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

Jose Manuel Fonseca Palacio
Eafit University
Colombia
jmfonsecap@eafit.edu.co

Santiago Puerta Florez
Eafit University
Colombia
spuertaf@eafit.edu.co

Mauricio Toro
Eafit University
Colombia
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.
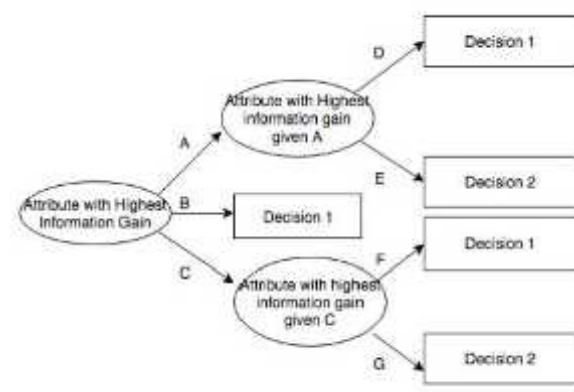
3.2 C4.5 Algorithm

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 10th	Name of High school board	CBSE,ICSE,HCSE
5	Board of 12th	Name of Senior secondary board	CBSE,ICSE,HISE
6	10th-Grade	Student 's Grades in class 10th	A,B,C
7	12th Grade	Student 's Grades in class 12th	A,B,C
8	1st -year-Grade	Aggregate grade of 1st and 2nd semester	A,B,C,F
9	2nd -year-Grade	Aggregate grade of 3rd and 4th semester	A,B,C,F
10	3rd -year Grade	Aggregate grade of 5th and 6th semester	A,B,C,F
11	Ag-G-3rd	Aggregate grade up to 6th semester	A,B,C,F
12	7th sem-Grade	Grade of 7th semester	A,B,C,F
13	Ag-G-7th	Aggregate grade up to 7th semester	A,B,C,F
14	Backlog-no	Total no of Backlogs (till 7th)	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 7th	YES,NO
18	FinalClass	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.

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

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		N	%
☐ Femenino	Nominal	2.941	55,6
☐ Masculino		2.346	44,4
Colegio de Enseñanza Media		N	%
☐ Privado	Nominal	2.013	38,1
☐ Público		322	6,1
☐ Subvencionado		2.894	54,7
Deserción		N	%
☐ No	Nominal	4.189	79,2
☐ Si		1.099	20,8
Total		5.288	100,0

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

		Predicción de Deserción		
		SI	No	Total
Deserción Real	Si	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.

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.