DECISION TREES TO PREDICT COLOMBIAN STUDENT'S SUCCESS ON SABER PRO BASED ON SABER 11 RESULTS

Juan Jose Escobar Bernal
Universidad EAFIT
Colombia
jjescobarb@eafit.edu.co

Simón Lopera Botero
Universidad EAFIT
Colombia
sloperab1@eafit.edu.co

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

ABSTRACT

The purpose of this report is to find a connection between the scores of Colombian bachelor students in Pruebas Saber 11 and Pruebas Saber Pro; as well as, socio-economic details regarding their personal and familiar lifestyle. Through this relationship, it will be possible to predict the results on the Pruebas Saber Pro resulting in scores above or below average.

Resulting predictions from the use of the decision tree can help solve the deficiency in education found in Colombia. It can help find a correlation between socio-economic conditions and results in Pruebas Saber Pro; allowing a hint and suggestions towards the most vulnerable stratum to destine more funds directly.

The solution for the problem was the implementation of a CART binary tree running on Gini Impurity comparison. It sorts the students through the tree according to information gained in the variable;finally, returning a True or False value regarding future success.

The final results had fast run times with all of the data sets;however, it also consumed large amounts of memory. The tree resulted in a 87% accuracy in predicting success compared to the actual results.

1.INTRODUCTION:

Education and technology are two areas that have been left behind in comparison with other countries in the region. Using decision trees to predict student results in the Pruebas Saber Pro, allows improvement in Colombian education. It helps distribute resources in the education area of society.

The information from each student that will be used includes the specific results on every area of the Pruebas Saber 11, information about their lifestyle, where they live, as well as parent occupations. The data includes more than 30 different factors that help the prediction on their success be as accurate as possible.

2.PROBLEM:

The problem consists in predicting if a student taking the Pruebas Saber Pro will have a score below or above the average using information from their lifestyle and their results on the Pruebas Saber 11 with the implementation of a decision tree.

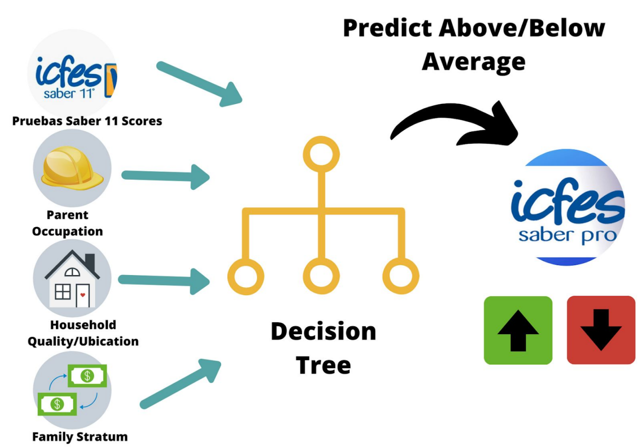


Figure 1. Visual representation of the problem [6][3]

3.POSSIBLE SOLUTIONS:

3.1. ID3

The iterative dichotomiser is a type of decision tree created by Ross Quinlan. It is considered to be one of the simplest but yet very efficient type of decision tree. It works by taking all of the attributes and sorting them by highest information gain. It tests the attribute with the highest information gain and then continues down a bunch of branches in a certain order until it gets to a final result. The algorithm works mostly with categorical attributes and is not the fastest since it has to check one attribute at a time, some others can check multiple attributes at the same time. [5]

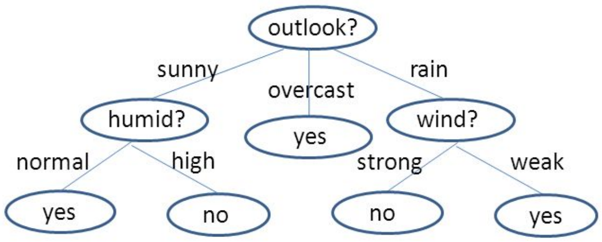


Figure 2: ID3 decision type tree [2]

3.2. C4.5

C4.5 is an algorithm that works as an extension to ID3 developed and created by Ross Quinlan. It is used for classification and works more effectively than ID3 because it works with continuous and discrete variables. It forms numerous reduced trees that are easier to understand and divides the problem multiple times. C4.5 is practical to use making its understanding easier while being able to use categorical and continuous values;however, it is not recommended for small sets.[5]

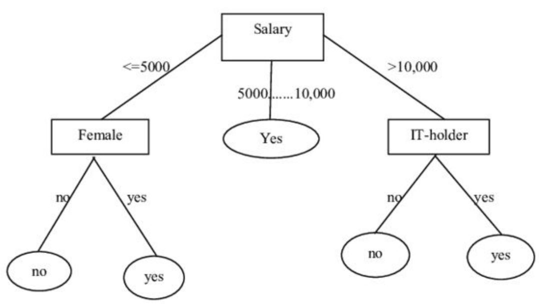


Figure 3: C4.5 decision type tree [1]

3.3. Random Forest

Random Forest is a type of decision tree created by Leo Breiman, consisting of numerous simple trees that return a response given certain predictor values declared at the start. It works for both classification and regression values and problems. This type of tree is one of the most accurate learning algorithms that gives an estimate of the most valuable variables classified. As a disadvantage, it is not easily interpreted by humans.[5]

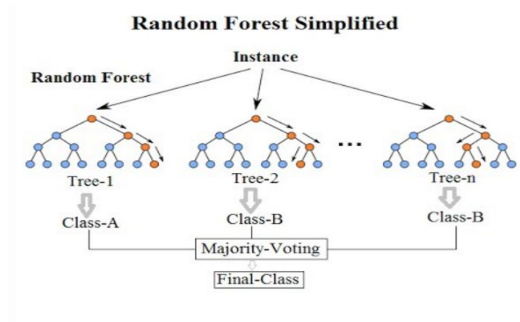


Figure 4: Random Forest decision type tree

```
To generate c classifiers:
    for i = 1 to c do
        Randomly sample the training data D with replacement to produce Di
        Create a root node,Ni containing Di
        Call BuildTree(Ni)
    end for

    BuildTree(N):
if N contains instances of only one class then
        return
    else
    Randomly select x% of the possible splitting features in N
    Select the feature F with the highest information gain to split on
    Create f child nodes of N,N1,NF, where F has f possible values(F1,Ff)
    for i = 1 to f do
        Set the content of Ni to Di,where Di is all intansces in N that match
        Fi
        Call BuildTree(Ni)
    end for
end if
```

Figure 5: Random Forest PseudoCode[7]

3.4. CART

CART, also known as Classification and Regression trees, is a type of decision tree that includes both classification and regression trees. One particular attribute about CART trees is that it only does binary separation, meaning every decision can only take two possible routes.[5]

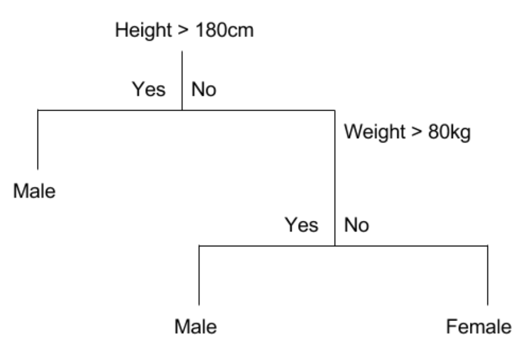


Figure 6: CART decision type tree [4]

DATA STRUCTURE DESIGN

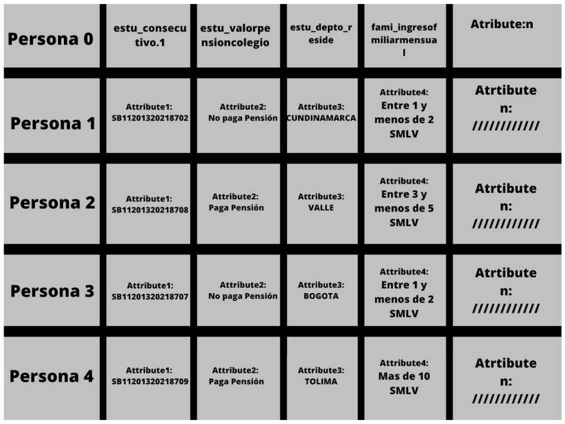


Figure 7. Visual representation of data structure. Each Persona contains an ID,Status,City,Salary,School and all the other ICFES questions.[6]

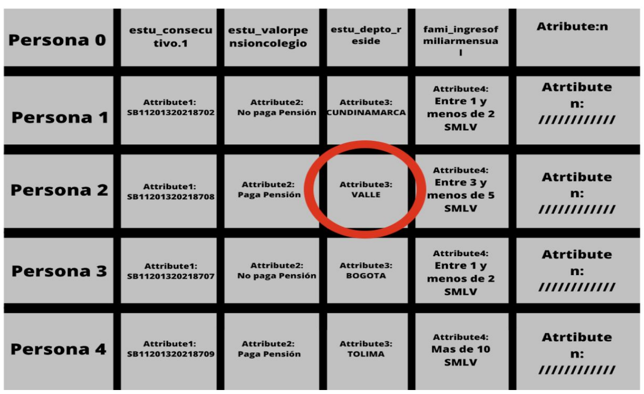


Figure 8. Visual Representation of Get Function on ArrayList[6]

DATA STRUCTURE COMPLEXITY

PERSONA 2D ARRAY FUNCTION	O COMPLEXITY
ADD	O(1)
GET	O(1)

DATA STRUCTURE DESIGN CRITERIA

In order to decide which data structure to use; complexity,memory usage and execution time were three factors taken into account in order to choose a structure suitable for the desired solution. Since ArrayList has a complexity of O(n) in some functions that were completely necessary for this project (add), this was not the ideal data structure for our executable time. LinkedList was also taken into consideration, but the problem that occurred with ArrayLists was also presented here.. Throughout the project it was discovered that the size needed before creating data structure was already given. This made it so it wasn't necessary to use ArrayList or LinkedList, Matrix/2D arrays could be used instead. This way everything can be done with a simple complexity of O(1) since the size is implemented from the start.

Average time in 100 repetitions for each of the datasets

Data	5000	35000	45000	135000
Average Time	100 ms	1532 ms	1579ms	2750ms

Figure 9. Chart showing Average time for data sets.

Memory Consumption for each of the datasets

Data	5000	35000	45000	135000
Memory Consumption	251.2 MB	490MB	490MB	1025MB

Figure 10. Chart showing Memory Consumption for data sets.

Data for time and memory used

Data	Best time	Worst time	Best memory	Worst memory
5000	93 ms	113 ms	250 mb	254 mb
35000	1492 ms	2204 ms	487 mb	490 mb
45000	1493 ms	1555 ms	488 mb	491 mb
135000	2701 ms	2771 ms	1022mb	1028mb

Figure 11. Chart showing best time and worst time, best memory and worst memory.

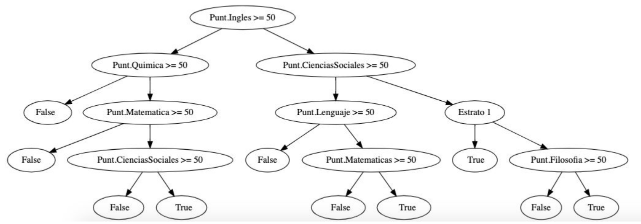


Figure 12. Binary tree representation.

Acknowledgments

This research was supported by Universidad EAFIT.

We thank Universidad EAFIT for assistance with equipment and resources necessary for the project.

Conclusion

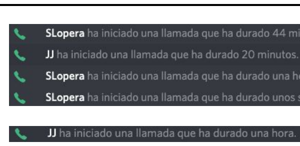
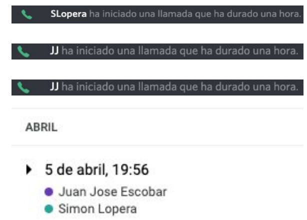
In this paper, a binary decision tree was developed in order to predict whether a student is going to be above average in their Saber Pro tests. The problem was established and analyzed through all different types of trees and solutions that could be used.CART tree was chosen as the optimal implementation and included the results of the project including the time and memory consumed.

After using a 2D array and a CART decision tree there were great results. An accurate tree was built that was able to predict 90% of the students' results. The time consumed was really low and that made the code really efficient. The only downside of the results was the memory consumption which was a bit high.

The first solution used an ArrayList that stored an object of type Person in each space. The final solution was much better since it managed to make the complexity of filling our data structure O(1) instead of O(n) by changing to a 2D array. This made the time consumption much more efficient.As a whole the creation of the tree was O(2^n).

After finishing this project there are a bunch of different things that could be implemented. One of the many ideas that could be very interesting is predicting whether a student will be successful in their Icfes using the grades they get at school.

Teamwork

First Deadline	
	
Second Deadline	
Three Deadline	

REFERENCES

[1] Anon. 2011. Decision Trees – C4.5. (November 2011). Retrieved February 9, 2020 from https://octaviansima.wordpress.com/2011/03/25/decision-trees-c4-5/

[2] Anon.Retrieved February 8, 2020 from https://www.philippe-fournier-viger.com/spmf/ID3.php

[3] Anon. Inscripciones abiertas para pruebas Saber 11 calendario A, Pre Saber y validación del bachillerato académico. Retrieved February 6, 2020 from https://www.valledelcauca.gov.co/publicaciones/62908/inscripciones-abiertas-para-pruebas-saber-11-calendario-a-pre-saber-y-validacion-del-bachillerato-academico/

[4] Jason Brownlee. 2019. Classification And Regression Trees for Machine Learning. (August 2019). Retrieved February 7, 2020 from https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/..

[5] Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora and Naresh Dhami. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications* 163(8):15-19, April 2017.

[6] Anon. Canva - Collaborate & Create Amazing Graphic Design for Free. Retrieved February 9, 2020 from https://www.canva.com/

[7] Luluah Alhusain and Alaaeldin M. Hafez. 2017. Cluster ensemble based on Random Forests for genetic data. (December 2017). Retrieved February 10, 2020 from https://link.springer.com/article/10.1186/s13040-017-0156-2

[8] AdarshAdarsh 6111 silver badge66 bronze badges, OldCurmudgeonOldCurmudgeon 57.9k1414 gold badges9898 silver badges184184 bronze badges, Zbynek Vyskovsky - kvr000Zbynek Vyskovsky - kvr000 15.6k22 gold badges2424 silver badges3737 bronze badges, and ControlAltDelControlAltDel 25.1k55 gold badges3838 silver badges6565 bronze badges. 1965. High Memory consumption for ArrayList object. (October 1965). Retrieved April 3, 2020 from https://stackoverflow.com/questions/34338695/high-memory-consumption-for-arraylist-object