Predicting Student Dropout Rates at Higher Degree Using Machine Learning and Dimensionality Reduction

Wanessa S. Bezerra[1], **Karliane M. O. Vale**[2], **Flavius L. Gorgônio**[2]
Fabrício V. A. Guerra[2], **Arthur C. Gorgônio**[3], **Anne M. P. Canuto**[3]

[1]Laboratório de Inteligência Computacional Aplicada a Negócios (LABICAN)
Universidade Federal do Rio Grande do Norte (UFRN)
Rua Joaquim Gregório, 296 – 59.300-000 – Caicó – RN – Brasil

[2]Departamento de Computação e Tecnologia (DCT)
Universidade Federal do Rio Grande do Norte (UFRN)
Rua Joaquim Gregório, 296 – 59.300-000 – Caicó – RN – Brasil

[3]Departamento de Informática e Matemática Aplicada (DIMAP)
Universidade Federal do Rio Grande do Norte (UFRN)
Campus Universitário – Lagoa Nova – 59.078-970 – Natal – RN – Brasil

wanessa.bezerra.126@ufrn.edu.br
{karliane.vale, flavius.gorgonio, fabricio.guerra}@ufrn.br
arthur.gorgonio.099@ufrn.edu.br, anne.canuto@ufrn.br

Abstract. *This study proposes four indices for predicting student dropout in a computing course at a higher education institution, using machine learning (ML) and dimensionality reduction. Ten classification algorithms were applied to three datasets, including versions with and without the proposed indices. The best performance was achieved by QDA. SHAP analysis highlighted persistence and number of enrollments as the most relevant predictors.*

1. Introduction

Student dropout is a global concern that affects both public and private universities [Fuentes et al. 2024]. This phenomenon is influenced by a variety of factors, including financial constraints, personal issues, academic difficulties, lack of motivation, and social pressures, which makes predicting which students are at risk of dropping out a challenging task [Abdi et al. 2024]. In addition, institutional factors, such as successive failures and the mismatch between academic training and demands from the labor market, also play an important role in the dropout process [Fukao et al. 2023]. Other factors influencing course withdrawal are the enrollment of students in a workload that exceeds their capacity and the difficulties faced in specific subjects [Rabelo and Zárate 2024].

Dropout in higher education is a critical problem with social, economic, and institutional consequences. Identifying and understanding the factors contributing to dropout is essential for developing effective strategies that promote academic retention, reducing the negative impacts on students, institutions, and society. Courses such as Computer Science and Information Systems face particular challenges due to the effort required to align academic training with labor market demands. Thus, early identification of at-risk students and the implementation of targeted interventions emerge as key measures to mitigate these adverse effects [Nurmalitasari et al. 2023,Naseem et al. 2019].

Although the use of ML techniques to predict dropout is already widespread, differences in educational policies and the difficulty of accessing updated data present significant challenges, requiring constant adaptations to predictive models [Krüger et al. 2023]. Given this, the present study aims to propose indices that influence the prediction of student dropout in higher education. These indices summarize patterns of progression, failure, and course load, obtained through the processing of students' academic records. Also, the indices were defined based on the literature and empirical evidence and they were analyzed using quantitative and predictive techniques to assess their relevance and impact. This research aims to provide support for practical interventions to reduce dropout in the program, as well as to offer a replicable approach for other educational contexts and advances in the literature on academic dropout.

This article is organized as follows: Section 2 discusses the main factors related to dropout in higher education and its early identification using data mining and ML techniques. Section 3 offers a review of related works. Section 4 details the methodology of this study. Section 5 describes the experiments, while Section 6 presents the obtained results. Finally, Section 7 provides the conclusions and suggestions for future work.

2. Theoretical Aspects

The study in [Fuentes et al. 2024] reports declining academic success at Cebu Technological University's Faculty of Computing, with only 42% of 2018-2019 students graduating by 2021-2022, and 33% for the 2019-2020 cohort. Dropout rates ranged from 11% to 24%, reaching 25% in the Information Systems program in 2021-2022. These figures reflect a broader global trend, especially in computing-related programs. In a computing program at a higher education institution Bachelor's program in Information Systems (BSI) at UFRN, dropout is a major concern. From 2009.1 to 2022.2, only 17.73% of students completed the program, with a dropout rate of 60.32% and 21.08% still enrolled. A review of 162 studies (2010–2022) on machine learning for academic prediction found most focused on identifying students at risk of failure (129 studies), while 33 addressed dropout or retention [Alalawi et al. 2023].

Machine learning-based solutions have shown potential in identifying and predicting factors influencing school dropout, such as social well-being, academic performance, learning conditions, demographic characteristics, and others [Prasanth and Alqahtani 2023]. Among these factors, academic aspiration stands out, reflecting students' motivation and commitment to achieving their educational goals. The alignment between academic performance and individual goals promotes healthier integration into the university environment, strengthening students' engagement and motivation, which, in turn, contributes to reducing the risk of dropout [Singh and Alhulail 2022].

Given the context of Brazilian higher education, for instance, notably, the federal network has experienced remarkable growth in its share of enrollments over the years [INEP and Teixeira. 2023]. Between 2012 and 2022, there was a significant 23.7% increase in enrollments in the federal network, showing the relevance of this sector to education in the country. In 2022, more than 1.3 million students graduated courses, reflecting the positive impact of higher education on professional training.

3. Related Work

Several studies explored dropout causes and early risk detection in higher education [Alyahyan and Düştegör 2020,Fernández-García et al. 2021,Freitas et al. 2020,Kim et al. 2023, Niyogisubizo et al. 2022, Shohag and Bakaul 2021]. Thus, those researches show the challenges faced by educational institutions and the relevance of using data mining and ML to improve dropout prediction.

In [Alyahyan and Düştegör 2020], a set of guidelines provides for educators to apply data mining to predict student success. Therefore, a literature review was conducted, and the state-of-the-art was compiled into a systematic process, where possible decisions and parameters are extensively addressed and explained. [Fernández-García et al. 2021] developed predictive systems capable of identifying high-risk dropout students in advance was examined, monitoring them from enrollment to the fourth semester at regular semester intervals. This aims to assist in formulating effective preventive policies, primarily applied to engineering students using real data. According to [Freitas et al. 2020], predicting school dropout based on socioeconomic data is essential for educational management, as it enables early interventions to reduce attrition. Hence, the author proposes automating the prediction process using an Internet of Things system to forecast dropout rates through ML and socioeconomic data.

[Shohag and Bakaul 2021] aimed to enhance the performance of an early warning system for dropout prediction. Therefore, the classifiers Decision Tree, Naive Bayes, Random Forest, SVM, and KNN were used. In [Niyogisubizo et al. 2022], a novel ensemble approach was proposed based on a hybrid method that combines Random Forest, XGBoost, Gradient Boosting, and Feedforward Neural Networks to predict student dropout. In [Kim et al. 2023] prescriptive analysis is utilized to assess the necessity of intervention programs for students at risk of dropping out. Therefore, a fuzzy logic-based algorithm was employed, which considers five linguistic variables identified from a previous study. This study aimed to determine the key factors influencing dropout risk and predict such risks based on input variables. Therefore, this study focuses on the application of data mining and machine learning techniques to predict dropout of higher degree, using a customized approach adapted to the specificities of the context of the field of computation. Additionally, this research stands out by emphasizing institutional factors such as failure rates, academic history, and specific course characteristics.

4. Methodology

As explained, the data used were obtained from the open data center of a higher education institution [1], which provides public information about undergraduate programs, including a course in the field of computing. The data collection covered the period from 2009.1, the year the program was created, to 2022.2, totaling 62 files used to construct a consolidated dataset that offers an overview of students' academic trajectories in the program.

Therefore, the data were organized from four main sources. The first corresponds to class information (30 files), which includes details such as class code, semester, schedule, and assigned professor. The second, curricular components (1 file), presents a list of courses offered in the program, with information such as course code, name,

[1] UFRN Open Data: https://dados.ufrn.br

Table 1. Description of dataset attributes

Attribute	Description	Type	Attribute	Description	Type
student	Student identifier.	object	estate_resid	Student's state of residence.	object
unit	Current unit of the academic term.	float64	city_resid	Student's city of residence.	object
grade	Grade obtained in the course.	object	final_grade	Final grade in the course.	float64
replacement	Grade replacement (yes/no).	object	description	Final status (approved, failed, etc.).	object
absences	Total absences in the course.	float64	year	Year and term of course offering.	int64
gender	Student's gender.	object	course_name	The name of course.	object
birth_year	Student's birth year.	int64	workload	Total course workload.	int64
estate_origin	Student's state of origin.	object	entry_year	Year of admission to the program.	int64
city_origin	Student's city of origin.	object	status	Student's overall status in the program.	object

workload, and prerequisites. The third, course enrollments (30 files), contains records of students' enrollments in each course, detailing status (approved, failed, withdrawn), grades, and attendance. The fourth, supplementary student data (1 file), provides sociodemographic and academic information of students, such as age, gender, program, and admission date.

To unify this dataset, these sources were integrated using primary and foreign keys. Course enrollments were linked to curricular components and classes using the course identifier and class code. Next, the supplementary student data were associated with enrollments through each student's unique identifier. Subsequently, the attributes were carefully selected to ensure that the final dataset was properly structured for further analysis. This stage included organizing and creating new datasets derived from the initial consolidated dataset, preparing the information for analysis and prediction steps. The characteristics of attributes used are presented in Table 1.

Creating the dataset, the records underwent a preprocessing to ensure the quality and consistency of the dataset. The data cleaning combined filling strategies guided by academic meaning and the removal of occurrences irrelevant to the study. First, duplicate records were identified and excluded, removing duplicates. Next, missing values were handled using appropriate filling or exclusion strategies, depending on their potential impact on the analysis. The attributes were standardized, ensuring uniformity in formats and nomenclature used. Records with critical errors, such as missing grades, attendance, and unit data, underwent a detailed review. Those that could not be corrected were excluded to minimize biases in the results. Similarly, records with inconsistencies were removed to preserve the integrity of the dataset.

4.1. Proposed Indices

The proposed indices were defined based on prior studies that associate academic persistence, repeated failures, and course progression with dropout risk [Krüger et al. 2023, Niyogisubizo et al. 2022, Alyahyan and Düştegör 2020]. These variables are consistently reported in the literature as significant predictors and reflect core aspects of student engagement and academic trajectory. In the transformation, variables were modified for classification algorithms. As defined in Equation 1, the "persistence" index is calculated as the ratio between the hours attended and the total course workload (1830 hours), calculated from the total hours completed by students in mandatory courses until their exit from the program. To obtain this value, the total hours per student were summed using the attribute "ch_total". This index reflects the student's progress toward program completion.

The index "persistence_per_semester", which quantifies a student's academic progress about the expected duration of the program, is obtained by the ratio between the number of semesters attended—calculated by summing the semesters in which the student was enrolled in courses, counted from the year of admission—and the number of semesters required for regular program completion, see Equation 2. The index “failure_rate,” determines the proportion of hours in which the student failed to the total “hours_attended,” assisting in evaluating the failure impact on academic performance, shown in Equation 3.

Besides, the “num_enrollments” index was included to count the number of enrollment attempts made by students, excluding enrollments classified as “denied” and “withdrawn,” as these do not have a direct impact on academic performance. Finally, the target variable “status” was transformed into a binary format to simplify the analysis: the value “-1” denotes dropouts, while “1” indicates completion or remain enrollment.

Other indicators such as grade point average - GPA and enrollment duration were considered but not incorporated, as the selected indices already cover core aspects of persistence and academic difficulty. To minimize bias, the indices were calculated from aggregated historical data prior to modeling, and their influence was validated through SHAP analysis, which confirmed their relevance independently of the target variable.

$$persistence = \frac{hours\_attended}{total\_workload} \tag{1}$$

$$persistence\_per\_semester = \frac{semesters\_attended}{expected\_duration} \tag{2}$$

$$failure\_rate = \frac{failed\_hours}{hours\_attended} \tag{3}$$

5. Experimental Design

To assess the predictive indices related to student dropout, three distinct datasets were developed, designed to enable a comparative analysis that enhances accuracy. The initial dataset, consisting of 12,618 records and 18 attributes (detailed in Table 1), served as the foundation for constructing datasets BD1, BD2, and BD3, each designed with specific characteristics to provide a more detailed understanding of the dropout phenomenon. These datasets were structured to capture different perspectives of the problem, ranging from raw data to transformed and synthesized structures, as shown in Table 2. The academic status of students was defined as the class for the predictive models, allowing the analyses to focus on identifying key factors in course completion or dropout.

BD1 consists of raw information extracted directly from the initial database, containing 12,618 records and 6 attributes (student, final_grade, description, year, workload, course_name, entry_year, and status). The remaining attributes were removed as they were not relevant for calculating the proposed indices. These data were maintained in their original format, without preprocessing, serving as the structural basis for developing the dropout prediction indices. The decision to preserve raw data ensured the integrity of the information. BD2 was derived from the manipulating, categorizing, and pivoting the initial dataset, which contained 12,618 records and 6 attributes, because

Table 2. Description of datasets used for prediction

Dataset	Description	Dropout Class	Non-Dropout Class	Total Instances	Total Features
BD1	Raw database, unprocessed, containing selected attributes from the original dataset, which will be used to construct the proposed indices in this study.	6,500	6,118	12,618	6
BD2	Dataset derived from the original data (BD1), containing cleaned, categorized, and pivoted data based on the initial structure.	410	248	658	327
BD3	Dataset derived from BD2, resulting from data manipulation to obtain the indices proposed in the research.	410	248	658	4

Table 3. Descriptive Statistics of BD3

Statistic	Persistence	Failure Rate	Persistence per Semester	Number of Enrollments	Status
Mean	0.69	0.34	0.7	19.36	-0.611
Std. Dev.	0.45	0.32	0.5	12.86	0.792
Min	0.03	0	0.12	1	-1
25%	0.21	0.09	0.25	6	-1
50% (Median)	0.64	0.25	0.62	17	-1
75%	1.08	0.48	1.12	31	-1
Max	1.75	1	2	50	1

in BD1, each student has an instance for each course. Therefore, the main transformation in BD2 was to merge all student information into a single instance. After transformation, each row came to represent the complete academic history of a student, from admission to program completion or dropout. This reorganization resulted in BD2, which consists of 658 records and 327 attributes, providing a detailed view of individual academic trajectories.

The BD3 dataset, derived from BD2, uses data manipulation to transform the 327 attributes of BD2 into four predictive indices of academic dropout proposed in this study and uses them as attributes of BD3, as described in Section 4.1. The process involved removing inconsistent records, transforming raw variables into aggregated indices, categorizing continuous variables into specific ranges, and using pivoting to reorganize the data into a structure suitable for predictive analysis. After these adjustments, the dataset was reduced from 12,618 to 658 records, consolidating each student's academic history into a single row while preserving essential information. In Table 3, descriptive statistics were used to summarize the main characteristics of the BD3 dataset, whose attributes correspond to the proposed indices.

Since it is a classification task with a wide variety of available algorithms, ten supervised learning algorithms were selected, with the attribute "status" (dropout and non-dropout) defined as the target class. The choice of this diverse set of algorithms was based on their distinct characteristics, which offer specific capabilities to handle challenges such as data noise, class imbalance, and high dimensionality. The algorithms used were: Logistic Regression (LR), *Random Forest* (RF), Decision Tree (DT), *AdaBoost* (AB), *Gaussian Processes* (GP), *Naive Bayes* (NB), Artificial Neural Network (NN), *Quadratic Discriminant Analysis* (QDA), *K-Nearest Neighbors* (K-NN), and *Support Vector Machine* (SVM).

In order to evaluate the performance of the models in each scenario, accuracy and F1-score metrics were used, complemented by the 10-fold cross-validation technique, repeated three times. The adopted cross-validation method was stratified, a recommended approach to preserve class proportions in each subset, which is essential for classification models with imbalanced distributions. This approach ensures more stable estimates, allowing each sample in the dataset to be used for both training and testing in different iterations. Additionally, it reduces bias in model evaluation and promotes better generalization of results, making it valuable in scenarios with limited datasets [Wong and Yeh 2020].

To assess the obtained results from a statistical point of view, the Friedman test is applied to verify whether statistical differences exist among all combiners. The Friedman test determines whether observations related to k methods derive from the same population (similar performance) or if the observed differences occurred by chance (performance superiority), for a complete discussion on Friedmann test, see [Theodorsson-Norheim 1987]). A significance level of 0.05 is used for this test.

If the p-value is smaller than the established threshold, the null hypothesis is rejected, with a confidence level exceeding 95%. If statistical differences are detected among the analyzed methods, a post-hoc test will be applied, and the results presented in the Critical Difference (CD) Diagram. In the CD diagram, the performance of one method is statistically different from another if the difference between their average rankings exceeds the critical difference calculated by the CD. When two methods are similar, a horizontal line connects them. The CD will be applied using the Nemenyi test, a multiple-comparison test aimed at pairwise comparisons to identify statistical significance.

6. Results

This section presents the results of the empirical analysis performed in this study, comparing the performance of ten classifiers across three scenarios: using the proposed indices as attributes (BD3), applying them directly to the original data (BD1 and BD2), and applying the PCA (Principal Component Analysis) feature extraction method in both datasets. For a fair comparison with the proposed dataset, PCA was employed to reduce the dimensionality of the data, generating a new dataset with four columns. The objective is to compare the performance of the algorithms using both the dataset reduced by PCA and the one constructed with the proposed indices. The models were evaluated based on two performance metrics (accuracy and F1-score), statistical analyses (critical difference plot), and a model interpretation tool (SHapley Additive exPlanations - SHAP).

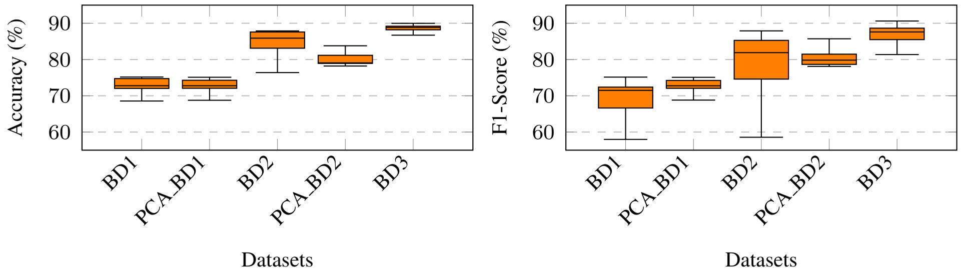
6.1. Performance Analysis

Table 4 presents the accuracy (Acc), standard deviation (std), and F1-score (F1) for different classifiers tested on two inicial datasets (BD1 and BD2), their respective PCA-transformed versions (PCA_BD1 and PCA_BD2) and the dataset using the proposed indices as attribute (BD3). The first column of the table lists the classifiers (class), identified by the abbreviations described in Section 5. The analysis of this table reveals that all classifiers achieved better performance with BD3, the dataset constructed using the indices proposed in this study, both in terms of accuracy and F1-score. Additionally, the standard deviation ranges from 0.03 to 0.05 in both metrics.

Still in Table 4, it can be observed that QDA achieves the highest accuracy (89.98%) and F1-score (90.62%) in BD3, consolidating itself as the best model in this

Table 4. Accuracy and F1-score results for different models and datasets.

Class	BD1		PCA_BD1		BD2		PCA_BD2		BD3	
	Acc ± std	**F1 ± std**	**Acc ± std**	**F1 ± std**	**Acc ± std**	**F1 ± std**	**Acc ± std**	**F1 ± std**	**Acc ± std**	**F1 ± std**
QDA	72.46 ± 0.01	72.46 ± 0.01	72.46 ± 0.07	72.46 ± 0.01	62.56 ± 0.05	63.64 ± 0.07	78.93 ± 0.05	78.72 ± 0.05	**89.98 ± 0.04**	**90.62 ± 0.03**
SVM	74.24 ± 0.01	35.03 ± 0.01	74.24 ± 0.06	74.17 ± 0.01	**87.89 ± 0.04**	47.85 ± 0.01	79.02 ± 0.04	78.82 ± 0.04	89.72 ± 0.04	82.65 ± 0.04
DT	68.56 ± 0.01	68.59 ± 0.01	68.19 ± 0.06	68.13 ± 0.01	82.47 ± 0.05	82.50 ± 0.05	78.21 ± 0.04	78.10 ± 0.04	89.19 ± 0.05	89.11 ± 0.05
GP	74.82 ± 0.01	74.79 ± 0.01	74.96 ± 0.01	74.94 ± 0.01	86.88 ± 0.04	84.67 ± 0.10	80.89 ± 0.04	80.84 ± 0.04	89.19 ± 0.04	81.37 ± 0.04
LR	73.08 ± 0.01	65.97 ± 0.01	73.08 ± 0.06	73.06 ± 0.01	86.22 ± 0.04	81.29 ± 0.04	78.93 ± 0.04	78.64 ± 0.04	88.99 ± 0.04	87.19 ± 0.04
AB	**75.18 ± 0.01**	**75.15 ± 0.01**	74.29 ± 0.05	74.25 ± 0.01	85.56 ± 0.04	85.49 ± 0.04	78.73 ± 0.05	78.54 ± 0.05	88.74 ± 0.04	88.68 ± 0.04
NN	74.91 ± 0.01	48.73 ± 0.14	**75.11 ± 0.06**	**75.09 ± 0.01**	**87.89 ± 0.04**	74.00 ± 0.10	81.57 ± 0.04	81.55 ± 0.04	88.34 ± 0.04	88.55 ± 0.04
RF	71.32 ± 0.01	71.16 ± 0.01	71.52 ± 0.04	71.51 ± 0.01	87.83 ± 0.03	**87.86 ± 0.04**	**83.79 ± 0.04**	**88.02 ± 0.04**	88.15 ± 0.04	88.02 ± 0.04
K-NN	71.98 ± 0.01	71.81 ± 0.01	71.97 ± 0.06	71.96 ± 0.01	85.05 ± 0.04	85.97 ± 0.04	81.25 ± 0.04	83.71 ± 0.04	87.62 ± 0.04	85.33 ± 0.04
NB	72.19 ± 0.01	72.19 ± 0.01	72.37 ± 0.06	72.37 ± 0.01	76.35 ± 0.05	76.38 ± 0.05	79.12 ± 0.04	81.28 ± 0.04	84.61 ± 0.05	85.97 ± 0.04



(a) Accuracy

(b) F-Score

Figure 1. Boxplot Results by Dataset

experiment. Additionally, other classifiers such as SVM, DT, and GP also obtained high accuracies in BD3, reaching 89.72%, 89.19%, and 89.19%, respectively. Regarding the F1-score, the classifiers DT (89.11%), AB (88.68%), NN (88.55%), and RF (88.02%) also demonstrated significant performance. These findings highlight the consistency of the proposed approach in this study, indicating that the proposed indices positively impact model performance. Moreover, it is important to note that although Principal Component Analysis (PCA) demonstrated efficiency, in the datasets where this technique was applied for dimensionality reduction (PCA_BD1 and PCA_BD2), the classifiers' performance did not surpass that observed in BD3. This emphasize that proposed indices achieves the best results across all classifiers, helping the classifiers to predict dropouts.

Additionally, Figure 1 present the boxplot graphs over the evaluated measures. These figures illustrate how the different datasets and their PCA-transformed versions influence classifier performance in terms of accuracy (Figure 1a) and F1-score (Figure 1b). The boxplot is a statistical tool that provides a clear visualization of data distribution, highlighting key information such as the median (represented by the central line in the box), quartiles (upper and lower box limits), and variability range (whiskers). The boxplot highlights the superior performance and lower variability of BD3 across classifiers, reinforcing the robustness of the proposed indices compared to other datasets. The based on the results achieved and the boxplot, it can be concluded that the identified pattern highlights that success is not exclusively linked to the type of classifier used but fundamentally depends on the careful organization and preparation of the data. The systematic and well-structured approach adopted in this study proved to be decisive in achieving superior results in terms of accuracy and F1-score. This organization ensures a solid foundation for future analyses, maximizing the potential of classifiers and reducing the influence of external factors such as data variability.

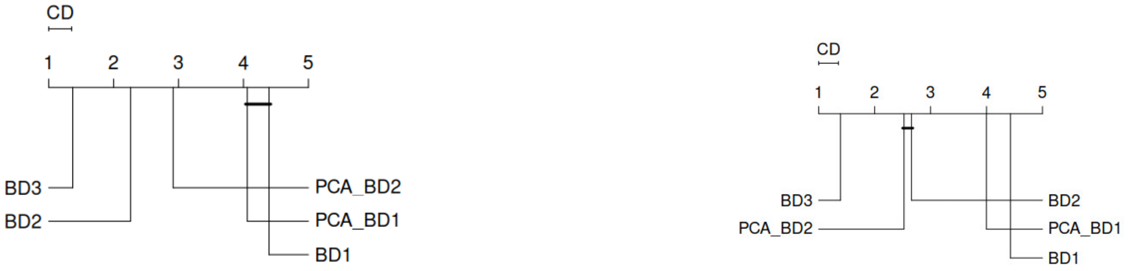
In summary, these findings demonstrate that the proposed methodology not only enhances the consistency of classifier performance but also establishes a methodological advantage. The quality of data organization emerges as a factor as critical as the choice of classifier, reaffirming its importance as a foundation for successful analyses and paving the way for more efficient and reliable applications.

6.2. Statistical Analysis

Figure 2 compare the rankings of datasets using a critical difference diagram (CD) with accuracy and F1-score metrics, respectively. In both cases, the results show that BD3 ranks first, exhibiting the lowest average ranking, making it the most efficient configuration among those evaluated. This analysis corroborates with the idea that the proposed indices are relevant for predicting students at risk of dropout. Although, BD1 and PCA_BD1 had the highest rankings, with BD1 standing out as the worst-performing dataset, with rankings between 4 and 5. However, with accuracy (Figure 2a) the connections between these datasets suggest that there is no statistically significant difference between them. This outcome was expected since BD1 and PCA_BD1 were based on raw data without any preprocessing. BD2 and PCA_BD2 ranked second and third, respectively with accuracy (Figure 2a), and their performances were statistically different, the opposite situation is observed with F1-Score (Figure 2b). Once again, this analysis reinforces that the success of predictive models depends not only on data organization but also on the appropriate selection of preprocessing methods and feature selection techniques.

6.3. SHAP Analysis

As explained, SHAP was used to enhance the interpretability of results, enabling a detailed analysis of the impact of predictive variables on academic dropout. SHAP is widely recognized for its applicability in both academic and practical contexts and is considered a reliable methodology for interpretable ML analyses [Salih et al. 2024]. Here, SHAP helped identify key variables and understand their interactions, as well as how these interactions influenced the model's predictions. This approach contributed to a deeper understanding of model behavior, promoting transparency in the results. Figure 3a shows the impact of predictive variables on the QDA model, which achieved the best performance. Given similar patterns in other classifiers, only QDA results are shown for simplicity. In this figure, the horizontal axis reflects the positive or negative impact of each variable on the classifier prediction, while the vertical axis ranks the variables in



(a) Accuracy **(b)** F-Score

Figure 2. CD Diagram for Statistical Evaluation

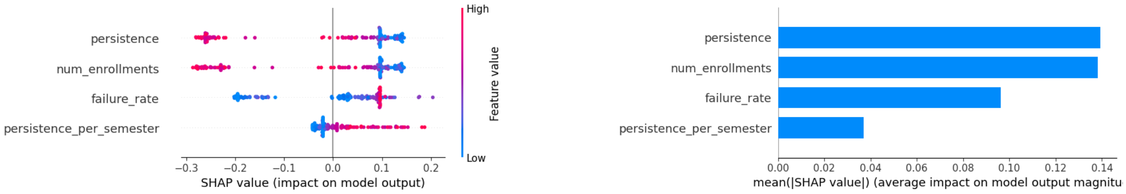
descending order of importance. Each point represents an observation from the dataset, with colors indicating variable values: red for high values and blue for low values. The dispersion of points within each row illustrates the variability of the individual variable impacts on the model.

The SHAP results reveal key patterns regarding the impact of predictive indices on the QDA model, highlighting the variable “persistence” as the most influential. The broad spread of points along the horizontal axis shows that different values of this index, represented by high (red) and low (blue) values, significantly affect the model’s predictions, both positively and negatively. High values, generally associated with a positive impact suggest that students with greater engagement and continuity have a lower risk of dropping out. However, a significant number of red points on the left side suggests that very high values may reflect specific situations, such as delays in course completion, lack of motivation, or burnout, ultimately increasing the likelihood of dropout.

The “num enrollments” index shows a similar pattern. High values on the right have a positive impact, suggesting that enrolling in an appropriate number of courses helps students persist. However, High values on the left suggest that excessive enrollment can cause exhaustion, increasing dropout risk. For the “failure rate” index, low values on the right have a positive impact, showing that students with fewer failures are more likely to persist. High values on the left indicate that a high failure rate negatively impacts persistence, contributing to dropout risk. Finally, the “persistence per semester” index shows that high values on the right are linked to retention, while red points on the left suggest that even highly persistent students may face issues like poor performance or poor course selection, raising their dropout risk.

Additionally, Figure 3b presents a bar diagram summarizing the average impact of each variable on the model, using the absolute mean SHAP values. This diagram allows for the identification of the most influential variables in aggregate form, ranking them in descending order of relevance. As highlighted, “persistence” and “num_enrollments” remain the most critical factors, reflecting their consistent influence on the QDA model. The indices “failure_rate” and “persistence_per_semester” have a lower average impact but still play a significant role in specific situations, demonstrating the importance of considering them in the modeling process.

In summary, the findings highlight the importance of academic factors like “persistence” and “num enrollments” in predicting dropout. Meanwhile, the “failure rate” and “persistence per semester” support the need for policies to enhance retention and reduce the effects of repeated failures and withdrawals. A key insight is the observed



(a) SHAP of the QDA Model

(b) Indices’ Contribution

Figure 3. Analysis of the Proposed Indices

impact of the number of enrollments on dropout rates. While the persistence metric aligns with common sense and prior literature, the connection between multiple enrollments and higher dropout risk is less intuitive and therefore noteworthy. SHAP analysis also underscores the model's ability to identify nonlinear relationships and complex interactions among variables.

7. Conclusion

This study proposed four predictive indices to support early detection of dropout risk in higher education. Among them, persistence had the greatest impact, followed by num_enrollments, failure_rate, and persistence_per_semester. These variables highlight the central role of academic performance and progression in student dropout.

All models achieved superior accuracy and F1-score using the BD3 dataset, which incorporates the proposed indices and reflects the students' complete academic history. Statistical tests confirmed BD3's effectiveness, with the QDA classifier delivering the best results. In contrast, dimensionality reduction with PCA (in PCA_BD1 and PCA_BD2) did not improve performance. These findings underscore that data quality and attribute selection are as critical as classifier choice. Less refined datasets, like BD1, showed lower predictive effectiveness.

The results offer practical insights for educational management. Monitoring key variables—such as failure rates and progression—can help identify at-risk students early. Institutions should also implement targeted academic and emotional support policies. Future work could include regression models to estimate dropout timing and the incorporation of new variables (e.g., socioeconomic or extracurricular data) to enhance predictions. Finally, applying this methodology in different contexts can help validate and generalize its use.

References

Abdi, H. M., Hassan, M. A., and Saralees, N. (2024). Predicting student dropout rates using supervised machine learning: Insights from the 2022 national education accessibility survey in somaliland. *Applied Sciences*, 14(17).

Alalawi, K., Athauda, R., and Chiong, R. (2023). Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. *Engineering Reports*, 5.

Alyahyan, E. and Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *Inter. Jour. of Educational Technology in Higher Education*, 17(1):3.

Fernández-García, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., and Sánchez-Figueroa, F. (2021). A real-life ml experience for predicting university dropout at different stages using academic data. *IEEE Access*, 9:133076–133090.

Freitas, F. A. d. S., Vasconcelos, F. F., Peixoto, S. A., Hassan, M. M., Dewan, M. A. A., Albuquerque, V. H. C. d., and Filho, P. P. R. (2020). Iot system for school dropout prediction using machine learning techniques based on socioeconomic data. *Electronics*, 9(10):1613.

Fuentes, N., Feliscuzo, L., and Sta Romana, C. L. (2024). Enhancing student retention in higher education: A fuzzy logic approach to prescriptive analytics. In *2024 IEEE 7th Inter. Conference on Big Data and Artificial Intelligence (BDAI)*, pages 41–48.

Fukao, A., Colanzi, T., Martimiano, L., and Feltrim, V. (2023). Study on evasion in computer science courses at the state university of maringá. In *Proceedings of the 3rd Brazilian Symposium on Computing Education*, pages 86–96, Porto Alegre, RS, Brazil. SBC.

INEP, N. I. f. E. S. and Teixeira., R. A. (2023). Higher education census 2022: Statistical notes. 2023.

Kim, S., Yoo, E., and Kim, S. (2023). Why do students drop out? university dropout prediction and associated factor analysis using machine learning techniques.

Krüger, J. G. C., Britto, A. S., and Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Apps*, 233:120933.

Naseem, M., Chaudhary, K., Sharma, B., and Lal, A. G. (2019). Using ensemble decision tree model to predict student dropout in computing science. In *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–8.

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., and Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3:100066.

Nurmalitasari, N., awang long, Z., and Mohd Noor, F. (2023). Factors influencing dropout students in higher education. *Education Research Inter.*, 2023:1–13.

Prasanth, A. and Alqahtani, H. (2023). Predictive modeling of student behavior for early dropout detection in universities using machine learning techniques. In *2023 IEEE 8th Int'l Conference on Engineering Technologies and Applied Sciences*, pages 1–5.

Rabelo, A. M. and Zárate, L. E. (2024). A model for predicting dropout of higher education students. *Data Science and Management.*

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., and Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems.*

Shohag, S. I. and Bakaul, M. (2021). A machine learning approach to detect student dropout at univ. *Int'l Journal Advanced Trends in Computer Science and Engineering.*

Singh, H. P. and Alhulail, H. N. (2022). Predicting student-teachers dropout risk and early identification: A four-step logistic regression approach. *IEEE Access*, 10:6470–6482.

Theodorsson-Norheim, E. (1987). Friedman and quade tests: Basic computer program to perform nonparametric two-way analysis of variance and multiple comparisons on ranks of several related samples. *Computers in Biology and Medicine*, 17(2):85–99.

Wong, T.-T. and Yeh, P.-Y. (2020). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions Knowledge and Data Engineering*, 32(8):1586–1594.