SD201

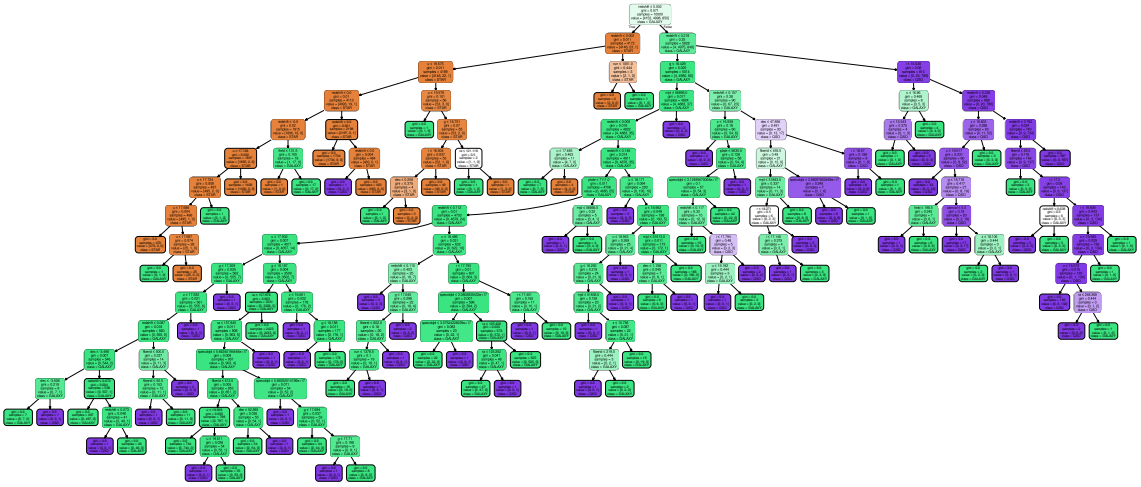
DECISION TREE

October 15, 2018

Student : Raphael REME
TELECOM-PARISTECH
Department of data sciences

1 Given the dataset we provided to you, build a decision tree using the default input parameters.

With the data provided I made a similar tree as the one given in exemple.



2 Compute the generalization error of the decision tree you built.

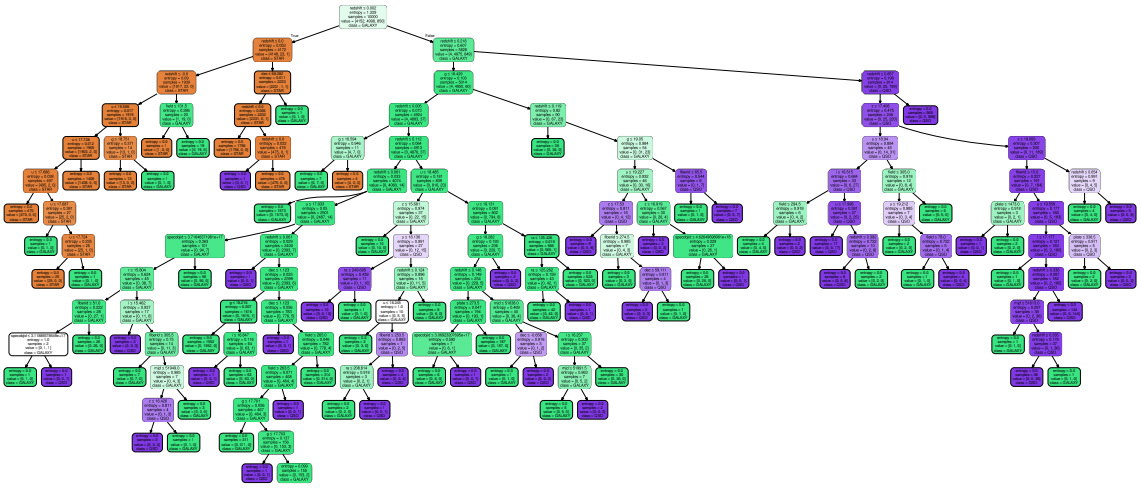
To compute it I use : $GenError = TrainError + 0.5 * NbLeaves$
As training error is zero, $GenError = 0.5 * NbLeaves$!
After computation I find $NbLeaves = 102$ => $GenError = \frac{51}{10000} = 0.51\%$

3 Build a decision tree with minimum generalization error.

The generalization error represents the expected error on a testing set. It takes in account the error on the testing set add to an error due to new data and the possibility of overfitting the training set (It is represent by the numer of leaves in the formula).
In order to minimize the generalization error we can therefore try to minimize the testing error or the added error (NbLeaves). But as the training error is already equal to zero, we can't minimize it more ! There is only the added error left.
To prevent overfitting (and to limit NbLeaves) we can use max_depth attribute to limit depth

of the tree and also min_samples_leaf attribute which set a minimal size to nodes : their size is then large enough to make a statistically significant decision. In both case NbLeaves decreases but the training error will probably increase. We can also try to change giny with entropy !
I tested all the possibility to compute the best generalization error : I had then $best\_gen\_error$ = 0.45% for the parameters :

1. Criterion = **entropy**
2. Max_depth = **14**
3. Min_samples_leaf = **1**



In fact not much has changed... if I had followed my instinct I would have reduced much more the depth and increase the min_samples_leaf ! As there are quite good small trees over the data (with good training error), one can think that theses huge trees over-fit the data...

4 Compare the decision trees you built in point 1 and the best one you obtained in point 2.

The second I got is I think better than the first one, as it is smaller and with less error. But as I said rigth before, may be a smaller tree with a little bit more generalization error would be better than these two !

5 PREDICT THE CLASS VALUE OF AN OBJECT OF YOUR CHOICE.

I choosed the objid : 1,23764870457714E+018 (the first one). Its redshift is negatif, u is lower than 19.589 and u is greater than 17.687 so it will be classed as a STAR !