PREDICTING ACADEMIC SUCCESS IN HIGHER EDUCATION

UNIVERSIDAD EAFIT®

Team Presentation

Isabella Serna

Mariana Quintero

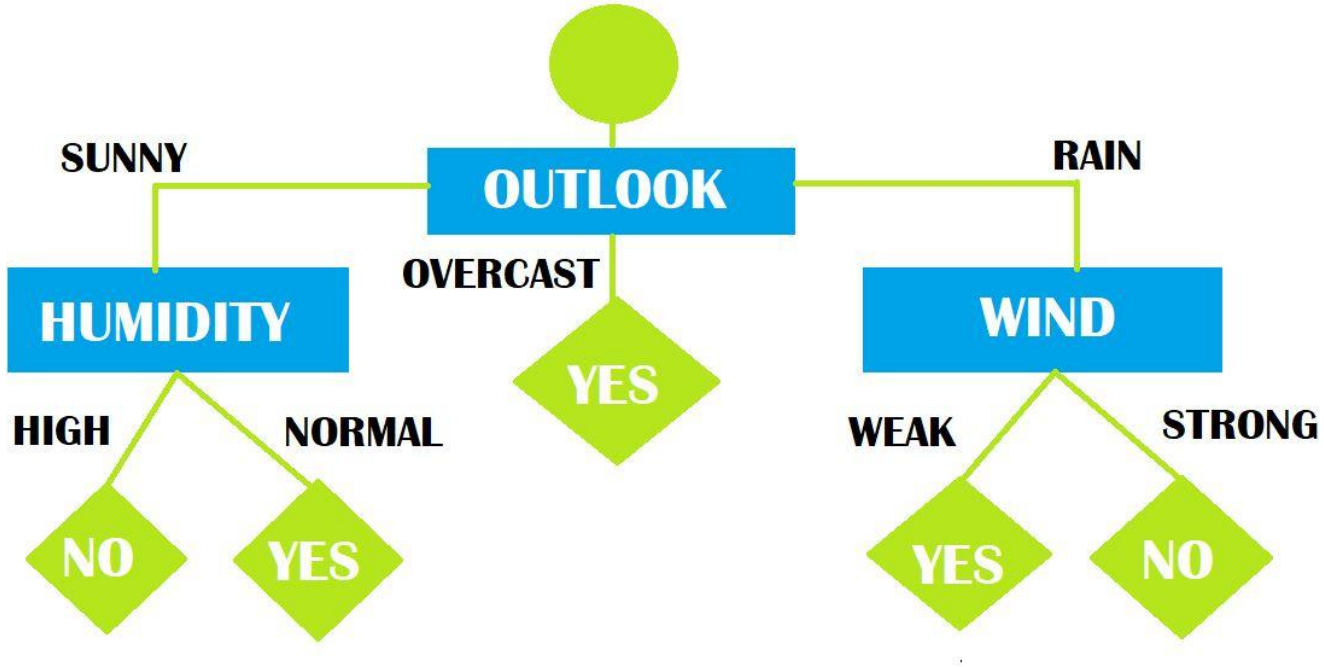
Miguel Correa

Mauricio Toro

https://github.com/Mariana13-31/ST0245-001/tree/master/proyecto

UNIVERSIDAD EAFIT

Algorithm Design



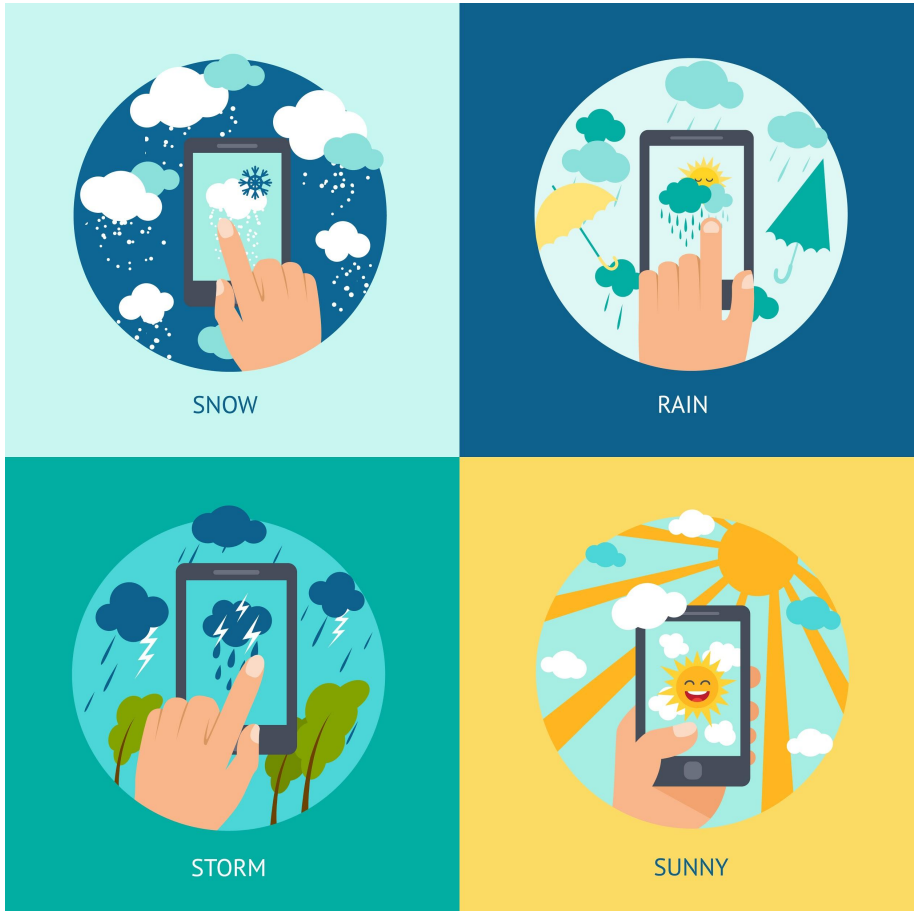


Figure 1. The algorithm used to build the binary decision tree was CART. In this figure we can see an example that processes different conditions in order to predict the weather using as root node the variable outlook, which has a lower impurity Gini so it is the best to divide the data that make up this example.

Node Splitting

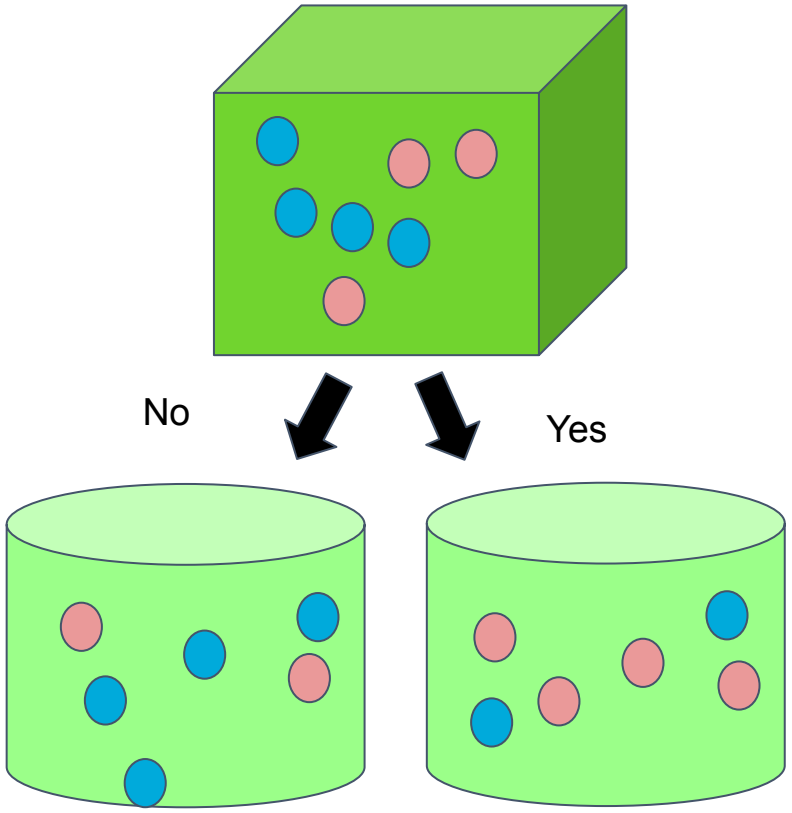


Figure 2. As an example, this split is based on the condition “English <=82.” For this case, left Gini impurity is 0.42, right Gini impurity is 0.41, and weigthed Gini impurity is 0.41.

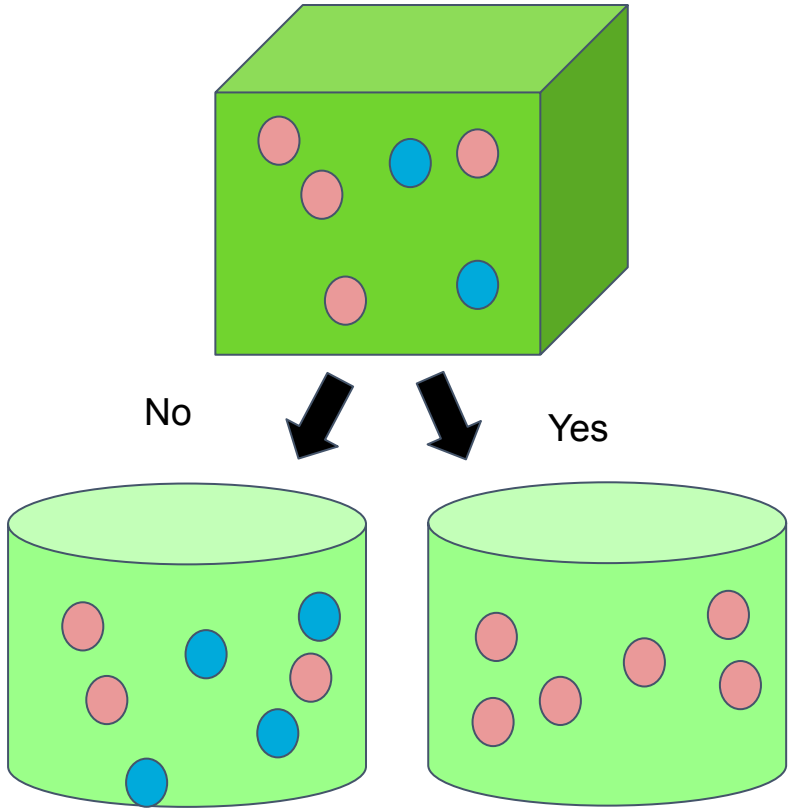


Figure 3. As an example, this split is based on the condition “Math <= 68.” For this case, left Gini impurity is 0.48, right Gini impurity is 0, and weigthed Gini impurity is 0.26.

Algorithm Design

Random Forest. To put it simply: the Random Forest creates multiple decision trees and combines them for a more accurate and stable prediction.

Decision-Tree Model

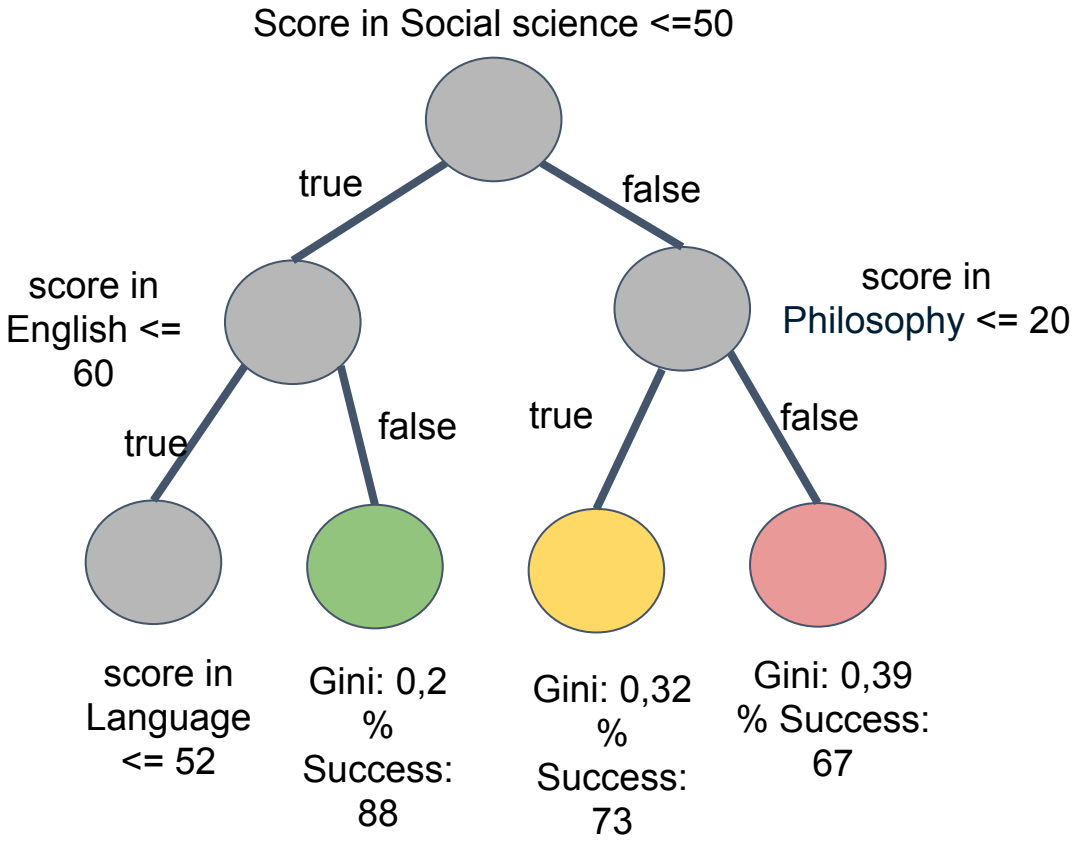


Figure 4. A binary decision tree to predict Saber Pro scores based on the results of Saber 11. Green nodes represent those with a high probability of success, yellow medium probability and red a low probability of success.

Most Relevant Features:

Social Sciences

English

Philosophy

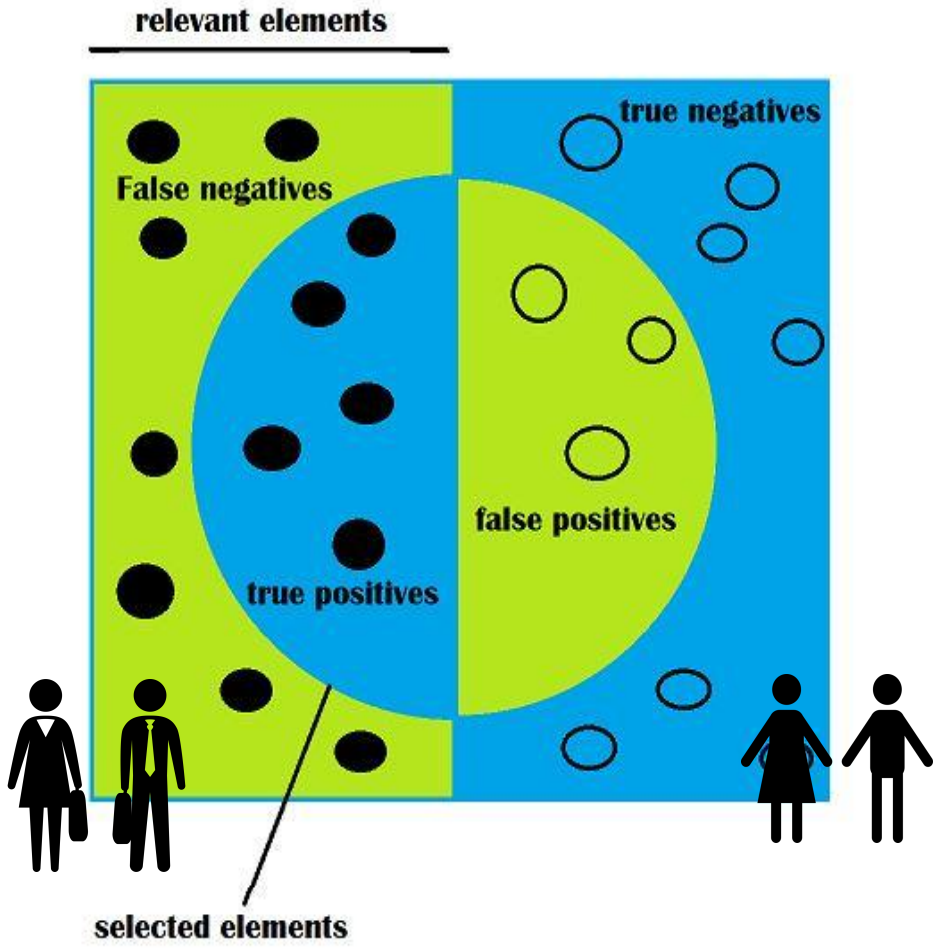
UNIVERSIDAD EAFIT

Algorithm Complexity

	Time Complexity	Memory Complexity
Training the model	O(m* t *nlogn)	O(m * n)
Testing the Model	O(n* t)	O(m * n)

Complexity in time and memory of the algorithm used that implements random forests based on the CART algorithm. Here we can interpret the variable **n** as the number of students, **m** the number of variables to be taken into account as relevant and **t** as the number of trees that make up the forest

Evaluation Metrics



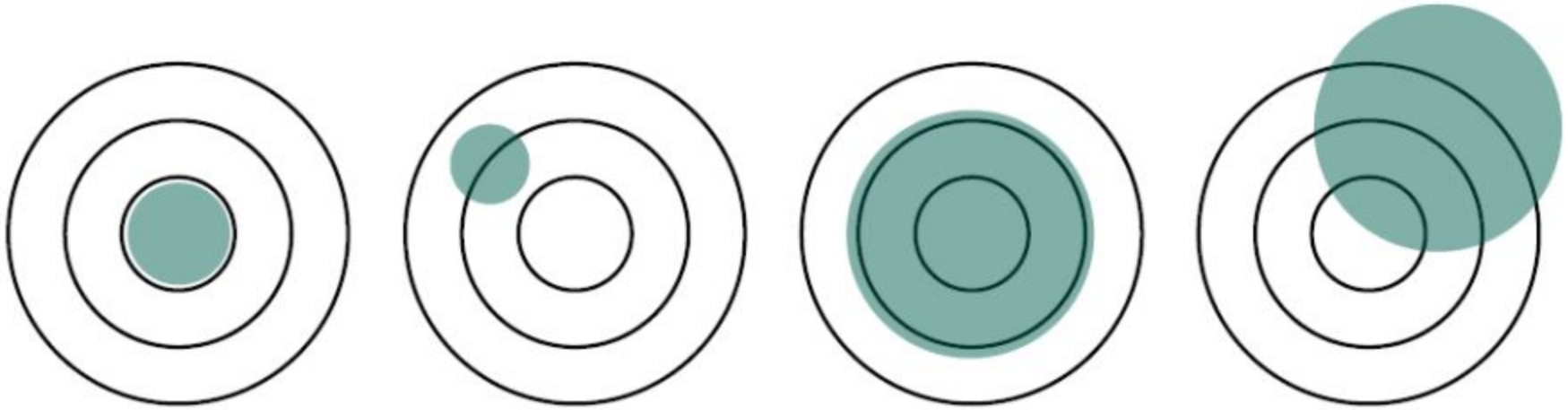
$\frac{\text{Success predictions}}{\text{Total predictions}}$	**Accuracy:** Total successful predictions over total data
$\frac{\text{Success predictions}}{\text{Total success}}$	**Precision:** Successful prediction of success over total success
$\frac{\text{Success predictions}}{\text{Total successful predictions}}$	**Recall:** Successful prediction of success over total successful predictions

Success predictions

Not Success predictions

UNIVERSIDAD EAFIT

Evaluation Metrics



High precision
Good trueness

High precision
Bad trueness

Low precision
Good trueness

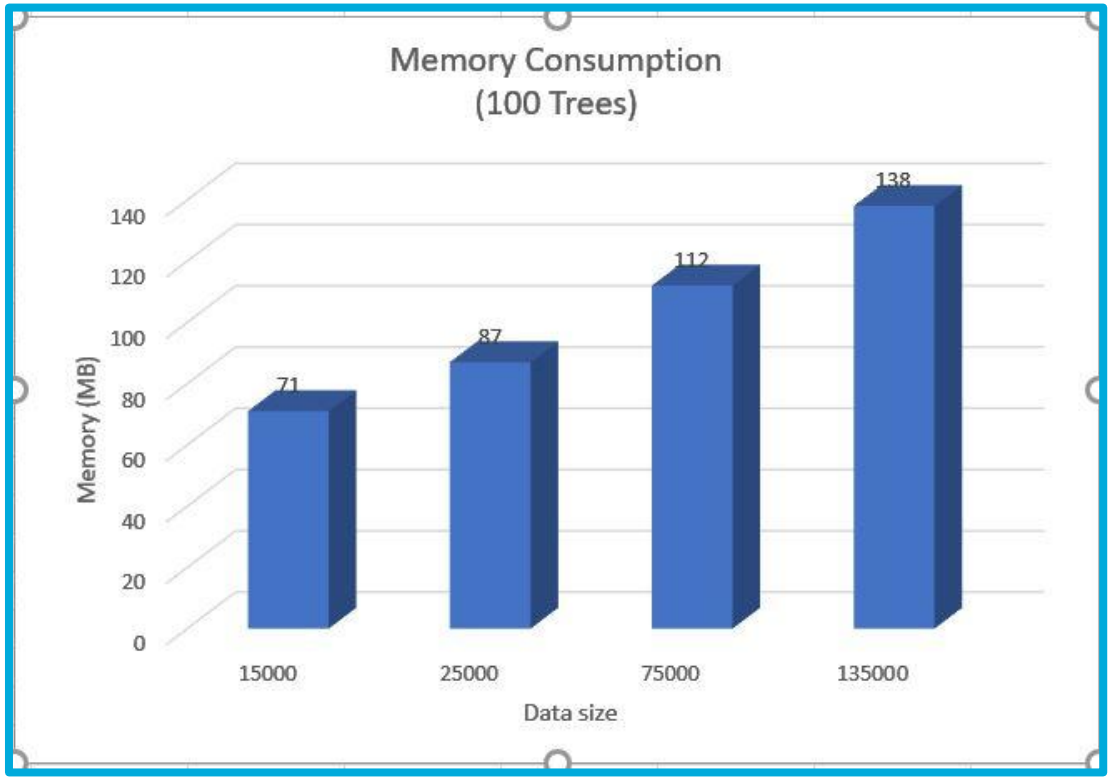
Low precision
Bad trueness

Evaluation Metrics

	Training data set	Testing data set
Accuracy	0.78	0.78
Precision	0.77	0.75
Recall	0.61	0.58

Evaluation metrics obtained with the training data set of 135,000 students and the testing data set of 25,000 students

Time and Memory Consumption



Time Consumption

Memory Consumption

UNIVERSIDAD EAFIT

THANK YOU!