PREDICTING ACADEMIC SUCCESS IN HIGHER EDUCATION USING DECISION TREES

Isabella Pérez	Mariana Quintero	Miguel Correa	Mauricio Toro
Eafit University	Eafit University	Eafit University	Eafit University
Colombia	Colombia	Colombia	Colombia
iperezs2@eafit.edu.co	mquintero3@eafit.edu.co	macorream@eafit.edu.co	mtorobe@eafit.edu.co

SUMMARY

The objective of this report is to predict the academic success in higher education taking into account the sociodemographic and academic variables of each student obtained previously. To do this, we seek to develop a program and data structure that allows us to read, save and organize the data from the files to be used, so that we can apply useful tools such as decision trees and mathematical algorithms, and with them predict the probability that a student will obtain an above-average result.

The solution to this problem is of great importance because, despite the fact that technology advances exponentially, little has been achieved in predicting academic success in higher education, and it would be incredibly useful to have more reliable predictive data in order to avoid taking exams with little probability of success. Therefore, the efficient development of this project can offer the ability to know with greater certainty what the probability of success would be for each person in a higher education test. There are several problems similar to the one raised in this report and some of these will be analyzed in order to find a more effective solution.

Keywords

Decision trees, automatic learning, academic success predicting test results.

1. INTRODUCTION

It is clear that the role of technology is a key factor in Colombian education that little by little becomes more a vital part of it, which inevitably leads us to want to take new steps or advances in the situation. Previous studies have shown the factors that influence the academic performance of students and other variables such as desertion and its causes or motivations, and based on that we have sought to obtain predictive results on this subject. However, very little has been achieved in terms of predicting academic success in higher education, since various ways of measuring success can be considered, such as the employability of graduates, the salary of graduates, and the happiness of the work of graduates, among others.

Taking into account the above, the purpose of this project is, then, to achieve predictions regarding academic success in higher education, taking this as the probability that a student has of obtaining a higher than average score in the Pro tests.

1.1. Problem

By using algorithms based on decision trees, and on academic and socio-demographic data of a student to design and implement a program that allows predicting the academic success of the student; with this prognosis the student knows definitively the level that he is with respect to the average, and thus to take decisions about his academic life (university), finally the society can prepare itself in an early way for these tests thanks to the result that the algorithm generates.

1.2 Solution

In this work, we focus on decision trees because they provide a great explanatory power, as it has been expressed in Revista d'Innovació i Recerca en Educació, where it points out that decision trees are an analysis tool that allows in principle to express in a graphic way, and later, under a mathematical schematization, the different paths, variables, causes and effects susceptible to materialize as a product of the actions derived by the participating individuals, despite recognizing in any case, that under conditions of uncertainty and risk, elements of stochastic or random type converge in each phase (Berlanga, Rubio and Vilà, 2013). We avoid black box methods such as neural networks, vector support machines and random forests because they lack explanability, since the use of a random selection of characteristics to divide each node produces high error rates that compare favorably to the AdaBoost algorithm (Freund and Schapire, 1996). The solution is an implementation of a decision tree algorithm to predict academic success. The CART algorithm was chosen because it is an algorithm that is used to generate decision trees, these can be used for classification, therefore, CART can process two discrete data as continuous, also distinguishes the variables

influential through the gini impurity, therefore this is the ideal algorithm to predict academic success.

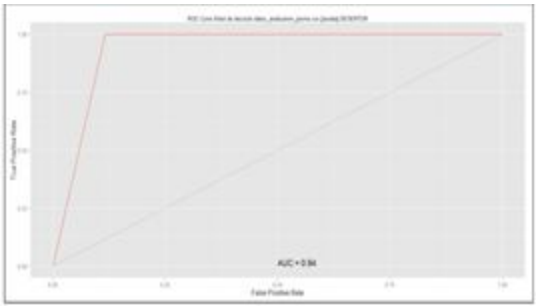
1.3Article structure

In the following, in section 2, we present the work related to the problem. Later, in section 3, we present the data sets and Subscribe to DeepL Pro to edit this document. Visit www.DeepL.com/Pro for more information. methods used in this research. In section 4, we present the design of the algorithm. Then, in section 5, we present the results. Finally, in section 6, we discuss the results and propose some directions for future work.

2. RELATED WORKS

2.1 Student Dropout Based on Decision Trees

This algorithm seeks to predict the probability that a student will abandon his or her academic program, through classification techniques, based on decision trees; by means of Classification and Regression Tree (CART) a tree was constructed with four levels of depth that evaluates possible dropouts. This leads to the conclusion that the level and grade variables have a greater influence on desertion. The following image (Figure 1) shows the ROC curve, where it is concluded that the model has a 94% effectiveness in prediction [1].



(figure 1).

2.2 Decision trees to predict factors associated with the academic performance of high school students on the Saber 11 tests

This algorithm seeks to detect factors associated with the academic performance of Colombian students in the eleventh grade of secondary education in the Saber 11° tests. This model allows us to predict which are the socioeconomic, academic, and institutional factors associated with the Saber 11° test score. A classification model based on decision trees was used, using the J48 algorithm of the WEKA tool, which implements the C.45 algorithm. The J48 algorithm is based on the use of the criterion of information gain, in this way it is possible to avoid that the variables with the greatest number of possible values benefit from the selection. The most important parameter that was taken into account for the pruning was the confidence level (C), which influences the size and predictability of the constructed tree. The lower that probability is made, the more significant the difference in prediction errors before and after pruning will be required for not pruning. Finally, the accuracy of 67% was obtained. [2]

2.3 Predicting academic performance by applying data mining techniques

This algorithm seeks to predict the academic performance of students in higher educational institutions, in order to predict the final classification (Failed or Passed) of the future students enrolled in the different courses. Decision tree and Bayesian network algorithms are used with the Weka program and with cross validation 10 folds; several TMDs (C4.5, ID3, CART and J48, Bayes' Nail, Neural Networks, k-media and k-neighbor) are compared to predict academic performance. Finally, data from the University of Thailand were used, which showed that the decision tree was 85.2% accurate. [3]

2.4 Prediction of student dropout in Private Higher Education.

This algorithm seeks to predict the desertion or abandonment in the Private Higher Education, the methodology CRIPSDM with spss clementine 12.0 of data mining decision trees was used. It was used 1761 data from students of César Vallejo Private University of the Professional School of Systems Engineering, for the analysis of these data 27 attributes were handled for each one of them that are related to the student's desertion. Finally, training, validation and testing were done with 100 new data where an accuracy of 89% was obtained. [4]

3.MATERIALS AND METHODS

This section explains how the data was collected and processed, and then how different solution alternatives were considered to choose a decision tree algorithm.

3.1 Data collection and processing

We obtained data from the Colombian Institute for the Promotion of Higher Education (ICFES), which is available online at ftp.icfes.gov.co. These data include anonymized results from Saber 11 and Saber Pro. Saber 11 results were obtained for all Colombian high school graduates, from 2008 to 2014, and Saber Pro results were obtained for all

Colombian undergraduate graduates, from 2012 to 2018. There were 864,000 records for Saber 11 and 430,000 for Saber Pro. Both Saber 11 and Saber Pro included not only the scores but also socioeconomic data of the students, collected by ICFES, before the test. In the next step, both sets of data were merged using the unique identifier assigned to each student. Therefore, a new data set was created that included students who took both standardized tests. The size of this new data set is 212,010 students. The binary predictor variable was then defined as follows: Is the student's score on the Saber Pro higher than the national average for the period in which he/she took the test? The data sets were found to be unbalanced. There were 95,741 above-average and 101,332 below-average students. We sub-sampled to balance the data set at a ratio of 50%- 50%. After sub-sampling, the final data set had 191,412 students. Finally, to analyze the efficiency and learning rates of our implementation, we randomly created subsets of the main data set, as shown in Table 1. Each data set was divided into 70% for training and 30% for validation. The data sets are available at https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets .

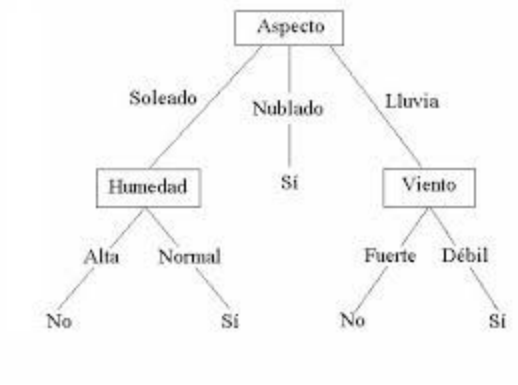
	Data sets 1	Data sets 2	Data sets 3	Sata set 4	Data sets 5
Training	15,000	45,000	75,000	105,000	135,000
Validation	5,000	15,000	25,000	35,000	45,000

Tabla 1. Number of students in each data set used for training and validation.

3.2 Decision tree algorithm alternatives

3.2.1 ID3 Algorithm

This type of decision tree is based on a greedy search, in which the most beneficial option is chosen, this algorithm continuously divides the properties into several groups. The greedy approach is used when each gain can be collected at each step, so that no option blocks another, finally the algorithm chooses as the best feature as the one that generates more current gain of information.



3.2.2 Classification And Regression Trees - CART

The model admits nominal, ordered and continuous input and output variables, so it can solve classification and regression problems, select the cut that leads to the greatest decrease in impurity. This algorithm uses the gini index to calculate the impurity measurement (Figure 2). The cut in each node is given by binary type rules.

$$G(A_i) = \sum_{j=i}^{M_i} p(A_{ij}) G(C/A_{ij})$$

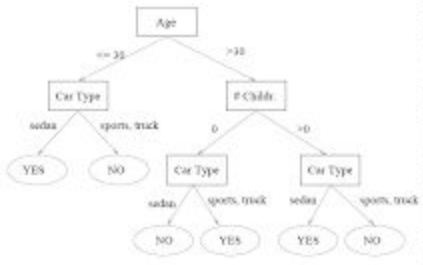
Siendo, $G(C/A_{ij})$ igual a:

$$G(A_{ij}) = -\sum_{j=i}^{M_i} p(C_k/A_{ij})(1 - p(C_k/A_{ij}))$$

- A_{ij} es el atributo empleado para ramificar el árbol,
- J es el número de clases,
- M_i es el de valores distintos que tiene el atributo A_i
- $p(A_{ij})$ constituye la probabilidad de que A_i tome su j-ésimo valor y
- $p(C_k/A_{ij})$ representa la probabilidad de que un ejemplo sea de la clase C_k cuando su atributo A_i toma su j-ésimo valor.

El índice de diversidad de Gini toma el valor cero cuando un grupo es completamente homogéneo y el mayor valor lo alcanza cuando todas las $p(A_{ij})$ son contantes, entonces el valor del índice es $\frac{(J-1)}{J}$.

(figure 2).

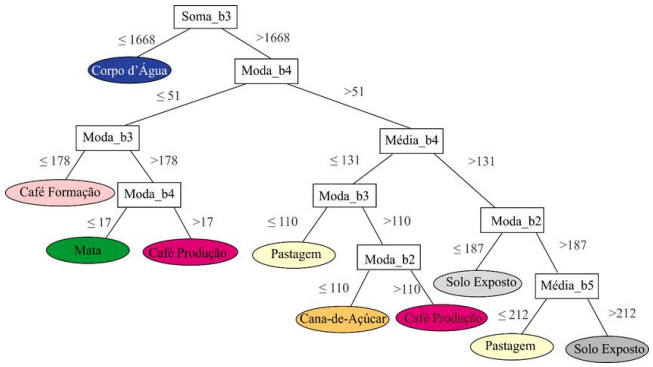


Car type	Age	Chd	Sub
sedan	23	0	yes
sports	31	1	no
sedan	36	1	no
truck	25	2	no
sports	30	0	no
sedan	36	0	no
sedan	25	0	yes
truck	36	1	no
sedan	30	2	yes
sedan	31	1	yes
sports	25	0	no
sedan	45	1	yes
sports	23	2	no
truck	45	0	yes

3.2.3Algorithm C4.5

This algorithm is an extension of the ID3 algorithm, which is used for statistical classification and generates a decision

tree from a set of training data (such as ID3); each example is a vector, which represents the attributes or characteristics of the example. This algorithm selects attributes that most effectively divide the sample into enriched subsets and selects the attribute with the highest information gain as a decision parameter.



3.2.4 Chi-square Automatic Interaction Detector (CHAID)

The CHAID algorithm is based on AID (Automatic Interaction Detection). The algorithm has two very important limitations; on one hand, it is due to the large number of sample elements needed to perform the analysis, and on the other hand, the lack of explanation or determination of dependent and explanatory variables. The grouping of categories is given by testing all possible binary combinations of the variables, the F statistical test is used to select the greatest possible difference. In this algorithm, the sample segmentation process is in dichotomous groups.

4. DESIGN OF THE ALGORITHMS

In what follows, we explain the structure of the data and the algorithms used in this work. The implementation of the algorithm and data structure is available on Github1 .

4.1 Data structure

The outputs are taken after processing the data set of the file and stored in a temporary structure and abstracted for processing, with the binary decision structure, this begins with the root node and branches in two with possible results, and so with each sub-node, until you reach a sheet that is its final result. Within the tree there are three types of nodes, the probability nodes (show the probability of the results), the decision nodes (represent the condition), and finally the terminals (are the final decision).

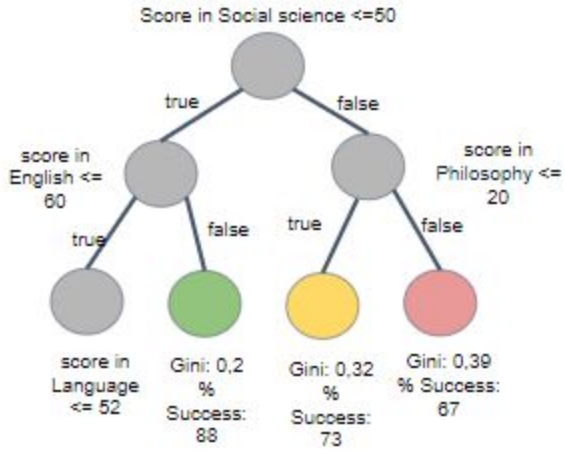


Figure 1: A binary decision tree to predict Saber Pro scores based on the results of Saber 11. Green nodes represent those with a high probability of success, yellow medium probability and red a low probability of success.

4.2 Algorithms

A random forest is a set of decision trees that are completely independent of each other and its main feature is that it is built using random data taken from the general data set. The final result, which is achieved through the establishment of a random forest, is decided by the majority of the trees that make it up, which is the democratic model. The decision tree can be understood as a representation of the process involved in the classification task. In the CART algorithm, the Gini impurity or Gini index is used to select the attribute.

4.2.1 Model training

The algorithm is responsible for dividing the data set worked according to the Gini impurity that is calculated, that condition with less impurity will determine the decision node, so it will run in a recursive manner until finding the best option. For the creation of the tree, the variable found that best divides the data worked on is used as the root, and two branches or arms of the left and right tree are created, and it continues in an analogous way with the subgroups created until the prediction sought is reached.

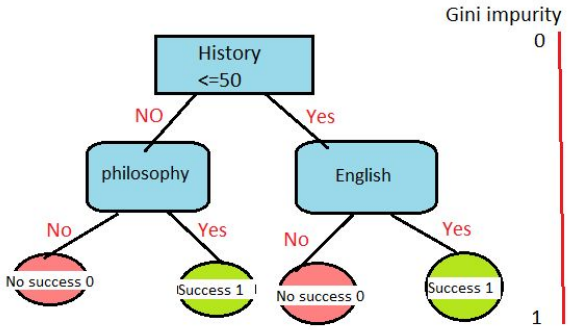


Figure 2: Training of a binary decision tree using CART.

4.2.2 Test Algorithm

This tree determines the label of each person based on each decision tree that predicts whether or not it is higher than the average in the known Pro results. Taking into account the results of each tree in the forest, the final prediction is determined Next , compare the results given by the learning algorithm with the real results and calculated the percentage of accuracy, precision and sensitivity of the algorithm.

4.3 Analysis of the complexity of the algorithms

Algorithm	The complexity of time
Train the decision tree	O(M*t*nlogn)
Validate the decision tree	O(N*t)

Table 2: Complexity in time and memory of the algorithm used that implements random forests based on the CART algorithm.

Here we can interpret the variable n as the number of students, m the number of variables to be taken into account as relevant and t as the number of trees that make up the forest

Algorithm	Memory complexity
Train the decision tree	O(N*M)
Validate the decision tree	O(M*N)

Table 3: The variable n as the number of students, m the number of variables to be taken into account as relevant and t as the number of trees that make up the forest

5. RESULTS

5.1 Model evaluation

In this section, we present some metrics to evaluate the model. Accuracy is the ratio of the number of correct predictions to the total number of input data. Accuracy. It is the proportion of successful students correctly identified by the model and successful students identified by the model. Finally, Sensitivity is the proportion of successful students correctly identified by the model and successful students in the data set.

5.1.1 Evaluation of the model in training

Below are the evaluation metrics for the training data sets in Table 3.

	Data set 1	*Data set 2*	*...Data set n*
Accuracy	0.7	0.75	0.9
Precition	0.7	0.75	0.9
Sensitivity	0.7	0.75	0.9

Table 3. Evaluation of the model with the training data sets.

5.1.2 Evaluation of the validation data sets

Below are the evaluation metrics for the validation datasets in Table 4.

	Data set1	*Dataset 2*	*...Data set n*
Accuracy	0.5	0.55	0.7
Precition	0.5	0.55	0.7
Sensitivity	0.5	0.55	0.8

Table 4. Model evaluation with validation data sets.

5.2 Execution times

Calculate the runtime of each data set in Github. Measure the runtime 100 times, for each data set, and report the average runtime for each data set.

	Data set 1	*Data set2*	*...Data set n*

Training time	10.2 s	20.4 s	5.1 s
Validation time	1.1 s	1.3 s	3.3 s

Tabla 5: Running time of the CART algorithm) for different data sets.

5.3 Memory consumption

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

	Data set 1	***Data set 2***	***...Data set n***
Memory consumption	10 MB	47 MB	106 MB

Tabla 6: Binary decision tree memory consumption for different datasets

6. DISCUSSION OF RESULTS

sensitivity has values of around 05, this means that the algorithm is not biased to better predict those who have succeeded and vice versa, i.e. this algorithm meets the objective. The implemented algorithm presents an adequate accuracy and precision when receiving the same set of training and validation data, which indicates that the model is not too adequate, that is, the data are not memorized. The accuracy and sensitivity do not change significantly with the number of trees in the forest or the amount of data they receive.

6.1 Future Work

The algorithm can be significantly improved by collecting large amounts of data and running it with less memory and time, adding different decision variables to identify the problem that increases the success rate and failure rate.

AGRADECIMIENTOS

We would like to thank Mauricio Toro, teacher of Data Structures and Algorithms, and Simón Marín Giraldo, instructor of the subject Data Structures and Algorithms 1. We recognize the help of Laura Zapata, partner of the EDA course, regarding the structure of the algorithm for a random forest and orientation of the information.

REFERENCIAS

1. Cuji, B.; Gavilanes, W. and Sanchez, R. 2017. Predictive model of student dropout based on decision trees. in Revista Espacios. Vol.38(N°55), 17-26. DOI: http://ww.revistaespacios.com/a17v38n55/a17v38n55p17.p df
2. Caicedo Zambrano, J; Timarán Pereira, R and Hidalgo Troya, A. 2019. Decision trees to predict factors associated with the academic performance of high school students in the Saber 11°. tests in Revista de Investigación, Desarrollo e Innovación. Vol.9(N°2), 363-378. DOI: https://dialnet.unirioja.es/servlet/articulo?codigo=6976577
3. Menacho Chiok, C. H. 2017. Prediction of academic performance applying data mining techniques. in La Molina Journals. Vol.78(N°1), 26-33. DOI: http://revistas.lamolina.edu.pe/index.php/acu/article/view/8 11/pdf_43
4. Daza, A. 2016. A decision-tree model for predicting student dropout in private higher education. Scientia. Vol.8(No.1), 59-73. DOI: https://dialnet.unirioja.es/servlet/articulo?codigo=6181472
5. Ramírez, Patricio E., and Grandón, Elizabeth E. 2018. Prediction of student dropout in a Chilean public university by means of classification based on decision trees with optimized parameters. Scielo. https://dx.doi.org/10.4067/S071 8-50062018000300003
6. Tedesco, A. and Felippe, A. 2015. Integration of OBIA, decision trees and hierarchical classification for throat mapping. ResearchGate. https://www.researchgate.net/figure/Figura-7-Arvoredecisao-gerada-com-o-algoritmoCART_fig4_276204366
7. Timarán-Pereira, R., Caicedo-Zambrano, J., and HidalgoTroya, A. 2019. Decision trees to predict factors associated with the academic performance of high school students on the Know 11 tests. Uptc. https://revistas.uptc.edu.co/index.php/investigacion_d uitama/article/view/9184
8. Vizcaino, P. 2008. Application of decision tree induction techniques to classification problems using weka (waikato environment for knowledge analysis). Docplayer. https://docplayer.es/5563204-Aplicacion-detecnicas-deinduction-of-decision-trees-problems-of-classificationbyuse-deweka-waikato-environment-for-knowledge.html