Forecasting Failure Risk in Early Mathematics and Physics Courses of a Bachelor's in Engineering Degree

Isaac Caicedo-Castro*†‡, Mario Macea-Anaya†‡§, Samir Castaño-Rivera*‡
*Socrates Research Team
†Research Team: Development, Education, and Healthcare
‡Faculty of Engineering
§CINTIA, Centre of INnovation in Technology of Information to support the Academia
University of Córdoba
Montería, Colombia
emails: {isacaic, mariomacea, sacastano}@correo.unicordoba.edu.co

***Abstract*—In this research, we study the functional mapping between university admission test scores and the risk of failing in initial mathematics and physics courses for students embarking on a Bachelor's degree in Systems Engineering. We assume that the admission test assesses students' competence and proficiency in natural sciences and mathematics, essential prerequisites for success in the foundational courses of this Systems Engineering program. A deficiency in these subjects might result in failure, leading to dropouts or an extended degree completion timeline. We harnessed machine learning techniques to probe this issue, focusing on the landscape of Colombian universities, specifically analysing the Systems Engineering program at the University of Córdoba. In this Colombian educational context, universities, including our case study institution, rely on the national standardized admission test known as Saber 11 to evaluate candidates for Bachelor's degree programs. By adopting machine learning methods to unveil underlying patterns that govern this functional mapping, we might proactively identify students at risk of struggling in the aforementioned courses based on their admission test scores. Early identification of these at-risk students opens the opportunity to pre-emptive measures, such as offering preparatory courses to fortify their prerequisites for success in these challenging subjects. Our research involved the examination of academic records from 56 anonymized students, using both 10-fold and 5-fold cross-validation. The outcomes from the 10-fold cross-validation reveal that the support vector machine method yields mean values of 71.33% for accuracy, 68.33% for precision, 60% for recall, and 62.05% for the harmonic mean (F_1). Therefore, we conclude that this method outperforms the others studied in this work.**

***Keywords*—machine learning; quantum machine learning; educational data mining; supervised learning; classification methods; failure forecasting.**

I. Introduction

This study is part of a broader research project called Course Prophet, whose goal is to design and implement an intelligent system to predict the risk of undergraduate students failing or dropping a course in the area of scientific computing in systems engineering at the University of Córdoba in Colombia. Scientific computing involves the use of mathematical models and computer simulations to solve complex engineering problems, such as, e.g., numerical methods and linear (or non-linear) programming. Our focus in this study is on predicting whether students are at risk of failing the foundational first courses of mathematics and physics in scientific computing, which are critical for success in advanced courses. The forecasting is based on the student's admission test outcomes. Early identification of at-risk students by the Course Prophet system has the potential to improve retention rates and support targeted interventions that enhance student success.

We delve into the details of the problem addressed in this study in Section I-A, while we motive this work in Section I-B. The assumptions and limitations of this work are presented in Section I-C. Finally, we summarize our contributions and outline the remainder of this paper in Section I-D.

A. Problem Statement

It is assumed that the high school experience and education process prepare college students to succeed in the endeavour of attaining an undergraduate degree. Nevertheless, other factors might influence their success in university, such as, e.g., personal circumstances, study habits, motivation, and so forth. Having a strong foundation in mathematics and natural sciences, particularly physics, might increase a student's chances of success in pursuing a Bachelor's degree in Systems Engineering. Therefore, it is important for the admission test to assess these subjects for candidates applying to the Bachelor's degree in Engineering.

Since 1968, the Saber 11 has been the standardized test used in Colombia to assess the competencies of high school students who are about to graduate. This test has been designed to be the official admission test for pursuing a Bachelor's degree in Colombian universities [1]. The same way as the Scholastic Assessment Test (SAT) is used for the same purpose in the United States. This study is focused on Systems Engineering students at the University of Córdoba in Colombia. Under Article 17 of the student code, candidates are admitted to the University of Córdoba based on their Saber 11 test scores [2].

The test Saber 11 evaluates five subjects: (i) mathematics, (ii) natural science, (iii) critical reading, (iv) social sciences, and (v) English language. The Colombian education ministry assumes these subjects are the foundation that all high school students must have learnt properly to pursue a bachelor's degree.

Thus, students with the highest scores in mathematics and natural science are better suited for engineering and science

 ISBN: 978-1-68558-089-6

undergraduate bachelor's programs. Proficiency in other subjects might also be beneficial for success at the university. For instance, students with good critical reading skills and proficiency in the English language may be better equipped to learn any topic and access a wider range of literature sources compared to students who have poor skills in these subjects.

Therefore, the research question to be addressed in this study is as follows: is it possible to forecast if a student is at risk of failing mathematics and physics courses in Bachelor's degree in Systems Engineering based on their scores in the admission test called Saber 11?

The reason to focus the study on the first courses in mathematics and physics is twofold. Firstly, these courses are typically more challenging than others in the Bachelor's degree in Systems Engineering. Secondly, these courses form the foundation of scientific computing, which is the primary focus of the project that this study is a part of. Dealing with other courses in natural sciences, such as chemistry or biology, is beyond the scope of this study, as these subjects are not included in the curriculum of the systems engineering major. In other words, the problem addressed in this study is finding the functional mapping between the student's risk of failing early mathematics and physics courses, which is the target variable, and the scores achieved by the student in each subject evaluated in the admission test, which are the independent variables or the student's features.

B. Motivation

Failing early courses in mathematics and physics causes several negative consequences, such as, e.g., students feeling demotivated to continue pursuing a Bachelor's degree in Systems Engineering, wasted financial resources, frustration, stress, or even losing student status due to a low overall grade, for instance, students at the University of Córdoba must maintain at least an overall grade point average (GPA) of 3.3, where is in the range of 0 to 5 in every Colombian university (see also the student's code [2]). This problem is commonly referred to as *student dropout*.

On the other hand, those students who dropout courses might take longer to fulfill the requirements to receive their Bachelor's degree. This problem is known as *long-term retention*.

Knowing in advance who are the students at risk of failure, allows the universities to take precautions to prevent those students from failing the first courses in mathematics and physics, which usually are the most challenging ones. For instance, those students at risk might attend preliminary courses to improve their proficiency in those subjects that are prerequisites to pass actual early university courses.

If the university helps the population at risk, eventually, students' dropout and long-term retention rates might decrease, considering that both problems are a serious concern in the higher education systems and for policy-making stakeholders at universities [3].

C. Key Assumptions and Limitations

In this study, we have considered the following key assumptions:

(i) We have assumed the Saber 11 test measures the knowledge and competencies required for pursuing a bachelor's degree, as stated in Article 17 of the University of Córdoba's student code, which states that admission is based on a candidate's Saber 11 global score [2].

(ii) We have assumed the student at academic risk fails one or more courses about mathematics or physics during the first term. The early courses about mathematics are Calculus I and Linear Algebra, while the Physics I is the first course about physical science. Failing courses will lower the student's overall grade and potentially affect their academic standing.

(iii) We have assumed the student at academic risk has an overall grade lower or equal to 3.3, which is the minimum requirement for maintaining the student status according to the student's code at University of Córdoba (cf., Article 16 in [2]). Bachelor's students at Colombian universities are graded in the range from 0 up to 5. If a student's overall grade falls between 3 and 3.3, they must improve it to at least 3.3 in the next semester, or risk being expelled, per Article 28. Any student who obtains an overall grade below 3 will be forced to withdraw from the University of Córdoba.

(iv) We have assumed that wrongly classifying students as being at risk when they are not (i.e., false positive) is just as problematic as failing at classifying students who are actually at risk (i.e., false negative). In the former case, both the students and the university might waste resources addressing an unfounded risk. In the latter case, students might not receive the support they need for succeeding in their studies, and the university might miss the opportunity to take the required precautions and help them stay on track.

(v) We assumed that each student is represented through a vector in a real-valued multidimensional euclidean space, where each entry of the vector corresponds to a Saber 11 score in a specific subject.

The limitations of this study are as follows:

(i) We did not aim at designing an artificial intelligent system that predicts the dropout rate nor the failure rate of a given course.

(ii) We did not consider additional input variables for the prediction, such as, e.g., gender, ethnicity or economic variables, because the students who took the survey are alike regarding these features. Figure 1 shows an evidence that most of the sampled students are male, do not consider themselves part of an ethnic group, belong to the first economical stratum, and more than half of the sample of the students' families earn less than two Colombian monthly minimum wages. Therefore, these features do not help to differentiate students, contributing little information to the forecasting process.

 ISBN: 978-1-68558-089-6

Furthermore, we are interested in studying the extent the admission test contributes to accurate forecasting.

D. Contributions and Paper Outline

The contributions of this research are as follows:

(i) A data set with 56 records, where each one contains the student's profile and academic history. These students have completed courses from their second to the ninth semester. Additionally, each record includes the student's score in every subject from the admission test.

(ii) The prototype of an intelligent system, written in Python, that forecasts if a recently admitted student might be at academic risk of failing any of the early courses in mathematics or physics, namely Linear Algebra, Calculus I, and Physics I.

(iii) An empirical study that reveals Support Vector Machine (SVMs) outperform the other evaluated classifiers in forecasting students' failure risk. During the evaluation through 10-fold cross-validation, SVMs achieved the mean values for accuracy, precision, recall, and harmonic mean (F_1) of 71.33%, 68.33%, 60%, and 62.05%, respectively.

The rest of this paper is organized as follows: in Section II, we review the literature and related work, whereas in Section III we describe the research methods adopted in this study. We present and discuss the results of this research in Section IV. Finally, we draw the conclusions and outline the directions for further research in Section V.

II. Prior Research

This study falls within the domain of educational data mining, which aims to apply machine learning methods to educational data sets to gain insights into students' learning behaviour. This includes the analysis of data, the exploration of pedagogical theories through data mining, understanding students' domain knowledge, and evaluating their engagement in learning tasks.

Related research endeavours have focused on predicting whether a student is at risk of failing or dropping out of a course based on their performance in prerequisite courses [4] [5] [6].

Prior research has used SAT scores to predict if students will withdraw from their bachelor's program [7] [8]. One approach to predict student withdrawals from bachelor's programs involves using SAT scores and first-year university performance as input data [7]. Unfortunately, predicting student withdrawals after the first year of university does not provide with insight into their long-term retention issues. Another similar approach also includes both demographic information and pre-university performance as input data in order to forecast student withdrawals [8]. While the previously-mentioned research studies share similarities with ours, our specific goal is to predict the risk of students failing their first courses in mathematics and physics based on their admission test scores.

Predicting the risk of bachelor's student withdrawal has also been based on factors such as the student's school performance [9] [10], cognitive abilities [9], and even measurements of emotional intelligence [11]. It is worth noting that the admission test has not been considered in the last two mentioned studies. In one study [9], forecasting accuracy is reported as unfeasible, while in another study [10], the prediction model is tailored to a specific context, making it non-reproducible in other contexts, such as Colombia.

In the Colombian context, a study has been conducted to predict bachelor's student withdrawal based on their academic and personal data [12]. This study focused on students enrolled in the bachelor's program of engineering at the University of Los Andes, majoring in systems engineering. Unfortunately, the results of this study cannot be reproduced because the collected data set is not publicly available. Forecasting the individual students' risk of withdrawal is more useful for decision-making and addressing at-risk students compared to simply predicting the overall withdrawal rate.

In another study conducted in the Colombian context, Saber 11 scores from four out of five subjects (excluding natural sciences) were used to predict the risk of withdrawal or long-term retention faced by recently admitted bachelor's students. The mean prediction accuracy was 72.5% based on a 10-fold cross-validation using a data set of 47 records collected from a survey of 86 systems engineering students at the University of Córdoba. For further details, please refer to [13].

To our knowledge, so far no prior research has aimed at predicting the student's risk of failing an early course of mathematics and physics given their outcomes in the admission test, which is the goal of our study.

III. Research Methods

The research methodology adopted in this study is quantitative. We collected a data set with 81 observations or records by conducting a survey using Google Forms. The survey was administered to students pursuing a Bachelor's degree in Systems Engineering at the University of Córdoba in Colombia during the second half of 2022. The participating students had completed courses from the second to the ninth semester. Detailed information about the data set is provided in Section III-A.

Once the data set has been collected, we applied machine learning methods to address the problem posed in this study, specifically classification methods, which are supervised learning algorithms. The evaluation of the classifiers used in this study is carried out with consideration that machine learning is an experimental discipline. We discuss these classification methods in Section III-B, and the evaluation approach is described in Section III-C.

A. Data set Description

Let $\mathcal{D}$ be the data set, defined as $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{Z}^D \wedge y_i \in 0, 1, \forall i = 1, \ldots, n\}$, where n and D represent the number of observations and independent variables, respectively. The resulting data set contains 56 observations out of the original 81 due to changes in the curriculum structure of the undergraduate program in 2018, i.e., $n = 56$. In this

 ISBN: 978-1-68558-089-6

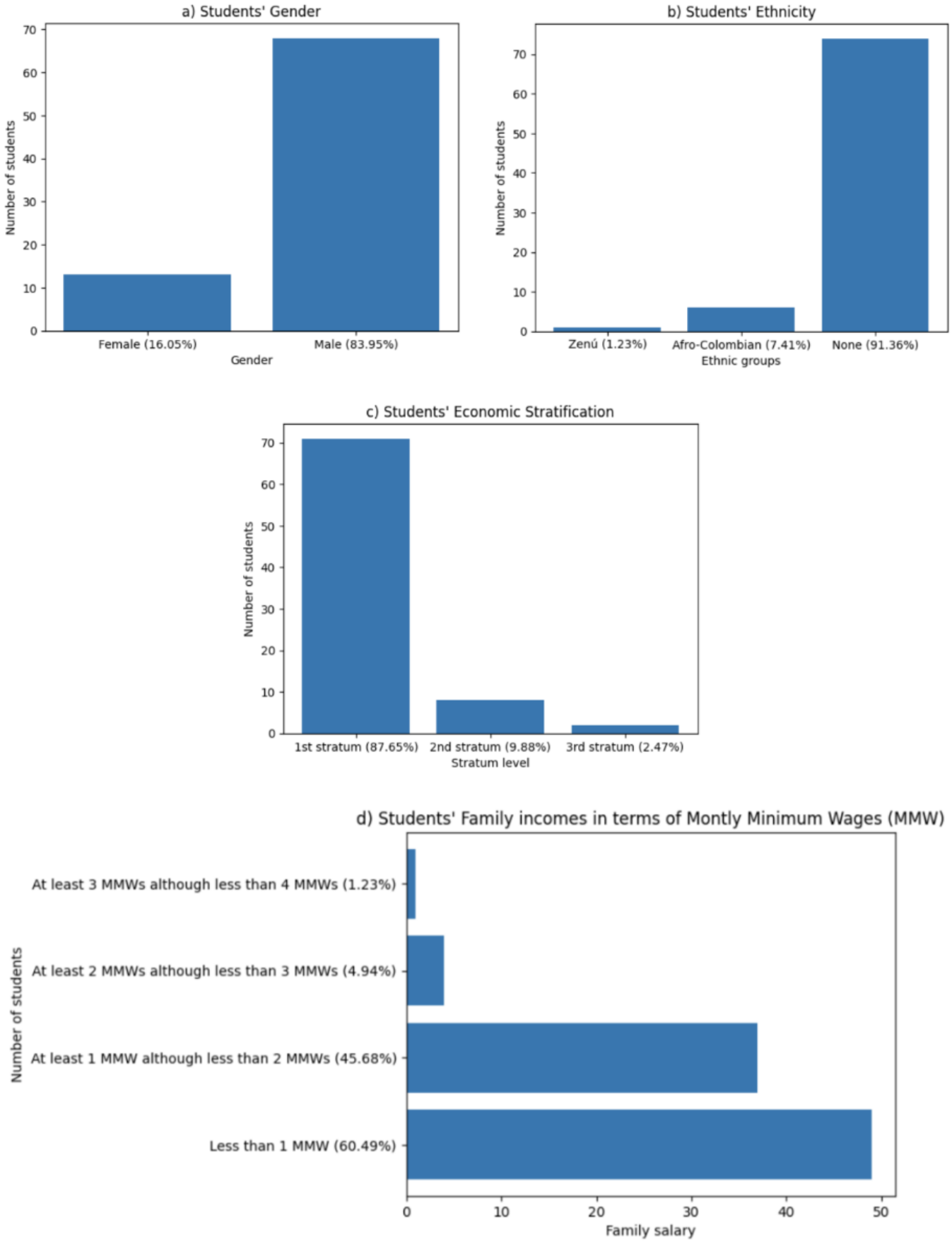

Figure 1. Sample distribution according to a) gender, b) ethnicity group, c) economical stratification, and d) family incomes.

 ISBN: 978-1-68558-089-6

context, the D-dimensional vector $\mathbf{x}_i$ represents the features or independent variables of the ith student, while y_i represents their corresponding target variable.

The target variable has one out of two possible values, i.e., $y_i = 1$ if the ith student has failed at least one of the early courses in mathematics or physics the first time the student enrolled them. These courses are calculus I, linear algebra, and physics I. In contrast, $y_i = 0$ otherwise.

On the other hand, there are five independent variables, where $D = 5$, representing the scores achieved by the ith student in each subject evaluated in the admission test. These scores range from 0 to 100. For a given ith student, the meaning of each component of their vector representation is explained as follows:

- x_{i1} is the score achieved by the ith student in the mathematics subject of the admission test.
- x_{i2} is the score achieved by the ith student in the natural science subject of the admission test.
- x_{i3} is the score achieved by the ith student in the social science subject of the admission test.
- x_{i4} is the score achieved by the ith student in the critical reading subject of the admission test.
- x_{i5} is the score achieved by the ith student in the social English proficiency evaluation of the admission test.

The proportion of classes is rather balanced in the data set, as it is illustrated in Figure 2.

The data set is available online to allow the reproduction of our study, and for further research [14].

B. Classification Methods

To find the functional mapping between the risk of failure and the performance in the admission test, we adopted supervised learning algorithms, specifically classification methods or classifiers. These algorithms identify patterns between students who have either failed or passed courses and their respective scores in every subject evaluated in the admission test. Formally, given the data set $\mathcal{D}$, the goal is to estimate the prediction function $g : \mathbb{R}^D \to \{0, 1\}$, such that $g(\mathbf{x}_i) \approx 1$ indicates that the function predicts the ith student is at risk, and $g(\mathbf{x}_i) \approx 0$ signifies otherwise.

To tackle the aforementioned problem, we used several classification methods, including Gaussian Process (GP). GP gets its name from the fact that it assumes the probability distribution of the target variable is Gaussian or normal [15] [16]. As a result, GP calculates the student's probability of failure risk, which is valuable for interpreting its forecasting outcome. One of the main advantages of GP is its ability to incorporate prior knowledge about the problem, improving forecasts even when the training data set is small. Another advantage is its suitability for solving non-linear classification problems. However, GP has the drawback of potentially high computational costs for fitting and forecasting, which can be problematic for large-scale data sets. In the context of this study, our data set is relatively small, so we chose to use this method, considering its advantages.

So far, Support Vector Machine (SVM) is considered one of the best theoretically motivated classification methods and amongst the most successful in the practice of modern machine learning [17, pg. 79]. Its objective function is convex, allowing for the discovery of a global maximum solution, which is its primary advantage. However, SVM is not particularly well-suited for interpretation in data mining, although it excels in training accurate intelligent systems. For a more comprehensive description of this algorithm, refer to the work by Cortes and Vapnik [18].

SVM is a linear classification method that assumes the input vector space is separable through a linear decision boundary or a hyperplane in multidimensional space. However, when this assumption is not satisfied, SVM can be used with kernel methods to handle non-linear decision boundaries. For further details, see Cortes and Vapnik [18].

In this study, we incorporated the Quantum Support Vector Machine (QSVM) method, which makes use of kernel methods. Our approach centres around the utilization of a quantum state space for the independent variables, as outlined in [19]. To achieve this, we employed the ZZ feature mapping, a well-implemented feature mapping in Qiskit, a prominent open-source software development kit. This mapping allows us to encode D input variables across D qubits. Qiskit provides a comprehensive toolkit with a wide range of quantum gates and circuits designed for various computational purposes [20]. For a deeper exploration of the ZZ feature mapping, we recommend consulting the documentation available on the Qiskit website [21]. In the context of qubit representation as normalized complex-value space vectors, we individually rescaled each variable, ensuring that the maximum value for each variable was standardized to 1.

We adopted the decision tree classifier, a commonly used model in data mining and knowledge discovery due to its tree-shaped hierarchical structure, which is easily interpreted and used for decision support. During training, a tree is created using the data set as input, with each internal node representing a test on an independent variable, branches representing the results of the test, and leaves representing estimated classes. The construction of the tree is carried out recursively, starting with the entire data set 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 can be stopped based on various criteria, such as when all the training data is classified or when the accuracy or performance of the classifier cannot be further improved.

The main drawback of decision trees arises from their fitting process, which relies on heuristic algorithms, such as greedy algorithms. These algorithms may lead to several local optimal solutions at each node, which is one of the reasons why there is no guarantee that the learning algorithm will converge to the most optimal solution. As a result, decision trees can exhibit different tree shapes due to small variations in the training data set. Breiman *et al*. introduced this method in 1984 [22]. Finally, we also adopted ensemble methods based on multiple decision trees such as, e.g., Adaboost (stands for adaptive

 ISBN: 978-1-68558-089-6

Figure 2. Distribution of students who have either failed or passed at least one of the early courses in Mathematics and Physics: calculus I, linear algebra, and physics I.

boosting) [23], Random forest [24], and XGBoost [25].

C. Evaluation Approach

We cannot determine in advance which machine learning method outperforms the others, as the *no free lunch theorem* states. Therefore, we need to conduct experiments to evaluate the quality of the machine learning methods, requiring multiple pairs of training and test data sets. To this end, we conducted experiments based on K-Fold Cross-Validation (KFCV), resulting in K pairs of training and test data sets derived from the original one. We selected values of $K = 10$ and $K = 5$, although K is typically set to 10 or 30. We opted for $K = 5$ in lieu of $K = 30$ due to the relatively small data set. Consequently, we tested each method K times using KFCV. Based on the test outcomes, we calculated the mean accuracy, mean precision, mean recall, and the average of the harmonic mean (F_1) to compare the learning methods and choose the best hyper-parameters for each of the previously mentioned methods.

IV. Evaluation

In this section, we delve into the details of the test bed and results obtained through the evaluation of the aforementioned classification methods. The experimental setting is explained in Section IV-A and Section IV-B presents and discusses the results of the evaluation.

A. Experimental Setting

The evaluation is conducted through K-Fold Cross-Validation, where $K = 10$ and $K = 5$, as it was mentioned in Section III-C. This procedure is performed on a data set containing 56 records or examples, each having 5 independent variables and the corresponding target variable (as described in Section III-A)..

We adopted Python to write the source code of the test beds and experiments, moreover, we used Scikit-Learn library [26], Google Colaboratory [27], Qskit library, and the quantum computing simulator called Aer [20].

The best hyper-parameter setting resulting from applying 10-fold cross-validation to tune each method is presented as follows:

- Gaussian Process (GP) with the radial basis function kernel, where the best values for σ and γ are 16 and 19, respectively. Both hyper-parameters are part of the following equation $k_G(\mathbf{x}_i, \mathbf{x}_j) = \gamma \exp(-||\mathbf{x}_j - \mathbf{x}_i||^2/2\sigma^2)$.
- GP with the Matern kernel, where the best values for nu, σ and γ are 1.3, 4 and 3.8×10^{-6}, respectively. These hyper-parameters belongs to the following equation $k_M(\mathbf{x}_i, \mathbf{x}_j) = \gamma \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}||\mathbf{x}_j - \mathbf{x}_i||^2}{\sigma}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}||\mathbf{x}_j - \mathbf{x}_i||^2}{\sigma}\right)$, where $K_{\nu}(\cdot)$ and $\Gamma(\cdot)$ are the modified Bessel function and the gamma function, respectively.
- GP with the dot product kernel, which is defined as follows: $k_d(\mathbf{x}_i, \mathbf{x}_j) = 1 + \langle \mathbf{x}_i, \mathbf{x}_j \rangle$.
- GP with the rational quadratic kernel, where σ and α are 1.56×10^{-2} and 6.1×10^{-5}, respectively. The kernel is defined as follows: $k_r(\mathbf{x}_i, \mathbf{x}_j) = (1 + ||\mathbf{x}_j - \mathbf{x}_i||^2/(2\alpha\sigma^2))^{-\alpha}$
- Support Vector Machines (SVM) with the radial basis function kernel, where γ and C are 1.22×10^{-4} and 65536, respectively. In this case, the kernel is defined as follows: $k_G(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_j - \mathbf{x}_i||^2)$.
- SVM with the polynomial kernel, where d (degree) and C are 4 and 7.8×10^{-3}, respectively. The kernel is defined as follows: $k_p(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^d$.
- SVM with the sigmoid kernel, where γ and C are 1.22×10^{-4} and 16, respectively. The kernel is defined as follows: $k_s(\mathbf{x}_i, \mathbf{x}_j) = \tanh \gamma \langle \mathbf{x}_i, \mathbf{x}_j \rangle$.
- Quantum SVM, where C is 12 and we adopted the full entanglement strategy, i.e., the qubits are entangled to each other.
- The decision trees were fitted using both the Gini and entropy indexes. The parameters used were given by default in Scikit-Learn API.
- XGBoost algorithm was fitted with a learning rate,

 ISBN: 978-1-68558-089-6

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*
Support Vector Machines with the polynomial kernel (degree = 4)	**71.33**		68.33		60		**62.05**	
Support Vector Machines with the sigmoid kernel	62.67	0.26	56.67	0.51	38.33	0.14	42.67	0.16
Support Vector Machines with the radial basis function kernel	65.33	0.48	**70.17**	0.9	58.33	0.9	58.07	0.74
Quantum Support Vector Machines	62	0.78	55	0.64	56.67	0.84	53.38	0.93
Decision tree with entropy index	46.67	**0.003**†	38.17	0.05	46.67	0.39	38.1	0.07
Decision tree with gini index	57.33	**0.04**†	49	0.21	45	0.27	42.76	0.11
Gaussian Process with the rational quadratic kernel	64	0.23	46.67	0.23	30	**0.03**†	33.67	0.05
Gaussian Process with the dot product kernel	44	**0.01**†	28.33	**0.02**†	21.67	**0.01**†	23.33	**0.01**†
Gaussian Process with the Matern kernel	58.67	0.07	60	0.57	45	0.24	46.67	0.16
Gaussian Process with the radial basis function kernel	62	0.12	61.67	0.65	48.33	0.37	49	0.24
Random forest with the gini index	60	0.17	63.5	0.73	**61.67**	0.89	56.5	0.59
Adaboost with the entropy index	50.33	0.23	35.67	0.15	51.67	0.43	40.43	0.79
XGBoost	50.33	**0.03**†	44.83	0.18	38.33	0.14	37.33	0.07

†Student's paired t-test reveals the difference between means is statistically significant

maximum depth, and number of estimators equal to 6.25×10^{-2}, 6, and 50, respectively. Besides, we used the entropy index in the trees.
- Adaboost algorithm was fitted with a learning rate and number of estimators equal to 0.5 and 170, respectively. Besides, we used the entropy index in the trees.
- Random forest was fitted with 15 trees (with gini index), at least one sample per leaf, at most five samples per split, and a maximum depth of fifth levels.

The best hyper-parameter setting resulting from applying 5-fold cross-validation to tune each method is presented as follows:

- GP with the radial basis function kernel, where the best values for σ and γ are 4 and 1.52×10^{-5}, respectively.
- GP with the Matern kernel, where the best values for nu, σ and γ are 1.3, 4 and 3.8×10^{-6}, respectively.
- GP with the rational quadratic kernel, where σ and α are 1 and 1.22×10^{-4}, respectively.
- SVM with the radial basis function kernel, where γ and C are 1.22×10^{-4} and 16384, respectively.
- SVM with the polynomial kernel, where d (degree) and C are 4 and 3.9×10^{-3}, respectively.
- SVM with the sigmoid kernel, where γ and C are 6.1×10^{-5} and 64, respectively.
- Quantum SVM, where C is 8 and we adopted the full entanglement strategy.
- The decision trees were fitted using both the Gini and entropy indexes. The parameters used were given by default in Scikit-Learn API.
- XGBoost algorithm was fitted with a learning rate, maximum depth, and number of estimators equal to 0.5, 5, and 80, respectively. Besides, we used the entropy index in the trees.
- Adaboost algorithm was fitted with a learning rate and number of estimators equal to 0.5 and 170, respectively. Besides, we used the entropy index in the trees.
- Random forest was fitted with 15 trees (with gini index), at least two sample per leaf, at most five samples per split, and a maximum depth of eighth levels.

B. Results and Discussion

According to the evaluation conducted through K-Fold Cross-Validation (KFCV) with both K = 10 (10FCV) and K = 5 (5FCV), Support Vector Machines (SVMs) consistently

TABLE II
FIVE-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*
Support Vector Machines with the polynomial kernel (degree = 4)	**71.82**		65.33		**60**		**60.9**	
Support Vector Machines with the sigmoid kernel	58.94	0.26	52.29	0.29	56	0.85	53.56	0.67
Support Vector Machines with the radial basis function kernel	67.88	0.71	**69.67**	0.76	56	0.83	60.88	0.99
Quantum Support Vector Machines	65.91	0.96	65	0.84	56	0.99	59.80	0.91
Decision tree with entropy index	53.33	0.07	49.9	0.2	44	0.38	45.78	0.31
Decision tree with gini index	51.52	0.07	49.33	0.23	40	0.27	43.27	0.25
Gaussian Process with the rational quadratic kernel	62.58	0.32	38.43	0.19	40	0.42	38.67	0.32
Gaussian Process with the dot product kernel	46.67	0.05	40	0.23	20	0.05	26.03	0.07
Gaussian Process with the Matern kernel	60.91	0.31	55	0.46	44	0.45	47.88	0.47
Gaussian Process with the radial basis function kernel	62.58	0.37	57	0.52	52	0.69	53.77	0.68
Random forest with the gini index	55	0.12	56	0.55	44	0.37	47.23	0.37
Adaboost with the entropy index	51.52	0.06	48	0.17	48	0.53	46.47	0.35
XGBoost	55	0.12	56	0.55	44	0.37	47.23	0.37

†Student's paired t-test reveals the difference between means is statistically significant

outperformed the other machine learning methods in nearly every measure. In the case of 10FCV, SVMs with the polynomial kernel excelled in terms of accuracy and the harmonic mean (F_1), while SVMs with the radial basis function achieved the highest mean precision.

On the other hand, Random Forest (RF) achieved a better mean recall. However, SVM with the polynomial kernel attained the third-best mean recall, and RF achieved the third-best mean precision. This explains why the SVM with the polynomial kernel outperformed the others in the harmonic mean. Besides, it reached the second-best values in both mean precision and mean recall. The results obtained through 10FCV are presented in Table I.

Table II shows the results of the 5FCV, where SVM performs better than the other machine learning methods in every measured metric. The outcomes are consistent with those achieved through 10FCV because, in both kinds of experiments, the trend reveals that SVM outperforms the other evaluated methods. SVM with the radial basis function outperformed SVM with the polynomial kernel in terms of mean precision, although the latter method is better in the other metrics and obtains the second-best place in terms of mean precision.

TABLE III
CONFUSION MATRIX FOR SUPPORT VECTOR MACHINES WITH THE POLYNOMIAL KERNEL DURING K-FOLD CROSS-VALIDATION WITH K = 10 AND K = 5

	Predicted class		
True class	*Student without risk*	*Student at risk*	*Total*
Student without risk	25	6	31
Student at risk	10	15	25
Total	35	21	56

Based on the results of 10FCV, there is strong statistical evidence (p-value < 0.05) that SVM with the polynomial kernel is more accurate than decision trees, GP with the dot product kernel, and XGBoost, with an accuracy score of 71.33%. Additionally, the same results demonstrate strong statistical evidence that the precision and harmonic mean of SVM with the polynomial kernel are greater than those achieved through the predictions of GP with the dot product kernel. Furthermore, the results also indicate statistical evidence that the recall of SVM with the polynomial kernel is greater than that achieved through the predictions of GP with the rational

 ISBN: 978-1-68558-089-6

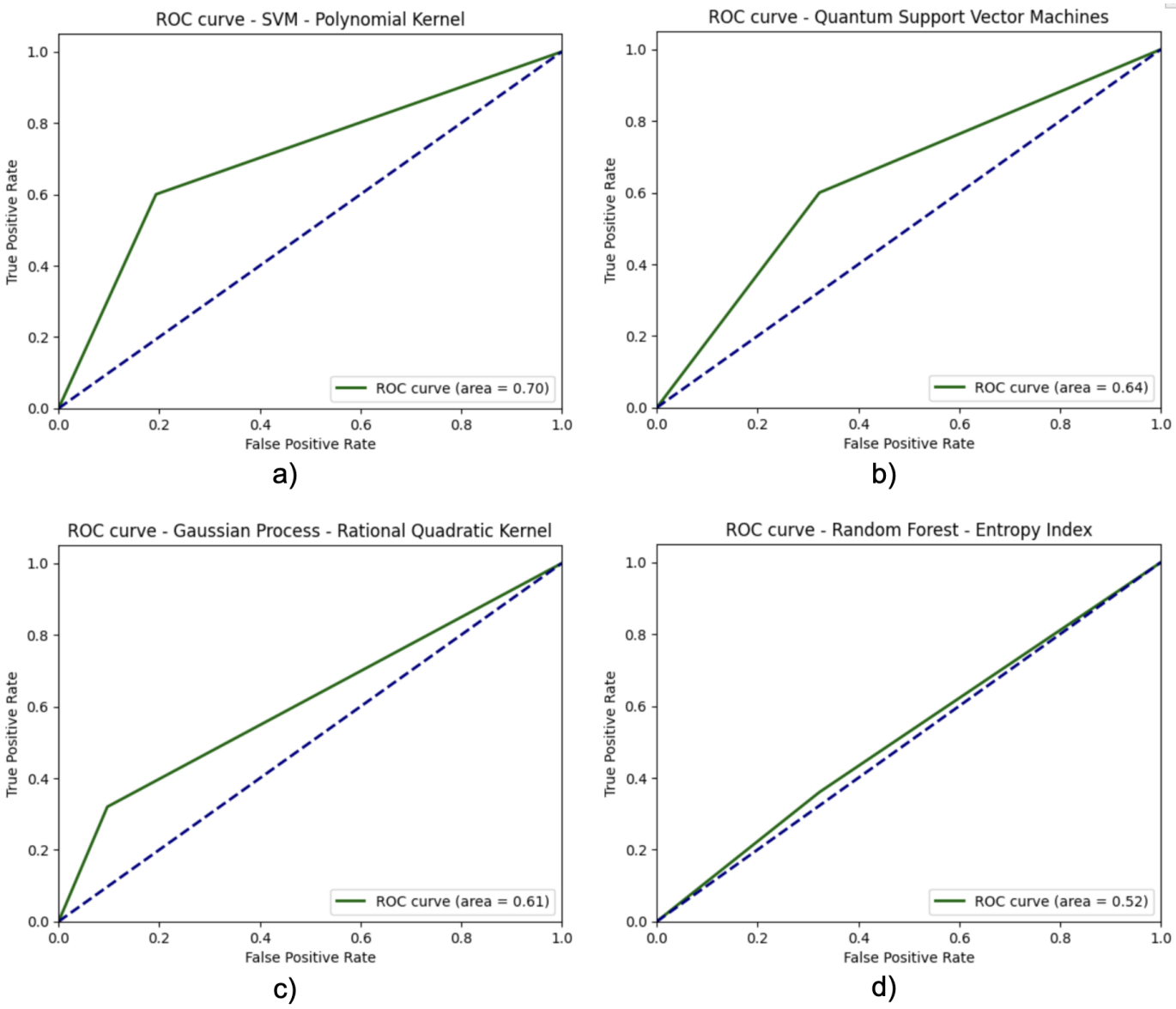

Figure 3. The highest ROC curve obtained from 10-Fold Cross-Validation for a) Support Vector Machine with the polynomial kernel, b) Quantum Support Vector Machine, c) Gaussian Process with the Rational Quadratic kernel, and d) Random forest

quadratic kernel. These results were obtained through a t-test comparing the performance of the classifiers, and the t-test results are reported in Table I. In contrast, the 5FCV results do not provide evidence of statistically significant differences between the classifiers' performance metrics.

Table III displays the confusion matrix obtained for both 10FCV and 5FCV, with the best-performing classifier being SVM with the polynomial kernel, as previously mentioned. The results obtained by this classifier align with the outcomes reported earlier, with 40 out of 56 students correctly classified, resulting in an accuracy of 71.42%. Only 6 out of 35 students at no risk were falsely identified as at risk, resulting in a false positive rate of 17%. Precision, which measures the proportion of correctly identified at-risk students out of all identified at-risk students, is a crucial metric in risk forecasting, as false positives can result in unnecessary expenditure of time and resources. In contrast, 10 out of 25 students at risk were not identified, resulting in a false negative rate of 40%. Recall, which measures the proportion of correctly identified at-risk students out of all at-risk students, is another important metric in risk forecasting, as false negatives can lead to students failing early courses in mathematics and physics. Both precision and recall exceed 60%, which is superior to random guessing.

Finally, the Receiver Operating Characteristic (ROC) curves shown in Figures 3 and 4 correspond to the classifiers with the highest areas under the ROC curves. Once again, these results reinforce that SVM outperforms the other methods. Besides, an area below the ROC curve of 0.7 is better than random guessing, although these classification methods were trained

 ISBN: 978-1-68558-089-6

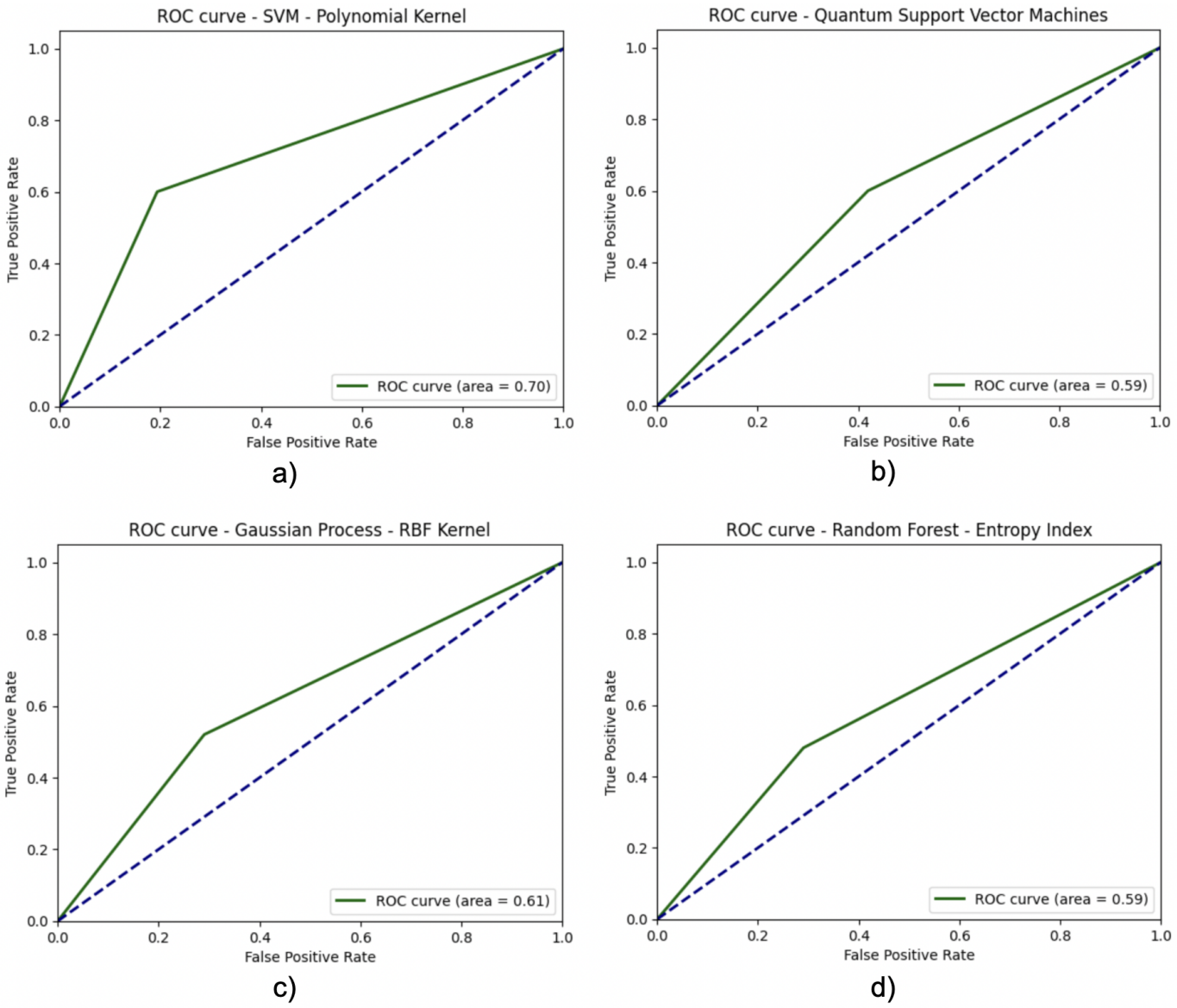

Figure 4. The highest ROC curve obtained from 5-Fold Cross-Validation for a) Support Vector Machine with the polynomial kernel, b) Quantum Support Vector Machine, c) Gaussian Process with the Radial Basis Function (RBF) kernel, and d) Random forest

on a small data set. Nevertheless, it is important to exercise caution when generalizing these results to larger data sets or different contexts, as classifier performance may vary. Further studies with larger and more diverse data sets are needed to confirm the robustness of these findings

V. Conclusions and Future Research

In this study, we explored the functional mapping between a student's risk of failing early courses in mathematics or physics and their performance on the admission test. Our contribution can be summarized in two parts: i) we have provided a data set for studying this relationship, and ii) we have developed a prototype intelligent system based on Support Vector Machine (SVM) that surpasses other machine learning methods, especially in terms of accuracy, as demonstrated by the conducted evaluation.

As a direction for further research, we shall collect more data to improve the accuracy, precision, recall, and harmonic mean of the intelligence system. Furthermore, with a greater data set, we shall evaluate models based on neural networks that tend to generalize well with large-scaled data sets.

Another research direction involves interpretable machine learning models that offer insights into the specific knowledge and skills that students need to succeed in these early courses. Such insights can be used to design courses that assist at-risk students in making a smooth transition from secondary school to the university. Moreover, we shall analyse the latent factors of the Saber 11 admission test to identify the most relevant factors contributing to success in these early mathematics and

 ISBN: 978-1-68558-089-6

physics courses, as well as for visualization purposes.

Furthermore, we shall continue our research on Quantum Support Vector Machine (QSVM) by exploring the ZZ feature mapping across diverse domains. Additionally, in the domain of our study, we aim to evaluate the performance of QSVM with other circuits for feature mapping, such as, e.g., angle encoding and amplitude encoding.

Finally, we aim at extending this study to other majors in engineering, such as, e.g., mechanical, environmental, food, and industrial engineering at the University of Córdoba in Colombia. To this end, we will collect data from those departments that offer the Bachelor's degree in those majors.

ACKNOWLEDGEMENT

Caicedo-Castro thanks the Lord Jesus Christ for blessing this project. We thank Universidad de Córdoba in Colombia for supporting the Course Prophet Research Project (grant FI-01-22). Particularly, we would like to express our deepest appreciation to Dr. Jairo Torres-Oviedo and Dr. Deivis Lujan-Rhenals, who serve as the president and research vice-president of the University of Córdoba. We thank all students who collaborated with us, answering the survey conducted for collecting the data set used in our research. Finally, we thank the anonymous referees for their comments that contributed to improve the quality of this article.

REFERENCES

[1] Colombian Institute for Education Assesment - ICFES. National System of Standardized Evaluation of the Education - Guideline of the Saber 11 test. (2013) https://www.icfes.gov.co/ [retrieved: October, 2023].

[2] I. Pacheco-Arrieta *et al.* (2004) 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: October, 2023]. University of Córdoba in Colombia.

[3] C. Demetriou and A. Schmitz-Sciborski, "Integration, Motivation, Strengths and Optimism : Retention Theories Past, Present and Future," in *Proceedings of the 7th National Symposium on Student Retention*. USA: The University of Oklahoma, 2011, pp. 300–312.

[4] 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.

[5] 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.

[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] J. Lin, P. Imbrie, and K. Reid, "Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results," in *Research in Engineering Education Symposium*. New York, USA: Curran Associates, Inc., 2009.

[8] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West. (2016) Predicting Student Dropout in Higher Education. https://arxiv.org/abs/1606.06364 [retrieved: October, 2023]. arXiv.

[9] J. B. Berger and J. F. Milem, "The Role of Student Involvement and Perceptions of Integration in a Causal Model of Student Persistence," *Research in Higher Education*, vol. 40, no. 6, pp. 641–664, 1999.

[10] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, "Predicting Students Drop Out: A Case Study." *Computers, Environment and Urban Systems*, pp. 41–50, 01 2009.

[11] J. Parker, M. Hogan, J. Eastabroo, A. Oke, and L. Wood, "Emotional intelligence and student retention: Predicting the successful transition from high school to university," *Personality and Individual Differences*, vol. 41, pp. 1329–1336, 2006.

[12] B. Pérez, C. Castellanos, and D. Correal, *Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study: First IEEE Colombian Conference, ColCACI 2018, Medellín, Colombia, May 16-18, 2018, Revised Selected Papers*. New York, USA: Institute of Electrical and Electronics Engineers, 05 2018, pp. 111–125.

[13] I. Caicedo-Castro, O. Velez-Langs, M. Macea-Anaya, S. Castaño-Rivera, and R. Catro-Púche, "Early Risk Detection of Bachelor's Student Withdrawal or Long-Term Retention," in *IARIA Congress 2022: International Conference on Technical Advances and Human Consequences*. Nice, France: International Academy, Research, and Industry Association, 2022, pp. 76–84.

[14] I. Caicedo-Castro. (2023) Dataset for Forecasting Failure Risk in Early Mathematics and Physical Science Courses in the Bachelor's Degree in Engineering. https://sites.google.com/correo.unicordoba.edu.co/isacaic/research [retrieved: October, 2023]. University of Córdoba in Colombia.

[15] 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. Cambridge, MA, USA: MIT Press, 1995, pp. 514–520.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[17] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[18] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[19] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.

[20] Qiskit Development Team. (2017) Qskit Development Kit. [Online]. Available: https://qiskit.org/[retrieved:October,2023]

[21] ——. (2017) ZZ Feature Mapping Library Documentation. https://qiskit.org/documentation/stubs/qiskit.circuit.library.ZZFeatureMap.html [retrieved: October, 2023].

[22] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

[23] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning*, vol. 96. Bari, Italy: Cambridge University Press, 1996, pp. 148–156.

[24] L. Breiman, "Random forests," in *Machine learning*, vol. 45, no. 1, Springer. USA: Springer, 2001, pp. 5–32.

[25] 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. New York, USA: Association for Computing Machinery, 2016, pp. 785–794.

[26] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] Colab. (2017) Google Colaboratory. https://colab.research.google.com/ [retrieved: October, 2023]. Google LLC.

 ISBN: 978-1-68558-089-6