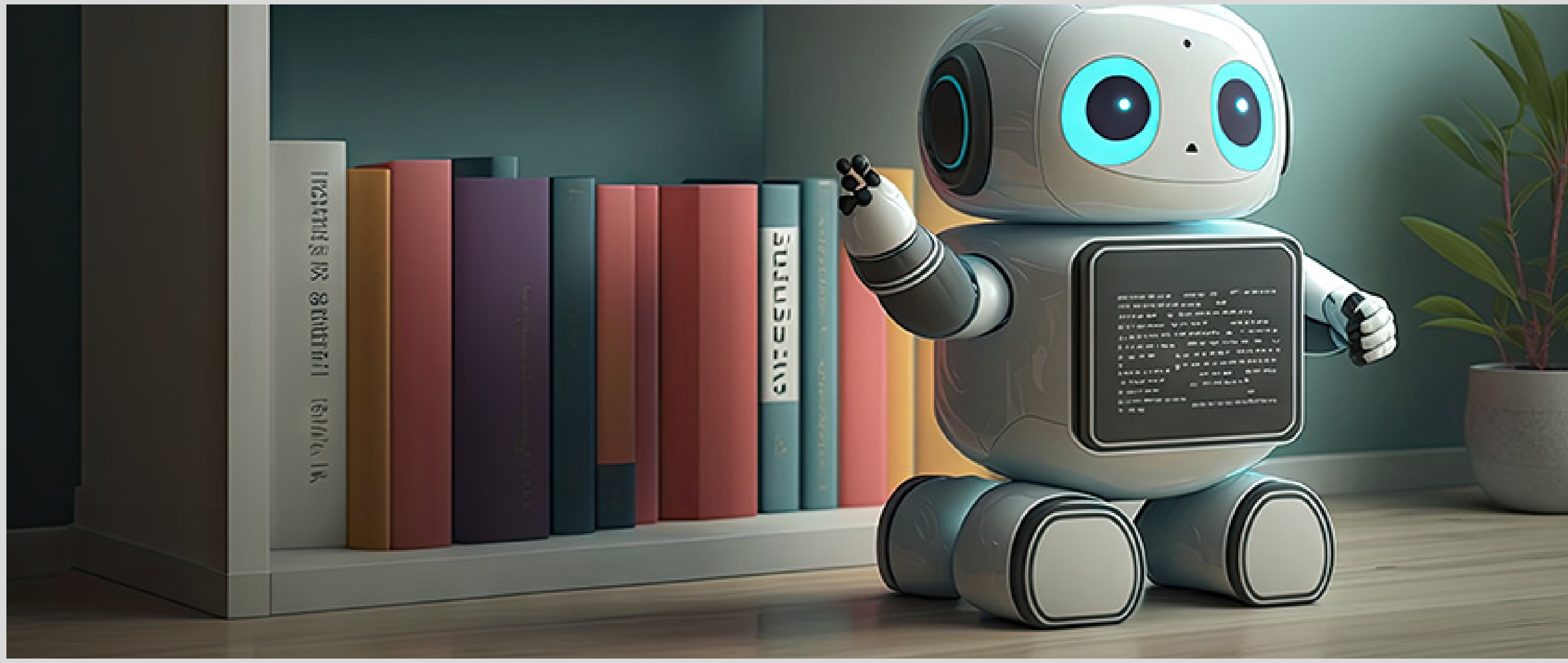
JOURNAL ON EFFICIENCY
AND RESPONSIBILITY IN
EDUCATION AND SCIENCE

The ERIES Journal is being managed by an international editorial board as a regular scientific journal. A rigorous process of papers' reviews (double-blind peer review) is fully supported by a web-based submission system. The journal is published electronically four times a year, on March 31, June 30, September 30 and December 31 of the current year.

The journal is indexed in:

- BASE - Bielefeld Academic Search Engine
- Directory of Open Access Journals
- EBSCO database
- Emerging Sources Citation Index - Web of Science™ Core Collection
- ERIC - Education Resources Information Center
- ERIH PLUS
- Google Scholar
- Open Academic Journals Index
- ResearchBib
- SCOPUS
- The list of reviewed periodicals in the Czech Republic



Special Issue

Predicting Students‘ Learning Outcomes Using Machine Learning

www.eriesjournal.com

Scientific journal of the Czech University of Life Sciences Prague JOURNAL ON EFFICIENCY AND RESPONSIBILITY IN EDUCATION AND SCIENCE, distributed by the Faculty of Economics and Management. Published quarterly. Executive editors: Ing. Martin Flégl, Ph.D., Ing. Jiří Fejfar, Ph.D., PhDr. Michaela Cocca and Armando Cocca, PhD. Editorial Office: ERIES Journal, Czech University of Life Sciences Prague, CZ 165 21 Prague 6 - Suchdol, Czech Republic, email: editor@eriesjournal.com, tel: +420 224 382 355.

volume 19
issue 1

2026

An international peer-reviewed journal published by
Faculty of Economics and Management
Czech University of Life Sciences Prague

editor@eriesjournal.com
www.eriesjournal.com
Online ISSN: 1803-1617
Printed ISSN: 2336-2375

EDITORIAL BOARD

EDITOR-IN-CHIEF

EXECUTIVE EDITORS

EDITORIAL BOARD MEMBERS

TECHNICAL EDITORS

EDITORIAL OFFICE

ERIES Journal, Czech University of Life Sciences Prague (CZU), CZ 165 00 Prague 6 - Suchdol, Czech Republic
email: editor@eriesjournal.com * tel: +420 224 382 355
Registration number: MK ČR E 21414 © Czech University of Life Sciences Prague, Czech Republic, 2026

INSTRUCTIONS FOR AUTHORS

The Journal on Efficiency and Responsibility in Education and Science publishes papers of the following categories: full research papers, short communications, review studies and book reviews (on invitation only).

- FULL RESEARCH PAPERS
- SHORT COMMUNICATION
- REVIEW STUDY

Papers are published in English. A paper may comprise an empirical study using an acceptable research strategy, such as survey, case study, experiment, archival analysis, etc. It may contain a theoretical study aimed at advancing current theory or adapting theory to local conditions or it may arise from theoretical studies aimed at reviewing and/or synthesizing existing theory. Concepts and underlying principles should be emphasized, with enough background information to orient any reader who is not a specialist in the particular subject area.

Submission checklist

The paper. The paper is carefully formatted according to the template of the journal (see bellow). Special attention is paid to the exact application of the Harvard referencing convention to both continuous citations and list of references. If an electronic source has the DOI number assigned, also it will be provided in the list of references. Manuscripts are submitted via the editorial system in the DOC.

Research highlights. The core results, findings or conclusions of the paper are emphasized in 2-4 bullet points (max. 150 characters per bullet point including spaces). The highlights are submitted as a text into the submission form in the editorial system.

Copyright form. The submission of a paper will imply that, if accepted for publication, it will not be published elsewhere in the same form, in any language, without the consent of the Publisher. The manuscript submitted is accompanied by the copyright form signed by the corresponding author who declares the agreement of all authors with the conditions in the Form. The Form is submitted into the editorial system in the PDF format.

Suggested reviewers. It is required to suggest two experts appropriate to evaluation of the paper. The experts should be out of the affiliation of the author(s), Czech University of Life Sciences Prague, and also both experts should be from different affiliations. The reviewers are submitted into the text fields in the submission form of the editorial system.

Preparation of the manuscript (technical notes)

Authors are responsible for applying all requirements that are specified in the journal's paper template in individual sections. Especially, the paper must provide a short review of current state in the area of the paper's aim in Introduction. The paper should refer significant sources, particularly scientific journals or monographs.
Papers must be closely scrutinized for typographical and grammatical errors. If English is not author's first language then the paper should be proof-read by a native English-speaking person, preferably one with experience of writing for academic use. Spelling should follow the Oxford English Dictionary.
Tables, graphs and illustrations should be drawn using a suitable drawing package. Colour may be used. Place all diagrams and tables where you wish them to appear in the paper. Ensure your diagrams fit within the margins and are resizable without distortion.

Review procedure

Following Editorial recommendation, papers are submitted to a double-blind peer review process before publication. Commentary by reviewers will be summarized and sent by email to authors, who can choose to revise their papers in line with these remarks. Re-submitted papers should be accompanied by the description of the changes and other responses to reviewers' comments (see above), so that the desk-editor can easily see where changes have been made.

Copyright

Authors are fully responsible for the paper's originality and for correctness of its subject-matter, language and formal attributes. Author's statement should be enclosed declaring that the paper has not been published anywhere else.

The submission of a paper will imply that, if accepted for publication, it will not be published elsewhere in the same form, in any language, without the consent of the Publisher. Before publication, authors will be asked to complete a copyright release, giving the publisher permission to publish the paper in a specific issue of this Journal. Overall copyright ownership of the paper, however, remains with the author/s. It is the authors' responsibility to obtain written permission to quote material that has appeared in another publication.

EDITORIAL

Education is simultaneously a process, an outcome, and an achievement across all the levels at which it exerts influence. From the moment we are born, human beings begin to learn—constructing experience and making sense of the external world through continuous interaction and adaptation. This process gradually takes form through institutions, learning environments, and organizational structures that provide coherence and progression to knowledge development. At its core, education is a deep social phenomenon, shaped not only by human interaction but also by the technological paradigms that define each era.

Throughout history, technology has consistently expanded the possibilities of teaching and learning. From clay tablets to abacuses, and from chalkboards to digital platforms, technological tools have enhanced the interaction between students, educators, and their environments. Each transformation has increased the capabilities of educational systems, enabling new forms of representation, communication, and understanding.

However, the management of the learning process and the information generated within educational settings has often remained secondary. For decades, education relied on extensive paperwork, fragmented records, and static grading systems—an experience familiar to many readers of this editorial. The advent of computing marked a fundamental shift, allowing information to be encoded, structured, and analyzed through digital systems. Texts, images, and records evolved into datasets, tables, and visualizations that could be updated dynamically, revealing patterns and trends that support more informed decision-making.

Today, we are witnessing yet another significant leap. The exponential growth of data in education, particularly following the global disruptions of the COVID-19 pandemic, has created unprecedented opportunities for analysis and innovation. This special issue and its corresponding call for papers are framed within this new paradigm: the application of statistical and mathematical methods to understand and leverage the vast amounts of data generated daily in educational contexts.

Among the most prominent trends, machine learning (ML) and artificial intelligence (AI) stand out as transformative tools for processing and analyzing large-scale educational data. These technologies enable the identification of patterns in student engagement (Khoudi et al., 2025), the prediction of learning outcomes (De la Hoz et al., 2023), and the enhancement of decision-making processes in education (Nieto et al., 2019). Furthermore, AI-driven tutoring systems, adaptive learning platforms, and automated assessment tools are reshaping traditional educational models by providing real-time feedback and personalized learning experiences.

In this context, this special issue collects seven papers to bring together cutting-edge research advancing our understanding of data-driven education. By integrating rigorous analytical methods with meaningful educational challenges, each author aims to contribute to a more responsive, inclusive, and effective educational ecosystem. By going through each of these beautiful pieces of research, readers will enjoy several approaches that incorporate ML and IA to create bridges, meaningful experiences, and innovative methods.

In the first paper, Anselmus Yata presents a structured overview of the use of machine learning (ML) to predict academic performance, based on a review of 58 empirical studies. The findings show that the most effective models integrate demographic, academic, behavioral, and psychosocial variables. Among the algorithms analyzed, Random Forest and Artificial Neural Networks achieve the highest performance, with accuracies ranging from 85% to 93%, supported by strong precision, recall, F1 score, and AUC metrics. The study highlights the potential of ML models to enhance early warning systems, inform decision-making, and enable personalized learning.

In the second paper, Enrique De La Hoz, Carlos Garcia-Yerena, and Ingrid Torres-Rojas develop an explainable machine learning framework to predict undergraduate performance in Colombia's SABER PRO examination. Using demographic and academic background variables alongside prior standardized test scores, the study formulates a binary classification problem and evaluates multiple models, including XGBoost, SVM, and GLMNET. The results highlight the strong predictive capability of non-linear models, complemented by explainability techniques such as SHAP to identify key drivers of student performance.

In the third paper, Marwan Nawae, Siripa Chankua, and Massaya Longsaman propose a hybrid Explainable AI (XAI) framework to address student dropout prediction while enhancing model transparency. Using AutoGluon, the authors develop high-performing multiclass classification models to predict whether students graduate, drop out, or remain enrolled. To overcome the black-box nature of these models, the study integrates global interpretability through a decision tree surrogate and local explanations via LIME. The results reveal that academic variables are the main drivers of outcomes, while socio-economic factors such as tuition fees also play a significant role.

In the fourth paper, Martin Flegl, Marketa Matulova, and Kristyna Vltavska examine disparities in student learning outcomes across Czech municipalities using machine learning and SHAP analysis. By incorporating demographic, economic, social, and housing variables, the study identifies educational structure as the most influential factor, particularly the proportion of individuals without secondary education and those with college degrees. Additionally, social stressors such as poverty and housing instability introduce nonlinear effects that help explain vulnerable subgroups. The model achieves an R^2 of 0.629, highlighting the combined impact of structural and contextual factors.

In the fifth paper, Andres Acero, Miguel Alejandro Garzón-Parra, and Jesús Isaac Vázquez-Serrano propose a novel two-stage analytical framework to evaluate academic efficiency in high-impact scholarship programs. The study combines Window Data Envelopment Analysis (DEA) to capture temporal efficiency dynamics with a Gaussian Mixture Model to identify latent student profiles. Using academic performance and contextual variables, the approach reveals heterogeneous trajectories of efficiency and uncovers distinct clusters of students with differentiated performance patterns. The results provide a nuanced understanding of how efficiency evolves over time and across groups, offering valuable insights for targeted interventions and resource allocation.

In the sixth paper, Ridwan Setiawan, Edi Noersasongko, Abdul Syukur, Fikri Budiman, and Dede Kurniadi present a systematic literature review on machine learning approaches for predicting student dropout and graduation under class imbalance. Analyzing 70 studies published between 2017 and 2025, the authors identify a predominance of binary classification tasks and highlight the importance of multiclass approaches to better reflect real educational scenarios. The review examines imbalance handling techniques, including resampling, class weighting, and ensemble methods, as well as diverse evaluation and validation strategies. The study emphasizes the need for transparent reporting and imbalance-aware metrics, proposing an integrative taxonomy to improve decision-making.

In the final paper, Enrique De La Hoz, Carlos Garcia-Yerena, and Rohemi Zuluaga-Ortiz analyze the academic productivity of Colombian social science programs between 2020 and 2023 through a PCA–Malmquist Index approach. Using data from 11,099 students, the study first applies Principal Component Analysis to identify performance profiles and then evaluates productivity changes through the Malmquist Index and its components. The findings show a growing proportion of high-performing students over time and reveal that technological change is the main driver of productivity improvements, accounting for most of the observed gains.

Together, the contributions in this special issue mark an important moment in education: data, analytics, and intelligent systems are not only tools for understanding learning but also instruments for transforming it. Beyond prediction and measurement, the true value of these approaches lies in their capacity to support equitable, transparent, and timely interventions. The challenge is not to build more accurate models, but to ensure that their insights translate into meaningful impact—advancing education systems that are more inclusive, adaptive, and responsive to the needs of every learner.

Sincerely

Andrés E. Acero G.

Andres Acero
Guest Editor
Tecnologico de Monterrey, Mexico
Institución Universitaria Politécnico
Grancolombiano, Colombia

REFERENCES

De La Hoz, E., Zuluaga, R. and Mendoza, A. (2021) 'Assessing and Classification of Academic Efficiency in Engineering Teaching Programs', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 14, No. 1, pp. 41–52. https://doi.org/10.7160/eriesj.2021.140104

Nieto, Y., García-Díaz, V., Montenegro, C., González, C. C. and González Crespo, R. (2019). 'Usage of machine learning for strategic decision making at higher educational institutions', *IEEE Access*, Vol. 7, pp. 75007–75017. https://doi.org/10.1109/ACCESS.2019.2919343

Khoudi, Z., Hafidi, N., Nachaoui, M., et al. (2025). 'New approach to enhancing student performance prediction using machine learning techniques and clickstream data in virtual learning environments', *SN Computer Science*, Vol. 6, 139. https://doi.org/10.1007/s42979-024-03622-6

CONTENT

Review study

A SYSTEMATIC APPROACH TO PREDICTING STUDENTS' ACADEMIC PERFORMANCE: A REVIEW OF RECENT LITERATURE

Anselmus Yata Mones✉

Sekolah Tinggi Pastoral St. Petrus
Keuskupan Atambua, Indonesia

✉ anselmojata@gmail.com

ABSTRACT

The rapid expansion of digital learning has generated large volumes of educational data, creating new opportunities to apply machine learning (ML) and data mining techniques to predict student academic performance. This study synthesizes 58 empirical studies that used Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, and Artificial Neural Networks to identify at-risk students and improve educational outcomes.

The review foc uses on predictor variables, validation methods, accuracy rates, and performance metrics. Findings suggest that the most effective predictive models combine four categories of variables: demographic factors, academic indicators, digital behavioral features, and psychosocial attributes. Among the algorithms examined, Random Forest and Artificial Neural Networks demonstrated the strongest predictive performance, achieving accuracy rates of 85%–93% across k-fold cross-validation and train-test split validation.

Performance measures such as precision, recall, F1 score, and AUC further confirm the robustness and generalizability of these models. ML-based academic prediction systems can strengthen early warning systems, support data-driven policymaking, and enable personalized learning interventions. The study concludes that combining multidimensional predictors with explainable AI can improve equity, personalization, operational efficiency, and accountability in educational decision-making.

KEYWORDS

Academic performance prediction, artificial neural networks, data mining, educational data analytics, machine Learning, operational efficiency, predictor variables, random forest, validation methods

HOW TO CITE

Yata Mones A. (2026) 'A Systematic Approach to Predicting Students' Academic Performance: A Review of Recent Literature', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 1–14. http://dx.doi.org/10.7160/eriesj.2026.190101

Article history
Received
June 6, 2025
Received in revised form
November 24, 2025
Accepted
February 6, 2026
Available on-line
March 31, 2026

Highlights

- *Reviews 58 empirical studies using machine learning to predict student academic performance and identify at-risk learners.*
- *Finds that the strongest models combine demographic, academic, digital behavioral, and psychosocial predictors.*
- *Shows Random Forest and Artificial Neural Networks achieved the highest predictive accuracy, typically between 85% and 93%.*
- *Highlights k-fold cross-validation and train-test split as the most common validation methods in the reviewed studies.*

INTRODUCTION

Education plays a fundamental role in human and social development. In the modern education system, the quality of learning and students' academic achievements are two primary indicators of the success of educational institutions. One major challenge educators and educational institutions face is identifying students at risk of academic failure early (Katarya, 2023; Nazir et al., 2023). Some of the challenges identified are difficulties in effectively managing and utilizing data (Baneres et al., 2019), the manual data collection process is time-consuming and prone to inaccuracies (Pek et al., 2022), and the use of personal data in predictive models raises privacy and ethical concerns, requiring strict data protection measures (Schmidt et al., 2025). This challenge encourages the development of data-driven predictive approaches to understand, anticipate, and improve students' academic performance (Wu et al., 2024).

With advances in information technology and data analysis techniques, new methods have emerged for understanding student behavior and performance, including predicting academic performance using machine learning algorithms and data mining techniques (Daza et al., 2022; De-La-Cruz et al., 2022; Roslan and Chen, 2022). This is where research on predicting students' academic performance gains significant urgency and relevance, as it can aid policymakers, teachers, and the students themselves in more targeted academic planning (Nazir et al., 2023).

However, the numerous prediction models and methodological differences in previous research create challenges in consistently unifying the findings. Therefore, a systematic literature review (SLR) is an essential approach for summarizing and synthesizing scientific evidence in a methodological, objective, and transparent manner. This approach ensures that the review is conducted thoroughly on relevant scientific publications, providing results that can serve as a foundation for policy development or further research.

Predicting academic performance is not merely a statistical or computational effort but rather a proactive strategy to identify risk factors and opportunities in education. Research shows that factors such as academic records, learning behavior, attendance, socio-economic conditions, and data from Learning Management Systems (LMS) are strongly correlated with students' academic outcomes (Roslan and Chen, 2022).

The application of algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVMs), and Neural Networks has proven effective for processing data to generate accurate predictions. In fact, recent research shows that ensemble learning methods such as boosting and bagging tend to achieve higher prediction accuracy than single algorithms (Wu et al., 2024).

More than just technology, these predictions impact real educational practices. With the presence of an accurate prediction system, educators can design more personalized learning strategies, improve student retention, and reduce dropout rates (Katarya, 2023)

With the increasing volume of scientific publications, narrative literature reviews are no longer sufficient to objectively summarize research findings. The SLR method is a systematic, standardized, and replicable approach that follows guidelines such as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). SLR allows researchers to filter the literature using strict inclusion and exclusion criteria, evaluate the methodological quality of each study, and synthesize findings thematically and quantitatively.

Several studies have conducted SLR in the context of academic prediction. For example, Daza et al. (2022) in their review found that the Knowledge Discovery in Databases (KDD) method is the most commonly used, with the CGPA attribute as the main prediction indicator, and Naïve Bayes, Neural Network, and Decision Tree as the most frequently applied algorithms (Daza et al., 2022; Karim-Abdallah et al., 2025). Meanwhile, Nawang et al. (2021) noted that academic data and student behavior are the most frequently appearing variables as determinants of academic performance.

Although the literature on predicting academic performance has developed rapidly, there are still several important gaps that need to be addressed:

- Many studies do not detail the procedures for attribute selection and model validation, thereby affecting the replicability and generalizability of results.
- The dominance of higher education contexts, while secondary and primary school contexts are relatively underexplored in depth (Rodrigues et al., 2022).
- There is a lack of studies evaluating the impact of implementing predictive models on real academic outcomes, such as student retention or grade improvement after intervention.
- The lack of commonly used benchmark datasets in research limits the ability to compare model performance across different studies (Balaji et al., 2021).

The identification of this gap emphasizes the need for a systematic literature review that not only presents a list of models and techniques but also assesses the methodological quality, contextual diversity, and practical potential of the research.

Over the past decade, developments in machine learning technology and big data analysis have broadened the range of methodologies for predicting academic success. Initially, the study focused solely on exam results or cumulative grade point averages (GPAs); however, prediction methodologies are currently evolving to become more intricate and multimodal. Recent research indicates that integrating academic, behavioral, demographic, and psychosocial data can substantially enhance the precision of prediction models (Baashar et al., 2021; Valdiviezo-Diaz and Chicaiza, 2024).

Commonly employed strategies in predictive research encompass:

- *Decision Trees and Random Forest*s are known for their superior interpretability and effective performance on educational data. Decision trees are known for their transparency and ease of understanding. They recursively partition the data into subspaces, making the decision-making process clear and interpretable (Rokach, 2016; Yassine and Mohammed, 2024). Random Forest is an ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and robustness (He et al., 2018; Shaik and Srinivasan, 2019). Studies consistently show that Random Forests achieve higher accuracy than Decision Trees. For instance, Random Forest achieved an accuracy of 91.7% compared to 88.2% for Decision Trees in predicting student success (Reddy et al., 2024). Another study found that Random Forests had higher R^2 scores and lower error metrics than Decision Trees (He et al., 2018).
- *Support Vector Machine (SVM)* is appropriate for binary classification tasks, such as pass/fail or promote/not promote. They work by finding the optimal hyperplane that separates data points of different classes with the maximum margin. SVMs are particularly useful in applications requiring high accuracy and robustness, such as structural health monitoring and fault diagnosis (Pisner and Schnyer, 2020; Wandekoken et al., 2011). They can handle large datasets and complex classification problems, although they may require significant computational resources for training (Gholami and Fakhari, 2017).

- *Artificial Neural Networks (ANNs)* exhibit exceptional efficacy in managing large, non-linear datasets. This is because these networks can model complex relationships and patterns in the data. They are widely used in various fields, including fog computing, remote sensing, and healthcare, to enhance data processing, classification, and decision-making (Abdolrasol et al., 2021; Abiodun et al., 2019). ANNs can adapt to noisy data and achieve high classification accuracy, making them ideal for applications that require robust, flexible models (Abdolrasol et al., 2021; Aghbashlo et al., 2015).
- *Naïve Bayes and K-Nearest Neighbors (KNN)* are often used as baseline models for comparison due to their simplicity and ease of implementation. Naïve Bayes is based on the Bayes theorem and assumes independence among features, making it computationally efficient but sometimes less accurate than more complex models (Kumar et al., 2024; Wickramasinghe and Kalutarage, 2021). KNN classifies data points based on the majority class of their nearest neighbors, which can be effective for small datasets but may struggle with large or high-dimensional data (Ortiz-Bejar et al., 2020; Ripan et al., 2021).

The influence of using student behavioral data from e-learning platforms is growing more substantial. Data such as the number of clicks, time spent completing assignments, daily login duration, and frequency of contact with learning materials have all proved to be critical markers of students' ultimate academic achievement (Daza et al., 2022; Namoun and Alshanqiti, 2020).

The ability to accurately predict student performance has strategic value for all stakeholders in education. For educational institutions, predictive models can be used to: 1) Identify low-risk students (Khan and Menezes, 2020; Seo et al., 2024) performance early and providing appropriate interventions; 2) Directing educational resources (teachers, guidance, facilities) more efficiently (Hoyos Osorio and Daza Santacoloma, 2023); 3) Improving student retention and reducing dropout rates (Deleña et al., 2025; Salloum et al., 2024); 4) Monitoring the impact of data-driven academic policies (Matz et al., 2023).

For students, prediction results can serve as a tool for self-reflection, allowing them to independently strategize their learning, manage their time better, and receive personalized guidance tailored to their needs. Research by Chakrapani and Chitradevi (2022) concluded that the use of data mining-based predictive models enables faster and more targeted academic interventions and can promote academic success despite students' diverse backgrounds (Chakrapani and Chitradevi, 2022).

The urgency of this study is growing, given that most schools and universities now gather substantial amounts of data through academic information systems and online learning platforms. Without a systematic framework for analyzing and utilizing this data, the potential for evidence-based early intervention will be lost.

In the global context, education is increasingly driven by the principles of evidence-based policy, in which important decisions on curriculum, teaching methods, resource allocation, and remedial programs are grounded in data and scientific findings (Salmi, 2015). One of the main drivers of this approach is Sustainable Development Goal (SDG) 4, which states, "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (García-Martín and Pérez Fernández, 2025; Reimers, 2024).

Predicting academic performance is an integral part of this effort, as it enables quick and accurate decision-making to prevent learning failures and dropouts. Educational institutions that implement predictive systems based on machine learning, as extensively researched in various scientific publications, can enhance the efficiency of academic and remedial guidance programs, adjust individual learning strategies based on performance predictions, and compile comprehensive student risk profiles (Bilquise et al., 2024; Deleña et al., 2025; Vaarma and Li, 2024). The use of predictive algorithms extends beyond performance classification to include the design of early interventions tailored to students' behavioral and academic profiles (Wu et al., 2024).

Moreover, in the post-pandemic era, many schools and universities have switched to hybrid learning systems, which generate large amounts of digital data—such as interactions with learning materials, forum participation, task completion speed, and online resource usage. All this data, if processed effectively, can become a strategic asset for mapping students' potential and challenges in real time. Considering the complexity of the theme, practical urgency, and the abundance of available literature, this research aims to:

- Conduct a systematic review of recent scientific literature (2015–2024) discussing the prediction of student academic performance.
- Identifying the main methods, algorithms, and attributes used in academic prediction research,
- Evaluating the methodological quality and publication trends in this field,
- Uncovering research gaps and recommending directions for further studies to develop more effective and inclusive prediction models.

The main contribution of this article lies in its effort to create a comprehensive knowledge map of student academic prediction approaches grounded in SLR principles. This map is expected to:

- Serve as a reference for novice researchers in selecting algorithms and predictive variables,
- Serving as a basis for decision-making for education policymakers,
- Providing practical insights for teachers and educational counsellors in implementing early detection systems in schools.

This research also addresses the calls of many previous studies that lament the lack of systematic synthesis in the literature on student performance prediction (Nabil et al., 2022; Waheed et al., 2020), as well as the limitations in developing benchmark predictive models that can be tested across educational contexts (Pelima et al., 2024).

The Systematic Literature Review (SLR) approach was chosen for this study as a robust framework to assess and synthesize scientific contributions to predicting students' academic performance. The author hopes to make scientific contributions that are not only academically relevant but also practically impactful in developing adaptive, predictive, and inclusive educational systems.

METHOD

This research uses the SLR approach as the primary method to systematically identify, evaluate, and synthesize scientific publications on predicting student academic performance (Fundoni et al., 2023; Krüger et al., 2020; Prill et al., 2021). SLR allows this research to be conducted transparently, structurally, and replicably so that the results can be academically accountable and serve as a reference for further research.

Research design

The design of this research aligns with the PRISMA framework, which consists of four primary stages:

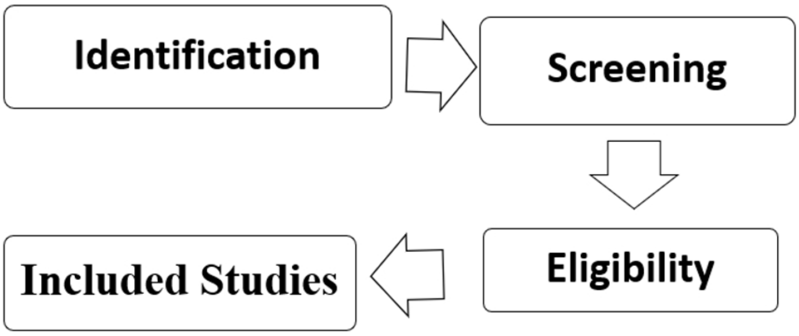


Figure 1: Stages of the PRISMA framework

The process illustrated in the image above follows a series of strict steps, beginning with the formulation of the research question and culminating in the thematic analysis of the study's findings.

Literature search strategy

The search was conducted across several reputable scientific databases, including Scopus, Publish or Perish, Google Scholar, ScienceDirect, and ACM Digital Library. The keywords employed include academic performance prediction, student performance, machine learning, educational data mining, and systematic literature review. Several criteria guide researchers in selecting articles for this study, including those published between 2015 and 2024, focusing on student academic performance through computational methods and data-driven empirical models, and written in English. Conversely, articles that do not meet the criteria include those with unclear writing style, those that are not peer-reviewed, and studies that focus solely on dropout rates without addressing academic performance. Following data extraction, the researchers conducted a thematic analysis based on the following categories: the prediction algorithms used; data types (academic, behavioral, demographic); the main objectives of the study; model validation (*cross-validation*, *hold-out*, *k-fold*); and the implications of the research findings. Additionally, assessing methodological quality is crucial as an indicator that warrants evaluation. These indicators include the clarity of research objectives and problems, data validity and source quality, clarity of analysis methods, comprehensive reporting of results, and relevance to the objectives of this research.

Study selection process

The literature identification and selection process in this study followed the PRISMA guidelines to ensure transparency and replicability of the results.

At the *identification stage*, the researcher collected data from various sources, namely databases (n = 2,000) and registers (n = 700), yielding a total of 2,700 records. Before the screening process, some records were removed due to duplication (n = 230), unsuitability as determined by automated tools (n = 500), and other reasons, such as incorrect format or incomplete data (n = 50). Thus, 1,920 records were advanced to the screening stage.

The *screening stage* involves reviewing titles and abstracts to assess their relevance to the research topic. A total of 1360 recordings were excluded as irrelevant, leaving 560 reports for further review.

Next, in the *eligibility stage*, the researcher attempts to obtain the full text of the remaining reports. However, 473 reports were inaccessible, leaving only 87 that could be successfully reviewed in their entirety. Of that number, 29 reports were excluded, with the following breakdown: 14 reports were not relevant to the research focus, 10 were not in English, and 5 had unclear research methods. Fourteen articles were excluded because they were not relevant to the research focus and did not directly relate to the research variables or parameters. These articles generally discussed topics outside the scope of the study and therefore could not provide the necessary data. This process was carried out to maintain data consistency and ensure that only literature that directly contributed to the primary analysis was included.

The *final stage*, inclusion, yielded 58 studies that met all selection criteria and were deemed suitable for this systematic review.

Overall, this selection process reflects a systematic, transparent, and replicable approach, aiming to ensure that the literature used is truly relevant and possesses adequate methodological quality to support the analysis of predicting students' academic performance.

The PRISMA diagram is displayed as follows:

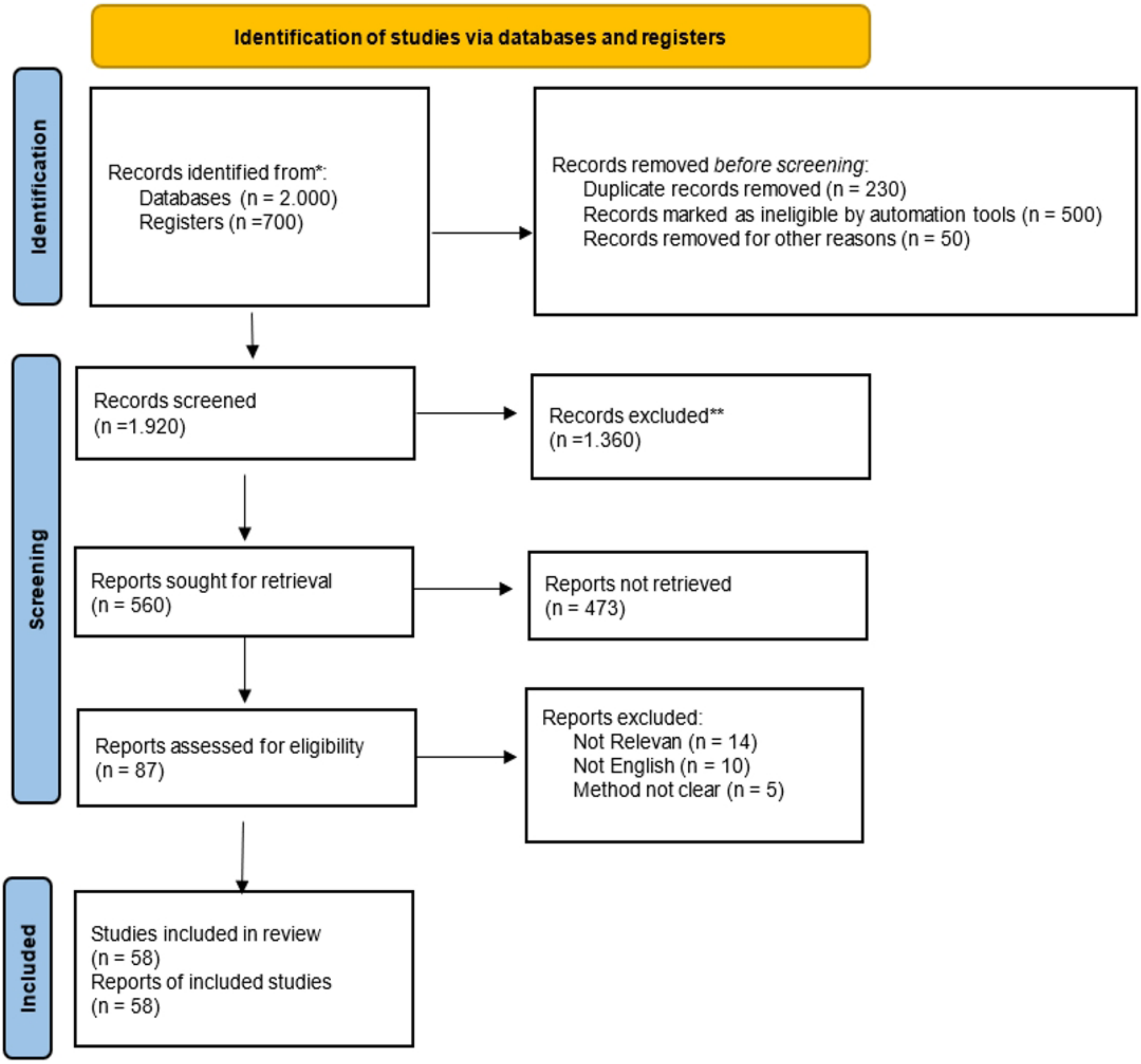

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from*:
Databases (n = 2.000)
Registers (n =700)
Records removed before screening:
Duplicate records removed (n = 230)
Records marked as ineligible by automation tools (n = 500)
Records removed for other reasons (n = 50)
Records screened
(n =1.920)
Records excluded**
(n =1.360)
Reports sought for retrieval
(n = 560)
Reports not retrieved
(n = 473)
Reports assessed for eligibility
(n = 87)
Reports excluded:
Not Relevan (n = 14)
Not English (n = 10)
Method not clear (n = 5)
Studies included in review
(n = 58)
Reports of included studies
(n = 58)

Figure 2: PRISMA flowchart

RESULTS

General trends in research related to predicting student academic performance

Research on predicting student academic performance has advanced rapidly over the past decade, primarily through ML and *Educational Data Mining* (*EDM*). An analysis of 58 selected scientific articles reveals a growing trend in publications on predicting student academic performance, particularly since 2020. The most frequently utilized techniques include *Decision Trees*, *Random Forests*, *Artificial Neural Networks, and Support Vector Machines* (De-La-Cruz et al., 2022; Molina and Cancell, 2021; Nazir et al., 2023; L. Zhang et al., 2021). These methods are favored for their ability to manage complex data and deliver accurate predictions. Additionally, ensemble methods such as Gradient Boosting and Naïve Bayes are often employed to enhance prediction accuracy (Chakrapani and Chitradevi, 2022; Kamal and Ahuja, 2019). This research indicates that machine learning models can effectively predict academic performance by leveraging various factors such as academic records, demographics, and student behavior (Valdiviezo-Diaz and Chicaiza, 2024). This progress is driven by the adoption of machine learning technology and the growing availability of digital educational data. Studies conducted by Hellas et al. (2018) and Roslan and Chen (2022) highlight that research in this area is not only increasing in quantity but also diversifying in the methods and variables utilized (Hellas et al., 2018; Roslan and Chen, 2022).

A comparison of the strengths and weaknesses of each Machine Learning Model can be shown in the following table.

Algorithm	Strengths	Use Cases
Decision Tree	High interpretability, easy to understand, and visualize	Identifying key factors in student performance (Niranjala et al., 2024)
Random Forest	High accuracy, robust to overfitting, handles large datasets, and missing data.	High accuracy, robust to overfitting, handles large datasets and missing data | Predicting student performance, identifying at-risk students (Abdelaziz et al., 2025; Salman et al., 2024)
SVM	High accuracy in binary classification, effective for well-separated classes	Pass/fail prediction, grade movement prediction (Niranjala et al., 2024; Strub, 2020)
ANN	Handles non-linear and extensive datasets, adaptable to noisy data	Requires large datasets, can be computationally intensive (Lutsenko and Zgonnikov, 2024)
Naïve Bayes	Simple, fast, and efficient for small datasets	Assumes feature independence, may be less accurate for complex relationships (Antonakis and Sfakianakis, 2009)
K-Nearest Neighbors (KNN)	Simple, intuitive, effective for small datasets	Computationally expensive for large datasets, sensitive to feature scaling (Ripan et al., 2021)

Table 1: A comparison of the strengths and weaknesses of each Machine Learning Model

Frequently used algorithms

Some commonly used algorithms include Decision Trees, Random Forests, SVMs, Logistic Regression, and ANNs. Decision Trees work by dividing data into branches based on specific variables to produce easily interpretable decisions. Random forests are an extension of Decision Trees that use multiple decision trees to improve prediction accuracy (Akhatkulov et al., 2024; Alhassan et al., 2020; Nazir et al., 2023). Meanwhile, SVM is used to determine the optimal separating hyperplane between data classes, which is suitable for classifying students as “at risk of failing” or “successful.” Logistic regression is often used to predict the probability of an event, such as the likelihood that a student will not graduate. At the same time, ANN mimics the way the human brain works to find complex patterns in educational data.

In practice, the analysis results from these algorithms can be used to provide personalized recommendations for each student. For example, the system can provide early warnings to teachers or parents if a student shows a pattern of declining performance based on their historical data. Additionally, schools can develop adaptive learning strategies that tailor the material and teaching methods to each student’s unique needs. Thus, the application of machine learning and data mining in education not only improves decision-making efficiency but also helps create a more inclusive and data-driven learning environment. Here is a quantitative table from various studies related to the use of machine learning algorithms to predict students’ academic performance:

Algorithm	Usage in Research (%)	Main Advantages	Main Disadvantages	References
Decision Tree	20%	Easy to interpret, suitable for categorical data	Prone to overfitting	(Al-Khlifeh et al., 2025)
Random Forest	25%	High accuracy, reduces overfitting by combining multiple trees	Hard to interpret	(Albreiki et al., 2021)
Support Vector Machine (SVM)	18%	Effective for high-dimensional data and clear class boundaries	Less efficient for large datasets	(Xu et al., 2019)
Logistic Regression	15%	Simple, fast, and easy to interpret	Not suitable for complex non-linear relationships	(Y. Zhang et al., 2021)
Artificial Neural Network (ANN)	22%	Can learn complex and non-linear patterns	Requires large amounts of data and a long training time	(Adejo and Connolly, 2018)

Table 2: Frequently used algorithms

Key algorithms and their applications

ML and data mining techniques have become powerful tools for predicting academic performance and identifying students at risk. This section highlights the key algorithms commonly used in educational data mining: Decision Trees, Random Forest, SVM, Logistic Regression, and ANN, and discusses their main applications supported by empirical studies.

The analysis of these algorithms indicates that ensemble and deep learning models, including Random Forest and ANN, often surpass classical models in prediction accuracy (85–93%). However, Decision Tree and Logistic Regression retain significance due to their interpretability and simplicity. In educational settings, the trade-off between accuracy and explainability is crucial, as transparent models are frequently favored for policy and intervention decisions.

Predictor variables in academic performance prediction

Understanding the characteristics that most strongly predict academic achievement is vital for building accurate, interpretable ML models in education. Across 58 empirical studies, four primary groups of predictor variables appear consistently: demographic, academic, digital behavioral, and psychological characteristics. Each category significantly adds to the model’s explanatory power and provides insights into educational policy and intervention design.

Algorithm	Description	Main Applications	Representative Studies
Decision Tree (CART, ID3, C4.5)	A rule-based model that recursively splits data into hierarchical nodes to make decisions.	Widely applied in early identification of at-risk students, dropout detection, and performance classification due to its interpretability.	*Predicting Academic Performance Using Machine Learning Techniques* (Akhatkulov et al., 2024; Bhimavarapu et al., 2025; Dai and Lu, 2024; Kumar et al., 2022)
Random Forest	An ensemble learning method that combines multiple decision trees to improve accuracy and robustness.	Used in student performance prediction, satisfaction analysis, and anomaly detection in educational datasets.	*Application of Machine Learning in Predicting Student Performance* (Boujmiraz et al., 2026; Cruz and Lumauag, 2024)
Support Vector Machine (SVM)	A supervised learning algorithm that separates classes using an optimal hyperplane with maximum margin.	Applied for classifying student outcomes, analyzing attendance patterns, and predicting graduation success.	*A Review on Predicting Student Academic Performance Using Data Mining Techniques* (Akhatkulov et al., 2024; Masangu et al., 2021; Pelima et al., 2024)
Logistic Regression	A statistical model that estimates the probability of a binary outcome (e.g., pass/fail, dropout/non-dropout).	Commonly used as a baseline model for academic success prediction based on demographic and socioeconomic variables.	*Using Data Mining Techniques to Predict Student Academic Performance* (Akhatkulov et al., 2024; Alhassan et al., 2020; Cruz and Lumauag, 2024)
Artificial Neural Network (ANN)	A computational model inspired by the human brain, consisting of interconnected layers of artificial neurons.	Highly effective in modeling complex nonlinear relationships in educational data, including e-learning behavior analysis and personalized learning recommendations.	*Analysis of Machine Learning Algorithms for Predicting Student Performance* (Chakrapani and Chitradevi, 2022; Daza et al., 2022; Rajendran et al., 2022; Santiketa et al., 2024)

Table 3: Key algorithms and their applications

Demographic predictors

Demographic variables form the foundation of most predictive models, as they describe the social and economic circumstances in which learning occurs. Variables such as gender, age, parental education, household income, and socioeconomic status (SES) are among the most often utilized. Multiple studies using Decision Tree, Random Forest, and Logistic Regression algorithms suggest that demographic characteristics, notably parental education and wealth, are strongly associated with student success (De-La-Cruz et al., 2022; Sarker et al., 2024; L. Zhang et al., 2021).

Students from wealthier socioeconomic backgrounds tend to achieve better academic performance due to access to superior educational resources, technological support, and favorable learning environments (Albreiki et al., 2021; Molina and Cancell, 2021; Nabil et al., 2022). For instance, decision tree models commonly highlight parental education as a significant node in the categorization of high- vs. low-performing kids (Balaji et al., 2021; Nawang et al., 2021). Similarly, Random Forest models boost prediction accuracy by mixing demographic factors with academic indicators such as attendance and past grades (He et al., 2018; Salman et al., 2024).

The implications are clear: incorporating demographic factors into predictive models enables institutions to identify structural inequities and design targeted interventions for disadvantaged groups, thereby promoting fairness and inclusivity in educational analytics.

Academic predictors

Academic factors provide the strongest and most consistent predictors of student achievement across practically all algorithms. These include past GPA, test results, attendance records, assignment submissions, and involvement in learning activities. Studies utilizing Random Forest and ANN models have repeatedly identified these traits as the most significant contributors to model accuracy (Abiodun et al., 2018; Shaik and Srinivasan, 2019; Yassine and Mohammed, 2024).

SVM and Logistic Regression algorithms also perform effectively when applied to academic indicators, especially for binary or ordinal classifications (e.g., pass/fail, dropout/non-dropout) (García-Martín and Pérez Fernández, 2025; Gholami and Fakhari, 2017). Academic predictors directly reflect prior achievement and learning consistency, making them the strongest statistical indicators of future success (Santiketa et al., 2024; Sarker et al., 2024).

Digital behavioral predictors

With the advent of e-learning environments and LMS, digital behavioral data have become more crucial for evaluating student engagement (Y. Zhang et al., 2021). Variables in this area include login frequency, time spent on learning platforms, number of forum conversations, resource downloads, video viewing duration, and quiz attempts (Farissi et al., 2023).

Machine learning models such as ANNs and SVMs excel in this domain because they can capture nonlinear and high-dimensional relationships in behavioral data (Baneres et al., 2019; Nguyen and Jones, 2022)especially in countries like Vietnam, with rich biodiversity and a high population growth rate. One of the main causes of biodiversity loss in Vietnam is the unsustainable bushmeat consumption rate in urban areas. To help mitigate the demand for bushmeat, this study aims to examine the associations between biodiversity loss perceptions, perception toward the prohibition of illegal wildlife consumption, and bushmeat consumption behaviors among urban residents in Vietnam. The investigation employed the Bayesian Mindsponge Framework (BMF). For example, an ANN model achieved over 90% accuracy in predicting final grades using patterns of LMS interaction, highlighting the role of online engagement as a proxy for motivation and effort (Al-Khlifeh et al., 2025; Daza et al., 2022).

These predictors provide educators with real-time insights into student engagement, enabling proactive interventions (e.g., early warnings for disengaged students or adaptive learning pathways). Moreover, integrating digital behavioral variables with academic records increases the accuracy and timeliness of predictive models.

Psychosocial predictors

Psychosocial factors are often overlooked but are gaining increased recognition; they encompass the emotional, motivational, and interpersonal aspects of learning (Karim-Abdallah et al., 2025). Variables such as motivation, self-efficacy, emotional intelligence, learning strategies, stress level, and family support enhance the interpretability of prediction models (Bilquise et al., 2024; Daza et al., 2022).

Decision tree and random forest models often identify motivation and study habits as major contributors to performance differentiation, particularly when combined with academic data (Reddy et al., 2024; Salman et al., 2024; Shaik and Srinivasan, 2019). Meanwhile, ANN models can represent more subtle, nonlinear dependencies between psychosocial factors and academic success, such as the indirect effects of stress and peer collaboration on performance (Aghbashlo et al., 2015; Hellas et al., 2018).

Validation and accuracy methods for predicting student academic performance

The dependability and applicability of ML models for forecasting academic achievement are fundamentally contingent on the validation techniques and accuracy measures used. Strong validation ensures that the predicted outcomes aren't solely the result of overfitting the model but instead reflect patterns that can be applied to new or unknown student data. The 58 papers used a wide range of validation methods and performance indicators. The most common ones were k-fold cross-validation, train-test splits, and hold-out procedures.

1. *Validation Techniques*

a) *K-Fold Cross-Validation*

The *k*-fold cross-validation method is the most widely adopted technique in educational prediction studies, particularly with Decision Tree, Random Forest, and SVM algorithms (Salman et al., 2024; Strub, 2020; Sushma and Sriramakrishnan, 2025).

- In this method, the dataset is divided into *k* equally sized folds (most commonly 5 or 10).
- The model is trained on *k–1* folds and tested on the remaining one, repeating the process *k* times.
- The final performance score is averaged across all folds to minimize bias and variance.

This approach ensures stability and prevents the model from being overly dependent on specific data partitions. Studies by Farissi et al. (2023), Lam et al. (2024), Salman et al. (2024), and Sarker et al. (2024) demonstrated that k-fold cross-validation enhances the consistency of predictive results, especially when datasets are limited in size.

b) *Train–Test Split Method*

The train–test split method remains common in larger datasets, particularly for ANN and Logistic Regression models. Typical ratios such as 70:30 or 80:20 are used to split the data into training and test sets (Bhimavarapu et al., 2025; Lou and Colvin, 2025; Pelima et al., 2024). While simpler than cross-validation, this approach provides a rapid estimation of model performance and is suitable when datasets exceed several thousand records (Albreiki et al., 2021; Cheng et al., 2024; De-La-Cruz et al., 2022). However, its disadvantage is the potential variance introduced by random data division, which can affect model reliability if not repeated or stratified properly.

c) *Hold-Out and Nested Validation*

A smaller proportion of studies utilized hold-out validation (where one subset is used exclusively for final evaluation) or nested cross-validation (for hyperparameter tuning); (Adejo and Connolly, 2018; Alalawi et al., 2023). These methods are particularly relevant in SVM and ANN models, which require extensive parameter optimization (Alhassan et al., 2020; Fundoni et al., 2023). Nested validation prevents data leakage during tuning and ensures a more realistic estimation of generalization performance.

2. *Accuracy Metrics and Evaluation Criteria*

To assess model performance, studies commonly employ a combination of accuracy-based, error-based, and probabilistic metrics. The selection of metrics depends on the model type (classification or regression) and research objective (e.g., predicting pass/fail outcomes vs. continuous GPA values).

a) *Accuracy and Classification Rate*

Accuracy, defined as the ratio of correctly classified instances to the total number of instances, remains the most widely reported metric (Abdelaziz et al., 2025; Daza et al., 2022). Across studies, Random Forest and ANN models consistently achieved the highest accuracy rates, ranging between 85% and 93% (Khan and Ghosh, 2021; Lam et al., 2024). Decision Trees and Logistic Regression models typically achieved accuracy between 75% and 85% (Meghji et al., 2023), while SVM models performed comparably in binary classification tasks (Syed Mustapha, 2023).

b) Precision, Recall, and F1-Score

Other studies evaluate precision (positive predictive value), recall (sensitivity), and F1 score (the harmonic mean of precision and recall) to assess a model's balance between false positives and false negatives (Dai and Lu, 2024; Deleña et al., 2025; Kamal and Ahuja, 2019). These metrics are important in academic prediction tasks where misclassification can lead to inappropriate interventions, for example, mislabeling successful students as "at risk" (Dai and Lu, 2024). ANN and SVM models tend to achieve higher F1 scores due to their ability to model nonlinear relationships and complex data distributions (Khan and Ghosh, 2021; Syed Mustapha, 2023).

c) ROC Curve and AUC

The Receiver Operating Characteristic (ROC) Curve and the Area Under the Curve (AUC) are frequently used to evaluate a model's discrimination ability, especially for binary outcome models (e.g., pass/fail, dropout/non-dropout) (Matz et al., 2023). Higher AUC values (> 0.85) were reported in Random Forest and ANN models, reflecting superior sensitivity-specificity trade-offs (Namoun and Alshanqiti, 2020; Nguyen and Jones, 2022) especially in countries like Vietnam, with rich biodiversity and a high population growth rate. One of the main causes of biodiversity loss in Vietnam is the unsustainable bushmeat consumption rate in urban areas. To help mitigate the demand for bushmeat, this study aims to examine the associations between biodiversity loss perceptions, perception toward the prohibition of illegal wildlife consumption, and bushmeat consumption behaviors among urban residents in Vietnam. The investigation employed the Bayesian Mindsponge Framework (BMF).

d) Error-Based Metrics

For regression-oriented prediction tasks (e.g., predicting continuous GPA), metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are widely employed (Khan and Menezes, 2020; Pek et al., 2022). Lower MAE and RMSE values indicate better model precision. ANN and Random Forest models consistently produce the lowest RMSE values across datasets, confirming their robustness in handling nonlinear and high-dimensional features (Pisner and Schnyer, 2020; Ripan et al., 2021; Salman et al., 2024).

3. Comparative Analysis of Model Accuracy

Algorithm	Common Validation Method	Reported Accuracy	Supporting Studies
Decision Tree (DT)	K-Fold CV (10-fold)	75–85%	(Akhatkulov et al., 2024; Bhimavarapu et al., 2025; Dai and Lu, 2024; Kumar et al., 2022)
Random Forest (RF)	K-Fold CV, Hold-Out	85–93%	(He et al., 2018; Salman et al., 2024)
Support Vector Machine (SVM)	K-Fold CV	80–88%	(Syed Mustapha, 2023)
Logistic Regression (LR)	Train–Test Split (80:20)	70–82%	(De-La-Cruz et al., 2022; Sarker et al., 2024; L. Zhang et al., 2021)
Artificial Neural Network (ANN)	Train–Test Split, Nested CV	88–92%	(Khan and Ghosh, 2021; Lam et al., 2024)

Table 4: Comparative analysis of model accuracy

Validation and accuracy assessment form the backbone of reliable educational data mining. The research data above demonstrates that Random Forest and ANN models, when validated through k-fold cross-validation or a nested approach, achieve the highest and most stable predictive performance. Furthermore, a multidimensional evaluation framework that combines accuracy, precision, recall, and AUC provides a more nuanced and ethical basis for the use of predictive analytics in education. This ensures that algorithms not only perform statistically well but also support fair, actionable, and student-centered outcomes.

DISCUSSION

The results of this study highlight that the effectiveness of ML models in predicting academic performance is not only determined by the complexity of the algorithm but also by the quality of the validation process, the diversity of predictor variables, and the suitability of evaluation metrics (Nazir et al., 2023; Patel and Patel, 2024). This research reveals several important theoretical and practical implications for educational research, institutional decision-making, and data-driven policy development.

One of the most significant benefits of accurate predictive modeling is the early detection of at-risk students, especially if teachers have the necessary personal and professional skills to effectively interpret and act on data-driven insights (Cocca et al., 2025). Models such as Random Forests and ANNs achieve accuracy rates above 90% (Aghbashlo et al., 2015; Khan and Ghosh, 2021; Lam et al., 2024). By combining academic, behavioral, and psychosocial predictors, institutions can intervene promptly and allocate resources to the most vulnerable students (Chakrapani and Chitradevi, 2022; Rodríguez-Hernández et al., 2021). The success of implementing predictive systems depends heavily on developing teachers' digital skills and adopting technology-based teaching methods (Can, Kerkez, and Manav, 2025). This predictive ability transforms machine learning models from solely diagnostic tools into proactive instruments for academic support and retention (Abdrakhmanov et al., 2024), ultimately enhancing graduate employability and the efficiency of higher education systems (Blašková and Staňková, 2023).

The ethical and validated use of accurate machine learning models has significant implications for educational equity and policy design. Integrating demographic and socioeconomic predictors can reveal systemic disparities among students from different backgrounds (Cheng, Liu, and Jia, 2024; Sarker et al., 2024). In this process, the construction of teachers' professional identities plays a crucial role; Giles and Yazan (2023) highlight that teacher collaboration in addressing student diversity is highly influential in shaping how interventions are implemented. Teachers do not simply act as algorithm executors but as moral agents who ensure data is used to uncover patterns of injustice without falling into deterministic predictions (Giles and Yazan, 2023). To avoid reinforcing bias, demographic data should not be used to predict outcomes deterministically but rather to reveal patterns of inequality and inform equitable interventions that respect the diverse backgrounds and learning needs of all students.

Although high-accuracy models like ANNs and Random Forests offer strong predictive performance, their interpretability is often limited compared to simpler models like Decision Trees or Logistic Regression (Hoyos Osorio and Daza Santacoloma, 2023; Kamal and Ahuja, 2019). In an educational environment where transparency and accountability are paramount, this exchange between accuracy and interpretation must be carefully managed (Haneem et al., 2017). Therefore, a hybrid approach combining interpretable models with ensemble methods is recommended. Analyzing feature importance and Shapley Additive exPlanations (SHAP) can help explain model predictions in human-understandable terms. Additionally, teachers must be trained to critically interpret model outputs, ensuring decisions remain pedagogically grounded rather than dictated by algorithms. This training should build upon teachers' existing professional skills while developing new competencies specific to data-driven decision-making in various educational contexts (Cocca et al., 2025).

The results of this study have several important implications for institutional implementation. Machine learning models must be retrained periodically with new data to account for evolving patterns and learning curricula. The model should also be adaptable across various educational contexts, including secondary schools, universities, and online learning platforms, to ensure scalability and maximize its contribution to graduates' employability outcomes at different educational levels (Blašková and Staňková, 2023).

As an implementation step, this predictive system must be embedded in the LMS to provide real-time monitoring and be updated regularly to remain relevant to the dynamic curriculum (Abdrakhmanov et al., 2024; Can, Kerkez, and Manav, 2025). This integration will only be successful if a human-in-the-loop approach is implemented, with educators remaining central to interpretation and action (Haneem et al., 2017). Thus, the synergy between technological sophistication, educators' collaborative identity, and the validation of professional skills will create an educational ecosystem that is not only technically accurate but also efficient, equitable, and responsible toward the future of its graduates.

CONCLUSION

The current study synthesizes evidence from 58 empirical investigations regarding the application of ML and data mining techniques to predict students' academic performance.

Some of the algorithms compared are Decision Tree, Random Forest, SVM, Logistic Regression, and ANN. The results of this analysis reveal a comprehensive understanding of how data-driven models can improve educational decision-making, identify at-risk students, and enhance learning outcomes.

Four main categories of predictor variables were identified as the most influential in modeling academic performance: (1) Demographic factors such as socioeconomic status and parental education, (2) Academic indicators such as GPA, previous grades, and attendance, (3) Digital behavioral variables derived from e-learning interactions, and (4) Psychosocial attributes including motivation, stress, and emotional well-being.

This heterogeneous predictor integration consistently yields superior model performance, with Random Forest and ANN algorithms achieving accuracy exceeding 90% when validated through rigorous cross-validation procedures.

In terms of validation and evaluation, the findings indicate that k-fold cross-validation and train-test split remain the most reliable approaches for ensuring generalization. Comprehensive performance evaluation using accuracy, precision, recall, F1-score, and AUC provides a more profound understanding of model behavior beyond simple correctness. Such methodological rigor prevents overfitting and strengthens confidence in the model's applicability in real-world educational contexts.

There are several practical implications of these findings, including:

- Enabling early detection of at-risk students through continuous monitoring of academic and behavioral data;
- Supporting evidence-based policy design to address structural inequalities; and
- Facilitating personalized learning interventions tailored to individual needs and learning styles.

One consideration is that ethical and pedagogical principles must guide the application of ML in education. Transparency, fairness, and student privacy must remain paramount in all model implementation processes.

The model must also remain understandable to educators, ensuring that predictions enhance, rather than replace, human pedagogical insights.

This synthesis also identifies some research gaps and future directions.

There is an urgent need for:

- Increased inclusion of psychosocial and affective factors in predictive modeling.
- Integration of explainable AI frameworks to make ML models more transparent and trustworthy; and
- Development of adaptive learning analytics systems capable of providing real-time feedback and personalized intervention recommendations.

Finally, machine learning provides a robust and evidence-based framework for transforming educational assessment and student support. When based on strong validation, ethical governance, and human-centered design, predictive models have the potential to make education more equitable, responsive, and effective, ensuring that every student receives the support needed to reach their full academic potential.

ACKNOWLEDGEMENT

No funding was received for this study.

REFERENCES

Abdrakhmanov, R., Zhaxanova, A., Karatayeva, M., Niyazova, G. Z., Berkimbayev, K. and Tuimebayev, A. (2024) 'Development of a Framework for Predicting Students' Academic Performance in STEM Education using Machine Learning Methods', *International Journal of Advanced Computer Science and Applications*, Vol. 15, No. 1, pp. 38–46. https://doi.org/10.14569/IJACSA.2024.0150105

Abdelaziz, A. A., Shaker, H., Tolba, A. S. and Abdelfattah, F. (2025) 'Predictive learning analytics: Analyzing student participation and performance on Moodle', in: Hassanien, A. E., Darwish, A., and El-Askary, H. (eds.), *The International Conference on Advanced Intelligent Systems and Informatics, Studies in Systems, Decision and Control*, Vol. 601, Cham: Springer, pp. 373–384. https://doi.org/10.1007/978-3-031-92240-4_34

Abdolrasol, M. G. M., Hussain, S. M. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S. and Milad, A. (2021) 'Artificial neural networks based optimization techniques: A review', *Electronics*, Vol. 10, No. 21, pp. 26–89. https://doi.org/10.3390/electronics10212689

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. and Arshad, H. (2018) 'State-of-the-art in artificial neural network applications: A survey', *Heliyon*, Vol. 4, No. 11, pp. 09–38. https://doi.org/10.1016/j.heliyon.2018.e00938

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., Arshad, H., Kazaure, A. A., Gana, U. and Kiru, M. U. (2019) 'Comprehensive review of artificial neural network applications to pattern recognition', *IEEE Access*, Vol. 7, pp. 158820–158846. https://doi.org/10.1109/ACCESS.2019.2945545

Adejo, O. W. and Connolly, T. (2018) 'Predicting student academic performance using multi-model heterogeneous ensemble approach', *Journal of Applied Research in Higher Education*', Vol. 10, No. 1, pp. 61–75. https://doi.org/10.1108/JARHE-09-2017-0113

Aghbashlo, M., Hosseinpour, S. and Mujumdar, A. S. (2015) 'Application of artificial neural networks (ANNs) in drying technology: a comprehensive review', *Drying Technology*, Vol. 33, No.12, pp. 1397–1462. https://doi.org/10.1080/07373937.2015.1036288

Akhatkulov, S., Yusupov, O. and Omonov, A. (2024) 'Predicting students' future final exam results using machine learning algorithms: A comparative analysis', *AIP Conference Proceedings*, Vol. 3244, p. 030071. https://doi.org/10.1063/5.0241786

Al-Khlifeh, E., Tarawneh, A. S., Almohammadi, K., Alrashidi, M., Hassanat, R. and Hassanat, A. B. (2025) 'Decision tree-based learning and laboratory data mining: an efficient approach to amebiasis testing', *Parasites and Vectors*, Vol. 18, No.1, pp. 1–18. https://doi.org/10.1186/s13071-024-06618-6

Alalawi, K., Athauda, R. and Chiong, R. (2023) 'Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review', *Engineering Reports*, Vol. 5, No. 12, pp. 1–25. https://doi.org/10.1002/eng2.12699

Albreiki, B., Zaki, N. and Alashwal, H. (2021) 'A systematic literature review of student' performance prediction using machine learning techniques', *Education Sciences*, Vol. 11, No. 9, p. 552. https://doi.org/10.3390/educsci11090552

Alhassan, A., Zafar, B. and Mueen, A. (2020) 'Predict students' academic performance based on their assessment grades and online activity data', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 4, pp. 185–194. https://doi.org/10.14569/IJACSA.2020.0110425

Antonakis, A. C. and Sfakianakis, M. E. (2009) 'Assessing naïve Bayes as a method for screening credit applicants', *Journal of Applied Statistics*, Vol. 36, No. 5, pp. 537–545. https://doi.org/10.1080/02664760802554263

Baashar, Y., Alkawsi, G., Ali, N., Alhussian, H. and Bahbouh, H. (2021) 'Predicting student's performance using machine learning methods: A systematic literature review', in: *2021 International Conference on Computer and Information Sciences (ICCOINS)*, pp. 357–362. https://doi.org/10.1109/ICCOINS49721.2021.9497185

Balaji, P., Alelyani, S., Qahmash, A. and Mohana, M. (2021) 'Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review', *Applied Sciences*, Vol. 11, No. 21, p. 10007. https://doi.org/10.3390/app112110007

Baneres, D., Rodríguez-Gonzalez, M. E. and Serra, M. (2019) 'An Early Feedback Prediction System for Learners At-Risk within a First-Year Higher Education Course', *IEEE Transactions on Learning Technologies*, Vol.12, No.2, pp. 249–263. https://doi.org/10.1109/TLT.2019.2912167

Bhimavarapu, N., Prasanthi, B. V., Lakshmi Veenadhari, C. H., Durga Satish, M., Matta, V. D. R. and Pradeep, I. K. (2025) 'Predicting student academic performance using machine learning: A comparison of classification algorithms', in: Smys, S., Tavares, J. M. R. S. and Balas, V. E. (eds.), *Springer Proceedings in Mathematics and Statistics*, Vol. 441, Singapore: Springer, pp. 703–716. https://doi.org/10.1007/978-3-031-51338-1_56

Bilquise, G., Ibrahim, S. and Salhieh, S. M. (2024) 'Investigating student acceptance of an academic advising chatbot in higher education institutions', *Education and Information Technologies*, Vol. 29, No. 5, pp. 6357–6382. https://doi.org/10.1007/s10639-023-12076-x

Blašková, V. and Staňková, M. (2023) 'Graduate Employability as a Key to the Efficiency of Tertiary Education', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 4, pp. 262–274. https://doi.org/10.7160/eriesj.2023.160401

Boujmiraz, S., Darhmaoui, H. and Drissi el Maliani, A. (2026) 'Predicting student performance: A comprehensive review of machine learning, deep learning, and explainable AI approaches', *Computers and Education: Artificial Intelligence*, Vol. 10, No. 1, p. 100548. https://doi.org/10.1016/j.caeai.2026.100548

Can, S., Kerkez, F. İlker and Manav, G. . (2025) 'Physical Education and Sports Teachers' Perceptions to Benefit from Web 2.0 Tools in Face-to-face Education after Emergency Remote Teaching Process: A Mixed Method Research', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 18, No. 1, pp. 1–12. https://doi.org/10.7160/eriesj.2025.180101

Chakrapani, P. and Chitradevi, D. (2022) 'Academic performance prediction using machine learning: A comprehensive and systematic review', in: *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, pp. 335–340. https://doi.org/10.1109/ICESIC53714.2022.9783512

Cheng, B., Liu, Y. and Jia, Y. (2024) 'Evaluation of students' performance during the academic period using the XG-Boost Classifier-Enhanced AEO hybrid model', *Expert Systems with Applications*, Vol. 238, p. 122136. https://doi.org/10.1016/j.eswa.2023.122136

Cocca, A., Ciesralová, M., Cocca, M., Greier, K., Uchytil, J. and Ruedl, G. (2025) 'Validation of the Teachers' Personal and Professional Skills Questionnaire in the Czech Physical Education Setting', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 18, No. 1, pp. 58–63. https://doi.org/10.7160/eriesj.2025.180107

Cruz, M. M. P. and Lumauag, R. G. (2024) 'Comparative analysis of machine learning algorithms for predicting student academic performance in higher education', in: *Proceedings of the 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS 2024)*, pp. 888–896. https://doi.org/10.1109/ICUIS64676.2024.10866086

Dai, Z. and Lu, P. J. (2024) 'The application of machine learning in student performance prediction in higher education institutions: A systematic literature review', in: *Proceedings of the 2024 13th International Conference on Computer Technologies and Development (TechDev)*, pp. 57–62. https://doi.org/10.1109/TechDev64369.2024.00019

Daza, A., Guerra, C., Cervera, N. and Burgos, E. (2022) 'Predicting Academic Performance through Data Mining: A Systematic Literature', *TEM Journal*, Vol. 11, No. 2, pp. 939–949. https://doi.org/10.18421/TEM112-57

De-La-Cruz, P., Rojas-Coaquira, R., Vega-Huerta, H., Pérez-Quintanilla, J. and Lagos-Barzola, M. (2022) 'A Systematic Review Regarding the Prediction of Academic Performance', *Journal of Computer Science*, Vol. 18, No.12, pp. 1219–1231. https://doi.org/10.3844/JCSSP.2022.1219.1231

Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T. and Gulam, S. B. (2025) 'Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors', *Systems and Soft Computing*, Vol. 7, p. 200352. https://doi.org/10.1016/j.sasc.2025.200352

Farissi, A., Dahlan, H. M. and Shah, Z. A. (2023) 'High accuracy feature selection using metaheuristic algorithm for classification of student academic performance prediction', in: Abraham, A., Gandhi, N., Hanne, T. and Hong, T. P. (eds.), *Proceedings of International Conference on Intelligent Systems Design and Applications, Lecture Notes on Data Engineering and Communications Technologies*, Vol. 179, Cham: Springer, pp. 399–409. https://doi.org/10.1007/978-3-031-36258-3_35

Fundoni, M., Porcu, L. and Melis, G. (2023) 'Systematic literature review: Main procedures and guidelines for interpreting the results', in: Bell, E., Bryman, A. and Harley, B. (eds.), *Researching and Analysing Business: Research Methods in Practice*, Abingdon: Routledge, pp. 55–74. https://doi.org/10.4324/9781003107774-5

García-Martín, J. and Pérez Fernández, L. M. (2025) 'A Review of Global Strategies for Achieving Sustainable Development Goal 4 in Higher Education (2020–2024): Key Actions in the Education for Sustainable Development Framework', *Sustainable Development*, Vol. 14, No. S2, pp. 843–856. https://doi.org/10.1002/sd.70374

Gholami, R. and Fakhari, N. (2017) 'Support vector machine: Principles, parameters, and applications', in: Saha, P., Saha, S. and Balas, V. E. (eds.), *Handbook of Neural Computation*, London: Academic Press, pp. 515–535. https://doi.org/10.1016/B978-0-12-811318-9.00027-2

Giles, A. and Yazan, B. (2023) 'Constructing teacher identity in teacher collaboration: What does it mean to be a teacher of culturally and linguistically diverse English learners?', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 1, pp. 36–45. https://doi.org/10.7160/eriesj.2023.160104

Haneem, F., Kama, N., Ali, R. and Selamat, A. (2017) 'Applying data analytics approach in systematic literature review: Master data management case study', *Frontiers in Artificial Intelligence and Applications*, Vol. 297, pp. 705–715. https://doi.org/10.3233/978-1-61499-800-6-705

Lingjun, H., Levine, R. A., Fan, J., Beemer, J. and Stronach, J. (2018) 'Random forest as a predictive analytics alternative to regression in institutional research', *Practical Assessment, Research and Evaluation*, Vol. 23, no. 1, pp. 1–16. https://doi.org/10.7275/1wpr-m024

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C. and Liao, S. N. (2018) 'Predicting academic performance: A systematic literature review', in: P*roceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018)*, pp. 175–199. https://doi.org/10.1145/3293881.3295783

Hoyos Osorio, J. K. and Daza Santacoloma, G. (2023) 'Predictive model to identify college students with high dropout rates', *Revista Electrónica de Investigación Educativa*, Vol. 25, No. e13, pp. 1–10 . https://doi.org/10.24320/redie.2023.25.e13.5398

Kamal, P. and Ahuja, S. (2019) 'An ensemble-based model for prediction of academic performance of students in undergrad professional course', *Journal of Engineering, Design and Technology*, Vol.17, No. 4, pp. 769–781. https://doi.org/10.1108/JEDT-11-2018-0204

Karim-Abdallah, B., Ayitey Junior, M., Appiahene, P., Harris, E. and Binful, D. K. (2025) 'Application of Machine Learning Algorithms in Predicting Academic Performance of Students in Higher Education Institutes (HEIS): A Systematic Review and Bibliographic Analysis', *African Journal of Applied Research*, Vol. 11, No. 1, pp. 536–559. https://doi.org/10.26437/ajar.v11i1.869

Katarya, R. (2023) 'A Systematic Review on Predicting the Performance of Students in Higher Education in Offline Mode Using Machine Learning Techniques', *Wireless Personal Communications*, Vol. 133, No. 3, pp. 1643–1674. https://doi.org/10.1007/s11277-023-10838-x

Khan, A. and Ghosh, S. K. (2021) 'Student performance analysis and prediction in classroom learning: A review of educational data mining studies', *Education and Information Technologies*, Vol. 26, No. 1, pp. 205–240. https://doi.org/10.1007/s10639-020-10230-3

Khan, S. and Menezes, J. (2020) 'Predictive modelling to illustrate factors influencing students at risk', *International Journal of Technology Transfer and Commercialisation*, Vol. 17, No. 1, pp. 68–75. https://doi.org/10.1504/IJTTC.2020.106574

Krüger, J., Lausberger, C., von Nostitz-Wallwitz, I., Saake, G. and Leich, T. (2020) 'Search. Review. Repeat? An empirical study of threats to replicating SLR searches', *Empirical Software Engineering*, Vol. 25, pp. 627–677. https://doi.org/10.1007/s10664-019-09763-0

Kumar, M., Singh, A. J., Sharma, B. and Cengiz, K. (2022) 'Evaluation of machine learning algorithms on academic big dataset by using feature selection techniques', in: Balas, V. E., Solanki, V. K. and Kumar, R. (eds.), *Intelligent Network Design Driven by Big Data Analytics, IoT, AI and Cloud Computing,* London: Institution of Engineering and Technology, pp. 61–92. https://doi.org/10.1049/PBPC054E_ch4

Kumar, R., Goswami, B., Mhatre, S. M. and Agrawal, S. (2024) 'Naive bayes in focus: a thorough examination of its algorithmic foundations and use cases', *International Journal of Innovative Science and Research Technology*, Vol. 9, No. 5, pp. 2078–2081. https://doi.org/10.38124/ijisrt/IJISRT24MAY1438

Lam, P. X., Mai, P. Q. H., Nguyen, Q. H., Pham, T., Nguyen, T. H. H. and Nguyen, T. H. (2024) 'Enhancing educational evaluation through predictive student assessment modeling', *Computers and Education: Artificial Intelligence*, Vol. 6, p. 100244. https://doi.org/10.1016/j.caeai.2024.100244

Lou, Y. and Colvin, K. F. (2025) 'Performance prediction using educational data mining techniques: a comparative study', *Discover Education*, Vol. 4, No. 112, pp. 1–14. https://doi.org/10.1007/s44217-025-00502-w

Lutsenko, V. and Zgonnikov, M. (2024) 'Fault tolerant system for data storage, transmission and processing in fog computing using artificial neural networks', in: Kotenko, I., Badica, C. and Taratukhin, V. (eds.), *Proceedings of International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes in Networks and Systems,* Vol. 744, Cham: Springer, pp. 199–212. https://doi.org/10.1007/978-3-031-64010-0_19

Masangu, L., Jadhav, A. and Ajoodha, R. (2021) 'Predicting student academic performance using data mining techniques', *Advances in Science, Technology and Engineering Systems*, Vol. 6, No.1, pp.153–163. https://doi.org/10.25046/aj060117

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A. and Stachl, C. (2023) 'Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics', *Scientific Reports*, Vol. 13, No. 1, pp. 5705. https://doi.org/10.1038/s41598-023-32484-w

Meghji, A. F., Shaikh, F. B., Wadho, S. A., Bhatti, S. and Ayyasamy, R. K. (2023) 'Using educational data mining to predict student academic performance', *VFAST Transactions on Software Engineering*, Vol. 11, No. 2, pp. 43–49. https://doi.org/10.21015/vtse.v11i2.1475

Molina, O. E., and Cancell, D. R. F. (2021) 'Is it possible to predict academic performance? An analysis from educational technology', *Revista Fuentes*, Vol. 3, No. 23, pp. 363–375. https://doi.org/10.12795/REVISTAFUENTES.2021.14278

Nabil, A., Seyam, M. and Elfetouh, A. A. (2022) 'Predicting students' academic performance using machine learning techniques: a literature review', *International Journal of Business Intelligence and Data Mining*, Vol. 20, No. 4, pp. 456–479. https://doi.org/10.1504/IJBIDM.2022.123214

Namoun, A. and Alshanqiti, A. (2020) 'Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review', *Applied Sciences*, Vol. 11, No. 1, p. 237. https://doi.org/10.3390/app11010237

Nawang, H., Makhtar, M. and Hamza, W. M. A. F. W. (2021) 'A systematic literature review on student performance predictions', *International Journal of Advanced Technology and Engineering Exploration*, Vol. 8, No. 84, pp. 1441–1453. https://doi.org/10.19101/ijatee.2021.874521

Nazir, M., Noraziah, A., Rahmah, M. and Sharma, A. (2023) 'Examining the potential of machine learning for predicting academic achievement: A systematic review', *Fusion: Practice and Applications*, Vol. 13, No. 2, pp.71–90. https://doi.org/10.54216/FPA.130207

Nguyen, M. H. and Jones, T. E. (2022) 'Predictors of support for biodiversity loss countermeasure and bushmeat consumption among Vietnamese urban residents', *Conservation Science and Practice*, Vol. 4, No. 12, p. e12822. https://doi.org/10.1111/csp2.12822

Niranjala, S. H., Alobaedy, M. M. and Goyal, S. B. (2024) 'A comparative study of machine learning techniques for predicting student academic performance', in: Kotenko, I., Badica, C. and Taratukhin, V. (eds.), *Proceedings of International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes in Networks and Systems,* Vol. 811, Cham: Springer, pp. 307–315. https://doi.org/10.1007/978-3-031-73318-5_31

Ortiz-Bejar, J., Tellez, E. S., Graff, M., Moctezuma, D. and Miranda-Jimenez, S. (2020) 'Improving k nearest neighbors and naïve Bayes classifiers through space transformations and model selection', *IEEE Access,* Vol. 8, pp. 221669–221688. https://doi.org/10.1109/ACCESS.2020.3042453

Patel, H. I. and Patel, D. (2024) 'Exploratory Data Analysis and Feature Selection for Predictive Modeling of Student Academic Performance Using a Proposed Dataset', *International Journal of Engineering Trends and Technology*, Vol. 72, No. 11, pp. 131–143. https://doi.org/10.14445/22315381/IJETT-V72I11P116

Pek, R. Z., Özyer, S. T., Elhage, T., Özyer, T. and Alhajj, R. (2022) 'The role of machine learning in identifying students at-risk and minimizing failure', *IEEE Access*, Vol. 11, pp. 1224–1243. https://doi.org/10.1109/ACCESS.2022.3232984

Pelima, L., Sukmana, Y. and Rosmansyah, Y. (2024) 'Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review', *IEEE Access*, Vol. 12, pp. 23451–23465. https://doi.org/10.1109/ACCESS.2024.3361479

Pisner, D. A. and Schnyer, D. M. (2020) 'Support vector machine', in: Mechelli, A. and Vieira, S. (eds.), *Machine Learning*, London: Academic Press, pp. 101–121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Prill, R., Karlsson, J., Ayeni, O. R. and Becker, R. (2021) 'Author guidelines for conducting systematic reviews and meta-analyses', *Knee Surgery, Sports Traumatology, Arthroscopy*, Vol.29, No. 9, pp. 2739–2744. https://doi.org/10.1007/s00167-021-06631-7

Rajendran, S., Chamundeswari, S. and Sinha, A. A. (2022) 'Predicting the academic performance of middle- and high-school students using machine learning algorithms', *Social Sciences and Humanities Open*, Vol. 6, No. 1, p. 100357. https://doi.org/10.1016/j.ssaho.2022.100357

Reddy, A. L., Sathish, T. and Sangeetha, N. (2024) 'Prediction of student results using novel random forest in comparison with decision tree to improve accuracy', *AIP Conference Proceedings*, Vol. 2853, p. 020053. https://doi.org/10.1063/5.0198498

Reimers, F. M. (2024) 'The sustainable development goals and education, achievements and opportunities', *International Journal of Educational Development*, Vol. 104, p. 102965. https://doi.org/10.1016/j.ijedudev.2023.102965

Ripan, R. C., Sarker, I. H., Hasan Furhad, M., Musfique Anwar, M. and Hoque, M. M. (2021) 'An effective heart disease prediction model based on machine learning techniques', in: Hassanien, A. E., Bhatnagar, R., Darwish, A. and Hameed, K. (eds.), *Proceedings of International Conference on Advanced Intelligent Systems and Informatics, Advances in Intelligent Systems and Computing*, Vol. 1375, Cham: Springer, pp. 280–288. https://doi.org/10.1007/978-3-030-73050-5_28

Rodrigues, L. S., Santos, M., De Araújo Costa, I. P. and Moreira, M. (2022) 'Student Performance Prediction on Primary and Secondary Schools-A Systematic Literature Review', *Procedia Computer Science*, Vol. 214, pp. 680–687. https://doi.org/10.1016/j.procs.2022.11.229

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E. and Cascallar, E. (2021) 'Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation', *Computers and Education: Artificial Intelligence*, Vol. 2, p. 100018. https://doi.org/10.1016/j.caeai.2021.100018

Rokach, L. (2016) 'Decision forest: Twenty years of research', *Information Fusion*, Vol. 27, No. 1, pp. 111–125. https://doi.org/10.1016/j.inffus.2015.06.005

Roslan, M. H. B. and Chen, C. J. (2022) 'Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021)', *International Journal of Emerging Technologies in Learning*, Vol. 17, No. 5, pp. 147–179. https://doi.org/10.3991/ijet.v17i05.27685

Salloum, S. A., Basiouni, A., Alfaisal, R., Salloum, A. and Shaalan, K. (2024) 'Predicting student retention in higher education using machine learning', in: Hassanien, A. E., Darwish, A. and El-Askary, H. (eds.), *Proceedings of International Conference on Advanced Intelligent Systems and Informatics, Communications in Computer and Information Science*, Vol. 2162, Cham: Springer, pp. 197–206. https://doi.org/10.1007/978-3-031-65996-6_17

Salman, H. A., Kalakech, A. and Steiti, A. (2024) 'Random Forest Algorithm Overview', *Babylonian Journal of Machine Learning*, Vol. 2024, pp. 69–79. https://doi.org/10.58496/BJML/2024/007

Salmi, J. (2015) 'Evidence-based policies in higher education: Data analytics, impact assessment and reporting', in: Curaj, A., Matei, L., Pricopie, R., Salmi, J. and Scott, P. (eds.), *The European Higher Education Area: Between Critical Reflections and Future Policies*, Cham: Springer International Publishing, pp. 807–813. https://doi.org/10.1007/978-3-319-20877-0_49

Santiketa, N., Chaikhan, S., Ninrutsirikun, U. and Wattanakitrungroj, N. (2024) 'Student academic performance prediction using machine learning with various features and scenarios', in: *International Computer Science and Engineering Conference (ICSEC 2024)*, pp. 1–6. https://doi.org/10.1109/ICSEC62781.2024.10770729

Sarker, S., Paul, M. K., Thasin, S. T. H. and Hasan, M. A. M. (2024) 'Analyzing students' academic performance using educational data mining', *Computers and Education: Artificial Intelligence*, Vol. 7, p. 100263. https://doi.org/10.1016/j.caeai.2024.100263

Schmidt, A., Cechinel, C., Queiroga, E. M., Primo, T., Ramos, V., Bordin, A. S., Mello, R. F. and Munoz, R. (2025) 'Analyzing Intervention Strategies Employed in Response to Automated Academic-Risk Identification: A Systematic Review', *IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje*, Vol. 20, pp. 77–85. https://doi.org/10.1109/RITA.2025.3540161

Seo, E.-Y., Yang, J., Lee, J.-E. and So, G. (2024) 'Predictive modelling of student dropout risk: Practical insights from a South Korean distance university', *Heliyon*, Vol. 10, No. 11, pp. 1–17. https://doi.org/10.1016/j.heliyon.2024.e30960

Shaik, A. B. and Srinivasan, S. (2019) 'A brief survey on random forest ensembles in classification model', in: Gunjan, V. K., Zurada, J. M. and Raman, B. (eds.), *Proceedings of International Conference on Recent Trends in Machine Learning, Lecture Notes in Networks and Systems*, Vol. 56, Singapore: Springer, pp. 253–260. https://doi.org/10.1007/978-981-13-2354-6_27

Strub, O. (2020) 'Optimal feature selection for support vector machine classifiers', in: *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM 2020)*, pp. 304–308. https://doi.org/10.1109/IEEM45057.2020.9309859

Sushma, P. G. and Sriramakrishnan, G. V. (2025) 'Exploring predictive algorithms: Linear regression and decision tree in student performance', *AIP Conference Proceedings*, Vol. 3270, p. 020154. https://doi.org/10.1063/5.0264437

Syed Mustapha, S. (2023) 'Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods', *Applied System Innovation*, Vol. 6, No. 5, p. 86. https://doi.org/10.3390/asi6050086

Vaarma, M. and Li, H. (2024) 'Predicting student dropouts with machine learning: An empirical study in Finnish higher education', *Technology in Society*, Vol. 76, p. 102474. https://doi.org/10.1016/j.techsoc.2024.102474

Valdiviezo-Diaz, P. and Chicaiza, J. (2024) 'Prediction of academic outcomes using machine learning techniques: A survey of findings on higher education', in: Hassanien, A. E., Darwish, A. and El-Askary, H. (eds.), *Proceedings of International Conference on Advanced Intelligent Systems and Informatics, Communications in Computer and Information Science*, Vol. 2049, Cham: Springer, pp. 206–218. https://doi.org/10.1007/978-3-031-58956-0_16

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J. and Nawaz, R. (2020) 'Predicting academic performance of students from VLE big data using deep learning models', *Computers in Human Behavior*, Vol. 104, p. 106189. https://doi.org/10.1016/j.chb.2019.106189

Wandekoken, E. D., Varejão, F. M., Batista, R. and Rauber, T. W. (2011) 'Support vector machine ensemble based on feature and hyperparameter variation for real-world machine fault diagnosis', in: Iliadis, L., Jayne, C. and Angelov, P. (eds.), *Engineering Applications of Neural Networks, Advances in Intelligent and Soft Computing*, Vol. 96, Berlin: Springer, pp. 271–282. https://doi.org/10.1007/978-3-642-20505-7_24

Wickramasinghe, I. and Kalutarage, H. (2021) 'Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation', *Soft Computing*, Vol. 25, No. 3, pp. 2277–2293. https://doi.org/10.1007/s00500-020-05297-6

Wu, M., Subramaniam, G., Zhu, D., Li, C., Ding, H. and Zhang, Y. (2024) 'Using machine learning-based algorithms to predict academic performance: A systematic literature review', in: *4th International Conference on Innovative Practices in Technology and Management (ICIPTM 2024)*, pp. 1–8. https://doi.org/10.1109/ICIPTM59628.2024.10563566

Xu, X., Wang, J., Peng, H. and Wu, R. (2019) 'Prediction of academic performance associated with internet usage behaviors using machine learning algorithms', *Computers in Human Behavior*, Vol. 98, pp.166–173. https://doi.org/10.1016/j.chb.2019.04.015

Yassine, E. A. and Mohammed, K. (2024) 'A Comparative Analysis of Decision Trees, Bagging, and Random Forests for Predictive Modeling in Monetary Poverty: Evidence from Morocco', *Applied Mathematics and Information Sciences*, Vol. 18, No. 2, pp. 233–240. https://doi.org/10.18576/amis/180203

Zhang, L., Li, K. F. and Bourguiba, I. (2021a) 'Recent advances in academic performance analysis', in: I*nternational Conference on Higher Education Advances (HEAd21)*, pp. 607–614. https://doi.org/10.4995/HEAd21.2021.13196

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H. and Shang, X. (2021b) 'Educational data mining techniques for student performance prediction: method review and comparison analysis', *Frontiers in Psychology*, Vol. 12, p. 698490. https://doi.org/10.3389/fpsyg.2021.698490

ACTIONABLE LEARNING ANALYTICS: PREDICTING UNIVERSITY PERFORMANCE LEVELS WITH INTERPRETABLE MACHINE LEARNING

Enrique De La Hoz[1]✉
Carlos Garcia-Yerena[1]
Ingrid Torres-Rojas[2]

[1]Universidad del Magdalena, Colombia

[2]Corporación Universitaria Autónoma del Cauca, Colombia

✉ enriquedelahoz@unimagdalena.edu.co

ABSTRACT

Higher education institutions need timely, explainable tools to identify students at risk of low performance on large-scale examinations and to guide targeted academic support strategies. In response to this challenge, this study proposes an explainable machine learning framework to predict undergraduate students' performance levels in Colombia's SABER PRO examination. Using student background variables (e.g., gender, region, school type, parental education, and occupation) and SABER 11 standardised test scores (Critical Reading, Mathematics, Citizenship Skills, Science, and English), we formulate a binary classification problem that distinguishes desirable outcomes (levels 3–4) from non-desirable outcomes (levels 1–2). We benchmark baseline models against non-linear learners, including XGBoost, GLMNET, SVM, DT, and LDA, using a 10-fold cross-validation protocol with systematic hyperparameter tuning. Model performance is assessed through confusion matrices and AUC scores. To support educational decision-making, we complement predictive results with explainability analyses, including global feature importance and individual-level explanations via SHAP, enabling transparent identification of the key drivers behind performance levels. The proposed approach provides actionable learning analytics to guide early academic support, promote responsible and transparent educational decision-making, and improve the likelihood of desirable SABER PRO achievement.

KEYWORDS

Academic performance, explainable artificial intelligence, learning analytics

HOW TO CITE

De La Hoz E., Garcia-Yerena C., Torres-Rojas I. (2026) 'Actionable Learning Analytics: Predicting University Performance Levels with Interpretable Machine Learning', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 15–27. http://dx.doi.org/10.7160/eriesj.2026.190102

Article history
Received
February 11, 2026
Received in revised form
March 24, 2026
Accepted
March 25, 2026
Available on-line
March 31, 2026

Highlights

- *The proposed framework supports actionable, interpretable, and responsible learning analytics for higher education.*
- *Combining prior academic achievement with socioeconomic context enables earlier and more equitable identification of students needing support.*
- *Explainability tools help institutions make transparent and auditable decisions when using predictive models.*
- *The proposed models show useful predictive performance and can support more efficient targeting of mentoring, tutoring, and academic reinforcement resources.*

INTRODUCTION

Predicting university learning outcomes has become a strategic priority for higher education systems that rely on standardised exit assessments to evaluate students' competencies near the end of undergraduate programs. In many countries, these assessments report results through ordered performance levels (e.g., four categories from low to high), which are directly used for institutional benchmarking, accreditation-related reporting, and the design of academic support strategies. This practical relevance aligns closely with the learning analytics agenda of converting educational records into actionable evidence to improve decision-making at scale (Long and Siemens, 2014). Therefore, beyond predictive accuracy, this study is positioned within a responsible learning analytics perspective. The purpose of the proposed framework is not merely to classify student performance but to support higher education institutions in making transparent, explainable, and contextually sensitive decisions. In this sense, the model's contribution lies in both its role in educational decision-making and its more efficient allocation of academic support resources, such as mentoring, tutoring, and reinforcement programs.

The academic variables are the most common predictors used in forecasting models, typically course grades, outputs from a learning management system, and standardised test results, because these indicators are closely related to academic management and are readily operationalised.

However, there is solid evidence that socioeconomic and contextual features, such as parental education and occupation, and economic resources, partially determine students' educational trajectories. Sirin (2005), in a systematic literature review, reveals a strong correlation between socioeconomic status and academic performance, suggesting the importance of contextual factors for both forecasting and classification. For instance, the international PISA tests integrate students' socioeconomic context through composite indices that represent parental education, employment status, and other factors, and serve as key variables for understanding and predicting learning outcomes (Lamichhane et al., 2021).

The main objective of a machine learning model is not only to predict a future value. The true value of machine learning modelling lies in the server as an objective support for decision-making, enabling the implementation of administrative and academic initiatives to improve educational quality in the long term. The philosophy of learning analytics has experienced rapid and sustained growth in recent years, leveraging diverse datasets and machine learning models to generate knowledge in the educational environment (Berens et al., 2019). However, there are common constraints in machine learning models applied to education, including difficulties in achieving generalizability across different scenarios and the need to include variables related to students' socioeconomic context (Delahoz-Domínguez and Hijón-Neira, 2025).

Consequently, one of the most common applications of learning analytics is the creation of early warning systems to identify students at risk of dropping out or being dissatisfied with the educational process. Previous research shows that predictive models can estimate a student's risk (Diaz Lema et al., 2024). However, they are not sufficiently self-explanatory to take effective action, since they do not seek to verify that the model correctly predicted a problem associated with the student, but rather to take actions to prevent the problem and to generalise the intervention scheme for future similar cases (Sušnjak, 2022). Multiple large-scale and meta-analytic studies find that socioecomic variables are a consistent predictor of academic performance, typically with moderate-to-medium effect sizes. Meta-analyses across primary/secondary education report strong correlations (Hasan et al., 2020). Consequently, machine learning models using socioeconomic and school-level variables (family income, parental education, school characteristics) achieve high accuracy in predicting success, with these SES-related features among the most important predictors (Sangsawang, 2025).

Based on the studies discussed above, predictive systems appear to be most valuable when they are (i) built from data available early enough to support action, and (ii) designed to produce outputs that can be operationalised within institutional processes. Motivated by these needs, this study develops a machine learning approach to predict students' performance levels in a final-year national standardised assessment using information available before graduation. Our feature set integrates two complementary sources: (i) student background variables and (ii) prior standardised test results from the end of high school. The background predictors include gender, region of residence, school sector (public/private), school calendar, and parental characteristics such as mother's/father's education and mother's/father's occupation. The academic preparation predictors correspond to competences assessed in the national exam in Colombia (critical reading, mathematics, citizenship skills, science, and English).

The combination of academic and socioeconomic predictors aims to improve predictive performance by capturing both intrinsic and extrinsic student characteristics, thereby enabling more equitable and context-sensitive educational decision-making. In the present research, we consider a binary formulation that distinguishes desirable outcomes (upper levels, 3–4) from non-desirable outcomes (lower levels, 1–2). This dual framing supports operational decision rules, such as prioritising students for mentoring, tutoring, or targeted academic reinforcement. In this way, the proposed framework contributes not only to prediction but also to more responsible and effective educational action.

Methodologically, we follow the recommendation that comparative modelling is essential because no single algorithm consistently dominates across educational contexts and feature types (Domínguez-Jiménez et al., 2020). We therefore benchmark interpretable baselines against non-linear machine learning methods that are frequently reported as strong performers in student outcome prediction. To improve robustness and reduce overfitting risk, models are evaluated under cross-validation and systematic hyperparameter tuning, consistent with best practices discussed in the Learning Analytics (LA) and Educational Data Mining (EDM) (Durairaj and Vijitha, 2014). Finally, educational prediction requires more than accuracy: institutions need transparent explanations to justify interventions and build trust among stakeholders. This study, therefore, incorporates explainable AI methods to provide both global interpretability (which predictors matter most) and local interpretability (why a particular student is predicted to fall into a given level). Model-agnostic local explanations and Shapley-value-based attribution are widely adopted approaches for interpreting complex models, especially when predictions may trigger real academic decisions (Messalas et al., 2019; Parisineni and Pal, 2023). In this way, the proposed framework aims to deliver not only predictive performance but also actionable, auditable insights consistent with the goals of the ERIES special issue on predicting learning outcomes using machine learning.

LITERATURE REVIEW

Predictive learning analytics in higher education

The development of EDM/LA in higher education aims to identify early signals that help students succeed. The key idea behind learning analytics is translating educational data into valuable insights to support decision-making and improve student satisfaction (Siemens, 2019). Predictive models play a fundamental role in the development of learning analytics, implementing everything from descriptive statistical techniques to advanced supervised and unsupervised machine learning models.

In the literature, there is a recurring emphasis on aligning predictive models with institutional constraints and contexts. Specifically, models must work with the information and data available in the environment to facilitate effective interventions and achieve interpretable, actionable results. Accordingly, Predictive Learning Analytics (PLA) most often predicts course grades, pass/fail status, retention, and dropout, as well as learning outcomes and engagement levels (Sghir et al., 2022). Besides, Common data sources include Learning Management Systems (LMS) logs (clicks, submissions), video analytics, assessment scores, demographics, prior Grade Point Average (GPA), and sometimes emotions or self-reports (Hasan et al., 2020; Umer et al., 2021). Supervised ML dominates: decision trees, random forests, Support Vector Machines (SVMs), K-Nearest Neighbours (k-NNs), logistic regression, neural networks, and ensembles (Edmond et al., 2025; Yağcı, 2022). Accordingly, ensemble and hybrid models (bagging, boosting, Random Forest (RF), gradient boosting) generally achieve the best accuracy and robustness (Chen et al., 2025). Reported accuracies often range 70–90%+ (e.g., ~70–75% with simple features; ~88–98% in richer or smaller datasets), but generalizability across contexts is uncertain (Pali and Verma, 2024).

Academic Predictors and the Role of Previous Standardised Tests

Across higher education prediction models, prior standardised/ entrance scores and institutional academic variables are typically integrated as complementary features in a single feature set, rather than as separate model stages (Rhaiem, 2017). In consequence, predictive higher-education models start from pre-admission data (e.g., high school grades, standardised entrance exams) and then add institutional academic variables, such as course grades, midterms, attendance, teacher quality, and program/department indicators, to form a single feature vector. These are fed simultaneously into ML models such as Random Forests, Gradient Boosting, SVMs, XGBoost, neural networks, or AutoML frameworks (Guevara-Reyes et al., 2025; Zeineddine et al., 2020). According to Ahmed et al. (2025), excluding entrance or standardised examination scores represents a missed opportunity, since the inclusion of university entrance exam data could further enhance predictive accuracy and robustness.

Accordingly, studies that quantify feature importance consistently find that historical academic performance (standardised/entrance exams, prior GPA, midterms) is a top predictor of later success, often alongside institutional variables such as attendance and parental education (Ahmed et al., 2025; Talajić et al., 2025) such as artificial intelligence (AI). However, some evidence warns that standardised test scores and highschool GPA are not universally reliable across diverse contexts, so models that also integrate institutional and contextual variables (teacher quality, infrastructure, student–teacher ratio) tend to generalise better and avoid overreliance on test scores alone (Guevara-Reyes et al., 2025).

Socioeconomic context as a key driver of educational outcomes

A substantial body of research in higher education consistently shows that family income, parental education, and parental occupation influence academic outcomes through multiple interrelated mechanisms involving resource availability, family support, and exposure to stress. In the Chinese context, mixed-method studies have found that higher parental education and greater family income are associated with better university GPA, as students from more affluent households tend to report stronger financial and emotional support, more favorable study environments, and, consequently, better academic performance; however, these studies also note that intrinsic motivation and institutional support play an important complementary role in shaping outcomes (Wang and Panicker, 2025). Similarly, research conducted in Türkiye shows that household income, need to work while studying, parental education, and region of residence significantly affect the chances of entering a desired university department and placement rank, indicating that regional economic disparities and the necessity of student employment channel inequality into selective higher education outcomes (Kutlu and Özer, 2024).

This pattern is reinforced by evidence from other national settings. Among Sudanese medical students, higher family income is significantly associated with higher cumulative GPA, even after controlling for age and other relevant characteristics, highlighting the persistent influence of socioeconomic background in academically demanding programs (Jaber et al., 2024). Likewise, Cross-country panel analysis finds that household access to credit significantly increases higher education participation, especially in developing countries, whereas macroeconomic uncertainty expands university enrollment in developed economies but reduces it in developing ones; in addition, the combination of uncertainty and household credit particularly harms women's tertiary outcomes (Koirala et al., 2024). In the same vein, during the COVID-19 pandemic in Colombia, parental education and household technological assets (e.g., computers, internet) were positively associated with test scores both before and during the pandemic, with the importance of technology and high-quality institutions increasing under remote instruction, thus magnifying the academic impact of household and institutional resources (Mena and Bulla, 2022). Taken together, these findings support the view that socioeconomic status should be understood not as a simple one-step predictor of grades, but rather as a structural background condition that shapes access to resources, psychosocial support, and vulnerability to academic risk.

Additionally, Frameworks for academic research efficiency emphasise individual, organisational, and contextual drivers; efficiency cannot be meaningfully compared without modelling all three levels (Rhaiem, 2017). Consequently, academic productivity is a latent construct, dependent on disciplinary norms and institutional capacities; single, decontextualised outputs (e.g., publications only) are conceptually inadequate (Martinez-Daza et al., 2024). Therefore, comprehensive Higher Education Institutions (HEI) evaluation frameworks explicitly integrate context, inputs, processes, and products, arguing that context is a formal dimension of performance rather than a nuisance to be averaged out (Chinta et al., 2016). Besides, ignoring context yields biased productivity scores, high misclassification rates, and inequitable resource allocation, particularly penalizing less selective, less resourced, or disadvantaged institutions (Agasisti et al., 2022; Guo and Ye, 2025). Thus, Context-aware models—through value-added metrics, conditional efficiency estimation, or multi-dimensional Context Items Processess Products (CIPP)-style frameworks—are necessary to produce accurate, interpretable, and fair productivity assessments in higher education (Horn and Lee, 2019; Rhaiem, 2017).

Datasets and predictors

Across higher education analytics, data quality and class/group imbalance are central determinants of both accuracy and fairness. Therefore, large higher-education datasets often have substantial missing responses; how these are imputed can affect both performance and group disparities. Several studies on college success prediction show that common imputation methods can increase bias when test data reflect historical (unequal) distributions, even when headline accuracy is acceptable (Nezami et al., 2024). In the case of early warning, models are typically trained on heavily imbalanced pass/fail or on-time/late graduation labels. Without correction, models become biased toward the majority class, missing many at risk students and sometimes disadvantaging minority groups. In course and graduation prediction tasks, oversampling the minority class substantially improves minority-class recall and F1, sometimes with only a negligible loss in overall accuracy (Sha et al., 2023). Consequently, unequal representation of gender, race, or first-language groups yields distribution and hardness biases—models are trained more on majority groups and on "easier" examples. Studies on course success and forum classification show that such data characteristics are strongly associated with systematic performance gaps across demographic groups (Sha et al., 2022).

Considering the variability in the model's predictors, it is evident that educational machine learning models for predicting student success are highly sensitive to which predictors are used and how they are engineered. For El-kenawy et al. (2025), careful feature selection and engineering consistently improve accuracy, reduce overfitting, and enable good performance with less data or fewer features. The systematic review by Alsariera et al. (2022) emphasises that prediction quality is determined by the traits or features used. Accordingly, the Academic variables (GPA/CGPA, grades, and attendance), internal assessments, and demographic/family attributes are repeatedly identified as high value predictors (Ahmed, 2024; Alsariera et al., 2022). Also, adaptive or ensemble selection methods reduce dimensionality while maintaining competitive cross-validation accuracy, simplifying models and speeding up training (Malik et al., 2025).

Explainability and responsible prediction

Models in social contexts should be transparent and explainable, as the decisions they make impact people's futures. From this perspective, Explainable AI (XAI) emerges as a framework for explaining the outputs of machine learning models. Consequently, an example of a model-agnostic technique is local surrogate explanations, which was created to help users understand individual predictions (Alonso and Casalino, 2019). From another perspective, methodologies based on Shapley values enable us to evaluate each predictor's contribution to the responses (Melo et al., 2022).

In the study by Johora et al. (2025), they incorporate XAI directly into the modelling flow for academic performance predictions, thereby transitioning from a model based on accuracy to one driven by decision support. Consequently, explainable AI (XAI) techniques such as SHAP and LIME are used to clarify global feature importance and local, per-student predictions, enabling trust and legal compliance (Oyedotun et al., 2025; Sušnjak, 2022). Thus, Responsible LA is framed as relational and "responseable": institutions must act appropriately on predictions rather than merely generate them (Khalil et al., 2023; Rets et al., 2023) facilitate effective teaching, highlight aspects of course content that might be adapted, and predict a range of possible outcomes, such as students registering for more appropriate courses, supporting students' self-efficacy, or redesigning a course's pedagogical strategy. It will do all these things based on the assumptions and rules that learning analytics developers set out. As such, learning analytics can exacerbate existing inequalities such as unequal access to support or opportunities based on (any combination of. From an operational aspect, Practical recommendations include ethics protocols, stakeholder involvement, ongoing monitoring, and an explicit equitybydesign framework (Mathrani et al., 2021). Accordingly, in this study, responsibility is understood as the combination of interpretability, transparency, and context-sensitive use of predictions in educational settings. The aim is not to automate decisions about students, but to provide evidence that can support fairer and more accountable institutional actions. At the same time, these predictions may improve institutional effectiveness by helping universities prioritise limited support resources according to student needs.

MATERIALS AND METHODS

Anticipating which supervised model will optimally fit the data a priori is challenging; this principle is referred to as the No Free Lunch theorem. Algorithms with superior theoretical predictive capabilities may sometimes fail to elucidate the links between input and output variables. Consequently, in light of the performance disparities among various algorithms, five methods will be employed: Extreme Gradient Boosting (XGBoost), Generalised Linear Model – Elastic Net Regularisation (GLMNET), SVM with linear kernel, Decision Trees (DT), and Linear Discriminant Analysis (LDA).

Proposed modeling structure

This study develops and evaluates five machine learning models for predicting university performance levels. The procedure begins by splitting the original dataset into 80% for training and 20% for testing (See Figure 1). Model training is conducted using a 10-fold cross-validation scheme, where the training set is randomly partitioned into 10 equal-sized folds. In each iteration, nine folds are used to train the model, and the remaining fold is used for validation, rotating the validation fold until all ten folds have been used once. The average predictive performance across the ten iterations is then reported as the overall evaluation of each model.

The model's input variables represent student information, grouped into three categories: high school standardised test results, high school socioeconomic data, and college standardised test results. The structure, category, and description of the variables are presented in Table 1.

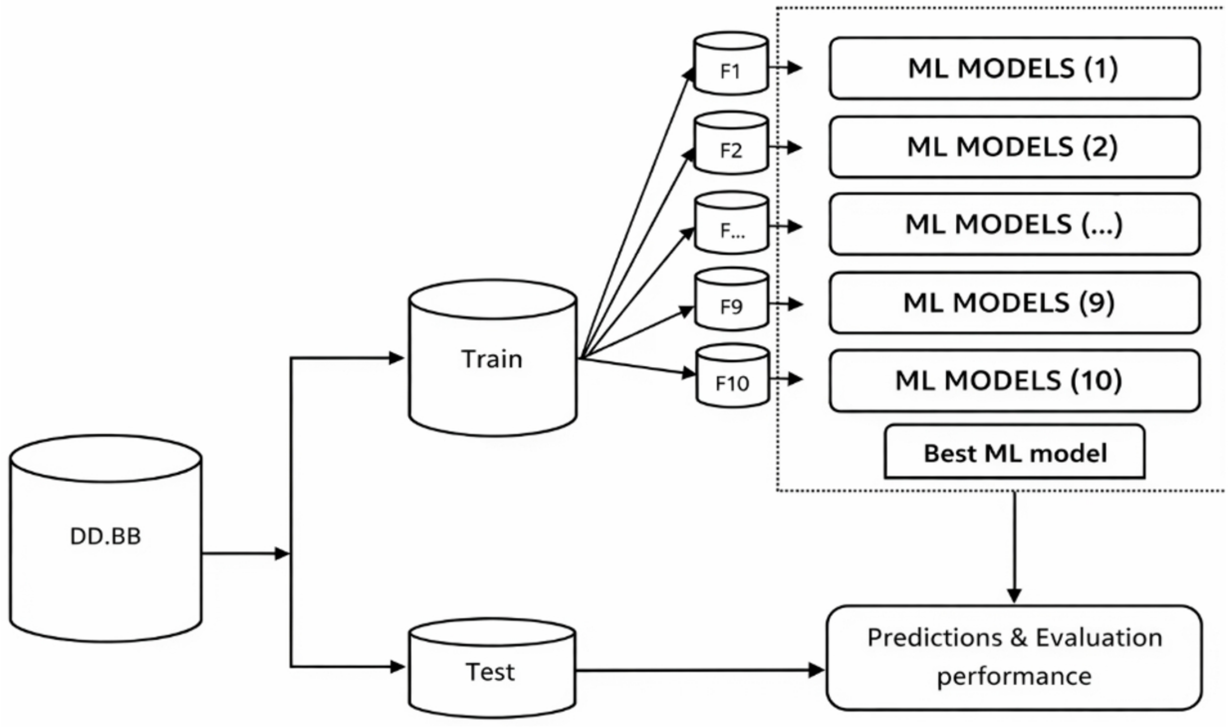


Figure 1: Machine Learning framework

Category	Name	Type	Levels/scale
Student Background	Gender (gen)	C	Male, female
	Department of Residence (dep.res)	C	Students' department of residence
	School type	C	Public, private
	School calendar (sch)	C	Calendar_A, Calendar_B
	Father's education (fedu) Mother education (medu)	C	Complete professional education, Incomplete professional education, None, Does not know, Postgraduate, complete secondary school, incomplete secondary school, complete Technical degree, incomplete technical degree.
	Father's occupation (focus) Mother's occupation	C	unemployed, general manager, auxiliary level employee, Domestic employee, businessman, Stay-at-home, day laborer, employee of a private company, government employee, Other activity or occupation, Little Businessman, Independent professional, Unpaid family worker, Self-employed, Worker without remuneration.
Standardised test at High School	Critical Reading (CR)	N	Score in the test (0-100)
	Math (Math)	N	Score in the test (0-100)
	Citizenship Skills (CS)	N	Score in the test (0-100)
	Science (sci)	N	Score in the test (0-100)
	English (ENG)	N	Score in the test (0-100)
Standardised test at University			
	Level of performance	N	1, 2, 3, 4

Table 1: Description of raw variables. In the Type column, the N denotes a numerical variable and C.

By feeding the model with a new student's social factors and academic performance on the Sabre Pro university exam, the model can operationalise predictions of which occupations the student is most likely to excel in. The operational flow of the recommendation system is depicted in Figure 2.

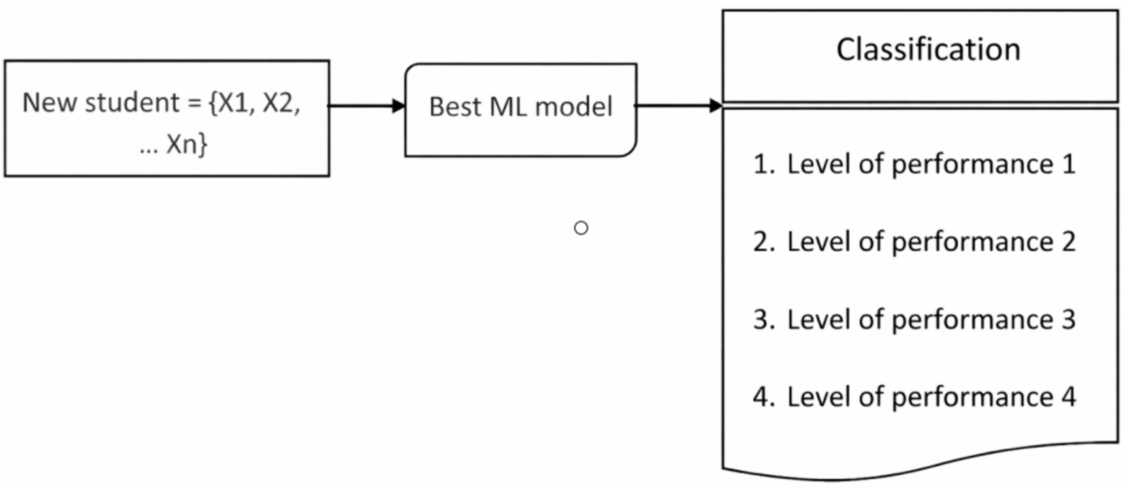

New student = {X1, X2, ... Xn}
Best ML model
Classification
1. Level of performance 1
2. Level of performance 2
3. Level of performance 3
4. Level of performance 4

Figure 2: Predictions development

Dataset

Data were collected from the Colombian Education Institute (ICFES) and included the records of undergraduate students who had participated in at least one national academic exam from 2008 to 2022. In total, 921,041 records containing the results of standardised tests for high school and university are present in the data. There are 102 degree focus options across multiple categories, including engineering, literature, science, and art. In addition, some socioeconomic information is included, such as parents' education and occupation, type of school, academic calendar, gender, and department of residence. Consequently, for each student, fourteen variables were defined, as described in Table 1. Eight of them are categorical variables derived from students' demographic information and school background. The above information was gathered from the ICFES repository. These variables were taken as inputs of the prediction model. The result of the standardised test at the university stage was defined as the target variable. Before analysis, the dataset was anonymised to protect any private or sensitive information.

Descriptive data analysis

Table 2 presents the descriptive analysis results for the 921,041 records in the database, adjusted by gender, over the 14 years of the study. As shown in Table 2, the gender distribution is 58% female and 42% male. Additionally, the median and mean values for all evaluated modules are higher in men than in women.

Variable	gender	*n*	Min	*q1*	Median	Mean	*q3*	Max	*sd*	IQR
CR	F	548046	0	47.1	53.0	53.5	59.3	102.6	9.5	12.1
	M	372995	0	47.3	53.2	54.1	60.0	113.2	9.6	12.7
	all	921041	0	47.2	53.0	53.7	59.4	113.2	9.5	12.2
MATH	F	548046	0	45.0	51.0	51.8	58.1	120.4	10.8	13.2
	M	372995	0	47.7	55.0	56.0	63.0	121.5	12.1	15.4
	all	921041	0	45.4	53.0	53.5	60.0	121.5	11.5	14.6
SCI	F	548046	0	45.6	51.0	52.0	57.8	122.2	9.6	12.2
	M	372995	0	47.5	53.2	54.5	61.0	123.0	10.7	13.5
	all	921041	0	46.3	51.9	53.0	58.6	123.0	10.1	12.3
CS	F	548046	0	45.9	52.0	52.4	58.2	108.3	9.7	12.3
	M	372995	0	47.7	54.0	54.1	60.3	107.0	10.3	12.6
	all	921041	0	46.1	53.0	53.1	59.8	108.3	10.0	13.6
ENG	F	548046	0	43.0	49.0	52.4	58.4	117.3	13.6	15.4
	M	372995	0	43.5	50.9	54.5	61.7	117.3	14.6	18.2
	all	921041	0	43.5	50.0	53.3	59.7	117.3	14.0	16.2
average.pro	F	548046	0	10.6	131.2	99.6	155.4	265.0	69.3	144.8
	M	372995	0	10.9	135.4	103.3	162.0	268.8	72.0	151.1
	all	921041	0	10.7	132.6	101.1	158.0	268.8	70.4	147.3

Model evaluation and Performance metrics

Table 2: Descriptive Statistics

The efficacy of the classification procedure is established by analysing the divergence between predicted outcomes and ground truth labels. This relationship is quantified using the fundamental metrics True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) (Fawcett, 2006). To provide a comprehensive assessment of the model's performance, we utilise the Correct Classification Rate (C) alongside the Receiver Operating Characteristic (ROC) curve analysis. The ROC curve serves as a graphical representation of the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate (1-Specificity) across a continuum of discrimination thresholds. The diagnostic utility of this curve is summarised by the Area Under the Curve (AUC) (Hanley and McNeil, 1982). Numerically, an AUC value of 1.0 denotes a model with flawless categorisation and perfect separability. In contrast, an AUC of 0.5 indicates a model with no predictive power, performing no better than random chance.

RESULTS

This section presents the 10-fold cross-validation results (See Table 3). Table 4 shows the tuning hyperparameters for all the models and the selected ones. Table 3 reports the AUC–ROC results across validation iterations for the five models. Overall, XGBoost achieved the strongest and most consistent discriminative performance, with a mean AUC of 0.85 and a relatively narrow range (0.77–0.92), indicating robust generalisation across data splits. GLMNET ranked second (mean AUC = 0.77; range: 0.66–0.85), suggesting that a regularised linear decision boundary captures part of the signal but does not reach the performance of the non-linear boosting approach. In contrast, the SVM model exhibited substantial variability, with a mean AUC of 0.66 and values ranging from 0.44 to 0.88, suggesting sensitivity to the specific training/validation partitions and potentially to hyperparameter settings or class distributions within folds.

Model	AUC		
	Min	Mean	Max
XGBOOST	0.77	0.85	0.92
GLMNET	0.66	0.77	0.85
SVML	0.44	0.66	0.88
DT	0.33	0.56	0.77
LDA	0.22	0.62	0.77

Table 3: Cross-validation scores for AUC-ROC

The single Decision Tree yielded lower average performance (mean AUC = 0.56; range: 0.33–0.77), consistent with the limited generalisation of un-ensembled trees, while LDA showed a moderate mean AUC of 0.62 but the widest instability at the lower end (0.22–0.77), suggesting that linear-discriminant assumptions may not hold uniformly across folds. Taken together, these results identify XGBoost as the most reliable model for predicting university performance levels in this setting, while GLMNET offers a competitive, more parsimonious alternative.

Model	Tuning parameters
XGBOOST	n.trees = 500, max_depth = c(1,4), eta = c(0.01; 0.1)
GLMNET	alpha = c(0,1), lambda = seq(0.001; 0.1 by 0.001)
SVML	cost = 2^c(0,5)
DT	cp = 2^c(-30, -22, -14, -8, -2)
LDA	NA

Table 4: Hyperparameters tuning

As explained in the materials and methods section, 20% of the data was reserved for the validation phase of the supervised learning models used. Figure 3 shows the ROC results for data that the algorithms did not previously know. In the model comparison, XGBoost shows the best performance, with an ROC value of 0.91. Thus, the ROC values of the five algorithms implemented are above 0.65. Those results indicate that these four models perform well at classifying academic performance. However, XGBoost outperforms the other models (XGBoost vs GLMNET, RF, KNN), achieving an ROC value of 0.91.

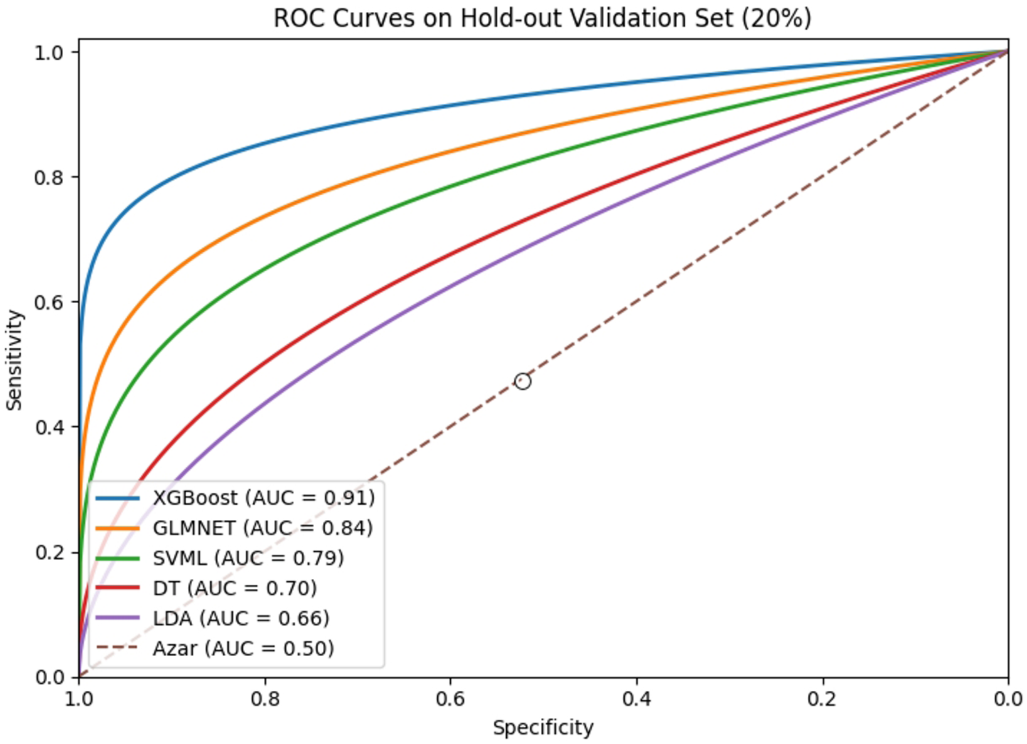


Figure 3: ROC curve for the validation set

Explainability through Global SHAP results

The Global SHAP results show that the most influential predictors of university performance are the prior academic competencies measured in SABER 11. In particular, Mathematics emerges as the strongest predictor, followed by Critical Reading, English, Science, and Citizenship Skills. This pattern indicates that university performance levels are primarily associated with the academic foundations students bring from secondary education. The direction of the effects is also consistent across these variables: higher prior scores are systematically associated with a greater probability of achieving higher performance levels. Among these predictors, Mathematics and Critical Reading stand out as the most decisive, suggesting that quantitative reasoning and academic literacy constitute the core competencies underpinning later achievement.

A second relevant finding is that importance values decrease gradually after the first five variables, suggesting a clear distinction between the predictive weight of prior academic preparation and that of background characteristics. While the standardised test competencies dominate the model, family-related variables still provide meaningful explanatory value. In particular, mothers' and fathers' education rank among the most important non-academic predictors, indicating that household educational capital shapes later academic outcomes. These variables likely capture differences in home-based academic support, expectations, and familiarity with educational trajectories. Therefore, the results suggest that performance is not explained exclusively by prior test scores, but also by the broader educational environment in which students develop before entering higher education.

The SHAP results also show that school type, department of residence, and school calendar contribute additional predictive information, even if their global importance is lower than that of academic competencies and parental education. These variables point to structural and territorial differences in students' pre-university trajectories. For example, school type may reflect variation in institutional quality and access to academic preparation. At the same time, the department of residence may capture geographic inequalities linked to educational opportunity, infrastructure, or socioeconomic conditions. Similarly, the school calendar may signal differences in school organisation and cohort composition. Although these variables are not the principal drivers of prediction, their presence confirms that educational performance is shaped not only by individual ability, but also by contextual conditions that precede university entry.

Finally, the lower-ranked variables, such as mother's occupation, father's occupation, and gender, should not be interpreted as irrelevant, but rather as factors with smaller average global effects than the dominant academic predictors. In particular, the relatively low global importance of gender suggests that it is not a primary determinant of overall prediction in the full sample. However, it may still be relevant for subgroup analysis and fairness monitoring. Taken together, the Global SHAP results support an explainable view of student success in which prior academic achievement is the central driver. At the same time, family background and contextual conditions provide complementary information that improves prediction and enriches the educational interpretation of the model. From a practical perspective, these findings justify interventions focused on strengthening quantitative reasoning, reading comprehension, and language skills, as well as supporting mentoring and context-sensitive institutional strategies for students from less advantaged backgrounds.

Rank	Feature (original variable)	Type	Mean |SHAP|	Typical direction toward higher performance*	Interpretation for decision-making
1	Math (SABER 11)	N	0.182	Higher Math -> higher level	Quantitative preparation is the strongest driver; it supports early reinforcement in quantitative reasoning.
2	Critical Reading (CR, SABER 11)	N	0.164	Higher CR -> higher level	Reading competence strongly differentiates levels; it points to academic literacy interventions.
3	English (ENG, SABER 11)	N	0.121	Higher ENG -> higher level	Language skills contribute to higher performance; they support strengthening English/academic language.
4	Science (sci, SABER 11)	N	0.107	Higher sci -> higher level	Scientific reasoning helps separate mid/high levels; indicates need for reasoning-focused support.
5	Citizenship Skills (CS, SABER 11)	N	0.093	Higher CS -> higher level	Civic competencies add signal beyond Math/CR; supports broad competency-building strategies.
6	Mother education (medu)	C	0.071	Higher education categories -> higher level	Proxy for educational capital; informs mentoring and structured academic guidance.
7	Father's education (fedu)	C	0.064	Higher education categories -> higher level	Similar to medu, it indicates differential support structures outside the university.
8	School type	C	0.058	Context-dependent	Captures pre-university institutional differences; suggests the need for differentiated onboarding.
9	Department of Residence (dep. res)	C	0.045	Context-dependent	Geographic disparities may reflect unequal opportunity; supports territory-sensitive support strategies.
10	School calendar (sch)	C	0.033	Context-dependent	Signals structural differences in prior schooling; helps interpret cohort heterogeneity.
11	Mother occupation (mocu)	C	0.028	Context-dependent	Socioeconomic proxy; may indicate time/resources available for academic support.
12	Father's occupation (focus)	C	0.024	Context-dependent	Socioeconomic proxy complements parental education as a background signal.
13	Gender (gen)	C	0.017	Context-dependent	Lower global impact; mainly relevant for subgroup monitoring and fairness diagnostics.

Table 5: Global SHAP summary for the XGBOOST model

DISCUSSION

This study aimed to determine whether early availability of academic and socioeconomic information can predict university performance in a final-year standardised exit assessment, under the premise that incorporating contextual variables should enhance the practical value of predictions for institutional decision-making. Thus, our development is aligned with the principles of Learning Analytics, which emphasise transforming educational data into knowledge to support timely interventions and continuous improvement (Long and Siemens, 2014). The results indicate that the selection and parameterisation of the model are critical, particularly for adjusting the nonlinear predictors and balancing discrimination and robustness.

In the cross-validation, XGBoost consistently achieved the best discriminative performance, with the highest mean AUC (0.85). This pattern suggests that the relationships between predictors and outcomes probably involve non-linearities and interactions that are not well captured by strict linear decision boundaries. This interpretation aligns with the established strengths of boosting methods for varied tabular datasets, where decision rules often depend on thresholds and complex feature combinations (Kolo et al., 2015). In comparison, GLMNET achieved a competitive but clearly lower performance level (mean AUC = 0.77), suggesting that a regularised linear model can capture some of the signal but might not fully capture the complex relationships between prior academic preparation and contextual factors.

Our results indicate that the model effectively predicts students' probability of belonging to higher or lower performance levels in the final-year standardised assessment, rather than academic success in a broad or undefined sense. This distinction is important because the dependent variable is operationalised as performance-level membership, and the Learning analytics interventions that explicitly predict actionable targets show improved pass rates, grades, and retention when educators use these outputs for targeted support (Alalawi, 2024a; Alalawi et al., 2024b). At the same time, the present results extend that perspective by showing that performance-level prediction can be meaningfully supported using information available before graduation, reinforcing the argument that actionable educational prediction depends on temporal usefulness as much as statistical precision (Pelima et al., 2024).

Among the predictors, Mathematics and Critical Reading are the most influential, followed by English, Science, and Citizenship Skills. This ordering suggests that later performance in the final-year assessment depends primarily on a broad academic preparation profile, with quantitative and literacy-related competencies occupying a central role. These findings are consistent with studies showing that prior academic achievement and standardised test competencies tend to be among the strongest predictors of subsequent university outcomes (Cerdeira et al., 2018). More specifically, the prominence of Mathematics and Critical Reading aligns with research emphasising that quantitative reasoning and academic literacy often structure performance not only in discipline-specific settings but also in general higher education assessments (Delahoz Dominguez et al., 2025). In contrast, some studies have suggested that institutional variables such as attendance, course grades, or early-semester assessments may dominate predictive models once university trajectory data are available (Parker et al., 2012). In the present case, however, the strong contribution of SABER 11 scores suggests that pre-university competencies remain highly informative even at later stages of the academic pathway.

The results also show that performance is not explained exclusively by prior academic scores. Variables such as mother's education, father's education, school type, department of residence, and school calendar provide additional predictive information, even if their global importance is lower than that of the academic competencies. This finding aligns with the literature, which shows that socioeconomic and contextual factors shape academic outcomes by shaping educational capital, resource availability, prior school quality, and territorial opportunity structures (Wang and Panicker, 2025). In particular, the contribution of parental education is consistent with studies linking family educational background to stronger academic guidance, higher expectations, and more supportive learning environments (Guerra and Braungart-Rieker, 1999). Similarly, the effects of school type and department of residence are compatible with research showing that institutional and regional inequalities influence students' preparation before entering higher education (Kutlu and Özer, 2024).

These findings are especially relevant when viewed through the lens of efficiency and responsibility in education. From an efficiency standpoint, the model can help institutions allocate limited academic support resources more strategically by identifying, before graduation, students who are more likely to fall into lower performance levels and the competency domains in which reinforcement may be most needed. This interpretation is consistent with prior studies arguing that predictive models create institutional value when they support earlier, more targeted, and more cost-conscious interventions (Delahoz-Domínguez and Hijón-Neira, 2024). From a responsibility standpoint, the use of an interpretable framework such as SHAP is essential because it enables universities to justify predictions, communicate the basis of model outputs, and reduce the risks associated with opaque algorithmic decision-making (Guevara-Reyes et al., 2025). Importantly, lower-ranked predictors such as gender should not be interpreted as irrelevant, but rather as variables whose role may be more visible in subgroup analysis, fairness monitoring, or interaction effects than in pooled global importance rankings (Delahoz-Domínguez and Hijón-Neira, 2025). In this sense, the present results also support recent calls for responsible learning analytics that combine predictive performance with transparency, contextual sensitivity, and attention to potential differential effects across student populations (Sangsawang, 2025).

Prior work emphasises that predictive tools are most useful when they identify students early enough for meaningful support and when evaluation reflects the real-world costs of false negatives and false positives (Mathrani et al., 2021). Complementary frameworks for identifying at-risk students further argue that models should be aligned with educators' decision-making needs rather than treated as purely technical forecasting tools (Pali and Verma, 2024). In our case, defining the target both as a binary "desirable vs non-desirable" outcome enables flexible operationalisation: the binary framing supports actionable triage under limited support capacity. Importantly, predicting performance levels rather than only continuous scores enhances practical adoption because they are easier to communicate to stakeholders.

Simultaneously, the application of predictive analytics in educational contexts requires a focus on both operational effectiveness and ethical considerations. Efficiency is demonstrated through the generation of precise and consistent predictions, leveraging readily accessible data from administrative and assessment platforms while avoiding the imposition of supplementary data-acquisition requirements (Oyedotun et al., 2025). Conversely, responsibility mandates preventing model outputs that exacerbate existing disparities or validate diminished expectations for marginalised populations. Because socioeconomic factors can be both predictive and ethically sensitive, their responsible use requires subgroup auditing and careful interpretation. Recent studies highlight the need to incorporate fairness assessments into student performance prediction systems and to examine the trade-off between accuracy and fairness across different groups (Valdivia et al., 2021). Therefore, a key implication of this research is that results should be reported not only as overall metrics (such as AUC) but also for specific groups (e.g., gender, school type, and region) to identify differences in error patterns that could affect the fairness of intervention strategies. Consequently, the strong performance of complex models such as XGBoost increases the importance of explainability. Educational stakeholders typically require transparent reasoning, particularly when predictions may influence support allocation or advising decisions. Explainable AI methods help bridge this gap by providing global and local interpretations—showing which variables drive overall predictions and why a specific student is predicted to fall into a given outcome category (Alonso and Casalino, 2019). In applied terms, interpretability enables institutions to move from "prediction as labeling" toward "prediction as guidance," supporting targeted academic actions. Accordingly, from an institutional perspective, these findings show that interpretability is not an accessory component of the model, but a necessary condition for responsible application in educational settings.

By identifying the predictors most strongly associated with lower performance levels, the framework can help universities design interventions that are both more transparent and more efficiently targeted, particularly when academic support resources are limited.

Finally, several limitations should be acknowledged. First, the study relies on administrative and testing variables; psychosocial constructs such as belonging, engagement, or institutional climate—often linked to achievement—are not included and may explain additional variance if consistently measured. Second, predictive performance may shift across cohorts due to curriculum changes or assessment redesign, necessitating temporal validation and periodic recalibration. Third, as with most LA/EDM prediction studies, the results are correlational rather than causal; models identify patterns useful for forecasting but do not establish causal mechanisms. Future work can extend this approach by incorporating explicit ordinal-learning objectives, probability calibration, and systematic fairness auditing under realistic intervention constraints.

CONCLUSION

The present research develops a predictive framework in learning analytics. Consequently, integrating standardised test scores with student socioeconomic variables. The proposed methodology facilitates the early identification of students at risk of low performance and those who could benefit from timely academic assistance. This approach addresses the shortcomings of prediction systems that rely solely on academic indicators, thereby neglecting broader contextual factors that influence university learning outcomes.

A significant contribution of this research lies in the empirical evaluation of diverse model families, conducted within a robust validation framework. Furthermore, this research highlights the practical value of modelling results as ordered performance levels, and also as a binary classification that distinguishes between good and poor achievement. This approach mirrors the typical ways institutions share assessment results and the threshold-based decision-making processes often used in academic support programs. Within this framework, predictive capabilities act as a decision-support tool, potentially improving early-warning systems by enabling universities to allocate resources for mentoring, tutoring, and targeted intervention programs where they are most needed. In conclusion, the proposed methodology demonstrates that incorporating prior test performance alongside socioeconomic and contextual variables enables precise, practically useful forecasting of university achievement. Beyond predictive accuracy, this study contributes to the development of responsible learning analytics by promoting explainable, auditable, and context-sensitive use of machine learning in higher education. In practical terms, the framework can support more effective early-warning systems and a more efficient allocation of academic support resources, helping institutions direct mentoring, tutoring, and reinforcement efforts to students most likely to benefit.

REFERENCES

Agasisti, T., Egorov, A. and Serebrennikov, P. (2023) 'Universities' efficiency and the socioeconomic characteristics of their environment-evidence from an empirical analysis', *Socio-Economic Planning Sciences*, Vol. 85, p. 101445. https://doi.org/10.1016/j.seps.2022.101445

Ahmed, E. (2024) 'Student Performance Prediction Using Machine Learning Algorithms', *Applied Computational Intelligence and Soft Computing*, Vol. 2024, No. 1, p. 4067721. https://doi.org/10.1155/2024/4067721

Ahmed, W., Wani, M. A., Pławiak, P., Meshoul, S., Mahmoud, A. and Hammad, M. (2025) 'Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions', *Scientific Reports*, Vol. 15, No. 1, p. 26879. https://doi.org/10.1038/s41598-025-12353-4

Alalawi, K., Athauda, R. and Chiong, R. (2024a) 'An Extended Learning Analytics Framework Integrating Machine Learning and Pedagogical Approaches for Student Performance Prediction and Intervention', *International Journal of Artificial Intelligence in Education*, Vol. 35, Né. 3, pp. 1239–1287. https://doi.org/10.1007/s40593-024-00429-7

Alalawi, K., Athauda, R., Chiong, R. and Renner, I. (2024b) 'Evaluating the student performance prediction and action framework through a learning analytics intervention study', *Education and Information Technologies*, Vol. 30, No. 3, pp. 2887–2916. https://doi.org/10.1007/s10639-024-12923-5

Alonso, J. M. and Casalino, G. (2019) '*Explainable artificial intelligence for human-centric data analysis in virtual learning environments*', in: Burgos, D., Cimitile, M., Ducange, P., Pecori, R., Picerno, P., Raviolo, P. and Stracke, C. M. (eds.), Higher Education Learning Methodologies and Technologies Online, Cham: Springer International Publishing, pp. 125–138. https://doi.org/10.1007/978-3-030-31284-8_10

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A. and Ali, N. (2022) 'Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance', *Computational Intelligence and Neuroscience*, Vol. 2022, p. 4151487. https://doi.org/10.1155/2022/4151487

Berens, J., Schneider, K., Gortz, S., Oster, S. and Burghoff, J. (2019) 'Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods', *Journal of Educational Data Mining*, Vol. 11, No. 3, pp. 1–41. https://doi.org/10.5281/zenodo.3594771

Cerdeira, J. M., Nunes, L. C., Reis, A. B. and Seabra, C. (2018) 'Predictors of student success in Higher Education: Secondary school internal scores versus national exams', *Higher Education Quarterly*, Vol. 72, No. 4, pp. 304–313. https://doi.org/10.1111/hequ.12158

Chen, J., Zhou, X., Yao, J. and Tang, S.-K. (2025) 'Application of machine learning in higher education to predict students' performance, learning engagement and self-efficacy: a systematic literature review', *Asian Education and Development Studies*, Vol. 14, No. 2, pp. 205–240. https://doi.org/10.1108/aeds-08-2024-0166

Chinta, R., Kebritchi, M. and Ellias, J. (2016) 'A conceptual framework for evaluating higher education institutions', *International Journal of Educational Management*, Vol. 30, No. 6, pp. 989–1002. https://doi.org/10.1108/ijem-09-2015-0120

Delahoz-Dominguez, E., Zuluaga, R. and García-Yerena, C. E. (2025) 'Evaluación predictiva de las habilidades en razonamiento cuantitativo en ingeniería', *Magis: Revista Internacional de Investigación en Educación*, Vol. 18, pp. 117. https://doi.org/10.11144/Javeriana.m18.ehrc

Delahoz-Domínguez, E. J. and Hijón-Neira, R. (2024) 'Recommender System for University Degree Selection: A Socioeconomic and Standardised Test Data Approach', *Applied Sciences*, Vol. 14, No. 18, p. 8311. https://doi.org/10.3390/app14188311

Delahoz-Domínguez, E. J. and Hijón-Neira, R. (2025) 'SAIL-Y: A Socioeconomic and Gender-Aware Career Recommender System', *Electronics*, Vol. 14, No. 20, p. 4121. https://doi.org/10.3390/electronics14204121

Diaz Lema, M., Vooren, M., Cannistrà, M. van Klaveren, C., Agasisti, T. and Cornelisz, I. (2024) 'Predicting dropout in Higher Education across borders', *Studies in Higher Education*, Vol. 49, No. 1, pp. 141–156. https://doi.org/10.1080/03075079.2023.2224818

Domínguez-Jiménez, J. A., Campo-Landines, K. C., Martínez-Santos, J. C., Delahoz, E. J. and Contreras-Ortiz, S. H. (2020) 'A machine learning model for emotion recognition from physiological signals', *Biomedical Signal Processing and Control*, Vol. 55, p. 101646. https://doi.org/10.1016/j.bspc.2019.101646

Durairaj, M. and Vijitha, C. (2014) 'Educational data mining for prediction of student performance using clustering algorithms', *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 4, pp. 5987–5991. Available at: https://www.semanticscholar.org/paper/Educational-Data-mining-for-Prediction-of-Student-Durairaj-Vijitha/892b0182c44c34a2ae68daec819eaec301c3bd9c

Edmond, U. V., Sada, S. M. and Osijirin, A. N. (2025) 'Predictive Analytics Using Machine Learning Models on Undergraduate Students' Performance of the Federal University of Allied Health Sciences, Enugu, Nigeria in Introduction to Computing Science', *Saudi Journal of Engineering and Technology*, Vol. 10, No. 7, pp. 324–332. https://doi.org/10.36348/sjet.2025.v10i07.004

El-Kenawy, E. M., Alharbi, A. H., Alhussan, A., Eid, M. M., Sobhi, M. and Khafaga, D. S. (2025) 'Optimizing student performance prediction through feature selection and machine learning models', in: *2025 International Telecommunications Conference (ITC-Egypt)*, pp. 226–231. https://doi.org/10.1109/itc-egypt66095.2025.11186576

Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern Recognition Letters*, Vol. 27, No. 8, pp. 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Guerra, A. L. and Braungart-Rieker, J. M. (1999) 'Predicting Career Indecision in College Students: The Roles of Identity Formation and Parental Relationship Factors', *The Career Development Quarterly*, Vol. 47, No. 3, pp. 255–266. https://doi.org/10.1002/j.2161-0045.1999.tb00735.x

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R. O., Cox-Riquetti, F. and Villegas-Ch, W. (2025) 'Machine learning models for academic performance prediction: interpretability and application in educational decision-making', *Frontiers in Education*, Vol. 10, p. 1632315. https://doi.org/10.3389/feduc.2025.1632315

Guo, R. and Ye, M. (2025) 'Input-output efficiency, productivity dynamics, and determinants in western China's higher education: A three-stage DEA, global Malmquist index, and Tobit model approach', *PLOS One*, Vol 20, No. 6, p. e0325901. https://doi.org/10.1371/journal.pone.0325901

Hanley, J. A. and McNeil, B. J. (1982) 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, Vol. 143, No. 1, pp. 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U. and Sattar, M. (2020) 'Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques', *Applied Sciences*, Vol. 10, No. 11, p. 3894. https://doi.org/10.3390/app10113894

Horn, A. S. and Lee, G. (2019) 'Evaluating the Accuracy of Productivity Indicators in Performance Funding Models', *Educational Policy*, Vol. 33, No. 5, pp. 702–733. https://doi.org/10.1177/0895904817719521

Jaber, M. H., Dafallah, I. A., Mohammed, A. Y., Eltahir Eltahir, R., Mohamed, M. A., Mohamed, T. A., Mudawi, M. H., Tayfour, D. O., Mohammed Ali, S. E. M., Ahmed, E. A. M., Osman, A. M., Kakoum, H. A., Bagadi, M. M. and Mohammed, A. O. (2024) 'Socioeconomic disparities and their effect on medical student academic attainment Sudanese Universities', *BMC Medical Education*, Vol. 24, No. 1, 929. https://doi.org/10.1186/s12909-024-05867-4

Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M. and Akter, F. (2025) 'An explainable AI-based approach for predicting undergraduate students academic performance', *Array*, Vol. 26, p. 100384. https://doi.org/10.1016/j.array.2025.100384

Khalil, M., Prinsloo, P. and Slade, S. (2023) 'Fairness, Trust, Transparency, Equity, and Responsibility in Learning Analytics', *Journal of Learning Analytics*, Vol. 10, No. 1, pp. 1–7. https://doi.org/10.18608/jla.2023.7983

Koirala, N., Koirala, D., Nyiwul, L. and Hu, Z. (2024) 'Economic uncertainty, households' credit situations, and higher education', *Journal of Macroeconomics*, Vol. 80, p. 103598. https://doi.org/10.1016/j.jmacro.2024.103598

Kolo, K. D., Adepoju Solomon A. and Alhassan, J. K. (2015) 'A Decision Tree Approach for Predicting Students Academic Performance', *International Journal of Education and Management Engineering*, Vol. 5, No. 5, pp. 12–19. https://doi.org/10.5815/ijeme.2015.05.02

Kutlu, M. and Özer, H. (2024) T'he Effect of Economic and Social Inequalities on Academic Success in Türkiye: Evidence from the Classical and Bayesian Discrete Choice Models', *Prague Economic Papers*, Vol. 33, No. 3, pp. 336–356. https://doi.org/10.18267/j.pep.860

Lamichhane, S., Eğilmez, G., Gedik, R., Bhutta, M. K. S. and Erenay, B. (2021) 'Benchmarking OECD countries' sustainable development performance: A goal-specific principal component analysis approach', *Journal of Cleaner Production*, Vol. 287, p. 125040. https://doi.org/10.1016/j.jclepro.2020.125040

Long, P. and Siemens, G. (2014) 'Penetrating the fog: analytics in learning and education', *Italian Journal of Educational Technology*, Vol. 22, No. 3, pp. 132–137. https://doi.org/10.17471/2499-4324/195

Malik, S., Patro, S. G. K., Mahanty, C., Hegde, R., Naveed, Q. N., Lasisi, A., Buradi, A., Emma, A. F. and Kraiem, N. (2025) 'Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution', *Scientific Reports*, Vol. 15, No. 1, p. 8738. https://doi.org/10.1038/s41598-025-92324-x

Martinez-Daza, M. A., Valencia-Quecano, L. I. and Guzmán-Rincón, A. (2024) 'Conceptual Model for the Assessment of Academic Productivity in Research Seedbeds From a Systematic Review', *European Journal of Educational Research*, Vol. 13, No. 2, pp. 813–833. https://doi.org/10.12973/eu-jer.13.2.813

Mathrani, A., Sušnjak, T., Ramaswami, G. and Barczak, A. (2021) 'Perspectives on the Challenges of Generalizability, Transparency and Ethics in Predictive Learning Analytics', *Computers and Education Open*, Vol. 2, p. 100060. https://doi.org/10.1016/j.caeo.2021.100060

Melo, E., Silva, I., Costa, D. G., Viegas, C. M. D. and Barros, T. M. (2022) 'On the Use of eXplainable Artificial Intelligence to Evaluate School Dropout', *Education Sciences*, Vol. 12, No. 12, p. 845. https://doi.org/10.3390/educsci12120845

Mena, N. P. and Bulla, J. F. A. (2022) 'Socioeconomic conditions and academic performance in higher education in Colombia during the pandemic', *Quality in Higher Education*, Vol. 29, No. 2, pp. 242–260. https://doi.org/10.1080/13538322.2022.2088564

Messalas, A., Kanellopoulos, Y. and Makris, C. (2019) 'Model-agnostic interpretability with Shapley values', in: *10th International Conference on Information, Intelligence, Systems and Applications (IISA),* pp. 1–7. https://doi.org/10.1109/iisa.2019.8900669

Nezami, N., Haghighat, P., Gándara, D. and Anahideh, H. (2024) 'Assessing Disparities in Predictive Modeling Outcomes for College Student Success: The Impact of Imputation Techniques on Model Performance and Fairness', *Education Sciences*, Vol. 14, No. 2, p. 136. https://doi.org/10.3390/educsci14020136

Oyedotun, S. A., Ejenarhome, O. P. and Oise, G. (2025) 'Learning Analytics and Predictive Modeling: Enhancing Student Success through Data-Driven Insights', *Journal of Science Research and Reviews*, Vol. 2, No. 3, pp. 42–51. https://doi.org/10.70882/josrar.2025.v2i3.77

Pali, P. and Verma, S. (2024) 'Predictive Analytics for Student Performance: A Machine Learning Model for Higher Education', *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 12, No. 5, pp. 8151–8158. https://doi.org/10.15680/ijircce.2024.1205366

Parisineni, S. R. A. and Pal, M. (2023) 'Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations', *International Journal of Data Science and Analytics*, Vol. 18, No. 4, pp. 457–466. https://doi.org/10.1007/s41060-023-00458-w

Parker, P. D., Schoon, I., Tsai, Y.-M., Nagy, G., Trautwein, U. and Eccles, J. S. (2012) 'Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: a multicontext study', *Developmental Psychology*, Vol. 48, No. 6, pp. 1629–1642. https://doi.org/10.1037/a0029167

Pelima, L. R., Sukmana, Y. and Rosmansyah, Y. (2024) 'Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review', *IEEE Access*, Vol. 12, pp. 23451–23465. https://doi.org/10.1109/access.2024.3361479

Rets, I., Herodotou, C. and Gillespie, A. (2023) 'Six Practical Recommendations Enabling Ethical Use of Predictive Learning Analytics in Distance Education', *Journal of Learning Analytics*, Vol. 10, No.1, pp. 149–167. https://doi.org/10.18608/jla.2023.7743

Rhaiem, M. (2017) 'Measurement and determinants of academic research efficiency: a systematic review of the evidence', *Scientometrics*, Vol. 110, pp. 581–615. https://doi.org/10.1007/s11192-016-2173-1

Sangsawang, T. (2025) 'Predicting Student Achievement Using Socioeconomic and School-Level Factors', *Artificial Intelligence in Learning*. https://doi.org/10.63913/ail.v1i1.4

Sghir, N., Adadi, A. and Lahmer, M. (2022) 'Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)', *Education and Information Technologies*, Vol. 28, No. 7, p. 8299–8333. https://doi.org/10.1007/s10639-022-11536-0

Sha, L., Gašević, D. and Chen, G. (2023) 'Lessons from debiasing data for fair and accurate predictive modeling in education', *Expert Systems with Applications*, Vol. 228, p. 120323. https://doi.org/10.1016/j.eswa.2023.120323

Sha, L., Raković, M., Das, A., Gašević, D. and Chen, G. (2022) 'Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education', *IEEE Transactions on Learning Technologies*, Vol. 15, No. 4, pp. 481–492. https://doi.org/10.1109/tlt.2022.3196278

Siemens, G. (2019) 'Learning analytics and open, flexible, and distance learning', *Distance Education*, Vol. 40, No. 3, pp. 414–418. https://doi.org/10.1080/01587919.2019.1656153

Sirin, S. R. (2005) 'Socioeconomic status and academic achievement: A meta-analytic review of research', *Review of Educational Research*, Vol. 75, No. 3, pp. 417–453. https://doi.org/10.3102/00346543075003417

Sušnjak, T. (2024) 'Beyond Predictive Learning Analytics Modelling and onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT', *International Journal of Artificial Intelligence in Education*, Vol. 34, No. 2, pp. 452–482. https://doi.org/10.1007/s40593-023-00336-3

Talajić, M., Matijević, R. and Morić, Z. (2025) 'Enhancing academic performance prediction through machine learning in cloud environments', *Edelweiss Applied Science and Technology*., Vol. 9, No. 6, pp. 370–395. https://doi.org/10.55214/25768484.v9i6.7814

Umer, R., Sušnjak, T., Mathrani, A. and Suriadi, L. (2021) 'Current stance on predictive analytics in higher education: opportunities, challenges and future directions', *Interactive Learning Environments*, Vol. 31, No. 6, pp. 3503–3528. https://doi.org/10.1080/10494820.2021.1933542

Valdivia, A., Sánchez-Monedero, J. and Casillas, J. (2021) 'How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness', *International Journal of Intelligent Systems*, Vol. 36, No. 4, pp. 1619–1643. https://doi.org/10.1002/int.22354

Wang, Y. and Panicker, C. M. V. (2025) 'An Examination Regarding The Academic Performance Of University Students In Relation To Their Parents' Socioeconomic Status In China', *Frontiers in Health Informatics*, Vol. 13, No. 6, pp. 4527–4533. https://doi.org/10.63682/fhi2584

Yağcı, M. (2022) 'Educational data mining: prediction of students' academic performance using machine learning algorithms', *Smart Learning Environments*, Vol. 9, No. 1. https://doi.org/10.1186/s40561-022-00192-z

Zeineddine, H., Braendle, U. and Farah, A. (2020) 'Enhancing prediction of student success: Automated machine learning approach', *Computers & Electrical Engineering*, Vol. 89, p. 106903. https://doi.org/10.1016/j.compeleceng.2020.106903

Full research paper

UNPACKING THE BLACK BOX: A HYBRID XAI FRAMEWORK FOR AUTOGLUON-BASED MULTICLASS STUDENT OUTCOME PREDICTION

Marwan Nawae✉
Siripa Chankua
Massaya Longsaman

Faculty of Education and Liberal Arts,
Hatyai University, Thailand

✉ marwan.n@hu.ac.th

ABSTRACT

High student dropout rates remain a significant impediment to achieving the United Nations SDG 4 (equitable education). While Artificial Intelligence (AI) offers robust early risk prediction, the intrinsic black-box nature of high-performing models constrains their transparency. This study designs and investigates a multi-layered Explainable AI (XAI)-based assessment framework to generate actionable insights for student retention. We utilized AutoGluon to construct high-performing multiclass classification models (Graduated, Dropout, or Enrolled) on a higher education dataset. To address the complexity of the AutoGluon-generated models, we employed a hybrid XAI framework that couples global interpretability via a decision tree surrogate model and local interpretability via LIME (Local Interpretable Model-agnostic Explanations). The analysis revealed that models from the Boosting family, particularly XGBoost with bagging level 2, achieved the highest predictive performance (exceeding 0.890 across all metrics). The global analysis demonstrated that academic factors were the primary drivers of prediction, but critical socio-economic factors, such as Tuition fees, also exerted significant influence. Local LIME analysis provided granular, case-specific insights, strongly linking dropout status to first-year academic challenges and to features such as age at enrollment. This integrated XAI approach transforms complex models into an interpretable system, supporting student retention and educational equity (SDG 4).

KEYWORDS

AutoGluon, educational efficiency, Explainable AI (XAI), machine learning, SDG 4, student dropout prediction

HOW TO CITE

Nawae M., Chankua S., Longsaman M. (2026) 'Unpacking the Black Box: A Hybrid XAI Framework for AutoGluon-Based Multiclass Student Outcome Prediction', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 28–39. http://dx.doi.org/10.7160/eriesj.2026.190103

Article history
Received
October 31, 2025
Received in revised form
March 8, 2026
Accepted
March 9, 2026
Available on-line
March 31, 2026

Highlights

- *A hybrid XAI framework interprets complex AutoML for multiclass prediction.*
- *Boosting model (XGBoost) achieved the highest performance (Accuracy > 0.890).*
- *Global analysis confirms that academic and socio-economic factors are the primary drivers.*
- *Local analysis provides granular, case-specific insights into each class for intervention.*

INTRODUCTION

The pursuit of equitable and inclusive education directly addresses the United Nations Sustainable Development Goal 4 (SDG 4). SDG 4 mandates high-quality, inclusive, and equitable education, promoting lifelong learning opportunities for all. However, achieving this remains challenging due to persistent inequalities in tertiary education. According to the Education at a Glance 2025: OECD Indicators report (OECD, 2025), unequal opportunities significantly hinder the educational attainment of learners from disadvantaged backgrounds. On average across OECD countries, only 26% of young adults whose parents did not complete upper secondary education hold a tertiary qualification. Low completion rates compound this access gap. Newly collected data from over 30 countries show that only 43% of new bachelor's students graduate within the expected duration. Even with three additional years, the completion rate only reaches 70% (OECD, 2025). This high attrition rate disproportionately affects marginalized groups and acts as a major impediment to inclusivity. Beyond academic failure, the socioeconomic consequences are profound. Research indicates that individuals without a degree have a 15% to 20% lower lifetime earning potential than graduates (Krüger et al., 2023). Furthermore, high attrition rates lead to workforce instability and increase the risk of long-term unemployment and intergenerational poverty cycles (Bandala and Andrade, 2017). These economic impacts prove that student retention is not just an academic issue, but a socio-economic necessity.

In this study, educational efficiency is defined as the ability of an institution to use its limited resources, such as time, funding, and teaching staff, to achieve the best possible student outcomes. To ensure high effectiveness, it is necessary to provide an evaluation that offers actionable, interpretable, and timely feedback. This forms the basis for designing better student participation and learning outcomes (Nagy and Molontay, 2024). Despite the growth of educational technology, many current systems still fail to identify at-risk students early enough to help them. Most traditional assessment methods are summative and retrospective, meaning they only look at past performance, often when it is too late to intervene (Ifenthaler and Yau, 2020). There is a clear need for more proactive, efficient tools that provide timely, actionable feedback to keep students motivated (Aulck et al., 2016).

In recent years, Machine Learning (ML) has become a popular tool for predicting student failure. ML models can analyze many factors, such as grades, attendance, and socio-demographic data, to find students who might quit (Realinho et al., 2022; Zanellati et al., 2024). However, most of these models are black boxes. This means they provide a prediction but do not explain why a student is at risk. This lack of transparency creates a trust gap between teachers and students and undermines the moral integrity of AI in assessment, raising concerns about algorithmic bias. Without clear reasons, institutions cannot fulfill their responsibility or decide on the best support, and models might unfairly penalize students based on their background (Arrieta et al., 2020; ElShawi et al., 2021).

To solve this, the field of Explainable Artificial Intelligence (XAI) can be addressed. XAI methods, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), help turn complex predictions into human-understandable explanations (Padmasiri and Kasthuriarachchi, 2024). By using XAI, educational institutions can improve operational efficiency by focusing limited teaching resources on each student's specific needs rather than a one-size-fits-all approach. XAI provides an enhanced understanding of the complex variables influencing student performance, facilitating early intervention in high-risk cases. Moreover, XAI promotes algorithmic responsibility by allowing educators to detect and mitigate hidden biases (Guevara-Reyes et al., 2025).

This study aims to design and investigate a multi-layered explainable AI-based assessment framework. We are particularly interested in examining how interpretability feedback, derived from this framework, can generate actionable insights to enhance students' learning outcomes, improve retention, and encourage equitable access to quality education.

LITERATURE REVIEW

AI in student performance prediction

Academic achievement and student dropout prediction are important areas in educational data mining. Researchers have used a range of datasets, machine learning models, and explainability techniques to address these issues.

Berens et al. (2019) developed an early detection system (EDS) to identify at-risk students at German universities. Their model, which used administrative data, showed that grades from the first few semesters are strong predictors of a student's likelihood of graduating. However, they warned that these models show patterns but do not explain the exact causes of student dropout. By 2022, Realinho et al. (2022) introduced a large dataset of 4,424 students from a Portuguese school. This dataset includes information about students' families, grades, and the economy. It was designed to help researchers compare different models. Following this, Martins et al. (2023) used this data to make predictions at different times of year. They found that data collected by the end of the first semester gives the most accurate results. They also suggested that models should be adjusted to fit different academic programs. Building on these studies, Villar and de Andrade (2024) compared different algorithms using the same Portuguese data. They found that boosting models, such as LightGBM and CatBoost, worked much better than older methods. These advanced models achieved accuracy scores over 85%, while older methods reached only 70-75%. They also used SHAP to explain which factors were most important. However, they noted that because they used only one dataset, their findings might not apply to every school.

Collectively, these studies show that AI is improving at predicting student success through increasingly complex models. However, a major problem remains: these models are often hard to understand. We need better ways to explain why a specific student is at risk so teachers can provide the right support.

Explainable AI in Classification

In important areas like education, AI models must be accurate and easy to understand. XAI helps people understand how complex black-box models make decisions (Panda and Mahanta, 2023). One common way to do this is by using surrogate models. These are simpler models, such as decision trees, that mimic the logic of complex systems to make them easier for humans to understand (Falvo and Cannataro, 2024). This allows schools to check how the AI works and ensure it is fair. Another popular tool is LIME. LIME explains individual predictions by looking at specific cases (Ribeiro et al., 2016; Falvo and Cannataro, 2024). However, LIME can sometimes produce unstable results due to its data sampling strategy (ElShawi et al., 2021).

In education, combining surrogate models with LIME can provide both a clear overall view and personalized feedback for each student. This helps teachers and students trust the AI system more and ensures it is used ethically.

MATERIALS AND METHODS

This section describes the experimental pipeline, dataset, data preprocessing, model construction, evaluation measures, and the explainability techniques employed in this work, as shown in Figure 1.

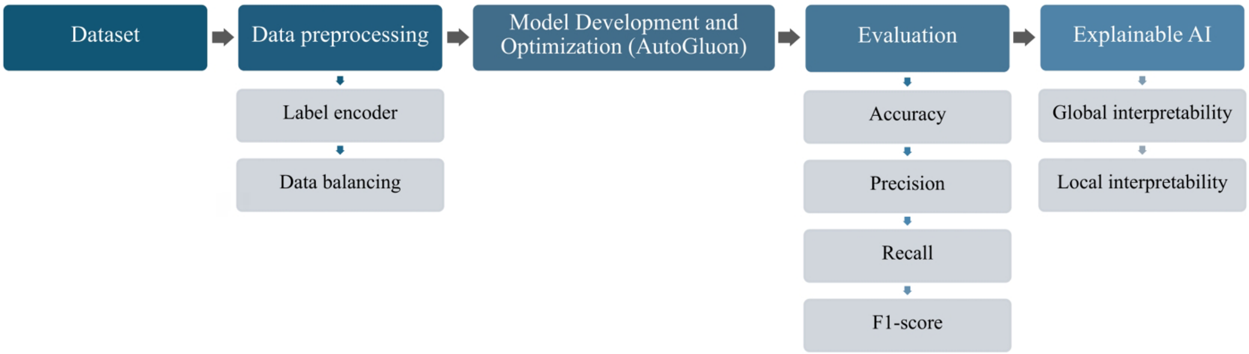


Figure 1: Workflow of the study, 2025 (source: own drawing)

Dataset

In this study, a publication dataset was used as the primary source of data for building machine learning models. The dataset is part of a learning analytics system used at the Polytechnic Institute of Portalegre. It has been used in previous studies to build machine learning models to forecast student academic achievement and dropout (Realinho et al., 2022). Data were acquired for the 2008/2009 to 2018/2019 cohorts. The dataset integrates information from multiple internal and external institutional systems, including the Academic Management System (AMS), Support System for the Teaching Activity (PAE), the General Directorate of Higher Education (DGES), and the Contemporary Portugal Database (PORDATA). The dataset includes 4,424 student records and 35 variables. It covers student demographics, socio-economic factors, and academic performance from the first two semesters. In this study, we use these data to predict three outcomes: Graduate, Dropout, or Enrolled. Table 1 summarizes the key features of the dataset. More specific details about the data characteristics and variables are available in the work of Realinho et al. (2022).

Category	Features
Personal & Demographic Data	Marital status Nationality Gender Age at enrollment International
Socio-Economic & Family Data	Mother’s qualification Father’s qualification Mother’s occupation Father’s occupation Tuition fees up to date Scholarship holder Debtor
Macro-Economic Indicators	Unemployment rate Inflation rate GDP
Academic & Application Factors	Application mode Application order Course Daytime/evening attendance Previous qualification Displaced Educational special needs
Academic Performance Metrics (1st Semester)	Curricular units 1st sem (credited) Curricular units 1st sem (enrolled) Curricular units 1st sem (evaluations) Curricular units 1st sem (approved) Curricular units 1st sem (grade) Curricular units 1st sem (without evaluations)
Academic Performance Metrics (2nd Semester)	Curricular units 2nd sem (credited) Curricular units 2nd sem (enrolled) Curricular units 2nd sem (evaluations) Curricular units 2nd sem (approved) Curricular units 2nd sem (grade) Curricular units 2nd sem (without evaluations)

Table 1: Summary of dataset features and categories, 2025 (source: Realinho et al., 2022)

Data preprocessing

We converted the three student status categorical features, graduate, dropout, and enrolled, into numerical labels (0, 1, and 2, respectively) with one-hot encoding as a preprocessing technique. The encoding is suitable for machine learning algorithms as it preserves the distinct identity of each class without implying an ordinal relationship among them (Panda and Mahanta, 2023).

To address the imbalance in the target class distribution, particularly the significantly lower numbers in the dropout class. We utilized the Synthetic Minority Over-sampling Technique (SMOTE). Previous experience with this dataset indicates that, although SMOTE and ADASYN perform similarly, SMOTE performs slightly better (Villar and de Andrade, 2024). SMOTE generates synthetic minority class examples by interpolating between close existing nearest-neighbor examples in the feature space. The technique is widely known to improve classifier performance, especially on educational datasets. With SMOTE used for class balancing, both classes now have an equal number of instances, as shown in Table 2. Particularly, each class now has 2209 samples, for a total of 6627 instances in the balanced dataset.

Label	Before SMOTE	After SMOTE
Graduate	2209	2209
Dropout	1421	2209
Enrolled	794	2209
Total	4424	6627

Table 2: Data distribution before and after SMOTE balancing, 2025 (source: own data)

Model development

For reliable model selection and optimization, we used AutoGluon, a state-of-the-art automated machine learning (AutoML) tool widely acclaimed for its cutting-edge performance on tabular data (Erickson et al., 2020). By automating the whole modeling pipeline, including diligent preprocessing, smart hyperparameter tuning, and advanced model stacking and ensembling, AutoGluon makes the construction of high-performing classification models both highly scalable and reproducible. Most importantly, this automation enables our study to bypass the tedious, often heuristic process of manual model selection and optimization. With AutoGluon's ability to reliably detect and build strong, often intricate, ensemble models, we ensure that the models analyzed by XAI are empirically the best possible black-box predictors. This provides a solid foundation for interpretability efforts, as the intrinsic complexity of these high-performing AutoML-generated models warrants the use of sophisticated post-hoc XAI tools. Although AutoGluon achieves state-of-the-art predictive accuracy, as an automated black box, it inherently obfuscates decision-making processes; a deficiency we directly overcome with our subsequent multi-XAI framework, which is specifically tailored to make these elaborate models transparent and interpretable.

To assess the predictive capability of the suggested model, we randomly split the dataset into a training set and a hold-out test set, with 80% of the data used for training and 20% for testing. This strategy aligns with best practices for moderately large datasets (about 4,500 instances), where computational efficiency must be traded off against robust performance estimation (Kuhn and Johnson, 2013).

AutoGluon was then fit to the specified 80% training split. In this process, it automatically performed end-to-end hyperparameter tuning, managed complex model ensembling, and conducted internal validation to select and configure the best model architecture. The ultimate, top-performing ensemble model, along with its optimized hyperparameters, was then evaluated only on the remaining, untouched 20% test set to provide an unbiased estimate of its capacity to generalize to new, unseen student data.

Evaluation metrics

To evaluate the model's performance in this multi-class classification, we employed four key metrics, consisting of accuracy, precision, recall, and F1-score, as given in Formulas 1–4 (Birchard et al., 2025). All these measures are based on the following fundamental counts for any class *k*.

True Positives ($TP_{k)}$: Samples assigned correctly as class *k*.

True Negatives (TN_k): Samples assigned correctly as not class *k*.

False Positives (FP_k): Samples that were incorrectly predicted to belong to class *k*.

False Negatives (FN_k): Samples incorrectly predicted as not of class *k*.

The metrics utilized are defined by:

Accuracy: This measure computes the ratio of correctly predicted instances to the total number of predictions, reflecting the overall model performance across all classes.

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN} \quad (1)$$

Precision: For a particular class, precision calculates the proportion of true positive predictions out of all instances predicted as positive for that class. It indicates the model's precision, or the quality of its positive predictions. In multiclass classification, we typically calculate Precision for each class and then average them.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: For a particular class, recall measures the proportion of true positive predictions among all actual positive cases. It is a measure of the model's completeness or its ability to include all positive, relevant examples. Similar to Precision, Recall is usually calculated for every class and then averaged.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score: The F1-score is the harmonic mean of precision and recall. It is a balanced measure that accounts for both false positives and false negatives and is especially useful for situations with imbalanced datasets. For multiclass problems, we compute the aggregate F1-score by averaging each class's F1-Score.

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Explainability framework

To address the complexity of machine learning black-box models, this work employs a robust explainability framework that provides both global and local interpretability by coupling a surrogate model with LIME. This framework allows stakeholders not only to understand the general patterns underlying predictions but also the specific reasoning behind each prediction. For this analysis, we selected the top-performing model from each major learning type in the AutoGluon.

Global interpretability

We used a decision tree classifier as a global surrogate model to explain the complex, black-box ensemble model generated by AutoGluon. This is based on the idea of model distillation, in which a simpler, more interpretable model is trained to mimic the decision-making structure of a more complex, opaque model (Guidotti et al., 2019). The first step was to generate a synthetic dataset for training the surrogate model. Specifically, for each instance in our original feature space, we obtained the predicted class label from the trained AutoGluon black-box model. This transformed dataset, which kept the original input features while using the black-box model's predictions as the new target variable, served as the training input for the decision tree.

Then the decision tree was trained to recursively split the feature space, developing a list of interpretable if-then rules. This model allowed the surrogate to capture the complex, often non-linear decision boundaries of the black-box model, clearly showing how interactions among features collectively affected the primary model's outputs. The training objective of this surrogate model was to minimize the variance between its own predictions and those of the original black-box model, typically quantified using a loss function such as cross-entropy. A key consideration in this surrogate modeling method is fidelity, which measures how well the interpretable model mimics the black-box model's predictions (Alangari et al., 2023). In classification problems, fidelity is defined as the surrogate model's prediction accuracy relative to the black-box model's output, not the true labels. High fidelity is essential; low fidelity indicates that the surrogate model does not accurately capture the black-box model's decision-making process, undermining the validity and trustworthiness of the derived explanations. When fidelity is convincingly high, the decision tree surrogate delivers a high-level, human-readable view of the prevailing patterns, feature importance, and interactive effects discovered by the complex ensemble model. This macro-level insight is especially useful for auditing, verifying model behavior, and informing governance in sensitive applications like educational policy (Arjunan, 2021).

Local interpretability

We enhanced the local interpretability of our black-box ensemble model by using LIME. LIME serves as a complementary technique to global explanation techniques by providing local, instance-specific explanations about the model's predictions (Ribeiro et al., 2016). This technique is particularly useful for local explanations, which are necessary for personalized interventions to improve student outcomes.

LIME uses a locally faithful surrogate model to predict a specific outcome. For example, LIME starts by generating a new dataset of perturbed samples around this instance. The perturbations are generated by slight changes to the original feature values, effectively creating synthetic data points that are theoretically close to the instance being analyzed.

For each newly generated perturbed instance, the original black-box model predictions are recorded. This process actually interacts with the black-box model to learn about its behavior in the local neighborhood of the target instance. Next, a simple and interpretable model, typically a linear regression or a shallow decision tree, is trained on these perturbed instances. Notably, each perturbed instance is weighted by its closeness to the original instance, ensuring the surrogate model prioritizes local accuracy. This weighting strategy ensures that the interpretable model captures the behavior of the black-box model correctly in the immediate context of the prediction (Ribeiro et al., 2016). The explanation obtained from this locally trained surrogate conveys feature weights or local rules that explain each input variable's contribution to the prediction of the specific case. This method is highly beneficial when decision-makers, such as university administrators or instructors, require quick, intuitive, and actionable explanations for individual student cases in real-time decision-making scenarios (ElShawi et al., 2021).

RESULTS

This section is structured into two parts. The first part presents the performance evaluation results of the models generated by the AutoGluon framework across various metrics. The second part provides explainability for the highest-performing black-box model of each model type at both global and local levels.

Model performance

As seen in Table 3, almost all models showed strong overall performance, with scores above 0.83 across all metrics. The top three models are XGBoost with bagging level 2 (referred to as XGBoost-B2), Weighted Ensemble with bagging level 3 (referred to as WeightedEnsemble-B3), and Light Gradient Boosting Machine with bagging level, achieving very similar results. They all reached an F1-score of 0.899. The XGBoost-B2 and WeightedEnsemble-B3 models recorded the highest precision (0.901), while all three top models had the same accuracy and recall scores at 0.899.

In the next group, the Neural Network model type also maintained high performance, with scores above 0.80 across all metrics. Neural Network implemented via FastAI with bagging level 2 (referred to as NeuralNetFastAI-B2) achieved high performance, with scores of 0.898, 0.899, 0.898, and 0.898 for accuracy, precision, recall, and the F1-score, respectively.

Conversely, models from the Random Forest and Extra Trees groups showed a decline in performance, with F1-scores falling in the range of 0.875 to 0.880. The Random Forest with bagging level 2 (referred to as RandomForestGini-B2) maintained the highest performance, scoring approximately 0.88 across all metrics. Lastly, the models with the lowest scores are below 0.80 but still above 0.72 for each metric.

Model	Accuracy	Precision	Recall	F1-score
XGBoost (bagging level 2)	0.899	0.901	0.899	0.899
WeightedEnsemble (bagging level 3)	0.899	0.901	0.899	0.899
LightGBM (bagging level 2)	0.899	0.900	0.899	0.899
NeuralNetFastAI (bagging level 2)	0.898	0.899	0.898	0.898
LightGBMXT (bagging level 2)	0.897	0.899	0.897	0.898
LightGBMLarge (bagging level 2)	0.893	0.896	0.893	0.893
CatBoost (bagging level 2)	0.893	0.895	0.893	0.893
RandomForestGini (bagging level 2)	0.881	0.882	0.881	0.880
RandomForestEntr (bagging level 2)	0.880	0.882	0.880	0.879
NeuralNetTorch (bagging level 2)	0.879	0.880	0.879	0.879
ExtraTreesGini (bagging level 2)	0.879	0.880	0.879	0.879
ExtraTreesEntr (bagging level 2)	0.876	0.877	0.876	0.875
LightGBMLarge (bagging level 1)	0.865	0.866	0.865	0.865
WeightedEnsemble (bagging level 2)	0.864	0.866	0.864	0.865
LightGBM (bagging level 1)	0.864	0.864	0.864	0.864
XGBoost (bagging level 1)	0.862	0.863	0.862	0.862
LightGBMXT (bagging level 1)	0.855	0.856	0.855	0.855
CatBoost (bagging level 1)	0.853	0.854	0.853	0.853
ExtraTreesEntr (bagging level 1)	0.850	0.853	0.850	0.850
RandomForestGini (bagging level 1)	0.848	0.851	0.848	0.848
ExtraTreesGini (bagging level 1)	0.847	0.850	0.847	0.848
CatBoost (bagging level 1)	0.845	0.847	0.845	0.845
RandomForestEntr (bagging level 1)	0.844	0.847	0.844	0.845
NeuralNetTorch (bagging level 1)	0.841	0.842	0.841	0.841
NeuralNetFastAI (bagging level 1)	0.833	0.834	0.833	0.834
KNeighborsDist (bagging level 1)	0.777	0.789	0.777	0.776
CatBoost (bagging level 1)	0.725	0.729	0.725	0.724
KneighborsUnif (bagging level 1)	0.722	0.729	0.722	0.722

Table 3: Predictive performance of all AutoGluon models on the test set, 2025 (source: own data)

Explainable AI

Based on the overall performance of all models, we selected the best model from three major learning categories (Boosting-based, Neural Network-based, and Tree-based) to perform the XAI analysis. The selected models are XGBoost-B2, NeuralNetFastAI-B2, and RandomForestGini-B2. The kNN-based models were excluded from this interpretability analysis because their performance scores fell below 0.8. The XAI results presented cover both global and local interpretability.

Global interpretability

As shown in Table 4, the decision tree surrogate model indicated that all three black-box models (XGBoost-B2, NeuralNetFastAI-B2, and RandomForestGini-B2) exhibited similar fidelity scores: 0.7773, 0.7689, and 0.7755, respectively.

This similarity suggests that the complex decision boundaries of the original models can be approximated by a simpler decision tree with comparable accuracy.

Furthermore, the global feature importance analysis derived from the surrogate decision tree model indicated that all three black-box models relied on the same top 10 features. These features, ranked in order of influence, are: curricular units 2nd sem (approved), curricular units 2nd sem (grade), tuition fees up to date, course, curricular units 1st sem (grade), age at enrollment, mother's occupation, unemployment rate, curricular units 2nd sem (evaluations), and GDP. Notably, the feature 'curricular units 2nd sem (approved)' was the most influential factor in the model's decision-making processes, clearly standing out from the other features (see Figures 3 and 4 in the Appendix).

Model	Fidelity
XGBoost (bagging level 2)	0.7773
NeuralNetFastAI (bagging level 2)	0.7689
RandomForestGini (bagging level 2)	0.7755

Table 4: Fidelity scores of surrogate decision trees approximating black-box models, 2025 (source: own data)

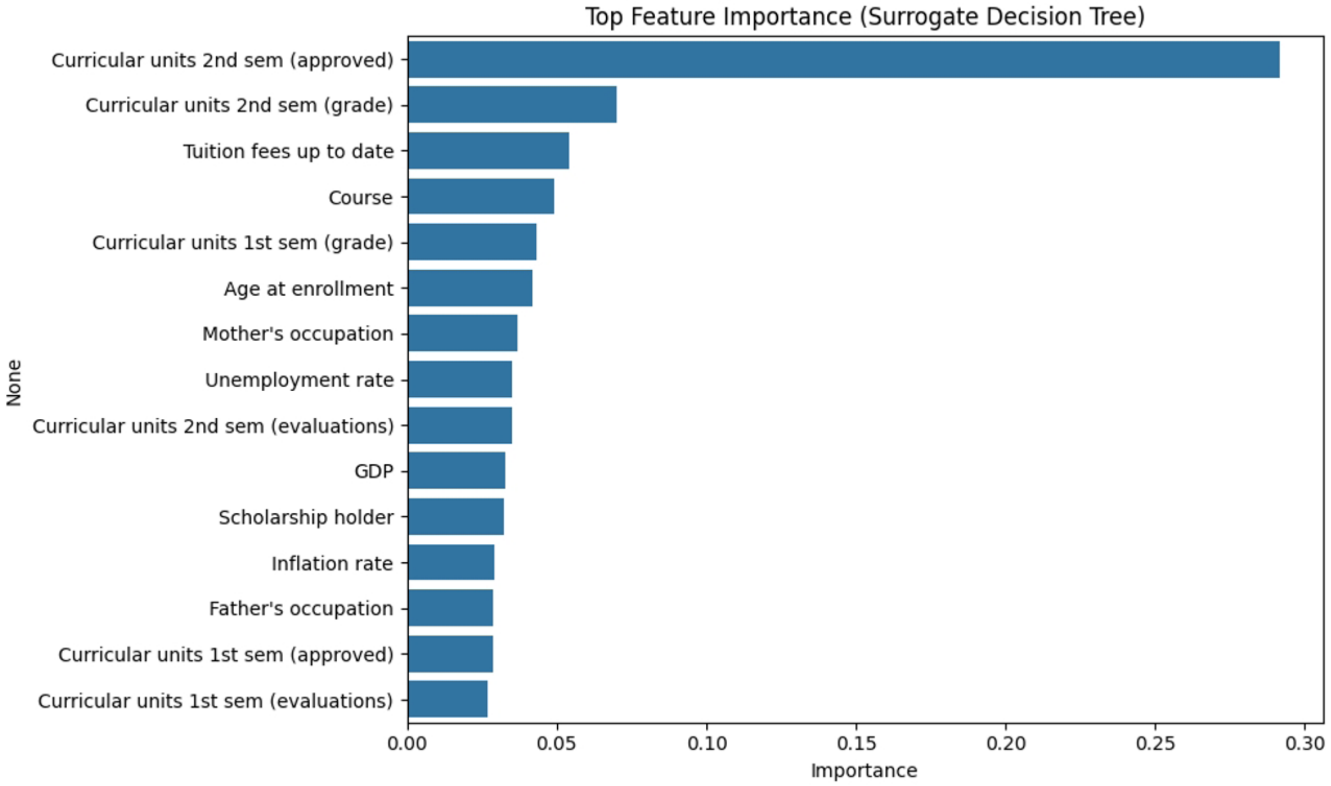


Figure 2: Top feature importance surrogate decision tree, 2025 (source: own data)

Local interpretability

Figure 4 (Appendix) displays the results of the sub-analysis for each classification class using LIME. Unlike the global XAI results, the local explanations showed greater diversity and complexity.

The XGBoost-B2 and NeuralNetFastAI-B2 models consistently achieved high confidence, with prediction probabilities exceeding 0.90 in nearly every instance across all three classes, especially in the graduate and dropout classes. The RandomForestGini-B2 model also performed well but generally showed lower certainty than the other two.

In terms of influential features for prediction, the three models largely shared similar feature sets across classes. The core features common across all three classes included curricular units 2nd sem (approved), curricular units 2nd sem (enrolled), curricular units 2nd sem (grade), curricular units 1st sem (approved), and curricular units 2nd sem (grade). However, specific classes involved additional or unique features: the dropout class also involved the additional features of the mother’s qualification and course. In contrast, the enrolled class included education, special needs, and course as additional influential features. These local differences highlight the unique, case-specific logic that each complex model employs when classifying individual data points.

DISCUSSION

Model performance

Based on the experimental results, the predictive performance of almost all models was highly consistent. This consistency is primarily due to the use of Bagging (Bootstrap Aggregating), an ensemble technique in the AutoGluon framework that enhances model stability and accuracy. Bagging works by training multiple base models on random subsets of the data and combining their outputs, effectively reducing prediction variance and preventing overfitting (Sisman et al., 2025).

Specifically, the Gradient Boosting Machines (GBMs) demonstrated superior performance in the dropout prediction task compared to other model categories. The XGBoost-B2 model highlighted the suitability of GBMs for this specific dataset. This finding aligns with the work of Sisman and team, who found that ensemble and boosting models yield strong results for tabular datasets (Sisman et al., 2025). This finding aligns with the core mechanism of the Boosting algorithm, which sequentially builds simple models (weak learners) to iteratively correct the errors of preceding models. This process enables GBMs to capture complex, non-linear relationships in the data (Nguyen and Ngo, 2025).

Concurrently, the NeuralNetFastAI-B2 model achieved performance metrics remarkably close to those of XGBoost-B2. This reflects a growing trend where deep learning architectures designed for tabular data are demonstrating competitive predictive power (Borisov et al., 2024). However, the marginal outperformance of GBMs in this study supports the consensus from several recent benchmark analyses that GBMs remain the state-of-the-art solution for most prediction problems involving tabular data (Grinsztajn et al., 2022).

In contrast, the models in the traditional Tree-based group, which showed reliable performance, exhibited lower overall efficacy when directly compared with the Boosting and Neural Network model groups.

Explainable AI

Global interpretability

When comparing the surrogate model fidelity scores for the three selected black-box models, the values were closely clustered around 0.77. It represents a significant reduction compared to the models' actual classification accuracy achieved by AutoGluon. The substantial gap between the high task accuracy and the lower explanation fidelity highlights the fundamental trade-off between model performance and interpretability. Since black-box models are highly complex, using a simpler decision tree as a surrogate imposes inherent limitations on the surrogate's predictive power. Consequently, as the original model's accuracy increases, its decision structure becomes more intricate, making it increasingly difficult for simple, interpretable models to mimic its behavior faithfully (Awad and Fraihat, 2023).

The analysis of features influencing dropout prediction across all three models revealed that the primary determinants are academic factors, such as Curricular units 2nd sem (approved), Curricular units 2nd sem (grade), Course, and Curricular units 1st sem (grade). This result confirms that academic performance is a central driver of the dropout problem, aligning strongly with previous research in educational data mining (Realinho et al., 2022; Olive et al., 2025).

In addition, socioeconomic factors were identified as highly influential, including Tuition fees up to date and the mother's occupation. The fact that Tuition fees up to date ranked as the third most important feature underscores the significant role of financial status in the model's prediction of student persistence (Olive et al., 2025).

As observed in the decision tree Surrogate model (see Figure 3 in the Appendix), the first partitioning of the data at the root node is defined by Curricular units 2nd sem (approved). Subsequent critical decision paths are governed by Curricular units 2nd sem (grade) and Tuition fees up to date. This hierarchical structure indicates that the model's classification logic systematically combines academic performance and financial status sequentially.

Local Interpretability

For the local interpretability analysis using LIME, we assessed the top features influencing individual predictions for the three best models across three distinct classes: graduate, dropout, and enrolled. While the specific ranking of features exhibits instance-by-instance volatility, the majority of the influential features remain consistent across all models and classes. This section details the key factors driving prediction for each status.

In predicting the graduate status, the models' decisions are predominantly influenced by positive academic performance indicators throughout the first and second semesters. These key factors include: number of curricular units registered, number of curricular units passed in each semester, and grade point average (GPA) for each semester. This result clearly indicates that a strong tendency toward graduation is strongly predicted by sustained academic achievement and consistency across all periods of study, a finding consistent with the previous work, which also identified educational factors as the primary determinant of graduation. (Apumayta et al., 2024; Villar and de Andrade, 2024). Additionally, factors related to age at enrollment within the normal, traditional range for higher education were found to positively influence the prediction of successful completion. Conversely, features such as international status and course type had minimal negative weight in the final prediction.

The prediction of dropout status is driven by a set of academic factors identical to those in the graduate class, but with an opposite directional influence. Specifically, the LIME explanations highlight that academic challenges in the first and second year (for example, failure to pass required credits and enrolled or low semester grades in either the 1st or 2nd semester) are strongly associated with a high likelihood of dropout. Furthermore, the course of study was found to be highly influential, with subjects perceived as complex or requiring a longer study duration significantly increasing the predicted probability of dropout. Intriguingly, socio-economic factors were also prominent, with mothers' occupation and qualifications contributing significantly to dropout prediction, suggesting that family background and economic stability are critical risk indicators (Apumayta et al., 2024). Moreover, age at enrollment outside the typical range for entry into higher education institutions was found to contribute positively to the dropout prediction.

The factors driving the enrolled status share similarities with the dropout factors but exhibit a unique pattern of challenges and resilience. The LIME analysis shows that negative socio-economic factors are influential in these instances, reflecting a context of adversity. However, this is critically counterbalanced by positive academic factors, which remain weighted in a positive direction. This suggests that the enrolled student group comprises individuals with the academic persistence and resilience to continue their studies and achieve academic success despite significant socio-economic barriers (Musaddiq et al., 2022).

Implications for Efficiency and Responsibility

Considering these specific predictors, the analysis presents significant implications for both efficiency and responsibility in education and science. Educational efficiency is directly enhanced because institutions can transition from generalized, resource-intensive support programs to highly targeted early interventions. By focusing on definitively identified academic predictors, such as second-semester grades and approved units, early detection systems optimize the allocation of limited institutional resources. This proactive shift reduces the waste of pedagogical efforts and ensures that support reaches students who need it most, thereby improving overall operational efficiency (Blašková and Staňková, 2023; Nagy and Molontay, 2024). Concurrently, identifying socio-economic vulnerabilities—such as the status of tuition fee payments—underscores deep institutional responsibility. By explicitly recognizing these systemic financial barriers, universities can implement equitable support mechanisms (Ferro and D'Elia, 2020). This ensures that predictive AI is used ethically as a supportive tool for disadvantaged students, rather than a black box that unfairly penalizes them based on their background (ElShawi et al., 2021). This dual approach actively aligns the XAI framework with the pursuit of sustainable, efficient, and responsible education, where data-driven insights serve both institutional performance and social justice.

However, this research faces limitations regarding the surrogate model's performance, which achieved a lower predictive score than the actual performance of the AutoGluon models.

Furthermore, this study is limited to a single open-source dataset from a single higher education institution in Portugal, thereby limiting the generalizability of the findings. Future work should therefore explore more complex surrogate modeling techniques to enhance explanation fidelity and validate the framework's generalizability by applying it to diverse, multi-institutional datasets from varied geographical regions. Crucially, the framework is adaptable for validation using an institution's internal, context-specific dataset to ensure its efficacy and relevance within a specific educational environment.

CONCLUSION

This paper successfully presents the development of student dropout prediction models using the AutoGluon framework. Further, it implements a multi-XAI approach to transform these top-performing black-box models into an interpretable system. The models from the Boosting family generally yielded the best results for the present dataset, with the XGBoost-B2 model achieving the highest overall performance metric for dropout prediction. The XGBoost-B2, NeuralNetFastAI-B2, and RandomForestGini-B2 represent the top-performing models for each learning type, with accuracies of 0.899, 0.898, and 0.881, respectively. This integrated XAI approach confirmed the fundamental trade-off between model accuracy and interpretability. Specifically, the fidelity of the Surrogate Model showed a notable decrease (approximately 0.77) when compared to the actual classification accuracy. The global analysis revealed that academic performance factors were the primary drivers of prediction, yet socio-economic factors, such as Tuition fees, also exerted significant influence. Furthermore, the LIME analysis provided granular, case-specific insights, indicating that both graduate and dropout statuses are strongly linked to academic performance challenges in the first year. Students who earn good grades across both semesters and consistently pass their required units have a high propensity to graduate. Crucially, socioeconomic factors also played a significant role in predicting dropout. Identifying key predictors can further assist in optimizing resource allocation and supporting vulnerable learners. This strategy enhances institutional accountability and aligns the analysis with SDG 4 objectives for global educational equity.

In future work, the prediction models will be refined and tailored to different academic programs within the institution to enhance accuracy and relevance to specific curricula. This targeted approach will enable the creation of actionable, timely feedback to support student retention and contribute to achieving the goal of educational equity (SDG 4).

REFERENCES

Alangari, N., El Bachir Menai, M., Mathkour, H. and Almosallam, I. (2023) 'Exploring Evaluation Methods for Interpretable Machine Learning: A Survey', *Information*, Vol. 14, No. 8, p. 469. https://doi.org/10.3390/info14080469

Apumayta, R. Q., Cayllahua, J. C., Pari, A. C., Choque, V. I., Valverde, J. C. C. and Ataypoma, D. H. (2024) 'University dropout: A systematic review of the main determinant factors (2020–2024)', *F1000Research*, Vol. 13, p. 253. https://doi.org/10.12688/f1000research.154263.2

Arjunan, G. (2021) 'Implementing Explainable AI in Healthcare: Techniques for Interpretable Machine Learning Models in Clinical Decision-Making', *International Journal of Scientific Research and Management (IJSRM)*, Vol. 9, No. 05, pp. 597–603. https://doi.org/10.18535/ijsrm/v9i05.ec03

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, Vol. 58, No. 1, pp. 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Aulck, L., Velagapudi, N., Blumenstock, J. and West, J. (2016) 'Predicting student dropout in higher education', *arXiv preprint*, arXiv:1606.06364. https://doi.org/10.48550/arXiv.1606.06364

Awad, M. and Fraihat, S. (2023) 'Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems', *Journal of Sensor and Actuator Networks*, Vol. 12, No. 5, p. 67. https://doi.org/10.3390/jsan12050067

Bandala, C. A. J. and Andrade, L. A. (2017) 'Education, Poverty and the Trap of Poor Countries in the Face of Development', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 10, No. 4, pp. 101–108. https://doi.org/10.7160/eriesj.2017.100402

Berens, J., Schneider, K., Gortz, S., Oster, S. and Burghoff, J. (2019) 'Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods', *Journal of Educational Data Mining*, Vol. 11, No. 3, pp. 1–41. https://doi.org/10.5281/ZENODO.3594771

Birchard, K., Boccia, C., Lounder, H., Colston-Nepali, L. and Friesen, V. (2025) 'Popfinder: A Highly Effective Artificial Neural Network Package for Genetic Population Assignment', *Molecular Ecology Resources*, Vol. 25, No. 1, p. e14096. https://doi.org/10.1111/1755-0998.14096

Blašková, V. and Staňková, M. (2023) 'Graduate Employability as a Key to the Efficiency of Tertiary Education', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 4, pp. 262–274. https://doi.org/10.7160/eriesj.2023.160401

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M. and Kasneci, G. (2024) 'Deep Neural Networks and Tabular Data: A Survey', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 6, pp. 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

ElShawi, R., Sherif, Y., Al-Mallah, M. and Sakr, S. (2021) 'Interpretability in healthcare: A comparative study of local machine learning interpretability techniques', *Computational Intelligence*, Vol. 37, No. 4, pp. 1633–1650. https://doi.org/10.1111/coin.12410

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M. and Smola, A. (2020) 'AutoGluon-Tabular: Robust and accurate AutoML for structured data', *arXiv preprint*, arXiv:2003.06505. https://doi.org/10.48550/arXiv.2003.06505

Falvo, F. R. and Cannataro, M. (2024) 'Explainability techniques for artificial intelligence models in medical diagnostic', in: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon: IEEE, pp. 6907–6913. https://doi.org/10.1109/BIBM62325.2024.10821826

Ferro, G. and D'Elia, V. (2020) 'Higher Education Efficiency Frontier Analysis: A Review of Variables to Consider', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 13, No. 3, pp. 140–153. https://doi.org/10.7160/eriesj.2020.130304

Grinsztajn, L., Oyallon, E. and Varoquaux, G. (2022) 'Why do tree-based models still outperform deep learning on typical tabular data?', in: *Advances in Neural Information Processing Systems (NeurIPS 2022)*, Vol. 35, pp. 507–520. https://doi.org/10.48550/arXiv.2207.08815

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F. and Villegas-Ch, W. (2025) 'Machine learning models for academic performance prediction: interpretability and application in educational decision-making', *Frontiers in Education*, Vol. 10, p. 1632315. https://doi.org/10.3389/feduc.2025.1632315

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2019) 'A Survey of Methods for Explaining Black Box Models', *ACM Computing Surveys*, Vol. 51, No. 5, pp. 1–42. https://doi.org/10.1145/3236009

Ifenthaler, D. and Yau, J. Y-K. (2020) 'Utilising learning analytics to support study success in higher education: a systematic review', *Educational Technology Research and Development*, Vol. 68, No. 4, pp. 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

Krüger, J. G. C., de Souza Britto Jr, A. and Barddal, J. P. (2023) 'An explainable machine learning approach for student dropout prediction', *Expert Systems with Applications*, Vol. 233, p. 120933. https://doi.org/10.1016/j.eswa.2023.120933

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*, New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3

Martins, M. V., Baptista, L., Machado, J. and Realinho, V. (2023) 'Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education', *Applied Sciences*, Vol. 13, No. 8, p. 4702. https://doi.org/10.3390/app13084702

Musaddiq, M. H., Sarfraz, M. S., Shafi, N., Maqsood, R., Azam, A. and Ahmad, M. (2022) 'Predicting the Impact of Academic Key Factors and Spatial Behaviors on Students' Performance', *Applied Sciences*, Vol. 12, No. 19, p. 10112. https://doi.org/10.3390/app121910112

Nagy, M. and Molontay, R. (2024) 'Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention', *International Journal of Artificial Intelligence in Education*, vol. 34, pp. 274–300. https://doi.org/10.1007/s40593-023-00331-8

Nguyen, N. and Ngo, D. (2025) 'Comparative analysis of boosting algorithms for predicting personal default', *Cogent Economics & Finance*, Vol. 13, No. 1, p. 2465971. https://doi.org/10.1080/23322039.2025.2465971

OECD (2025) *Education at a Glance 2025: OECD Indicators*, Paris: OECD Publishing. https://doi.org/10.1787/1c0d9c79-en

Olive, U., Bosco, M. and Enan, N. (2025) 'Predicting Student Dropout in Higher Education: An Ensemble Learning Approach with Feature Importance Analysis', *Journal of Information and Technology*, Vol. 5, No. 4, pp. 31–40. https://doi.org/10.70619/vol5iss4pp31-40

Padmasiri, P. and Kasthuriarachchi, S. (2024) 'Interpretable prediction of student dropout using explainable AI models', in: *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Colombo: IEEE, pp. 1–7. https://doi.org/10.1109/SCSE61872.2024.10550525

Panda, M. and Mahanta, S. R. (2023) 'Explainable artificial intelligence for healthcare applications using random forest classifier with LIME and SHAP', in: Balas, V. E., Kumar, R. and Srivastava, S. (eds.), *Explainable, Interpretable, and Transparent AI Systems*, Boca Raton: CRC Press, pp. 89–105. https://doi.org/10.1201/9781003442509-6

Realinho, V., Machado, J., Baptista, L. and Martins, M. V. (2022) 'Predicting Student Dropout and Academic Success', *Data*, Vol. 7, No. 11, p. 146. https://doi.org/10.3390/data7110146

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) '"Why should I trust you?": Explaining the predictions of any classifier', in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, San Francisco: ACM, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

Sisman, S., Kara, A. and Aydinoglu, A. C. (2025) 'Leveraging spatial data infrastructure for machine learning based building energy performance prediction', *PLOS One*, Vol. 20, No. 1, p. e0335531. https://doi.org/10.1371/journal.pone.0335531

Villar, A. and de Andrade, C. R. V. (2024) 'Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study', *Discovery Artificial Intelligence*, Vol. 4, No. 1, pp. 1–24. https://doi.org/10.1007/s44163-023-00079-z

Zanellati, A., Zingaro, S. P. and Gabbrielli, M. (2024) 'Balancing performance and explainability in academic dropout prediction', *IEEE Transactions on Learning Technologies*, Vol. 17, pp. 2086–2099. https://doi.org/10.1109/TLT.2024.3425959

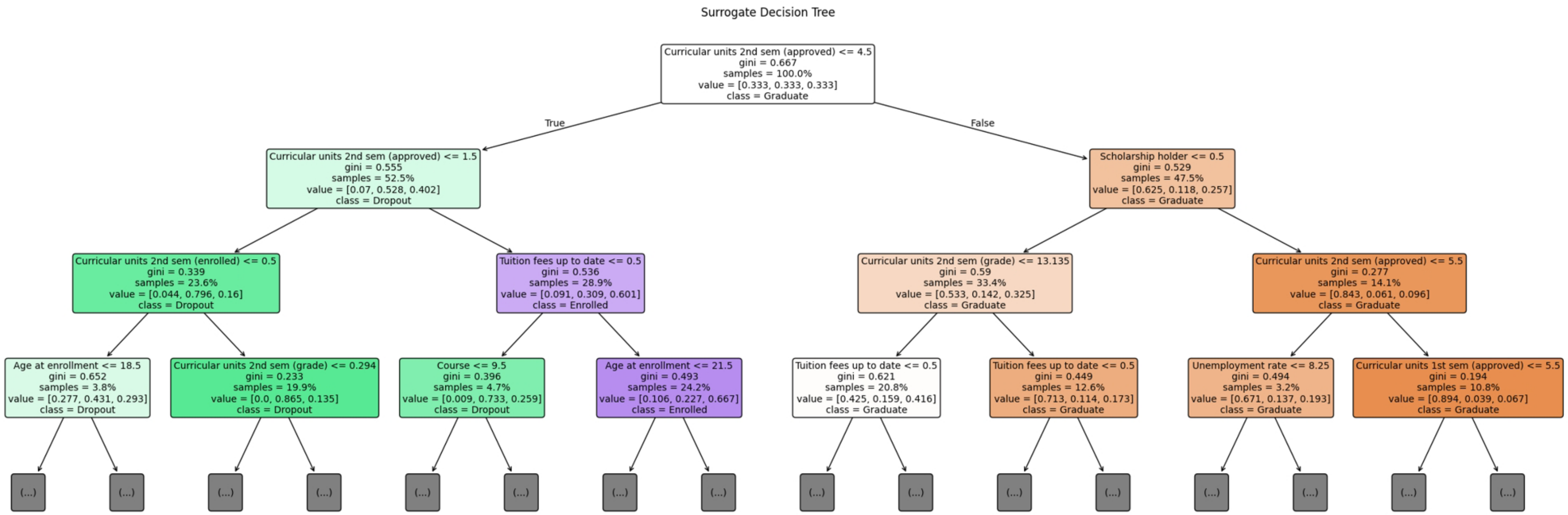
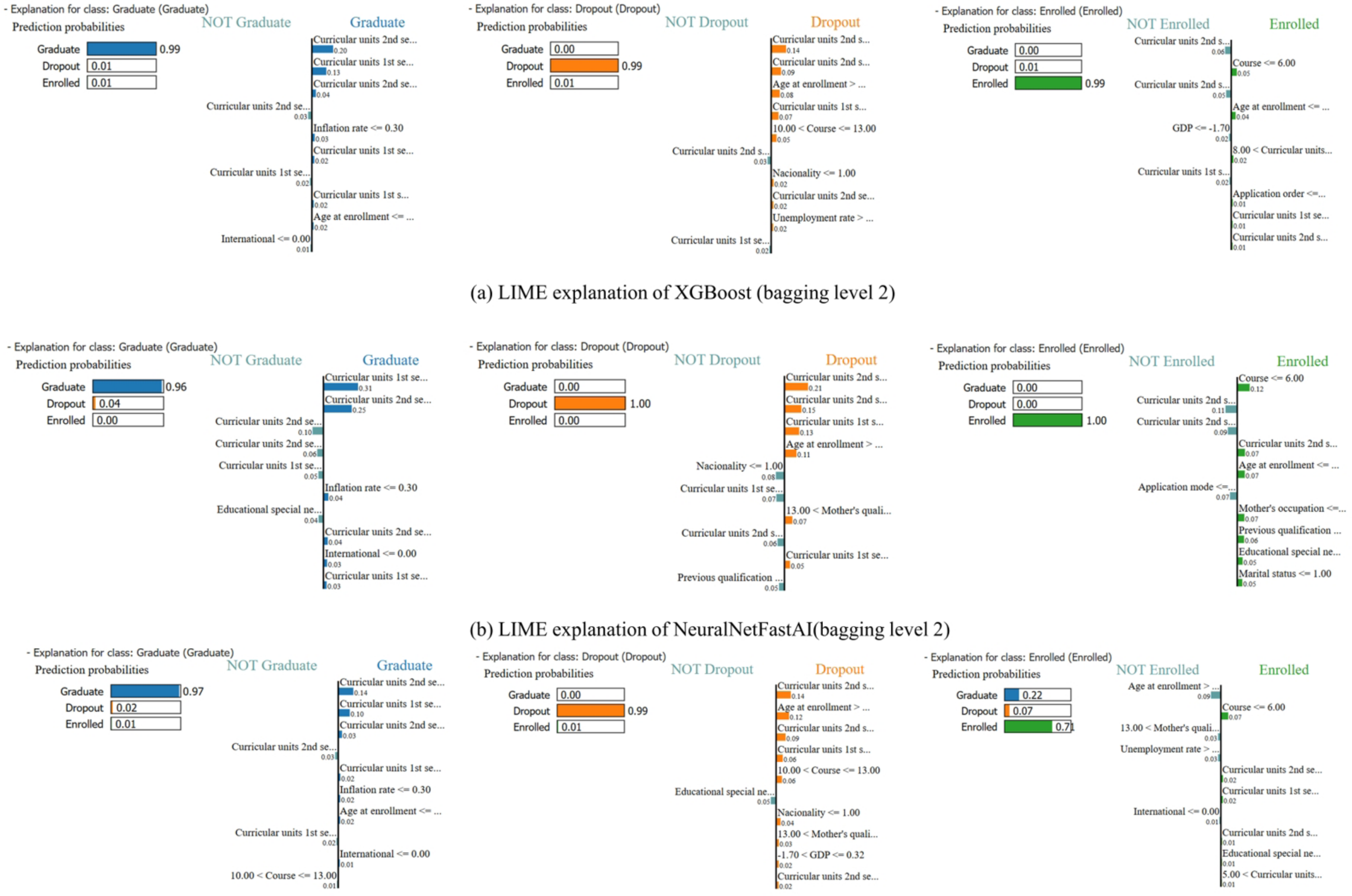

Surrogate Decision Tree
Curricular units 2nd sem (approved) <= 4.5
gini = 0.667
samples = 100.0%
value = [0.333, 0.333, 0.333]
class = Graduate
True
False
Curricular units 2nd sem (approved) <= 1.5
gini = 0.555
samples = 52.5%
value = [0.07, 0.528, 0.402]
class = Dropout
Scholarship holder <= 0.5
gini = 0.529
samples = 47.5%
value = [0.625, 0.118, 0.257]
class = Graduate
Curricular units 2nd sem (enrolled) <= 0.5
gini = 0.339
samples = 23.6%
value = [0.044, 0.796, 0.16]
class = Dropout
Tuition fees up to date <= 0.5
gini = 0.536
samples = 28.9%
value = [0.091, 0.309, 0.601]
class = Enrolled
Curricular units 2nd sem (grade) <= 13.135
gini = 0.59
samples = 33.4%
value = [0.533, 0.142, 0.325]
class = Graduate
Curricular units 2nd sem (approved) <= 5.5
gini = 0.277
samples = 14.1%
value = [0.843, 0.061, 0.096]
class = Graduate
Age at enrollment <= 18.5
gini = 0.652
samples = 3.8%
value = [0.277, 0.431, 0.293]
class = Dropout
Curricular units 2nd sem (grade) <= 0.294
gini = 0.233
samples = 19.9%
value = [0.0, 0.865, 0.135]
class = Dropout
Course <= 9.5
gini = 0.396
samples = 4.7%
value = [0.009, 0.733, 0.259]
class = Dropout
Age at enrollment <= 21.5
gini = 0.493
samples = 24.2%
value = [0.106, 0.227, 0.667]
class = Enrolled
Tuition fees up to date <= 0.5
gini = 0.621
samples = 20.8%
value = [0.425, 0.159, 0.416]
class = Graduate
Tuition fees up to date <= 0.5
gini = 0.449
samples = 12.6%
value = [0.713, 0.114, 0.173]
class = Graduate
Unemployment rate <= 8.25
gini = 0.494
samples = 3.2%
value = [0.671, 0.137, 0.193]
class = Graduate
Curricular units 1st sem (approved) <= 5.5
gini = 0.194
samples = 10.8%
value = [0.894, 0.039, 0.067]
class = Graduate
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Figure 3: Surrogate decision tree plot, 2025 (source: own data)



(a) LIME explanation of XGBoost (bagging level 2)

(b) LIME explanation of NeuralNetFastAI(bagging level 2)

(c) LIME explanation of RandomForestGini (bagging level 2)

Figure 4: LIME explanations in each class for XGBoost, NeuralNetFastAI, and RandomForestGini, 2025 (source: own data)

Full research paper

MACHINE LEARNING PREDICTIONS OF STUDENT OUTCOMES: THE ROLE OF EDUCATIONAL STRUCTURE AND SOCIAL STRESSORS IN CZECH MUNICIPALITIES

Martin Flegl[1]✉
Marketa Matulova[2]
Kristyna Vltavska[3]

[1]School of Engineering and Sciences, Tecnologico de Monterrey, Mexico

[2]Faculty of Economics and Administration, Masaryk University, Czech Republic

[3]Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic

✉ martin.flegl@tec.mx

ABSTRACT

Persistent disparities in student learning outcomes across Czech municipalities highlight the challenge of ensuring equitable access to quality education. These disparities are not only associated with demographic and economic conditions but also with the responsibility of municipalities and institutions to address structural inequalities. This study applies machine learning and SHAP analysis to predict student learning outcomes across municipalities with extended jurisdiction (MEJs), using demographic, economic, social, and housing indicators. Results highlight the dominant role of educational structure, with the share of people without secondary education and the proportion of younger adults holding college degrees emerging as the most influential predictors. Social and housing stressors, including parental executions, poverty destabilization, and housing allowances, further moderate outcomes, revealing nonlinear threshold effects that refine the explanatory narrative. The combined model achieved an R^2 of 0.629, confirming that while demographic and educational indicators explain most of the variance, contextual vulnerabilities add interpretive richness by identifying vulnerable subgroups. These findings underscore the dual influence of structural educational attainment and social stressors on student performance, while emphasizing educational responsibility as a key dimension in promoting equity and sustainable development.

KEYWORDS

Czech municipalities, educational responsibility, educational structure, machine learning, predictive analytics, social stressors

HOW TO CITE

Flegl M., Matulova M., Vltavska K. (2026) 'Machine Learning Predictions of Student Outcomes: The Role of Educational Structure and Social Stressors in Czech Municipalities', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 40–56. http://dx.doi.org/10.7160/eriesj.2026.190104

Article history
Received
February 11, 2026
Received in revised form
March 19, 2026
Accepted
March 30, 2026
Available on-line
March 31, 2026

Highlights

- *Demographic and educational structure strongly predict student learning outcomes.*
- *Social and housing stressors reveal vulnerable subgroups and nonlinear threshold effects.*
- *Combined model supports multi-level strategies for equity in Czech municipalities.*

INTRODUCTION

In the Czech Republic, schools have a strong tradition but face systemic challenges across all levels. Governance differs—municipalities oversee primary schools, regions manage secondary schools, and higher education institutions operate under the Higher Education Act (Eurydice, 2024)—yet common issues include chronic underfunding, high pupilteacher ratios, and pronounced social selectivity (OECD, 2020; OECD, 2023a). These factors contribute to persistent inequalities in educational outcomes and limit the system's capacity to adapt to demographic and social pressures.

Primary education is compulsory and widely accessible, with strengths such as school autonomy, decentralized governance, and curriculum modernization (MEYS, 2020). Czech pupils perform above international averages in Trends in International Mathematics and Science Study (TIMSS) 2023: fourth graders scored 511 in mathematics and 515 in science, while eighth graders achieved 520 and 518, respectively (European Commission, 2025). These results place Czech students among the stronger performers in the EU. Nevertheless, challenges remain, including teacher shortages, insufficient support for pupils with special needs, reliance on memorization, and heavy administrative burdens (OECD, 2025a; Mazouch and Fischer, 2024).

Secondary education also faces persistent difficulties. Social selectivity, outdated teaching methods, curricular overload, and the low prestige of vocational programmes continue to limit progress (European Commission, 2025). Regional inequalities exacerbate these issues: pupils in Karlovy Vary and Ústí nad Labem scored 20–25 points below the national mean in PISA 2022, placing them under the OECD average (OECD, 2023b). TIMSS results show similar gaps compared to peers in Prague and South Moravia (European Commission, 2025). Teacher quality and resources also reflect disparities, with 12% of teachers in Ústí nad Labem lacking full qualifications, compared to the national average of 6% (SGI, 2024), and pupil-teacher ratios exceeding 20:1, compared to the Czech average of 18:1 (OECD, 2023b).

Educational trajectories in disadvantaged regions are strongly shaped by social background and parental attainment. In Ústí nad Labem, over 60% of pupils enter vocational programmes compared to less than 40% in Prague, reinforcing inequalities and limiting mobility (European Commission, 2025). Strategy 2030+ seeks to reduce these disparities through targeted funding and stronger methodological support (MEYS, 2020). Persistent regional gaps highlight the challenge of ensuring equitable opportunities. Despite strong traditions and aboveaverage international performance, systemic weaknesses—social selectivity, uneven teacher quality, and regional underperformance—continue to shape student outcomes. Addressing these inequalities requires policies that account for demographic, social, and economic stressors to strengthen both academic achievement and civic participation.

At the tertiary level, Czech higher education offers tuitionfree study in Czechlanguage programmes and benefits from strong academic traditions. Yet participation rates remain relatively low, dropout levels high, and access continues to reflect social selectivity, shaped by parental resources and attainment (Hauschildt et al., 2024). Current reforms emphasize evaluating teaching quality and integrating practical experience into curricula and instruction (European Commission, 2025).

Research on these issues has so far been addressed mainly at the national level. For example, using PISA data, Simonová and Soukup (2013) examined how primary and secondary effects of social origin influence transitions to tertiary education. Primary effects generate class differences in academic achievement, while secondary effects shape educational choices and transitions regardless of prior performance. Šťastný (2021) analyzed the extent and characteristics of private tutoring among lowersecondary students, highlighting the factors driving its use across different educational tracks.

Building on the systemic challenges and regional disparities outlined above, this article identifies which social, demographic, and economic factors are associated with student outcomes. To achieve this, the analysis employs data from 206 municipalities with extended jurisdiction (MEJs) in the Czech Republic, enabling a macro-level territorial perspective that goes beyond individual or school-level studies. By integrating these variables into predictive models, the study highlights how local educational structures and contextual stressors predict performance. The novelty of this work lies in its territorial application. Rather than proposing a methodological breakthrough, it develops a predictive framework for Czech MEJs that informs policy interventions and institutional practices, ultimately supporting a more equitable and effective education system. Accordingly, this study addresses the following research question: Which social, demographic, and economic factors most strongly predict student learning outcomes across Czech municipalities?

PREDICTIVE ANALYTICS AND MACHINE LEARNING IN EDUCATION

Evolution of predictive analytics

Predictive analytics in education has progressed from simple descriptive statistics to advanced machine learning approaches. Early applications relied on basic statistical summaries that offered limited insight into the causes of student performance or its future trajectory (Romero and Ventura, 2010). With the expansion of educational data and the development of more sophisticated analytical techniques, predictive analytics has become a central tool for guiding educational practice and policy (Ferguson, 2012). Contemporary approaches employ algorithms that process large datasets to forecast outcomes, enabling proactive, personalized interventions (Deleña et al., 2025). These methods can identify students at risk before failure occurs, thereby improving success rates (Arnold and Pistilli, 2012) and enhancing efficiency by directing resources to areas of greatest need (Herodotou et al., 2019).

Applications in institutional and macro-level contexts

At the institutional level, numerous studies have demonstrated the potential of machine learning models to predict dropout and performance outcomes. For example, Khan et al. (2025) used a hybrid model combining Convolutional Neural Networks (CNNs) and Random Forests (RFs) with XGBoost to identify key predictive factors, including studied credits, number of previous attempts, entrance results, and geographical region. The analysis also classified students into three groups based on performance and background characteristics. Rabelo and Zárate (2025) proposed an ensemble model to improve dropout prediction by combining logistic regression, neural networks, and decision trees. Similarly, Chung and Lee (2019) developed an early warning system using a random forest model to predict high school student dropouts in Korea. Cheng et al. (2024) evaluated student academic performance using machine learning and metaheuristic algorithms, comparing five classification methods: Random Forest, Decision Tree, K-Nearest Neighbors, MLP, and XGBoost. Bravo Sanzana et al. (2015) applied classification and regression trees (CART) together with Random Forests to predict and characterize profiles of Chilean eighth-grade elementary students based on mathematics performance, using features related to individual attributes and family behavior.

Beyond institutional settings, research has increasingly shifted

Electronic ISSN 1803-1617 Printed ISSN 2336-2375

toward macro-level analyses to capture geographical and structural drivers of education and their broader economic effects. Bertoletti et al. (2022) examined how higher education systems influence regional economic development across 649 NUTS-3 regions in 29 European countries from 2014–2016, combining econometric and machine-learning approaches to capture nonlinear relationships. However, the capacity to leverage such data varies significantly across borders. Nouri et al. (2019) analyzed the state of learning analytics across seven European countries, revealing that, despite high levels of digitalization, unified national strategies for data-driven education are largely absent, resulting in fragmented implementation. Tsai and Gašević (2017) further argue that, without cohesive state-level policy frameworks, the ability to systematically apply predictive tools across regions remains limited due to privacy and ethical inconsistencies.

Current research landscape and gaps

The current landscape of predictive analytics in education is predominantly focused on the binary identification of at-risk students. According to a comprehensive systematic review by Umer et al. (2023), the majority of research in higher education utilizes classification tasks to predict student dropout or course failure. These models are essential for timely institutional intervention; however, they often lack the granularity needed to understand the full spectrum of student achievement. In contrast, studies employing machine learning to predict specific learning outcomes or final grades remain less common. For instance, Hussain et al. (2018) utilized regression models to forecast specific levels of students' engagement in virtual learning environment activities, while Conijn et al. (2017) demonstrated the utility of regression techniques in predicting final scores across diverse course types. Further expanding this scope, Asif et al. (2017) integrated pre-university data with early performance metrics to predict overall undergraduate success, and Kotsiantis (2012) used key demographic characteristics to predict students' marks.

While the literature is heavily weighted toward higher education due to data accessibility, predictive analytics in the K-12 (primary and secondary) sector is an area of growing importance. In their systematic review of 145 studies, Shafiq et al. (2022) found that although 46% of the research focused on undergraduate students, only 16% focused on school-level education. This disparity highlights a significant gap in the application of predictive modeling within primary and secondary settings, where early identification of academic risk is equally critical. Unlike higher education, where individual engagement metrics are paramount, K-12 models frequently highlight the profound impact of contextual and family-level stressors.

MATERIALS AND METHODS

Data

The analysis uses data on social conditions and the demographic and economic structure of 206 Czech municipalities with extended jurisdiction (MEJs). Municipalities with extended jurisdiction (so-called third-level municipalities in the Czech Republic, abbreviated ORP - Obec s rozšířenou působností) are an intermediate link in the delegation of self-government powers between regional authorities and other municipal authorities (the lower link is the authorized municipal authorities, and the lowest link is all other municipal authorities). Municipal authorities of MEJs thus have some additional areas of competence compared to other municipal authorities, not only for their own basic administrative district, but usually also for other municipalities in the surrounding area. The distribution of the 206 MEJs is shown in Figure 1.

Given the analysis's objective, the dependent variable is the testing results index for each municipality with extended jurisdiction. This index is calculated from the results of students in the Czech School Inspectorate (Česká školní inspekce) testing at the 5th and 9th grades, as well as outcomes from the unified entrance examination[1]. Figure 1 illustrates the distribution of testing results in 2025. Information on testing results, social conditions, and the demographic and economic structure of each MEJ was obtained from the DataPAQ regional data viewer tool (https://www.datapaq.cz), which systematically collects and integrates data on education and social conditions across the Czech Republic.

The social conditions component of the analysis consists of 49 indicators, grouped into the following categories: Housing shortage (5 indicators), Executions (7), Unemployment (2), Social exclusion (1), State social support (14), Social support – Help in material need (13), and Social support – Other allowances (4). The demographic and economic structure component includes 85 indicators, divided into the following: Population and municipalities (25), Population movement (18), Educational structure (19), Labor market (13), and Commuting (10). Tables 1 and 2 in the Appendix summarize all included indicators, along with brief explanations for each.

The Index of testing results corresponds to 2025, whereas the other indicators capture the situation between 2021 and 2024, depending on data availability. In all cases, the most recent published data was used to minimize the temporal gap with the testing results. Due to the testing results index methodology, data were unavailable for 23 MEJs (11.17% of the sample). In these cases, missing values were imputed using the median. Similarly, 13 missing cases (6.31%) for the indicator SC-HS-Children_in_housing_shortage were imputed using the median.

[1] The index of testing results is composed of 3 input indicators: the share of 9th-grade students who took the unified entrance exam and placed in the top fifth of those tested, and 2 summary indicators describing the results of a sample survey of 5th- and 9th-grade elementary school students. All three input indicators are strongly correlated with each other (correlation of 0.7 or higher) and can therefore be combined into a single index while retaining a large portion of the total variance. For each MEJ, these three indicators are weighted differently in the index. The weight of the unified entrance exam results is the same for all MEJs and is the output of the principal components analysis over the 3 input indicators. The weight of the 2 summary indicators is then calculated for each MEJ based on the ratio of the number of tested students in 5th and 9th grades. The index is not calculated for MEJs with fewer than 250 tested students, with CSI testing participation of less than 50% of all students in the given classes, and for whom the CSI testing results do not correspond to the results of the unified entrance exam. Please consult PAQresearch (2026) for further details.

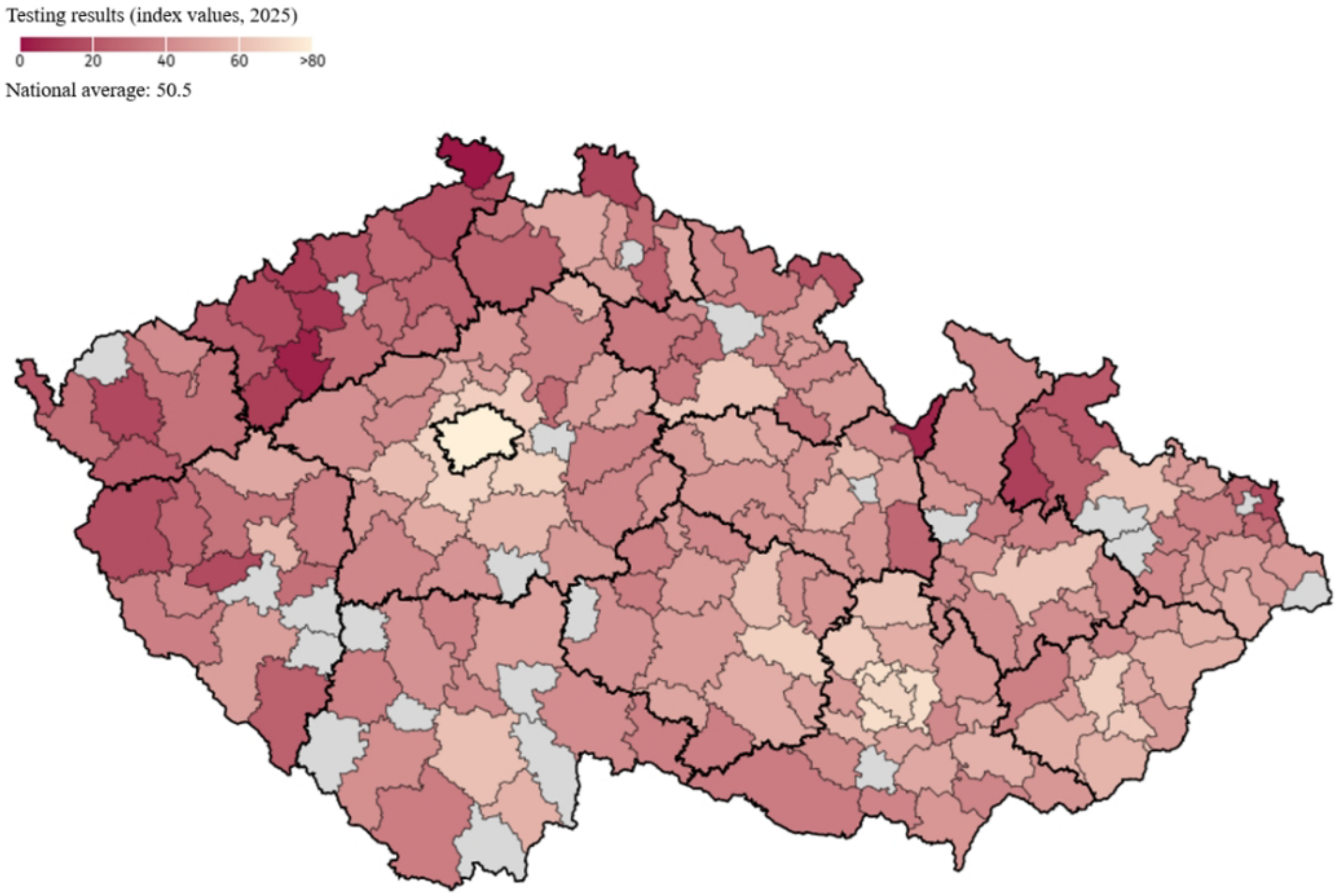


Figure 1: Czech MEJs and their corresponding testing results in 2025 (source: elaborated using DataPAQ data).

Random forest

Random Forest is an ensemble learning method introduced by Breiman (2001) that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. The algorithm works by creating a large number of decision trees, each trained on a bootstrapped sample of the data. At each split in a tree, only a random subset of features is considered, which introduces additional randomness and helps to decorrelate the trees (Zuluaga et al., 2023). For regression tasks, the final prediction is the average of all trees' outputs, whereas for classification tasks, the majority vote determines the outcome.

The strength of Random Forest lies in its ability to handle complex, nonlinear relationships and variable interactions without requiring extensive preprocessing (Bertoletti et al., 2022; Jafari et al., 2025; Jiang et al., 2024). It is robust to noise, can manage large datasets with mixed data types, and provides measures of feature importance that allow researchers to interpret which variables contribute most to predictions. One of the biggest advantages of Random Forest is its versatility, as it can be used for regression and classification tasks and for assessing the relative importance of input features (Cheng et al., 2024). These characteristics make it particularly useful in applied fields such as education, economics, medicine, and environmental science, where data often exhibit nonlinear patterns and multicollinearity.

Model training and evaluation

The dataset was split into training (80%) and test (20%) subsets, with the random seed set to 42 to ensure reproducibility. Model performance was evaluated using 5fold crossvalidation with shuffling. This means the dataset was divided into five equal parts (folds) and the model was trained and validated five times, each time using a different fold as the validation set and the remaining folds as the training set. Before splitting into folds, the data was shuffled randomly to prevent bias that could arise if the dataset had an inherent order (e.g., sorted by time or grouped by category). Crossvalidation process is widely recognized as a robust method for estimating generalization performance in machine learning (Arlot and Celisse, 2010; Browne, 2000).

A RandomForestRegressor with 200 trees (n_estimators = 200) was trained using scikitlearn. Default hyperparameters were retained unless otherwise specified, and random seeds were fixed to control stochastic variation. Performance metrics included Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 (coefficient of determination). These were reported as crossvalidated scores to avoid insample bias. This procedure helped to control overfitting, which occurs when a model learns patterns that are too specific to the training data and fails to generalize to unseen data. Overfitting is a common challenge in machine learning, particularly when the number of predictors is large relative to the sample size (Hawkins, 2004; Zhang and Yang, 2017). By averaging performance across folds, cross-validation provides a more robust estimate of predictive accuracy, especially given the relatively small sample size (206 MEJs) and the large number of predictors.

SHapley Additive exPlanations (SHAP)

Despite its advantages, Random Forest is less interpretable than linear models because it does not produce straightforward coefficients. However, interpretability can be enhanced through methods such as SHAP, which explain each feature's contribution to individual predictions. SHAP is a unified framework for interpreting machine learning models that builds on cooperative game theory. It extends the concept of Shapley values, originally developed by Shapley (1953), which provides a fair distribution of contributions among players in a coalition. In predictive modeling, SHAP assigns each feature a value representing its contribution to a particular prediction, thereby offering a consistent and accurate explanation of model outputs (Lundberg and Lee, 2017).

Unlike traditional importance measures, SHAP explains individual predictions by quantifying how much each variable increases or decreases the predicted outcome. This makes it particularly valuable for complex, nonlinear models such as Random Forests, Gradient Boosting Machines, and Neural Networks, where interpretability is often limited (Bertoletti et al., 2022; Jafari et al., 2025). Recent work has demonstrated SHAP's ability to move from local explanations to global understanding, enabling researchers to identify both individual-level drivers and broader structural patterns in data (Lundberg et al., 2020). Its popularity also stems from its model-agnostic nature and intuitive visualizations, which make it accessible across disciplines (Molnar, 2025).

RESULTS

This section is organized into three parts. First, we examine the impact of social condition indicators on student testing results, with particular attention to the most influential variables. Second, we evaluate the role of demographic and economic structure indicators across the analyzed MEJs, highlighting those that exert the strongest influence on outcomes. Third, we analyze the combined model that incorporates both groups of indicators and describe how its performance differs from the individual models. For each of the three models, performance is assessed using cross-validated RMSE, MAE, and R^2 scores, ensuring that estimates reflect out-of-sample prediction rather than in-sample bias. This approach provides a consistent and reliable measure of predictive accuracy across all stages of the analysis.

Social conditions indicators

The RF model trained on social condition indicators achieved an RMSE of 9.191, an MAE of 6.886, and an R^2 of 0.522. These values suggest that, while the model captures a significant portion of the variability in student testing outcomes, its explanatory power is rather moderate. The R^2 indicates that social condition indicators alone explain approximately half of the variance in the student testing results. Such a result underscores both their relevance and also highlights their limitations as sole predictors. The relatively high error metrics RMSE and MAE reflect the complexity and heterogeneity of social conditions, which may interact with other factors not included in the analysis. Still, the analysis reveals that indicators such as executions, poverty destabilization, and housing conditions manifest measurable effects on educational outcomes. In general, the results obtained provide important contextual insights into the environments in which students learn.

The analysis indicates that executions were the most influential predictors of the student testing outcomes (Figure 2). Parents in multiple executions (*SCEXParentsmulti_ex*), estimated as the number of people aged 30-49 with 2 or more executions, emerged as the single strongest variable (importance 0.222). This means that MEJs experiencing multiple parental executions face heightened educational vulnerability, as such financial instability likely disrupts household environments and student learning conditions. In relation to this, Parents in execution (*SC-EX-Parents_execution*, 0.093), Juveniles executions (*SC-EX-Juvenils_execution*, 0.045), the proportion of people aged 15 to 29 with at least one foreclosure, out of all people in that age group, and People in execution (*SC-EX-People-execution*, 0.025), Share of people with at least one execution (including children) out of all residents, also appeared within the most influential indicators.

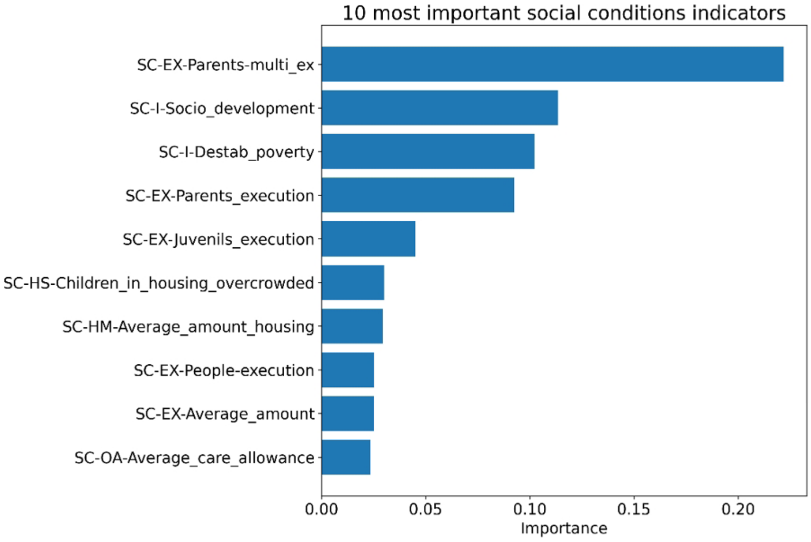


Figure 2: 10 most important social conditions indicators on students' testing outcomes.

The second most important factor was the socio-economic development of a region (*SC-I-Socio_development*). This index consists of the share of people with higher education, employment in the region, and the share of employees in the highest wage quantile, with an importance of 0.114. We can consider this factor as broader community-level resources and opportunities that shape educational attainment. The third most important factor was Destabilizing poverty (*SCIDestab_poverty*, 0.102), an index that primarily includes executions, housing shortages, and socially excluded locations. The destabilizing poverty affects families and children directly because it is related to stress and insecurity, housing losses, the breakdown of social ties and aspirations, and the absence of positive role models.

Although their relative importance was lower, housing-related indicators, such as overcrowding (*SC-HS-Children_in_housing_overcrowded*, 0.030) and the average amount of care allowance (*SC-OA-Average_care_allowance*, 0.023), contributed consistently across the model. These variables highlight the role of living conditions in shaping educational outcomes, where overcrowded households may limit study space and concentration, and inadequate housing support may exacerbate stressors.

Figure 3 displays the 10 most influential social condition factors using SHAP analysis. In these plots, the horizontal axis shows the values of an independent variable. In contrast, the vertical axis shows the SHAP values indicating its positive or negative effect on predicted test results. The color of each observation reflects the relative value of the independent variable (as shown on the right color scale), illustrating how low versus high values influence the prediction. Dependence plots revealed clear nonlinear threshold effects, underscoring the complex ways in which social stressors shape educational outcomes. For example, MEJs with elevated execution rates exhibited relatively stable predictions up to a certain point (between 8% and 10% in the case of *SCEXParentsmulti_ex*). However, once critical thresholds were exceeded, sharp declines in student performance became evident. This pattern suggests that financial instability may exert compounding effects, where moderate levels of stress can be absorbed, but extreme conditions trigger rapid deterioration in outcomes.

Similarly, Socio-economic development of a region (*SCISocio_development*) and Destabilizing poverty (*SCIDestab_poverty*) showed nonlinear impacts with specific thresholds. In the case of socio-economic development, once the index level exceeds 55, a significant impact on student performance can be observed. Destabilizing poverty shows an opposite pattern: student performance decreases linearly until a threshold of around 60, after which a significant drop is clearly observable.

The last observation is linked to the average amount of housing allowance (*SCHMAverage_amount_housing*), which represents the average value of a single housing supplement. This allowance is provided to households in material need whose income (including the allowance itself) is insufficient to cover housing costs and basic living expenses. The analysis revealed a largely linear relationship: higher levels of housing support were associated with improved student learning outcomes, with the effect's slope remaining relatively stable.

When the predictor Children in housing need in precarious housing (*SCHSChildren_in_housing_preca*) is considered, however, the relationship shifts to a nonlinear trend. Children in precarious housing constitute a particularly vulnerable group, as their families cannot secure adequate housing without assistance and often live in unstable or inadequate conditions. Such circumstances frequently generate elevated household stress, limited space or quiet for study, and recurrent disruptions due to moves or insecure tenancy. The interaction analysis shows that when housing supplements are below 5,000 CZK, testing results decline sharply among children in precarious housing. Conversely, when supplements exceed 5,000 CZK, student outcomes improve significantly, in some cases surpassing those of peers in less precarious housing situations. This finding suggests that sufficient housing support can mitigate, and even reverse, the educational disadvantages associated with precarious living conditions.

Nevertheless, this effect must also account for whether children live in overcrowded housing (*SC-HS-Children_in_housing_overcrowded*). A negative nonlinear effect was observed among children in housing need residing in overcrowded apartments, where the compounded stress of inadequate space and precarious living conditions further undermines educational performance. This finding emphasizes that while sufficient housing support can mitigate disadvantages, its effectiveness is constrained when combined with severe housing overcrowding.

Demographic and economic structure

The predictive block based on demographic and economic indicators achieved an RMSE of 8.087, an MAE of 6.130, and an R^2 of 0.633. These values indicate a moderate level of explanatory power: while the model captures a substantial portion of the variance in testing outcomes, residual error remains non-negligible. The relatively balanced RMSE and MAE suggest that extreme outliers do not dominate prediction errors; rather, they reflect consistent deviations across municipalities. Further, the demographic and economic structure predictors themselves show higher predictive power compared to social conditions ($R^2 = 0.522$). The most influential predictors within this block were the Share of people without secondary education (*DEESPeople_without_sec_education*, 0.123), the Share of people without secondary education in the 40–44 age group (*DEESPeople_without_sec_education_4044*, 0.089), and the Share of people with a college degree in the 30–34 age group (*DEESPeople_with_college_degree_3034*, 0.088). As illustrated in Figure 4, the strongest predictors are consistently linked to the educational structure of the population, which outweighs variables related to population distribution, migration, labor market participation, and commuting patterns.

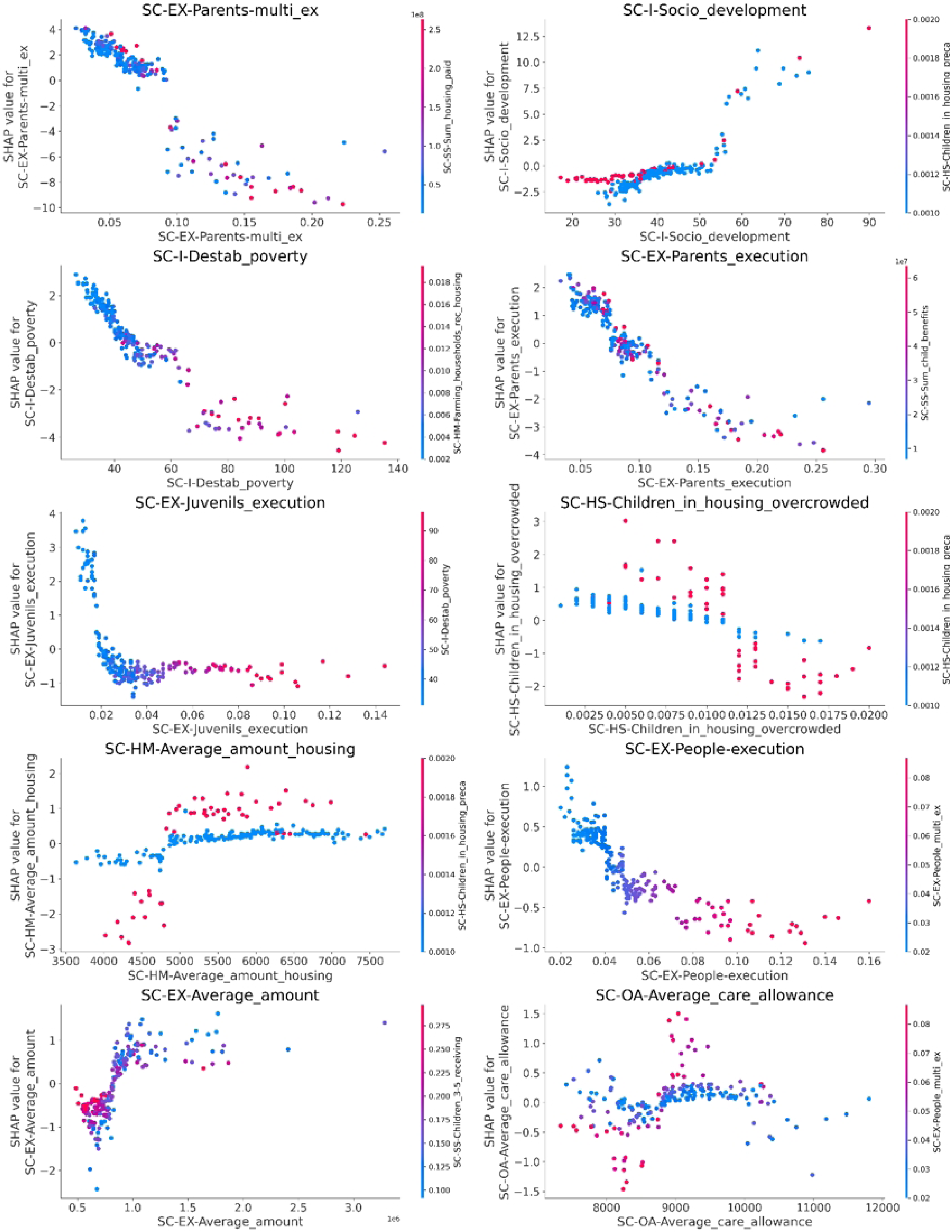


Figure 3: SHAP dependence of the top 10 most influential social conditions indicators

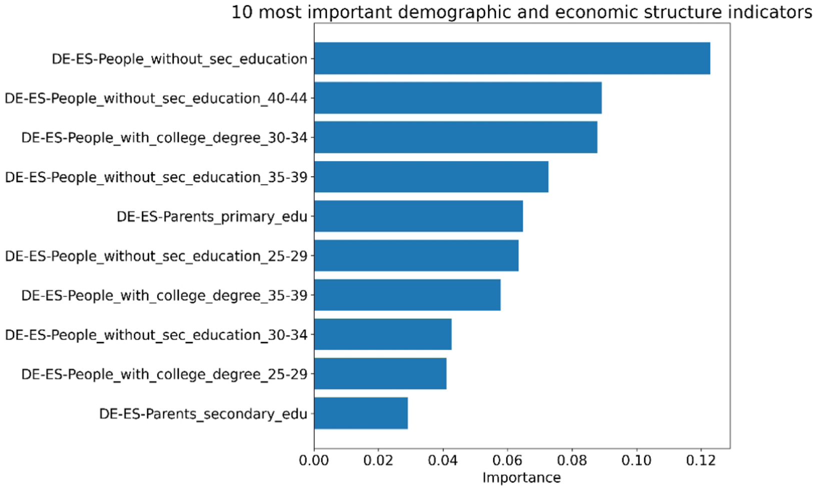


Figure 4: 10 most important demographics and economic structure indicators on students' testing outcomes.

The SHAP analysis (Figure 5) further revealed nonlinear relationships, indicating threshold effects rather than simple linear trends. As expected, higher proportions of the population with completed education at specific levels exert a positive influence on test outcomes. In contrast, incomplete or missing education is associated with lower test scores. In addition, testing results appear to be moderated not only by parental education level but also by the prevailing educational composition within each municipality. In general, the higher the educational level within a given MEJ, the stronger the positive effect on student testing outcomes. This dual influence underscores the importance of both family background and community-level educational attainment in shaping learning performance, suggesting that individual disadvantages may be amplified or mitigated depending on the broader educational environment.

Combined model

The integrated model, which incorporates demographic, economic, social, and housing indicators, achieved an RMSE of 8.133, an MAE of 6.147, and an R^2 of 0.629. These values indicate that the combined specification explains a substantial proportion of the variance in testing outcomes, though predictive accuracy remains comparable to the individual blocks. The relatively stable RMSE and MAE suggest that errors are evenly distributed across municipalities rather than driven by extreme outliers.

Interestingly, the explanatory power of the combined model does not markedly exceed that of the demographic and economic block alone ($R^2 = 0.633$), implying that while social and housing stressors add interpretive richness, their incremental predictive contribution is modest. This finding highlights the complexity of educational outcomes: structural demographic and educational composition variables capture much of the variance, while contextual stressors such as foreclosure, poverty destabilization, and housing conditions refine the narrative by identifying vulnerable subgroups and nonlinear threshold effects.

The combined model confirms that the educational structure of the population remains the dominant driver of student testing outcomes (Figure 6). Indicators such as the share of people without secondary education (*DE-ES-People_without_sec_education*) and the proportion of younger adults with college degrees (*DE-ES-People_with_college_degree_30-34*) consistently rank at the top of the importance list, underscoring the central role of both parental and community-level educational attainment. These findings highlight that municipalities with stronger educational composition provide a more supportive environment for student achievement, while gaps in secondary education exert a persistent negative influence.

At the same time, social stressors and housing conditions—such as parental executions (*SC-EX-Parents-multi_ex and SC-EX-Parents_execution*), poverty destabilization (*SC-I-Destab_poverty*), and average amount of housing supplement paid (*SC-HM-Average_amount_housing*)—appear as secondary but meaningful predictors. Their presence in the top 20 features suggests that while demographic and educational variables explain most of the variance, contextual vulnerabilities moderate outcomes in important ways. Taken together, the combined model illustrates that student performance is shaped by a dual structure: the foundational impact of educational attainment and the amplifying or buffering effects of social and housing stressors.

Considering the SHAP analysis (Figure 7 - see next page), one notable difference compared to the separate models is the predictor Share of people without secondary education in the 30–34 age group (*DE-ES-People_without_sec_education_30-34*), which negatively influences testing outcomes. However, this effect is attenuated in MEJs with a high Share of people working in advanced services (*DE-LM-People_working_advanced_services*), including information and communication activities, finance and insurance, and real estate. When the share of employment in these advanced services declines, the negative impact of missing secondary education becomes more pronounced.

A similar moderating pattern is observed with the Occurrence of juveniles in execution (*SC-EX-Juvenils_execution*), which constrains the otherwise positive effect of a higher Share of young people with a college degree (*DE-ES-People_with_college_degree_25-29*). This interaction highlights how social stressors can diminish the benefits of educational attainment, reinforcing the importance of considering both structural and contextual factors in explaining student performance.

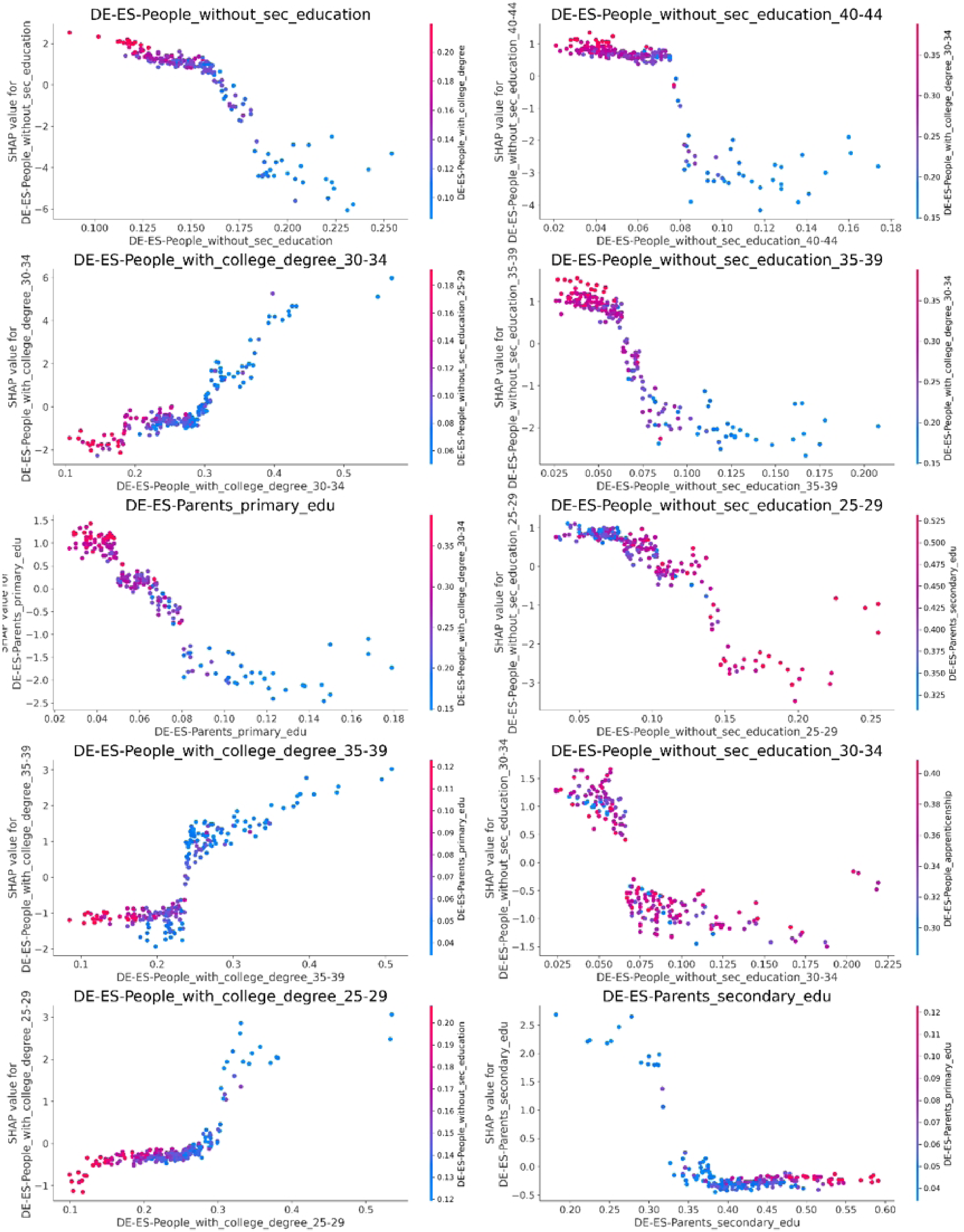


Figure 5: SHAP dependence of the top 10 most influential demographic and economic structure indicators

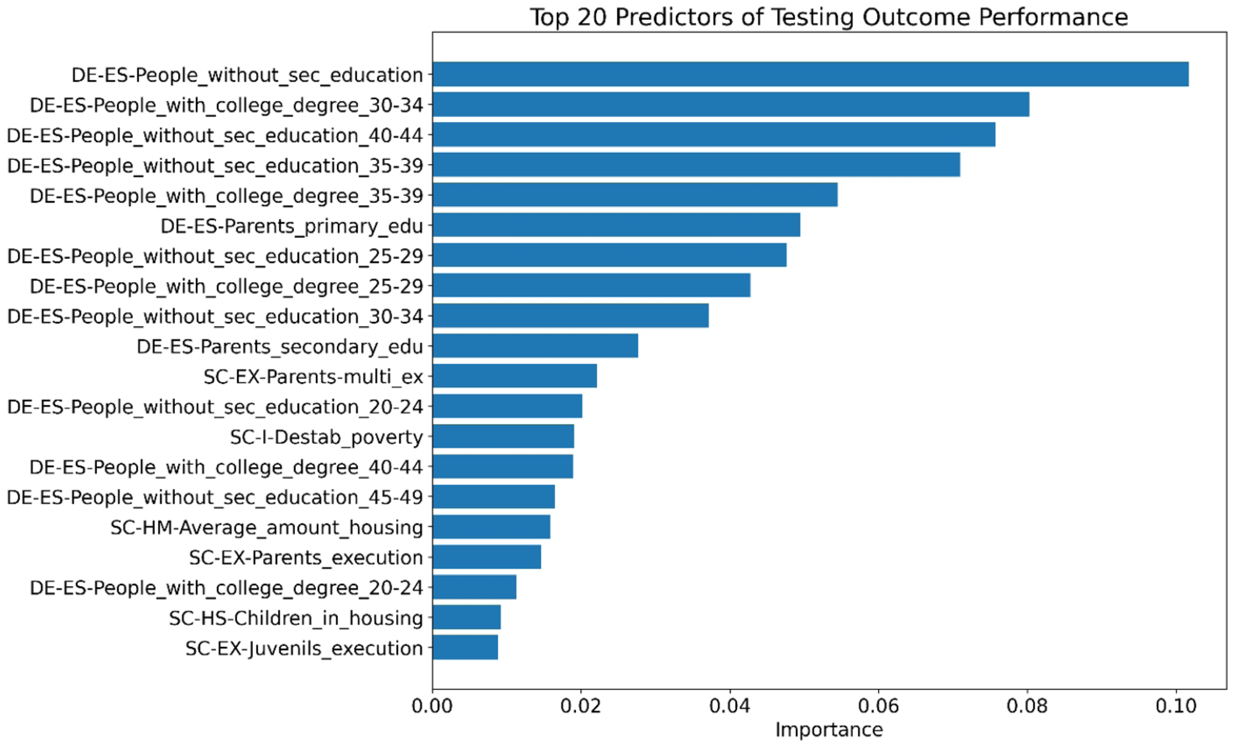


Figure 6: 20 most important indicators on students' testing outcomes.

DISCUSSION

The analysis across the three blocks highlights the central role of demographic and educational structure in predicting student learning outcomes. Indicators such as the share of people without secondary education and the proportion of younger adults with college degrees consistently emerged as the strongest predictors. This underscores the importance of educational attainment not only at the individual level but also as a collective characteristic of municipalities. These findings are consistent with Tan (2024), who confirmed that socioeconomic status and educational attainment are among the strongest predictors of student learning outcomes, with evidence across 48 meta-analyses. Similarly, Clément and Piaser (2022) provided municipal-level evidence that education inequality is spatially distributed and strongly associated with income disparities. From a policy perspective, these results reinforce the relevance of prioritizing investments in secondary and higher education pathways, particularly targeting groups with persistently low completion rates. This aligns with OECD (2025b), which demonstrates that higher educational attainment is linked to improved labour market participation and stronger social cohesion. Raising completion rates is therefore not merely an educational objective but a broader social and economic imperative, as municipalities with stronger collective educational composition are better positioned to support both student success and community development (Nieuwenhuis and Hooimeijer, 2016; Veerman et al., 2021).

The social and housing block adds nuance by showing how contextual vulnerabilities are associated with educational outcomes. Parental executions, poverty destabilization, and housing allowances illustrate that structural disadvantages are linked to variations in the benefits of education. For example, sufficient housing support is associated with reduced negative impacts of precarious living conditions, while overcrowding is linked to increased stress and lower performance. Consistent with prior evidence that family socioeconomic stressors (poverty, instability) are predictors of academic achievement and attainment (Song et al., 2025; Xu, 2020), these findings suggest that housing assistance should be designed not only to provide financial relief but also to address qualitative aspects of housing stability and adequacy.

In the Czech context, disadvantaged populations often concentrate in specific regions that function as "poverty traps," where low educational attainment and a high concentration of social stressors are linked to one another. This dynamic mirrors findings from regional case studies by Lourens and Bleazard (2016), who documented similar cycles of disadvantage. Our analysis further identified strong nonlinear effects: there appear to be critical tipping points at which educational outcomes decline sharply once institutional or family resilience is exceeded. This observation aligns with Bird et al. (2021), who emphasize that the stability and transparency of predictive models are essential for detecting vulnerable clusters that traditional linear approaches may overlook.

From a policy perspective, these findings provide a robust empirical basis for targeted interventions that go beyond simple financial redistribution. Such measures could include expanding school social work and specialized counseling services in areas where socioeconomic stressors are associated with lower educational outcomes. In addition, the predictive framework developed here could support the design of Early Warning Systems (EWS), enabling municipalities to deploy proactive tutoring or mentoring programs specifically tailored to communities with lower educational capital.

The combined model demonstrates that while demographic and educational indicators account for most of the variance, social and housing stressors refine the explanatory narrative by identifying

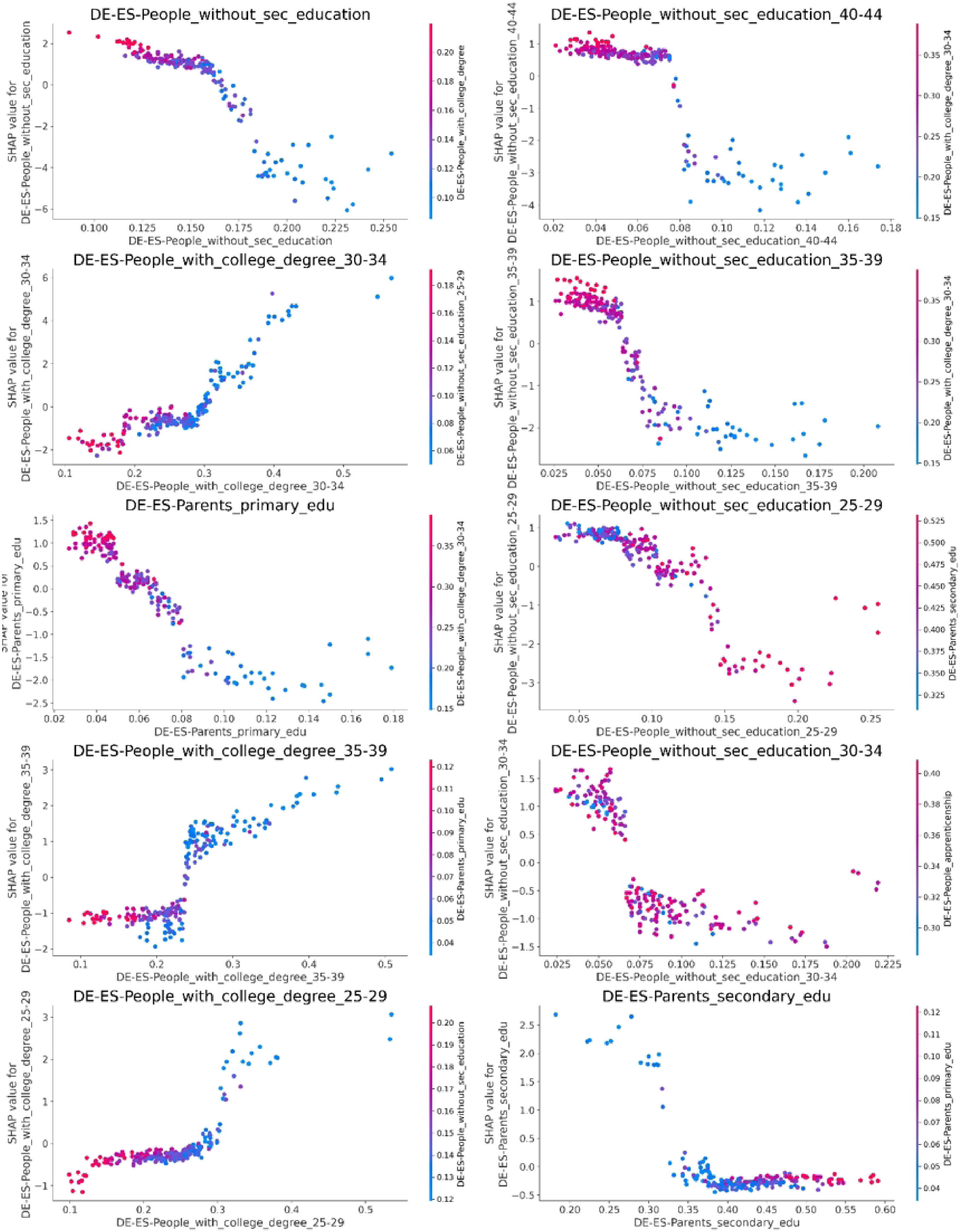


Figure 7: SHAP dependence of the top 10 most influential indicators

vulnerable subgroups and nonlinear threshold effects. Although the overall fit remains similar across specifications, integrating contextual stressors enhances interpretive value and highlights the importance of multi-level strategies. These results are consistent with broader evidence on the societal role of education. Flégl et al. (2025) showed that inequalities in Czech primary education are associated not only with student performance but also with civic engagement and governance efficiency. Their findings, together with the presented analysis, emphasize that improving educational attainment and reducing socioeconomic disparities are linked to benefits that extend beyond the classroom, strengthening both human capital and democratic participation. Importantly, this also underscores the dimension of educational responsibility: municipalities and institutions are accountable for addressing structural disadvantages and ensuring that vulnerable groups are not left behind. Recognizing responsibility as a guiding principle reinforces the need for policies that integrate educational pathways with social support, thereby promoting equity, resilience, and sustainable community development.

Study limitations

While the analysis provides valuable insights into the role of demographic, educational, and social factors in shaping student outcomes across Czech municipalities, several limitations should be acknowledged. First, the dependent variable—the index of test results—captures performance at specific grade levels and on entrance examinations, which may not fully reflect broader dimensions of student achievement, such as creativity, problem-solving, or socio-emotional skills. Second, the explanatory indicators are drawn from data published between 2021 and 2024, whereas the testing results correspond to 2025. However, the most recent data were used to minimize temporal gaps; some lag effects may remain. Third, median imputation for missing values, while methodologically sound, may reduce variability and obscure localized extremes. Fourth, we acknowledge the imbalance between the number of observations (206 MEJs) and the number of predictors (49 social indicators, 85 demographic/economic indicators, and 134 combined). This challenge is common in socio-economic and educational research, where datasets often contain many explanatory variables but relatively few cases (Han et al., 2021; Hawkins, 2004). Random Forests were selected because they are well-suited to high-dimensional data, adaptively selecting splits and effectively ignoring irrelevant predictors, which reduces the risk of overfitting (Breiman, 2001). To further mitigate this issue, dimensionality reduction techniques (e.g., Principal Component Analysis (PCA) or feature selection) can be applied in future work. Finally, the analysis is limited to municipalities with extended jurisdiction in the Czech Republic, which constrains the generalizability of findings to other national contexts. Future research could address these limitations by incorporating longitudinal data, alternative measures of student success, and comparative analyses across different educational systems.

CONCLUSION

This study examined how social, demographic, and economic factors are associated with student learning outcomes across Czech municipalities with extended jurisdiction. By operationalizing testing results as the dependent variable and integrating indicators of educational attainment, social stressors, and housing conditions, the analysis showed that demographic and educational structure are the strongest predictors of student performance. At the same time, social and housing stressors refine the explanatory narrative, identifying vulnerable subgroups and nonlinear threshold effects that highlight the importance of contextual resilience.

The findings indicate that municipalities with stronger collective educational composition are more likely to be linked to higher student success and community development. Conversely, regions marked by concentrated disadvantage are associated with "poverty traps," where low educational attainment and social stressors co-occur. These insights highlight the relevance of multilevel strategies that combine investments in secondary and higher education pathways with targeted social support measures. Beyond immediate educational outcomes, the results have broader societal implications. Higher educational attainment is consistently associated with stronger human capital formation, social cohesion, civic engagement, and democratic participation. Addressing disparities in education and social conditions is, therefore, both an educational and societal imperative, underscoring the dimension of educational responsibility as municipalities and institutions remain accountable for fostering resilience, equity, and sustainable community development.

REFERENCES

Arlot, S. and Celisse, A. (2010) 'A survey of cross-validation procedures for model selection', *Statistics Surveys*, Vol. 4, pp. 40–79. https://dx.doi.org/10.1214/09-SS054

Arnold, K. E. and Pistilli, M. D. (2012) 'Course signals at Purdue: Using learning analytics to increase student success', in: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK 2012),* pp. 267–270. https://doi.org/10.1145/2330601.2330666

Asif, R., Merceron, A., Ali, S. A. and Haider, N. G. (2017) 'Analyzing undergraduate students' performance using educational data mining', *Computers & Education*, Vol. 113, pp. 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Bertoletti, A., Berbegal-Mirabent, J. and Agasisti, T. (2022) 'Higher education systems and regional economic development in Europe: A combined approach using econometric and machine learning methods', *Socio-Economic Planning Sciences*, Vol. 82, p. 101231. https://doi.org/10.1016/j.seps.2022.101231

Bird, K. A., Castleman, B. L., Mabel, Z. and Song, Y. (2021) 'Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education', *AERA Open*, Vol. 7, No. 1, pp. 1–19. https://doi.org/10.1177/23328584211037630

Bravo Sanzana, M., Salvo Garrido, S. and Muñoz Poblete, C. (2015) 'Profiles of Chilean students according to academic performance in mathematics: An exploratory study using classification trees and random forests', *Studies in Educational Evaluation*, Vol. 44, pp. 50–59. http://dx.doi.org/10.1016/j.stueduc.2015.01.002

Breiman, L. (2001) 'Random Forests', *Machine Learning*, Vol. 45, pp. 5–32. https://doi.org/10.1023/A:1010933404324

Browne, M. W. (2000) 'Cross-validation methods', *Journal of Mathematical Psychology*, Vol. 44, No. 1, pp. 108–132. https://doi.org/10.1006/jmps.1999.1279

Cheng, B., Liu, Y. and Jia, Y. (2024) 'Evaluation of students' performance during the academic period using the XG-Boost Classifier-Enhanced AEO hybrid model', *Expert Systems with Applications*, Vol. 238, p. 122136. https://doi.org/10.1016/j.eswa.2023.122136

Chung, J. Y. and Lee, S. (2019) 'Dropout early warning systems for high school students using machine learning', *Children and Youth Services Review*, Vol. 96, pp. 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030

Clément, M. and Piaser, L. (2022) 'Geography of Income and Education Inequalities in Mexico: Evidence from Small Area Estimation and Exploratory Spatial Analysis', *The European Journal of Development Research*, Vol. 34, No. 2, pp. 703–732. https://doi.org/10.1057/s41287-021-00386-0

Conijn, R., Snijders, C., Kleingeld, A. and Matzat, U. (2017) 'Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS', *IEEE Transactions on Learning Technologies*, Vol. 10, No. 1, pp. 17–29. https://doi.org/10.1109/TLT.2016.2616312

Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T., and Gulam, S. B. (2025) 'Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors', *Systems and Soft Computing*, Vol. 7, p. 200352. https://doi.org/10.1016/j.sasc.2025.200352

European Commission (2025) *European Commission: Directorate-General for Education, Youth, Sport and Culture, Education and training monitor 2025 – Czechia*. Publications Office of the European Union. Available at: https://data.europa.eu/doi/10.2766/6444520 [Accessed 2 February 2026].

Eurydice (2024) *National education systems: Czech Republic*, European Commission. Available at: https://eurydice.eacea.ec.europa.eu/eurypedia/czechia/overview [Accessed 11 February 2026]

Ferguson, R. (2012) 'Learning Analytics: Drivers, Developments and Challenges', *International Journal of Technology Enhanced Learning*, Vol. 4, No. 5/6, pp. 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Flegl, M., Vltavská, K. and Acero, A. (2025) 'Towards evaluation of the Czech primary education and its effect on civic engagement and governance', in: *Proceedings of the 43rd International Conference on Mathematical Methods in Economics (MME 2025)*, Zlín, Czech Republic, pp. 192–197.

Han, S., Williamson, B. D. and Fong, Y. (2021) 'Improving random forest predictions in small datasets from two-phase sampling designs', *BMC Medical Informatics and Decision Making*, Vol. 21, No. 1, p. 322. https://doi.org/10.1186/s12911-021-01688-3

Hauschildt, K., Gwosc, C., Schirmer, H., Mandl, S. and Menz, C. (2024) *Social and economic conditions of student life in Europe: Eurostudent 8 synopsis of indicators 2021–2024*, Bielefeld: wbv Media. https://doi.org/10.3278/6001920ew

Hawkins, D. M. (2004) 'The problem of overfitting', *Journal of Chemical Information and Computer Sciences*, Vol. 44, No. 1, pp. 1–12. https://doi.org/10.1021/ci0342472

Herodotou, C., Rienties, B., Boroeca, A., Zdrahal, Z. and Hlosta, M. (2019) 'A Large-scale Implementation of Predictive Learning Analytics in Hospitality and Healthcare Courses', *Educational Technology Research and Devwlopment*, Vol. 67, No. 5, pp. 1273–1306. https://doi.org/10.1007/s11423-019-09685-0

Hussain, M., Zhu, W., Zhang, W. and Abidi, S. M. R. (2018) 'Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores', *Computational Intelligence and Neuroscience*, Vol. 2018, No. 1, p. 6347186. https://doi.org/10.1155/2018/6347186

Jafari, A., Aghsami, A. and Rabbani, M. (2025) 'Selecting the best way to forecast income in the banking industry using data mining methods, a case study', *OPSEARCH*, Vol. 62, No. 3, pp. 1383–1422. https://doi.org/10.1007/s12597-024-00852-3

Jiang, X., Du, Y. and Zheng, Y. (2024) 'Evaluation of physical education teaching effect using Random Forest model under artificial intelligence', *Heliyon*, Vol. 10, No. 1, e23576. https://doi.org/10.1016/j.heliyon.2023.e23576

Khan, S., Mazhar, T., Shahzad, T., Khan, M.A., Waheed, W., Waheed, A. and Hamam, H. (2025) 'Predictive analytics in education- enhancing student achievement through machine learning', *Social Sciences & Humanities Open*, Vol. 12, p. 101824. https://doi.org/10.1016/j.ssaho.2025.101824

Kotsiantis, S. B. (2012) 'Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades', *Artificial Intelligence Review*, Vol. 37, No. 4, pp. 331–344. https://doi.org/10.1007/s10462-011-9234-x

Lourens, A. and Bleazard, D. (2016) 'Applying predictive analytics in identifying students at risk: A case study', *South African Journal of Higher Education*, Vol. 30, No. 2, pp. 129–150. https://doi.org/10.20853/30-2-583

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S. I. (2020) 'From Local Explanations to Global Understanding with Explainable AI for Trees', *Nature machine intelligence*, Vol. 2, No. 1, pp. 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M. and Lee, S. I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4765–4774. https://doi.org/10.48550/ARXIV.1705.07874

Mazouch, P. and Fischer, J. (2024) *Více času na pedagogické vedení školy prostřednictvím efektivního zajištění nepedagogických činností* [More time for pedagogical management of the school through effective provision of non-pedagogical activities], Prague: Prague University of Economics and Business. Available at: https://partnerstvi2030.cz/wp-content/uploads/Vice_casu_na_pedagogicke_vedeni_skoly_VSE.pdf [Accessed 3 March 2026]

MEYS (2020) *Strategy for the education policy of the Czech Republic up to 2030+*, Prague: Ministry of Education, Youth and Sports. Available at: https://msmt.gov.cz/uploads/brozura_S2030_en_fin_online.pdf [Accessed 2 February 2026]

Molnar, C. (2022) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Self-Published. https://christophm.github.io/interpretable-ml-book/

Nieuwenhuis, J. and Hooimeijer, P. (2016) 'The association between neighbourhoods and educational achievement, a systematic review and meta-analysis', *Journal of Housing and the Built Environment*, Vol. 31, No. 2, pp. 321–347. https://doi.org/10.1007/s10901-015-9460-7

Nouri, J., Ebner, M., Ifenthaler, D., Saqr, M., Malmberg, J., Khalil, M., Bruun, J., Viberg, O., González, M. Á. C., Papamitsiou Z. and Berthelsen, U. D. (2019) 'Efforts in Europe for Data-Driven Improvement of Education: A Review of Learning Analytics Research in Seven Countries', *International Journal of Learning Analytics and Artificial Intelligence for Education (iJAI)*, Vol. 1, No. 1, pp. 8–27. https://doi.org/10.3991/ijai.v1i1.11053

OECD (2020) *Education policy outlook in the Czech Republic, OECD Education Policy Perspectives*, No. 11, Paris: OECD Publishing. https://doi.org/10.1787/6363ab1d-en

OECD (2023a) *Education at a Glance 2023: OECD Indicators*, Paris: OECD Publishing. https://doi.org/10.1787/e13bef63-en

OECD (2023b) *PISA 2022 results (Volume I): The state of learning and equity in education*, Paris: OECD Publishing. https://doi.org/10.1787/53f23881-en

OECD (2025a) *OECD economic surveys: Czechia 2025*, Paris: OECD Publishing. https://doi.org/10.1787/7a70af5c-en

OECD (2025b) *Education at a Glance 2025: OECD Indicators*, Paris: OECD Publishing. https://doi.org/10.1787/1c0d9c79-en

PAQresearch (2026) *Mapa vzdělávání [Education map]*, Available at: https://mapavzdelavani.cz/ [Accessed 26 January 2026]

Rabelo, A. M. and Zárate, L. E. (2025) 'A model for predicting dropout of higher education students', *Data Science and Management*, Vol. 8, No. 1, pp. 72–85. https://doi.org/10.1016/j.dsm.2024.07.001

Romero, C. and Ventura, S. (2010) 'Educational Data Mining: A Review of the State of the Art', *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 40, No. 6, pp. 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

SGI (2024) *Sustainable governance indicators – Czech Republic, Bertelsmann Stiftung*. Available at: https://www.sgi-network.org/2024/Czechia [Accessed 11 February 2026]

Shafiq, D. A., Marjani, M., Habeeb, R. A. A. and Asirvatham, D. (2022) 'Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review', *IEEE Access*, Vol. 10, pp. 72480–72503. https://doi.org/10.1109/ACCESS.2022.3188767

Shapley, L. (1953) 'A value for n-person games', in: Kuhn, H. and Tucker, A. (eds.), *Contributions to the Theory of Games II*, Princeton: Princeton University Press, pp. 307–317. https://doi.org/10.1515/9781400881970-018

Simonová, N. and Soukup, P. (2015) 'Impact of primary and secondary social origin factors on the transition to university in the Czech Republic', *British Journal of Sociology of Education*, Vol. 36, No. 5, pp. 707–728. https://doi.org/10.1080/01425692.2013.854690

Song, Q., Liu, Y. and Tan, C. Y. (2025) 'Effects of Family Socioeconomic Status on Educational Outcomes in Primary and Secondary Education: A Systematic Review of the Causal Evidence', *Educational Psychology Review*, Vol. 37, No. 29. https://doi.org/10.1007/s10648-025-10004-8

Šťastný, V. (2023) 'Shadow education in the context of early tracking: between-track differences in the Czech Republic', *Compare: A Journal of Comparative and International Education*, Vol. 53, No. 3, pp. 380–398. https://doi.org/10.1080/03057925.2021.1922271

Tan, C. Y. (2024) 'Socioeconomic Status and Student Learning: Insights from an Umbrella Review', *Educational Psychology Review*, Vol. 36, No. 4. https://doi.org/10.1007/s10648-024-09929-3

Tsai, Y. and Gašević, D. (2017) 'Learning analytics in higher education - challenges and policies: A review of eight learning analytics policies', in: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK 2017)*, pp. 233–242. https://doi.org/10.1145/3027385.3027400

Umer, R., Susnjak, T., Mathrani, A. and Suriadi, L. (2023) 'Current stance on predictive analytics in higher education: opportunities, challenges and future directions', *Interactive Learning Environments*, Vol. 31, No. 6, pp. 3503–3528. https://doi.org/10.1080/10494820.2021.1933542

Veerman, G. J. and Denessen, E. (2021) 'Social cohesion in schools: A non-systematic review of its conceptualization and instruments', *Cogent Education*, Vol. 8, No. 1, pp. 1–14. https://doi.org/10.1080/2331186X.2021.1940633

Xu, Y. (2020) 'Foreclosed American Dream? Parental Foreclosure and Young Adult Children's Homeownership', *Journal of Family and Economic Issues*, Vol. 41, No. 3, pp. 458–471. https://doi.org/10.1007/s10834-020-09665-0

Zhang, Y. and Yang, Q. (2022) 'A Survey on Multi-Task Learning', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 12, pp. 5586–5609. https://dx.doi.org/10.1109/TKDE.2021.3070203

Zuluaga, R., Camelo-Guarín, A. and De La Hoz, E. (2023) 'Assessing the Relative Impact of Colombian Higher Education Institutions Using Fuzzy Data Envelopment Analysis (Fuzzy-DEA) in State Evaluations', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 4, pp. 299–312. http://dx.doi.org/10.7160/eriesj.2023.160404

APPENDIX

Area	Indicator	Code
Social condition	Destabilizing poverty	SC-I-Destab_poverty
	Socioeconomic development	SC-I-Socio_development
	Socioeconomic disadvantage	SC-I-Socio_disadvantage
Housing shortage	Children in housing need	SC-HS-Children_in_housing
	Children in housing need in short-term housing contracts	SC-HS-Children_in_housing_short
	Children in housing need in precarious housing	SC-HS-Children_in_housing_preca
	Children in housing need living in shelters and hostels	SC-HS-Children_in_housing_shelters
	Children in housing need living in overcrowded apartments	SC-HS-Children_in_housing_overcrowded
Execution	Children in execution	SC-EX-Children_execution
	People with multiple foreclosures	SC-EX-People_multi_ex
	People in execution	SC-EX-People-execution
	Juveniles in execution	SC-EX-Juvenils_execution
	Average amount recovered per executed	SC-EX-Average_amount
	Parents in execution	SC-EX-Parents_execution
	Parents in multiple foreclosures	SC-EX-Parents-multi_ex
Unemployment	Long-term unemployment	SC-UN-Long_term_unemp
	Unemployment	SC-UN-Unemployment
Social exclusion	Social exclusion index	SC-SE-Social_exclusion_index
Social support	Number of child benefits paid in an average month	SC-SS-Number_child_benefits
	Number of child benefits paid in an average month per 1,000 inhabitants under 15 years of age	SC-SS-Number_child_benefits_per_1000
	Number of housing benefits paid in an average month	SC-SS-Number_housing_benefits
	Number of housing benefits paid in an average month per 1,000 inhabitants over 15 years of age	SC-SS-Number_housing_benefits_per_1000
	Share of children under 15 receiving child benefit	SC-SS-Children_under_15_receiving
	Share of children aged 0 to 2 receiving child benefit	SC-SS-Children_under_2_receiving
	Share of children aged 3 to 5 receiving child benefit	SC-SS-Children_3-5_receiving
	Share of children aged 6 to 14 receiving child benefit	SC-SS-Children_6-14_receiving
	Share of farm households receiving housing allowance	SC-SS-Farm_households_receiving
	Share of dependent children receiving child benefit	SC-SS-depen_children_receiving
	Average amount of child benefit	SC-SS-Average_amount_child
	Average amount of housing allowance	SC-SS-Average_amount_housing
	Sum of child benefit payments	SC-SS-Sum_child_benefits
	Sum of housing allowances paid	SC-SS-Sum_housing_paid
	Number of emergency immediate assistance benefits paid per 1,000 population over 15 years of age	SC-HM-Emergency_15+
	Number of housing allowances paid in an average month	SC-HM-Housing_paid
	Number of housing allowances paid in an average month per 1,000 inhabitants over 15 years of age	SC-HM-Housing_paid_per_1000
	Number of living allowances paid in an average month	SC-HM-Living_paid
	Number of living allowances paid in an average month per 1,000 inhabitants over 15 years of age	SC-HM-Living_paid_per_1000
	Share of farming households receiving housing allowance	SC-HM-Farming_households_rec_housing
	Share of farming households receiving subsistence allowance	SC-HM-Farming_households_rec_subsistence
	Average amount of emergency immediate assistance benefit	SC-HM-Average_amount_emergency
	Average amount of housing supplement paid	SC-HM-Average_amount_housing
	Average amount of living allowance	SC-HM-Average_amount_living
	Sum of paid amounts of emergency immediate assistance benefits	SC-HM-Sum_amount_emergency
	Sum of housing supplement paid	SC-HM-Sum_amount_housing
	Sum of paid amounts of living allowances	SC-HM-Sum_amount_living
	Number of care allowances paid in an average month	SC-OA-Number_care_allowance
	Number of care benefits paid in an average month per 1,000 inhabitants over 15 years of age	SC-OA-Number_care_allowance_per_1000
	Average amount of care allowance	SC-OA-Average_care_allowance
	Sum of paid care allowances	SC-OA-Sum_care_allowance

Table 1: Overview of analyzed indicators in the social conditions area

Area	Indicator	Code
Population and municipalities	Single-parent households – number	DE-PM-Single-parent_households_num
	Single-parent households – share	DE-PM-Single-parent_households_share
	Population density	DE-PM-Population_density
	Population – number	DE-PM-Population
	Population by age – proportion 0-14	DE-PM-Population_share_0-14
	Population by age – proportion 0-17	DE-PM-Population_share_0-17
	Population by age – proportion 15-64	DE-PM-Population_share_15-64
	Population by age – proportion 18-29	DE-PM-Population_share_18-29
	Population by age – proportion 30-39	DE-PM-Population_share_30-39
	Population by age – proportion 40-49	DE-PM-Population_share_40-49
	Population by age – proportion 50-64	DE-PM-Population_share_50-64
	Population by age – proportion 65+	DE-PM-Population_share_65+
	Number of school-age children - 3-5	DE-PM-School-age_children_3-5
	Number of school-age children - 6-10	DE-PM-School-age_children_6-10
	Number of school-age children - 11-14	DE-PM-School-age_children_11-14
	Number of school-age children - 15-19	DE-PM-School-age_children_15-19
	Number of municipalities	DE-PM-Number_municipalities
	Number of municipalities by population - 0-500	DE-PM-Number_municipalities_by_pop_0-500
	Number of municipalities by population - 501-1000	DE-PM-Number_municipalities_by_pop_501-100
	Number of municipalities by population - 1001+	DE-PM-Number_municipalities_by_pop_1001+
	Share of residents in small municipalities - 0-500	DE-PM-Residents_small_muni_0-500
	Share of residents in small municipalities - 0-1000	DE-PM-Residents_small_muni_0-1000
	Average age of population - total	DE-PM-Age_population_total
	Average age of population - men	DE-PM-Age_population_men
	Average age of population - women	DE-PM-Age_population_women
Population movement	Total population growth	DE-MO-Population_growth
	Total population growth per 1000 inhabitants	DE-MO-Population_growth_per_1000
	Number of children born	DE-MO-Children_born
	Number of births per 1000 population	DE-MO-Children_born_per_1000
	Number of immigrants	DE-MO-Number_immigrants
	Number of immigrants per 1000 inhabitants	DE-MO-Number_immigrants_per_1000
	Number of divorces	DE-MO-Number_divorces
	Number of divorces per 1000 inhabitants	DE-MO-Number_divorces_per_1000
	Number of marriages	DE-MO-Number_marriages
	Number of marriages per 1000 inhabitants	DE-MO-Number_marriages_per_1000
	Number of emigrants	DE-MO-Number_emigrants
	Number of emigrants per 1000 inhabitants	DE-MO-Number_emigrants_per_1000
	Number of deaths	DE-MO-Number_deaths
	Number of deaths per 1000 inhabitants	DE-MO-Number_deaths_per_1000
	Natural population growth	DE-MO-Natural_popu_growth
	Natural population growth per 1000 inhabitants	DE-MO-Natural_popu_growth_per_1000
	Population growth through migration	DE-MO-Popu_growth_migration
	Population growth through migration per 1000 inhabitants	DE-MO-Popu_growth_migration_per_1000
Educational structure	Share of people without secondary education	DE-ES-People_without_sec_education
	Share of people without secondary education in age groups 20-24	DE-ES-People_without_sec_education_20-24
	Share of people without secondary education in age groups 25-29	DE-ES-People_without_sec_education_25-29
	Share of people without secondary education in age groups 30-34	DE-ES-People_without_sec_education_30-34
	Share of people without secondary education in age groups 35-39	DE-ES-People_without_sec_education_35-39
	Share of people without secondary education in age groups 40-44	DE-ES-People_without_sec_education_40-44
	Share of people without secondary education in age groups 45-49	DE-ES-People_without_sec_education_45-49
	Proportion of people with a high school diploma	DE-ES-People_with_high_school
	Share of people with a college degree	DE-ES-People_with_college_degree
	Share of people with a college degree in the 20-24 age group	DE-ES-People_with_college_degree_20-24
	Share of people with a college degree in the 25-29 age group	DE-ES-People_with_college_degree_25-29
	Share of people with a college degree in the 30-34 age group	DE-ES-People_with_college_degree_30-34
	Share of people with a college degree in the 35-39 age group	DE-ES-People_with_college_degree_35-39
	Share of people with a college degree in the 40-44 age group	DE-ES-People_with_college_degree_40-44
	Share of people with a college degree in the 45-49 age group	DE-ES-People_with_college_degree_45-49
	Share of people with an apprenticeship certificate	DE-ES-People_apprenticeship
	Share of parents with at most secondary education without a high school diploma	DE-ES-Parents_secondary_edu
	Share of parents with at most primary education	DE-ES-Parents_primary_edu
	Proportion of parents with higher education	DE-ES-Parents_higher_edu

Area	Indicator	Code
Labor market	Total share of people working in advanced services	DE-LM-People_working_advanced_services
	Total share of people working in essential services	DE-LM-People_working_essential_services
	Total share of people working in agriculture, industry, construction, etc.	DE-LM-People_working_agriculture_constr
	Total share of people working in public administration, education and science	DE-LM-People_working_public_administration
	Share of employees in the highest quintile by employment income	DE-LM-Employees_highest_income
	Share of employed at skill level 1 (lowest) by gender - men	DE-LM-Employees_skill_level1_men
	Share of employed at skill level 1 (lowest) by gender - women	DE-LM-Employees_skill_level1_women
	Share of employed at skill level 2 by gender - men	DE-LM-Employees_skill_level2_men
	Share of employed at skill level 2 by gender - women	DE-LM-Employees_skill_level2_women
	Share of employed at skill level 3 by gender - men	DE-LM-Employees_skill_level3_men
	Share of employed at skill level 3 by gender - women	DE-LM-Employees_skill_level3_women
	Share of employed at skill level 4 (highest) by gender - men	DE-LM-Employees_skill_level4_men
	Share of employed at skill level 4 (highest) by gender - women	DE-LM-Employees_skill_level4_women
Commuting	Share of workers who commute to another municipality for work	DE-CO-Workers_commute_to_municipality
	Share of workers who commute to work in another region	DE-CO-Workers_commute_to_region
	Share of workers who commute to work in another district	DE-CO-Workers_commute_to_district
	Share of workers who commute abroad for work	DE-CO-Workers_commute_to_abroad
	Share of workers who work in the municipality of residence	DE-CO-Workers_commute_to_residence
	Share of pupils and students who attend school in another municipality in the same district	DE-CO-Students_commute_to_municipality
	Share of pupils and students who attend school in a region other than their place of residence	DE-CO-Students_commute_to_region
	Proportion of pupils and students who attend school in another district in the same region	DE-CO-Students_commute_to_district
	Proportion of pupils and students who attend school abroad	DE-CO-Students_commute_to_abroad
	Share of pupils and students who attend school in the same municipality as their residence	DE-CO-Students_commute_to_residence

Table 2: Overview of analyzed indicators in the demographic and economic structure

Full research paper

MEASURING ACADEMIC EFFICIENCY IN HIGH-IMPACT SCHOLARSHIPS: A TWO-STAGE WINDOWS DEA AND GAUSSIAN MIXTURE MODEL APPROACH

Andres Acero[1,2]
Miguel Alejandro Garzón-Parra[1]
Jesús Isaac Vázquez-Serrano[1]

[1]Tecnologico de Monterrey, Mexico

[2]Institución Universitaria Politécnico Grancolombiano, Colombia

andres.acero@tec.mx

ABSTRACT

Evaluating the effectiveness of social support programs in higher education requires moving beyond homogeneous assessments of student performance. This study integrates intersectionality with dynamic efficiency analysis to examine how academic efficiency evolves across diverse student profiles within the Líderes del Mañana full-scholarship program in Mexico. Using a longitudinal dataset of 1,796 students (22,718 student–term observations), we apply a two-stage approach. First, Window Data Envelopment Analysis (DEA) estimates relative academic efficiency over time. Second, Gaussian Mixture Modeling identifies intersectional student profiles based on efficiency trajectories and contextual characteristics. Results reveal five distinct efficiency trajectories. While most students converge toward high-efficiency levels, one cluster exhibits a clear negative efficiency slope, greater variability, and limited institutional alignment, indicating it is a priority for intervention. Other clusters display stable high performance, continuous improvement, or moderate but non-accelerating trajectories. Findings demonstrate that efficiency differences are not explained by single demographic factors but by configurations of social background and institutional context. This study provides a scalable, data-driven framework for aligning equity and efficiency objectives in higher education policy and scholarship programs.

KEYWORDS

Educational efficiency, intersectionality, Data Envelopment Analysis (DEA), Gaussian Mixture Models (GMM), higher education, learning analytics

HOW TO CITE

Acero A., Garzón-Parra M. A., Vázquez-Serrano J. I. (2026) 'Measuring Academic Efficiency in High-Impact Scholarships: A Two-Stage Windows DEA and Gaussian Mixture Model Approach', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 57–71. http://dx.doi.org/10.7160/eriesj.2026.190105

Article history
Received
January 12, 2026
Received in revised form
February 27, 2026
Accepted
March 10, 2026
Available on-line
March 31, 2026

Highlights

- *Combining Window DEA and Gaussian Mixture Modeling identifies five intersectional efficiency profiles, revealing academic trajectories hidden by conventional single-axis analyses.*
- *Academic efficiency is dynamic and cumulative: students converge toward the efficiency frontier in later semesters, but trajectories diverge significantly across intersectional profiles.*
- *Institutional context buffers or amplifies structural disadvantage: semi-urban students are the most vulnerable due to a policy blind spot in urban-rural support frameworks.*
- *Efficiency and equity are complementary objectives: differentiated, intersectionality-informed interventions outperform uniform or single-demographic targeting strategies.*

INTRODUCTION

Higher education institutions worldwide face the dual challenge of promoting access for underrepresented students while ensuring efficient use of limited resources to support their academic success. In Mexico, programs like Líderes del Mañana at Tecnológico de Monterrey exemplify efforts to address educational inequity by providing full scholarships to high-achieving youth who demonstrate exceptional academic performance, social leadership, and financial need. The program seeks not only to facilitate access to higher education but also to cultivate transformational leaders committed to social impact in their communities. Participants receive comprehensive support, including full tuition coverage, medical insurance, financial aid for materials, and access to mentoring networks, with the expectation that they will later contribute as agents of social transformation (*Programa LDM | Líderes Del Mañana, n.d.*).

Understanding how effectively educational resources translate into student outcomes is critical for both program sustainability and social equity. However, traditional approaches to evaluating educational efficiency often treat student populations as homogeneous or segment them by single characteristics such as socioeconomic status or geographic origin (Johnes, 2006; Borgonovi and Pokropek, 2019). This overlooks the reality that students navigate educational systems with multiple, intersecting identities, including gender, race, class, disability status, and geographic background, that interact to shape their experiences and outcomes in complex ways.

Intersectionality has become a key framework for analyzing how multiple social identities interact to produce differentiated experiences in educational contexts (Agosto and Roland, 2018; Nichols and Stahl, 2019). In higher education, studies show that gender remains the most examined axis, while other identities receive less attention (Harris and Patton, 2019; Nichols and Stahl, 2019). Moreover, in K–12 and teacher education, intersectionality is mostly applied at the micro-level, focusing on individual experiences with limited attention to systemic resource allocation (Agosto and Roland, 2018; Leckie and Buser De, 2020; Pugach et al., 2019).

Parallel to this theoretical development, Data Envelopment Analysis (DEA) and related efficiency measures have been widely applied to assess educational performance across schools, secondary education systems, and universities in multiple countries (Chiariello et al., 2022; Muniz et al., 2024; Sun et al., 2023; Taleb et al., 2023; Temoso et al., 2023; Tran et al., 2022; Ulkhaq et al., 2024; Zhou et al., 2024). These studies identify factors associated with higher efficiency, such as infrastructure, ICT integration, teacher qualifications, and institutional resources, while highlighting regional or systemic disparities in performance.

Despite these advances, a critical gap remains: existing efficiency studies treat student populations as homogeneous or segment them by single demographic characteristics. At the same time, intersectionality research has yet to be integrated with efficiency measurement. This disconnect limits our understanding of whether educational systems efficiently serve all students equally or whether resource utilization varies systematically across students with different combinations of intersecting identities. As postsecondary institutions become increasingly diverse, and as quantitative methods for intersectional analysis mature (Keller et al., 2023; Prior et al., 2025; Slominski et al., 2024), the need to bridge these research streams has never been more pressing. This study addresses the following research questions:

- How does educational efficiency vary across students with different intersectional profiles?
- Which combinations of social identities are associated with more or less efficient conversion of educational inputs into academic outcomes?
- What factors explain differences in efficiency across intersectional student groups over time?

This paper addresses these questions by integrating quantitative intersectional methods with dynamic efficiency analysis within the Líderes del Mañana program. We employ a multi-stage analytical framework that combines person-centered clustering techniques with DEA to examine how efficiently educational resources are utilized across intersectional student profiles.

Our approach unfolds as follows. First, we prepare and clean program data, defining relevant input and output variables. Second, we apply Gaussian Mixture Models (GMMs) to identify distinct intersectional student profiles across multiple identity dimensions and characteristics. And third, we employ dynamic DEA to evaluate efficiency over time for each identified cluster, assessing how effectively students convert program inputs (financial support, infrastructure, mentoring) into outputs (academic performance, persistence, leadership development).

This integrated approach makes three key contributions. Methodologically, it demonstrates how person-centered intersectional analysis can be combined with efficiency measurement to produce insights invisible to conventional approaches. Substantively, it reveals whether and how educational efficiency varies across multiply marginalized student groups, identifying potential inequities in resource utilization. In practice, it provides evidence-based guidance for designing targeted interventions that are responsive to students' complex, multidimensional identities.

The remainder of this paper is organized as follows. Section 2 reviews the theoretical foundations of intersectionality and its application in educational research, discusses quantitative methods for intersectional analysis, and examines the literature on efficiency measurement in education. Section 3 describes the context of the Líderes del Mañana program, our data sources, and the five-stage analytical framework. Section 4 presents results from clustering analysis, dynamic DEA, and predictive modeling. Section 5 discusses the implications of our findings for educational equity and resource allocation. Section 6 concludes with policy recommendations and directions for future research.

LITERATURE REVIEW

This section presents a comprehensive literature review organized thematically by applications of intersectionality, quantitative methods for analyzing heterogeneity, and efficiency measurement in education. The literature review culminates with Table 1, which summarizes relevant studies since 2014. Additionally, each subsection identifies research gaps and outlines the contributions of this study.

Intersectionality in education research

Intersectionality has emerged as a critical framework for understanding how multiple social identities interact to shape educational experiences and outcomes. In higher education, Nichols and Stahl (2019) conducted a systematic review of 50 studies examining inclusion and exclusion through an intersectional lens, finding that gender dominates as the primary axis combined with other identities, and that most studies rely on qualitative case studies. Harris and Patton (2019) traced the application of intersectionality across 97 higher education articles, highlighting tensions between the framework's radical social justice origins and its increasingly academicized use.

Beyond higher education, intersectionality has been applied across educational contexts. Agosto and Roland (2018) reviewed 15 studies in K–12 educational leadership and found that intersectional analyses primarily focus on individual leaders' experiences rather than systemic inequities. In teacher education, Pugach et al. (2019) synthesized 25 years of research and found that identity is often treated unidimensionally, with limited attention to intersecting social markers. Leckie and Buser De (2020) demonstrated how intersectionality-informed professional development can integrate teachers' lived experiences of privilege and oppression into classroom practice. Macias and Stephens (2019) examined how race and gender intersect to create compounded challenges for women of color in education workplaces, particularly among Latina educators.

Intersectionality has also been adopted in specialized educational domains. Robert and Yu (2018) evaluated its use in transnational education policy research, arguing that its deployment in non-Western contexts yields new analytic insights. Maina-Okori et al. (2018) examined intersectionality in environmental and sustainability education, critiquing the field's limited engagement with overlapping systems of marginalization. Bešić (2020) advocated for an intersectional approach to inclusive education in Austria, arguing that recognizing multiple identity factors is essential to identifying discriminatory processes affecting diverse student groups.

Gap: While intersectionality is widely applied theoretically in education research, quantitative operationalization remains limited. Most studies employ qualitative methods and struggle to capture the complexity of intersecting identities through statistical approaches that move beyond examining single identity dimensions or simple two-way interactions.

Quantitative methods for intersectional analysis

Recent methodological advances have begun bridging the gap between intersectional theory and quantitative research. Latent class analysis (LCA) has emerged as a person-centered approach for identifying unobserved subpopulations. Slominski et al. (2024) introduced LCA to STEM education research as a mixture-modeling technique that can uncover heterogeneous student experiences obscured by variable-centered methods. Garnett et al. (2014) applied LCA to examine intersecting forms of discrimination and bullying among ethnically diverse adolescents, identifying four latent classes with differential mental health outcomes. Bauer et al. (2022) systematically reviewed 16 quantitative health studies that employed clustering methods aligned with intersectional theory, finding limited engagement with the intersectional methodology literature despite widespread theoretical citation.

Clustering techniques have also been applied to educational contexts. Hanauer et al. (2025) used hierarchical cluster analysis with data from 2,082 STEM students to identify four underlying identity orientations: heritage, health, self-expression, and career, revealing the complexity of student positionalities. Reinwald and Annen (2023) employed k-medoid clustering to create intersectional employee profiles and examine how professional development affects job satisfaction across groups over 5 years.

Multilevel modeling represents another quantitative approach for intersectional analysis. Byrd et al. (2015) used multilevel models with data from the Education Longitudinal Study to examine how adolescent school misconduct relates to social control across racial, ethnic, and gender groups in different contexts. Robson et al. (2014) applied multilevel multinomial logistic regression within an intersectionality framework to examine postsecondary trajectories of students with special education needs in Toronto.

The Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA) has recently been introduced as a purpose-built quantitative intersectional method. Keller et al. (2023) demonstrated MAIHDA using German data on 5,451 students across 40 intersectional strata defined by gender, immigrant background, and parental characteristics, highlighting advantages including scalability, parsimony, and precision-weighted estimates for small strata. Prior et al. (2025) applied MAIHDA to examine sociodemographic inequalities in student achievement in London across 144 intersectional strata, finding that between-stratum variation is driven largely by additive rather than interactive effects.

Gap: Despite these methodological advances, no studies have combined intersectional analysis with efficiency measurement in education. Quantitative intersectional methods, such as clustering and MAIHDA, have been applied to examine outcomes and experiences but not to assess how efficiently educational resources are utilized across intersectional student groups or whether resource allocation varies by students' multiple, overlapping identities.

Efficiency analysis in education

DEA has become the predominant method for evaluating educational efficiency across multiple levels and contexts. At the K–12 level, Chiariello et al. (2022) used stochastic frontier analysis to evaluate regional efficiency in primary and secondary education in Italy from 2011–2018, revealing significant North–South disparities driven by contextual factors, including GDP, poverty, and institutional quality. Muniz et al. (2024) applied the Slacks-Based Measure (SBM) DEA model to Brazilian schools in Sobral, identifying libraries, computer labs, and sports courts as key infrastructure elements associated with student performance. Kounetas et al. (2023) examined 643 Greek secondary schools over 18 years, finding persistent inefficiencies across regions with limited reform impact and identifying school-level factors, such as science laboratories and class size, as determinants of efficiency.

In higher education, DEA applications have grown increasingly sophisticated. Taleb et al. (2023) applied super-efficiency SBM models to 41 Taiwanese universities, identifying 25 institutions as super-efficient. Tran et al. (2022) conducted one of Vietnam's most comprehensive efficiency assessments, covering 172 institutions and revealing that public universities are less efficient than private ones, while internationally engaged institutions are more efficient. Temoso et al. (2023) introduced a network-based DEA framework for South African universities that separates teaching and research processes, finding that research units lag substantially behind teaching units (efficiency of 0.782 vs. 0.942). Sun et al. (2023) developed a double-frontier parallel

Electronic ISSN 1803-1617 Printed ISSN 2336-2375

DEA model to jointly assess ordinary and vocational education subsystems across 30 Chinese provinces, incorporating both optimistic and pessimistic perspectives.

Cross-national efficiency studies have also emerged. Ulkhaq et al. (2024) evaluated schools across six South-East Asian countries using 2018 PISA data, combining super-efficiency DEA with bootstrapped quantile regression to examine ICT-related determinants. Zhou et al. (2024) assessed the efficiency of China's educational science and technology industry across 31 provinces using the DEA Malmquist index, identifying technological progress as the primary driver of efficiency change.

Gap: Existing efficiency studies in education treat student populations as homogeneous or segment them by single characteristics (e.g., public vs. private institutions, regional location). None of them examines efficiency through an intersectional lens that acknowledges how students' multiple, overlapping identities may relate to resource utilization and educational outcomes. This oversight limits understanding of whether educational systems efficiently serve all students or whether disparities exist across intersectional groups.

Research contributions

This paper addresses the identified gaps by integrating intersectionality with efficiency analysis in education, a combination absent from existing literature, as summarized in Table 1. Specifically, this study employs clustering methods to identify intersectional student profiles across multiple identity dimensions and then applies DEA to assess educational efficiency across these profiles. This approach contributes to three areas:

Methodological innovation: Demonstrates how quantitative intersectional methods (clustering) can be combined with efficiency measurement (DEA) to produce actionable insights about resource allocation and student outcomes across multiply marginalized groups.

Substantive understanding: Reveals whether educational efficiency varies across intersectional student profiles, identifies which combinations of identities are associated with efficient resource utilization, and highlights potential inequities invisible in traditional analyses.

Policy relevance: Provides evidence for targeted interventions by identifying specific intersectional groups that may require additional support or different resource configurations to achieve equitable outcomes, moving beyond one-size-fits-all approaches to educational improvement.

By bridging intersectionality and efficiency analysis, this study advances both the theoretical sophistication of educational equity research and the practical capacity to design interventions responsive to students' complex, multidimensional identities.

Paper	Scope	Intersection-ality	Latent class analysis	Multilevel modeling	Clusters	DEA	Efficiency in education
Leckie and Buser De (2020)	Teacher	X					
Slominski et al. (2024)	Students	X	X				
Byrd et al. (2015)	Students	X		X			
Prior et al. (2025)	Students	X		X			
Robson et al. (2014)	Students	X		X			
Garnett et al. (2014)	Students	X	X		X		
Hanauer et al. (2025)	Students	X			X		
Reinwald and Annen (2023)	Students	X			X		
Alvarez-Hernandez (2021)	Students	X				X	X
Zhou et al. (2024)	Students					X	X
Chiariello et al. (2022)	Students					X	X
Muniz et al. (2024)	Students					X	X
Taleb et al. (2023)	Universities					X	X
Tran et al. (2022)	Universities					X	X
Kounetas et al. (2023)	Secondary school					X	X
Temoso et al. (2023)	Universities					X	X
Sun et al. (2023)	Universities					X	X
Ulkhaq et al. (2024)	Schools					X	X
This paper	Students	X			X	X	X

Table 1: Literature review

MATERIALS AND METHODS

Dataset description

The dataset used in this study comes from the IFE Living Lab and Data Hub of Tecnológico de Monterrey, Mexico. This initiative was designed to support research on the academic progress and social commitment of students participating in the Líderes del Mañana program, a long-standing scholarship and leadership development effort that aims to expand access to higher education for academically talented young people with limited socioeconomic resources. The program's mission is to foster social mobility, prepare students to be positive agents of community change, and reduce educational inequality.

The dataset contains 22,718 observations and 47 variables, organized at the student–academic period level, which allows the same student to appear multiple times across different academic terms (Table 2). The dataset comprises information on undergraduate students enrolled in the Líderes del Mañana program from 2014 to 2023. These students represent multiple cohorts over nine academic cycles, and the data covers a broad range of categories relevant to both academic performance and contextual background. Sociodemographic variables include measures such as gender, age range, and type of residential environment; admission data include pre-university achievement indicators such as high school grade point average and standardized assessment scores; and academic records capture term- and program-level grades, credit load, and curricular status. Additionally, the dataset contains indicators of students' experiences in leadership and community projects, extracurricular involvement, and other measures linked to their engagement and retention in the university context.

The dataset includes three psychometric instruments collected at admission. The DISC assessment measures four personality dimensions: Dominance (decisiveness and leadership), Influence (communication and teamwork), Steadiness (empathy and stability), and Conscientiousness (analytical thinking and precision). The Values Index captures seven motivational drivers: Aesthetic, Economic, Individualistic, Political, Altruistic, Regulatory, and Theoretical. CV levels reflect pre-college participation intensity across eight domains (sports, cultural activities, student organizations, community service, leadership, work experience, academic achievements, and international experience). Finally, Full-Time Equivalent (FTE) is a continuous variable indicating the proportion of courses a student is enrolled in relative to the expected full load for their semester.

Category	Variable	Nature	Measurement Scale
Intersectional and Demographic	Gender	Categorical	Nominal (Female, Male)
	Age Group	Categorical	Ordinal (18 and below, 19-21, 22+)
	Zone Type	Categorical	Nominal (Urban, Rural, SemiUrban)
	First Generation	Categorical	Binary (Yes, No)
Prior Academic Background	High School GPA	Continuous	Ratio (90-100)
	Admission Test Score	Discrete	Interval (0-1600)
	Origin School	Categorical	Binary (0,1)
Psychometric and Leadership	DISC Scores	Continuous	Ratio (0-100)
	Values Index	Continuous	Ratio (0-100)
	CV Levels	Discrete	Ordinal (1,2,3,4)
Institutional and Performance	Term GPA	Continuous	Ratio (0-100)
	Term GPA Program	Continuous	Ratio (0-100)
	Graduation Status	Categorical	Binary (0,1)
	FTE (Academic Load)	Continuous	Ratio (<0)

Table 2: Description of the dataset

Data preparation and preprocessing

The methodological process began with a structured data preparation stage to ensure internal consistency, robustness, and suitability for longitudinal efficiency analysis. Given the dataset's educational nature and temporal structure, special attention was paid to preserving the number of observations across academic periods.

Missing values in numerical variables related to students' academic engagement were handled using forward-fill imputation. This decision was motivated by the need to avoid systematic bias that could arise from deleting incomplete observations, particularly in datasets where data are temporarily associated across transitional academic periods.

Variables not directly associated with academic performance, student context, or profiling objectives were excluded from the analysis. This dimensionality reduction step helped to improve model interpretability, reduce noise, and ensure alignment between the theoretical framework and the analytical models. After preprocessing, the resulting dataset was balanced and coherent, suitable for dynamic efficiency assessment and subsequent clustering.

Window DEA

To evaluate student academic performance over time, this study employed Window Data Envelopment Analysis (Window DEA). In this framework, each student was conceptualized as a decision-making unit (DMU), observed across multiple academic terms. DEA is particularly appropriate in educational contexts due to its ability to accommodate multiple inputs and outputs without imposing restrictive assumptions on the underlying production process (Charnes et al., 1978; Mergoni et al., 2025). This interpretation follows the educational production function perspective, in which prior academic preparation represents the initial endowment of academic capital, and university academic outcomes represent the outputs of the learning process.

In the DEA model, two variables capturing students' pre-university academic preparation were used as inputs: the standardized admission test score and the high school grade point average (GPA) from the origin school. These variables represent the academic resources or prior knowledge students bring into the university system.

Three variables were considered as outputs representing academic performance during each academic period: the student's term GPA, the program-specific term GPA, and the FTE indicator reflecting the student's academic load. In this context, efficiency reflects how effectively students transform their prior academic preparation into academic outcomes and sustained academic engagement throughout the program.

For each window, efficiency scores were computed by solving the following linear programming problem:

$$\begin{aligned} \min_{\theta,\lambda} \quad & \theta \\ \text{s.t.} \quad & Y\lambda \geq y_j, \\ & \theta x_j - X\lambda \geq 0, \\ & \lambda \geq 0, \end{aligned} \tag{1}$$

where x_j and y_j represent the vectors of academic inputs and outputs for student *j*, *X* and *Y* denote the corresponding matrices for all students within the window. An efficiency score of 1 indicates that the student lies on the efficiency frontier, while values below 1 reflect relative inefficiency.

Unlike static DEA, Window DEA applies this formulation to overlapping time-series subsets. If the dataset spans *T* academic periods and a window width of *w* is selected, the number of overlapping windows is:

$$T - w + 1. \tag{2}$$

This approach treats each student–period observation as a distinct unit within each window, thereby increasing discriminatory power and enabling the identification of performance dynamics over time. Window DEA is particularly well-suited for educational settings, where academic performance evolves gradually and cannot be fully captured by single-period efficiency measures. Because academic performance evolves gradually across semesters, a moderate window size (3) is particularly appropriate in educational contexts, where abrupt efficiency shifts are less common than gradual performance improvements.

The application of the Window DEA yields a sequence of efficiency scores for each student across academic periods. These sequences represent efficiency trajectories, reflecting not only the level of academic efficiency but also its evolution over time.

To transform these trajectories into analytically tractable representations, several summary indicators were derived. These indicators were designed to capture complementary dimensions of performance behavior, including overall efficiency, stability, temporal direction, persistence, and recent outcomes.

Let $\theta_{j,t}$ denote the efficiency score of student *j* in period *t*. The following indicators were computed:

- **Average efficiency**, capturing the overall level of performance:

$$\bar{\theta}_j = \frac{1}{T_j}\sum_{t=1}^{T_j}\theta_{j,t} \tag{3}$$

- **Efficiency trend**, estimated through a linear regression of efficiency on time:

$$\theta_{j,t} = \alpha_j + \beta_j t + \varepsilon_{j,t} \tag{4}$$

where β_j indicates improvement or deterioration over time.

- **Efficiency variability**, measuring stability across periods:

$$\sqrt{\frac{1}{T_j - 1}\sum_{t=1}^{T_j}(\theta_{j,t} - \bar{\theta}_j)^2} \tag{5}$$

Additional indicators were computed to capture cumulative performance, the frequency of improvements between consecutive periods, and the most recent efficiency outcome. Together, these metrics provide a compact yet informative description of each student's longitudinal efficiency profile.

Gaussian Mixture Modeling

To identify unidentified groups of students exhibiting similar efficiency trajectories and contextual characteristics, a Gaussian Mixture Model (GMM) was employed to operationalize intersectional profiles. This clustering approach is appropriate when group boundaries are not sharply defined and when probabilistic membership is desirable.

The GMM assumes that the observed data arise from a mixture of *K* multivariate normal distributions, with the probability density function given by:

$$p(x) = \sum_{k=1}^{K}\pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k), \tag{6}$$

Where π_k denotes the mixing proportion of cluster *k*, while μ_k and Σk represent its mean vector and covariance matrix. Parameters were estimated using the Expectation–Maximization algorithm, which maximizes the log-likelihood to ensure an optimal probabilistic fit for the intersectional student profiles:

$$\mathcal{L} = \sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)\right) \tag{7}$$

The clustering model incorporated both efficiency-based indicators derived from Window DEA and a set of contextual and demographic characteristics reflecting students' backgrounds and institutional environments. Continuous variables were standardized before clustering to ensure comparability across scales.

Cluster assignment was based on posterior membership probabilities, allowing students to be assigned to the cluster with the highest likelihood while preserving uncertainty information. This probabilistic approach enables a more nuanced interpretation than hard-clustering methods, particularly in heterogeneous educational populations.

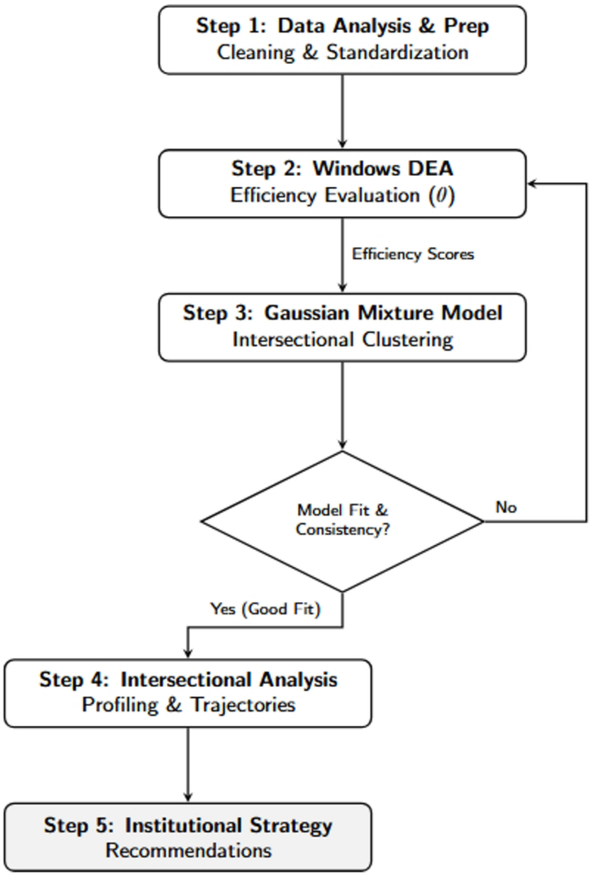

Step 1: Data Analysis & Prep
Cleaning & Standardization
Step 2: Windows DEA
Efficiency Evaluation (θ)
Efficiency Scores
Step 3: Gaussian Mixture Model
Intersectional Clustering
Model Fit & Consistency?
No
Yes (Good Fit)
Step 4: Intersectional Analysis
Profiling & Trajectories
Step 5: Institutional Strategy
Recommendations

Figure 1: Methodological framework

RESULTS

Window DEA

Window DEA results present a clear temporal structure in students' academic efficiency trajectories, as shown in Figure 1. The evolution of average efficiency across academic semesters indicates that students enter the program operating relatively close to the efficiency frontier, with initial average values around 0.94. This suggests that, even in the early semesters, students can transform their academic inputs into outcomes with greater effectiveness than their peers during the same period. However, the persistence of a measurable efficiency gap during the initial semesters reflects an adjustment phase associated with the transition to higher education.

As semesters progress, average efficiency increases gradually and stabilizes within an approximate range of 0.95 to 0.96 during the middle stage of the academic trajectory. This phase is characterized by steady but modest gains, indicating a process of consolidation rather than rapid improvement. From an educational perspective, this pattern is consistent with the progressive acquisition of academic routines, study strategies, and familiarity with institutional expectations, while students simultaneously face increasing curricular demands (Garnet et al., 2014).

A pronounced shift in efficiency dynamics is observed in the later stages of the trajectory. From the advanced semesters onward, average efficiency rises substantially, approaching values close to 1.00 and ultimately converging to the efficiency frontier. This convergence suggests that most students achieve comparable levels of academic performance as they approach program completion. The stabilization of efficiency at frontier levels in the final semesters indicates that differences in relative performance diminish substantially at this stage of the academic trajectory.

To further examine differences in efficiency trajectories across student profiles, Figure 2 was extended to include both the overall average efficiency trajectory and the cluster-level trajectories identified through the Gaussian Mixture Model. While the global trend shows a gradual increase in academic efficiency across the program, cluster-specific lines reveal important heterogeneity in how students approach the efficiency frontier over time.

Clusters 0 and 3 exhibit the highest efficiency levels throughout most of the trajectory, remaining consistently close to the frontier. Cluster 4 shows a clear upward trajectory, indicating progressive improvement as students advance through the program. Cluster 1 maintains relatively stable efficiency levels with limited variation over time. In contrast, Cluster 2 shows the lowest trajectory and greatest instability, reflecting gradual deterioration in efficiency across periods.

These differentiated trajectories highlight that academic efficiency evolves differently across the intersectional configurations of students' characteristics and institutional context. While the overall trend suggests convergence toward high efficiency levels in later

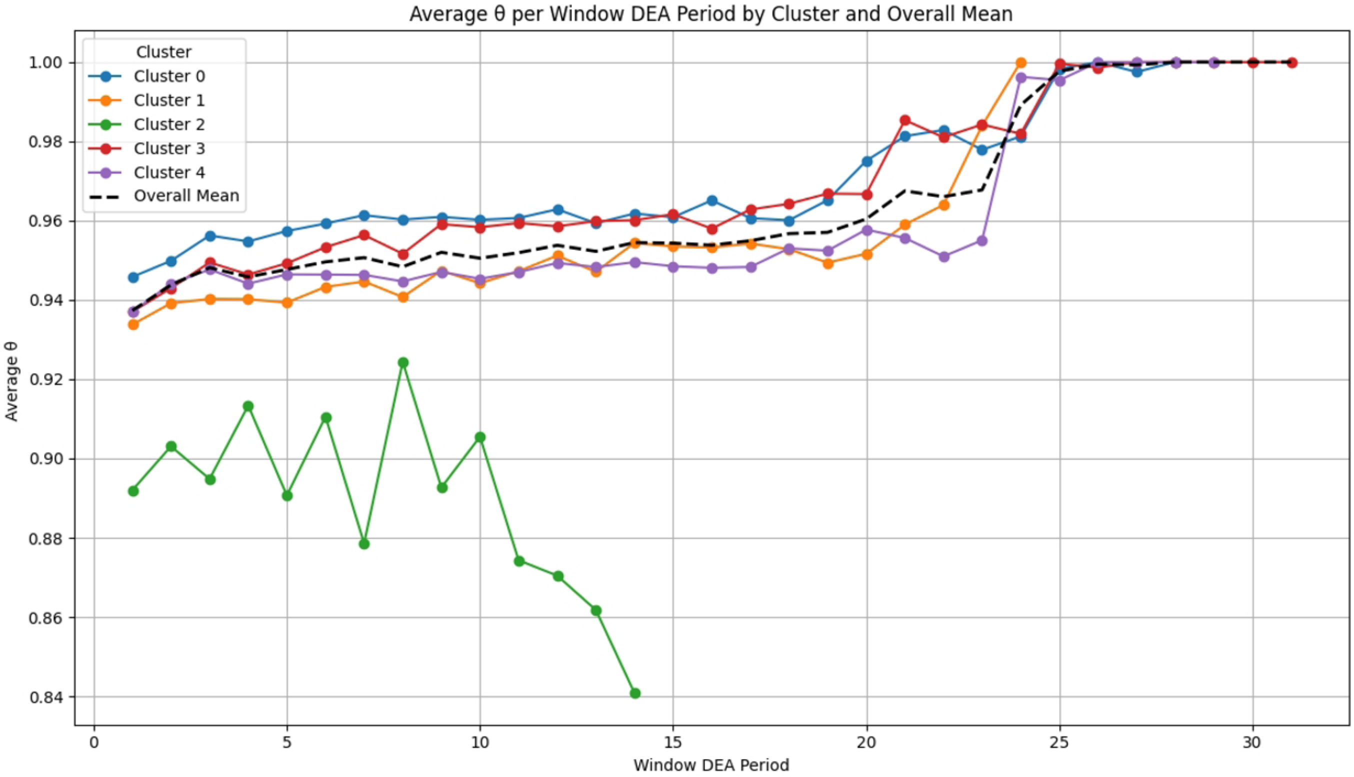


Figure 2: Evolution of average efficiency across Window DEA periods and clusters

stages of the program, the cluster-level analysis reveals persistent structural differences in the pace and stability of this convergence. In general, Window DEA results indicate that academic efficiency among students in the program is a cumulative and dynamic process. While most students exhibit sustained improvement and eventually converge toward high-efficiency levels, meaningful differences in individual trajectories persist throughout the academic lifecycle. These findings underscore the value of a longitudinal efficiency framework for understanding student progression in higher education and motivate further analysis to identify and characterize differentiated academic trajectories.

Cluster description

The information provided by the Window DEA constitutes the empirical foundation for subsequent trajectory-based analysis. By preserving students' relative positions within overlapping time windows, the method captures not only whether academic efficiency improves, but also the consistency and pace of that improvement. The resulting efficiency sequences enable the identification of distinct patterns of academic progression, which are later summarized and differentiated through cluster-level analyses.

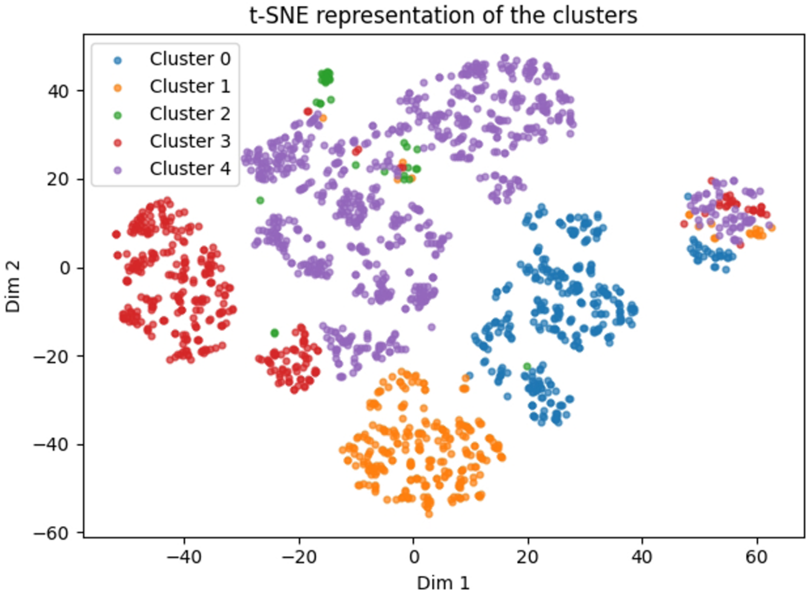


Figure 3: t-SNE visualization of student clusters based on efficiency trajectories and contextual characteristics

Figure 3 presents a two-dimensional visualization of the clustering results using t-distributed Stochastic Neighbor Embedding (t-SNE). Clusters were previously identified using a Gaussian Mixture Model applied to indicators summarizing students' efficiency trajectories derived from the Window DEA analysis, together with contextual variables. The t-SNE projection does not generate the clusters but provides a visual representation of the multidimensional similarity structure of the data. Dim1 and Dim2 correspond to the two coordinates of the t-SNE embedding, which compresses the high-dimensional feature space into two dimensions while preserving local relationships between observations.

The map shows that the identified clusters occupy distinguishable regions of the embedded space, providing qualitative support for the clustering solution. Students assigned to Cluster 3 are primarily located on the left-hand side of the map, indicating strong similarity in their efficiency trajectories and contextual characteristics. Cluster 1 is concentrated in the lower region of the embedding, forming a relatively distinct group. Cluster 0 appears mainly in the center-right portion of the map, while Cluster 4 spans a broader central area, suggesting a more heterogeneous profile that partially overlaps with neighboring groups. Cluster 2 is represented by a smaller and more sparsely distributed set of observations located in the upper-central region of the map.

Some degree of spatial overlap between clusters is expected because the Gaussian Mixture Model assigns probabilistic rather than deterministic cluster membership. Observations near cluster boundaries may exhibit characteristics of multiple profiles, as reflected in the partial mixing visible in the t-SNE representation. Despite this overlap, the overall spatial configuration indicates that the clusters capture meaningful differences in academic efficiency trajectories and contextual student characteristics.

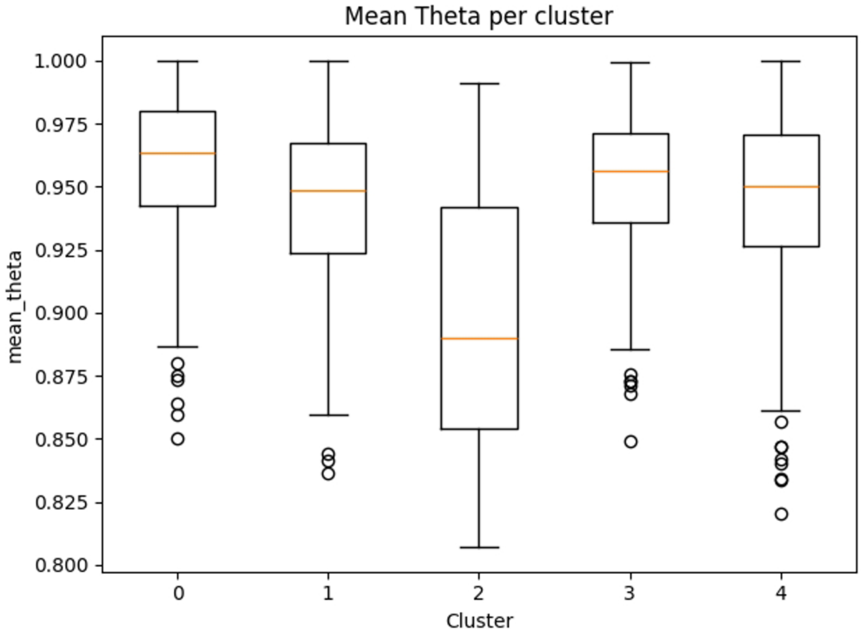


Figure 4: Boxplot of the average variation of the theta parameter

Figure 4 presents the distribution of average efficiency scores across clusters. While clusters 0 and 3 exhibit the highest median efficiency levels, Cluster 2 shows the lowest performance and the greatest variability. Clusters 1 and 4 occupy intermediate positions but differ in dispersion and trajectory direction.

Figure 5 summarizes the standardized characteristics associated with each cluster. The heatmap reveals how demographic, academic background, and institutional variables interact with efficiency trajectories to form distinct intersectional profiles.

Based on Figures 4 and 5, the distinct intersectional profiles were identified and defined. The description of each group is as follows:

Cluster 0: High and stable academic trajectory

Cluster 0 is characterized by consistently high academic performance (Mean: 0.96, Trend: 0.003) and low variability across semesters (Std. Deviation:0.01). Students in this cluster maintain efficiency levels close to the upper bound of the cohort, with a slight but steady improvement over time, indicating early adaptation to academic demands and sustained effective study practices.

From a social and institutional perspective, this cluster shows a distinct composition. Male students are under-represented, while students from rural backgrounds and first-generation students are over-represented. Despite these characteristics, which are often associated with increased educational risk, students in this cluster exhibit stable, high academic trajectories. This pattern is accompanied by a strong concentration in a specific academic school and a moderate association with campus regions, suggesting that institutional and program-level contexts play a compensatory role in supporting sustained academic performance.

Overall, Cluster 0 represents students who require limited academic intervention and benefit from well-aligned institutional environments that enable consistent academic success.

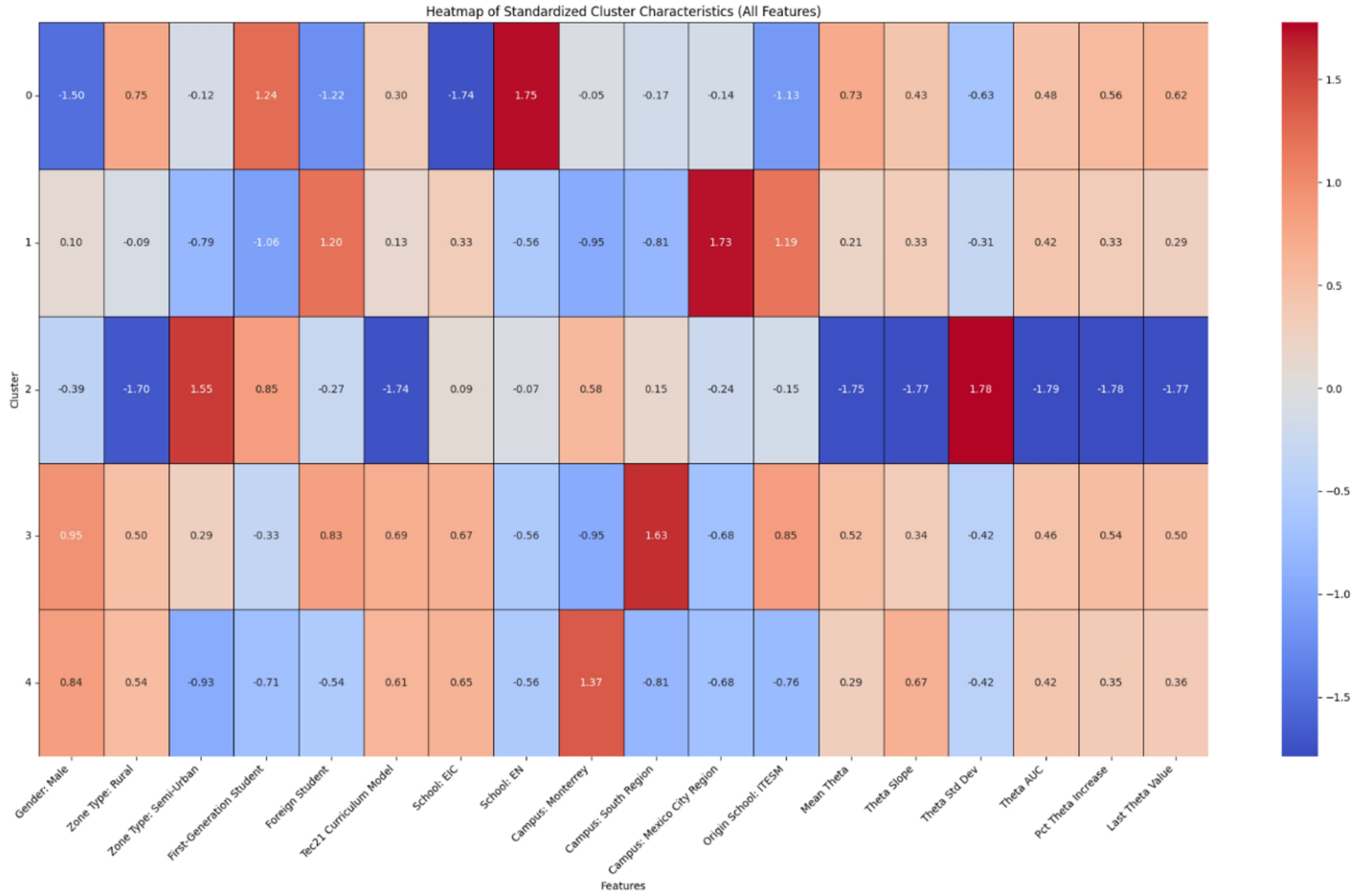


Figure 5: Characteristics heatmap of the features included

Cluster 1: Stable academic outcomes without acceleration

Cluster 1 is characterized by moderately high academic performance (Mean = 0.94, Trend = 0) that remains largely stable over time (Std. Deviation = 0.015). Students in this group maintain efficiency levels above the population average, though consistently below those of the highest performing clusters. Their academic trajectories are essentially flat, showing little systematic improvement or decline across semesters, with moderate variability and some lower-performing cases.

The social and institutional profile of Cluster 1 is marked by an over-representation of out-of-state students and a slight under-representation of students from rural and semi-urban backgrounds. In addition, this cluster shows a strong concentration in specific campus regions and academic schools, indicating a high degree of institutional alignment.

From an educational perspective, this combination of stable academic outcomes and strong institutional concentration suggests that students in Cluster 1 operate within supportive environments that sustain performance but do not appear to promote further academic acceleration. These students consistently meet academic expectations but show limited evidence of progressive improvement, suggesting the potential value of enrichment strategies that foster deeper learning and academic growth rather than remedial support.

Cluster 2: Declining and vulnerable academic trajectory

Cluster 2 represents the most critical profile when academic and social dimensions are considered jointly. From an academic perspective, students in this cluster exhibit the lowest average performance across all groups (Mean = 0.91), the highest variability (Std. Deviation = 0.025), and a clearly negative progression over time (Trend = -0.004). Their academic trajectories deteriorate as semesters advance, and outcomes at the end of the program remain the lowest observed among all clusters. This pattern reflects persistent difficulties in sustaining academic performance under increasing curricular demands.

The social and institutional configuration of this cluster further reinforces its vulnerability. Male students are slightly under-represented, while students from semi-urban backgrounds are strongly over-represented. This prevalence highlights a critical “vulnerability limbo” because, unlike rural students, who are often the primary focus of targeted social interventions, semi-urban populations may receive diluted institutional support. Unlike other groups, Cluster 2 does not display compensatory over-representation of first-generation or out-of-state students that might buffer academic risk. Institutionally, this cluster is markedly under-represented across multiple academic schools and campus regions, indicating limited alignment with supportive academic environments.

From an educational standpoint, the combination of declining academic trajectories, high instability, and unfavorable institutional positioning suggests restricted access to effective academic support structures. The absence of social or institutional buffering effects makes this group particularly vulnerable to academic disengagement and attrition. Consequently, Cluster 2 constitutes the primary target for early identification, intensive academic support, and retention-focused interventions.

Cluster 3: High and continuously improving academic trajectory

Cluster 3 is characterized by high academic performance combined with sustained positive progression over time. Students in this group maintain performance levels comparable to those of the highest performing cluster (mean = 0.97), while continuing to improve as they advance through the program (Trend = 0.004). Variability in academic outcomes is low (Std. Deviation = 0.009), indicating consistent learning and stable academic behavior across semesters. Final academic outcomes remain high and are sustained through program completion.

The social composition of Cluster 3 is relatively balanced, with a clear over-representation of male students and a slight over-representation of students from rural backgrounds, alongside a moderate presence of first-generation students. Institutionally, this cluster is over-represented in specific campus regions associated with stronger academic outcomes, suggesting alignment with supportive learning environments.

From an educational perspective, the convergence of stable social conditions and favorable institutional contexts appears to facilitate continuous academic consolidation. Students in Cluster 3 not only perform well but also refine their academic skills and learning strategies over time, demonstrating effective responses to increasing curricular complexity. This profile reflects a resilient and self-reinforcing learning trajectory within the program.

Cluster 4: Improving academic trajectory

Cluster 4 shows significant academic improvement over time. Although students in this group do not begin with the highest performance levels, their trajectories display the most pronounced positive progression among all clusters (Trend = 0.006). Average academic outcomes are moderately high (Mean = 0.95), variability is moderate to low (Std. Deviation = 0.012), and final performance levels exceed the cohort average, indicating convergence toward stronger academic outcomes by program completion.

The social profile of Cluster 4 shows an over-representation of male students, an under-representation of students from semi-urban backgrounds, and a slight under-representation of first-generation students. Institutionally, this cluster exhibits clear concentration within specific academic schools and campus regions, suggesting effective alignment with environments that facilitate academic development.

From an educational perspective, students in Cluster 4 appear to progressively benefit from academic exposure, institutional resources, and learning experiences. Their improvement over time reflects successful adaptation and sustained engagement, highlighting the importance of developmental trajectories in which students close initial performance gaps through continuous participation.

Intersectional Analysis

The five clusters are primarily distinguished by how students' academic performance evolves, rather than by initial performance levels alone. Clusters 0 and 3 both represent high-performing students, but with different patterns. Cluster 0 is defined by early academic adaptation and stable performance across semesters, while Cluster 3 combines high performance with continued improvement. Together, these clusters highlight two forms of academic success: sustained excellence from early stages and progressive consolidation over time.

Clusters 1 and 4 occupy intermediate positions but exhibit distinct trajectories. Cluster 1 is characterized by stable but largely unchanged performance, with students consistently meeting academic expectations without notable improvement. This makes the cluster noteworthy for its potential responsiveness to academic enrichment rather than remediation. In contrast, Cluster 4 is defined by strong academic growth. Although students in this group do not start at the highest performance levels, they show the greatest improvement over time and converge toward above-average outcomes by the end of the program.

Cluster 2 stands out as the most critical group. It is the only cluster showing a clear decline in academic performance over time, combined with high variability and limited institutional alignment. Unlike other clusters, students in this group do not benefit from stabilizing or compensatory conditions, making them particularly vulnerable to academic disengagement and attrition. This cluster, therefore, represents the primary focus for early identification and targeted academic support.

In terms of intersectionality, cluster structure highlights how academic performance emerges from the intersection of personal characteristics and institutional context, rather than from individual attributes in isolation. Clusters 0 and 3 illustrate that high academic performance can be sustained under different intersectional configurations. Cluster 0 combines high and stable performance with a higher presence of first-generation and rural students, suggesting that supportive institutional environments can offset structural disadvantages. Cluster 3, in contrast, pairs high performance with more advantaged contextual positioning, where social stability and favorable institutional placement jointly reinforce continuous academic improvement.

Clusters 1 and 4 show how similar performance levels can arise from different intersectional pathways. Cluster 1 is characterized by stable but non-progressive academic outcomes within strongly structured institutional settings and a higher presence of out-of-state students, indicating adequacy without acceleration. Cluster 4, however, demonstrates that students with fewer structural disadvantages and strong institutional alignment can translate sustained engagement into marked academic improvement over time, even if their initial performance is not among the highest.

Cluster 2 represents the most critical intersectional configuration, where multiple disadvantages converge. Students in this group combine declining academic trajectories with unfavorable institutional positioning and limited compensatory social characteristics. The interaction of these factors amplifies vulnerability, leading to high variability and worsening outcomes over time. This cluster illustrates how academic risk is produced not by single characteristics, but by the cumulative and interacting effects of personal background and institutional context.

DISCUSSION

The integration of quantitative intersectional analysis with dynamic efficiency measurement reveals patterns of educational performance that conventional analytical approaches miss. By combining Window DEA with Gaussian Mixture Modeling, this study demonstrates that academic efficiency in higher education is neither uniformly distributed nor determined by isolated demographic characteristics. Instead, efficiency emerges from the complex interplay between students' intersecting social identities and their institutional environments, producing distinct trajectories that evolve differentially over time.

Theoretical implications

This study makes three important theoretical contributions to the intersection of educational equity and efficiency research. First, it operationalizes intersectionality as a quantitative framework for efficiency analysis, moving beyond single-axis demographic comparisons to reveal how multiple identities interact to shape resource utilization patterns. The identification of five distinct efficiency clusters demonstrates that students with similar demographic profiles may follow markedly different academic trajectories depending on their institutional positioning and the specific configuration of their intersecting identities.

These findings align with and extend prior quantitative intersectional work. Keller et al. (2023) and Prior et al. (2025), both employing MAIHDA, found that between-stratum variation in student achievement is largely driven by additive rather than interactive effects of demographic characteristics. The present study complements this by showing that when efficiency trajectories, rather than static achievement scores, are the outcome of interest, meaningful heterogeneity persists across intersectional profiles even after accounting for additive effects. Similarly, Hanauer et al. (2025) identified four distinct identity orientations among STEM students using hierarchical clustering, concluding that student positionalities are far more complex than single-axis analyses suggest. Our five-cluster solution echoes this finding, though it goes further by linking those positionalities directly to resource utilization patterns over time, a dimension absent from Hanauer et al.'s analysis.

Second, the findings challenge the assumption that educational efficiency is primarily a function of individual characteristics or institutional quality alone. The divergent trajectories of Clusters 0 and 2, both containing students from structurally disadvantaged backgrounds but experiencing opposite academic outcomes, illustrate that efficiency arises from the interaction between student characteristics and institutional context. This suggests that traditional efficiency models that treat student populations as homogeneous or control for demographics through simple dummy variables may systematically misidentify sources of inefficiency.

This finding resonates with Temoso et al. (2023), who, using a network DEA framework, demonstrated that teaching and research efficiency differ substantially within the same institution, underscoring that efficiency is not a uniform property of an institution but varies across processes and contexts. Likewise, Tran et al. (2022) found that institutional type, public versus private, moderates efficiency in Vietnamese universities, suggesting that organizational context shapes how inputs are converted into outcomes. The present study extends this logic to the student level: just as institutional type conditions efficiency at the system level, the intersection of students' social characteristics and their program placement conditions efficiency at the individual level. Critically, however, unlike Tran et al. (2022) and Temoso et al. (2023), who treat student populations as largely homogeneous within institutional categories, our results show that even within a single program, students occupying different intersectional positions experience systematically different efficiency trajectories.

Third, the temporal dimension of efficiency revealed through Window DEA highlights the inadequacy of static, cross-sectional approaches to educational evaluation. The distinction between stable high performance (Cluster 0), continuous improvement (Clusters 3 and 4), stable adequacy (Cluster 1), and progressive decline (Cluster 2) demonstrates that efficiency is a dynamic process that unfolds over time. This has important implications for when and how institutions should intervene to support student success.

The dynamic patterns observed here contrast with the predominantly static designs that dominate efficiency research in education. Kounetas et al. (2023), examining 643 Greek secondary schools over 18 years, found persistent inefficiencies with limited reform impacts, but their analysis aggregated efficiency at the school level across years rather than tracing individual-level trajectories. Zhou et al. (2024) used the Malmquist index to capture efficiency change over time across Chinese provinces, identifying technological progress as the primary driver; however, this approach collapses individual heterogeneity into regional averages. By contrast, Window DEA applied at the student level reveals that efficiency change is not uniform: some students improve continuously (Clusters 3 and 4), others plateau (Cluster 1), and others deteriorate (Cluster 2). This trajectory-level granularity represents a meaningful advancement over aggregate temporal analyses and suggests that the timing and targeting of institutional interventions matter as much as their content.

Practical implications and policy recommendations

The intersectional efficiency profiles identified in this study have direct implications for educational policy and institutional practice, particularly for programs serving underrepresented and economically disadvantaged students. The five clusters require fundamentally different institutional responses.

Cluster 2 emerges as the highest-priority intervention, requiring early identification systems and intensive, sustained academic support. The combination of declining trajectories, high variability, and limited institutional alignment suggests that students in this cluster need comprehensive support that addresses both academic skills and institutional integration. Critically, the over-representation of semi-urban students in this cluster highlights a policy blind spot: unlike rural students, who often receive targeted support, semi-urban populations may fall through the cracks of binary urban-rural intervention frameworks.

The identification of semi-urban students as a particularly vulnerable group within Cluster 2 is a finding with few direct parallels in the efficiency literature, which typically segments populations by institutional type or geographic region rather than by students' origin characteristics. Chiariello et al. (2022), studying Italian primary and secondary schools, found significant North–South efficiency disparities driven by contextual factors including poverty and institutional quality, but did not examine how students' social backgrounds interact with these regional effects. Muniz et al. (2024), focusing on Brazilian schools, identified infrastructure elements such as libraries and computer labs as key determinants of efficiency, yet did not consider how these resources differentially benefit students from distinct social backgrounds. Our results suggest that the policy-relevant unit of analysis should not be the institution or the region alone, but the intersection of student characteristics and institutional environment, since the same institutional resources may produce very different efficiency outcomes depending on who is being served.

Cluster 1 presents a different challenge: these students consistently meet academic expectations but show limited growth over time. Rather than remediation, this group would benefit from academic enrichment initiatives designed to promote deeper engagement and skill development. The strong institutional concentration within this cluster suggests that program-level interventions such as enhanced research opportunities, advanced coursework options, or international experiences may be more effective than individual-level support. Clusters 3 and 4 demonstrate that institutional alignment and supportive environments can facilitate continued academic development, even among students who do not start at the highest levels of performance. This suggests that strategic placement and early connection to effective academic communities may be as important as direct academic support services.

The findings have important implications for how institutions allocate limited support resources. Traditional approaches that distribute resources uniformly across all scholarship recipients or target students solely on the basis of single demographic characteristics, such as first-generation status, may be inefficient (Adamecz-Völgyi et al., 2020; Herbaut and Geven, 2020). The intersectional profiles reveal that vulnerability and need for support are determined by configurations of characteristics rather than isolated attributes. For example, Cluster 0 demonstrates that first-generation and rural students can achieve sustained high performance when positioned within supportive institutional environments. This suggests that investments in institutional capacity, such as strengthening academic programs, developing mentoring networks, and creating inclusive campus communities, may be as important as direct student services. Conversely, Cluster 2's under-representation across multiple academic schools and regions indicates that some students lack access to these supportive contexts, suggesting the need for deliberate efforts to ensure equitable distribution of students across institutional resources.

The finding that first-generation and rural students in Cluster 0 achieve sustained high performance when positioned within supportive institutional environments challenges a common assumption in the efficiency literature that these demographic groups are inherently associated with lower educational efficiency. Ulkhaq et al. (2024), analyzing schools across six South-East Asian countries using PISA data, found that ICT-related resources are positively associated with efficiency, but treated student populations as homogeneous within national and school-level categories. Taleb et al. (2023), in their examination of Taiwanese universities, focused exclusively on institutional-level super-efficiency, without considering student composition. Neither study would be able to detect the compensatory dynamic observed here, where institutional context offsets the risk factors typically associated with first-generation and rural status. This has a direct policy implication: efficiency-improving investments directed at institutional environments, academic programs, mentoring structures, and campus communities may yield higher returns than equivalent investments in student-level remediation alone, particularly for students whose intersectional profiles combine structural disadvantage with high potential.

The temporal patterns revealed by Window DEA suggest that early identification is critical but insufficient on its own. While Cluster 2 students show relatively high initial efficiency (around 0.94), their trajectories begin to diverge by the middle semesters and continue to deteriorate thereafter. This indicates that intervention systems should monitor trajectory patterns rather than just absolute performance levels and should be designed to identify students experiencing declining efficiency even when their current performance remains acceptable. Moreover, the intersection-based clustering approach suggests that risk prediction models should move beyond single demographic indicators to consider configurations of characteristics. A first-generation student from a rural background enrolled in a strongly supportive academic program (likely Cluster 0) faces very different risks than a semi-urban student with weak institutional alignment (likely Cluster 2), even though traditional risk models might flag both.

The over-representation of certain clusters within specific academic schools and campus regions raises questions about whether program structures inadvertently concentrate risk or advantage. If particular academic programs or campus locations consistently produce better outcomes for students with certain intersectional profiles, institutions should examine what features of these environments are protective and whether they can be deliberately cultivated elsewhere. Similarly, the finding that institutional context can buffer structural disadvantages (as in Cluster 0) suggests that program design should prioritize not just admitting diverse students but ensuring they have equitable access to high-quality academic environments. Admissions decisions, program placement, and campus assignment should be informed by understanding which environments best support students with particular intersectional profiles.

Limitations

Several limitations should be acknowledged when interpreting these findings. First, the analysis is based on data from a single scholarship program at one institution in Mexico. While the Líderes del Mañana program serves a diverse, economically disadvantaged population, the specific intersectional configurations and their relationships to efficiency may differ across other institutional contexts or national settings. The transferability of findings to other contexts should be empirically verified rather than assumed. Second, the intersectional profiles identified through GMM reflect the particular set of identity dimensions and institutional characteristics available in the dataset. Other important axes of identity and marginalization, including race, ethnicity, indigenous status, disability, and language background, were not available for analysis but may play critical roles in shaping educational efficiency in other contexts. The clusters identified here represent patterns within the observed data rather than exhaustive categories of intersectional experience.

Third, the efficiency scores derived from Window DEA are inherently relative, comparing students to their peers within specific time windows rather than to absolute performance standards. While this approach is methodologically appropriate for identifying differential resource utilization, it does not directly address questions about whether overall resource levels are adequate or whether all students are achieving desired learning outcomes in absolute terms. Fourth, the causal mechanisms underlying the observed efficiency patterns cannot be definitively established through the clustering and efficiency analysis employed here. While the findings reveal systematic associations between intersectional profiles and efficiency trajectories, the specific processes through which these patterns emerge, whether through differential access to resources, variations in institutional support quality, differences in academic preparation, or other mechanisms, require further investigation through complementary research designs.

Finally, the study focuses on students who remained enrolled in the program across multiple semesters, potentially introducing survivorship bias. Students who left the program early, whether due to academic difficulty, financial constraints, or other reasons, are underrepresented in the later stages of the efficiency analysis. This may lead to underestimation of efficiency gaps and over-optimistic assessments of program effectiveness, particularly for the most vulnerable groups. However, it is important to note that the Líderes del Mañana program has a very low attrition rate, with approximately 1% of students leaving before completion during the study period.

CONCLUSION

This study demonstrates that integrating quantitative intersectional methods with dynamic efficiency analysis reveals patterns of educational performance that remain hidden when student populations are treated as homogeneous. By applying Window DEA and Gaussian Mixture Modeling to longitudinal data from the Líderes del Mañana scholarship program, we identified five distinct intersectional efficiency profiles characterized by different patterns of academic trajectory, institutional positioning, and social background. Critically, structural disadvantages do not deterministically produce poor outcomes; institutional environments can buffer or amplify disadvantage depending on the specific configuration of intersecting identities and institutional factors.

These findings carry direct policy implications. Promoting educational equity requires moving beyond uniform interventions or single-axis targeting toward differentiated support strategies responsive to students' complex, multidimensional positionalities. Efficiency in higher education emerges from the dynamic interaction between students' multiple identities and the institutional contexts they navigate. Several directions remain open for future research, including mixed-methods investigation of the mechanisms behind these patterns, expansion to additional identity axes such as race, ethnicity, and disability status, and cross-institutional comparative studies. Methodological advances, including Bayesian methods, panel DEA, and machine learning approaches, could further refine intersectional efficiency analysis. Ultimately, this study provides both a methodological template and an empirical foundation for demonstrating that efficiency and equity are not competing values but complementary objectives that require an intersectional understanding to be achieved simultaneously.

REFERENCES

Adamecz-Völgyi, A., Henderson, M. and Shure, N. (2020) 'Is 'first in family' a good indicator for widening university participation?', *Economics of Education Review*, Vol. 78, p. 101974. https://doi.org/10.1016/j.econedurev.2020.102038

Agosto, V. and Roland, E. (2018) 'Intersectionality and educational leadership: A critical review', *Review of Research in Education*, Vol. 42, No. 1, pp. 255–285. https://doi.org/10.3102/0091732X18762433

Alvarez-Hernandez, L. R. (2021) 'Teaching note—Teaching intersectionality across the social work curriculum using the intersectionality analysis cluster', *Journal of Social Work Education*, Vol. 57, No. 1, pp. 181–188. https://doi.org/10.1080/10437797.2020.1713944

Bauer, G. R., Mahendran, M., Walwyn, C. and Shokoohi, M. (2022) 'Latent variable and clustering methods in intersectionality research: Systematic review of methods applications', *Social Psychiatry and Psychiatric Epidemiology*, Vol. 57, No. 2, pp. 221–237. https://doi.org/10.1007/s00127-021-02195-6

Bešić, E. (2020) 'Intersectionality: A pathway towards inclusive education?', *Prospects*, Vol. 49, No. 3–4, pp. 111–122. https://doi.org/10.1007/s11125-020-09461-6

Pokropek, A., Borgonovi, F. and Jakubowski, M. (2015) 'Socio-economic disparities in academic achievement: A comparative analysis of mechanisms and pathways', *Learning and Individual Differences*, Vol. 42, pp. 105–119. https://doi.org/10.1016/j.lindif.2015.07.011

Byrd, K. M., Kahle, L. L., Peguero, A. A. and Popp, A. M. (2015) 'Social control and intersectionality: A multilevel analysis of school misconduct, location, race, ethnicity, and sex', *Sociological Spectrum*, Vol. 35, No. 2, pp. 109–135. https://doi.org/10.1080/02732173.2014.1000552

Charnes, A., Cooper, W. W. and Rhodes, E. (1978) 'Measuring the efficiency of decision making units', *European Journal of Operational Research*, Vol. 2, No. 6, pp. 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

Chiariello, V., Rotondo, F. and Scalera, D. (2022) 'Efficiency in education: Primary and secondary schools in Italian regions', *Regional Studies*, Vol. 56, No. 10, pp. 1729–1743. https://doi.org/10.1080/00343404.2021.2005245

Garnett, B. R., Masyn, K. E., Austin, S. B., Miller, M., Williams, D. R. and Viswanath, K. (2014) 'The intersectionality of discrimination attributes and bullying among youth: An applied latent class analysis', *Journal of Youth and Adolescence*, Vol. 43, No. 8, pp. 1225–1239. https://doi.org/10.1007/s10964-013-0073-8

Hanauer, D. I., Zhang, T., Graham, M. and Hatfull, G. (2025) 'Who is in our STEM courses and how do we know? Student self-descriptions, intersectionality and inclusive education', *CBE—Life Sciences Education*, Vol. 24, No. 1, pp. ar9(1–20). https://doi.org/10.1187/cbe.24-02-0078

Harris, J. C. and Patton, L. D. (2019) 'Un/doing intersectionality through higher education research', *The Journal of Higher Education*, Vol. 90, No. 3, pp. 347–372. https://doi.org/10.1080/00221546.2018.1536936

Herbaut, E. and Geven, K. (2020) 'What works to reduce inequalities in higher education? A systematic review of the (quasi-) experimental literature on outreach and financial aid', *Research in Social Stratification and Mobility*, Vol. 65, p. 100442. https://doi.org/10.1016/j.rssm.2019.100442

Johnes, J. (2006) 'Measuring teaching efficiency in higher education: An application of data envelopment analysis to economics graduates from UK universities 1993', *European Journal of Operational Research*, Vol. 174, No. 1, pp. 443–456. https://doi.org/10.1016/j.ejor.2005.02.044

Keller, L., Lüdtke, O., Preckel, F. and Brunner, M. (2023) 'Educational inequalities at the intersection of multiple social categories: An introduction and systematic review of the multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) approach', *Educational Psychology Review*, Vol. 35, No. 1, p. 31. https://doi.org/10.1007/s10648-023-09733-5

Kounetas, K., Androulakis, G., Kaisari, M. and Manousakis, G. (2023) 'Educational reforms and secondary school's efficiency performance in Greece: A bootstrap DEA and multilevel approach', *Operational Research*, Vol. 23, No. 1, p. 9. https://doi.org/10.1007/s12351-023-00764-y

Leckie, A. and Buser De, M. (2020) 'The power of an intersectionality framework in teacher education', *Journal for Multicultural Education*, Vol. 14, nN. 1, pp. 117–127. https://doi.org/10.1108/JME-07-2019-0059

Macias, A. and Stephens, S. (2019) 'Intersectionality in the field of education: A critical look at race, gender, treatment, pay, and leadership', *Journal of Latinos and Education*, Vol. 18, No. 2, pp. 164–170. https://doi.org/10.1080/15348431.2017.1383912

Maina-Okori, N. M., Koushik, J. R. and Wilson, A. (2018) 'Reimagining intersectionality in environmental and sustainability education: A critical literature review', The Journal of Environmental Education, Vol. 49, No. 4, pp. 286–296. https://doi.org/10.1080/00958964.2017.1364215

Mergoni, A., Emrouznejad, A. and De Witte, K. (2025) 'Fifty years of data envelopment analysis', *European Journal of Operational Research*, Vol. 326, no. 3, pp. 389–412. https://doi.org/10.1016/j.ejor.2024.12.049

Muniz, R. D. F., Andriola, W. B., Muniz, S. M. M. and Thomaz, A. C. F. (2024) 'The use of data envelopment analysis (DEA) to estimate the educational efficiency of Brazilian schools', *Journal of Applied Research on Industrial Engineering*, Vol. 11, No. 1, pp. 93–102. https://doi.org/10.22105/jarie.2021.308815.1388

Nichols, S. and Stahl, G. (2019) 'Intersectionality in higher education research: A systematic literature review', *Higher Education Research and Development*, Vol. 38, No. 6, pp. 1255–1268. https://doi.org/10.1080/07294360.2019.1638348

Prior, L., Evans, C., Merlo, J. and Leckie, G. (2025) 'Sociodemographic inequalities in student achievement: An intersectional multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA)', *Sociology of Race and Ethnicity*, Vol. 11, No. 3, pp. 351–369. https://doi.org/10.1177/23326492241267251

Programa LDM | Líderes del Mañana (n.d.) Programa LDM, Available at: https://lideresdelmanana.tec.mx/programa-ldm [Accessed 9 December 2025]

Pugach, M. C., Gomez-Najarro, J. and Matewos, A. M. (2019) 'A review of identity in research on social justice in teacher education: What role for intersectionality?', *Journal of Teacher Education*, Vol. 70, No. 3, pp. 206–218. https://doi.org/10.1177/0022487118760567

Reinwald, S. and Annen, S. (2023) 'Influence of gender and prior education intersectionality on further education investments and job satisfaction', *SAGE Open*, Vol. 13, No. 2, p. 21582440231180095. https://doi.org/10.1177/21582440231180095

Robert, S. A. and Yu, M. (2018) 'Intersectionality in transnational education policy research', *Review of Research in Education,* Vol. 42, No. 1, pp. 93–121. https://doi.org/10.3102/0091732X18759305

Robson, K. L., Anisef, P., Brown, R. S. and Parekh, G. (2014) 'The intersectionality of postsecondary pathways: The case of high school students with special education needs', *Canadian Review of Sociology*, Vol. 51, No. 3, pp. 193–215. https://doi.org/10.1111/cars.12044

Slominski, T., Odeleye, O. O., Wainman, J. W., Walsh, L. L., Nylund-Gibson, K. and Ing, M. (2024) 'Calling for equity-focused quantitative methodology in discipline-based education research: An introduction to latent class analysis', *CBE—Life Sciences Education*, Vol. 23, No. 4, p. es11. https://doi.org/10.1187/cbe.24-01-0023

Sun, Y., Wang, D., Yang, F. and Ang, S. (2023) 'Efficiency evaluation of higher education systems in China: A double frontier parallel DEA model', *Computers and Industrial Engineering*, Vol. 176, p. 108979. https://doi.org/10.1016/j.cie.2023.108979

Taleb, M., Khalid, R., Attallah, M., Kareem, Q. A. and Ramli, R. (2023) 'Assessing and ranking the performance of higher education institutions: A non-radial super efficiency DEA approach', *International Journal of Education Economics and Development*. Available at: https://www.inderscienceonline.com/doi/10.1504/IJEED.2023.129890 [Accessed 3 March 2026]

Temoso, O., Tran, C. - T. T. D. and Myeki, L. (2023) 'Network DEA efficiency of South African higher education: Evidence from the analysis of teaching and research at the university level', *Journal of Further and Higher Education*, Vol. 47, No. 8, pp. 1009–1026. https://doi.org/10.1080/0309877X.2023.2209799

Tran, T.-V., Pham, T. P., Nguyen, M.-H., Do, L.-T. and Pham, H.-H. (2022) 'Economic efficiency of higher education institutions in Vietnam between 2012 and 2016: A DEA analysis', *Journal of Applied Research in Higher Education*, Vol. 15, No. 1, pp. 199–212. https://doi.org/10.1108/JARHE-06-2021-0238

Ulkhaq, M. M., Oggioni, G. and Riccardi, R. (2024) 'Two-stage super-efficiency model for measuring efficiency of education in South-East Asia', *Decisions in Economics and Finance*, Vol. 47, No. 2, pp. 513–543. https://doi.org/10.1007/s10203-024-00453-1

Zhou, G., Zhang, F. and Fei, Y. (2024) 'Research on the development efficiency of educational science and technology in China: An approach from three-stage DEA Malmquist model', *Technology Analysis and Strategic Management*, Vol. 36, No. 6, pp. 1067–1082. https://doi.org/10.1080/09537325.2022.2069556

IMBALANCED MULTI-CLASS PREDICTION OF STUDENT DROP-OUT AND GRADUATION: A SYSTEMATIC LITERATURE REVIEW

Ridwan Setiawan[1,2]✉
Edi Noersasongko[2]
Abdul Syukur[2]
Fikri Budiman[2]
Dede Kurniadi[1]

[1]Institut Teknologi Garut, Indonesia

[2]Universitas Dian Nuswantoro, Indonesia

✉ ridwan.setiawan@itg.ac.id

ABSTRACT

Student study status prediction, including drop-out and graduation, is a widely studied topic in higher education. Yet, evidence across studies remains difficult to compare due to differences in targets, imbalance treatment, metrics, and validation strategies. This systematic literature review synthesizes 70 peer reviewed articles published between 2017 and 2025 that apply machine learning or deep learning to predict study outcomes under class imbalance. Results reveal a strong dominance of binary targets, while multi class experiments are relatively rare, though they better reflect institutional categories and expose larger performance gaps across classes. Reported imbalance handling includes data level resampling, algorithm level class weighting, and ensemble or hybrid designs, but many studies lack sufficient procedural detail. Evaluation practices vary considerably; studies reporting per-class measures and imbalance-aware metrics, such as macro F1 and balanced accuracy, provide more decision-relevant evidence than those relying mainly on accuracy. Validation strategies range from hold out and stratified cross validation to nested validation, temporal splits, and external testing, shaping the credibility of reported performance for deployment. We propose an integrative taxonomy linking target formulation, imbalance degree, handling strategy, and evaluation design to enhance intervention efficiency through capacity aware prioritization, while strengthening responsibility through transparent reporting, defensible validation, and explicit attention to minority class performance.

KEYWORDS

Higher education, imbalanced classification, learning analytics, machine learning, responsibility, student dropout

HOW TO CITE

Setiawan R., Noersasongko E., Syukur A., Budiman F., Kurniadi D. (2026) 'Imbalanced Multi-class Prediction of Student Drop-out and Graduation: A Systematic Literature Review', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 72–90. http://dx.doi.org/10.7160/eriesj.2026.190106

Article history
Received
January 12, 2026
Received in revised form
February 27, 2026
Accepted
March 12, 2026
Available on-line
March 31, 2026

Highlights

- *Synthesizes 70 open-access Scopus journal studies from 2017 to 2025 on predicting student drop-out and graduation using machine learning and deep learning.*
- *Finds strong dominance of binary targets, while multi-class targets remain rare yet better reflect institutional status categories and reveal larger per-class performance gaps.*
- *Shows imbalance handling is often under-reported; when reported, options include resampling, class weighting, cost-sensitive learning, and ensemble or hybrid designs, with implications for transparent and accountable study design.*
- *Highlights that accuracy alone is insufficient under imbalance; per-class metrics and imbalance-aware summaries improve decision relevance for efficient interventions and improve responsibility through clearer minority class reporting and more defensible evidence.*

INTRODUCTION

Student dropout and delayed graduation remain persistent challenges in higher education across different countries and institutional contexts. Their consequences extend beyond individual students and entail economic, social, and managerial implications for institutions and policymakers, particularly regarding service sustainability, academic quality, and accountability in higher education management (Csalódi and Abonyi, 2021; Véliz Palomino and Ortega, 2023; Quimiz-Moreira et al., 2025). Early identification of students at risk is therefore a strategic necessity because it enables interventions to be delivered in a timelier, more targeted, and more evidence-based manner (Setiawan et al., 2025). In this review, this practical need is closely linked to *efficiency*, understood as the extent to which predictive evidence can support the prioritization of interventions under constrained institutional resources. At the same time, *responsibility* refers to the defensibility, transparency, and fairness of the evidence used in educational decision support and scientific reporting. Accordingly, this review is positioned not merely as an inventory of modelling approaches, but as an examination of the study design decisions that shape whether research findings are useful for intervention prioritization and defensible for educational and scientific use.

In line with the growing availability of academic records and learning activity data, learning analytics and educational data mining are increasingly used to understand risk patterns and support data-driven decision-making. A systematic review by de Oliveira et al. (2021) showed that dropout risk is influenced by a combination of academic and nonacademic factors, while Andrade-Girón et al. (2023) confirmed the relevance of machine learning and deep learning for building pattern-based predictions and early warning systems. However, interpretability and variation in evaluation practices remain important methodological concerns.

Several review studies have mapped predictive approaches in the education domain, including the use of artificial neural networks in educational data mining by Okewu et al. (2021), data mining practices for academic performance prediction by Daza et al. (2022), and a review of graduation prediction that highlighted limitations in algorithm coverage, database selection, and the transparency of data collection procedures in earlier systematic literature reviews by Pelima et al. (2024). In the context of dropout prediction, Salinas-Chipana et al. (2024) reported a PRISMA-based review that showed the dominance of random forest models and confirmed the importance of academic, demographic, economic, and health-related attributes as predictors. Although the review literature has enriched the understanding of predictive methods and feature sets, class imbalance and evaluation consistency, especially in multiclass scenarios, have not yet been addressed as central issues in an integrated manner.

The distribution of labels in student retention data is commonly imbalanced. Cases of dropout are often substantially fewer than cases of persistence or graduation. As a result, a model may appear strong on aggregate metrics while remaining weak at detecting minority groups, which are often the primary targets of intervention. Budiman et al. (2022) and Villar and de Andrade (2024) emphasized that class imbalance is a recurring issue in educational data mining and requires both appropriate imbalance handling strategies and the use of more representative evaluation metrics. In this respect, Martins et al. (2021) noted that balance-sensitive metrics based on precision and recall, including the F1 score, are more informative than accuracy alone, particularly when the objective is to identify at-risk students in a fairer and more actionable way.

Beyond the challenge of imbalance, the complexity of academic status often requires richer target formulations than binary labels alone (Kurniadi et al., 2021). Categories such as still enrolled, graduated, delayed graduation, and dropout reflect different academic pathways and may call for different forms of intervention. Within this context, *efficiency* concerns whether predictive models can support the targeted allocation of mentoring and intervention resources so that unnecessary alarms do not consume institutional capacity and missed risk cases are managed in line with policy priorities. In parallel, *responsibility* concerns the adequacy of minority-class evaluation, methodological transparency, and the avoidance of biased conclusions arising from inappropriate validation designs or performance metrics that obscure failures to detect students at risk. Thus, evaluation quality is not merely a technical matter, but one with direct implications for the effectiveness of retention policies and the fairness of institutional services.

This study addresses this gap by providing a systematic literature review of the application of machine learning and deep learning techniques to predict student academic status in higher education, with particular attention to dropout and graduation outcomes and to the methodological challenges of multiclass classification under class imbalance. It makes three specific contributions. First, it offers a systematic framework for examining how prior studies formulate prediction targets as binary or multiclass problems and for clarifying the methodological consequences of these choices for model evaluation in higher education contexts (de Oliveira et al., 2021; Pelima, Sukmana, and Rosmansyah, 2024; Salinas-Chipana et al., 2024). Second, it comparatively synthesizes the class imbalance mitigation strategies reported in the literature, at both the data and algorithmic levels, and relates them to evaluation and validation practices that determine whether performance on minority classes can meaningfully support intervention prioritization and responsible educational decision support (Andrade-Girón et al., 2023; Martins et al., 2023; Villar and de Andrade, 2024). Third, as an integrative outcome, it proposes a taxonomy of study design decisions that links target formulation, imbalance severity, mitigation strategies, and the selection of performance metrics and validation schemes. This taxonomy is intended both as an audit framework for reviewing prior studies and as a guide for designing future research that is more transparent, accountable, and methodologically aligned with the complementary aims of *efficiency* and *responsibility* in educational science (Page et al., 2021; Rethlefsen et al., 2021).

The purpose of this manuscript is to systematically synthesize empirical evidence on predicting student dropout and graduation in higher education, with particular emphasis on the relationships among binary and multiclass target formulations, class imbalance handling strategies, and evaluation and validation practices. This synthesis leads to a taxonomy of study design decisions intended to support intervention efficiency and the responsible use of predictive models. In line with this objective, the review addresses six research questions: RQ1, what machine learning and deep learning algorithms have been used to predict student dropout and graduation; RQ2, how have targets been formulated as binary or multiclass classifications, and what implications do these choices have for modelling and reporting; RQ3, what class imbalance handling strategies have been reported at the data, algorithmic, and decision levels; RQ4, to what extent have ensemble and hybrid methods been used, and what application patterns emerge in this corpus; RQ5, which evaluation metrics and validation methods have been used, and how suitable are they for imbalanced data; and RQ6, what methodological gaps and challenges remain, and what recommendations can strengthen future research. The systematic review procedure and the reporting of the study selection flow follow the principles of Systematic Literature Review (SLR) and PRISMA to ensure procedural traceability and replicability (Kitchenham, 2004; Page et al., 2021).

This manuscript is organized as follows. The Materials and Methods section presents the systematic review procedure, search strategy, selection criteria, and data extraction and synthesis processes. The Results section presents the main findings in response to the research questions. The Discussion section relates these findings to the broader literature, explains their implications for intervention efficiency and responsible model use, and discusses the limitations of the review. Finally, the Conclusion section summarizes the principal contributions and outlines directions for future research.

MATERIALS AND METHODS

This study employs a systematic literature review to synthesize empirical evidence on predicting student dropout and graduation in higher education, with an emphasis on multi-class formulations under class imbalance conditions (Kitchenham, 2004; Page et al., 2021). The scope of the study and research questions were determined using the PICOC framework to maintain consistency in selection and extraction decisions and to link the population, intervention, comparison, outcome, and context elements to the research questions. The operationalization of Population, Intervention, Comparison, Outcomes, and Context (PICOC) and its mapping to the research questions are presented in Table 1.

PICOC Element	Operationalization	RQ	Motivation
Population	Student academic data in higher education institutions includes imbalanced binary and multi-class scenarios.	RQ1; RQ2; RQ5	Mapping data characteristics and variations in the number of classes to identify multi-class research gaps.
Intervention	Machine learning and deep learning algorithms, including single, ensemble, and hybrid configurations.	RQ1; RQ3; RQ4	Inventorying modelling approaches and consistency in handling imbalanced data.
Comparison	Baseline without handling imbalance; resampling technique variations; cost sensitivity; ensemble; hybrid; and target formulation variations.	RQ1; RQ2; RQ3; RQ4	Mapping variations in experimental design without making it a single claim of superiority.
Outcome	Aggregate metrics and per-class metrics, the confusion matrix, and computational efficiency are reported.	RQ5	Assessing the representativeness of metrics for minority classes and potential evaluation bias.
Context	Formal higher education, public or institutional datasets, focusing on dropout rates, academic performance, graduation, and early warning systems.	RQ1; RQ2; RQ3; RQ4; RQ5; RQ6	Maintaining domain relevance and facilitating cross-institutional application mapping.

Table 1: PICOC framework, motivation, dan RQ

The process of identifying, screening, and reporting the study selection flow follows PRISMA. A summary of the selection process and reasons for exclusion are presented in Figure 1 (Page et al., 2021; Rethlefsen et al., 2021).

Data source and search strategy

A literature search was conducted using Scopus as the sole database to ensure query repeatability and selection consistency across a single, cross-publisher, cross-disciplinary index. Scopus was selected because it provides standardized bibliographic metadata and DOI linkage across peer-reviewed journals, which supports protocol auditability and replicable retrieval under a fixed query string. This choice improves procedural consistency, yet it is also a limitation because reliance on a single index may introduce coverage bias and may under represent relevant journal outlets that are more visible through other curated indexes or discipline specific libraries (Mongeon and Paul-Hus, 2016; Baas et al., 2020) To enhance independent verifiability, the corpus was limited to open access journal articles with DOIs. This restriction enables readers to access full texts and verify data extraction, but it may exclude relevant evidence available outside open access or outside Scopus indexing. Future extensions of this review may integrate additional sources, such as Web of Science and discipline-focused libraries, to quantify overlap and assess whether distributional findings differ across indexes (Helbach et al., 2022).

The inquiries aim to encompass three dimensions: the modelling methodology, the higher education context, and outcomes on dropout or graduation rates. The inquiry was executed utilizing the advanced search function within the title, abstract, and keywords boxes, imposing a temporal constraint from 2015 to 2025, specifying journal source type, English language, and final publication status. The search was conducted on August 31, 2025, and updated on December 1, 2025. The Scopus query string used is listed in this section to ensure the reproducibility of the procedure.

> TITLE-ABS-KEY ((“machine learning” OR “deep learning” OR “neural network*” OR “artificial intelligence” OR “predictive analytics” OR “educational data mining”) AND (“student dropout” OR “dropout prediction” OR “student retention” OR “graduation prediction” OR “student graduation” OR “academic performance” OR “student success” OR “student attrition”) AND (“university” OR “college” OR “higher education”)) AND (LIMIT-TO (SRCTYPE,

"j")) AND (LIMIT-TO (OA, "all")) AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (LANGUAGE, "english"))

Inclusion and exclusion criteria

Inclusion and exclusion criteria are established before the selection process begins to prevent ad hoc changes. In summary, studies were included if they were conducted in a higher education context, used student academic data, had a prediction target related to institutional study status, employed machine learning or deep learning as the primary approach, and reported evaluations with clear metrics and procedures. Studies were excluded if they focused on a non-institutional context, such as MOOCs, courses, training, or semester completion; were not aligned with the definition of study status outcome; did not use ML or DL; did not report evaluations; or were from discontinued journals. The criteria are presented in Table 2.

Type	Criteria
Inclusion	(1) the context of higher education and the use of student academic data. (2) the prediction or classification target relates to institutional study status, including dropout or study cessation, retention, graduation, timeliness, or other academic risk categories explicitly mapped to student study status. (3) Using ML or DL as the primary approach. (4) Providing sufficient information on the target formulation, features, or dataset for synthesis. (5) Reporting model performance with clear evaluation metrics or procedures.
Exclusion	(1) Focus on levels apart from student study status. (2) Focus on MOOCs, courses, training, or semester completion. (3) the study status outcomes are not aligned with the definition provided by the institution. (4) Does not use ML or DL. (5) Does not report evaluation. (6) the study was published in a journal that has since been discontinued.

Table 2: Inclusion and exclusion criteria

Study selection process and records management

The selection was conducted in stages: title and abstract screening, followed by full-text assessment. Out of 845 records screened at the title and abstract stage, 268 studies proceeded to full-text assessment. Subsequently, 198 studies were excluded for documented reasons, leaving 70 for inclusion in the final corpus. The selection process and reasons for exclusion are summarized in Figure 1.

At the record management stage, records exported from Scopus are checked for duplicate entries using a combination of DOI, title, and bibliographic metadata. Entries identified as duplicates are retained as the most complete record to maintain consistency in corpus calculations at subsequent filtering stages. Screening and extraction were performed by one reviewer. Then decisions at critical points, especially exclusions at the full-text stage and the appropriateness of the exclusion reasons, were checked by a second reviewer. If there are differences in assessment at critical decision points, the final decision is determined through discussion until consensus is reached.

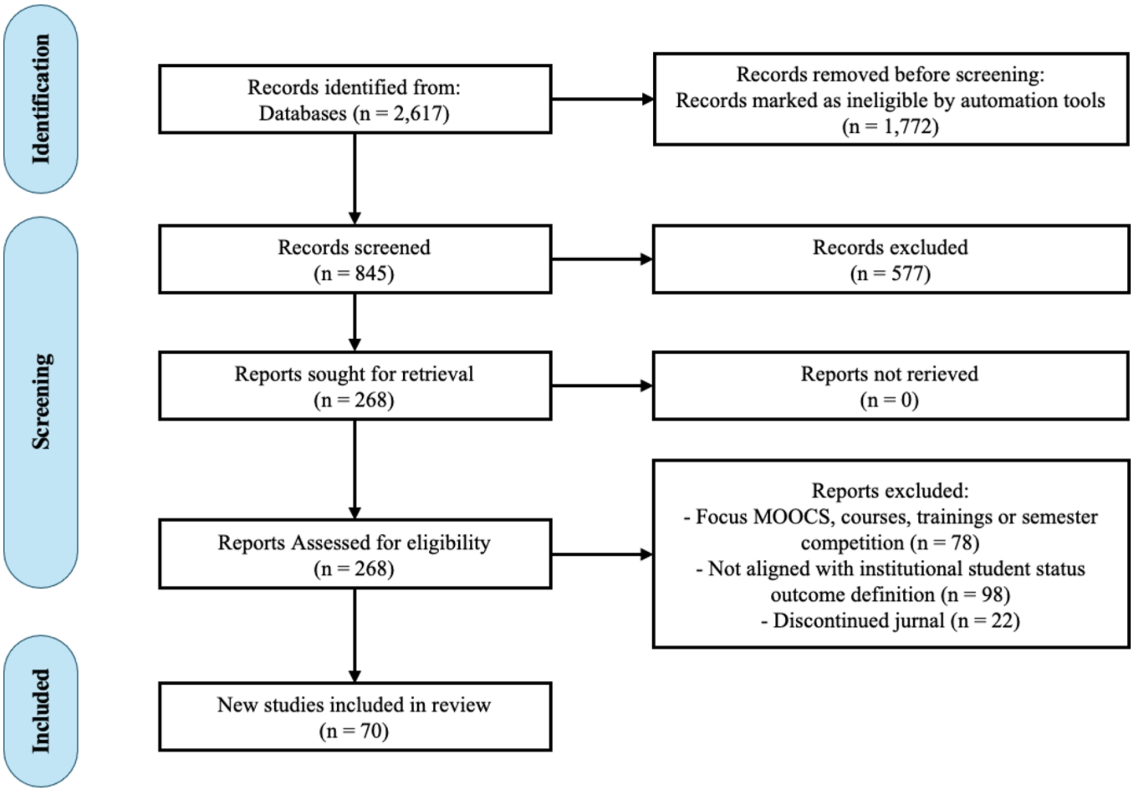

Identification
Records identified from:
Databases (n = 2,617)
Records removed before screening:
Records marked as ineligible by automation tools
(n = 1,772)
Screening
Records screened
(n = 845)
Records excluded
(n = 577)
Reports sought for retrieval
(n = 268)
Reports not rerieved
(n = 0)
Reports Assessed for eligibility
(n = 268)
Reports excluded:
- Focus MOOCS, courses, trainings or semester competition (n = 78)
- Not aligned with institutional student status outcome definition (n = 98)
- Discontinued jurnal (n = 22)
Included
New studies included in review
(n = 70)

Figure 1: PRISMA flow diagram of study identification

Corpus Study Description and Data Characteristics

The final corpus consists of 70 journal articles that meet the inclusion criteria. The publication range is from 2017 to 2025. The distribution of articles by year of publication is shown in Figure 2 to illustrate the temporal distribution of research on predicting student study status within the corpus.

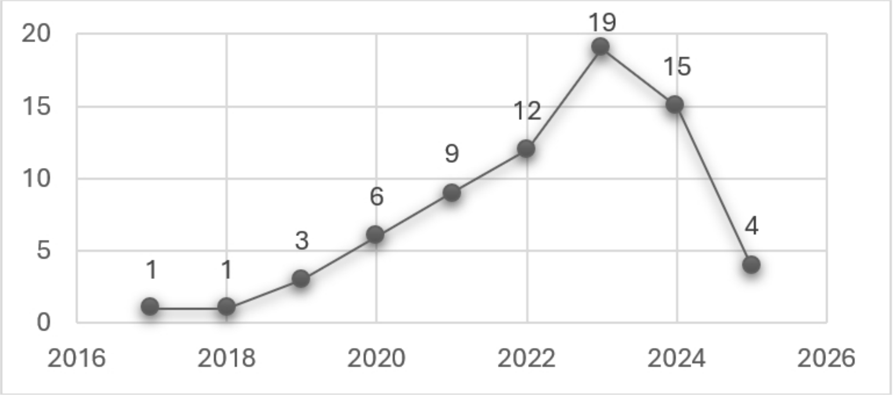


Figure 2: Distribution of articles by year of publication, 2017-2025

The distribution of articles by journal quartile is presented in Figure 3. This visualization describes the corpus, not the quality of each study, which is assessed through the primary studies' methodological components.

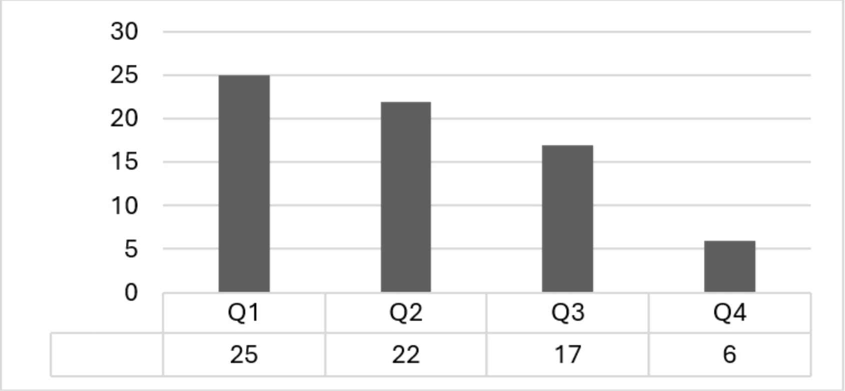


Figure 3: Article distribution across journal quartiles 2017-2025

The characteristics of the corpus are also reviewed based on the dataset's source, as this aspect affects the traceability of the experiment and the possibility of replication. Across the 70 studies, 62 (88.5%) relied on private institutional datasets, while 8 (11.5%) used public or open-access datasets. To support transparency in the reviewed corpus, a complete list of the 70 included studies is provided in Table 4.

The eight studies utilizing public or open-access datasets are not widely distributed across many sources but rather concentrated on a few reference datasets. Three studies used the Instituto Politécnico de Portalegre, Portugal dataset published through the UCI Machine Learning Repository by Realinho et al. (2021), namely the research by Goran et al. (2024), Martins et al. (2023), and Villar and de Andrade (2024). Two studies used the Tecnológico de Monterrey, Mexico dataset available through the open data repository by Alvarado-Uribe et al. (2022), namely Cañete-Sifuentes et al. (2023) and Gonzalez-Nucamendi et al. (2023). Additionally, Rovira et al. (2017) and Deleña et al. (2025) used datasets published alongside their articles, whereas Csalódi and Abonyi (2021) referenced the dataset's public location in the sources they cited. The implications of private dataset dominance for experiment reproducibility and generalization of findings are discussed in the Discussion section.

Data extraction and synthesis procedures

Data were extracted using a structured form to maintain traceability to the primary source. Extraction includes bibliographic information, data context, target definition, and number of classes, indication of class imbalance if reported, algorithms used, strategies for handling imbalance at both the data and algorithm levels, the use of ensembles or hybrid approaches, evaluation metrics, and data validation and splitting schemes. If available, information regarding hyperparameter tuning and computational efficiency is also recorded. This summary of extraction variables is used as the basis for descriptive and thematic synthesis.

The synthesis was conducted in two stages. The first stage involved a descriptive summary to map the distribution of findings according to the research questions. The second stage involved a cross-research question thematic synthesis to link study design decisions, including target formulation, degree of imbalance, mitigation strategies, and evaluation and validation plans.

No	Reff	Year	No	Reff	Year
1.	(Rabelo and Zárate, 2025)	2025	36.	(Haerani et al., 2023)	2023
2.	(Deleña et al., 2025)	2025	37.	(Mouchantaf and Chamoun, 2023)	2023
3.	(Oqaidi, Aouhassi, and Mansouri, 2025)	2025	38.	(Hoyos Osorio and Daza Santacoloma, 2023)	2023
.4	(Hooper, Ragland, and Artemiou, 2025)	2025	39.	(Niyogisubizo et al., 2022)	2022
5.	(Roslan et al., 2024)	2024	40.	(Vidal et al., 2022)	2022
6.	(Delogu et al., 2024)	2024	41.	(Segura, Mello and Hernández, 2022)	2022
7.	(Nguyen Thi Cam, Sarlan and Arshad, 2024)	2024	42.	(Barramuño, Meza-Narváez and Gálvez-García, 2022)	2022
8.	(Villar and de Andrade, 2024)	2024	43.	(Moreira da Silva et al., 2022)	2022
9.	(Vaarma and Li, 2024)	2024	44.	(Cannistrà et al., 2022)	2022
10.	(Goran et al., 2024)	2024	45.	(Nuanmeesri et al., 2022)	2022
11.	(Zanellati, Zingaro and Gabbrielli, 2024)	2024	46.	(Hammoodi and Al-Azawei, 2022)	2022
12.	(Okoye et al., 2024)	2024	47.	(Vega et al., 2022)	2022
13.	(Nagy and Molontay, 2024)	2024	48.	(Canto, De Oliveira and De Mattos Veroneze, 2022)	2022
14.	(Ndunagu et al., 2024)	2024	49.	(Yaqin, Rahardi, and Abdulloh, 2022)	2022
15.	(Setiadi et al., 2024)	2024	50.	(Rose and Mary.T, 2022)	2022
16.	(Herianto et al., 2024)	2024	51.	(Fernandez-Garcia et al., 2021)	2021
17.	(Darenoh, Bachtiar, and Perdana, 2024)	2024	52.	(Opazo et al., 2021)	2021
18.	(Sayed, 2024)	2024	53.	(Csalódi and Abonyi, 2021)	2021
19.	(Anagnostopoulos et al., 2024)	2024	54.	(Uliyan et al., 2021)	2021
20.	(Delen, Davazdahemami and Rasouli Dezfouli, 2024)	2024	55.	(Palacios et al., 2021)	2021
21.	(Cho, Yu and Kim, 2023)	2023	56.	(Fontana et al., 2021)	2021
22.	(Phan, De Caigny and Coussement, 2023)	2023	57.	(Nanglae et al., 2021)	2021
23.	(Hammoudi Halat et al., 2023)	2023	58.	(Cuizon, 2021)	2021
24.	(Matz et al., 2023)	2023	59.	(Yaqin, Laksito, and Fatonah, 2021)	2021
25.	(Villegas-Ch, Govea, and Revelo-Tapia, 2023)	2023	60.	(Tsai et al., 2020)	2020
26.	(Cañete-Sifuentes et al., 2023)	2023	61.	(Sandoval-Palis et al., 2020)	2020
27.	(Kaensar and Wongnin, 2023)	2023	62.	(Freitas et al., 2020)	2020
28.	(Salam and Zeniarja, 2023)	2023	63.	(Alvarez, Callejas and Griol, 2020)	2020
29.	(Gutierrez-Pachas et al., 2023)	2023	64.	(Sani et al., 2020)	2020
30.	(Won et al., 2023)	2023	65.	(Bedregal-Alpaca et al., 2020)	2020
31.	(Kim et al., 2023)	2023	66.	(Rodríguez-Muñiz et al., 2019)	2019
32.	(Martins et al., 2023)	2023	67.	(Ortigosa et al., 2019)	2019
33.	(Anggrawan, Hairani, and Satria, 2023)	2023	68.	(Febro, 2019)	2019
34.	(Gonzalez-Nucamendi et al., 2023)	2023	69.	(Arumugam, Vinodhini, and Chandrasekaran, 2018)	2018
35.	(Song et al., 2023)	2023	70.	(Rovira, Puertas and Igual, 2017)	2017

Table 4: List of studies included in the SLR 2017-2025

Quality of reporting assessment

Quality assessment is positioned as an evaluation of the quality of reporting relevant to the auditability of the methodology, rather than as an absolute quality assessment. A study is considered adequate if it includes at least five components: the modeling method, class balancing strategies or an explicit statement regarding imbalance, evaluation metrics, validation design or data splitting, and the use or absence of hybrid or ensemble approaches if claimed as a contribution (Page et al., 2021). The assessment was conducted using a binary checklist for each component, and a summary of the results provided context for interpreting the findings in the results and discussion sections.

RESULTS

RQ1: What machine learning and deep learning algorithms are used to predict student dropout and graduation?

To answer RQ1, this review synthesizes the machine learning and deep learning algorithms used in the study corpus to predict students' academic status in higher education, including dropout, retention, and graduation. A summary of algorithm usage distribution at the corpus level is presented in Table 5. One study can report more than one model, so the frequencies in the table represent the number of studies reporting that model and are not mutually exclusive.

Algorithm / Family	Number of Studies	Percentage
Random Forest	34	48.6%
Decision Tree (C4.5/CART)	28	40.0%
Logistic Regression	27	38.6%
Support Vector Machine (SVM)	27	38.6%
KNearest Neighbors (KNN)	16	22.9%
Naïve Bayes	10	14.3%
MLP / Deep Neural Network	10	14.3%
XGBoost	15	21.4%
LightGBM	6	8.6%
CatBoost	5	7.1%
Gradient Boosting lain	9	12.9%
AdaBoost	6	8.6%
CNN	4	5.7%
LSTM / RNN	4	5.7%
Ensemble eksplisit (stacking, dsb)	6	8.6%

Table 5: Distribution of model usage in the study corpus

Based on Table 5, tree-based and tree ensemble models are the most frequently reported, particularly Random Forest and Decision Tree. Baseline models are also reported in a significant proportion, including logistic regression and SVM. Within the boosting group, XGBoost is reported in several studies, along with other boosting variations such as Gradient Boosting, LightGBM, CatBoost, and AdaBoost. In the deep learning group, MLPs or deep neural networks are reported in a portion of the studies, while CNNs, LSTMs, or RNNs are reported in smaller proportions. Additionally, some studies report the use of explicit ensembles such as stacking as a model combination configuration.

To provide context for reporting at the study level, representative study examples, classification scenarios, main models, and reported metrics are presented in Table 6. The examples in Table 6 are used to illustrate the variations in modeling scenarios and selected metrics, not to compare performance across studies.

Main Scenarios and Models	Summary of study metrics and context notes
Binary, Random Forest	Accuracy: 95.93 percent; F1 score: approximately 0.88. Dropout prediction in the B40 context (Sani et al., 2020).
Binary, XGBoost	AUC is approximately 0.97; accuracy is approximately 94.1 percent (Canto, De Oliveira, and De Mattos Veroneze, 2022; Haerani et al., 2023; Kim et al., 2023)
Binary, Decision Tree C4.5	Accuracy is approximately 89 percent. Predicting dropout in the Malaysian context (Roslan et al., 2024).
Binary, Stacking SMLOS with SMOTE and Optuna	Accuracy is 95.5 percent. Configuration using resampling and hyperparameter tuning (Herianto et al., 2024).
Binary, Voting ensemble LR, DT, and ANN	Recall dropout is about 98 percent. Reporting emphasizes dropout metrics (Rabelo and Zárate, 2025).
Multiclass: 3 classes; XGBoost-tuned AGbSCHO.	Accuracy 88.00 percent; Cohen's kappa 0.666. Metaheuristic tuning in multi-class scenarios (Goran et al., 2024).
Multiclass for three classes, LightGBM and CatBoost tuned with Optuna	F1 dropout 0.88; F1 graduate 0.86; F1 enrolled 0.83. Reporting metrics per class (Villar and de Andrade, 2024).
Multiclass for three classes, Random Forest with SVMSMOTE	Balanced accuracy: 74.8 percent; global F1: 0.745. Reporting balanced accuracy and global F1 in multi-class scenarios (Martins et al., 2023).
Multiclass for three classes, C4.5	F-measure continues at 99.6 percent; dropout at 72.0 percent; change at 44.4 percent. The F-measure is reported for each class, including the change class (Rodríguez-Muñiz et al., 2019).

Table 6: Representative studies, classification scenarios, main models, and metrics reported on predicting student academic status, 2017 to 2025

In line with Table 6, in the binary scenario, the corpus includes Random Forest and XGBoost reporting with metrics such as accuracy, F1, and AUC, as well as Decision Tree C4.5 with the accuracy metric. The corpus also includes studies that report ensemble configurations through stacking and voting, with metric reporting emphasizing dropout classes such as recall. In a three-class multiclass scenario, the corpus includes studies reporting XGBoost with Cohen's kappa, studies reporting LightGBM and CatBoost with F1 per-class, and studies reporting Random Forest with SVMSMOTE using balanced accuracy and global F1. In C4.5-based multiclass studies, F-measure reporting varies across classes and lists values for each status.

RQ2: How are targets formulated as binary or multiclass classification in predicting student study status, and what implications follow for modelling and reporting?

To answer RQ2, this review synthesizes how studies in the SLR corpus frame the target of predicting student study status as binary or multiclass classification, including those that test both scenarios within a single experimental design. A summary of the target formulation distribution at the study level is presented in Table 7. Quantitatively, 64 out of 70 studies only conducted binary experiments, 3 studies only conducted multi-class experiments, and 3 studies ran hybrid binary and multi-class scenarios (Rodríguez-Muñiz et al., 2019; Alvarez, Callejas, and Griol, 2020; Uliyan et al., 2021; Martins et al., 2023; Goran et al., 2024; Villar and de Andrade, 2024). Of the experiments conducted, 67 evaluated binary scenarios, while 6 evaluated multi-class scenarios.

Study Category	Number of Studies	Percentage
Binary only (experiment with only 2 classes)	64	91.4%
Multiclass only (experiments with more than 2 classes)	3	4.3%
Hybrid (runs binary and multi-class)	3	4.3%
Total number of studies conducting binary experiments	**67**	**95.7%**
Total number of studies conducting multi-class experiments	**6**	**8.6%**

Table 7: Distribution of target class formulations in the SLR corpus

Based on Table 7, the majority of studies conducted binary experiments, either as the sole scenario or as part of a hybrid scenario. In binary formulations, the target is typically expressed as dropout versus non-dropout, or graduation versus non-graduation, within a specific time horizon. Examples of the dropout-versus-non-dropout binary formulation are reported in the studies by Roslan et al. (2024) and Sani et al. (2020), which model dropout as the primary output. Examples of binary formulations for graduation or dropout output are also found in the studies by Canto et al. (2022), Haerani et al. (2023), and Kim et al. (2023), which report graduation or dropout predictions using metrics such as AUC and accuracy.

In contrast, multi-class formulations represent the target as multiple states, each corresponding to the student's study path in greater detail. In the corpus, a frequently occurring example is three study status classes, such as Graduate, Enrolled, and Dropout, which are evaluated by reporting metrics per-class (Villar and de Andrade, 2024), as well as three classes of Dropout, Enrolled, and Graduate in another multi-class configuration (Goran et al., 2024). Variations in the definition of multi-class labels are also found in other contexts, such as Continue, Dropout, and Change (Rodríguez-Muñiz et al., 2019); Promotion, Repetition, and Dropout (Alvarez, Callejas, and Griol, 2020); and learning progress-based labels such as Ongoing or Normal, At Risk, and Unsurpassed (Uliyan et al., 2021).

RQ3: What class imbalance handling strategies are reported, and what application patterns emerge?

To answer RQ3, this review examines the class imbalance handling strategies reported in studies predicting student study status, including interventions at the data, algorithm, and decision levels. The distribution of strategies reported at the corpus level is presented in Table 8. In this synthesis, the category "not reported" refers to studies that do not explicitly address class imbalance, making it impossible to trace mitigation strategies from the primary studies.

Reported strategies	Number of Studies	Percentage
Not reported	33	47.1%
Oversampling	16	22.9%
Undersampling	6	8.6%
Combined sampling and cost-sensitive sampling	5	7.1%
Cost-sensitive or class-weighted sampling	3	4.3%
Hybrid resampling and cleaning	2	2.9%
Decision threshold adjustment	2	2.9%
Balanced or stated design	2	2.9%
Ensemble-based sampling	1	1.4%

Table 8: Distribution of class imbalance handling strategies reported in the SLR corpus, 2017-2025

Based on Table 8, almost half of the studies did not explicitly report on the treatment of class imbalance. Among the studies reporting strategies, resampling at the data-level is the most frequently mentioned approach, particularly oversampling and undersampling. Additionally, the corpus also includes strategies at the algorithm and decision level, such as cost-sensitive learning or class weighting, decision threshold adjustment, and sampling-based ensembles.

At the data level, the studies by Song et al. (2023) and Yaqin et al. (2021, 2022) reported the use of family-based oversampling techniques such as SMOTE and its variations. In the study by Villar and de Andrade (2024) on imbalanced multiclass data, the corpus also included comparisons of oversampling techniques like SMOTE and ADASYN. Undersampling was reported in several studies to reduce the dominance of the majority class or to explore class ratios in rare dropout conditions (Opazo et al., 2021; Cañete-Sifuentes et al., 2023).

At the algorithm and decision level, the corpus includes approaches such as class weighting or cost-sensitive learning, probability threshold adjustment, and sampling-based ensembles (Barramuño, Meza-Narváez, and Gálvez-García, 2022; Gonzalez-Nucamendi et al., 2023; Villegas-Ch, Govea, and Revelo-Tapia, 2023; Delen, Davazdahemami, and Rasouli Dezfouli, 2024). Additionally, studies by Alvarez et al. (2020) and Martins et al. (2023) report problem transformation through class merging to reduce label imbalance in extreme minority conditions. In certain configurations, the corpus also includes hybrid techniques that combine resampling and cleaning (Kim et al., 2023).

RQ4: To what extent are ensemble and hybrid methods used, and what application patterns emerge in this corpus?

To answer RQ4, this review examines how studies in the corpus apply ensemble and hybrid approaches to predicting student study status and identifies the application patterns that emerge across study designs. In this RQ, the term "ensemble" refers to the strategy of combining multiple models to produce a final prediction, such as stacking, voting, bagging, or boosting (Cuizon, 2021). The term "hybrid" refers to designs that combine different methods within a single workflow, such as combining deep learning and machine learning, or building a staged pipeline based on clustering or feature selection (Phan, De Caigny, and Coussement, 2023).

A summary of the ensemble and hybrid approach categorizations identified in the corpus is presented in Table 9 to illustrate the variety of configurations and examples of studies within each category.

Approach Categories in RQ4	Number of Studies	Recorded Studies
Ensemble Stacking	2	(Niyogisubizo et al., 2022; Herianto et al., 2024)
Ensemble Voting	4	(Cuizon, 2021; Cañete-Sifuentes et al., 2023; Okoye et al., 2024; Rabelo and Zárate, 2025)
Weighted and Cascading Voting	1	(Fernandez-Garcia et al., 2021)
Random Forest-Based Ensemble Bagging	4	(Sani et al., 2020; Palacios et al., 2021; Kaensar and Wongnin, 2023; Matz et al., 2023)
Boosting as an Ensemble	1	(Hammoudi Halat et al., 2023)
Ensembles for Imbalanced Data	1	(Martins et al., 2023)
Hybrids of Deep Learning and Machine Learning or Model Combination	3	(Kim et al., 2023; Delen, Davazdahemami, and Rasouli Dezfouli, 2024; Nguyen Thi Cam, Sarlan, and Arshad, 2024)
Hybrid Stepwise Clustering or Feature Selection-Based Hybrids	2	(Nanglae et al., 2021; Nuanmeesri et al., 2022)
Hybrids Across Analytical Paradigms	1	(Csalódi and Abonyi, 2021)
Dual Classification and Survival Modeling	1	(Gutierrez-Pachas et al., 2023)

Table 9: Summary of ensemble and hybrid approach categories in the study corpus, 2017 To 2025

Based on Table 9, the most frequently occurring categories are voting ensembles and random forest-based bagging, each appearing in 4 studies. The corpus also includes hybrid configurations that combine modeling paradigms or stages, such as integrating deep learning and machine learning, as well as cluster-based or feature-selection-based staged pipelines. To provide context for reporting at the study level, the following description summarizes the configuration, comparators, and performance findings as reported in the primary studies.

Herianto et al. (2024) reported hybrid stacking via SMLOS, which combines multiple base models with a meta model and compares it to individually optimized base models, with the highest reported accuracy in the tested configuration. Cañete-Sifuentes et al. (2023) reported on a voting ensemble based on machine learning automation, where VotingEnsemble combines multiple tree-based models and is compared to single models and models specifically designed to handle class imbalance, reporting the combination of true positive rate and false positive rate at a specific ratio.

Fernandez-Garcia et al. (2021) reported a proportional weighted ensemble combining gradient boosting, random forests, and support vector machines, with individual model comparisons. They observed changes in recall and precision in the first semester. Martins et al. (2023) reported on ensembles for imbalanced data by comparing Balanced Random Forest and Easy Ensemble with resampling pipelines and standard models, namely SMOTE with Random Forest and SVMSMOTE with Random Forest. They reported F1 scores and balanced accuracy for the minority class in a multi-class scenario. Rabelo and Zárate (2025) reported a classic voting ensemble that combines CART, logistic regression, and artificial neural networks, with individual model comparisons, and found higher prediction stability in the study context. Hammoudi Halat et al. (2023) reported boosting as an ensemble with the comparators used in the study and reported performance results on the tested configurations.

RQ5: Which evaluation metrics and validation methods are used, and how suitable are they for imbalanced data?

To answer RQ5, the corpus shows that evaluating the prediction of student study status becomes less representative when relying solely on accuracy, especially when class distributions are imbalanced. Several studies improve accuracy using confusion matrix-based metrics, including precision, recall, and F1 score, as well as metrics that are more sensitive to imbalance, such as balanced accuracy and G-mean (Yaqin, Rahardi, and Abdulloh, 2022; Kim et al., 2023; Martins et al., 2023).

To complement the tabular synthesis, Figure 4 presents two representative confusion matrices from a primary study to illustrate class-wise error patterns in a multi-class setting. This visualization supports the argument that per-class reporting and confusion matrix reading are essential under class imbalance, particularly when minority class detection is central to intervention decisions. The figure is provided as a representative example and does not imply an aggregated benchmark across heterogeneous studies.

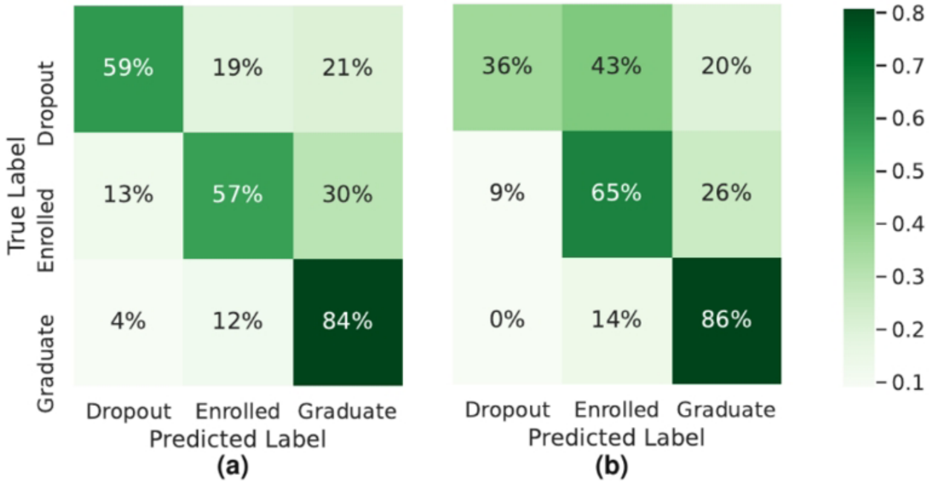


Figure 4: Representative confusion matrices for multi-class study status prediction illustrating class-wise error patterns. (A) Example from dataset S0. (B) Example from dataset S2. Reproduced from (Martins et al., 2023)

Some studies report correlation and agreement metrics, such as the Matthews correlation coefficient and Cohen's kappa, as performance summaries that account for all parts of the confusion matrix (de la Cruz Huayanay, Bazán, and Russo, 2024). In the corpus, this metric appears as a summary alternative when studies want to present a performance measure that does not rely on simple aggregation.

In area-based metrics, the corpus includes reporting of AUC-ROC and precision-recall-based metrics, including AUC-PRPR (Luque et al., 2019; Palacios et al., 2021; Martins et al., 2023; Delogu et al., 2024; Vaarma and Li, 2024). Several studies highlight the precision-recall curve and AUC PR as important complements when the evaluation focus is directed toward rare positive classes, while AUC ROC is still reported to maintain comparability with more common reporting in the literature (Luque et al., 2019; Palacios et al., 2021; Delogu et al., 2024).

The corpus also includes error metric reporting, such as false-positive and false-negative rates, to describe more specific error patterns within the institution's classes of interest. In this context, some studies emphasize the false negative rate when the evaluation focus is on the risk of failing to detect at-risk students (Cañete-Sifuentes et al., 2023; Okoye et al., 2024).

Beyond the metrics, the validation designs used in the primary studies varied. The corpus reports the use of hold-out, stratified k-fold cross-validation, cross-validation with external validation, nested cross-validation, and temporal validation (Moreira da Silva et al., 2022; Niyogisubizo et al., 2022; Kim et al., 2023; Martins et al., 2023; Phan, De Caigny, and Coussement, 2023). Stratified k-fold is said to keep the representation of minority classes in each fold, while nested cross-validation is used in studies that combine evaluation with more systematic hyperparameter tuning (Martins et al., 2023; Phan, De Caigny, and Coussement, 2023). Temporal validation is also reported when studies test the consistency of models across cohorts or academic periods (Moreira da Silva et al., 2022; Kim et al., 2023).

In some studies, the class balancing technique is also described as part of the evaluation pipeline. The corpus includes reports indicating that resampling techniques, such as SMOTE, were applied to the training data within cross-validation schemes by implementing them in each training fold, which allows the evaluation procedure to be traced without merging the training and test data during the balancing stage (Kim et al., 2023; Martins et al., 2023; Song et al., 2023).

RQ6: What methodological gaps and challenges remain, and what recommendations strengthen future research?

To answer RQ6, the corpus synthesis shows that research on predicting student study status is developing rapidly, but still leaves methodological gaps that can affect the validity of conclusions, especially when the problem is formulated as multiclass classification with imbalanced label distributions. In the corpus, binary formulations remain dominant, while studies explicitly testing multi-class classification are relatively limited. In available multi-class studies, the prominent challenges are not only the decline in aggregate performance but also the performance disparity between classes, especially when one class becomes an extreme minority or when classes are conceptually close and easily confused. Therefore, reporting metrics by class and analyzing error patterns emerged as a crucial methodological need for understanding performance readability at relevant study statuses for intervention. Confusion matrix analysis should be used to identify systematic confusions between conceptually adjacent statuses and to quantify minority-class errors that aggregate summaries may hide. Figure 4 is included as a representative example to illustrate how such confusions can be inspected in multi-class settings (Rodríguez-Muñiz et al., 2019; Uliyan et al., 2021; Martins et al., 2023; Goran et al., 2024; Villar and de Andrade, 2024).

The next gap concerns the traceability of methodological decisions for handling class imbalance. Several studies have reported data-level strategies, including resampling (e.g., SMOTE and its derivatives), hybrid approaches that combine oversampling and cleaning, and undersampling. Other studies have reported algorithm-level strategies such as class weighting and cost-sensitive learning. However, the corpus also shows that reporting on the treatment of imbalance is not always explicit, making it difficult for readers to assess whether improvements in minority classes stem from balancing strategies, model selection, or procedural consequences of the evaluation. The key concern at this juncture is the placement of resampling within the assessment pipeline, as implementing it before data partitioning may yield excessively optimistic performance estimates due to information leakage. Therefore, this review recommends that resampling techniques such as SMOTE be applied strictly to the training folds within the cross-validation scheme, after the split is created. The validation fold and any held-out test set must remain untouched to prevent information leakage and overly optimistic estimates (Cañete-Sifuentes et al., 2023; Kim et al., 2023; Song et al., 2023).

To improve auditability and to avoid overly optimistic estimates caused by information leakage, Figure 5 summarizes a leakage-free evaluation workflow for imbalanced multi-class classification. The workflow emphasizes that preprocessing and resampling must be performed only within each training fold after splitting, while the validation fold remains untouched until evaluation. Procedural guidance for leakage-free evaluation under class imbalance is as follows. First, define the target mapping rules and report the class distribution. Second, select a validation design that matches the intended deployment scenario, such as stratified cross-validation, temporal validation, or external validation. Third, within each training fold only, fit preprocessing steps, apply resampling, and then fit the model, while keeping the validation fold untouched. Fourth, when hyperparameter tuning is performed, use a nested cross-validation design or an inner loop to prevent information leakage into the evaluation. Fifth, evaluate on the untouched validation fold using per-class precision, recall, and F1, together with imbalance-sensitive summaries such as macro F1 and balanced accuracy, and interpret results through confusion matrix diagnostics.

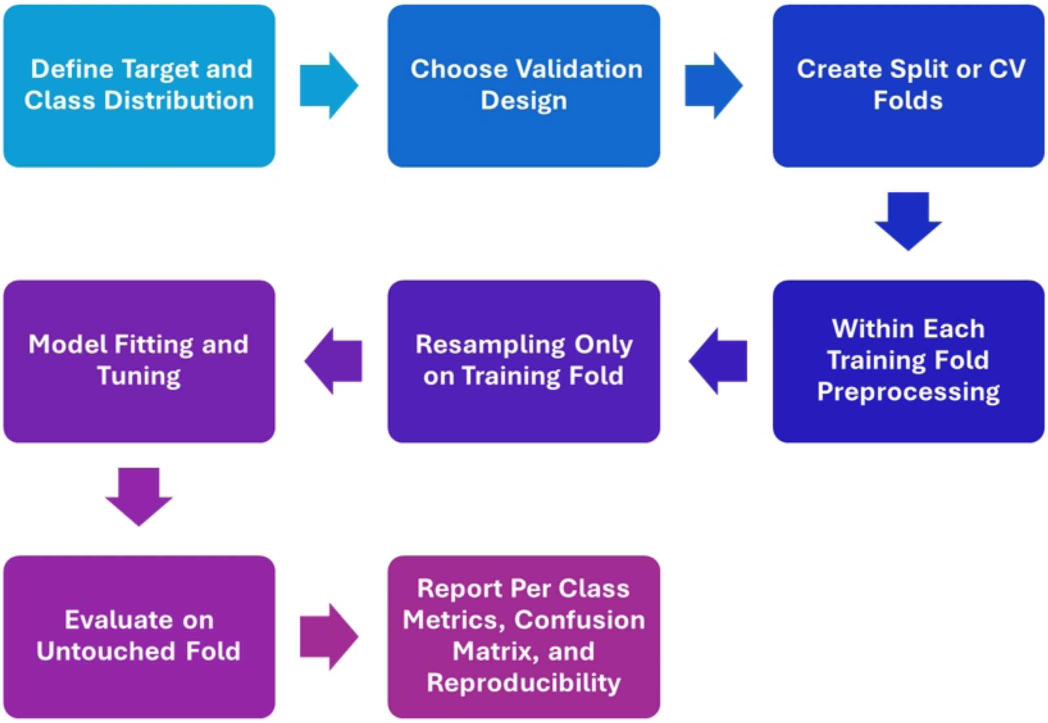

Define Target and Class Distribution
Choose Validation Design
Create Split or CV Folds
Within Each Training Fold Preprocessing
Resampling Only on Training Fold
Model Fitting and Tuning
Evaluate on Untouched Fold
Report Per Class Metrics, Confusion Matrix, and Reproducibility

Figure 5: Workflow for leakage-free evaluation under multi-class class imbalance, emphasizing preprocessing and resampling only within training folds, while validation folds remain untouched until evaluation

In answering RQ5, the corpus indicates that research closely linked to the traits of imbalanced data typically improves accuracy by using measures derived from confusion matrices, imbalance-sensitive metrics, and area-based and error-based metrics. This variation is closely associated with the model's intended application, particularly when its output informs risk rating and intervention prioritization. The corpus presents a variety of validation methodologies, including hold-out, stratified k-fold cross-validation, external validation, nested cross-validation, and temporal validation. This design variation demonstrates that the reliability of performance estimates is significantly affected by the data partitioning, the management of hyperparameter tuning, and the inclusion of cross-cohort or cross-temporal assessments (Moreira da Silva et al., 2022; Niyogisubizo et al., 2022; Kim et al., 2023; Martins et al., 2023; Phan, De Caigny, and Coussement, 2023).

Another recurring challenge in the corpus concerns the repeatability of experiments and opportunities for cross-institutional replication. The corpus demonstrates a dominance of internal private institutional datasets over public or open-access datasets. This condition enriches the institutional context but, at the same time, limits the exact replication of experiments by other researchers due to data access, the definition of study status labels, and feature structures that are often tied to local academic policies. In the corpus, the distribution of dataset sources shows a very high proportion of private datasets relative to public or open-access datasets, making replication and generalizability issues important contexts for interpreting cross-study findings.

To promote a more open future research agenda despite private data constraints, this review recommends standardizing feature structures and reporting artifacts so that models developed on internal datasets can still be tested for cross-institutional generalizability. Concretely, studies should publish a feature schema listing feature names, definitions, types, allowable ranges, missing-value handling, and encoding rules, along with a mapping from local variables to the shared schema. In addition, studies should document target mapping rules and prediction horizons using consistent terminology and provide evaluation scripts or pseudocode that enable external teams to replicate preprocessing, splitting protocols, and metric computation on their own institutional data. These steps do not require releasing identifiable student records, yet they enable reproducible comparisons and cross-institutional validation through a shared representation of the problem.

To strengthen traceability and auditability, the corpus also indicates the need for consistent minimal reporting standards across studies. In this manuscript, quality assessment is positioned as an evaluation of reporting quality relevant to methodological auditability, with minimal components including modeling methods, strategies for handling imbalance or explicit statements regarding imbalance, evaluation metrics, validation designs or data splitting, and clarity of use or the absence of hybrid or ensemble approaches when claimed as contributions (Page et al., 2021). Minimal reporting checklist for auditability and reproducibility: (1) Target definition and mapping rules, including the institutional policy assumptions used to label dropout or graduation related outcomes; (2) Class distribution reported for the overall dataset and for each split or fold, not only at the full dataset level; (3) Data splitting and validation protocol, including whether the split is stratified, temporal, or uses an external test set, and the rationale for the choice; (4) Pipeline specification that explicitly states the sequence of preprocessing, resampling, and model fitting, and explicitly states that resampling is performed only within training folds; (5) Hyperparameter tuning protocol, including whether nested cross-validation or an inner loop is used, and which data are used for selection; (6) Evaluation metrics including per-class precision, recall, and F1, together with at least one imbalance-sensitive summary such as macro F1 or balanced accuracy, and an additional summary such as MCC or kappa when applicable; (7) Confusion matrix analysis describing systematic confusions between conceptually adjacent classes and the error profile of the minority class; (8) Reproducibility artifacts including random seed control, software libraries, and a description of the feature schema sufficient for replication or cross-institutional testing. This framework helps read the findings of RQ1-RQ5 in a more structured way, while also highlighting the methodological sections that are most often a source of uncertainty in the interpretation of primary studies. As a cross-RQ1 to RQ6 summary, the dimensions of target formulation, degree of imbalance, handling strategies, and evaluation and validation can be mapped as a study design taxonomy synthesized in the manuscript and presented in Figure 6. This taxonomy is presented as a summary of findings that facilitates pattern reading and as a framework for formulating a methodological strengthening agenda for subsequent research, without making any single configuration a claim of universal superiority across all institutional contexts.

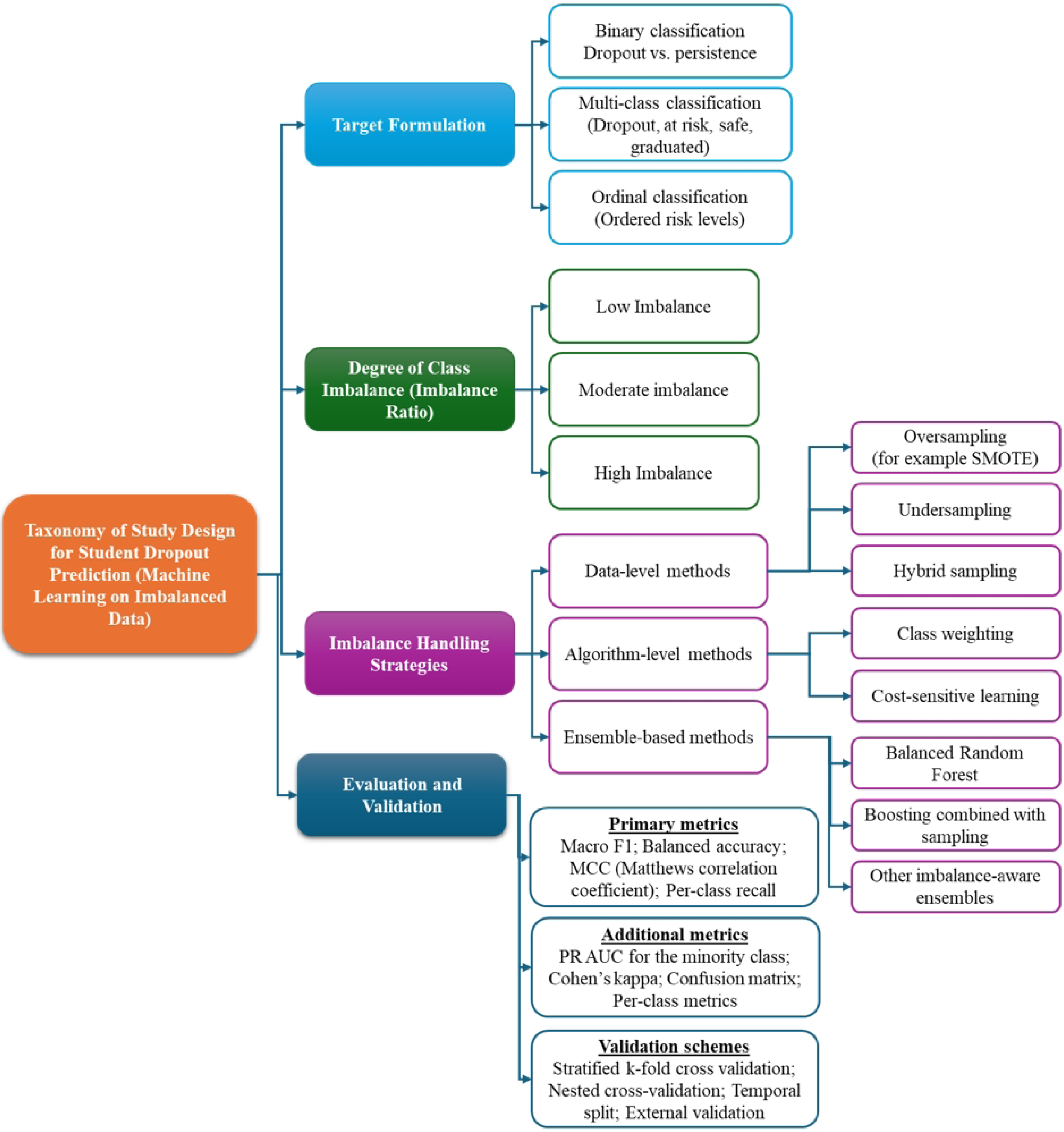

Taxonomy of Study Design for Student Dropout Prediction (Machine Learning on Imbalanced Data)
Target Formulation
Binary classification Dropout vs. persistence
Multi-class classification (Dropout, at risk, safe, graduated)
Ordinal classification (Ordered risk levels)
Degree of Class Imbalance (Imbalance Ratio)
Low Imbalance
Moderate imbalance
High Imbalance
Imbalance Handling Strategies
Data-level methods
Oversampling (for example SMOTE)
Undersampling
Hybrid sampling
Algorithm-level methods
Class weighting
Cost-sensitive learning
Ensemble-based methods
Balanced Random Forest
Boosting combined with sampling
Other imbalance-aware ensembles
Evaluation and Validation
Primary metrics
Macro F1; Balanced accuracy; MCC (Matthews correlation coefficient); Per-class recall
Additional metrics
PR AUC for the minority class; Cohen's kappa; Confusion matrix; Per-class metrics
Validation schemes
Stratified k-fold cross validation; Nested cross-validation; Temporal split; External validation

Figure 6: Study design taxonomy for predicting student dropout on imbalanced data, synthesized from findings RQ1 to RQ6

DISCUSSION

The synthesis in the Results section indicates that predicting students' study status in higher education has become an established research theme, characterized by a variety of publication venues and the widespread use of machine learning, including tree-based models, boosting, and some forms of deep learning. However, across the RQs, it also appears that the maturity of modeling is not always matched by that of evaluation or the traceability of study design decisions. This inequality is important within the framework of efficiency and responsibility in education and science. From an efficiency perspective, indefensible evaluation designs can lead to inefficient allocation of limited support capacity due to misleading error trade-offs. From a responsibility perspective, the same indefensible designs can lead to conclusions that are not defensible for educational decision support and not accountable as scientific evidence (Vidal et al., 2022; Villegas-Ch, Govea, and Revelo-Tapia, 2023). Accordingly, this Discussion interprets the results in terms of operational implications for efficient interventions and methodological implications for responsible evidence, and it concludes with actionable recommendations derived from the taxonomy, including reporting, evaluation, and reproducibility priorities.

The findings in Table 5 and the case studies in Table 6 confirm the dominance of the pragmatic tabular approach, particularly tree-based methods such as tree families and tree ensembles, with boosting. This pattern aligns with the generally administrative and tabular nature of academic data and the relatively easy implementation requirements. However, for cross-institutional generalization, the primary value of this SLR is better characterized as mapping the patterns of study design decisions that determine the strength of performance claims, rather than ranking cross-study models. Variations in the definition of study status and prediction horizon, and differences in curriculum structure and academic policies, often make performance figures between studies incomparable, so consistency in defining targets and the defensibility of evaluations are prerequisites for interpretation (Rodríguez-Muñiz et al., 2019; Palacios et al., 2021). The private dataset dominance observed in the corpus further constrains cross-institutional generalizability because target definitions, feature construction, and academic policies are often institution-specific. Therefore, an open research agenda in this domain should prioritize the portability of research artifacts rather than that of raw data. A practical path is to standardize feature structures through shared feature schemas and data dictionaries that preserve privacy while enabling cross-institutional testing. Under this agenda, different institutions can implement a compatible feature schema locally and evaluate the same modelling and validation protocol on their own cohorts, which supports stronger external validation evidence without requiring sensitive record sharing.

From the perspective of target formulation, Table 7 shows that binary experiments dominate, while multi-class experiments are relatively limited. The dominance of binary experiments can be understood as a simplification of the early warning task. Still, multi-class studies in the corpus show a complexity closer to the reality of the institutional study status. In multi-class scenarios, performance differences between classes become more apparent, especially when there are extreme minority classes or conceptually close status categories that are easily confused (Rodríguez-Muñiz et al., 2019; Martins et al., 2023; Goran et al., 2024; Villar and de Andrade, 2024). The implication is that a single-number performance summary based on aggregate data becomes increasingly risky for multiple classes because it can mask weaknesses in the status most relevant for intervention.

Regarding class imbalance, Table 8 shows that the reported handling strategies included resampling, class weighting, cost-sensitive learning, and other variations, but the proportion of studies that did not explicitly state the imbalance treatment remained high. This limitation is not marginal: Table 8 shows that 33 of 70 studies (47.1%) did not explicitly report class-imbalance handling strategies. This lack of procedural reporting is a significant obstacle to auditability and reproducibility because readers cannot verify whether reported improvements in the minority class arise from the intended imbalance strategy, from model choice, or from evaluation design side effects. This is a critical reporting issue because the lack of procedural information makes it difficult to assess whether performance is driven by the appropriate strategy or by estimation bias arising from a loose evaluation design. The literature emphasizes that the key issue is not just the use of resampling but its placement within the evaluation pipeline, specifically, whether it is applied only to the training data, including to each training fold during cross-validation, to prevent information leakage and overly optimistic estimates (Kim et al., 2023; Martins et al., 2023; Song et al., 2023). Therefore, defensible reporting needs to explicitly state the sequence of procedures, including data separation, validation schemes, hyperparameter tuning, and resampling positions.

On the metric side, the findings of RQ5 reinforce the conclusion that accuracy is an inadequate single summary measure for class imbalance. More responsible practices are evident in studies that complement accuracy with precision, recall, F1, and imbalance-sensitive metrics such as balanced accuracy and G-mean (Yaqin, Rahardi, and Abdulloh, 2022; Kim et al., 2023; Martins et al., 2023). The corpus also includes summary metrics that consider all components of the confusion matrix, such as the Matthews correlation coefficient and Cohen's kappa (de la Cruz Huayanay, Bazán, and Russo, 2024). In curve-based evaluation, the literature confirms that in extreme imbalances, ROC can appear satisfactory even though the performance of the rare positive class remains weak, making precision-recall-based metrics, including AUC PR, more informative when the institution's goal is to rank dropout risk (Luque et al., 2019; Palacios et al., 2021; Martins et al., 2023; Delogu et al., 2024; Vaarma and Li, 2024). At the implementation level, the cost of errors is reflected in the reporting of false-positive and false-negative rates, especially when the focus is on the risk of failing to detect at-risk students (Cañete-Sifuentes et al., 2023; Okoye et al., 2024). This confirms that the efficiency of interventions depends on managing the trade-off between errors and academic service capacity, not solely on average performance.

RQ4 shows that ensembles and hybrid configurations are present with varying intensities (Table 9). The emergence of voting ensembles and bagging based on Random Forest can be interpreted as a pragmatic strategy to improve prediction stability on tabular data. However, adding complexity through stacking, weighted voting, or hybrid designs does not always make implementation more efficient unless there is strict validation and clear reporting. Some studies show benefits in their own test settings, such as hybrid stacking (Herianto et al., 2024) and changes in the trade-off for automation-based voting (Cañete-Sifuentes et al., 2023). Other studies focus on changes in recall and precision in the early stages that are important for early warning (Fernandez-Garcia et al., 2021). Therefore, ensembles and hybrids are more accurately positioned as tools for managing trade-offs in specific contexts, rather than as guarantees of universal improvement (Vidal et al., 2022; Villegas-Ch, Govea, and Revelo-Tapia, 2023).

As a cross-RQ synthesis, Figure 6 summarizes the most critical design decisions for result reliability, namely target formulation, degree of imbalance, imbalance-handling strategies, and evaluation and validation. This taxonomy confirms that algorithm selection cannot be separated from more fundamental decisions, particularly the definition of labels, the prediction horizon, and the validation scheme. Thus, Figure 6 can serve as an audit framework for reporting prior studies and as a checklist for designing new studies, thereby ensuring greater consistency and comparability.

Implications for efficiency and responsibility can be drawn directly from the taxonomy. For efficiency, the evidence base becomes more actionable when studies report class-specific error patterns and explicitly relate metric choices to intervention capacity, because early warning systems operate under constrained mentoring and support resources. For responsibility, the evidence base becomes more defensible when studies transparently report target definitions, validation designs, and minority class performance, because these elements determine whether results can be trusted for educational decision support and whether findings are reproducible and accountable as scientific contributions. Accordingly, this review emphasizes that efficiency-oriented deployment and responsibility-oriented research practice depend on consistent reporting and evaluation choices, not solely on algorithm selection. This link clarifies how study design decisions translate into both operational value and scientific accountability. A further implication of using a single database is that the descriptive distributions reported in this review, such as the relative prevalence of specific algorithms, imbalance-handling strategies, or validation designs, may be sensitive to index coverage. For example, education-oriented journals and applied analytics outlets may be indexed differently from engineering and computing venues, which could shift the observed proportions of methods even when the substantive methodological issues remain similar. Importantly, the main conclusions of this review emphasize study design transparency, leakage-free evaluation, and reporting on minority classes rather than absolute performance rankings. Therefore, while the inclusion of additional databases may alter some frequency-based summaries, the central recommendations on auditability and defensible evaluation are expected to remain applicable. Nevertheless, future work should replicate the protocol across multiple sources, such as Web of Science and discipline-specific libraries, and compare overlaps to assess the robustness of the observed patterns.

This study has three main limitations. First, the search relies on a single database, Scopus, which may introduce index coverage bias (Mongeon and Paul-Hus, 2016; Baas et al., 2020). Scopus was selected as a practical proxy for high-quality peer-reviewed journal literature because it provides broad multidisciplinary journal indexing with consistent metadata, enabling a reproducible and auditable protocol. However, relevant studies may still appear in other curated indexes or discipline-specific libraries, and their inclusion could shift some descriptive distributions, such as the relative frequencies of algorithms or validation designs. The main conclusions of this review are primarily methodological and focus on transparency, leakage-free evaluation, and minority class reporting, so they are less dependent on the exact distribution of methods. However, future work should extend retrieval to additional sources, such as the Web of Science and discipline-specific libraries, to quantify overlap and test robustness. Second, the protocol restricts the corpus to open-access journal articles with DOIs and English language, which improves verifiability but may omit relevant evidence that is not open-access, not written in English, or disseminated in alternative publication formats. Third, the dominance of private datasets limits exact replication and cross-institutional comparison, as target definitions, feature construction, and academic policies are often context-specific. To mitigate this limitation while respecting privacy constraints, future work should standardize feature structures through shared feature schemas and mapping documentation, so that models and evaluation protocols can be tested across institutions even when the underlying datasets cannot be released (Rodríguez-Muñiz et al., 2019; Palacios et al., 2021; Hooper, Ragland, and Artemiou, 2025).

The implied future research agenda is to strengthen evaluation and reporting standards for predicting study status, especially in imbalanced multi-class scenarios. Priorities include reporting metrics per-class and stable summaries such as the Matthews correlation coefficient or Cohen's kappa, affirming validation procedures and the position of resampling in the pipeline, and using more conservative validation when intensive hyperparameter tuning is performed, including nested cross-validation, temporal validation, or external validation on different cohorts (Niyogisubizo et al., 2022; Martins et al., 2023; Phan, De Caigny, and Coussement, 2023; Song et al., 2023; de la Cruz Huayanay, Bazán, and Russo, 2024).

CONCLUSIONS

This study synthesizes 70 studies on predicting student academic status in higher education using machine learning and deep learning approaches from 2017 to 2025, with an emphasis on addressing class imbalance and ensuring defensible evaluation. The findings indicate that the contribution of research in this field cannot be assessed solely on the basis of algorithm selection or aggregate performance metrics. Its reliability and usability are more determined by the study design decisions, particularly the formulation of the target, the degree of imbalance, the strategies for handling imbalance, and the evaluation and validation design. Binary formulation dominance still stands out, while multi-class studies are relatively limited but closer to the reality of institutional study status and tend to reveal clearer performance gaps between classes. At the same time, variations and incompleteness in reporting treatment imbalances, metrics, and validation procedures limit cross-study comparability and reduce confidence in cross-institutional generalizability.

Within the framework of efficiency and responsibility in education and science, the practical value of a prediction system depends on the transparency of error trade-offs, the reporting of metrics representing minority classes, and validation aligned with operational scenarios, because these elements determine whether the evidence supports capacity-aware interventions and whether the results remain defensible, reproducible, and accountable. Further research is needed to strengthen minimum reporting standards, correctly place resampling within the training and validation process, and expand class-wise metric reporting in multi-class scenarios. More conservative validation designs, including nested, temporal, or external validation when possible, are also needed to support stable implementation across cohorts.

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. and Ceballos, H. G. (2022) 'Student dataset from Tecnologico de Monterrey in Mexico to predict dropout in higher education', *Data*, Vol. 7, No. 9, p. 119. https://doi.org/10.3390/data7090119

Alvarez, N. L., Callejas, Z. and Griol, D. (2020) 'Predicting computer engineering students' dropout in Cuban higher education with pre-enrollment and early performance data', *Journal of Technology and Science Education*, Vol. 10, No. 2, pp. 241–258. https://doi.org/10.3926/jotse.922

Anagnostopoulos, T., Papakyriakopoulos, D., Psaromiligkos, Y. and Retalis, S. (2024) 'Exploiting LSTM neural network algorithm potentiality for early identification of delayed graduation in higher education', *WSEAS Transactions on Information Science and Applications*, Vol. 21, pp. 524–532. https://doi.org/10.37394/23209.2024.21.48

Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodriguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., Villarreal-Torres, H. and Angeles-Morales, J. (2023) 'Predicting student dropout based on machine learning and deep learning: A systematic review', *EAI Endorsed Transactions on Scalable Information Systems*, Vol. 10, No. 5, pp. 1–11. https://doi.org/10.4108/eetsis.3586

Anggrawan, A., Hairani, H. and Satria, C. (2023) 'Improving SVM classification performance on unbalanced student graduation time data using SMOTE', *International Journal of Information and Education Technology*, Vol. 13, No. 2, pp. 289–295. https://doi.org/10.18178/ijiet.2023.13.2.1806

Arumugam, S., Vinodhini, G. and Chandrasekaran, R. M. (2018) 'Predicting students' academic performance in the university using meta decision tree classifiers', *Journal of Computer Science*, Vol. 14, No. 5, pp. 654–662. https://doi.org/10.3844/jcssp.2018.654.662

Baas, J., Schotten, M., Plume, A., Côté, G. and Karimi, R. (2020) 'Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies', *Quantitative Science Studies*, Vol. 1, No. 1, pp. 377–386. https://doi.org/10.1162/qss_a_00019

Barramuño, M., Meza-Narváez, C. and Gálvez-García, G. (2022) 'Prediction of student attrition risk using machine learning', *Journal of Applied Research in Higher Education*, Vol. 14, No. 3, pp. 974–986. https://doi.org/10.1108/JARHE-02-2021-0073

Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zárate-Valderrama, J. and Yanque-Churo, P. (2020) 'Classification models for determining types of academic risk and predicting dropout in university students', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 1, pp. 266–272. https://doi.org/10.14569/IJACSA.2020.0110133

Budiman, F., Saputro, I. A., Purwanto, P. and Andono, P. N. (2022) 'Optimization of classification results by minimizing class imbalance on decision tree algorithm', in: *International Seminar on Machine Learning, Optimization, and Data Science (ISMODE 2021)*, pp. 6–11. https://doi.org/10.1109/ISMODE53584.2022.9743062

Cañete-Sifuentes, L., Robles, V., Menasalvas, E. and Monroy, R. (2023) 'Comparing automated machine learning against an off-the-shelf pattern-based classifier in a class imbalance problem: Predicting university dropout', *IEEE Access*, Vol. 11, pp. 139147–139156. https://doi.org/10.1109/ACCESS.2023.3336596

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T. and Paganoni, A. M. (2022) 'Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques', *Studies in Higher Education*, Vol. 47, No. 9, pp. 1935–1956. https://doi.org/10.1080/03075079.2021.2018415

Canto, N. G., De Oliveira, M. A. and De Mattos Veroneze, G. (2022) 'Supervised learning applied to graduation forecast of industrial engineering students', *European Journal of Educational Research*, Vol. 11, No. 1, pp. 325–337. https://doi.org/10.12973/eu-jer.11.1.325

Cho, C. H., Yu, Y. W. and Kim, H. G. (2023) 'A study on dropout prediction for university students using machine learning', *Applied Sciences*, Vol. 13, No. 21, p. 12004. https://doi.org/10.3390/app132112004

Csalódi, R. and Abonyi, J. (2021) 'Integrated survival analysis and frequent pattern mining for course failure-based prediction of student dropout', *Mathematics*, Vol. 9, No. 5, p. 463. https://doi.org/10.3390/math9050463

Cuizon, J. C. (2021) 'Ensemble predictive model for academic churn risk using plurality voting', Mindanao *Journal of Science and Technology*, Vol. 19, No. 1, pp. 224–235. https://doi.org/10.61310/mndjsteect.1028.21

Darenoh, N. V., Bachtiar, F. A. and Perdana, R. S. (2024) 'Prediction of on-time student graduation with deep learning method', *Journal of ICT Research and Applications*, Vol. 18, No. 1, pp. 1–20. https://doi.org/10.5614/itbj.ict.res.appl.2023.18.1.1

Daza, A., Guerra, C., Cervera, N. and Burgos, E. (2022) 'Predicting academic performance through data mining: A systematic literature review', *TEM Journal*, Vol. 11, No. 2, pp. 939–949. https://doi.org/10.18421/TEM112-57

Delen, D., Davazdahemami, B. and Rasouli Dezfouli, E. (2024) 'Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework', *Information Systems Frontiers*, Vol. 26, No. 2, pp. 641–662. https://doi.org/10.1007/s10796-023-10397-3

Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T. and Gulam, S. B. (2025) 'Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors', *Systems and Soft Computing*, Vol. 7, p. 200352. https://doi.org/10.1016/j.sasc.2025.200352

de la Cruz Huayanay, A., Bazán, J. L. and Russo, C. M. (2024) 'Performance of evaluation metrics for classification in imbalanced data', *Computational Statistics*, Vol. 39, No. 3, pp. 1447–1473. https://doi.org/10.1007/s00180-024-01539-5

Delogu, M., Lagravinese, R., Paolini, D. and Resce, G. (2024) 'Predicting dropout from higher education: Evidence from Italy', *Economic Modelling*, Vol. 130, p. 106583. https://doi.org/10.1016/j.econmod.2023.106583

Febro, J. D. (2019) 'Utilizing feature selection in identifying predicting factors of student retention', *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 9, pp. 269–274. https://doi.org/10.14569/IJACSA.2019.0100934

Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M. and Sanchez-Figueroa, F. (2021) 'A real-life machine learning experience for predicting university dropout at different stages using academic data', *IEEE Access*, Vol. 9, pp. 133076–133090. https://doi.org/10.1109/ACCESS.2021.3115851

Fontana, L., Masci, C., Ieva, F. and Paganoni, A. M. (2021) 'Performing learning analytics via generalised mixed-effects trees', *Data*, Vol. 6, No. 7, p. 74. https://doi.org/10.3390/data6070074

Freitas, F. A. D. S., Vasconcelos, F. F. X., Peixoto, S. A., Hassan, M. M., Ali Akber Dewan, M., de Albuquerque, V. H. C. and Rebouças Filho, P. P. (2020) 'IoT system for school dropout prediction using machine learning techniques based on socioeconomic data', *Electronics*, Vol. 9, No. 10, p. 1613. https://doi.org/10.3390/electronics9101613

Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V. and García-Castelán, R. M. G. (2023) 'Predictive analytics study to determine undergraduate students at risk of dropout', *Frontiers in Education*, Vol. 8, p. 1244686. https://doi.org/10.3389/feduc.2023.1244686

Goran, R., Jovanovic, L., Bacanin, N., Stanković, M. S., Simic, V., Antonijevic, M. and Zivkovic, M. (2024) 'Identifying and understanding student dropouts using metaheuristic optimized classifiers and explainable artificial intelligence techniques', *IEEE Access*, Vol. 12, pp. 122377–122400. https://doi.org/10.1109/ACCESS.2024.3446653

Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G. and Gomez-Nieto, E. (2023) 'Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the Latin American context', *Education Sciences*, Vol. 13, No. 2, p. 154. https://doi.org/10.3390/educsci13020154

Haerani, E., Syafria, F., Lestari, F., Novriyanto, N. and Marzuki, I. (2023) 'Classification academic data using machine learning for decision making process', *Journal of Applied Engineering and Technological Science (JAETS)*, Vol. 4, No. 2, pp. 955–968. https://doi.org/10.37385/jaets.v4i2.1983

Hammoodi, M. S. and Al-Azawei, A. (2022) 'Using socio-demographic information in predicting students' degree completion based on a dynamic model', *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 2, pp. 107–115. https://doi.org/10.22266/ijies2022.0430.11

Hammoudi Halat, D., Abdel-Salam, A.-S. G., Bensaid, A., Soltani, A., Alsarraj, L., Dalli, R. and Malki, A. (2023) 'Use of machine learning to assess factors affecting progression, retention, and graduation in first-year health professions students in Qatar: A longitudinal study', *BMC Medical Education*, Vol. 23, No. 1, p. 909. https://doi.org/10.1186/s12909-023-04887-w

Helbach, J., Pieper, D., Mathes, T., Rombey, T., Zeeb, H., Allers, K. and Hoffmann, F. (2022) 'Restrictions and their reporting in systematic reviews of effectiveness: An observational study', *BMC Medical Research Methodology*, Vol. 22, No. 1, p. 230. https://doi.org/10.1186/s12874-022-01710-w

Herianto, H., Kurniawan, B., Hartomi, Z. H., Irawan, Y. and Anam, M. K. (2024) 'Machine learning algorithm optimization using stacking technique for graduation prediction', *Journal of Applied Data Sciences*, Vol. 5, No. 3, pp. 1272–1285. https://doi.org/10.47738/jads.v5i3.316

Hooper, S. E., Ragland, N. and Artemiou, E. (2025) 'Random forest models reveal academic and financial factors outweigh demographics in predicting completion of a year-round veterinary program', *Journal of the American Veterinary Medical Association*, Vol. 263, No. 2, pp. 1–9. https://doi.org/10.2460/javma.24.08.0501

Hoyos Osorio, J. K. and Daza Santacoloma, G. (2023) 'Predictive model to identify college students with high dropout rates', *Revista Electrónica de Investigación Educativa*, Vol. 25, pp. 1–10. https://doi.org/10.24320/redie.2023.25.e13.5398

Kaensar, C. and Wongnin, W. (2023) 'Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning', *Eurasia Journal of Mathematics, Science and Technology Education*, Vol. 19, No. 12, p. 2369. https://doi.org/10.29333/ejmste/13863

Kim, S., Choi, E., Jun, Y.-K. and Lee, S. (2023) 'Student dropout prediction for university with high precision and recall', *Applied Sciences*, Vol. 13, No. 10, p. 6275. https://doi.org/10.3390/app13106275

Kitchenham, B. (2004) *Procedures for performing systematic reviews*, Keele: Keele University, pp. 1–26.

Kurniadi, D., Abdurachman, E., Warnars, H. L. H. S. and Suparta, W. (2021) 'Predicting student performance with multi-level representation in an intelligent academic recommender system using backpropagation neural network', *ICIC Express Letters, Part B: Applications*, Vol. 12, No. 10, pp. 883–890. https://doi.org/10.24507/icicelb.12.10.883

Luque, A., Carrasco, A., Martín, A. and de las Heras, A. (2019) 'The impact of class imbalance in classification performance metrics based on the binary confusion matrix', *Pattern Recognition*, Vol. 91, pp. 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Martins, M. V., Baptista, L., Machado, J. and Realinho, V. (2023) 'Multi-class phased prediction of academic performance and dropout in higher education', *Applied Sciences*, Vol. 13, No. 8, p. 4702. https://doi.org/10.3390/app13084702

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T. and Realinho, V. (2021) 'Early prediction of student's performance in higher education: A case study', in: Rocha, Á., Adeli, H., Reis, L. P. and Costanzo, S. (eds.), *Trends and Applications in Information Systems and Technologies*, Cham: Springer, pp. 166–175. https://doi.org/10.1007/978-3-030-72657-7_16

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A. and Stachl, C. (2023) 'Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics', *Scientific Reports*, Vol. 13, No. 1, p. 5705. https://doi.org/10.1038/s41598-023-32484-w

Mongeon, P. and Paul-Hus, A. (2016) 'The journal coverage of Web of Science and Scopus: A comparative analysis', *Scientometrics*, Vol. 106, No. 1, pp. 213–228. https://doi.org/10.1007/s11192-015-1765-5

Moreira da Silva, D. E., Solteiro Pires, E. J., Reis, A., de Moura Oliveira, P. B. and Barroso, J. (2022) 'Forecasting students dropout: A UTAD university study', *Future Internet*, Vol. 14, No. 3, p. 76. https://doi.org/10.3390/fi14030076

Mouchantaf, N. and Chamoun, M. (2023) 'Predicting student dropout with minimal information', *Iraqi Journal of Science*, Vol. 64, No. 10, pp. 5265–5279. https://doi.org/10.24996/ijs.2023.64.10.33

Nagy, M. and Molontay, R. (2024) 'Interpretable dropout prediction: Towards XAI-based personalized intervention', *International Journal of Artificial Intelligence in Education,* Vol. 34, No. 2, pp. 274–300. https://doi.org/10.1007/s40593-023-00331-8

Nanglae, L., Iam-On, N., Boongoen, T., Kaewchay, K. and Mullaney, J. (2021) 'Determining patterns of student graduation using a bi-level learning framework', *Bulletin of Electrical Engineering and Informatics*, Vol. 10, No. 4, pp. 2201–2211. https://doi.org/10.11591/eei.v10i4.2502

Ndunagu, J. N., Oyewola, D. O., Garki, F. S., Onyeakazi, J. C., Ezeanya, C. U. and Ukwandu, E. (2024) 'Deep learning for predicting attrition rate in open and distance learning (ODL) institutions', *Computers*, Vol. 13, No. 9, p. 229. https://doi.org/10.3390/computers13090229

Nguyen Thi Cam, H., Sarlan, A. and Arshad, N. I. (2024) 'A hybrid model integrating recurrent neural networks and the semi-supervised support vector machine for identification of early student dropout risk', *PeerJ Computer Science*, Vol. 10, p. e2572. https://doi.org/10.7717/peerj-cs.2572

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E. and Nshimyumukiza, P. C. (2022) 'Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization', *Computers and Education: Artificial Intelligence*, Vol. 3, p. 100066. https://doi.org/10.1016/j.caeai.2022.100066

Nuanmeesri, S., Poomhiran, L., Chopvitayakun, S. and Kadmateekarun, P. (2022) 'Improving dropout forecasting during the COVID-19 pandemic through feature selection and multilayer perceptron neural network', *International Journal of Information and Education Technology*, Vol. 12, No. 9, pp. 851–857. https://doi.org/10.18178/ijiet.2022.12.9.1693

Okewu, E., Adewole, P., Misra, S., Maskeliunas, R. and Damasevicius, R. (2021) 'Artificial neural networks for educational data mining in higher education: A systematic literature review', *Applied Artificial Intelligence*, Vol. 35, No. 13, pp. 983–1021. https://doi.org/10.1080/08839514.2021.1922847

Okoye, K., Nganji, J. T., Escamilla, J. and Hosseini, S. (2024) 'Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education', *Computers and Education: Artificial Intelligence*, Vol. 6, p. 100205. https://doi.org/10.1016/j.caeai.2024.100205

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J. and Moreira, F. (2021) 'How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review', *Big Data and Cognitive Computing*, Vol. 5, No. 4, p. 64. https://doi.org/10.3390/bdcc5040064

Opazo, D., Moreno, S., Álvarez-Miranda, E. and Pereira, J. (2021) 'Analysis of first-year university student dropout through machine learning models: A comparison between universities', *Mathematics*, Vol. 9, No. 20, p. 2599. https://doi.org/10.3390/math9202599

Oqaidi, K., Aouhassi, S. and Mansouri, K. (2025) 'Predicting graduation in Moroccan open-access bachelors: Early indicators and re-enrollment data', *Bulletin of Electrical Engineering and Informatics*, Vol. 14, No. 1, pp. 524–532. https://doi.org/10.11591/eei.v14i1.8580

Ortigosa, A., Carro, R. M., Bravo-Agapito, J., Lizcano, D., Alcolea, J. J. and Blanco, Ó. (2019) 'From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system', *IEEE Transactions on Learning Technologies*, Vol. 12, No. 2, pp. 264–277. https://doi.org/10.1109/TLT.2019.2911608

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P. and Moher, D. (2021) 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', *BMJ*, p. n71. https://doi.org/10.1136/bmj.n71

Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V. and Marchant, C. (2021) 'Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile', *Entropy*, Vol. 23, No. 4, p. 485. https://doi.org/10.3390/e23040485

Pelima, L. R., Sukmana, Y. and Rosmansyah, Y. (2024) 'Predicting university student graduation using academic performance and machine learning: A systematic literature review', *IEEE Access*, Vol. 12, pp. 23451–23465. https://doi.org/10.1109/ACCESS.2024.3361479

Phan, M., De Caigny, A. and Coussement, K. (2023) 'A decision support framework to incorporate textual data for early student dropout prediction in higher education', *Decision Support Systems*, Vol. 168, p. 113940. https://doi.org/10.1016/j.dss.2023.113940

Quimiz-Moreira, M., Delgadillo, R., Parraga-Alava, J., Maculan, N. and Mauricio, D. (2025) 'Factors, prediction, explainability, and simulating university dropout through machine learning: A systematic review, 2012–2024', *Computation*, Vol. 13, No. 8, p. 198. https://doi.org/10.3390/computation13080198

Rabelo, A. M. and Zárate, L. E. (2025) 'A model for predicting dropout of higher education students', *Data Science and Management*, Vol. 8, No. 1, pp. 72–85. https://doi.org/10.1016/j.dsm.2024.07.001

Realinho, V., Martins, M. V., Machado, J. and Baptista, L. (2021) Predict students' dropout and academic success [Dataset], UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., Blunt, H., Brigham, T., Chang, S., Clark, J., Conway, A., Couban, R., de Kock, S., Farrah, K., Fehrmann, P., Foster, M., Fowler, S. A., Glanville, J., Harris, E., Hoffecker, L., Isojarvi, J., Kaunelis, D., Ket, H., Levay, P., Lyon, J., McGowan, J., Murad, M. H., Nicholson, J., Pannabecker, V., Paynter, R., Pinotti, R., Ross-White, A., Sampson, M., Shields, T., Stevens, A., Sutton, A., Weinfurter, E., Wright, K. and Young, S. (2021) 'PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews', *Systematic Reviews*, Vol. 10, No. 1, p. 39. https://doi.org/10.1186/s13643-020-01542-z

Rodríguez-Muñiz, L. J., Bernardo, A. B., Esteban, M. and Díaz, I. (2019) 'Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques?', *PLoS ONE*, Vol. 14, No. 6, p. e0218796. https://doi.org/10.1371/journal.pone.0218796

Rose, A. L. P. J. and Mary, A. C. (2022) 'An early intervention technique for at-risk prediction of higher education students in cloud-based virtual learning environment using classification algorithms during COVID-19', *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 1, pp. 612–621. https://doi.org/10.14569/IJACSA.2022.0130174

Roslan, N., Jamil, J. M., Shaharanee, I. N. M. and Sultan Alawi, S. J. (2024) 'Prediction of student dropout in Malaysian's private higher education institute using data mining application', *Journal of Advanced Research in Applied Sciences and Engineering Technology*, Vol. 45, No. 2, pp. 168–176. https://doi.org/10.37934/araset.45.2.168176

Rovira, S., Puertas, E. and Igual, L. (2017) 'Data-driven system to predict academic grades and dropout', *PLoS ONE*, Vol. 12, No. 2, p. e0171207. https://doi.org/10.1371/journal.pone.0171207

Salam, A. and Zeniarja, J. (2023) 'Classification of deep learning convolutional neural network feature extraction for student graduation prediction', *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 32, No. 1, p. 335. https://doi.org/10.11591/ijeecs.v32.i1.pp335-341

Salinas-Chipana, J., Obregon-Palomino, L., Iparraguirre-Villanueva, O. and Cabanillas-Carbonell, M. (2024) 'Machine learning models for predicting student dropout—a review', in: Yang, X.-S., Sherratt, R. S., Dey, N. and Joshi, A. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Singapore: Springer Nature Singapore, pp. 1003–1014. https://doi.org/10.1007/978-981-99-3043-2_83

Sandoval-Palis, I., Naranjo, D., Vidal, J. and Gilar-Corbi, R. (2020) ‘Early dropout prediction model: A case study of university leveling course students’, *Sustainability*, Vol. 12, No. 22, p. 9314. https://doi.org/10.3390/su12229314

Sani, N. S., Fikri, A., Ali, Z., Zakree, M. and Nadiyah, K. (2020) ‘Drop-out prediction in higher education among B40 students’, *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 11, pp. 550–559. https://doi.org/10.14569/IJACSA.2020.0111169

Sayed, M. (2024) ‘Student progression and dropout rates using convolutional neural network: A case study of the Arab Open University’, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 28, No. 3, pp. 668–678. https://doi.org/10.20965/jaciii.2024.p0668

Segura, M., Mello, J. and Hernández, A. (2022) ‘Machine learning prediction of university student dropout: Does preference play a key role?’, *Mathematics*, Vol. 10, No. 18, p. 3359. https://doi.org/10.3390/math10183359

Setiadi, H., Sanjaya, K., Wijayanto, A., Wardhani, D. W. and Cahyono, H. D. (2024) ‘Comparative analysis of classification algorithms using feature selection techniques to predict on-time student graduation’, *Ingénierie des systèmes d’information*, Vol. 29, No. 4, pp. 1365–1379. https://doi.org/10.18280/isi.290412

Setiawan, R., Nursasongko, E., Syukur, A., Budiman, F. and Kurniadi, D. (2025) ‘Handling class imbalance in student success prediction using machine learning: A comparison of SMOTE and SMOTETomek’, in: *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, pp. 1–6. https://doi.org/10.1109/SIML65326.2025.11081128

Song, Z., Sung, S.-H., Park, D.-M. and Park, B.-K. (2023) ‘All-year dropout prediction modeling and analysis for university students’, Applied Sciences, Vol. 13, No. 2, p. 1143. https://doi.org/10.3390/app13021143

Tsai, S.-C., Chen, C.-H., Shiao, Y.-T., Ciou, J.-S. and Wu, T.-N. (2020) ‘Precision education with statistical learning and deep learning: A case study in Taiwan’, International Journal of Educational Technology in Higher Education, Vol. 17, No. 1, p. 12. https://doi.org/10.1186/s41239-020-00186-2

Uliyan, D., Aljaloud, A. S., Alkhalil, A., Amer, H. S. Al, Mohamed, M. A. E. A. and Alogali, A. F. M. (2021) ‘Deep learning model to predict students retention using BLSTM and CRF’, *IEEE Access*, Vol. 9, pp. 135550–135558. https://doi.org/10.1109/ACCESS.2021.3117117

Vaarma, M. and Li, H. (2024) ‘Predicting student dropouts with machine learning: An empirical study in Finnish higher education’, *Technology in Society*, Vol. 76, p. 102474. https://doi.org/10.1016/j.techsoc.2024.102474

Vega, H., Sanez, E., De La Cruz, P., Moquillaza, S. and Pretell, J. (2022) ‘Intelligent system to predict university students dropout’, International Journal of Online and Biomedical Engineering (iJOE), Vol. 18, No. 7, pp. 27–43. https://doi.org/10.3991/ijoe.v18i07.30195

Véliz Palomino, J. C. and Ortega, A. M. (2023) ‘Dropout intentions in higher education: Systematic literature review’, *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 2, pp. 149–158. https://doi.org/10.7160/eriesj.2023.160206

Vidal, J., Gilar-Corbi, R., Pozo-Rico, T., Castejón, J.-L. and Sánchez-Almeida, T. (2022) ‘Predictors of university attrition: Looking for an equitable and sustainable higher education’, *Sustainability*, Vol. 14, No. 17, p. 10994. https://doi.org/10.3390/su141710994

Villar, A. and de Andrade, C. R. V. (2024) ‘Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study’, *Discover Artificial Intelligence*, Vol. 4, No. 1, p. 2. https://doi.org/10.1007/s44163-023-00079-z

Villegas-Ch, W., Govea, J. and Revelo-Tapia, S. (2023) ‘Improving student retention in institutions of higher education through machine learning: A sustainable approach’, *Sustainability*, Vol. 15, No. 19, p. 14512. https://doi.org/10.3390/su151914512

Won, H.-S., Kim, M.-J., Kim, D., Kim, H.-S. and Kim, K.-M. (2023) ‘University student dropout prediction using pretrained language models’, *Applied Sciences,* Vol. 13, No. 12, p. 7073. https://doi.org/10.3390/app13127073

Yaqin, A., Laksito, A. D. and Fatonah, S. (2021) ‘Evaluation of backpropagation neural network models for early prediction of student’s graduation in XYZ University’, *International Journal on Advanced Science, Engineering and Information Technology*, Vol. 11, No. 2, pp. 610–617. https://doi.org/10.18517/ijaseit.11.2.11152

Yaqin, A., Rahardi, M. and Abdulloh, F. F. (2022) ‘Accuracy enhancement of prediction method using SMOTE for early prediction student’s graduation in XYZ University’, *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 6, pp. 418–424. https://doi.org/10.14569/IJACSA.2022.0130652

Zanellati, A., Zingaro, S. P. and Gabbrielli, M. (2024) ‘Balancing performance and explainability in academic dropout prediction’, IEEE Transactions on Learning Technologies, Vol. 17, pp. 2086–2099. https://doi.org/10.1109/TLT.2024.3425959

ACADEMIC PRODUCTIVITY DYNAMICS IN COLOMBIAN SOCIAL SCIENCE PROGRAMS: A PCA–MALMQUIST INDEX APPROACH (2020–2023)

Enrique De La Hoz[1]✉
Carlos Garcia-Yerena[1]
Rohemi Zuluaga-Ortiz[3]

[1]Universidad del Magdalena, Colombia

[2]Universidad Tecnológico de Comfenalco, Colombia

[3]Institución Universitaria Mayor de Cartagena, Colombia

✉ enriquedelahoz@unimagdalena.edu.co

ABSTRACT

Higher education institutions need timely, explainable tools to identify students at risk of low performance on large-scale examinations and to guide targeted academic support strategies. In response to this challenge, this study proposes an explainable machine learning framework to predict undergraduate students' performance levels in Colombia's SABER PRO examination. Using student background variables (e.g., gender, region, school type, parental education, and occupation) and SABER 11 standardised test scores (Critical Reading, Mathematics, Citizenship Skills, Science, and English), we formulate a binary classification problem that distinguishes desirable outcomes (levels 3–4) from non-desirable outcomes (levels 1–2). We benchmark baseline models against non-linear learners, including XGBoost, GLMNET, SVM, DT, and LDA, using a 10-fold cross-validation protocol with systematic hyperparameter tuning. Model performance is assessed through confusion matrices and AUC scores. To support educational decision-making, we complement predictive results with explainability analyses, including global feature importance and individual-level explanations via SHAP, enabling transparent identification of the key drivers behind performance levels. The proposed approach provides actionable learning analytics to guide early academic support, promote responsible and transparent educational decision-making, and improve the likelihood of desirable SABER PRO achievement.

KEYWORDS

Academic efficiency, educational analysis, Malmquist Index, social sciences

HOW TO CITE

De La Hoz E., Garcia-Yerena C., Zuluaga-Ortiz R. (2026) 'Academic Productivity Dynamics in Colombian Social Science Programs: A PCA–Malmquist Index Approach (2020–2023)', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 91–100. http://dx.doi.org/10.7160/eriesj.2026.190107

Article history
Received
February 20, 2026
Received in revised form
March 18, 2026
Accepted
March 24, 2026
Available on-line
March 31, 2026

Highlights

- *PCA improved dimensionality reduction in academic productivity assessment.*
- *Regional disparities shaped productivity in Colombian Social Science programs.*
- *Technological change was the main driver of productivity growth.*
- *Productivity gaps reflect unequal regional capacity for academic modernization.*

INTRODUCTION

Higher education systems face growing pressure to demonstrate effectiveness and equity while operating under heterogeneous regional constraints (Baciu et al., 2025). In Colombia, academic outcomes and institutional capacities vary substantially across departments, partly reflecting differences in digital infrastructure, teaching resources, and financing. These disparities motivate the use of objective, reproducible methods to evaluate academic efficiency and to understand how productivity evolves at the regional level.

Most empirical efficiency studies in higher education rely either on static frontier estimates or on fragmented indicator sets that limit comparability across time and space (Jain and Gulati, 2025). When the objective is to evaluate intertemporal performance, the Malmquist productivity index (MI) provides a well-established decomposition of productivity change into shifts in the best-practice frontier (technological change) and movements toward or away from that frontier (efficiency change). However, MI results are sensitive to input and output selection, and rich educational databases may introduce redundancy and multicollinearity among competency indicators. Consequently, estimating educational productivity enables the identification of gaps and opportunities that lead to better planning and resource allocation. Besides, academic skills are not uniform across different university programs. For an educational decision-maker, it would be important to know how efficiency varies over time, allowing for the assessment of the quality of education transversally (Sánchez-Sánchez et al., 2024; Zuluaga-Ortiz et al., 2023).

In Colombia, higher education presents a multifaceted problem with at least three fundamental causes. First, institutional resources are not evenly distributed across regions. Several studies report academic gaps between regions, driven by factors such as technological infrastructure, teacher training, and financing, which generate significant differences in academic productivity (Metzger and Shenai, 2021; Sierra-González and Ramos-Pérez, 2021). Accordingly, Gallimore (2025) shows that peripheral regions like Chocó and La Guajira have 40% less access to digital platforms than metropolitan areas like Bogotá and Antioquia, restricting their capacity to implement educational innovations. Second, there are methodological and structural gaps in the evaluation of academic efficiency. The absence of standardised metrics for measuring productivity leads to inaccurate diagnoses. Several studies identify that 78% of Ibero-American institutions use incongruous indicators, distorting comparisons between Higher Education Institutions (HEIs) (Agaronnik et al., 2022; Segbenya et al., 2024).

Third, there is a low integration of the factors that drive high productivity. Recent research shows that educational strategies and policies often fail to integrate variables such as technology and human talent. Delahoz-Dominguez et al. (2020) state that considering only one dimension of productivity increases productivity by just 0.3%, whereas considering all factors could boost it by up to 12%. Finally, these challenges are intensified by regressive funding mechanisms. Authors quantify that departments with low productivity levels receive at least 34% fewer resources, thereby widening the aforementioned gap each year (Barbosa-Camargo et al., 2021; Cai and Lönnqvist, 2022).

Consequently, in the research by Agasisti and Johnes (2015), the methodology was applied to European universities, identifying technological change (TC) as the main driver of productivity gains, while pure technical efficiency (PECH) explains system stability. Gao et al. (2022) combined the Malmquist Index with Principal Component Analysis (PCA) to successfully identify groups with diverse performance levels.

Thus, the study contributes not only to the measurement of efficiency but also to the discussion of responsibility in higher education, particularly in relation to regional equity and accountable resource allocation. Therefore, this paper addresses these challenges by combining PCA with a DEA-based Malmquist framework to investigate how academic productivity in Colombian social science programmes changes across departments over the 2020–2023 period, and whether these changes are primarily explained by technological change or by shifts in efficiency. Its theoretical contribution lies in reducing "indicator fragmentation" in academic productivity assessment by explicitly linking (i) the latent structure of competencies identified through PCA with (ii) a parsimonious and reproducible Data Envelopement Analysis (DEA) production specification for dynamic productivity measurement.

LITERATURE REVIEW

Efficiency in higher education

Analysing several reviews of the higher education efficiency literature, non-parametric DEA emerges as the dominant methodological approach, accounting for roughly 50% to 70% of published studies, while Stochastic Frontier Analysis (SFA) constitutes most of the remainder (Naderi, 2022; Rella et al., 2025; Ye et al., 2025)and faculties within colleges of a comprehensive university in Iran, we simultaneously evaluate efficiency scores of departments, faculties, colleges, and the university. It has been shown that: (1. These studies typically rely on cross-sectional designs or short panel datasets and employ single-frontier models that provide a "snapshot" of technical or cost efficiency relative to a best-practice frontier at a given point in time (Alvarez-Sández et al., 2023; Dipierro and Witte, 2024; Jain and Gulati, 2025) like any organization, must attend to the needs of the environment to provide quality services. Among the various aspects related to educational quality, administrative efficiency management has gained interest in recent times. This is due to the need to optimize resources and streamline the daily operations of an educational institution. This scoping review examines how efficiency is being measured in HEIs. By contrast, dynamic productivity approaches, such as the Malmquist index, as well as models that distinguish persistent inefficiency, remain comparatively less common. However, more recent contributions suggest growing interest in capturing intertemporal changes in institutional performance (Ye et al., 2025) pure technical efficiency (PTE).

Reviews of the higher education efficiency literature show a marked reliance on single indicators or a limited set. In the teaching dimension, the most common outputs are degrees completed or the number of graduates, while value-added measures and learning quality indicators remain comparatively rare (Rella et al., 2025). In research, studies usually rely on either publication counts or grant income, with ongoing debate over which proxy is more appropriate; both tend to produce highly correlated efficiency rankings while capturing different dimensions of performance (Oliveira-Melo et al., 2025). Likewise, only a small minority of studies explicitly incorporate quality-related variables or third-mission outputs, such as patents or community engagement activities (Liu et al., 2024). Although the literature increasingly acknowledges the importance of multi-mission frameworks and institutional heterogeneity—for example, through multilevel frontiers, meta-frontiers, and strategic clustering—most empirical applications still operationalize efficiency through relatively narrow and fragmented combinations of inputs and outputs (Agasisti and Berbegal-Mirabent, 2020; Ferro and D'Elia, 2020; Liu et al., 2024).

Malmquist and productivity estimation

The Malmquist Productivity Index (MPI) has become a widely used tool for analysing productivity change in HEIs, particularly in multi-output settings where teaching, research, and third-mission activities jointly define institutional performance. Evidence from diverse national contexts—including Europe, China, Malaysia, New Zealand, Spain, and Colombia—shows that the MPI is well suited to tracking changes in total factor productivity over time and decomposing them into efficiency change and technological change components (Brintseva, 2024; Parteka and Wolszczak-Derlacz, 2013; Wang et al., 2020; Xiao et al., 2024). At the same time, the validity of MPI-based estimates depends critically on the careful specification of inputs and outputs, as well as on the treatment of potential sources of bias such as environmental conditions, technological heterogeneity, and statistical uncertainty (Guo and Ye, 2025; Thanassoulis et al., 2011). Recent studies using more advanced extensions, including the global Malmquist index, metafrontier approaches, bootstrapping procedures, and three-stage DEA models, suggest that these methods improve both robustness and comparability across time periods and institutional groups (Mehrolhassani et al., 2019). Overall, the literature supports the MPI as a valuable framework for dynamic productivity analysis in HEIs, provided that its application is grounded in strong methodological rigour.

Inputs and Outputs in Malmquist Productivity Analysis

Malmquist Productivity Index (MPI)–DEA results in higher education are highly dependent on how the production process is specified: which inputs/outputs are chosen, how technology is modelled over time, and what assumptions are made about scale and the environment (Pourmahmoud and Bagheri, 2023). In western China's higher education, Guo and Ye (2025) build a three-stage DEA + global Malmquist model using province-level inputs (faculty, expenditure structures, human capital) and outputs (student and research performance). They show that adjusting for environmental variables and random shocks in a threestage framework changes the level and decomposition of efficiency and TFP, revealing that unadjusted models overestimate efficiency and obscure the role of scale vs. pure technical efficiency. In a similar approach, Arbona et al. (2022) apply a metafrontier Malmquist–Luenberger to education systems, allowing for heterogeneous technologies and "good" and "bad" outputs (e.g., performance and inequality). They show that incorporating bad outputs and metafrontiers changes the interpretation: part of what would be labeled "technical change" in a simpler MPI becomes technologygap and qualityorientation effects across groups.

Consequently, outputs restricted to volume (graduates, enrolments) tend to show scale-driven efficiency change, while including research outputs, quality, or equity indicators often reallocates variation into the technological change term or reveals regression in quality despite expansion (Brintseva, 2024; Xue et al., 2021). In contrast, contemporaneous Malmquist (frontier by year) can yield noncircular and inconsistent productivity paths; shifts in the frontier between t and $t + 1$ are very sensitive to sample composition and outliers (Afsharian and Ahn, 2015). Besides, Robust, uncertain, or bootstrapbased MPI extensions show that small perturbations in inputs/outputs can materially change efficiency and MPI scores, indicating sensitivity to data uncertainty and to the particular index formulation used to compute EC and TC (Peykani et al., 2025).

Malmquist and Principal Component Analysis

(Bo-xin et al., 2007) propose an "improved DEAbased MPI" that explicitly incorporates PCA to deal with multicollinearity among variables and weak links between DMUs and input/output indices 9. Using panel data from 10 Chinese container ports, they show that the PCA-enhanced MPI provides a more practical dynamic performance evaluation. In contrast, conventional MPI suffers from correlated variables and poorly structured index sets.

More broadly, multiple DEA–PCA studies show that PCA improves discrimination (fewer spurious efficient units) and stabilises efficiency scores when dimensionality is high, which is precisely the same DEA frontier that underlies MPI. For instance, Adler and Golany (2002) reduce the "curse of dimensionality" by using PCA to derive assurance-region constraints on DEA weights; three PCA–DEA formulations are shown to "noticeably improve the strength of DEA models". Alternatively, using Monte Carlo simulations, PCA–DEA consistently yields more accurate classification of efficient/inefficient units than variable-reduction methods, reducing misclassification across all basic DEA models (Adler and Yazhemsky, 2010). Overall, empirical PCA–DEA applications (corporate performance, public services) find far fewer "efficient" units and lower average efficiency, interpreted as a more realistic measure of performance rather than inflated efficiency due to too many variables (Liang et al., 2009; Lim et al., 2018).

Since MPI is built from DEA scores over time, improving the frontier's stability and discrimination via PCA is expected to yield more robust MPI trajectories, as illustrated by the improved DEA-based MPI (Bo-xin et al., 2007). This is direct evidence that PCA–DEA integration can improve the robustness of MPI-type productivity tracking.

MATERIALS AND METHODS

The original database contained 11,099 individual student-level observations collected between 2020 and 2023. However, the decision-making units (DMUs) in the DEA–Malmquist analysis were not individual students, but 23 Colombian departments observed over time, as the study aimed to compare territorial dynamics in academic productivity. Accordingly, the microdata was aggregated by department and year, generating a balanced panel for the study period. For each department-year, the variables associated with prior academic conditions (Sabre 11) and university outcomes (Sabre Pro) were summarised using central tendency indicators consistent with the latent dimensions identified through PCA. This aggregation strategy preserved the informational richness of the original database while aligning the empirical specification with the study's regional-comparative objective.

For the construction of the results, the public database of the Colombian Institute for the Evaluation of Education Quality (ICFES) is used (ICFES, 2022). This database relates the assessment of state competencies in high school (Saber 11) and university (Saber PRO) (See Table 1). Furthermore, it is important to note that the competencies assessed in Saber 11 are the inputs to the research's academic production function, while those assessed in Saber PRO are its outputs. Finally, the DMUs in the study are the departments of Colombia.

Component	Module	Brief description of what is assessed	Scale
Saber 11	Critical Reading	Ability to understand, interpret, and critically evaluate texts from everyday and academic contexts, including taking a critical position on their content.	0-300
	Mathematics	Ability to address situations that can be solved using mathematical tools, using them to understand situations, transform information, justify statements, and solve problems.	
	Social and Citizenship	Knowledge and skills to understand the social world, analyse events, arguments, concepts, and contexts, and issue critical judgments in civic and social situations.	
	Natural Sciences	Ability to understand and use concepts, notions, and theories from the natural sciences to solve problems, and to explain phenomena in nature based on observations, patterns, and scientific concepts.	
	English	Reading comprehension and communicative abilities in English, focused on language use at the reading level.	
Saber Pro	Critical Reading	Ability to understand, interpret, and critically evaluate written texts, including identifying explicit and implicit ideas, argumentative structure, and purpose.	0-100
	Quantitative Reasoning	Ability to interpret quantitative information, apply mathematical reasoning, and solve problems using numerical, graphical, and statistical representations.	
	Written Communication	Ability to produce a coherent and relevant written argument in response to a prompt, with attention to structure, clarity, and communicative purpose.	
	Citizenship Skills	Ability to analyse social situations and civic issues, reason about coexistence, rights, and democratic participation, and evaluate arguments in public-life contexts.	
	English	Reading comprehension and language use in English, aligned with CEFR/MCER proficiency levels (A1-B2).	

Table 1*:* **Variable's description**

The methodological design of the research is structured in two stages: first, a Principal Component Analysis (PCA) is carried out using an orthogonal linear transformation to identify performance profiles based on the DMUs' location in the quadrants. Second, a model is built using the Malmquist Index analysis methodology to evaluate academic efficiency between 2020 and 2023. This analysis is conducted using the production function shown in Figure 1.

The DEA model was specified with an output-oriented perspective, given that the analysis aimed to evaluate the extent to which departments transformed initial academic conditions into higher education outcomes. This orientation is appropriate in the educational context because territorial units have limited capacity to modify incoming students' characteristics in the short term. In contrast, institutional and regional policies are expected to improve academic outcomes by better leveraging available educational conditions and support structures. In this sense, the model assesses the potential expansion of academic results conditional on the observed input structure. Thus, the model was estimated under variable returns to scale (VRS) to account for structural heterogeneity across Colombian departments.

To preserve the discriminatory capacity of DEA, the dimensionality of the production function was kept parsimonious relative to the number of DMUs. In line with standard recommendations in the DEA literature, the number of observations must be sufficiently larger than the total number of inputs and outputs included in the model (Dyson et al., 2001; Khezrimotlagh, Cook and Zhu, 2021). For that reason, PCA was used as a preliminary step to organize the original indicators into coherent dimensions and avoid overparameterization. This procedure reduced redundancy among variables and contributed to a more stable and reproducible frontier estimation, preventing the excessive concentration of efficiency scores near unity that often arises when too many correlated indicators are introduced into the model (Cinca and Molinero, 2004). Therefore, in methodological terms, PCA was not used as a substitute for DEA, but as a complementary procedure to structure the information space, reduce redundancy among indicators, and support a more parsimonious specification of the academic production function. This improves the transparency and reproducibility of the subsequent DEA–Malmquist estimation.

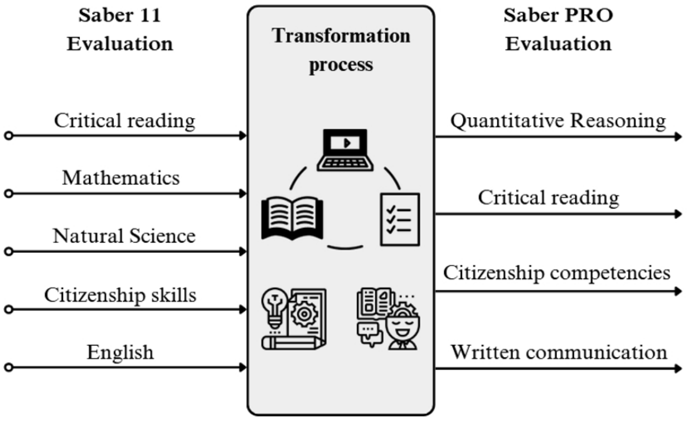


Figure 1*:* **Academic production function**

RESULTS

The first stage of the method involves data exploration through Principal PCA. For this, a graphical representation of the PCA in two dimensions is constructed (See Figure 2). The relationship of the variables with the principal components shows that they are grouped as follows: the first quadrant (Q1) groups the competencies Saber 11 Quantitative Reasoning (RC_11), Saber 11 Citizenship Competencies (CC_11), and Saber PRO Quantitative Reasoning (RC_PRO). The second and third quadrants (Q2, Q3) have no academic competencies. Finally, the fourth quadrant (Q4) includes the competencies Saber 11 Mathematics (MAT_11) and Natural Sciences (NAT_11), Saber PRO Critical Reading (LC_PRO), Citizenship Competencies (CC_PRO), English and Written Communication (CE_PRO), and Saber 11 English (ING_11). It is noteworthy that these competencies determine the characterization of the quadrants; thus, if an observation falls in the fourth quadrant, it indicates high performance related to the competencies there.

The distribution of observations across the quadrants is as follows: the second quadrant accounts for the majority (32%), followed by the first (25%), then the third (24%), and finally the fourth (20%). This distribution may reflect how competencies align based on their statistical similarities and their weight within the model. The lower percentage in the fourth quadrant could imply that the competencies in that quadrant show greater dispersion or lower weight in the first principal component.

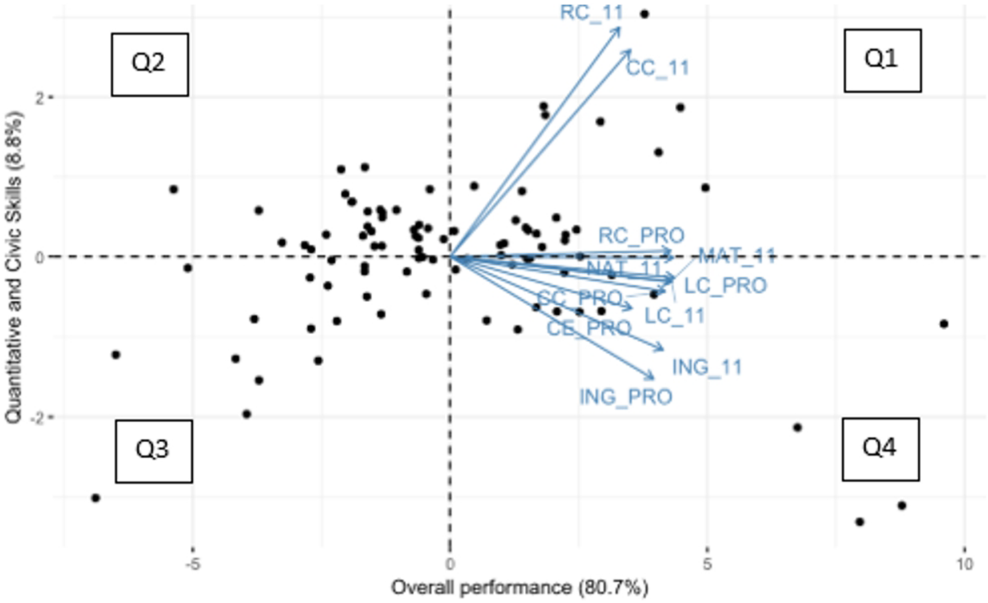


Figure 2*: PCA two-dimensional biplot of academic competencies*

Consequently, Table 2 is constructed to break down the PCA results by year. Consequently, in 2020, observations were concentrated in the third quadrant (43%), and combining this with the concentration in the second quadrant (14%) yields 57%, indicating that over half of the study's observations showed low performance. In 2021, the concentration in the second and third quadrants was 61%, up from 2020, indicating a deterioration in performance. Similarly, in 2022 and 2023, the concentration in quadrants two and three was 52%, showing a decrease of 9% from 2021 to 2022, which then remained stable in the transition to 2023.

Furthermore, the behaviour in the fourth quadrant is important to note as it has the highest share of academic competencies. This quadrant had the lowest concentration of observations across all years. However, in 2022 and 2023, it increased to 27%, suggesting an improvement in the population's overall performance.

Year	Quadrant			
	1	2	3	4
2020	7 (30%)	3 (14%)	10 (43%)	3 (13%)
2021	6 (26%)	6 (26%)	8 (35%)	3 (13%)
2022	5 (22%)	10 (43%)	2 (9%)	6 (26%)
2023	5 (22%)	10 (43%)	2 (9%)	6 (26%)

Table 2: Distribution of Observations by Year in the PCA 2D Biplot.

The second stage of the method involves analysing the academic production function using the Malmquist Index. Table 3 shows the results of the Malmquist Index indicators (MI: Malmquist Index; TC: technological change; PECH: pure technical efficiency; SECH: scale efficiency). For the first period (2020-2021), the productivity indicator MI is less than 1 (MI = 0.96), indicating a decrease in overall system performance. This is associated with the technological change indicator (TC = 0.95); however, pure technical efficiency (PECH = 1.01) showed a slight improvement, and its scale efficiency (SECH = 1.0) remained constant. For the second period (2021-2022), a recovery in productivity is observed (MI = 1.01), with an increase in the technological change indicator (TC = 1.01), and stability in both the technical efficiency indicator (PECH = 1.0) and the scale efficiency indicator (SECH = 1.0). Finally, for the third period (2022-2023), the overall system productivity level remained constant (MI = 1.0). Meanwhile, there was an improvement in the technological change indicator (TC = 1.01), but a regression in the pure technical efficiency indicator (PECH = 0.99), while scale efficiency remained stable (SECH = 1.0).

Period	MI	TC	PECH	SECH
2020-2021	0.96	0.95	1.01	1.00
2021-2022	1.01	1.01	1.00	1.00
2022-2023	1.00	1.01	0.99	1.00

Table 3: Overall Malmquist Index Results

On the other hand, Table 4 presents the Malmquist indicators by department. Overall, departments with improvements in their academic productivity (MI > 1) are: Atlántico (1.01), Bolívar (1.01), and Magdalena (1.03). These three departments show progress in their technological indicator, and Bolívar also shows an increase in technical efficiency. Among departments with stable productivity (MI = 1.0) are Caldas, Cauca, Cundinamarca, La Guajira, and Risaralda. In contrast, departments with lower productivity levels (MI < 1) include Huila, Quindío, Tolima, Santander, Nariño, and Cesar. It is seen that in some cases, their technological change and pure technical efficiency indices are also below 1.

Consequently, significant differences between regions are evident. For instance, the department of Huila has the lowest overall productivity performance (MI = 0.95), with regressions in its technological change indicator (TC = 0.96) and technical efficiency (PECH = 0.99), suggesting that efforts are insufficient to improve educational quality. In contrast, for the department of Magdalena, its productivity index (MI = 1.03) is the highest, supported by its level of technological change (TC = 1.03), while the other indicators remain constant. In the cases of Caldas and Cauca, we observe stable productivity indicators (MI = 1.00), while improvements occur in pure technical efficiency (PECH = 1.01) and technological change decreases (TC = 0.99). Lastly, in the case of Córdoba and La Guajira, despite the productivity levels (MI = 0.99; MI = 1.00), the respective scale efficiency values both indicate an upward trend (SECH = 1.01), suggesting that both may be employing effective quality improvement strategies.

DMU	MI	TC	PECH	SECH
Antioquia	0.99	0.99	1.00	1.00
Atlántico	1.01	1.01	1.00	1.00
Bogotá	0.99	0.99	1.00	1.00
Bolívar	1.01	1.00	1.01	1.00
Boyacá	0.99	0.99	1.00	1.00
Caldas	1.00	0.99	1.01	1.01
Cauca	1.00	0.99	1.01	1.00
Cesar	0.98	0.98	1.00	1.00
Choco	0.99	0.99	1.00	1.00
Córdoba	0.99	0.99	1.00	1.01
Cundinamarca	1.00	1.00	1.00	1.00
Huila	0.95	0.96	0.99	1.00
La guajira	1.00	0.99	1.00	1.01
Magdalena	1.03	1.03	1.00	1.00
Meta	0.99	0.99	1.00	1.00
Nariño	0.97	0.98	0.99	1.00
Norte Santander	0.99	0.98	1.00	1.00
Quindío	0.96	0.97	0.99	1.00
Risaralda	1.00	1.00	1.00	1.00
Santander	0.97	0.98	0.99	1.00
Sucre	0.98	0.99	0.99	1.00
Tolima	0.97	0.98	0.99	1.00
Valle	0.99	0.98	1.00	1.00

Table 4: Malmquist Index Results by Department

Table 5 presents the results of the Malmquist indicators by period and department. In the first period (2020-2021), most departments experienced a decline in productivity (MI < 1), likely due to insufficient effort in the technological aspect (TC < 1). On the contrary, in the second period (2021-2022), some departments improved their productivity levels (MI ≥ 1), and technological transition (TC > 1) enabled them to do so. For instance, departments such as Antioquia, Atlántico, Bolívar, Caldas, Cauca, and especially Cundinamarca (TC = 1.12) showed improvements, suggesting that more effective strategies were used. This recovery, which was present in the second period, allowed a consolidation and recovery in the third period (2022-2023), as the department of Magdalena presented an improvement in its overall productivity level (MI = 1.15) and its technological transition indicator (TC = 1.15), which evidently positions this department as a benchmark in the region. The behaviours of each department are differentiated, finding marked trends. For instance, Magdalena shows sustained and exemplary growth in the last period. Cundinamarca shows a disruptive improvement in the second period (2021-2022), followed by a regression in the third period (2022-2023). Atlántico keeps a stable increasing behaviour with MI = 0.99 in the first period, and then MI = 1.02 in the second and third periods. The department of Bolívar shows sustained growth in all its indicators; similarly, the department of Boyacá moved from a decline in 2020 to positive behaviour in 2022. In contrast, the department of Huila shows low productivity performance in all periods. Departments like Risaralda and Caldas show continuous improvement, while Tolima shows an irregular trajectory. The department of Quindío keeps a constant level of productivity.

DMU	MI			TC			PECH			SECH		
	2020-2021	2021-2022	2022-2023	2020-2021	2021-2022	2022-2023	2020-2021	2021-2022	2022-2023	2020-2021	2021-2022	2022-2023
Antioquia	0.95	1.02	1.00	0.95	1.03	0.99	1.00	0.99	1.01	1.00	1.00	1.00
Atlántico	0.99	1.02	1.02	0.99	1.04	0.99	1.00	0.98	1.02	1.00	1.00	1.00
Bogotá	0.98	1.00	0.99	0.97	1.00	1.00	1.01	1.00	0.99	1.00	1.00	1.00
Bolívar	0.97	1.00	1.05	0.94	1.02	1.03	1.02	0.98	1.02	1.01	1.00	1.00
Boyacá	0.93	1.01	1.02	0.93	1.01	1.02	1.00	1.00	1.00	1.00	1.00	1.00
Caldas	0.98	1.02	1.01	0.94	1.02	1.01	1.02	1.00	1.00	1.01	1.00	1.00
Cauca	0.98	1.01	1.01	0.94	1.01	1.01	1.04	1.00	1.00	1.00	1.00	1.00
Cesar	0.94	1.00	1.00	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Choco	0.94	1.05	0.98	0.94	1.02	1.00	1.00	1.00	1.00	1.00	1.03	0.97
Córdoba	0.96	1.03	0.99	0.94	1.00	1.01	0.99	1.02	0.98	1.02	1.00	1.00
Cundinamarca	0.98	1.12	0.93	0.98	1.12	0.93	1.00	1.00	1.00	1.00	1.00	1.00
Huila	0.94	0.96	0.97	0.94	0.97	0.97	1.00	0.99	0.99	1.00	1.00	1.00
La guajira	0.98	0.98	1.02	0.96	0.98	1.02	1.00	1.00	1.00	1.03	1.00	1.00
Magdalena	0.94	1.02	1.15	0.94	1.02	1.15	1.00	1.00	1.00	1.00	1.00	1.00
Meta	0.97	0.99	1.00	0.97	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.01
Nariño	0.94	0.99	0.98	0.93	0.99	1.03	1.00	1.00	0.96	1.01	1.00	0.99
Norte Santander	0.99	0.98	0.98	0.95	1.00	1.00	1.04	0.99	0.98	1.01	0.99	1.00
Quindío	0.92	1.00	0.96	0.92	1.00	0.99	1.00	1.00	0.97	1.00	1.00	1.00
Risaralda	0.95	1.05	1.02	0.95	1.04	1.01	1.01	1.00	1.00	0.99	1.00	1.01
Santander	0.94	0.99	0.97	0.94	1.02	0.99	1.00	0.98	0.99	1.00	1.00	0.99
Sucre	0.99	0.97	0.98	0.98	0.97	1.00	1.00	1.00	0.98	1.01	1.00	1.00
Tolima	0.91	1.00	0.99	0.94	1.00	1.00	0.97	1.01	0.99	1.00	1.00	1.00
Valle	0.95	1.02	0.99	0.94	1.02	0.99	1.01	1.00	1.00	1.00	1.00	1.00

Table 5: Malmquist Index Results by Year and Department

DISCUSSION

This research contributes to the frontier of knowledge in the field of efficiency analysis in education by using two data science tools to evaluate the dynamics of productivity in social science programs in Colombia. The combined use of the tools addresses two related drawbacks in the literature: the lack of indicators and the predominance of static assessments. In this order of ideas, the research is aligned with recent works that conclude that the analysis of academic productivity generates valuable information because it analyzes efficiency dynamically and not as isolated indicators (Arbona et al., 2022; Fu and See, 2022)

However, in terms of results, the most consolidated performance profile is in quadrant IV. This quadrant indicates that the productivity of the evaluated programs is related to the areas of critical reading, English, and written communication. It is important to note that, although the concentration of DMUs in this quadrant increased in the last period analysed, the concentration of DMUs in the lower performing profiles was maintained. This result is supported by the literature, which evidences that academic performance is strongly related to environmental inequalities derived from the inequitable distribution of state resources (Arbona et al., 2022; Timmis and Valladares-Celis, 2025)

In line with the above, the results of the technological change indicator indicate that improvements in academic productivity depend on the ability to modernize teaching, learning, and organizational conditions. This result aligns with recent studies that conclude that factors such as technical progress, digitalization of processes, and institutional modernization play a fundamental role in academic productivity indicators (Fu and See, 2022; Gao and Wang, 2023; Hieu and Niem, 2024; Liu et al., 2023).

On the other hand, in a global analysis, Magdalena stands out as a point of reference for the last period of analysis. In addition, Bolívar and Atlántico also presented favorable trajectories, while Quindío, Huila, and Nariño remained below the productivity frontier. In this sense, the results show that differences in performance are linked to unequal territorial capacity rather than only to institutional efforts.

Finally, this study has several limitations that should be acknowledged. First, although the DEA–Malmquist framework allows the analysis of productivity change over time, the results remain sensitive to the specification of inputs and outputs, even when PCA is used to reduce redundancy and improve parsimony. In line with standard DEA recommendations, the number of observations must be sufficiently larger than the total number of inputs and outputs included in the model, which requires maintaining a simplified production structure at the departmental level. Second, aggregating 11,099 student-level observations into 23 departments improves comparability across territories, but it may also mask within-department heterogeneity and institutional differences among higher education providers. Third, the study covers the 2020–2023 period, which is analytically relevant but still relatively short for capturing longer-term structural changes in academic productivity. Future research could address these limitations by extending the time horizon, incorporating additional contextual variables related to institutional quality and regional investment, and conducting complementary analyses at the university or program level to better capture intra-regional variation. Such extensions would help strengthen the explanatory power and policy relevance of the findings.

CONCLUSION

This study examined academic productivity in Colombian Social Science programs as a dynamic and multidimensional expression of educational performance. The findings show that productivity is not determined solely by final academic outcomes, but by the interaction among students' incoming competency profiles, the academic conditions of higher education institutions, and the broader regional context in which these processes occur. The increase in the high-performance quadrant over the study period suggests that some territories improved their academic positioning, while the Malmquist decomposition showed that technological change was the main driver of productivity growth. At the same time, the departmental analysis confirmed that academic productivity in Colombia is markedly heterogeneous, with some regions displaying stronger trajectories of innovation and adjustment than others. These results reinforce the idea that the quality of higher education should be assessed through dynamic, territorially sensitive approaches rather than isolated, static indicators.

From a methodological perspective, integrating Principal Component Analysis with Malmquist productivity analysis provides a coherent framework for addressing indicator fragmentation and evaluating changes in academic efficiency over time. In particular, PCA helped organize the dataset's information structure through an unsupervised learning approach. At the same time, the DEA–Malmquist framework enabled identification of whether productivity changes were associated with technological progress, pure efficiency gains, or scale effects. Beyond its technical contribution, the study also has implications for educational policy and responsibility in higher education. The results suggest the need to strengthen academic quality through differentiated regional strategies to improve digital infrastructure, pedagogical capacity, and institutional support systems, especially in departments facing less favorable productivity dynamics. In this sense, promoting academic productivity should not be understood only as an efficiency objective, but also as part of a broader commitment to equity, accountability, and balanced territorial development in higher education.

REFERENCES

Adler, N. and Golany, B. (2002) 'Including principal component weights to improve discrimination in data envelopment analysis', *Journal of the Operational Research Society*, Vol. 53, No. 9, pp. 985–991. https://doi.org/10.1057/palgrave.jors.2601400

Adler, N. and Yazhemsky, E. (2010) 'Improving discrimination in data envelopment analysis: PCA-DEA or variable reduction', *European Journal of Operational Research.*, Vol. 202, No. 1, pp. 273–284. https://doi.org/10.1016/j.ejor.2009.03.050

Afsharian, M. and Ahn, H. (2015) 'The overall Malmquist index: a new approach for measuring productivity changes over time', *Annals of Operations Research*, Vol. 226, No. 1, pp. 1–27. https://doi.org/10.1007/s10479-014-1668-5

Agaronnik, N., Xiong, G. X., Uzosike, A., Crawford, A. M., Lightsey, H. M., Simpson, A. K. and Schoenfeld, A. J. (2022) 'The role of gender in academic productivity, impact, and leadership among academic spine surgeons', *The spine journal: official journal of the North American Spine Society*, Vol. 22, No. 5, pp. 716–722. https://dx.doi.org/10.1016/j.spinee.2021.12.003

Agasisti, T. and Berbegal-Mirabent, J. (2020) 'Cross-country analysis of higher education institutions' efficiency: The role of strategic positioning', *Science and Public Policy*, Vol. 48, No. 1, pp. 66–70. https://doi.org/10.1093/scipol/scaa058

Agasisti, T. and Johnes, G. (2015) 'Efficiency, costs, rankings and heterogeneity: the case of US higher education', *Studies in Higher Education*, Vol. 40, No. 1, pp. 60–82. https://doi.org/10.1080/03075079.2013.818644

Alvarez-Sández, D., Velázquez-Victorica, K., Mungaray-Moctezuma, A. and López-Guerrero, A. (2023) 'Administrative Processes Efficiency Measurement in Higher Education Institutions: A Scoping Review', *Education Sciences*, Vol. 13, No. 9, p. 855. https://doi.org/10.3390/educsci13090855

Arbona, A., Giménez, V., López-Estrada, S. and Prior, D. (2022) 'Efficiency and quality in Colombian education: An application of the metafrontier Malmquist-Luenberger productivity index', *Socio-Economic Planning Sciences*, Vol. 79, p. 101122, https://doi.org/10.1016/j.seps.2021.101122

Baciu, E.-L., Lazăr, T.-A. and Totan, R. I. (2025) 'Social goals under a neoliberal agenda: measures to promote equality in European higher education read through a Foucauldian lens', *Frontiers in Sociology*, Vol. 10, p. 1492863. https://doi.org/10.3389/fsoc.2025.1492863

Barbosa-Camargo, M. I., García-Sánchez, A. and Ridao-Carlini, M. L. (2021) 'Inequality and Dropout in Higher Education in Colombia. A Multilevel Analysis of Regional Differences, Institutions, and Field of Study', *Mathematics*, Vol. 9, No. 24, p. 3280. https://doi.org/10.3390/math9243280

Bo-xin, F., Xiang-qun, S., Zi-jian, G. and Zhuo, F. (2007) 'Study on Improved DEA-Based MPI and Its Application in Dynamic Performance Evaluation', in: *2007 International Conference on Management Science and Engineering*, pp. 446–451, https://dx.doi.org/10.1109/icmse.2007.4421888

Brintseva, O. (2024) 'A DEA-based Malmquist Productivity Index for Analysing University Performance and Competitiveness', *Krakow Review of Economics and Management/Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, Vol. 4, No. 1006, pp. 5–22. https://doi.org/10.15678/krem.17685

Cai, Y. and Lönnqvist, A. (2022) 'Overcoming the Barriers to Establishing Interdisciplinary Degree Programmes: The Perspective of Managing Organisational Innovation', *Higher Education Policy*, Vol. 35, No. 4, pp. 946–968. https://doi.org/10.1057/s41307-021-00242-0

Cinca, C. S. and Molinero, C. M. (2004) 'Selecting DEA specifications and ranking units via PCA', *Journal of the Operational Research Society*, Vol. 55, No. 5, pp. 521–528. https://doi.org/10.1057/palgrave.jors.2601705

Delahoz-Dominguez, E. J., Fontalvo, T. and Zuluaga, R. (2020) 'Evaluation of academic productivity of citizen competencies in the teaching of engineering by using the Malmquist index', *Formación Universitaria*, Vol. 13, No. 5, pp. 27–34. http://dx.doi.org/10.4067/S0718-50062020000500027

Dipierro, A. R. and Witte, K. D. (2024) 'The underlying signals of efficiency in European universities: a combined efficiency and machine learning approach', *Studies in Higher Education*, Vol. 50, No. 6, pp. 1306–1325. https://doi.org/10.1080/03075079.2024.2370948

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S. and Shale, E. A. (2001) 'Pitfalls and protocols in DEA', *European Journal of Operational Research*, Vol. 132, No. 2, pp. 245–259. https://doi.org/10.1016/S0377-2217(00)00149-1

Ferro, G. and D'Elia, V. (2020) 'Higher Education Efficiency Frontier Analysis: A Review of Variables to Consider', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 13, No. 3, pp. 140–153. https://doi.org/10.7160/eriesj.2020.130304

Fu, T.-T. and See, K. F. (2022) 'An integrated analysis of quality and productivity growth in China's and Taiwan's higher education institutions', *Economic Analysis and Policy*, Vol. 74, pp. 234–249. https://doi.org/10.1016/j.eap.2021.12.013

Gallimore, A. (2025) 'Missing Piece of the Pie: Public Expenditure on Education in Colombia', *Economics*, Vol. 8. https://doi.org/10.54014/8ZVV-XQXZ

Gao, Q. and Wang, Q. (2023) 'A Study on the Spatial–Temporal Evolution of Innovation Efficiency in Chinese Universities in the Context of the Digital Economy', *Sustainability*, Vol. 15, No. 1, p. 39. https://doi.org/10.3390/su15010039

Gao, S., Sun, H. and Wang, R. (2022) 'Audit Evaluation and Driving Force Analysis of Marine Economic Development Quality', *Sustainability*, Vol. 14, No. 11, p. 6822. https://doi.org/10.3390/su14116822

Guo, R. and Ye, M. (2025) 'Input-output efficiency, productivity dynamics, and determinants in western China's higher education: A three-stage DEA, global Malmquist index, and Tobit model approach', *PLOS One*, Vol. 20, No. 6, p. e0325901. https://doi.org/10.1371/journal.pone.0325901

Hieu, N. T. and Niem, L. D. (2024) 'Autonomy Acquisition and Performance within Higher Education in Vietnam—A Road to a Sustainable Future?', *Sustainability*, Vol. 16 No. 3, p. 1336. https://doi.org/10.3390/su16031336

ICFES (2022) *Resultados de la evaluación Saber PRO*. Available at: https://www.icfes.gov.co/web/guest/acerca-del-examen-saber-pro [Accessed 26 October 2023]

Jain, I. and Gulati, R. (2025) 'Efficiency in Higher Education: A Review of Research From 1977 to 2022', *Higher Education Quarterly*, Vol. 79, No. 1, p. e70010. https://doi.org/10.1111/hequ.70010

Khezrimotlagh, D., Cook, W. D. and Zhu, J. (2021) 'Number of performance measures versus number of decision making units in DEA', *Annals of Operations Research*, Vol. 303, No. 1, pp. 529–562. https://doi.org/10.1007/s10479-019-03411-y

Liang, L., Li, Y. and Li, S. (2009) 'Increasing the discriminatory power of DEA in the presence of the undesirable outputs and large dimensionality of data sets with PCA', *Expert Systems with Applications*, Vol. 36, No. 3, pp. 5895–5899. https://doi.org/10.1016/j.eswa.2008.07.022

Lim, J., Jang, Y., Chang, H., Park, J. and Lee, J. (2018) 'Role of multi-response principal component analysis in reliability-based robust design optimization: an application to commercial vehicle design', *Structural and Multidisciplinary Optimization*, Vol. 58, No. 2, pp. 785–796. https://doi.org/10.1007/s00158-018-1908-4

Liu, J., Jungyin, K., Jaewoo, S. W., Heechul, L. and Shah, W. U. H. (2024) 'Evaluating the efficiency, productivity change, and technology gaps of China's provincial higher education systems: A comprehensive analytical framework', *PLOS One*, Vol. 19, No. 1, p. e0294902. https://doi.org/10.1371/journal.pone.0294902

Liu, Z., Xiong, H. and Sun, Y. (2023) 'Will Online MOOCs Improve the Efficiency of Chinese Higher Education Institutions? An Empirical Study Based on DEA', *Sustainability*, Vol. 15 No. 7, p. 5970. https://doi.org/10.3390/su15075970

Mehrolhassani, M., Goudarzi, R., Feyzabadi, Y., Pourhosseini, S. and Darvishi, A. (2019) 'Efficiency and Productivity Measurement in Research Sector of Iranian Medical Sciences Universities Using Data Envelopment Analysis and Malmquist Index', *Iranian Journal of Epidemiology*, Vol. 14, pp. 1–11. Available at: https://journals.tums.ac.ir/irje/article-1-6139-en.html&sw=Research

Metzger, N. and Shenai, V. (2021) 'Economic growth and human development in OECD countries: a twenty-year study of data 2000-2019', *Journal of European Economy*, Vol. 20, No. 4, pp. 585–631. https://doi.org/10.35774/jee2021.04.585

Naderi, A. (2022) 'Efficiency measurement of higher education units using multilevel frontier analysis', *Journal of Productivity Analysis*, Vol. 57, No. 1, pp. 79–92. https://dx.doi.org/10.1007/s11123-021-00621-0

Oliveira-Melo, F. G., Barbosa, A. S. and Sant'Anna, Â. M. O. (2025) 'Efficiency of Higher Education Systems Toward Sustainable Development Goal 4: Cross-Country Analysis Based on a Bootstrap DEA Model', *IEEE Access*, Vol. 13, pp. 191097–191113. https://dx.doi.org/10.1109/access.2025.3629526

Parteka, A. and Wolszczak-Derlacz, J. (2013) 'Dynamics of productivity in higher education: cross-european evidence based on bootstrapped Malmquist indices', *Journal of Productivity Analysis*, Vol. 40, No. 1, pp. 67–82. https://doi.org/10.1007/s11123-012-0320-0

Peykani, P., Soltani, R., Tanasescu, C., Shojaie, S. E. and Jandaghian, A. (2025) ‘The Robust Malmquist Productivity Index: A Framework for Measuring Productivity Changes over Time Under Uncertainty’, *Mathematics*, Vol. 13, No. 11, p. 1727. https://doi.org/10.3390/math13111727

Pourmahmoud, J. and Bagheri, N. (2023) ‘Uncertain Malmquist productivity index: An application to evaluate healthcare systems during COVID-19 pandemic’, *Socio-Economic Planning Sciences*, Vol. 87, p. 101522. https://doi.org/10.1016/j.seps.2023.101522

Rella, A., Guillamón, M.-D., Benito, B. and Vitolla, F. (2025) ‘Efficiency determinants for higher education institutions: an empirical study of Italian public universities’, *International Journal of Productivity and Performance Management*, Vol. 75, No. 1, pp. 108–131. https://doi.org/10.1108/IJPPM-09-2024-0649

Sánchez-Sánchez, A.M., Mello-Román, J.D., Segura, M. and Hernández, A. (2024) ‘Identifying the Determinants of Academic Success: A Machine Learning Approach in Spanish Higher Education’, *Systems, Multidisciplinary Digital Publishing Institute*, Vol. 12, No. 10, p. 425. https://doi.org/10.3390/systems12100425

Segbenya, M., Senyametor, F., Aheto, S.-P. K., Agormedah, E. K., Nkrumah, K. and Kaedebi-Donkor, R. (2024) ‘Modelling the influence of antecedents of artificial intelligence on academic productivity in higher education: a mixed method approach’, *Cogent Education*, Vol. 11, No. 1, p. 2387943. https://doi.org/10.1080/2331186X.2024.2387943

Sierra-González, J. H. and Ramos-Pérez, C. E. (2021) ‘Science, Technology, Innovation, and Inclusive Development: A Country Comparison Between Colombia and Mexico’, in Orozco, L.A., Ordóñez-Matamoros, G., Sierra-González, J.H., García-Estévez, J. and Bortagaray, I. (eds.), *Science, Technology, and Higher Education: Governance Approaches on Social Inclusion and Sustainability in Latin America*, Cham: Springer International Publishing, pp. 287–343. https://doi.org/10.1007/978-3-030-80720-7_11

Thanassoulis, E., Kortelainen, M., Johnes, G. and Johnes, J. (2011) ‘Costs and efficiency of higher education institutions in England: a DEA analysis’, *Journal of the Operational Research Society*, Vol. 62, No. 7, pp. 1282–1297. https://doi.org/10.1057/jors.2010.68

Timmis, S. and Valladares-Celis, M. C. (2025) ‘Digital inequalities and the COVID legacy in higher education in the global South and North: intersecting inaccessibilities and institutional assumptions’, *Compare: A Journal of Comparative and International Education*, pp. 1–19. https://doi.org/10.1080/03057925.2025.2483691

Wang, C.-N., Tibo, H., Nguyen, V. and Duong, D. (2020) ‘Effects of the Performance-Based Research Fund and Other Factors on the Efficiency of New Zealand Universities: A Malmquist Productivity Approach’, *Sustainability*, Vol. 12, No. 15, p. 5939. https://doi.org/10.3390/su12155939

Xiao, S., Sheng, J. and Zhang, G. (2024) ‘Rising Tides of Knowledge: Exploring China’s Higher Education Landscape and Human Capital Growth’, *Journal of the Knowledge Economy*, Vol. 16, No. 1, pp. 4392–4421. https://doi.org/10.1007/s13132-024-02102-9

Xue, W., Li, H., Ali, R., Rehman, R. and Fernández-Sánchez, G. (2021) ‘Assessing the Static and Dynamic Efficiency of Scientific Research of HEIs China: Three Stage DEA–Malmquist Index Approach’, *Sustainability*, Vol. 13, No. 15, p. 8207. https://doi.org/10.3390/su13158207

Ye, Z., Khanal, R. and Cao, Z. (2025) ‘Studying on the efficiency of higher education resource allocation and its influencing factors in the western China using DEA-Malmquist and Tobit models’, *PLOS One*, Vol. 20, No. 10, p. e0334090. https://doi.org/10.1371/journal.pone.0334090

Zuluaga-Ortiz, R., Camelo-Guarin, A. and Delahoz-Domínguez, E. (2023) ‘Efficiency analysis trees as a tool to analyze the quality of university education’, *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 13, No. 4, pp. 4412–4421. https://dx.doi.org/10.11591/ijece.v13i4.pp4412-4421