An Empirical Study of Machine Learning for Course Failure Prediction: A Case Study in Numerical Methods

Isaac Caicedo-Castro
Socrates Research Team
Research Team: Development, Education, and Healthcare
Faculty of Engineering
University of Córdoba
Carrera 6 No. 76-103, 230002, Montería, Colombia
ORCID: 0000-0002-7567-3774
e-mail: isacaic@correo.unicordoba.edu.co

***Abstract*—In this paper, we address the problem of predicting whether a student might fail a course before it starts, based on their academic history. This study is centered on predicting failure in the numerical methods course, which is part of the curriculum for the bachelor's degree in systems engineering at the University of Córdoba in Colombia. To tackle this problem, we adopt classification methods from supervised machine learning. To this end, we utilize a dataset initially collected in [1] and subsequently expanded in [2]. This dataset is used to fit and validate the machine learning methods employed in this study. Our work contributes to improving the quality of the forecasting task compared to prior research [1], [2]. This improvement has been achieved by modifying the vector representation of the student's academic history, considering only the student's performance in mathematics, as evidenced in the admission test called Saber 11 and prerequisite courses. The results of the experimental validation reveal that the method based on Gaussian processes with the Radial Basis Function achieves mean values of accuracy, precision, recall, and harmonic mean of 83%, 80.67%, 77%, and 76.70%, respectively. This method has outperformed the others studied in this work. Moreover, the prediction outcome of Gaussian processes is the probability that a given student will fail the course, which is convenient for designing an intervention plan to help them succeed. Therefore, the conclusion of this study is twofold. Firstly, Gaussian processes are the best choice to implement an intelligent system for the prediction task studied herein. Secondly, this study finds a clear correlation between the probability of succeeding in the numerical methods course and the student's competencies in mathematics obtained before enrolling in this course. This suggests that good training in mathematics courses is required to succeed in the numerical methods course.**

***Index Terms*—Machine learning; educational data mining; classification algorithm.**

I. Introduction

The aim of this study is to leverage machine learning accuracy to design an intelligent system for predicting student failure in the numerical methods course before it even starts. The input variables are derived from the student's performance in prerequisite courses that are assumed necessary for succeeding in the numerical methods course within the bachelor's degree program in systems engineering at the University of Córdoba, Colombia.

Herein and in several literature references, failing a course is often referred to as either dropping out or not passing the course successfully. Identifying students at risk of failing a specific course is crucial, as it enables stakeholders such as lecturers, students, academic policymakers, and others to take necessary precautions to prevent failure. This proactive approach helps students avoid psychological stress, frustration, and financial loss.

Prior research has explored various machine learning approaches to identify students at risk of course failure. These methods include classification methods, such as artificial neural networks or multilayer perceptron [1]–[7], support vector machines [1]–[4], [8], quantum-enhanced support vector machine [9], logistic regression [1], [4], [8], [10], decision trees [1], [2], [4], [8], [10], [11], ensemble methods with different classification methods [3], [6], random forest [1], [2], [4]–[6], gradient boosting [5], extreme gradient boosting (XGBoost) [1], [2], [5], [6], variants of gradient boosting [5], [8], such as CatBoost [12] and LightGBM [13], and Gaussian processes for classification [1], [2].

Much of the previously referenced literature has centered on online courses [3], [4], [6], [7], [10], covering various topics such as computer networking and web design [3], mathematics [10], and STEM (science, technology, engineering, and mathematics) in general [7]. It is noteworthy that the primary goal of these research endeavors is not to predict the risk of failure before students begin their courses; instead, they are focused on forecasting risk during the course development phase. This forecasting relies on students' activities, including the number of course views, content downloads, and grades achieved in assignments, tests, quizzes, projects, and other assessments.

The objectives pursued in [1], [2], [9] align with the goals of this study, operating within the same context of predicting student failure in the numerical methods course based on performance in prerequisite courses within the undergraduate program in systems engineering at the University of Córdoba in Colombia. However, it is worth noting that [1] utilized a smaller dataset compared to the one employed in this study, which corresponds to the dataset collected in [2] and used in [9].

In [1], the student's performance in prerequisite courses, including mathematics, physics, and computer programming,

as well as their outcomes in the admission test, are used as input variables for predicting whether they might fail the numerical method course. In contrast, in [2], it is observed that excluding the student's outcomes in the admission test leads to increased accuracy, precision, and recall in the prediction. Furthermore, in [9], quantum machine learning is adopted; however, this approach does not outperforms the performance of Gaussian processes for classification as utilized in [2].

A 10-Fold Cross-Validation conducted in [2] indicates that the Gaussian process with the Matern kernel achieves mean values of accuracy, precision, recall, and harmonic mean of 80.45%, 83.33%, 66.5%, and 72.52%, respectively. In contrast, our study has found far better results in terms of accuracy and harmonic mean by reducing the input variables to only include the student's performance in the prerequisite mathematics course and the outcomes of the admission test in the same subject. With this input configuration, the Gaussian process with the Radial Basis Function kernel yields mean values of accuracy, precision, recall, and harmonic mean of 83%, 80.67%, 77%, and 76.79%, respectively.

The remainder of this paper is organized as follows: Section II formalizes the problem, presents key assumptions, outlines the representation of the input space, and introduces the target variable. This section also presents the machine learning methods adopted in this study, along with the validation method. Section III provides details on the dataset features, programming language, software library, and computing environment used during the validation process. In Section IV, the results of the experiments are presented, followed by a discussion in Section V. Finally, Section VI concludes the discussion of the results, highlights the novelty of this study, and suggests directions for future research.

II. Methods

The problem addressed in this study is to identify regular patterns between failing the numerical methods course and student's performance in prerequisite courses, as well as their performance in the admission test. To cope with this problem, a quantitative approach is adopted. This approach involves quantifying student's performance through their grades in prerequisite courses, their scores on the admission test, and considering the frequency of course enrollments due to previous failures.

In this study, we use a dataset initially collected in [14], which was subsequently expanded with additional instances in [2] and further utilized in [9]. The dataset describes input variables for each prerequisite course as follows: the variables $x_{i,j}$ and $x_{i,j+2}$ represent the highest and lowest final grades obtained by the ith student in the jth course, respectively, while $x_{i,j+1}$ denotes the number of semesters the ith student has enrolled in the jth course. If the ith student does not fail the jth course upon the first enrollment, $x_{i,j}$ and $x_{i,j+2}$ will have the same value, and $x_{i,j+1}$ will be equal to one. Additionally, in this study, $x_{i,1}$ represents the score achieved by the ith student in the subject of mathematics in the admission test.

This dataset has been collected through a survey conducted among students of systems engineering at the University of Córdoba in Colombia. To safeguard the privacy of the participants, the dataset has been anonymized, retaining only students' grades and admission scores. Personal information, including identification numbers, names, gender, and economic stratum, has been omitted.

In this study, it is assumed that only mathematics courses are prerequisite for success in the numerical methods course. Therefore, the prerequisite courses considered are linear algebra, calculus I, II, and III. Conversely, in previous studies [2], [9], [14], physics and computer programming courses are also considered prerequisites, although the outcomes of the admission test are not utilized in [2], [9], while all admission test outcomes are used in [14]. As a result, the input space in our study has 13 dimensions (three per prerequisite course and the mathematics subject score of the admission test), compared to 33 dimensions in [2], [9], and 38 dimensions in [14].

Mathematically, the ith student is represented by a D-dimensional vector $\mathbf{x}_i \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^D$ and $D = 13$, accounting for four prerequisite courses and one additional dimension for the mathematics subject score in the admission test. Each component of the vector $\mathbf{x}_i$ corresponds to a specific input variable, detailed as follows:

- x_{i1}: The score obtained by the student in the admission test for mathematics, ranging from 0 to 100 ($x_{i1} \in \mathbb{Z}$, $0 \leq x_{i1} \leq 100$).
- x_{i2} to $x_{i,13}$: Grades and enrollment information for the prerequisite courses, with x_{ij}, $x_{i,j+2}$, and $x_{i,j+1}$ representing the highest final grade, lowest final grade, and number of semesters enrolled for each course j, respectively. Here, j takes values of 2, 5, 8, and 11, corresponding to calculus I, II, III, and linear algebra courses.
- In Colombian universities, students are graded in the range from 0 up to 5 (see student's code of the University of Córdoba [15]), i.e., $x_{ij} \in \mathbb{R}$ and $0 \leq x_{ij} \leq 5$ for $j = 2, 4, 5, 7, 8, 10, 11, 13$.
- The number of semesters enrolled for each course j is a natural number or zero if a the ith student has never enrolled it, i.e., $x_{ij} \in \mathbb{N} \cup \{0\}$ for $j = 3, 6, 9, 12$.

On the other hand, y_i represents the target variable associated with the ith student, where $y_i = 1$ if the student failed the numerical methods course, and $y_i = 0$ otherwise (i.e., $y_i \in \{0, 1\}$). Thus, the $\mathcal{D}$ denotes the dataset defined as $\mathcal{D} = \{(x_i, y_i) | x_i \in \mathcal{X}, y_i \in \{0, 1\}, \text{for all } i = 1, \ldots, n\}$, where n represents the size of the dataset.

Given the dataset $\mathcal{D}$ described earlier, the problem in this study is to determine the function g, which maps the input variables to the target variable, i.e., $g : \mathcal{X} \to \{0, 1\}$. Once this function is established, it may be used to predict whether a new student, represented by the vector $x' \in \mathcal{X}$, might fail the numerical methods course. Specifically, if $g(x') = 1$, the function predicts that the student might fail the course; otherwise, if $g(x') = 0$, the prediction is that the student will

not fail. This problem falls under supervised learning and is addressed using classification methods.

Before training various classifiers, the dataset undergoes preprocessing. This involves centering each input variable by removing the mean and scaling to unit variance. It is important to note that the dataset contains no missing data, eliminating the need for any imputation methods.

After preparing the dataset, K-fold cross-validation is employed to evaluate each classification method. This process involves splitting the dataset $\mathcal{D}$ into K equal parts, denoted as $\mathcal{D}_i$, where i ranges from 1 to K. During each iteration, one of the K parts is set aside as the validation set, denoted as $\mathcal{V}_k = \mathcal{D}_j$, while the remaining $K-1$ parts are used for training the classifier, represented as $\mathcal{T}_k = \cup_{i=1,i\neq j}^{K}\mathcal{D}_j$. This partitioning is carried out as follows:

$$\begin{aligned} \mathcal{V}_1 &= \mathcal{D}_1, & \mathcal{T}_1 &= \mathcal{D}_2 \cup \mathcal{D}_3 \cup \cdots \cup \mathcal{D}_K \\ \mathcal{V}_2 &= \mathcal{D}_2, & \mathcal{T}_2 &= \mathcal{D}_1 \cup \mathcal{D}_3 \cup \cdots \cup \mathcal{D}_K \\ &\vdots & &\vdots \\ \mathcal{V}_K &= \mathcal{D}_K, & \mathcal{T}_K &= \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_{K-1} \end{aligned}$$

In this study, various classification methods were validated, including logistic regression. The latter assumes the existence of a hyperplane that separates vectors into two classes in a multidimensional real-valued space. This assumption is reasonable given that the input space is a multidimensional real-valued vector space (i.e., $D = 13$).

Nevertheless, other classification methods more suitable for nonlinear classification problems are also utilized in this study, under the assumption that there might be a more effective input representation in a higher-dimensional space. Probabilistic methods such as Gaussian processes (GPs) have been adopted. Based on Bayesian inference, GPs assume that the probability distribution of the target variable is drawn from a Gaussian or normal distribution, hence the name of the method [16], [17]. The main advantage of this method is its ability to incorporate prior knowledge about the problem, contributing to improved forecasting accuracy, even with a small training dataset, as is the case in the context of this study.

In this study, we used several kernels (a.k.a., covariant functions) with GPs. For instance, the Radial Basis Function kernel, which is defined as follows:

$$k_G(\mathbf{x}_i, \mathbf{x}_j) = \gamma \exp\left(-\frac{||\mathbf{x}_j - \mathbf{x}_i||^2}{2\sigma^2}\right), \tag{1}$$

where $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$ are two D-dimensional vectors in real-valued space, and the hyperparameters $\gamma, \sigma \in \mathbb{R}$ are real numbers that corresponds to the weight and length scale of the kernel, respectively.

In addition, we used the Matern kernel, which is defined as follows:

$$k_M(\mathbf{x}_i, \mathbf{x}_j) = \frac{\gamma(\sqrt{2\nu}||\mathbf{x}_j - \mathbf{x}_i||^2)^\nu}{\Gamma(\nu)2^{\nu-1}\sigma^\nu} K_\nu\left(\frac{\sqrt{2\nu}||\mathbf{x}_j - \mathbf{x}_i||^2}{\sigma}\right), \tag{2}$$

where $K_\nu(\cdot)$ and $\Gamma(\cdot)$ are the modified Bessel function and the gamma function, respectively. The hyperparameter $\nu \in \mathbb{R}$ controls the smoothness of the kernel function.

Moreover, a rational quadratic kernel is utilized, which defined as follows:

$$k_r(\mathbf{x}_i, \mathbf{x}_j) = \left(1 + \frac{||\mathbf{x}_j - \mathbf{x}_i||^2}{2\alpha\sigma^2}\right)^{-\alpha}, \tag{3}$$

where σ is used for the same purpose as in (1), while $\alpha \in \mathbb{R}$ is the scale mixture parameter, such that $\alpha > 0$.

Furthermore, Matern kernel and radial basis function are combined by summing both as follows:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \gamma_G k_G(\mathbf{x}_i, \mathbf{x}_j) + \gamma_M k_M(\mathbf{x}_i, \mathbf{x}_j), \tag{4}$$

where γ_G and γ_M are the weights assigned to the kernels.

On the other hand, the classification method based on support vector machines (SVMs) is considered one of the most theoretically motivated and successful in modern machine learning practices ([18], p. 79). SVMs are based on convex optimization, allowing for the identification of a global maximum solution, which is their main advantage. However, SVMs are not well-suited for interpretation in data mining; nevertheless, they excel in training accurate machine learning systems. For a detailed description of this method, refer to [19].

Both SVMs and logistic regression are linear classification methods, operating under the assumption that the input vector space can be separated by a linear decision boundary or, in the case of multidimensional input spaces, by a hyperplane. However, when this assumption is not met, SVMs may be used alongside kernel methods to handle nonlinear decision boundaries (see [19] for further details). In this study, we utilize the radial basis function kernel, similar to the one presented in (1), defined as follows:

$$k_G(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_j - \mathbf{x}_i||^2), \tag{5}$$

where γ controls the radius of this spherical kernel, whose center is $\mathbf{x}_j$. Additionally, polynomial and Sigmoid kernels are used, which defined in (6) and (7), respectively. In (6), $d \in \mathbb{N}$ is the degree of the kernel, and $\gamma \in \mathbb{R}$ is the coefficient in (7).

$$k_p(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^d \tag{6}$$

$$k_s(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma\langle \mathbf{x}_i, \mathbf{x}_j \rangle) \tag{7}$$

Although support vector machines (SVMs) are considered one of the most successful methods in modern machine learning, multilayer perceptrons (MLPs) and their variants, namely artificial neural networks, have emerged as the most successful in deep learning and big data applications, particularly in tasks such as speech recognition, computer vision, and natural language processing ([20], p. 3). In this study, MLPs are trained using the back-propagation algorithm with cross-entropy error minimization [21], along with the optimization algorithm known as Adam [22]. Specifically, MLPs with one and five hidden layers are adopted in this work.

MLPs offer a significant advantage as they can approximate any function for both classification and regression tasks, making them universal approximators. However, they also have a major drawback: the objective function, typically based on cross-entropy error, is not convex. Consequently, the synaptic weights obtained during training might not converge to the optimal solution due to the presence of multiple local minima in the objective function. The solution heavily relies on the random initialization of synaptic weights. Additionally, MLPs require tuning more hyperparameters compared to other learning methods such as SVMs or naive Bayes, which presents another disadvantage.

Among the methods mentioned above, logistic regression stands out for its interpretability. However, for the remaining methods, interpretability is a challenge. To address this, decision trees are adopted, as they are classification algorithms commonly used in data mining and knowledge discovery. During decision tree training, a tree is constructed using the dataset as input, where each internal node represents a test on an independent variable, each branch represents the result of the test, and leaves represent predicted classes. The tree is built recursively, starting with the entire dataset as the root node. At each iteration, the fitting algorithm selects the next attribute that best separates the data into different classes. The fitting algorithm may halt based on various criteria, such as when all training data have been classified or when the classifier's accuracy or performance can no longer be improved.

Decision trees are constructed using heuristic algorithms, often employing greedy strategies. At each node, these algorithms may identify several local optimal solutions, leading to no guarantee that the learning process will converge to the most optimal solution. This issue is not unique to decision trees but is also present in other algorithms, such as multilayer perceptrons. However, it remains a primary drawback of decision trees, as small variations in the training dataset can cause significant changes in the tree structure.

The method of decision trees is introduced in 1984, in [23] is delved into its details. To improve the performance of decision trees, ensemble methods based on multiple decision trees have been developed. These methods include Adaboost (adaptive boosting) [24], random forest [25], and extreme gradient boosting, which is also known as XGBoost [26].

Ultimately, so far there is no analytical method to definitively determine the best machine learning approach, as demonstrated by the No Free Lunch Theorem [27]. According to this theorem, the predictive quality of machine learning methods hinges on the unknown distribution of the dataset enshrined to fit them. Consequently, experimental validation becomes the only mean of identifying the most effective method for addressing the problem studied in this work. The following section details the experimental setup adopted to empirically validate these methods.

III. Experimental Setting

To fit and validate the machine learning methods outlined in the previous section, we employed a dataset containing 103 examples. Each example consists of 38 independent variables along with its corresponding dependent or target variable. However, in this study, only 13 out of the 38 independent variables are utilized, as explained earlier. Notably, the dataset is the same used in [2] to compare the results of both studies.

Figure 1 illustrates the proportion of positive and negative instances in the dataset. Positive instances represent examples where students failed the numerical methods course, while negative instances denote cases where students passed. The pie chart depicts that the dataset is reasonably balanced, with a slightly higher number of negative examples due to more students successfully passing the course.

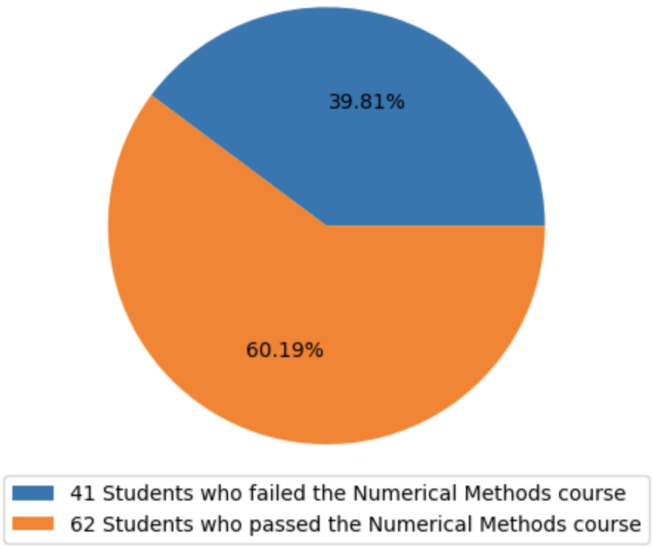

Fig. 1. Distribution of student outcomes in the numerical methods course dataset. The figure illustrates that 41 out of 103 students who participated in the study failed the numerical methods course (39.81% of the surveyed students), while 62 out of 103 students passed the course (60.19% of the sample).

The dataset has been collected through a survey conducted on students enrolled in courses from the fifth to ninth semester of the bachelor's degree program in systems engineering at the University of Córdoba, Colombia. Due to changes in the curriculum structure in 2018, data collection before that year was not feasible, resulting in the dataset's limited size.

As explained in the previous section, the students' outcomes from the Saber 11 test are included in the dataset used in this study. Figure 2 shows that students who failed the numerical methods course scored lower in the mathematics section of the Saber 11 test than those who succeeded in the course. Indeed, the notches of the boxplots in the figure do not overlap, indicating that the median score of students who succeeded in the course is significantly higher than that of students who failed.

The number of times each student enrolls in a prerequisite course is one of the variables in the dataset. If students succeed in the prerequisite course on the first enrollment, this value is equal to one; otherwise, it is greater. Students who have failed the numerical methods course tend to enroll in prerequisite

courses more times than those who succeed in the numerical methods course (see Figure 3).

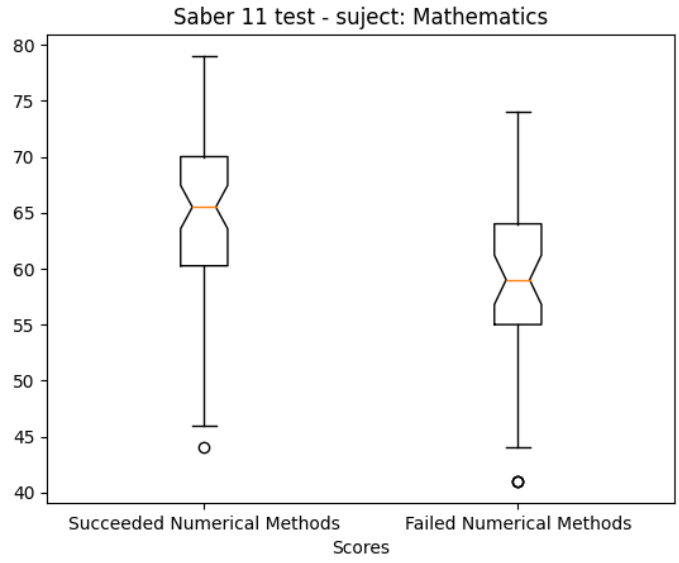

Fig. 2. Boxplots showing the scores obtained by the students in the mathematics subject of the Saber 11 test, categorized based on success or failure in the numerical methods course.

Enrolling in a prerequisite course several times means that a given student has failed it with the same frequency. However, some students who have passed prerequisite courses on their first try have failed the numerical methods course as well. This explains why variables related to the number of enrollments are not sufficient to predict success in the numerical methods course, as evidenced in Figure 3. The boxplots in this figure illustrate that the median of each of these aforementioned variables is equal to one for students who have failed numerical methods.

The lowest grades attained in prerequisite courses are lower for students who have failed the numerical methods course compared to those who have succeeded in it (see Figure 4). The notches of the boxplots do not overlap for most prerequisite courses, except for calculus III. This suggests that performance in calculus III might not significantly contribute to accurate predictions, while the median of the lowest grades in other courses varies depending on whether students have succeeded or failed the numerical methods course.

This observation aligns with the histogram of the lowest grades achieved in calculus III shown in Figure 5, where students who failed this prerequisite course also failed numerical methods, whereas several students who succeeded in calculus III on the first enrollment still failed numerical methods. Recall that if a student succeeds in a prerequisite course on the first enrollment, their lowest grade is at least 3; otherwise, the lowest grade is lower than this value (see Section II).

Similarly, the highest grades in prerequisite courses for students who succeeded in the numerical methods course are better than the highest grades in prerequisite courses for those who failed the numerical methods, as illustrated in Figures 6 and 7.

Therefore, the statistics indicate that students who failed the numerical methods course perform less effectively in prerequisite courses compared to those who succeeded. Indeed, this observation constitutes a key assumption of our study.

Besides, the validation of each classification method with the aforementioned dataset is conducted using the Python programming language and the open-source library Scikit-learn [28]. Scikit-learn provides comprehensive support for various machine learning tasks, including supervised and unsupervised methods. The validation tests are implemented in notebooks within the Google Colaboratory platform [29].

Furthermore, grid search and K-fold Cross-Validation (K-Fold CV) are both utilized in combination to explore various hyperparameter value combinations and tune the hyperparameters of each model.

Finally, during K-Fold CV, a value of $K = 10$ is chosen, although $K = 30$ is also common. However, the larger K is, the smaller the validation set becomes, this potentially limits the ability to test hypotheses concerning the performance of the methods. To assess the validation outcomes, a paired t-test is employed. The results of the validation are presented in the next section, while an analysis of their significant differences, based on paired t-test, is discussed in Section V.

IV. Results

The results obtained from the Ten-Fold Cross-Validation (10-Fold CV) reveal that the Gaussian process (GP) with the Radial Basis Function (RBF) kernel attained the highest accuracy, recall, and harmonic mean (F_1). While the GP with the RBF kernel occupies third place in terms of precision, the method called Support Vector machines (SVM) with the same kernel (RBF) ranks among the top three most accurate methods, alongside the GP with the rational quadratic kernel. Details of these results can be found in Table I.

The GP with the RBF kernel has the best trade-off between precision and recall. This is evident in its F_1 score, which is a desirable feature for an intelligent system predicting student failure in the numerical methods course. For instance, while SVMs with the RBF kernel achieved the highest precision, they had lower recall compared to the GP with the same kernel. This means that a system based on SVMs is more likely to miss students at risk of failing the course compared to one based on GPs.

The results regarding the F_1 metrics are aligned with the confusion matrix shown in Table II, where GP with the RBF kernel predicted that 8 out of 62 students would fail the numerical methods course although they never did, resulting in 8 false positive examples. Moreover, the GP predicted that 10 out of 41 students would not fail the course, although they did, resulting in 10 false negative instances. According to the same confusion matrix, 85 out of 103 students are classified properly during the validation of GP.

TABLE I
TEN-FOLD CROSS-VALIDATION RESULTS

Machine learning method	*Mean Accuracy (%)*	*p-value*	*Mean Precision (%)*	*p-value*	*Mean Recall (%)*	*p-value*	*Mean* $\mathbf{F_1}$ *(%)*	*p-value*
Gaussian process with the radial basis function kernel	**83.00**		80.67		**77.00**		**76.79**	
Gaussian process with the Matern kernel	81.00	0.80	81.67	0.92	69.50	0.55	73.49	0.77
Gaussian process with a sum of radial basis function and Matern kernel	79.00	0.61	76.00	0.63	72.00	0.69	71.62	0.64
Gaussian process with the dot product kernel	76.00	0.33	70.67	0.28	72.00	0.69	68.51	0.42
Gaussian process with the rational quadratic kernel	80.09	0.69	77.33	0.79	67.50	0.54	69.27	0.54
Support vector machines with the radial basis function kernel	80.09	0.69	**85.00**	0.74	55.00	0.13	63.48	0.31
Support vector machines with the sigmoid kernel	78.00	0.48	75.83	0.60	67.00	0.41	69.25	0.46
Support vector machines with the polynomial kernel (degree = 3)	78.00	0.49	79.17	0.91	52.50	0.09	60.55	0.21
Decision tree with gini index	72.91	0.18	75.00	0.59	59.50	0.16	61.90	0.17
Decision tree with entropy index	66.18	**0.04**†	59.67	0.07	54.50	0.07	55.91	0.06
XGBoost	75	0.27	76.83	0.76	52.50	0.06	59.60	0.14
Adaboost with the entropy index	62.27	**0.04**†	56.33	**0.05**†	52.50	0.06	53.41	0.06
Random forest with the entropy index	72.09	0.16	72.50	0.46	52.00	**0.05**†	58.29	0.10
Logistic regression	69.82	0.07	70.00	0.53	27.50	**0.0006**†	37.90	**0.005**†
Multilayer perceptron with a single hidden layer	69.09	**0.04**†	62.76	0.06	69	0.56	61.01	0.12
Multilayer perceptron with five hidden layers	75.09	0.32	65.33	0.21	67.50	0.51	64.33	0.33

†Paired t-test reveals the difference between means is statistically significant

TABLE II
CONFUSION MATRIX ILLUSTRATING THE PERFORMANCE OF GAUSSIAN PROCESS CLASSIFICATION USING THE RADIAL BASIS FUNCTION KERNEL.

	***F*orecasted class**		
True class	*Student might not fail*	*Student might fail*	*Total*
Student did not fail	54	8	62
Student failed	10	31	41
Total	64	39	103

It is noteworthy that the accuracy of the GP with the RBF kernel aligns with the area under the Receiver Operating Characteristics (ROC) curve, as illustrated in Figure 8. With an area of 0.81, this result indicates that the classification method performs much better than random guessing. The dataset's richness contributes to accurately predicting the probability of course failure, demonstrating the classifier's robust discriminatory power, which is well-suited for this predictive task.

The optimal hyperparameter settings for each classifier is determined through 10-Fold CV and grid search. To facilitate the reproducibility of the results in future research, the hyperparameter settings corresponding to the outcomes presented in Table I are as follows:

- Gaussian processes for classification:
  - Radial Basis Function kernel: $\gamma = 0.125$, $\sigma = 0.5$.
  - Matern kernel: $\gamma = 2.44 \times 10^{-4}$, $\sigma = 0.5, \nu = 1.3$.
  - Combination of Radial Basis Function and Matern kernel: $\gamma_G = 0.125$, $\gamma_M = 2.44 \times 10^{-4}$
  - Rational Quadratic kernel: $\gamma = 32$, $\sigma = 0.25$
- Support Vector Machines:

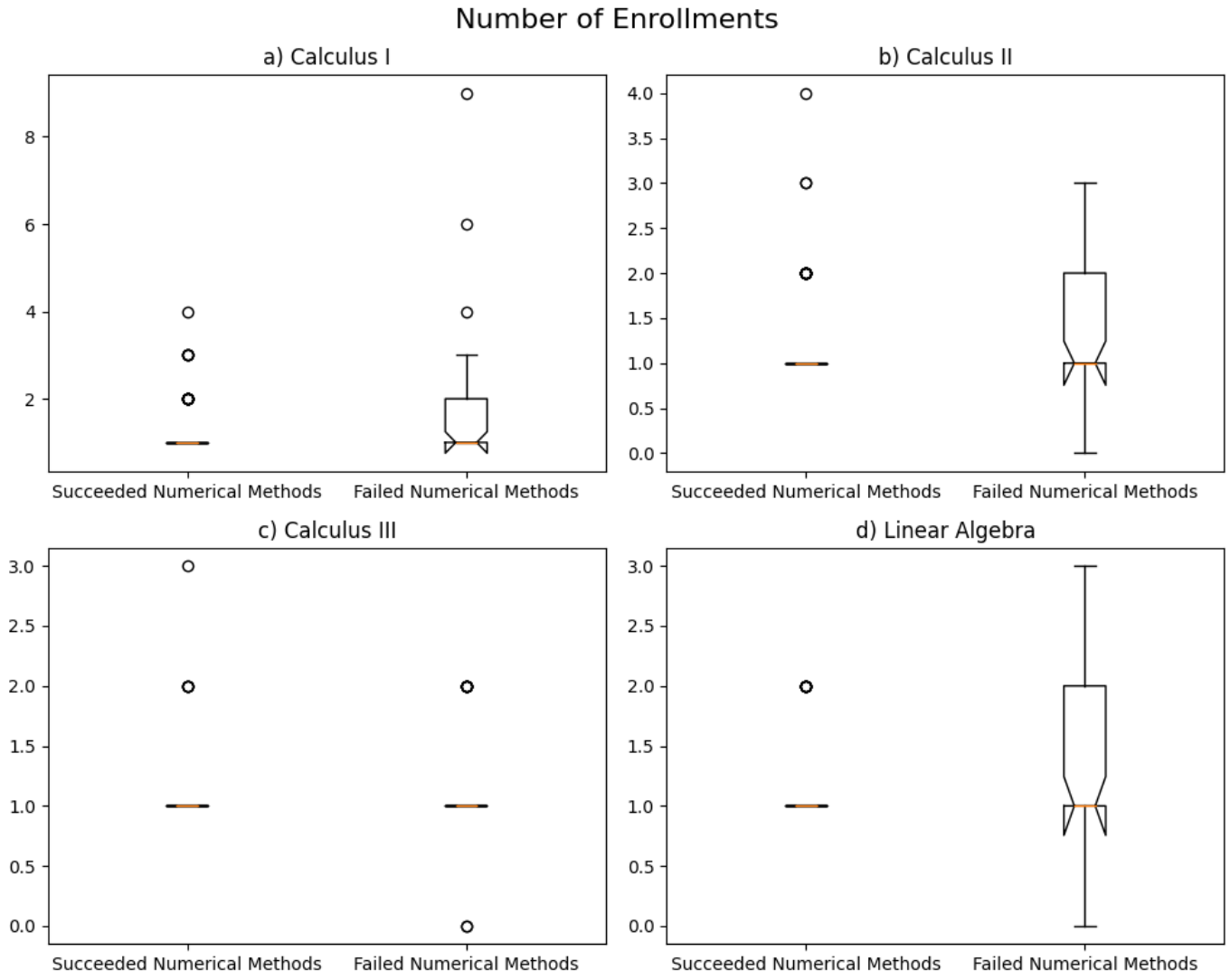

Fig. 3. Boxplots depicting the number of enrollments in each prerequisite course, categorized based on success or failure in the numerical methods course.

  - Radial Basis Function kernel: $C = 1$, $\gamma = 0.25$
  - Polynomial kernel: Regularization parameter $C = 8$, $d = 3$
  - Sigmoid kernel: Regularization parameter $C = 8192$, $\gamma = 3.05 \times 10^{-5}$
- Logistic regression: Regularization parameter $C = 0.01$
- Decison trees: Gini and Entropy indexes.
- XGBoost: Learning rate equal to 0.0625, maximum depth of 5 levels, 80 estimators, and Entropy index.
- Adaboost: Learning rate equal to 0.124, 50 estimators , and Entropy index.
- Random forest: 15 trees, minimum 2 samples per leaf, minimum 4 samples per split, maximum depth of 8 levels.

V. Discussion

Based on the validation results (see Table I), Gaussian processes with the Radial Basis Function kernel (GPRBF) emerged as the top-performing machine learning method in this study. This outcome is due to the fact that there is no hyperplane decision boundary that separates the original input space between the two classes (i.e., students at risk of failing and those not at risk). There is a regular pattern in the academic history of students who fail numerical methods, but no single variable is sufficient to accurately predict their likelihood of failure. For instance, some students who have failed the numerical methods course succeeded in prerequisite courses on their first enrollment, as shown in Figure 4. Therefore, a nonlinear method such as GPRBF is well-suited to the problem addressed in this study.

Besides, paired t-tests are conducted on the means of each metric obtained during validation, revealing the following insights:

- The mean accuracy of GPRBF is far greater than one attained through extreme gradient boosting (or XGBoost), Adaboost, and Multilayer perceptron with a single hidden layer.

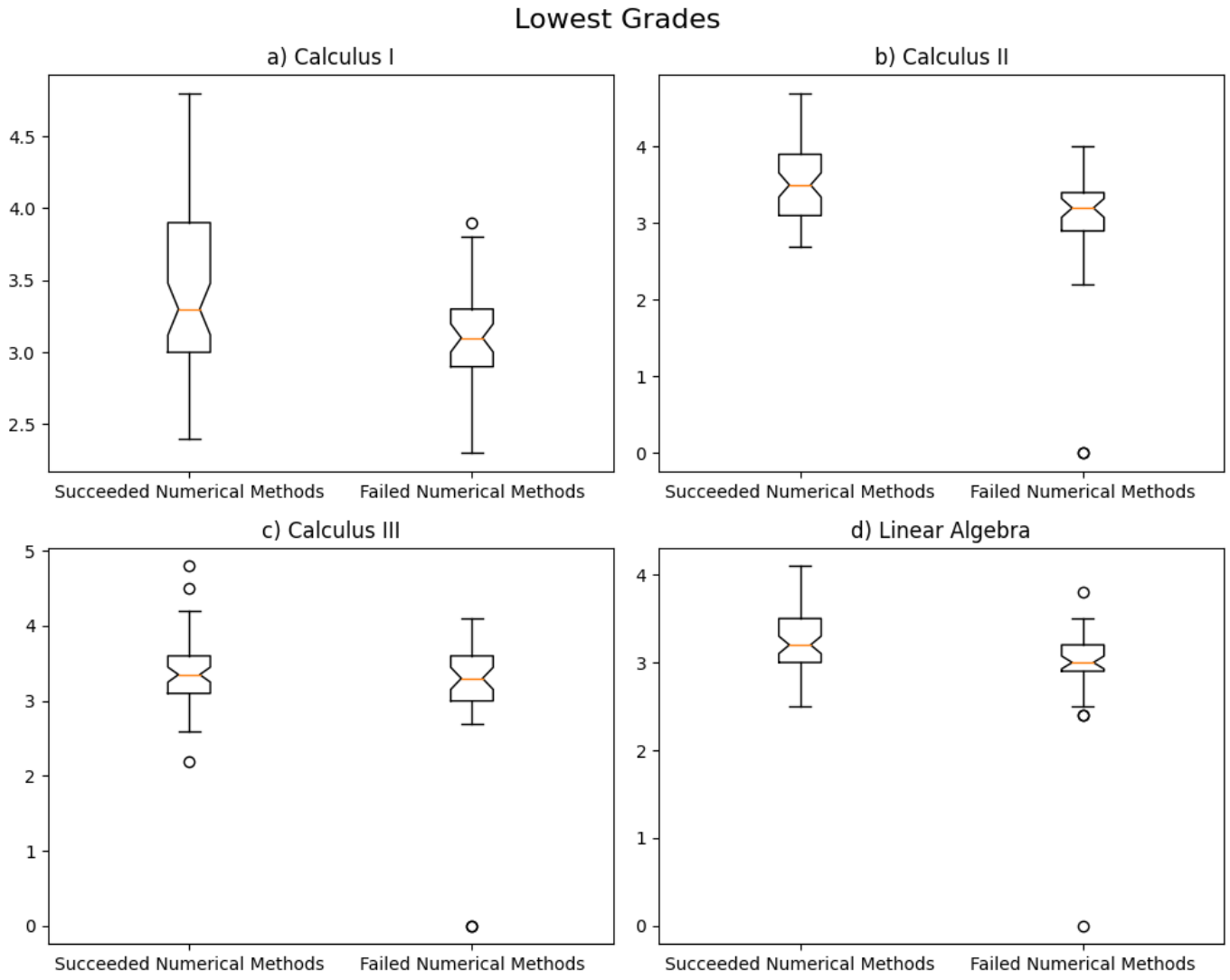

Fig. 4. Boxplots illustrating the lowest grades attained in each prerequisite course, categorized based on success or failure in the numerical methods course.

- A notable difference in mean precision is observed between GPRBF and Adaboost, indicating that the former method is considerably more precise.
- The means of precision between GPRBF and Support Vector Machines with the Radial Basis Function kernel (SVMRBF) do not show significant differences.

Moreover, classification methods with a probabilistic nature, such as Gaussian processes (GP), logistic regression, and Multilayer Perceptron, offer the advantage of providing the user with information about the probability of failing the numerical methods course. For instance, it becomes evident that precautions may be necessary to prevent failure for a specific student, especially when their probability of failing is as high as 80%, compared to another student whose probability is approximately 58%. This underscores the suitability of GP for implementing intelligent systems aimed at the predictive task addressed in this study.

Students identified as being at high risk of failing the course might benefit from support services [30]. These services may include access to course advisors, psychologists, learning and writing advisors, counselors, librarians, disability specialists, and so forth. By directing these resources toward students facing a significant risk of failure, it is possible to optimize cost-effectiveness and ensure targeted support where it is most needed.

Implementing pedagogical contracts between lecturers and students presents another intervention method to support students at high risk of failing the course. These contracts include personalized agreements tailored to each at-risk student's unique needs, strengths, weaknesses, and learning style. They establish clear goals aligned with the student's capabilities and offer flexibility to adapt to individual circumstances throughout the course. By continuously adjusting the contract, educators can provide targeted support to facilitate the student's progress and improve their chances of success.

Thus, students with a moderate probability of failing the

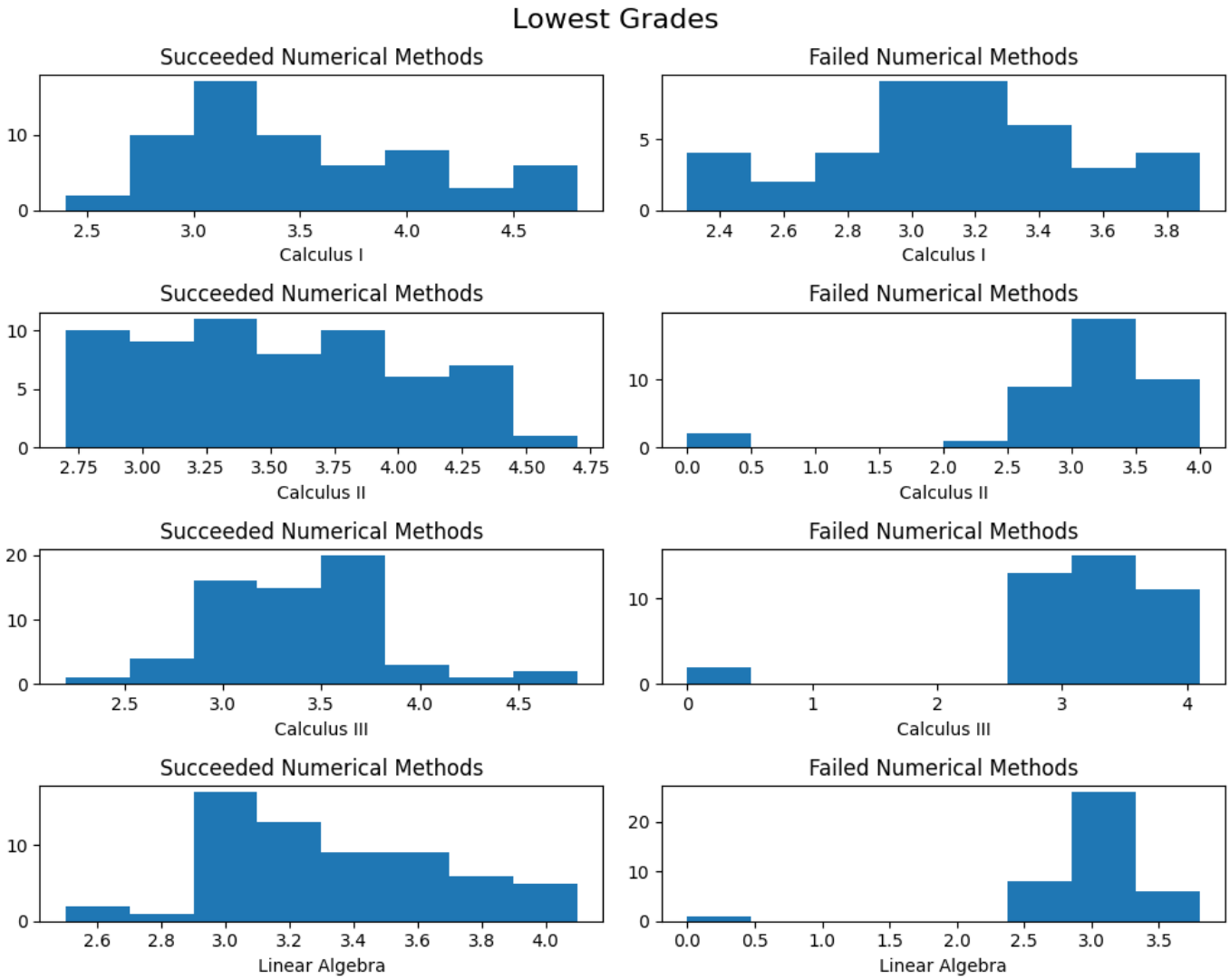

Fig. 5. Histograms displaying the lowest grades obtained in each prerequisite course, categorized based on success or failure in the numerical methods course.

course may benefit from less intensive intervention strategies. For example, providing a variety of instructional approaches, materials, and activities can help to engage these students and address their learning needs effectively. It is also essential to monitor their progress closely and refer them to available support services within the university if necessary, ensuring they receive the assistance required to succeed.

Furthermore, lecturers may leverage probability information to implement differentiated instruction strategies tailored to individual student needs. By understanding the probability of each student failing the course, lecturers may identify areas of weakness and adjust their teaching approach accordingly. For instance, conducting pre-tests and quizzes allows lecturers to determine student comprehension levels and tailor instruction to address specific misconceptions or gaps in understanding. Timely feedback and targeted remediation further support student learning by providing opportunities for reinforcement and mastery of prerequisite competencies. Research in educational psychology has shown that personalized learning approaches may lead to improved student outcomes and engagement [31]. Therefore, by incorporating probability-based insights into instructional planning, instructors can create a more inclusive and effective learning environment for all students.

Utilizing the probability of failing as a metric for student differentiation opens up avenues for fostering collaborative learning environments within the classroom. By identifying at-risk students based on their probability scores, lecturers may orchestrate peer-to-peer instructional sessions and problem-solving activities tailored to address the specific needs of these individuals. This approach facilitates the formation of balanced study groups or teams, where students proficient in prerequisite competences and skills might provide mentorship and support to their at-risk peers. Through collaborative engagement, at-risk students not only receive targeted assistance but also benefit from exposure to diverse perspectives and collective problem-solving, enhancing their overall learning outcomes.

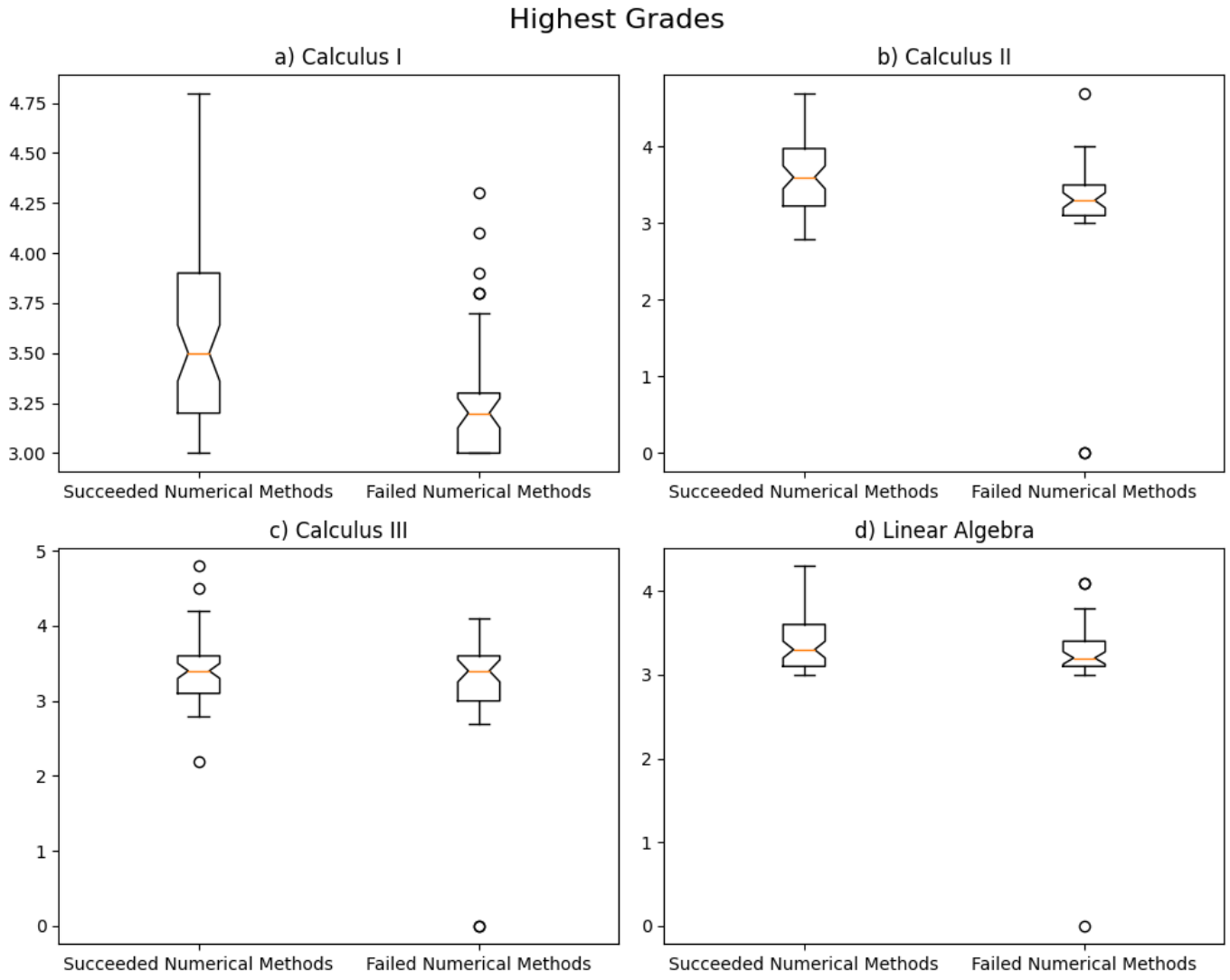

Fig. 6. Boxplots showing the highest grades attained in each prerequisite course, classified based on success or failure in the numerical methods course.

Additionally, at-risk students can benefit from instruction in metacognitive strategies, including techniques for goal-setting, self-monitoring, and reflection. By equipping the student with these cognitive tools, they can enhance their study habits and develop into more strategic and self-regulated learners. These strategies empower the student to take ownership of their learning process, identify areas for improvement, and implement targeted interventions to address challenges they might encounter.

In the context of this study, reporting the probability of failing offers distinct advantages for policymakers and other stakeholders, enabling them to design more effective intervention plans than merely identifying at-risk students in advance. This aspect positions GPRBF as superior to SVMRBF, as the latter does not inherently provide probability information with its predictions. However, this limitation can be addressed by employing Platt scaling [32], which estimates probabilities from the decision values of SVMRBF. Notably, this capability is internally implemented in the Scikit-Learn library.

Incorporating the student's performance in prerequisite mathematics courses, along with their scores on the admission test in this subject, yields improved predictive quality compared to using all scores from the admission test or performance in prerequisite courses related to computer programming or general science (e.g., physics) as input variables. This suggests a significant relationship between the student's proficiency in mathematics and their probability of failing the numerical methods course. It implies that the effectiveness of training in prerequisite mathematics courses such as calculus and linear algebra directly impacts performance in numerical methods.

This relationship between prerequisite mathematics courses and the numerical methods course requires a thorough review of the content covered in prerequisite courses. This review aims to identify key concepts, competencies, abilities, and techniques essential for success in the numerical methods

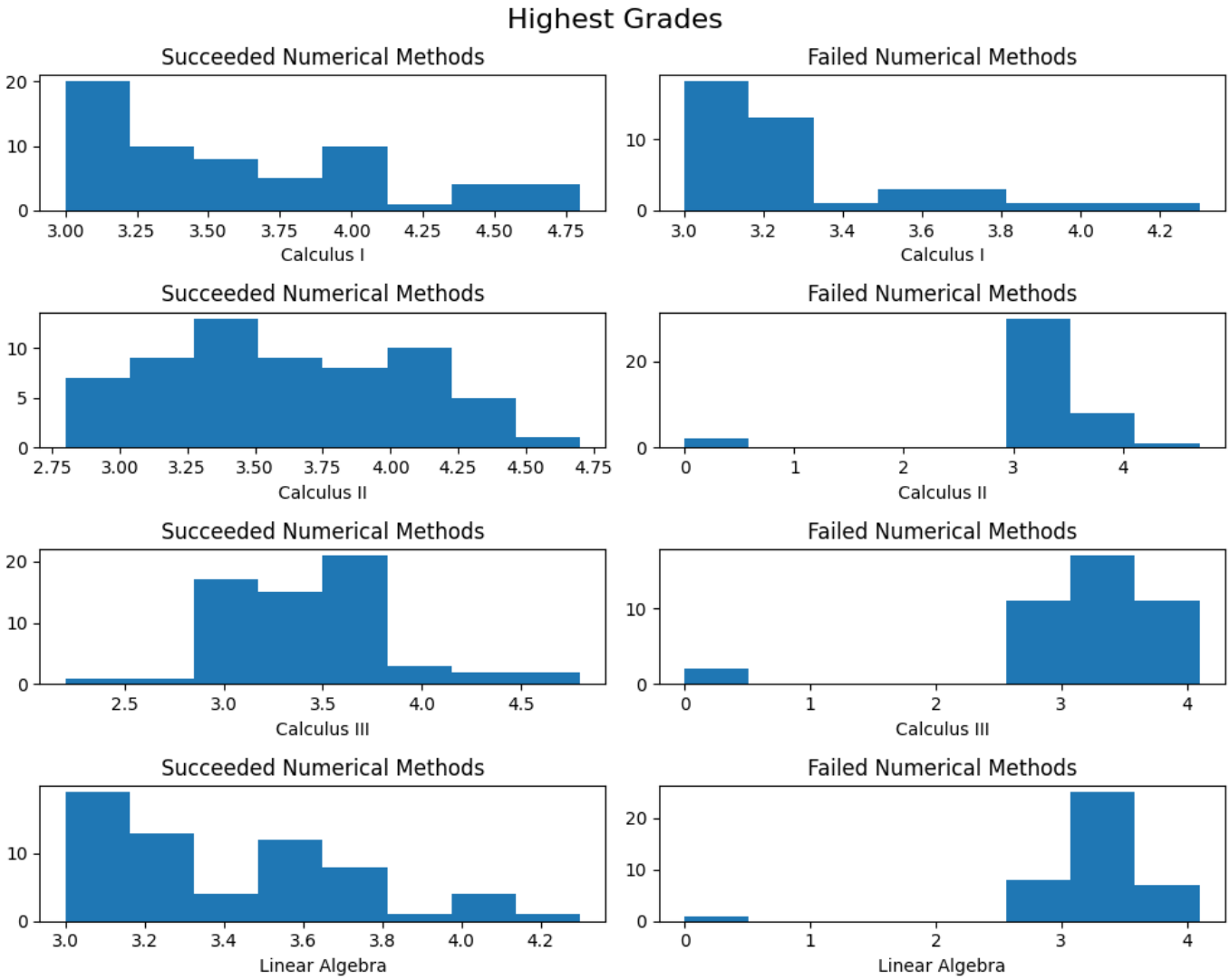

Fig. 7. Histograms illustrating the highest grades achieved in each prerequisite course, categorized based on success or failure in the numerical methods course.

course.

Besides, this process entails identifying common topics, principles, and mathematical techniques that are relevant to both the prerequisite courses and the numerical methods course. This may encompass subjects such as, e.g., differentiation, integration, matrix operations, and so forth. Additionally, it may be beneficial to include in the curriculum, preliminary courses in the first semesters, which introduce foundational mathematical concepts to facilitate the transition from high school to university. Subjects such as set theory, number theory, basic algebra, and analytical geometry might be reviewed to ensure students are adequately prepared.

Finally, this process culminates in mapping the learning outcomes between the numerical methods course and prerequisite courses. This involves articulating the learning outcomes of the numerical methods course and specifying the competences, skills, and knowledge that students are expected to attain. Subsequently, these learning outcomes are aligned with the concepts and goals outlined in the prerequisite mathematics courses.

VI. Conclusions

In conclusion, Gaussian processes for classification with the Radial Basis Function kernel emerged as the top-performing method for the predictive task at hand. Significantly outperforming XGBoost, Adaboost, and Multilayer perceptrons with a single hidden layer, this approach demonstrates superior predictive accuracy. Leveraging machine learning methods, the study forecasts student failure in the numerical methods course based on their performance in prerequisite mathematics courses and admission test scores in the same subject.

The Gaussian processes classification method offers distinct advantages due to its probabilistic nature, providing predictions in the form of probabilities for failing the course. This

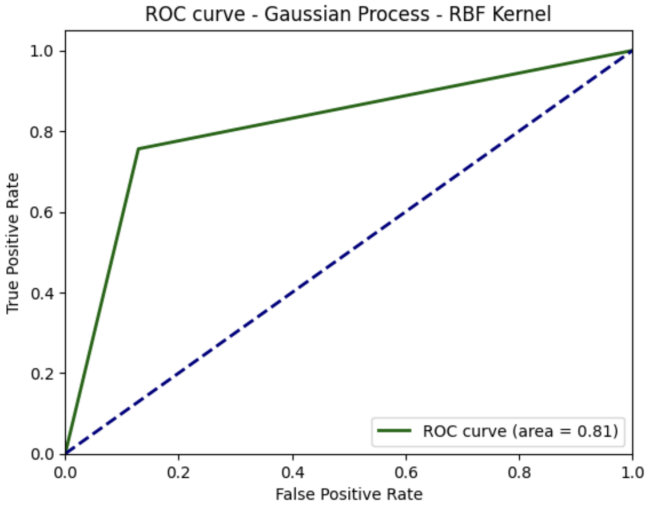

Fig. 8. Receiver operating characteristics (ROC) curve for the Gaussian process with the Radial Basis Function kernel. The diagonal dashed line represents the performance of random guessing, where the classifier performs better than random chance.

feature enables the design of tailored intervention plans for at-risk students while optimizing resource allocation and time efficiency. Future research endeavors could focus on validating intervention plans aimed at supporting students identified as at-risk.

The primary contribution of this work lies in enhancing the prediction quality compared to prior research [1], [2]. This improvement was achieved by modifying the vector representation used in previous studies. Specifically, the focus was narrowed to the student's performance in mathematics, excluding consideration of other subjects such as computer programming and natural science (physics). This underscores the importance of a strong foundation in mathematics for success in the numerical methods course, a finding not previously emphasized in the literature. Nevertheless, further research is recommended to investigate the causal relationship between a student's mathematical skills acquired prior to enrolling in the numerical methods course and their probability of passing it.

Furthermore, the machine learning methodology employed in this study could be expanded to explore the correlation between prerequisite courses across various bachelor's degree programs. This extension has the potential to facilitate curriculum design by mapping the corresponding learning outcomes of prerequisite courses more efficiently.

In future research, we aim to adopt ensemble methods to improve the predictive performance of Gaussian processes. By combining multiple models, such as those with different hyperparameters, kernels, or subsets of the training dataset, we might reduce variance and potentially enhance predictive accuracy. Ensemble techniques may involve averaging predictions of multiple models or using a regression method to weigh their predictions. Although ensemble methods offer the potential for significant improvement in prediction quality, they also introduce computational complexity and require careful hyperparameter tuning.

Acknowledgment

The author thanks the Lord Jesus Christ for blessing this project. He thanks Universidad de Córdoba in Colombia for supporting the Course Prophet Research Project (grant FI-01-22). He also thanks all students who collaborated by answering the survey conducted for collecting the dataset used in this study. Finally, the author thanks the anonymous reviewers for their comments that contributed to improve the quality of this article.

References

[1] I. Caicedo-Castro, M. Macea-Anaya, and S. Rivera-Castaño, "Early Forecasting of At-Risk Students of Failing or Dropping Out of a Bachelor's Course Given Their Academic History - The Case Study of Numerical Methods," in *PATTERNS 2023: The Fifteenth International Conference on Pervasive Patterns and Applications*, ser. International Conferences on Pervasive Patterns and Applications. IARIA: International Academy, Research, and Industry Association, 2023, pp. 40–51.

[2] I. Caicedo-Castro, "Course Prophet: A System for Predicting Course Failures with Machine Learning: A Numerical Methods Case Study," *Sustainability*, vol. 15, no. 18, 2023.

[3] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers and Education*, vol. 53, no. 3, pp. 950–965, 2009.

[4] J. Kabathova and M. Drlik, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *Applied Sciences*, vol. 11, p. 3130, 04 2021.

[5] D. E. M. da Silva, E. J. S. Pires, A. Reis, P. B. de Moura Oliveira, and J. Barroso, "Forecasting Students Dropout: A UTAD University Study," *Future Internet*, vol. 14, no. 3, pp. 1–14, February 2022.

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

[7] V. Čotić Poturić, A. Bašić-Šiško, and I. Lulić, "Artificial neural network model for forecasting student failure in math course," in *ICERI2022 Proceedings*, ser. 15th annual International Conference of Education, Research and Innovation. IATED, 2022, pp. 5872–5878.

[8] S. Zihan, S.-H. Sung, D.-M. Park, and B.-K. Park, "All-Year Dropout Prediction Modeling and Analysis for University Students," *Applied Sciences*, vol. 13, p. 1143, 01 2023.

[9] I. Caicedo-Castro, "Quantum Course Prophet: Quantum Machine Learning for Predicting Course Failures: A Case Study on Numerical Methods," in *26th International Conference on Human-Computer Interaction*. Springer, July 2024, to appear.

[10] V. Čotić Poturić, I. Dražić, and S. Čandrlić, "Identification of Predictive Factors for Student Failure in STEM Oriented Course," in *ICERI2022 Proceedings*, ser. 15th annual International Conference of Education, Research and Innovation. IATED, 2022, pp. 5831–5837.

[11] S. Merchán-Rubiano, A. Beltrán-Gómez, and J. Duarte-García, "Formulation of a predictive model for academic performance based on students' academic and demographic data," in *IEEE Frontiers in Educacion Conference*, Texas, USA, 2015, pp. 1–7.

[12] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," *CoRR*, vol. abs/1810.11363, 2018.

[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[14] I. Caicedo-Castro, M. Macea-Anaya, and S. Castaño-Rivera, "Forecasting Failure Risk in Early Mathematics and Physical Science Courses in the Bachelor's Degree in Engineering," in *IARIA Congress 2023: International Conference on Technical Advances and Human Consequences*. International Academy, Research, and Industry Association, 2022, pp. 177–187.

[15] I. Pacheco-Arrieta *et al.*, "Agreement No. 004: Student's code at the University of Córdoba in Colombia," http://www.unicordoba.edu.co/wp-content/uploads/2018/12/reglamento-academico.pdf [retrieved: February, 2024], 2004.
[16] C. Williams and C. Rasmussen, "Gaussian Processes for Regression," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds., vol. 8. MIT Press, 1995, pp. 514–520.
[17] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.
[18] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. The MIT Press, 2018.
[19] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.
[20] C. C. Aggarwal, *Neural Networks and Deep Learning*. Springer, 2018.
[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," http://arxiv.org/abs/1412.6980 [retrieved: February, 2024], 2014.
[23] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.
[24] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *ICML*, vol. 96, 1996, pp. 148–156.
[25] L. Breiman, "Random forests," in *Machine Learning*, vol. 45, no. 1. Springer, 2001, pp. 5–32.
[26] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.
[27] D. Wolpert and W. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, April 1997.
[28] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[29] "Google Colaboratory," https://colab.research.google.com/ [retrieved: February, 2024], 2004.
[30] R. Ajjawi, D. Boud, N. Zacharias, M. Dracup, and S. Bennett, "How Do Students Adapt in Response to Academic Failure?" *Student Success*, vol. 10, pp. 84–91, 12 2019.
[31] E. T. Khor and M. K, "A Systematic Review of the Role of Learning Analytics in Supporting Personalized Learning," *Education Sciences*, vol. 14, no. 1, 2024.
[32] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.