Detecting At-Risk Students for Dropout: A Comparative Study of Longitudinal vs. Aggregated Data Models

BRUNO LUZA*, University of Caxias do Sul, Brazil
ANDRE GUSTAVO ADAMI, University of Caxias do Sul, Brazil

Student dropout in higher education is a critical, multifaceted global challenge with significant implications for students, institutions, and society at large. Traditional monitoring and intervention approaches often struggle to capture the complexity and early signals associated with student disengagement. In response, Machine Learning has emerged as a strategic approach for the early identification of students at risk of dropping out, but its practical application requires addressing the challenges inherent in educational data. Based on a raw longitudinal dataset, where each student is represented by multiple samples corresponding to the enrolled semesters, we compare two modeling approaches: (1) the original longitudinal representation, and (2) an aggregated representation, where each student is represented by a single sample. The results demonstrated a relative improvement of 10.22% in the area under the Precision-Recall Curve (AUPRC). Additionally, we found that balancing the data offers no performance gains and, in some cases, leads to a decrease in model performance. The results suggest that while the classifiers are robust to class imbalance, how students are represented remains a critical challenge. Finally, the model interpretation using SHAP highlighted the importance of numerical and categorical feature processing and identified the key factors associated with dropout, transforming the model into a practical tool for institutional decision support.

Reference:
Bruno Luza and Andre Gustavo Adami. 2026. Detecting At-Risk Students for Dropout: A Comparative Study of Longitudinal vs. Aggregated Data Models. *Bachelor's Thesis in Computer Science* (July 2026), 23 pages.

1 Introduction

Dropout in higher education is a global phenomenon and represents one of the most significant challenges for contemporary educational policies. According to data from the Organization for Economic Co-operation and Development (OECD), among its member countries, only 43% of students complete their programs within the expected duration, while 13% drop out during the first year (OECD 2025). In Brazil, this reality is even more concerning: one in four students drops out in the first year and only 38% complete their programs on time (INEP 2024; OECD 2025). Far from being merely a statistical figure, dropout is a complex phenomenon resulting from the interaction of multiple socioeconomic, institutional, and academic factors that extend beyond the individual sphere and produce negative consequences not only for students but also for higher education institutions (HEIs) and society (Aina et al. 2022; Moreno and Támara 2024; Silva et al. 2022).

In this context, Machine Learning (ML) emerges as a valuable and strategic approach, enabling the early identification of students who are either at high risk of dropout or more likely to achieve academic success (Cardona et al. 2023). Furthermore, the interpretability of ML models allows revealing factors associated with these academic outcomes, thereby promoting transparency in understanding these patterns (Nagy and Molontay 2024). This interpretative capability enables ML to function as a decision-support tool, providing various stakeholders

*Corresponding Author.

Authors' Contact Information: Bruno Luza, bbluza@ucs.br, University of Caxias do Sul, Caxias do Sul, Rio Grande do Sul, Brazil; Andre Gustavo Adami, ORCID: 0000-0001-8105-7698, andre.adami@ucs.br, University of Caxias do Sul, Caxias do Sul, Rio Grande do Sul, Brazil.

with the empirical evidence needed to inform institutional interventions focused on student monitoring and support (Nagy and Molontay 2024; Silva et al. 2022). Nevertheless, fully realizing this potential requires addressing the challenges inherent to the educational context of this task.

As noted by Baker and Yacef (2009), most data in educational contexts commonly display temporal dependencies or hierarchical structures, meaning that when we refer to students, their representation consists of multiple observations over time. This characteristic adds a significant layer of complexity, as each observation is not an independent sample, but rather belongs to a group of samples representing a larger entity. Compounding this challenge is data imbalance, resulting from the low incidence of dropout cases compared with the overall number of retained students (Albreiki et al. 2021; Rastrollo-Guerrero et al. 2020). These challenges have direct implications for the entire ML pipeline, affecting every stage from data preparation to model training and evaluation.

The primary objective of this study is to build a Machine Learning model capable of detecting dropout students. Driven by the inherent challenges of this domain, we aim to assess how different student representation strategies and class imbalance affect the performance of these predictive models. To ground this analysis in a practical setting, our study begins with a raw, unlabeled dataset featuring a longitudinal structure, where each student is represented by multiple observations, composed of academic features typically available in most HEIs. We then introduce our approaches to prepare the dataset, including a methodology to label students for supervised learning, feature engineering, and considerations for managing high-cardinality categorical variables, a frequent characteristic of academic datasets. Finally, we provide an interpretability analysis, identifying the factors associated with dropout. The contributions of this study extend beyond the educational domain, as the proposed approaches can be adapted to other classification problems with similar challenges.

2 Predicting Dropout Students Task

We address the task of predicting student dropout in higher education as a binary classification problem, which consists of discriminating between two possible outcomes: retention or dropout. This classification task can be formally conceptualized as a hypothesis testing problem. The null hypothesis (H_0) represents the default state of student retention, which does not require institutional action. The alternative hypothesis (H_1) represents student dropout, the critical event that the predictive model is designed to identify. Evidence supporting H_1, leading to the rejection of H_0, provides the necessary empirical evidence to implement preventive measures and prevent students from dropping out.

The statistical hypotheses are formally mapped to the class labels, where H_0 (retention) corresponds to the negative class (N) and H_1 (dropout) defines the positive class (P). A confusion matrix is used to quantify the four possible classification scenarios that arise from these definitions:

(1) **True Positives (TP):** *dropped* sample that is correctly predicted as *dropped* (correct rejection of H_0).
(2) **False Negatives (FN):** *dropped* sample that is incorrectly predicted as *retained* (failure to reject H_0).
(3) **False Positives (FP):** *retained* sample that is incorrectly predicted as *dropped* (incorrect rejection of H_0).
(4) **True Negatives (TN):** *retained* sample that is correctly predicted as *retained* (correct non-rejection of H_0).

Given that FP and FN errors carry distinct institutional consequences, we adopt a terminology that better reflects their practical impact. An FP error is characterized as a *false alarm*, as it involves incorrectly classifying a retained student as a potential dropout, leading to unnecessary interventions and inefficient use of institutional resources. Conversely, an FN error represents a *missed opportunity*, where a student at risk of dropping out is incorrectly classified as retained, resulting in a lost opportunity to provide timely support. The central challenge for predictive models, therefore, is to minimize these *missed opportunities* without generating an excessive number of *false alarms*.

2.1 Metrics of Interest

The metrics used to evaluate model performance must be sensitive to class imbalance to prevent the true performance of the model from being masked by the large number of negative class examples (Luque et al. 2019). A more informative approach is to evaluate performance by examining the distinct institutional impacts of *false alarms* and *missed opportunities*. For this reason, our analysis focuses on metrics that effectively quantify this trade-off.

2.1.1 Threshold-Based Metrics. The model evaluation is based on a specific decision threshold. Recall (Equation 1) is the most critical metric in this context (Fernández-García et al. 2021), as it measures the proportion of students who actually dropped out and were correctly identified by the model. A high recall value indicates that the model is effective in capturing the majority of dropouts, minimizing *missed opportunities*. This ensures that the institutional intervention strategy will reach most of the at-risk population.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{1}$$

On the other hand, Precision (Equation 2) evaluates the accuracy of positive predictions. It measures, among all students that the model classified as dropouts, the proportion who actually dropped out. High precision is crucial to ensure the efficient use of institutional resources. A low precision value means that many of the students flagged as at-risk are actually *false alarms*, which results in resources being spent on unnecessary interventions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

While recall and precision focus on the ability to correctly handle the positive class, Specificity (Equation 3) evaluates performance on the negative class. This metric quantifies the proportion of retained students who are correctly identified by the model. A high specificity value indicates that the model is effective at correctly recognizing students who are not at risk, minimizing *false alarms* and misallocation of institutional resources.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

The F1-score (Equation 4) combines recall and precision into a single metric, which is particularly useful for navigating their inherent trade-off. It calculates the harmonic mean between the two measures, providing a single value that summarizes the performance. A high F1-score indicates that the model achieves an effective balance between minimizing *missed opportunities* (high recall) and simultaneously avoiding *false alarms* (high precision).

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

2.1.2 Overall Performance Metrics. Previous metrics are useful, but they evaluate performance at a single cut-off point. For a more robust evaluation, performance should be assessed independently of any specific threshold, especially when the objective is to compare distinct ML models. The difference can be understood through an analogy: the F1-score provides a snapshot of performance, whereas a global metric offers the complete picture of the behavior of the model across all possible thresholds.

The Area Under the Receiver Operating Characteristic Curve (AUROC) is a global metric frequently referenced in the literature. However, its effectiveness is limited in scenarios with class imbalance. One of the main components of the ROC curve is the False Positive Rate (FPR), which is calculated as 1 − Specificity (Fawcett 2006). When the negative class dominates the dataset, specificity becomes an uninformative (or misleading) metric because its value is heavily influenced by the number of TNs (see Equation 3). When the number of TNs is high, the specificity value remains high even as the number of FPs increases (i.e., the model makes more errors in

identifying positive cases), thereby masking the poor performance of the model in identifying the minority class (Saito and Rehmsmeier 2015).

Due to this limitation, the Area Under the Precision-Recall Curve (AUPRC) is a more robust metric for this context (Davis and Goadrich 2006; Lee and Chung 2019; Saito and Rehmsmeier 2015; Vaarma and Li 2024), as it quantifies the trade-off between precision and recall, metrics that are calculated independently of the number of TNs. This makes the AUPRC a more informative and reliable indicator for evaluating model performance on imbalanced data. Furthermore, we are also interested in this metric because it consequently quantifies the trade-off between *false alarms* and *missed opportunities* more faithfully.

3 Related Work

Predicting student dropout has been a central challenge in numerous recent studies. These works can be divided between those that focus their analyses on primary and secondary education and those that concentrate on higher education. In this analysis, we examine how these studies represent the students' trajectories, whether they use only data commonly available at most HEIs (termed academic features), and what techniques they apply to address class imbalance.

A relevant study for temporal dropout analysis was conducted by Fernández-García et al. (2021), which proposed a cascade modeling approach to predict dropout across successive semesters, using 1,418 engineering student records from a Spanish public university based solely on academic data. A separate model was built for each semester, where each student was represented by a single observation whose variables reflected cumulative academic performance up to that point, and the prediction from the previous model was incorporated as a new feature for training the next one. To address class imbalance, the authors used a combination of the Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links methods. The results showed a progressive improvement in performance as information accumulated over the semesters, with the Support Vector Machine achieving 91.55% recall, 89.04% precision, and an F1-score of 90.28% in the fourth (last) semester.

Aouifi et al. (2024) proposed a hybrid architecture combining Long Short-Term Memory (LSTM) and Deep Neural Networks (DNN). The study used academic and demographic data from 4,424 students from a Portuguese institution, where each student was represented by a single observation, and temporal information from the first and second semesters was included as columns rather than as multiple observations per student. To address class imbalance, SMOTE was applied. In the proposed architecture, temporal features are first processed by an LSTM layer, and the outputs of this layer are concatenated with the remaining features and used as input for subsequent dense layers of the DNN. The hybrid model achieved an F1-score of 89.00%, with 95.00% precision and 83.00% recall. The authors also evaluated a standalone LSTM model, achieving a comparable F1-score of 88.00%. Additional models were evaluated for comparison, including a Multilayer Perceptron (MLP) with an F1-score of 83.00% and a Random Forest model with an F1-score of 77.00%.

Vaarma and Li (2024) adopted a different temporal perspective, creating a new predictive model for each month based on demographic, academic, and engagement data from 8,813 students at a Finnish university. By evaluating performance monthly over 60 months, the best-performing model, Categorical Boosting (CatBoost), achieved an average AUPRC of 0.8530, with an average recall of 47.00%, precision of 81.00%, and F1-score of 59.00%. The study revealed the dynamic nature of dropout, showing that the importance of features changes over time and simultaneously identifying engagement data as one of the strongest predictors.

Complementarily, Song et al. (2023) presented a comparison of four strategies to handle the longitudinal nature of academic data, using a dataset of academic and demographic records from a South Korean university. The models were trained with (i) only the first semester observation, (ii) the last semester observation, (iii) the mean of historical values, and (iv) the median of historical values. To address class imbalance, they applied SMOTE. The approach based on the mean of historical values combined with tree-based classifiers and gradient boosting

methods achieved the best performance, with Light Gradient Boosting (LightGBM) Machine reaching 78.00% recall, 81.00% precision, and an F1-score of 79.00%. The authors concluded that aggregating students' data from previous semesters up to the current point, by averaging their historical values, yielded the best predictive performance, aligning with the results reported in other studies.

Nagy and Molontay (2024) focused on model interpretability, employing Explainable Artificial Intelligence (XAI) techniques to explain predictions at both global and individual levels. The study used pre-enrollment academic data from 8,508 students at a public Hungarian university, where each student was represented by a single observation. Differing from prior research, they considered graduation as the target outcome. The data imbalance was not addressed, but it's important to note that the positive class in this case was the majority class. CatBoost achieved the best performance, with an AUROC of 0.7740, but its main contribution was the interpretation of model predictions using methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). The global analysis showed that the average high school grade and the number of years between the end of high school and university enrollment were the most relevant features. In addition, the authors demonstrated how SHAP and LIME can be applied to explain the importance of each feature at the individual level, providing stakeholders with clear and interpretable information to support personalized interventions.

Rabelo and Zárate (2025) proposed a different approach by restricting the dataset to students who had dropped out before completing 85% of the program, explicitly excluding students who never dropped out, using academic data from a Brazilian private university. The authors applied a domain-driven data preparation and variable analysis process to reduce the original set of variables. To address class imbalance, random undersampling was employed, resulting in a final training dataset with 762 instances. Using a model with 18 features, an ensemble based on Decision Trees, Logistic Regression, and a Multi-Layer Perceptron achieved the best performance, reaching 90.50% recall, 97.30% precision, and an F1-score of 93.80%. By restricting the analysis to dropout cases and excluding students with more than 85% program completion, the study adopts a targeted problem formulation that differentiates it from related works.

Expanding the scope to dropout prediction in secondary education, previous research has identified challenges similar to those encountered in higher education. Psyridou et al. (2024) explored the early prediction of secondary education students in Finland, using primary school data. Two models were created, where each student was represented by a single observation summarizing their information until Grade 6 and Grade 9, respectively. The dataset comprised 311 features describing academic performance, cognitive skills, family background, and other factors. To address class imbalance, they used random undersampling. The Random Forest model trained with data up to Grade 9 achieved an AUROC of 0.6500, recall of 59.90%, precision of 58.20%, and an F1-score of 57.30%, whereas the model using data only up to Grade 6 reached an AUROC of 0.6100, recall of 58.70%, precision of 56.90%, and an F1-score of 55.40%. These results suggest that the risk factors associated with dropout in secondary education can be detected early in the school trajectory, and that predictive performance improves as more recent data, closer to the time of dropout, are incorporated. A closer look at the most relevant features reveals that, among the twenty most important predictors in the Grade 9 model, nineteen were academic variables and one reflected behavioral aspects, underscoring the predominance of academic indicators on dropout prediction.

Krüger et al. (2023) created models by school stage, including preschool, basic school, and secondary school, as well as models at different points in the year, aggregating student information into a single observation following the same approach used in related works. The data were collected from 19 Brazilian schools and comprised 137 academic and demographic features. In the analysis by educational stage, the best result was obtained for secondary education, with Extreme Gradient Boosting (XGBoost), achieving an AUPRC of 0.6878, recall of 72.78%, precision of 92.28%, and an F1-score of 81.38%. When evaluated at different points in the year, the best performance was achieved in the third (last) trimester, reaching an AUPRC of 0.8950, recall of 93.93%, precision of 95.23%, and an F1-score of 94.58%. In line with other studies, the results indicated that information

aggregated in the last trimester generated better performance, and the feature importance analysis revealed that academic features were the most relevant.

In order to optimize an early warning system for school dropout, Lee and Chung (2019) focused on addressing the challenge of data imbalance using SMOTE. The authors used an academic dataset of 165,715 high school students from South Korea, where each student was represented by a single observation corresponding to one academic year. Each classifier was trained in two distinct ways, one using the original (imbalanced) dataset and another using the training dataset after applying SMOTE. The results showed varying effects on performance for each classifier. For the Boosted Decision Tree model, applying SMOTE resulted in a performance decline, causing the AUPRC to decrease from 0.8980 (without SMOTE) to 0.7240 (with SMOTE). Conversely, for the Random Forest model, applying SMOTE led to a slight improvement, increasing the AUPRC from 0.6340 to 0.6430. The authors also emphasized the limitations of the ROC curve in imbalanced datasets, as all models achieved excellent AUROC performance (ranging from 0.9860 to 0.9900). Compared with the PR curve metrics, it became clear how the ROC curve was masking the errors of the models in handling the minority class.

The literature review highlights two critical methodological aspects in dropout prediction. First, students are typically represented by a single observation, but studies diverge in how this representation is defined. Some develop models at different temporal levels, such as monthly or per-semester, while others encode information from multiple samples as new features or aggregate all available samples into a single representation per student. Second, class imbalance is a recurrent characteristic. While some works address it through data balancing techniques, others prefer to leave the natural distribution unchanged. Table 1 summarizes the key findings.

Table 1. Summary of related work.

Study	Academic Features Only	Data Balancing	Classifier	F1-score	AUC
Rabelo and Zárate (2025)	✓	Undersampling	Ensemble	93.80%	-
Fernández-García et al. (2021)	✓	SMOTE + Tomek Links	Support Vector Machine	90.28%	-
Lee and Chung (2019)	✓	SMOTE	Boosted Decision Tree	-	0.8980
Nagy and Molontay (2024)	✓	-	CatBoost	-	0.7740
Song et al. (2023)	-	SMOTE	LightGBM	79.00%	-
Aouifi et al. (2024)	-	SMOTE	LSTM-DNN	89.00%	-
Psyridou et al. (2024)	-	Undersampling	Random Forest	57.30%	0.6500
Vaarma and Li (2024)	-	-	CatBoost	59.00%	0.8530
Krüger et al. (2023)	-	-	XGBoost	94.58%	-

4 Educational Data Source

The educational dataset used in this study comprises the academic records of a cohort of undergraduate students from the University of Caxias do Sul, a non-profit community university located in the southern region of Brazil. Data access was granted under institutional authorization, ensuring full compliance with Brazil's General Data Protection Law. The study specifically focuses on the segment of their academic trajectories from the first semester of 2021 through the second semester of 2024, including their enrollment status in the first term of 2025. The dataset is structured longitudinally, where each student is represented by one or more observations, one for each academic semester. All records pertain exclusively to students enrolled in on-campus (in-person) undergraduate programs. For students enrolled in more than one program, only the most recent enrollment was considered. In cases of simultaneous enrollments, the enrollment corresponding to the lowest completion rate was selected. To protect individual privacy, all processing procedures included anonymization of student identifiers and the removal of personally identifiable information (PII). The dataset is intentionally composed of academic features commonly available at most HEIs. Table 2 provides a summary of the 29 initial dataset features.

Table 2. Initial dataset features.

Type	Feature	Description
Numerical	id	unique identifier of the student
	program_id	unique identifier of the student's academic program
	program_name	name of the academic program in which the student is enrolled
	age	student age
	semester	curricular semester of the student in the program (1–10)
	elapsed_terms	number of academic terms elapsed since admission
	start_year	calendar year in which the student's semester started (2021–2024)
	end_year	calendar year in which the student's semester ended (2021–2024)
	start_term	term of the year in which the student's semester started (1 or 2)
	end_term	term of the year in which the student's semester ended (1 or 2)
	admission_year	year in which the student was admitted to the HEI (1996–2024)
	admission_term	academic term of admission to the HEI (1 or 2)
	completion_rate	percentage of the degree program completed
	months_between	number of months between finishing secondary education and admission
	debt_amount	total outstanding debt with the HEI
	debt_negotiations	number of debt negotiations between the student and the HEI
	enrolled_load	total load of enrolled courses in hours
	absences	number of recorded absences in hours
	enrolled_courses	number of enrolled courses
	canceled_courses	number of canceled courses
	failed_courses_by_grade	number of courses failed due to grades
	failed_courses_by_absence	number of courses failed due to absences
	gpa	grade point average (0–10)
Categorical	degree_type	type of degree program
	enrollment_type	type of enrollment
	admission_type	type of admission
	external_transfer_request	whether the student applied for a transfer to another HEI
	course_request	whether the student requested new courses
	academic_support	whether the student received academic support

4.1 Data Cleaning and Target Definition

Our first step was to ensure consistency. This process involved removing inconsistent samples and handling missing values by either removing or replacing them when possible. Two methodologies were used for this treatment (one for removal and one for replacement), both grounded in the understanding that each student is represented by one or more observations:

- **Ensuring sample integrity**: if any observation for a student is removed due to inconsistencies, all other observations for that same student must also be discarded. This rule treats each student as an indivisible analytical unit, preventing fragmented trajectories.
- **Ensuring temporal consistency**: prohibits the use of future information to impute or modify past records. This constraint is essential to prevent data leakage and to accurately simulate a real-world predictive scenario where only past information is available.

Having achieved consistency, we proceeded to define the target label, since the original dataset did not explicitly indicate the outcome (e.g., dropped or retained). Formally, let $u_{i,j}$ and $w_{i,j}$ denote the start and final academic terms of the j-th chronological observation for student i, respectively. Furthermore, let $T_{\max}$ represent the final academic term within the observation window, and let R_{post} be the set of student IDs known to be enrolled in the term immediately following the window $T_{\max} + 1$. The label $y_{i,j}$ is determined by checking for enrollment continuity. An observation is labeled *retained* if a subsequent enrollment record k exists within the dataset (i.e., $u_{i,k} = w_{i,j} + 1$), or if the observation ends exactly at the boundary of the observation window ($w_{i,j} = T_{\max}$) and the student is known to be active afterward ($i \in R_{\text{post}}$). This is formalized in Equation 5.

$$y_{i,j} = \begin{cases} \text{retained}, & \text{if } (\exists k \text{ such that } u_{i,k} = w_{i,j} + 1) \text{ or } (w_{i,j} = T_{\max} \text{ and } i \in R_{\text{post}}) \\ \text{dropped}, & \text{otherwise} \end{cases} \tag{5}$$

The definition in Equation 5 leads to two considerations. First, it may result in students with multiple *dropped* labels. To focus the analysis on the first dropout event, these students were subsequently removed. Secondly, the inclusion of the R_{post} set is essential to correctly classify students who are still active at the end of the observation window. Without this information, these students could not be labeled, as they lack subsequent observations.

4.2 Dataset Analysis

The resulting dataset comprises 40,168 samples corresponding to 17,788 students. Each student is represented by at least one and at most n observations, where n is limited by the observation window (in this study, 8 semesters). The number of samples will determine which case type the student belongs to, and the cases can be divided according to academic status:

- **Fully retained**: *retained* samples only, with the number of samples per student ranging from 1 to n.
- **Dropped out**: a *dropped* sample, preceded by k *retained* samples, where k ranges from 0 to $n - 1$.

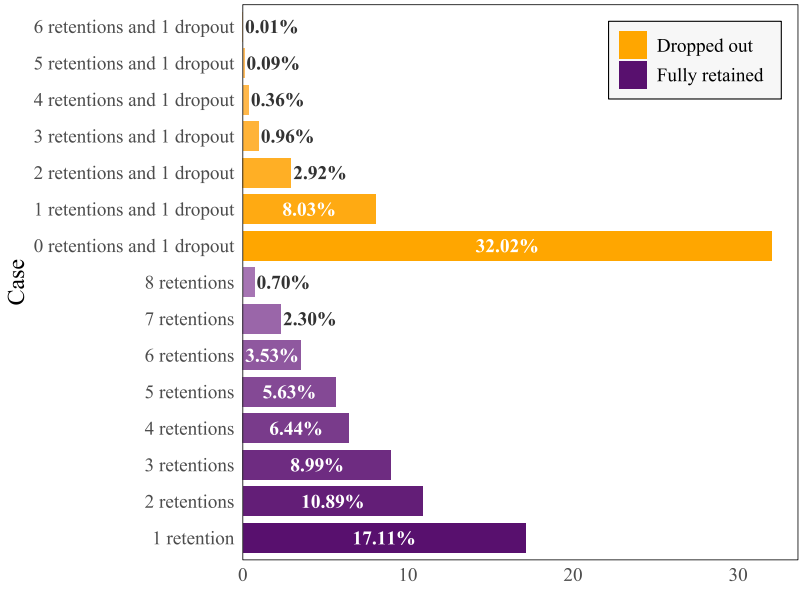

Fig. 1. Student distribution across different case types.

Figure 1 shows the distribution of students by their corresponding cases. The distribution reveals that cases with a single observation are the most frequent, both for students who remained retained (1 retention - 17.11%) and for those who dropped out (0 retentions and 1 dropout - 32.02%). These retained students are first-semester students observed only at the last semester of the dataset, while dropout students are likewise recent entrants, but they can be observed in different semesters, as dropout is defined by discontinuation after the first semester. It is also noted that students with continuous retention tend to have cases with a higher number of samples. In contrast, dropping out after the completion of one curricular semester is less common (8.03%) and becomes rare after completing two curricular semesters ($\leq$ 2.92%), indicating that, predominantly, students take no more than one or two academic semesters to decide on dropping out.

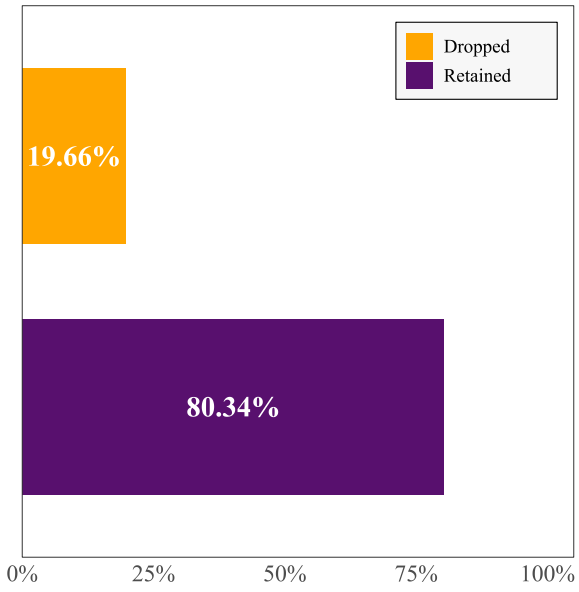

Fig. 2. Class distribution.

The dataset is imbalanced (Figure 2), meaning that one class contains substantially more examples than the other. In the dropout prediction task, the minority class is precisely the one we aim to detect. In our case, samples labeled as dropped represent only 19.66% of the dataset, whereas retained samples account for 80.34%. This corresponds to an approximate 1:4 ratio, indicating a substantial imbalance between the two classes, which is consistent with Brazilian higher education statistics. This imbalance is further influenced by the longitudinal nature of the dataset, since all samples that precede the student's final term are labeled as retained, with the dropped label occurring only at the end of the trajectory.

5 Proposed Approach

The proposed approach can be divided into three components, as illustrated in Figure 3: categorical feature processing, numerical feature processing, and classification. The categorical and numerical processing stages are executed in parallel, and their outputs are subsequently integrated to support the final classification decision.

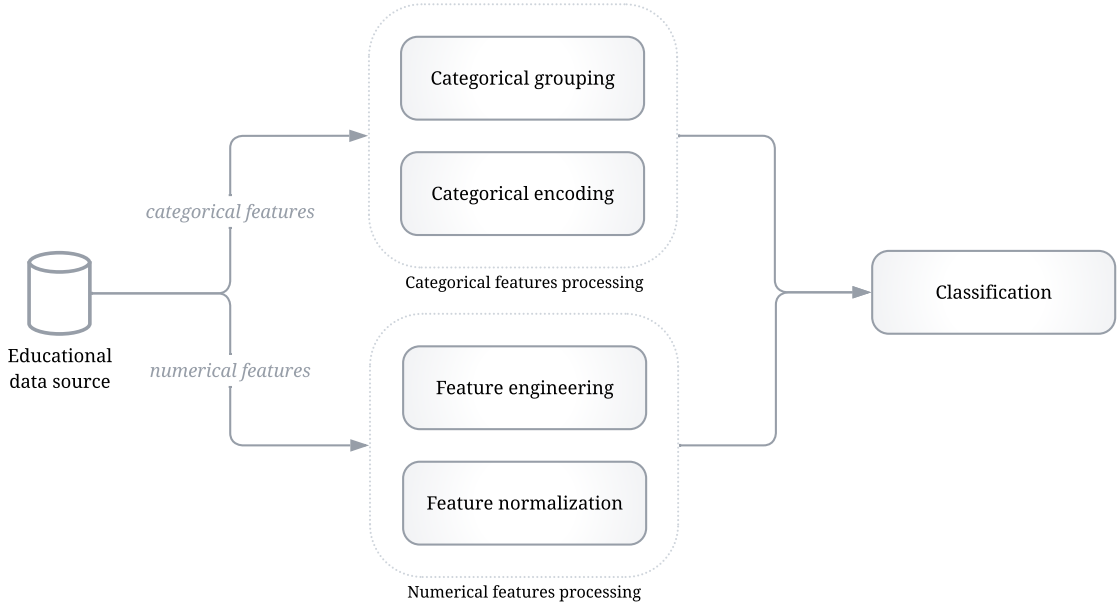

Fig. 3. Pipeline of the proposed approach for data processing.

5.1 Categorical Features Processing

The strategy for encoding a categorical variable is determined by its cardinality (K), defined as the number of unique categories it contains. Variables with only two categories ($K = 2$) are represented by a single binary feature, typically with values of 0 and 1 or -1 and 1 (Hastie et al. 2009). While this straightforward method is suitable for most categorical features in our dataset, a different approach is required for variables with more than two categories ($K > 2$). For these, such as *degree type* ($K = 4$), *enrollment type* ($K = 9$), and *admission type* ($K = 29$), a common strategy is to expand them into binary features, with one-hot encoding generating K and dummy coding $K - 1$ binary variables. However, the effectiveness of these expansion methods is highly dependent on the cardinality of the variable and the distribution of its categories.

Categorical variables with high cardinality and rare categories can significantly increase the dimensionality of the dataset and simultaneously introduce near-zero variance predictors, as the features derived from rare categories will consist almost entirely of zeros (Kuhn and Johnson 2019). Among the variables with $K > 2$, *degree type* ($K = 4$) is the only one that would not lead to a considerable dataset expansion, as its four categories show a balanced distribution across the dataset. On the other hand, *enrollment type* ($K = 9$) and *admission type* ($K = 29$) would not only expand the dataset by 36 new features combined, but 22 of those would be near-zero variance predictors. This is a result of the presence of many rare categories in these features, since only five of *enrollment type* and ten categories of *admission type* account for more than 98.07% and 94.08% of the samples, respectively. Therefore, an alternative strategy is required for both variables.

To devise a more effective encoding strategy, a closer examination of the conceptual structure of *enrollment type* and *admission type* is necessary. The *enrollment type*, which refers to the different registration models offered by the HEI, reveals that several categories are essentially subtypes of broader ones. For example, the *custom fixed-based* option is a minor variation of the main *fixed-based* option, yet it is treated as entirely separate. A similar pattern is observed in *admission type*, which represents the various pathways for student admission.

Categories that represent essentially the same phenomenon, such as those referring to reentry students, are often subdivided into finer distinctions, for example *reentry to the same program* and *reentry to a different program*. Similarly, scholarship admissions are split into categories for PROUNI (University for All Program) and FIES (Student Financing Fund), two of Brazil's main federal funding programs. This fragmentation explains the large number of rare categories, as many of them represent conceptually coherent groups that have been separated due to minor terminological distinctions.

A conventional approach to handling high-cardinality categorical variables is to group rare categories into an *other* group, typically based on a frequency criterion (Kuhn and Johnson 2019). The authors themselves caution that such pooling must be sensible, but in practice, the direct application of an automatic cutoff may overlook this fundamental premise. In our context, this indiscriminate approach would be counterproductive, as it would group conceptually related subcategories of *admission type* and *enrollment type*, leading to a loss of information.

Our approach is designed to address this challenge, operationalizing the sensible pooling principle through a process called *categorical grouping*, which is formalized as follows. Starting with the set $C = \{c_1, \ldots, c_K\}$ of all original categories, the first step uses domain knowledge to define M disjoint conceptual subsets ($G_a \cap G_b = \emptyset, \forall a \neq b$), where each G_a groups conceptually similar categories associated with a unified class name L_a. Next, the residual categories, those not belonging to any conceptual group, are assigned to a set C_{other}, defined by the set difference (Equation 6). The final transformation is given by a mapping function $f(c)$, which assigns a representative label L_a to each original category $c \in G_a$ (Equation 7).

$$C_{\text{other}} = C - \left(\bigcup_{a=1}^{M} G_a \right) \tag{6}$$

$$f(c) = \begin{cases} L_a & \text{if } c \in G_a, \quad \text{for } a = 1, \ldots, M \\ \text{Other} & \text{if } c \in C_{\text{other}} \end{cases} \tag{7}$$

For example, all subcategories in G_{reentry} are mapped to the new label *reentry*. This procedure ensures that the category pooling is always conceptually valid, and the outcome is a more concise and robust set of predictors that maintains the conceptual integrity of the original data.

A naive application of dummy coding to the original *degree type* ($K = 4$), *enrollment type* ($K = 9$), and *admission type* ($K = 29$) features would have expanded the dataset by 39 binary features. Instead, by applying the *categorical grouping* approach, we reduced the cardinality of *enrollment type* to $K = 6$ and *admission type* to $K = 7$, while *degree type* remained unchanged ($K = 4$). As a result, the subsequent dummy encoding of these transformed variables yielded only 14 new binary features, representing a 64.10% reduction compared to the 39 initially expected. Crucially, this optimization was achieved without discarding any information, as all original categories were preserved through intelligent regrouping.

5.2 Numerical Features Processing

The temporal nature of the student representations motivates the development of features that describe their academic progression over time. Inspired by Krüger et al. (2023), who used a grade sum and a trimester variation delta, we adopted a similar approach. We applied these two metrics (sum and delta) to the grades and additionally extended them to the course load. Formally, let $v_{i,j}$ be the value of a variable associated with student i at the j-th observation, where j indexes the observations chronologically. We define the cumulative value $V_{i,j}$ as the sum of the variable from the first observation ($k = 1$) up to the j-th, as given by Equation 8.

$$V_{i,j} = \sum_{k=1}^{j} v_{i,k} \tag{8}$$

The corresponding delta value $\Delta v_{i,j}$, defined in Equation 9, captures the sequential variation of the variable. This metric calculates the change between the variable's value at the current observation j and its value at the immediately preceding observation $j-1$. For the initial observation ($j = 1$), where no prior data point is available for comparison, the delta is defined as 0.

$$\Delta v_{i,j} = \begin{cases} 0, & j = 1, \\ v_{i,j} - v_{i,j-1}, & j > 1 \end{cases} \tag{9}$$

Beyond these cumulative and sequential metrics, we engineered features that capture multiple dimensions of the student's journey, including academic achievement, engagement, and progression. The *success rate* (Fernández-García et al. 2021), denoted by $r_{i,j}$, expresses the proportion of courses in which student i achieved approval in their j-th observation, and it is defined by

$$r_{i,j} = \frac{a_{i,j}}{e_{i,j}}, \tag{10}$$

where $a_{i,j}$ represents the count of approved courses for student i at observation j, and $e_{i,j}$ is the corresponding total count of enrolled courses.

To further characterize student engagement, we define the *absences rate*, denoted by $b_{i,j}$, as a relative measure of attendance behavior. We note that the absolute number of absence hours becomes meaningful only in relation to the total enrolled course load. This feature can be defined as

$$b_{i,j} = \frac{h_{i,j}}{l_{i,j}}, \tag{11}$$

where $h_{i,j}$ represents the total absence hours, and $l_{i,j}$ is the corresponding total course load in hours.

To capture the pace of academic advancement, we formulated the *progression ratio*, denoted by $p_{i,j}$. Let $t_{i,j}$ represent the student's official curriculum semester, $d_{i,j}$ be the total number of elapsed terms (duration) since their admission, and $t_{i,1}$ be their first observation semester. The ratio compares the actual number of terms elapsed ($d_{i,j}$) with the number of curriculum semesters the student has advanced since their entry point ($t_{i,j} - t_{i,1} + 1$). This effectively distinguishes between regular ($p_{i,j} = 1$), delayed ($p_{i,j} > 1$), and accelerated ($p_{i,j} < 1$) academic trajectories, as defined by

$$p_{i,j} = \frac{d_{i,j}}{t_{i,j} - t_{i,1} + 1}. \tag{12}$$

This feature engineering process introduced seven new features based on the five metric types previously defined. This total resulted from applying the cumulative ($V_{i,j}$) and delta ($\Delta v_{i,j}$) metrics to both *gpa* and *course load*, combined with the single-variable metrics (*success rate*, *absence rate*, and *progression ratio*). The creation of these derived features allowed for the removal of six original variables whose information was effectively encapsulated by the new metrics. Specifically, we removed features such as *elapsed terms*, *absences*, *enrolled courses*, *canceled courses*, *failed courses by absences*, and *failed courses by grades*. Furthermore, a second set of features was removed, as they were either irrelevant for the classification task or could hinder the generalization capabilities of the model. This set included identifiers (e.g., *id*, *program id*, *program name*) and specific temporal markers (e.g., *start year*, *end year*, *start term*, *end term*, and *admission year*). The final dataset expanded to 33 features after the processing of categorical and numerical features. This represents an increase of 4 features from the 29 in the original dataset. Table 3 presents an overview of the final feature set.

Table 3. Final dataset features.

Type	Features		
Numerical	absences_rate	debt_amount	months_between
	admission_term	debt_negotiations	progression_ratio
	age	delta_enrolled_load	semester
	completion_rate	delta_gpa	success_rate
	cumulative_enrolled_load	enrolled_load	
	cumulative_gpa	gpa	
Binary	academic_support	admission_transfer	enrollment_credit_based
	admission_prior_degree	course_request	enrollment_days_based
	admission_program_change	degree_medical	enrollment_fixed_based
	admission_reentry	degree_teaching	enrollment_flat_fee
	admission_regular	degree_technology	external_transfer_request
	admission_scholarship	enrollment_course_based	

As the final processing step, the numerical features were standardized. The categorical features, already encoded as binary indicators, did not require this transformation. The z-score approach, also referred to as standard scaling (Amorim et al. 2023), was applied as defined in Equation 13.

$$z = \frac{x - \overline{x}}{s} \tag{13}$$

where x is the observed value, ($\overline{x}$) is the mean of the training set, and (s) is the corresponding standard deviation. The same statistics were then applied to transform all data subsets.

5.3 Classification

Supervised learning consists of building a prediction model that maps a set of input features to an outcome based on labeled data (Hastie et al. 2009). In our context, this corresponds to the classification stage, where the model combines the previously processed features to determine whether a student is more likely to have dropped out or retained, assigning one of these labels to each sample based on the learned decision boundaries. To implement this classification stage, three classifiers were selected following previous work in the literature:

- **Random Forest**: proposed by Breiman (2001), is an ensemble method that builds a forest of decision trees, designed to be significantly more robust and diverse than individual models. First, through the *bootstrap aggregating* (bagging) process, each tree in the forest is trained on a different subsample of the training data. Second, when constructing each tree, whenever a node needs to be split, the algorithm does not evaluate all available features; instead, it selects a random subset of features and searches for the best split within that subset. This process results in a robust ensemble model, where new data is evaluated by all independent trees, and the final output is determined by majority vote or averaging.
- **XGBoost**: proposed by Chen and Guestrin (2016), is an optimized implementation of the Gradient Boosting algorithm over decision trees. In contrast to Random Forest, where trees are built in parallel and independently, XGBoost uses a process called *boosting*, building the trees sequentially. Each new tree is trained to correct the errors left by the current model by focusing on the instances that are hardest to predict. What distinguishes it from traditional boosting methods are its system optimizations, its integrated regularization

controls to prevent overfitting, and its native ability to handle missing data. The resulting model is a highly optimized ensemble where the final prediction is calculated as a weighted sum of the trees.

- **Multi-Layer Perceptron**: originating as an extension of the Perceptron proposed by Rosenblatt (1958), is a feedforward artificial neural network (Goodfellow et al. 2016), whose objective is to overcome the main limitation of the Single-Layer Perceptron, which can only solve linearly separable problems. Its architecture introduces one or more hidden layers positioned between the input and output layers, utilizing non-linear activation functions that allow the network to learn complex and non-linear relationships. To train this architecture, the backpropagation algorithm (Rumelhart et al. 1986) is commonly employed. This algorithm calculates the prediction error at the output, propagating it back through the layers and iteratively adjusting the weights to minimize the prediction error. Once trained, the network effectively becomes a complex non-linear function, where new data is passed forward through the optimized weights to generate a final prediction at the output layer.

Random Forest and XGBoost were chosen to represent tree-based ensembles, given their strong effectiveness demonstrated in the literature for this task. XGBoost, in particular, serves as the representative for modern boosting variations (such as CatBoost and LightGBM) due to its well-known optimizations. As these two models are non-parametric, the Multi-Layer Perceptron was selected as the parametric model, serving as a classic benchmark for neural networks. This selection also allows for a direct comparison between the behavior of the non-parametric models and a classic parametric approach.

6 Results

We begin by describing the experimental setup used for all comparisons. Next, we present our analysis in three main stages: (1) an evaluation on longitudinal and (2) aggregated datasets, followed by (3) a balanced data analysis.

6.1 Experimental Setup

The dataset was partitioned into training (2/3) and test (1/3) subsets. This split was carefully structured to ensure a robust evaluation by enforcing three fundamental constraints:

(1) **Student-level Independence**: all samples belonging to a specific student were allocated entirely to either the training or the test set. This constraint is crucial to prevent data leakage, ensuring that the classifier is evaluated on unseen students.
(2) **Class Proportion Preservation**: the split was stratified to maintain the original imbalanced class distribution (detailed in Figure 2) in both subsets. This ensures that the classifier is trained and tested on the same class proportions found in the original dataset.
(3) **Case Proportion Preservation**: following the same stratification principle, the proportions of case types (from Figure 1) were identically preserved. This ensures the classifier is exposed to all cases representatively.

All our objectives require comparing the performance between distinct models. We first ensured that all classifiers operated under optimal conditions by tuning their hyperparameters using a 10-fold cross-validation procedure, repeated five times, on the training data. The AUPRC is the chosen metric for this comparison, as it effectively summarizes the overall performance of a model across all decision thresholds. However, simply observing a difference in AUPRC scores is insufficient. The primary importance of a statistical significance test is to provide a rigorous method to determine whether an observed difference is statistically significant or simply due to random chance. This step is crucial to confidently claim that one model truly outperforms another.

A common approach to verify if two AUPRC measures are statistically different is to use the bootstrap method (Boyd et al. 2013; Efron and Tibshirani 1993), where m iterations are generated over the test sets. In each iteration, a replicate of the test set is created, the AUPRC measures are recorded for both models, and the difference between them is stored, generating an empirical distribution of these differences. This distribution allows for both the

generation of a confidence interval and the calculation of a p-value. Considering a significance level α (typically 0.05, corresponding to a 95% confidence level), the superiority of a model is considered statistically significant if the calculated p-value is less than or equal to α. For our analyses, we adopt $m = 10000$ iterations and $\alpha = 0.05$.

6.2 Evaluation on the Longitudinal Dataset

To establish a baseline performance, we first evaluated the three classifiers using the original longitudinal dataset, in which each student is represented by one or more observations (n), with $1 \leq n \leq 8$. The class proportion in this dataset corresponds to Figure 2. The comparative performance metrics are detailed in Table 4.

Table 4. Comparative performance of the classifiers on the longitudinal dataset.

Classifier	AUPRC	Recall	Precision	F1-score
XGBoost	**0.8403**	**67.21%**	**84.88%**	**75.02%**
Random Forest	0.8287	66.38%	82.64%	73.62%
Multi-Layer Perceptron	0.8211	64.97%	82.69%	72.77%

The XGBoost classifier achieved the best performance, with an AUPRC of 0.8403 and an F1-score of 75.02%. The Random Forest and Multi-Layer Perceptron models performed similarly, with AUPRCs of 0.8287 and 0.8211, and F1-score of 73.62% and 72.77%, respectively. When applying the bootstrap method, a statistically significant difference was observed between the performances of Random Forest and Multi-Layer Perceptron ($p \approx 0.0331 \leq \alpha$). The superior performance of both XGBoost and Random Forest suggests that their non-parametric nature was more effective at modeling the patterns than the Multi-Layer Perceptron (parametric). Furthermore, the fact that XGBoost outperformed Random Forest indicates that its specific architecture, based on gradient boosting, was more effective than the independent tree generation approach of Random Forest. Figure 4 further illustrates the superior performance of XGBoost over Random Forest and of Random Forest over the Multi-Layer Perceptron.

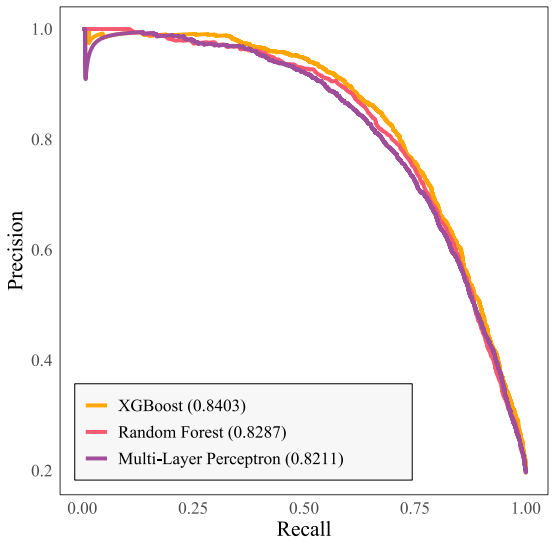

Fig. 4. Precision–Recall curves on the longitudinal dataset.

A clear trade-off is observed, where the models prioritize precision (e.g., 84.88% for XGBoost) over recall (67.21%). This behavior is reflected in the overall performance scores, such as the AUPRC of 0.8403 and F1-score of 75.02%. The models are generating few *false alarms* while at the same time having many *missed opportunities*. For the HEI, this is not an ideal situation. The benefit of wasting few resources on *false alarms* comes at the significant cost of failing to identify approximately one-third of the dropouts.

6.3 Evaluation on the Aggregated Dataset

We derived an alternative dataset from our longitudinal version by aggregating all available samples for each student into a single representative instance ($n = 1$). In contrast to Song et al. (2023), we applied different aggregation operations depending on the semantic nature of each feature, allowing the aggregation process to better preserve the underlying information. This formulation enables us to investigate the scenario where students are represented by a single sample and to directly compare the results with the longitudinal approach.

The aggregation operations applied to the longitudinal student data and their corresponding features are:

- **Sum**: computes the cumulative sum across all available observations up to the current semester. Applied to: *debt negotiations*, *enrolled load*.
- **Mean**: calculates the arithmetic mean of the values up to the current semester. Applied to: *absences rate*, *delta enrolled load*, *delta gpa*, *progression ratio*, *success rate*.
- **Weighted Mean**: computes the weighted mean across all available observations, using the count of non-cancelled enrolled courses as the weight, up to the current semester. Applied to: *gpa*.
- **Max**: selects the highest recorded value from all observations up to the current semester. Applied to: all binary features, *debt amount*.
- **Any**: selects any value from the student's observations. Applied to: *admission term*, *months between*.
- **Last**: selects the last recorded value up to the current semester. Applied to: *age*, *completion rate*, *cumulative enrolled load*, *cumulative gpa*, *semester*.

Figure 5 presents the class distribution obtained after applying the aggregation rules.

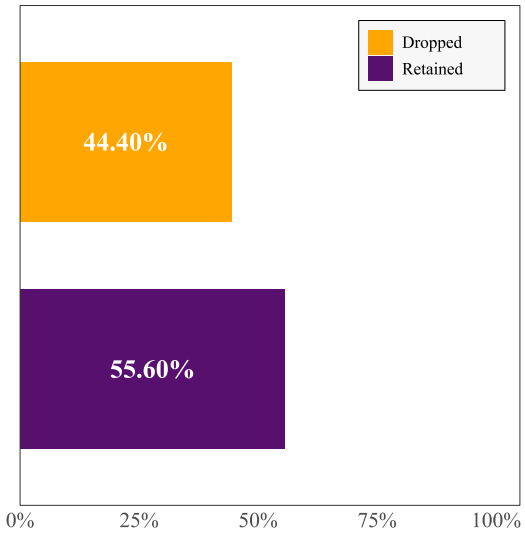

Fig. 5. Class distribution in the aggregated dataset.

According to Table 5, the XGBoost classifier achieved the best overall performance, with an AUPRC of 0.9262 and an F1-score of 83.26%. In contrast to the longitudinal approach, the significance test indicated that there was no statistically significant difference between the performances of Random Forest and Multi-Layer Perceptron ($p \approx 0.0764 > \alpha$). This result suggests that the non-parametric nature alone was not sufficient to provide a clear performance advantage over the parametric approach in this scenario. Instead, the superior performance achieved by XGBoost indicates that its gradient boosting mechanism was the main factor contributing to its advantage over the other classifiers. Figure 6 further highlights the superior performance of XGBoost, while the curves of Random Forest and Multi-Layer Perceptron exhibit closely aligned behaviors, reinforcing the absence of a statistically significant difference between their performances.

Table 5. Comparative performance of the classifiers on the aggregated dataset.

Classifier	AUPRC	Recall	Precision	F1-score
XGBoost	**0.9262**	**81.34%**	**85.26%**	**83.26%**
Random Forest	0.9128	81.61%	83.00%	82.30%
Multi-Layer Perceptron	0.9074	79.41%	83.40%	81.35%

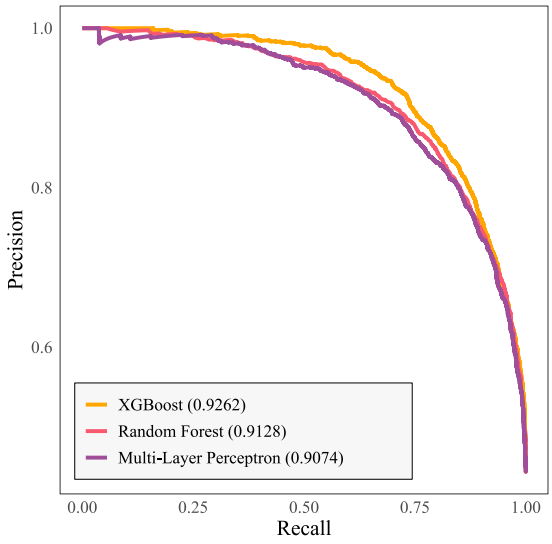

Fig. 6. Precision–Recall curves on the aggregated dataset.

Results on the aggregated dataset yielded a relative improvement of approximately 10.22% in AUPRC and 10.98% in F1-score. The substantial gain in F1-score indicates a more effective balance between precision and recall (*false alarms* vs. *missed opportunities*). For the HEI, this is an ideal scenario, as it represents an improved capability to identify dropout students while optimizing the allocation of institutional resources. This contrasts with the longitudinal approach, where the institution faced a trade-off between allocating additional resources to detect more students at risk or preserving efficiency at the cost of identifying fewer dropouts.

6.4 Evaluation on Balanced Datasets

Although the aggregated dataset achieved superior performance, it also presented a more favorable class distribution than the longitudinal dataset, making it unclear whether the observed improvement resulted from the aggregation process itself or from the reduced class imbalance. Therefore, balancing the data enables a fair comparison between the two approaches, regarding the trade-off between *false alarms* and *missed opportunities*.

Due to its predominant adoption in the literature, SMOTE (Chawla et al. 2002) was selected as the technique applied for data balancing. This method operates directly in feature space to create new synthetic examples for the minority class. Given a minority-class sample, the algorithm first identifies its k nearest neighbors, from which one is randomly selected. A new synthetic sample is then interpolated at a random point along the line segment connecting the original sample to its chosen neighbor. This is achieved by taking the difference between their feature vectors, multiplying that difference by a random number (between 0 and 1), and adding the result to the feature vector of the original sample. The result is a synthetic sample positioned along the line segment connecting the original instance and its neighbor, expanding the decision region for the minority class.

Table 6 summarizes the performance of XGBoost with and without SMOTE on both datasets, whereas Figure 7 presents the corresponding Precision-Recall curves.

Table 6. Classifier performance on the balanced datasets.

Dataset	Classifier	AUPRC	Recall	Precision	F1-score
Longitudinal	XGBoost	0.8403	67.21%	84.88%	75.02%
	XGBoost-SMOTE	0.8328	70.55%	79.56%	74.79%
Aggregated	XGBoost	0.9262	81.34%	85.26%	83.26%
	XGBoost-SMOTE	0.9282	83.81%	83.43%	83.62%

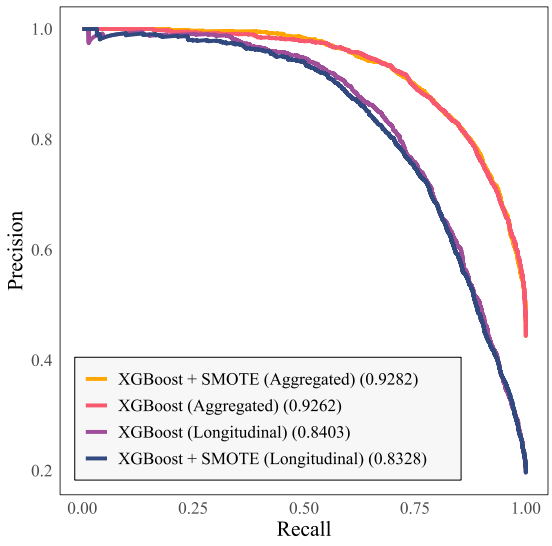

Fig. 7. Precision–Recall curves on the balanced dataset.

The application of SMOTE in the longitudinal dataset led to a small but statistically significant decrease in AUPRC (from 0.8403 to 0.8328, $p \approx 0.0004 \leq \alpha$). In contrast, in the aggregated dataset, the performance remained statistically unchanged (from 0.9262 to 0.9282, $p \approx 0.1373 > \alpha$). This outcome results from a trade-off in which SMOTE increases recall at the expense of precision, a behavior also observed by Tantithamthavorn et al. (2020), leading to no improvement in AUPRC. Similarly, Lee and Chung (2019) reported heterogeneous effects of SMOTE, with a performance decline for Boosted Decision Trees and an AUPRC improvement for Random Forest, although no statistical significance analysis was reported in their study.

While the increase in recall at the default threshold might seem beneficial for the HEI's goal of detecting more students, this perspective is misleading. The AUPRC is the critical metric as it represents the model's total predictive power across all possible thresholds, not just the default. A decrease in the overall AUPRC, as seen in the complete dataset, indicates the model is fundamentally less effective. Adjusting the threshold on this model will only select a different point on this inferior curve.

These results clarify the ambiguity previously observed when comparing the longitudinal and aggregated approaches, where it was unclear whether the performance gain was due to a more favorable class balance. Even under equally balanced conditions, the aggregated approach (AUPRC 0.9282) still substantially outperformed the longitudinal approach (AUPRC 0.8328). This provides clear evidence that the superior performance observed in the aggregated approach is not an artifact of class distribution, but rather a result of the aggregation process itself, which effectively reduces the noise introduced by the multiple samples per student.

6.5 Model Interpretation

To understand the factors that contribute to dropout, we used SHAP (Lundberg and Su-In Lee 2017), an approach based on game theory that provides a unified framework for interpreting the model predictions, which has also been used in the literature by Nagy and Molontay (2024). It assigns each feature an importance (a SHAP value) based on its average contribution to the output of the model. We present the interpretations from the XGBoost model for the aggregated approach, where the 20 most important features are displayed in Figure 8.

The SHAP global importance analysis reveals that the features created during the feature engineering process were crucial for dropout detection. The three most important features were *cumulative enrolled load*, *success rate*, and *cumulative gpa*. Collectively, these features show that students at a high risk of dropping out tend to accumulate a lower enrolled load, exhibit a lower success rate (reflected in more failures and cancellations), and maintain lower academic performance throughout their trajectory. From an institutional perspective, these results suggest that early-warning systems should prioritize monitoring students with consistently low success rates, reduced academic load accumulation, and low cumulative academic performance.

Furthermore, *progression ratio* and *delta gpa* are among the top 10 most relevant features. The analysis shows that dropouts not only exhibit lower *progression ratio* values, but also present low *delta gpa* values, which are strong indicators of dropout. The low values indicated by SHAP suggest that dropouts tend to progress more slowly through the curriculum, enrolling in fewer courses than the standard curriculum semester offers and consequently taking longer than expected to complete their studies. In addition, they exhibit a lower average grade variation than other students. In other words, they demonstrate more pronounced or consistent performance declines throughout their trajectory, which differentiates them from students who may also experience temporary negative fluctuations but ultimately remain retained.

Financial indicators also proved to be relevant. The *debt negotiations* and *debt amount* features reveal a clear pattern, where detected dropouts tend to have debts and, at the same time, show no negotiation activity. This combination of factors suggests that students with this profile represent a risk group, as this inertia in resolving the financial issue may be a strong behavioral indicator of general disengagement from the HEI. Therefore,

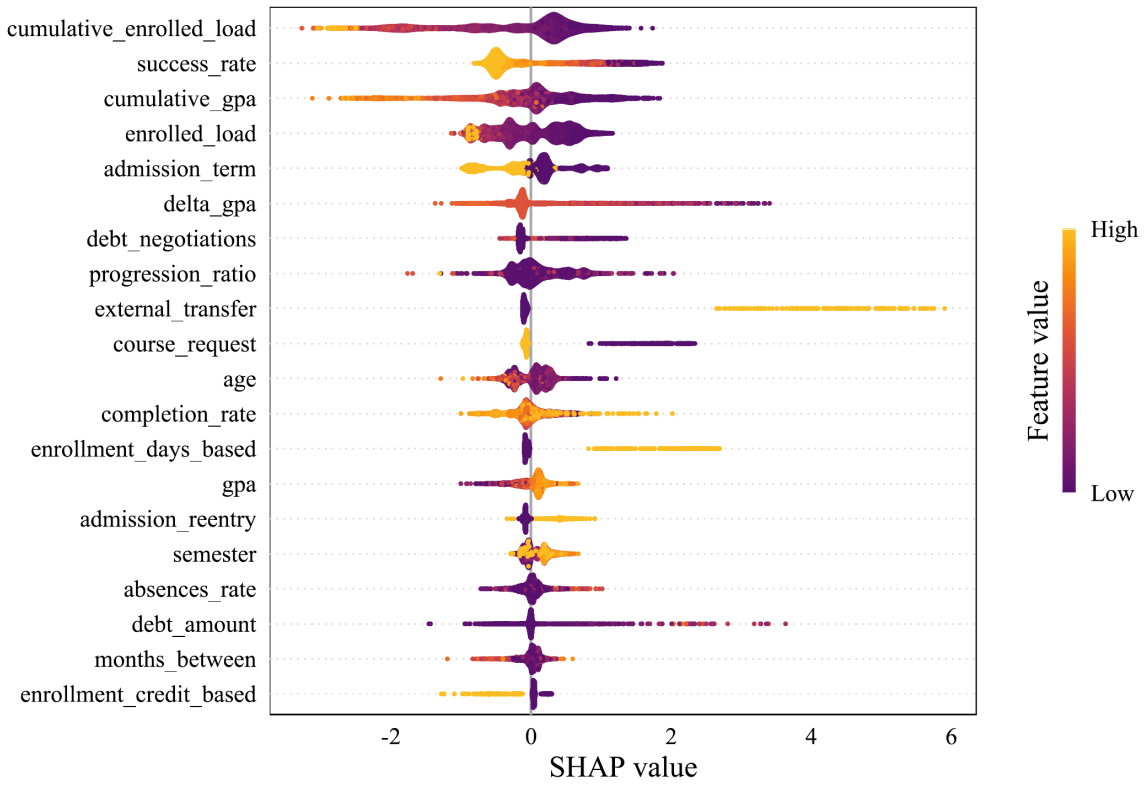

Fig. 8. Feature importance for the XGBoost classifier trained on the aggregated dataset.

implementing proactive policies to facilitate the renegotiation of these debts emerges as an intervention strategy with the potential to reduce dropout rates.

The analysis also identifies features that serve as clear warning signs for intervention and groups that do not require as much attention. In the first case, *absences rate* and *course request* emerged as important predictors associated with engagement. Students with a high rate of absence or low activity in course requests (whether for the current or upcoming semester) showed a significantly higher propensity for dropout. Even more critically, the *external transfer request* feature reveals that students who initiate an external transfer process have a high probability of dropping out, demanding immediate intervention. In contrast, the *months between* feature indicates that students who take longer to enroll in higher education after completing secondary education tend to present a lower probability of dropout. This pattern suggests that a longer transition period before entering an HEI may be associated with greater persistence throughout the academic trajectory.

Other features also indicate risks present from the beginning of the student's trajectory. The *enrollment days based* and *enrollment credit based* modalities (specific enrollment types at this HEI), student reentry (*admission reentry*), and admission in the first term of the year (*admission term*) proved to be strong indicators of dropout. These findings can direct the HEI to review the enrollment modalities with a higher tendency for dropout or, more immediately, to implement support actions for these students from the moment they are admitted.

Finally, the categorical grouping process proved to be crucial in processing the categorical features. Features like *enrollment days based* and *admission reentry* are particularly illustrative. Before this treatment, both represented rare or fragmented categories that, in a standard categorical variable treatment, could have been hidden in a generic *other* category or removed. Their presence among the top 20 most important features demonstrates the value of this technique in preserving relevant information that would otherwise have been lost. This reinforces the importance of a rigorous analysis of categorical features, which can contain valuable insights for HEIs.

7 Conclusions

In this study, we developed a ML model to detect student dropout in higher education, addressing the inherent challenges that the longitudinal data structure and class imbalance introduce to this task. To manage these complexities, our research analyzed and compared different student representation strategies and the use of a balancing technique to assess the impact of class imbalance on model performance.

The results showed that the proposed aggregation approach significantly improved the performance of all classifiers. When analyzing the XGBoost results in particular, the model trained on the aggregated dataset achieved a relative improvement of approximately 10.22% in AUPRC and 10.98% in F1-score compared to the longitudinal dataset. These improvements highlight a substantial gain in overall model performance, as well as a better balance between precision and recall metrics. This scenario is especially desirable for HEIs, as the model exhibits comparable levels of *false alarms* and *missed opportunities*. From a practical perspective, this balance supports more efficient use of institutional resources while maintaining the ability to identify dropout students.

Meanwhile, considering the inherently imbalanced nature of the dataset, the use of class balancing techniques did not improve the proposed approach. The results indicate that applying SMOTE did not lead to statistically significant gains in overall performance (AUPRC) or F1-score. Despite the observed increase in recall ($\approx 3.04\%$), the concurrent decrease in precision ($\approx 2.19\%$) led to no practical improvement in performance. In other words, the number of *missed opportunities* decreased, but this reduction was accompanied by an increase in *false alarms*. Importantly, the application of data balancing was essential to confirm that the performance gains observed between the longitudinal and aggregated approaches were driven by the sample aggregation strategy itself, rather than by differences in class proportions across the two scenarios.

The interpretability analysis reveals that dropout is not an isolated event, but a multidimensional phenomenon resulting from the combination of academic performance, engagement, and enrollment characteristics. Specifically, our findings show that at-risk students exhibit not only lower academic performance but also a declining trajectory, marked by slower academic progression, higher absence rates, and a consistent pattern of failures and course cancellations. We also identified a critical behavioral inertia regarding financial issues, where the lack of debt negotiation attempts serves as a strong indicator of broader institutional disengagement. Furthermore, the study underscores the importance of monitoring early warning signs, revealing that specific enrollment and admission types carry inherently higher or lower dropout risks from the beginning of the student's trajectory. Ultimately, these insights demonstrate how interpretable ML can serve as a valuable decision-support tool, providing empirical evidence for stakeholders to design targeted, data-driven interventions.

In summary, this study reinforces that the success of dropout prediction critically depends on how student data is prepared and represented. As future work, we suggest exploring more sophisticated ways to represent students that simultaneously make optimal use of their longitudinal data, as well as incorporating additional sources of information that may capture other dimensions related to student behavior and engagement. Furthermore, future studies could investigate strategies to translate predictive models into practical decision-support approaches, enabling institutions to better identify risk scenarios and support timely interventions. Ultimately, advancing in this direction can strengthen the role of ML as a reliable tool for supporting institutional policies aimed at promoting student success and reducing dropout rates.

References

C. Aina, E. Baici, G. Casalone, and F. Pastore. June 2022. "The Determinants of University Dropout: A Review of the Socio-economic Literature." *Socio-Economic Planning Sciences*, 79, (June 2022), 101102. doi:10.1016/j.seps.2021.101102.

B. Albreiki, N. Zaki, and H. Alashwal. Sept. 2021. "A Systematic Literature Review of Student's Performance Prediction using Machine Learning Techniques." *Education Sciences*, 11, 9, (Sept. 2021), 552. doi:10.3390/educsci11090552.

L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz. Jan. 2023. "The Choice of Scaling Technique Matters for Classification Performance." *Applied Soft Computing*, 133, (Jan. 2023), 109924. doi:10.1016/j.asoc.2022.109924.

H. E. Aouifi, M. E. Hajji, and Y. Es-Saady. Oct. 2024. "A Hybrid Approach for Early-identification of At-risk Dropout Students using LSTM-DNN networks." *Education and Information Technologies*, 29, 14, (Oct. 2024), 18839–18857. doi:10.1007/s10639-024-12588-0.

R. S. J. d. Baker and K. Yacef. Oct. 2009. "The State of Educational Data Mining in 2009: A Review and Future Visions." *Journal of Educational Data Mining*, 1, 1, (Oct. 2009), 3–17. doi:10.5281/zenodo.3554657.

K. Boyd, K. H. Eng, and C. D. Page. 2013. "Area Under the Precision-Recall Curve: Point Estimates and Confidence Intervals." In: *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013*. Lecture Notes in Computer Science. Vol. 8190. Ed. by H. Blockeel, K. Kersting, S. Nijssen, and F. Železný. Springer, Berlin, Heidelberg, 451–466. doi:10.1007/978-3-642-40994-3_29.

L. Breiman. Oct. 2001. "Random Forests." *Machine Learning*, 45, 1, (Oct. 2001), 5–32. doi:10.1023/A:1010933404324.

T. Cardona, E. A. Cudney, R. Hoerl, and J. Snyder. May 2023. "Data Mining and Machine Learning Retention Models in Higher Education: A Systematic Review." *Journal of College Student Retention: Research, Theory & Practice*, 25, 1, (May 2023), 51–75. doi:10.1177/1521025120964920.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. June 2002. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, 16, (June 2002), 321–357. doi:10.1613/jair.953.

T. Chen and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 785–794. doi:10.1145/2939672.2939785.

J. Davis and M. Goadrich. 2006. "The Relationship Between Precision-Recall and ROC Curves." In: *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*. ACM, Pittsburgh, PA, USA, 233–240. doi:10.1145/1143844.1143874.

B. Efron and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Vol. 57. Chapman & Hall/CRC, New York, NY. doi:10.1201/9780429246593.

T. Fawcett. June 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters*, 27, 8, (June 2006), 861–874. doi:10.1016/j.patrec.2005.10.010.

A. J. Fernández-García, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, and F. Sánchez-Figueroa. Sept. 2021. "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data." *IEEE Access*, 9, (Sept. 2021), 133076–133090. doi:10.1109/ACCESS.2021.3115851.

I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press. ISBN: 9780262035613.

T. Hastie, R. Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (2nd ed.). Springer Series in Statistics. Springer, New York, NY. doi:10.1007/978-0-387-84858-7.

INEP. 2024. *Higher Education Census — National Institute of Educational Studies and Research (INEP)*. Retrieved August 19, 2025 from app.powerbi.com/view?r=eyJrIjoiMGJiMmNiNTAtOTY1OC00ZjUzLTg2OGUtMjAzYzNiYTA5YjliIiwidCI6IjI2ZjczODk3LWM4YWMtNGIxZS05NzhmLWVhNGMwNzc0MzRiZiJ9&pageName=ReportSection4036c90b8a27b5f58f54.

J. G. C. Krüger, A. de Souza Britto, and J. P. Barddal. Nov. 2023. "An Explainable Machine Learning Approach for Student Dropout Prediction." *Expert Systems with Applications*, 233, (Nov. 2023), 120933. doi:10.1016/j.eswa.2023.120933.

M. Kuhn and K. Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman & Hall/CRC Press, New York, NY. doi:10.1201/9781315108230.

S. Lee and J. Y. Chung. July 2019. "The Machine Learning-based Dropout Early Warning System for Improving the Performance of Dropout Prediction." *Applied Sciences*, 9, 15, (July 2019), 3093. doi:10.3390/app9153093.

S. M. Lundberg and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates, Inc., Long Beach, California, USA, 4765–4774. ISBN: 9781510860964.

A. Luque, A. Carrasco, A. Martín, and A. de las Heras. 2019. "The Impact of Class Imbalance in Classification Performance Metrics Based on the Binary Confusion Matrix." *Pattern Recognition*, 91, 216–231. doi:10.1016/j.patcog.2019.02.023.

S. P. B. Moreno and L. G. Támara. Nov. 2024. "Complexities of Student Dropout in Higher Education: a Multidimensional Analysis." *Frontiers in Education*, 9, (Nov. 2024). doi:10.3389/feduc.2024.1461650.

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

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

M. Psyridou, F. Prezja, M. Torppa, M.-K. Lerkkanen, A.-M. Poikkeus, and K. Vasalampi. June 2024. "Machine Learning Predicts Upper Secondary Education Dropout as Early as the End of Primary School." *Scientific Reports*, 14, 1, (June 2024), 12956. doi:10.1038/s41598-024-63629-0.

A. M. Rabelo and L. E. Zárate. Jan. 2025. "A Model for Predicting Dropout of Higher Education Students." *Data Science and Management*, 8, 1, (Jan. 2025), 72–85. doi:10.1016/j.dsm.2024.07.001.

J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez. Feb. 2020. "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review." *Applied Sciences*, 10, 3, (Feb. 2020), 1042. doi:10.3390/app10031042.

F. Rosenblatt. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review*, 65, 6, 386–408. doi:10.1037/h0042519.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Oct. 1986. "Learning Representations by Back-Propagating Errors." *Nature*, 323, 6088, (Oct. 1986), 533–536. doi:10.1038/323533a0.

T. Saito and M. Rehmsmeier. Mar. 2015. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLOS ONE*, 10, 3, (Mar. 2015), 1–21. doi:10.1371/journal.pone.0118432.

L. M. H. D. Silva, I.-A. Chounta, M. J. Rodríguez-Triana, E. R. Roa, A. Gramberg, and A. Valk. Aug. 2022. "Toward an Institutional Analytics Agenda for Addressing Student Dropout in Higher Education: Insights from Institutional Stakeholders." *Journal of Learning Analytics*, 9, 2, (Aug. 2022), 179–201. doi:10.18608/jla.2022.7507.

Z. Song, S.-H. Sung, D.-M. Park, and B.-K. Park. Jan. 2023. "All-Year Dropout Prediction Modeling and Analysis for University Students." *Applied Sciences*, 13, 2, (Jan. 2023), 1143. doi:10.3390/app13021143.

C. Tantithamthavorn, A. E. Hassan, and K. Matsumoto. 2020. "The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models." *IEEE Transactions on Software Engineering*, 46, 11, 1200–1219. doi:10.1109/TSE.2018.2876537.

M. Vaarma and H. Li. Feb. 2024. "Predicting Student Dropouts With Machine Learning: An Empirical Study in Finnish Higher Education." *Technology in Society*, 76, (Feb. 2024), 102474. doi:10.1016/j.techsoc.2024.102474.