PREDICTION ALGORITHM FOR THE ACHIEVEMENT OF TEST "SABERPRO"

Juan Felipe Lopez Gutierrez
Universidad Eafit
Colombia
jflopezg@eafit.edu.co

David Restrepo Ramirez
Universidad
Colombia
drestrepor@eafit.edu.co

Mauricio Toro
Universidad Eafit
Colombia
mtorobe@eafit.edu.co

ABSTRACT

The Colombian social context is broad and plural and in turn it manages to be a key part of the spectrum of possible decisions that a person can make throughout his life. Many of them can be classified and ordered through a study of the same person, in most cases being able to even predict future behaviors or situations. The student phenomenon of high desertion within higher education in the country is added to the study, this is a serious problem given that it is added to and integrally related to the low reception of young people with the opportunity to enter quality university education and causing within society a deficit of graduates in many of the branches of knowledge.

It is there where one enters then to the study of available variables and to see if it is possible to achieve a correlation between these data and the possibility that any student can be successful or not within the university environment.

Keywords

Matrices, decision trees, Gini impurity, data structure, operations, big O notation.

Keywords of the ACM classification

Theory of computation → Design and analysis of algorithms → Design and analysis of data structures→ data compression.

1. INTRODUCTION

The document seeks to address a solution to a situation as worrisome within the educational system as dropout or failure during and after university studies and thus, according to the student's general data, be able to give a timely prediction through an algorithm and thus, to guide in a more optimal way the future of Colombian youth within the education system. In addition to this, the document will take a look at related problems in the context of data structures.

The tool that we will use to tackle this problem will be the implementation of decision trees, given their advantages over other data structures; such as software "machine learning"; since in this way it will be able to obtain faster responses when any of the conditions is met. In addition, the implementation of this type of structure does not require as much memory, nor as much computing capacity as neural networks do.

But what is a decision tree?
A decision tree is a prediction model in which, given a set of data, these are categorized based on an "objective" criterion, so that we can make one decision or another depending on the situation.

We have some attributes with different possibilities; Depending on the possibility that we find ourselves, we will take one way or another until a decision is made.

Finally, the algorithm, its structure, formulation and justification will be explained, which, with the datasets granted, seek to provide the most optimal solution to the prediction of desertion through the use of decision trees.

1. TROUBLE

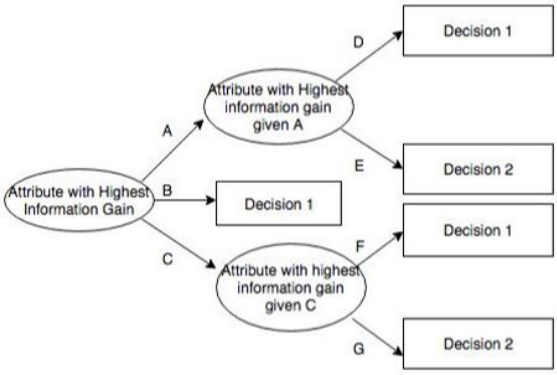
The basic problem to solve is, to find a prediction about the future success of a student based on the study of variables raised in their evaluation of tests to know. Thanks to the great variety of data belonging to different areas of the student concerned, it is possible to achieve the resolution of a virtual approach to the future reality of the same through the application of an algorithm. The implementation of this type of dropout solution would be of great help and thus be able to better guide the future of Colombian higher education, helping to reduce the worrying registered dropout rates that reach 42% in the country (1). Using the algorithm, a correlation could be made in advance between the student's data and thus predict their success, in order to ensure as much as possible the success of the student and simultaneously of the Colombian society, having as a

framework of achievement and achievement within of their studies be above the average of their corresponding test to know pro.

2. RELATED WORKS

2.1 ID3

This is one of the pioneering decision tree algorithms. It works by hierarchically organizing the attributes that you have, and creating nodes in them; being the first or "best" those with the least entropy. So, what you do is give an answer based on the data you have, first evaluating the case in the attributes with the least entropy (to see if you can get an immediate answer) and then going to those that follow in the hierarchical order; this, repeating it throughout the decision tree.



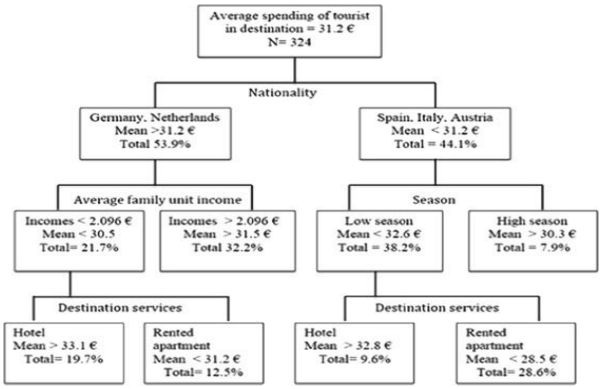
2.2 C4.5

It is the evolution of the ID3, but with some improvements that optimize the use of memory. It differs from its predecessor in that it can create multiple nodes of the same attribute. In addition, it "prunes" the branches of the tree that do not provide extra information, such as whether all the alternatives in an attribute are the same.

Additionally, it is easier to implement, since it has few base cases.

3.3 CHAID

This is also a classification algorithm. It seeks to find criteria for the identification or differentiation of sets, based on a categorical variable (that separates them into two groups). From the criteria found, choose the best based on how much the sets that generate these criteria can be differentiated based on the categorical variable. This, he does with a formula. Arguably, it works similarly to CART. It is an algorithm widely used in statistics



4. DATA STRUCTURE - CART

CART decision trees create child nodes with two branches for each conditional evaluated going through the tree. The height will be the same as the number of variables in the dataset.

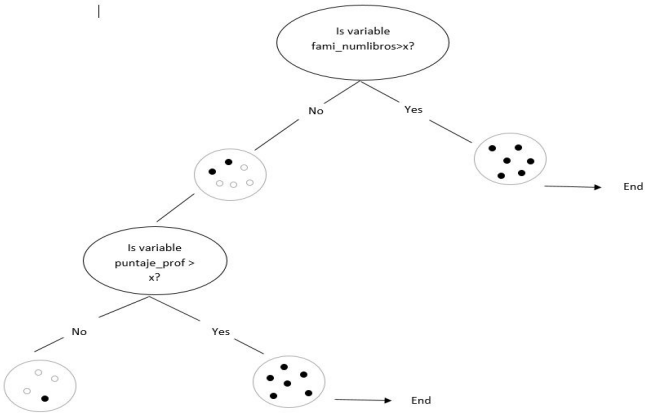


Figure 1: CART data structure evaluates and creates new child nodes based on the results it obtains.

4.1 Operations of the data structure

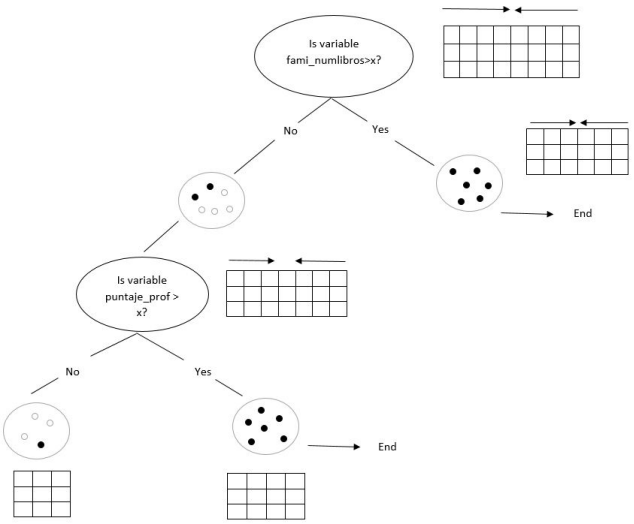


Figure 2: CART tree algorithm is going through the data set. as it finds the lowest gini impurity which divides the data, it creates nodes with conditions and takes the variables processed out the matrix to proceed with the next variables.

4.2 Design criteria of the data structure

Our data structure is a CART binary tree. We built the algorithm functions that call themselves recursively until all the data is processed. This will ease the process because we can then make it work with any problem size. Also, recursion provides better Big-O complexity.

4.3 Complexity analysis

Operation	Complexity
main(String [] args)	**O(n.m)**
leerArchivo(String data)	**O(n.m)**
seleccionarDataset()	**O(n)**
llenarMatriz(double [][][] matriz)	**O(n.m)**
llenarImpureza(double [][][] matriz)	**O(n.m)**
menores(double [][][] matriz)	**O(n)**
addChild(Node childNode, int position)	**O(1)**
addNewNode(Node u, Object info, int i)	**O(1)**
numberOfNodesInTree(Node rootNode)	**O(n)**

Table 1: Table to report complexity analysis

4.4 Execution time

Operation	Execution time
leerArchivo(String data)	**0.0ms**
seleccionarDataset()	**0.0ms**
llenarMatriz(double [][][] matriz)	**0.0ms**
llenarImpureza(double [][][] matriz)	**0.0ms**
menores(double [][][]matriz)	**0.0ms**
addChild(Node childNode, int position)	**0.0ms**
addNewNode(Node u, Object info, int i)	**0.0ms**
numberOfNodesInTree(Node rootNode)	**0.0ms**

Table 2: Execution time of the operations of the data structure for each data set

4.5 Memory used

Data set	Memory used
data_set_train.csv	11.0MB
data_set_test.csv	3.7MB

Table 3: Memory used for each operation of the data structure and for each dataset

4.6 Results analysis

Finally we solved the problem and the decision tree could finally determine if a student is going to have a success in the career.

6. CONCLUSIONS

In this research we found out that decision trees can determine thins and accelerate the solution time for a kind of problem.

Our data structure could finally determine if a student had the possibility of study success

Decision trees can have many applications to real life problems. They can be used for predicting and preventing many negative things.

6.1 Future work

We would like to adapt our algorithm for can work on many other kind of problems, maybe renew the algorithm for better performance and times when a major quantity of data are going to be used.

ACKNOWLEDGEMENTS

We are especially grateful for the help and advice given by the Data Structures monitor Simon Marin, student of systems engineering at EAFIT, who helped us extensively with our questions about the project.

We also thank to Isabel, Daniel and Miguel, students from EAFIT University in Systems Engineering and Mathematics for the support during the course, the resolution of our questions and the unconditional help during the course

REFERENCES

1. Wikipedia. 2018. C4.5. (9 April 2018). Retrieved *May 11, 2020* from https://es.wikipedia.org/wiki/C4.5

2. Wikipedia.2019.ID3 algorithm. (22 may 2019). retrieved *May 12, 2020* from https://en.wikipedia.org/wiki/ID3_algorithm

3. Synergy37AI. 2019. Tree algorithms: ID3, C4.5, C5.0 and CART.(20 February 2019). Retrieved May 12, 2020 from https://medium.com/datadriveninvestor/tree-algorithms-id3c4-5-c5-0-and-cart-413387342164