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

Isaac Caicedo-Castro, Oswaldo Vélez-Langs, Mario Macea-Anaya,
Samir Castaño-Rivera, and Rubby Castro-Púche

emails: {isacaic, oswaldovelez, mariomacea, sacastano,
rubycastro}@correo.unicordoba.edu.co

IARIA Congress 2022

University of Córdoba in Colombia

July 24, 2022 to July 28, 2022

# Who are we? 1/5

- Isaac Caicedo-Castro
- Full Professor in the Faculty of Engineering at the University of Córdoba in Colombia
- Ph.D. in Informatics - University of Grenoble Alpes in France
- Ph.D. in Engineering - National University of Colombia

# Who are we? 2/5

- Oswaldo Vélez-Langs
- Full Professor in the Faculty of Engineering at the University of Córdoba in Colombia
- Ph.D. in software engineering, systems, and languages - Polytechnic University of Madrid in Spain

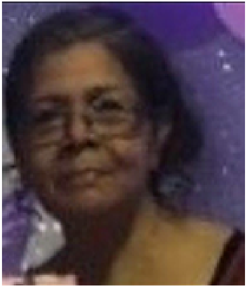
# Who are we? 3/5

- Mario Macea-Anaya
- Full Professor in the Faculty of Engineering at the University of Córdoba in Colombia
- Chief of CINTIA, Center of INnovation in Technology of Information to support the Academia at the University of Córdoba
- Sc.D. in management - Dr. Rafael Belloso Chacín University in Venezuela

# Who are we? 4/5

- Samir Castaño-Rivera
- Head of the Systems Engineering Department at the University of Córdoba in Colombia
- M.Sc. in Free Software - Autonomous University of Bucaramanga in Colombia

# Who are we? 5/5



- Rubby Castro-Púche
- Full Professor in the Faculty of Humanity and Social Science at the University of Córdoba in Colombia
- M.Sc. in Education - La Salle University in Colombia

# Outline

1. Introduction
2. Prior Research
3. Research Method
4. Evaluation
5. Conclusions and Perspectives
6. Acknowledgment

# Outline

1 Introduction

2 Prior Research

3 Research Method

4 Evaluation

5 Conclusions and Perspectives

6 Acknowledgment

# Introduction

We aim to forecast if a student recently admitted to a bachelor's career might face the risk of:

- Withdrawal
- Long-term retention

# Research Context and Problem Statement

Here forecasting means classify:

- Class 1: student at risk of withdrawal or long-term retention
- Class 2: the student is not at risk

# Research Context and Problem Statement

Students enrolled:

- Bachelor of science in engineering.
- Major: systems engineering.
- University of Córdoba in Colombia.
- Saber 11 is the standardized and official test adopted for bachelor program admission.

The test called Saber 11 evaluates

(i) mathematics ($0 \leq x_1 \leq 100$),

(ii) critical reading ($0 \leq x_2 \leq 100$),

(iii) social science ($0 \leq x_3 \leq 100$), and

(iv) English language ($0 \leq x_4 \leq 100$).

# Research Context and Problem Statement

- Given the dataset $\{(x^t, r^t)\}_{t=1}^N$ ($t$ is a super index).
- Where the $d$-dimensional vector $x^t$ represents the $t$-th student's outcomes obtained from the test Saber 11 ($d = 4$). This vector contains the independent variables (i.e., student's features).
- $r^t \in \{0, 1\}$ represents the target variable (a.k.a, independent variable), where $r^t = 1$ if the $t$-th student is at risk of withdrawal or long-term retention. $r^t = 0$ means otherwise.
- The problem is to find the functional dependency between the student's features and the target variable, i.e., to find the function $g$ such that $g : \mathbb{R}^d \rightarrow \{0, 1\}$.

# Research Context and Problem Statement

## 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?

# Motivation

Student's withdrawal or long-term retention cause

- psychological issues,
- frustration, and
- financial loss.

Anticipating the student's risk allows the universities to take precautions to prevent these issues, such as, e.g.,

- psychological advice, and
- courses that let the student overcome the associated risk.

# Key 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 (c.f., article 17-th of the student code at the University of Córdoba)

(ii) We assume that a student at academic risk leaves at least one course without completing the first semester

(iii) We assume that a student at academic risk fails at least one course the first semester

(iv) We assume the student at academic risk obtains a global average grade lower or equal to the required for keeping the student status (c.f., article 16-th of the student code at the University of Córdoba).

(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.

# Key Assumptions

(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).

# Limitations

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. candidate.

# Contribution

(i) A dataset with 47 records.

(ii) The proof-of-concept of an intelligence system.

(iii) An empirical study that reveals the multilayer perceptrons algorithm outperforms the other studied learning algorithms, by reaching a mean accuracy of about 72.5%

# Outline

1. Introduction
2. Prior Research
3. Research Method
4. Evaluation
5. Conclusions and Perspectives
6. Acknowledgment

# Prior Research

## Berger and Milem (1999)

- Goal: forecasting the student's persistence in a bachelor's career.
- Input variables: the student's performance at school, and cognitive abilities.
- Drawback: The prediction accuracy was unfeasible.

## Dekker, Pechenizkiy, and Vleeshouwers (2009)

- Goal: forecasting if the recently admitted student might be at risk of leaving a bachelor's program without completion
- Input variables: the student's performance at school.
- Drawback: this research is fitted to the particular context of the Dutch University educational system.

# Prior Research

## Prediction based on the test SAT

- Similarities between SAT and Saber 11 → evaluate mathematics knowledge and communication skills.
- Differences between SAT and Saber 11 → Saber 11 evaluates social science knowledge as well as communication competencies in two tongues, i.e., the English and Spanish languages.
- Differences between SAT and Saber 11 → SAT is designed to evaluate just the communication skills in the English language.
- Lin, Imbrie, and Reid (2009)
- Aulck *et al*. (2016)

# Prior Research

## Lin, Imbrie, and Reid (2009)

- Goal: forecasting if a student will withdraw after the first year in the bachelor's program.
- Input variables: the student's SAT outcome and the freshman performance.
- Drawback: forecasting after the first year might be too late for anticipating the student's long-term retention issues.

# Prior Research

## Aulck *et al*. (2016)

- Goal: forecasting if a student will withdraw after the first year in the bachelor's program.
- Input variables: the student's SAT outcome, their performance at school, their demographic information, and their freshman performance.
- Drawback: forecasting after the first year might be too late for anticipating the student's long-term retention issues.
- Using demographic information for forecasting purposes is beyond our research scope.
- The authors conducted data imputation for completing 40% of the missing admission outcomes while we used the actual values.

# Prior Research

## Parker *et al.* (2006)

- Goal: predicting bachelor students’ withdrawal.
- Input variables: emotional intelligence measurements.
- 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.

## Perez, Castellanos, and Correal (2018)

- Goal: predicting the bachelor’s student withdrawal rate in the context of a Colombian university.
- Input variables: academic and personal data.
- Drawback: the final dataset isn’t publicly available.
- we don’t aim at estimating the withdrawal rate.

# Outline

1 Introduction

2 Prior Research

3 Research Method

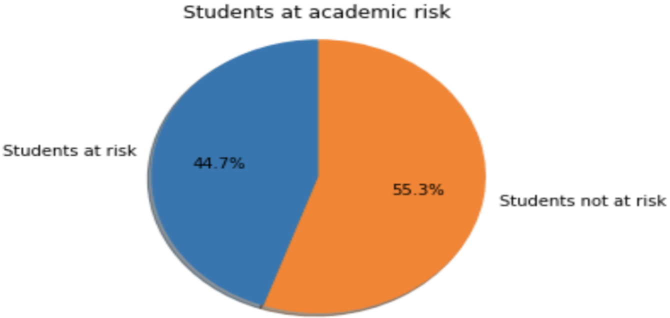
4 Evaluation

5 Conclusions and Perspectives

6 Acknowledgment

# Research Method

- Quantitative research approach
- Collect the dataset
  - We surveyed 86 students (bachelor’s program of Systems Enginnering).
  - We removed those records with inconsistent data, resulting in a dataset with 47 records.
  - In the final dataset, 21 out of 47 surveyed students are at risk.



# Research Method

Machine learning algorithms for classification:

- Support vector machines algorithm: it's the best theoretical motivated and the most successful one in the practice of modern machine learning (Mohri *et al*., 2018)
- Multilayer perceptrons net: it's an artificial neural network, is the most successful algorithm in the practice of deep learning and big data (Aggarwal, 2018)
- Decision trees: it's simple to interpret.
- Logistic regression

# Outline

1 Introduction

2 Prior Research

3 Research Method

4 Evaluation

5 Conclusions and Perspectives

6 Acknowledgment

# Experimental Setting

- $K$-Fold Cross-Validation ($K = 10$). Thus, we get $K$ pairs of training and test datasets.
- With support vector machines → two kernels: polynomial and Gaussian kernel.
- With decision trees → two impurity functions: entropy and Gini function.
- With multilayer perceptrons → three activation functions: ReLU, hyperbolic tangent, and sigmoid.
- With multilayer perceptrons and logistic regression → regularization parameter.
- Scikit-Learn libray + Google Colaboratory.

# Results

## Machine Learning Algorithm Performance

| 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.

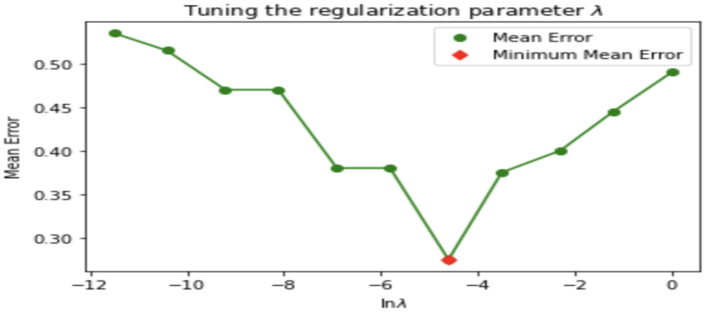
[d] LR stands for Logistic Regression.

[e] DTE stands for Decision Tree with Entropy impurity function.

[f] DTGI stands for Decision Tree with Gini impurity function.

# Results

## Best Setting for the Multilayer Perceptrons Net (1/2)



- The regularization parameter → $10^{-2}$.
- Weight decay → early stopping.
- Hidden Layer: ReLU activation function → 600.
- Output Layer: sigmoid function.

# Results

## Best Setting for the Multilayer Perceptrons Net (2/2)

- Training Algorithm → Adam algorithm
- Initial learning rate → $10^{-2}$.
- Exponential decay rate for estimating the $1th$ 'n' $2nd$ moment vectors → 0.9 'n' 0.999, respectively
- The numerical stability → $10^{-8}$.
- Batch size → 8 examples.

## Best Setting for Decision Trees and Logistic Regression

- Logistic regression → best regularization parameter → $10^{-2}$
- Decision trees → Entropy impurity function

# Results

## Best Setting for Support Vector Machines with a Polynomial Kernel

$$k(x^t, x^s) = [(x^t)^T x^s + 1]^p \tag{1}$$

- The best degree (i.e., $p$)→ 2.
- The best regularization parameter → 0.5. (i.e., $C = 0.5$)

## Best Setting for Support Vector Machines with a Gaussian Kernel

$$k(x^t, x^s) = exp(-\gamma||x^s - x^t||^2) \tag{2}$$

- The best gamma parameter (i.e., $\gamma$) → $10^{-4}$.
- The best regularization parameter → $32 \times 10^4$ (i.e., $C = 32 \times 10^4$).

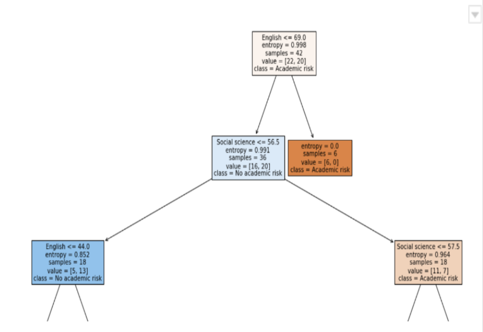
# Results

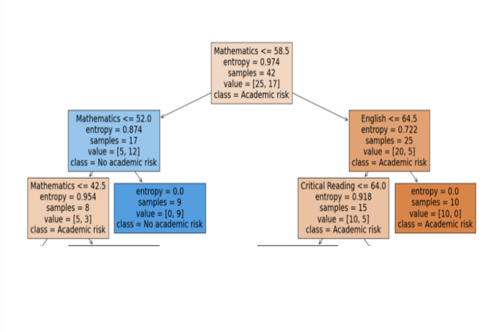
### Student’s paired t-test on mean error

| 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 |

# Discussion

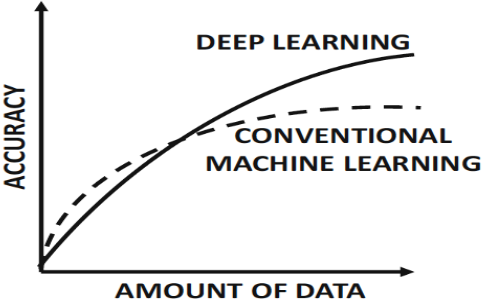
- Mean accuracy 72.5% (Mean error 27.4%) → Tossing a coin (the expected accuracy of about 50%)
- Mean Error of the Multilayer Percetrons net < Mean Error of the other algorithms ($p$-value > 0.05)
- The tree shape changes drastically → This does not allow generalizing the rules for forecasting





# Discussion

Why do we recommend the Multilayer Perceptrons net?



Taken from Aggarwal, C. C. (2018). Neural Networks and Deep Learning.

## The sixth assumption

We assume accuracy is more relevant for improving the user's experience than the interpretation of the forecasting algorithm.

# Discussion

Final remarks:

- 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.
- Using the pre-trained synaptic weights → transfer knowledge to similar contexts (e.g., forecasting based on the test SAT)

# Outline

1 Introduction

2 Prior Research

3 Research Method

4 Evaluation

5 Conclusions and Perspectives

6 Acknowledgment

# Conclusions

(i) Support vector machine with a polynomial kernel is a better choice than Support vector machines with a Gaussian kernel

(ii) Support vector machines and logistic regression have the same mean precision.

(iii) Support vector machine with a polynomial kernel and logistic regression have the same mean recall.

(iv) The decision tree with the entropy impurity function performs better than the one with the Gini impurity function.

(v) The multilayer perceptrons algorithm outperforms the other studied learning algorithms.

(vi) Concerning the research question, given the student's outcomes from the test Saber 11, the student's risk forecasting accuracy of a machine learning algorithm is about 72.5%

# Conclusions

(vii) 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.

(viii) 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.

# Perspectives

(i) We shall collect more data, including more variables.

(ii) 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?

(iii) Will the accuracy increase as more areas are included in the test Saber 11?

# Perspectives

(iv) 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)?

(v) 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. 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?

# Outline

1. Introduction
2. Prior Research
3. Research Method
4. Evaluation
5. Conclusions and Perspectives
6. Acknowledgment

# Acknowledgment

- Caicedo-Castro thanks the Lord Jesus Christ for blessing this project.
- We thank the University of Córdoba in Colombia for the financial support (grant FE-06-17)
- Our deepest appreciation goes to Dr. Jairo Torres-Oviedo, University of Córdoba president.
- We thank the anonymous reviewers for their comments that contributed to improve the quality of this article.

# The end

## That’s all folks

## Thanks for your attention!!!

**Praise the name of God forever and ever, for he has all wisdom and power. He controls the course of world events; he removes kings and sets up other kings. He gives wisdom to the wise and knowledge to the scholars. He reveals deep and mysterious things... (Daniel 2:20-22)**