mathematics

MDPI

Article

Improving the Automatic Detection of Dropout Risk in Middle and High School Students: A Comparative Study of Feature Selection Techniques

Daniel Zapata-Medina, Albeiro Espinosa-Bedoya and Jovani Alberto Jiménez-Builes *

Department of Computer and Decision Sciences, Faculty of Mines, Universidad Nacional de Colombia, Medellín 050034, Colombia; dzapatame@unal.edu.co (D.Z.-M.); aespinos@unal.edu.co (A.E.-B.)
* Correspondence: jajimen1@unal.edu.co

Abstract: The dropout rate in underdeveloped and emerging countries is a pressing social issue, as highlighted by studies conducted by The Organization for Economic Co-operation and Development. This study compares five feature selection techniques to address this challenge and improve the automatic detection of dropout risk. The methodological design involves three distinct phases: data preparation, feature selection, and model evaluation utilizing machine learning algorithms. The results demonstrate that (1) the top features identified by feature selection techniques, i.e., those constructed through feature engineering, proved to be among the most effective in classifying student dropout; (2) the F-score of the best model increased by 5% with feature selection techniques; and (3) depending on the type of feature selection, the performance of the machine learning algorithm can vary, potentially increasing or decreasing based on the sensitivity of features with higher noise. At the same time, metaheuristic algorithms demonstrated significant precision improvements, but there was a risk of increasing errors and reducing recall.

Keywords: middle and high school; dropout; feature engineering; feature selection; metaheuristic algorithms; machine learning

MSC: 68T05

Citation: Zapata-Medina, D.; Espinosa-Bedoya, A.; Jiménez-Builes, J.A. Improving the Automatic Detection of Dropout Risk in Middle and High School Students: A Comparative Study of Feature Selection Technique. *Mathematics* **2024**, *12*, 1776. https://doi.org/10.3390/math12121776

Academic Editor: Mingbo Zhao

Received: 4 May 2024
Revised: 1 June 2024
Accepted: 5 June 2024
Published: 7 June 2024

1. Introduction

There is an initial preprocessing stage when implementing computational models since the data require an adequate treatment that allows knowing their behavior and their influence on the techniques' selection, training, and results [1]. This stage includes cleaning (incorrect, atypical, or empty data), integration, transformation (discretization, normalization), feature selection, feature reduction, and class balancing. In some studies, it is not mentioned in detail, but there is a need for information reliability that is mediated by data quality [2].

Variations have been identified in the final results of the methods used for prediction, which have been associated with shortcomings in preprocessing. Eckert and Suénaga (2015) [1] state, "*Preprocessing is considered an extensive and, at the same time, fundamental stage, because the subsequent results obtained in the training and evaluation of the algorithms depend on it*". In other words, it is a crucial stage before implementing any educational data mining technique. Márquez-Vera et al. (2013) [3] mention two problems to be addressed in educational data mining: (i) the dimension of the number of features and (ii) class imbalance. Moreover, a study by Delen et al. (2020) [4] showed that the reduction in data dimensionality and the use of synthetic data (for unbalanced classes) favored a 6% increase in the precision of machine learning algorithms. By the aforementioned, data quality significantly influences the outcomes of machine learning algorithms.

The authors Kuhn and Johnson (2019) [5] point out that there are different ways to represent predictors in a model and that some of these representations are more effective than others, namely kernel function and PCA (Principal Component Analysis), ICA (Independent Component Analysis), NNMF (Non-Negative Matrix Factorization), PLS (Partial Least Squares), autoencoders, spatial sign transformation, distance measures, and y depth measures, which are supposed to increase the effectiveness of a model. In the present study, we utilized the features developed during the feature engineering stage as described in [6]. We aim to compare feature selection methods aimed at enhancing the performance of automatic detection. To achieve this objective, we conducted experiments employing diverse types of features (personal, socioeconomic, and academic performance), five different feature selection techniques (Boruta, mRMR, LASSO, as well as metaheuristic algorithms such as genetic algorithm (GA) and Particle Swarm Optimization (PSO)), along with three machine learning algorithms (Support Vector Machines, random forest, and Gradient Boosting). Among the various feature selection techniques, metaheuristic algorithms have gained attention due to their ability to efficiently explore the feature space and find solutions close to the optimal ones. Inspired by natural processes or phenomena, these algorithms offer a flexible and adaptable approach to addressing complex optimization problems, such as feature selection [7].

Finally, for class imbalance, other studies use synthetic data. However, in this research, we only used data from a real scenario of a public educational institution at the middle and high school levels. The aim is to enhance the precision of the automatic detection of school dropout risk through feature selection techniques. We conducted the research design focused on educational data mining at a public educational institution covering secondary and high school levels in a traditional classroom setting. The study collected demographic and academic data from 1865 students in 2016–2019. All data were labeled (dropout, non-dropout). Therefore, to achieve the proposed objective, it was validated through the following steps:

(1) A comparison of feature selection techniques using demographic, academic features.
(2) A comparison of machine learning algorithms used in the dropout risk detection task with the feature inputs mentioned in the previous items.

The paper is organized as follows: Section 2 introduces the materials and methods and contains the proposed methodology for automatically detecting students at risk of dropping out of school through appropriate feature selection. Section 3 describes the results of the experimental design, comparing five feature selection techniques, and presents the findings. Section 4 presents the discussion, limitations, and future research directions. Finally, the conclusions are provided in Section 5.

2. Materials and Methods

School dropout has been the subject of multiple investigations, which seek to identify the factors that influence this problem. The study by Hernández-Blanco et al. (2019) [8] proposes a taxonomy of thirteen tasks related to educational data mining. Two of these tasks are considered the most important: predicting students' academic performance and detecting school dropout. Berens et al. (2019) [9] concur with other studies that one cannot reduce the determinants of school dropout to a single factor. Their results emphasize using academic performance data to enhance outcomes rather than solely relying on demographic data. It is worth noting that the datasets vary across related works.

2.1. Related Work

Several studies have identified that students' specific socioeconomic and personal features are related to their dropout risk. For example, in Maheshwari et al. (2020) [10], a detailed characterization is conducted alongside specific experiments to identify predominant relationships in school dropout, considering both male and female students. The study emphasizes the identification of the most influential factors in dropout through clustering techniques. For prediction, random forests and naive Bayes algorithms are

primarily employed. These prediction methods achieve an accuracy performance of 66.17% and 82.35%, respectively. Furthermore, this performance refers to the dataset utilized in the study, which encompasses various variables related to girls' school dropout, such as the number of female teachers and the percentage of schools with facilities exclusively for girls.

In another study [11], a cost matrix is applied to a dataset with a very high class imbalance, where only 4% of the students are dropouts. The results suggest that, considering the significantly higher cost of FN (false negative) error compared to FP (false positive) error, a recall performance exceeding 90% is achieved using random forests, followed by neural networks.

It is crucial to perform a careful selection of features in educational data mining, given the dimensionality of the datasets and the relevance of these features for model training [12]. It is essential to avoid the inclusion of noise or factors that may affect the decisions made by machine learning algorithms. For example, in the study conducted by Sansone (2019) [13], various models were used to estimate the probability of high school dropout among students. These models ranged from basic methods such as logit and probit to machine learning algorithms such as support vector machines and boosted regression. Different input variables were employed for each model. The best performance was achieved with SVM, which achieved an accuracy of 89.1% and a recall of 21.7%, followed by regression, with an accuracy of 88.8% and a recall of 20.6%. These results underscore the significant improvement gained from applying machine learning techniques, as well as the features selected using LASSO and boosted regression methods. Notable among the selected features are academic average, date of birth, math test scores, and participation in science and math courses.

Some studies highlight the diversity of features available in school information systems and the limitations in utilizing certain personal or censored data of students. Therefore, some of the features utilized in [14] are associated with student absences and the time spent on work or activities, which were deemed more significant based on their research. Furthermore, the random forests model achieved an accuracy of 95%, a specificity of 95%, and a sensitivity of 85%, indicating its superior prediction of the non-dropout class. However, the prediction of the dropout class may be considered within the reasonable ranges. In another study [15] that also utilized variables such as absences, travel time to school, gender, grades, assignments, and school and class size, among others, it was demonstrated that the random forest algorithm achieved an accuracy performance of 93.5%, compared to the SVM, which attained 90.4%. Lastly, the study suggests the inclusion of additional features or new ones to observe the behavior of machine learning algorithms.

The study by Pradeep et al. (2015) [2] found that the most influential features on performance were having a low socioeconomic status, being older than 15 years, having a low motivation to study, and considering mathematics as a complex subject, which allowed the detection model in their study to achieve an accuracy of 88–94%. da Cunha et al. (2016) [16] found that being a public school student with a family income of up to minimum wage, living with parents, being unemployed, or being a minor were also related to the risk of dropping out. Delen et al. (2020) [4] identified that parental financial and educational background and the student's prospects before starting school also influence performance, achieving an accuracy of 81–87% in their study using Bayesian networks. However, the most commonly employed features in dropout identification are those related to academic performance, such as course grades and and overall grade point average (GPA) [13,15]. Also, in [17], math test scores, credits taken, the ratio of credits earned to credits taken, age, and surplus score were considered. The above could be due to the efficiency of training a machine learning algorithm with these features and the availability and accessibility of such academic data.

On the other hand, other studies of higher education Pérez et al. (2018) [18] identified features related to the place of residence, student's career, high school cumulative average, an d UST (University Selection Test) score in mathematics; in particular, "course loss" and "low GPA" are highlighted. Furthermore, Refs. [19,20] identified the relationship

between course repetition, gender, income level, and mathematics skills with dropout. The latter features were also related to algebra and calculus courses identified by Dekker et al. (2009) [21] concerning the low scores found in the nodes of the random forest and decision tree algorithms, using a cost matrix to improve the accuracy of the algorithms in identifying the dropout class.

As mentioned above, considering the availability and accessibility of academic performance data within both secondary and higher education contexts, the GPA features have been widely studied and were an influential factor in school dropout in multiple investigations [22–24]. Furthermore, in other studies [25,26], low GPA and absenteeism have been identified as standard features in students who drop out, while Barros et al. (2019) [27] and Bedregal-Alpaca et al. (2020) [28] identified the number of credits the student enrolls in.

In addition, other studies [29,30] identified grades, exams, and course results, particularly the loss of a course, to be influential factors in dropout. Aulck et al. (2019) [31] proposed the following question: What types of data from registrar records are most useful in predicting attrition? The answer was that academic features are more effective than demographic data in predicting dropout, particularly in academic summaries of semester time sequences. However, as future research, it was suggested to delve deeper into understanding the relationships among the features used to predict dropout by combining demographic and academic features.

Regarding the use of different types of features in educational data, Manrique et al. (2019) [32] proposed an approach that considers two types: global feature-based (GFB) and local feature-based (LFB) types. This approach allows a comparative analysis of various predictive methods based on a proposed classification strategy to predict dropout. The GFB is achieved by selecting demographic features. In contrast, the LFB is based on a vector of features extracted from academic performance, using the grades obtained (considering the best grade in case of grade repetition), demonstrating that the latter is the most suitable for predicting dropout. The authors conclude that it is possible to accurately predict dropout risk using grades from a few introductory courses without a complex feature selection process.

In the context of educational data mining, it has been observed that selecting appropriate features significantly impacts the results obtained in prediction and classification models [5]. Recent studies have shown that results vary according to the features used, and the challenge of adequately incorporating demographic features remains [9,23].

Currently, educational databases include relevant information on student demographics to identify their social environment. The availability of this information can improve communication between the educational institution, the family, and the community, which in turn can positively impact the quality of school life [33]. On the other hand, academic features are related to different aspects of the student's learning process, including cognitive, procedural, and attitudinal aspects. The evaluation of the educational process is defined from the integral measurement of these three aspects. Consequently, the appropriate use of academic features is fundamental for accurately assessing students' academic performance.

Several features can be identified in student data, among which demographic (socioeconomic and personal) and academic features stand out, differing in their heterogeneous nature. The former are collected during the enrollment process, while the latter are obtained during the learning process. In [6], new features were proposed as an alternative, incorporating aspects implicit in student data but not explicitly present in the dataset, such as grade repetition, coverage, and distance to the school, among others.

The findings suggest that the performance achieved in the reviewed studies falls within reasonable ranges for correctly classifying dropout students. This is a promising result, indicating that automatic classification could be practically implemented in expert systems or school information systems to detect students at risk of dropping out early, potentially leading to more effective intervention strategies. The primary gap identified in the reviewed literature on the automatic classification of students at risk of school dropout lies in the difficulty of determining the most relevant and suitable predictors for training machine learning algorithms. Additionally, using different feature selection techniques

across various studies, or the absence of such techniques in some cases, leads to variability in the machine learning algorithm's performance outcomes. There has yet to be a consensus on which feature selection methods are most effective.

2.2. Research Gap

Feature selection is a technique that can improve algorithm performance in some cases: (1) It can reduce computational complexity and enhance the model's generalization capability, as demonstrated in several studies [3,34]. (2) Increasing the model's efficiency is crucial, particularly in imbalanced class scenarios, especially for the minority class [11]. Focusing on more significant features facilitates model interpretation [5]. Based on previous studies, the identified research gap is the need to improve performance in the minority class without compromising precision in the majority class. Implementing class balancing techniques and more rigorous feature selection can be an effective strategy to address this challenge and enhance the model's overall performance, according to recent studies [3].

2.3. Methodology

The traditional preprocessing method generally involves a contextualized and multi-stage process. In this research, we performed (i) data cleaning, (ii) data transformation, and (iii) feature selection. Data analysis is often affected by record inconsistencies; therefore, a deep preprocessing stage can reduce representation problems and facilitate the understanding of the features. Thus, a three-stage methodology was proposed: data preparation, feature selection, and model evaluation (see Figure 1).

The first step of the methodology began with data collection, which the educational institution provided. These data included demographic features such as age, gender, and socioeconomic status and academic features such as grades in various subjects. After collecting the data, we performed preprocessing, which involved cleaning and preparing the data for analysis. This process involved removing missing values and transforming features using normalization and potential transformations, such as the Box–Cox transformation. Afterward, we conducted an exploratory data analysis to identify critical relationships. Based on this analysis, we performed a feature engineering process, which involved creating new features from existing ones. Five different feature selection techniques were then applied to identify features that could provide relevant and suitable information for machine learning algorithms. The next step involved implementing the classification task, utilizing three different algorithms trained with both the initial features and those selected by each method for subsequent performance comparison. Cross-validation techniques, grid search, and performance metrics such as precision, recall, and F-score were employed to assess algorithm performance.

Figure 1. Methodological scheme proposed. Source: Authors.

2.4. Data Preparation

Some studies presented data cleaning and transformation preprocessing [1,34–36]. Kiss et al. (2019) [17] used the techniques k-NN, Miss Forest, MICE for empty data, and matrix factorization for outliers. Dharmawan et al. (2018) [37] used filters in WEKA, such as ReplaceMissingValues and InterquartileRange. On the other hand, transforming features

includes converting to the same scale to ensure that the values are in the same range. We applied data scaling techniques, thereby transforming the features to a predefined range, typically (0,1). Based on the theory of feature engineering, there are two crucial aspects. The first refers to the design, and the second to feature selection. In this research, we used the records of the dataset of a real scenario of a public educational institution. This data collection includes personal, socioeconomic, and academic information on students from 2016 through 2019. However, this study only consisted of middle school and high school.

First, we cleaned the data and then performed feature engineering. After the data were cleaned, 1865 records were obtained, with twenty-nine student features (sixteen demographic and twelve academic), including the label (class) for each record; see Table 1. The demographic features gathered the personal and socioeconomic attributes of the student body, and the academic features corresponded to grades in different courses (academic performance) with reference to a specific year in which the student was enrolled. To conduct further analysis, we needed to transform the place of residence, neighborhood, and city of residence into geographic coordinates. To achieve this, we utilized the Google API to geocode the addresses and locations of residence, which allowed for the conversion of each address into its corresponding latitude and longitude coordinates.

Table 1. Demographic and academic features of the initial dataset.

Feature		Description
Socio-economic	Latitude	Latitude coordinate of the student's place of residence
	Longitude	Longitude coordinate of the student's place of residence
	SISBEN_Score	Score at SISBEN system
	Siblings [a] _School	Number of siblings at school
	Social_Class	Socioeconomic level
	SISBEN_Category [a]	Level SISBEN III (Extreme_poverty, Moderate, Non-poverty)
Personal	Grade_n	Student's grade of entry to the educational institution
	Sector_n	Zone of residence
	Marital_Status_Father	Marital status of student's father
	Marital_Status _Mother	Marital status of student's mother
	Gender_n	Student's Gender (Male, Female)
	Course_in	Student's current grade level
	Rh_Factor	Student's Rh_blood group
	Age	Student's age
	Ethnicity	Type of student's ethnicity
	Year	Student's year of entry to the educational institution
Academic	Natural_Sciences	Student's point average in Natural Sciences
	Mathematics	Student's point average in Mathematics
	Entrepreneurship	Student's point average in Entrepreneurship
	Technology	Student's point average in Technology
	Sports	Student's point average in Sports
	English	Student's point average in English
	Religion	Student's point average in Religion
	Peace_Lecture	Student's point average in Peace lecture
	Social_Sciences	Student's point average in Social Sciences
	Arts	Student's point average in Arts
	Ethics	Student's point average in Ethics and Values
	Spanish	Student's point average grade in Spanish

[a] Sisbén: System for the Identification of Potential Beneficiaries of Social Programs. Source: Authors, adapted from [33].

Additionally, further features derived from the initial features were incorporated. Previous studies established specific criteria, which are outlined in Table 2:

- **Distance:** The distance metric was calculated from the school's geographic location, using the *Haversine* equation to measure the distance from the student's point of residence to the school's location.
- **Age_frac:** The age_frac metric utilized the student's date of birth (day, month, and year) to calculate their age based on the school enrollment year.
- **Overage** The overage metric focused on overage, the difference between the student's current age and the theoretical age defined by the Ministry of National Education in Colombia—MEN [38], for the school grade they attend.
- **Repetition:** The repetition metric focused on grade repetition based on the year of the study as a criterion, which was conducted in 2020.
- **Cox_hec:** As for the Cox_hec, the number of siblings the student has in the same school was used as a criterion, and the Box–Cox transformation was applied to avoid bias.

Table 2. Features derived from initial features.

Feature	Equation
Distance	$d = 2r\,sen^{-1}\left(\sqrt{sen^2\left(\frac{lat2-lat1}{2}\right) + cos(lat1)\,cos(lat2)\,sen^2\left(\frac{lon2-lon1}{2}\right)}\right)$
Age_frac	$Age\_frac = date_{\_ofbirth}(dd/mm/yyyy) - date(15/06/year_{entry})$
Overage	$theoretic\_age = grade\_entry + 5$ $Overage = age\_entry - theoretic\_age$
Repetition	$Repetition = (current\_year - year + grade\_entry) - current\_grade$
Cox_hec	$x = \begin{cases} \frac{x^\lambda - 1}{\lambda \tilde{x}^{\lambda-1}} & \lambda \neq 0 \\ \tilde{x} \log x & \lambda = 0 \end{cases}$

Source: Authors, adapted from [33].

2.5. Feature Selection

In addition to the data cleaning and transformation process, other studies have employed feature selection to favor dimensionality reduction and enhance the efficiency, accuracy, and interpretability of machine learning algorithms. This approach focuses on the most relevant features for the issue of school dropout [35,39]. Hegde and Prageeth (2018) [40] used the InfoGainAttributeEval algorithm in WEKA for feature selection. Feature selection methods can be classified into three categories: (1) Filtering methods: correlation coefficient, Chi-square coefficient, discriminant analysis, cluster analysis, mutual information, among others. (2) Wrapper methods: Recursive Feature Elimination (RFE), Forward Selection, Backward Selection, Metaheuristic Algorithms, Boruta's method, etc. (3) Embedded methods: Lasso, Ridge Regression, etc.

2.5.1. Discriminant Analysis

The discriminant analysis allows separating the data into two groups and determining to which of the groups a new data point should be assigned: "*The discriminant analysis reduces this overlap by maximizing the variation between clusters and minimizing the variation within each cluster*" [41]. On the other hand, using correlation analysis alone can ignore any dependency relationship that some features may have with the target variable.

Ding and Peng (2003) [42] first proposed the Minimum Redundance Maximum Relevance Feature Selection (mRMR) method. It enables the determination of the most relevant features with good discriminant capacity. So the mRMR method was initially designed to find the features that maximize the relevance between them and the target class while minimizing the redundancy between features. The mRMR method for discrete features is defined as follows. Given two variables x and y, their mutual information is defined in

terms of their probabilistic density function $p(x)$, $p(y)$ and $p(x, y)$. In this study, the mRMR method is applied to select the subset of features most influential in school dropout. To achieve this discrimination, the quotient of relevance $(c|x_i)$ over redundancy $(x|x_{i-1})$ is calculated for each class–feature pair to select the feature whose score is higher [33].

2.5.2. Boruta Method

This algorithm, as described by Kursa and Rudnicki (2010) [43], was developed as a wrapper selection method around a random forest classification algorithm. Following the concept of the algorithm proposed by Guyon et al. (2002) [44], Boruta's method resembles recursive feature elimination (RFE), which iteratively fine-tunes a supervised algorithm from a set of features. As mentioned in [45], "*At each stage of the search, the least important predictors are iteratively eliminated before rebuilding the model*". Therefore, Boruta's primary objective is to iteratively adjust a supervised algorithm (random forest) and to rank and eliminate less relevant features using a statistical test based on the score obtained for each feature.

2.5.3. LASSO Regression

LASSO—Least Absolute Shrinkage and Selection Operator—is a regularization method that penalizes using the L1-norm (Manhattan). Feature selection is performed simultaneously with regularization to improve the model (value of the coefficients), and the penalty is proportional to the absolute value of the coefficients. As a result, reducing the coefficients of certain unwanted features to zero eliminates unnecessary variables [45]. Friedman et al. (2010) [46] stated that "*LASSO is not sensitive to highly correlated predictors, and will tend to choose one and ignore the rest*".

2.5.4. Metaheuristic Algorithms

Metaheuristic algorithms are notable for their capacity for iterative improvement, which involves a thorough exploration of various areas within the solution space to identify optimal candidate solutions [7]. This capability is based on their stochastic nature, characterized by incorporating random or probabilistic components in the search process, enabling them to bypass local optima. While they do not guarantee the identification of the global optimum, metaheuristics aim to find satisfactory solutions within a reasonable timeframe. Flexibility is another distinctive feature, as these algorithms are adaptable and applicable to a wide range of optimization problems [7].

In this study, we employ two metaheuristic algorithms for feature selection in school data. The first is a genetic algorithm (GA), belonging to the category of evolution-based algorithms, inspired by natural evolution. These algorithms maintain a population of candidate feature subsets and apply genetic operators such as selection, crossover, and mutation to evolve toward optimal solutions. The second algorithm, belonging to the category of swarm-intelligence-based algorithms: the Particle Swarm Optimization (PSO) technique, developed by Kennedy and Eberhart [47], inspired by the behavior of bird flocking.

2.6. Model Evaluation

The problem of dropout risk detection has been extensively investigated in the literature using supervised learning techniques, such as the random forest. These studies have achieved high accuracy levels above 80% [13,15]. In addition, the simplicity of parameterizing these algorithms and the less complex preprocessing stage compared to other algorithms, such as the SVM or ANN, have been highlighted. However, previous studies have reported that the SVM algorithm, which has also been frequently used to address the problem of dropout risk detection, exhibits low performance with unbalanced data and overlapping classes. The above is due to the generalization ability affecting the minority class [37,48,49]. Therefore, using a nonlinear kernel for this type of problem and taking into account the penalty to compensate (metrics), using, for example, the parameter

weight = "balanced" have been suggested. In this way, during the algorithm's training, an adequate balance for the minority class is guaranteed.

Machine Learning Algorithms

In dropout risk detection, some previous studies have used traditional rule-based algorithms [21,50], which can be complex to describe and can present difficulties in decision-making. In comparison, learning machines can automatically learn rules from known samples. After training on these samples, the resulting model is used to classify the unknown inputs. Some studies have proposed using ensemble methods to address the problem of unbalanced classes in dropout risk detection [13,51,52]. These methods, such as Bagging, Boosting, and Stacking, combine several basic classifiers to improve classification performance. The random forest algorithm has also been utilized and has achieved notable performance in accuracy, ranging between 66% and 90% [10,14,15].

In the current study, we compared three algorithms for detecting students at risk of dropout: a support vector machine (SVM), a random forest (RF), and a gradient boosting tree (GBT). Our review of this prior work revealed that among the 77 studies analyzed, 16% utilized random forest models, while 10% utilized support vector machines and gradient boosting models. Likewise, Serra et al. (2018) [53] used the LR, NB, and SVM algorithms; the SVM algorithm achieved the best performance, with 83% precision. On the other hand, Sangodiah et al. (2015) [54] implemented three algorithms, the SVM, decision trees, and a rule induction algorithm, which achieved precision rates of 89.84%, 86.32%, and 81.98%, respectively.

2.7. Ethical Considerations

The study was conducted with informed consent and under the approval and supervision of a university scientific ethics committee. The purpose of data use was clearly defined, and a robust policy of confidentiality and privacy was established to safeguard the collected personal information. The fair and equitable treatment of all involved students was ensured. These ethical considerations are supported by Colombian Law 1266 of 2008 on the Right of Habeas Data and Statutory Law 1581 of 2012, which protects the privacy of personal data. Additionally, data pseudonymization was implemented to prevent the individual identification of students at any point during the study.

3. Results

The most relevant features were selected using the mRMR (discriminant analysis), Boruta, LASSO, GA, and PSO (Table 3). The mRMR method obtained importance scores for each demographic and academic feature. The results obtained indicate that twelve academic features were selected (Figure 2). In this way, we seek to enhance the particular importance of the features with the relationships found among them. The results of demographic features show that the six features selected by the mRMR method were *grade_n*, *repetition*, *distance*, *course_in*, *sector_n*, and *sisben*.

Implementing the mRMR method followed the variants proposed by Hanchuan Peng et al. (2005) [55] under the feature type, whether quantitative or qualitative. The method calculates the quotient between the F-test and Pearson's correlation coefficient. However, for this study, a variant of the method was used, which involved using the coefficient of the Kruskal–Wallis test and Spearman's correlation due to the non-normal distribution of the data. The Kruskal–Wallis test (H-test) was used because it is a hypothesis-testing technique applicable to multiple independent samples and does not depend on the normal distribution of the data as long as they are on an ordinal or continuous scale. Furthermore, the simulation study by Lix et al. (1996) [56] describing a meta-analysis of different alternative statistical tests to the analysis of variance (ANOVA) concludes that the selection of an alternative test procedure should be considered when there are consequences under violations of the assumptions of normality and homogeneity of variance; for example,

it is appropriate to test the equality of distributions with the Kruskal–Wallis procedure. However, it is recommended to take a large number of samples.

Table 3. Features selected by each technique (mRMR, Boruta, LASSO, PSO, GA). An 'X' indicates the features selected by each technique.

	Feature	mRMR	Boruta	LASSO	PSO	GA
Socio-economic	Latitude			X	X	X
	Longitude					X
	Sisben_Score				X	X
	Siblings_School				X	
	Social_Class			X		
	Sisbén_Category					
	Distance	X		X	X	X
Personal	Grade_n	X	X	X		
	Sector_n	X		X	X	X
	Marital_Status_Father				X	X
	Marital_Status _Mother					X
	Gender _n			X		
	Course_in	X	X			
	Rh_Factor				X	X
	Age		X	X	X	X
	Ethnicity				X	X
	Year _entry			X	X	X
	age _frac		X		X	X
	Overage		X	X	X	
	Repetition	X	X	X		
	Cox _hec				X	X
Academic	Natural_Sciences	X	X	X	X	X
	Mathematics	X	X	X	X	X
	Entrepreneurship	X	X	X		
	Technology	X	X	X		
	Sports	X	X	X	X	X
	English	X	X	X	X	
	Religion	X	X	X		
	Peace_Lecture	X	X	X	X	
	Social_Sciences	X	X		X	X
	Arts	X	X	X		
	Ethics	X	X	X		
	Spanish	X	X		X	

Source: Authors.

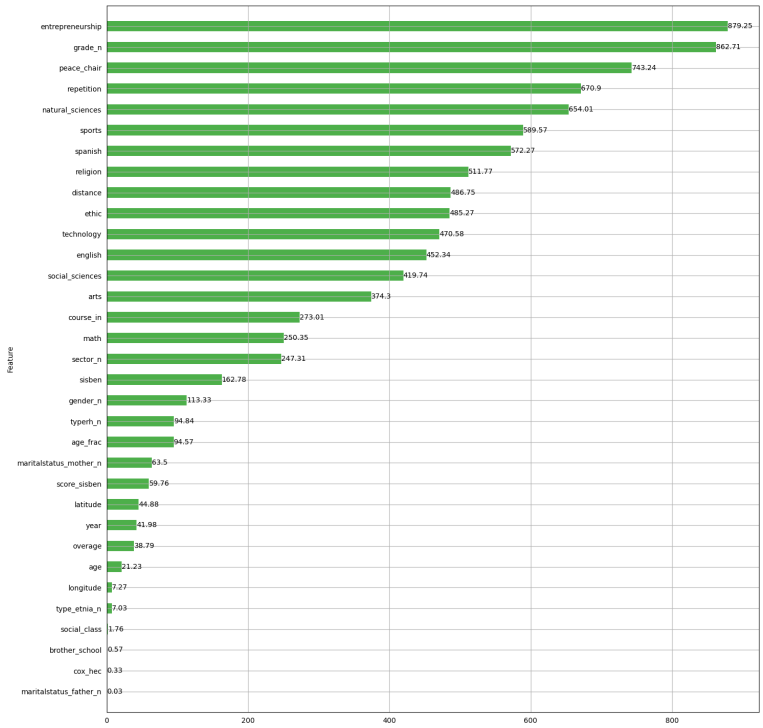

Figure 2. Feature importance—mRMR. Source: Authors.

The mRMR method sorts the features independently for each iteration; i.e., in the first iteration, the order of the features is not necessarily the same for the second iteration,n and so on, which indicates that the method allows discriminating in independent analyses, combining the analysis of a statistical test concerning the target variable and the correlation among predictor features. The Boruta and LASSO methods selected six and ten demographic features, respectively (Figures 3 and 4). Those two methods selected all relevant features in a single iteration using an automatic classification model to choose the features (random forest and Logistic Regression). It was observed that the mRMR and Boruta coincide in selecting fifteen features, while the LASSO selects two not selected by the others. The Boruta and LASSO methods selected twelve and ten academic features, respectively.

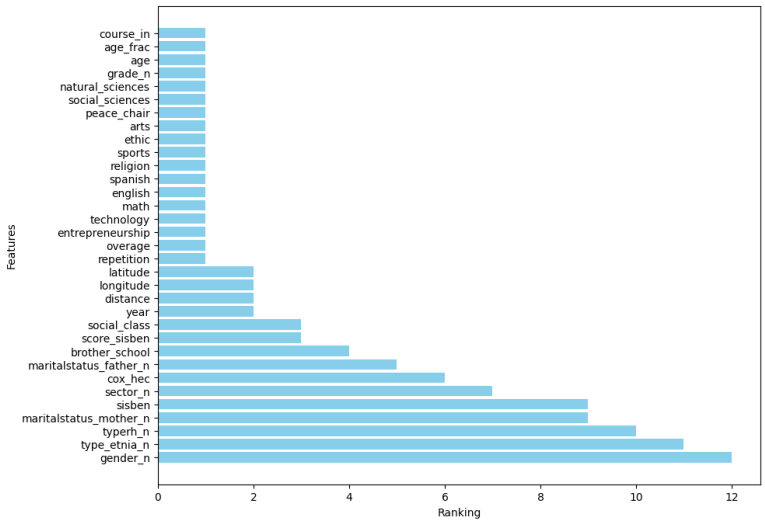

Figure 3. Feature importance—Boruta. Source: Authors.

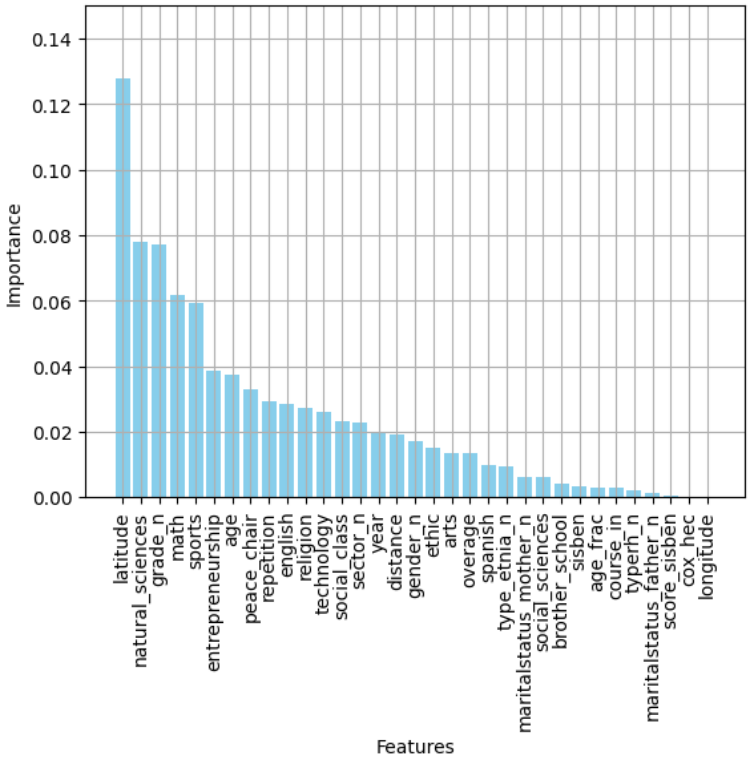

Figure 4. Feature importance—LASSO. Source: Authors.

The Boruta method is a technique that relies on the random forest algorithm to produce a set of shadow features in each iteration from the predictors, where each of the predictors is copied, and the elements of each column created are permuted. The LASSO method is based on the Logistic Regression algorithm with the "L1-norm" penalty. In this way, to obtain a complete view of the importance of the predictor variables, the hyperparameters used in the random forest classifier were selected by a five-fold cross-validation grid search (taking into account the best scores achieved in terms of precision and recall). Finally, for the genetic algorithm (GA), the random forest algorithm was employed, while the Particle Swarm Optimization (PSO) algorithm utilized Logistic Regression. The hyperparameters for the GA included the number of features, the population size, the number of iterations, the percentages of the elite population, and the total genes that mutate. As for the PSO, the hyperparameters comprised the number of features, population size, number of iterations, and acceleration coefficients.

The metaheuristic algorithms PSO and GA selected thirteen demographic features. Regarding academic features, they selected seven and four, respectively. It was observed that all five methods coincide on two features, "Math" and "Natural_sciences", and nearly always, at least four methods coincide on ten features. This demonstrates the relevance of features such as overage and grade repetition, distance, year of enrollment (year_entry), current grade level (current_grade), gender, and the average score in the natural sciences course. Finally, we observed that three features, "Sisbén_Category", "Social_Class", and "Gender_n", were exclusively selected by one of the three methods or none at all, indicating that they are the least relevant to the compared techniques.

Evaluation

In this study, we conducted eight experiments to train and test three classifiers, RF, GBT, and SVM-RBF (radial basis function kernel), using different types of features (see Table 1). The first experiment consisted of training the three classifiers with twenty-one demographic features (DF), the second with twelve academic features (AF), and the third with thirty-three features, both demographic and academic (DAF). The fourth experiment used fifteen features that were selected using the mRMR (FS_mRMR), while the fifth experiment used

nineteen features that were selected using the Boruta (FS_Boruta) method. In the sixth experiment, twenty-five features were selected using the LASSO (FS_LASO) method. In the seventh experiment, nineteen features were selected using the PSO algorithm (FS_PSO). In the eighth experiment, eighteen features were selected by the algorithm GA (FS_GA).

To evaluate the feature selection and representation activity, the random forest (RF), gradient boosting (GBT), and support vector machine (SVM) algorithms were trained to automatically detect dropout risk. The dataset was divided into 70% for training and 30% for testing; see Table 4.

Table 4. Distribution of dataset for training and evaluation.

Class	Train	Test
Non-Dropout	870	435
Dropout	435	125
Total	1305	560

Source: Authors.

The hyperparameters used in the RF, GBT, and SVM classifiers were selected by a five-fold cross-validation grid search. Then, to measure the performance of the classifier, the following was used: precision $Pr = \frac{TP}{TP+FP}$, recall $R = \frac{TP}{TP+FN}$ and F-measure $F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision+Recall}$. Table 5 shows the variations in performance during the test, which shows that the GBT algorithm is superior in most experiments compared to RF and SVM. The results indicate that SVM is sensitive to the type of feature used. At the same time, RF works well for any feature type, especially when using feature selection.

Table 5. Classification performance on the testing set with the initial demographic features (DF), academic features (AF), all initial features (DAF), features selected by the Boruta method (FS_Boruta), features selected by the mRMR method (FS_mRMR), features selected by the LASSO method (FS_LASSO), features selected by the Particle Swarm Optimization (FS_PSO) algorithm, and features selected by the genetic algorithm (FS_GA). Bold values indicate the best performance metrics.

Algorithm / Features	SVM			RF			GBT		
	Pr	**R**	**F1**	**Pr**	**R**	**F1**	**Pr**	**R**	**F1**
DF	0.69	0.58	0.63	0.79	0.61	0.69	0.82	0.63	0.72
AF	0.55	0.71	0.62	0.72	0.57	0.63	0.81	0.54	0.65
DAF	**0.79**	0.58	0.67	**0.92**	0.57	0.70	0.84	0.60	0.70
FS_mRMR	0.64	0.75	0.69	0.80	0.53	0.64	0.85	0.56	0.68
FS_Boruta	0.63	**0.79**	**0.70**	0.85	**0.62**	**0.71**	0.84	0.59	0.70
FS_LASSO	0.65	0.75	0.70	0.87	0.57	0.69	**0.92**	**0.63**	**0.75**
FS_GA	0.61	0.65	0.63	0.88	0.50	0.64	0.90	0.53	0.67
FS_PSO	0.59	0.69	0.64	0.88	0.54	0.67	0.88	0.54	0.67

Source: Authors.

However, we found that implementing the learning algorithms with only demographic or academic features proved to be complex because the parameters of the RF and SVM presented a cost for correctly classifying the dropout class. In other words, obtaining a low variance resulted in a high bias and vice versa, making classifying both classes challenging.

On the other hand, we found that the feature selection methods increased the performance of the machine learning algorithms in detecting students at risk of dropout. The smaller number of features reduced the computation time, and similar results were obtained. However, the results of the test dataset show that the number of selected features had an impact during the evaluation of the algorithms with new unknown samples. Table 5 shows the performances achieved by the algorithms with the features selected by each technique in every training experiment, demonstrating the importance of selecting the

demographic and academic features. The recall values decreased in the evaluation of the experiment with the thirty-three initial features, contrary to what happened when it used feature selection methods. Regarding this aspect, the recall value was significant, determining how many positive classes were correctly predicted. That is, it risks false positives, but it is more important to avoid false negatives. The above is crucial in the case of detecting a dropout student.

Table 5 shows the optimal performance results regarding the precision, recall, and F1-score obtained using different classifiers and features in the test dataset. During training, the SVM, RF, and GBT exhibited high performance; however, when testing unknown samples, they made errors in too many instances, resulting in a lack of model generalization when using all features. Therefore, selecting the appropriate features helps the models improve their performance and capacity for generalization. In particular, the GBT algorithm achieves 92% precision and 63% recall with features selected by LASSO, indicating a significant improvement compared to using initial features, where 84% precision and 60% recall were obtained. In addition, applying feature selection techniques, specifically the minimal–optimal method, allowed identifying the optimal features to train the model, reducing the computational cost. We must note that we addressed class imbalance during model training using a penalty strategy without synthetic data.

The best model is difficult to choose because three classifiers demonstrated different results depending on the feature selection technique. SVM performed well with features selected by Boruta, achieving a precision of 64% and recall of 75%. Gradient boosting trees (GBTs) achieved the best performance with features selected by LASSO, with a precision of 92% and recall of 63%. Random forest, with features selected by Boruta, achieved a precision of 85% and recall of 62%. Overall, the classifiers obtained an average performance of over 88% concerning the accuracy metric and achieved better performances with feature selection, demonstrating the effectiveness of the process in improving the performance of the machine learning algorithm in detecting students at risk of dropout. Furthermore, good results were obtained with the RF and GBT algorithms when only academic features were utilized. However, combining both academic and demographic features yielded even better results.

4. Discussion, Limitations, and Future Research

The results demonstrated that depending on the feature selection, varying performances in machine learning algorithms can be observed, which may increase or decrease depending on how sensitive the algorithm is to noisier features. While the metaheuristic algorithms allow for significantly increasing precision, there is a risk of making too many errors and obtaining a low recall.

It is relevant to note that most studies on the phenomenon of school dropout rely on datasets from higher or university education, unlike our study, which focused on data from middle and high schools. School dropout is a multifaceted phenomenon influenced by various factors, including individual student features and the school and social environment. By addressing multiple educational levels, we can capture a broader range of these influences and better understand how they vary across different educational stages. Moreover, since academic performance results may directly impact students' progression from one grade to the next, it is crucial to consider how these results relate to the likelihood of school dropout at each educational level. We contend that the joint analysis of data from middle and high schools provides a more comprehensive view of the school dropout phenomenon, enabling us to develop more robust predictive models applicable to a variety of educational contexts.

For example, Berens et al. ([9]) compared two university data sets, one from public universities and the other from private universities, and concluded that although demographic features help predict student dropout during enrollment and in the first semester, academic performance becomes a more robust and precise indicator as students progress in their studies. Therefore, academic data gain greater significance in predicting dropout

in subsequent semesters. These results are consistent with our study, which found that academic features improve machine learning algorithms' predictions. However, it was necessary to incorporate some demographic features to improve the precision of the GBT algorithm from 81% to 84%, the RF from 72% to 92%, and the SVM from 55% to 79%.

Our results indicate that including demographic features significantly improves the predictive ability of machine learning algorithms for school dropouts. These findings are relevant for middle school and high school. Furthermore, these results are consistent with other studies highlighting the importance of considering academic and demographic features in predicting dropout.

According to Aulck et al. (2019) [31], academic performance features are critical in predicting dropout risk rather than relying exclusively on demographic data and manually creating features. The latter could save time and effort in modeling. Additionally, understanding why a response is reached is as important as obtaining an output or response from the algorithms.

On the other hand, Manrique et al. (2019) [32] conducted a similar study and proposed three different representations based on global, local, and time series features. Of these representations, the results showed that the second representation was the most suitable, indicating that predicting dropout accurately is possible using grades from a few introductory courses without requiring a complex feature extraction process. Manrique et al. (2019) [32] achieved the best results with the RF and GBT algorithms, similar to the results obtained in this investigation. However, this study required demographic and academic features to improve the predictions.

Concerning the above, we developed some of these proposals; in particular, a dataset with demographic and academic features was included to generalize the results more. To achieve this goal, we used common features in other studies, such as distance, grade repetition, and overage. Combining different features was essential to improve the precision of the predictions, as algorithms using only one type of feature were found to be ineffective.

A reviewed study [4] showed that proper feature selection improved machine learning algorithm performances by 6%; compared to our research, the algorithm performances increased by 10% in terms of sensitivity. Furthermore, in that study, it was observed that the best performance in terms of sensitivity (70%) was achieved using all fifteen features and with synthetic data (Synthetic Minority Oversampling Technique (SMOTE) algorithm). It also achieved a precision of 76% and a specificity of 82%. Feature selection supported the improvement in classification performance, achieving a 10% increase in recall of the machine learning algorithm compared to using all features.

In general, feature selection significantly impacts the models' precision, recall, and F1 Score: GBT performs exceptionally well with the FS_Lasso feature set, achieving the highest precision of 92%, recall of 63%, and F1 score of 75%. The SVM model performs best with Boruta's selected features, achieving the highest values (79% and 70%, respectively). Finally, the RF model shows a reasonable recall with the features selected by Boruta (62%) and achieves a high F1 Score (71%).

The variability in the results suggests that certain feature selection algorithms may be more suitable for specific models. Also, using all features (DAF) can lead to a more complex model prone to overfitting, even though some models like RF show significant improvements in precision. In addition, feature selection is crucial, but not all selection methods guarantee significant improvements across all models. On the other hand, class imbalance (more examples of the majority class) can affect the models' ability to learn and predict the minority class correctly.

4.1. Limitations

The educational dataset of this study presented limitations concerning class imbalance (dropout class with a 3:1 ratio), and it was necessary to use the penalty strategy during the training of the machine learning algorithm. In addition, feature selection generated more

time during data preprocessing, and it is crucial to consider the type of distribution of the data under study when using the mRMR method.

Feature engineering is a time-consuming process. A robust preprocessing stage involving various processes such as feature cleaning, transformation, extraction, and selection must be performed to perform the transformation or reworking of new features. Each dataset is unique, so understanding the relationships and behaviors of the data may yield different results depending on the statistical tests performed. Therefore, only a few metrics, such as distance, level of repetition, and overage, can be used generally. Currently, there are automated methods for feature engineering, but the manual process, although slower due to the trial-and-error strategy, allows an understanding of the behavior of the data and interpretation, to some extent, of the response of machine learning algorithms. Hence, explanatory methods are needed, given that the current work provides only a partial answer.

It is worth noting that techniques such as GA and PSO can be computationally intensive in terms of time and resources. However, the results they produce may only sometimes generalize well to other datasets due to the inherent variability in the data and selected features. This trade-off between computational intensity and generalizability is an important consideration in the context of feature selection for machine learning models.

Some educational data are collections of records created by persons, which are susceptible to errors during data collection. Initially, the complexity of educational data lies in the outdated nature of records, lack of maintenance, and failure to track real-time changes in the data.

In general, real-world educational data suffer from quality problems such as missing data or data lacking features, records containing incorrect or atypical values (i.e., data noise), and inconsistency in some records. In addition, detecting the risk of student dropout presents specific challenges that depend on each information system and how it records data. Another challenge is the class imbalance, with more records in the "Non-Dropout" class.

4.2. Future Research

To select features, one must consider that one should aim at the best n-features. The sentence described above was demonstrated by the example described in 1976 by Cover (1974) [57], who stated, "*The two best independent measures are not the two best*". Other techniques for feature selection can be explored, such as those proposed by Lima et al. (2018) [58], by using the K-means algorithm, and Fernández et al. (2018) [34], by using a centered alignment kernel to match demographic, academic, and biopsychosocial features to the target variable (Dropout/Non-Dropout). In this regard, exploring techniques to identify which features have significantly influenced the model predictions, such as Shapley Additive Explanations (SHAP), would also be pertinent. As evidenced by the results obtained, there is variability in the model's performance depending on the feature selection technique employed. In this context, merely evaluating the importance of features alone may not suffice. Therefore, it is imperative to investigate and consider which features have contributed to the increase or decrease in prediction errors in the models to discern which ones provide genuine benefits rather than merely being relevant.

According to the comparison of feature selection techniques, it is necessary to implement the feature engineering used here with other educational data collections and to validate the responsiveness of these with other data or to generate new features that have points in common to establish them in a general way on the subject of school dropout, as is the case with distance.

Regardless of the model to implement in future research, a better interpretation of the results will always be necessary. Most machine learning algorithms are complex to interpret, so it is necessary to continue emphasizing strategies that help users of the systems, in this case, in the education sector, understand the results—for example, using unsupervised learning algorithms to segment and represent feature types at the personal, social, and academic levels. Therefore, in future work, Aulck et al. (2019) [31] suggest deepening the

understanding of the features used to predict attrition risk and finding the best combination of features across subsets. For future research, Manrique et al. (2019) [32] proposed conducting new experiments with several data sets and including the sociodemographic data omitted in their study to generalize the results.

In the present study, using academic and demographic features achieved satisfactory results. However, using other features such as class attendance, psychological and health issues of the students, access to resources necessary for their educational process, or interactivity in social networks can improve these results. One of the most relevant points in this type of dropout research is the use of a real-time model, i.e., obtaining as much data as possible at the beginning of enrollment and exploring the periodic behavior of the students, generating immediate alerts and strategies that increase the student's interest in remaining in the educational system. Moreover, researchers can explore the possibility of analyzing pattern variations in different subjects, both high and low, in hourly intensity. This analysis could provide a more detailed and specific understanding of how factors might influence student dropout, taking into account their academic performance.

To enhance the identification of students at risk of dropout, we proposed implementing new features that allow a complete view of the student's environment and needs. These features comprise attendance records, transportation time to the institution, and the modes of transportation utilized. In addition, socio-economic factors such as participation in school feeding programs can serve as motivating factors for economically disadvantaged students to persist in attendance, despite their academic performance falling below expectations.

5. Conclusions

In this paper, we compare five feature selection methods to determine the most relevant features or provide suitable information for training machine learning algorithms and improving their performance, emphasizing feature engineering and the use of metaheuristic algorithms, which can be used to enhance dropout detection in a real scenario of educational data obtained from the school. As demonstrated, there is a consistent occurrence where at least three feature selection techniques coincide on 10 features, which achieved the best performances compared to fully utilizing the initial features. According to this study, the features that have a strong relevance were the distance between the student's residence and the educational institution, overage, grade repetition, and academic performance. Furthermore, the wrapper or integrated methods select all relevant features; these methods may select more than the required features because the technique uses a supervised algorithm, which may overestimate the features' importance due to frequency sensitivity or high cardinality.

The selection by wrapper methods (in this case Boruta and LASSO) could be an alternative for the selection of student features, which allows the evaluation of the importance of a feature as a whole for different iterations, as opposed to a statistical analysis by correlation or Chi-square separately, which in some cases completely ignores the dependence relationships that some features may have as a whole because each feature is evaluated individually.

Feature selection functions by reducing the dataset's dimensionality and eliminating irrelevant or redundant features, a process that can significantly enhance the algorithm's precision, recall, and F1 score. In this study, methods such as Boruta and LASSO have emerged as potent tools, substantially elevating the performance of specific selected models, thereby reinforcing the efficacy of feature selection. While metaheuristic algorithms such as GA and PSO may not always achieve optimal results compared to other feature selection methods, they hold promise. Their effectiveness is tied to the complexity of these algorithms and the need for more precise adjustments, suggesting they could be the best option with adequate parameter optimization. Similarly, the mRMR method, while not always producing the best outcomes, showed improvements compared to previous results. This suggests that while mRMR is a helpful method, it may need to sufficiently capture the interactions between features across all models. Ultimately, the use of all features can yield

commendable precision performance, as evidenced by RF and the DAF set. However, it also amplifies the model's complexity and the peril of overfitting, underscoring the crucial need for meticulous feature selection and the exploration of different techniques to pinpoint the most effective combination.

Author Contributions: Conceptualization, D.Z.-M., A.E.-B. and J.A.J.-B.; Methodology, D.Z.-M., A.E.-B. and J.A.J.-B.; Software, D.Z.-M.; Validation, D.Z.-M., A.E.-B. and J.A.J.-B.; Formal analysis, D.Z.-M., A.E.-B. and J.A.J.-B.; Investigation, D.Z.-M., A.E.-B. and J.A.J.-B.; Resources, D.Z.-M., A.E.-B. and J.A.J.-B.; Data curation, D.Z.-M., A.E.-B. and J.A.J.-B.; Writing—original draft, D.Z.-M.; Writing—review & editing, D.Z.-M., A.E.-B. and J.A.J.-B.; Visualization, D.Z.-M. and J.A.J.-B.; Supervision, A.E.-B. and J.A.J.-B.; Project administration, J.A.J.-B.; Funding acquisition, D.Z.-M., A.E.-B. and J.A.J.-B.. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Data Availability Statement: The data presented in this study are available upon request from the corresponding authors. The data are not publicly available due to the privacy, sensitive, and personal nature of the information provided by the student population.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Eckert, K.B.; Suénaga, R. Analysis of attrition-retention of college students using classification technique in data mining [Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos]. *Form. Univ.* **2015**, *8*, 3–12. [CrossRef]
2. Pradeep, A.; Das, S.; Kizhekkethottam, J.J. Students dropout factor prediction using EDM techniques. In Proceedings of the Proceedings of the IEEE International Conference on Soft-Computing and Network Security, ICSNS 2015, Coimbatore, India, 25–27 February 2015. [CrossRef]
3. Márquez-Vera, C.; Romero Morales, C.; Ventura Soto, S. Predicting school failure and dropout by using data mining techniques. *Rev. Iberoam. Tecnol. Aprendiz.* **2013**, *8*, 7–14. [CrossRef]
4. Delen, D.; Topuz, K.; Eryarsoy, E. Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. *Eur. J. Oper. Res.* **2020**, *281*, 575–587. [CrossRef]
5. Kuhn, M.; Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*; CRC Press: New York, NY, USA, 2019.
6. Zapata, M.D.; Espinosa, B.A.; Jiménez, B.J. Transformación de Características Basada en Métricas para apoyar la detección del Riesgo de Deserción Escolar vía SVM. In Proceedings of the XXX Simposio Internacional de Estadística—UNAL, [Symposium Presentation], Zoom, Virtual, 21–24 September 2021; p. 12.
7. Agrawal, P.; Abutarboush, H.F.; Ganesh, T.; Mohamed, A.W. Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009–2019). *IEEE Access* **2021**, *9*, 26766–26791. [CrossRef]
8. Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Complexity* **2019**, *2019*, 1306039. [CrossRef]
9. Berens, J.; Schneider, K.; Görtz, S.; Oster, S.; Burghoff, J. Early Detection of Students at Risk—Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *J. Educ. Data Min.* **2019**, *11*, 1–41. [CrossRef]
10. Maheshwari, E.; Roy, C.; Pandey, M.; Rautray, S.S. Prediction of Factors Associated with the Dropout Rates of Primary to High School Students in India Using Data Mining Tools. *Adv. Intell. Syst. Comput.* **2020**, *1013*, 242–251. [CrossRef] [PubMed]
11. Orooji, M.; Chen, J. Predicting louisiana public high school dropout through imbalanced learning techniques. In Proceedings of the Proceedings—18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, Boca Raton, FL, USA, 16–19 December 2019; pp. 456–461. [CrossRef]
12. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Mousa Fardoun, H.; Ventura, S. Early dropout prediction using data mining: A case study with high school students. *Expert Syst.* **2016**, *33*, 107–124. [CrossRef]
13. Sansone, D. Beyond Early Warning Indicators: High School Dropout and Machine Learning. *Oxf. Bull. Econ. Stat.* **2019**, *81*, 456–485. [CrossRef]
14. Chung, J.Y.; Lee, S. Dropout early warning systems for high school students using machine learning. *Child. Youth Serv. Rev.* **2019**, *96*, 346–353. [CrossRef]
15. Şara, N.B.; Halland, R.; Igel, C.; Alstrup, S. High-school dropout prediction using machine learning: A Danish large-scale study. In Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015—Proceedings, Bruges, Belgium, 22–23 April 2015; pp. 319–324.
16. da Cunha, J.A.; Moura, E.; Analide, C. Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. *Adv. Intell. Syst. Comput.* **2016**, *445*, 189–198. [CrossRef]

17. Kiss, B.; Nagy, M.; Molontay, R.; Csabay, B. Predicting dropout using high school and first-semester academic achievement measures. In Proceedings of the ICETA 2019—17th IEEE International Conference on Emerging eLearning Technologies and Applications, Starý Smokovec, Slovakia, 21–22 November 2019; pp. 383–389. [CrossRef]
18. Pérez, A.; Grandón, E.E.; Caniupán, M.; Vargas, G. Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs. Decision Trees. In Proceedings of the International Conference of the Chilean Computer Science Society, SCCC, Santiago, Chile, 5–9 November 2018. [CrossRef]
19. Da Fonseca Silveira, R.; Holanda, M.; De Carvalho Victorino, M.; Ladeira, M. Educational data mining: Analysis of drop out of engineering majors at the UnB—Brazil. In Proceedings of the 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, Boca Raton, FL, USA, 16–19 December 2019; pp. 259–262. [CrossRef]
20. Salazar-Fernandez, J.P.; Sepúlveda, M.; Munoz-Gama, J. Influence of student diversity on educational trajectories in engineering high-failure rate courses that lead to late dropout. In Proceedings of the IEEE Global Engineering Education Conference, EDUCON, Dubai, United Arab Emirates, 8–11 April 2019; pp. 607–616. [CrossRef]
21. Dekker, G.W.; Pechenizkiy, M.; Vleeshouwers, J.M. Predicting Students Drop Out: A Case Study. In Proceedings of the International Working Group on 712 Educational Data Mining, Cordoba, Spain, 1–3 July 2009.
22. Martins, M.P.G.; Migueis, V.L.; Fonseca, D.S.B.; Gouveia, P.D.F. Prediction of academic dropout in a higher education institution using data mining [Previsão do abandono académico numa instituição de ensino superior com recurso a data mining]. *RISTI—Rev. Iber. Sist. E Tecnol. Inf.* **2020**, *2020*, 188–203.
23. Peralta, B.; Poblete, T.; Caro, L. Automatic feature selection for desertion and graduation prediction: A chilean case. In Proceedings of the International Conference of the Chilean Computer Science Society, SCCC, Valparaiso, Chile, 10–14 October 2016. [CrossRef]
24. Wan Yaacob, W.F.; Mohd Sobri, N.; Nasir, S.A.M.; Wan Yaacob, W.F.; Norshahidi, N.D.; Wan Husin, W.Z. Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques. *J. Phys. Conf. Ser.* **2020**, *1496*, 012005. [CrossRef]
25. Sari, E.Y.; Sunyoto, A. Optimization of weight backpropagation with particle swarm optimization for student dropout prediction. In Proceedings of the 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019, Yogyakarta, Indonesia, 20–21 November 2019; pp. 423–428. [CrossRef]
26. Shiratori, N. Derivation of Student Patterns in a Preliminary Dropout State and Identification of Measures for Reducing Student Dropouts. In Proceedings of the 2018 7th International Congress on Advanced Applied Informatics, IIAI-AAI 2018, Yonago, Japan, 8–13 July 2018; pp. 497–500. [CrossRef]
27. Barros, T.M.; Silva, I.; Guedes, L.A. Determination of dropout student profile based on correspondence analysis technique. *IEEE Lat. Am. Trans.* **2019**, *17*, 1517–1523. [CrossRef]
28. Bedregal-Alpaca, N.; Cornejo-Aparicio, V.; Zarate-Valderrama, J.; Yanque-Churo, P. Classification models for determining types of academic risk and predicting dropout in university students. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 266–272. [CrossRef]
29. Cuji Chacha, B.R.; Gavilanes López, W.L.; Vicente Guerrero, V.X.; Villacis Villacis, W.G. Student Dropout Model Based on Logistic Regression. *Commun. Comput. Inf. Sci.* **2020**, *1194*, 321–333. [CrossRef] [PubMed]
30. Nuankaew, P. Dropout situation of business computer students, University of Phayao. *Int. J. Emerg. Technol. Learn.* **2019**, *14*, 117–131. [CrossRef]
31. Aulck, L.; Nambi, D.; Velagapudi, N.; Blumenstock, J.; West, J. Mining university registrar records to predict first-year undergraduate attrition. In Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019), Montreal, QC, Canada, 2–5 July 2019.
32. Manrique, R.; Nunes, B.P.; Marino, O.; Casanova, M.A.; Nurmikko-Fuller, T. An analysis of student representation, representative features and classification algorithms to predict degree dropout. In Proceedings of the 9th International Learning Analytics & Knowledge Conference, Tempe, AZ, USA, 4–8 March 2019. [CrossRef]
33. Medina, D.Z.; Builes, J.A.J.; Bedoya, A.E. Automatic detection of students at risk of dropping out of school using mRMR and Late Fusion. In Proceedings of the 2022 XII International Conference on Virtual Campus (JICV), Arequipa, Peru, 29–30 September 2022; pp. 1–4. [CrossRef]
34. Fernández, J.; Rojas, A.; Daza, G.; Gómez, D.; Álvarez, A.; Orozco, Á. Student desertion prediction using kernel relevance analysis. In *Progress in Artificial Intelligence and Pattern Recognition*; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2018; Volume 11047, pp. 263–270. [CrossRef]
35. Aguilar-Gonzalez, S.; Palafox, L. Prediction of Student Attrition Using Machine Learning. In *Advances in Soft Computing*; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2019; Volume 11835, pp. 212–222. [CrossRef]
36. Hutagaol, N.; Suharjito. Predictive modelling of student dropout using ensemble classifier method in higher education. *Adv. Sci. Technol. Eng. Syst.* **2019**, *4*, 206–211. [CrossRef]
37. Dharmawan, T.; Ginardi, H.; Munif, A. Dropout Detection Using Non-Academic Data. In Proceedings of the 2018 4th International Conference on Science and Technology, ICST 2018, Yogyakarta, Indonesia, 7–8 August 2018. [CrossRef]
38. Ministerio de Educación Nacional—República de Colombia. Sistema Nacional de Indicadores Educativos para los Niveles de Preescolar, Básica y Media en Colombia. Available online: https://www.mineducacion.gov.co/1759/articles-363305_recurso_1.pdf (accessed on 11 January 2024).

39. Hegde, V. Dimensionality reduction technique for developing undergraduate student dropout model using principal component analysis through R package. In Proceedings of the 2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016, Chennai, India, 15–17 December 2016. [CrossRef]
40. Hegde, V.; Prageeth, P.P. Higher education student dropout prediction and analysis through educational data mining. In Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018, Coimbatore, India, 19–20 January 2018; pp. 694–699. [CrossRef]
41. Fenton, N.; Bieman, J. *Software Metrics: A Rigorous and Practical Approach*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2014; p. 618.
42. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. In Proceedings of the 2003 IEEE Bioinformatics Conference, CSB2003, Stanford, CA, USA, 11–14 August 2003; pp. 523–528. [CrossRef]
43. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
44. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
45. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; Volume 26.
46. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef] [PubMed]
47. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95—International Conference on Neural Networks, Perth, Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948. [CrossRef]
48. Ahmed, S.A.; Khan, S.I. A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019, Kanpur, India, 6–8 July 2019. [CrossRef]
49. De Santos, K.J.O.; Menezes, A.G.; De Carvalho, A.B.; Montesco, C.A.E. Supervised learning in the context of educational data mining to avoid university students dropout. In Proceedings of the IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019, Maceio, Brazil, 15–18 July 2019; pp. 207–208. [CrossRef]
50. Meedech, P.; Iam-On, N.; Boongoen, T. Prediction of Student Dropout Using Personal Profile and Data Mining Approach. In *Proceedings of the Intelligent and Evolutionary Systems*; Lavangnananda, K., Phon-Amnuaisuk, S., Engchuan, W., Chan, J.H., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 143–155. [CrossRef]
51. Ahmad Tarmizi, S.S.; Mutalib, S.; Abdul Hamid, N.H.; Abdul-Rahman, S.; Md Ab Malik, A. A Case Study on Student Attrition Prediction in Higher Education Using Data Mining Techniques. *Commun. Comput. Inf. Sci.* **2019**, *1100*, 181–192. [CrossRef] [PubMed]
52. Da Silva, P.M.; Lima, M.N.; Soares, W.L.; Silva, I.R.; De Fagundes, R.A.; De Souza, F.F. Ensemble regression models applied to dropout in higher education. In Proceedings of the 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019, Salvador, Brazil, 15–18 October 2019; pp. 120–125. [CrossRef]
53. Serra, A.; Perchinunno, P.; Bilancia, M. Predicting student dropouts in higher education using supervised classification algorithms. In *Computational Science and Its Applications—ICCSA 2018*; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2018; Volume 10962, pp. 18–33. [CrossRef]
54. Sangodiah, A.; Beleya, P.; Muniandy, M.; Heng, L.E.; Ramendran Spr, C. Minimizing student attrition in higher learning institutions in Malaysia using support vector machine. *J. Theor. Appl. Inf. Technol.* **2015**, *71*, 377–385.
55. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
56. Lix, L.M.; Keselman, J.C.; Keselman, H.J. Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *Rev. Educ. Res.* **1996**, *66*, 579–619. [CrossRef]
57. Cover, T.M. The Best Two Independent Measurements Are Not the Two Best. *IEEE Trans. Syst. Man Cybern.* **1974**, *SMC-4*, 116–117. [CrossRef]
58. Lima, J.; Alves, P.; Pereira, M.; Almeida, S. Using academic analytics to predict dropout risk in engineering courses. In Proceedings of the European Conference on e-Learning, ECEL, Athens, Greece, 1–2 November 2018; pp. 316–321.