Exercise 5.1: Examine overfitting

Task description:

- Measure the accuracy of the previously generated Decision Tree (3 levels) on the training data

- Construct two other Decision Trees with (maximum) depth 5 resp. 10 and measure their accuracy on the test data as well as on the training data

→ **Do you observe overfitting?**

Group 24: Christopher Schmidt, Gerd Mund, Maren Pielka

Our approach

- We used Python 3 to implement the TDIDT algorithm and the library pygraphviz for easier visualization.
- To store the tree internally we utilized nested python dictionaries.
- Each dictionary represents a node and contains the corresponding attribute, calculated gain, number of samples, threshold and the left and right child of the node.

Results:

Maximum tree depth	Dataset	Accuracy
3	Training	0.8616
	Test	0.8671
5	Training	0.9387
	Test	0.8844
10	Training	1.0
	Test	0.9133

→ **No, there is no measurable overfitting effect.**

On the contrary, the Decision Tree's performance increases on the training data as well as on the test data with increasing maximum depth.

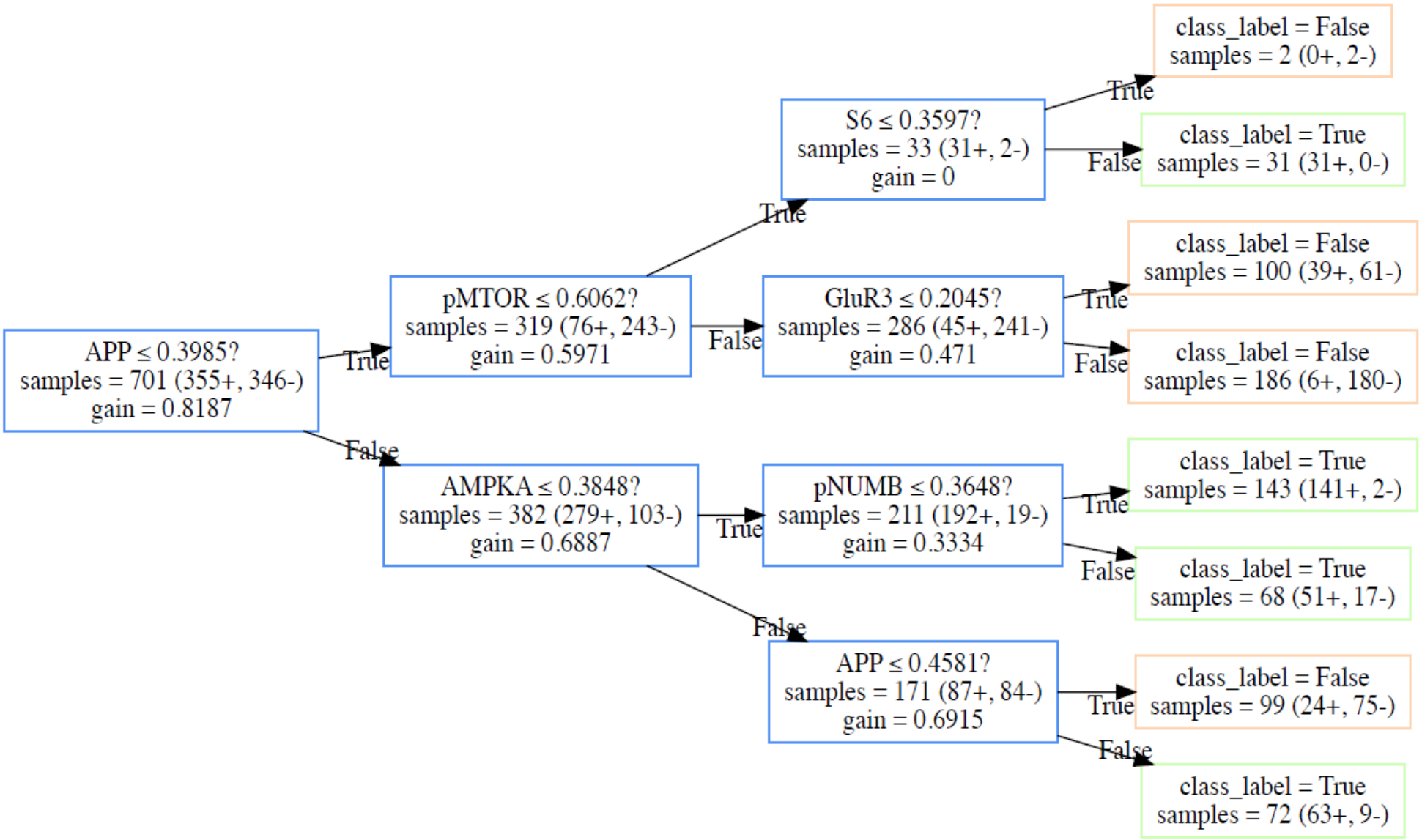
Other observations:

- The performance on the test data is slightly better than on the training data for a maximum depth of 3.

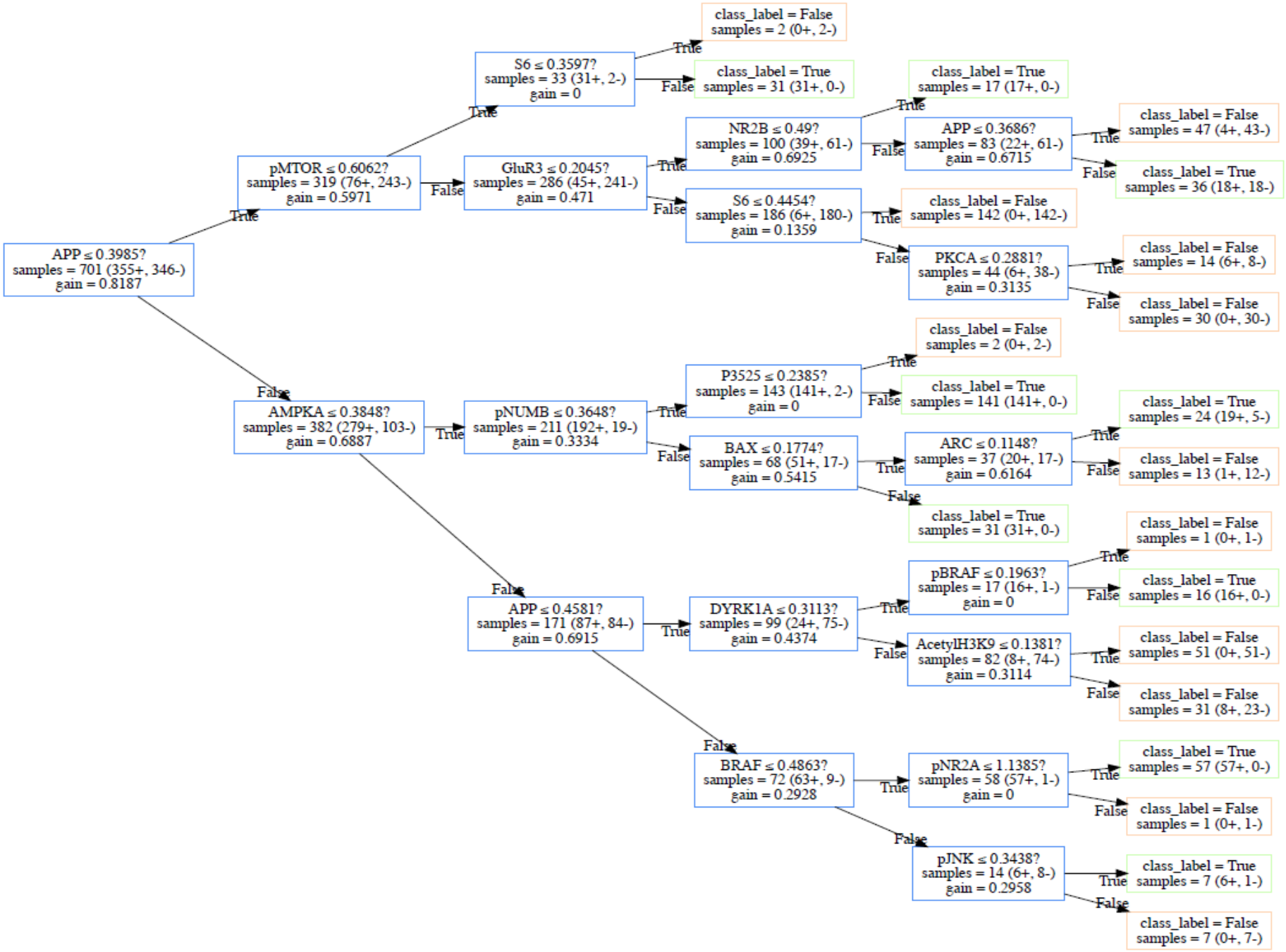
  → This small tree seems to be well generalizable.

- The tree with maximum depth of 10 fits the training data exactly, indicating that this tree is fully grown and it would not exceed a depth of 10 without pruning.
- The fact that the best performance on the test data is also achieved by this fully grown tree, shows that pruning does not yield any accuracy improvement for this problem, and overfitting is not an issue here.

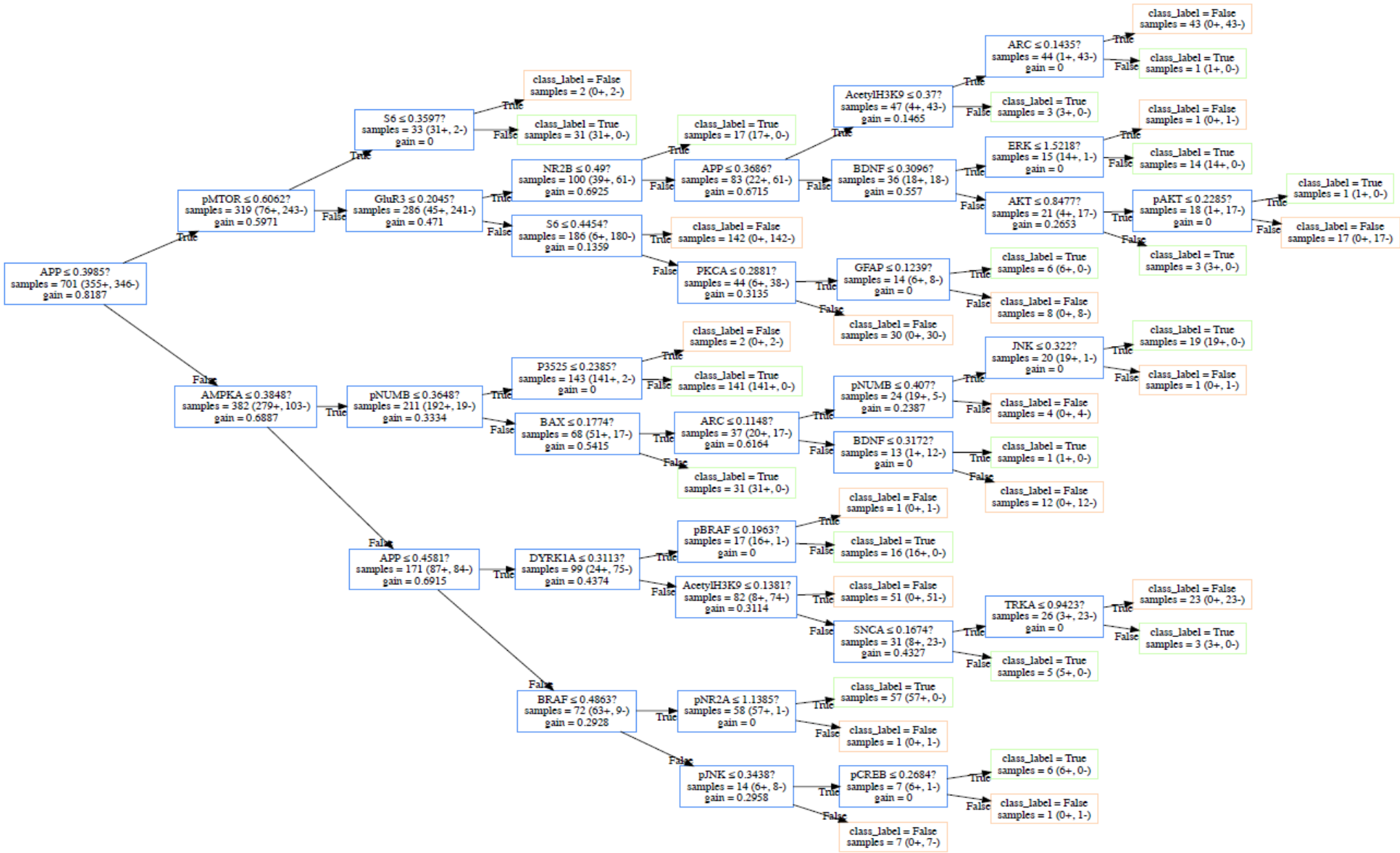
Maximum depth 3



Maximum depth 5



Maximum depth 10



Thank you!