Could we predict a student's success by using statistics?

David Jose Cardona Nieves Universidad EAFIT Colombia djcardonan@eafit.edu.co	Juan David Valencia Torres Universidad EAFIT Colombia jdvalencit@eafit.edu.co	Kevin Mauricio Loaiza Arango Universidad EAFIT Colombia kmloaizaa@eafit.edu.co	Miguel Correa Universidad EAFIT Colombia macorream@eafit.edu.co	Mauricio Toro Universidad EAFIT Colombia mtorobe@eafit.edu.co

Black text = Miguel and Mauricio's contribution

Green text = To complete for the 1st deliverable

Blue text = To complete for the 2nd deliverable

Violet text = To complete for the 3rd deliverable

ABSTRACT

The problem to be solved is the lack of mechanisms that allow predicting the academic success of a student in the Saber Pro colombian tests based on the results obtained previously and their sociodemographic and economic statutes. It is an issue that must be addressed since it is here where future inequalities in society begin, where many times people with great potential are limited by the conditions (often precarious) in which they live. This development could also contribute to the solution of problems of the same nature through the strengthening of programs such as Generación E, Familias en Accion, Ingreso Solidario, etc...

We decided to use the ID3 algorithm, because we consider it to be an easy one to understand, as well as fulfilling its task, perhaps not in the most optimal way, but safe, The results obtained when using the algorithm were 80% accurate, based on this it can be concluded that the development of new forms of development can be crucial to obtain effective results that are totally independent of biased opinions dictated by experts. In this project we obtained a run time of 30 seconds for training the algorithm and 3 seconds for testing it, with 75k data for training and 35k for testing around 900mb, for the training data we have an accuracy of 0.79 for the first set, 0.8 for the second and an approximate accuracy of 0.81 in the n datasets, also we have an accuracy of 0.82 for the dataset one and two, for the n sets we approximate an accuracy of 0.8, finally we have a sensitivity of 0.82 for the first two datasets and 0.8 for the n cases of datasets

Keywords

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

1. INTRODUCTION

Colombia lives under immense socio-economic inequality due to the high dropout rate by students who day after day see how they must face difficulties with almost non-existent aid from the government.

1.1. Problem

Drop-out and poor academic success are a very important problem and must be studied, taking into account the factors and problems that young people in the country live with, to find the causes of such low academic success

We must consider the development of a solution for this problem as a high priority objective since it would reduce the huge social gap that exists in Colombia, thus achieving results that would favor communities that are in terrible conditions.

1.2 Solution

In this work, we focused on decision trees because they provide great explainability (*A citation for this argument is missing!*). We avoid black-box methods such as neural networks, support-vector machines and random forests because they lack explainability. (*Another citation for this argument is missing!*)

Our solution will be the implementation of the ID3 algorithm to have a better prediction of academic success in Saber Pro using data from Saber 11 presented by highschool graduates. This algorithm is really helpful when trying to divide data sets, improving efficiency in the execution time.

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

3.1 Decision trees for predicting the academic success of students this paper classified students into one of two categories, depending on their success at the end of their first academic year, with the purpose of finding meaningful variables affecting student success, three algorithms were used: j48, randomForest and REPTree, showing that the REPTree algorithm had the highest accuracy rate of 79.35%.

3.2 Predicting Student Performance using Classification and Regression Trees Algorithm

The aim of this study was to test the ability of CART analysis to predict success in web-based blended learning environments, by using online interactions stored in the system log file, the CART technique achieved very high accuracy (99.1 %) in classifying students into those who successfully passed the class and those who failed to do so.

3.3 Mining Student Data Using Decision Trees The main objective of this paper is an attempt to use data mining methodologies to study students' performance in the courses, using ID3, c4.5 and Naive Bayes algorithm, showing an accuracy of 38.4%, 35.8%, 33.3% respectively.

3.4 Predicting students' performance using ID3 and C4.5 classification algorithms

This paper analysed the data of students enrolled in first year of engineering and then applied the ID3 and C4.5 algorithms after pruning the dataset to predict the results of these students in their first semester as precisely as possible, showing that the accuracy achieved is 75.145% for both ID3 and C4.5 algorithms.

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. *(In this semester, examples of such algorithms are ID3, C4.5 and CART).*

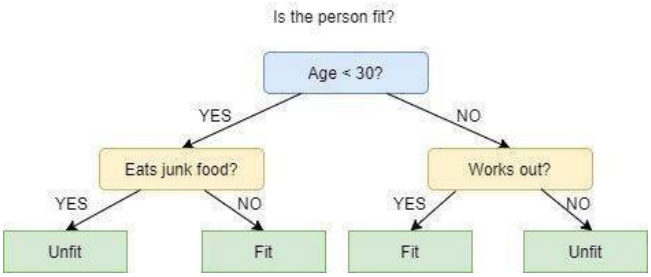
3.2.1 ID3 algorithm

The ID3 algorithm selects the best feature at each step while building a Decision tree.

The initial node is called the root node (colored in blue), the final nodes are called the leaf nodes (colored in green) and the rest of the nodes are called intermediate or internal nodes.

Complexity: **O(m*n)**

The root and intermediate nodes represent the decisions while the leaf nodes represent the outcomes.



3.2.2 C4.5 algorithm

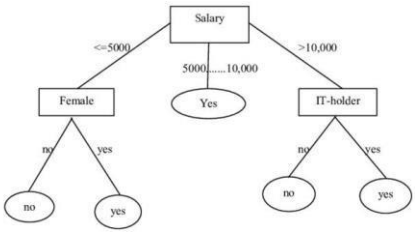
C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. In The C4.5 algorithm made a number of changes to improve ID3 algorithm. Some of these are: • Handling training data with missing values of attributes

• Handling differing cost attributes

• Pruning the decision tree after its creation

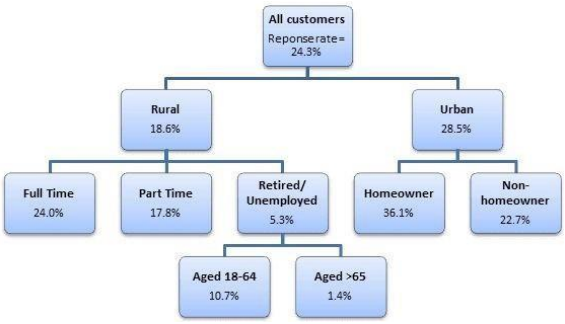
• Handling attributes with discrete and continuous values

Complexity: **$O(m*n^2)$**



3.2.3 CHAID algorithm

Chi-square automatic interaction detection (CHAID) is a decision tree technique, based on adjusted significance testing (Bonferroni testing). The technique was developed in South Africa and was published in 1980 by Gordon V. Kass, who had completed a PhD thesis on this topic. CHAID can be used for prediction (in a similar fashion to regression analysis, this version of CHAID being originally known as XAID) as well as classification, and for detection of interaction between variables.

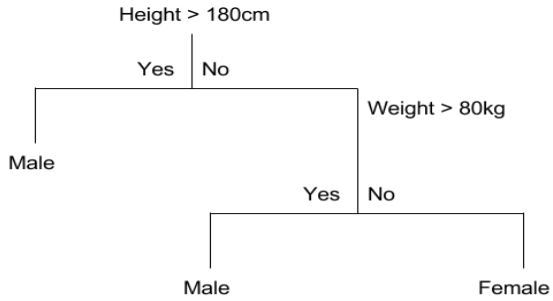


3.2.4 CART algorithm

CART can be used to build both Classification and Regression Decision Trees. When a decision tree is used to separate a dataset into two classes, the model is a classification tree but, when the target variable is numeric or continuous, the predictive taskis regression.

When a classification tree is used, the aim is to split the dataset at hand into two parts using the homogeneity of data as criterion.

Complexity: **O(v ,nlogn)**



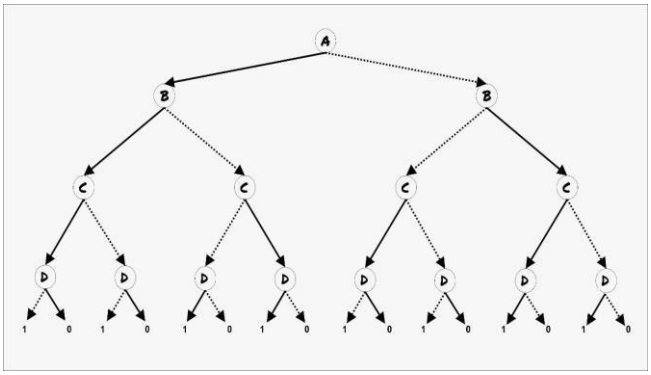
4. ALGORITHM DESIGN AND IMPLEMENTATION

In what follows, we explain the data structure and the algorithms used in this work. The implementation of the [1] data structure and algorithm is available at Github .

4.1 Data Structure

We will be using a binary decision tree implementing the ID3 algorithm, with the intention of optimizing performance seeking to divide the larger data set into subsets that are easier to work with and with greater testing capacity.

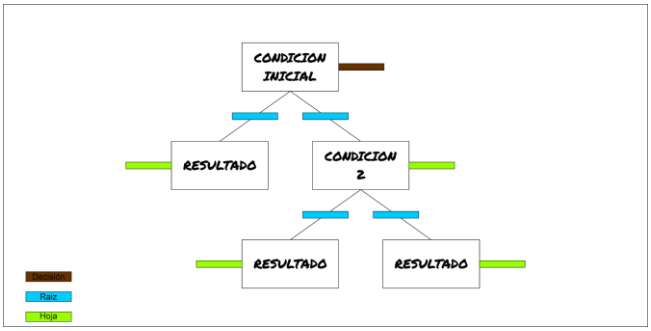
Figure 1: A binary decision tree to predict Saber Pro based on the results of Saber 11. Violet nodes represent those with a high probability of success, green medium probability and red a low probability of success.



4.2 Algorithms

The classification algorithm works by selecting certain variables (and conditions that these variables must meet). When iterating through each of these variables and obtaining different values, the Gini index will be calculated (lower is better).

The development algorithm is in charge of processing these classified data and returning a result (the predicted success).



4.2.1 Training the model

[1]http://www.github.com/ jdvalencit /proyecto/

The creation of the tree (and generation of its leaves) will be carried out under the premise that the "decisions" that have / return a lower Gini index are the most correct for the continuation of the tree; thus discarding the nodes that have a higher index. In addition, the algorithm will evaluate each of the nodes and then (in case of obtaining a satisfactory result) divide them into "subnodes" and optimize the results.

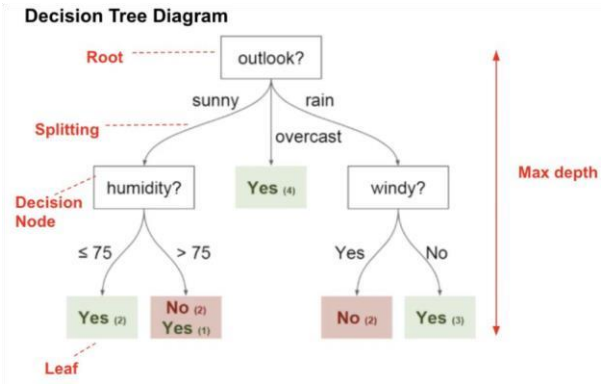


Figure 2: Training a binary decision tree using *(In this semester, one could be CART, ID3, C4.5… please choose)* . In this example, we show a model to predict whether or not to play Golf, according to weather.

4.2.2 Testing algorithm

Explain, briefly, how did you test the model: This is equivalent to explain how does your algorithm classifies new data after the tree is built.

4.3 Complexity analysis of the algorithms

The worst case using O notation.

Algorithm	Time Complexity
Train the decision tree	$O(N^3+ M^2+k)$
Test the decision tree	$O(2^n+M^2)$

Table 2: Time Complexity of the training and testing algorithms.

Algorithm	Memory Complexity
Train the decision tree	$O(N^2*M^2)$
Test the decision tree	O(n)

Table 3: N is the number of examples, M is the number of conditions and K is the constants such as the Hash methods

4.4 Design criteria of the algorithm

The algorithm was designed based on the social concept for the algorithm. That is why the conditions that were considered discriminatory were eliminated and the algorithm was improved so that it could only be used in certain contexts (when the Gini is very high). In addition, the decision to use the ID3 was preceded by the intention to make an easy to understand algorithm based on linear regression.

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. **5.1.1 Evaluation on training datasets**

In what follows, we present the evaluation metrics for the training datasets in Table 3.

	Dataset 1	***Dataset 2***	***...Dataset n***
Accuracy	0.79	0.8	0.81
Precision	0.82	0.82	0.8
Recall	0.82	0.82	0.81

Table 3. Model evaluation on the training datasets.

5.1.2 Evaluation on test datasets

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

	Dataset 1	***Dataset 2***	***...Dataset n***
Accuracy	0.69	0.7	0.71
Precision	0.78	0.88	0.9
Recall	0.78	0.88	0.9

Table 4. Model evaluation on the test datasets.

5.2 Execution times

Compute execution time for each dataset in github. Measure execution time 100 times for each dataset and report average execution time for each dataset.

	Dataset 1	***Dataset 2***	***...Dataset n***
Training time	13 s	20 s	31 s
Testing time	2 s	2.3 s	3.3 s

Table 5: Execution time of the *(Please write the name of the algorithm, C4.5, ID3)* algorithm for different datasets.

5.3 Memory consumption

We present memory consumption of the binary decision tree, for different datasets, in Table 6.

	Dataset 1	***Dataset 2***	***...Dataset n***
Memory consumption	600 MB	730 MB	910 MB

Table 6: Memory consumption of the binary decision tree for different datasets.

To measure memory consumption, you should use a profiler. An very good one for Java is VisualVM, developed by Oracle, http://docs.oracle.com/javase/7/docs/technotes/guides/visualvm/profiler.html For Python, use C Profiler.

6. DISCUSSION OF THE RESULTS

Explain the results obtained. Is precision, accuracy and sensibility appropriate for this problem? Is the model overfitting? Is memory consumption and time consumption appropriate? *(In this semester, according to the results, can this be applied to give scholarships or to help students with low probability of success? For which one is better?)*

6.1 Future work

The algorithm could be improved by adding methods that allow working with more data types and adding the option to return more a binary response, a possibility.

ACKNOWLEDGEMENTS

Identify the kind of acknowledgment you want to write: for a person or for an institution. Consider the following

guidelines: 1. Name of teacher is not mentioned because he is an author. 2. You should not mention websites of authors of articles that you have not contacted. 3. You should mention students, teachers from other courses that helped you.

As an example: This research was supported/partially supported by [Name of Foundation, Grant maker, Donor].

We thank for assistance with [particular technique, methodology] to [Name Surname, position, institution name] for comments that greatly improved the manuscript.

REFERENCES

http://www.ijitee.org/wp-content/uploads/papers/v9i3/C8964019320.pdf
https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/
https://en.wikipedia.org/wiki/ID3_algorithm
https://arxiv.org/ftp/arxiv/papers/1310/1310.2071.pdf

1. Mesarić, Josip., Šebalj, Dario. Decision trees for predicting the academic success of students. *Croatian*

Operational Research Review.7(2016), 367–388. Recuperado de http://bib.irb.hr/datoteka/853222.clanak.pd f

2. Krishna, M.,Bandlamudi, Rani, *et al.* Predicting Student Performance using Classification and Regression Trees Algorithm. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(3). Recuperado de http://www.ijitee.org/wp-content/uploads/papers/v9i3/C8964019320.pdf

3. Al-Radaideh, Qasem, Al-Shawakfa, Emad, Al-Najjar, Mustafa, Mining Student Data Using Decision Trees. *The 2006 International Arab Conference on Information Technology (ACIT'2006)*. Recuperado de https://www.acit2k.org/ACIT2006/Proceeding/131.pdf

4. Adhatrao, Kalpesh, Gaykar, Aditya, *et al.* Predicting students' performance using ID3 and C4.5 classification algorithms, International Journal of Data Mining & Knowledge Management Process (IJDKP), 3(5). 39-52. Recuperado de https://arxiv.org/ftp/arxiv/papers/1310/1310.2071.pdf