Early Risk Detection of Bachelor's Student Withdrawal or Long-Term Retention

Isaac Caicedo-Castro[123], Oswaldo Vélez-Langs[23], Mario Macea-Anaya[235],
Samir Castaño-Rivera[23], Rubby Castro-Púche[14]
[1]*Research Group: Development, Education, and Healthcare*
[2]*Socrates Research Group*
[3]*Faculty of Engineering*
[4]*Faculty of Humanity and Social Science*
[5]*CINTIA, Center of INnovation in Technology of Information to support the Academia*
University of Córdoba
Carrera 6 No. 76-103, 230002, Montería, Colombia
emails: {isacaic, oswaldovelez, mariomacea, sacastano, rubycastro}@correo.unicordoba.edu.co

***Abstract*—In this research, we study the problem of forecasting recently admitted students at risk of withdrawing from the university or being long-term retained in a bachelor's program. We conduct research to study the case of students enrolled in courses up to the ninth semester, in the Department of Systems Engineering at the University of Córdoba in Colombia. At most universities throughout Colombia, including the University of Córdoba, the standardized and official admission test Saber 11 has been adopted for bachelor's program admissions. Therefore, we address the following research question: Might the admission test Saber 11 be used to forecast if the recently admitted student will be at either withdrawal or long-term retention risk, in the foreseeable future, before starting the first semester? We are motivated to solve the previously mentioned question because once the admitted students at risk have been identified, the University might make choices to help such students. To this end, we collected a dataset from 86 surveyed students. Although the original dataset has 86 records, after cleaning the dataset, and removing records with missing or inconsistent values, the final version of the dataset contains records of 47 students. According to the results of this research, given the student's test admission outcomes, machine learning algorithms learn regular patterns for forecasting if a recently admitted student is at withdrawal or long-term retention risk with a mean accuracy of about 72.5% (i.e., mean error of approximately 27.5%).**

***Keywords*—machine learning; educational data mining; classification algorithm; University admission test; student withdrawal; student long-term retention.**

I. Introduction

Universities offer bachelor programs that provide people, who have finished school, with higher education or vocational training for contributing to society in several sectors such as, e.g., healthcare, education, agronomy, industry, building, business management, government, and so forth. The education quality at schools (besides other factors) influences the student's performance at university. Moreover, university resources are limited, hence, each cannot admit an unlimited number of students. As a consequence, universities perform a selection process, where applications are usually studied according to the candidate's performance during the admission test, interviews, and other criteria. With the admission test, the goal is to evaluate if the candidate has reached the appropriate level to pursue a bachelor's program. Nevertheless, some students lack the required competencies, skills, or knowledge to succeed in the bachelor's program, albeit they have passed the admission test.

Those students who are not properly prepared, either might fail courses or might abandon them. In the former case, such students face the risk of losing their student status, when their performance is lower than required according to the university rules. This problem is known as *student withdrawal*. On the other hand, those students who leave courses without completion will eventually take more time than required to finish the bachelor's program. This problem is known as *long-term retention*. In this research, we study the problem of forecasting recently admitted students at risk of withdrawing from the university or being long-term-retained in a bachelor's program.

In Section I-A, we state the problem and research context. Section I-B discusses the arguments that motivate us to conduct this research. The key assumptions and motivations considered in this research are mentioned in Section I-C. In Section I-D, we present the contributions of this research and outline the rest of this article.

A. Research Context and Problem Statement

The problem addressed in this research is to predict if an admitted student might be at risk of withdrawing from the university or being long-term-retained in the bachelor's program, before starting the first semester. Herein, predicting means to classify the admitted student according to two classes as follows: (i) student at risk or (ii) student at no risk. Therefore, the problem is to classify the admitted students according to the previous two classes given the student's admission test outcomes.

The target variable is the class of students, whereas the student's admission test outcomes are input variables. Thus, in order to classify a student, the problem is to find the functional dependency between the target variable and input variables

 ISBN: 978-1-68558-017-9

from the history of previously admitted students, who have finished at least the first semester. In machine learning, this is a classification problem, because the target variable is discrete.

We conducted this research, by studying the case of students enrolled in the bachelor of science in engineering, who chose the major in systems engineering, in the context of the University of Córdoba in Colombia, which is a public university.

In Colombia, Saber 11 is the standardized and official test adopted for bachelor program admission, as well as Scholastic Assessment Test (SAT), is used for the same purpose in the United States. Therefore, candidates at the University of Córdoba are admitted or rejected, taking into account their outcomes obtained in the Saber 11 test.

The test Saber 11 evaluates four areas as follows: (i) mathematics, (ii) critical reading, (iii) social sciences, and (iv) English language. The Colombian education ministry assumes these areas are the foundation that every school student must learn properly to pursue a bachelor's program.

The problem is formally defined as follows: let $\{(\mathbf{x}^t, r^t)\}_{t=1}^{N}$ be the training dataset, where $\mathbf{x}^t \in \mathbb{R}^d$ and $r^t \in \{0, 1\}$. Henceforth, t is a super index rather than an exponent, for $t = 1, \ldots, N$. The d-dimensional vector $\mathbf{x}^t$ represents the t-th student's admission test outcomes. For instance, the j-th component, i.e., x_j^t, represents the resulting score in the mathematics area achieved by the t-th student in the admission test. $r^t = 1$ means the t-th student is at academic risk, whereas $r^t = 0$ means otherwise.

Given the previously described dataset, the learning problem is to find the functional dependency between the (independent) variables in $\mathbf{x}^t$ (or the student's features) and the target variable r^t (a.k.a., dependent variable). In other words, the problem is to find the function g such that $g : \mathbb{R}^d \rightarrow \{0, 1\}$. Thus, once the function g is found, given the input variables in the d-dimensional vector $\mathbf{x}$, corresponding to a new student, we can classify the student as follows: $g(\mathbf{x}) = y$, where y is the output variable, and $y = 1$ if the function g classifies the new student as one at risk, otherwise $y = 0$.

The above-described problem leads us to ponder the following research question: Might the student's outcome, achieved from the admission test, be used to forecast if the recently admitted student will be at either withdrawal risk, or long-term retention risk, in the foreseeable future, before starting the first semester?

B. Motivation

We are motivated to conduct this research to help universities (in particular the University of Córdoba in Colombia) at identifying those students at risk, who might leave their academic programs without completion, in the foreseeable future, as well as those students who might be retained in their bachelor's programs, beyond the expected time. Both cases are caused because such students were admitted lacking key competencies, or knowledge, to attain the required performance, which allows them to keep their student status, and finish their programs in the expected time. As a consequence, this causes students psychological issues, frustration, and financial loss.

If stakeholders at the university know in advance, who are those students at academic risk, they can carry out plans of action and strategies to handle the above-mentioned issues (e.g., the student's frustration, and financial loss), in order to help students, before starting their bachelor career, to keep their student status, and complete their programs within the expected time.

Strategies for coping with the risk might be such as, e.g., psychological support or extra courses to cover those topics that such students did not learn properly before being admitted to the university. Thus, eventually, students' withdrawal 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 (cf., [1]).

C. Key Assumptions and Limitations

In this research, we have considered the following assumptions:

(i) We assume the test called Saber 11 actually measures the knowledge and competencies, which students ought to attain for pursuing a bachelor's degree. Indeed, article 17-th of the student code at the University of Córdoba states that candidates are admitted according to their score achieved in the test Saber 11.

(ii) We assume that a student at academic risk leaves at least one course without completing the first semester because such courses might be prerequisites for attending further ones, or the student might face a high workload later, in another semester, enrolling unfinished courses (or equivalent courses to fulfill the graduation requirements). Therefore, eventually, the overwhelmed student will need more time than required to conclude the program.

(iii) We assume that a student at academic risk fails at least one course the first semester because this causes the same issues faced by another student who leaves at least one course without completion starting the bachelor's career. Moreover, there is a chance the student's global average grade decreases below the minimum required, compromising its student status after finishing the first semester, or later.

(iv) We assume the student at academic risk obtains a global average grade lower or equal to the required for keeping the student status. Bachelor students at Colombian universities are graded in the range from 0 up to 5. In the specific case of the University of Córdoba, according to the student's code (cf., article 16-th in [2]), each student ought to achieve a global average grade equal to or greater than 3.3, otherwise, this one might be dropped out from the university. According to article 28-th of the same code, if a student's global average grade is between 3 and 3.3, this one must increase the global average grade at least up to 3.3 the next semester, otherwise, the student is dropped out. Finally, if any student achieves a grade lower than 3, this one is withdrawn from the university.

 ISBN: 978-1-68558-017-9

(v) We assume the student might be at academic risk if this one might lose the student status, or the student takes more time in the academic program than the expected time.
(vi) We assume accuracy is more relevant for improving the user's experience than the interpretation of the forecasting algorithm.
(vii) We assume that classifying students at risk, who are not at risk whatsoever (i.e., false positive) is as inconvenient as classifying them without risk, though they are at an actual risk (i.e., false negative). In the first case, both students and the University will spend unnecessary resources. In the second case, students at risk will face the consequences of poor preparation for pursuing the bachelor's degree, and the University will not be able to plan how to deal with such students.

The scope limitations of this research are as follows:

(i) We shall not predict the student's grades in bachelor courses given their admission test performance.
(ii) We shall not aim at interpreting the functional dependency between the academic risk, i.e., the target variable, and student's performance in the admission test, i.e., the input variables.

D. Contributions and Outline

The contributions of this research are as follows:

(i) A dataset with 47 records. This includes the student's profile and academic history. These students have attended courses from the second up to the ninth semester. Besides, the dataset includes their respective outcomes from the standard admission evaluation Saber 11, which is taken into account in Colombia, to study applications for bachelor's degrees.
(ii) The proof-of-concept of an intelligence system, written in Python, that learns regular patterns from the outcome achieved by the students, during the University admission test, and their performance during the first semester. Such regular patterns are learned in order to forecast if a recently admitted student might be at academic risk of leaving the bachelor's career, due to low performance, or taking more time to finish the bachelor's career, than the expected one.
(iii) An empirical study that reveals the multilayer perceptrons algorithm outperforms support vector machines, logistic regression, and decision trees. The multilayer perceptrons net reaches a mean accuracy of about 72.5% (i.e., mean error about 27.5%)

The remainder of this article is outlined as follows: in Section II, we discuss the prior research. In Section III, we explain the research method we adopted for conducting this study. In Section IV, we present the experimental setting, including dataset features, adopted evaluation procedure, and which hyper-parameters are tuned for each model. Moreover, in the same section, we present and discuss the results of the experiments. Finally, Section V concludes the article with the findings drawn from the results and we discuss perspectives for further research.

II. Prior Research

This research belongs to the domain of educational data mining, which is a discipline whose goal is to adopt machine learning algorithms to large-scale datasets collected from educational settings in order to better understand students and the way they learn. Educational data mining includes (although not limited to) the following research direction: analyzing educational datasets, studying pedagogical theories through data mining, contributing to understanding the students' domain representations, evaluating the students' engagement in the learning tasks, and so forth.

Herein, we are focused on analyzing an educational dataset by training machine learning algorithms to find regular patterns, in order to classify a recently admitted student according to two classes, i.e., i) student at withdrawal or long-term-retention risk, and ii) student at no risk.

In this research direction, the performance of American students at school, and their cognitive abilities, have been used for predicting the student's persistence in a bachelor's career, unfortunately, the prediction accuracy was unfeasible [3].

Another related research has been taken into the student's performance, during the Dutch pre-university secondary education, for prediction purposes. The prediction is done before the student starts the first semester. The goal is to forecast if the student might be at risk of leaving a bachelor's program without completion later. [4]. The drawback of this research approach is, that it is fitted to the particular Dutch pre-university educational system, hence, it is not feasible to be reproduced in other contexts, such as, e.g., the Colombian one.

The outcomes of the standard American admission test, known as SAT, have been used in prior research to predict if students will withdraw from the bachelor's program [5], [6]. tests, SAT and Saber 11, evaluate mathematics knowledge and communication skills. However, the test Saber 11 evaluates social science knowledge as well as communication competencies in two tongues, i.e., the English and Spanish languages. SAT is designed to evaluate just the communication skills in the English language.

So, the test SAT and the performance during the freshman year at university, have been used to predict if a student will withdraw from the bachelor's program [5]. However, predicting student withdrawal after the freshman year does not aid in anticipating the student's long-term retention issues. A similar approach also includes the student's demographic information, besides the pre-university student's performance information, for forecasting purposes. [6]. This research endeavor is similar to our study, although our goal is to carry out the prediction before the student starts the freshman year.

Moreover, using demographic information for prediction is beyond our research scope because it is not related to our research question. Another relevant difference is that we use the actual admission test outcomes for training and prediction, whereas Lovenoor et al. carried out data imputation

 ISBN: 978-1-68558-017-9

for completing 40% of the missing admission outcomes in their dataset [6].

Another research direction is taking into account emotional intelligence measurements for predicting bachelor students withdrawal [7]. However, our research is rather focused on the relationship between the test Saber 11 outcomes and the risk of long-term retention and bachelor's career withdrawal.

On the other hand, academic and personal data have been used for predicting the bachelor's student withdrawal rate in the context of a Colombian university [8]. Unfortunately, the final dataset used in this research is not publicly available, for experimental reproduction purposes. Besides, we do not aim at estimating the withdrawal rate, instead, we are focused on the risk prediction of each recently admitted student in a bachelor's program.

As far as we know, no prior research has studied if only the admission test is sufficient for forecasting whether a recently admitted student might face the risk of withdrawing from the bachelor's program or being long-term retained beyond the expected time.

III. Research Method

We adopted a quantitative research approach, using machine learning algorithms for predicting if a bachelor's degree student will be at academic risk, given their outcomes in the admission test called Saber 11. To this end, we collected a dataset for training these algorithms. The procedure to collect this dataset is described in Section III-A.

The machine learning algorithms used for prediction in this research are supervised learning algorithms for classification. We discuss them in Section III-B.

A. Collecting the Dataset

We collected the required information for this study through Google Form. We surveyed 86 students enrolled in the bachelor's program of Systems Engineering at the University of Córdoba in Colombia. These students are attending courses from the second up to the eighth semester. The information collected from each student includes the outcome achieved from the admission test. Thus, a t-th student's features are represented through a four-dimensional vector, i.e., $\mathbf{x}^t \in \mathbb{R}^n$ (here $n = 4$ and t is a super index instead of an exponent), where its components correspond to the following areas of the test: (i) mathematics, whose student's score is denoted as x_1^t, (ii) critical reading, whose student's score is denoted as x_2^t, (iii) social sciences, whose student's score is denoted as x_3^t, and (iv) English language, whose student's score is denoted as x_4^t. These variables do not depend on other ones (i.e., independent variables), where each one is in the range from 0 up to 100.

We also collected the following information for each student: (i) the number of students' failed courses in the first semester, (ii) the number of the students' canceled courses in the first semester, and (iii) the student's global average grade achieved the first semester. These variables are used to determine the target variable r^t (once again, t is a super index) considering the following conditions:

- If the t-th student does not approve all the courses the first semester, then this one might be at risk of being retained or losing the student status due to poor performance, i.e., $r^t = 1$ as long as the t-th student fulfills this condition, otherwise $r^t = 0$.
- The t-th student is at risk of being retained in the program if this one cancels at least one course since the first semester, i.e., $r^t = 1$ as long as the student t-th fulfills this condition, otherwise $r^t = 0$.
- The student might be at risk of being withdrawn from the university as well, if this one achieves a global average grade lower than the minimum required to keep the student status according to the rules of the University of Córdoba in Colombia, in this case, the t-th student is at risk of being dropped out if this one achieves an average grade lower than 3.3, where grades are in the range from 0 up to 5, i.e., $r^t = 1$ as long as the t-th student fulfills this condition, otherwise $r^t = 0$.

Once we collected the dataset, we removed those records with inconsistent data such as, e.g., those records whose sum of the score per area is different from the total score. After this procedure, the dataset contains 47 records, furthermore, each student was de-identified to keep their identity anonymous. Currently, the dataset is available on the web, to allow the reproduction of our study, and for further research [9].

Figure 1 depicts the proportion of students at academic risk from the remaining records. The final dataset is rather balanced due to almost half of the records corresponding to students at risk, while the remainder dataset does not.

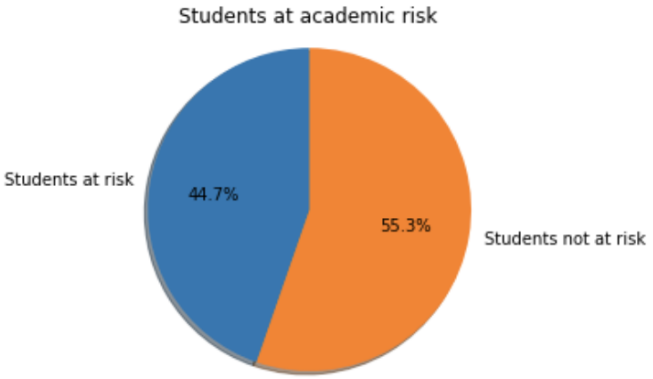

Figure 1. In the final dataset, 21 out of 47 surveyed students are at academic risk.

B. Classification Algorithms

We have adopted four supervised machine learning algorithms for predicting if an admitted student will be at risk, i.e., support vector machines, logistic regression (a.k.a., logistic discrimination), multilayer perceptrons, and decision trees. These algorithms carry out the prediction by classifying the student according to two classes, namely i) student at risk or ii) student at no risk.

 ISBN: 978-1-68558-017-9

So far, the support vector machines algorithm is the best theoretical motivation and the most successful one in the practice of modern machine learning [10, pg. 79]. This algorithm is based on convex optimization, as a consequence, there is a global maximum solution to be found, i.e., there is only one optimal solution, which is its main advantage. Nonetheless, this algorithm does not suit for interpretation in data mining, hence, this is not appropriate for discovering knowledge but for training accurate intelligence systems. A broader description of this algorithm is provided by Cortes and Vapnik [11].

With both classification algorithms, support vector machines, and logistic regression, it is assumed the input vector space can be separated through a linear decision boundary (or a hyperplane in the case of a multidimensional space), thereby, these algorithms are known as linear discrimination algorithms. Nevertheless, when this assumption is not satisfied the support vector machines algorithm is used with kernel methods (see Cortes and Vapnik [11] for further details).

In the case of logistic regression, the input space can be mapped to another vector space, where this assumption is set. Another option is adopting artificial neural networks, where each neuron is actually a logistic discriminator. The neuron outputs in the middle of the network become inputs of the neurons that actually classify. Thus, the original input variables are mapped into a new vector space, through the neurons in the middle, where the previously mentioned assumption is fulfilled. Anderson and McLachlan delve into the details of logistic regression [12], [13], besides, we trained the logistic regression classifier through Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [14], [15].

Although support vector machines is considered the most successful algorithm in the practice of modern machine learning, the multilayer perceptrons algorithm, which is an artificial neural network, is the most successful algorithm in the practice of deep learning and big data [16, pg. 3]. In this research, we have adopted the multilayer perceptrons algorithm trained through back-propagated cross-entropy error [17], and the optimization algorithm known as Adam [18]. We used one hidden layer due to the high time complexity of the back-propagation algorithm.

The multilayer perceptrons algorithm is a universal approximator (i.e., this is able to approximate any function for either classification or regression), which is its main advantage, whereas its main disadvantage is the objective function (a.k.a., loss function) based on the cross-entropy error is not convex, therefore, the synaptic weights obtained through the training process might not converge in the most optimum solution because there are several local minimums in the objective function. Thus, finding a solution depends on the random initialization of the synaptic weights. Furthermore, the multilayer perceptrons have more hyper-parameter to be tuned than other learning algorithms (e.g., support vector machines or naive Bayes), which is an additional shortcoming.

Finally, the decision tree algorithm is the most common learning algorithm adopted for mining data or knowledge discovery because this one is simple to interpret. It is possible to visualize trees, which is a desirable feature for making decisions, and its best advantage. Decision trees are trained through heuristic algorithms, such as greedy algorithms, where there are several local optimal solutions at each node. Therefore, there is no guarantee the learning algorithm converges in the most optimal solution, as well as the multilayer perceptrons algorithm. So, this is the main drawback of the decision trees, and it also causes completely different tree shapes due to small variations in the training dataset (as we shall see in Section IV-C). The decision tree algorithm was proposed in 1984, Breiman *et al*. delve into its details (cf., [19]).

IV. Evaluation

A. Experimental Setting

To evaluate the machine learning algorithms used for predicting if a student is at withdrawal or long-term retention risk, we need several pairs of training and test datasets. To this end, we carried out experiments based on K-Fold Cross-Validation (KFCV), thus, from the original dataset, we get K pairs of training and test datasets. We chose $K = 10$, where it is usually 10 or 30. We did not choose $K = 30$ because the dataset is small. Thus, we test each algorithm K times through KFCV. With the test outcomes, we calculate the mean error to compare the learning algorithms, and choosing the algorithm hyper-parameters (e.g., the regularization parameter in the multilayer perceptrons and logistic regression). Besides the mean error, we also measure the mean of precision and recall.

With support vector machines, we tested two kernels, namely, polynomial and Gaussian kernel (a.k.a., radial basis function kernel).

We tested two decision trees with two impurity functions, namely, entropy and Gini function.

Moreover, we tested multilayer perceptrons with several neurons within one hidden layer. We evaluated three activation functions in the hidden layer, i.e., ReLU (Rectified Linear Unit), hyperbolic tangent, and sigmoid function. Besides, we tested various regularization parameter values for logistic regression and multilayer perceptrons net. Both algorithms have been trained for minimizing the sum of cross-entropy errors. The sigmoid function is the activation function in the output layer of the multilayer perceptions net. By definition, the same function is the generalization function in logistic regression.

Finally, we have programmed all the experiments with Python, using the Scikit-Learn library [20], in Google Colaboratory [21].

B. Results

According to the results shown in Table I, the multilayer perceptrons algorithm outperforms the other tested learning algorithms, despite the t-test revealing there is no statistical evidence that the mean error of the multilayer perceptrons algorithm is far lower than the one obtained through the other algorithms, i.e., the resulting p-value is greater than 0.05 (see Table II).

 ISBN: 978-1-68558-017-9

TABLE I
PERFORMANCE OF THE MACHINE LEARNING ALGORITHMS ADOPTED IN THIS RESEARCH

Machine Learning Algorithm	Mean error (%)	Mean precision (%)	Mean recall (%)
MP[a]	**27.5**	55	43.33
SVMPK[b]	30	**63.33**	**65.83**
SVMGK[c]	34.5	**63.33**	62.5
LR[d]	32.5	**63.33**	**65.83**
DTE[e]	34	46.33	58.33
DTGI[f]	40.5	48.33	46.67

[a]MP stands for Multilayer Perceptrons.
[b]SVMPK stands for Support Vector Machine and polynomial kernel.
[c]SVMGK stands for Support Vector Machine and Gaussian kernel.
[d]LR stands for Logistic Regression.
[e]DTE stands for Decision Tree with Entropy impurity function.
[f]DTGI stands for Decision Tree with Gini impurity function.

Thus, according to the experiments, the multilayer perceptrons algorithm achieved the lowest mean error with the following setting:

- The regularization parameter selection for decaying the synaptic weights in the multilayer perceptrons algorithm is sketched in Figure 2, where the best setting is obtained when the regularization parameter is equal to 10^{-2}.
- Another weight decay method used is early stopping.
- The lowest mean error was achieved using ReLU activation function with 600 neurons within the hidden layer, whereas we use sigmoid function with the neuron in the output layer.
- We used the Adam algorithm for training, where the initial learning rate is equal to 10^{-2}. The exponential decay rate for estimating the first and second moment vectors are equal to 0.9 and 0.999, respectively. The numerical stability in Adam is equal to 10^{-8}.
- We used a batch size of 8 examples.

TABLE II
STUDENT'S PAIRED T-TEST ON MEAN ERROR TO COMPARE THE MULTILAYER PERCEPTRONS ACCURACY WITH OTHER MACHINE LEARNING ALGORITHMS ADOPTED IN THIS RESEARCH

Machine Learning Algorithm	Mean error (%)	p-value
Multilayer Perceptrons	27.5	–
Support Vector Machine with Polynomial Kernel	30	0.82
Support Vector Machine with Gaussian Kernel	34.5	0.55
Linear Regression	32.5	0.68
Decision Tree with Entropy impurity function	34	0.53
Decision Tree with Gini impurity function	40.5	0.24

Support vector machines with a polynomial kernel is the next best choice according to the experiments. The best results for this learning algorithm is achieved with the following setting:

- The best degree value for the polynomial kernel is equal to 2.

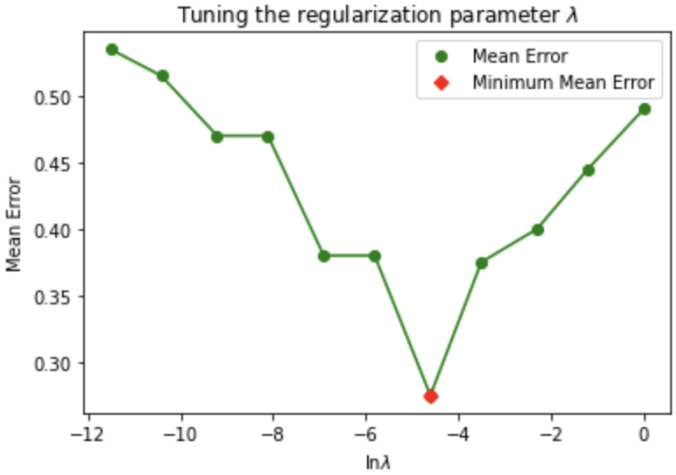

Figure 2. Tuning the multilayer perceptrons through 10-Fold Cross-Validation for choosing the regularization parameter λ according to the elbow rule. The minimum mean error is achieved when $\lambda = 10^{-2}$

- The best regularization parameter of the support vector machines algorithm with polynomial kernel is equal to 0.5. (i.e., $C = 0.5$)

On the other hand, the best setting for the support vector machines algorithm with a Gaussian kernel is as follows:

- The best gamma parameter for the Gaussian kernel is equal to 10^{-4}.
- The best regularization parameter of the support vector machines algorithm with Gaussian kernel is equal to 32×10^4 (i.e., $C = 32 \times 10^4$).

Finally, with logistic regression, the best regularization parameter is equal to 10^{-2}, whereas the entropy impurity function in decision trees performs better than Gini impurity function.

C. Discussion

The results reveal that, given the student's test admission outcomes, machine learning algorithms learn regular patterns for forecasting if a recently admitted student is at withdrawal or long-term-retention risk with a mean accuracy of about 72.5% (i.e., mean error of approximately 27.5%), which is much more accurate than tossing an unbiased coin, despite the dataset containing few instance numbers. Therefore, it is expected that the bigger dataset is, the better the mean accuracy will be.

On the other hand, the t-test that reveals there is no statistical evidence to prove that the multilayer perceptrons algorithm is significantly far more accurate than the other machine learning algorithms tested in this research because the p-value greater is than 0.05 (see Table II). This might lead us to think that in this case, adopting the decision trees algorithm is the right choice due to this is simple to interpret. Nevertheless, the variations in the training dataset caused by the 10-fold cross-validation, the tree shape changes drastically, as shown in Figure 3. This does not allow generalizing the rules for estimating the student's

 ISBN: 978-1-68558-017-9

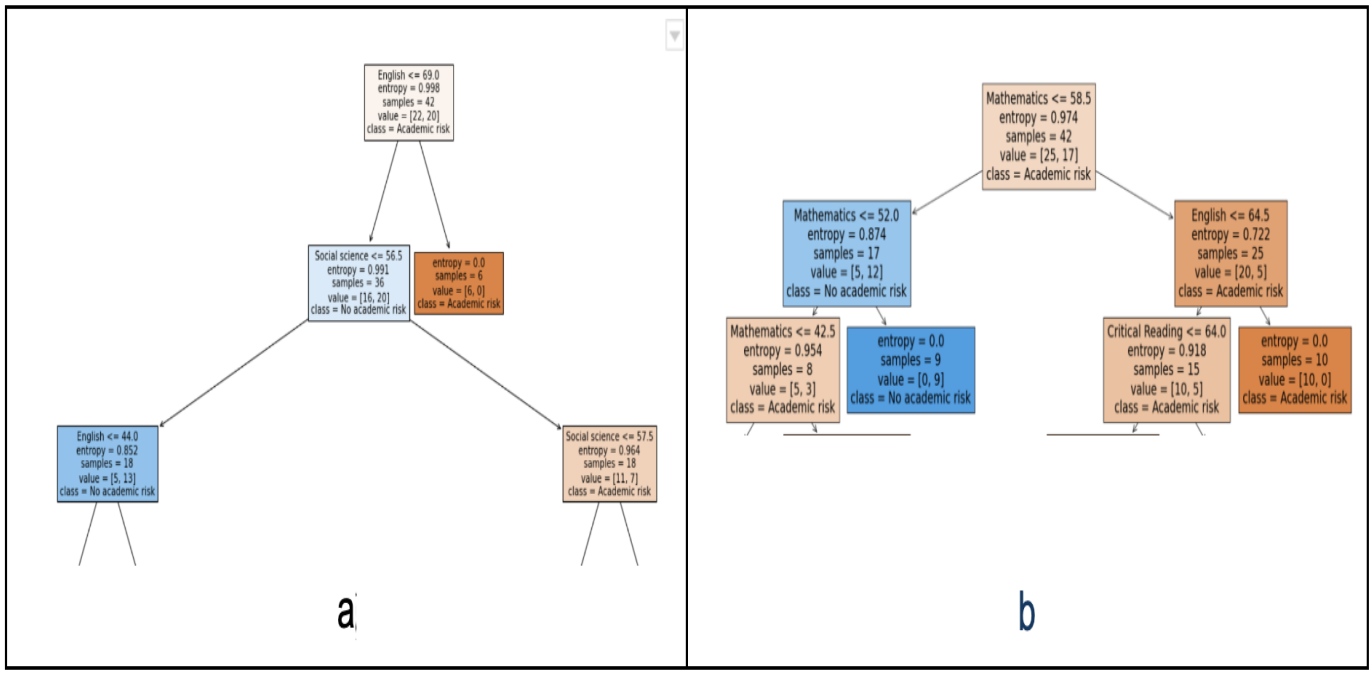

Figure 3. a) Part of the resulting decision tree shape during the first iteration of the 10-Fold Cross-Validation. b) Part of the resulting decision tree shape during the second iteration of the 10-FCV.

risk. Therefore, the reason for recommending the multilayer perceptrons algorithm to tackle the problem addressed in our research is twofold:

(i) Experiments in other domains have evidenced that the bigger the dataset is, the more accurate the multilayer perceptrons algorithm is, even more than other machine learning algorithms [16, pg. 3]. As a consequence, we expect significant improvement of the multilayer perceptrons accuracy, compared with the other tested learning algorithms, as we collect more examples for training it.

(ii) Taking into account the sixth assumption mentioned in Section I-C, the multilayer perceptrons algorithm is the better choice than decision trees, according to the results,because the prediction accuracy is more desirable than an interpretative prediction, that is less accurate.

Finally, regarding the test Saber 11 is similar to SAT, the outcomes of this research might be extended to the context of American Colleges or Universities. Indeed, by adopting the multilayer perceptrons algorithm, the knowledge it attains might be transferred to similar contexts, using the pre-trained synaptic weights, so it is not required to train a new multilayer perceptrons net from scratch likewise this is done in other domains, such as, e.g., computer vision.

V. Conclusions and Perspectives

In this research, we addressed the following question: Might the student's outcome, achieved from the admission test called Saber 11, be used to forecast if the recently admitted student will be at either withdrawal or long-term retention risk, in the foreseeable future, before starting the first semester?

Anticipating the student's risk might allow the Universities to take precautions necessary to prevent the issues related to these risks, such as, e.g., the student's frustration, financial loss, and so forth.

Some precautions might be such as, e.g., psychological advice, and courses that let the student overcome the associated risk.

Herein, the addressed problem is to find the functional dependency between the admission test outcomes achieved by the student, and its withdrawal or long-term-retention risk. We have tackled this problem, by using supervised machine learning algorithms for classification, i.e., multilayer perceptrons, support vector machines, logistic regression, and decision trees.

To train and evaluate machine learning algorithms, we collected a dataset by surveying 86 students. After cleaning the dataset, we removed 39 records, resulting in a dataset containing 47 records.

We draw the following conclusions from the experimental evaluation (through K-fold cross-validation):

(i) The polynomial kernel is a better choice than the Gaussian kernel for adopting the support vector machines algorithm.

(ii) Support vector machines and logistic regression have the same mean precision, while the former algorithm with the polynomial kernel has the same mean recall that the latter.

(iii) The decision tree with the entropy impurity function

 ISBN: 978-1-68558-017-9

performs better than the one with the Gini impurity function.

(iv) The multilayer perceptrons algorithm outperforms the other studied learning algorithms, despite the t-test revealing there is no statistical evidence that the mean error of the support vector machines algorithm is far lower than the one obtained through the other algorithms, i.e., the resulting p-value is greater than 0.05.

(v) Concerning the research question, with machine learning, it is possible to predict if a recently admitted student in a bachelor's program will be at withdrawal or long-term-retention risk with a mean accuracy of about 72.5% (i.e., a mean error of approximately 27.5%).

(vi) The results reveal that the multilayer perceptrons algorithm is the best choice for facing the problem addressed in this research, regarding also the experience in other domains, where the bigger the dataset is, the more accurate deep neural networks based on the multilayer perceptrons algorithm are, even far more accurate than other learning algorithms [16, pg. 3]

(vii) The multilayer perceptrons algorithm is a better choice than decision trees, according to the results, because it is more desirable accurate forecasting than a less accurate prediction based on an interpretative model.

For further research, we shall collect more data, including more variables, such as, e.g., demographic, economic, emotional, psychological, environmental variables, and so forth. Thus, we can study their influence on the student's performance. On the other hand, a dataset with more records will reduce the classification error and improve the forecasting accuracy.

Finally, we propose other research directions based on the following open questions:

(i) Might the admission test Saber 11 be used for suggesting bachelor's degrees, according to the risk faced by the student in pursuing such bachelor's careers? Arguably, a candidate who has poor performance in the mathematics area of the admission test might be at risk if, for instance, this person pursues a bachelor of engineering. Nevertheless, if the same candidate has a good outcome in the critical reading area, might not be at risk, as long as this person chooses a bachelor's degree that does not require advanced quantitative competencies such as, e.g., a bachelor's degree in literature.

(ii) Will the accuracy increase as more areas are included in the test Saber 11? For instance, if general science is evaluated, this might help to predict the student performance in bachelor of science with majors in either science (e.g., physics, chemistry, biology, and so forth) or engineering (e.g., computer science, electrical and electronic engineering, etc.).

(iii) Might the accuracy of the learning algorithms increase above 90% by training them with more examples, without including more variables (e.g., demographic data or emotional measurements)? If so, might variables such as, e.g., demographic, psychological, emotional, economic, and so forth, be latent factors that can be inferred from the test Saber 11 outcomes? For example, recommender systems might infer latent factors such as, e.g., movie genre from the rating given by the user to movies.

(iv) In Colombia, there is a standardized test called Saber Pro, which is taken by bachelor's students before fulfilling the requirements to receive a bachelor's degree. The test Saber Pro is similar to Saber 11, and it is designed to evaluate the critical reading, quantitative reasoning, citizenship competencies, Spanish written communication, and English communication skills. Might the test Saber Pro be used to forecast if a recently admitted graduate student (e.g., enrolled in either a master's or a Ph.D. program) will be at risk of withdrawing from the University, or being long-term-retained in the graduate program?

ACKNOWLEDGMENT

Caicedo-Castro thanks the Lord Jesus Christ for blessing this project. We thank Universidad de Córdoba in Colombia for the financial support to this research (grant FE-06-17). In special, our deepest appreciation goes to Dr. Jairo Torres-Oviedo, University of Córdoba president, who has always helped us and supported us throughout this research. We thank all students who collaborated with us, answering the survey conducted for collecting the dataset, used to train the learning algorithms adopted in this research. Finally, we thank the anonymous reviewers for their comments that contributed to improve the quality of this article.

REFERENCES

[1] 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*, 2011, pp. 300–312.

[2] I. Pacheco-Arrieta *et al.* (2004) Agreement No. 004: Student's code at the University of Córdoba in Colombia. Retrieved on June 14. [Online]. Available: http://www.unicordoba.edu.co/wp-content/uploads/2018/12/reglamento-academico.pdf

[3] 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.

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

[5] 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*. Curran Associates, Inc., 2009.

[6] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West. (2016) Predicting Student Dropout in Higher Education. Retrieved on December 10. [Online]. Available: https://arxiv.org/abs/1606.06364

[7] 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.

[8] 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*. IEEE, 05 2018, pp. 111–125.

 ISBN: 978-1-68558-017-9

[9] I. Caicedo-Castro. (2022) Dataset for Early Risk Detection of Bachelor Student Withdrawal or Long-Term-Retention. Retrieved on March 15. [Online]. Available: https://sites.google.com/correo.unicordoba.edu.co/isacaic/research
[10] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. The MIT Press, 2018.
[11] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.
[12] J. A. Anderson, *Logistic Discrimination*. North Holland, 1982, pp. 169–192.
[13] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, 1992.
[14] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 3, (Ser. B), pp. 503–528, 1989.
[15] R. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.
[16] C. C. Aggarwal, *Neural Networks and Deep Learning*. Springer, 2018.
[17] 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.
[18] D. P. Kingma and J. Ba. (2014) Adam: A method for stochastic optimization. Retrieved on December 10. [Online]. Available: http://arxiv.org/abs/1412.6980
[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.
[20] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[21] (2017) Google Colaboratory. Retrieved on June 14. [Online]. Available: https://colab.research.google.com/

 ISBN: 978-1-68558-017-9