

Article

Data Mining to Identify Factors Associated with University Student Retention

Yuri Reina Marín ¹, Lenin Quiñones Huatangari ², Judith Nathaly Alva Tuesta ¹, Omer Cruz Caro ^{1,*},
Jorge Luis Maicelo Guevara ³, Einstein Sánchez Bardales ⁴ and River Chávez Santos ⁵

¹ Oficina de Gestión de la Calidad, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Chachapoyas 01001, Peru; yuri.reina@untrm.edu.pe (Y.R.M.); judith.alva@untrm.edu.pe (J.N.A.T.)

² Facultad de Ingeniería Zootecnista, Biotecnología, Agronegocios y Ciencia de Datos, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Chachapoyas 01001, Peru; lenin.quinones@untrm.edu.pe

³ Escuela de Economía, Pontificia Universidad Católica del Perú, Lima 15088, Peru; a20206051@pucp.edu.pe

⁴ Programa de Doctorado en Ciencias para el Desarrollo Sustentable, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Chachapoyas 01001, Peru; einstein.sanchez@untrm.edu.pe

⁵ Facultad de Educación y Ciencias de la Comunicación, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Chachapoyas 01001, Peru; river.chavez@untrm.edu.pe

* Correspondence: omer.cruz@untrm.edu.pe

Abstract

Student retention has become a major challenge for higher education institutions due to the influence that academic, socioeconomic, family, and motivational factors exert on students' academic continuity. In this context, understanding the determinants that explain university persistence is essential for designing effective retention strategies. Based on the analysis of factors related to motivation, commitment, attitude, academic integration, and social and economic conditions, retention patterns were examined in a population of 532 university students, of whom 57.7% showed high retention, 38.2% medium retention, and 4.1% low retention. To identify the factors with the greatest influence on academic continuity, educational data mining techniques and supervised classification models were applied and evaluated using stratified 10-fold cross-validation. Tree-based ensemble models showed the most consistent predictive performance, with Random Forest achieving the best results (accuracy = 0.729 ± 0.058 ; F1-macro = 0.636 ± 0.136). Model interpretability was examined through SHAP analysis, which revealed that transportation conditions (0.249), task completion (0.170), absence of work obligations (0.168), and course completion (0.164) were the most influential predictors in the classification of retention levels. In addition, sensitivity analysis indicated that academic commitment accounts for 41.6% of the predictive impact, followed by motivation (23.5%). These findings demonstrate that student retention is shaped by the interaction of academic, motivational, and contextual factors and provide practical implications for the development of **early warning systems, personalized tutoring programs, psychosocial support initiatives, and financial assistance policies aimed at strengthening university retention.

Keywords: educational data mining; student engagement; academic persistence; predictive analytics; psychosocial factors; socioeconomic conditions



Academic Editor: Antony Bryant

Received: 5 February 2026

Revised: 12 March 2026

Accepted: 25 March 2026

Published: 27 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Student dropout in higher education is a phenomenon with social and academic impact that affects both institutions and students [1]. It is not merely a statistical issue but a

complex process involving economic, social, academic, and cultural factors that influence students' decisions to discontinue their studies [2,3]. Its consequences include a reduction in the graduation rate, a negative impact on international rankings, weakening of institutional reputation, and jeopardized financial sustainability [4]. At a social level, dropout represents a loss of human capital, delays in achieving national educational goals, and an increased risk of youth unemployment [5]. These challenges are particularly pronounced in developing countries, where structural limitations already constrain higher education systems and intensify dropout rates [6].

The concept of student retention has evolved from being understood as simply “not dropping out” to being conceived as a dynamic process of integration, commitment, and academic success [7]. Pimentel et al. [8] and Garcés et al. [9] emphasize that retention should be analyzed from a systemic perspective that considers the interaction between personal, institutional, and contextual factors. This conceptual shift reflects the recognition that university students often follow non-linear academic pathways characterized by interruptions, such as changes in major or hybrid learning modalities, which require more flexible and predictive support systems [10–13].

Research on academic retention and permanence has therefore focused on identifying the conditions that facilitate or hinder the completion of educational trajectories [14–16]. Understanding why some students successfully graduate while others drop out remains a central question for higher education systems [17,18]. In this sense, the central question “Why do some manage to graduate while others drop out?” is not solely a matter of academic curiosity, but rather a response to the need to guarantee equity and efficiency in educational systems [19,20]. Identifying these determinants enables institutions to design strategic interventions aimed at reducing educational gaps and strengthening university education, particularly in contexts characterized by socioeconomic vulnerability [15,21].

Theoretical models have provided relevant explanations, such as Tinto's longitudinal model, which argues that students drop out of their studies when they fail to integrate socially and academically into the university, thus weakening their academic and institutional commitment [22–25]. In addition, Bean and Metzner's model focuses on non-traditional students (workers, over 24 years old, with family responsibilities) and highlights the influence of external factors such as time, family, money, or self-efficacy [21,26]. Both approaches are essential, although recent literature emphasizes the need to adapt them to the profile of the 21st-century student [21]. Investigations such as those of Almalawi et al. [27] and Lamprooulos & Evangelidis [28] have developed hybrid models that integrate psychological, sociological, and technological elements, where self-efficacy, resilience, and institutional support function as predictors of academic success. Additionally, academic and social capital are increasingly recognized as mediating variables influencing student retention, particularly in vulnerable contexts [29,30].

Empirical studies indicate that the factors associated with student retention are highly diverse, ranging from academic performance to sociodemographic characteristics [31]. Variables such as academic history, age, gender, socioeconomic status, ethnicity, sense of belonging, and social support networks influence students' continuity in higher education [17,18,26]. These variables interact with each other rather than acting independently, generating heterogeneous scenarios that characterize retention as a multidimensional phenomenon [32,33]. Additional factors include academic performance indicators, such as study strategies and academic progress [34], intrinsic motivation, well-being, and self-regulation, and institutional climate [35], including quality of teachers, tutoring, support services, and retention policies [36]. They require analysis through predictive models and systematic reviews, as these tools allow estimating the risk of dropout, prioritizing the factors with the greatest explanatory weight, and guiding evidence-based interventions [36].

In this way, the understanding, monitoring, and evaluation of these factors are strengthened within the educational and social contexts in which students develop [36,37].

Within this analytical context, Educational Data Mining (EDM) has emerged as a powerful methodological approach for processing large volumes of educational data and identifying patterns associated with student retention [17,38]. Unlike traditional approaches, which tend to be reactive, EDM allows for proactive action through predictive models capable of identifying at-risk students and generating personalized interventions [6]. Machine learning techniques facilitate the extraction of meaningful insights from institutional data, strengthening retention programs and improving decision-making processes within universities [39–42]. In addition, learning analytics contributes to identifying learning patterns and providing feedback that enhances the monitoring of educational processes [43–45].

There has been remarkable progress in the use of supervised and unsupervised algorithms, such as Decision Trees, Random Forest, Support Vector Machines, Artificial Neural Networks, and XGBoost [1,2,6,26]. These methods have shown encouraging results in various international contexts, including Europe, Latin America, and Asia, confirming their global applicability for improving student retention and engagement [18,46,47]. For developing countries, the adoption of these analytical techniques represents a strategic opportunity to reduce educational inequalities and strengthen graduation indicators, which are increasingly relevant in accreditation and institutional quality assurance processes [19,48].

Despite these advances, the implementation of EDM approaches still faces several challenges. Institutional datasets are often fragmented, incomplete, or difficult to access, limiting the reliability and generalizability of predictive models [49,50]. Furthermore, models developed in universities located in highly digitalized contexts may not easily transfer to smaller institutions or those operating under resource constraints, creating challenges related to model adaptability and external validity [51]. Another critical issue concerns potential biases in predictive models when contextual variables such as gender, ethnicity, or socioeconomic background are insufficiently considered [32,52–54]. These limitations highlight the need to develop data-driven methodologies that are sensitive to institutional contexts and capable of integrating local realities into predictive modeling frameworks [55,56].

In Latin America, recent studies such as those by De la hoz et al. [57] and Flores & Nuñez [58] demonstrate the applicability of data mining in contexts with technological limitations. For example, Mexico and Colombia have used clustering techniques and neural networks to identify risk profiles and design personalized tutoring programs; likewise, in Chile and Peru, the integration of the Cross Industry Standard Process for Data Mining (CRISP-DM) model into institutional academic tracking systems has resulted in significant reductions in dropout rates [41]. These findings demonstrate that the integration of EDM techniques into higher education analytics not only advances academic research but also supports the development of strategic interventions aimed at promoting equity and sustainability within higher education systems.

In this context, understanding student retention from an analytical and predictive perspective constitutes a strategic necessity for higher education institutions, particularly in contexts characterized by structural inequalities and heterogeneous educational trajectories. This study analyzes the factors associated with university student retention using EDM techniques, with the aim of modeling and explaining how academic, socioeconomic, and institutional variables influence the continuation of studies. By integrating data-driven predictive models into the analysis of student retention, this research contributes to expanding the understanding of this phenomenon in the Latin American context, where data fragmentation and technological gaps hinder the implementation of early warning

systems. In this way, the study provides empirical evidence to strengthen institutional retention policies and support decision-making aimed at reducing student dropout in higher education.

To provide clarity and guide the reader through the development of the study, the remainder of this paper is structured as follows. The Abstract summarizes the main objective, methodological approach, and key findings of the research. Section 1 presents the context of the problem, relevant background, and the justification for the study. Section 2 describes the research design, sample characteristics, instruments, and analytical procedures employed. Section 3 presents the main empirical findings derived from the data analysis. Subsequently, Section 4 interprets these findings considering the existing scientific literature. Finally, Section 5 summarizes the principal contributions of the study and highlights their implications for future research and institutional practice.

2. Materials and Methods

2.1. Methodology

The CRISP-DM approach was used for the study; it represents a structured framework used in practical and academic settings for data exploration. This methodological scheme (Figure 1) consists of 6 stages: analysis of the business environment, familiarization with available information, data conditioning, model building, validation of results, and deployment of solutions [19].

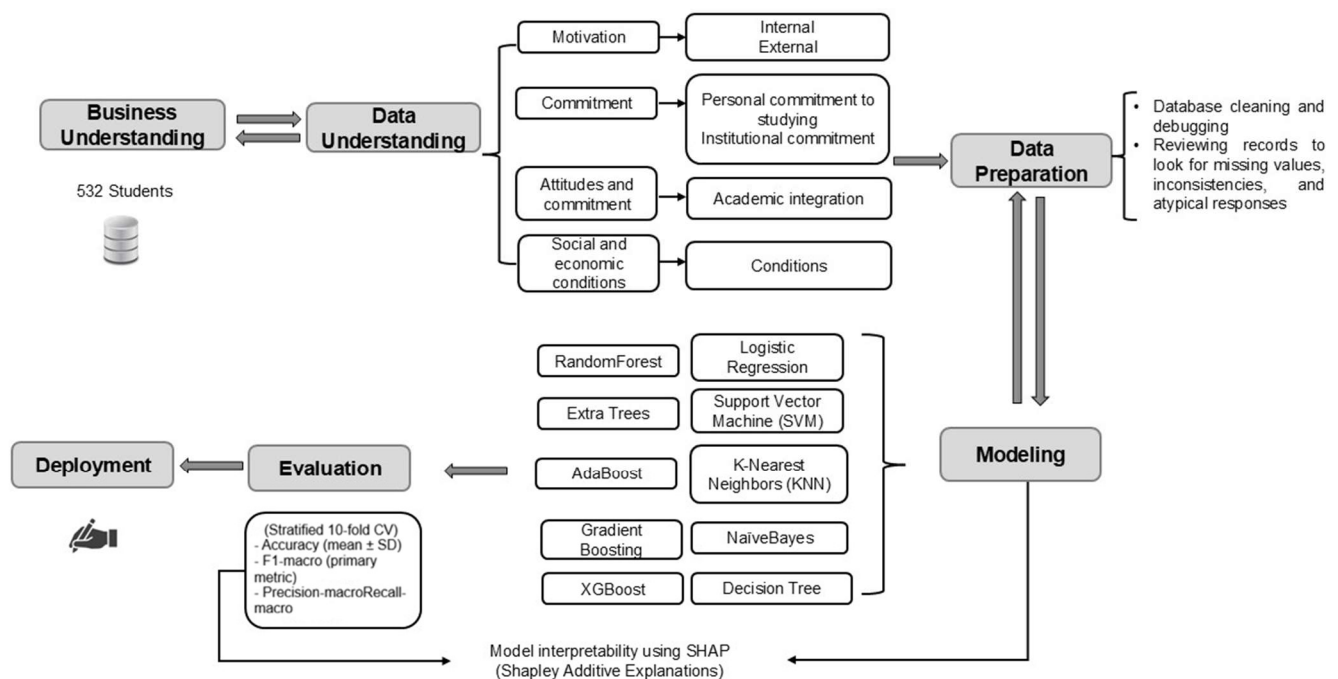


Figure 1. Methodological design of the study with the CRISP-DM approach.

2.1.1. Business Understanding

The target population consisted of 4603 undergraduate students enrolled during the 2025-II academic semester at a public university located in the Amazonas region of Peru. This institution provides higher education access to students from diverse socioeconomic and geographic backgrounds within the Peruvian Amazon. The population included students from the first to the twelfth academic semesters, encompassing both first-year entrants and students in intermediate and advanced stages of their programs.

From this population, a sample of 532 students was obtained through non-probabilistic convenience sampling based on voluntary participation. Students who agreed to participate

completed the survey after providing informed consent in accordance with the institutional ethical guidelines for research involving human subjects. Although the sampling strategy was non-probabilistic, the resulting dataset was considered adequate for the analytical procedures applied in this study, particularly the use of data mining and machine learning techniques, which require sufficiently large datasets to identify patterns and generate predictive models.

2.1.2. Data Understanding

To predict the factors that determine student retention, a review and systematization process was carried out on the relevant variables reported in the literature on academic persistence, retention, student success, and continued education in higher education. This phase allowed for the delimitation of the theoretical constructs established in Table 1 that explain retention and the organization of the analytical dimensions subsequently used in the modeling.

Table 1. Description of the factors associated with student retention.

Factors	Description	Authors
Motivation	It includes internal motivation (personal goals, expectations of success, self-concept) and external motivation (influence of the teacher in the classroom).	Bravo et al. [39] Martínez [59] Aisenson et al. [16] Jadue [60] Fonseca & García [61]
Commitment	It is divided into personal commitment to study (self-efficacy, academic performance, perception of difficulty) and commitment to the institution perceived (quality of the career, services, tutoring).	Velasquez et al. [15] Pascarella & Terenzini [62] Fonseca & García [61]
Attitude and commitment	Related to academic integration: sense of belonging, relationship with school authorities, relationship with peers.	Fishbein & Ajzen [63] Velasquez et al. [15] Pascarella & Terenzini [62] Fonseca & García [61]
Social and economic conditions	It includes social and family interaction (moral support, communication, respectful relationships) and economic conditions (financial resources, scholarships, transportation, etc.).	Bean [64] Gonzales [42] Velázquez & González [32] Fonseca & García [61]
Retention	Expected outcome of the process: timely approval of subjects, regular attendance, and uninterrupted continuity of university studies.	Rubén & García [14] Spady [65] Fonseca & García [61]

The survey instrument was designed based on theoretical constructs and indicators reported in the literature on student retention in higher education. The questionnaire items were adapted to the institutional context of higher education in the Amazonas region of Peru to ensure contextual relevance. To ensure content validity, the instrument was evaluated by three experts in higher education and educational research, who assessed the clarity, coherence, and relevance of the indicators. Based on their recommendations, minor adjustments were made to the wording and consistency of the items before applying the final version of the questionnaire.

Furthermore, to structure the collected information and translate the theoretical constructs into measurable variables, Table 2 defines a set of categories and indicators associated with each retention factor. These indicators enabled the operationalization of the previously identified conceptual dimensions and facilitated their incorporation into the statistical analysis and predictive modeling. The data collection instrument used in

this study (survey questionnaire) is provided in Supplementary Material (Instrument S1), allowing full transparency and reproducibility of the measurement process.

Table 2. Factors of the student retention model.

Factors	Categories	Indicators
Motivation	Internal	Personal Goals
		Expectations of success
		Self-concept
	External	By the teacher inside the classroom
Commitment	Personal commitment to studying	Self-efficacy
		Academic performance
		Perception of difficulty
	Commitment to the institution	Career quality
Attitude and commitment	Academic integration	Services
		Sense of belonging
		Relationship with university authorities
Socioeconomic conditions	Conditions	Peer relationships
		Social and family interaction
		Economic Conditions

Note. The indicators were selected and adapted based on theoretical and empirical references on student retention [33].

2.1.3. Data Preparation

In this stage, the database was cleaned and refined prior to analysis. This process involved reviewing the records to identify missing values, inconsistencies, and outliers, ensuring that only complete and valid cases were retained for the study. Additionally, the structure and format of the variables were verified, item names were standardized, and the dataset was prepared for subsequent statistical analyses. Regarding the dependent variable, the original Likert-scale responses (1–5) associated with student retention were recoded into three categorical levels (low, medium, and high) in order to facilitate their interpretation and use in the classification analysis.

Figure 2 summarizes the results of the Exploratory Factor Analysis (EFA) applied to the items corresponding to the five theoretical constructs of the student retention model. The suitability of the dataset for factor analysis was confirmed by a KMO value of 0.966, considered “excellent” according to multivariate psychometric criteria, and by Bartlett’s test of sphericity ($\chi^2 = 9997.6$; $p < 0.001$), which rejected the identity matrix hypothesis and confirmed the existence of a robust latent structure. It is important to note that the EFA was used solely to assess the internal structure and validity of the indicators associated with the theoretical factors. The validated indicators were subsequently used as input variables for the predictive modeling stage.

- Scree plot of eigenvalues. The horizontal axis represents the factor number extracted from the correlation matrix, while the vertical axis shows the eigenvalue associated with each factor. The dashed horizontal line indicates the Kaiser criterion (eigenvalue = 1) used to determine the number of retained factors. The sharp decline in the first components followed by an inflection point supports the retention of five latent factors.
- Heatmap of factor loadings for the retained factors. The horizontal axis represents the extracted latent factors (Factor 1–Factor 5), while the vertical axis lists the observed questionnaire items grouped by theoretical constructs (Motivation, Commitment,

Attitude and Commitment, Social and Economic Conditions, and Retention). The color scale represents the magnitude of the factor loadings, where darker tones indicate stronger associations between items and factors. Only items with the highest loadings (≥ 0.40 – 0.50) are displayed to highlight those with the greatest explanatory contribution and to improve visual interpretability.

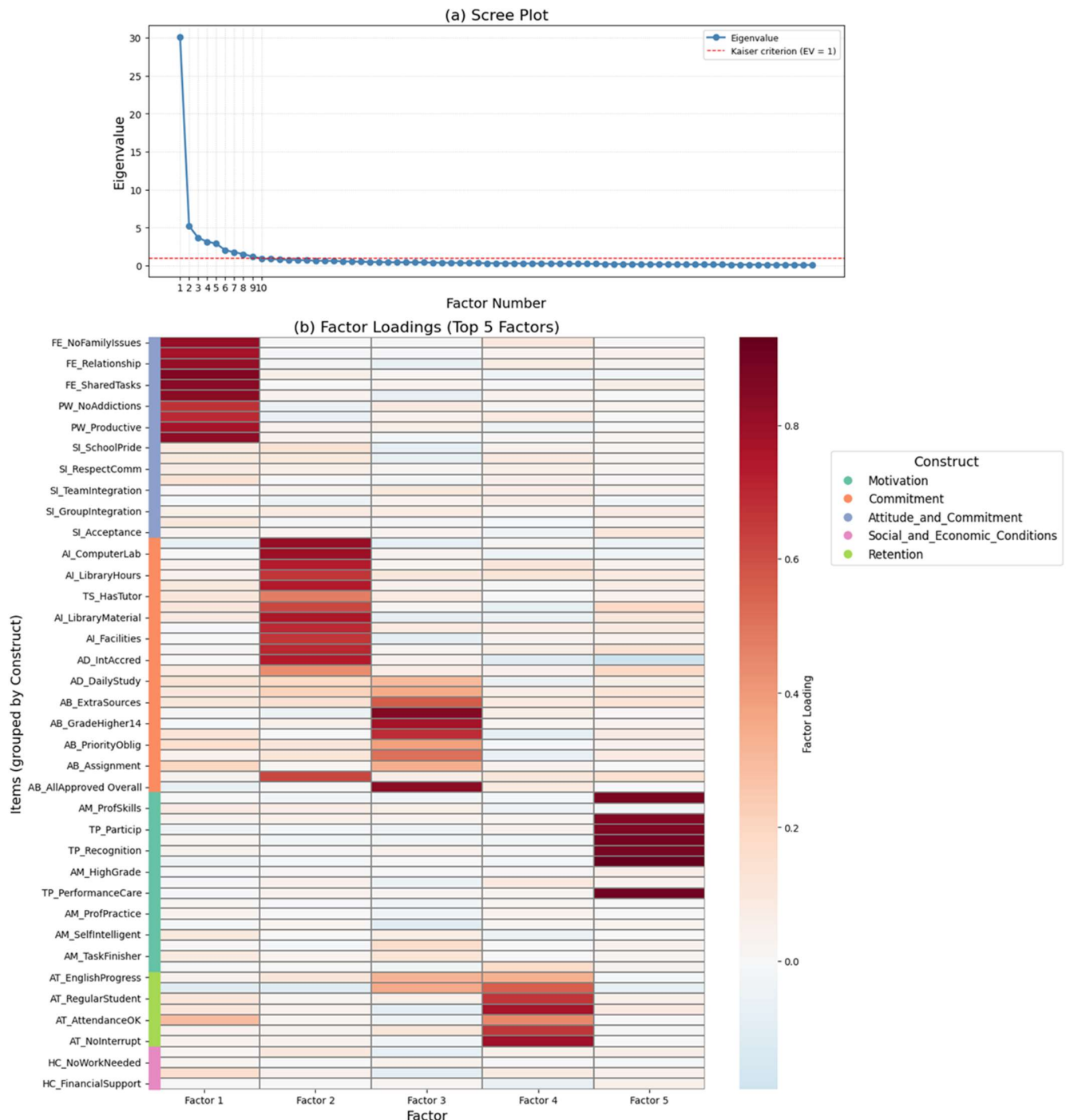


Figure 2. Exploratory factor analysis: Screen plot and factor loadings by dimension. Note. The adequacy of the data for factor analysis was confirmed by a KMO: 0.966 (≥ 0.6 adequate); Bartlett $\chi^2 = 9997.6$; and $p < 0.001$, indicating significant correlations among variables. Internal consistency reliability (Cronbach's alpha) by construct was as follows: Motivation (0.95), Commitment (0.95), Attitude and Commitment (0.97), Social and Economic Conditions (0.83), and Retention (0.87). (a) scree plot showing eigenvalues by factor number, including the Kaiser criterion (EV = 1); (b) heatmap of factor loadings for the top five factors, with items grouped by construct and color intensity representing loading magnitude.

2.1.4. Modeling

The collected data on student retention factors were processed and analyzed using multiple data mining algorithms, enabling their systematic coding and classification [66]. In order to identify the most relevant attributes of the dataset, various classification and prediction algorithms were applied, including: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, and XGBoost.

Logistic Regression: It is a supervised statistical model used to predict binary or multi-categorical categories using a logistic function that models the probability of belonging to a class [67]. This is one of the basic algorithms in data mining; it was used as a reference model to classify retention levels and compare its performance with more complex techniques.

Support Vector Machine (SVM): SVMs are a machine learning method that aims to accurately separate data into distinct classes. To achieve this, they construct a hyperplane that maximizes the distance between classes, allowing for greater accuracy and generalizability. This approach can be applied to both linear and more complex, nonlinear problems [68]. In this research, it was used to evaluate its ability to distinguish between levels of permanence with more precise decision boundaries.

K-Nearest Neighbors (KNN): It is an algorithm based on the similarity between observations that, using distance metrics such as Euclidean distance, predicts results by considering the k nearest neighbors; in regression, it averages their values, and in classification, it assigns the majority class [66]. This was applied here to assess how the proximity between academic and socioeconomic profiles influences the ranking of permanence.

Naïve Bayes: It is a probabilistic classifier based on Bayes' theorem and the assumption of conditional independence between predictors, allowing for efficient performance on high-dimensional data [69]. In this study, a simple probabilistic model was included to compare it against the other algorithms.

Decision Tree: These models progressively divide the dataset into smaller groups, following rules based on the attributes of each observation. This process facilitates the interpretation of the results and allows for the discovery of hierarchical relationships between variables, making them a useful and understandable tool for analyzing different types of problems [70]. Its value in educational research lies in its clarity and interpretability, allowing the identification of combinations of factors associated with each level of permanence.

Random Forest: It is a machine learning method that uses a set of decision trees, each built from random samples of data and attributes. By combining the results of these multiple trees, the model achieves greater stability and reduces the risk of overfitting, becoming a robust and reliable tool for tackling complex classification and prediction problems [71]. In this study, it was applied to obtain more stable predictions and to evaluate the relevance of retention factors.

Extra Trees: this method constructs multiple decision trees in which data splits are performed randomly at each node. This strategy increases diversity within the set of trees and, as a result, helps reduce the model's variance [66]. In this study, a randomized ensemble was compared against Random Forest, and its gains in accuracy were measured.

AdaBoost: Adaptive impulse response is an ensemble method that sequentially combines weak classifiers, such as decision trees, to improve model accuracy. In each iteration, it adjusts the weights of the observations, focusing on previous errors, thus enabling the construction of a robust and adaptive final model [66]. It was used to verify whether adaptive weighting increased the accuracy of permanence classification.

Gradient Boosting: It is a technique that builds models progressively, adding predictors that correct the errors made by previous ones. To achieve this, it uses the gradient descent

method, which allows for step-by-step improvements in the model's accuracy [72]. It allowed for modeling non-linear interactions between factors associated with permanence.

XGBoost: It is a supervised learning algorithm based on decision trees that uses regularization and iterative optimization techniques to avoid overfitting and improve computational efficiency. Its ability to capture complex and nonlinear relationships between variables makes it a tool for predicting student performance [66]. In this research, it demonstrated competitive performance and was selected for interpretability analysis using SHAP (Shapley Additive Explanations) due to its robustness and suitability for nonlinear relationships.

2.1.5. Evaluation

Model performance was evaluated using stratified 10-fold cross-validation to ensure robustness and stability across class distributions. In this procedure, the dataset was partitioned into ten folds preserving class proportions, and each fold was iteratively used as validation while the remaining nine served as training data. This resampling-based validation strategy provides a robust estimation of generalization performance and mitigates the risk of overfitting associated with single hold-out validation [73,74]. Accuracy and macro-averaged F1-score were computed for each fold, and results are reported as mean $\pm$ standard deviation across folds. The F1-macro metric was prioritized due to the multiclass and moderately imbalanced nature of the retention levels.

Additionally, SHAP was used, an interpretability technique based on cooperative game theory that assigns to each variable a marginal contribution value to the prediction [75]. This allows each model decision to be broken down into positive or negative contributions of attributes, ensuring consistent and comparable interpretations across features. Its importance lies in providing both a global view (average importance of each variable across the entire model) and a local view (explanation of individual predictions), thus improving the transparency of tree-based algorithms, as well as identifying factors that influence the classification of tenure levels [75].

3. Results

3.1. Descriptive Analysis

Figure 3 presents a comprehensive descriptive characterization of the sociodemographic and academic variables of the student population, allowing for the identification of patterns associated with retention rates. Each subfigure provides complementary evidence to characterize the student profile, offering preliminary insights into factors that may influence academic continuity.

Figure 3a shows the gender distribution, revealing a virtually balanced composition between men (50.2%) and women (49.8%). This symmetry indicates that the sample does not present significant gender bias, thus ensuring that subsequent analyses of retention are not influenced by male or female overrepresentation.

Figure 3b, which shows the distribution of retention levels, reveals that the largest proportion of students are at the high level (307), followed by the medium level (203), while the low level comprises only 22 students. This distribution suggests that, overall, the student population maintains mostly stable academic trajectories, although the presence of a minority group with low retention highlights the need for targeted interventions.

Figure 3c analyzes the relationship between scholarships and retention. Among scholarship recipients, high retention rates (69%) and low retention rates (6%) are predominant. In contrast, students without scholarships show greater heterogeneity in their retention distribution, with 56% high, 40% medium, and 3% low. This pattern suggests that financial support functions as a protective factor, promoting academic continuity and mitigating the vulnerability associated with financial limitations.

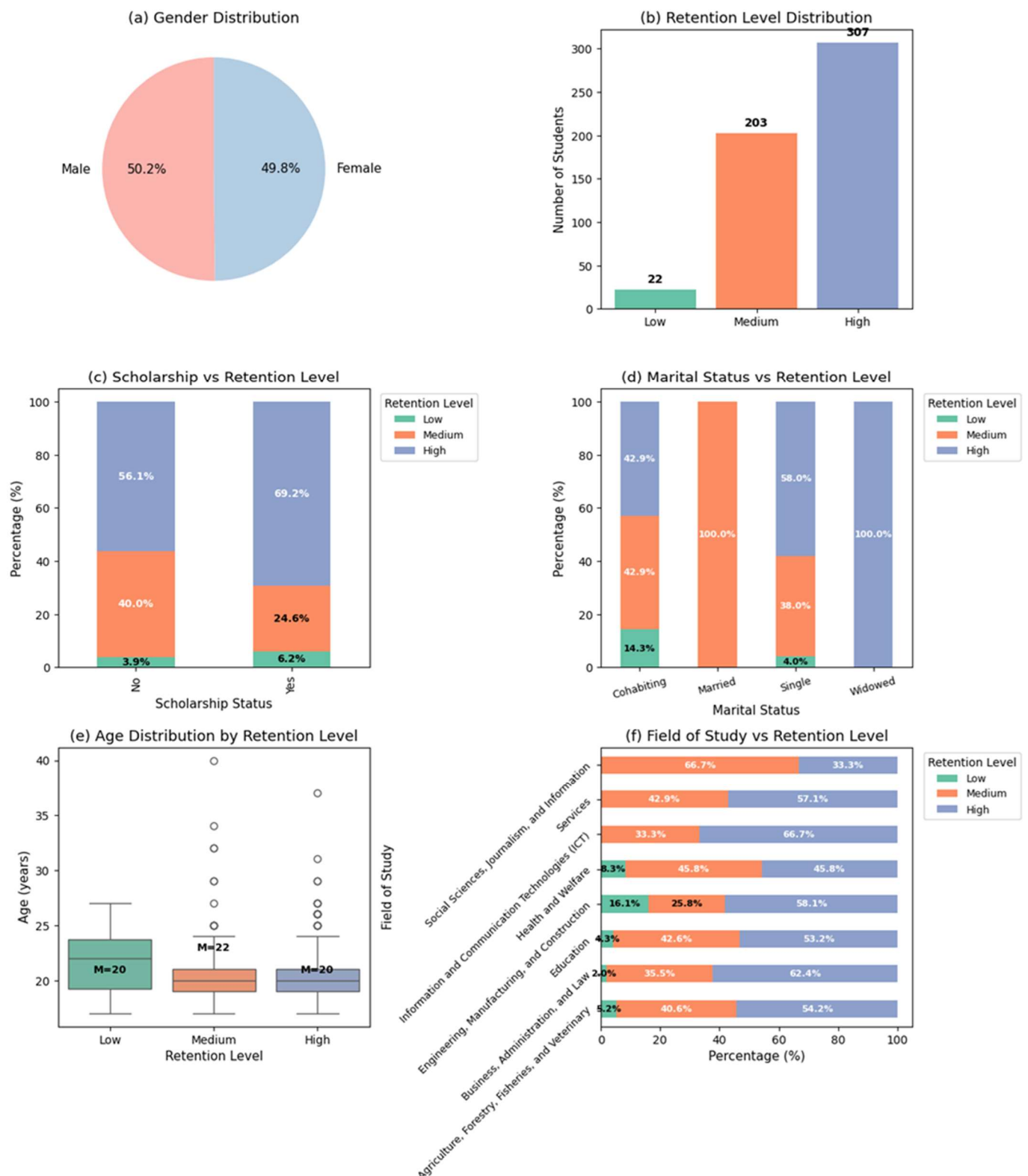


Figure 3. Descriptive distribution of sociodemographic and academic variables associated with the level of student retention. Note. (a) gender distribution (% male/female); (b) retention level distribution (number of students); (c) retention level by scholarship status (%); (d) retention level by marital status (%); (e) age distribution by retention level (boxplot); (f) retention level by field of study (%).

Figure 3d examines marital status in relation to retention levels. Single students have the highest proportion of high retention (57%), while those living with a partner show a balanced distribution between high (42%) and medium (42%) retention levels. For married

students, retention is exclusively at the medium level, while for widowed students, it is entirely at the high level. These results indicate that family responsibilities, household stability, and financial burdens may influence academic trajectories.

In Figure 3e, the age distribution by retention level reveals similar medians for the low and high levels (20 years), while the average retention level shows a higher median (22 years) and greater dispersion. This variability suggests that older students may face greater challenges, such as work or family obligations, that affect the stability of their continued academic progress.

Finally, Figure 3f shows the relationship between field of study and retention. Substantial differences are observed between academic areas: disciplines such as Social Sciences, Journalism and Information (66%), and Business Administration and Law (54%) have higher retention rates, while Education and Engineering show high proportions at the medium and low levels. This suggests that factors related to curriculum design, academic demands, and institutional support may influence the observed retention patterns.

Figure 4 analyzes how psychosocial variables such as motivation, commitment, attitude, and socioeconomic conditions are related to student retention levels. Motivation levels show clear differences between the groups. Students with high motivation exhibit a marked tendency toward academic continuity, with 69% at high retention and only 1% at low retention. This contrasts with those with medium motivation, among whom medium retention predominates (55.6%), suggesting more unstable performance. In the case of low motivation, the proportion of low retention increases significantly (38%), and only 19% manage to remain at a high level, indicating that a lack of motivation becomes a considerable risk factor for retention.

Academic engagement reflects an even more pronounced pattern. High engagement is almost exclusively associated with stable academic trajectories: 74% high retention and no cases of low retention. When engagement is moderate, the distribution is balanced between high (41.9%) and moderate (54%) retention, suggesting sustained, though not yet fully consolidated, effort. However, low engagement shows a critical trend: 66.7% of students fall into the low retention category, and only 11% achieve high retention, reaffirming its role as one of the strongest predictors of dropout.

The combined variable of attitude and commitment reinforces these trends. Students with high levels achieve a 65.8% high retention rate, reflecting a positive disposition toward learning and academic responsibility. Among those with a medium level, retention is concentrated at the medium level (68%), indicating acceptable performance but susceptible to fluctuations. The low level of this variable is directly associated with greater academic risk: 82% are at low retention, making it the critical indicator among the factors analyzed.

Regarding socioeconomic conditions, while their effect is less pronounced than that of motivational and attitudinal variables, relevant differences are observed. Students from favorable backgrounds show an 88% high retention rate, suggesting that economic stability facilitates continued education. In middle-income brackets, retention is mostly at the middle level (45.8%). In low-income brackets, retention is divided between the middle (46%) and low (27%) levels, indicating that economic limitations can hinder continued enrollment.

3.2. Decision Tree

Figure 5 illustrates the complete structure of the trained decision tree. The root node corresponds to the variable *HC_NoWorkNeeded*, indicating that work-related constraints represent the primary splitting criterion in differentiating retention levels. This position at the top of the hierarchy highlights the central role of external responsibilities in shaping academic continuity.

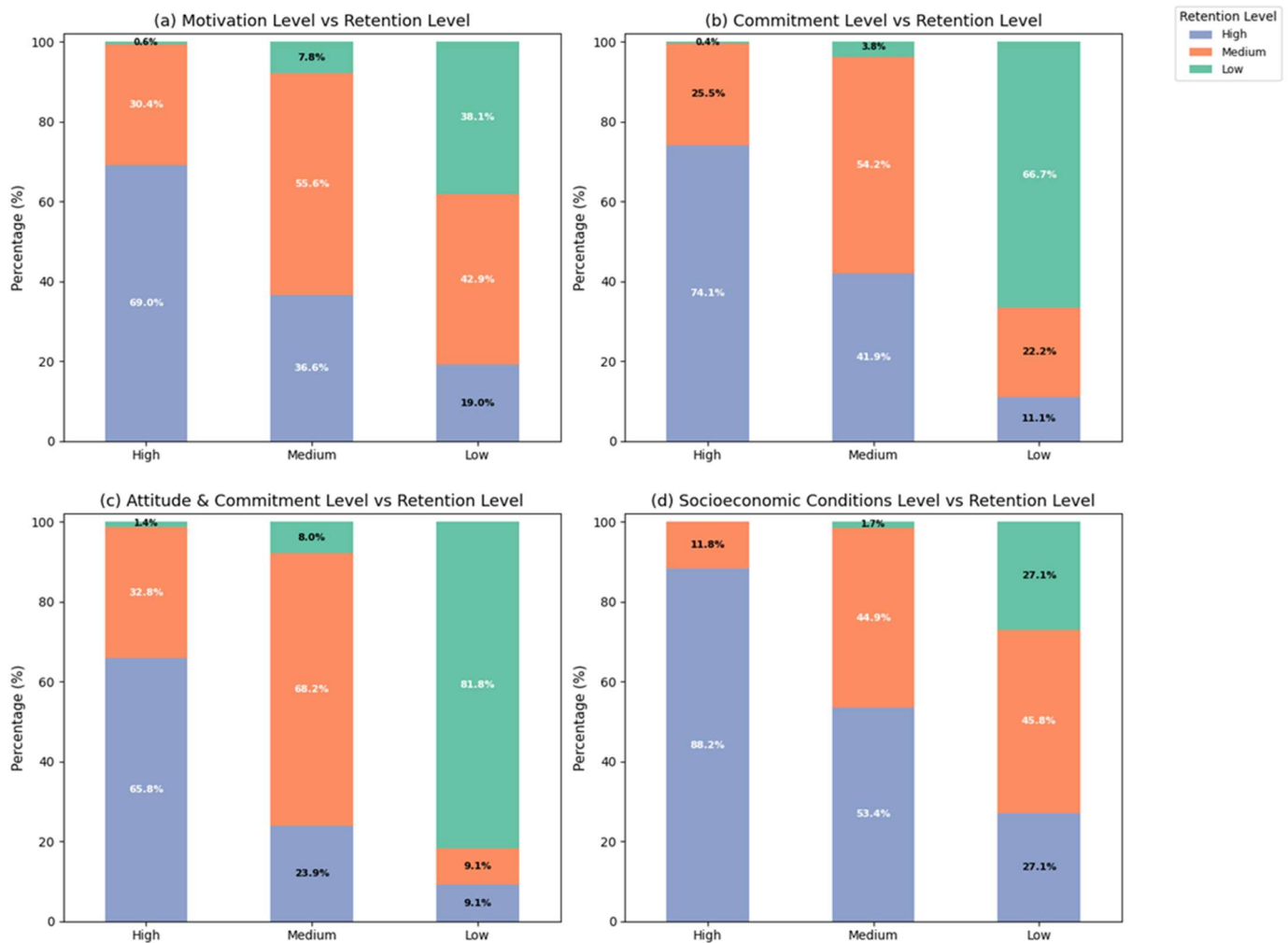


Figure 4. Association between levels of motivation, commitment, attitude-commitment, and socioeconomic conditions with the level of student retention. Note. (a) motivation level vs. retention level (%); (b) commitment level vs. retention level (%); (c) attitude and commitment level vs. retention level (%); (d) socioeconomic conditions level vs. retention level (%).

In the left branch, where students report fewer work constraints, the model further incorporates academic and attitudinal variables such as task completion and course approval. These splits refine the classification toward medium and high retention levels, suggesting that academic engagement acts as a reinforcing factor when external burdens are limited.

Conversely, the right branch groups students experiencing greater work or socioeconomic constraints. In this pathway, the probability of classification as low retention increases, especially when combined with lower academic performance indicators. These terminal nodes are predominantly associated with class 0 (low retention), reflecting the model's greater sensitivity in identifying academically vulnerable students.

The tree also reveals the interaction between academic performance and contextual variables. For instance, favorable academic outcomes partially mitigate the negative effect of external constraints, whereas the coexistence of financial limitations and weaker academic indicators tends to direct the classification toward lower retention levels. This hierarchical structure reinforces the multidimensional nature of retention, where external, academic, and motivational components interact dynamically.

Although the Decision Tree model provides valuable interpretability, its cross-validated macro performance remains below that of ensemble-based methods. This dif-

ference is expected, as single-tree structures are more sensitive to sample partitioning and may not fully capture complex nonlinear interactions within the dataset.

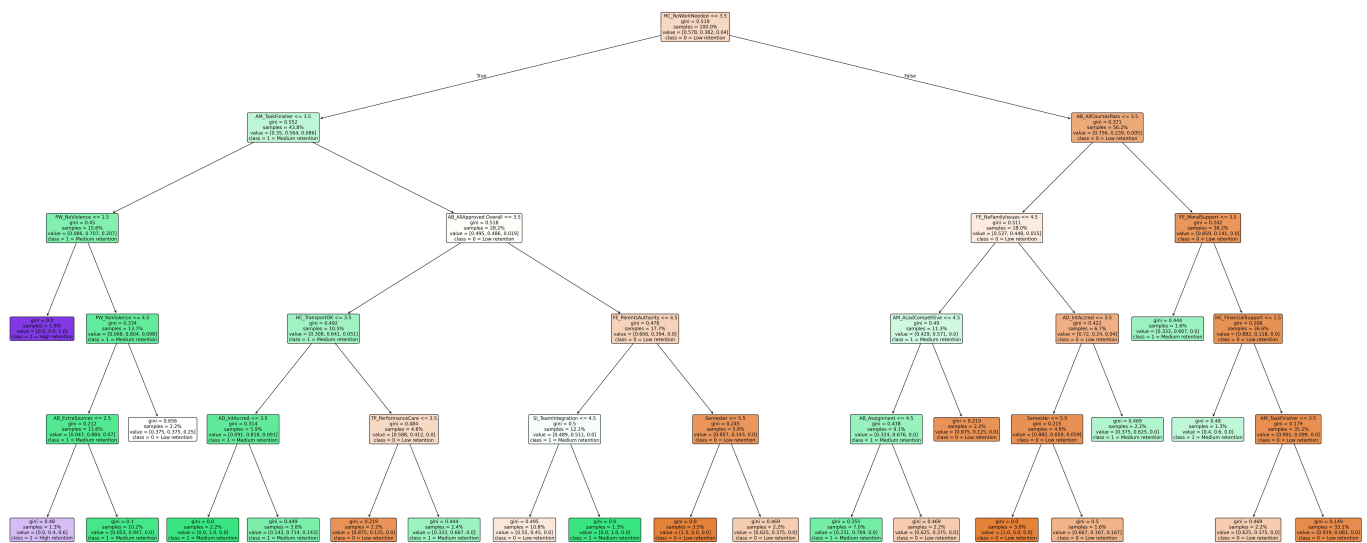


Figure 5. Decision tree structure for student retention classification. The diagram represents the hierarchical partitioning learned by the model. Each node displays the splitting criterion, Gini impurity, number of samples, and class distribution. Model performance metrics correspond to stratified 10-fold cross-validation results. Colors indicate the predominant class at each node, and color intensity reflects node purity.

3.3. Classification of Factors

The results in Table 3 present the performance of the different classification algorithms evaluated under a stratified 10-fold cross-validation framework. All metrics are reported as mean $\pm$ standard deviation across folds to ensure robustness and to provide a reliable assessment of generalization performance. In addition to overall accuracy, macro-averaged precision, recall, and F1-score were computed to account for the multiclass and moderately imbalanced nature of the retention levels.

Table 3. Stratified 10-Fold Cross-Validation Performance of Classification Models for Student Retention Prediction.

Model	Accuracy	Precision-Macro	Recall-Macro	F1-Macro
Random Forest	0.729 ± 0.058	0.678 ± 0.170	0.623 ± 0.134	0.636 ± 0.136
Extra Trees	0.720 ± 0.061	0.681 ± 0.180	0.601 ± 0.115	0.625 ± 0.134
XGBoost	0.714 ± 0.047	0.639 ± 0.161	0.593 ± 0.121	0.606 ± 0.132
KNN	0.692 ± 0.070	0.666 ± 0.174	0.586 ± 0.132	0.605 ± 0.140
Logistic Regression	0.679 ± 0.062	0.620 ± 0.139	0.617 ± 0.147	0.605 ± 0.132
SVM	0.703 ± 0.063	0.657 ± 0.172	0.586 ± 0.141	0.603 ± 0.147
Gradient Boosting	0.705 ± 0.064	0.610 ± 0.158	0.595 ± 0.142	0.592 ± 0.139
Decision Tree	0.692 ± 0.048	0.603 ± 0.168	0.557 ± 0.115	0.568 ± 0.130
Naïve Bayes	0.635 ± 0.063	0.518 ± 0.096	0.594 ± 0.139	0.516 ± 0.096
AdaBoost	0.699 ± 0.037	0.457 ± 0.029	0.463 ± 0.023	0.454 ± 0.023

Among the evaluated models, tree-based ensemble methods demonstrated the most stable and competitive performance. Random Forest achieved the highest macro-averaged F1 score (0.636 ± 0.136), accompanied by an average accuracy of 0.729 ± 0.058 , precision-macro of 0.678 ± 0.170 , and recall-macro of 0.623 ± 0.134 . These results indicate a balanced capacity to correctly classify the three retention levels while maintaining moderate stability

across folds. Although the variability observed in the standard deviations reflects sensitivity to data partitioning, the model consistently ranked among the top performers across all evaluation metrics.

Extra Trees followed closely, obtaining an F1-macro of 0.625 ± 0.134 and an accuracy of 0.720 ± 0.061 , with precision-macro and recall-macro values of 0.681 ± 0.180 and 0.601 ± 0.115 , respectively. Its performance is comparable to that of Random Forest, suggesting that ensemble approaches based on randomized decision trees are well suited to capture nonlinear relationships within the retention dataset. XGBoost also exhibited competitive results (F1-macro = 0.606 ± 0.132 ; accuracy = 0.714 ± 0.047 ; precision-macro = 0.639 ± 0.161 ; recall-macro = 0.593 ± 0.121), confirming that gradient-based boosting techniques remain effective, although without a clear statistical separation from other tree-based models.

Models such as SVM, Gradient Boosting, Logistic Regression, and KNN displayed intermediate performance levels, with F1-macro values ranging from 0.592 to 0.605. While their accuracy values remain relatively stable, the dispersion observed in precision-macro and recall-macro suggests moderate sensitivity to class imbalance, particularly when predicting the minority retention level.

The Decision Tree model showed lower overall macro performance (F1-macro = 0.568 ± 0.130), reflecting its higher variance compared to ensemble-based approaches. Naïve Bayes and AdaBoost obtained the lowest macro-averaged scores (F1-macro = 0.516 ± 0.096 and 0.454 ± 0.023 , respectively), indicating reduced sensitivity to class-level differences and limited ability to model complex nonlinear interactions present in the data.

Overall, the results indicate that tree-based ensemble models provide the most consistent predictive performance for multiclass student retention prediction. However, differences between the top-performing models remain moderate when considering cross-validation variability across accuracy, precision, recall, and F1-macro metrics.

Figure 6 presents the average permutation-based feature importance computed within a stratified 10-fold cross-validation framework for each of the evaluated algorithms. By estimating feature relevance across folds and averaging the results, the figure provides a more robust assessment of variable influence while reducing the effect of partition-specific variability. This approach allows the identification of factors that consistently contribute to the classification of student retention levels across models with different structural assumptions.

In linear algorithms such as Logistic Regression and SVM, variables associated with academic self-regulation, commitment, and perceived support tend to appear among the most relevant predictors. These results suggest that attitudinal and motivational dimensions exhibit a proportional and systematic relationship with retention levels, which can be captured even by models with linear decision boundaries.

In the KNN model, factors related to timely completion of academic tasks, personal productivity, respectful academic climate, and availability of financial support show relatively higher importance values. Similarly, in the Naïve Bayes model, variables associated with successful completion of courses, absence of work obligations during studies, and identification with the academic program appear among the most influential predictors. These results indicate that, even under simpler modeling assumptions, these algorithms capture meaningful differences in retention patterns associated with motivational, academic, and contextual dimensions of student permanence.

Tree-based models, including Decision Tree, Random Forest, Extra Trees, Gradient Boosting, and XGBoost, display a clearer hierarchical structure of feature relevance. Across these ensemble approaches, variables related to socioeconomic conditions, academic performance, and work or financial constraints frequently appear among the top predictors. The consistency of these factors across multiple folds and across different ensemble strategies reinforces their discriminative power in differentiating retention levels.

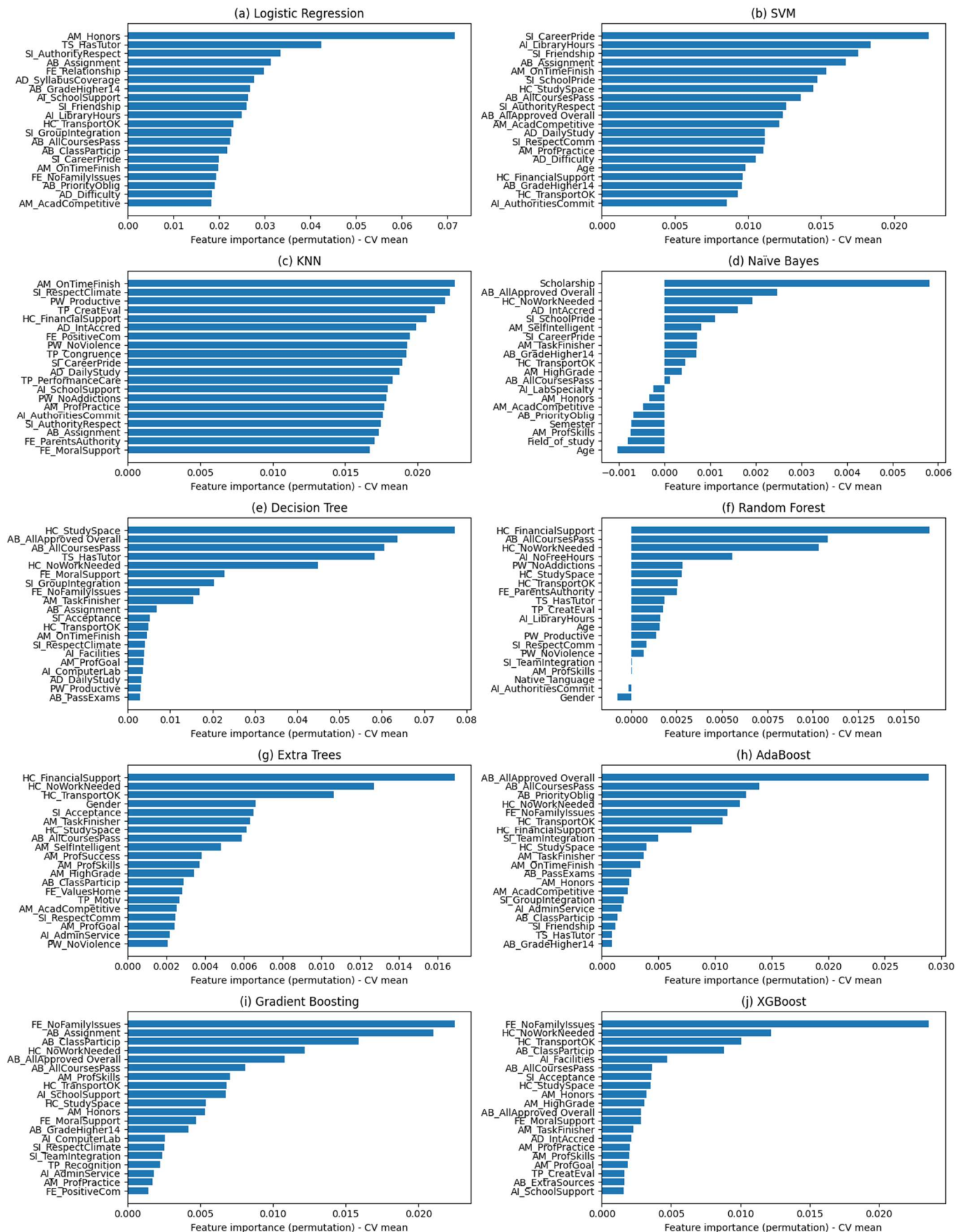


Figure 6. Mean permutation feature importance across stratified 10-fold cross-validation for each classification algorithm. (a) Logistic Regression; (b) Support Vector Machine (SVM); (c) K-Nearest Neighbors (KNN); (d) Naïve Bayes; (e) Decision Tree; (f) Random Forest; (g) Extra Trees; (h) AdaBoost; (i) Gradient Boosting; (j) XGBoost.

(i) Gradient Boosting; (j) Extreme Gradient Boosting (XGBoost). Note. Feature relevance was estimated using permutation importance computed on the validation fold within each split of the stratified 10-fold cross-validation procedure (scoring = F1-macro; n_repeats = 10). The values displayed correspond to the mean importance across folds for the top 20 predictors per model.

Although Random Forest obtained the numerically highest mean F1-macro score under cross-validation, performance differences between top-performing ensemble models remain moderate within cross-validation variability ranges. This convergence suggests that student retention is influenced by a set of interacting academic, socioeconomic, and psychosocial factors rather than by isolated predictors. In this sense, retention emerges as a multidimensional phenomenon shaped by the combined effect of performance indicators, contextual constraints, and individual motivational attributes.

3.4. SHAP Analysis

Figures 7 and 8 show the results of the SHAP analysis applied to the XGBoost model, with the purpose of interpreting the influence and behavior of the predictor variables in the classification of student retention levels. The SHAP approach allows us to explain the individual contribution of each characteristic to the prediction, as well as to identify interaction patterns and nonlinear effects among the factors considered. In this context, the impact of each variable is measured using the mean absolute SHAP value (mean $|\text{SHAP value}|$), which represents the average magnitude of the contribution of each predictor to the model output. Higher values therefore indicate variables with a stronger overall influence on the prediction of student retention levels.

Figure 8, which shows the average impact, reveals that the variables with the greatest overall contribution to the model are those related to household conditions, such as the availability of a suitable study space and access to reliable transportation, and to academic performance, represented by indicators such as passing all courses, passing exams, and maintaining an average grade above 14. These factors emerge as the main determinants driving predictions toward higher retention rates. Other elements, such as field of study, age, availability of a tutor, and financial support, have an intermediate influence, while aspects related to family well-being and personal habits show a smaller impact, although still relevant in certain cases.

Figure 8, which shows the dispersion of SHAP values, illustrates how high or low values for each variable affect the direction of the prediction. High values for academic performance, general motivation, task completion, and institutional support are associated with increased likelihood of belonging to medium or high retention levels. Conversely, a lack of material resources or family difficulties shifts the prediction toward lower levels. This visualization also reveals the presence of nonlinear relationships and internal variability among students, showing that not all characteristics have a uniform influence: in some cases, an intermediate value can have different effects depending on its combination with other variables.

Table 4 presents the contribution of the variables to the XGBoost model using SHAP values broken down by retention class and their overall impact. The variable HC_TransportOK shows the greatest average impact with an overall SHAP value of 0.2494, also being the most influential variable in class 0 (0.367) and maintaining significant contributions in classes 1 (0.202) and 2 (0.179). This behavior indicates that adequate transportation conditions constitute a cross-cutting determinant of retention. The variable AM_TaskFinisher, associated with the student's ability to complete tasks, has an overall impact of 0.170, with significant contributions in both class 0 (0.258) and class 1 (0.150), decreasing in class 2 (0.103).

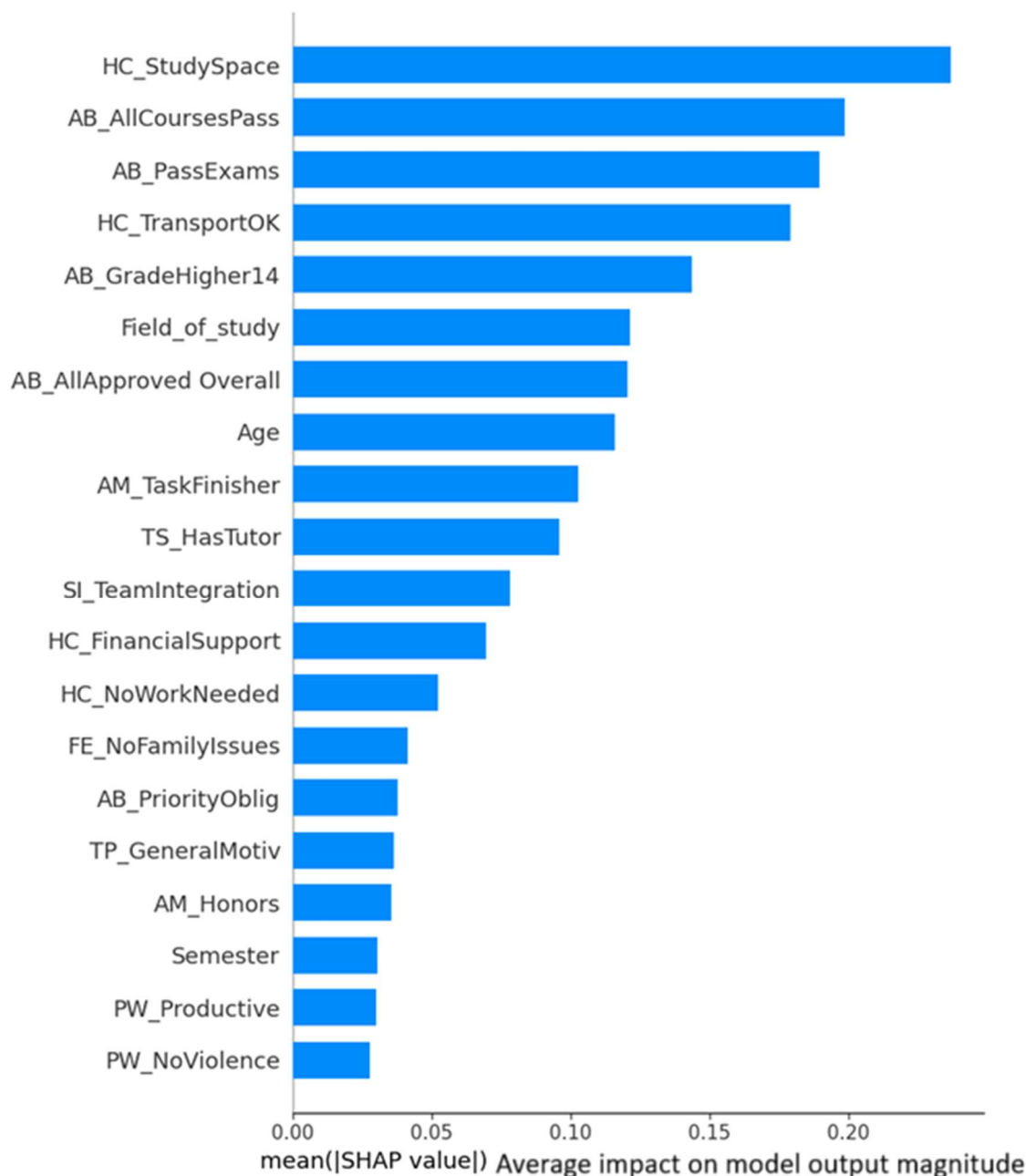


Figure 7. Importance of SHAP-based characteristics for the final student retention model.

Similarly, HC_NoWorkNeeded has an overall impact of 0.168, with values of 0.234 for class 0 and 0.217 for class 1, while its influence decreases in class 2 (0.052). The need to work thus appears to be a significant barrier for students at risk or at an intermediate level of retention. Within the group of academic variables, AB_AllCoursesPass has an overall impact of 0.1640, with a balanced contribution in classes 0 (0.147) and 1 (0.146), and a higher contribution in class 2 (0.199), indicating that continuous course completion is a key factor for achieving high retention. A similar pattern emerges with AB_AllApprovedOverall (overall impact 0.145), which exerts the greatest influence in class 0 (0.225), the least in class 1 (0.091), and a moderate influence in class 2 (0.120). The average above 14 (AB_GradeHigher14) also shows a significant overall impact (0.140), with contributions distributed among the three classes: 0.167 in class 0, 0.111 in class 1, and 0.144 in class 2, indicating that academic performance contributes to differentiating retention patterns across the three retention levels.

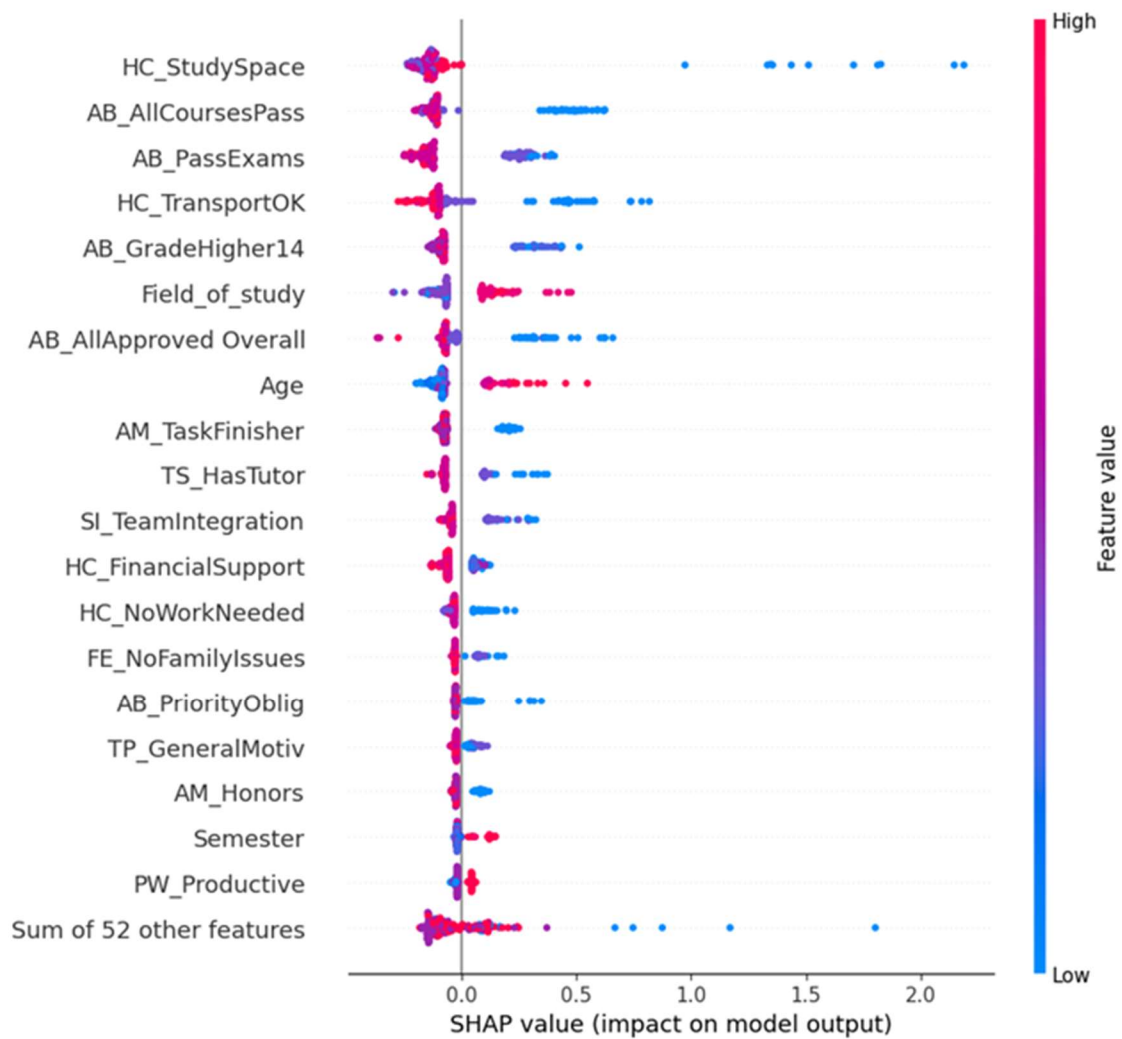


Figure 8. Impact and variation in characteristics on the model according to SHAP value distribution.

Table 4. Top variables by global mean ($|SHAP|$).

Items	Class_0	Class_1	Class_2	Global Shape
HC_TransportOK	0.367	0.202	0.179	0.249
AM_TaskFinisher	0.258	0.150	0.103	0.170
HC_NoWorkNeeded	0.234	0.217	0.052	0.168
AB_AllCoursesPass	0.147	0.146	0.199	0.164
AB_AllApproved Overall	0.225	0.091	0.120	0.145
AB_GradeHigher14	0.167	0.111	0.144	0.140
FE_NoFamilyIssues	0.110	0.247	0.041	0.133
Semester	0.172	0.190	0.030	0.131
AM_OnTimeFinish	0.194	0.150	0.007	0.117
HC_StudySpace	0.051	0.041	0.237	0.110
HC_FinancialSupport	0.157	0.084	0.069	0.103
Field of study	0.119	0.043	0.121	0.095
Age	0.037	0.129	0.116	0.094
AB_PassExams	0.021	0.032	0.190	0.081
SI_TeamIntegration	0.100	0.061	0.078	0.080
TS_HasTutor	0.036	0.098	0.096	0.077
AB_PriorityOblig	0.084	0.068	0.038	0.063
AD_IntAccred	0.081	0.082	0.009	0.057
PW_NoViolence	0.044	0.100	0.027	0.057
AM_AcadCompetitive	0.108	0.042	0.000	0.050

Class 1 identifies variables with greater relative weight compared to the other classes. This is the case with FE_NoFamilyIssues, which achieves a contribution of 0.247 in Class 1, one of the highest within this category, and an overall impact of 0.133, indicating that family stability is crucial for maintaining average levels of retention. The Semester also stands out, with an overall impact of 0.131, but with greater weight in Class 1 (0.172), indicating that the student's position in their academic trajectory modulates the probability of remaining at an intermediate level.

In class 2, HC_StudySpace stands out, with the highest value within this category (0.237) and an overall impact of 0.110, revealing that having an adequate study space is a fundamental component of high retention. A similar pattern is observed with AB_PassExams, whose contribution in class 2 is 0.190, despite showing reduced values in classes 0 and 1 (0.021 and 0.032, respectively). This confirms that the consistent passing of assessments specifically distinguishes students with greater academic continuity.

Other variables, such as TS_HasTutor, SI_TeamIntegration, PW_NoViolence, and Age, show moderate global impacts between 0.0938 and 0.0765 distributed heterogeneously among the three classes, reflecting their contextual and complementary influence within the model.

3.5. Sensitivity Analysis

Figure 9 shows the sensitivity analysis, which illustrates the relative contribution of the four groups of factors evaluated to student retention. The results indicate that the commitment dimension has the greatest influence, representing 41.6% of the total impact on the model. This finding confirms that the components associated with the student's level of involvement with their academic goals and continued education are those that generate the most significant variations in the probability of retention.

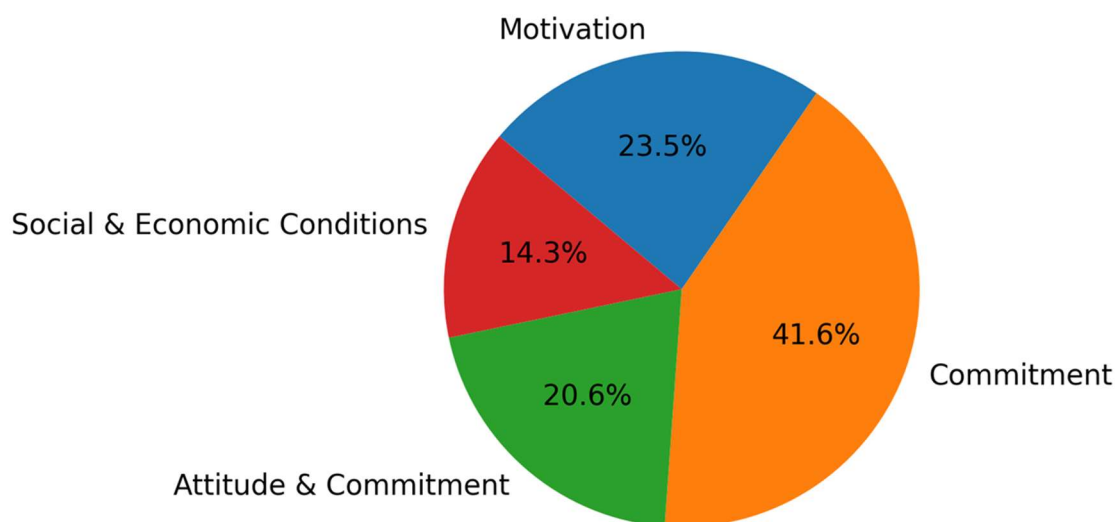


Figure 9. Sensitivity analysis of the factors and their impact on the dimensions in predicting the level of permanence.

Motivation contributes 23.5%, demonstrating that personal disposition, interest in learning, and self-perceived efficacy also play a relevant role in predicting the model, although with less weight than commitment. Additionally, the attitude and commitment dimension contributes 20.6%, showing that the way students approach their responsibilities and study habits acts as an important modulator, reinforcing or weakening the impact of the primary commitment factor.

Finally, social and economic conditions contribute 14.4%, a smaller contribution compared to motivational and attitudinal factors. This suggests that, while these conditions influence academic stability, their predictive effect is less decisive within the model, operating more as a supporting element than as a direct driver of retention.

4. Discussion

The results allow us to understand student retention as a phenomenon conditioned by academic, motivational, attitudinal, familial, and socioeconomic factors, consistent with the literature reviewed. The predominance of students with high and medium retention rates aligns with research indicating that continuity in educational trajectory depends on both intrinsic motivation and the sustained commitment students develop toward their academic obligations [17,18]. This pattern observed in the sample confirms that the student's internal dimensions, such as personal interest, discipline, self-confidence, and clarity of goals, constitute structuring elements of retention, as evidenced by Aisenso et al. [16] and Fonseca & García [61]. The results show that student retention does not depend solely on favorable external conditions, but also on the student's ability to consistently sustain their academic goals.

The predominance of high and medium levels of retention, along with a balanced distribution by gender, coincides with studies that highlight the integrated role of individual and contextual factors in educational continuity [16,32]. The higher retention rate among scholarship students confirms that financial support acts as a protective factor, in line with evidence that warns that economic insecurity forces many students to reduce study hours or work extensively, affecting their academic progress [76]. Therefore, institutional policies focused on financial support and integrated services are essential to promote the persistence and completion of studies [77]. Likewise, the lower stability observed in older students or those with family responsibilities suggests the influence of external burdens on retention.

The patterns observed in motivation, commitment, and attitude show that psychosocial factors play a critical role in retention. Students with high motivation and greater commitment consistently exhibit better retention rates, which aligns with research highlighting the importance of engagement and self-confidence for sustaining academic progress [41,60]. Furthermore, the literature underscores that mentoring and structured academic guidance strengthen institutional integration, improve performance, and promote retention, especially among underrepresented or unevenly prepared students [78,79]. These programs, both peer-to-peer and teacher-led, have proven effective in supporting academic success and improving continuity in resource-limited contexts [80,81]. Overall, the results suggest that strengthening student motivation, academic support, and mentoring constitutes a key pathway for consolidating student retention. These findings are consistent with Tinto's integration model, as they demonstrate that students' active engagement with their academic responsibilities and with the institutional environment remains a central component of persistence.

The results derived from data mining complement these findings. Tree-based ensemble models, particularly Random Forest, Extra Trees, and XGBoost, exhibited comparatively stronger and more stable predictive capacity under cross-validation, consistent with studies highlighting the potential of ensemble methods to identify complex retention patterns [1,19]. However, the low representation of students in the low-retention category created difficulties for their correct classification, a behavior that has been observed in research where class imbalance affects predictive performance [3,53]. This situation highlights the importance of continuing to refine models and applying balancing strategies to improve sensitivity in populations with rare characteristics.

The SHAP analysis allowed us to visualize the relevance of each variable in predicting retention rates. Academic variables such as educational services, passing all courses, and maintaining a good GPA emerged as highly influential factors, which aligned with research linking sustained performance with continued enrollment [42]. Likewise, family variables such as the absence of problems at home and economic variables such as the availability of scholarships or resources to study showed a significant impact, coinciding with evidence that highlights how family tensions and economic limitations can become risk factors for dropping out [32]. The simultaneous relevance of academic, family, and material variables confirms that student retention is shaped at the intersection of academic performance, social support, and the concrete conditions that enable students to study.

Sensitivity analysis shows that academic engagement is the strongest predictor, contributing 41.6%, thus reinforcing its central role in continued education. The research included in the file underscores that academic engagement, manifested in responsibility for assignments, consistency, and study planning, is directly associated with retention and academic achievement [11,33]. Additionally, motivation is positioned as a decisive factor, coinciding with studies that identify motivation as the axis of persistence and as a fundamental psychological resource for coping with difficulties and sustaining academic effort [8,56].

Attitudinal factors also showed a significant relationship with retention rates. Positive attitudes toward studying and favorable expectations of academic success were associated with medium and high levels of continuation, which aligns with studies linking emotional readiness and personal beliefs to the ability to cope with educational challenges [21]. This coincidence reinforces the notion that permanence is not solely a result of cognitive skills, but also of emotional stability, optimism, and perceived self-efficacy. Attitudes operate as an interpretive filter through which students confront academic demands and assess their possibilities for continuing their studies.

Additionally, the identified limitations, such as class imbalance, the cross-sectional nature of the study, and the single-institution origin of the sample, highlight the need to extend the analysis through longitudinal designs, comparative samples, and the inclusion of additional institutional variables. In particular, the relatively small number of students classified in the low-retention category may have influenced the sensitivity of the classification models, making it more difficult to accurately identify the highest-risk profiles. Likewise, the cross-sectional nature of the data limits the ability to examine changes in academic trajectories over time, which would be essential for understanding the dynamic processes underlying student persistence and dropout. Similarly, although machine learning models allow the identification of complex patterns and the estimation of retention probabilities, their predictive nature does not permit the establishment of definitive causal relationships among the analyzed variables. Despite these limitations, the findings provide valuable empirical evidence for the development of context-sensitive predictive models and for the design of institutional retention strategies based on real student risk profiles.

5. Conclusions

The study demonstrates that student retention is a multidimensional phenomenon influenced by academic, motivational, attitudinal, socioeconomic, and familial factors. Academic commitment was identified as the most decisive variable, followed by motivation and attitude toward studying, highlighting that internal student factors are fundamental to ensuring educational continuity. Furthermore, the majority of students exhibit high or medium retention rates, confirming that academic commitment, intrinsic motivation, integration into the educational environment, and emotional stability are central factors for sustaining a continuous educational trajectory. Additionally, the differences observed

according to economic conditions, age, and family responsibilities show that personal context shapes the likelihood of continued enrollment.

The classification results indicate that data mining constitutes an effective analytical tool for identifying risk profiles and retention patterns. Tree-based ensemble models showed the most stable predictive performance under cross-validation. The SHAP interpretation confirmed that passing courses, maintaining adequate performance levels, having a stable family environment, and having sufficient financial resources are factors that significantly strengthen retention rates. The sensitivity analysis reinforced that academic commitment is the factor with the greatest impact on prediction, followed by motivation and attitudes toward learning.

The study provides empirical evidence that reinforces the integrative view of student retention models, demonstrating that the interaction between individual, institutional, and socioeconomic factors offers more comprehensive explanations than one-dimensional approaches. The incorporation of data mining techniques and interpretability through SHAP enriches the traditional analysis, allowing us to understand not only which variables predict retention, but also how they do so. On a practical level, the results provide an objective basis for designing early warning systems, prioritizing interventions, and targeting institutional resources for student support.

Based on these findings, it is recommended that institutional programs be implemented to strengthen academic engagement, motivation, and positive attitudes toward learning through personalized tutoring, psychological counseling, peer mentoring, and emotional regulation strategies. Likewise, it is a priority to consolidate financial aid policies, improve student welfare services, expand suitable study spaces, and establish data-driven mechanisms for continuous monitoring. The use of predictive models should be integrated into institutional monitoring processes to identify at-risk students early and provide timely and differentiated interventions.

Future research should address the limitations of this study. In particular, longitudinal designs are needed to analyze the evolution of student retention over time, and broader samples from different institutions and sociocultural contexts should be considered to improve the generalizability of the results. Furthermore, future studies could incorporate additional variables related to institutional climate, mental health, learning strategies, and teaching quality, as well as empirically evaluate the effectiveness of interventions derived from predictive models.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/informatics13040050/s1>, Instrument S1: Data collection instrument.

Author Contributions: Conceptualization, Y.R.M., L.Q.H., J.N.A.T., O.C.C., J.L.M.G., E.S.B. and R.C.S.; methodology, L.Q.H. and O.C.C.; software, O.C.C.; validation, Y.R.M., L.Q.H., J.N.A.T., O.C.C., J.L.M.G., E.S.B. and R.C.S.; formal analysis, Y.R.M., L.Q.H., J.N.A.T., O.C.C., J.L.M.G., E.S.B. and R.C.S.; investigation, Y.R.M., L.Q.H., J.N.A.T., O.C.C., J.L.M.G., E.S.B. and R.C.S.; resources, Y.R.M.; data curation, Y.R.M., L.Q.H., J.N.A.T., O.C.C., J.L.M.G., E.S.B. and R.C.S.; writing—original draft preparation, Y.R.M., L.Q.H., J.N.A.T., O.C.C., J.L.M.G., E.S.B. and R.C.S.; writing—review and editing, L.Q.H. and O.C.C.; visualization, Y.R.M., L.Q.H., J.N.A.T., O.C.C., J.L.M.G., E.S.B. and R.C.S.; supervision, J.N.A.T.; project administration, J.N.A.T.; funding acquisition, Y.R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical approval for this study was obtained from the Institutional Research Ethics Committee (Comité Institucional de Ética en la Investigación, CIEI) of the Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Peru. The study was approved under ethics approval code CIEI-N° 00267, granted on 6 November 2025.

Informed Consent Statement: Written informed consent was obtained from all participants before they participated in the study.

Data Availability Statement: The original contributions presented in this study are included in the article and Supplementary Material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Carballo-Mendivil, B.; Arellano, A.; Ríos, N.; Lizardi, M.d.P. Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data. *Appl. Sci.* **2025**, *15*, 9202. [\[CrossRef\]](#)
2. Dia, N.J.; Sieras, J.C.; Khalid, S.A.; Macatotong, A.H.T.; Mondejar, J.M.; Genotiva, E.R.; Delena, R.D. EduGuard RetainX: An advanced analytical dashboard for predicting and improving student retention in tertiary education. *SoftwareX* **2025**, *29*, 102057. [\[CrossRef\]](#)
3. Bird, K.A.; Castleman, B.L.; Mabel, Z.; Song, Y. Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. *AERA Open* **2021**, *7*, 23328584211037630. [\[CrossRef\]](#)
4. Mutka, A.; Mutka, F.Ž.; Žagar, M.; Tolić, D. Enhancing Student Retention in Introductory Programming Courses: Leveraging Advanced Learning Validation Tools and Educational Data Mining. *IEEE Access* **2025**, *13*, 153614–153626. [\[CrossRef\]](#)
5. Araka, E.; Wario, R.; Maina, E. Promoting University Students' Self-Regulated Learning Skills on E-Learning Platforms Using Educational Data Mining. In Proceedings of the 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 28–30 May 2025; pp. 1–12.
6. Tang, Z.; von Seekamm, K.; Colina, F.E.; Chen, L. Enhancing Student Retention with Machine Learning: A Data-Driven Approach to Predicting College Student Persistence. *J. Coll. Stud. Retent.* **2025**, 15210251251336372. [\[CrossRef\]](#)
7. Klačnja-Milićević, A.; Ivanović, M.; Vesin, B.; Satratzemi, M.; Wasson, B. Editorial: Learning Analytics—Trends and Challenges. *Front. Artif. Intell.* **2022**, *5*, 856807. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Pimentel, M.; Villamar, M.; Andrade, D.; Zambrano, B. Estrategias para evitar la deserción universitaria. *RECIAMUC* **2023**, *7*, 273–280. [\[CrossRef\]](#)
9. Garcés, M.; De la Ossa, S.; Arellano, W.; Alvis, J.; Figueroa, L. ¿Volver o no volver a las clases presenciales? Motivaciones y temores que influyen en la deserción universitaria en Colombia en tiempos de postpandemia. *Salud Uninorte* **2024**, *40*, 52–68. [\[CrossRef\]](#)
10. French, A. Toward a New Conceptual Model: Integrating the Social Change Model of Leadership Development and Tinto's Model of Student Persistence. *J. Leadersh. Educ.* **2017**, *16*, 97–117. [\[CrossRef\]](#)
11. Savage, M.W.; Strom, R.E.; Hubbard, A.S.E.; Aune, K.S. Commitment in College Student Persistence. *J. Coll. Stud. Retent.* **2019**, *21*, 242–264. [\[CrossRef\]](#)
12. Elturki, E.; Liu, Y.; Hjeltness, J.; Hellmann, K. Needs, Expectations, and Experiences of International Students in Pathway Programs in the United States. *J. Int. Stud.* **2019**, *9*, 192–210. [\[CrossRef\]](#)
13. Álvarez-Santana, C.; Caicedo-Montesdeoca, D. La intervención social y la tutoría estudiantil como medida de disminución de los índices de deserción de las universidades de la provincia de Manabí, periodo 2015–2019. *Rev. Cient. Multidiscip. Arbitr. Yachasun* **2021**, *5*, 36–50. [\[CrossRef\]](#)
14. Rubén, E.; García, A. Clima y Compromiso Organizacional. 2007. Available online: <https://www.eumed.net/libros-gratis/2007c/340/> (accessed on 3 September 2025).
15. Velasquez, M.; Posada, P.M.; Gomez, D.N.C.; Lopez, N.; Vallejo, G.F.; Ramirez, P.A.; Hernandez, E.C.; Vallejo, A. Acciones Para Favorecer La Permanencia. Universidad de Antioquía, 2011. Available online: <https://revistas.utp.ac.pa/index.php/clabes/article/view/856> (accessed on 3 September 2025).
16. Aisenson, G.; Valenzuela, V.; Celeiro, R.; Bailac, K.; Legaspi, L. El significado del estudio y la motivación escolar de jóvenes que asisten a circuitos educativos diferenciados socioeconómicamente. *Anu. Investig.* **2010**, *12*, 109–119.
17. Castrillón-Gómez, O.D.; Sarache, W.; Ruiz-Herrera, S. Predicción de las principales variables que conllevan al abandono estudiantil por medio de técnicas de minería de datos. *Form. Univ.* **2020**, *13*, 217–228. [\[CrossRef\]](#)
18. Castro, L.; Esperanza, E.; Romero, R. Análisis de características que influyen en la deserción estudiantil en el contexto de una universidad latinoamericana. *Rev. EIA* **2023**, *20*, 4002. [\[CrossRef\]](#)
19. Deleña, R.D.; Dia, N.J.; Sacayan, R.R.; Sieras, J.C.; Khalid, S.A.; Macatotong, A.H.T.; Gulam, S.B. Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors. *Syst. Soft Comput.* **2025**, *7*, 200352. [\[CrossRef\]](#)
20. Pérez, M.; Navarrete, D.; Baldeon-Calisto, M.; Guerrero, Y.; Sarmiento, A. Unlocking Student Success: Applying Machine Learning for Predicting Student Dropout in Higher Education. In Proceedings of the 2025 13th International Symposium on Digital Forensics and Security (ISDFS), Boston, MA, USA, 24–25 April 2025; pp. 1–6.

21. Torres, C.Z.; Ramos, C.A.; Moraga, J.L. Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos. *Cienc. Amaz.* **2016**, *6*, 73. [\[CrossRef\]](#)
22. Eckert, K.B.; Suénaga, R. Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. *Form. Univ.* **2015**, *8*, 3–12. [\[CrossRef\]](#)
23. Miranda, M.; Guzmán, J. Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Form. Univ.* **2017**, *10*, 61–68. [\[CrossRef\]](#)
24. González, A.; Alonso, M.A.; Gómez, M.d.L.Á.; Aliagas, I. Peer mentoring, university dropout and academic performance before, during, and after the pandemic in Spain. *Eval. Program Plan.* **2025**, *113*, 102676. [\[CrossRef\]](#)
25. Grijalva, P.; Freire, V.; Real, K.; Arellano, A.; Cornejo, G. Aplicación de Técnicas de Minería de Datos para el Análisis de la Eficiencia Académica. *Rev. Cient. Hallazgos* **2018**, *3*, 1–16. [\[CrossRef\]](#)
26. Dórame, D.L. Factores asociados a la permanencia estudiantil de la universidad de Sonora. *Rev. Psicol. Univ. Auton. Estado México* **2022**, *11*, 70–96. [\[CrossRef\]](#)
27. Almalawi, A.; Soh, B.; Li, A.; Samra, H. Predictive Models for Educational Purposes: A Systematic Review. *Big Data Cogn. Comput.* **2024**, *8*, 187. [\[CrossRef\]](#)
28. Lampropoulos, G.; Evangelidis, G. Learning Analytics and Educational Data Mining in Augmented Reality, Virtual Reality, and the Metaverse: A Systematic Literature Review, Content Analysis, and Bibliometric Analysis. *Appl. Sci.* **2025**, *15*, 971. [\[CrossRef\]](#)
29. Khalid, F.; Javed, A.; Ain, Q.-U.; Ilyas, H.; Irtaza, A. DFGNN: An interpretable and generalized graph neural network for deepfakes detection. *Expert Syst. Appl.* **2023**, *222*, 119843. [\[CrossRef\]](#)
30. López-Meneses, E.; Mellado-Moreno, P.C.; Herrerías, C.G.; Pelicano-Piris, N. Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics. *Computers* **2025**, *14*, 68. [\[CrossRef\]](#)
31. Reina, Y.; Huatangari, L.Q.; Caro, O.C.; Guevara, J.L.M.; Tuesta, J.N.A.; Bardales, E.S.; Santos, R.C. Data Mining to Identify University Student Dropout Factors. *Appl. Sci.* **2025**, *15*, 11911. [\[CrossRef\]](#)
32. Murillo-Zabala, A.M.; Santos, P.J.-D.L. Permanencia estudiantil: Factores que inciden en el Politécnico Internacional de Bogotá, Colombia. *Rev. Electron. Educ.* **2021**, *25*, 1–25. [\[CrossRef\]](#)
33. Narváez, Y.V.; Medina, M.A.G. Factores asociados a la permanencia de estudiantes universitarios: Caso uamm-uat. *Rev. Educ. Super.* **2017**, *46*, 117–138. [\[CrossRef\]](#)
34. Cusquillo, E.J.L.; Cambell, D.C.V.; Vera, M.A.L.; Morán, N.Y.B.; Santander, K.M.A. Estrategias Activas de Aprendizaje: Incidencia en el Rendimiento Académico de Estudiantes de Básica Superior. *Cienc. Lat. Rev. Cient. Multidiscip.* **2025**, *9*, 6469–6480. [\[CrossRef\]](#)
35. Torres-Garagundo, V.; Quispe-Chero, C. Aprendizaje autorregulado y motivación intrínseca en estudiantes de la UNMSM. *PsiqueMag/Rev. Cient. Digit. Psicol.* **2021**, *11*, 18–27. [\[CrossRef\]](#)
36. Vaarma, M.; Li, H. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technol. Soc.* **2024**, *76*, 102474. [\[CrossRef\]](#)
37. Resendiz, J.E.L.; de Oca, E.R.M.; Jiménez, L.P.L. Importancia de la tutoría en la formación académica de estudiantes de agronomía. *RIDE Rev. Iberoam. Para Investig. Desarro. Educ.* **2025**, *15*, e883. [\[CrossRef\]](#)
38. Constate-Amores, A.; Martínez, E.F.; Asencio, E.N.; Fernández-Mellizo, M. Factores asociados al abandono universitario. *Educ. XX1* **2020**, *24*, 17–44. [\[CrossRef\]](#)
39. Bravo, A.; Gonzáles, D.; Maytorena, M. Motivación De Logro En Situaciones De Éxito Y Fracaso Académico De Estudiantes Universitarios. Available online: https://www.comie.org.mx/congreso/memoriaelectronica/v10/pdf/area_tematica_01/ponencias/0762-F.pdf (accessed on 3 September 2025).
40. Casanova, J.; Cervero, A.; Núñez, J.; Almeida, L.; Bernardo, A. Factors that determine the persistence and dropout of university students. *Psicothema* **2018**, *4*, 408–414. [\[CrossRef\]](#)
41. Palacios, C.A.; Reyes-Suárez, J.A.; Bearzotti, L.A.; Leiva, V.; Marchant, C. Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile. *Entropy* **2021**, *23*, 485. [\[CrossRef\]](#)
42. Gonzales, T. El alumno ante la escuela y su propio aprendizaje: Algunas líneas de investigación en torno al concepto de implicación. *REICE Rev. Iberoam. Sobre Calid. Efic. Cambio Educ.* **2010**, *8*, 10–31. Available online: <https://www.redalyc.org/articulo.oa?id=55115064002> (accessed on 3 September 2025).
43. Contreras-Bravo, L.-E.; Tarazona-Bermúdez, G.-M.; Rodríguez-Molano, J.-I. Tecnología y analítica del aprendizaje: Una revisión a la literatura. *Rev. Cient.* **2021**, *41*, 150–168. [\[CrossRef\]](#)
44. Viberg, O.; Hatakka, M.; Bälter, O.; Mavroudi, A. The current landscape of learning analytics in higher education. *Comput. Hum. Behav.* **2018**, *89*, 98–110. [\[CrossRef\]](#)
45. Khalil, M.; Prinsloo, P.; Slade, S. The use and application of learning theory in learning analytics: A scoping review. *J. Comput. High. Educ.* **2023**, *35*, 573–594. [\[CrossRef\]](#)
46. Choque, V.M.; Jauregui, V.D.S. Análisis del Diseño curricular como factor de deserción académica utilizando Minería de Datos. *Yachay—Rev. Cient. Cult.* **2022**, *11*, 551–555. [\[CrossRef\]](#)

47. Sifuentes, M.S.G.C.; Pérez, L.G.V.; Cantabrana, M.G.N.; Acosta, I.I.F.O.; Santana, F.A.Á.; Fierro, M.d.L.Á.S. Modelo Predictivo de la Deserción Escolar en Educación Superior: Una Aproximación desde la Minería de Datos Utilizando la Metodología CRISP-DM. *Cienc. Lat. Rev. Cient. Multidiscip.* **2023**, *7*, 7797–7812. [\[CrossRef\]](#)
48. Bakariwie, A.; Asamoah, D.; Duwieuah, A.B. Prevention of student attrition: A data-backed approach to school counselling using Delphi technique and multiple classification algorithms. *Discov. Educ.* **2025**, *4*, 259. [\[CrossRef\]](#)
49. Deng, P.-S. Using Affinity Analysis-Driven Adaptive Data Mining Life Cycle for the Development of a Student Retention DSS. *WSEAS Trans. Adv. Eng. Educ.* **2021**, *18*, 135–147. [\[CrossRef\]](#)
50. Xu, T.; Hsu, H.-Y.; Wang, Y.; Li, X.-B. Applying Text Mining to Identify Critical Factors Contributing to the Retention Rate of First-Year Students at Historically Black Colleges and Universities. *J. Coll. Stud. Retent.* **2025**. [\[CrossRef\]](#)
51. Attiya, W.M.; Bin Shams, M. Predicting Student Retention in Higher Education Using Data Mining Techniques: A Literature Review. In Proceedings of the 2023 International Conference on Cyber Management and Engineering (CyMaEn), Bangkok, Thailand, 26–27 January 2023; pp. 171–177.
52. Cardona, T.; Cudney, E.A.; Hoerl, R.; Snyder, J. Data Mining and Machine Learning Retention Models in Higher Education. *J. Coll. Stud. Retent.* **2023**, *25*, 51–75. [\[CrossRef\]](#)
53. Shafiq, D.A.; Marjani, M.; Habeeb, R.A.A.; Asirvatham, D. Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. *IEEE Access* **2022**, *10*, 72480–72503. [\[CrossRef\]](#)
54. Yu, C.H.; DiGangi, S.; Jannasch-Pennell, A.; Kaprolet, C. A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. *J. Data Sci.* **2021**, *8*, 307–325. [\[CrossRef\]](#)
55. Kang, K.; Wang, S. Analyze and Predict Student Dropout from Online Programs. In Proceedings of the ICCDA 2018: 2018 The 2nd International Conference on Compute and Data Analysis, DeKalb, IL, USA, 23–25 March 2018; pp. 6–12.
56. Navarro, M.M.; Utreras, E.G.; Ugarte, C.G.B.; Vidal, C.L. Factores psicológicos asociados a la permanencia de estudiantes beneficiados por el Programa de Acceso-Acompañamiento Efectivo a la Educación Superior. *Rev. Electron. Investig. Educ.* **2023**, *25*, 1–13. [\[CrossRef\]](#)
57. De la Hoz-Granadillo, E.J.; Reyes-Ruiz, L.; Sanchez-Villegas, M. Cluster analysis and artificial neural networks to assess and diagnosis suicide ideation in school adolescents. *Rev. Interam. Psicol./Interam. J. Psychol.* **2023**, *57*, e1360. [\[CrossRef\]](#)
58. Jaramillo, J.D.F.; Olivera, N.R.N. Aplicación de Inteligencia Artificial en la Educación de América Latina: Tendencias, Beneficios y Desafíos. *Rev. Veritas Difus. Cient.* **2024**, *5*, 1–21. [\[CrossRef\]](#)
59. Martínez, J. Auto-motivación y rendimiento académico en el Espacio Europeo de Educación Superior. *Cuad. Educ. Desarro. Laguna EUMED* **2011**, *3*, 1–12. Available online: <https://dialnet.unirioja.es/servlet/articulo?codigo=6372719> (accessed on 3 September 2025).
60. Jadue, G. Hacia una mayor permanencia en el sistema escolar de los niños en riesgo de bajo rendimiento y de deserción. *Estud. Pedagog.* **1999**, *25*, 83–90. Available online: <https://www.redalyc.org/pdf/1735/173513845005.pdf> (accessed on 3 September 2025). [\[CrossRef\]](#)
61. Fonseca, G.; García, F. Permanencia y abandono de estudios en estudiantes universitarios: Un análisis desde la teoría organizacional. *Rev. Educ. Super.* **2016**, *45*, 25–39. [\[CrossRef\]](#)
62. Pascarella, E.T.; Terenzini, P.T. Predicting Freshman Persistence and Voluntary Dropout Decisions from a Theoretical Model. *J. High. Educ.* **1980**, *51*, 60–75. [\[CrossRef\]](#)
63. Fishbein, M.; Ajzen, I. *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*; Addison-Wesley: Reading, MA, USA, 1975; Available online: <https://people.umass.edu/ajzen/f&a1975.html> (accessed on 3 September 2025).
64. Bean, J.P. Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Res. High. Educ.* **1980**, *12*, 155–187. [\[CrossRef\]](#)
65. Spady, W.G. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange* **1970**, *1*, 64–85. [\[CrossRef\]](#)
66. Ahmed, W.; Wani, M.A.; Plawiak, P.; Meshoul, S.; Mahmoud, A.; Hammad, M. Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. *Sci. Rep.* **2025**, *15*, 26879. [\[CrossRef\]](#)
67. Nadkarni, P. Core Technologies: Data Mining and ‘Big Data’. In *Clinical Research Computing*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 187–204. [\[CrossRef\]](#)
68. Quan, Z.; Pu, L. An improved accurate classification method for online education resources based on support vector machine (SVM): Algorithm and experiment. *Educ. Inf. Technol.* **2023**, *28*, 8097–8111. [\[CrossRef\]](#)
69. Kalra, M.; Kumar, V.; Kaur, M.; Idris, S.A.; Öztürk, Ş.; Alshazly, H. Attribute weighted naïve bayes classifier. *Comput. Mater. Contin.* **2022**, *71*, 1945–1957. [\[CrossRef\]](#)
70. Blockeel, H.; Devos, L.; Frénay, B.; Nanfack, G.; Nijssen, S. Decision trees: From efficient prediction to responsible AI. *Front. Artif. Intell.* **2023**, *6*, 1124553. [\[CrossRef\]](#) [\[PubMed\]](#)
71. Schönlau, M.; Zou, R.Y. The random forest algorithm for statistical learning. *Stata J. Promot. Commun. Stat. Stata* **2020**, *20*, 3–29. [\[CrossRef\]](#)
72. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967. [\[CrossRef\]](#)

73. Bates, S.; Hastie, T.; Tibshirani, R. Cross-Validation: What Does It Estimate and How Well Does It Do It? *J. Am. Stat. Assoc.* **2024**, *119*, 1434–1445. [[CrossRef](#)]
74. Allgaier, J.; Pryss, R. Cross-Validation Visualized: A Narrative Guide to Advanced Methods. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1378–1388. [[CrossRef](#)]
75. Zhan, Z.; Shen, T. Development of a prediction model for student teaching satisfaction based on 10 machine learning algorithms. *Sci. Rep.* **2025**, *15*, 36547. [[CrossRef](#)]
76. Goldrick-Rab, S. *Paying the Price: College Costs, Financial Aid, and the Betrayal of the American Dream*; University of Chicago Press: Chicago, IL, USA, 2016.
77. Tinto, V. *Completing College: Rethinking Institutional Action*; University of Chicago Press: Chicago, IL, USA, 2012.
78. Kuh, G.D.; Cruce, T.M.; Shoup, R.; Kinzie, J.; Gonyea, R.M. Unmasking the Effects of Student Engagement on First-Year College Grades and Persistence. *J. High. Educ.* **2008**, *79*, 540–563. [[CrossRef](#)]
79. Nabi, G.; Walmsley, A.; Mir, M.; Osman, S. The impact of mentoring in higher education on student career development: A systematic review and research agenda. *Stud. High. Educ.* **2025**, *50*, 739–755. [[CrossRef](#)]
80. Merisotis, J.P.; McCarthy, K. Retention and student success at minority-serving institutions. *New Dir. Institutional Res.* **2005**, *2005*, 45–58. [[CrossRef](#)]
81. Ortiz, D.G. State Repression and Mobilization in Latin America. In *Handbook of Social Movements Across Latin America*; Almeida, P., Cordero Ulate, A., Eds.; Springer: Dordrecht, The Netherlands, 2015; pp. 43–59. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.