Who Fails and Why: An Analysis of Student Trajectories and the Prediction of Undergraduate Performance in Programming Courses

Rodrigo Gutiérrez-Benítez , Andrea Vásquez-Guerra , José Luis Carrasco-Sáez *

Posted Date: 23 March 2026

doi: 10.20944/preprints202603.1704.v1

Keywords: introductory programming (CS1); higher education; academic trajectories; early risk prediction; educational data mining; machine learning; at-risk students

 doi:10.20944/preprints202603.1704.v1

Article

Who Fails and Why: An Analysis of Student Trajectories and the Prediction of Undergraduate Performance in Programming Courses

Rodrigo Gutiérrez-Benítez [1], Andrea Vásquez-Guerra [2] and José Luis Carrasco-Sáez [1,*]

[1] Departamento de Electrónica e Informática, Universidad Técnica Federico Santa María, Concepción 4030000, Chile

[2] Departamento de Informática, Universidad Técnica Federico Santa María, Valparaiso 2390123

* Correspondence: jose.carrascosa@usm.cl

Featured Application

The proposed two-stage early warning framework can be applied in introductory programming courses to identify students at risk of failure before the semester begins and after the first major written exam (C1), thereby supporting timely interventions such as leveling programs, tutoring, and targeted academic counseling.

Abstract

This study examines failure in introductory programming courses (CS1) in Chilean higher education by combining an analysis of academic trajectories with early-risk prediction models. We analyzed a cohort of 994 students from a Chilean technical university (2025–1), with a 46% failure rate, integrating pre-university academic and admission variables (e.g., mathematics and language indicators, as well as baseline diagnostic measures when available), sociodemographic information, and within-semester performance indicators. Group differences were assessed using non-parametric tests, and predictive performance was evaluated under two realistic information-availability scenarios: (i) pre-university variables only and (ii) variables available up to the first major written examination (C1). The results show statistically significant differences between students who passed and those who failed, with indicators of quantitative preparedness and, most notably, C1 performance emerging as the strongest signals of risk. In the pre-university scenario, models achieved acceptable discrimination (AUC ≈ 0.77), whereas incorporating C1 substantially improved discriminative performance (AUC ≈ 0.92) and increased precision in identifying at-risk students while reducing false positives. These findings support a staged institutional strategy: broad, low-cost preventive support before the semester begins, followed by more targeted and intensive interventions after C1, thereby enabling more efficient early-warning systems in high-stakes first-year courses.

Keywords: introductory programming (CS1); higher education; academic trajectories; early risk prediction; educational data mining; machine learning; at-risk students

1. Introduction

In Chilean higher education, undergraduate programs face persistently high failure rates in gateway courses such as Introductory Programming (CS1) in Computer Science curricula, which in some universities exceed 50%. Student failure in CS1 is not an isolated outcome but a multifactorial phenomenon shaped by academic preparation, socio-educational conditions, and patterns of engagement with institutional learning environments (e.g., Learning Management Systems, LMS). Given this context, early identification of at-risk students has become a priority for institutions seeking to implement timely support and reduce adverse academic outcomes. This urgency is driven

 doi:10.20944/preprints202603.1704.v1

not only by the economic costs associated with attrition—estimated at USD 1.169 billion between 2017 and 2023 in Chile [1]—but also by the loss of human capital that dropout may entail.

Despite this urgency, traditional prediction approaches often fail to provide sufficiently early or actionable signals to enable effective interventions [2]. As a result, Educational Data Mining (EDM) and Machine Learning (ML) have become central to leveraging historical and early-course data to anticipate academic risk [3–5]. While prior studies have shown that ML models can predict CS1 performance with substantial accuracy, two challenges remain: (i) identifying which academic variables provide the strongest and most timely signals of failure, and (ii) assessing their predictive value under realistic early-warning windows, particularly when risk emerges through non-linear interactions among heterogeneous predictors [6].

This study addresses these gaps by: (1) statistically characterizing the academic profiles and trajectories associated with CS1 failure, and (2) conducting a rigorous comparison of predictive models under two early-intervention scenarios: before the course begins (pre-university information) and after the first major written examination (C1).

The study was guided by the following hypotheses:

Hypothesis 1 (H1): Students who fail CS1 exhibit academic trajectories characterized by lower prior performance, a heavier academic workload when taking programming, and greater lag relative to their incoming cohort, compared to students who pass.

Hypothesis 2 (H2): ML models can predict the risk of CS1 failure with meaningful recall—exceeding a simple baseline—using historical academic variables available at early stages of the semester.

The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 describes the methodology and experimental design; Section 4 reports results for H1 and H2; Section 5 discusses the findings; and Section 6 presents conclusions and directions for future work.

2. Related Work

The prediction of academic performance in Computer Science–related courses, including Introductory Programming (CS1), has been widely studied using Data Mining (DM) and Machine Learning (ML) methods. Broadly, the literature can be organized into two methodological streams: (i) traditional ML classifiers, which often provide robust baselines with relatively high interpretability and low computational cost, and (ii) Deep Learning (DL) architectures, which aim to capture complex non-linear patterns in student data and longitudinal academic trajectories.

2.1. Machine Learning Models

A substantial body of research has validated classical ML models for early warning systems. The systematic literature review (SLR) by Brito Correia et al. [7] reports that such models are frequently adopted due to their interpretability and lower computational demands. Among these approaches, Decision Trees (DT) and their variants remain popular; for example, Khan et al. [8] showed that the J48 algorithm can classify students into risk categories (e.g., red, yellow, green) to support instructional interventions. Similarly, Gollapalli et al. [9] reported that probabilistic models such as Naïve Bayes (NB) achieved up to 90.54% accuracy in a training-related context, outperforming other simple classifiers. Consistent with these findings, Support Vector Machines (SVM) and NB have also been reported as strong classifiers for identifying at-risk students in programming courses [10,11].

Beyond single-model approaches, recent evidence suggests that ensemble techniques often yield improved generalization. Xuan Lam et al. [12] compared multiple classifiers (including Random Forest, SVM, and k-Nearest Neighbors) and found that a stacking-based ensemble consistently outperformed individual models by effectively combining complementary predictive signals.

 doi:10.20944/preprints202603.1704.v1

2.2. Deep Learning Models

While traditional ML provides a strong benchmark, recent studies indicate that DL architectures may achieve superior performance when modeling complex student behaviors and academic trajectories. Brdesee et al. [5] applied Long Short-Term Memory (LSTM) networks and addressed class imbalance using SMOTE, reporting improved performance for early prediction of retention and graduation risks. Importantly, the benefits of neural networks often emerge in metrics beyond overall accuracy. Rodríguez-Hernández et al. [4], using a large dataset, found that Artificial Neural Networks (ANNs) outperformed traditional ML models particularly in recall and F1-score—metrics that are critical in early warning contexts where minimizing false negatives is a priority. Similarly, Ilić et al. [13] showed that combining ANNs trained with the Levenberg–Marquardt algorithm with feature filtering techniques (e.g., minimum redundancy–maximum relevance) can improve predictive accuracy in programming tutoring systems.

2.3. Variables and Data Sources

The performance of both ML and DL models is strongly influenced by the nature and timing of input features. Liao et al. [3] emphasized that the predictive value of data sources varies across contexts: prerequisite grades may be more predictive in upper-division courses, whereas early engagement signals and immediate feedback can be more informative in CS1. This is consistent with An et al. [2], who argued that granular behavioral metrics (e.g., work habits and submission timing) can provide deeper insights than interim grades alone. Mai et al. [14] further advanced this line of work by applying community detection to behavioral data to identify groups of students with similar learning patterns, improving outcome prediction.

As models become more complex—particularly with DL—the need for interpretability has intensified. Choi et al. [15] reviewed Explainable Artificial Intelligence (XAI) in education and highlighted methods such as SHAP (SHapley Additive exPlanations) as essential for moving beyond black-box predictions toward actionable pedagogical insights. This interpretability requirement remains central when comparing high-performing DL approaches against more transparent traditional models [16].

Overall, the literature supports three conclusions: (i) traditional ML models provide strong baselines for early warning, (ii) ensembles and DL models can better capture non-linear patterns and often improve sensitivity-oriented metrics such as recall, particularly when behavioral data are available, and (iii) a practical gap persists between predictive accuracy and prediction timeliness, partly due to feature availability early in the semester and limited clarity on which historical variables yield actionable early warnings.

This study addresses these limitations in the Chilean higher education context by statistically characterizing at-risk students and rigorously comparing early-stage predictive models under realistic intervention windows. Ultimately, the goal is to support timely pedagogical interventions that may reduce dropout risk and its associated social and economic costs.

3. Materials and Methods

The overall methodological workflow is summarized in Figure 1 and described in the remainder of this section.

 doi:10.20944/preprints202603.1704.v1

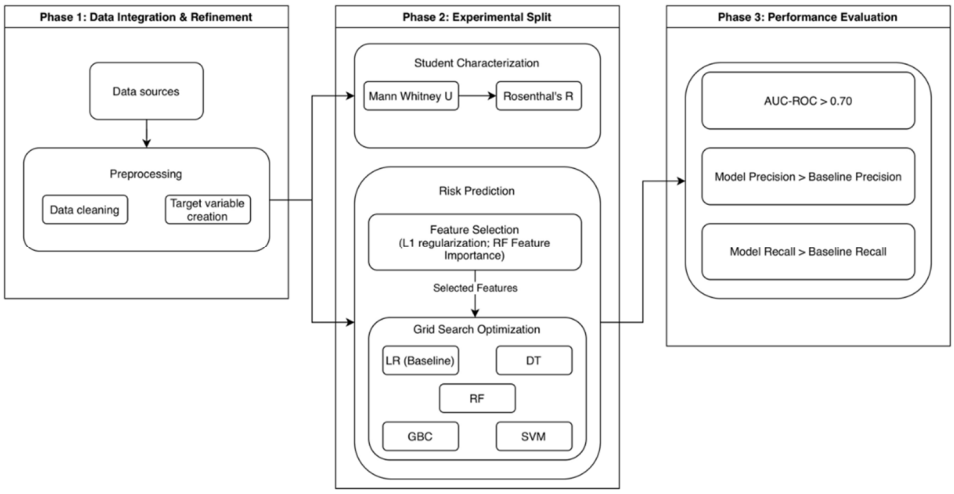

Figure 1. Methodological workflow followed in the study for each prediction scenario.

3.1. Data Collection and Preprocessing

3.1.1. Data Collection

All the data used in this study correspond to a cohort of 994 students from Universidad Técnica Federico Santa María (USM), Chile, enrolled in the CS1 programming course during the first semester of 2025. The course exhibits a 46% failure rate for this cohort. CS1 is a mandatory first-year course delivered at scale (typically over 1,200 students per semester) over a 17-week semester.

The assessment framework comprises multiple graded components: written examinations (PC), in-class quizzes (PCt), assignments (PT), formative assessments (Pef), self-graded assessments (Psm), and teaching-assistant-led quizzes, which are incorporated into the in-class quiz component through a weighted average. In this course, PC is computed as the arithmetic mean of the two major written exams (C1 and C2). Course completion requires a final grade ≥ 55 on a 1–100 scale, computed as:

$$\boldsymbol{Final\ Grade = PC * 0.5 + PCt * 0.2 + Pef * 0.05 + PT * 0.2 + Psm * 0.05}$$

A key milestone for this study is the first major written exam (C1), typically administered around week 8 of the semester. Because PC aggregates C1 and C2, C1 represents the earliest high-stakes summative checkpoint that meaningfully constrains final performance and enables realistic early-intervention decisions. This timing motivates the evaluation of two information-availability scenarios: a prediction based solely on pre-university information (before the semester starts) and a prediction incorporating performance information available up to C1.

To construct a dataset that reflects students' academic trajectories more comprehensively, course performance variables were complemented with demographic information and pre-university academic history. In this study, "pre-university" variables refer to information available before the beginning of the semester and prior to C1, including admission scores (e.g., mathematics, language, and science components, as well as school-based indicators such as NEM and ranking) and institutional baseline measures when available (e.g., an entry diagnostic test). These variables are analyzed both descriptively (pass vs. fail comparisons) and within the predictive scenario that simulates preventive intervention before the course begins.

Admission variables are derived from national university selection scores and school records. Specifically, Mathematics M1 corresponds to the common mathematics component, whereas Mathematics M2 corresponds to the advanced mathematics component. Some components may not be available for all students (e.g., History may be optional depending on the admission pathway), which must be considered when interpreting availability and predictive contribution. All admission scores are reported on a standardized scale with a maximum of 1000 points, facilitating interpretation of distributions and group comparisons.

 doi:10.20944/preprints202603.1704.v1

For transparency and replicability regarding the sequence and timing of assessments throughout the semester, the course assessment timeline is consistent with prior institutional descriptions reported in Vásquez et al. [17].

3.1.2. Preprocessing

An exploratory analysis was conducted to summarize distributions, detect atypical values, identify potential data-quality issues, and characterize the main features of the dataset. This stage included (i) data preparation and consistency checks, (ii) univariate exploration of key variables, (iii) bivariate analyses to inspect relationships among academic background and course performance indicators, and (iv) assessment of missingness patterns to determine variable availability.

Following exploratory analysis, the data were cleaned and structured for subsequent descriptive and predictive analyses. This process included variable harmonization across sources, detection and correction of invalid or out-of-range values, and the treatment of missing data. Variables exhibiting high levels of missingness and limited coverage were excluded from modeling to avoid instability, whereas remaining missingness was handled using standard procedures appropriate to each model and analysis step. Finally, the dependent variable Pass was defined from the final grade computed in Eq. (1): students with a final grade ≥ 55 were labeled as Pass (1), and those with a final grade < 55 were labeled as Fail (0).

3.2. Experimental Design

The preprocessed dataset was split into three mutually exclusive subsets: training (70%), validation (15%), and test (15%). Model development—including feature selection and hyperparameter tuning—was performed using the training and validation subsets only. Hyperparameters were optimized via grid search with stratified cross-validation conducted on the training data to improve robustness and mitigate sampling variance. The validation set was used to select the best configuration after tuning, and final generalization performance was reported on the independent test set, which remained unseen throughout model selection.

3.2.1. Characterization of Students (E.H1)

To address H1, we conducted two complementary analyses comparing students in the Pass and Fail groups. First, we performed group comparisons using the Mann–Whitney U test, interpreting statistical significance under a conservative threshold of $p < 0.01$, to assess whether distributions differed significantly between the two groups. Second, we quantified the magnitude of these differences using Rosenthal's r effect size, enabling interpretation beyond statistical significance.

3.2.2. Risk Prediction (E.H2)

To address H2, we developed and evaluated classification models under two realistic information-availability scenarios aligned with the academic timeline: (i) before the course begins and (ii) after the first major written exam (C1). This design supports early-warning use cases by explicitly linking model inputs to feasible intervention windows.

3.2.2.1. Feature Selection

Feature selection was conducted to identify informative and non-redundant predictors and to assess potential multicollinearity. We used two complementary methods to capture both linear and non-linear relevance:

Linear importance: Logistic Regression (LR) with L1 regularization was used to estimate linear predictive relevance while shrinking redundant coefficients toward zero.

Non-linear importance: Random Forest (RF) feature importance was used to capture non-linear relationships and interaction effects beyond linear models.

 doi:10.20944/preprints202603.1704.v1

The final feature set for model training was derived from predictors consistently identified as relevant across both methods.

3.2.3. Prediction Scenarios

To evaluate predictive performance and practical feasibility, we defined two data-availability scenarios based on the student academic timeline, allowing identification of the earliest actionable intervention point:

Pre-University scenario: includes only features available before university entry (e.g., admission and pre-university academic history).

Information up to C1 scenario: includes features available up to and including C1, simulating a realistic post-C1 intervention window and enabling assessment of early within-semester signals.

For both scenarios, hyperparameter optimization was performed via grid search with stratified cross-validation, restricted to training/validation data to prevent leakage. The final performance metrics were computed on the independent test set, which remained unseen during feature selection and tuning. Table 1 summarizes the candidate models and the hyperparameter grids evaluated.

Table 1. Selected models and their hyperparameters for prediction scenarios experiments.

Model	Hyperparameters
Logistic Regression (baseline)	max_iter = 1000
Decision Tree	max_depth = [5, 10, None]; min_samples_leaf = [1, 50, 100]
Random Forest	n_estimators = [50, 100, 150, 200]; max_depth = [5, 10, None]
Gradient Boosting	n_estimators = [50, 100, 150, 200, 250]; max_depth = [3, 5, 10]; learning_rate = [0.01, 0.05, 0.1]
Support Vector Machine	C = [0.001, 0.01, 0.1, 1, 10, 100]; kernel = [linear, rbf]; gamma = [scale, auto]

3.3. Evaluation Metrics

Evaluation of H2 relied on three complementary performance criteria designed to balance early-warning usefulness with institutional feasibility. In this study, the baseline model corresponds to a logistic regression classifier, and all reported gains are interpreted relative to its held-out performance.

Discriminative power (AUC–ROC): the selected model must achieve an AUC–ROC > 0.70, indicating discrimination substantially above chance across classification thresholds. AUC–ROC summarizes the model's ability to rank students by estimated risk of failure independent of a specific decision threshold.

Incremental precision (control of false positives): the selected model must improve precision for the Fail class relative to a defined baseline by at least 5 percentage points. In operational terms, this criterion reduces the number of false positives (FP)—students incorrectly flagged as at risk—thereby helping to prevent inefficient allocation of tutoring and support resources.

Risk sensitivity (recall and F1-score): the model must outperform the baseline in recall for the Fail class, minimizing false negatives (FN) (i.e., at-risk students who remain undetected). Because precision and recall trade off, model selection also considered a balanced F1-score to avoid improvements driven by extreme threshold choices.

Overall, the optimal model was chosen to minimize the expected institutional cost associated with FN (missing students who require support) while maintaining an FP volume that remains feasible for the available tutoring capacity.

 doi:10.20944/preprints202603.1704.v1

4. Results

4.1. Exploratory Data Analysis

4.1.1. Data Preprocessing and Missing Value Imputation

Initial data profiling indicated high completeness for most demographic and institutional variables. Specifically, the missingness rate was approximately 0.30% for features such as Gender, Region, Major, and School Funding Type. In contrast, the Lawson Pre assessment exhibited substantial missingness (63.48%). Given this level of sparsity and the risk of unstable estimates if imputed, Lawson Pre was excluded from downstream analyses.

For the remaining variables with minor missingness, we applied Multivariate Imputation by Chained Equations (MICE) to preserve multivariate relationships among predictors. To avoid information leakage, imputation was performed as part of the modeling pipeline using only the training data, and the fitted imputation model was subsequently applied to validation and test splits.

4.1.2. Institutional and Demographic Analysis

Figure 2 summarizes pass rates by school funding type and gender. Overall, students from private-funded and subsidized private schools show higher pass rates than students from municipal schools. Differences across funding categories suggest systematic variation in CS1 outcomes aligned with prior evidence on socio-educational stratification; however, these comparisons are descriptive and should be interpreted with caution. Notably, the Local Education Service category exhibits a visible gender gap, with higher pass rates for female students than for male students. The wider confidence intervals for this category indicate a smaller sample size, limiting the strength of any subgroup-level interpretation.

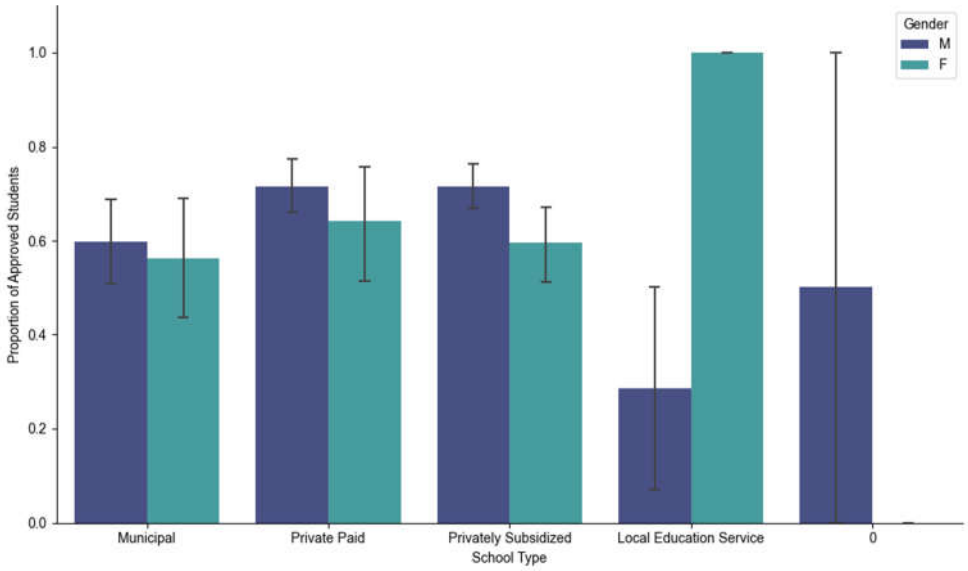

Figure 2. Proportion of approved students by gender and school funding type.

4.1.3. Entrance Scores and Academic Progression

The distribution of weighted entrance scores is concentrated around ~700 points (on a 0–1000 scale) and shows substantial overlap between students who eventually Pass (1) and those who Fail (0). As illustrated by the probability density estimates (Figure 3), the two distributions are highly overlapping, suggesting that admission scores, while informative for selection, provide limited discriminative value for identifying individual failure risk once students are enrolled.

 doi:10.20944/preprints202603.1704.v1

In contrast, within-semester academic progression offers clearer early signals of risk. Analysis across assessment cycles (Figure 4) shows that students who ultimately fail tend to exhibit lower median performance from the earliest milestones. The first major written exam (C1) presents the most pronounced separation: most students in the failing group score below the passing threshold (55 points), whereas the passing group displays a distribution shifted toward higher values. Together, these patterns motivate the inclusion of early-course performance indicators—particularly C1—as critical inputs for early-warning prediction.

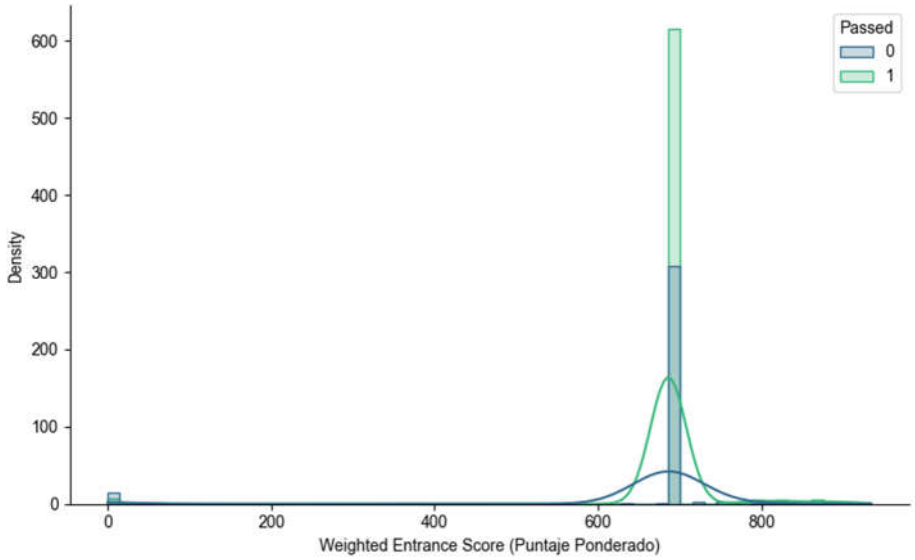

Figure 3. Probability Density of Entrance Scores.

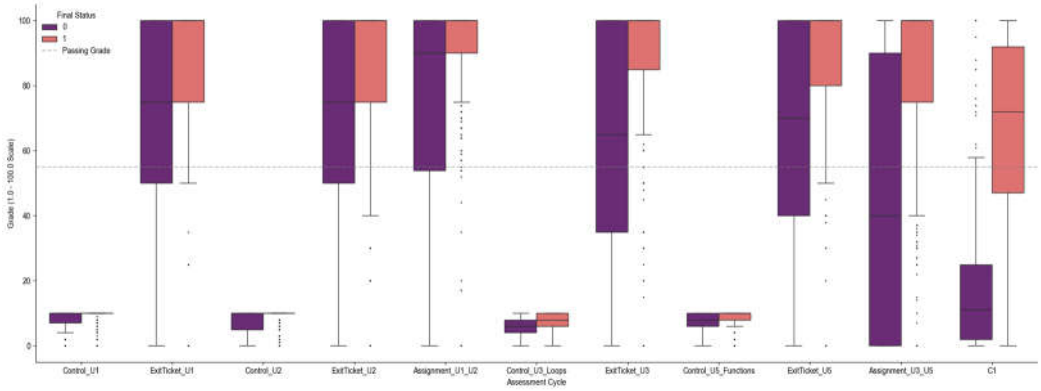

Figure 4. Grade Progression Across Semester Milestones.

4.1.4. Multicollinearity and Variable Selection

To assess the extent of redundant linear information among predictors and to improve the stability of model estimation, we conducted a multicollinearity analysis using the Variance Inflation Factor (VIF). As shown in Table 2, several admission-related variables exhibited severe multicollinearity, with VIF values far exceeding the conventional threshold of 10. In particular, NEM (VIF ≈ 2146), Ranking_Score (VIF ≈ 1618), and MATH_M1_Score (VIF ≈ 200) displayed extreme linear dependence, indicating that these predictors carry highly overlapping information.

This pattern is expected because the Weighted Entrance Score is computed as a linear combination of multiple admission components. Consequently, feature selection was warranted to reduce redundancy and mitigate coefficient instability in linear models, while retaining a set of representative predictors for downstream modeling. In contrast, within-semester assessment

 doi:10.20944/preprints202603.1704.v1

variables (e.g., ExitTicket_U2, ExitTicket_U3, and C1) showed VIF values near or below 10, supporting their use as distinct temporal markers of academic performance within the predictive framework.

Table 2. Variance inflation factors (VIF) for selected predictors.

Feature	VIF
NEM	2146.08
NEM_Score	1951.02
Ranking_Score	1618.65
Graduation_Year	1256.42
Math_M1_Score	200.53
Language_Score	98.81
Math_M2_Score	65.79
Final_Grade	61.58
Weighted_Score	40.96
Entrance_Test_Score	20.39
Control_U1	15.68
Science_Score	13.85
Control_U2	11.32
Control_U5_Functions	10.94
Approved	10.38
EXITTICKET_U1	10.05
ASSIGNMENT_U1_U2	10.05
C1	9.80
Control_U3_Cycles	9.50
LAWSON PRE	8.32
EXITTICKET_U5	8.14
EXITTICKET_U2	7.79
EXITTICKET_U3	7.76
ASSIGNMENT_U3_U5	4.26
History_Score	2.28

4.1.4. Normality Testing

To inform the choice of statistical procedures for group comparisons, we assessed distributional properties for key continuous variables, with emphasis on the first major written exam (C1). Normality was evaluated using the Shapiro–Wilk test and complemented with visual inspection given the large sample size. For C1, the Shapiro–Wilk test indicated a departure from normality ($W = 0.912$, $p < 0.001$), which is consistent with the asymmetric distributions and the presence of outliers observed in the grade progression across assessment milestones (Figure 4).

Because the outcome variable (Pass/Fail) is binary, normality testing is not applicable to the target itself. Accordingly, non-parametric methods were prioritized for inferential comparisons between Pass and Fail groups, and predictive modeling relied on classifiers that do not assume normally distributed predictors. Model evaluation emphasized threshold-independent discrimination (AUC–ROC) and sensitivity-oriented metrics (recall and F1-score) to support early-warning use cases.

4.2. Experiment 1: Characterization of Students

The comparative analysis between students in the Pass and Fail groups revealed statistically significant differences across most evaluated dimensions. The main results are summarized below and reported in detail in Table 3, which includes both Mann–Whitney U test results and Rosenthal's r effect sizes.

 doi:10.20944/preprints202603.1704.v1

Table 3. Characterization of differences between Pass and Fail groups using the Mann–Whitney U test and Rosenthal's r effect size. Features are sorted in descending order by Rosenthal's r. Effect size categories follow conventional thresholds: small (≥ 0.10), medium (≥ 0.30), and large (≥ 0.50).

Feature	U	P_Value	Rosenthal_r	Effect Size
C1	202585.5	< 0.01	0.6802	Large
Puntaje_Math_M2_Score	169294.5	< 0.01	0.4337	Medium
Puntaje_Math_M1_Score	167662.5	< 0.01	0.4216	Medium
ASSIGNMENT_U3_U5	163577.0	< 0.01	0.3914	Medium
EXITTICKET_U5	161656.5	< 0.01	0.3772	Medium
Science_Score	159730.5	< 0.01	0.3629	Medium
Entrance_Test_Score	155856.5	< 0.01	0.3342	Medium
EXITTICKET_U3	155235.5	< 0.01	0.3296	Medium
ASSIGNMENT_U1_U2	150044.5	< 0.01	0.2912	Small
Control_U3_Cycles	148046.0	< 0.01	0.2764	Small
Language_Score	147549.0	< 0.01	0.2727	Small
Control_U5_Functions	146964.5	< 0.01	0.2684	Small
EXITTICKET_U2	145771.5	> 0.01	0.2596	Small

4.2.1. Academic and Admission Background

Several admission-related variables showed statistically significant differences in favor of the Pass group. In particular, indicators of quantitative preparation exhibited the strongest separations, with Mathematics M2 (r = 0.4337) and Mathematics M1 (r = 0.4216) showing medium effect sizes, followed by Science_Score (r = 0.3629). Language_Score (r = 0.2727) also differed significantly between groups, although with a small effect size. Overall, these results indicate that pre-university preparation—especially in mathematics and science—meaningfully differentiates students who eventually pass from those who fail in CS1.

4.2.2. In-course Performance

Within-semester assessments yielded the clearest separation between groups. The first major written examination (C1) exhibited the largest effect size in the entire analysis (r = 0.6802, p < 0.01), indicating the strongest group separation among all evaluated variables. This result is consistent with the central role of C1 as the first high-stakes summative checkpoint in the course.

Other in-course assessments also showed meaningful differences. In particular, ASSIGNMENT_U3_U5 (r = 0.3914), EXITTICKET_U5 (r = 0.3772), and EXITTICKET_U3 (r = 0.3296) displayed medium effect sizes, whereas ASSIGNMENT_U1_U2, Control_U3_Cycles, and Control_U5_Functions showed small effects. Taken together, these patterns suggest that stronger and more consistent academic performance during the early stages of the semester is associated with a higher likelihood of passing CS1. Although EXITTICKET_U2 showed a small effect size estimate, it did not meet the predefined significance threshold and should therefore be interpreted with caution.

4.2.3. Sociodemographic Factors and Routes of Entry

Several sociodemographic and entry-route variables were also associated with course outcomes, although their differences were notably smaller than those observed for academic variables. Students from the Metropolitan Region were more represented in the Pass group, whereas students from the Valparaíso Region and those admitted through the PACE pathway were more represented in the Fail group.

Differences were also observed by major, with students in Computer Engineering showing a slight tendency toward success, whereas students in Metallurgical Engineering and Mechanical Engineering were somewhat more represented in the Fail group. These results should be interpreted

 doi:10.20944/preprints202603.1704.v1

as descriptive associations within this cohort, and they appear to provide weaker practical differentiation than academic performance indicators.

4.3. Experiment 2

4.3.1. Pre-University Scenario

The feature selection process was conducted to assess the relative contribution of candidate predictors and to guide the subsequent modeling phase under the pre-university scenario. Using a dual approach—Logistic Regression with L1 regularization to capture linear relevance and Random Forest (RF) feature importance to capture non-linear effects and interactions—we identified a set of core predictors available prior to course start.

As reported in Table 4, Mathematics M1 emerged as the most consistently prioritized predictor across both approaches, showing the largest coefficient among retained predictors in the L1-regularized logistic model ($\beta = 1.574$) and the highest RF importance score (0.124). This agreement indicates that quantitative preparedness measured at admission provides a strong early signal associated with CS1 outcomes.

A notable pattern is the sparsity induced by L1 regularization: several coefficients were shrunk to zero (e.g., Science and NEM-related indicators), whereas RF assigned these variables non-trivial importance (approximately 0.058–0.119). This divergence suggests that some admission and school-trajectory variables may contribute limited independent linear signal, yet still add predictive value through non-linear relationships and multivariate dependencies captured by tree-based models. Finally, high-school graduation year received a negative coefficient in the linear model ($\beta = -0.593$), indicating that a longer gap between graduation and university entry is associated with a lower probability of passing, holding other variables constant.

Table 4. Top 15 selected features for the pre-university scenario using L1 and RF feature importance, sorted by RF feature importance in descending order.

Feature	Importance L1	Importance Random Forest
Math_M1_Score	1.574	0.124
Math_M2_Score	0.000	0.119
Science_Score	0.000	0.116
Language_Score	0.000	0.076
Ranking_Score	0.000	0.071
NEM_Score	0.000	0.061
NEM	0.000	0.058
Entrance_Test_Score	0.030	0.058
History_Score	0.000	0.039
Degree_Civil Engineering in Computer Science	0.561	0.019
Region_METROPOLITAN REGION OF SANTIAGO	0.182	0.016
Graduation_year	-0.593	0.015
Region_Valparaiso Region	-0.344	0.014
Degree_Civil Engineering Common Plan	0.000	0.013
Degree_Civil Mechanical Engineering	0.000	0.012

Candidate models were tuned via grid search with stratified cross-validation on the training data. Table 5 summarizes the best cross-validated AUC achieved during training and the corresponding Fail-class precision on the validation set. Although SVM obtained the highest cross-validated training AUC (0.8244), Gradient Boosting (GBC) was selected because it achieved the best validation precision for the Fail class (0.5682). This selection criterion prioritizes reducing false

 doi:10.20944/preprints202603.1704.v1

positives in an early-warning setting, where over-flagging students may lead to inefficient allocation of tutoring resources.

Table 5. Model optimization results and validation performance for the pre-university scenario.

Model	Best CV AUC (Training)	Precision (Failed)—Validation
SVM	**0.8244**	0.5088
RF	0.8201	0.5600
GBC	0.8127	**0.5682**
DT	0.7583	0.4688

The selected GBC configuration used a learning rate of 0.05, max depth of 3, and 100 estimators. To address class imbalance, SMOTE oversampling was applied within the training process only (i.e., fitted on the training split/folds), preventing information leakage into validation or test data.

Final generalization performance was evaluated on an independent test set (n = 149) and compared against a logistic regression baseline (Table 6). Across models, discrimination remained acceptable (AUC-ROC in the range 0.766–0.789), with Random Forest achieving the highest AUC (0.789). In line with the operational selection criterion, GBC delivered the highest Fail-class precision on the test set (0.638), improving upon the baseline precision (0.577) by 6.1 percentage points, while maintaining a recall of 0.60 and an F1-score of 0.619. These results suggest that, under pre-university information alone, the selected model can improve targeting efficiency (fewer false positives at comparable sensitivity) without materially reducing the ability to identify at-risk students.

Table 6. Classification metrics summary for the pre-university scenario (test set).

Model	AUC-ROC	Precision (Failed)	Recall (Failed)	F1 (Failed)
Logistic Regression (baseline)	0.758	0.577	0.60	0.588
Decision Tree	0.774	0.574	**0.70**	**0.631**
Random Forest	**0.789**	0.612	0.60	0.606
Gradient Boosting	0.766	**0.638**	0.60	0.619
Support Vector Machine	0.767	0.593	0.64	0.615

4.3.2. Information up to C1 Scenario

The feature selection process in this scenario was designed to assess how the inclusion of early university performance data affects predictive signal strength. Using the same dual approach—Logistic Regression with L1 regularization for linear relevance and Random Forest (RF) feature importance for non-linear effects—we identified the core predictors available up to the first major written examination (C1).

As shown in Table 7, C1 emerged as the most dominant feature across both methods, showing the largest retained coefficient in the L1-regularized logistic model ($\beta = 4.329$) and the highest RF importance score (0.172). This pattern indicates that early in-course performance provides a substantially stronger signal for final outcomes than pre-university indicators alone.

A notable result in this scenario is the pronounced sparsity induced by L1 regularization. The L1 model shrank most coefficients to zero, including all standardized test scores (e.g., Science, Math M1, and Math M2), while retaining only C1 and ASSIGNMENT_U3_U5 ($\beta = 0.444$). In contrast, RF assigned non-trivial importance to a broader set of variables, including Science_Score (0.069) and early in-course indicators such as EXITTICKET_U3 and EXITTICKET_U5. This divergence suggests that, although C1 captures the strongest direct linear relationship with the outcome, pre-university background and other early-course signals still contribute meaningful predictive value through non-linear interactions and multivariate dependencies.

 doi:10.20944/preprints202603.1704.v1

Table 7. Top 15 selected features for the information up to C1 scenario using L1 and RF feature importance, sorted by RF feature importance in descending order.

Feature	Importance L1	Importance Random Forest
C1	4.329	0.172
Science_Score	0.000	0.069
Math_M2_Score	0.000	0.061
ASSIGNMENT_U3_U5	0.444	0.060
Math_M1_Score	0.000	0.055
EXITTICKET_U3	0.000	0.050
EXITTICKET_U5	0.000	0.049
ASSIGNMENT_U1_U2	0.000	0.040
EXITTICKET_U2	0.000	0.035
Ranking_Score	0.000	0.031
NEM	0.000	0.031
Language_Score	0.000	0.030
NEM_Score	0.000	0.029
Entrance_Test_Score	0.000	0.027
Control_U3_Cycles	0.000	0.026

Predictive performance under this scenario was evaluated through the systematic comparison summarized in Table 8. During grid-search optimization, all models benefited substantially from the inclusion of C1, reaching markedly higher training cross-validated AUC values than in the pre-university scenario. Although SVM achieved the highest cross-validated AUC (0.9339), RF was selected as the final model because it yielded the highest Fail-class precision on the validation set (0.7674), which was the pre-defined criterion for model selection given the institutional priority of limiting false positives in targeted interventions.

Table 8. Model optimization results and validation performance for the information up to C1 scenario.

Model	Best CV AUC (Training)	Precision (Failed)—Validation
SVM	**0.9339**	0.7018
RF	0.9272	**0.7674**
GBC	0.9213	0.7660
DT	0.8907	0.7455

The selected RF configuration used 100 estimators with no maximum depth constraint (max_depth = None). To address class imbalance, SMOTE oversampling was applied within the training process only, thereby avoiding information leakage into validation or test data.

Final generalization performance was then assessed on an independent test set (n = 149) and compared against a logistic regression baseline (Table 9). Consistent with the addition of C1, all models showed strong predictive performance, with AUC-ROC values ranging from 0.916 to 0.926. Under the pre-defined validation-based selection criterion, the chosen RF model achieved a Fail-class precision of 0.787, improving upon the baseline (0.695) and indicating a lower false-positive burden than the pre-university scenario.

Table 9. Classification metrics summary for the information up to C1 scenario.

Model	AUC-ROC	Precision (Failed)	Recall (Failed)	F1 (Failed)
Logistic Regression (baseline)	**0.926**	0.695	0.820	0.752
Decision Tree	0.916	0.695	0.820	0.752
Random Forest	0.916	0.787	0.740	0.763
Gradient Boosting	0.923	**0.796**	0.780	0.788
Support Vector Machine	0.925	0.733	**0.880**	**0.800**

 doi:10.20944/preprints202603.1704.v1

At the same time, the held-out test results show that performance differences among the top models are relatively small and depend on the metric considered. GBC achieved slightly higher precision (0.796), while SVM obtained the highest recall (0.880) and the highest F1-score (0.800) for the Fail class. Although RF was not the top-performing model on every test metric, it remained competitive and aligned with the validation-based selection criterion, which prioritized precision for the Fail class under an operational early-warning setting. More importantly, these results indicate that the inclusion of C1 creates a substantially more informative intervention window, in which multiple non-linear models achieve strong and practically useful performance for institutional decision-making.

4.3.3. Comparative Analysis of Scenarios

To quantify the effect of incorporating early university performance data, we compared the validation-selected models across the two prediction scenarios: Pre-University and Up to C1 (Table 10). This comparison captures the transition from a baseline early-screening model based solely on pre-university information to a substantially more informative early-warning model that incorporates the first major in-course performance signal.

Table 10. Comparative performance summary across scenarios (validation-selected models evaluated on the test set).

Metric	Pre-University (GBC)	Up to C1 (RF)	Absolute Improvement (percentage points)
AUC-ROC	0.766	**0.916**	+15.0 pp
Precision (Failed)	0.638	**0.787**	+14.9 pp
Recall (Failed)	0.600	**0.740**	+14.4 pp
F1-Score (Failed)	0.619	**0.763**	+14.4 pp
Overall Accuracy	0.750	**0.850**	+10.0 pp

The transition from the Pre-University scenario to the Up to C1 scenario yields three main findings. First, the inclusion of C1 is associated with a marked improvement across all reported metrics. AUC-ROC increases from 0.766 to 0.916 (an absolute gain of 15.0 percentage points), indicating a substantial increase in the model's discriminative capacity. Similar improvements are observed in precision, recall, and F1-score for the Fail class. These results reinforce the idea that, while pre-university data provide a meaningful predictive baseline, early in-course performance captures a much stronger signal of academic risk.

Second, the precision for the Fail class improves from 0.638 to 0.787 (an absolute gain of 14.9 percentage points). From an institutional perspective, this means that interventions triggered after C1 can be targeted more efficiently: a larger proportion of students flagged as at risk are truly in need of support, reducing the burden of false positives relative to the pre-university scenario.

Third, the comparative results support a feature-combination effect rather than a simple replacement of historical predictors by in-course performance. As shown in the feature selection analysis, the Up to C1 scenario is strongly dominated by C1 in both linear and non-linear importance rankings. However, the strong performance of ensemble models suggests that pre-university background variables—particularly quantitative admission indicators such as Mathematics M1 and broader academic preparation variables such as Science_Score—still contribute contextual information that helps refine the decision boundary when combined with early university performance.

From an institutional standpoint, the Pre-University scenario remains valuable for broad preventive screening before the semester begins, enabling low-cost support actions such as leveling, orientation, or early monitoring. In contrast, the Up to C1 scenario provides a substantially higher-fidelity basis for more intensive and resource-demanding interventions, such as personalized tutoring or academic counseling, with greater confidence in the identification of at-risk students.

 doi:10.20944/preprints202603.1704.v1

5. Discussion

This study addresses persistently high failure rates in Introductory Programming (CS1) courses in Chilean higher education, a phenomenon that can exceed 50% in some institutions. Using a cohort of 994 students (46% failure rate), we pursued two complementary goals: (i) to characterize the academic profiles and trajectories associated with failure, and (ii) to compare predictive models under two realistic early-intervention windows—before the semester starts (pre-university scenario) and after the first major written examination (up to C1).

The results support Hypothesis 1 (H1) by showing consistent differences between students who pass and those who fail, particularly in indicators of prior quantitative preparation and admission-related measures. This pattern suggests that a meaningful portion of academic risk does not emerge solely "within" the course, but is partly rooted in pre-existing gaps that may constrain early progress in programming and the capacity to sustain the course pace. In addition, within-semester evidence indicates that students who ultimately fail tend to exhibit early and persistent disadvantages, consistent with prior work on early warning signals in CS1 contexts.

Regarding Hypothesis 2 (H2), the predictive experiments show that machine learning models can identify at-risk students with practically meaningful performance, exceeding a simple historical baseline. Importantly, the comparison between intervention windows reveals a clear operational trade-off: the pre-university scenario provides a useful predictive floor (i.e., acceptable discrimination using only pre-entry information), whereas incorporating performance information up to C1 yields a substantial improvement in both overall predictive quality and identification of the failing class. From an institutional perspective, this supports a staged intervention strategy: pre-semester predictions can guide broad, low-cost preventive supports, whereas post-C1 predictions can trigger more targeted and intensive interventions with greater confidence and a lower false-positive burden.

These findings align with recent literature highlighting the value of early warning systems in programming education. Consistent with prior studies (e.g., [8,10]), our results indicate that conventional classifiers such as Decision Trees and Support Vector Machines can provide strong baseline performance for identifying at-risk students. At the same time, our findings reinforce evidence that ensemble approaches often offer competitive or superior performance in operational settings. While some individual models achieved stronger performance on specific metrics, ensemble classifiers—particularly Random Forest (RF) and Gradient Boosting (GBC)—tended to provide more stable performance and stronger precision for the failing class across model-selection stages, consistent with trends reported by Xuan Lam et al. [12].

A further contribution of this study is the dual feature-selection perspective (L1 regularization vs. RF feature importance), which provides insight into the structure of academic risk. First, L1-based selection induced strong sparsity, retaining only a small number of dominant predictors, whereas RF assigned meaningful importance to a broader set of variables. Rather than a contradiction, this divergence suggests that several admission and high-school trajectory indicators (e.g., standardized scores and NEM/Ranking) may not exhibit strong independent linear relationships with failure, yet still contribute predictive value through non-linear interactions and multivariate dependencies. Second, in the "up to C1" scenario, both approaches converged on the dominance of C1 as the most influential predictor. This result is coherent with the course structure, where PC is computed as the mean of C1 and C2: C1 therefore functions as the earliest high-stakes summative checkpoint, substantially constraining final performance and providing a particularly informative signal for early warning.

From a pedagogical standpoint, the prominence of C1—despite the presence of multiple smaller assessments—has direct implications for instructional design and support allocation. Frequent low-stakes activities (e.g., quizzes and assignments) may serve an essential formative role by promoting practice and feedback, yet they may not function as equally strong diagnostic signals for risk prediction in this context. Two actionable directions follow: (i) strengthening the diagnostic value of early assessments through improved item design, tighter alignment with prerequisite skills, and

 doi:10.20944/preprints202603.1704.v1

higher-quality formative feedback, and (ii) defining explicit post-C1 alert thresholds that activate structured tutoring, academic coaching, and remedial supports.

Overall, the evidence supports a two-stage intervention framework for CS1 and similar first-year gateway courses: (a) pre-semester segmentation using pre-university variables to prioritize leveling actions, particularly in quantitative skills and study strategies; and (b) higher-fidelity early warning after C1 to deploy targeted interventions more efficiently. This framework offers a practical way to balance coverage and cost, and it may support institutional decision-making in contexts where failure is closely linked to academic progression and student retention.

6. Conclusions

6.1. Conclusions and Contribution

The primary objective of this study was to characterize the academic trajectories associated with failure in an introductory programming course (CS1) and to evaluate early prediction models under two realistic intervention windows: (i) before the semester begins and (ii) after the first major written examination (C1). In the current educational landscape—where Generative Artificial Intelligence (GenAI) tools provide rapid access to code generation and explanations—the early detection of gaps in computational thinking and quantitative foundations becomes increasingly important. In this context, the risk extends beyond course failure to the possibility of developing forms of technological dependence that may obscure structural learning difficulties.

This study contributes to the field in three main ways. First, it demonstrates a clear predictive duality: pre-university variables provide a meaningful baseline for early risk estimation (AUC ≈ 0.77), whereas incorporating performance information up to C1 yields a substantial improvement, enabling a markedly higher-fidelity model (AUC ≈ 0.92). Second, it identifies critical predictors of academic risk: model-based analyses consistently highlight quantitative background indicators and, most notably, C1 performance as dominant signals, reinforcing the close relationship between programming success and prior quantitative preparation. Third, it shows that ensemble models (particularly RF and GBC) provide strong and operationally useful performance in this context, supporting more efficient allocation of academic support resources by improving early targeting of at-risk students.

6.2. Limitations

Several limitations should be considered when interpreting these results. First, the analysis is based on a single institution and a single cohort, which may limit generalizability to other universities, curricular structures, or higher education systems. Second, although ensemble models achieved strong predictive performance, they offer less immediate interpretability than simpler classifiers; accordingly, translating predictions into actionable pedagogical explanations may require additional Explainable AI (XAI) techniques. Third, the dataset does not include fine-grained behavioral indicators (e.g., LMS traces, lab interaction logs, or time-on-task measures) that could enrich early-warning signals. Finally, this study does not incorporate qualitative or attitudinal variables (e.g., motivation, study strategies, or perceptions of GenAI), which may help explain why some students struggle and how they respond to support interventions.

6.3. Future Work

Future research should extend and strengthen this line of work in at least four directions. First, it should examine the impact of GenAI tools (e.g., ChatGPT, GitHub Copilot) on student learning behaviors and performance—particularly around high-stakes milestones such as C1—and assess whether GenAI use alters the success/failure patterns detected by the model. Second, it should incorporate XAI methods (e.g., SHAP, LIME) to provide instructors and tutors with clearer student-level explanations that support personalized feedback and targeted remediation. Third, it should

 doi:10.20944/preprints202603.1704.v1

pursue multi-cohort and multi-institution validation in order to assess robustness across contexts and move toward more scalable early-warning systems. Fourth, it should explore longitudinal outcomes, including retention and progression into second-year coursework, to evaluate whether early interventions informed by these models are associated with improved persistence and reduced dropout.

Finally, this study contributes evidence specific to the Chilean higher education context, where high failure rates in CS1 can constrain academic progression. By combining a statistical characterization of at-risk students with a rigorous comparison of early prediction scenarios, the proposed approach supports timely, data-informed pedagogical decision-making and provides a foundation for broader institutional—and potentially national—implementation.

Author Contributions: Conceptualization, R.G.-B. and J.C.-S.; methodology, R.G.-B., and J.C.-S.; software, R.G.-B.; formal analysis, R.G.-B.; investigation, R.G.-B. A.V.-G and J.C-S; writing—original draft preparation, R.G.-B.; writing—review and editing, A.V.-G. and J.C.-S.; supervision, J.C-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FIC-R IA 40036152-0 Project of the Regional Government of Biobío, Chile.

Institutional Review Board Statement: Ethical review and approval were not required for this study because it used retrospective, fully anonymized academic records provided by the university, with no access to direct identifiers and no possibility of reidentifying individual students. The use of the anonymized records for research purposes was authorized by the corresponding institutional authority at Universidad Técnica Federico Santa María.

Informed Consent Statement: Informed consent was not required because the study used retrospective anonymized academic records, and no identifiable personal data were available to the authors.

Data Availability Statement: The data are not publicly available due to institutional and privacy restrictions. The analyses were conducted using retrospective anonymized academic records provided by Universidad Técnica Federico Santa María. Because the dataset was delivered in anonymized form and is subject to institutional restrictions, access requests must be directed to the corresponding institutional authority.

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

References

1. Ganga Contreras, F.; Améstica-Rivas, L.; Ramírez González, V.; et al. Deserción estudiantil, el costo económico para las universidades chilenas. Universidad, Ciencia y Tecnología 2023, 27, 17–28. https://doi.org/10.47460/UCT.V27I118.683.
2. An, M.; Zhang, H.; Savelka, J.; et al. Are working habits different between well-performing and at-risk students in online project-based courses? In Proceedings of the Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE), pp. 324–330. https://doi.org/10.1145/3430665.3456320.
3. Liao, S.N.; Zingaro, D.; Alvarado, C.; et al. Exploring the value of different data sources for predicting student performance in multiple CS courses. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE 2019), pp. 112–118. https://doi.org/10.1145/3287324.3287407.
4. Rodríguez-Hernández, C.F.; Musso, M.; Kyndt, E.; Cascallar, E. Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. Computers and Education: Artificial Intelligence 2021, 2, 100018. https://doi.org/10.1016/j.caeai.2021.100018.
5. Brdesee, H.; Alsaggaf, W.; Aljohani, N.; Hassan, S.U. Predictive model using a machine learning approach for enhancing the retention rate of students at risk. International Journal of Swarm Intelligence and Evolutionary Computation 2022, 18, 1–21. https://doi.org/10.4018/IJSWIS.299859.
6. Qushem, U.B.; Oyelere, S.S.; Akçapınar, G.; et al. Unleashing the power of predictive analytics to identify at-risk students in computer science. Technology, Knowledge and Learning 2024, 29, 1385–1400. https://doi.org/10.1007/s10758-023-09674-6.

 doi:10.20944/preprints202603.1704.v1

7. Brito Correia, J.P.J.; Gomes, F.; Borges, A.; et al. Predicting student performance in introductory programming courses. Computers 2024, 13, 219. https://doi.org/10.3390/computers13090219.
8. Khan, I.; Ahmad, A.R.; Jabeur, N.; Mahdi, M.N. Machine learning prediction and recommendation framework to support introductory programming course. International Journal of Emerging Technologies in Learning (iJET) 2021, 16, 42–59. https://doi.org/10.3991/IJET.V16I17.18995.
9. Gollapalli, M.; Rahman, A.; Alkharraa, M.; et al. SUNFIT: A machine learning-based sustainable university field training framework for higher education. Sustainability 2023, 15, 8057. https://doi.org/10.3390/su15108057.
10. Veerasamy, A.K.; Laakso, M.J.; D'Souza, D.; Salakoski, T. Predictive models as early warning systems: A Bayesian classification model to identify at-risk students of programming. In Intelligent Computing: Proceedings of the 2021 Computing Conference, pp. 174–195. https://doi.org/10.1007/978-3-030-80126-7_14.
11. Morampudi, M.K.; Gonthina, N.; Reddy, V.D.; Rao, K.S. Analyzing student performance in programming education using classification techniques. In Proceedings of the International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC 2022). https://doi.org/10.1109/ASSIC55218.2022.10088377.
12. Xuan Lam, P.; Mai, P.Q.H.; Nguyen, Q.H.; et al. Enhancing educational evaluation through predictive student assessment modeling. Computers and Education: Artificial Intelligence 2024, 6, 100244. https://doi.org/10.1016/j.caeai.2024.100244.
13. Ilić, M.; Keković, G.; Mikić, V.; et al. Predicting student performance in a programming tutoring system using AI and filtering techniques. IEEE Transactions on Learning Technologies 2024, 17, 1931–1945. https://doi.org/10.1109/TLT.2024.3431473.
14. Mai, T.T.; Bezbradica, M.; Crane, M. Learning behaviours data in programming education: Community analysis and outcome prediction with cleaned data. Future Generation Computer Systems 2022, 127, 42–55. https://doi.org/10.1016/j.future.2021.08.026.
15. Choi, W.C.; Lam, C.T.; Pang, P.C.I.; Mendes, A.J. A systematic literature review of explainable artificial intelligence (XAI) for interpreting student performance prediction in computer science and STEM education. In Proceedings of the Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE), 2025, 1, 221–227. https://doi.org/10.1145/3724363.3729027.
16. De Oliveira, C.F.; Sobral, S.R.; Ferreira, M.J.; Moreira, F. Interpretable success prediction in a computer networks curricular unit using machine learning. Procedia Computer Science 2024, 239, 598–605. https://doi.org/10.1016/j.procs.2024.06.212.
17. Vásquez, A.; Meza, F.; Godoy, P.G. Emergency remote teaching model for massive programming classes. In 2021 XLVII Latin American Computing Conference (CLEI), Cartago, Costa Rica, 2021, pp. 1–9. https://doi.org/10.1109/CLEI53233.2021.9640145.