The problem of this project is to create an algorithm to

predict if a student will get grades above the average in the

"Pruebas Saber Pro". Based in the past grades of the test and

some sociodemographic variables as age, gender, the time spent

on the internet, between others. We consider this problem is

important because of the huge impact it could make in the prediction

of the variables that can influence the success of a student in

the future.

For the solution we will create a decision tree that takes the

information mentioned above and that will predict if the student

in mention will get a grade above the average or not. For this

first part of the project, we will analyze some decision trees

as they follow.

Decision trees

Decision tree algorithm can be classified under the supervised learning category, and can be used for both regression (numerical) and classification (classes) problems. We are going to analyze the following four:

ID3, C4.5, C5.0 and CART

Id3

El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, ( el atributo a clasificar ) es el objetivo, el cual es de tipo [binario](https://es.wikipedia.org/wiki/Binario) ( positivo o negativo, sí o no, válido o inválido, etc. ).

De esta forma el [algoritmo](https://es.wikipedia.org/wiki/Algoritmo) trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo.

ID3 realiza esta labor mediante la construcción de un [árbol de decisión](https://es.wikipedia.org/wiki/%C3%81rbol_de_decisi%C3%B3n).

In this algorithm the set of values must be a series of tuples, each of them named attributes. Inside the attributes there is one whom is the objective. This objective must be of binary type.

By this way the algorithm tries to set the hypothesis that classifies new entries of tuples in true or false.

Entropy:

Is the amount of uncertainty in the data.

Information gain

Is the measure on the difference in entropy from before to after the set was split on attribute A.

Steps of the algorithm

1. Calculating the entropy of all the attributes
2. Split the set into subsets by the attribute who minimizes entropy.
3. Make a decision node containing the attribute
4. Use recursion on subsets using the remaining attributes.

C4.5

C4.5 builds decision trees from a set of training data in the same way as [ID3](https://en.wikipedia.org/wiki/ID3_algorithm), using the concept of [information entropy](https://en.wikipedia.org/wiki/Entropy_(information_theory)). The training data is a set

S

=

s

1

,

s

2

,

.

.

.

{\displaystyle S={s\_{1},s\_{2},...}}

 of already classified samples. Each sample

s

i

{\displaystyle s\_{i}}

 consists of a p-dimensional vector

(

x

1

,

i

,

x

2

,

i

,

.

.

.

,

x

p

,

i

)

{\displaystyle (x\_{1,i},x\_{2,i},...,x\_{p,i})}

where the

x

j

represent attribute values or [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) of the sample, as well as the class in which

s falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized [information gain](https://en.wikipedia.org/wiki/Information_gain_in_decision_trees) (difference in [entropy](https://en.wikipedia.org/wiki/Entropy_(information_theory))). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then [recurses](https://en.wikipedia.org/wiki/Recursion_(computer_science)) on the [partitioned](https://en.wikipedia.org/wiki/Partition_of_a_set) sublists.

This algorithm has a few [base cases](https://en.wikipedia.org/wiki/Base_case_(recursion)).

* All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
* None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

[https://www.statisticshowto.com/decision-tree-definition-and-exam [1]ples](https://www.statisticshowto.com/decision-tree-definition-and-examples)

[https://www.elmundo.es/sociedad/2016/02/10/56bb2b56268e3e357d8b4629.ht [2]ml](https://www.elmundo.es/sociedad/2016/02/10/56bb2b56268e3e357d8b4629.html)

<ftp://ftp.boulder.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/14.0/TREE-CART.pdf>

[https://www.hindawi.com/journals/cin/2018/6347 [3]186/](https://www.hindawi.com/journals/cin/2018/6347186/)

[https://en.wikipedia.org/wiki/C4.5\_al [4]gorithm](https://en.wikipedia.org/wiki/C4.5_algorithm)

[https://es.wikipedia.org/wik [5]i/Algoritmo\_ID3](https://es.wikipedia.org/wiki/Algoritmo_ID3)

[https://en.wikipedia.org/wiki/Supervise [6]d\_learning](https://en.wikipedia.org/wiki/Supervised_learning)

[https://en.wikipedia.org/wiki/Chi-square\_automatic\_interaction\_detect [7]ion](https://en.wikipedia.org/wiki/Chi-square_automatic_interaction_detection)

Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J., 1984. *Classification and Regression Tree* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Cali

# Bibliografía

|  |  |
| --- | --- |
| [1] | Stephanie, «Decision Tree: Definition and Examples,» 13 09 2015. [En línea]. Available: https://www.statisticshowto.com/decision-tree-definition-and-examples. |
| [2] | EFE, «Perú, Chile y EEUU, entre los 11 países con más desigualdad educativa,» 10 02 2016. [En línea]. Available: https://www.elmundo.es/sociedad/2016/02/10/56bb2b56268e3e357d8b4629.html. |
| [3] | H. Mushtaq, Z. Wenhao, Z. Wu y R. A. Syed Muhammad, «Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores,» António D. P. Correia, 02 10 2018. [En línea]. Available: https://www.hindawi.com/journals/cin/2018/6347186/. [Último acceso: 10 08 2020]. |
| [4] | Wikipedia, «C4.5 Algorithm,» Wikipedia, 29 02 2020. [En línea]. Available: https://en.wikipedia.org/wiki/C4.5\_algorithm. [Último acceso: 11 08 2020]. |
| [5] | Wikipedia, «Algoritmo ID3,» Wikipedia, 25 04 2020. [En línea]. Available: https://es.wikipedia.org/wiki/Algoritmo\_ID3. [Último acceso: 10 08 2020]. |
| [6] | Wikipedia, «Supervised learning,» Wikipedia, 12 08 2020. [En línea]. Available: https://en.wikipedia.org/wiki/Supervised\_learning. [Último acceso: 13 08 2020]. |
| [7] | Wikipedia, «Chi-square automatic interaction detection,» Wikipedia, 05 01 2019. [En línea]. Available: https://en.wikipedia.org/wiki/Chi-square\_automatic\_interaction\_detection. [Último acceso: 13 08 2020]. |

fornia.