# PREDICTING PRUEBAS SABER PRO GRADES USING DECISSION TREES.

| Vicente Aristizábal<br>Universidad Eafit<br>Colombia<br>varisti7@eafit.edu.co | | Miguel Correa<br>Universidad Eafit<br>Colombia<br>macorream@eafit.edu.co | Mauricio Toro<br>Universidad Eafit<br>Colombia<br>mtorobe@eafit.edu.co |
|---|---|---|---|

## ABSTRACT

The project's goal is to create an algorithm to predict if a student will get grades above the average in the "Pruebas Saber Pro". Based in the past grades of the test and some sociodemographic variables as age, gender, the time spent on the internet, between others. We consider this problem is important because of the huge impact it could make in the prediction of the variables that can influence the success of a student in the future. The algorithm I proposed was the cart algorithm, the results achieved by the algorithm were satisfactory with results of, 72.58%, 72.53% and 72.54 in precision, recall and accuracy respectively. The time for the algorithm to train with 135.000 lines and to process with 45.000 was of 11:02.

Keywords
Decision trees, machine learning, academic success, standardized student scores, test-score prediction

## 1. INTRODUCTION

Latin America is one of the places with the gratest inequality in access to resources for the education. By the prediction of the results based on socidemographic variables we could find which one are the ones with the greater impact in the results and focus in a way of reducing the impact so the students can have more posibilities of success.

### 1.1. Problem

We have to create an algorithm capable of predicting if the grades of a student will be above the average by creating a decision tree that analyzes sociodemographic variables of and the results in the "Pruebas Saber 11" of the same student. This can have a very positive impact in society because it would help to determine which are the variables that affect the most the success in students.

### 1.2 Solution

In this work, we focused on decision trees because they provide great explainability. We avoid black-box methods such as neural networks, support-vector machines and random forests because they lack explainability.

For the solution of this problem we are going to use the CART algorithm; because it has it bases in the Gini index, which we have studied during the course.

### 1.3 Article structure

In what follows, in Section 2, we present related work to the problem. Later, in Section 3 we present the datasets and methods used in this research. In Section 4, we present the algorithm design. After, in Section 5, we present the results. Finally, in Section 6, we discuss the results and we propose some future work directions.

## 2. RELATED WORK

### 2.1 Estimating student retention and Degree-Completion time: Decision trees and neural networks vis a vis regression.

They solved via three algorithms (C&RT, CHAID-based and C5.0) the problem of analyzing and predicting the freshmen retention and the degree completition time in a study of data from 1995 through 2005 of more tan 20.000 of students, and with an accuracy of 93% with the C5.0 algorithm. Herzog, S. (2006). Estimating student retention and degree- completion time: Decision trees and neural networks vis- à- vis regression. New directions for institutional research, 2006(131), 17-33.

### 2.2 Towards freshman retention prediction: a comparative study.

They solved via the C4.5 algorithm the freshman retention prediction in 7.800 students with an acurracy of 86% overall.. Djulovic, A., & Li, D. (2013). Towards freshman retention prediction: a comparative study. International Journal of Information and Education Technology, 3(5), 494-500.

### 2.3 Performance prediction using classification

They solved the prediction of "At-Risk" status students in their first semester of an undergraduate degree program with the ID3 algorithm with an accuracy over 80%. MOOLIYIL, G. (2019). Performance Prediction Using Classification (Doctoral dissertation, The British University in Dubai (BUiD)).

### 2.4 Predicting students retention.

They solved the predictoin of students who are at risk of dropping of a graduate business program with the CHAID and C&RT algorithm. Eshghi, A., Haughton, D., Li, M., Senne, L., Skaletsky, M., & Woolford, S. (2011). Enrolment Management in Graduate Business Programs: Predicting Student Retention. Journal of Institutional Research, 16(2), 63-79.

## 3. MATERIALS AND METHODS

In this section, we explain how the data was collected and processed and, after, different solution alternatives considered to choose a decision-tree algorithm.

### 3.1 Data Collection and Processing

We collected data from the *Colombian Institute for the Promotion of Higher Education* (ICFES), which is available online at ftp.icfes.gov.co. Such data includes anonymized Saber 11 and Saber Pro results. Saber 11 scores of all Colombian high schools graduated from 2008 to 2014 and Saber Pro scores of all Colombian bachelor-degree graduates from 2012 to 2018 were obtained. There were 864,000 records for Saber 11 and records 430,000 for Saber Pro. Both Saber 11 and Saber Pro, included, not only the scores but also socio-economic data from the students, gathered by ICFES, before the test.

In the next step, both datasets were merged using the unique identifier assigned to each student. Therefore, a new dataset that included students that made both standardized tests was created. The size of this new dataset is 212,010 students. After, the binary predictor variable was defined as follows: Does the student score in Saber Pro is higher than the national average of the period?

It was found out that the datasets were not balanced. There were 95,741 students above average and 101,332 students below average. We performed undersampling to balance the dataset to a 50%-50% ratio. After undersampling, the final dataset had 191,412 students.

Finally, to analyze the efficiency and learning rates of our implementation, we randomly created subsets of the main dataset, as shown in Table 1. The dataset was divided into 70% for training and 30% for testing. Datasets are available at https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets .

| | **Dataset 1** | **Dataset 2** | **Dataset 3** | **Dataset 4** | **Dataset 5** |
|---|---|---|---|---|---|
| **Train** | 15,000 | 45,000 | 75,000 | 105,000 | 135,000 |
| **Test** | 5,000 | 15,000 | 25,000 | 35,000 | 45,000 |

**Table 1.** Number of students in each dataset used for training and testing.

### 3.2 Decision-tree algorithm alternatives

In what follows, we present different algorithms to solve to automatically build a binary decision tree.

#### 3.2.1 ID3

In this algorithm the set of values must be a series of tuples, each of them named attributes. Inside the attributes there is one whom is the objective. This objective must be of binary type. By this way the algorithm tries to set the hypothesis that classifies new entries of tuples in true or false.

Steps of the algorithm: 1. Calculating the entropy of all the attributes. 2. Split the set into subsets by the attribute who minimizes entropy. 3.Make a decision node containing the attribute. 4. Use recursion on subsets using the remaining attributes.

#### 3.2.2 CART

A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning simple, base don the Gini Index classification as metric.

#### 3.2.3 C4.5

It builds the decision tree in the same way the ID3 with the information entropy. At each node the algorithm chooses the attribute that better splits the samples based on normalized information gain. It has the following base cases: all the samples in the list belong to the same class, none of the features provide any information gain and instance of previously-unseen class encountered

#### 3.2.4 Chi square automatic interaction detection (CHAID)

Is based on the adjusted significance testing (Bonferroni Testing).

## 4. ALGORITHM DESIGN AND IMPLEMENTATION

In what follows of the project, the algorithm can be found in https://github.com/varisti7/ST0245-001.

### 4.1 Data Structure

The Data structure we are going to use is a binary tree to make the decisions based on a training method.

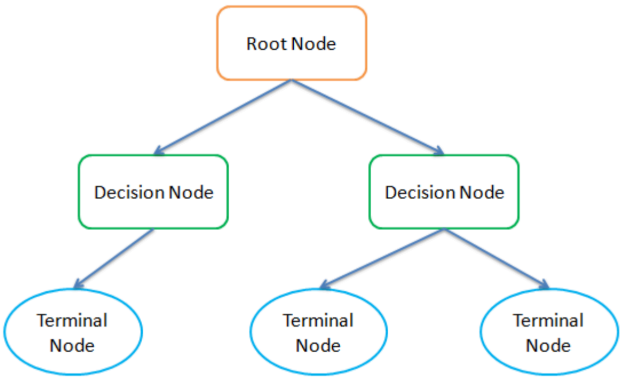

Figure 1: A binary decision tree to predict Saber Pro based on the results of Saber 11.

### 4.2 Algorithms

The algorithm will read some sociodemographic variables from each student, analyze it by some parameters we have chosen that reduce the most the Gini Index, and predict the success of each student.

### 4.2.1 Training the model

We analyzed all the variables, choose the ones that reduced more the Gini Index and that WERE ETHICAL; variables as gender and ethnics were not taken in account.

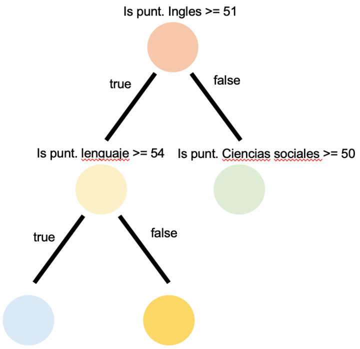

**Figure 2:** Training a binary decision tree CART.

### 4.2.2 Testing algorithm

After the tree is built, the algorithm starts to classify the new information we enter it with the questions that most reduce the gini index, until it has no more questions to ask and classifies it.

### 4.3 Complexity analysis of the algorithms

For the worst case of the algorithm in complexity, it will validate for each level of the tree, in each node, that would be N-1 nodes, understanding N as the number of students to evaluate, all the possible questions and see which would reduce the gini index the most. So this part would be N*M with M as all the possible questions, times log2(N) that would be the height of the tree, and the number of nodes.

| | **Time Complexity** | **Memory complexity** |
|---|---|---|
| Training the model | O(M*N*log(N)) | O(log(N)) |
| Testing the Model | O(N*log(N)) | O(1) |

### 4.4 Design criteria of the algorithm

This algorithm was designed this way because it evaluates all the possible combinations of questions, of the data, and via the Gini index, picks the one or the ones that reduce the most the uncertainty, having very good results: above the 70%.

## 5. RESULTS

### 5.1 Model evaluation

In this section, we present some metrics to evaluate the model. Accuracy is the ratio of number of correct predictions to the total number of input samples. Precision. is the ratio of successful students identified correctly by the model to successful students identified by the model. Finally, Recall is the ratio of successful students identified correctly by the model to successful students in the dataset.

### 5.1.2 Evaluation on test datasets

In what follows, we present the evaluation metrics for the test datasets in Table 4.

The evaluation are for the 135.000 training dataset with the 45.000 test and the 105.000 training set and 35.000 test.

| | ***Dataset 1*** | ***Dataset 2*** |
|---|---|---|
| *Accuracy* | 0.7258 | 0.7249 |
| *Precision* | 0.7253 | 0.7232 |
| *Recall* | 0.7254 | 0.5636 |

**Table 4.** Model evaluation on the test datasets.

### 5.2 Execution times

The execution times are for the 135.000 training dataset with the 45.000 test and the 105.000 training set and 35.000 test.

| | ***Dataset 1*** | ***...Dataset n*** |
|---|---|---|
| *Training time* | 662 | 489 |
| *Testing time* | 0.6 | 0.45 |

**Table 5:** Execution time of the *CART* algorithm for different datasets.

The results obtained were satisfactory for the purpose of the course.

### 6.1 Future work

For the future I would like to implement this algorithm in real world work.

### REFERENCES

Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J., 1984. Classification and Regression Tree Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.

Stephanie, «Decision Tree: Definition and Examples,» 13 09 2015. [En línea].

Available: https://www.statisticshowto.com/decision-tree-definition- and-examples.

EFE, «Perú, Chile y EEUU, entre los 11 países con más