data, and the concentration of evidence in specific contexts limit the transferability of the findings and highlight the need for future comparative, longitudinal, and context-sensitive research.

References

- Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodríguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., Villarreal-Torres, H., & Angeles-Morales, J. (2023). Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review. *EAI Endorsed Transactions on Scalable Information Systems*, 10(5), 1-11. <https://doi.org/10.4108/eetsis.3586>
- Barragán Moreno, S. P., & González Támara, L. (2024). Complexities of student dropout in higher education: a multidimensional analysis. *Frontiers in Education*, 9. <https://doi.org/10.3389/educ.2024.1461650>
- Bean, J. P. (1979). *Dropouts and Turnover: The Synthesis and Test of a Causal Model of Student Attrition* (Issue ED174873). <https://eric.ed.gov/?id=ED174873>
- Bentaouet Kattan, R., & Székely, M. (2015). Patterns, Consequences, and Possible Causes of Dropout in Upper Secondary Education in Mexico. *Education Research International*, 2015, 1–12. <https://doi.org/10.1155/2015/676472>
- Cabrera, A. F., Nora, A., & Castañeda, M. B. (1992). The Role of Finances in the Persistence Process: A Structural Model. *Research in Higher Education*, 33(5), 571–593. <https://doi.org/10.1007/BF00973759>
- Chen, Y., Zheng, Q., Ji, S., Tian, F., Zhu, H., & Liu, M. (2020). Identifying at-risk students based on the phased prediction model. *Knowledge and Information Systems*, 62(3), 987-1003. <https://doi.org/10.1007/s10115-019-01374-x>
- de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing*, 5(4). <https://doi.org/10.3390/bdcc5040064>

- Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *Proceedings of the 2nd International Conference on Educational Data Mining*, 41–50.
- Delen, D., Davazdahemami, B., & Rasouli Dezfooli, E. (2024). Predicting and Mitigating Freshmen Student Attrition: A Local-Explainable Machine Learning Framework. *Information Systems Frontiers*, 26(2), 641–662. <https://doi.org/10.1007/s10796-023-10397-3>
- Elibol, S., & Bozkurt, A. (2023). Student Dropout as a Never-Ending Evergreen Phenomenon of Online Distance Education. *European Journal of Investigation in Health, Psychology and Education*, 13(5), 906–918. <https://doi.org/10.3390/ejihpe13050069>
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15(90), 3133–3181. <http://jmlr.org/papers/v15/delgado14a.html>
- Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, 9, 133076–133090. <https://doi.org/10.1109/ACCESS.2021.3115851>
- Herzog, S. (2006). Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks vis-à-vis Regression. *New Directions for Institutional Research*, 2006(131), 17–33. <https://doi.org/10.1002/ir.185>
- Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a Real-Time Classroom Orchestration Tool to Support Teacher–AI Complementarity. *Journal of Learning Analytics*, 6(2), 27–52.
- Instituto Nacional de Estadística y Geografía (INEGI). (2025). *Tabulados Interactivos - INEGI*. <https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=9171df60-8e9e-4417-932e-9b80593216ee>
- Jiménez-Gutiérrez, A. L., Mota-Hernández, C. I., Mezura-Montes, E., & Alvarado-Corona, R. (2024). Application of the performance of machine learning techniques as support in the prediction of school dropout. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-53576-1>
- Kocsis, Á., & Molnár, G. (2025). Factors influencing academic performance and dropout rates in higher education. *Oxford Review of Education*, 51(3), 414–432.
- Kumar, M., Singh, N., Wadhwa, J., Singh, P., Kumar, G., & Qtaishat, A. (2024). Utilizing Random Forest and XGBoost Data Mining Algorithms for Anticipating Students' Academic Performance. *International Journal of Modern Education and Computer Science (IJMECS)*, 16(2), 29–44.
- Organización de las Naciones Unidas. (2024a). *Declaración Universal de los Derechos Humanos*. <https://www.un.org/es/about-us/universal-declaration-of-human-rights>
- Organización de las Naciones Unidas. (2024b). *Educación - Objetivos de Desarrollo Sostenible*. <https://www.un.org/sustainabledevelopment/es/education/>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372. <https://doi.org/10.1136/bmj.n71>
- Parra-Sánchez, J. S., Pardo, I. D. T., & de Merino, C. Y. M. (2023). Factores explicativos de la deserción universitaria abordados mediante inteligencia artificial [Explanatory Factors of University Dropout Explored Through Artificial Intelligence]. *Revista Electronica de Investigacion Educativa*, 25. <https://doi.org/10.24320/REDIE.2023.25.E18.4455>
- Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., Munoz, R., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: insights from moodle log data. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-93918-1>
- Secretaría de Educación Pública. (2020). *Programa Sectorial de Educación 2020–2024*. https://www.gob.mx/cms/uploads/attachment/file/562380/Programa_Sectorial_de_Educaci_n_2020-2024.pdf
- Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition* (2nd ed.). University of Chicago Press.
- UNESCO. (2024). *Educación | UNESCO*. <https://www.unesco.org/es/education>
- Urteaga, I., Siri, L., & Garófalo, G. (2020). Predicción temprana de deserción mediante aprendizaje automático en cursos profesionales en línea [Early dropout prediction via machine learning in professional online courses]. *RIED-Revista Iberoamericana de Educacion a Distancia*, 23(2), 147–167. <https://doi.org/10.5944/ried.23.2.26356>
- Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, 76. <https://doi.org/10.1016/j.techsoc.2024.102474>
- Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology*, 45(3), 223–235. <https://doi.org/10.1080/17439884.2020.1798995>