PREDICTING THE ACADEMIC FUTURE OF COLOMBIAN STUDENTS USING DECISION TREES

Juan Pablo Cortés Gonzalez
Universidad Eafit
Colombia
jpcortesg@eafit.edu.co

Julián Rojas Gallego
Universidad Eafit
Colombia
jrojasg1@eafit.edu.co

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

ABSTRACT
The objective of this project is predict the success academic in the saber Pro test, defined success academic like score for over of cohort score. Previous research used decision trees with ID3 and C4.5 algorithm and concepts of mining data, We will implement decision trees with CART (Classification and Regression Trees) algorithm, starting from the Environment Variables and results than have every student in Saber 11 test, and finally we will get the students to have academic success so that we can see the variables that most affect a student's academic success, in this way pay attention to help them achieve better academic performance .

Keywords
Data structures; prediction; decision tree; Complexity; Big O;Academic success; Data; Algorithms; Machine Learning.

ACM CLASSIFICATION Keywords

Computing methodologies → Machine learning

Applied computing → Physical sciences and engineering → Mathematics and statistics

Theory of computation → Design and analysis of algorithms → Data structures design and analysis

Theory of computation → Computational complexity and cryptography → Oracles and decision trees

1. INTRODUCTION
Pruebas Saber 11º, also known as the *ICFES* test is an evaluation conducted by the ministry of national education in Colombia. The aim of this evaluation is to measure the skills and knowledge that students acquired through their school years before moving on to superior studies. The subjects that are tested are critical reading, math, social and citizen skills, natural sciences and English. The results of the test fall within a range of 1 to 500, the latter being a perfect score. This result can predict the academic success of a student, and their result in the *Saber Pro*, the equivalent test but after completing superior studies. The result is not the only variable to consider, and thanks to the public availability of the data of people who present these tests, a more educated approach to a guess can be made.

2. PROBLEM
There is a lot of rich data inside every *Pruebas Saber 11º* ever taken. A lot of demographical and social indicators can be found within every test taken, alongside the results. The reason we want to try to predict the successful outcome of a student in the posterior tests, is because we want to know what living conditions or characteristics of the test taker can deeply influence this result. By successful we mean the student gets a score above the mean of his graduating class. If a clear influence can be detected, it can be reported back to the community as means of a pointer into what the Colombian government can tackle to improve the quality of education of its citizens.

It is concerning that Colombia has such low results in the Program for International Student Assessment (PISA), ranking in the lower 30% of the countries in all categories. Worst of all, the results have been dropping since Colombia started taking part in this assessment, in a period where the Colombian population has been fragmented and taking steps back in the development of capable, educated citizens. It is important that a clear and data-backed solution can be presented, to stop arbitrary decisions from taking the country in the wrong direction.

3. RELATED WORK
The field of machine learning has a wide approach, depending on your goal the algorithms you choose will vary a lot. From simple linear regressions, to more complex vector support machines, or even deep learning algorithms which transform data into large neural networks which are intertwined in different layers with layers which allow a great variety of data in parallel. These last examples are quite complex for the task at hand, so we are going to develop a decision tree algorithm. These algorithms are easy to interpret, are quick to run, and give out solid information.

3.1 ID3
ID3 is an algorithm used for Decision Tree, the principal elements it contains the algorithm ID3 are:

- Searching algorithms (greedy algorithm, heuristic search, hill climbing, alpha-beta pruning)
- Logic (OR, AND rules)
- Probability (Dependent and Independent)
- Information Theory (Entropy)

It is precursor to the C4.5 algorithm, was invented by Ross Quinlan and the process used for made the Decision Tree is:

- Take all unused attributes and calculates their entropies.
- Chooses attribute that has the lowest entropy is minimum or when information gain is maximum
- Makes a node containing that attribute

pseudo code of ID3 is as follows:

```
# x is examples in training set
# y is set of attributes
# labels is labeled data
# Node is a class which has properties values, childs, and next
# root is top node in the decision tree

# Declare:
x = # Multi dimensional arrays
y = # Column names of x
labels = # Classification values, for example {0, 1, 0, 1}
         # correspond that row 1 is false, row 2 is true, and so on
root = ID3(x, y, label, root)

# Define:
ID3(x, y, label, node)
  initialize node as a new node instance
  if all rows in x only have single classification c, then:
    insert label c into node
    return node
  if x is empty, then:
    insert dominant label in x into node
    return node
  bestAttr is an attribute with maximum information gain in x
  insert attribute bestAttr into node
  for vi in values of bestAttr:
    // For example, Outlook has three values: Sunny, Overcast, and
Rain
    insert value vi as branch of node
    create viRows with rows that only contains value vi
    if viRows is empty, then:
      this node branch ended by a leaf with value is dominant label
in x
    else:
      newY = list of attributes y with bestAttr removed
      nextNode = next node connected by this branch
      nextNode = ID3(viRows, newY, label, nextNode)
  return node
```

Applications:

- Operational Research
- Finance
- Scheduling problems

3.2 Information Gain

Information gain is a statistical property that measures how well a given attribute separates the training examples according to their target classification.

To define information gain, we need to first define a measure commonly used in information theory called entropy that measures the level of impurity in a group of examples. Mathematically, it is defined as:

$$Entropy: \sum_{i=1} -p * log_2(p_i)$$
$$p_i = Probabilityofclassi$$

3.2 CART

Classification models are used when the target values have discrete nature, and when the values are of continuous values, usually numbers you use Regression models. Utilizing these two models together create the CART decision tree algorithms which use GINI index, instead of information gain or other methods to select attributes for creating the tree.

$$\text{Gini} = 1 - \sum_i p(i|t)^2$$

3.3 Hunt's Algorithm

Hunt's algorithm grows the decision tree in a recursive way, by partitioning the training set into purer subsets.

.

4. Title of the first data structure designed

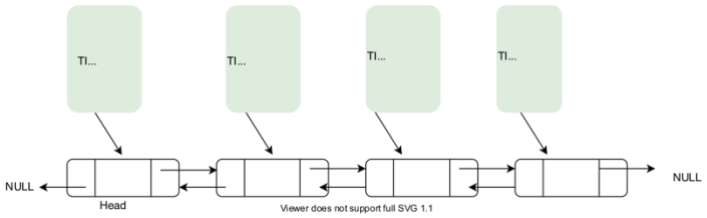

Figure 1: Doubly Linked List of students, one student is a class with your attributes(columns of file .csv),the linked list always insert at the end.

4.1 Operations of the data structure

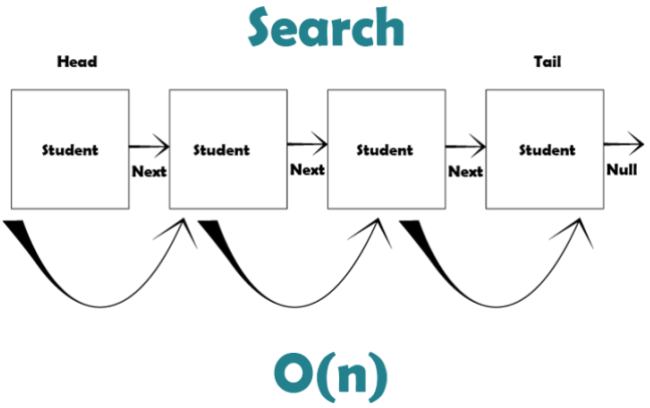

Figure 2: the figure shows the Search a data and your complexity.

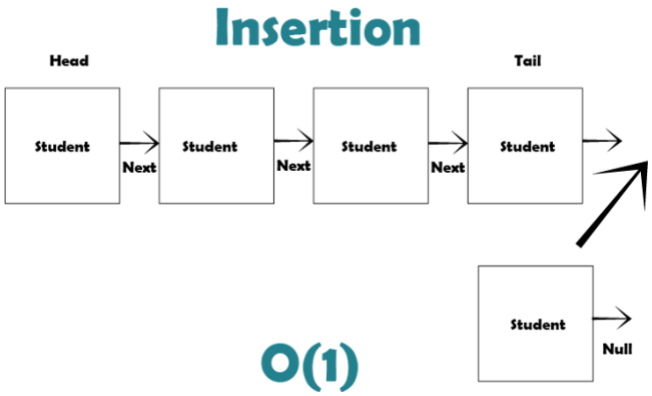

Figure 3: the figure shows the insertion a data and your complexity.

4.2 Design criteria of the data structure

We choose this data structure because facilities the loading of data since complexity in this data structure is O(1), which provides good performance when loading data. Additionally, in comparation whit others data structures , this a good shape for represent a tree, since the nodes of linked list, represent the sheets at the tree. This data structure is simple and effectivity for what it want to achieve, because there a lot of data.

4.3 Complexity analysis

The next table shows the operations implemented.

Operation	Complexity
Search	**O(n * m)**
Insertion	**O(1)**

Table 1: The table shows the operations implemented.

4.4 Result analysis

Table 2: Results of used resource of system in time and memory.

No. Rows	Time	Memory used
15000	11 seg	221MB
45000	18 seg	378MB
75000	30 seg	403MB
105000	60 seg	597MB
135000	80 seg	644MB
57765	25 seg	390MB

5. Title of the last data structure designed

We implemented a new data structure to improve response times.

Figure 4: Show the represent the new data structure implemented.

5.1 Operations of the data structure

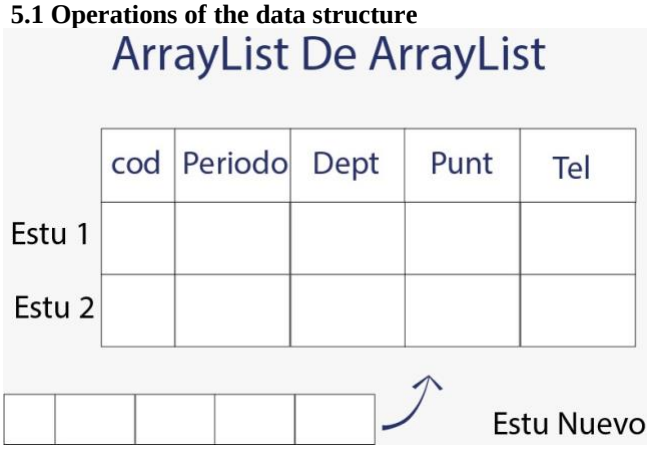

Figure 5: the image shows the operation of insertion.

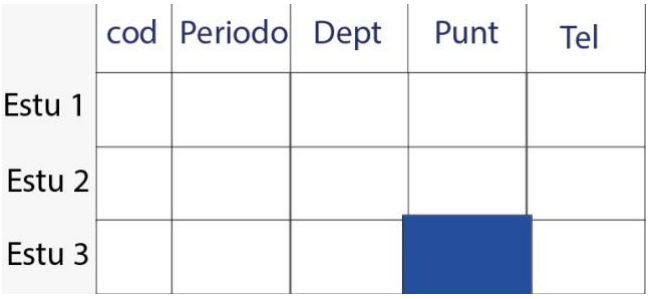

Figure 6: the image shows the operation of access to one position of the matrix.

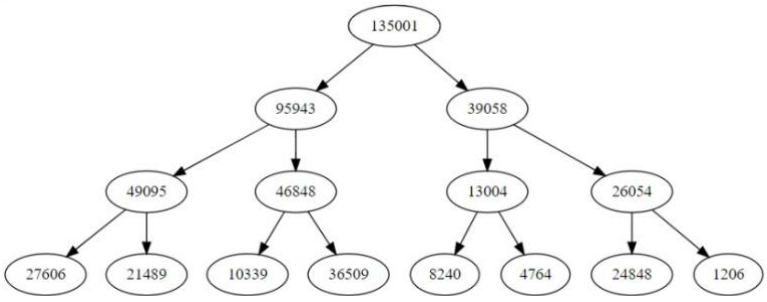

Figure 7: The image shows the decision tree, where each node is marked with the data number of the data set.

5.2 Design criteria of the data structure

We changed the data structure to ArrayList of ArrayList because we said that access is important than insert, then we changed to ArrayList of ArrayList (which is also equal to an array). Since when starting it, the complexity to access a position is O (1) and improves execution times.

5.3 Complexity analysis

Operation	Complexity
Search	O(n * m)
Access	O(1)
Insertion	O(n*m)

Operation	Complexity
dividir	O(n^2)
Calculary gini	O(n^2)
Buscar valores de un varible	O(n)
Crear nodo	T(c) = 2^c + c^2

Tabla 3: show the complexity of the operations implemented.

5.4 Execution time

DataSet	load dataset(insertion)	Access
15000	4	1
45000	9	1
75000	15	1
105000	25	3
135000	54	3
57765	12	1

Table 4: the table show the time for every operation in seconds.

5.5 Memory used

DataSet	Memory(MB)
15000	186
45000	387
105000	645
135000	790
57765	280

Table 5: Memory used for load every dataset.**5.6 Result analysis**

operation	Best time	Worst time	average
Load dataset	4 sec	125 sec	62 sec
Acces position	1 sec	3sec	4 sec

Tabla 6: time comparison, from best to worst time.

6. CONCLUSIONS

The prediction of the academic success of a group of students can be done with a simple and easy algorithm to implement, on the other hand it is of great help for psychosocial studies that are done to assess the influence of the family context and social territory.

Second, the sequence of variables to create the decision tree vary according to the amount of data examined, on the other hand, the best data structure for load the data is a matrix or Arraylist of Arraylist, given an initial value for rows and columns of the matrix.

6.1 Future work

Finally we plan is make a random forest and in this way avoid data overfitting and have a better prediction of academic success.

ACKNOWLEDGEMENTS

We want to thank Sapiencia for providing us with all the resources and tools that were necessary to carry out the project development process.

Finally, we want to thank all our colleagues and our family for supporting me even when my spirits fell. In particular, I want to mention our parents, who were always there to give us words of support to recharge.

REFERENCES

1.Aprendemachinelearning. 2017. Principales Algoritmos usados en Machine Learning. Retrieved november 4, 2017 from https://www.aprendemachinelearning.com/principales-algoritmos-usados-en-machine-learning/

2. Abbas Rizvi.2010. ID3 Algorithm. Retrieved Spring 2010 from http://athena.ecs.csus.edu/~mei/177/ID3_Algorithm.pdf

3. Hafidz Jazuli.2018. An Introduction to Decision Tree Learning: ID3 Algorithm. Retrieved march 18,2018 from

https://medium.com/machine-learning-guy/an-introduction-to-decision-tree-learning-id3-algorithm-54c74eb2ad55

4.Brownlee, J. How To Implement The Decision Tree Algorithm From Scratch In Python. Machine Learning Mastery, 2020. https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/.