PREDICTION OF COLOMBIAN STUDENTS' ACADEMIC SUCCESS USING DECISION TREES ALGORITHMS

Juliana Restrepo	Andrea Carvajal	David Vergara	Mauricio Toro
Universidad EAFIT	Universidad EAFIT	Universidad EAFIT	Universidad EAFIT
Colombia	Colombia	Colombia	Colombia
jrestrepot@eafit.edu.co	acarvajalm@eafit.edu.co	dvergarap@eafit.edu.co	mtorobe@eafit.edu.co

ABSTRACT

The paper uses a decision trees algorithm to predict the performance of Colombians' in the Saber Pro national standardized test based on data from their Saber 11. All this to identify which variables are influential in students' academic success and conclude about Colombia's education system.

Keywords

Decision tree, Saber Pro, machine learning, CART, C5, DataFrame, academic success, prediction algorithm.

ACM CLASSIFICATION Keywords

Theory of computation ->Design and analysis of algorithms-> machine learning theory.

1. INTRODUCTION

Colombia's Ministry of Education has established several standardized tests that are taken by students from different age ranges throughout the whole country. During his/her academic life, a Colombian student must undergo three different standardized exams: "Pruebas Saber" performed by students in fifth and ninth grade, "Prueba saber 11" also known as "ICFES" carried out by students in eleventh grade or their last year of high school, and finally "ECAES" or "Prueba Saber Pro" which focuses in the field of study of university students who are about to finish their professional career. These exams allow the government to easily identify strengths, weaknesses, and progress within the education system.[1]

A decision tree is a technique used in programming to separate observations in a specific problem. It classifies each element of the observation in different branches depending on the attribute in question. The process of classifying is repeated with the remaining attributes using the algorithm of recursion. This technique constructs a tree made up of nodes and branches. Nodes are the points in the diagram where the pathways branch and are subdivided into three different types: the root node, the internal nodes, and the leaf nodes. The root node is the initial node placed on the top of the diagram and is the origin of all ramifications. The internal nodes are all the nodes inside the diagram, that is to say, all the nodes that come from the root node and lead to other nodes. The leaf nodes are the nodes at the end of the diagram, these are the ones that output the prediction of the results. [2]

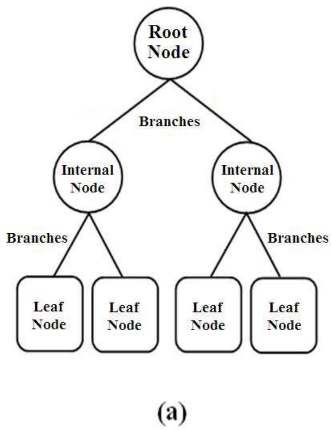


Figure 1.

[9]

Each node has a variable that is compared to the dataset using a separation algorithm. The number of branches that fall of each node depends on the separation algorithm used and the variables selected. [2]

The decision tree method was chosen over other techniques such as neural networks because it is more practical, agile in terms of processing time, and has a higher accuracy predicting results for academic success. [4]

2. PROBLEM

The main purpose of this project is to predict whether a student will be successful in the exam "Saber Pro" or not. For the development of the project, academic success was defined as obtaining a score that is above the average score of the student's respective cohort. To get a prediction that resembles reality, data from the exam "Prueba saber 11" has been collected providing information regarding the individuals' social, economic and academic context, in addition to the score obtained in this previous exam.

3. RELATED WORK

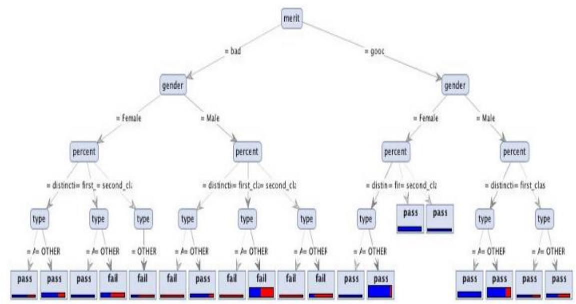
3.1. ID3

ID3 is an algorithm invented by Ross Quinlan. It produces a decision through the construction of a decision tree that classifies each tuple in the database. ID3 is commonly applied in problems related to machine learning and processing of natural language. [3]

Firstly, the algorithm calculates the entropy of every attribute of the dataset. Then, it splits the set into subsets using the attribute with maximum information gain (minimal entropy). Next, it makes a decision tree node containing that attribute. Finally, it recurses on subsets using the remaining attributes. [8]

The study conducted by Adhatrao, Gaykar, Dhawan, Jha, and Honrao (2013) [3], whose objective was to predict the general and singular performance in future academics of enrolled students, developed prediction systems using decision trees.

The researchers used a training dataset consisting of information about the admitted students' background and percentage marks obtained in their board examinations. The relevant information was fed into the database and then ID3 was applied to the training data, resulting in the following decision tree.



[3]

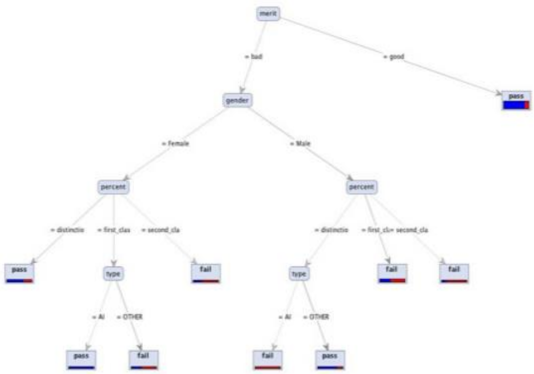
Figure 2.

3.2. C4.5

It is an improved version of the ID3 algorithm. This statistical classifier handles training data with missing values and attributes with discrete and continuous values. Furthermore, it prunes unnecessary branches of the decision tree after its construction. [3]

C4.5 selects the attribute that splits the data in the most effective way to classify each sample into one class or the other. Next, the algorithm makes the decision taking into account the attribute with the highest information gain, and then repeats this process in a recursive way. Finally, it prunes the tree. [5]

Adhatrao, Gaykar, Dhawan, Jha, and Honrao (2013) [3] also used this algorithm to solve the problem mentioned above. The same training data was used, but this time it was applied to C4.5. This new tree had fewer decision nodes. However, both algorithms had the same effectiveness (75.145%).



[3]

Figure 3.

3.3. C5

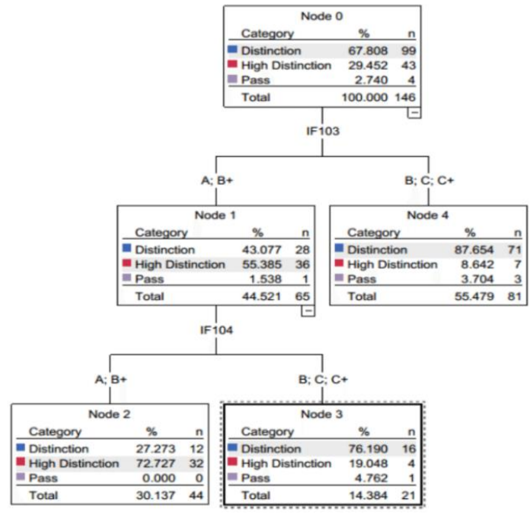
C5 is an extension of C4.5. This updated version produces tree models that are easily transformed into mathematical expressions and less biased, more understandable outputs, making it one of the best options for the solution of machine learning related problems. [6]

Sena and Ucarb (2012) conducted a study that used data from 3047 records taken by Karabük University Computer Engineering Department. The dataset included students' background information and whether the students studying in distance education or regular education. C5 algorithm and neural networks were used to compare the achievements of Computer Engineering Department students in Karabük University based on the given data. The decision tree was the most accurate algorithm, with 97.8107% of precision on 10 fold holdout dataset. [4]

3.4. CART

CART stands for Classification and Regression Trees. It is a non-parametric statistical method for multivariable data that uses binomial splits to correctly classify members of the population. Each variable is analyzed and split for sensitivity and specificity in the classification. Then, the resulting tree is pruned to minimize its size and inaccuracy rate. [6]

Kasih, Ayub, and Susanto (2013) [7] did a work aimed to help the academic advisors by predicting alumni's final results. The study used academic transcripts from 146 students as input data and used the CART algorithm, which produced the following tree.



[7]

Figure 4.

4. Used Data Structure

DataFrame:

	Key	Key Value		
0	**Student´s code**	**Period**	**Studied abroad?**	**…**
1	**SB11201220492225**	**20152**	**YES**	**…**
2	**SB11201220492224**	**20131**	**NO**	**…**

3	**SB11201220492226**	**20151**	**NO**	**…**

Figure 5: DataFrame

A data frame is a two-dimensional data structure with a mutable size and capacity to store different types of data. It can be compared to the hash table since the first position in a row acts like a key, and the rest of the row acts like the value associated with it.

GitHub Link:
https://github.com/jrestrepot/ST0245-032/blob/master/proyecto/codigo/proyecto_final_datos1.py

4.1 Operations of Data Structure

The only two operations with data frames that will be used in this project are the following:

Access:

Accessing a data frame's position has a time complexity of O(1) because the structure knows its location in the memory.

For instance:

data.iloc[0][0]

data.loc[“SB11201220492226”][0]

SB11201220492226	**20152**	**NO**
SB11201220492226	**20161**	**YES**

Figure 6: Accessing a data frame

Add:

In this case, concatenating two data frames has a time complexity of O(1) since it always adds the new data frame at the end.

For instance:

data=pandas.concat([dataframe1, dataframe2)

SB11201220492226	20152	NO
SB11201220492226	20161	YES

Figure 7: Data frame 1

SB11201220492225	20152	YES
SB11201220492224	20131	NO

Figure 8: Data frame 2

SB11201220492226	20152	NO
SB11201220492226	20161	YES
SB11201220492225	20152	YES
SB11201220492224	20131	NO

Figure 9: Concatenating dataframe1 and dataframe2.

4.2 Design criteria of the data structure.

Since the format of the input data is a matrix, it is very convenient to store it in a table-like structure. Besides, this data structure allows storing different data types, which is an advantage given that the CSV has data in forms of String, Integer, among others. Additionally, the data frame is one of the most used data structures when working with big data and AI due to its efficient and easy way of organizing and managing a large volume of data.

4.3 Complexity analysis:

	Complexity	
Method	**Best case**	**Worst case**
iloc/loc	O(1)	O(1)
pandas.concat	O(1)	O(1)

Figure 10: Complexity table

4.4 Execution time:

	Creation	Access	Concat
Train set 0	0.63s	0.000271s	(All data frames were concatenated together) 0.77s
Train set 1	1.27s	0.000242s	
Train set 2	1.99s	0.000238s	
Train set 3	2.48s	0.000238s	
Train set 4	3.16s	0.000237s	
Train set 5	1.49s	0.000237s	

Figure 11: Execution time of main operations

4.5 Memory used

	Memory consumption
Train set 0	59869502 bytes
Train set 1	179671900 bytes
Train set 2	299473951 bytes
Train set 3	419297982 bytes
Train set 4	539116653 bytes
Train set 5	229654982 bytes

Figure 12: Memory consumption.

4.6 Results analysis:

The time and space occupied depend on how large the volume of data is.

Structure	**Data Frame**
Space	179671900 bytes - 59869502 bytes

Time of creation	0.41s-4.43s
Time of access	0.000216s-0.000648s
Time of concatenation	0.75s-0.80s

Figure 13: Table of execution values.

In the worst case, creating a DataFrame will take 4.43s, and in the best case, it will take 0.41s. This difference is due to the size of the DataFrame created and other programs that the computer may be running at the same time. The time it takes to concatenate or access a DataFrame is extremely short, which is beneficial for the development of the project because these two methods are the most frequently used. A DataFrame will occupy between 179671900 and 59869502 bytes of memory.

REFERENCES

[1] Anon. 2006. Las distintas Pruebas. (2006). Retrieved February 6, 2020 from https://www.mineducacion.gov.co/1621/article-107522.html

[2] Al-Radaideh, Qasem & Al-Shawakfa, Emad & Al-Najjar, Mustafa. (2006). Mining Student Data Using Decision Trees. The International Arab Journal of Information Technology - IAJIT. https://www.researchgate.net/profile/Qasem_Al-Radaideh/publication/242076740_Mining_Student_Data_Using_Decision_Trees/links/53e279ab0cf275a5fdd876c8/Mining-Student-Data-Using-Decision-Trees.pdf

[3] Anon. 2013. P STUDENTS PERFORMANCE USING ID3 C4.5 CLASSIFICATION ... (September 2013). Retrieved February 8, 2020 from https://arxiv.org/pdf/1310.2071

[4] Baha Sen and Emine Ucar. 2012. Evaluating the achievements of computer engineering department of distance education students with data mining methods. (May 2012). Retrieved February 9, 2020 from https://www.sciencedirect.com/science/article/pii/S2212017312000540

[5] Anon. 2020. C4.5. (February 2020). Retrieved February 9, 2020 from https://es.wikipedia.org/wiki/C4.5#Algoritmo

[6] Anon. 2012. Comparison of C5.0 & CART Classification algorithms using ... (June 2012). Retrieved February 8, 2020 from https://www.ijert.org/research/comparison-of-c5.0-cart-classification-algorithms-using-pruning-technique-IJERTV1IS4104.pdf

[7]Kasih, Julianti & Ayub, Mewati & Susanto, Sani. (2013). Predicting students' final passing results using the Classification and Regression Trees (CART) algorithm. World Transactions on Engineering and Technology Education Vol.11, No.1, 2013. 11. https://www.researchgate.net/publication/275950590_Predicting_students'_final_passing_results_using_the_Classification_and_Regression_Trees_CART_algorithm

[8]Anon. 2020. ID3 algorithm. (February 2020). Retrieved February 9, 2020 from https://en.wikipedia.org/wiki/ID3_algorithm#Algorithm

[9]Silva da Sá, Almeida, Pereira da Rocha & Santos Mota. 2016. Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy. (Mar. 2016), 1-5. DOI: 10.21528/CBIC2011-27.1