DECISION TREES IN ACADEMIC SUCCESS IN SABER PRO

Santiago Ochoa Universidad Eafit Colombia sochoac1@eafit.edu.co	Miguel Angel Zapata Universidad Eafit Colombia mazapataj@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

ABSTRACT

Considering new strategies to predict academic success during Saber 11 would have beneficial effects in search of improving education system in Colombia. Unfortunately, the technological advances with these purposes have not gone far away in the last years. However, nowadays there are a vast variety of tools and widgets able to do estimations and predictions over different variables. In this case, decision trees will be functional to compare sociodemographic and academic variables related to identify success average on students during Saber 11.

Keywords

Decision trees, machine learning, academic success, standardized student scores, test-score prediction

1. INTRODUCTION

Colombia and technology haven´t had a close relation in terms of education. Development of new widgets, based on data study, would improve the way to estimate possible predictions to find out the main reasons of student dropout. Students during the last course of their bachelor's degree are assessed to qualify their knowledge with a test known as Saber Pro. There are many causes involve in having a score above the required average. Through a specific study about this causes, it is possible to predict academic success in bachelor's degrees in the country.

1.1. Problem

The problem to deal with is to implement an algorithm based in a decision tree adjust to different variables defined in a data structure. According to this, variables are related with sociodemographic and academic information such as: age, parents' income, career, Saber 11 previous scores, gender, among others. Besides, for each student there is a variable which consist in their average score in relation with the total score in Saber Pro.

In this way, the objective to achieve is designing different decision trees concluding prediction statements that define academic success as the probability of a student of having a score above their own previous average.

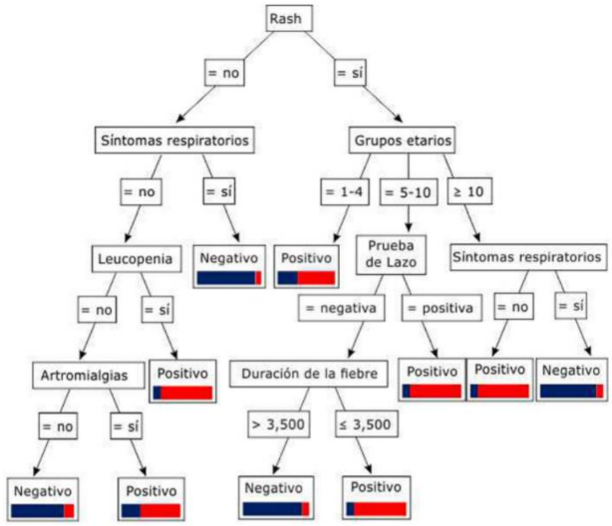
1.3 Article structure

In what follows, in Section 2, we present related work to the problem. Later, in Section 3 we present the datasets and methods used in this research. In Section 4, we present the algorithm design. After, in Section 5, we present the results. Finally, in Section 6, we discuss the results and we propose some future work directions.

2. RELATED WORK

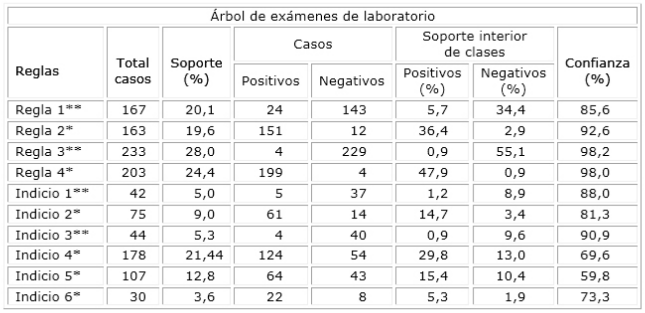
2.1 Decision trees in the dengue diagnosis

Basically, the main purpose was to identify symptoms and signs around a group of patients and concluded they vulnerability to suffer dengue according to decision trees. The study was divided in two kind of variables: 25 numerical and 21 categorical. The accuracy average rounded 80%. Processes were carried out by the software RapidMiner.



Nodes represented a specific rule made up by the antecedent (branches) and the consequent (leaves). Each rule calculated two indicators. The first one meant the percentage of how many times antecedents and consequent were together, and the second pointed out the conditional probability to accomplish a rule.

Tabla 1. Indicadores que apoyan las reglas del *Árbol de síntomas y signos*



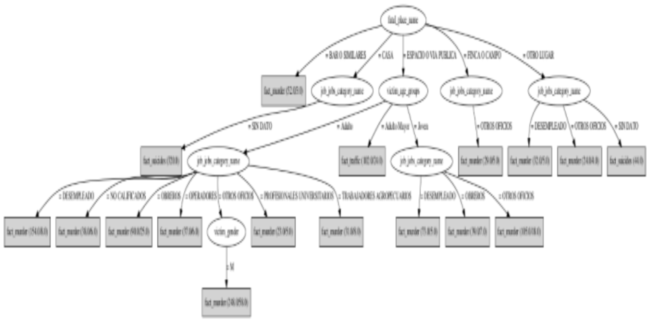
Árbol de exámenes de laboratorio							
Reglas	Total casos	Soporte (%)	Casos		Soporte interior de clases		Confianza (%)
			Positivos	Negativos	Positivos (%)	Negativos (%)	
Regla 1**	167	20,1	24	143	5,7	34,4	85,6
Regla 2*	163	19,6	151	12	36,4	2,9	92,6
Regla 3**	233	28,0	4	229	0,9	55,1	98,2
Regla 4*	203	24,4	199	4	47,9	0,9	98,0
Indicio 1**	42	5,0	5	37	1,2	8,9	88,0
Indicio 2*	75	9,0	61	14	14,7	3,4	81,3
Indicio 3**	44	5,3	4	40	0,9	9,6	90,9
Indicio 4*	178	21,44	124	54	29,8	13,0	69,6
Indicio 5*	107	12,8	64	43	15,4	10,4	59,8
Indicio 6*	30	3,6	22	8	5,3	1,9	73,3

Regla/indicio positivo*; regla/indicio negativo**.

Finally, results were evaluated by a confusion matrix to show the algorithm performance. Besides, a diagnostic scale justified the algorithm efficiency.

2.2 Decision trees in violence prevention

In summary, different death patterns were discovered by using decision trees. Consequently, the study used the Cross-industry standard process for data mining (CRISP-DM) dividing the problem in six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. A big amount of information was provided by the Colombian Observatory of Organized Crime and by taking advantage of data mining different decision trees were designed. Decisions trees were implemented using an easy algorithm known as J48 by WEKA data mining tool.

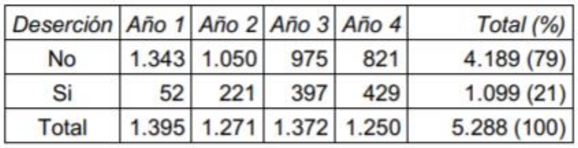


Keeping in mind CRISP-DM phases, the study concluded with different estimations for each death pattern between 2003 until 2013.

2.3 Decision tree to predict academic dropout in a Chilean university

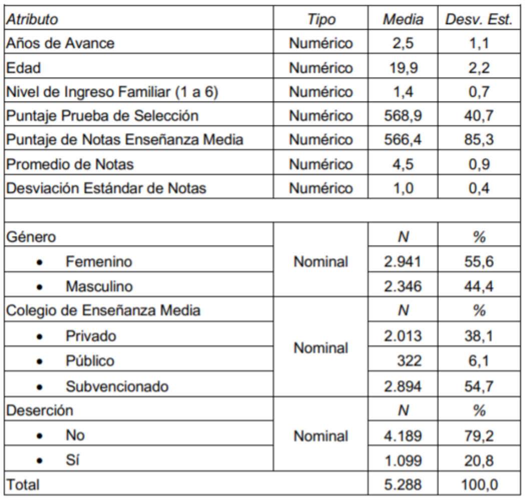
The main objective was to give a deep vision of some reasons that cause students dropout. The university could identify the trigger in the dropout using a decision tree. The study implements the optimization (A method to introduce some parameters with a range and a specific result) to get a certain percentage and to get the maximum depth of the tree. In the process they used three tables: The first one describes the academic dropout, the second one describes the attributes used to the analysis and finally the third one shows the parameters used to express the optimization.

Tabla 1: Deserción académica de la muestra



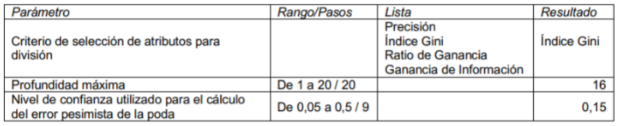
Deserción	*Año 1*	*Año 2*	*Año 3*	*Año 4*	*Total (%)*
No	1.343	1.050	975	821	4.189 (79)
Si	52	221	397	429	1.099 (21)
Total	1.395	1.271	1.372	1.250	5.288 (100)

Tabla 2: Atributos para el análisis de la CBAD



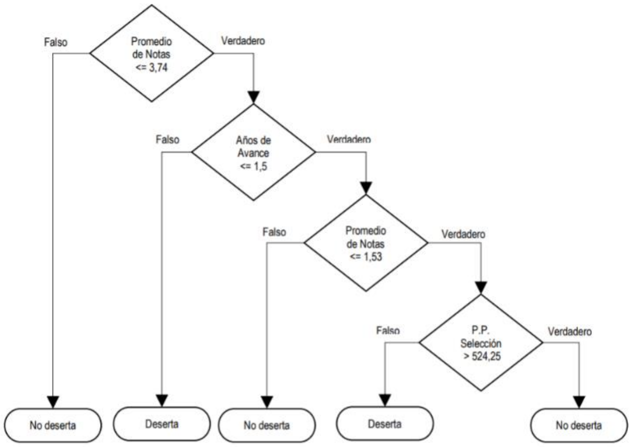
Atributo	*Tipo*	*Media*	*Desv. Est.*
Años de Avance	Numérico	2,5	1,1
Edad	Numérico	19,9	2,2
Nivel de Ingreso Familiar (1 a 6)	Numérico	1,4	0,7
Puntaje Prueba de Selección	Numérico	568,9	40,7
Puntaje de Notas Enseñanza Media	Numérico	566,4	85,3
Promedio de Notas	Numérico	4,5	0,9
Desviación Estándar de Notas	Numérico	1,0	0,4
Género	Nominal	*N*	*%*
• Femenino		2.941	55,6
• Masculino		2.346	44,4
Colegio de Enseñanza Media	Nominal	*N*	*%*
• Privado		2.013	38,1
• Público		322	6,1
• Subvencionado		2.894	54,7
Deserción	Nominal	*N*	*%*
• No		4.189	79,2
• Sí		1.099	20,8
Total		5.288	100,0

Tabla 3: Parámetros optimizados



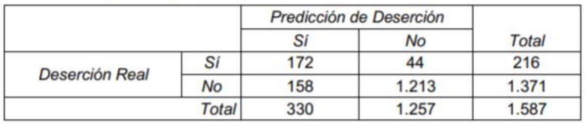
Parámetro	*Rango/Pasos*	*Lista*	*Resultado*
Criterio de selección de atributos para división		Precisión Índice Gini Ratio de Ganancia Ganancia de Información	Índice Gini
Profundidad máxima	De 1 a 20 / 20		16
Nivel de confianza utilizado para el cálculo del error pesimista de la poda	De 0,05 a 0,5 / 9		0,15

The decision tree implemented was based in three factors: Student average, years coursing bachelor's degree and score in the selection test.



Finally, results were evaluated by a confusion matrix to show the performance and the accuracy average rounded the 87.27%.

Tabla 4: Matriz de confusión para la predicción de deserción

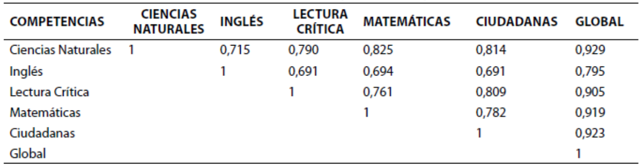


		Predicción de Deserción		
		Sí	*No*	*Total*
Deserción Real	*Sí*	172	44	216
	No	158	1.213	1.371
	Total	330	1.257	1.587

2.4. Decision tree to predict Pruebas saber 11

In search to find factors which have a predictive value associated with academic performance across Pruebas Saber 11, different algorithms based on decision trees were carried out. In this way, to proceed along the study was required the CRISP-DM methodology. Furthermore, variables were related with socio-economic, academic, and institutional information. Decision trees were developed using the WEKA data mining tool to identify patterns associated with academic failure or success.

Tabla 4. Matriz de correlaciones de las competencias de las pruebas Saber 11°.



COMPETENCIAS	CIENCIAS NATURALES	INGLÉS	LECTURA CRÍTICA	MATEMÁTICAS	CIUDADANAS	GLOBAL
Ciencias Naturales	1	0,715	0,790	0,825	0,814	0,929
Inglés		1	0,691	0,694	0,691	0,795
Lectura Crítica			1	0,761	0,809	0,905
Matemáticas				1	0,782	0,919
Ciudadanas					1	0,923
Global						1

In conclusion, decision trees were able to show solid information about the different stablished parameters.

3. MATERIALS AND METHODS

In this section, we explain how the data was collected and processed and, after, different solution alternatives considered to choose a decision-tree algorithm.

3.1 Data Collection and Processing

We collected data from the *Colombian Institute for the Promotion of Higher Education* (ICFES), which is available online at ftp.icfes.gov.co. Such data includes anonymized Saber 11 and Saber Pro results. Saber 11 scores of all Colombian high schools graduated from 2008 to 2014 and Saber Pro scores of all Colombian bachelor-degree graduates from 2012 to 2018 were obtained. There were 864,000 records for Saber 11 and records 430,000 for Saber Pro. Both Saber 11 and Saber Pro, included, not only the scores but also socio-economic data from the students, gathered by ICFES, before the test.

In the next step, both datasets were merged using the unique identifier assigned to each student. Therefore, a new dataset that included students that made both standardized tests was created. The size of this new dataset is 212,010 students. After, the binary predictor variable was defined as follows: Does the student score in Saber Pro is higher than the national average of the period?

It was found out that the datasets were not balanced. There were 95,741 students above average and 101,332 students below average. We performed undersampling to balance the dataset to a 50%-50% ratio. After undersampling, the final dataset had 191,412 students.

Finally, to analyze the efficiency and learning rates of our implementation, we randomly created subsets of the main dataset, as shown in Table 1. The dataset was divided into 70% for training and 30% for testing. Datasets are available at https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets .

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Train	15,000	45,000	75,000	105,000	135,000
Test	5,000	15,000	25,000	35,000	45,000

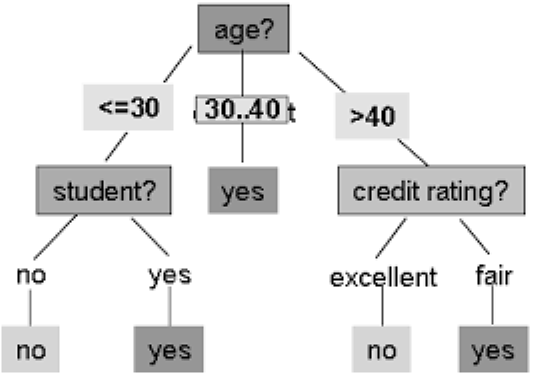
Table 1. Number of students in each dataset used for training and testing.

3.2.1 ID3 Algorithm

It is the most used algorithm in decision trees at present. The ID3 algorithm is simple and powerful at the same time. Its main idea is to approximate discrete values.

The algorithm starts with the first node (root) and keeps generating nodes without using the attribute of the set and determine the entropy (information gain) of the attribute. It continues the process until is no more instance there. The require choosing the attribute is based on choose the largest value of the information gain between the attributes left.

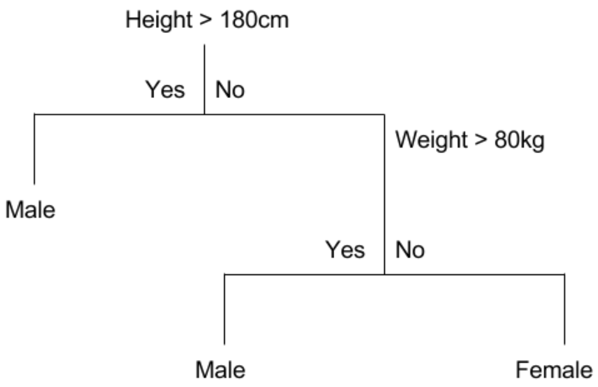
The set need to be order in series, where each of the values are categorized as attributes, being this the objective. The classification is a binary decision (yes or no).



3.2.2 CART (Classification And Regression Trees)

This model allows to have input and output variables such as nominal, ordinal and continuous. This algorithm is used to predict categories of objects and to predict continuous values. At the time of implement the decision tree, the nodes can only be divided in two groups since this is the restriction that binary method imputes. CART uses the Gini index as a measure to identify the node with the greater reduction of impurity and it is used to split the node records.

CART only accept two values: number or categories.



3.2.3. C4.5 Algorithm

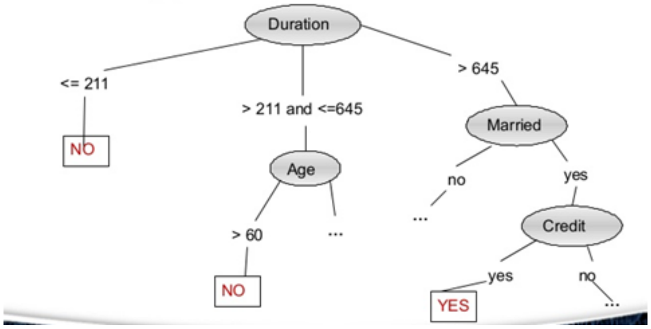
This algorithm is an improved version from the ID3. A decision tree generated (The tree is based in the depth-first strategy) by the C4.5 algorithm is used to for classification (Frequently known as a statistic classifier). The algorithm starts with data distribution performed recursively.

In the process it selects the all samples and identify which can be the most efficiently at the time of divides the samples set in a sort of subsets uses to take decision, basically in each node the system need to decide which sample take to do a data distribution. The entropy criterion (The attribute with more information gain) is used to generate new nodes as ID3 algorithm does. Finally, the attribute with the highest information gain is stablished as the decision parameter.

The C4.5 algorithm propose two tentative tests:

- Standard test: For discrete variables with a specific result and branches for each value.

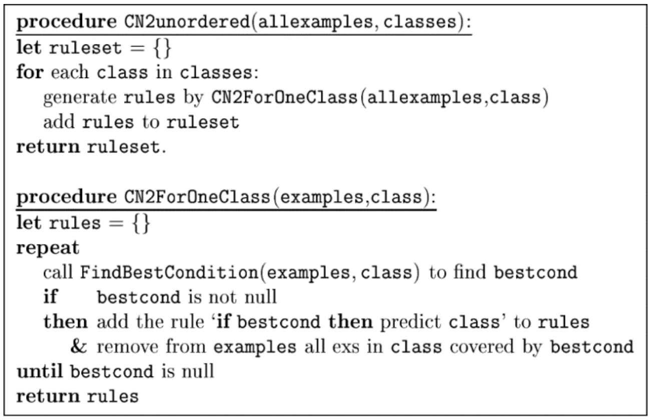
- Complex test: As the standard test use discrete variables but the result is just for a specific group where the variables are assigned.



3.2.4. CN2 algorithm

It is an algorithm based in the induction system (method uses to classify the variables present in the problem). It collects ideas of ID3 and AQ algorithms. The CN2 algorithm was designed to work with problems where are lack of information and in some time a poor description of the language implemented, basically when the "training data" is imperfect.

The processes begin when it identifies a set of examples and started to do a complex search to stablish the classification rules with the objective of organize all the variables that are present. In the classification the program can adding new conjunctive terms or removing a disjunctive element in one of its selectors. During the learning process the algorithm must take some decisions to evaluate the functions that have given the condition set.



```
procedure CN2unordered(allexamples, classes):
let ruleset = {}
for each class in classes:
    generate rules by CN2ForOneClass(allexamples,class)
    add rules to ruleset
return ruleset.

procedure CN2ForOneClass(examples,class):
let rules = {}
repeat
    call FindBestCondition(examples, class) to find bestcond
    if     bestcond is not null
    then add the rule 'if bestcond then predict class' to rules
        & remove from examples all exs in class covered by bestcond
until bestcond is null
return rules
```

REFERENCES

1. Acosta, J., Oller, L., Sokol, N., Balado, Sardiñas, J., Montero, D., Balado, Sansón, R. and Sardiñas, M.E. Revista Cubana de Pediatría, 88(4). 441-453. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0034-75312016000400005&lng=en&tlng=en

2. Clark. P and Niblett. T. 1989 The CN2 induction algorithm. The Turing Institute, Kluwer Academic Publishers, United Kingdom. https://link.springer.com/content/pdf/10.1023/A:1022641700528.pdf

3. Timarán, R., Calderón, A. and Hidalgo, A. Universidad y Salud. 19(3). 359-365.

   DOI: http://dx.doi.org/10.22267/rus.171903.101

4. Timarán, R., Calcedo, J. and Hidalgo,A. Rev.investig.desarrollo.innov.9(2).363-378.

   DOI: 10.19053/20278306.v9.n2.2019.9184

5. Martinez, Z. and Menendez, J. 2004. Predicción de crisis empresariales en seguros no-vida. una aplicación del algoritmo See5. Universidad Complutense. Spain. https://eprints.ucm.es/6834/1/04010.pdf

6. Ramírez, P.E. and Grandón, E.E. Formación Universitaria. 11(3), 3-10.

   https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-50062018000300003&lng=en&tlng=en

7. Wikipedia. 2019. ID3 algortihm. (22 May 2019). https://es.wikipedia.org/wiki/Algoritmo_ID3