Predictive Modeling for Student Retention: Evaluation of Machine Learning Algorithms with Temporal Validation

José Thiago Holanda de Alcântara Cabral

[1] Cabedelo Centro Advanced Campus
Federal Institute of Paraíba (IFPB)
João Pessoa, Paraíba, Brazil
jose.alcantara@ifpb.edu.br
ORCID: 0009-0009-5843-303X

Abstract. *This study aims to develop and evaluate predictive models to identify students at risk of dropout at a campus of the Federal Institute of Paraíba (IFPB), using administrative data from 2017 to 2023 obtained from the Nilo Peçanha Platform. Several supervised machine learning algorithms were applied, including interpretable models such as Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN), as well as more complex models like Support Vector Machine (SVM), Random Forest, and XGBoost. Model performance was assessed through stratified cross-validation and a prospective test with unseen data from 2023. The Random Forest model achieved the best overall results, particularly in AUC-ROC and Recall, offering a balanced trade-off between generalization and sensitivity. These findings demonstrate the feasibility of integrating predictive models into institutional decision-support systems to strengthen student retention strategies. Future work involves deploying the model in a monitoring system and incorporating concept drift detection techniques to maintain long-term reliability in dynamic educational contexts.*

Resumo. *Este estudo tem como objetivo desenvolver e avaliar modelos preditivos para identificar estudantes com risco de evasão em um campus do Instituto Federal da Paraíba (IFPB), utilizando dados administrativos de 2017 a 2023 provenientes da Plataforma Nilo Peçanha. Foram aplicados diversos algoritmos de aprendizado de máquina supervisionado, incluindo modelos interpretáveis, como Regressão Logística, Árvore de Decisão e K-Nearest Neighbors (KNN), além de modelos mais complexos, como Support Vector Machine (SVM), Random Forest e XGBoost. A avaliação dos modelos foi realizada por meio de validação cruzada estratificada e de um teste prospectivo com dados inéditos de 2023. O modelo Random Forest apresentou o melhor desempenho geral, destacando-se em AUC-ROC e Recall, oferecendo um equilíbrio adequado entre generalização e sensibilidade. Os resultados demonstram a viabilidade de integrar modelos preditivos aos sistemas institucionais de apoio à decisão, fortalecendo as estratégias de permanência estudantil. Como trabalhos futuros, propõe-se implantar o modelo em um sistema de monitoramento e incorporar técnicas de detecção de concept drift, visando garantir a confiabilidade do modelo em ambientes educacionais dinâmicos.*

DOI: 10.5753/sbie.2025.11981

1. Introduction

Student dropout is a critical challenge for educational institutions, with pedagogical, institutional, and social impacts. Understanding the factors that influence dropout is essential to support retention strategies that promote more inclusive and high-quality education. In this context, Educational Data Mining (EDM) has emerged as a relevant tool, capable of uncovering hidden patterns in educational data and providing valuable insights to support evidence-based decision-making [Hegazi and Abugroon 2016, Lynn and Emanuel 2021, Alturki et al. 2022]. The use of machine learning algorithms in EDM has enabled the development of robust predictive models that assist educational management in anticipating dropout risks [Kotsiantis 2012, Colpo et al. 2024].

The literature highlights the application of various algorithms, such as Logistic Regression, Decision Trees, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and ensemble-based models like Random Forest and XGBoost, as well as Deep Learning techniques. These models, when combined with proper data preprocessing, have proven effective in predicting student dropout [Krüger et al. 2023, Khedr and El Seddawy 2015].

Given this scenario, this study aims to evaluate different machine learning models and select the one that demonstrates the best performance in predicting student dropout, considering socioeconomic, academic, and institutional data of students from a specific campus of IFPB, covering the period from 2017 to 2023. The analysis was conducted using data from the Nilo Peçanha Platform, applying a prospective testing scenario in which the models were trained with data from 2017 to 2022 and evaluated with data from 2023.

Among the algorithms analyzed, Random Forest stood out for its robustness, ability to handle multiple variables, and greater resistance to overfitting—characteristics also emphasized in the literature [Krüger et al. 2023, Khedr and El Seddawy 2015, Hassan et al. 2024]. Previous studies have shown its effectiveness in dropout prediction, with performance superior to most evaluated models.

While numerous studies have applied machine learning to predict student dropout, this work addresses a specific practical gap. First, it presents a detailed and transparent case study of predictive modeling within a specific campus of a Brazilian Federal Institute, a context underrepresented in the literature. Second, it demonstrates the feasibility of building a robust model using only the type of aggregated administrative data available from national public platforms, a common constraint for many educational institutions.

Methodologically, the study makes two further contributions. Third, it explicitly analyzes the impact of a major real-world anomaly—the COVID-19 pandemic—on model training and selection, offering insights into building resilient models in dynamic environments. Finally, it employs a rigorous prospective validation methodology, providing a more realistic and reliable assessment of the model's performance in a real-world deployment scenario.

The following sections present the theoretical background, methodological procedures, results, and conclusions of this study.

2. Theoretical Background

To provide the conceptual and methodological foundation for this study, a literature review was conducted on student dropout prediction using machine learning techniques. The search was carried out in the Scopus database using a query string (Table 1) that combined terms related to dropout, literature reviews, and computational techniques, restricting the results to publications from 2014 onwards. Twelve review articles selected from this search were incorporated and cited throughout the development of this study, offering a comprehensive and up-to-date foundation on the topic.

Table 1. Query string used for the systematic search in the Scopus database.

```
TITLE-ABS-KEY(("dropout" OR "student attrition" OR "school dropout" OR
"educational dropout") AND ("literature review" OR "systematic review" OR
"survey") AND ("machine learning" OR "artificial intelligence" OR "data mining"
OR "predictive modeling")) AND PUBYEAR >= 2014.
```

Educational Data Mining (EDM) is an established field that aims to extract useful knowledge from educational data [Guleria and Sood 2014, Romero and Ventura 2010]. It supports educational management through predictive analytics, recommendations, and data visualization. Reviews such as [Alturki et al. 2022] and [Alnasyan and et al. 2024] highlight the growing range of EDM applications, from performance analysis to dropout prediction, particularly in virtual learning environments and MOOCs.

Student dropout is a complex problem with significant institutional and social impacts [Colpo and et al. 2024]. EDM has proven to be an effective tool for anticipating dropout cases, enabling more assertive interventions. Studies such as [Colpo and et al. 2024] and [Albreiki et al. 2021] emphasize the importance of Learning Analytics in identifying risk factors, while [Estrada-Molina et al. 2024] reinforces the use of Deep Learning to predict dropout in online learning environments.

Various machine learning algorithms are applied in EDM, including Logistic Regression, Decision Trees, Neural Networks, Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) [Shahiri et al. 2015, Colpo and et al. 2024]. Ensemble-based models, such as Random Forest, are widely used due to their robustness and generalization capabilities [Colpo and et al. 2024]. More recent techniques, such as Deep Learning (DNNs, CNNs, RNNs, and LSTMs), also demonstrate high accuracy in predicting performance and dropout [Alnasyan and et al. 2024]. However, the effectiveness of these models depends not only on the chosen algorithm but also on the quality of data preprocessing—a crucial step to ensure the reliability of the results [Romero and Ventura 2010].

Data preprocessing, as highlighted by [Romero and Ventura 2010], is a fundamental step in EDM since it ensures data quality and directly impacts the performance of predictive models. The literature reinforces that the combination of well-prepared data, appropriate algorithms, and a solid understanding of the institutional context is essential to develop effective dropout prediction models. Therefore, this study is grounded in these well-established practices, applying machine learning techniques to support student retention strategies.

3. Methodology

The problem definition and the objectives were carried out following the guidelines of the first step of the CRISP-DM model.

3.1. Business Understanding

Student dropout is a relevant challenge in public education, with direct impacts on the institutional mission. Data from the Nilo Peçanha Platform[1] reveal significant variations in dropout rates between 2017 and 2023, as shown in Table 2. Notably, the year 2020 presented an atypically low dropout rate (8.88%), reflecting institutional adaptations during the COVID-19 pandemic, such as emergency remote teaching and flexible academic criteria.

Table 2. Dropout rates between 2017 and 2023.

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

The increase in dropout in 2021 (42.85%) reflects the resumption of academic requirements and the return of FIC courses, which had been interrupted or made flexible in 2020. This context highlights the need for a predictive model capable of considering temporal variations, changes in student profiles, and shifts in the educational landscape.

Therefore, the proposed model must handle typical challenges in the educational domain, such as temporal changes, the emergence of new courses, shifts in student demographics, and external factors impacting the educational process. The clear definition of this problem guides the subsequent steps of data preparation, model development, and evaluation.

3.2. Data Understanding

The data were obtained from "Academic Indicators" module, in Nilo Peçanha Platform, under the covering the period from 2017 to 2023. The initial dataset contains 3737 records, including academic, sociodemographic, and institutional variables, such as racial classification, family income, gender, age group, course type and name, educational modality, and course shift, in addition to year, enrollment numbers, and dropout counts.

The data collection process involved accessing the Nilo Peçanha Platform and navigating to the *Academic Indicators* section. The data were selected according to breakdowns by Racial Classification, Family Income, Gender, Age Group, Course Type, Course Name, Offer Type, Educational Modality, Course Shift, and Year. Subsequently, the information was filtered for the CACC/IFPB campus, and data on Year, Enrollments, and Dropouts were manually collected. These records were then stored in a spreadsheet format (`.csv`) and validated through cross-checking to ensure consistency.

[1] `https://www.gov.br/mec/pt-br/pnp`

Table 3 presents the collected variables along with their respective categories and formats. The data collection was manual, stored in spreadsheets, followed by thorough review and validation. Table 4 provides a sample of the collected data, illustrating the organization of variables and their distribution over time. A small occurrence of missing data was identified (0.26% in the gender variable), which was later addressed during data preparation. The tabular structure of the dataset enables predictive analysis considering different student profiles and academic contexts.

The categorical variables feature a diverse range of categories compatible with the institutional profile, including the distinction between technical courses (TÉC) and initial and continuing education courses (FIC), as well as segmentation by shift and modality. This rich set of information allows for the construction of predictive models considering academic, socioeconomic, and institutional aspects.

The dataset used in this study has a tabular structure where each row represents the count of enrolled and dropout students for a specific context, defined by the combination of categorical variables and the reference year. A sample of this aggregated data is presented in Table 4. The purpose of this table is solely to provide a better understanding of the data structure before the preparation phase and is not intended for a detailed exploration or interpretation of the presented values.

Table 3. Description of the variables in the dataset.

Variable	Type	Description
RACIAL CLASSIFICATION	Categorical	Self-declared racial group
FAMILY INCOME	Categorical	Family income range
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

Table 4. 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

3.3. Data Preparation

The data preparation followed the CRISP-DM model, transforming the aggregated raw data from the Nilo Peçanha Platform into an individualized dataset suitable for supervised learning. Initially, the aggregated dataset provided, for each combination of categorical variables and year, the total number of enrolled and dropout students. To enable machine learning algorithms, the dataset was expanded by creating one record for each student, following this procedure:

- For rows with multiple enrollments, individual records were created.

- Non-dropout students were labeled as `DR = 0`.
- Dropout students were labeled as `DR = 1`.

The following snippet illustrates the Python logic used for this expansion.

```
# List to store the expanded data
expanded_data = []

# Assuming 'df_aggregated' is the original DataFrame
for index, row in df_aggregated.iterrows():
    # Create records for non-dropout students
    num_non_dropouts = row['Enrollments'] - row['Dropouts']
    for _ in range(num_non_dropouts):
        new_row = row.copy()
        new_row['DR'] = 0  # Target: Not Dropped Out
        expanded_data.append(new_row)

    # Create records for dropout students
    for _ in range(row['Dropouts']):
        new_row = row.copy()
        new_row['DR'] = 1 # Target: Dropped Out
        expanded_data.append(new_row)
```

Listing 1. Python snippet for expanding the aggregated dataset into individual student records.

This expansion resulted in a dataset with 5,718 individual records, maintaining a dropout rate of approximately 32%. After applying the other transformations mentioned—such as removing the racial classification variable and converting the age group to a numerical average—the final structure of the dataset is ready for the modeling phase. To ensure compatibility with the machine learning algorithms, final preprocessing steps were applied. Numerical variables, such as AVG-AGE, were standardized using the StandardScaler, while categorical variables were transformed via one-hot encoding (OneHotEncoder), which was configured to ignore unseen categories in future data. Table 5 illustrates the result of this entire preparation process when applied to the curated sample shown previously in Table 4.

Table 5. Sample of the dataset after initial expansion.

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

3.4. Modeling

In this stage, the prepared data were temporally split into training (2017–2022) and testing (2023) sets, ensuring prospective validation of the models, as recommended in educational contexts [Romero and Ventura 2010, Alturki et al. 2022]. The task addressed is a binary classification to predict student dropout (dropout = 1, retention = 0).

Robust and interpretable algorithms widely used in educational research were evaluated: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). RF, SVM, and XGBoost stand out for their ability to capture nonlinear relationships and handle imbalanced data [Colpo and et al. 2024, Albreiki et al. 2021, Mduma et al. 2019], while Logistic Regression and Decision Tree were included for their high potential for interpretability, a crucial factor for institutional adoption. KNN approximates prediction to human reasoning by comparing similar students.

Hyperparameter tuning was performed via grid search with stratified cross-validation to ensure stability and avoid overfitting [Kohavi 1995, Bergstra and Bengio 2012]. Class imbalance was addressed through class weighting (`class_weight='balanced'`) for RF, SVM, Logistic Regression, and Decision Tree, and through the hyperparameter `scale_pos_weight` in XGBoost. KNN models were adjusted to optimize performance despite lacking a native balancing parameter.

Due to the anomaly in dropout rates in 2020 caused by the COVID-19 pandemic, models were trained and evaluated under two scenarios — including and excluding that year. The best results for predicting 2023 during validation were obtained excluding 2020 data, which justifies this choice for the final analyses. This approach ensures robust, interpretable models suitable for the educational context, providing effective support for academic coordination decision-making. Table 6 presents the hyperparameters tuned via grid search for each evaluated algorithm, specifying the values obtained under the two training scenarios considered.

The differences in the optimal hyperparameters between the two scenarios (Table 6) are a direct result of the influence of the 2020 data on the training set's distribution. The grid search process, identical for both scenarios, identified different optimal settings because each model adapted to a distinct set of underlying patterns.

For instance, when the anomalous 2020 data was included, the grid search favored a simpler Decision Tree (max_depth=5) compared to the model trained without 2020 data (max_depth=10). This suggests that a less complex model was required to prevent overfitting to the atypical patterns of the pandemic year, thereby improving generalization. Similarly, the change in KNN's distance metric from 'manhattan' to 'euclidean' reflects how the underlying data distribution was altered by the 2020 data, making a different distance function more effective.

3.5. Model Evaluation

The model evaluation employed specific metrics for each phase. During the training and validation phase (2017–2022), recall and AUC-ROC were prioritized due to data imbalance and the strategic importance of correctly identifying students at risk of dropout. In particular, AUC-ROC was considered as the key metric for hyperparameter tuning, ensuring a balance between sensitivity and specificity during model optimization.

In the final testing phase (2023), a more comprehensive evaluation was adopted using the following metrics: accuracy, precision, recall, F1-score, and AUC-ROC. This choice aims to provide a balanced analysis of performance, considering both the ability to correctly identify cases and the minimization of errors.

Table 6. Hyperparameters tuned via grid search for the evaluated models, with and without the year 2020 in the training set.

Model	Training Data	Tuned Hyperparameters
Logistic Regression	Without 2020	C=0.4, max_iter=100, penalty='l2', solver='sag'
	With 2020	C=0.4, max_iter=100, penalty='l1', solver='saga'
Decision Tree	Without 2020	criterion='gini', max_depth=10, max_features='log2', min_samples_leaf=3, min_samples_split=10
	With 2020	criterion='gini', max_depth=5, max_features=None, min_samples_leaf=5, min_samples_split=2
Random Forest	Without 2020	bootstrap=True, n_estimators=100, criterion='entropy', max_depth=3, max_features=None, min_samples_leaf=3, min_samples_split=10
	With 2020	bootstrap=True, n_estimators=300, criterion='entropy', max_depth=3, max_features=None, min_samples_leaf=5, min_samples_split=2
XGBoost	Without 2020	gamma=0.1, learning_rate=0.01, max_depth=3, n_estimators=250, reg_alpha=0.1, reg_lambda=1
	With 2020	gamma=0, learning_rate=0.01, max_depth=3, n_estimators=250, reg_alpha=0, reg_lambda=1
K-Nearest Neighbors (KNN)	Without 2020	metric='manhattan', n_neighbors=7, p=1, weights='uniform'
	With 2020	metric='euclidean', n_neighbors=9, p=1, weights='uniform'
Support Vector Classifier (SVC)	Without 2020	C=0.1, degree=4, gamma='scale', kernel='poly'
	With 2020	C=0.1, degree=2, gamma='auto', kernel='sigmoid'

To reflect a realistic application scenario, a temporal data split was adopted: data from 2017 to 2022 were used for training and validation, while data from 2023 were reserved exclusively for the final, unseen test set. This approach reinforces the challenge for the models by simulating a real-world future prediction task. It introduces a significant level of difficulty, as it prevents the model from accessing examples from the same period as the test data during training. Consequently, the model must be robust enough to handle variations in data patterns across different years—such as changes in course offerings, new student cohorts, or curricular modifications—thus emphasizing the need for robust metrics for both validation and testing.

This approach also makes the process more demanding, requiring the model to be sufficiently generalizable to handle previously unseen events, simulating a real-world future prediction scenario. Therefore, it was expected that this data split would impose an additional level of difficulty in the learning process, which reinforces the importance of robust metrics such as recall and AUC-ROC during the validation phase, along with a comprehensive analysis using multiple metrics during the test phase. Based on the results obtained through cross-validation, the models were then evaluated on the 2023 test set, whose results and comparative analyses are presented in Section 4.

4. Results

This section presents the main findings of the study, starting with the exploratory data analysis and student profiling. Then, it shows the results of training and cross-validation, with and without the 2020 data. Afterwards, the 2023 test results are reported, highlighting the final performance of the models. Finally, a comparative analysis between the algorithms is conducted, emphasizing their predictive behavior and generalization capacity.

4.1. Training and Testing Datasets Analysis

This subsection presents a comparative analysis of the datasets used for training and testing, focusing on dropout rates and the general characteristics of the variables. The objec-

tive is to identify differences that may affect the predictive models' ability to generalize to future data, highlighting phenomena such as *data drift* and *concept drift*, which are common in longitudinal educational datasets.

Dropout Rate

Table 7 shows the dropout rates in the training sets (with and without 2020 data) and the test set (2023). The inclusion of 2020 — an atypical year — reduced the training set's dropout rate from 38.5

Table 7. Dropout Rates by Dataset.

Dataset	Total Records	Dropouts (DR=1)	Dropout Rate
Training (Without 2020)	3813	1469	0.3852
Training (With 2020)	4905	1566	0.3193
Test (2023)	814	242	0.2973

The differences in dropout rates and variable distributions indicate changes in student profiles over time, impacting model performance. These variations illustrate the presence of *data drift* and *concept drift*, which challenge the models' ability to generalize to 2023 data. Therefore, we assessed the impact of including or excluding the year 2020 in training to determine the most robust approach.

Moreover, using appropriate metrics such as F1-score and AUC-ROC is essential to address data imbalance and accurately evaluate predictive performance. This analysis reinforces the importance of monitoring and understanding temporal changes in educational data to improve predictive modeling. Table 8 summarizes the differences in variables between the training and test sets, highlighting changes that may affect model performance, such as shifts in gender distribution and the appearance of new categories in the program type variable in the test set.

Table 8. Summary of Independent Variables Relationships with Dropout and Differences Across Datasets.

Variable	Relationship with Dropout	Differences Across Datasets	Impact on Modeling
Gender	Higher dropout among males; lower among females	Female dropout is lower and male dropout is higher in the test set compared to training	Indicates concept drift in gender-dropout relationship
Age	Similar across datasets; higher minimum age in test set	Test set excludes younger age groups present in training	May affect accuracy if dropout is sensitive to age
Course Type	Lower dropout in FIC courses in 2023; changes in course proportions	More EST courses in test; FIC dropout decreased in test set	Changes in distribution hinder generalization
Program Type	New categories appear in test set	NSA, PROINT, PROSUB only present in test set	Models trained without these categories may lose information
Course Shift	Changes in distribution and categories	New categories in test not seen in training	May impact prediction due to lack of training examples
RC	Differences in categories and associated dropout	Absence or different representation of "ND" in test set	May cause difficulty in pattern recognition

4.2. Performance Analysis

This section presents the results from the *cross-validation* and *final test* phases for the predictive models aimed at identifying students at risk of dropout at IFPB specific Campus. The evaluation metrics were AUC-ROC and recall, prioritizing the identification of dropouts, especially given the imbalanced data scenario.

The decision to exclude the year 2020 from the training set was based on consistent evidence observed during the cross-validation process. Models trained without 2020 systematically achieved higher AUC-ROC scores across all algorithms, indicating better generalization capacity and more reliable discrimination between classes. Additionally, the results without 2020 exhibited lower variance, suggesting greater stability and robustness during training. Although models trained with 2020 sometimes achieved higher recall — as observed particularly in Random Forest and SVM — this came with a noticeable increase in the variability of results, indicating reduced reliability. In contrast, the recall obtained without 2020 presented a more balanced trade-off between sensitivity and stability. Therefore, prioritizing a more consistent AUC-ROC performance combined with stable recall outcomes justified the adoption of the training scenario without the year 2020.

Table 9. Cross-validation Results (k=10) — Average AUC-ROC and Recall for Models with and without 2020.

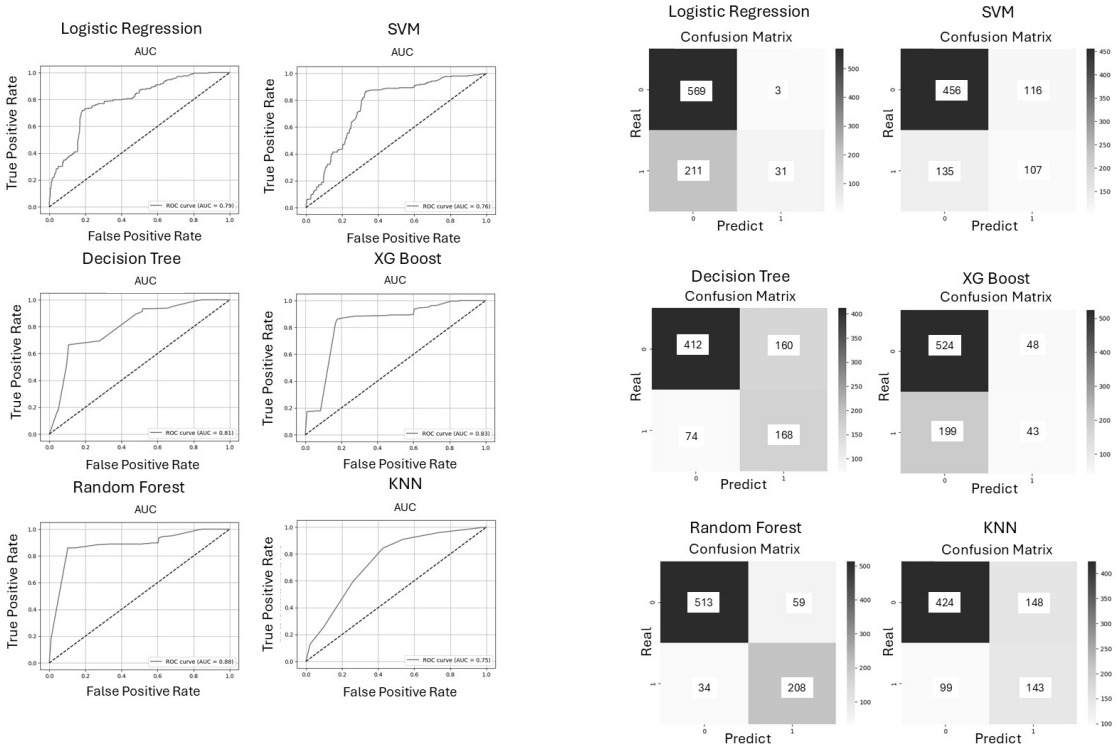
Model	Training	Average AUC (CV)	Average Recall (CV)
Logistic Regression	Without 2020	**0.76**	0.64
	With 2020	0.69	**0.72**
Decision Tree	Without 2020	**0.73**	**0.61**
	With 2020	0.67	0.58
Random Forest	Without 2020	**0.75**	0.63
	With 2020	0.68	**0.83**
XGBoost	Without 2020	**0.77**	**0.87**
	With 2020	0.69	0.84
KNN	Without 2020	**0.73**	**0.52**
	With 2020	0.64	0.28
SVM	Without 2020	**0.76**	0.66
	With 2020	0.65	**0.90**
Mean (SD)	Without 2020	**0.75 (0.02)**	0.66 (0.12)
	With 2020	0.67 (0.02)	**0.69 (0.23)**

In the test phase, using the 2023 data, Random Forest achieved the best overall performance (AUC=0.88 and recall=0.86), demonstrating a high capacity for generalization. On the other hand, XGBoost, despite maintaining a high AUC (0.87), showed a significant drop in recall (0.18), suggesting that the decision threshold may not be optimal for the test scenario. This behavior indicates that, although the model is capable of ranking instances correctly (as reflected by the high AUC), its ability to capture positive cases at the default threshold is limited. Therefore, future adjustments in the decision threshold could be considered to balance precision and recall according to the operational needs of the academic monitoring environment.

Table 10 presents the models' performance on the test set. Figures 1a and 1b illustrate the ROC curves and confusion matrices results, respectively, providing a visual analysis of model performance.

Table 10. Test Results (2023) — Model Performance.

Model	AUC	F1-score	Recall	Precision	Accuracy
Logistic Regression	0.79	0.22	0.13	**0.91**	0.74
Decision Tree	0.81	0.59	0.69	0.51	0.71
Random Forest	**0.88**	**0.82**	**0.86**	0.78	**0.89**
Support Vector Machine	0.76	0.46	0.44	0.48	0.69
XGBoost	0.83	0.26	0.18	0.47	0.70
KNN	0.75	0.54	0.59	0.49	0.70



(a) *ROC* curves of the models evaluated on the 2023 test set.

(b) Confusion matrices of the models evaluated on the 2023 test set.

Figure 1. Comparison between ROC curves and confusion matrices for the models evaluated on the 2023 test set.

4.3. Comparative Model Analysis

The comparative analysis shows that excluding 2020 from training resulted in more stable models with better generalization capabilities. Random Forest stood out with the best performance, achieving an AUC-ROC of 0.88 and recall of 0.86 in the test, demonstrating strong robustness to data changes.

Models such as XGBoost, SVM, and KNN proved more sensitive to temporal variations, indicating difficulties in handling the changes in data profiles for 2023. This sensitivity manifested clearly in the test phase results, leading to poor F1-scores. For instance, XGBoost's recall dropped sharply from 0.87 during validation to just 0.18 on the test set. The Logistic Regression model, despite high precision (0.91), was rendered ineffective by its extremely low recall (0.13), compromising its ability to identify dropout cases. The SVM likewise achieved only intermediate performance.

This collective underperformance underscores a critical challenge in real-world educational data mining known as temporal concept drift: even with a comprehensive hyperparameter tuning process using grid search, with the resulting optimal hyperparameters shown in Table 6, the optimal settings learned from historical data failed to generalize.

This difficulty was likely caused by structural and institutional changes in the 2023 data, which altered the optimal decision boundary. Despite generally high *AUC-ROC* values (ranging from 0.75 to 0.88), metrics such as *recall* and *F1-score* were affected by changes in data distribution—a common characteristic in educational contexts. Therefore, the Random Forest without the 2020 data was selected as the most suitable model, balancing both discriminative performance and sensitivity.

The Random Forest model demonstrated robust and consistent performance in both the validation phase and the test set, outperforming all other models in key metrics such as *AUC* and *recall*. In contrast, algorithms like Logistic Regression showed inferior performance, especially in terms of sensitivity (*recall*), highlighting their limitations for this type of problem. This result underscores the effectiveness of random tree-based ensemble methods in the studied educational context, demonstrating their ability to capture complex and variable patterns without overfitting. Furthermore, the use of ensemble models proved advantageous compared to simpler and more easily interpretable models, which, although closer to intuitive human decision-making, lack the accuracy and adaptability offered by machine learning.

Finally, the analysis reinforces the importance of continuous monitoring, periodic model updates, and the adoption of *concept drift* detection techniques to ensure the effectiveness of predictive systems in academic management.

5. Conclusion

Based on the results, the Random Forest model trained without 2020 data showed the best generalization and predictive performance on the 2023 test set, supporting its choice as the institutional predictive model for the specific IFPB campus. With the release of 2024 data, the model can be updated by including these data while keeping class balancing methods. This will enable validation of predictions for the new period and monitoring of potential concept drifts in the institutional context.

The future work will focus on the following directions:

- **System Implementation:** Develop an institutional tool integrating the predictive model with an academic dashboard for practical use.
- **Dataset Enrichment:** Incorporate more detailed academic and behavioral variables (e.g., grades and attendance) to enhance predictive power.
- **Impact Evaluation:** Conduct longitudinal studies to assess the real-world effectiveness of model-guided interventions in reducing dropout.
- **Methodological Enhancement:** Address *concept drift* using strategies such as sliding windows and dynamic ensemble approaches for long-term robustness.
- **Interpretation Analysis:** Apply techniques such as feature importance and SHAP values to provide actionable insights on dropout risk factors.

These initiatives aim to strengthen evidence-based academic management at IFPB and improve student retention policies.

References

Albreiki, B., Zaki, N., and Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, 11(9).

Alnasyan, K. and et al. (2024). Deep learning techniques for predicting student performance in virtual learning environments: A systematic review. *IEEE Access*.

Alturki, S., Hulpuș, I., and Stuckenschmidt, H. (2022). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2):281–305.

Colpo, A., Silva, M. L., and Oliveira, R. (2024). Análise preditiva da evasão escolar no ensino técnico federal. *Revista Brasileira de Informática na Educação*, 32(1):120–138.

Colpo, G. and et al. (2024). Predicting student dropout using machine learning: A systematic review. *Journal of Educational Data Science*.

Estrada-Molina, O., Mena, J., and López-Padrón, A. (2024). The use of deep learning in open learning: A systematic review (2019 to 2023). *International Review of Research in Open and Distributed Learning*, 25(3).

Guleria, S. and Sood, M. (2014). Data mining in education: Review and future directions. *International Journal of Computer Applications*.

Hassan, M., Zhang, Y., and Liu, T. (2024). Predicting student dropout with random forest: A case study in technical education. In *Proceedings of the 2024 International Conference on Artificial Intelligence in Education (ICAIED)*, pages 101–110.

Hegazi, M. O. and Abugroon, M. A. (2016). The state of the art on educational data mining in higher education. *International Journal of Computer Trends and Technology (IJCTT)*, 31(1):46–55.

Khedr, A. E. and El Seddawy, A. I. (2015). Using random forests technique for early stage students' performance prediction. *International Journal of Computer Applications*, 113(5):1–5.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI)*, volume 2, pages 1137–1143. Morgan Kaufmann.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4):331–344.

Krüger, C. F., Dias, J., and Souza, A. P. (2023). Aplicação de técnicas de mineração de dados para predição da evasão escolar. *Revista de Estatística Aplicada*, 19(2):45–67.

Lynn, N. D. and Emanuel, A. W. R. (2021). Using data mining techniques to predict students' performance. a review. In *IOP Conference Series: Materials Science and Engineering*, volume 1096, page 012083. IOP Publishing.

Mduma, N., Kalegele, K., and Machuve, D. (2019). A survey of machine learning techniques for student dropout prediction. *International Journal of Advanced Computer Science*.

Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state-of-the-art. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6):601–618.

Shahiri, A., Husain, W., and Rashid, N. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*.