

## Article

# Low-Overhead Learning Analytics: A Parsimonious Early Warning System for At-Risk Students in Chilean STEM Courses

Camilo Pérez , Christopher Nikulin  and Bernardo Madariaga \* 

Faculty of Engineering, Universidad Alberto Hurtado, Santiago 8340575, Chile; cperez@uahurtado.cl (C.P.); cnikulin@uahurtado.cl (C.N.)

\* Correspondence: bmadariaga@uahurtado.cl

## Abstract

Early warning systems grounded in learning analytics offer a practical route to identifying students at risk of failing, yet their adoption in Chilean undergraduate STEM courses is constrained by limited infrastructure and data availability. This study proposes a parsimonious methodology based on univariate logistic regression with the first partial grade as the sole predictor—a deliberate choice prioritising interpretability, actionable timing, and replicability without specialised analytical infrastructure. The approach was tested on two introductory mathematics courses (Linear Algebra and Mathematics I) at two Chilean universities across 2022 and 2023 cohorts. Models trained on 2022 data were externally validated on 2023 data without retraining. Discrimination was high at both institutions (AUC 0.85–0.93 in external validation). At the institution where approval prevalence remained stable across cohorts, operational metrics (PPV, NPV, balanced accuracy) were consistent between years. At the institution where prevalence shifted markedly—because the second semester is composed mainly of students who previously failed—the model preserved its discriminatory capacity but PPV and NPV inverted substantially, illustrating that operational metrics are functions of prevalence rather than indicators of model failure. A gender equity audit revealed systematically lower sensitivity for female students, indicating under-identification of at-risk women under a uniform threshold and underscoring the need for periodic equity monitoring in any deployment. The methodology offers a replicable, low-overhead floor for early warning practice in comparable courses.

**Keywords:** early warning systems; logistic regression; student retention; STEM education; learning analytics



Academic Editor: Hengtao Tang

Received: 19 May 2026

Revised: 1 July 2026

Accepted: 7 July 2026

Published: 11 July 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

## 1. Introduction

### 1.1. Early University Dropout in Chile

Early university dropout is a persistent problem in the Chilean higher education system. According to the latest undergraduate retention report from the Servicio de Información de Educación Superior, the first-year retention rate for the 2022 cohort reached 76% nationally, with marked differences between accredited (76.4%) and non-accredited (59.9%) institutions ([Servicio de Información de Educación Superior, 2023](#)). In the field of Technology, retention is the second lowest in the system (73.3% for women, 72.3% for men). Behind these figures lie social and economic costs borne by students, their families, and institutions ([Améstica-Rivas et al., 2021](#)), which the literature associates with a combination of academic, sociodemographic, and structural factors ([López-Aguilar et al., 2021](#); [Von Hippel & Hofflinger, 2021](#)). [Olivares-Rodríguez et al. \(2024\)](#) document that STEM course

configurations in the foundational cycle (Algebra and Calculus among them) constitute a significant share of dropout decisions. In the Chilean context specifically, pre-enrolment records are only weakly predictive of persistence, whereas within-semester grade data substantially improve prediction accuracy (Von Hippel & Hofflinger, 2021). Recent evidence from learning analytics studies in Chilean and Latin American universities reinforces this pattern, confirming the value of intra-course performance data for early identification of at-risk students (Henríquez Cabezas & Vargas Escobar, 2022; Villalobos Díaz et al., 2024).

### 1.2. Theoretical Anchor

Performance in early assessments operates as an observable signal of academic integration and, therefore, as an ecologically valid predictor of course persistence (Tinto, 1975). This link is the foundational justification for the early warning systems subfield, which operationalises academic engagement through quantitative partial performance indicators.

### 1.3. Early Warning Systems in STEM Courses

Institutional deployment of early warning systems was consolidated with initiatives such as Course Signals at Purdue University, where Arnold and Pistilli (2012) reported reductions in D, F, and withdrawal rates by combining prediction with timely notification to instructors and students. Jayaprakash et al. (2014), within the Open Academic Analytics Initiative, showed that predictive models trained at one institution can transfer to others with acceptable performance, albeit with sensitivity costs. In first-year engineering courses, Marbouti et al. (2016) compared seven prediction techniques (logistic regression, naive Bayes, SVM, and k-NN among them) on within-semester assessment data, reaching accuracies around 79% at week 2 and 90% at week 4, and explicitly identifying the asymmetric cost of false negatives as a model selection criterion. These antecedents establish three field regularities: (a) partial performance predictors are competitive against models with dozens of variables; (b) actionability depends on the temporal interval between observation and the opportunity to intervene; and (c) threshold decisions must consider the asymmetry between error types. A systematic review of predictive studies in computing and STEM education confirms that early partial grades consistently outperform demographic variables as predictors of academic success (Hellas et al., 2018).

### 1.4. Success Prediction and Algorithmic Bias Risks

Gardner and Brooks (2018b), focused on MOOCs, and Gardner and Brooks (2018a), in a broader evaluation framework, systematise the critical methodological decisions in student success prediction: outcome variable definition, feature engineering, model evaluation, and deployment architecture. Their shared methodological conclusion relevant to this work is that many field evaluations use inadequate metrics or ignore the dependence between apparent performance and sample composition. In parallel, Baker and Hawn (2022) document that educational prediction models can exhibit differential biases by subgroup (gender, race, socioeconomic status) even when global metrics are satisfactory, and propose equity audits as standard practice. These two antecedents frame the methodological decisions of the present study: comparing against baselines, reporting balanced metrics, and conducting a gender equity audit where data permit.

### 1.5. Research Questions and Hypotheses

Drawing on the framework above, this study formulates three research questions:

RQ1. Can a univariate logistic regression with the first partial grade as the sole predictor identify students at risk of failing an introductory STEM course with acceptable accuracy?

RQ2. Under what distributional conditions does predictive performance remain stable when a model trained on one cohort is applied to the following cohort of the same course?

RQ3. How do changes in the approval prevalence between cohorts affect the operational metrics (PPV, NPV) of the model?

With respect to RQ1, the study predicts that AUC will reach or exceed 0.80 in external validation for both courses. With respect to RQ3, the study predicts that PPV and NPV will shift in opposite directions when prevalence changes between cohorts. RQ2 is treated as exploratory: the study examines whether discrimination metrics remain stable when models are applied to a subsequent cohort, and under what distributional conditions stability breaks down.

To address these questions, the study employs a univariate logistic regression with the first partial examination grade as the sole predictor (Note1 at University A and Certamen1 at University B). This predictor was chosen for its systematic availability across both institutions' records, its actionability as an early intervention signal available 10–12 weeks before the final examination, and its interpretability for teaching staff without specialised analytical infrastructure. The model is logistic regression rather than a more complex classifier, not because it achieves maximum predictive accuracy, but because it produces a transparent, documentable decision rule that an instructor can implement and audit without data science support. The study's intended locus of transferability is methodological rather than populational: the results are addressed to teaching staff and academic managers who possess basic statistical knowledge and access to partial examination records, and who seek a functional early warning floor they can implement, monitor, and periodically recalibrate.

## 2. Methods

### 2.1. Design and Data

The study adopts an observational, comparative design based on two institutional case studies. Data come from academic records of two Chilean universities belonging to the Consejo de Rectores de Universidades Chilenas, anonymised in this report as University A and University B. University A is a larger institution whose Linear Algebra course draws students from multiple STEM programmes across faculties; University B is a mid-sized private university in Santiago serving primarily engineering students through a common first year. At each institution, one introductory mathematics course taught to STEM students was analysed: Linear Algebra at University A and Mathematics I at University B. Four semesters were included for both courses: 2022-1, 2022-2, 2023-1, and 2023-2.

Course assessments at both universities followed a stable grading scheme throughout the observation period. At University A, the course included three non-recoverable partial examinations, each weighted at 20%, in addition to complementary laboratory activities weighted at 10% and a mandatory comprehensive final examination weighted at 30%. The course was passed with a grade of 3.95, rounded to 4.0, on a 1.0 to 7.0 grading scale. At University B, the three partial examinations had progressive weightings of 20%, 25%, and 30%, respectively, in addition to three minor assessments administered before each partial examination, which together accounted for 25% of the final grade. The course was passed with a minimum score of 55 out of 100. If a student did not reach the minimum score of 55, they were required to take a comprehensive final examination; in such cases, the final course grade was recalculated by incorporating this examination as 30% of the total course grade. This operational stability allows comparisons to be made between cohorts within the same university, without differences in assessment instruments explaining variations in student performance.

## 2.2. Variables and Predictor Choice

The outcome variable is final course approval (1 = pass, 0 = fail), defined according to each institution's assessment regulations. The sole predictor corresponds to the first partial examination grade in each course. For University A, this variable is identified as Note1, whereas for University B, the equivalent variable is identified as Certamen1. This parsimonious choice reflects three operational considerations. First, availability: both institutions' records systematically contain this variable, whereas other indicators (attendance, platform usage) show irregular coverage. Second, in terms of actionability, both Note1 and Certamen1 are available several weeks before the comprehensive final examination, thereby providing a usable intervention window. Third, interpretability for teaching staff: a single-variable rule is transparent and reproducible without specialised analytical infrastructure.

This parsimony has a recognised cost. Models with multiple predictors documented in the literature (Marbouti et al., 2016) typically achieve marginally higher accuracy. The defence of the univariate choice is not maximum performance but minimum viable replicability: what an instructor can implement without specialised institutional support. Comparative benchmarks indicate that logistic regression achieves performance within a few percentage points of more complex machine learning classifiers in academic prediction tasks (Tomasevic et al., 2020), while single early-assessment indicators have been shown to predict course failure effectively in introductory STEM settings (Alzen et al., 2018). The adoption of learning analytics in Latin American universities is additionally constrained by limited data governance capacity and analytical infrastructure (Hilliger et al., 2020), which reinforces the case for a replicable, low-overhead methodology.

This decision is also consistent with prior systematic evidence in educational data mining and learning analytics. Namoun and Alshanjiti (2021) found that term assessment scores and examination grades are among the most recurrent predictors of student performance, while also noting that many predictive studies lack baseline comparisons or validation across multiple datasets. In this sense, the present study contributes to this literature by testing a deliberately parsimonious predictor under external validation conditions and by comparing its performance against simple deterministic baselines.

## 2.3. Treatment of Withdrawn Students

Records include students who withdrew from the course during the semester. These cases were excluded from the main analysis because they lack the outcome variable (no final pass or fail). We acknowledge that this exclusion introduces a potential selection bias, as the decision to withdraw may be endogenous to academic risk. The scope and implications of this exclusion are discussed in the Limitations Section.

## 2.4. Model, Training, and Validation

For each subject and each 2022 semester, a binary logistic regression was fitted with Note1 as the sole predictor. Training used 80% of the corresponding 2022 cohort data with five-fold cross-validation; the remaining 20% served as the internal test set, on which the metrics in the tables of the Section 3 are reported. External validation was performed by applying the 2022-trained model to the equivalent 2023 cohort data (2022-1 → 2023-1; 2022-2 → 2023-2) without retraining or recalibration. The classification threshold was determined from the training partition and applied without modification to both the internal test set and the external validation cohorts.

### 2.5. Performance Metrics and Baselines

Reported metrics include: overall accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under the ROC curve (AUC), balanced accuracy, and F1-score. For internal validation, tables report sensitivity, specificity, and AUC, which are sufficient to characterise threshold selection and internal discriminatory capacity; the full set of metrics is reported for external validation, where comparative judgement between models is the primary analytical task. Performance metrics are reported as point estimates; interval estimation is not pursued given the deterministic nature of the external validation design.

Recognising that accuracy can be misleading under class imbalance (Hicks et al., 2022; Mduma, 2023), the primary metrics for comparative judgement are balanced accuracy and F1; overall accuracy is reported as a secondary measure. AUC is additionally reported for the logistic model in the external validation tables as a threshold-independent measure of discriminatory capacity. Two baselines are additionally reported in each external validation table: (1) a majority-class classifier that always predicts the modal class of the training set; and (2) a heuristic rule that classifies as passing those who exceed the administrative approval threshold on the first assessment, which differs by institution scale ( $\text{Note1} \geq 4.0$  on a 1–7 scale for University A;  $\text{Certamen1} \geq 55$  points out of 100 for University B), without probabilistic estimation. Including the heuristic baseline directly addresses the concern that the logistic model may not add value over the simple “look at the first grade” rule (see Section 4.2).

### 2.6. Classification Threshold Selection

For each trained model, an optimal Note1 threshold was identified using the Youden criterion applied to the training partition, which maximises the difference between the true positive rate and the false positive rate. This threshold was then held fixed for all subsequent evaluations.

### 2.7. Gender Equity Audit

For University A, where the gender variable is available, sensitivity, specificity, PPV, and NPV are reported separately for male and female subgroups in the external validation cohorts. This audit addresses the algorithmic bias risk documented by Baker and Hawn (2022). Data from University B do not consistently contain the gender variable, so the audit is not replicable for Case 2.

### 2.8. Actionability Window

The first partial assessment in both subjects is administered around weeks 4–5 of the academic semester. Considering a semester of approximately 16 weeks until the global exam, this leaves a 10–12 week window during which an academic intervention (tutoring, reinforcement sessions, psychosocial support referral) could be applied.

### 2.9. Supplementary Non-Parametric Tests

Two non-parametric tests were applied to examine group differences in the predictor variable, supplementing the logistic regression analysis. At University A, a Wilcoxon rank-sum test was applied to compare Note1 distributions between male and female students within each semester. This test was selected because Note1 does not follow a normal distribution and the research interest concerns the location of the distributions across subgroups. At University B, a Kruskal–Wallis test was applied to compare Certamen1 distributions across programmes of origin, given the multi-group structure of the course and the non-normality of the predictor. Additionally, likelihood ratio tests were applied

to evaluate whether adding programme membership as a categorical covariate improved model fit over Certamen1 alone. These tests compare the log-likelihood of the full model against the restricted model under a chi-squared distribution, with degrees of freedom equal to the number of additional parameters introduced by the programme variable. The likelihood ratio test results supported the retention of the univariate specification, as programme membership did not improve prediction over Certamen1 alone.

### 3. Results

#### 3.1. Case 1: University A—Linear Algebra

Table 1 summarises the four cohorts analysed at University A. The course has a distinctive composition: the first semester is taken by a large group of Civil Engineering students in the common cycle, while the second semester brings together students from diverse STEM programmes. The approval prevalence remained within a narrow band between 74% and 82% across the four semesters.

**Table 1.** Enrolled students and approval prevalence by cohort (University A).

|                     | 2022-1 | 2022-2 | 2023-1 | 2023-2 |
|---------------------|--------|--------|--------|--------|
| Enrolled students   | 858    | 440    | 876    | 405    |
| Prevalence (% pass) | 78%    | 78%    | 82%    | 74%    |

The logistic regression models trained on 2022 data produced positive and highly significant coefficients for Note1 in both semesters (Table 2). The optimal classification threshold was 3.9 for 2022-1 and 4.3 for 2022-2, with internal AUC values of 0.87 and 0.90 respectively (Table 3).

**Table 2.** Logistic regression coefficients (University A).

| Semester | Variable  | Beta   | Std. Error | Z Value | p      |
|----------|-----------|--------|------------|---------|--------|
| 2022-1   | Note1     | 1.505  | 0.123      | 12.25   | <0.001 |
|          | Intercept | −4.499 | 0.442      | −10.17  | <0.001 |
| 2022-2   | Note1     | 1.679  | 0.211      | 7.98    | <0.001 |
|          | Intercept | −5.904 | 0.865      | −6.83   | <0.001 |

**Table 3.** Internal validation on the 2022 test set (University A).

| Semester | Threshold | Sensitivity | Specificity | AUC  |
|----------|-----------|-------------|-------------|------|
| 2022-1   | 3.9       | 0.752       | 0.816       | 0.87 |
| 2022-2   | 4.3       | 0.868       | 0.842       | 0.90 |

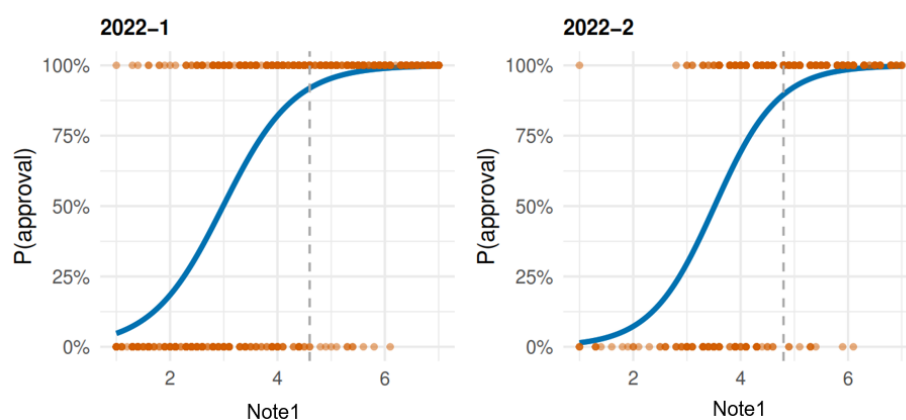
AUC computed on the held-out 20% test set. Sensitivity and specificity correspond to predictions at the Youden threshold determined from the training partition.

Figure 1 shows the sigmoid probability curves for both 2022 semesters, and Figure 2 shows the corresponding ROC curves.

Applied to the 2023 cohort, the models showed overall accuracy of 0.798 in the first semester and 0.672 in the second (Table 4). PPV remained high in both semesters (0.974 and 0.956), while NPV was moderate (0.471 and 0.437). Sensitivity dropped between semesters (0.774 → 0.583), although specificity remained stable near 0.90. External AUC was 0.926 in 2023-1 and 0.884 in 2023-2, confirming that discriminatory capacity is maintained without retraining. The logistic model outperforms the majority classifier in balanced accuracy in both cohorts. Against the heuristic, the model and the heuristic perform equivalently in

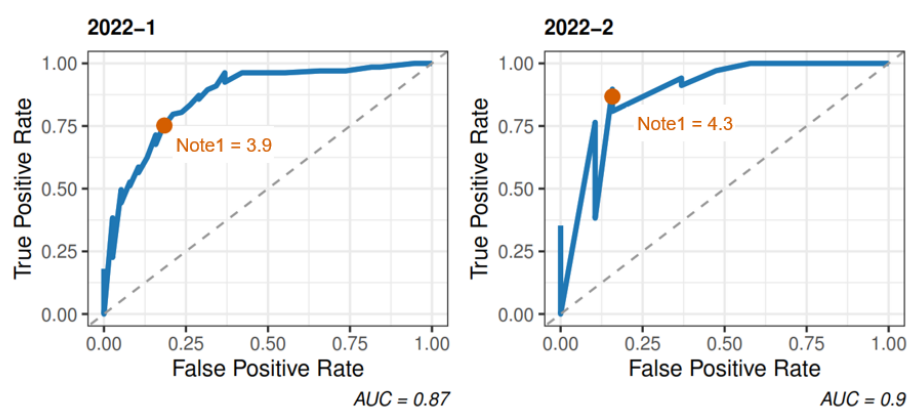
2023-1 (balanced accuracy 0.840 for both), while in 2023-2 the heuristic outperforms the model (0.801 vs. 0.754), a result discussed in Section 4.2.

#### Sigmoid probability of course approval (University A)



**Figure 1.** Sigmoid probability of course approval as a function of Note1 for semesters 2022-1 (left) and 2022-2 (right), University A. The dashed vertical line marks the median of Note1 in each training set.

#### ROC curves for the logistic model (University A)



**Figure 2.** ROC curves for the logistic models evaluated on the 2022 test set—2022-1 (left) and 2022-2 (right), University A. The orange dot marks the point corresponding to the Youden threshold selected from the training partition. The dashed line represents chance performance.

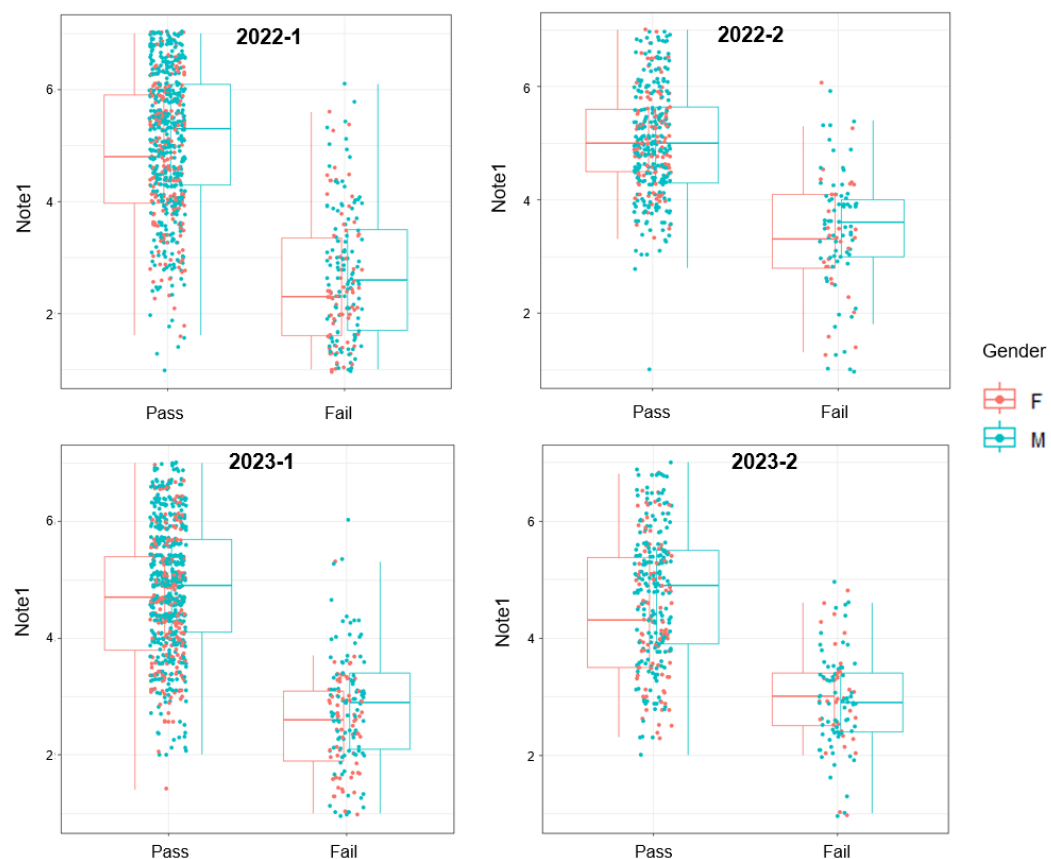
**Table 4.** External validation on the 2023 cohort with comparative baselines (University A).

| Cohort/Model                   | Acc.  | Sens. | Spec. | PPV   | NPV   | Bal. Acc/F1 | AUC   |
|--------------------------------|-------|-------|-------|-------|-------|-------------|-------|
| 2023-1—Logistic model          | 0.798 | 0.774 | 0.906 | 0.974 | 0.471 | 0.840/0.862 | 0.926 |
| 2023-1—Majority classifier     | 0.818 | 1.000 | 0.000 | 0.818 | n/a   | 0.500/0.900 | —     |
| 2023-1—Heuristic $N1 \geq 4.0$ | 0.798 | 0.774 | 0.906 | 0.974 | 0.471 | 0.840/0.862 | —     |
| 2023-2—Logistic model          | 0.672 | 0.583 | 0.924 | 0.956 | 0.437 | 0.754/0.725 | 0.884 |
| 2023-2—Majority classifier     | 0.741 | 1.000 | 0.000 | 0.741 | n/a   | 0.500/0.851 | —     |
| 2023-2—Heuristic $N1 \geq 4.0$ | 0.756 | 0.707 | 0.895 | 0.951 | 0.516 | 0.801/0.811 | —     |

PPV = positive predictive value; NPV = negative predictive value; Bal. Acc = balanced accuracy; n/a = metric undefined when classifier produces no predictions in that class. AUC is reported only for the logistic model, as baseline classifiers are deterministic rules that do not produce probability rankings. Baseline classifiers are deterministic rules.

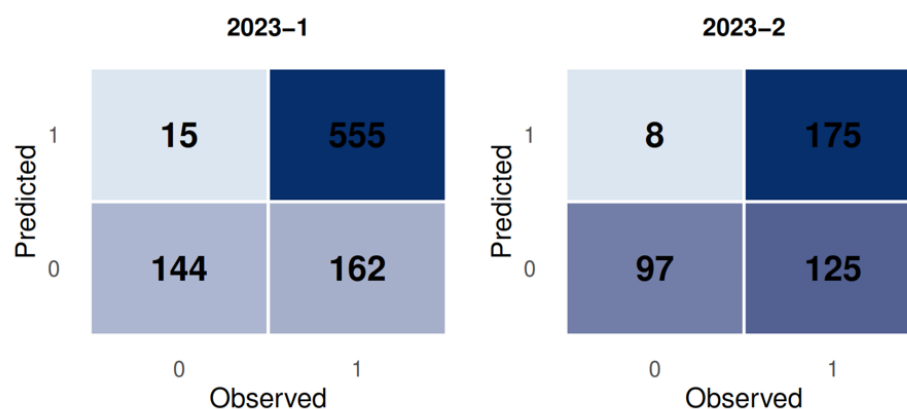
Figure 3 shows the distribution of Note1 by approval status and gender for all four semesters of University A, and Figure 4 shows the confusion matrices for the 2023 external validation. Table 5 reports model performance disaggregated by gender for the two external validation cohorts.

**Distribution of Note1 by approval status and gender (University A)**



**Figure 3.** Box plots of Note1 distribution by approval status and gender for semesters 2022-1, 2022-2, 2023-1, and 2023-2 (University A). A Wilcoxon test indicated a statistically significant difference between genders in semesters 2022-1 and 2023-1 but not in 2022-2 or 2023-2.

**Confusion matrices – External validation 2023 (University A)**



**Figure 4.** Confusion matrices for the external validation on semesters 2023-1 (left) and 2023-2 (right), University A. Positive class: pass (1). Darker squares indicate higher values.

**Table 5.** Gender equity audit: model performance metrics disaggregated by gender, external validation cohorts 2023 (University A).

| Cohort | Gender | <i>n</i> | Sens. | Spec. | PPV   | NPV   |
|--------|--------|----------|-------|-------|-------|-------|
| 2023-1 | Female | 282      | 0.714 | 0.985 | 0.994 | 0.508 |
| 2023-1 | Male   | 594      | 0.800 | 0.851 | 0.966 | 0.444 |
| 2023-2 | Female | 132      | 0.480 | 0.912 | 0.940 | 0.378 |
| 2023-2 | Male   | 273      | 0.634 | 0.930 | 0.962 | 0.471 |

Gender data available only for University A. The same classification threshold was applied to both subgroups without recalibration. In 2023-1, sensitivity is lower for female students (0.714 vs. 0.800) while specificity is substantially higher (0.985 vs. 0.851), indicating that the model flags fewer female students as at-risk and generates fewer false positives in this group. In 2023-2, sensitivity decreases for both subgroups relative to 2023-1, and the gap widens (0.480 vs. 0.634). These patterns should be interpreted cautiously given the unequal subgroup sizes; a formal bias analysis would require larger and more balanced samples.

### 3.2. Case 2: University B—Mathematics I

University B presents a distinct situation. Mathematics I brings together students from several engineering and related science programmes, and the second academic semester is composed mainly of students who failed in the first semester of the same year. This characteristic produces marked variation in approval prevalence between cohorts: 49% and 60% in first semesters versus 28% and 27% in second semesters (Table 6).

**Table 6.** Enrolled students and approval prevalence by cohort (University B).

|                     | 2022-1 | 2022-2 | 2023-1 | 2023-2 |
|---------------------|--------|--------|--------|--------|
| Enrolled students   | 688    | 270    | 795    | 267    |
| Prevalence (% pass) | 49%    | 28%    | 60%    | 27%    |

The logistic regression models trained on 2022 data produced positive and highly significant coefficients for Certamen1 in both semesters (Table 7). The optimal classification threshold was 22 for 2022-1 and 49 for 2022-2, with internal AUC values of 0.83 and 0.92 respectively (Table 8). Figures 5 and 6 show the sigmoid curves and ROC curves for the University B models.

**Table 7.** Logistic regression coefficients (University B).

| Semester | Variable  | Beta   | Std. Error | Z Value | <i>p</i> |
|----------|-----------|--------|------------|---------|----------|
| 2022-1   | Certamen1 | 0.111  | 0.010      | 11.71   | <0.001   |
|          | Intercept | −2.516 | 0.223      | −11.27  | <0.001   |
| 2022-2   | Certamen1 | 0.102  | 0.016      | 6.55    | <0.001   |
|          | Intercept | −5.888 | 0.854      | −6.90   | <0.001   |

**Table 8.** Internal validation on the 2022 test set (University B).

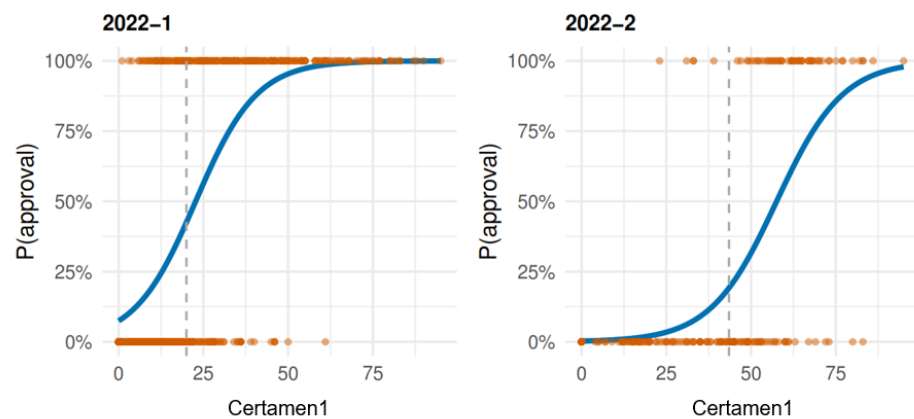
| Semester | Threshold (pts) | Sensitivity | Specificity | AUC  |
|----------|-----------------|-------------|-------------|------|
| 2022-1   | 22              | 0.716       | 0.771       | 0.83 |
| 2022-2   | 49              | 0.800       | 0.872       | 0.92 |

AUC computed on the held-out 20% test set. The markedly different optimal thresholds (22 vs. 49 points) reflect the direct effect of the prevalence difference between semesters.

In external validation on 2023 (Table 9), sensitivity was high in both semesters (0.873 and 0.918), but predictive values inverted between cohorts: PPV dropped from 0.765 to 0.472 and NPV rose from 0.765 to 0.952. Overall accuracy was 0.765 in the first semester and 0.697 in the second. External AUC was 0.850 in 2023-1 and 0.872 in 2023-2, indicating

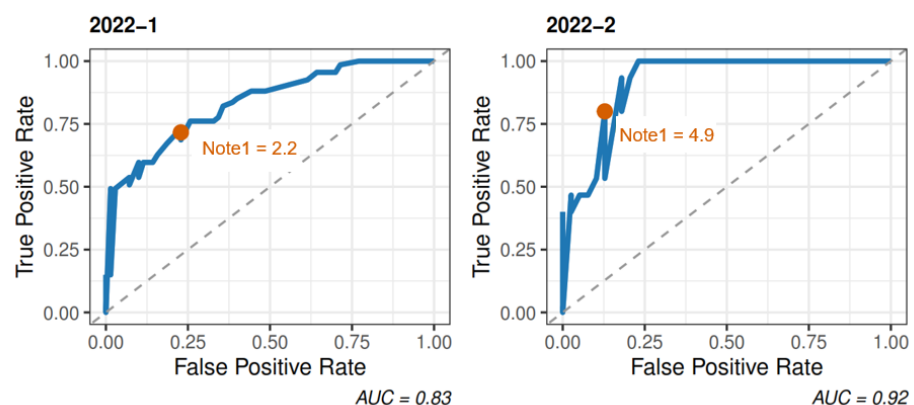
that the model's discriminatory capacity was preserved despite the pronounced prevalence shift between cohorts. The logistic model outperforms the majority classifier in balanced accuracy in both cohorts. The heuristic ( $\text{Certamen1} \geq 55$ ) shows very low sensitivity in 2023-1 (0.389), indicating that this threshold is inadequate for this group.

#### Sigmoid probability of course approval (University B)



**Figure 5.** Sigmoid probability of course approval as a function of Certamen1 for semesters 2022-1 (left) and 2022-2 (right), University B. The dashed vertical line marks the median of Certamen1 in each training set.

#### ROC curves for the logistic model (University B)



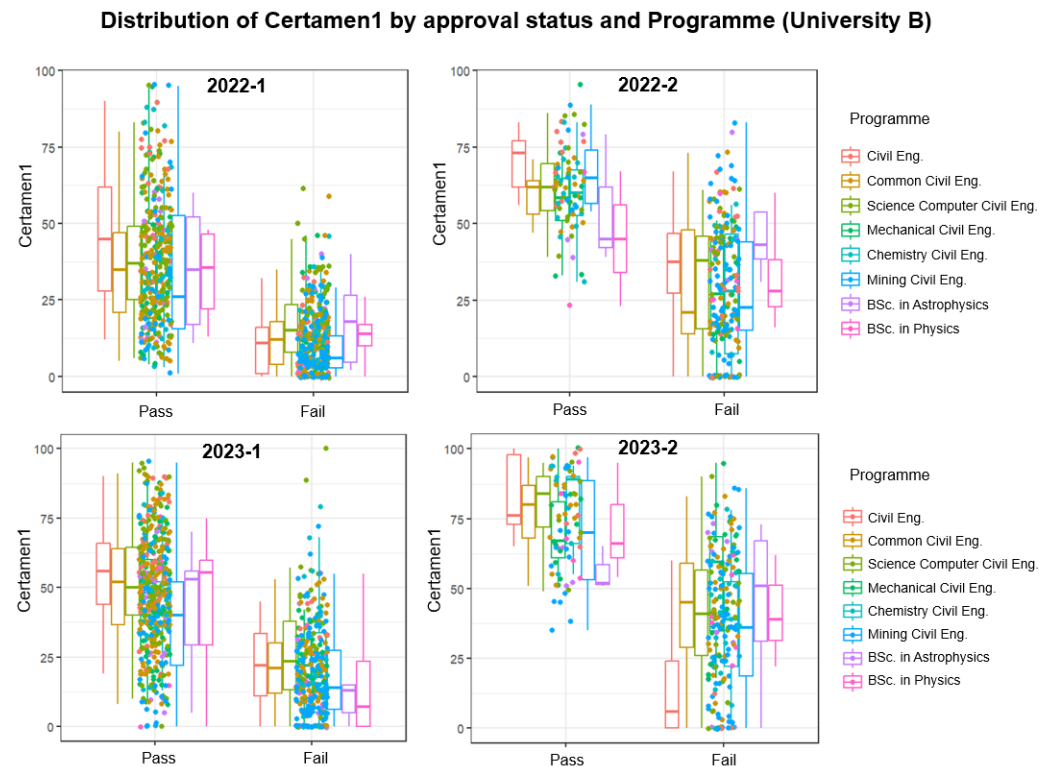
**Figure 6.** ROC curves for the logistic models evaluated on the 2022 test set—2022-1 (left) and 2022-2 (right), University B. The orange dot marks the point corresponding to the Youden threshold selected from the training partition. The dashed line represents chance performance.

**Table 9.** External validation on the 2023 cohort with comparative baselines (University B).

| Cohort/Model                  | Acc.  | Sens. | Spec. | PPV   | NPV   | Bal. Acc/F1 | AUC   |
|-------------------------------|-------|-------|-------|-------|-------|-------------|-------|
| 2023-1—Logistic model         | 0.765 | 0.873 | 0.606 | 0.765 | 0.765 | 0.739/0.815 | 0.850 |
| 2023-1—Majority classifier    | 0.405 | 0.000 | 1.000 | n/a   | 0.405 | 0.500/n/a   | —     |
| 2023-1—Heuristic $C1 \geq 55$ | 0.621 | 0.389 | 0.963 | 0.939 | 0.518 | 0.676/0.550 | —     |
| 2023-2—Logistic model         | 0.697 | 0.918 | 0.613 | 0.472 | 0.952 | 0.766/0.623 | 0.872 |
| 2023-2—Majority classifier    | 0.727 | 0.000 | 1.000 | n/a   | 0.727 | 0.500/n/a   | —     |
| 2023-2—Heuristic $C1 \geq 55$ | 0.723 | 0.836 | 0.680 | 0.496 | 0.917 | 0.758/0.622 | —     |

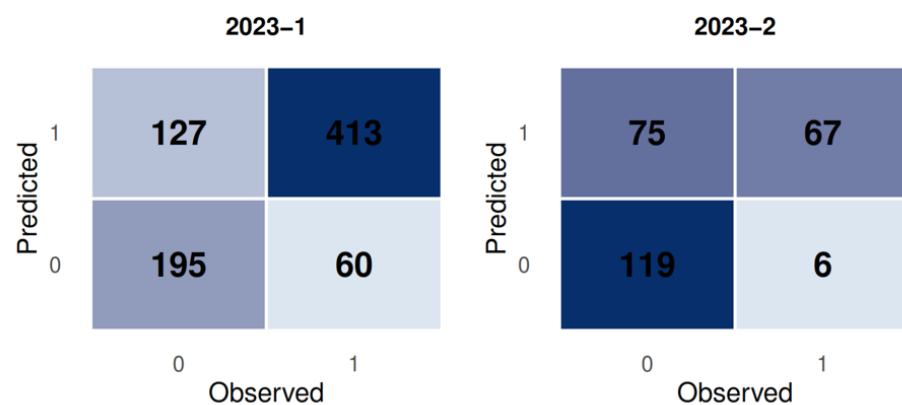
Heuristic based on the administrative approval threshold for University B ( $\text{Certamen1} \geq 55$  points out of 100). The majority classifier predicts the modal class of the training set; for University B the modal class is “fail” in both training semesters, hence sensitivity is 0 and specificity is 1. AUC is reported only for the logistic model, as baseline classifiers are deterministic rules that do not produce probability rankings. n/a = metric undefined. Baseline classifiers are deterministic rules.

Figure 7 shows the distribution of Certamen1 by approval status and programme, and Figure 8 shows the confusion matrices for the 2023 external validation.



**Figure 7.** Box plots of Certamen1 distribution by approval status and programme of origin for all four semesters, University B. A Kruskal–Wallis test indicated differences between programmes in 2022-1; however, likelihood ratio tests showed that programme did not improve prediction over Certamen1 alone, supporting the parsimonious single-predictor model.

#### Confusion matrices – External validation 2023 (University B)



**Figure 8.** Confusion matrices for the external validation on semesters 2023-1 (left) and 2023-2 (right), University B. Positive class: pass (1).

## 4. Discussion

### 4.1. Main Findings Against the Hypotheses

The prediction for RQ1: AUC values between 0.85 and 0.93 in external validation demonstrate that discriminatory capacity is maintained across cohorts without retraining, with a sole predictor. The prediction for RQ3 is confirmed: University B results show that PPV and NPV shift in opposite directions when prevalence changes, while University A, whose prevalence remained stable, shows relatively conserved operational metrics. Notably,

the external AUC values at University B (0.850 and 0.872) remain well above 0.80 despite the pronounced prevalence shift between cohorts, confirming that the model's underlying discrimination is robust to compositional changes even when operational metrics (PPV, NPV) are not. Regarding the exploratory question posed in RQ2, cohort stability holds for University A but breaks down for University B, where the structural composition of the second semester—predominantly students who failed in the first—produces a prevalence shift that substantially alters operational metrics while leaving AUC largely intact. This result clarifies the conditions under which stability can be expected: it requires not only the same course but a comparable composition of the student population across cohorts, as evidenced by the prevalence figures in Tables 1 and 6.

#### 4.2. On the Circularity Inherent to the Predictor

A reasonable objection to the design is that predicting a final grade from a partial grade is partially circular: both measure the same underlying construct (course content mastery) at different moments. We acknowledge this circularity as real. The model's contribution, however, is not epistemic but operational, and it manifests most clearly in three ways.

First, the probabilistic output enables threshold decisions adjusted to the asymmetric cost of errors, regardless of whether the predicted probabilities are individually precise. A raw grade provides an ordinal ranking; the logistic model translates it into a monotone probability score that can be shifted to reflect institutional risk tolerance—for instance, by lowering the classification threshold when the cost of a missed at-risk student outweighs the cost of a false alarm. This threshold-adjustment capacity is absent in the heuristic, which applies a single fixed cut-off. Whether the resulting probabilities are well-calibrated in an absolute sense is a question identified as a limitation in Section 5 and reserved for future work; for the binary classification task addressed here, the ordinal properties of the score suffice. Second, the comparison between universities demonstrates that the same Note1 value carries different predictive weight depending on the distributional context: the model trained on University A 2022-1 assigns a different probability of approval to a grade of 4.0 than the model trained on 2022-2, because the approval prevalence and student composition differ. The heuristic ignores this context and applies a fixed rule that may be appropriate for one cohort but not another. The University B 2023-1 case illustrates this directly: the fixed 55-point threshold collapses (sensitivity 0.389) because many students who ultimately pass begin the course with low partial grades and recover; the logistic model, with a threshold calibrated from the ROC curve (22 points), captures this intermediate zone. Third, and most practically, the model provides a decision support tool that an instructor can apply systematically and document—two requirements that the informal heuristic does not satisfy.

Reviewing the four external validation cohorts jointly, the logistic model outperforms the heuristic in one case (University B 2023-1, where the fixed administrative threshold collapses under distributional shift), performs equivalently in two (University A 2023-1 and University B 2023-2), and is outperformed in one (University A 2023-2). This distribution of outcomes is informative rather than damaging: the logistic model adds operational value precisely when the student population deviates from the cohort used to set the administrative threshold. When the population is sufficiently homogeneous and the administrative threshold is well-aligned with historical outcomes, the additional degree of freedom from threshold optimisation does not improve out-of-sample performance—a result that defines the boundary conditions of practical utility rather than constituting a failure of the method.

#### 4.3. Prevalence and Operational Metrics

The contrast between the two universities constitutes, in our judgement, the most useful contribution of the study. University A maintained stable prevalence between cohorts, and the model behaved consistently between semesters and years. University B presented a prevalence that oscillated between 49% and 27% depending on the semester, and there the operational metrics (especially PPV and NPV) varied drastically. This pattern should not be read as a model failure but as a well-known property of classifiers: PPV and NPV are functions of prevalence, and a model whose intrinsic discrimination (measured by AUC) remains stable can nevertheless deliver very different operational metrics when group composition changes. The practical implication is that the alert threshold should not be fixed once with historical data but re-evaluated when the composition of the target group changes, in line with [Gardner and Brooks' \(2018a\)](#) observations on the dependence between model evaluation and distributional context.

#### 4.4. Practical Implications and Risk of Self-Fulfilling Prophecy

For teaching staff and academic managers, the results suggest that a simple model can operate as a replicable early warning floor under three conditions of use. First, the window between the first assessment and the global exam must be wide enough to allow for intervention, and the institution must have support resources (tutoring, reinforcement sessions, psychosocial referral) to which identified students can be directed. Second, the classification threshold is not a fixed parameter but an institutional policy decision about the relative cost of Type I and Type II errors. Third, identifying at-risk students does not automatically improve retention: the intervention literature ([Arnold & Pistilli, 2012](#); [Jayaprakash et al., 2014](#)) shows positive effects when identification is accompanied by well-designed interventions, but the causal effect between identification and retention requires specific evaluation.

Additionally, the algorithmic labelling of a student as at-risk is not neutral. [Baker and Hawn \(2022\)](#) document that automated classifications can reproduce subgroup biases and generate self-fulfilling prophecy effects when information is transmitted without context. The gender equity audit (Table 5) illustrates this concern concretely: female students are flagged as at-risk less frequently than male students (lower sensitivity, higher specificity), which could mean that genuinely at-risk female students are systematically missed. Any deployment of the methodology should therefore include: (a) periodic equity audits by available subgroups; (b) clear protocols on who receives the information and how it is communicated to the student; and (c) periodic threshold re-evaluation as cohort composition changes.

### 5. Limitations

This study has five main limitations that bound the scope of its conclusions.

First, the design covers two introductory mathematics courses at two Chilean universities; any generalisation to other disciplines, educational levels, or national contexts requires independent validation.

Second, the model uses a sole predictor by design choice. Models with multiple predictors could achieve higher accuracy, especially in cohorts where prevalence varies significantly.

Third, the study does not include a control group without intervention, so it cannot make any claim about the causal effect of early detection on actual retention. The results reported here concern identification capacity, not intervention efficacy.

Fourth, the gender equity audit was only possible at University A and should be interpreted cautiously given the unequal subgroup sizes. A complete audit for University B remains pending and requires access to demographic data not available in the records used.

Fifth, students who withdrew from the course during the semester were excluded from the analysis because they lack a final pass or fail outcome. This exclusion may introduce upward bias in performance metrics, as students with the lowest academic engagement are removed from the sample. The reported metrics should therefore be understood as an optimistic bound on real-world performance. Additionally, model calibration—evaluating whether the predicted probabilities are reliable as such and not only as a ranking—was not assessed in this study and represents a relevant direction for future work, particularly if the probabilistic output is used as input for individualised pedagogical decisions.

## 6. Conclusions

Three findings from this study are relevant for practitioners and researchers working on early warning in Latin American STEM higher education. First, a univariate logistic regression with the first partial grade as the sole predictor achieves an AUC between 0.85 and 0.93 in external validation without retraining, demonstrating that a single universally available assessment variable can sustain a functional early identification system in resource-constrained institutional contexts. Second, the contrast between the two institutions clarifies the distributional conditions that govern threshold stability: when course composition is comparable across cohorts, operational metrics (PPV, NPV) remain consistent; when prevalence shifts—as occurs in courses dominated by students repeating the subject—the classification threshold must be re-evaluated at each deployment rather than fixed from historical data, even though the model’s discriminatory capacity (AUC) is preserved. Third, the gender equity audit at University A reveals consistently lower sensitivity for female students across both external validation cohorts, indicating systematic under-identification of at-risk women under a uniform threshold; this finding supports the recommendation to embed periodic equity monitoring in any institutional deployment of the methodology.

Together, these results provide a replicable, transparent, and low-overhead design that teaching staff and academic managers can implement without specialised infrastructure. Future work should evaluate the causal effect of targeted interventions on actual retention, replicate the methodology in other STEM disciplines and institutional contexts, assess absolute model calibration, and examine whether the inclusion of additional variables improves equity without sacrificing interpretability for end users.

**Author Contributions:** Conceptualization, C.P. and C.N.; methodology, C.P., C.N. and B.M.; software, C.P. and B.M.; validation, C.P.; formal analysis, C.P. and B.M.; investigation, C.P., C.N. and B.M.; data curation, C.P.; writing—original draft preparation, C.P. and C.N.; writing—review and editing, B.M.; visualization, C.P., C.N. and B.M.; supervision, C.N. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Ethical review and approval were waived for this study because the analysis was based exclusively on anonymised secondary data from institutional administrative academic records. Data collection was part of the regular administrative management processes of the participating universities during 2022 and 2023. No direct contact with students was made and no additional instruments were applied. The institutions are reported in anonymised form (University A and University B) to preserve institutional confidentiality.

**Informed Consent Statement:** Informed consent was waived due to the same reason above.

**Data Availability Statement:** The individual-level data underlying the findings of this study are not publicly available due to restrictions arising from the institutional ownership of academic records and the terms under which they were originally accessible to the authors. Aggregated data and analysis code may be requested from the corresponding author upon reasonable justification.

**Acknowledgments:** The authors thank the academic registrar offices of both participating institutions for providing access to the anonymised records used in this study.

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

## Abbreviations

|      |                                                  |
|------|--------------------------------------------------|
| AUC  | Area Under the ROC Curve                         |
| PPV  | Positive Predictive Value                        |
| NPV  | Negative Predictive Value                        |
| ROC  | Receiver Operating Characteristic                |
| STEM | Science, Technology, Engineering and Mathematics |

## References

- Alzen, J. L., Langdon, L. S., & Otero, V. K. (2018). A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. *International Journal of STEM Education*, 5(1), 56. [CrossRef] [PubMed]
- Améstica-Rivas, L., King-Domínguez, A., Sanhueza, D., & Ramírez, V. (2021). Efectos económicos de la deserción en la gestión universitaria: El caso de una universidad pública chilena. *Hallazgos*, 18(35), 209–231. [CrossRef]
- Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267–270). ACM. [CrossRef]
- Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4), 1052–1092. [CrossRef]
- Gardner, J., & Brooks, C. (2018a). Evaluating predictive models of student success: Closing the methodological gap. *Journal of Learning Analytics*, 5(2), 105–125. [CrossRef]
- Gardner, J., & Brooks, C. (2018b). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, 28(2), 127–203. [CrossRef]
- Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175–199). ACM. [CrossRef]
- Henríquez Cabezas, N., & Vargas Escobar, D. (2022). Modelos predictivos de rendimiento y deserción académica en estudiantes de primer año de una universidad pública chilena. *Revista de Estudios y Experiencias en Educación*, 21(45), 299–316. [CrossRef]
- Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1), 5979. [CrossRef] [PubMed]
- Hilliger, I., Ortiz-Rojas, M., Pesántez-Cabrera, P., Scheihing, E., Tsai, Y.-S., Muñoz Merino, P. J., Broos, T., Whitelock-Wainwright, A., & Pérez-Sanagustín, M. (2020). Identifying needs for learning analytics adoption in Latin American universities: A mixed-methods approach. *The Internet and Higher Education*, 45, 100726. [CrossRef]
- Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1), 6–47. [CrossRef]
- López-Aguilar, D., Álvarez-Pérez, P. R., & Garcés-Delgado, Y. (2021). El engagement académico y su incidencia en el rendimiento del alumnado de grado de la Universidad de La Laguna. *Revista Electrónica de Investigación y Evaluación Educativa*, 27(1), art. 5. [CrossRef]
- Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103, 1–15. [CrossRef]
- Mduma, N. (2023). Data balancing techniques for predicting student dropout using machine learning. *Data*, 8(3), 49. [CrossRef]
- Namoun, A., & Alshantiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), 237. [CrossRef]
- Olivares-Rodríguez, C., Moreno-Marcos, P. M., Scheihing García, E., Muñoz Merino, P. J., & Delgado-Kloos, C. (2024). An actionable learning path-based model to predict and describe academic dropout. *Ingeniería e Investigación*, 44(2), e109389. [CrossRef]
- Servicio de Información de Educación Superior. (2023). *Informe de retención de 1er año de pregrado, Cohortes 2018–2022* (Tech. Rep.). Ministerio de Educación de Chile. Available online: [https://www.mifuturo.cl/wp-content/uploads/2023/09/Retencion\\_de\\_Pregrado\\_2023\\_SIES.pdf](https://www.mifuturo.cl/wp-content/uploads/2023/09/Retencion_de_Pregrado_2023_SIES.pdf) (accessed on 15 October 2025).

- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89–125. [\[CrossRef\]](#)
- Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, 143, 103676. [\[CrossRef\]](#)
- Villalobos Díaz, E., Hilliger, I., Gonzalez, C., Celis, S., Pérez-Sanagustín, M., & Broisin, J. (2024). The mediating role of learning analytics: Insights into students' approaches to learning and academic achievement in Latin America. *Journal of Learning Analytics*, 11(1), 6–20. [\[CrossRef\]](#)
- Von Hippel, P. T., & Hofflinger, A. (2021). The data revolution comes to higher education: Identifying students at risk of dropout in Chile. *Journal of Higher Education Policy and Management*, 43(1), 2–23. [\[CrossRef\]](#) [\[PubMed\]](#)

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