PREDICTING THE ACADEMIC FUTURE OF COLOMBIAN STUDENTS USING DECISION TREES

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

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

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.

.

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