sustainability

MDPI

Article

Fuzzy Artificial Intelligence—Based Model Proposal to Forecast Student Performance and Retention Risk in Engineering Education: An Alternative for Handling with Small Data

Adriano Bressane [1,*], Marianne Spalding [1], Daniel Zwirn [1], Anna Isabel Silva Loureiro [2], Abayomi Oluwatobiloba Bankole [2], Rogério Galante Negri [1], Irineu de Brito Junior [1], Jorge Kennety Silva Formiga [1], Liliam César de Castro Medeiros [1], Luana Albertani Pampuch Bortolozo [1] and Rodrigo Moruzzi [1]

[1] Environmental Engineering Department, Institute of Science and Technology, São Paulo State University, São José dos Campos 12245-000, Brazil
[2] Civil and Environmental Engineering Graduate Program, Faculty of Engineering, São Paulo State University, Bauru 17033-360, Brazil
* Correspondence: adriano.bressane@unesp.br

Abstract: Understanding the key factors that play an important role in students' performance can assist improvements in the teaching-learning process. As an alternative, artificial intelligence (AI) methods have enormous potential, facilitating a new trend in education. Despite the advances, there is an open debate on the most suitable model for machine learning applied to forecast student performance patterns. This paper addresses this gap, where a comparative analysis between AI methods was performed. As a research hypothesis, a fuzzy inference system (FIS) should provide the best accuracy in this forecast task, due to its ability to deal with uncertainties. To do so, this paper introduces a model proposal based on AI using a FIS. An online survey was carried to collect data. Filling out a self-report, respondents declare how often they use some learning strategies. In addition, we also used historical records of students' grades and retention from the last 5 years before the COVID pandemic. Firstly, two experimental groups were composed of students with failing and passing grades, compared by the Mann-Whitney test. Secondly, an association between the 'frequency of using learning strategies' and 'occurrence of failing grades' was quantified using a logistic regression model. Then, a discriminant analysis was performed to build an Index of Student Performance Expectation (SPE). Considering the learning strategies with greater discriminating power, the fuzzy AI-based model was built using the database of historical records. The learning strategies with the most significant effect on students' performance were lesson review (34.6%), bibliography reading (25.6%), class attendance (23.5%), and emotion control (16.3%). The fuzzy AI-based model proposal outperformed other AI methods, achieving 94.0% accuracy during training and a generalization capacity of 91.9% over the testing dataset. As a practical implication, the SPE index can be applied as a tool to support students' planning in relation to the use of learning strategies. In turn, the AI model based on fuzzy can assist professors in identifying students at higher risk of retention, enabling preventive interventions.

Keywords: engineering education; artificial intelligence; students' performance

check for updates

Citation: Bressane, A.; Spalding, M.; Zwirn, D.; Loureiro, A.I.S.; Bankole, A.O.; Negri, R.G.; de Brito Junior, I.; Formiga, J.K.S.; Medeiros, L.C.d.C.; Pampuch Bortolozo, L.A.; et al. Fuzzy Artificial Intelligence—Based Model Proposal to Forecast Student Performance and Retention Risk in Engineering Education: An Alternative for Handling with Small Data. *Sustainability* **2022**, *14*, 14071. https://doi.org/10.3390/su142114071

Academic Editor: Firoz Alam

Received: 15 September 2022
Accepted: 25 October 2022
Published: 28 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The high level of retention is an undesirable fact, but very common in engineering, with consequences on dropouts, which are among the highest of all disciplines of knowledge. When the student does not give up, the engineering degree lasts for many years, resulting in higher costs for both the educational institutions and the students' family. According to [1], STEM students (Science, Technology, Engineering and Mathematics) are especially vulnerable in the initial years of their academic programs with more than 60%

of the dropouts occurring within the first two years. Therefore, early identification of at-risk students is crucial for a focused intervention if institutions are to support students towards completion.

Although the causes of student retention and evasion are multiple and complex, performance is one of the factors often affected [2,3]. In turn, academic performance of students has been correlated with the use of learning strategies [4]. Thereby, advances in knowledge about the most effective strategies can assist faculty and students, to improve academic performance and facilitate the successful completion of their studies.

Although there are many ways to understand the effects of different learning strategies on students' performance, the most common approaches are those which distinguish the strategies between cognitive and metacognitive [5]. Cognitive strategies are a set of behaviors that influence the learning process allowing information to be stored more efficiently, and metacognitive strategies are procedures used by the individual for planning, monitoring and regulating their own learning [6].

As verified by Séllei et al. [7], personality traits, emotional intelligence, and engagement, among other psychological factors, also affect university performance, illuminating the importance of stress management for students. Therefore, learning strategies must be considered not only as a set of cognitive abilities but also as behavioral skills and self-regulatory (emotional) control used by the learners to control their own psychological learning processes [8]. On the other hand, there are few validated instruments and standardized procedures for comparatively evaluating the effects of different learning strategies [9–11].

The ability to predict student performance creates opportunities to improve educational outcomes and avoid dropouts [12,13]. Therefore, the interest in establishing measures that identify the possibility of student retention has been the focus of some studies in recent years [14–16].

Over the last years, AI-based approaches have increasingly provided predictive modeling methods. This allows educators to forecast student-learning performance and retention risk, facilitating a new trend in education [17,18]. Although, this approach is quite new and relatively unknown to many researchers and educators, it has enormous potential in education [12,19].

Despite the wide variety of methods already investigated, there are gaps to which this study sought to contribute:

i. a comparative analysis of different methods of AI, widely recognized for their performance in forecast tasks;
ii. a search in the main databases did not return studies focused on AI methods using fuzzy inference systems (FIS) to forecast retention risk in engineering education;
iii. a large sample is often needed for the machine learning process, as a small dataset may be insufficient to build an accurate predictive model [20].

To address these issues, this paper aims at introducing a fuzzy AI-based model to answer the following research questions: (i) what are the impact of different learning strategies on the students' performance? (ii) can a small dataset be sufficient to build an accurate predictive model using a fuzzy inference system?

The hypothesis is that fuzzy inference system should provide the best accuracy, due to its ability to deal with uncertainties, which tends to increase as the amount of information decreases (small dataset size). The theory of fuzzy sets was introduced by Lotfi Zadeh, to deal with the uncertainties that arise in complex systems [21]. The inference systems based on fuzzy logic operate with non-linear functions, limits of a gradual transition between intervals and degrees of certainty (μ), which admit partial membership in more than one set of values [22].

The remainder of this text is organized as follows: Section 2 presents a brief literature review; Section 3 lists methods and materials; Section 4 offers paper results and discussions, and the concluding remarks are in Section 5.

2. Related Works

A search in the main databases, including Science Direct, IEEE Computer Society, Springer, and MDPI Journals, return few studies focused on AI methods using fuzzy inference systems (FIS) to forecast retention risk in engineering education. This search was carried out in September 2022, applying the keywords "Predicting students' performance", "Predicting algorithm students", "Machine learning prediction students", "Algorithms analytics students", and "Students analytics prediction performance".

Jayagowri and Karpagavalli [23] developed a Learning Fuzzy Cognitive Map (LFCM) to predict the academic performance of students in an online master's program. The authors used the LFCM technique to build a causal map demonstrating the links between the factors assessed and how these links influence student performance. Petra and Aziz [24] proposed a fuzzy expert system to investigate and analyze the performance of civil engineering students at a university in Malaysia, considering the students' cognitive, affective and psychomotor characteristics.

Mansouri et al. [25] proposed an approach to predicting student performance using the LFCM approach considering a longitudinal quantitative and qualitative dataset of students in an online program for three consecutive semesters, intending to predict every individual student's performance for the fourth semester. Nosseir and Fath [26] developed a mobile application for prediction of student performance combining Fuzzy Logic and Artificial Neural Networks. Among other AI methods that have been investigated to develop accurate models to forecast student retention are Decision Tree, Artificial Neural Networks Naive Bayesian, Support Vector Machine, and Genetic Algorithm [18,27–36].

Ahmed et al. [37] showed that peer mentoring and coaching are activities with the potential to avoid dropping out by first-generation biomedical engineering students. Huerta-Manzanilla et al. [38] proposed a metric to quantify the retention risk for students of engineering programs. Using the proposed metric, the authors predicted the dropout risk through a logistic regression model.

Similarly, Bello et al. [39] adopted machine learning methods to identify significant variables useful to predict dropout by engineering students. In summary, the Random Forest method was first employed to select the most relevant feature, consequently defining an effective decision tree-based predictor.

Ortiz-Lozano et al. [40] built prediction models using information collected at three different moments throughout the students' first university year. Beyond highlighting the academic performance data as a good predictor for dropout risk, the author also reinforces the need for an early first-year intervention to prevent non-completion.

Alvarez et al. [41] employed decision trees and neural networks combined with distinct features to predict the promotion, repetition, and dropout cases of Computer Engineering students. In a correlated study, Alvarez et al. [42] concluded that variables such as the province of origin, the grades obtained for the entrance examination in Mathematics, and academic performance in Mathematics and Programming, are influential factors when predicting student dropout in the first year of Computer Engineering.

Hassan et al. [43] developed a students' performance prediction model by identifying students' features and suitable ensemble learning techniques. The result shows that behavior-related features and the majority voting strategy outperform other ensemble methods.

Acero et al. [44] adopted data mining techniques to analyze the possible causes behind the high dropout rates of Colombian engineering students. The results show that the number of subjects viewed and the semester the student entered significantly impact the dropout probability. Lastly, Ravikumar et al. [45] proposed an early-alert system to help identify undergraduate students under retention risk, followed by appropriate peer-tutoring and academic support.

3. Material and Method

A case study was developed to introduce and validate our fuzzy AI-based proposal. We consider the Differential and Integral Calculus 1, which is a mandatory subject for all

engineering students and has historically been responsible for high rates of failure and evasion in many degree courses [46], as well as being "*extremely efficient at lowering student confidence*" [47].

An online survey was conducted for data collection. As an inclusion criterion, we consider respondents who attended the subject from 2015 to 2019 (prior to the Covid-19 pandemic), totaling a population of 136 students. From this population, the sample size of 111 respondents provided a test power $(1 - \beta)$ of 0.80 for a significance level (α) of 0.05 and a minimum detectable effect size (ρ) of 0.5. Regarding the interviewed students profiling, 63.9% came from a private high school, 59.5% are female, 33.3% are under 19 years old, and 41.4% correspond to failing grade students.

It is worth noting that this database covers a restricted number of observations (small data). However, as an exploratory study, the present study focused on verifying the potential of the proposed approach (fuzzy AI), to handle conditions where big data is not available.

Filling out a self-report, respondents were inquired on how often they used the following learning strategies, based on the studies of Warr and Downing [8] and Martins and Zerbini [10]: cognitive (lessons review, bibliography reading, and exercise solving), self-regulatory (contextualized motivation and emotion control), and behavioral (help-seeking and class attendance). A discrete five-point scale was used as an answer option for each question: very small (SS), small (S), medium (M), large (L), and very large (LL). In addition, the final grades and class attendance of students were collected in the institution's database (Figure 1).

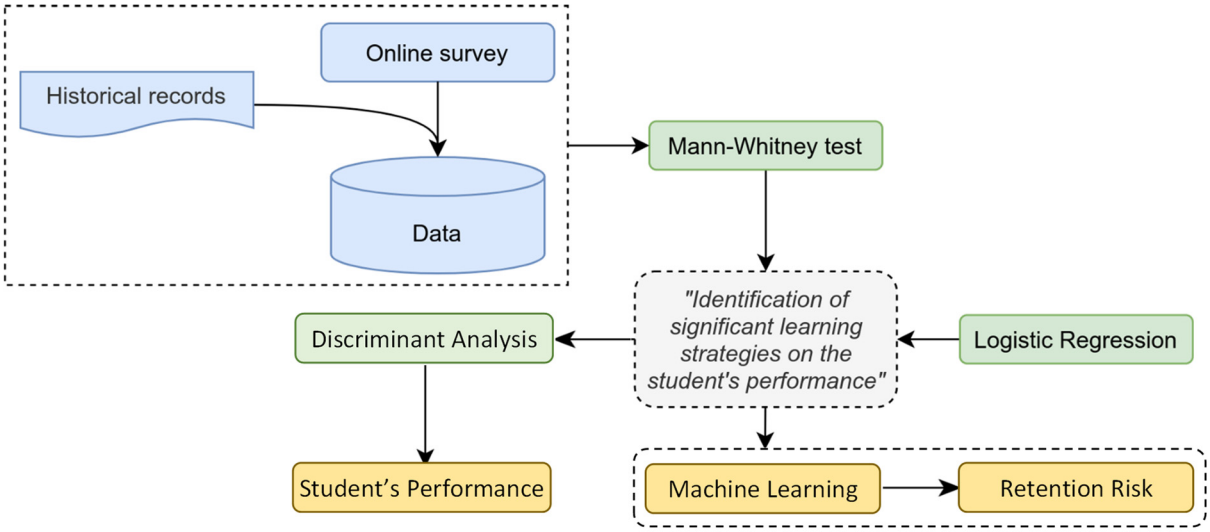


Figure 1. The data analysis process workflow.

Firstly, to analyze differences in usage of learning strategies by the students, two experimental groups were composed: students with a failing grade (<5), and the ones with a passing grade ($\geq$5). Then, a comparative assessment between these groups was performed by means of the Mann-Whitney test [48]. Moreover, we performed a logistic regression as a second analysis. This aimed at quantifying the association between the events: (e_1) frequency of using learning strategies; and (e_2) occurrence of failing grade (student's retention). If the odds ratio (OR) is equal to 1, then the events are independent. On the other hand, if the regression is significant and the OR is less than 1, then the events are negatively associated, meaning the presence of event e_1 reduces the chances of event e_2 [49]. These two initial combined analyzes allowed us to identify what are the learning strategies with significant effects on student performance/retention.

Secondly, the learning strategies with significant effects were subjected to a quadratic discriminant analysis [50]. This analysis aimed to rank learning strategies according to

their importance, i.e., their discriminating power to differentiate students with different levels of performance: low (grade < 5), medium (5 ≤ grade < 7), and high (grade ≥ 7). This analysis generates standardized canonical functions, based on weighted combinations of learning strategies. The function with the greatest discriminating power was adopted as an Index of Student Performance Expectation (SPE). All the statistical analyzes considered a significance level (α) of 0.05, after verifying parametric assumptions for determining the applicable test. Finally, an AI-based model capable of recognizing patterns in historical databases was built using machine learning to predict the probability of retention [51]. The database for this study was composed of responses collected by the online survey, and historical records of student retention in the last 5 years before the COVID-19 pandemic (2015 to 2019).

Taking into account the uncertainties associated with data based on students' perceptions, we used an artificial intelligence method specially developed for modeling Fuzzy Inference Systems (FIS), the Wang & Mendel algorithm (WM). This algorithm was made available by Guillaume et al. [52] in the package 'FisPro', for the R programming language. The machine learning process behind the FIS method encompasses three main steps. In the first (fuzzification step), the range of input variables is partitioned using membership functions.

Several partition techniques (regular, k-means, hierarchical fuzzy partitioning—hfp) and membership function types (triangular, trapezoidal, gaussian, other) can be used in the fuzzification step. In the second step (induction), the rule base is built. This step includes defining the rule induction technique (FPA, OLS, TREE, WM), logical conjunction (minimum, product, Lukasiewicz), and disjunction (sum, max) operators, which results in relational propositions according the following Theorem 1:

Let x_i be an input variable (*learning strategy*), y_j be an output variable (*retention risk*), a_{ij} be a linguistic value belonging to {*SS*, *S*, *M*, *L*, *LL*} of the x_i, and c_{ij} be a linguistic value of the y_i belonging to {*low*, *high*}, with $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$.

$$\text{If } x_1 \text{ is } a_{1j} \text{ and } x_2 \text{ is } a_{2j} \text{ and } \ldots \text{ and } x_n \text{ is } a_{mj}, \text{ then } y_j = c_{ij}.$$

Fuzzy logic operators play a key role in the inference process, allowing the model to deal with partial degrees of truth (uncertainties) that may be, for example, associated with missing information or small data [51].

The main fuzzy logic operators include (Figure 2): minimum, given by $O_{min}(a, b) = \min\{a, b\} = a \wedge b$, which results in the lowest membership among those evaluated; product, given by $O_{prod}(a, b) = a.b$, which results from the multiplication of the assessed membership; Lukasiewicz, given by $O_{luk}(a, b) = \{0, a + b - 1\}$, which results from an arithmetic operation between the assessed membership; and maximum, given by $O_{max}(a, b) = \max\{a, b\} = a \vee b$, which results in the highest membership among those evaluated [51].

In the third step, the fuzzy output can be converted into a crisp output using defuzzification techniques (maximum crisp, Sugeno). In addition to the uncertainty associated with data based on students' perceptions (subjectivity), the small amount of data (n = 111) also represents an important source of uncertainty.

It is worth noting that big data is generally not available in the case of new courses, with few classes already completed. In this scenario, a method capable of dealing with uncertainties, such as a fuzzy inference system (FIS), can be a promising alternative.

To verify this research hypothesis, the performance provided by the FIS was compared with distinct forecast algorithms: cascade-correlation network (CCN), multilayer perceptron network (MLP), polynomial neural network (GMDH), probabilistic neural network (PNN), radial basis function network (RBFN), gene expression programming (GEP), decision tree forest (Tree Boost), and support vector machine (SVM). All experiments were performed using the *R* software [53], an open-source simulation tool.

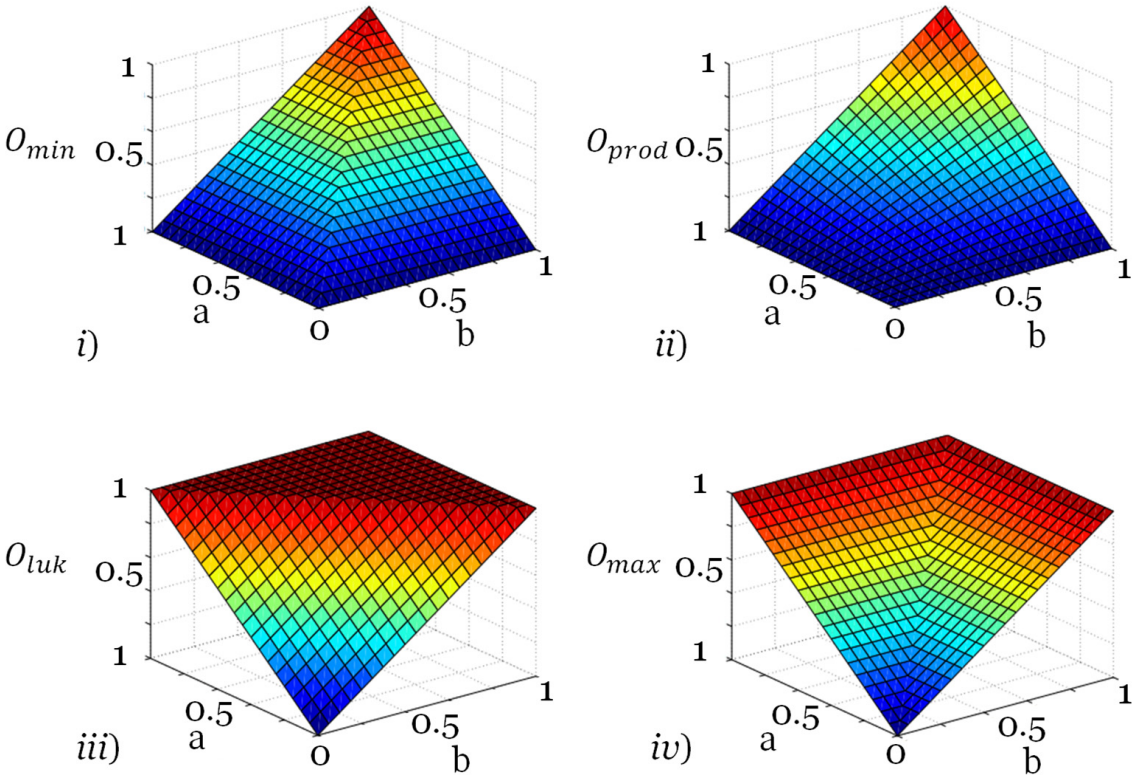


Figure 2. Fuzzy logic operators: (**i**) minimum, (**ii**) product, (**iii**) Lukasiewicz, and (**iv**) maximum.

The overall accuracy (θ) was employed to measure the performance of each analyzed algorithm and compare them. The parameters were optimized using a grid search procedure, and selecting the hyperparameters configuration that achieved the best performance on the test data, using 3-fold cross-validation, with one third of the data used for training and the rest for testing.

4. Results and Discussion

Concerning the student's profiling factors, differences in the final grade were statistically non-significant (Table 1). Both gender (p-value = 0.921) and age (p-value = 0.239) groups achieved approximately equal averages in the final grades. It is also worth highlighting that there was no statistical difference in the performance between students from public and private schools (p-value = 0.426). In other words, the final grade of students who came from private high schools is not statistically superior, when compared to the students who came from public schools. This finding indicates that students who came from public schools do not tend to perform worse. In Brazil, the average income of students in public schools is one-third of the average income of students in private schools [54]. Thereby, this finding is in contrast to what is expected according to Tinto [55], who stated that lower-income students are academically underprepared and their retention rates are higher when compared to those from the private school.

Analyzing Table 2, the learning strategies with the most significant correlation with student retention risk were lessons review (p-value < 0 .001) and class attendance (p-value = 0.002). Regarding the cognitive strategies, the only activity that showed no difference, between the groups of failing and passing grade students, was the exercise solving activity (p-value = 0.616). Nevertheless, this does not mean that it is a less important learning activity; it just indicates that both groups of students were engaged in such cognitive strategy.

Table 1. Effects of student's profiling factors on final grade.

Profiling Factors		Final Grade					
		obs.	**min.**	**max.**	**mean**	**sd.**	***p***
gender	female	66	1.10	9.80	5.23	2.06	0.921
	male	45	0.40	10.00	5.13	2.47	
high school	private	30	0.40	9.80	5.26	2.22	0.426
	public	41	1.10	10.00	5.06	2.27	
age group	young (<19)	37	0.90	10.00	4.84	2.30	0.239
	young adult (≥19)	74	0.40	9.80	5.36	2.18	

obs.: observations; min.: minimum; max.: maximum; sd.: standard deviation; *p*: *p*-value.

Table 2. Spearman's [1] correlation between learning strategies and retention risk.

		Cognitive			Self-Regulatory		Behavioral	
		ES	**LR**	**BR**	**CM**	**EC**	**HS**	**CA**
retention risk	ρ	0.05	0.34 ***	0.24 *	0.08	0.26 **	0.25 **	0.29 **
	p	0.614	<0.001	0.010	0.390	0.005	0.008	0.002

ES: exercise solving; LR: lessons review; BR: bibliography reading; CM: contextualization motivation; EC: emotion control; HS: help seeking; CA: class attendance; * statistically significant; ** significant; *** very significant; *p*: *p*-value; ρ: Spearman's rho. [1] Spearman rank-order correlation coefficient is a nonparametric measure of association, used when at least one of the variables is categorical ordinal.

Bibliography reading (*p*-value = 0.011) and lesson review (*p*-value < 0.001) showed a statistically significant difference. Especially the last one cognitive strategy, which had a greater frequency of use by the passing grade students. This finding reveals strong evidence about the importance of these strategies to improve students' performance and for preventing their retention. Martins and Zerbini [10] also found similar results for hybrid courses when the students use the bibliography and support material to organize their studies and to solve exercises.

Among self-regulatory strategies, contextualized motivation (*p*-value = 0.388) showed no significant difference in the frequency of use by failed and passed students. Contextualized motivation is a teaching-learning strategy of recognized importance, especially in the context of active methodologies. On the other hand, while contextualized motivation brings evident contribution in professionalizing and specific subjects, in the case of basic subjects, as is the case of Differential and Integral Calculus 1, this importance was not so evident.

In turn, the emotion control was extremely statistically significant (*p*-value = 0.006), as a self-regulatory strategy. Emotion and motivation control are increasingly recognized as soft skills. In the workplace, leaders and head-hunters seek professionals capable of working under pressure and keeping calm. In line with Warr and Downing [8] and Jakubowski and Dembo [56], our results reveal that emotional control is also a key issue to improve students' performance, especially the highly anxious ones.

Class attendance (*p*-value = 0.002) and help-seeking (*p*-value = 0.008), also proved to be quite significant, as behavioral strategies capable of affecting students' performance. Despite the quick access to a large amount of information available on digital platforms, which provides greater autonomy to students. Our results reveal evidence that class attendance remains an extremely significant factor in students' performance.

Help-seeking, which also relates to soft skills, including a willingness to face difficulties through social interaction and group work, has also been revealed as a behavioral strategy with positive effects on student performance. As verified by Tseng et al. [57], our findings indicate a positive association between academic performance and social skills, reinforcing results from previous studies on the role of social skills in the learning process [58,59].

Aligned with the aforementioned results, the logistic regression model also verified the effects with statistical significance. As shown in Table 3, variations in the frequency of use of learning strategies have significant effects on the likelihood of retention (failing grade). It is interesting to note that the students with failing grades seek more help, when compared to those who pass in the subject, as indicated by the odds ratio greater than 1 (OR = 1.42).

Table 3. Effects of using learning strategies on likelihood of retention based on logistic regression.

Learning Strategies	*p*-Value		OR	Likelihood of Retention	
	(χ^2)	(β)		From *	To **
exercise solving	0.665	0.666	Δ	Δ	Δ
bibliography reading	0.003	0.005	0.553	26.9%	79.7%
lessons review	0.001	0.001	0.504	24.7%	83.6%
contextualized motivation	0.319	0.322	Δ	Δ	Δ
emotion control	0.007	0.010	0.639	21.9%	62.7%
class attendance	0.003	0.006	<0.001	0.06%	31.2%
help seeking	0.005	0.006	1.424	59.6%	26.4%

χ^2. Goodness-of-fit of data to the model; β. Significance of the regression coefficient; Δ. No significance; * higher/** lower: frequency of use of the learning strategies.

Only exercise solving and contextualized motivation did not have goodness-of-fit of data to the logistic model (χ^2 not significant, *p*-value > 0.05). Regarding bibliography reading, the likelihood of retention increases from 26.9% to 79.7%, when the frequency of using this strategy is low (OR = 0.5532 and *p*-value = 0.0052). In the case of lesson review, this likelihood increases from 24.7% to 83.6%. In turn, it increases from 21.9% to 62.7% and from 0.06% to 31.2%, for low frequency of emotion control and class attendance, respectively (Figure 3).

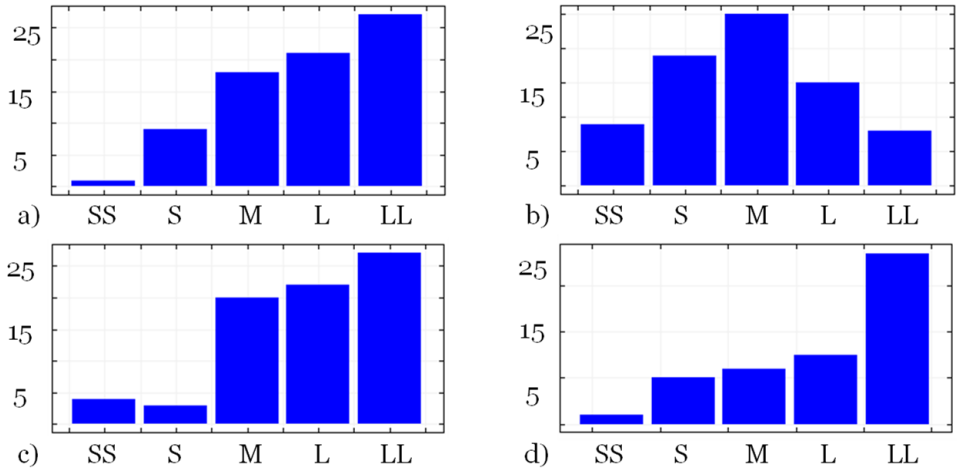


Figure 3. Frequency of using the most significant learning strategies: (**a**) lesson review, (**b**) emotion control, (**c**) bibliography reading, and (**d**) class attendance; SS: very small; S: small, M: medium, L: large, and LL very large.

As summarized in Figure 1, from the discriminant analysis we found that the strategies with the most significant effect on student performance and retention risk were, from the most to least important, lesson review (34.6%), bibliography reading (25.6%), class attendance (23.5%), and emotion control (16.3%). It is worth mentioning that such percentages associated to the factors (x_i), correspond to the amount of information capable of explaining students' performance and retention, which each strategy provides.

From this outcome, the Index of Student Performance Expectation (SPE) was built, equivalent to the first canonical function ($f_1(x)$), available for download at https://l1nq.

com/SPEindex, accessed on 24 October 2022. The SPE can contribute to the student's awareness of his own deficiencies and, from there, support the adequacy of the strategies used for learning (Figure 4).

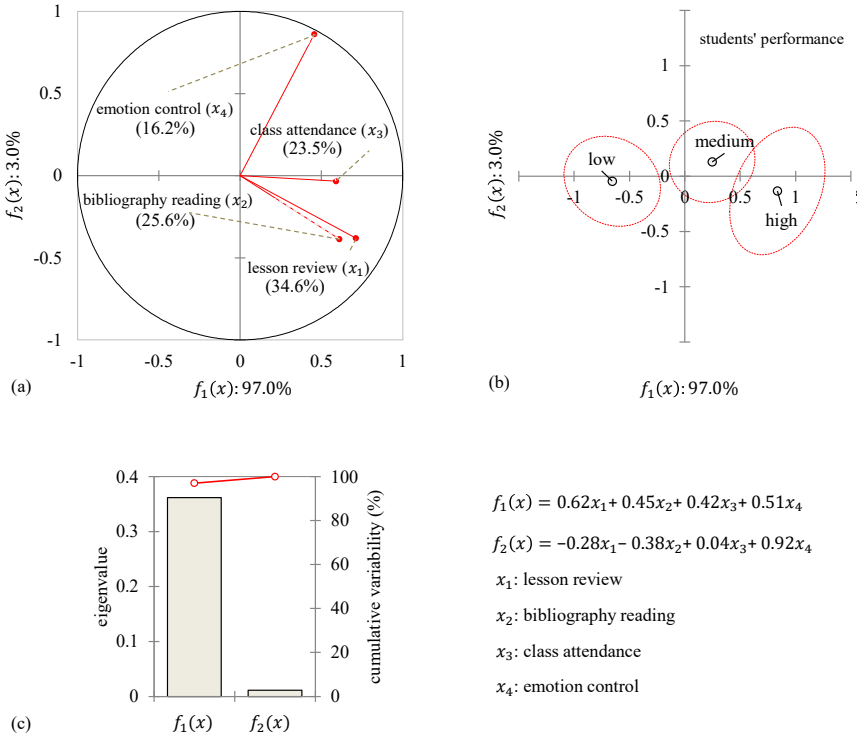


Figure 4. Importance of each learning strategy for student performance and risk retention: (**a**) scoring plot, (**b**) loading plot, and (**c**) importance explained.

Based on the knowledge provided by these prior analyses, the AI-based models were built considering the learning strategies identified as statistically significant (lesson review, bibliography reading, class attendance, and emotion control). The accuracy in forecasting the risk of retention in engineering education provided by each AI method is presented in Table 4.

As previously indicated, the uncertainty tends to increase as the amount of information decreases, so recognizing patterns in small data can be a hard challenge for the most AI methods, impacting its accuracy on test data. Due to its capability of dealing with partial degrees of certainty (μ), the FIS outperformed other AI methods, confirming the research hypothesis. For this, such degrees of certainty were modeled using the membership functions shown in Figure 5.

Table 4. Overall accuracy in forecasting the risk of retention.

Machine Learning Algorithm	Parameterization Setting	Accuracy (%)	
		Train	Test
Cascade-correlation network (CCN)	kernel: sigmoid, gaussian; neurons: [0, 10^3]; candidates: 10^2; epoch: 10^3; overfitting control: cross validation	89.91	63.3
Decision Tree Forest (Tree Boost)	number of trees: 300, min. size node to split: 10; depth: 8; prune to min. error; min. trees in series: 10	86.24	70.64
Fuzzy Inference System (FIS)	type of model: WM; type of function: triangular; terms: 5, t-norm: prod; disjunction operator: maximum	94.02	91.92

Table 4. *Cont.*

Machine Learning Algorithm	Parameterization Setting	Accuracy (%)	
		Train	Test
Gene expression programming (GEP)	population: 50; max. tries: 10^4; genes: 4; gene head length: 8; maximum: 2000; gen. without improvement: 10^3	74.31	65.14
Multilayer perceptron network (MLP)	number of layers: 03; type of function: logistic; train: scaled conjugate gradient; overfitting control: min. error over test	67.89	66.97
Polynomial neural network (GMDH)	maximum network layers: 20; max. polynomial order: 16; number of neuros per layer: 20	76.15	62.39
Probabilistic neural network (PNN)	type of kernel function: gaussian; steps: 20; sigma: each var. [10^{-4}, 10]; prior probability: frequency distribution	86.24	69.72
Radial basis function network (RBFN)	max. neurons: 10^3; radius: [10^{-2}, 10^3]; population size: 200; maximum generations: 20; max. gen. flat: 5	73.39	58.72
Support vector machine (SVM)	type: C-CSV; stopping criteria: 10^{-3}; kernel function: RBF; optimize: minimize total error	70.64	66.97

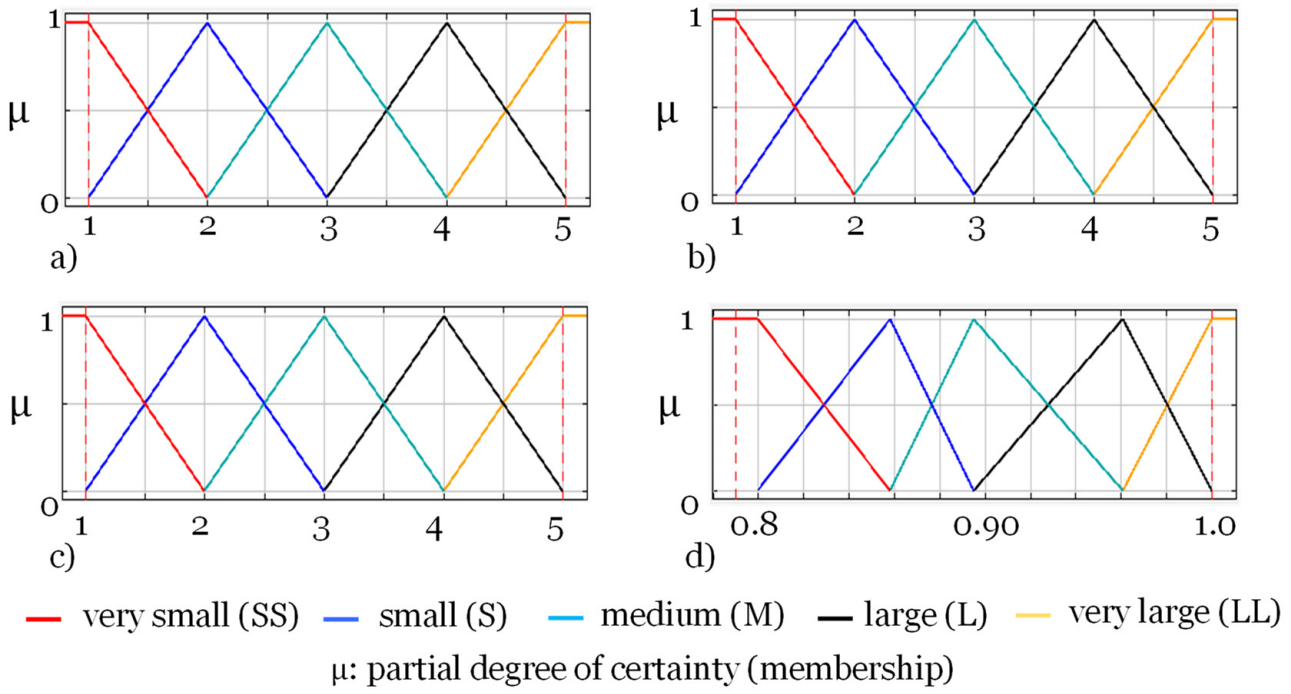


Figure 5. Fuzzification of the predictor variables in the input space: (**a**) lesson review, (**b**) emotion control, (**c**) bibliography reading, and (**d**) class attendance.

By building the FIS using the WM rule induction technique, the fuzzification of the linguistic values of each variable was performed exclusively with triangular membership functions. Considering the values and ranges shown in Figure 5, some examples of rules generated by AI-model during machine learning are as follows (Figure 6):

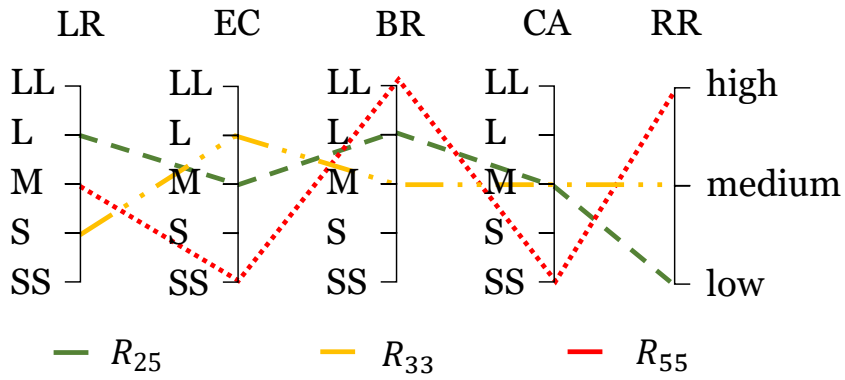


Figure 6. Parallel coordinates plot of the rules (R_{25}, R_{33} and R_{55}): LR—lesson review; EC—emotion control; BR—bibliography reading; CA—class attendance; RR—retention risk; SS—very small; S—small; M—medium; L—large; and LL—very large.

R_{25}: if lesson review is *L* and emotion control is *M* and *bibliography reading* is *L* and *class attendance* is *M* then retention risk is low.

R_{33}: if lesson review is *S* and emotion control is *L* and *bibliography reading* is *M* and *class attendance* is *M* then retention risk is medium.

R_{55}: if lesson review is *M* and emotion control is *SS* and *bibliography reading* is *LL* and *class attendance* is *SS* then retention risk is high.

In comparison with other partitioning methods (regular and hfp), k-means procedure provided the best performance during training. By considering the data distribution, k-means was able to obtain stronger partitioning, that is, with greater discriminative power between retention risk groups.

For the product conjunction operator, five antecedent terms were enough (SS, S, M, L, and LL) for the model to reach the best performance over the training dataset (94.02%). Whilst the minimum operator considers only the lowest membership in the conjunction of antecedents, the product t-norm takes into account all membership values.

Similarly, the maximum disjunction operator considers only the largest value, while the sum considers all memberships, which provides better performance in the classification task [60,61]. In combination, these parameterization settings result in 86 rules, enabling the model to achieve high accuracy during testing (91.92%). Thus, the model was able to achieve a generalization capacity, which can be considered an excellent degree of accuracy, given the complexity involved. This predictive model of risk retention can help professors in identifying students from higher-risk groups and support the direction of efforts, and the most appropriate approaches, case-by-case.

5. Conclusions

From our study, we observed that different learning strategies provided a significant impact on the students' performance, but not with the same strength, once some showed bigger influence. Initially, according to the Mann-Whitney test, significant differences in the final grades were not found when segmenting the students by gender (p-value = 0.921), age (p-value = 0.239), and public/private school origin (p-value = 0.426). Additionally, although the absence of statistical evidence to prove that exercise solving (p-value = 0.616) and contextualized motivation (p-value = 0.388) are effective during the learning process, the bibliographic reading (p-value = 0.011) and lesson review (p-value < 0.001) were identified as significant strategies. Moreover, the emotion control (p-value = 0.006), class attendance (p-value = 0.002) and help-seeking (p-value = 0.008) are highlighted as determinant elements for satisfactory final grades. A logistic-based coefficient investigation was carried out to verify the impact of each learning strategy on the retention outcomes, corroborating the previous significance analyses. At last, the relevant factors were adopted to model

the retention risk through distinct AI-based models. The FIS showed the best prediction performance with an overall accuracy above 90%.

The fuzzy AI-based model was capable of providing an accurate prediction, confirming the research hypothesis. As verified in the present study, due to the lower amount of information available in small data, the prediction task was more difficult for the other AI methods, impairing its accuracy on test data. Therefore, the FIS can be considered a promising alternative for handling with small data in forecasting student performance and retention risk in engineering education. The used database intentionally covers a restricted number of observations only (small data) and, hence, does not aim to offer a ready-to-use technology. One could argue this as a main limitation. However, as an exploratory study, the potential of the proposed approach (fuzzy AI) to handle small data was assessed and results have shown a good performance.

As a practical implication, the SPE index can be applied as a tool to support students' planning in relation to the use of learning strategies. In turn, the AI model based on fuzzy can assist professors in identifying students at higher risk of retention, enabling preventive interventions. As an important limitation of this current work, it is important to highlight that the presentation and validation of the model was based on a single case study. Future studies should consider the effectiveness of this approach for other subjects and knowledge areas as well.

Author Contributions: Conceptualization, A.B.; data curation, A.B., D.Z., A.I.S.L., R.G.N., J.K.S.F., L.A.P.B. and R.M.; formal analysis, D.Z., A.I.S.L., A.O.B., I.d.B.J., J.K.S.F., L.C.d.C.M. and R.M.; investigation, R.G.N., I.d.B.J., L.C.d.C.M. and L.A.P.B.; methodology, A.B., D.Z., M.S., A.O.B., R.G.N., J.K.S.F., L.C.d.C.M., L.A.P.B. and R.M.; project administration, A.B.; software, D.Z., I.d.B.J., M.S. and L.A.P.B.; validation, A.B., A.O.B., I.d.B.J.; writing, A.B., D.Z., A.I.S.L., A.O.B., R.G.N., I.d.B.J., J.K.S.F., M.S., L.C.d.C.M., L.A.P.B. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: No funding was obtained for this study. The APC was funded by São Paulo State University (PROPG Notice #53/2022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Chen, Y.; Johri, A.; Rangwala, H. Running out of STEM: A Comparative Study across STEM Majors of College Students At-Risk of Dropping Out Early. In Proceedings of the International Conference on Learning Analytics and Knowledge, Sydney, NSW, Australia, 7–9 March 2018; pp. 270–279.
2. Buttcher, C.; Davies, C.; Highton, M. *Designing Learning*; Routledge: London, UK, 2020.
3. Nandeshawar, A.; Menzies, T.; Nelson, A. Learning patterns of university student retention. *Expert Syst. Appl.* **2011**, *38*, 14984–14996. [CrossRef]
4. Bartoszewski, B.L.; Gurung, R.A.R. Comparing the relationship of learning techniques and exam score. *Scholarsh. Teach. Learn. Psychol.* **2015**, *1*, 219–228. [CrossRef]
5. Souza, L.F.N.I. Learning strategies and related motivational factors. *Educar* **2010**, *36*, 95–107.
6. Dembo, M.H.; Seli, H. *Motivation and Learning Strategies for College Success: A Focus on Self-Regulated Learning*; Routledge: New York, NY, USA, 2012.
7. Séllei, B.; Stumphauser, N.; Molontay, R. Traits versus Grades-The incremental Predictive Power of Positive Psychological Factors over Pre-Enrollment Achievement Measures on Academic Performance. *Appl. Sci.* **2021**, *11*, 1744. [CrossRef]
8. Warr, P.; Downing, J. Learning strategies, learning anxiety and knowledge acquisition. *Br. J. Psychol.* **2000**, *91*, 311–333. [CrossRef]
9. Lu, K.; Yang, H.H.; Shi, Y.; Wang, X. Examining the key infuencing factors on college students' higher-order thinking skills in the smart classroom environment. *Int. J. Educ. Technol. High. Educ.* **2021**, *18*, 1. [CrossRef]
10. Martins, L.B.; Zerbini, T. Scale of learning strategies: Validity evidences in hybrid higher education. *Psico* **2014**, *19*, 317–328.
11. Bressane, A.; Bardini, V.S.S.; Spalding, M. Active learning effects on students' performance: A methodological proposal combining cooperative approaches towards improving hard and soft skills. *Int. J. Innov. Learn.* **2021**, *29*, 154–165. [CrossRef]

12. Rastrollo-Guerrero, J.L.; Gómez-Pulido, J.A.; Durán-Domínguez, A. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Appl. Sci.* **2020**, *10*, 1042. [CrossRef]
13. Hellas, A.; Ihantola, P.; Petersen, A.; Ajanovski, V.V.; Gutica, M.; Hynnien, T.; Knutas, A.; Leinonen, J.; Messom, C.; Liao, C. Predicting Academic Performance: A Systematic Literature Review. In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Lanarca, Cyprus, 2–4 July 2018.
14. Csalódi, R.; Abonyi, J. Integrated Survival Analysis and Frequent Pattern Mining for Course Failure-Based Prediction of Student Dropout. *Mathematics* **2021**, *9*, 463. [CrossRef]
15. Kiss, B.; Nagy, M.; Molontay, R.; Csabay, B. Predicting Dropout Using High School and First-semester Academic Achievement Measures. In Proceedings of the 17th International Conference on Emerging eLearning Technologies and Applications, Starý Smokovec, Slovakia, 21–22 November 2019.
16. Tight, M. Tracking the scholarship of teaching and learning. *Policy Rev. High. Educ.* **2018**, *2*, 61–78. [CrossRef]
17. Bird, K.A.; Castleman, B.L.; Mabel, Z.; Song, Y. Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. *AERA Open* **2021**, *7*, 1–19. [CrossRef]
18. Bertolini, R.; Finch, S.J.; Nehm, R.H. Quantifying variability in predictions of student performance: Examining the impact of bootstrap resampling in data pipelines. *Comput. Educ. Artif. Intell.* **2022**, *3*, e100067. [CrossRef]
19. Chen, L.; Chen, P.; Lin, Z. Artificial Intelligence in Education: A Review. *IEEE Open Access J.* **2020**, *8*, 75264–75278. [CrossRef]
20. Lateh, M.A.; Muda, A.K.; Yusof, Z.I.M.; Muda, N.A.; Azmi, M.S. Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review. *J. Phys. Conf. Ser.* **2017**, *892*, e012016.
21. Zadeh, L.A. *Computing with Words*; Springer: Berlin/Heidelberg, Germany, 2012.
22. Barros, L.C.; Bassanezi, R.C.; Lodwick, W.A. *A First Course in Fuzzy Logic, Fuzzy Dynamical Systems, and Biomathematics*; Springer: Berlin/Heidelberg, Germany, 2017.
23. Jayagowri, P.; Karpagavalli, K. A LFCM approach to student performance prediction based on learning fuzzy cognitive map. *Adv. Appl. Math. Sci.* **2021**, *21*, 3953–3965.
24. Petra, T.Z.H.T.; Aziz, M.J.A. Analysing student performance in Higher Education using fuzzy logic evaluation. *Int. J. Sci. Technol. Res.* **2021**, *10*, 322–327.
25. Mansouri, T.; ZareRavazan, A.; Ashrafi, A. A Learning Fuzzy Cognitive Map (LFCM) approach to predict student performance. *J. Inf. Technol. Educ. Res.* **2021**, *20*, 221–243. [CrossRef]
26. Nosseir, A.; Fathy, Y.M. A mobile application for early prediction of student performance using Fuzzy Logic and Artificial Neural Networks. *Int. J. Interact. Mob. Technol.* **2020**, *14*, 4–18. [CrossRef]
27. Lin, J.J.; Malyscheff, A.; Imbrie, P.K. Artificial Intelligence Methods to Forecast Engineering Students' Retention Based on Cognitive and Non Cognitive Factors. 2008. Available online: https://peer.asee.org/4315 (accessed on 7 April 2021).
28. Yadav, S.; Bharadwaj, B.; Pal, S. Mining Education Data to Predict Student's Retention: A comparative Study. *Int. J. Comput. Sci. Inf. Secur.* **2012**, *10*, 113–117.
29. Kolo, D.K.; Adepoju, S.A.; Alhassan, V.K. A Decision Tree Approach for Predicting Students Academic Performance. *Int. J. Educ. Manag. Eng.* **2015**, *5*, 12–19.
30. Mesarić, J.; Šebalj, D. Decision trees for predicting the academic success of students. *Croat. Oper. Res. Rev.* **2016**, *7*, 367–388. [CrossRef]
31. Dhanalakshmi, V.; Bino, D.; Saravanan, A. Opinion mining from student feedback data using supervised learning algorithms. In Proceedings of the 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, Oman, 15–16 March 2016; pp. 1–5.
32. Raut, A.B.; Nichat, A.A. Students Performance Prediction Using Decision Tree Technique. *Int. J. Comput. Intell. Res.* **2017**, *13*, 1735–1741.
33. Rao, P.V.V.S.; Sankar, S.K. Survey on Educational Data Mining Techniques. *Int. J. Eng. Comput. Sci.* **2017**, *4*, e12482.
34. Kavitha, G.; Raj, L. Educational Data Mining and Learning Analytics-Educational Assistance for Teaching and Learning. *Int. J. Comput. Organ. Trends* **2017**, *41*, 21–25. [CrossRef]
35. Lu, O.H.; Huang, A.Y.; Huang, J.C.; Lin, A.J.; Ogata, H.; Yang, S.J. Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educ. Technol. Soc.* **2018**, *21*, 220–232.
36. Buenaño-Fernández, D.; Gil, D.; Luján-Mora, S. Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study. *Sustainability* **2019**, *11*, 2833. [CrossRef]
37. Ahmed, M.; Muldoon, T.J.; Elsaadany, M. Employing Faculty, Peer Mentoring, and Coaching to Increase the Self-Confidence and Belongingness of First-Generation College Students in Biomedical Engineering. *J. Biomech. Eng.* **2021**, *143*, 121001. [CrossRef]
38. Huerta-Manzanilla, E.L.; Ohland, M.W.; Peniche-Vera, R.D.R. Co-enrollment density predicts engineering students' persistence and graduation: College networks and logistic regression analysis. *Stud. Educ. Eval.* **2021**, *70*, 101025. [CrossRef]
39. Bello, F.A.; Kohler, J.; Hinrechsen, K.; Araya, V.; Hidalgo, L.; Jara, J.L. Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout. In Proceedings of the 39th International Conference of the Chilean Computer Science Society (SCCC), Coquimbo, Chile, 16–20 November 2020.
40. Ortiz-Lozano, J.M.; Rua-Vieites, A.; Bilbao-Calabuig, P.; Casadesús-Fa, M. University student retention: Best time and data to identify undergraduate students at risk of dropout. *Innov. Educ. Teach. Int.* **2020**, *57*, 74–85. [CrossRef]

41. Alvarez, N.L.; Callejas, Z.; Griol, D. Predicting Computer Engineering Students' Dropout In Cuban Higher Education With Pre-Enrollment and Early Performance Data. *J. Technol. Sci. Educ.* **2020**, *10*, 241–258. [CrossRef]
42. Alvarez, N.L.; Callejas, Z.; Griol, D. Factors that affect student desertion in careers in Computer Engineering profile. *Rev. Fuentes* **2020**, *22*, 105–126.
43. Hassan, H.; Anuar, S.; Ahmad, N.B.; Selamat, A. Improve student performance prediction using ensemble model for higher education. *Front. Artif. Intell. Appl.* **2019**, *1529*, 52041.
44. Acero, A.; Achury, J.C.; Morales, J.C. University dropout: A prediction model for an engineering program in bogotá, Colombia. In Proceedings of the 8th Research in Engineering Education Symposium, REES 2019-Making Connections, Cape Town, South Africa, 10–12 July 2019.
45. Ravikumar, R.; Aljanahi, F.; Rajan, A.; Akre, V. Early Alert System for Detection of At-Risk Students. In Proceedings of the ITT 2018-Information Technology Trends: Emerging Technologies for Artificial Intelligence, Dubai, United Arab Emirates, 28–29 November 2019.
46. Fontes, L.S.; Gontijo, C.H. Active learning methodologies and their contribution to differential and integral calculus teaching. In Proceedings of the 12th Summer Workshop in Mathematics, Brasilia, Brazil, 10–14 February 2020.
47. Bressoud, D.; Rasmussen, C. Seven characteristics of successful calculus programs. *Not. AMS* **2015**, *62*, 144–146. [CrossRef]
48. Zimmerman, D.W. A note on consistency of non-parametric rank tests and related rank transformations. *Br. J. Math. Stat. Psychol.* **2012**, *65*, 122–144. [CrossRef] [PubMed]
49. Szumilas, M. Explaining Odds Ratios. *J. Can. Acad. Child Adolesc. Psychiatry* **2010**, *19*, 227–229. [PubMed]
50. McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*; Wiley-Interscience: Hoboken, NJ, USA, 2005.
51. Bressane, A.; Silva, P.M.; Fiore, F.A.; Carra, T.A.; Ewbank, H.; de-Carli, B.P.; Mota, M.T. Fuzzy-based computational intelligence to support screening decision in environmental impact assessment: A complementary tool for a case-by-case project appraisal. *Env. Impact Assess. Rev.* **2020**, *85*, 106446. [CrossRef]
52. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
53. Guillaume, S.; Charnomordic, B.; Lablée, J.; Jones, H.; Desperben, L. *FisPro: Fuzzy Inference System Design and Optimization*; R Package Version 1.1.1; R Core Team: Auckland, New Zealand, 2022; Available online: https://CRAN.R-project.org/package=FisPro (accessed on 24 October 2022).
54. R Core Team. *R: A Language and Environment for Statistical Computing, Version 4.0*; R Packages Retrieved from Mran Snapshot 2021-04-01; R Core Team: Auckland, New Zealand, 2021; Available online: https://cran.r-project.org (accessed on 1 January 2020).
55. Sampaio, B.; Guimarães, J. Diferenças de eficiência entre ensino público e privado no Brasil. *Econ. Appl.* **2009**, *13*, 45–68. [CrossRef]
56. Tinto, V. Research and practice of student retention: What next? *J. Coll. Stud. Retent. Res. Theory Pract.* **2006**, *8*, 1–19. [CrossRef]
57. Jakubowski, T.G.; Dembo, M.H. Social Cognitive Factors Associated with the Academic Self-Regulation of Undergraduate College Students in a Learning and Study Strategies Course. In Proceedings of the Annual Meeting of the American Educational Research Association, New Orleans, LA, USA, 1–5 April 2002.
58. Tseng, H.; Xiang, Y.; Yeh, H.T. Learning-related soft skills among online business students in higher education: Grade level and managerial role differences in self-regulation, motivation, and social skill. *Comput. Hum. Behav.* **2019**, *95*, 179–186. [CrossRef]
59. Anthony, S.; Garner, B. Teaching soft skills to business students: An analysis of multiples pedagogical methods. *Bus. Prof. Commun. Q.* **2016**, *79*, 360–370. [CrossRef]
60. Broadbent, J. Comparing online and blended learner's self-regulated learning strategies and academic performance. *Internet High. Educ.* **2017**, *33*, 24–32. [CrossRef]
61. Ghodousian, A.; Naeeimi, M.; Babalhavaeji, A. Nonlinear optimization problem subjected to fuzzy relational equations defined by Dubois-Prade family of t-norms. *Comput. Ind. Eng.* **2018**, *119*, 167–180. [CrossRef]