OmniaScience

Journal of Technology and Science Education

JOTSE, 2026 – 16(2): 554-575 – Online ISSN: 2013-6374 – Print ISSN: 2014-5349

https://doi.org/10.3926/jotse.4362

CHARACTERIZATION AND PREDICTION OF ACADEMIC PERFORMANCE IN HIGHER EDUCATION IN ENGINEERING USING MACHINE LEARNING TECHNIQUES

Gustavo Sosa-Cabrera[1*], Rossana Martínez[2]

[1]Facultad Politécnica - Universidad Nacional de Asunción, Asunción (Paraguay)

[2]Universidad Nacional de Caaguazú, Caaguazú (Paraguay)

Corresponding author: gdsosa@pol.una.py

rmartinez@unca.edu.py

Received February 2026

Accepted July 2026

Abstract

Today, academic performance in higher education engineering programs remains a key indicator of student retention and educational quality. The early identification of at-risk students poses an institutional challenge for universities due to the multiple factors involved in each case. The main objective of this study is to compare the predictive power of two categories of information: (1) early academic variables and (2) socioeconomic variables. The methodology employed was a quantitative approach based on educational data mining techniques. Thus, predictive models were built using Random Forest with cost-sensitive classification, which made it possible to increase sensitivity in identifying at-risk students. In addition, the relative importance of the variables was quantified, and clustering techniques were applied to analyze the structural coherence of the institutionally defined target variable. The results show that first-year academic performance has greater discriminatory power than the socioeconomic variables considered in isolation. Furthermore, performance in basic science subjects was identified as the most relevant structural predictor. Consequently, this study has shown that early academic data constitute a solid and reliable source of information for identifying academic risk in engineering. Finally, this study provides empirical evidence for the design of early-warning systems in engineering schools.

Keywords – Academic performance, Early academic risk, Early warning systems, Engineering education, Machine learning, Educational data mining.

To cite this article:

Sosa-Cabrera, G. &Martínez, R. (2026). Characterization and prediction of academic performance in higher education in engineering using machine learning techniques. *Journal of Technology and Science Education, 16*(2), 554–575. https://doi.org/10.3926/jotse.4362

---

1. Introduction

Today, higher education is a strategic factor in the scientific and technological development of nations. In particular, engineering programs play a central role in fostering innovation and competitiveness. However, student retention in these programs remains a structural challenge for institutions, as noted by Martínez and Sosa-Cabrera (2024).

In this regard, academic performance is one of the main indicators of college success and is typically measured by grades earned on formal assessments. However, as Lázaro-Alvarez et al. (2020) point out, dropout and failure to graduate in engineering programs stem from a multifactorial phenomenon involving early academic performance, admission characteristics and conditions, and contextual factors. A large number of scholars have recently suggested that it is possible to obtain early estimates of the risk of dropping out using data available as early as the first semester, achieving acceptable levels of accuracy (Bañeres et al., 2020).

It is worth noting that, from a theoretical perspective, the study of college performance has traditionally been approached through descriptive models based on academic and social integration. Tinto (1993) argues that retention in higher education depends largely on the degree of a student's academic and social integration within the institution. Today, this concept has been extended to include a sense of belonging in hybrid learning environments (Kahu & Nelson, 2018). Complementarily, Bean (1980) proposes that the decision to drop out of college is influenced by a combination of academic, psychological, and environmental variables, thereby shaping a model of persistence based on the interaction between performance and context.

Against this backdrop, specifically in the field of higher education in engineering, performance in basic science courses during the first year has been identified as a critical predictor of academic career (Veenstra et al., 2009). Particularly in engineering programs, courses in the basic sciences (i.e., mathematics, physics, and chemistry) play a foundational role in the educational career, as they provide the conceptual foundations upon which the professional courses in higher-level years are built. Various studies have shown that difficulties experienced during this initial period significantly increase the likelihood of falling behind in coursework and even of dropping out entirely (Johnson & Parker, 2025).

Furthermore, Maquen-Niño et al. (2025) demonstrated that prior performance is one of the most significant factors in classifying student performance using machine learning models. Their results show that early academic data has high discriminatory power when supervised classification techniques are used.

Furthermore, findings from previous studies have suggested that the cultural capital approach provides an interpretive framework for understanding the influence of socioeconomic variables. In (Bourdieu, 2018), it is argued that cultural capital inherited from the family environment influences educational trajectories, shaping dispositions and resources that affect academic performance. Variables such as parental educational level or age of entry may act as structural determinants; however, today they are mediated by the digital literacy and AI gap (Baidoo-Anu et al., 2024), which adds a new layer of complexity to the cultural capital required for technical success.

In recent decades, the development of what is known as Educational Data Mining and Learning Analytics (Romero & Ventura, 2020) has made it possible to apply these conceptual frameworks to the systematic analysis of large volumes of educational data with the aim of understanding learning processes, predicting academic trajectories, and supporting evidence-based decision-making. According to Romero and Ventura (2020), data mining techniques applied to educational contexts facilitate the identification of hidden patterns and the prediction of academic behaviors.

In this regard, machine learning has established itself as a tool capable of modeling nonlinear relationships and capturing complex interactions between academic and contextual variables (Xing & Du, 2019). Numerous studies have shown that predictive models make it possible to identify at-risk students early enough to implement preventive interventions that reduce academic failure and college dropout rates.

However, a gap remains in the literature, as most of the studies reported in the literature simultaneously incorporate academic and socioeconomic variables into a single predictive model, without comparatively evaluating their relative discriminatory power. Thus, uncertainty remains regarding which dimension—that is, the academic or the contextual—provides greater predictive power in specific engineering education

contexts. Consequently, there is a need for research that isolates and compares the weight of these sets of informative attributes.

This study aims to address the gaps in research mentioned earlier. In this regard, the main contribution consists of empirically comparing two distinct informational dimensions within a real institutional context of higher education in engineering: (1) early academic variables and (2) socioeconomic entry variables. Unlike previous studies focused exclusively on maximizing predictive accuracy, this research seeks to determine what type of institutional information provides the greatest value for the development of early warning systems in higher education.

As far as this study is concerned, its purpose is to systematically compare the relative usefulness of different sources of educational information within the same institutional context, providing empirical evidence that can support the design of strategies for academic monitoring and student retention in engineering programs.

2. Theoretical Framework

2.1. Academic Performance in Higher Education in Engineering

Academic performance is one of the main indicators of the quality of the teaching-learning process in higher education. From an institutional perspective, it reflects the degree to which expected learning outcomes are achieved, while from the student's perspective, it serves as an indicator of their academic progress, retention, and likelihood of graduation. In engineering programs, the analysis of academic performance takes on special relevance due to the high failure and dropout rates recorded during the first years of study, particularly in courses in the area of basic sciences (Veenstra et al., 2009).

These courses provide the mathematical and scientific foundations necessary for the development of future professional competencies; therefore, difficulties encountered at this stage often have repercussions throughout a student's entire academic career. Various theoretical models have explained student retention as a multifactorial phenomenon in which individual, academic, institutional, and social factors interact simultaneously. Among these, Tinto's (1993) model of academic and social integration remains one of the most influential conceptual frameworks for understanding college retention, while the contributions of Bean (1980) and Kahu and Nelson (2018) emphasize the interaction between student engagement, institutional integration, and socioeconomic context as determinants of academic success. These approaches agree that performance during the first semesters serves as an early indicator of the risk of falling behind in coursework or dropping out.

2.2. The Importance of Performance in Basic Sciences During the First Year

Performance in basic science courses during the first year is critical in engineering programs because it underpins subsequent understanding of technical subjects and serves as an early indicator of academic progress (Johnson & Parker, 2025; Maquen-Niño et al., 2025). Strong initial performance facilitates retention and reduces the risk of failing and dropping out in later stages. The importance of these courses should not be interpreted solely as a result of their intrinsic difficulty. From a curricular perspective, performance in basic sciences simultaneously reflects prior mathematical competencies, study habits, problem-solving skills, and the ability to adapt to the demands of higher education. Therefore, using this performance as an early indicator is particularly useful for designing institutional strategies for academic support.

2.3. Learning Analytics y Educational Data Mining

Over the past decade, the development of Learning Analytics (LA) and Educational Data Mining (EDM) has made it possible to incorporate data analysis methodologies to understand and optimize educational processes (Romero & Ventura, 2020). Although both disciplines share similar computational techniques, they differ somewhat in their purpose. While Learning Analytics is primarily geared toward supporting educational decision-making through the interpretation of academic data, Educational Data

Mining focuses on the automatic discovery of behavioral patterns using statistical and machine learning techniques. Currently, both fields are converging on the development of institutional early-warning systems capable of identifying students at academic risk before failure or dropout occurs. These systems use information from academic platforms, academic records, and demographic variables to generate predictive models that support preventive interventions aimed at improving student retention. Recent studies show that the practical utility of these models depends not only on their predictive accuracy but also on the interpretability of the identified factors and their ability to guide specific institutional actions.

2.4. Predictive Models of the Academic Performance

Predicting academic performance is one of the most well-established lines of research within Educational Data Mining (Xing & Du, 2019). The literature reports the use of traditional statistical techniques, decision trees, support vector machines, neural networks, and ensemble methods such as Random Forest and Gradient Boosting. However, available evidence indicates that increased predictive accuracy does not always translate into greater value for educational management, especially when models have low interpretability or use a large number of variables that are difficult to obtain. Consequently, a recent trend involves evaluating not only the performance of the algorithms but also the relative contribution of different dimensions of institutional information, such as academic background, socioeconomic variables, demographic characteristics, or indicators of interaction with educational platforms. This approach is particularly relevant for the design of early warning systems, where the primary objective is to identify vulnerable students in a timely manner using variables available during the early stages of their college career.

2.5. Technics for Automatic Learning

In order to compare the predictive power of different information dimensions, the literature on Educational Data Mining supports the combined use of supervised and unsupervised approaches. Within this theoretical framework, the integration of Decision Trees, Random Forests, attribute selection techniques based on information gain, and the K-Means clustering algorithm is justified by their proven ability to model nonlinear relationships, estimate the relative importance of complex variables, and maintain an optimal balance between computational accuracy and pedagogical interpretability. To support this conceptual analysis, the standard score (Z-score) serves as a key statistical measure that describes the position of a value relative to a group's mean; in the field of educational analytics, this standardization is theoretically essential for homogenizing grades across different courses, neutralizing the bias associated with the intrinsic difficulty of each course, and enabling a fair comparison of students' relative performance (Sholeh & Nurnawati, 2024). Based on this same principle of equivalence, the K-Means algorithm theory supports the feasibility of clustering observations based on their proximity to their centroids; a method widely validated in the analysis of higher academic performance (Mohamed-Nafuri et al., 2022) to verify whether students naturally cluster according to latent institutional risk profiles.

From the perspective of supervised learning, decision trees provide a logical foundation by representing sets of classification rules through hierarchical structures of nodes and branches that are easy for decision-makers to read (Segal & Xiao, 2011). However, to overcome the limitations of overfitting in a single model, ensemble theory supports the use of Random Forests, which combine multiple trees using random subsamples (bagging) and determine the final prediction by majority vote—a crucial property for capturing complex interactions between socioeconomic background and academic success (Segal & Xiao, 2011). The theoretical relevance of these explanatory variables is measured using Information Gain, a metric based on entropy in information theory that quantifies the reduction in uncertainty regarding the target variable (Sosa-Cabrera et al., 2024), allowing us to identify which factors provide greater conceptual clarity in distinguishing between performance groups. Finally, the literature acknowledges that educational settings exhibit a marked class imbalance, where students in extreme situations tend to be a minority that skews conventional algorithms (Radwan & Cataltepe, 2017). To

correct this conceptual asymmetry, the theory of the asymmetric cost matrix proposes assigning differentiated weights to classification errors (Fernández et al., 2018). This approach is ethically robust for engineering education, as it recognizes that the institutional consequences of a false negative (failing to identify a student at real risk of dropping out) are significantly greater than those of a false positive. To make the techniques described easier to understand, Table 1 provides a concise summary of the seven conceptual approaches analyzed.

Approach / Technique	Purpose in Educational Analytics	Key Advantage	Theoretical Limitation
Z-Score	Relative Performance Normalization.	Eliminates the disparity in difficulty among different courses.	It requires that the data follow an approximately normal distribution.
K-Means	Student Profile Discovery. Automatically and unsupervised, it groups students based on similar characteristics.	It automatically and unsupervised clusters data based on similar features.	Sensitive to the initial choice of the number of clusters (k).
Decision Trees	Generation of risk rules.	High visual interpretability for academic administrators.	Tendency toward over-adjustment if the tree is very deep.
Random Forest	Robust prediction of dropout.	High precision and effective handling of nonlinear interactions.	It functions like a "black box" that is less transparent than a single tree.
Information Gain	Identification of critical predictors.	Filter out irrelevant variables (e.g., demographic and academic variables).	It tends to favor attributes with many different numerical values.
Unbalance classes	Correction of asymmetry in the license plate.	Prevent the model from ignoring the at-risk minority.	It can artificially inflate the false-positive rate if it is not calibrated.
Cost Matrix	Mitigation of institutional impact.	Failing to identify a student at risk (false negative) is penalized.	It requires the institution to provide a subjective definition of the cost.

Table 1. A Comparative Analysis of Educational Data Mining Approaches: Objectives, Methodological Advantages, and Conceptual Limitations for Institutional Decision-Making

3. Methodology

3.1. Research Design

The methodological approach adopted in this study is based on a quantitative, explanatory–predictive design aimed at modeling academic performance in higher education in engineering. Thus, it combines statistical analysis and supervised machine learning techniques with the aim of: (1) identifying patterns associated with academic performance; (2) building interpretable predictive models; and (3) generating useful evidence for institutional decision-making regarding engineering programs.

3.2. Data

The research data in this study are drawn from two main sources: (1) the Academic Data Management Information System of a publicly administered technological university located in the Paraguay region from 2010 to 2023, and (2) the results of a socioeconomic survey of students and graduates of the Civil Engineering program conducted in 2025.

The academic records were disaggregated for analysis into two independent datasets: (1) the Academic Dataset, comprising 1,414 students from four engineering programs (Civil, Electrical, Electronic, and Computer Engineering), and (2) the Socioeconomic Dataset, comprising 410 Civil Engineering students. In both cases, 10-fold stratified cross-validation was applied. The records were anonymized. The dataset

contained no missing values, and no outliers or significant anomalies were detected during the exploratory analysis and preliminary validation stages.

3.3. Variables

The observation window spans a period of 13 (thirteen) years, from the 2010 to 2023 school years, during which the variables were categorized according to their nature and function within the predictive model. The target variable was treated as a binary nominal categorical variable in both sets of experiments. No pre-balancing techniques were applied. Any potential imbalance was addressed during the modeling phase using cost-sensitive classification. The definition adopted does not imply causality. It represents a functional operationalization of the concept of early academic risk within the institutional context analyzed.

3.3.1. Target Variable

It is defined based on an institutional criterion aimed at the early identification of students at academic risk. In both datasets, a binary variable called "Early Academic Risk" was constructed, although its operationalization differed depending on the nature of the available information.

3.3.2. Target Variable in the Academic Dataset

It was defined based on indicators derived from first-year performance. A composite criterion was adopted to capture different dimensions of risk. A student was classified as Risk = 1 if he or she met at least one of the following conditions:

- Pass rate below 0.60.
- Absenteeism rate above 0.20.
- Early grade point average below the institutional threshold (2.55 on a 1–5 scale).

Otherwise, the student was classified as Risk = 0. The use of a composite criterion made it possible to integrate academic performance and course participation, avoiding reliance solely on the arithmetic mean. This construct is intended to reflect early operational risk rather than isolated underperformance. The variables used in constructing the criterion were excluded from the set of predictors to avoid information leakage.

3.3.3. Target Variable in the Objective of Dataset

The institutional composite index was not directly available in the socioeconomic dataset. Therefore, academic performance was operationalized by standardizing the grade point average using z-scores. The relative performance of each student compared to the cohort was calculated. The result was then binarized into two categories:

- Risk = 1: students whose performance is below the institutionally defined cutoff point. Success $\Rightarrow\Rightarrow z \geq -0.073$ (Integrate Average, Good and Excellent).
- Risk = 0: students whose performance is equal to or above the cutoff score. Risk Level: $\Rightarrow\Rightarrow z < -0.073$ (Includes the critical segment).

This procedure made it possible to maintain conceptual consistency with the notion of early-stage risk without directly incorporating derived academic variables into the set of predictors.

3.3.4. Predictive Variables

The selection and construction of the predictor variables were based on a theoretical and methodological framework designed to capture two complementary dimensions of the phenomenon of early academic performance in engineering: the structural academic dimension and the socioeconomic-biographical dimension. Two independent datasets were designed, each with specific variables tailored to the corresponding analytical objective.

3.3.4.1. Academic Dataset (n = 1414)

The predictor variables in the Academic dataset were derived from institutional records for the first year of study. The unit of analysis is a student, represented by a single consolidated row. The variables used in supervised modeling were as follows:

Basic Demographic Variables

- Gender: Nominal categorical variable (F, M).
- Age at Enrollment: A continuous numerical variable calculated as the difference between the year of enrollment and the year of birth.
- Years in College: Discrete numerical variable indicating cumulative length of enrollment.
- Admission Status: Nominal categorical variable (CPI, DIR).

Academic Variables by Knowledge Area

In order to avoid zero values resulting from differences in the curricula of different degree programs, the courses were grouped into homogeneous areas. For each student, the first-year average in each area was calculated:

- Basic Science: Average grades in math and physics courses.
- Computing: Average grade in programming and computer tools courses.
- General Education: Average grade in cross-disciplinary courses.
- Management: Average grade in courses related to administration and economics.
- Applied Engineering: Average grade in the initial technical courses specific to each degree program.

These variables capture structural performance in distinct curricular domains.

Variables Excluded from the Model

The variables listed below were used to construct the target variable but were excluded from the modeling to prevent information leakage.

- Courses Taken.
- Courses Passed.
- Failed Courses.
- Absences.
- Early Grade Point Average.
- Pass Rate.
- Fail Rate.
- Absenteeism Rate.
- Early Performance Index.
- Student ID.

3.3.4.2. Socioeconomic Dataset (n = 410)

The Socioeconomic Dataset includes 15 biographical attributes collected upon admission to the institution. These variables are designed to capture cultural capital, study conditions, and socioeconomic context.

Sociodemographic Variables

- Gender. Nominal categorical variable (F, M).
- Age. Continuous numerical variable.
- Income Level. (CPI, DIR).
- Children. Binary variable (Yes, No).

Family-Friendly Cultural Capital

- Academic degree of Mother : (High School, College).
- Academic degree of Father : (High School, College).

These variables serve as proxies for the educational capital of the family environment.

Working and Economic Conditions

- Currently Working: (Yes, No).
- Average Hours Worked: (None, <10, 11–20, >20 hours).
- Educational Expenses: (Part-time job, Family support, Scholarships).

Previous Educational Background

- High School GPA: (three, four, five).
- Type of School: (public, private, charter).

Study Strategies and Resources

- Study Method: (Alone, Groups, Both).
- Average Study Hours: (<5 hours, ≥5 hours).
- Teaching Tool: (Yes, No).
- Available Educational Resources: (Yes, No).

These variables make it possible to model patterns of organization. These variables make it possible to model patterns of academic organization and institutional support.

To facilitate understanding of the analyzed data ecosystem, Figure 1 systematically details the taxonomy of the variables classified into the study's two key informational dimensions. This conceptual map clearly illustrates the logical relationship between early academic factors and socioeconomic variables.

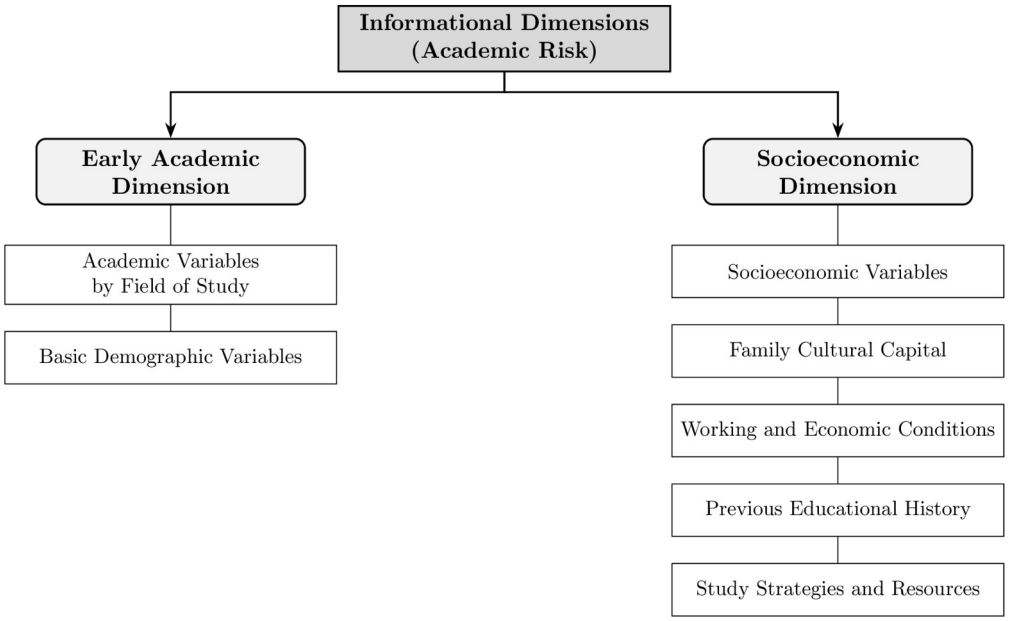

Figure 1. Taxonomy of the informational dimensions and variables analyzed for predicting Academic Risk

3.4. Grading and Standardization System

The courses analyzed are part of the first year of the engineering programs in the School of Science and Technology. All of them are governed by institutional academic regulations, which establish an

official grading scale from 1 to 5, with 2 being the minimum passing grade. The evaluation process follows a continuous assessment model, in which the final grade for each course is determined based on performance in midterm exams, practical assignments, and a final exam. This model aims to progressively assess the achievement of the intended learning outcomes throughout the academic semester.

The dataset used in this study includes information on incoming students from 2010 to 2023, covering multiple academic cohorts. Although institutional regulations maintained the same grading scale and general passing criteria throughout the period analyzed, it is reasonable to consider the existence of variations inherent to the educational context, such as differences in teaching staff, curricular adjustments, modifications to assessment tools, or changes in the level of difficulty of certain courses. These variations may introduce heterogeneity in absolute grades without necessarily representing equivalent differences in students' relative performance.

In this context, the use of standardized scores obtained through Z-score transformation is based on the need to improve the comparability of grades among students from different academic cohorts. Standardization allows individual performance to be expressed relative to the grade distribution of their reference cohort, reducing the influence of contextual variations associated with the assessment process and facilitating the identification of comparable performance patterns throughout the study period.

From a machine learning perspective, this transformation constitutes a preprocessing step aimed at improving the statistical consistency of the data used during the training of predictive models. By working with a homogeneous representation of grades, the algorithms reduce the likelihood of learning patterns associated with incidental differences between cohorts and focus the learning process on more stable relationships between the explanatory variables and academic performance. Consequently, standardization helps increase the robustness and stability of models developed using historical institutional data.

It is important to note that this procedure does not alter the pedagogical significance of the grades awarded by the institution, nor does it replace the assessment system originally applied. The transformation is used exclusively for analytical and predictive modeling purposes, preserving the educational interpretation of the results. Under this approach, the conclusions drawn continue to be based on students' actual academic performance, while standardization serves as a methodological mechanism designed to ensure the longitudinal comparability of the data and strengthen the validity of the inferences derived from the predictive model. Consequently, the use of standardized scores does not imply replacing an assessment based on the achievement of learning outcomes with a relative assessment of the student; it constitutes solely a methodological strategy to standardize the representation of grades during the modeling process, while keeping the academic meaning of the originally administered assessments unchanged.

3.5. Data Preprocessing

3.5.1. Preprocessing of Academic Dataset

The original dataset was in long format, with multiple rows per student due to repeated courses and exam periods. The processing steps are described below.

1. Temporal filtering. Only records corresponding to the first year of study were selected.
2. Consolidation by student. A single row was generated for each student by aggregating records.
3. Removal of academic duplicates. In cases where the same course was taken multiple times, the consolidated final grade was retained.
4. Handling of exam absences. Grades with a value of 0 were coded as academic absences and used to calculate rates.
5. Calculation of Derived Variables: Pass Rate. The proportion of courses passed out of the total number of courses taken.
    a) Failure rate. The proportion of failed courses out of the total number of courses taken.

b) Absenteeism rate. The frequency of absences from exams.
c) Early grade point average. The average grade earned exclusively during the first academic term to assess initial performance.
d) Composite performance index. An integrated metric that weights the academic average with the success rate (pass rate) to provide a multidimensional assessment of performance.

6. Grouping by curricular areas. The courses were mapped to five homogeneous domains to reduce dimensionality and avoid dispersion due to curricular differences among degree programs.
7. Internal standardization. For complementary analyses, standardization was applied using z-scores for continuous variables.
8. Prevention of data leakage. The variables used to define the target variable were excluded from the set of predictors.

3.5.2. Preprocessing of Socioeconomic Dataset

The biographical dataset contained categorical variables with multiple levels and some imbalanced distributions. The data processing is described below.

- Semantic consistency review. Equivalent categories in educational and occupational variables were standardized.
- Categorical encoding. Nominal variables were transformed using binary (one-hot) encoding when required by the algorithm.
- Grouping of infrequent levels. Categories with low marginal frequency were merged to avoid model instability.
- Supervised discretization. For variables such as hours worked and hours studied, institutionally defined intervals were used.
- Standardization of the continuous variable "Age." It was kept on a numerical scale to preserve its discriminatory power.
- Class imbalance control. Neither undersampling nor oversampling was applied. The imbalance was addressed during the modeling phase using a cost matrix.

3.6. Predictive Modeladling Strategy

A comparison of the algorithms' discriminatory power is provided. The design evaluates the structural interpretability of the data. The analysis measures sensitivity to class imbalance. The team implemented a supervised approach for classification. This methodology also incorporates additional unsupervised validation.

3.6.1. General Modeling Framework

Two independent datasets were used. They were not combined into a single model in order to evaluate the relative predictive power of each information dimension. In both cases, stratified cross-validation with 10 (ten) folds was applied. This procedure allowed us to estimate generalizable performance while avoiding partitions that depended on a single training/test split. The metrics evaluated are listed below.

- Accuracy. The total percentage of instances correctly classified by the model out of the total dataset analyzed.
- Kappa statistic. A measure of agreement that adjusts the model's accuracy by eliminating the effect of chance, thereby evaluating the classifier's true robustness.
- Area under the ROC curve (AUC). A probability metric that measures the algorithm's ability to correctly distinguish between the "success" and "risk" classes.
- Recall per class. The proportion of actual cases in a specific category that the model was able to correctly identify without omission.
- Precision per class. The probability that an instance classified into a category actually belongs to that class, measuring the reliability of the prediction.

To increase the reliability of the measurement, it was proposed that it is necessary to evaluate not only overall performance but also the detection capability for the class of interest (Risk).

3.6.2. Implemented Models

3.6.2.1. Random Forest

Random Forest was used as the primary model due to its well-known strong performance in modeling the scenarios listed below.

- Capture nonlinear relationships.
- Reduce overfitting through ensemble methods.
- Estimate the relative importance of attributes.

Based on the existing literature, the model was configured with 500 trees and default parameters for depth and random attribute selection.

The importance of variables was estimated using the average reduction in impurity (Gini) in the Academic dataset and using the model's internal ranking in the Socioeconomic dataset.

3.6.2.2. Decision Tree (J48)

J48 was implemented with the goal of obtaining an interpretable model. The tree made it possible to identify root variables and explicit decision rules. No additional manual pruning was performed beyond the algorithm's standard parameters.

3.6.2.3. Information Gain

InfoGainAttributeEval was used to perform univariate attribute ranking. This analysis made it possible to compare the consistency between individual statistical importance and the structural relevance observed in the Random Forest model.

3.6.3. Addressing Class Imbalance

In the Socioeconomic Dataset, an imbalance between classes was identified. Instead of applying resampling techniques (SMOTE or subsampling), cost-sensitive classification was chosen. A 5:1 cost matrix was defined, penalizing false negatives (failing to detect a student at risk) five times more than false positives. This methodological decision is based on an institutional criterion of preventive prioritization.

In the Academic dataset, no additional penalty was applied because the class distribution allowed for stable performance without adjustments.

3.6.4. Non-Supervised Validation

To analyze the structural consistency of the target variable, the SimpleKMeans algorithm was applied with k=3. Essentially, clustering was used to identify relationships in the data described below.

- Identify natural groupings.
- Evaluate the correspondence between supervised segmentation and latent structure.
- Analyze patterns of demographic and academic variables within each cluster.

In other words, unsupervised analysis was not used to redefine the target variable, but rather as a complementary tool for structural validation.

3.6.5. Reproducible Considerations

The modeling was implemented using the open-source software Weka (Frank et al., 2016) with documented settings. No exhaustive optimization of hyperparameters was performed using grid search, since the primary objective was to compare the explanatory power of the explanatory variables rather than

to marginally maximize accuracy. The strategy adopted prioritizes interpretability, institutional coherence, and predictive capability over intensive parametric adjustments.

To clarify the technical architecture of the proposed predictive model, Figure 2 illustrates the workflow implemented in this research in a sequential manner.

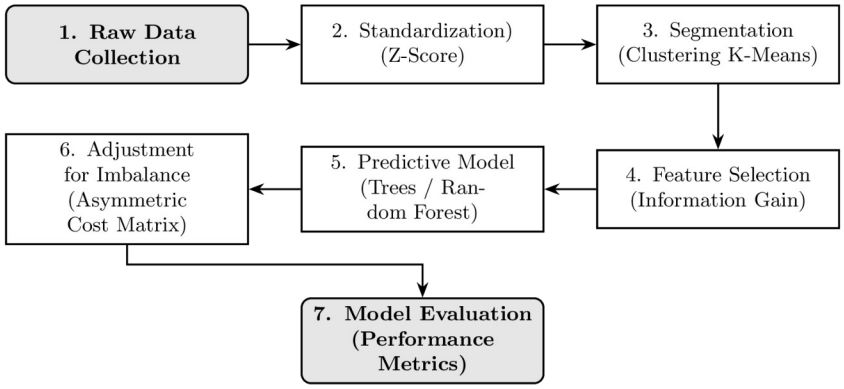

Figure 2. Sequential flowchart of the implemented data preparation, predictive modeling, and cost optimization process

3.7. Unsupervised Characterization Experiments

The goal of unsupervised clustering was to identify latent structures in the data without using the target variable as a guide. This phase made it possible to analyze whether the natural clustering patterns were consistent with the institutionally defined segmentation of "Early Academic Risk." The clustering was not intended to redefine the target variable, but rather to evaluate structural consistency and internal segmentation. Using SimpleKMeans (k=3), it was confirmed that students with $z<-0.073$ exhibit structural cohesion defined by advanced age and heavy workload, thereby validating the target variable from an unsupervised perspective.

The SimpleKMeans algorithm was implemented due to its stability, geometric interpretability, and suitability for standardized numerical variables. The preprocessing steps are listed below.

- Standardization using z-scores for continuous variables.
- Exclusion of the target variable.
- Verification that there is no extreme multicollinearity.

The number of clusters was set to k=3 based on the factors listed below.

- Educational interpretability.
- Preliminary empirical separation.
- Comparability with institutional segmentations.

k (the number of clusters) was not optimized using automatic methods (elbow or silhouette) because the objective was to analyze structural coherence rather than geometric optimization.

3.8. Validation Procedure

Predictive performance validation was designed to estimate the generalizability and stability of the models under different data partitions. A stratified cross-validation approach was adopted, supplemented by class-specific metric analysis.

3.8.1. Stratified Cross-Validation

Ten-fold stratified cross-validation was applied to both datasets. The procedure is described in the following list of steps.

1. Divide the dataset into ten subsets of similar size.
2. Preserve the class proportions in each fold.
3. Use 9 subsets for training and one for testing.
4. Repeat the process ten times, rotating the test subset.
5. Average the metrics obtained across the ten iterations.

Stratification made it possible to maintain stability in the presence of moderate imbalance, particularly in the Socioeconomic Dataset. This approach reduces the variance associated with a single training/test split and provides a more robust estimate of expected performance on unobserved data.

3.8.2. Metrics of Assessment

Complementary metrics were used to avoid relying solely on overall accuracy; specifically, the Kappa statistic was used to assess the model's consistency while taking into account the distribution of classes. The AUC provided a threshold-independent measure of discrimination.

3.8.3. Validation Under Cost-Sensitive Classification

Cost-sensitive classification was implemented in the Socioeconomic Dataset. In this context, validation took into account the defined cost matrix (5:1). The following points are worth noting regarding each validation fold.

- The model was trained using the penalty matrix.
- Predictions were evaluated by considering the differential impact of false negatives.

Thus, this procedure made it possible to estimate performance consistent with the institutional goal of preventive prioritization.

3.8.4. Stability Analysis

The consistency of the results was verified using the procedures listed below.

- Comparison of average metrics across models.
- Analysis of variation across folds.
- Consistency in the ranking of variable importance.

It should be noted that no abrupt variations were observed across folds that would suggest structural instability.

3.8.5. Methodological Considerations

We chose not to perform an additional independent hold-out split due to the moderate size of the datasets and the exploratory, comparative nature of the study. Nor was external validation performed using a different cohort. Therefore, the results should be interpreted as internal evidence of predictive performance. Furthermore, the validation procedure adopted seeks to strike a balance between statistical rigor and institutional feasibility of implementation.

3.9. Model Interpretability

It is widely accepted that interpretability is a central component of models applied to educational contexts. In institutional settings, the ability to explain decisions is just as important as predictive performance. In this regard, the analytical strategy we adopted combined ensemble models with explicit structural interpretation techniques.

3.9.1. Interpretability in Random Forest

Although Random Forest is an ensemble model and does not provide direct decision rules, it allows us to estimate the relative importance of attributes using the average impurity reduction (Gini). Thus, in the Academic dataset, the variables with the greatest contribution are listed below.

- Performance in Basic Sciences.
- Age at admission.
- General Education.
- Years in college.

It is particularly important that the predominance of Basic Sciences confirms its structural role in early- stage risk segmentation in engineering. Thus, relative importance does not imply causality, but it does indicate discriminatory contribution within the model.

Furthermore, in the Socioeconomic Dataset, the most relevant variables are listed below.

- Age.
- Mother's educational level.
- Father's educational level.
- Hours worked.

It can be seen that these variables act as structural determinants that precede formal academic performance. This is why the stability of the ranking of importance across validation runs suggests that the model has internal consistency.

3.9.2. Interpretability through Decision Tree (J48)

Essentially, the J48 algorithm made it possible to extract explicit classification rules. In the Academic dataset, the root variable was Basic Sciences, indicating that the first split in the attribute space is based on performance in this area. The derived rules show that combinations of low performance in Basic Sciences and General Education increase the probability of belonging to the Risk class. In the Socioeconomic Dataset, the first splits in the tree were associated with Age and Parental Educational Level, reinforcing their structural relevance. The moderate depth of the tree indicates that the segmentation does not depend on excessively complex interactions.

3.9.3. Consistency Among Interpretive Methods

The series of analyses we conducted allowed us to observe the following convergences.

- Random Forest importance ranking.
- Information gain (InfoGain).
- Root variables in J48.
- Emerging patterns in clustering.

The result is that this convergence strengthens the internal validity of the main finding, which is that early academic risk in engineering is structured primarily around performance in core areas and specific demographic variables.

3.9.4. Scope of the Interpretation

We can say that the interpretation presented is limited to structural associations within the predictive model. No causal relationship is inferred between the variables and Risk Academic. Thus, the interpretive objective is to provide understandable insights for institutional decision-making, while maintaining consistency with the quantitative results. The combination of models with high predictive power and interpretable models allows for a balance between technical performance and explainability in educational contexts.

3.10. Personal Data Protection

For the purposes of this study, the datasets used were anonymized prior to analytical processing. Any direct identifiers of the students were removed. The analyses were conducted using internal codes that could not be traced publicly. No additional sensitive data beyond the reported academic and

socioeconomic information was used. Access to the data was restricted to authorized researchers within the framework of the institutional project.

4. Results

This section presents only the empirical results obtained from supervised and unsupervised experiments.

It does not include extensive theoretical interpretations.

4.1. Results of the Academic Dataset (n = 1414)

4.1.1. Random Forest and Decision Trees (J48)

Two (2) supervised algorithm implementations were evaluated on the Academic dataset (Random Forest and J48). The Random Forest model with 500 trees and the Decision Tree (J48) model with 56 leaves— both using stratified cross-validation with 10 (ten) folds—showed a predominance of true positives and true negatives with a low relative proportion of false negatives.

Metric	Random Forest	Decision Tree
Accuracy	0.8946	0.8720
Kappa	0.7818	0.7351
AUC (ROC)	0.9640	0.9080
Recall (Risk)	0.9150	0.8840
Recall (No Risk)	0.8660	0.8520

Table 2. Confusion Matrix and Performance Metrics

As shown in Table 2, both models based on academic variables achieved high discriminatory power, with the Random Forest model demonstrating greater discriminatory power than the J48 model. The difference is evident in the overall accuracy, the Kappa value, and the area under the ROC curve. However, the J48 model showed consistent and stable performance, with competitive metrics.

Furthermore, the decision tree generated by J48 had 56 leaves. The variable at the root was the average score in Basic Sciences, indicating its structural role in the segmentation of the "Academic Risk" category. The main branches of the tree show that low scores in Basic Sciences significantly increase the probability of being classified in the "Risk" category.

4.1.2. Ranking of Key Variables (Information Obtained)

Demographic variables showed lower values than those observed in academic areas, according to the univariate ranking of attributes generated using the information-theory-based measure known as Information Gain.

The analysis of variable importance (see Figure 3) reveals a markedly heterogeneous hierarchy, in which Basic Sciences emerges as the dominant predictor with a relative weight of 0.488. Noteworthy is the quantitative gap ($\Delta = 0.268$) observed between this primary factor and the second level of importance, which consists of a competitive group that includes General Education (0.220), Computer Science (0.210), and Applied Engineering (0.200). This discontinuity in the ranking suggests that core competencies have twice the explanatory power of any other curricular category. In contrast, demographic or administrative variables, such as Gender (0.001) and Age at Enrollment (0.000), have marginal or no significance, indicating that the performance or impact analyzed is intrinsically linked to the student's academic trajectory rather than to their initial conditions upon enrollment.

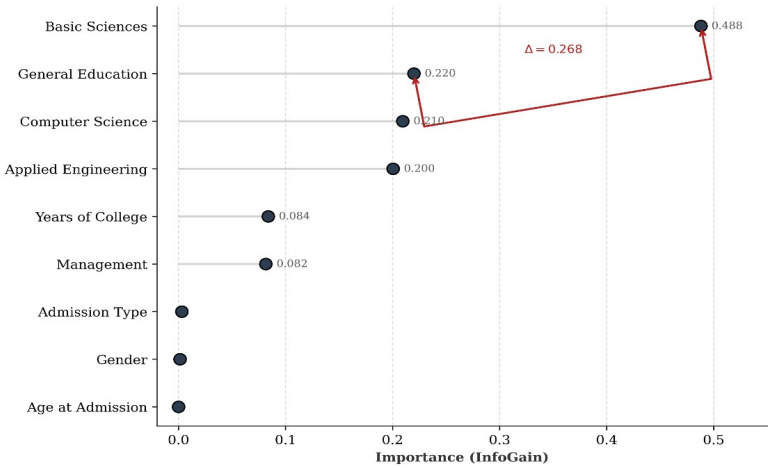

Figure 3. Ranking by Importance – Academic Dimension

4.2. Results of the Socioeconomic Dataset (n = 410)

4.2.1. Random Forest with Cost-Sensitive Classification (5:1)

The experiment achieved an ROC area of 0.733 and a recall of 82.5% for the "Risk" class; the model's validation values are shown in Table 3.

Metrics	Risk Level	Success(Avg/Good)
TP Rate (Recall)	0.825	0.385
FP Rate	0.615	0.175
Precision	0.490	0.754

Table 3. Asymmetric Cost Matrix

Similarly, the confusion matrix showed 141 true positives, 30 false negatives, and 147 false positives.

4.2.2. Ranking of Key Variables (Random Forest)

In the context of Random Forest, variable importance is known as structural weight factors (Feature Importance), which quantify the relative contribution of each predictor variable to reducing impurity or error within the ensemble of trees. In this regard, the importance ranking placed Age (0.41), Sex (0.32), and the Mother's Academic Degree (0.32) as the factors with the highest structural weight.

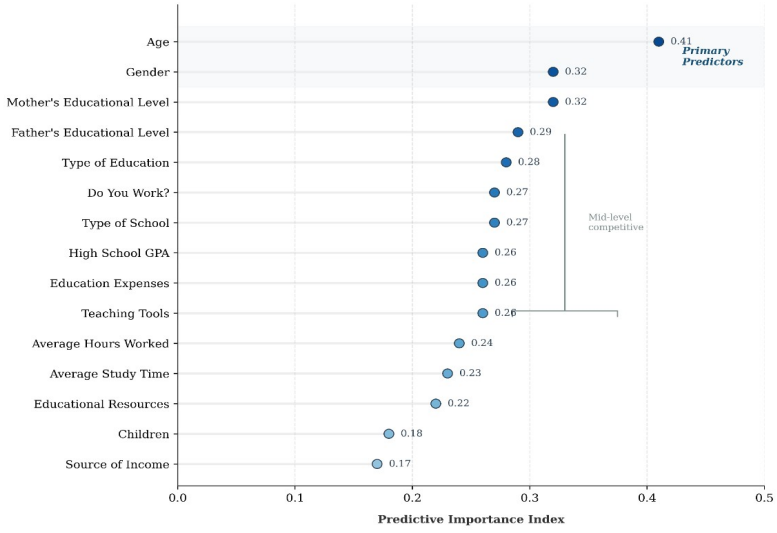

Figure 4. Structural Weight Factors for Random Forest

As shown in Figure 4, the significance profile reveals a model characterized by multifactorial convergence, in which Age (0.41) leads the ranking but without establishing absolute dominance. The parity observed between Gender (0.32) and the Mother's Academic Degree (0.32) suggests that the model relies on a balanced set of demographic and socio-family variables to make its predictions. Unlike distributions with abrupt drops, this set exhibits a gradual decline in weights, where even mid-range variables (such as Mode of Study and Employment Status) maintain a competitive relevance greater than 0.25. This density in the upper block indicates that the phenomenon under analysis is complex in nature and is conditioned by an interconnected network of factors, where no single predictor overshadows the importance of the whole.

4.2.3. Ranking of Key Variables (Information Gain)

Institutional teaching tools and educational resources scored lower than the demographic characteristics, according to the univariate ranking of attributes generated using the information-theory-based measure known as Information Gain.

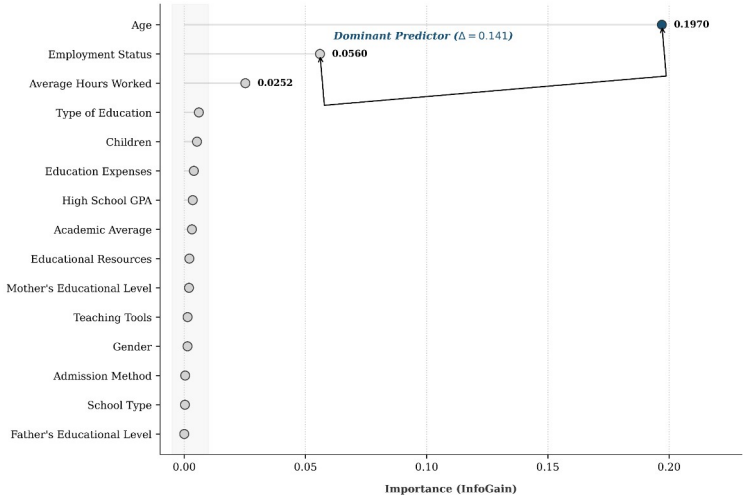

Figure 5. Ranking by Importance – Socioeconomic Dimension

The variable importance analysis (see Figure 5) reveals a hierarchical structure with exceptional dominance of the Age factor (0.197), which exhibits a critical quantitative gap ($\Delta = 0.141$) relative to the second predictor, Employment Status (0.056). This discontinuity suggests that the subject's chronological maturity acts as the gravitational center of the model, possessing an explanatory power nearly four times greater than that of active employment status. Below this threshold, an exponential decline in the relevance of the factors is observed; variables related to the socio-family environment (e.g., fathers' educational attainment) and prior educational history show only marginal importance (<0.01). These results indicate that, in the context analyzed, basic demographic characteristics almost completely overshadow the influence of pedagogical tools and institutional educational resources.

4.3. Unsupervised Characterization Results

4.3.1. Dataset Academic

The SimpleKMeans algorithm with k = 3 generated 3 (three) distinct clusters based on averages by curricular areas and demographic variables. The cluster with the lowest average in Basic Sciences had the highest proportion of students classified as "Risk" by the supervised model.

4.3.2. Socioeconomic Dataset

The clustering identified 3 (three) profiles primarily associated with differences in age and parental educational capital. It is worth noting that the separation between clusters was less pronounced compared to the Academic dataset.

4.4. Performance Comparison Across Information Dimensions

The results discussed here are based on a dual relevance analysis. The Information Gain (IG) algorithm identified Age as the primary predictor. Working Conditions also scored highly in the univariate ranking. The Random Forest model generated a different hierarchical order (see Figure 6). The ranking based on Gini impurity highlighted Sex as a relevant variable. The mother's academic degree showed significant structural importance in the tree ensemble.

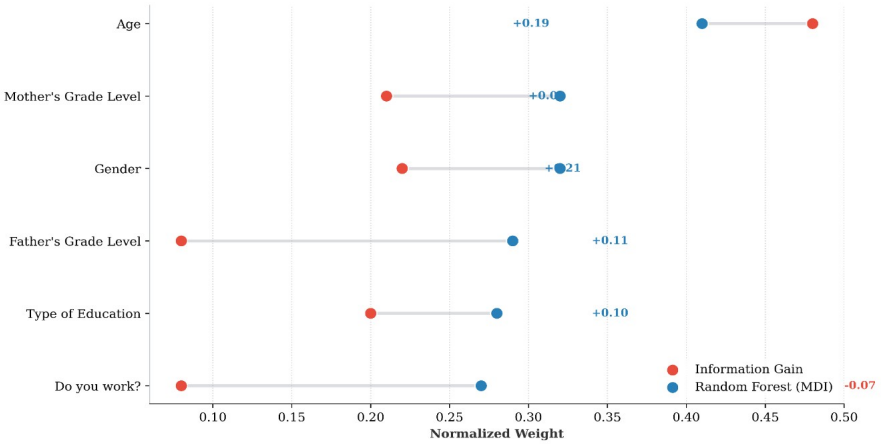

Figure 6. Comparative Importance of Features: InfoGain vs. Random Forest Structural Weights

5. Discussion

This study set out to compare the predictive power of two (2) informational dimensions (i.e., early academic performance and socioeconomic status) in identifying academic risk in engineering programs. An important finding is that the results show consistent differences in discriminatory performance, as well as convergent structural patterns between supervised models and unsupervised analysis.

5.1. Early Academic Excellence

This study found that the model based on first-year academic variables achieved high levels of accuracy and AUC. This result is consistent with the evidence reported in the international literature on Educational Data Mining, where prior performance is typically the most robust predictor of future performance.

Given these results, we can conclude that, in the specific context of engineering education, the centrality of the Basic Sciences area—as a root variable and the attribute with the greatest information gain—reinforces the structural role that mathematics and physics courses play in a student's academic trajectory. This is an interesting finding, as it not only confirms their evaluative weight but also demonstrates their ability to identify risk early on.

Furthermore, the convergence among Random Forest, J48, and Information Gain suggests that the detected pattern is internally stable. Thus, this methodological consistency reinforces the validity of the result regardless of the algorithm used.

5.2. Predictive Power of Socioeconomic Variables

It is perhaps not so surprising that the model built exclusively with socioeconomic variables showed lower overall accuracy. However, by incorporating an asymmetric cost matrix, we were able to increase sensitivity in identifying at-risk students.

These findings suggest that, in general, socioeconomic variables function as structural determinants rather than as direct predictors of academic performance. Factors such as age and parental educational level

emerged as relevant attributes, which can be interpreted in terms of prior educational trajectories and available cultural capital.

Furthermore, the smaller geometric separation observed in the socioeconomic clustering is consistent with the reduction in supervised discriminatory power. Therefore, this consistency between supervised and unsupervised analyses lends structural robustness to the interpretation.

5.3. Sensitivity versus Accuracy in Educational Contexts

The most obvious finding to emerge from this study is that the use of cost-sensitive classification highlights a relevant methodological issue in educational analytics, where overall accuracy is not always the optimal metric when the institutional objective is preventive.

Given that in early warning systems, the critical error is failing to identify a vulnerable student, in this study, differential penalization allowed us to prioritize sensitivity over accuracy, aligning the model's behavior with academic management criteria. In practical terms, this approach highlights the importance of adapting the evaluation strategy to institutional objectives and not limiting the comparison to traditional metrics.

5.4. The Multicausal Approach

Taken together, the discrepancy observed between the methods used to measure the importance of the variables suggests that Risk Academic does not follow a simple linear causal relationship. While the IG prioritizes individual variables with high information content, such as age, the Random Forest highlights bioChart factors and cultural capital factors that act as critical moderators.

In light of these results, it is worth noting specifically that the significance of the mother's educational level indicates that social and family support influences students' resilience in the face of the demands of engineering. These findings are consistent with previous studies that identify socioeconomic background as a catalyst for academic performance, even more so than initial technical aptitude. Furthermore, the inclusion of gender and cultural capital allows the model to capture complex interactions that traditional statistical analyses often overlook.

5.5. Convergence Between Characterization and Prediction

By applying clustering, we were able to verify the structural consistency of the institutionally defined target variable. Thus, in the Academic dataset, the natural clustering largely replicated the segmentation by Risk obtained through supervised classification. Therefore, this convergence suggests that the operational definition of Risk early on captures latent patterns present in the attribute space. On the other hand, in the Socioeconomic dataset, the segmentation was less pronounced, which reinforces the interpretation that these variables represent baseline conditions and are not direct determinants of early performance.

5.6. Contribution of the Study

The main contribution of this study lies in the systematic empirical comparison of informational dimensions within the same institutional engineering context. Unlike studies that indiscriminately combine academic and socioeconomic variables into a single model, this approach allowed us to evaluate the relative discriminatory power of each dimension. Furthermore, the integration of supervised models with unsupervised validation provides a more robust analytical framework for characterizing early academic risk.

Overall, the findings show that early academic risk in engineering is primarily structured around initial academic performance, while socioeconomic variables provide complementary contextual information. The evidence obtained supports the feasibility of implementing institutional early-warning systems based on data analytics with explicit preventive criteria.

5.7. Analytical Limitations

Readers should bear in mind that this study has limitations that must be taken into account. Although it would be interesting to examine this in greater detail, the validation performed is internal and does not include an independent external cohort. It is well known that access to student information is difficult, which is why the Socioeconomic Dataset is smaller than the Academic Dataset; we understand that this may influence the stability of the model.

Furthermore, the definition of the target variable depends on specific institutional thresholds. Different contexts may require adjustments to the operationalization of risk. It is recognized that the direct transferability of the findings to other countries may be affected by these institutional variables. While these limitations do not invalidate the results, it can be inferred that they limit the scope of their generalizability. Consequently, a promising future line of research will involve conducting comparative cross-national analyses to assess the model's adaptability across different academic engineering.

6. Conclusion

This study aimed to comparatively determine the predictive power of early academic variables and socioeconomic variables at the time of admission in identifying academic risk in engineering programs.

The results of this research demonstrate that first-year academic performance is the most robust predictor of early academic risk. In particular, performance in basic science courses emerges as the structural core of segmentation. This finding is consistent across different supervised algorithms and is confirmed by unsupervised analysis.

The second most important finding was that socioeconomic variables have less discriminatory power when considered in isolation. However, attributes such as age and parental education level provide relevant contextual information and improve the model's sensitivity when cost-sensitive classification is incorporated.

From a practical and methodological perspective, the study highlights the importance of selecting metrics aligned with institutional goals. The practical implications of prioritizing sensitivity over overall accuracy are particularly relevant in early warning systems designed to promote student retention.

In light of these results, it could be argued that the main contribution of this work lies in the differentiated evaluation of informational dimensions within a single institutional engineering context, as well as in the integration of supervised and unsupervised analysis to validate the structural coherence of the concept of risk.

It is clear from this study that, in practical terms, the results support the feasibility of implementing predictive models as tools to support academic support strategies. However, these models should be used as probabilistic, non-deterministic instruments.

The authors of this article recommend that research be conducted in the areas listed below.

- Incorporate dynamic longitudinal variables that capture changes in performance over time.
- Evaluate the model's stability in external cohorts.
- Analyze potential algorithmic biases and equity criteria.
- Integrate hybrid models that combine academic and socioeconomic dimensions.

Based on the data, we can conclude that early academic risk in engineering can be characterized and modeled with significant precision using structured institutional data. Finally, the evidence obtained provides an empirical foundation for the development of educational analytics systems aimed at continuous improvement and student retention.

Declaration of Conflicting Interests

The authors declare that they have no potential conflicts of interest regarding the research, authorship, and/or publication of this article.

Funding

This project, INIC01-292, is funded by CONACYT through the PROCIENCIA Program with resources from FONACIDE's Fund for Excellence and Research (FEEI).

Authors' contributions

Gustavo Sosa-Cabrera: conceptualization, methodology, research, literature review, formal analysis, experiments, interpretation of results, writing – original draft, writing – review and editing, supervision.

Rossana Martínez: data curation, data preparation, data processing, interpretation of institutional regulations, and creation of calculated variables based on the regulations of the university from which the data were obtained, acquisition of funds, project management.

Data availability

Data subject to third-party restrictions

Use of Artificial Intelligence

The authors declare that the content of the article has not been developed using Artificial Intelligence.

References

Baidoo-Anu, D., Asamoah, D., Amoako, I., & Mahama, I. (2024). Exploring student perspectives on generative artificial intelligence in higher education learning. *Discover Education, 3*(1), 98. https://doi.org/10.1007/s44217-024-00173-z

Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E., & Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences, 10*(13), 4427. https://doi.org/10.3390/app10134427

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education, 12*(2), 155–187. https://doi.org/10.1007/BF00976194

Bourdieu, P. (2018). The forms of capital. In *The sociology of economic life* (pp. 78-92). Routledge. https://doi.org/10.4324/9780429494338

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Cost-sensitive learning. In *Learning from imbalanced data sets* (pp. 63-78). Springer International Publishing. https://doi.org/10.1007/978-3-319-98074-4_4

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann. https://www.cs.waikato.ac.nz/ml/weka/

Johnson, M. T., & Parker, J. M. (2025). Engineering student success based on performance in first semester foundational courses. *Proceedings of the 2025 ASEE Annual Conference & Exposition*. https://doi.org/10.18260/1-2--56392

Kahu, E. R., & Nelson, K. (2018). Student engagement in the educational interface: understanding the mechanisms of student success. *Higher Education Research & Development, 37*(1), 58–71. https://doi.org/10.1080/07294360.2017.1344197

Lázaro-Alvarez, N., Callejas, Z., & 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 (JOTSE), 10*(2), 241-258. https://doi.org/10.3926/jotse.922

Maquen-Niño, G., Miguel-Flores, M., Aurich-Mio, L., Adrianzén-Olano, I., de la Cruz-Vélez, P., & Castro-Cárdenas, D. (2025). Machine learning model for classifying high school students' academic performance in mathematics amidst the COVID-19 context. *Journal of Technology and Science Education, 15*(2), 322-334. https://doi.org/10.3926/jotse.2945

Martínez, R., & Sosa-Cabrera, G. (2024). Explorando el rendimiento académico: Un análisis basado en áreas de conocimiento. In *Libro de Actas del XXX Congreso Argentino de Ciencias de la Computación (CACIC)* (pp. 603–607). Universidad Nacional de La Plata. https://doi.org/10.35537/10915/172755

Mohamed-Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences, 12*(19), 9467. https://doi.org/10.3390/app12199467

Radwan, A. M., & Cataltepe, Z. (2017). Improving performance prediction on education data with noise and class imbalance. *Intelligent Automation & Soft Computing*, *24*(4), 777–783 . https://doi.org/10.1080/10798587.2017.1337673

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Segal, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley interdisciplinary reviews: Data mining and knowledge discovery, 1*(1), 80-87. https://doi.org/10.1002/widm.12

Sholeh, M., & Nurnawati, E. K. (2024). Comparison of Z-score, min-max, and no normalization methods using support vector machine algorithm to predict student's timely graduation. In *AIP Conference Proceedings* (Vol. 3077, No. 1, p. 040003). AIP Publishing LLC. https://doi.org/10.1063/5.0202505

Sosa-Cabrera, G., Gómez-Guerrero, S., García-Torres, M., & Schaerer, C. E. (2024). Feature selection: A perspective on inter-attribute cooperation. *International Journal of Data Science and Analytics, 17*(2), 139-151. https://doi.org/10.1007/s41060-023-00439-z

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). University of Chicago Press.

Veenstra, C. P., Dey, E. L., & Herrin, G. D. (2009). A model for freshman engineering retention. *Advances in Engineering Education, 1*(3), n3.

Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research, 57*(3), 547–570. https://doi.org/10.1177/0735633118757015

Published by OmniaScience (www.omniascience.com)

Journal of Technology and Science Education, 2026 (www.jotse.org)

Article's contents are provided on an Attribution-Non Commercial 4.0 Creative commons International License. Readers are allowed to copy, distribute and communicate article's contents, provided the author's and JOTSE journal's names are included. It must not be used for commercial purposes. To see the complete licence contents, please visit https://creativecommons.org/licenses/by-nc/4.0/.