Machine Learning - Project 1

Task 1 : Decision tree with 2 max leaf nodes

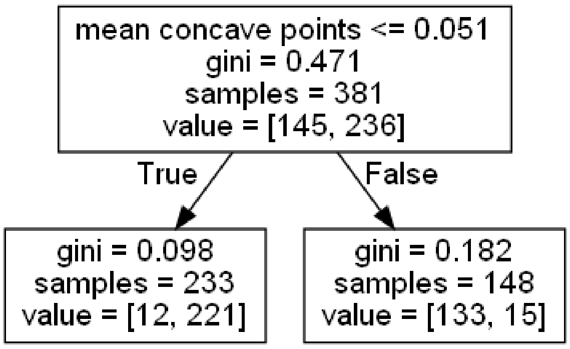

Figure 1: Decision tree with 2 max leaf nodes

As shown in figure 1, a decision tree with max leaf nodes = 2 allows one to move from a Gini index of 0.471 on the first level (which is roughly equivalent to flipping a coin to determine the class of the tumor) to a Gini index of 0.098 and 0.182 on the second one.

By discriminating on a single feature, the model performs significantly better than luck with a score of 92% on the training set and 88% on the testing one.

Regarding interpretability, the dataset is rather well distributed (about 61% of the samples in the first box and 39% in the other one) and the logic behind the algorithm is understandable by everyone.

Task 2 : Decision tree with 30 max leaf nodes

By increasing the max leaf nodes metaparameter, it changes the number of leaves (nodes that don't split further), it will also indirectly change the depth of the tree.

Regarding the score, the accuracy of this model rises at 100% on the training set and 90% on the testing one.

Regarding interpretability, this model is much more complex than the previous one. One can't understand the logic behind it in a simple way. The dataset is not well distributed between each leaf (2 of the 20 leaves cover 86% of the samples while 13 of them have 4 samples or less, see figure 2).

In conclusion, by increasing the max leaf node metaparameter, the decision tree has become more complex at the expense of interpretability and readability. One can say this model is overfitting because even if it has a perfect accuracy on the testing set, it performs barely better than the previous one (which was much simpler).

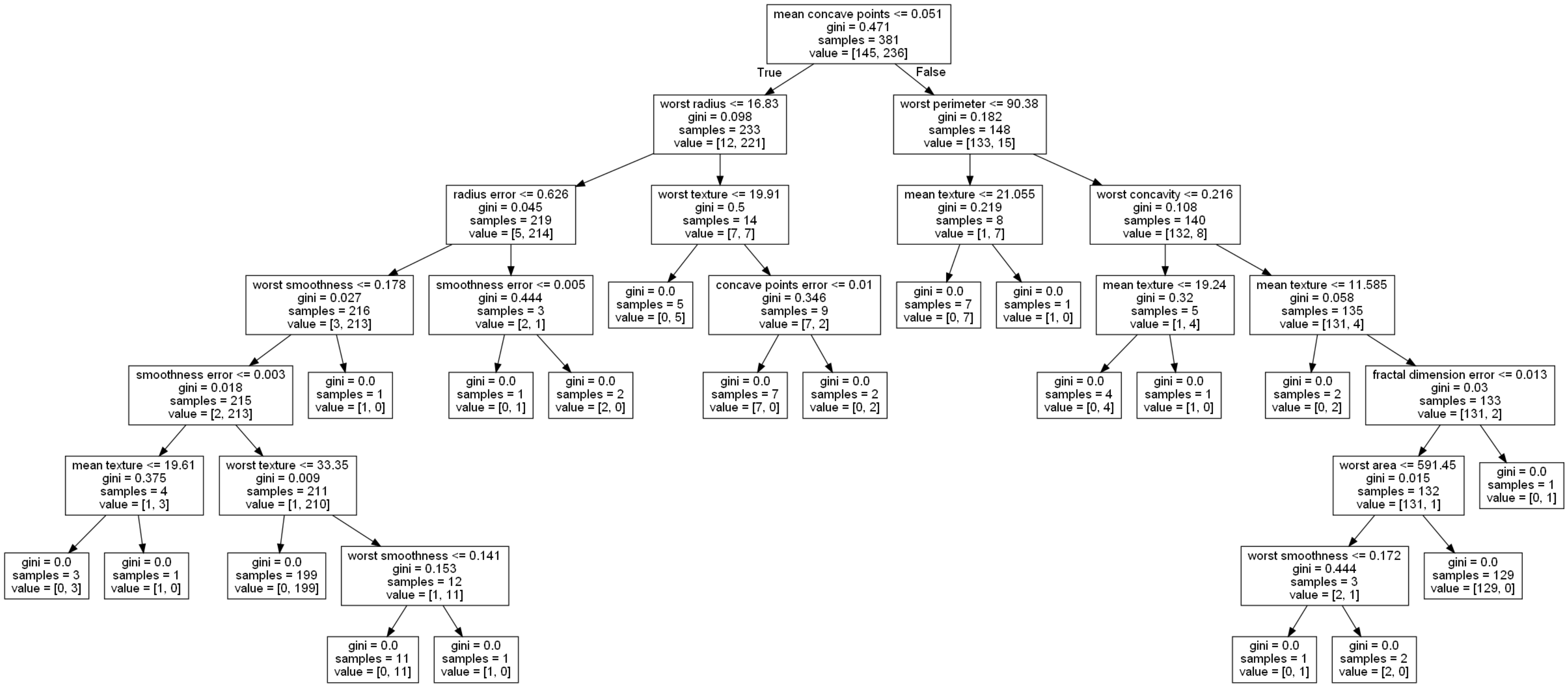

Figure 2: Decision tree with 30 max leaf nodes

Task 3 : Decision trees with 2 to 30 max leaf nodes

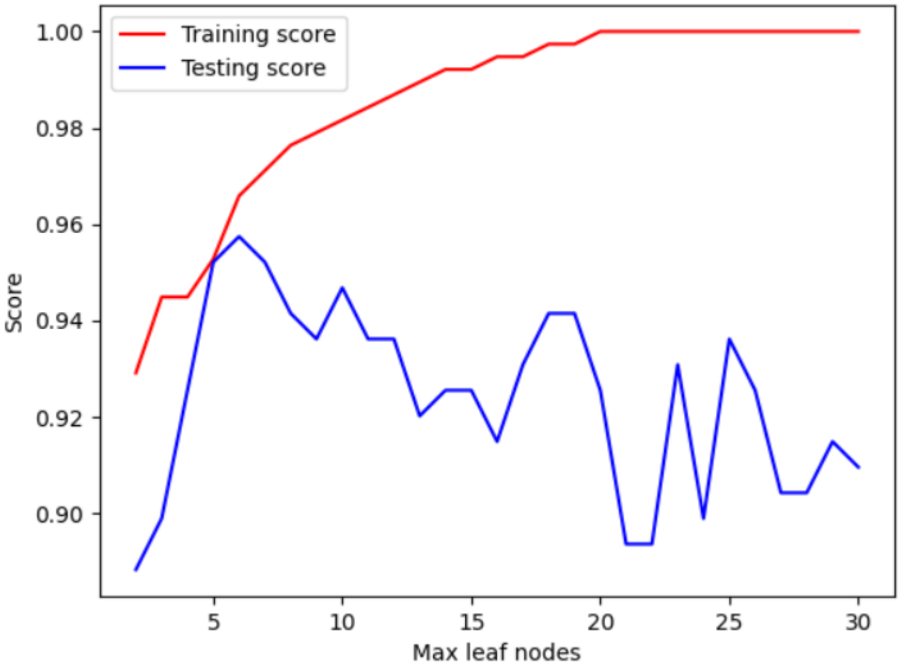

Figure 3: Accuracy of model when increasing the max leaf nodes metaparameter

As shown on the plot above, the model is underfitting before max leaf nodes = 5 (as the rising curve of testing set accuracy shows). The model is overfitting when max leaf nodes >=6 because even if the accuracy on the training set rises, the precision on new data gets worst. The model achieves perfect accuracy on the training set with 20+ max leaf nodes.

The optimum value for the max leaf nodes metaparameter is 5-6 (depending on the execution) as it gets the best precision on new data (95-96%).