Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 17/06/2024; 1st round notif.: 09/10/2024;
Camera ready: 20/12/2024; Edition review: 03/04/2025; Available online: 03/04/2025; Published: 03/04/2025;

Exploring Feature Reduction for Dropout Predicting in Higher Education in Brazil

André Menolli
Northern Paraná State University
State University of Londrina
ORCID: 0000-0002-4755-8031
menolli@uenp.edu.br

Gustavo Marcelino Dionísio
State University of Londrina
ORCID: 0000-0003-0924-2666
gugamd93@gmail.com

Alan Silva da Paz Floriano
Federal Institute of Paraná
ORCID: 0000-0001-7493-6192
alan.floriano@ifpr.edu.br

Thiago Adriano Coleti
Northern Paraná State University
University of São Paulo
ORCID: 0000-0002-1078-4334
thiago.coleti@uenp.edu.br

Abstract
Dropout within the higher education system is a prevalent and intricate phenomenon characterized by a multitude of reasons that can differ significantly from one context to another. Developing machine learning models and discerning the key features for diverse contexts poses a considerable challenge. In this paper, we propose a process based on feature selection to create and evaluate machine learning models for predicting dropout in the higher education system. The approach not only outlines the essential steps for model development in any context but also emphasizes the identification of the most critical features. We conducted a comprehensive study across five distinct contexts within Brazilian higher education, specifically focusing on face-to-face courses. Through this process, we identified the most important features for predicting dropout. The results highlight that the correlation between a student's enrollment duration and the percentage of the course completed emerges as the primary predictor of dropout. However, we noticed the fundamental role of context in predicting dropout. Moreover, in all scenarios explored, it was possible to create more accurate models with a reduced set of features compared to the original models.
Keywords: *ropout, Feature Selection, Machine Learning, Higher Education System.*

Cite as: Menolli, A. L. A., Dionisio, G. M., Floriano, A. da P. S., & Coleti, T. A. (2025). Exploring feature reduction for dropout predicting in higher education in Brazil. Revista Brasileira de Informática na Educação, vol. 33, pp. 106–129. https://doi.org/10.5753/rbie.2025.4619.

1 Introduction

Higher education plays a crucial role in the development of countries. It offers not only technical learning but also serves as a valuable cultural and scientific asset, fostering personal growth and driving social and economic change.

The number of students in higher education has grown significantly over the past decade, driven by factors such as increased enrolment, student mobility, diverse offerings, research developments, and advancements in technology. Currently, an estimated 235 million students are enrolled in universities worldwide (Unesco, 2023).

However, despite the rising demand, the global enrollment ratio remains at 40%, with significant disparities across countries and regions (Unesco, 2023). Improving this enrollment ratio is a key challenge for Higher Education Institutions (HEIs), which can be addressed by both increasing student intake and reducing dropout rates.

Dropping has economic and social consequences both for dropouts themselves and for the country as a whole (Rumberger, 2020), and it is a complex phenomenon with multifactorial causes such as personal and individual issues, academic and pedagogical aspects, and university management (Costa et al., 2018). Among the main factors influencing dropout in HEI are: study conditions at university; external conditions; information and admission requirements; prior academic achievement in school; personal characteristics of the student; and sociodemographic background of the student (Kehm et al., 2019). Furthermore, the students and course contexts should be considered in dropout analysis, such as regional aspects and course area (Lobo, 2012).

Understanding the causes of dropout is not an easy task. Different countries provide indicators that have been used by government agencies and researchers to measure the number and rate that students dropout, like the United States of America and Brazil (Rumberger et al., 2017). Yet dropout rates alone may not be sufficient to reveal the extent of the problem. For this cause, this issue has been addressed by several works from different countries over the past few years, e.g., (Demeter et al., 2022; Menolli et al., 2020; Musso et al., 2020; Perez et al., 2018).

Therefore, gaining knowledge about the main factors that contribute to dropout is essential. At the HEI level, it is possible for to management start to act with prevention routines. At the government level, there are more important consequences, such as reducing spending on education and improving social and economic aspects by increasing qualified labor.

Studies addressing Machine Learning (ML) techniques in dropout, in general focus on a specific context (Country, HEI, or course) and generate prediction models just considering ML metrics like precision and recall (Demeter et al., 2022; Jiménez et al., 2023; Musso et al., 2020; Zhang et al., 2022).

Due to this reason, is important not only to create prediction models but also to achieve knowledge about the main factors that contribute to dropout. Considering this, the goal of this work is to propose a dropout prediction process, based on machine learning and feature selections, applicable to any context. We consider the context of the circumstances that form the scenario for the dropout evaluation. For instance, if we want to analyze the dropout in a specific course in a single state, this is the context. Thus, employing the proposed process, we analyzed the most significant features for predicting dropout in Brazilian face-to-face higher education courses

besides examining the variations in feature importance for dropout prediction across different contexts.

2 Background

In this section, we offer a comprehensive introduction to the key concepts employed in our study, along with a review of relevant prior research.

2.1 Feature Selection and Data Classification

This study selected several important algorithms traditionally used in tabular data classification, chosen for their robustness, versatility, and proven performance across diverse datasets and problem domains (Hassan et al., 2018). These algorithms include:

- **Support Vector Machine (SVM):** Identifies the optimal hyperplane to effectively separate data into distinct classes while maximizing the margin between them, making it particularly effective in high-dimensional contexts and binary classifications.
- **Random Forest:** An ensemble learning method that combines multiple decision trees to enhance predictive accuracy and reduce overfitting.
- **Logistic Regression:** A statistical technique that models the relationship between a dependent variable and one or more independent variables using the logistic function, offering clear probability interpretations.
- **K-Nearest Neighbors (KNN):** A straightforward method that classifies an unknown data point based on the majority class among its k-nearest neighbors in the feature space.
- **Naive Bayes:** A probabilistic classification algorithm grounded in Bayes' theorem, known for its speed and efficiency, especially with large datasets.

Regarding feature selection, numerous techniques have been developed in response to the abundance of data featuring hundreds of variables, resulting in high-dimensional datasets. The focus of feature selection is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results (Guyon & Elisseeff, 2003). A feature represents a distinct and quantifiable attribute of the observed process (Chandrashekar & Sahin, 2014). Machine learning algorithms leverage sets of features to facilitate classification tasks.

For this purpose, various techniques have been devised to tackle the issue of diminishing irrelevant and redundant variables, which can impose a significant challenge in complex tasks. One of the main techniques is the filter method, which uses variable ranking techniques as the principle criteria for variable selection by ordering (Chandrashekar & Sahin, 2014). The filter methods can be based on correlation criteria (Battiti, 1994; Guyon & Elisseeff, 2003) or mutual information (Battiti, 1994; Guyon & Elisseeff, 2003; Lazar et al., 2012), which uses the measure of dependency between two variables to rank criteria.

Another technique is the wrapper method, which uses the predictor as a black box and the predictor performance as the objective function to evaluate the variable subset (Chandrashekar & Sahin, 2014). Among the main strategies to implement wrapper techniques are the sequential selection algorithms, which start with an empty set and add one feature for the first step which gives the highest value for the objective function. Another approach to creating wrapper methods is the heuristic search algorithms, such as Genetic Algorithm (GA) (Sastry et al., 2013). Finally, embedded methods (Guyon & Elisseeff, 2003) are used to reduce the computation time taken up for reclassifying different subsets which is done in wrapper methods.

Besides the vast techniques and algorithms of feature selection, the comparison between feature selection algorithms can only be done using a single dataset since each underlying algorithm will behave differently for different data (Chandrashekar & Sahin, 2014). Therefore, selecting the appropriate feature selection algorithm is not an easy task. Regarding this, (Chandrashekar & Sahin, 2014) identified some important factors in selecting a feature selection algorithm: simplicity, stability, number of reduced features, classification accuracy, storage, and computational requirements. Moreover, we have tested some approaches, such as correlation and wrapper methods, and analyzed the feature selection used in other studies in related areas, such as (Melo & Souza, 2023). Hence, taking into account our internal tests, the factors defined by (Chandrashekar & Sahin, 2014) and the analysis of other studies, the Boruta algorithm (Kursa & Rudnicki, 2010) was chosen in this work. Boruta is designed as a wrapper around a Random Forest classification algorithm and it iteratively removes the features that are proved by a statistical test to be less relevant than random probes. Moreover, boruta package is most widely used for feature selection (Anand et al., 2021) and applied on a dataset with a large number of attributes provides greater accuracy, precision and recall values (Bhalaji et al., 2018).

2.2 Related Works

In the literature, several studies focus on analyzing student dropout. These studies range from more specific contexts, such as courses at a particular higher education institution (HEI), to more comprehensive approaches that examine the situation of a course across the country (Menolli et al., 2020). Among these studies, several focus on dropout prediction and some use Artificial Intelligence techniques, such as Machine Learning (ML) and Data Mining (DM).

The use of AI, ML and DM to study phenomena related to school dropout is a constant for developers and researchers. The variety of methods, techniques, tools, research objectives, metrics and other research questions can be identified in Sistematic Literature Mappings/Reviews, such as some works cited below. Nascimento et al. (2024) conducted a Systematic Literature Review and analyzed 65 papers to provide a state-of-the-art overview of the use of methods, techniques, and tools for Artificial Intelligence and Machine Learning in addressing dropout. This work, with a technical objectives, allows them to identify and understand a wide set of programming languages, algorithms, evaluation metrics, inducing factors and databases used by the researched evaluated.

Colpo et. al. (2024) analyzed, in addition to technical factors, the objectives of studies on dropout, as well as the levels of education to which it were applied. As research objectives, the search for patterns or attributes that impacted dropout phenomena stood out; and the construction of predictive models to identify students who could potentially drop out of courses. Regarding to the level of study at which the research was carried out, graduation stood out and the authors

justified these numbers by the reason that dropout from graduation is a major concern for managers and that researchers are mostly conducted at universities. Furthermore, the authors highlighted the existence of deficiencies in the application of predictive models and in making their predictions available to academic managers, which indicates an underutilization of the efforts and potential of most of these studies in educational practice.

Jesus and Gusmão (2024) presented as research questions the characterization and evolution of the state of the art of AI and ML in studies on school dropout. The authors analyzed 71 articles and concluded that studies related to higher education, in-person education, are the most recurrent, confirming what was presented in the previous work. The authors present numbers that provide information regarding an increase in studies in this area from 2015 onwards, with a higher value in 2020. Several algorithms are discussed and those based on Decision Trees were the most used.

In addition to bibliographic studies, related works can be cited, research that applied Machine Learning methods to identify or classify information about evasion. For example: Berka & Marek (2021), for example, analyzed school dropout at the degree program level, and found credit lost in the last semester to be the most important characteristic for predicting dropout. The study conducted by Cannistrà et al. (2022) analyzed dropouts in higher education in a more comprehensive context, considering demographic data, academic history, and information about students.

Another work that undertook a comprehensive analysis of the phenomenon of dropout in higher education, adopting data mining approaches to investigate determining factors and predict student dropout is presented in (Djulovic & Li, 2013). In the work Martins et. al. (2023), the authors developed models to predict academic performance and dropout rates using data from 4433 students, between 2009 and 2017, from a polytechnic university in Portugal.

Matz et. al. (2023) explore the prediction of student retention using sociodemographic data and app engagement metrics at four US universities. In the study conducted by Nagy & Molontay (2023), they employing machine learning classifiers to predict student dropout, using demographic data and academic history of 6,398 students from a Hungarian university.

In work focused on university students in South Korea, Song et. al. (2023) used six machine learning classifiers to predict these students' dropout rate. In the study performed by Vaarma & Li (2024), the prediction of dropout rates in higher education was explored using data from a Finnish university focusing on the virtual learning environment. Teodoro & Kappel (2020) studied the phenomenon of school dropout in the context of Brazilian HEIs, to identify the most determining characteristics for students to dropout and, thus, try to predict the possible dropout of other students.

Considering the related works, it is observed that among the main factors leading students to dropout are: lost credits, academic performance, age at enrollment, years elapsed between high school graduation and university enrollment, extracurricular activities, age, total course workload, among others.

The researches presented in this section sought to build models to identify or test attributes in the construction of prediction algorithms for school dropout. Our research differs from related works in that we propose a process based on feature selection to create and evaluate machine learning models for predicting dropout in the higher education system. The approach not only outlines the essential steps for model development in any context but also emphasizes the identification of

the most critical features.

3 The Feature Reduction Process

The proposed process presented in Figure 1 needs two inputs. The first one is the initial model, which serves as the foundation for further analysis and is subjected to a classification task. The second is a stop criterion, which defines the maximum acceptable loss of accuracy regarding the initial model. Since several models are generated with a reduced number of features, this criterion serves to define which models are acceptable. Once the inputs are specified, the process for establishing models with a reduced number of features is outlined in Figure 1.

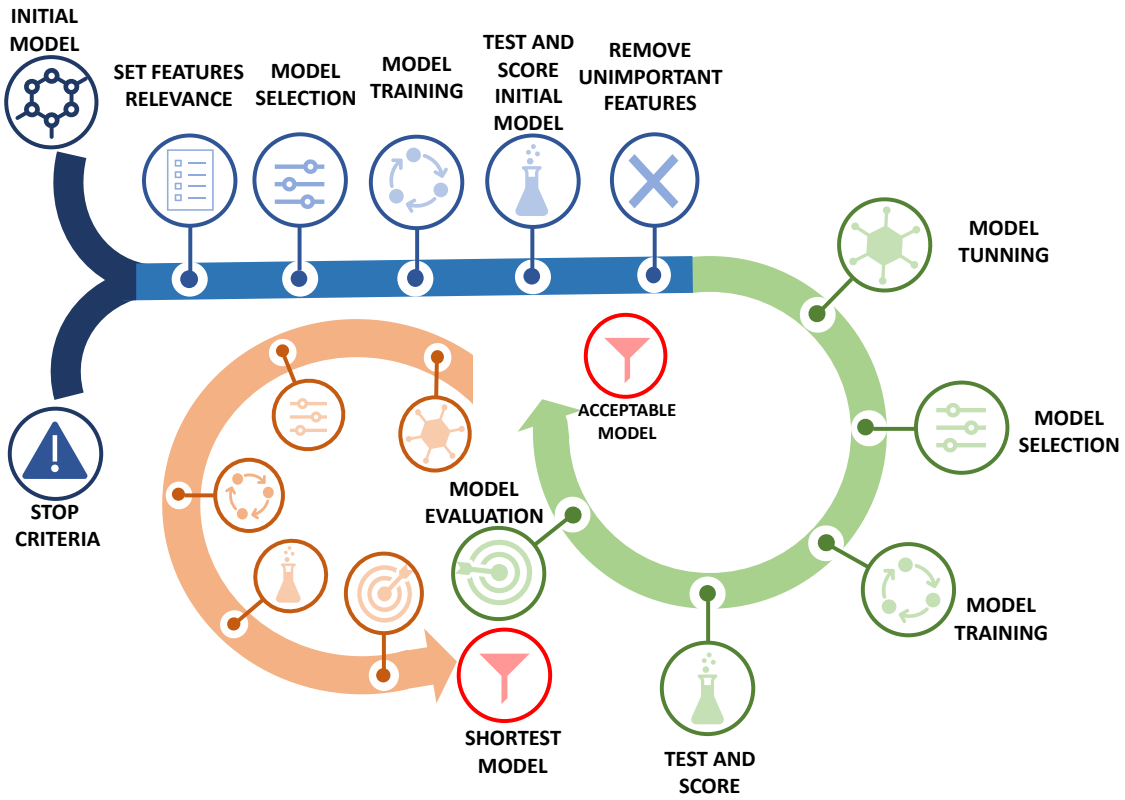

Figure 1: The process to define new models with a reduced number of features.

The initial stage (depicted in blue) involves testing and scoring the initial model. The process commences with the calculation of feature relevance, employing the Boruta feature selection algorithm (Kursa & Rudnicki, 2010). The outcomes of this calculation guide subsequent activities aimed at diminishing the number of features in the model.

Afterward, data from the initial model is loaded, encompassing both target and features. Subsequently, the sampling method is determined. In our approach, we employ cross-validation k-folds technique with 10 folders. This technique is widely used technique in machine learning and statistics for assessing the performance of a predictive model. K-fold cross-validation is often used for hyperparameter tuning, model selection, and assessing the overall performance of a machine learning algorithm. It helps in gaining a better understanding of how a model generalizes

to different subsets of data and can provide more confidence in the model's performance estimates.

Following this, the next activity is the model selection, where algorithms are chosen. In our approach were used five ML algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes. These algorithms are then applied to classify the training data, followed by the classification of the test data, yielding the corresponding test scores. To scores the following metrics were used:

- The Area Under the ROC Curve (AUC): is a crucial metric particularly for binary classification tasks. It assesses the performance of a classification model by measuring its ability to distinguish between positive and negative classes across different threshold settings.
- Accuracy: The most basic metric and represents the ratio of correctly predicted instances to the total instances. It's suitable for balanced datasets but may be misleading for imbalanced ones.
- Precision: The ratio of true positive predictions to the total positive predictions. It helps evaluate the model's ability to make accurate positive predictions.
- Recall: Calculates the ratio of true positive predictions to the total actual positives in the dataset. It assesses how well the model identifies positive instances.
- F1 Score: The score is the harmonic mean of precision and recall. It provides a balanced measure of a model's accuracy.

The last activity in this stage involves the removal of unimportant features, taking into account the Boruta statistics data. Subsequently, a new model is generated and used as input for the iterative cycle (green part).

The green cycle starts executing the model tunning, where 10% of the most unimportant features are removed. For the first iteration, this step is not performed. After that, the algorithms are chosen, and the model is trained and tested, generating scores.

The final stage of the green cycle involves comparing the current model scores with those of the initial model. To decide whether to proceed with additional iterations, the assessment includes the following metrics: AUC, Accuracy, Precision, Recall, and F1 Score. If the difference between the metrics is smaller than the threshold established in the stop criteria, the model is deemed acceptable, and a new iteration is initiated.

Thus, within the green cycle, each iteration produces an acceptable model with a reduced number of features until an unacceptable model is generated, prompting the cessation of further iterations.

The orange cycle receives as input the unacceptable model. In each iteration, the most important feature previously removed is reintroduced in the unacceptable model. The model is then trained, tested, and compared with the initial model. If the model meets the stopping criteria, it becomes the final model; otherwise, the cycle repeats. Thus, the orange cycle is repeated until get a satisfactory model, which is the shortest model.

3.1 Experimental Setup

In this section, we provide a concrete example of how the process is applied to enhance comprehension of our work and its application in one of the five contexts employed in this study, specifically focusing on all face-to-face courses in Brazil. The same methodology described in this section was applied to all other contexts utilized as experiments in this study.

We used the Census of Higher Education (Rumberger et al., 2017) of the year 2019. This dataset consists of distinct CSV files, and we specifically utilized the student, course, and institution files for our study.

In the original dataset, the student file contains 105 features, course 112, and institution 48. This dataset covers various types of higher education programs in Brazil, including undergraduate and technology degree courses. In the study, we used just undergraduate courses.

However, for our study, we exclusively considered undergraduate courses, concentrating our analysis on this specific subset to align with our research focus.

In the data extraction phase, we began by importing files into a database and carrying out essential ETL (Extract, Transform, Load) and data cleaning tasks. In the course of these activities, we accomplished the following:

1. Data Integration: We combined data from various sources into a single table, consolidating our dataset for streamlined analysis.
2. External Data Augmentation: To enhance our dataset, we incorporated information from external databases, such as the Brazilian Institute of Geography and Statistics (IBGE) for geo-referential data, and the International Standardized Classification of Education Adapted for Undergraduate Courses and Specific Training Sequences (Cine Brasil) for comprehensive insights into course areas.

In the data selection phase, we removed duplicate features and eliminated bad features, ultimately resulting in a refined dataset comprising 61 essential features. Following that, we defined the context for our initial study, which involved examining undergraduate face-to-face courses nationwide. Within this context, our dataset was streamlined to consist of 50 fields.

Furthermore, we crafted a balanced dataset, totaling 400 thousand records, equally distributed between students who dropped out and those who did not.

In the set process input, the first step was to define the initial model, where dropout was defined as the target and the other 49 fields as features.

After that, we defined the stop criteria. A restrictive rule was established to assess the acceptability of a model. This criterion required that no metric should exhibit a deterioration of more than 2% compared to the metrics obtained from the initial model. Therefore, we have established a criterion where the maximum allowable difference between the current model and the initial model for all metrics is set at 2%. This threshold guides our assessment of model performance.

Following this, we progressed with the subsequent steps, as depicted by the light blue section. The first process activity involves determining feature relevance. To achieve this, we em-

ployed the Boruta algorithm on the model. The Boruta algorithm executed 99 iterations, identifying 46 attributes confirmed as important, 3 attributes confirmed as unimportant. A summary of the feature statistics from this subprocess can be found in Table 1.

The second task within this phase involved model selection, where we employed five ML algorithms. Subsequently, we trained and tested the initial model, yielding the results presented in Table 2.

Subsequently, the iteration defined in the green part of the process of Figure 1 was performed. This cycle creates new models with a reduced number of features, and its scores are then compared with those of the initial model. If the difference between the metrics falls below the threshold defined in the stopping criteria, the cycle persists.

In this cycle, the first iteration removes all features that have been flagged as rejected by the Boruta algorithm. Following this, in each subsequent iteration, we proceed to remove 10% of the less important features. Given that there are a total of 49 features, we opt to remove 5 features during each iteration.

Once the green process produces an unsatisfactory model, this model is used as input to the orange cycle (Figure 1). In this cycle, the most important features should be put back at the model one by one, until reaching an acceptable model, that is the shortest model generated.

Table 3 provides a comprehensive summary of the outcomes achieved in each iteration. To enhance clarity, only the best model results from each iteration are presented, with Random Forest consistently yielding the best performance in all instances. The initial iteration is generated by the light blue part of the process depicted in Figure 1. Subsequent iterations, from the first (boruta) to the eighth, are provided by the green cycle. The eighth iteration yielded an unsatisfactory model, which is highlighted in Table 3 alongside the corresponding metric. This iteration serves as input for the orange cycle, which produces the final iteration, denoted as the shortest.

It is important to highlight that when we analyzing the results, this occurred in two aspects. Firstly, we evaluated the model in its entirety, considering the metrics over all classes. This comparison aimed to assess the model's overall quality, i.e., how well it performed in predicting both dropout and non-dropout cases. However, as we are particularly interested in dropout, we also compare recall, F1, and precision for dropout class.

4 Data Analysis and Results

In this section, we present the results of experiments conducted using the proposed approach. The experiments were designed with two primary objectives: (1) To assess the approach's effectiveness across various contexts; (2) To gain insight into the key features influencing dropout prediction in different scenarios. Considering this, we have employed the approach in five distinct contexts. The initial context pertained to undergraduate face-to-face courses in Brazil. In the second context, all face-to-face computing courses in Brazil were taken into consideration, encompassing over 15 different courses in which dropout rates were analyzed. The remaining three contexts are centered around nursing courses. The third context involves face-to-face nursing degrees across Brazil as a whole, providing a comprehensive overview of dropout trends within the nursing discipline

Table 1: Attributes statistics of the set features relevance.

Feature	mean	median	min	max	normHits	decision
tp_modality_teaching	0.00	0.00	0.00	0.00	0	Rej.
tp_academic_level	0.00	0.00	0.00	0.00	0	Rej.
co_country_origin	2.16	2.13	-0.63	5.21	0.34	Rej.
in_academic_mobility	5.45	5.48	3.68	6.76	1	Confir.
tp_attribute_admission	7.97	7.93	5.93	10.08	1	Confir.
reservation_type	11.26	11.29	10.03	12.65	1	Confir.
in_reserve_vacancies	11.87	11.84	10.73	13.22	1	Confir.
tp_nationality	17.41	17.42	14.86	20.17	1	Confir.
in_help_disabled	19.50	19.46	16.48	22.93	1	Confir.
in_free	24.55	24.55	23.25	26.02	1	Confir.
in_internet_service	23.13	23.16	19.64	27.04	1	Confir.
in_scholarship	26.62	26.57	24.29	28.61	1	Confir.
tp_academic_grade	23.48	23.21	17.39	28.83	1	Confir.
in_has_laboratory	30.75	30.48	24.69	35.83	1	Confir.
tp_period	34.11	33.94	32.03	36.42	1	Confir.
in_deficiency	29.78	29.91	25.30	36.75	1	Confir.
tp_high_school_completion	32.16	31.94	28.71	37.00	1	Confir.
in_social_support	36.56	36.65	33.81	39.59	1	Confir.
tp_academic_organization	37.45	37.40	34.44	41.05	1	Confir.
in_signs_another_base	36.61	36.34	32.21	41.31	1	Confir.
in_discipline_libras	35.73	35.64	31.69	41.81	1	Confir.
tp_race_color	38.94	39.08	34.59	43.81	1	Confir.
in_online_catalog	40.08	40.15	34.90	44.24	1	Confir.
in_translator_libras	42.71	42.60	39.77	45.86	1	Confir.
co_large_area	40.80	40.81	35.67	46.09	1	Confir.
offers_semi-face-to-face_disc	43.21	43.01	40.86	46.52	1	Confir.
tp_administrative_category	43.61	43.69	41.21	46.62	1	Confir.
in_capital	42.71	42.49	35.44	48.39	1	Confir.
co_cine_general_area	42.72	42.53	35.55	49.58	1	Confir.
in_access_portal_capes	47.24	47.07	43.94	50.33	1	Confir.
admission_type	44.69	44.42	38.70	51.46	1	Confir.
area	46.99	46.74	40.63	52.46	1	Confir.
time_course_range	47.56	47.27	41.53	53.38	1	Confir.
no_course	46.16	45.63	40.06	53.51	1	Confir.
co_region	44.68	44.80	36.43	54.56	1	Confir.
in_total_entry	50.97	50.93	47.08	55.23	1	Confir.
curse_workload	50.58	50.65	46.13	55.30	1	Confir.
in_institutional_repository	54.04	53.86	48.05	59.00	1	Confir.
co_state	51.28	50.88	42.05	61.26	1	Confir.
age	58.20	58.48	54.30	62.37	1	Confir.
in_integrated_search	58.76	58.91	53.48	62.72	1	Confir.
co_cine_label	57.60	57.58	50.48	63.91	1	Confir.
in_activity_extracurricular	60.98	60.86	57.24	64.72	1	Confir.
in_admission_process	74.81	74.72	70.04	80.38	1	Confir.
co_city	73.00	73.12	64.85	82.32	1	Confir.
in_concluding	79.59	79.70	75.36	84.22	1	Confir.
co_hei	78.87	78.96	70.23	85.84	1	Confir.
course_entry_time	89.46	89.73	81.42	98.17	1	Confir.
perc_completed_range	101.67	101.49	94.44	106.44	1	Confir.

Table 2: Test Scores for Initial Model.

	Total Average over classes					Dropout Class		
Model	AUC	CA	F1	Prec	Recall	F1	Prec	Recall
Random Forest	0.91	0.84	0.84	0.84	0.84	0.87	0.84	0.91
Logistic Regression	0.52	0.61	0.46	0.37	0.61	0.76	0.61	1.00
SVM	0.50	0.61	0.46	0.68	0.61	0.76	0.61	1.00
Naive Bayes	0.73	0.68	0.67	0.67	0.68	0.75	0.72	0.78
kNN	0.81	0.75	0.75	0.75	0.75	0.81	0.76	0.86

Table 3: Test Scores for all iterations of the all courses context .

	Total Average over classes					Dropout Class		
Iter.	AUC	CA	F1	Prec	Recall	F1	Prec	Recall
Inital	0.908	0.839	0.837	0.839	0.839	0.873	0.840	0.909
1(Boruta)	0.908	0.839	0.837	0.839	0.839	0.873	0.839	0.910
2	0.907	0.839	0.837	0.839	0.839	0.873	0.841	0.906
3	0.903	0.835	0.833	0.835	0.835	0.869	0.841	0.900
4	0.898	0.832	0.830	0.831	0.832	0.867	0.839	0.896
5	0.899	0.834	0.832	0.833	0.834	0.868	0.841	0.898
7	0.897	0.833	0.831	0.832	0.833	0.867	0.841	0.894
8	0.891	0.826	0.824	0.825	0.826	0.861	0.837	0.887
Shortest	0.897	0.833	0.831	0.832	0.833	0.867	0.842	0.893

nationwide. The fourth context is a specialized extension, narrowing its focus to private Higher Education Institutions (HEI) within the state of São Paulo. Lastly, the fifth context delves into nursing courses within public HEI in São Paulo. Table 4 summarizes the main information about each experiment and the dataset used.

As mentioned in Section 3, if any of the metrics (AUC, Accuracy, Precision, Recall, and F1 Score) falls below the established threshold, the feature removal process stops. In this experiment, we set the threshold at 2%. In the iteration where any metric was 2% worse than defined in the initial model, the feature removal would stop, and that model would be considered unsatisfactory.

To facilitate our analysis, we provide Table 5, which outlines the metrics of the initial model and the percentage differences in metrics for each iteration relative to the metrics of the initial model for context A, representing the analysis on all face-to-face courses in Brazil. The table also presents the number of features used in each iteration. In the eighth iteration, the metrics that deviate by 2% or more from the initial model are highlighted. The final iteration provides the performance differences between the shortest model and the initial model for each metric. In this context, the model's accuracy with 13 features was only 0.75% lower than that of the initial model. The same method was used for the other contexts, however for the reason of space limitation we just present Figure 2, which provides an overview of the results across all contexts, showcasing the accuracy of the initial model, along with the model generated with features confirmed as important by Boruta, best and shortest models identified for each specific context, including the number of features utilized in each model.

For the second context (B), computer science face-to-face courses in Brazil, we have at-

Table 4: Information about the contexts and dataset used in the experiments.

Experiment Context Information					DataSet Information		
Context	**Coverage**	**Nº Courses**	**Nº HEI**	**HEI Type**	**DataSet Size**	**Dropout Students**	**Non Dropout Students**
A	Country (Brazil)	251	2381	public and private	400.000	200.000	200.000
B	Country (Brazil)	15	661	public and private	72.000	36.000	36.000
C	Country (Brazil)	1	904	public and private	72.000	36.000	36.000
C1	One State (São Paulo)	1	144	private	40.000	20.000	20.000
C2	One State (São Paulo)	1	13	public	700	350	350

Table 5: The difference (%) in performance between the best model in each iteration and the initial model for all face-to-face courses in Brazil.

	Total Average over classes					**Dropout Class**			
Iter.	AUC	CA	F1	Prec	Recall	F1	Prec	Recall	**N.Feat.**
Initial	0.908	0.839	0.837	0.839	0.839	0.873	0.840	0.909	49
1 (boruta)	0.00%	-0.01%	0.00%	-0.02%	-0.01%	-0.02%	0.08%	-0.14%	46
2	0.13%	0.06%	0.05%	0.09%	0.06%	0.08%	-0.12%	0.30%	51
3	0.05%	0.01%	-0.01%	0.03%	0.01%	0.04%	-0.17%	0.27%	36
4	0.53%	0.46%	0.42%	0.53%	0.46%	0.43%	-0.07%	0.97%	31
5	1.03%	0.83%	0.79%	0.92%	0.83%	0.73%	0.11%	1.40%	26
6	0.91%	0.58%	0.53%	0.65%	0.58%	0.54%	-0.07%	1.18%	21
7	1.14%	0.74%	0.67%	0.83%	0.74%	0.70%	-0.15%	1.59%	16
8	1.89%	1.58%	1.51%	1.69%	1.58%	1.36%	0.41%	2.36%	11
Shortest	1.19%	0.76%	0.68%	0.86%	0.76%	0.73%	-0.24%	1.77%	13

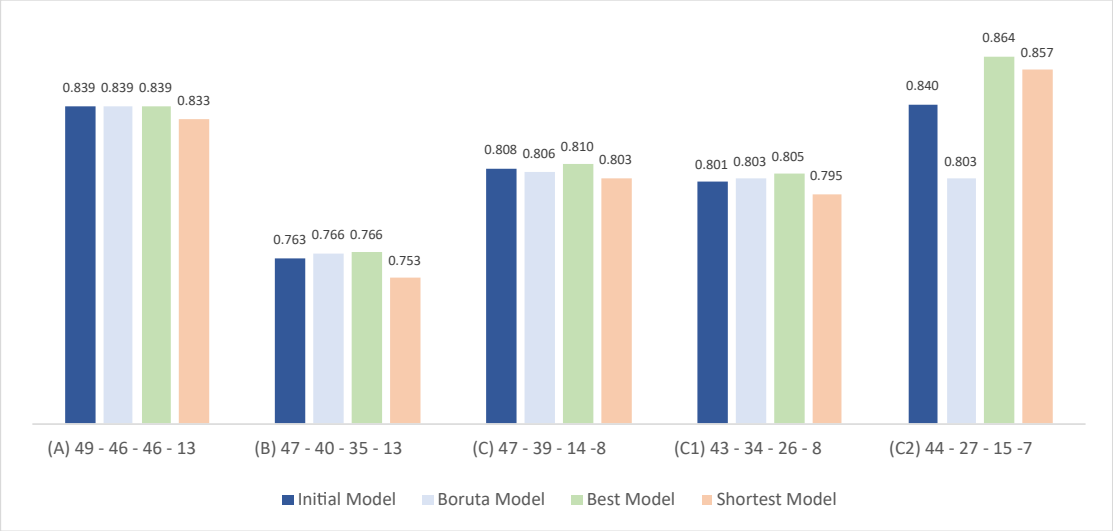

Figure 2: Accuracy and the number of features for each context in the experiment, considering the Initial, Boruta, Best, and Shortest models..

tained a model that exhibits a 1.34% decrease in performance compared to the initial model while utilizing only 28.2% of the initial features.

In the third context (C), the shortest model exhibited slight decrements until 0.5% in six metrics, improvements in one metric, and a single metric that deteriorated by more than 1%. However, it's crucial to highlight that this model operated with just 17% of the initial features, representing a reduction of nearly 85%.

In the fourth context (C1), the shortest model exhibited slight decrements of less than 1.0% in six metrics, improvements in one metric, and a single metric that deteriorated by more than 1%. However, it's crucial to highlight that this model operated with just 18.6% of the initial features, representing a reduction of more than 80%.

In the last context (C2), in contrast to the others, the shortest model consistently achieved superior results across all metrics compared to the initial model. In this specific context, a model was created with an accuracy that surpassed the initial model by nearly 3%, despite using only 34% of the original features.

4.1 Comparison with Other Feature Selection Algorithms

Another result presented involves the comparing the results obtained from the proposed process with those from other feature reduction algorithms. For this, we used five commonly adopted filter-based feature selection algorithms (Theng & Bhoyar, 2024; Vora & Yang, 2017):

- Chi-Square Score (Liu & Setiono, 1995) is a statistical measure used to test correlation between two variables. In feature selection, it assesses whether a class label is independent of a feature in a labeled dataset.
- Information Gain (Cover, 1999), based on information entropy, calculates how much information a feature adds to differentiate between classes. Higher values indicate better distinguishing features.

- Gini Index (Liu & Setiono, 1995) is a common measure in tree-based classifiers (e.g., Random Forests) that selects features to best separate samples across classes.
- ReliefF (Kononenko, 1994) extends the Relief (Kira & Rendell, 1992) algorithm for feature selection in two-class datasets, aiming to identify features that differentiate instances across classes.
- Fast Correlation-Based Filter (FCBF) (L. Yu & Liu, 2003) is a supervised method evaluating both feature-class and inter-feature correlations. Using entropy measures, it first selects features highly relevant to the class label, then applies heuristics to remove redundancies.

To perfume the comparison, the best model obtained in each Context was compared with the five algorithms. The following steps were taken:

1. Create a reduced model using the same number of features as the best model generated by the proposed process.
2. Apply the *Random Forest* algorithm using the same training and testing configurations used to generate the results of the proposed process.
3. Compare the results obtained with those from the proposed process.

Figure 3 presents a comparison of AUC, accuracy, and F1 metrics between the proposed approach and the results from other algorithms. For Contexts A and B, all methods showed very similar results. However, for Contexts C, C1, and C2, our approach achieved better results.

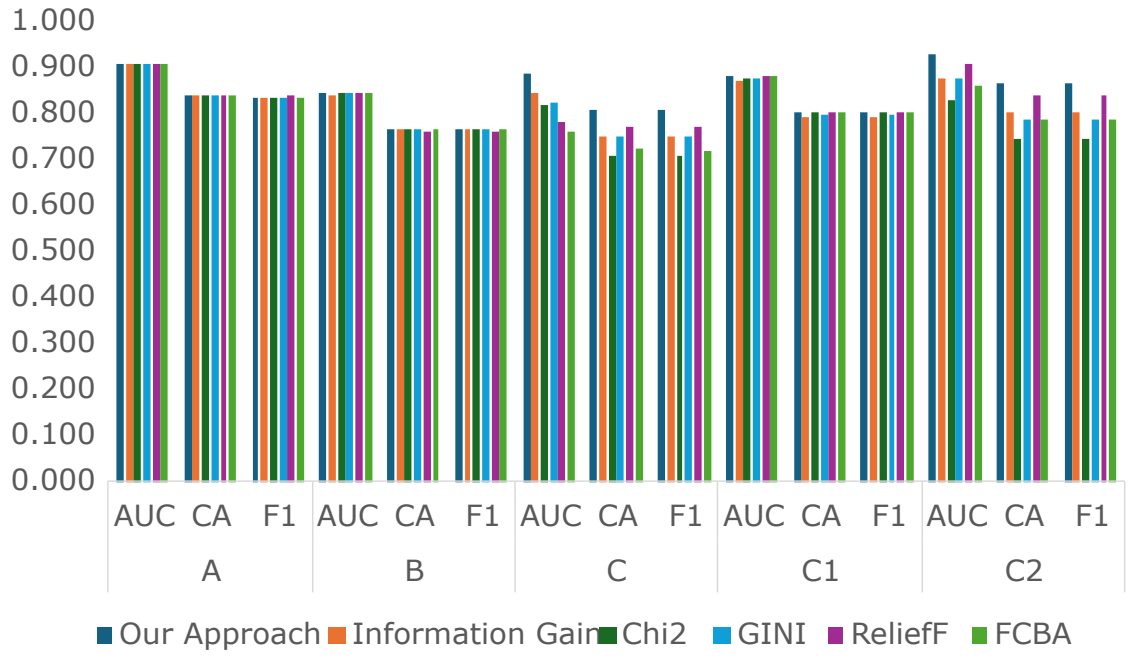

Figure 3: Metrics produced by the proposed approach across all contexts, compared to other feature selection algorithms..

5 Discussion

In this section, we focus the discussion on two main aspects, model effectiveness and the most important features to predict dropout.

5.1 Discussion on Proposed Model

The results depicted in Figure 2 reveal that across all contexts, the optimal model utilized a reduced number of features compared to the initial model. Notably, in four out of the five contexts, the top-performing model employed fewer features than Boruta, suggesting promising outcomes for feature reduction.

Furthermore, as depicted in Figure 4, three metrics are visually presented, showcasing their behavior regarding the number of features across two contexts. In Context C2, the results exhibit an improvement as the number of features decreases, reaching its peak at 15 features. However, beyond this threshold, the accuracy starts to decline, albeit experiencing a slight improvement with 7 features. Conversely, in Context C, the accuracy of models remains relatively stable until it reaches 14 features, where it achieves its optimal performance.

Regarding the results, we emphasize the significance of the process in creating high-accuracy models through feature reduction. This is particularly crucial as relying solely on feature selection algorithms may not always guarantee the generation of the best models. Our process yielded superior results compared to using feature selection algorithms alone.

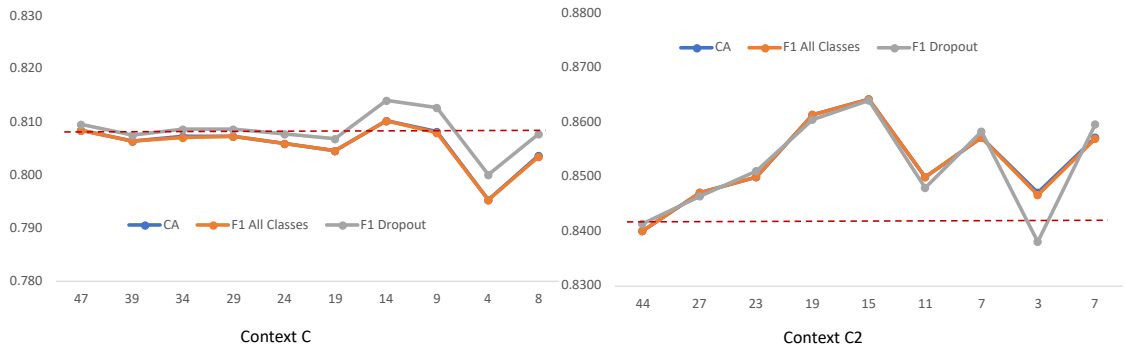

Figure 4: Results of metrics for C and C2 contexts regarding the number of features.

The Figures show the results of CA, F1 over all classes and F1 over dropout class considering the interactions for each one of the four contexts. The red dashed line indicates the result of the initial model and axe x the number of features.

5.2 Important Features to Predicting Dropout

In this study, we implemented a process aimed at minimizing the number of features while maintaining model accuracy comparable to that of the original model containing all features. This process was applied across five distinct contexts, and in none of them, the resulting models exhibited an accuracy decrease of more than 1.34% compared to the initial model. This methodology

empowered us to develop predictive models utilizing only the most pertinent features essential for dropout prediction.

Regarding the most important features to predict dropout, first of all, as shown in Figure 5, there is a difference in the number of features in the shortest and better models among the contexts. Thus, it is a clear indication that there are differences in the importance of the features in distinct contexts. Furthermore, in the more heterogeneous contexts, in the best and shortest models, there are a greater number of features, indicating that the broader the context more difficult it is to predict the factors that lead to dropout.

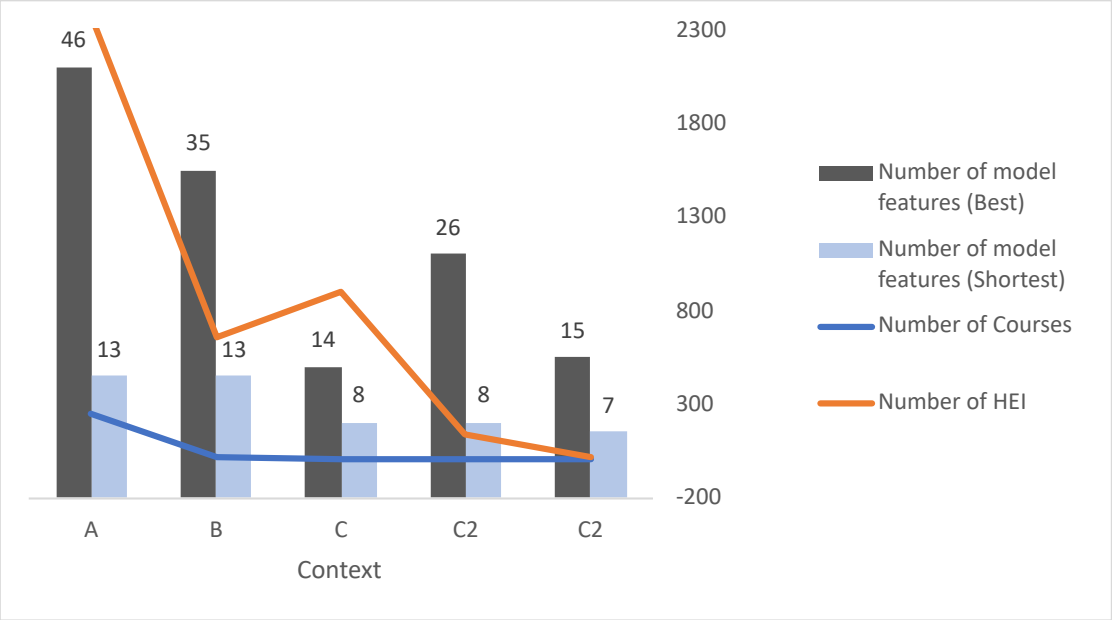

Figure 5: Results of metrics for each context regarding the number of features.

Given the variable nature of the number of features in the shortest models across different contexts, a direct comparison becomes unfeasible. Hence, Figure 6 shows the correlation matrix, to depict the relation of the same important features between contexts. Thus, different contexts present distinct weights in the features for dropout prediction. Even similar contexts present significant differences in the most important characteristics. The importance of the feature is related to several factors.

Contexts	A	B	C	C1	C2
A	1.00	0.69	0.87	0.75	0.71
B	0.69	1.00	1.00	0.88	0.86
C	0.53	0.62	1.00	0.75	0.86
C1	0.46	0.54	0.75	1.00	0.57
C2	0.38	0.46	0.38	0.50	1.00

Figure 6: Correlation matrix of the most important features in different contexts.

Regarding the most important features, Table 6 provides an overview of the most important

features considering the five contexts explored. According to the study's findings, the correlation between the duration of a student's enrollment and the percentage of the course completed emerges as the most important feature for predicting dropout. Moreover, extracurricular activities are essential for predicting dropout, although to a lesser extent in contexts where exclusively private HEI were involved. The student's age also appears as an important feature as well as how long the course exists.

Features related to HEI are also highly significant. This includes the HEI code and integrated_search, a feature indicating the infrastructure level of the HEI. Moreover, in country contexts, features related to locality are also crucial.

Finally, it is observed that in more focused contexts, features linked to students play a more significant role and contribute to the creation of more accurate models. These student features can be divided into two categories: those related to their progress in the course (course_entry_time, perc_completed_range, in_activity_extracurricular, in_concluding, admission_type, study_period) and personal characteristics (age, race_color). The results indicate that race_color is important in private context while study_period in public.

A final discussion concerns the findings from other dropout studies and compare with this study. Table 7 summarizes the data from several studies focusing on dropout in a face-to-face context, highlighting the accuracy and key features of each study.

In our study, the metrics exhibited the following values: AUC ranging from 0.85 to 0.93; CA from 0.77 to 0.86; F1 from 0.77 to 0.86; Precision from 0.77 to 0.86; and Recall from 0.77 to 0.86. When comparing the results of our study with others found in the literature, the metrics obtained in our study are either equivalent to or superior to those of other studies. However, it is important to acknowledge that many studies have contexts that differ significantly from ours, including variations in the data used. Several studies incorporate financial and academic performance data, aspects which were not addressed in our study. The only work utilizing the same database as our study is presented by (Teodoro & Kappel, 2020), and in four out of the five contexts we analyzed, we achieved superior accuracy.

For this particular study (Teodoro & Kappel, 2020), we compared all the metrics from the work with those obtained in our study. As shown in Table 8, the only metric with a comparable result is recall. All other metrics in our study demonstrated superior performance.

Regarding the most significant features, as indicated by the studies in Table 7, factors such as academic performance, and in the case of private HEI, financial data, can contribute to enhancing dropout prediction, alongside the features considered important in our work.

It is important to highlight that this work presents two key differences compared to the studies found in the literature. Firstly, it introduces a process that enables the reduction of the number of features to the minimum necessary while still producing models with accuracy equal to or greater than other works. Secondly, it stands as the only study to analyze dropout in various contexts, indicating that reasons that lead to dropout vary according to context.

Table 6: Most important features and the contexts in which they appeared in the shortest model.

Feature	Description	Contexts
course_entry_time	Total time in years that the student enrolled in the course	A;B;C;C1;C2
perc_completed_range	Describes the student's total percentage completed of the course in ranges.	A;B;C;C1;C2
in_activity_extracurricular	Informs whether the student participates in any type of extracurricular activity	A;B;C;C2
in_admission_process	Determines whether the student gained entry to the course via a selection process or through alternative means.	A;B;C;C1
in_concluding	Informs whether the student is a graduate	A;B;C1;C2
co_hei	Unique HIE identification code	A;B;C;C1
age	student age range	A;B;C1;C2
in_integrated_search	Informs whether the HEI libraries offer services over the internet	A;B;C
co_state	IBGE code of the federation unit where the in-person course is offered	A;B;C
time_course_range	Time which the course works in ranges	B;C;C1
new_student	Informs whether the student is new student, regardless of the form of entry used.	B;C2
curse_workload	Total course workload categorized into ranges	A
in_institutional_repositor	Determine whether the HEI possesses an online database that systematically collects and organizes its scientific production	A
co_cine_label	Course identification code, as adapted from the International Standard Classification of Cine/Unesco Education	A
co_city	IBGE code of the municipality where the in-person course is offered	A
no_course	Course Name	B
co_region	IBGE code for the region of the federation where the in-person course is offered	B
tp_period	Defines the course period to which the student is enrolled	C2
tp_race_color	Defines the student's color/race	C1

5.3 Threats to Validity

In this section, we discuss the main threats to the validity and limitations of this study. For this, it was used the classification of threats presented in Petersen & Gence (2013), that follows the definition of Maxwell (1992).

Theoretical Validity - What are confounding factors (uncontrollability)? **Do we capture what we intend to capture?** Regarding theoretical validity, two aspects were important. Firstly, whether it was feasible to construct models with a reduced number of features, enabling the identification of the most crucial features while maintaining accuracy comparable to or exceeding that of the original model. Secondly, whether there existed variations in the most important features according to the explored context. We believe that by exploring five contexts, we were able to

Table 7: Summary of previous research that analyzed the most important features of dropout predictions in face-to-face higher education.

Author	Amount of data	Performance	Feature importance
(Berka & Marek, 2021)	3339	CA around 80%	the percentage of lost credit vouchers in the last semester
(Cannistrà et al., 2018)	31071	AUC de 0.87 to 0.96	accumulated credits in the first year
(Delen, 2010)	16066	prec. de 75% to 87%	earned hours divided by registered hours, student loan at spring, fall GPA
(Djulovic & Li, 2013)	7800	prec. 66% to 74%, recall 24% to 52%	academic performance
(Kiss et al., 2019)	10196	prec. 67% to 86% recall 74% to 81%	credits index, credits earned (accumulated credits), age at the enrollment
(Martins et al., 2023)	4433	f1-scores 58% to 66%	accumulated credits, but varies with time at 3 points within the first semester
(Nagy & Molontay, 2023)	6398	AUC 0.774	high school GPA, math score, years elapsed between high school graduation and university enrollment
(Song et al., 2023)	36000	prec. 72% to 83%	number of scholarships, tuition fee, access year
(R. Yu et al., 2021)	93457	prec. 84%, recall 54%	features gender, first-generation college student, underrepresented minority and high financial need are not important
(Teodoro & Kappel, 2020)	376746	CA around 80%	extracurricular activity, age, total course workload

capture our intended and achieve theoretical validity.

Interpretive Validity (Objective Researcher) **Are the conclusions/inferences drawn reasonable given the data representing an objective/ subjective truth?** It is evident that the interpretation relies on the perspective of the researcher. Nevertheless, we presented the data in various formats, enabling readers to comprehend the process behind our conclusions and inferences. Furthermore, to avoid introducing biased conclusions or inferences, we carried out a series of diverse experiments, employing different contexts at varying levels. Another concern related to interpretive validity is the quality of data. To mitigate this threat, we implemented several measures, including data cleaning, manual feature elimination, and, whenever feasible, the removal of outliers. Additionally, we made an effort to employ large datasets to diminish the influence of flawed data on our analysis.

Repeatability (Reproducibility/Dependability) **Data and analysis methods/ instruments should be defined and enable repeatability**. To begin with, the experiment is highly replicable, given that we utilized publicly available data. Moreover, we detailed our approach, elucidating each step and substep in the process, and we transparently disclosed the algorithms and met-

Table 8: Comparison of the metrics obtained in our study with the study presented by Teodoro and Kappel, 2020.

	Metrics				
Work	AUC	CA	F1(Dropout)	Prec(Dropout)	Recall(Dropout)
Our Study	0.90	0.83	0.87	0.84	0.91
Teodoro and Kappel, 2020	≈0.88	≈0.80	≈0.78	≈0.81	≈0.91

rics employed. Considering this, we understand that there is no great dependability, and other researchers should be able to reproduce it.

6 Final Considerations and Future Research

The main goal of this work was to propose a dropout prediction process employing machine learning techniques and feature selection, which prioritize the most relevant features. We have applied the process in five different contexts, using the national database of higher education courses in Brazil, we constructed multiple datasets. These datasets vary in size, ranging from the largest to the smallest compared to related studies.

The findings suggest that in more restricted contexts, models with fewer features tend to yield more accurate predictions of evasion. Typically, these restricted models are based on student data. Conversely, in heterogeneous contexts, models utilize more course and HEI data. The importance of the feature is related to several factors, such as coverage, area, course, type of institution, and teaching modality among others.

Considering the presented findings, this work contributes in several ways. Firstly, it introduces a novel process centered around feature selection. In the conducted experiments, this process yielded reduced models exhibiting higher accuracy than both the original models and those generated only by removing unimportant features identified by the feature selection algorithm.

In contrast to previous related studies, this research delves into dropout across various scenarios, which range from broader and heterogeneous contexts to restricted and homogeneous contexts. This enables us to offer evidence highlighting the contextual dependency of dropout and the necessity of specific features in each context for dropout prediction. Nonetheless, despite this context dependency, our study indicates that for face-to-face courses in Brazil, there is a core set of features critical for predicting dropout rates across all contexts.

Despite employing a Brazilian database to generate the five contexts for this study, we believe the approach can be applied to any dataset containing features and a dropout target field. Therefore, we consider this to be a versatile methodology that facilitates the identification of key features within dropout contexts.

Thus, the contributions of the presented research can be summarized as follows:

- To present an ML approach that helps identify the most important features in dropout prediction in different contexts.
- To explore how the contexts are important in the dropout predictions.

- To present strong evidence that the better defined the analysis contexts, the better the models for predicting dropout.

While creating predictive models for dropout is undeniably important, it is equally crucial to grasp the primary factors associated with it to make informed decisions aimed at preventing dropout. In this context, our study offers a valuable approach that not only aids in predicting dropout but also offers decision-makers a concise set of critical features that influence it. This dual benefit of predictive power and streamlined feature selection enhances the practical utility of our research.

As the main future works we intend to explore other contexts to better to gain a more comprehensive understanding of dropout prediction. Additionally, as future work it is intended to incorporate academic performance data into the generation of models, as considering related work, it is believed that these features possess the potential to enhance the effectiveness of the models.

References

Anand, N., Sehgal, R., Anand, S., & Kaushik, A. (2021). Feature selection on educational data using Boruta algorithm [GS Search]. *International Journal of Computational Intelligence Studies*, *10*(1), 27–35. https://doi.org/10.1504/IJCISTUDIES.2021.113826

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning [GS Search]. *IEEE Transactions on Neural Networks*, *5*(4), 537–550. https://doi.org/10.1109/72.298224

Berka, A., & Marek, M. (2021). Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil [GS Search]. *Revista Brasileira de Informática na Educação*, *28*, 838–863. https://doi.org/10.5753/rbie.2020.28.0.838

Bhalaji, N., Kumar, K. S., & Selvaraj, C. (2018). Empirical study of feature selection methods over classification algorithms [GS Search]. *International Journal of Intelligent Systems Technologies and Applications*, *17*(1-2), 98–108. https://doi.org/10.1504/IJISTA.2018.091590

Cannistrà, T. C., Silva, J. C., & Cortes, O. A. C. (2018). Técnicas de mineração de dados: Um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão [GS Search]. *Revista Brasileira de Computação Aplicada*, *10*(3), 11–20. https://doi.org/10.5335/rbca.v10i3.8427

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods [GS Search]. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Colpo, M. P., Primo, T. T., Aguiar, M. S., & Cechinel, C. (2024). Mineração de dados educacionais na predição da evasão estudantil: Tendências, oportunidades e desafios. *Revista Brasileira de Informática na Educação*, *32*, 220–256. https://doi.org/10.5753/rbie.2024.3559

Costa, F. J., Bispo, M. S., & Pereira, R. C. F. (2018). Dropout and retention of undergraduate students in management: A study at a Brazilian Federal University [GS Search]. *RAUSP Management Journal*, *53*, 74–85. https://doi.org/10.1016/j.rauspm.2017.12.007

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management [GS Search]. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Demeter, E., Dorodchi, M., Al-Hossami, E., Benedict, A., Slattery Walker, L., & Smail, J. (2022). Predicting first-time-in-college students' degree completion outcomes [GS Search]. *Higher Education*, *84*, 589–609. https://doi.org/10.1007/s10734-021-00790-9

Djulovic, A., & Li, D. (2013). Towards freshman retention prediction: A comparative study [GS Search]. *International Journal of Information and Education Technology*, *3*(5), 494–500. https://www.ijiet.org/papers/324-K045.pdf

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection [GS Search]. *Journal of machine learning research*, *3*(Mar), 1157–1182. https://www.jmlr.org/papers/v3/guyon03a.html

Hassan, C. A. U., Khan, M. S., & Shah, M. A. (2018). Comparison of machine learning algorithms in data classification [GS Search]. *2018 24th International Conference on Automation and Computing (ICAC)*, 1–6. https://doi.org/10.23919/IConAC.2018.8748995

Jesus, J. A., & Gusmão, R. P. (2024). Investigação da evasão estudantil por meio da mineração de dados e aprendizagem de máquina: Um mapeamento sistemático [GS Search]. *Revista Brasileira de Informática na Educação*, *32*. https://doi.org/10.5753/rbie.2024.3466

Jiménez, O., Jesús, A., & Wong, L. (2023). Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine [GS Search]. *2023 33rd Conference of Open Innovations Association (FRUCT)*, 116–124. https://doi.org/10.23919/FRUCT58615.2023.10143068

Kehm, B. M., Larsen, M. R., & Sommersel, H. B. (2019). Student dropout from universities in Europe: A review of empirical literature [GS Search]. *Hungarian Educational Research Journal*, *9*(2), 147–164. https://doi.org/10.1556/063.9.2019.1.18

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection [GS Search]. In *Machine learning proceedings 1992* (pp. 249–256). Elsevier. https://doi.org/10.1016/B978-1-55860-247-2.50037-1

Kiss, B., Nagy, M., Molontay, R., & Csabay, B. (2019). Predicting dropout using high school and first-semester academic achievement measures [GS Search]. *2019 17th international conference on emerging eLearning technologies and applications (ICETA)*, 383–389. https://doi.org/10.1109/ICETA48886.2019.9040158

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief [GS Search]. *European conference on machine learning*, 171–182. https://doi.org/10.1007/3-540-57868-4_57

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package [GS Search]. *Journal of statistical software*, *36*, 1–13. https://doi.org/10.18637/jss.v036.i11

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis [GS Search]. *IEEE/ACM transactions on computational biology and bioinformatics*, *9*(4), 1106–1119. https://doi.org/10.1109/TCBB.2012.33

Liu, H., & Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes [GS Search]. *Proceedings of 7th IEEE international conference on tools with artificial intelligence*, 388–391. https://doi.org/10.1109/TAI.1995.479783

Lobo, M. B. C. M. (2012). Panorama da evasão no ensino superior brasileiro: Aspectos gerais das causas e soluções [GS Search]. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, *25*, 14.

Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). Multi-class phased prediction of academic performance and dropout in higher education [GS Search]. *Applied Sciences*, *13*(8), 4702. https://doi.org/10.3390/app13084702

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics [GS Search]. *Scientific Reports*, *13*(1), 5705. https://doi.org/10.1038/s41598-023-32484-w

Maxwell, J. (1992). Understanding and validity in qualitative research [GS Search]. *Harvard educational review*, *62*(3), 279–301. https://doi.org/10.17763/haer.62.3.8323320856251826

Melo, E. C., & Souza, F. S. H. (2023). Improving the prediction of school dropout with the support of the semi-supervised learning approach [GS Search]. *iSys-Brazilian Journal of Information Systems*, *16*(1), 10:1–10:26. https://doi.org/10.5753/isys.2023.2852

Menolli, A., Horita, F., Dias, J. J. L., & Coelho, R. (2020). Bi–based methodology for analyzing higher education: A case study of dropout phenomenon in information systems courses [GS Search]. *XVI Brazilian Symposium on Information Systems*, 1–8. https://doi.org/10.1145/3411564.3411636

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach [GS Search]. *Higher Education*, *80*, 875–894. https://doi.org/10.1007/s10734-020-00520-7

Nagy, M., & Molontay, R. (2023). Interpretable dropout prediction: Towards XAI-based personalized intervention [GS Search]. *International Journal of Artificial Intelligence in Education*, 1–27. https://doi.org/10.1007/s40593-023-00331-8

Nascimento, F. F., Dantas, L. C. O., Castro, A. F., & Queiroz, P. G. G. (2024). Técnicas de mineração de dados e aprendizado de máquina aplicados à evasão estudantil: Um mapeamento sistemático da literatura [GS Search]. *Revista Brasileira de Informática na Educação*, *32*, 270–294. https://doi.org/10.5753/rbie.2024.3296

Perez, B., Castellanos, C., & Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study [GS Search]. *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*, 1–6. https://doi.org/10.1109/ColCACI.2018.8484847

Petersen, K., & Gencel, C. (2013). Worldviews, research methods, and their relationship to validity in empirical software engineering research [GS Search]. *2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, 81–89. https://doi.org/10.1109/IWSM-Mensura.2013.22

Rumberger, R. W. (2020). The economics of high school dropouts [GS Search]. *The economics of education*, 149–158. https://doi.org/10.1016/B978-0-12-815391-8.00012-4

Rumberger, R. W., Addis, H., Allensworth, E. M., Balfanz, R., Bruch, J., Dillon, E., Duardo, D., Dynarski, M., Furgeson, J., Jayanthi, M., et al. (2017). *Preventing dropout in sec-*

ondary schools [GS Search]. National Center for Education Evaluation; Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/PracticeGuide/24

Sastry, K., Goldberg, D., & Kendall, G. (2013). Genetic algorithms [GS Search]. In E. K. Burke & G. Kendall (Eds.), *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques* (pp. 97–125). Springer US. https://doi.org/10.1007/0-387-28356-0_4

Song, Z., Sung, S.-H., Park, D.-M., & Park, B.-K. (2023). All-year dropout prediction modeling and analysis for university students [GS Search]. *Applied Sciences*, *13*(2), 1143. https://doi.org/10.3390/app13021143

Teodoro, L. A., & Kappel, M. A. A. (2020). Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil [GS Search]. *Revista Brasileira de Informática na Educação*, *28*, 838–863. https://doi.org/10.5753/rbie.2020.28.0.838

Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: A survey of more than two decades of research [GS Search]. *Knowledge and Information Systems*, *66*(3), 1575–1637. https://doi.org/10.1007/s10115-023-02010-5

Unesco. (2023, April). What you need to know about higher education. https://www.unesco.org/en/higher-education/need-know

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education [GS Search]. *Technology in Society*, *76*, 102474. https://doi.org/10.1016/j.techsoc.2024.102474

Vora, S., & Yang, H. (2017). A comprehensive study of eleven feature selection algorithms and their impact on text classification [GS Search]. *2017 Computing Conference*, 440–449. https://doi.org/10.1109/SAI.2017.8252136

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution [GS Search]. *Proceedings of the 20th international conference on machine learning (ICML-03)*, 856–863. https://cdn.aaai.org/ICML/2003/ICML03-111.pdf

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? [GS Search]. *Proceedings of the eighth ACM conference on learning@scale*, 91–100. https://doi.org/10.1145/3430895.3460139

Zhang, W., Wang, Y., & Wang, S. (2022). Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in China [GS Search]. *Education and Information Technologies*, *27*(9), 13051–13066. https://doi.org/10.1007/s10639-022-11170-w