PREDICTION OF ACADEMIC SUCCESS, IN SUPERIOR EDUCATION TEST, USING DECISION TREES

Jose Manuel Fonseca Palacio	Santiago Puerta Florez	Mauricio Toro
Eafit University	Eafit University	Eafit University
Colombia	Colombia	Colombia
jmfonsecap@eafit.edu.co	spuertaf@eafit.edu.co	mtorobe@eafit.edu.co

ABSTRACT

The aim of this study is to present a classification based on decision trees to predict whether a student that is going to present the "Pruebas Saber Pro" is going to be over the average. The study analyzes 135000 cases of students belonging to Colombian universities, which include parameters such as gender, results of "Pruebas Saber 11", socioeconomical aspects of the student and the family of the student, the studies they realized, results while they studied, if they studied abroad, if they waited before starting university and other type of data. We expect that with the amount of data, and using the classification based on decision trees, our predictions are mostly accurate, so they can use this information for the improvement of student's studies, the improvement of their possible results and the decision based in such results.

KEYWORDS

Decision trees, algorithm, artificial intelligence, grade prediction, ID3, Data mining, nominal data.

1. INTRODUCTION

The prediction of academical success with decision trees has been a really concurred topic with different researches. For starters, the one involving the ID3 decision algorithm for the analysis and prediction of the student's grades, or the one that used C4.5 algorithm as its base to predict grades as well. All these trees and researches that are being mentioned will be introduced and shown later on this document, but the question to be asked is why are these predictions important? Well, in our case we are predicting student's Saber Pro grades based on different socio-economical standards, like capital, parents' income, social stratum, gender, number of hours spent on the internet. The solution to this problem will let us know if the factors mentioned before really have an important impact on the probability of scoring above the average in the Saber pro test, therefore helping different students to achieve a score above the medium.

2. THE PROBLEM

The problem that we are facing, is the design of an algorithm based in decision trees and based as well in the data of the Saber 11 test. This algorithm will let us predict the probability of a certain student to score above the standard in the Saber Pro examinations, based on socio-economical standards.

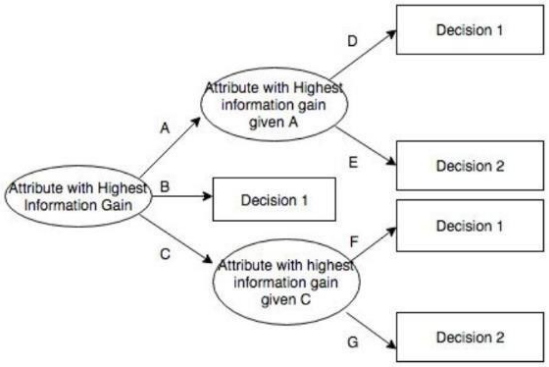
3. RELATED PROJECTS

3.1 ID3 Decision Tree Algorithm

This type of decision tree is based on a top-down greedy search, in which the tree at each iteration decides the best feature now to create a node. ID3 stands for Iterative Dichotomizer 3 and is named such because the algorithm repeatedly dichotomizes (divides) features into two or more groups at each step.

How does the ID3 decides the best feature?

The algorithm uses Information Gain or just Gain to find the best feature. The feature with the best information gain is selected as the best one.



3.2 C4.5 Algorithm

This algorithm is an extension of the ID3 algorithm developed by Ross Quinlan, differing with the ID3, this algorithm is used for statistical classification. C4.5 generates decision trees from a data training set such as ID3 does. Each example is a vector were the example's attributes or characteristics are represented. C4.5 picks the attribute that most efficiently divides the samples in enriched subsets. The attribute with the highest information gain is picked as the decision parameter.

3.3 C5.0 Algorithm

C5.0 it is an upgrade to the c4.5 algorithm with minor changes. In relation to C4.5, C5.0 is faster, has a more efficient use of space, uses smaller decision trees with similar results. Supports Boosting which gives decision trees more precision. It also has Winnowing which is a classification algorithm which eliminates those attributes that are of little help.

3.4 Chi-square Automatic Interaction Detector (CHAID)

CHAID is a tool used to discover the relationship between variables. CHAID analysis builds a predictive model, or tree, to help determine how variables best merge to explain the outcome in the given dependent variable. This type of decision tree allows multiple types of data to be used such as, analysis, nominal, and ordinal. CHAID creates all possible cross tabulations for each categorical predictor until the best outcome is achieved and no further splitting can be done.

3.5 Student Performance analysis and prediction

S.N	Name	Description	Possible Values
1	Gender	Student Gender	Male, Female
2	Branch	Student Branch	CSE,IT,MECH,ECE
3	Age	Age of student	22,23,24,25,26
4	Board of 10^{th}	Name of High school board	CBSE,ICSE,HCSE
5	Board of 12^{th}	Name of Senior secondary board	CBSE,ICSE,HBSE
6	10^{th}-Grade	Student 's Grades in class 10^{th}	A,B,C
7	12^{th} -Grade	Student 's Grades in class 12^{th}	A,B,C
8	1^{st} –year-Grade	Aggregate grade of 1^{st} and 2^{nd} semester	A,B,C,F
9	2^{nd} –year-Grade	Aggregate grade of 3^{rd} and 4^{th} semester	A,B,C,F
10	3^{rd} –year-Grade	Aggregate grade of 5^{th} and 6^{th} semester	A,B,C,F
11	Ag-G-3^{rd}	Aggregate grade up to 6^{th} semester	A,B,C,F
12	7^{th} sem-Grade	Grade of 7^{th} semester	A,B,C,F
13	Ag-G-7^{th}	Aggregate grade up to 7^{th} semester	A,B,C,F
14	Backlog-no.	Total no of Backlogs (till 7^{th})	0,1-5,6-10,>10
15	Gap	Gap in study(in years)	0,1,2
16	Region	Region from where a student belongs to.	NCR,FARIDABAD, OUTER ZONE
17	Backlog	Backlogs (till 7^{th})	YES,NO
18	Final/Class	Prediction Class	A,B,C,F

The student performance analysis and prediction is a project made by a student of Manav Rachna College of engineering, where, like this project, presented a model based on decision trees that described the possible performance a student could have on the Btech exams based on some factors related to the student's life during secondary as well as some personal factors such as gender, age, gap in study measured in years after graduating, the region they belong to, the name of the school board, and their grades during certain classes and semesters. The results for this prediction are given in possible grades in the exams earlies mentioned, and are classified from F to A.

For the model creation a C4.5 decision tree was used, which is based in gain ratio as attribute selection ratio. This allowed the study to determine certain factors from the start of the tree since they had more gain ratio than others. For example, if they had an A in the Ag-G-7th (which is the root of the tree) the prediction for it would be an A right away due to the gain ratio it has.

The project after the predictions and cross validation with the actual information the study had an accuracy of 80.15% and 82.58% in which they used j48 as the java version for the C4.5 decision tree. This concluded that it was a good project for forecasting grades of students. [2]

3.6 Student Grade Analysis and Prediction Using ID3 Algorithm

The student grade analysis and prediction was a project conducted by Khin Khin Lay and San San Nwe, this project had the purpose of predicting the performance of students by feeding the Algorithm nominal data such as: Attendance, aptitude, assignment, test, and presentation by employing the ID3 Decision Tree.

For measuring (best information gain to worst information gain) the information received they applied an information gained metric, the idea with this kind of metric was to split the criteria to the determine the criteria with the best information gain for the creation of a particular node in the tree. Then the tree goes on by deciding which criteria has got the best information and with the information retrieved making the rules. For example, they fed the algorithm the table shown below.

Sr. no.	Roll no.	Attend-ance	Apti- tute	Assign-ment	Test	Presentation	Grade
1	IT1	Good	Avg	Yes	Pass	Good	Excellent
2	IT2	Good	Avg	Yes	Pass	Good	Excellent
3	IT 3	Good	Avg	Yes	Pass	Good	Excellent
4	IT4	Good	Avg	Yes	Pass	Good	Excellent
5	IT5	Good	Avg	Yes	Pass	Good	Excellent
6	IT6	Avg	Avg	Yes	Pass	Avg	Good
7	IT7	Poor	Good	Yes	Pass	Avg	Good
8	IT8	Avg	Good	Yes	Pass	Avg	Good
9	IT9	Avg	Good	Yes	Pass	Avg	Good
10	IT10	Poor	Poor	No	Fail	Poor	Fail
11	IT11	Poor	Poor	No	Fail	Poor	Fail
12	IT12	Avg	Age	Yes	Pass	Age	Good
13	IT13	Good	Good	Yes	Pass	Good	Excellent
14	IT14	Good	Good	Yes	Pass	Good	Excellent
15	IT15	Good	Good	Yes	Pass	Good	Excellent

And the Decision Tree generated the set of rules shown below. [3]

IF Presentation = "Good" AND Attendance = " Good" THEN Grade = "Excellent"
IF Presentation = "Average" AND Test = " Pass" THEN Grade = "Good"
IF Presentation = " Poor" AND Test = " Fail" THEN Grade = "Fail"

3.7 Prediction of Student Dropout in a Chilean Public University through Classification based on Decision Trees with Optimized Parameters

The study presents a classification based on decision trees (DTBC) with optimized parameters to be able to predict the dropout in Chilean Public Universities. The study analyzed the cases of 5288 students at a Chilean public university. The parameters were optimized to improve the accuracy of

the software for predictions they used called RapidMiner, which with they managed to achieve an accuracy of the 87.27%. With this they concluded that the usage of optimization of parameters in applications as DTBC results in a better precision in comparison to other research with a similar amount of data. They used a technique of Data mining to find a pattern on what influenced the most on university dropout and with that they used to predict whether the student would drop out or not with the software RapidMiner Studio 7.5, which used the algorithm C4.5 which was selected as the most influential algorithm for data mining and Decision trees-based classification.[7]

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

This was the data that was collected and analyzed for the study.

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

The data after the prediction and after the actual results.

4. ABOUT GINI IMPURITY

With the type of data structure that we are dealing with (decision trees) first we must find the variable that divides better our tree. One method to find this variable is called “Gini impurity”.

The Gini impurity is a metric that will tell us how much a chosen data will be labeled wrongly. This Gini calculation will be done in all nodes of the corresponding tree, finally and by calculating the Gini impurity of the right and left node and the pondered impurity of the root we will know the variable that best divides our data set.

You might ask, all the explained above for what? Well, by doing so, we will have a tree that predicts with more accuracy.

4.1 CHOSEN TREE:

The chosen data structure is a CART decision tree which will help us in its implementation in random forests and will help us ease the time complexity of the algorithm.

4.2 COMPLEXITY ANALYSIS:

Broadly the time complexity for the training of the CART decision tree would be O (v * m log(n)), where v is the number of rows and m is the number of columns of the data matrix. Then we got that the time complexity to get a result is O(k) being k the depth or number of nodes of the tree.

Also let us not forget about the random forest, with the implementation of these the training time complexity escalates to O(n*log(n)*d*m), being m the number of trees that we want our forest to have, and d being how many variables we want to sample at each node of the trees that craft our forest. We end by having an O(k*m) time complexity.

To put it in measurable quantities using random forests and with an input of 135000 data matrix the accuracy percentage was 80%.

In terms of memory…

In terms of memory used the CART decision tree has a space complexity of O(p) being p the number of nodes in the tree.

The random forest in the other hand, hand us a space complexity of O(p*m).

4.3 RESULT ANALYSIS:

Finally, the tree determines if a given person is likely to score above the average or not in the saber pro tests.

5. CONCLUSIONS

In this research we found out that decision trees can determine things and accelerate the solution time for various problems.

Our data structure could finally determine if a given person is likely to score above the average or not in the saber pro tests.

Decision trees can have many applications to real life problems. They can be used for predicting and preventing many negative things.

REFERENCES:

[1.] towardsdatascience. 2020. Decision Trees: ID3 Algorithm Explained. (31 March 2020). Retrieved *August 8, 2020* from https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1

[2.] Wikipedia. 2020. C4.5 algorithm. (29 February 2020). Retrieved *August 11, 2020* from https://en.wikipedia.org/wiki/C4.5_algorithm

[5.] pdfs.sematicscholar.org. 2013. A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction. (January 2013). Retrieved *August 10, 2020* from https://pdfs.semanticscholar.org/a62b/5b8e3c34984ca22e4f9ea5f1a58770fce7c8.pdf

[6.] zenodo.org. 2019. Using ID3 Decision Tree Algorithm to the Student Grade Analysis and Prediction. (20 July 2019). Retrieved *August 11, 2020* from https://zenodo.org/record/3590845#.XzK3jyhKhPZ

[7.] Ramírez, P. E., & Grandón, E. E. (2018). Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados / Prediction of Student Dropout in a Chilean Public University through Classification based on Decision Trees with Optimized Parameters. Formación Universitaria, 11(3), 3–10.