**Final project of data structures and algorithms:**

Antonio Carmona Gaviria

acarmonag@eafit.edu.co

Juan Esteban Cardona Ospina

[jecardonao@eafit.edu.co](mailto:jecardonao@eafit.edu.co)

**Introduction:**

The purpose of this project was to analyze de results of the ICFES exam, with the objective to see what are the factors that determine de success of a student in the future in the life. But the principal challenge for us, to sort out the information, we created a program that builds a tree using different data sets of dates that contains the results of the ICFES exam in a year. With that characteristics we can compare these results with the results of previous years, and then we can determine what’s going to help the students to obtain better results in the future.

Then, with that information we will try to make a program that can predict and help us to achieve better results un the ICFES. As long of the semester we studied the different types of algorithms to choose the perfect one and create our tree, that make these calculus efficient, and also this can aid and serve as a model to the educational system to take care of the different areas that will need to reinforce.

**Problem and Statistics:**

We did a search of the different problems or causes that have influences in the exit of the students when they present the test. We found different problems that can explain the bad results of the ICFES exam, all of that it is rounded in one problem, the inequality of the schools, the average shows that the private institutions has more probability of success than the public or rural schools, even the public schools they don't have such low results, but the percentage demonstrate a decrease than years ago. The percentage shows a calculus of 226 points for rural schools and for private schools and public shows 263 points. This is the principal problem that the teachers explain the breach of the difference of results, also we don’t count with the people that born in the camp, and the formation of the different teachers of the country.

Then we pass to the statistics, according to the world bank the percentage of drop out rate 42%, that means that Colombia is the second country in Latin America with this dropout rate. So, this will need to take in a count for the improvement of the country.

**Algorithm to consider:**

As long of the semester we saw a lot of forms of algorithms to search the better one to carry out this project, but we are among 3 clear algorithms with which we think that we can understand each other very well and at the same time work efficiently, the CART, the C4.5 and the CHAID.

Here we are going to explain each one, and then we show the algorithm that we selected, so the first is the CART, this decision tree learning, this algorithm it is used to predict models approach in data meaning and in matching learning, this algorithm the structure of the tree is, the leaves represent the classes, and the branches represent conjunctions of features to represents the labels.

The second one is the C4.5, this algorithm is as the CART, this one create classification trees but with a variant, is often referred to statistical classifier, also uses the concept of information entropy (Random information), each one consist in a p-dimensional vector.

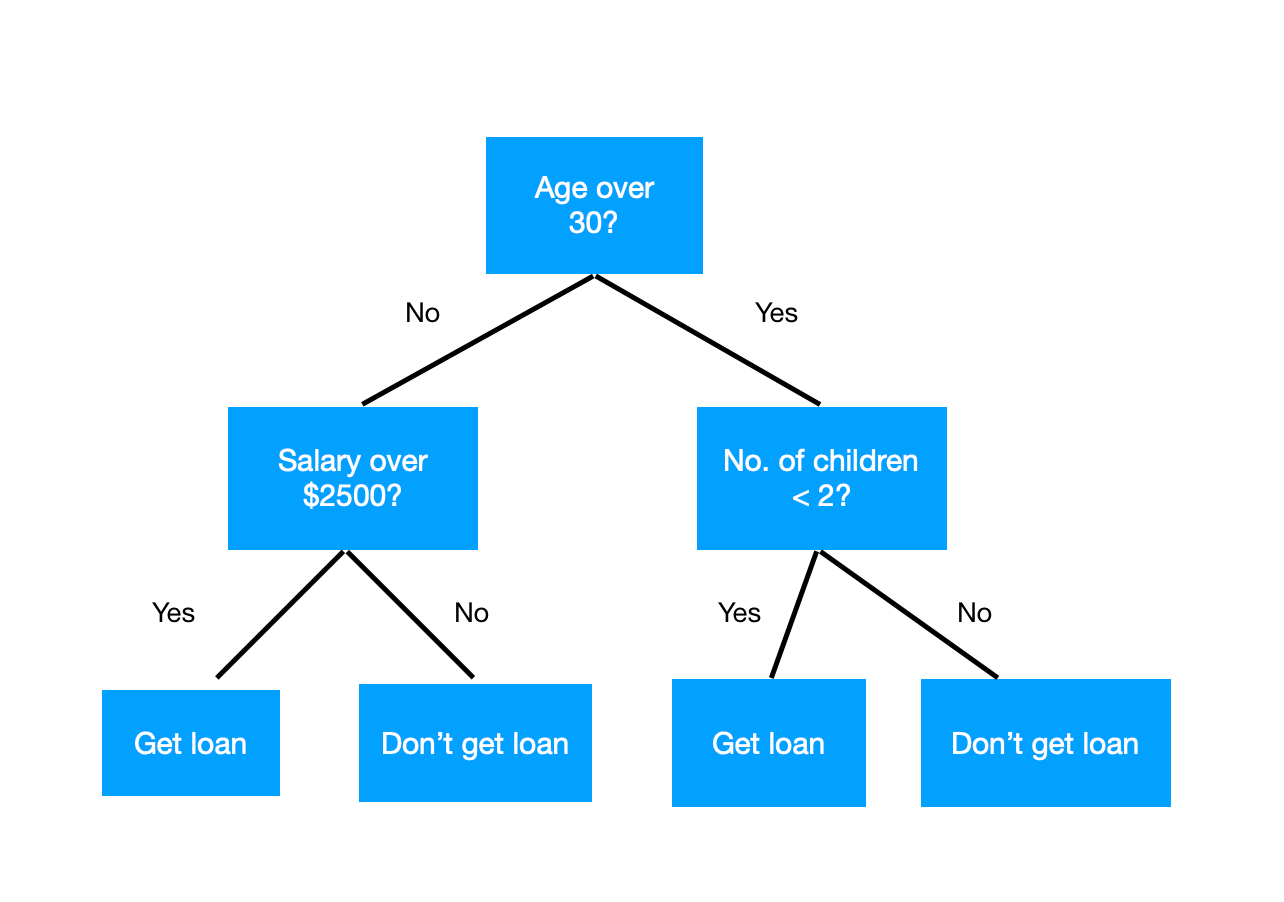
The last one, is the CHAID is based on adjusted significance testing on, that means that the algorithm compare the others labels to counteract the problem or the search that the tree need to do, it is used for predictions and regression analysis and has a difference between the others, this one is used more for marketing than matching learning or the management of the dates.

**Algorithm that we choose:**

Analyzing the problem, they posed, how we could predict whether a person could do well in the saber pro tests, and what algorithms we could work with to find the solution to this problem. We find that there are 1 main algorithm with which we can work and classify the datasets and table data well, in the end we conclude that we are going to use 1 main algorithm: the CART algorithm.

The algorithm CART (classification and regression trees), this algorithm uses the Gini impurity to able to learn based on decisions trees, this code generate binary decisions tree

* The classification trees predict objects, also predict continuous objects.



Steps to consider developing the operation of the decision tree:

* Organize the questions according to the Gini and its importance.
* How to calculate the Gini.
* Identify the types of data.
* Calculate the accuracy of the tree.
* Method for each person with their results.

With this we would start with the next part of the project, later we will see the possibility of adding more methods or creating different ones to improve the operation of the program and thus develop a better job.

**Matrices:**

In this project, first we created a program that read the data sets and the we proceeded to create the classification tree. To start with this, we must use the panda’s library o python, this library has access to the .CVS files, and with this the pandas transform the dates in objects, in data frames. And then here we can show what is the reason for which we decided to use an array, working with the panda’s library we found that is tiresome, so the matrices or arrays can order, search, delete, modified and insert what ever we want, so we work with the library pandas to covert the data frames in list, with the tolist() that pandas has, and with that it becomes much more manageable for us to be able to create the decision tree and make it work in the best way.

**Process:**

At the beginning we explored and make the pandas library like more normal to work and also to known the different functions, and we make the first table, that read the dates of the data sets and then organize the different types of signatures and periods and the other things that the ICFES exam has in dates.

Then in group we choose the algorithm that for us are the better one, also explain the function of the array that we choose, the algorithm CART, and we start the code to create the tree of decisions that can help us to determine the percentage of success of the students.

In the last part we complete the code to graph the tree and the different tables of the file “TEST.csv”, we saw that the algorithm that we did take time to analyze all of the dates but is optimal and also can give us the dates that we need. In the same array we create an option of success that can show us the certainty in the classes of the algorithm, “Exito” or “No Exito”, this one can show us the different points of the students where they haves success or not.

And finally, we can show in the tree of decisions that we did that the algorithm works:

Diagrama

Descripción generada automáticamente

**Conclusions:**

We can see that our tree can predict que results of what we were looking for, it can be useful to predict how well students can be in the future based on the grade or points that they obtain in the ICFES exam. Also, we saw that the CART algorithm was he best option for us to develop the project, since this showed us all of the alternatives that we can have to work in the decision’s trees in the future.

We need to take in a count that this tree only has an average a percentage that show the probability of a student will do very well in the test, but each student can make a difference.

To conclude, this project was based on the statistics of the ICFES exam that is the exam of the government in the last year of school, but we can apply this types of algorithms in other camps of work for develop or study something new, there are so important and useful for our lives as a programmers.