

Student Dropout Prediction under Temporal Drift: An Empirical Evaluation of Random Forest Performance Across Academic Years

José Thiago Holanda de Alcântara Cabral

¹ Cabedelo Centro Campus
Federal Institute of Paraíba (IFPB)
João Pessoa, Paraíba, Brazil
jose.alcantara@ifpb.edu.br

Abstract. *Student dropout is a critical challenge in the Federal Network of Professional, Scientific, and Technological Education. This study evaluates the temporal robustness of dropout prediction using longitudinal data (2017–2024) from the Nilo Peçanha Platform. Using rolling temporal validation with Random Forest and Logistic Regression, predictive performance and classification threshold stability were evaluated. Results revealed substantial variability, with AUC varying from 0.34 to 0.83 and optimal thresholds ranging from 0.01 to 0.54. The findings are consistent with the possible occurrence of concept drift in educational data. Student dropout prediction systems require continuous monitoring and periodic recalibration to maintain their effectiveness.*

Keywords: *Student Dropout, Educational Data Mining, Random Forest, Concept Drift, Temporal Validation.*

Resumo. *A evasão estudantil é um desafio crítico na Rede Federal de Educação Profissional, Científica e Tecnológica. Este estudo avalia a robustez temporal da predição de evasão a partir de dados longitudinais (2017–2024) da Plataforma Nilo Peçanha. Utilizando validação temporal deslizante com Random Forest e Regressão Logística, avaliaram-se o desempenho preditivo e a estabilidade dos limiares de classificação. Os resultados revelaram variabilidade substancial, com AUC variando entre 0,34 e 0,83 e limiares ótimos entre 0,01 e 0,54. Os achados são consistentes com a possível ocorrência de concept drift em dados educacionais. Sistemas de predição de evasão estudantil requerem monitoramento contínuo e recalibração periódica para manter sua eficácia.*

Palavras-chave: *Evasão Estudantil, Mineração de Dados Educacionais, Random Forest, Concept Drift, Validação Temporal.*

1. Introduction

Student dropout remains one of the most significant challenges faced by institutions within the Federal Network of Professional, Scientific, and Technological Education in Brazil. Beyond affecting institutional performance indicators, student dropout compromises the educational and social mission of these institutions. In this context, Educational Data Mining (EDM) techniques have increasingly been employed to identify students at risk of dropout and support early intervention strategies.

Previous research conducted at a campus of the Federal Network demonstrated the potential of supervised machine learning for dropout prediction using administrative

educational data extracted from the Nilo Peçanha Platform. In particular, [Cabral 2025] identified the *Random Forest* model as the most effective approach among several predictive algorithms, reinforcing the suitability of tree-based ensemble methods for structured institutional data.

Despite these promising results, an important challenge remains insufficiently explored: the temporal stability of predictive models. Student dropout is influenced by socioeconomic, institutional, pedagogical, and external factors that evolve over time. Consequently, predictive relationships learned from historical data may gradually lose relevance when models are applied to future cohorts. Such temporal changes are commonly associated with *concept drift*, in which the underlying data distribution or the relationship between predictors and the target changes over time [Widmer and Kubat 1996, Gama et al. 2014]. These dynamics raise concerns regarding the long-term reliability of predictive models trained on historical data.

Rather than focusing exclusively on predictive accuracy, it is therefore essential to investigate how model performance evolves under realistic temporal validation scenarios. Understanding whether predictive models remain stable over time is fundamental for supporting institutional decision-making and ensuring the practical applicability of learning analytics systems.

The main hypothesis of this work is that the predictive performance of student dropout models is strongly affected by temporal variations in educational data, resulting in substantial fluctuations in predictive accuracy, Recall, F1-score, and optimal classification thresholds across academic years. Accordingly, this study investigates the following research questions:

- **RQ1:** Does the predictive performance of student dropout models remain stable when applied to future academic cohorts?
- **RQ2:** Do optimal classification thresholds remain consistent across different academic years?
- **RQ3:** To what extent do the observed results provide evidence consistent with possible concept drift in educational dropout prediction?

To answer these questions, a series of rolling temporal validation experiments were conducted.

2. Theoretical Background

Educational Data Mining (EDM) has become an important field for extracting actionable knowledge from educational data, supporting tasks such as student performance prediction and decision support [Romero and Ventura 2010, Kotsiantis 2012, Romero and Ventura 2013]. Among its various applications, student dropout prediction has received considerable attention due to its potential to support early intervention strategies and improve student retention rates [Alturki et al. 2022]. The increasing availability of institutional educational data has enabled the development of predictive models capable of identifying students at risk of dropout before disengagement occurs.

Several studies have investigated student dropout prediction using supervised machine learning techniques, including ensemble methods, linear models, support vector machines, neural networks, and boosting algorithms. Systematic reviews and recent empirical studies consistently report that ensemble methods, particularly Random

Forest, achieve robust predictive performance on heterogeneous educational datasets [Alturki et al. 2022, Vaarma and Li 2024]. Previous research conducted within the Brazilian Federal Network of Professional, Scientific, and Technological Education also identified Random Forest as one of the most effective approaches for predicting student dropout using data extracted from the Nilo Peçanha Platform [Cabral 2025]. More recent studies have shown that predictive performance can be further improved through hyperparameter optimization and the incorporation of demographic, socioeconomic, and institutional variables [Ahmed et al. 2026, Reina Marín et al. 2025]. Collectively, these findings support the adoption of Random Forest as a robust predictive model for heterogeneous educational datasets.

Although previous studies consistently demonstrate the effectiveness of machine learning techniques for student dropout prediction, most evaluate predictive models using conventional random train–test partitions or cross-validation, implicitly assuming that the statistical properties of the data remain stable over time. Consequently, the temporal robustness of these models under realistic deployment conditions remains comparatively underexplored. Since student dropout is influenced by socioeconomic conditions, institutional policies, pedagogical practices, and external events that evolve over time, predictive relationships learned from historical data may gradually lose relevance when applied to future cohorts. Such temporal changes are commonly associated with *concept drift*, in which the statistical properties of the data or the relationship between predictor variables and the target evolve over time, potentially reducing the validity of previously learned models [Widmer and Kubat 1996, Gama et al. 2014, Lu et al. 2022].

Although concept drift has been extensively investigated in data stream mining and adaptive machine learning, its application to Educational Data Mining remains comparatively limited. In particular, few studies have analyzed how predictive performance evolves across successive academic cohorts despite the dynamic nature of educational environments. This limitation also raises methodological concerns regarding how predictive models should be evaluated when temporal dependencies are present.

Models developed from temporally structured educational data require evaluation strategies that preserve the chronological relationships among observations. Conventional random train–test partitions or cross-validation may allow information from future cohorts to influence model development, leading to overly optimistic estimates of predictive performance. Roberts et al. [Roberts et al. 2017] emphasized that evaluation strategies should preserve the temporal structure of the data whenever the objective is to estimate model generalization under future conditions.

Building upon this methodological perspective, temporal validation has increasingly been recognized as a more appropriate evaluation strategy for predictive models deployed in evolving environments, since conventional random train–test partitions assume stationary data distributions and may overestimate predictive performance [Bifet and Gavaldà 2007, Gama et al. 2014]. Likewise, studies on concept drift recommend continuous model monitoring, recalibration, and adaptation when the underlying data-generating process changes over time [Widmer and Kubat 1996, Gama et al. 2014, Lu et al. 2022]. These principles provide the theoretical foundation for the rolling temporal validation framework adopted in this study.

3. Methodology

The problem definition and research objectives were established following the CRISP-DM framework. However, unlike traditional predictive studies that focus exclusively on maximizing classification performance, this work emphasizes the temporal stability of student dropout prediction models under realistic deployment conditions. To achieve this objective, a multi-year temporal validation strategy was adopted, allowing the assessment of model performance across different academic periods.

3.1. Business Understanding

Student dropout remains a major challenge for institutions within the Federal Network of Professional, Scientific, and Technological Education, affecting educational outcomes and institutional planning. Previous research identified the *Random Forest* algorithm as an effective approach for dropout prediction [Cabral 2025]. However, an important question remains: do predictive models maintain their effectiveness when applied to future academic cohorts?

This study focuses on the temporal stability of dropout prediction models by analyzing how predictive performance, classification thresholds, and generalization capacity vary across different academic years. Data obtained from the Nilo Peçanha Platform¹ reveal substantial fluctuations in dropout rates between 2017 and 2024, as shown in Table 1.

Table 1. Dropout rates between 2017 and 2024.

Year	Dropout Rate (%)
2017	41.85
2018	32.34
2019	26.93
2020	8.88
2021	42.85
2022	40.90
2023	29.77
2024	25.68

The observed variation suggests that dropout patterns may change over time, potentially affecting the reliability of models trained on historical data. Therefore, this study adopts a rolling temporal validation strategy to evaluate whether predictive patterns learned from previous cohorts remain valid when applied to future academic years and to investigate the potential presence of concept drift.

3.2. Data Understanding

The data were obtained from the Academic Indicators module of the Nilo Peçanha Platform and comprise academic, sociodemographic, and institutional variables collected between 2017 and 2024. The dataset includes attributes such as racial classification, family income, gender, age group, course type, educational modality, and course shift.

The records were extracted for the selected campus, consolidated into a single dataset, and validated to ensure consistency across academic years. The validation process included verifying annual enrollment and dropout counts, checking for duplicated

¹<https://www.gov.br/mec/pt-br/pnp>

records after consolidation, and confirming that the aggregated totals matched the official indicators published by the Nilo Peçanha Platform.

Only the *Family Income* attribute contained missing values (35% of the records), corresponding to students who chose not to declare this information.

The temporal structure of the dataset enables the evaluation of predictive models using chronologically ordered academic cohorts. Therefore, the *YEAR* attribute was used exclusively to define temporal partitions and was not included as a predictive feature, preventing temporal leakage.

Table 2 presents the variables collected from the Nilo Peçanha Platform and their respective data types.

Table 2. Description of the variables in the dataset.

Variable	Type	Description
RACIAL CLASSIFICATION	Categorical	Self-declared racial group
FAMILY INCOME	Numerical (float)	Family income range
FAMILY INCOME NOT DECLARED	Categorical (bool)	True or False
GENDER	Categorical	Biological sex (M or F)
AGE GROUP	Categorical	Age group in 5-year intervals
COURSE TYPE	Categorical	FIC, TECHNICAL, or SPECIAL TECHNICAL
COURSE NAME	Categorical	Course acronym
OFFER TYPE	Categorical	Subsequent, Integrated, etc.
EDUCATIONAL MODALITY	Categorical	In-person, Distance Learning (EAD)
COURSE SHIFT	Categorical	Morning, Afternoon, Evening, NSA
YEAR	Numerical (Temporal)	Reference year
ENROLLMENTS	Numerical (int)	Number of enrolled students
DROPOUTS	Numerical (int)	Number of dropout students

3.3. Data Preparation

The Nilo Peçanha Platform provides academic indicators in an aggregated format rather than individual student records. Therefore, the data preparation procedure followed the expansion methodology proposed by [Cabral 2025], which reconstructs a student-level dataset suitable for supervised learning. For each aggregated record, the reported number of enrollments was expanded into individual observations. Dropout cases were labeled as $DR=1$, whereas the remaining observations were labeled as $DR=0$, resulting in one record per student while preserving the statistical distribution of the original aggregated data.

Missing values in the family income attribute were preserved and complemented by the binary variable *RF_ND*, indicating non-declared income. Age groups were converted to numerical values using the midpoint of each interval, categorical variables were transformed using one-hot encoding, and the encoded datasets were aligned to ensure consistency across temporal partitions.

The *YEAR* attribute was used exclusively for temporal partitioning and removed from the predictive features to prevent temporal leakage. The final dataset contains 6,820 student records spanning the period from 2017 to 2024, with an overall dropout rate of approximately 33%, enabling the construction of multiple temporal train-validation-test scenarios.

Table 3 presents a sample of the aggregated data.

Table 4 illustrates the expanded student-level dataset prior to feature transformation. The predictive dataset comprises ten predictor attributes plus one target variable

Table 3. Sample of the collected dataset.

RC	FI	Gender	AG	CT	CN	OT	EM	CS	Year	Enrollments	Dropouts
PRE	ND	M	20-24	TEC	GT	SUB	P	AFT	2018	1	0
PAR	0.5	F	20-24	TEC	SJ	SUB	P	AFT	2017	2	1
BRA	2.5	M	60-64	TEC	GT	SUB	P	AFT	2018	1	1
IND	ND	M	25-29	SPEC	THL	NSA	D	NSA	2023	2	2
BRA	3.5	F	35-39	TEC	SJ	SUB	D	NSA	2023	2	0

(DR).

Table 4. Sample of the dataset after initial expansion.

FI	FI-ND	Gender	AVG-AGE	CT	CN	OT	EM	CS	Year	DR
NaN	1	M	22	TEC	GT	SUB	P	AFT	2018	0
0.5	0	F	22	TEC	SJ	SUB	P	AFT	2017	1
0.5	0	F	22	TEC	SJ	SUB	P	AFT	2017	0
2.5	0	M	62	TEC	GT	SUB	P	AFT	2018	1
NaN	1	M	27	SPEC	THL	NSA	D	NSA	2023	1
NaN	1	M	27	SPEC	THL	NSA	D	NSA	2023	1
3.5	0	F	37	TEC	SJ	SUB	D	NSA	2023	0
3.5	0	F	37	TEC	SJ	SUB	D	NSA	2023	0

3.4. Modeling

Random Forest is an ensemble learning technique that combines multiple decision trees through bootstrap aggregation, reducing variance and improving generalization performance [Breiman 2001]. The algorithm was selected based on previous findings reported by [Cabral 2025], which identified Random Forest as one of the most effective approaches for student dropout prediction using institutional educational data. Since the objective of this study is to investigate the temporal stability of a previously validated predictive model rather than compare different machine learning algorithms, Random Forest was adopted as the primary model.

Hyperparameter optimization was performed through a grid-search procedure involving different combinations of the following parameters: number of trees, maximum tree depth, minimum samples required for node splitting, minimum samples required at leaf nodes, split criterion, number of features evaluated at each split, and bootstrap sampling strategy. To mitigate class imbalance, the parameter *class_weight* = "balanced" was adopted in all experiments.

To identify the most appropriate model configuration, the hyperparameter search space presented in Table 5 was evaluated for each temporal validation scenario, and the best-performing configuration on the validation dataset was selected for subsequent testing.

Random Forest also provides probabilistic outputs, enabling the analysis of classification thresholds, which constitutes one of the research questions investigated in this study.

To support the interpretation of the temporal experiments, feature importance was computed only for the Random Forest model using the impurity-based importance measure (Mean Decrease in Impurity) provided by the scikit-learn implementation. Since Logistic Regression was included exclusively as a predictive baseline, feature importance analysis was not performed for this model.

Table 5. Hyperparameter search space evaluated for the Random Forest model.

Hyperparameter	Values Evaluated
n_estimators	100, 300, 500
max_depth	5, 10, 20, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 5
criterion	gini, entropy, log_loss
max_features	sqrt, log2
bootstrap	True
class_weight	balanced

For each temporal validation scenario, the hyperparameter optimization procedure selected a potentially different Random Forest configuration according to the corresponding validation dataset.

Table 6 summarizes the best hyperparameters selected for each temporal experiment. Although the selected configurations varied slightly across the evaluation windows, the differences were relatively small, suggesting that the observed performance variation is primarily associated with temporal changes in the data rather than differences in model complexity.

Table 6. Best Random Forest hyperparameters selected for each temporal validation experiment.

Exp.	Train	Val.	N_est	Depth	Split	Leaf	Criterion	Feat.	Boot.	Weight
1	2017	2018	500	5	2	5	gini	log2	True	balanced
2	2018	2019	100	10	5	5	gini	log2	True	balanced
3	2019	2020	300	10	10	1	entropy	log2	True	balanced
4	2020	2021	100	10	2	1	entropy	log2	True	balanced
5	2021	2022	500	5	2	1	log_loss	sqrt	True	balanced
6	2022	2023	100	5	5	5	gini	log2	True	balanced

To provide a baseline for comparison, Logistic Regression (LR) was also evaluated in the temporal validation experiments. LR is a widely adopted linear classifier in educational data mining due to its interpretability and competitive predictive performance in binary classification tasks. Unlike Random Forest, Logistic Regression models linear relationships between predictor variables and the log-odds of the target class. Therefore, it serves as a simple and interpretable baseline for assessing whether the observed temporal degradation is model-specific or inherent to the prediction problem.

Logistic Regression was evaluated using the same temporal train-validation-test windows, model selection procedure, and validation-based threshold optimization adopted for Random Forest.

3.5. Temporal Validation Strategy

A rolling temporal validation strategy was adopted to evaluate the stability and generalization capacity of the predictive models across successive academic years. Unlike conventional random train–test splits, temporal validation preserves the chronological order of observations by training models on past cohorts and evaluating them on future cohorts, thereby reflecting realistic deployment conditions. Six sequential temporal windows were constructed using the period from 2017 to 2024, in which each model was trained on one academic year, validated on the subsequent year, and tested on the following unseen year.

For each temporal scenario, all candidate hyperparameter combinations were evaluated on the corresponding validation dataset, and the best-performing configuration was selected for subsequent testing. After model selection, the optimal classification threshold was determined by maximizing the standard binary F1-score provided by scikit-learn. This threshold was then applied unchanged to the corresponding test year, preventing information leakage and providing a realistic assessment of model performance under future operating conditions.

The comparison between validation and test results across successive academic years enables the investigation of temporal performance degradation, threshold variability, and the potential effects of concept drift on student dropout prediction.

3.6. Model Evaluation

Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics were selected because they provide complementary perspectives on classification performance, particularly in scenarios involving class imbalance.

The AUC-ROC metric was used to assess the model's ability to discriminate between dropout and non-dropout students independently of a specific classification threshold. Precision and Recall were employed to evaluate the trade-off between correctly identifying dropout cases and minimizing false positive predictions. Particular emphasis was placed on Recall, since failing to identify students at risk of dropout may reduce the effectiveness of institutional intervention strategies. The F1-score was adopted as the primary criterion for threshold optimization, balancing Precision and Recall into a single measure.

Beyond conventional predictive performance, this study also evaluated the temporal stability of the models by analyzing the variation of the evaluation metrics across successive validation and test periods. In addition, the variability of the optimal classification thresholds obtained during validation was investigated as an indicator of changes in the underlying dropout patterns over time.

3.7. Threats to Validity

Several factors may influence the validity of the findings. First, the study was conducted using data from a single institution within the Federal Network of Professional, Scientific, and Technological Education, which may limit the generalizability of the results to other campuses, regions, or educational systems.

Second, substantial changes in educational policies, teaching modalities, institutional practices, and external socioeconomic conditions may alter the relationships associated with student dropout over time. Such changes may affect the performance of models trained exclusively on historical data when applied to future cohorts.

Third, the available attributes consist primarily of administrative and demographic information. Although routinely available in institutional databases and suitable for operational predictive systems, they do not capture academic, behavioral, psychological, and other contextual factors that may also influence student dropout.

Finally, although the observed temporal degradation is consistent with possible concept drift, this study did not employ dedicated drift detection algorithms (e.g., AD-

WIN, DDM, or Page-Hinkley), and only eight academic cohorts were available for analysis. Consequently, the findings should be interpreted as evidence consistent with possible concept drift rather than as a formal confirmation of its occurrence.

4. Results

This section presents the findings from the experimental validation and analyzes the temporal performance of the evaluated predictive models.

4.1. Temporal Validation Results

Table 7 summarizes the results obtained across all temporal validation experiments. For each scenario, the Random Forest model was trained on one academic year, validated on the subsequent year, and tested on an unseen future cohort. The table presents the Accuracy, Precision, Recall, F1-score, and AUC achieved during the validation and testing stages, together with the optimal classification threshold determined during the validation phase.

Table 7. Results obtained across the temporal validation experiments.

Years	Acc _V	Acc _T	Prec _V	Prec _T	Rec _V	Rec _T	F1 _V	F1 _T	AUC _V	AUC _T	Thr _V - Thr _O
2017-2018-2019	0.63	0.36	0.40	0.30	0.25	1.00	0.31	0.46	0.56	0.58	0.22 - 0.33
2018-2019-2020	0.65	0.20	0.33	0.10	0.28	0.98	0.30	0.18	0.59	0.49	0.21 - 0.21
2019-2020-2021	0.78	0.33	0.12	0.26	0.23	0.30	0.15	0.28	0.52	0.34	0.13 - 0.01
2020-2021-2022	0.48	0.43	0.29	0.42	0.15	1.00	0.20	0.59	0.35	0.67	0.01 - 0.22
2021-2022-2023	0.61	0.51	0.60	0.38	0.16	0.95	0.25	0.54	0.68	0.83	0.39 - 0.54
2022-2023-2024	0.55	0.58	0.39	0.56	0.90	0.39	0.54	0.46	0.85	0.58	0.54 - 0.31
Mean	0.62	0.40	0.35	0.33	0.33	0.77	0.29	0.42	0.59	0.58	0.25 - 0.27
Std. Dev.	0.10	0.13	0.16	0.16	0.28	0.33	0.14	0.16	0.17	0.16	0.19 - 0.17
CV	0.16	0.34	0.45	0.46	0.87	0.43	0.47	0.38	0.28	0.28	0.76 - 0.65
Min	0.48	0.20	0.12	0.10	0.15	0.30	0.15	0.18	0.35	0.34	0.01 - 0.01
Max	0.78	0.58	0.60	0.56	0.90	1.00	0.54	0.59	0.85	0.83	0.54 - 0.54

Overall, Accuracy and AUC exhibited lower relative variability than Precision, Recall, and F1-score. Nevertheless, their absolute values still differed considerably across the temporal scenarios, particularly in the test cohorts. The optimal decision threshold also varied substantially, indicating that both predictive performance and classification cutoffs are sensitive to temporal changes.

Test AUC values ranged from 0.34 to 0.83 despite the use of the same modeling strategy and hyperparameter optimization procedure in all experiments. Similar variability was observed for Recall and F1-score, indicating that predictive performance depended strongly on the temporal characteristics of the academic cohorts. These findings suggest that models calibrated on historical cohorts may not generalize reliably to future academic years, reinforcing the importance of temporal validation under evolving educational conditions.

The following subsections further analyze these findings through a comparison with the Logistic Regression baseline, an investigation of threshold variability, and a discussion of temporal changes in feature importance.

4.2. Logistic Regression Baseline

To assess whether the observed temporal variability was specific to the Random Forest model or also present in alternative predictive approaches, Logistic Regression was

included as a baseline model. The baseline was evaluated on the same sequence of future-year test cohorts, with results summarized in Table 8. The Logistic Regression model achieved test AUC values ranging from 0.29 to 0.62, while the Random Forest model ranged from 0.34 to 0.83 across the corresponding temporal experiments. Although the Logistic Regression baseline exhibited a narrower AUC range, its overall discriminative performance remained lower than that of Random Forest in most scenarios. Moreover, substantial temporal variation was also observed in its predictive metrics, with Recall ranging from 0.28 to 1.00 and F1-score from 0.16 to 0.58. These findings indicate that temporal instability is not exclusive to Random Forest but also reflects changes in the underlying data-generating process across academic cohorts, although its magnitude and pattern differ between models.

Table 8. Results obtained for the Logistic Regression baseline across the temporal validation experiments.

Years	Acc _V	Acc _T	Prec _V	Prec _T	Rec _V	Rec _T	F1 _V	F1 _T	AUC _V	AUC _T	Thr _V - Thr _O
2017-2018-2019	0.65	0.30	0.45	0.24	0.39	0.76	0.42	0.37	0.59	0.42	0.32 - 0.17
2018-2019-2020	0.27	0.09	0.27	0.09	1.00	1.00	0.42	0.16	0.57	0.54	0.39 - 0.39
2019-2020-2021	0.57	0.37	0.10	0.27	0.48	0.28	0.17	0.27	0.58	0.29	0.39 - 0.16
2020-2021-2022	0.43	0.41	0.43	0.41	1.00	1.00	0.60	0.58	0.50	0.57	0.39 - 0.39
2021-2022-2023	0.63	0.37	0.63	0.32	0.22	1.00	0.33	0.49	0.73	0.62	0.19 - 0.21
2022-2023-2024	0.62	0.57	0.42	0.54	0.71	0.50	0.53	0.52	0.73	0.59	0.48 - 0.21
Mean	0.53	0.35	0.38	0.31	0.64	0.76	0.41	0.40	0.62	0.51	0.26 - 0.36
Std. Dev.	0.15	0.16	0.18	0.15	0.32	0.31	0.15	0.16	0.09	0.13	0.11 - 0.10
CV	0.28	0.45	0.47	0.49	0.51	0.41	0.37	0.40	0.15	0.25	0.42 - 0.27
Min	0.27	0.09	0.10	0.09	0.22	0.28	0.17	0.16	0.50	0.29	0.16 - 0.19
Max	0.65	0.57	0.63	0.54	1.00	1.00	0.60	0.58	0.73	0.62	0.39 - 0.48

4.3. Threshold Variability Analysis

One of the most notable findings of this study is the substantial variability observed in the optimal classification thresholds obtained during the validation phase. Across the six temporal experiments, as shown in the last column of Table 7, the optimal thresholds varied from 0.01 to 0.54, representing a variation of more than fifty times between the lowest and highest values.

This variability indicates that the decision boundary separating dropout and non-dropout students is highly sensitive to temporal changes in the data. While some academic years required very low probability thresholds to maximize predictive performance, others required substantially higher cutoffs, suggesting that the optimal operating point changes over time.

From an operational perspective, these findings indicate that decision thresholds require periodic recalibration, since values optimized for one cohort may become sub-optimal when applied to subsequent academic years. Consequently, predictive systems deployed in educational institutions should not rely on a fixed decision threshold but instead periodically reassess the probability cutoff to maintain predictive effectiveness over time.

4.4. Discussion of Concept Drift Effects

The temporal experiments revealed a coherent pattern across the three analyses conducted in this study. Predictive performance varied substantially across academic years despite the use of the same modeling strategy, threshold analysis demonstrated that decision

boundaries require periodic recalibration, and feature importance analysis showed that the relative influence of the predictors evolved over time. Taken together, these findings suggest that the relationships between predictors and dropout outcomes are not temporally stationary, affecting the model's ability to generalize to future cohorts.

The experiment trained on 2019 data, validated on 2020, and tested on 2021 provides a particularly illustrative example of temporal instability. This scenario achieved the lowest test AUC (0.34). The validation year corresponded to 2020, when the dropout rate dropped to 8.88%, followed by a sharp increase to 42.85% in 2021 (Table 1). Such abrupt changes likely affected the underlying dropout patterns, limiting the model's ability to generalize to the subsequent cohort. Conversely, the 2021–2022–2023 experiment achieved the highest test AUC (0.83), suggesting that greater temporal consistency contributes to improved model stability. A similar effect was observed in the 2022–2023–2024 scenario, where a validation AUC of 0.85 decreased to 0.58 on the unseen cohort, indicating that strong validation performance does not necessarily translate into equivalent future performance.

A possible explanation for the degradation observed in the 2019–2020–2021 experiment is the disruption caused by the COVID-19 pandemic. Emergency remote teaching, temporary academic regulations, greater flexibility regarding attendance and course progression, and changes in institutional support policies substantially altered educational dynamics and dropout patterns. These contextual changes may also explain the temporal shifts observed in feature importance, as socioeconomic variables became less dominant while demographic and institutional characteristics gained predictive relevance in later years.

Although no formal drift detection algorithm was employed, the combined evidence obtained from temporal performance degradation, threshold instability, and changes in feature importance is consistent with the presence of possible concept drift. From a practical perspective, these findings indicate that predictive systems deployed by educational institutions should not be considered static decision-support tools. Instead, model monitoring should become part of the institutional analytics workflow, including periodic performance evaluation, threshold recalibration, and, when necessary, model retraining. Such procedures may help maintain predictive reliability despite temporal changes in student populations and institutional policies.

The Logistic Regression baseline provides additional evidence that temporal instability is not exclusive to the Random Forest architecture. Although Random Forest consistently achieved higher AUC values across the evaluated cohorts, Logistic Regression also exhibited substantial fluctuations in Recall, F1-score, and AUC over time. These findings suggest that part of the observed degradation reflects temporal changes in the underlying data distribution rather than solely the choice of predictive algorithm, although model architecture still influences the magnitude of performance variation.

4.5. Feature Importance Analysis

Feature importance was estimated using the impurity-based Mean Decrease in Impurity (MDI) measure provided by the Random Forest implementation in scikit-learn. The scores were analyzed across the temporal experiments to investigate whether the contribution of the predictors remained stable over time.

Table 9 presents the five most important features across the temporal experiments.

Table 9. Top five most important features across the temporal experiments.

Feature	2019	2020	2021	2022	2023	2024	Mean
AGE	0.07	0.17	0.15	0.24	0.16	0.14	0.15
RI	0.22	0.14	0.09	0.14	0.09	0.06	0.12
RI_LND	0.20	0.11	0.05	0.07	0.08	0.05	0.09
OFFER_TYPE_INTEGRATED	0.03	0.03	0.03	0.07	0.15	0.10	0.07
COURSE_SHIFT_INTEGRAL	0.01	0.02	0.02	0.08	0.15	0.001	0.05

The results reveal substantial temporal variation in predictor importance. Family-income-related variables (*RI* and *RI_LND*) were the most influential predictors in the earlier experiments but gradually lost importance over time. In contrast, *AGE* became the most consistently important predictor, while *OFFER_TYPE_INTEGRATED* and *COURSE_SHIFT_INTEGRAL* gained relevance in the later experiments, suggesting that institutional characteristics became progressively more influential than socioeconomic variables.

These temporal shifts complement the performance and threshold analyses by indicating that the predictors associated with student dropout evolved across academic years. The increasing importance of *AGE* is consistent with the predominance of younger students in integrated technical programs, which historically present lower dropout rates. Likewise, the growing relevance of *OFFER_TYPE_INTEGRATED* and *COURSE_SHIFT_INTEGRAL* reflects institutional characteristics of the educational offerings. Although *RI* and *RI_LND* remained relevant throughout the study, impurity-based feature importance quantifies predictive contribution rather than the direction of the relationship between family income and dropout.

5. Conclusion

This study investigated the temporal robustness of Random Forest and Logistic Regression models for student dropout prediction using longitudinal educational data.

The results addressed the three research questions. Regarding RQ1, both Random Forest and Logistic Regression exhibited substantial temporal variability in predictive performance across academic cohorts, although Random Forest consistently achieved superior discrimination. Regarding RQ2, the optimal classification thresholds for Random Forest varied considerably across academic years, indicating that threshold recalibration is necessary for reliable deployment. Regarding RQ3, the combined evidence from performance degradation, threshold instability, and temporal changes in feature importance is consistent with the presence of possible concept drift.

Overall, these findings suggest that student dropout prediction should be treated as a dynamic rather than static classification task. Educational information systems should therefore incorporate periodic model monitoring, threshold recalibration, and regular re-training to maintain predictive effectiveness under changing educational conditions.

Future work will extend the analysis to additional machine learning algorithms, incorporate dedicated concept drift detection techniques, evaluate adaptive learning strategies, and investigate whether richer academic and behavioral variables (e.g., attendance, scholarship participation, and student engagement) improve temporal robustness.

Declaration of Generative AI Use

Generative Artificial Intelligence tools were used during the development of this study to support writing revision, language refinement, structure organization, and programming activities. AI-assisted tools were also employed to support the implementation and documentation of data preprocessing, model training, hyperparameter optimization, and evaluation scripts. All generated content, code, analyses, methodological decisions, experimental results, and final interpretations were reviewed, validated, and approved by the authors. The authors assume full responsibility for the accuracy, integrity, and originality of the submitted work.

References

- Ahmed, R., Fahad, N., Miah, M. S. U., Hossen, M. J., and Bhattacharjee, K. (2026). Hyoptensemble: Custom-weighted soft voting hyperparameter optimization ensemble model, explainable-ai for predicting mental state among university students. *Discover Artificial Intelligence*, 6(1):136.
- Alturki, S., Hulpuş, I., and Stuckenschmidt, H. (2022). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*, 27(2):275–307.
- Bifet, A. and Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In *Proceedings of the Seventh SIAM International Conference on Data Mining (SDM)*, pages 443–448. SIAM.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Cabral, J. T. H. d. A. (2025). Predictive modeling for student retention: Evaluation of machine learning algorithms with temporal validation. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 99–112, Curitiba, PR, Brasil. Sociedade Brasileira de Computação (SBC).
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4).
- Kotsiantis, S. B. (2012). Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4):331–344.
- Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2022). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2494–2513.
- Reina Marín, Y., Quiñones Huatangari, L., Cruz Caro, O., Maicelo Guevara, J. L., Alva Tuesta, J. N., Sánchez Bardales, E., and Chávez Santos, R. (2025). Data mining to identify university student dropout factors. *Applied Sciences*, 15(22):11911.
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.
- Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(6):601–618.

- Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.
- Vaarma, M. and Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in finnish higher education. *Technology in Society*, 76:102474.
- Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101.