Exercise 3:
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

- Repeat the task with the pruned trees and classification rules

→ **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.

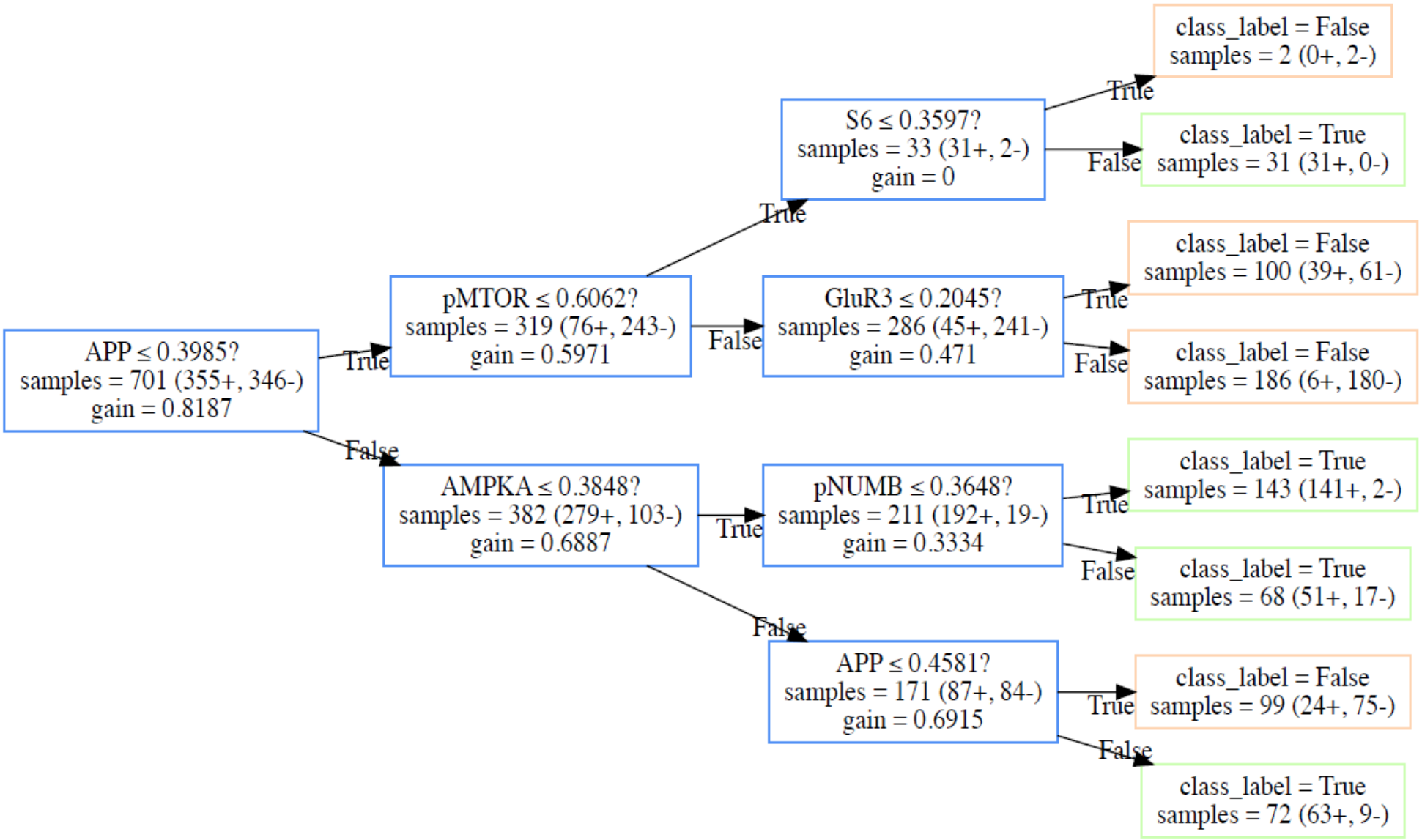
Results

	Original	Postpruning (accuracy)	Postpruning (pess. error)	Class rules (origin)	Class rules (pess. error)
Depth 3					
Test data	86.71%	86.71%	86.71%	86.71%	82.66%
Train data	86.16%	86.16%	86.16%	86.16%	81.6%
Depth 5					
Test data	88.44%	89.02%	89.02%	89.02%	84.39%
Train data	93.87%	92.15%	92.15%	92.15%	89.16%
Depth 10					
Test data	91.33%	91.62%	91.62%	91.62%	87.28%
Train data	100%	95.29%	95.29%	95.29%	95.29%

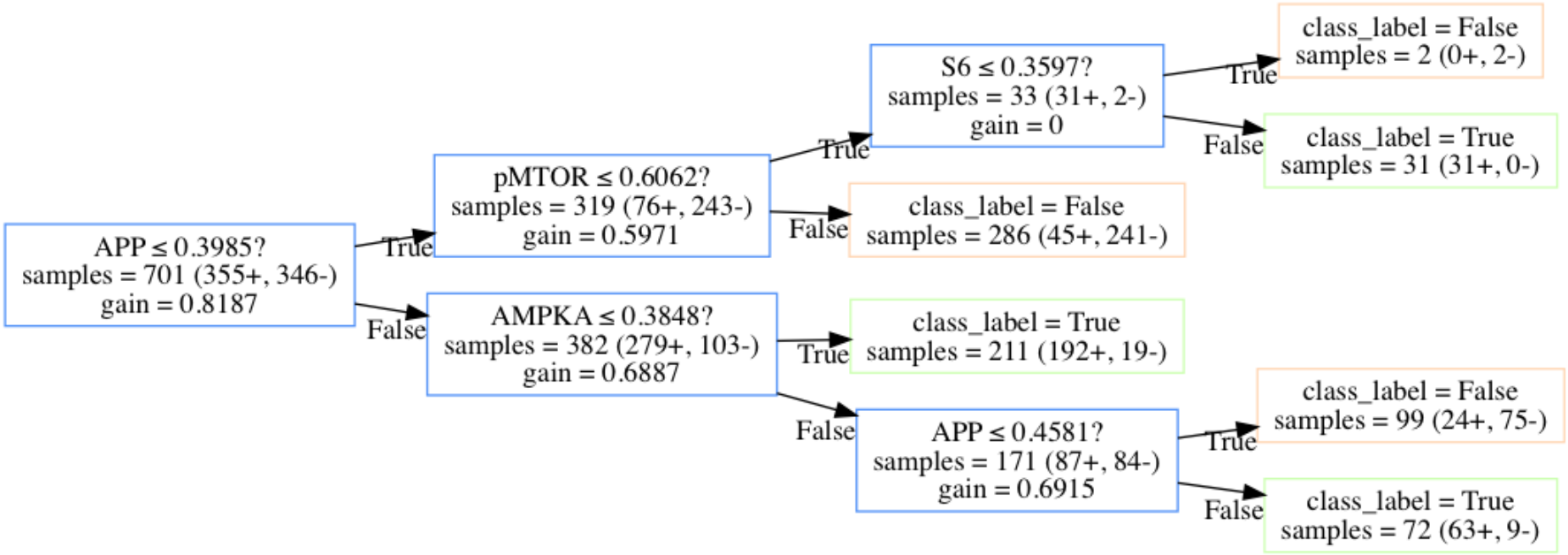
Other observations:

- The results of pruning by accuracy and pessimistic error are identical
- The accuracy of the classification rules gets worse after pruning, but is also slightly worse without pruning

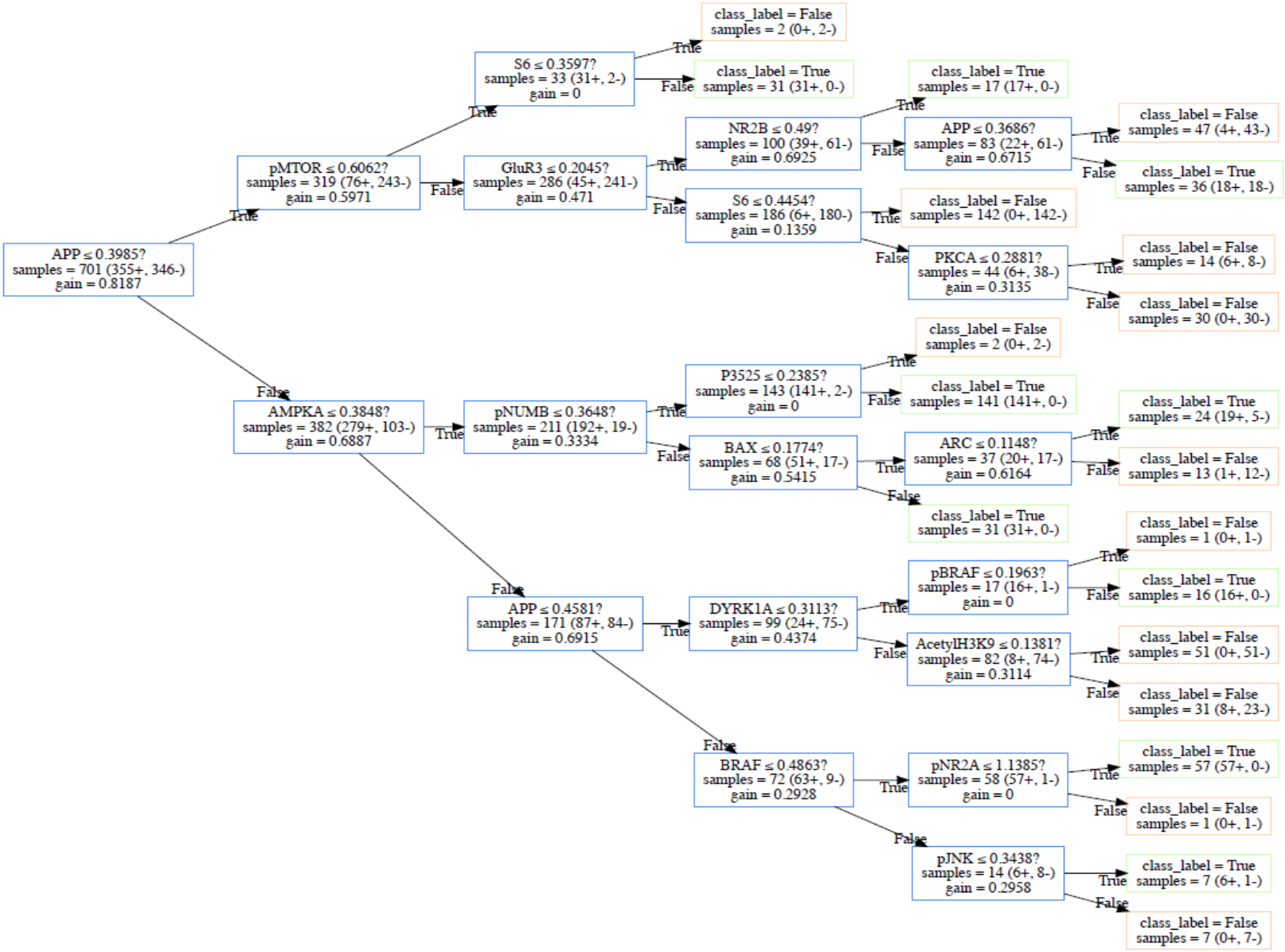
Maximum depth 3



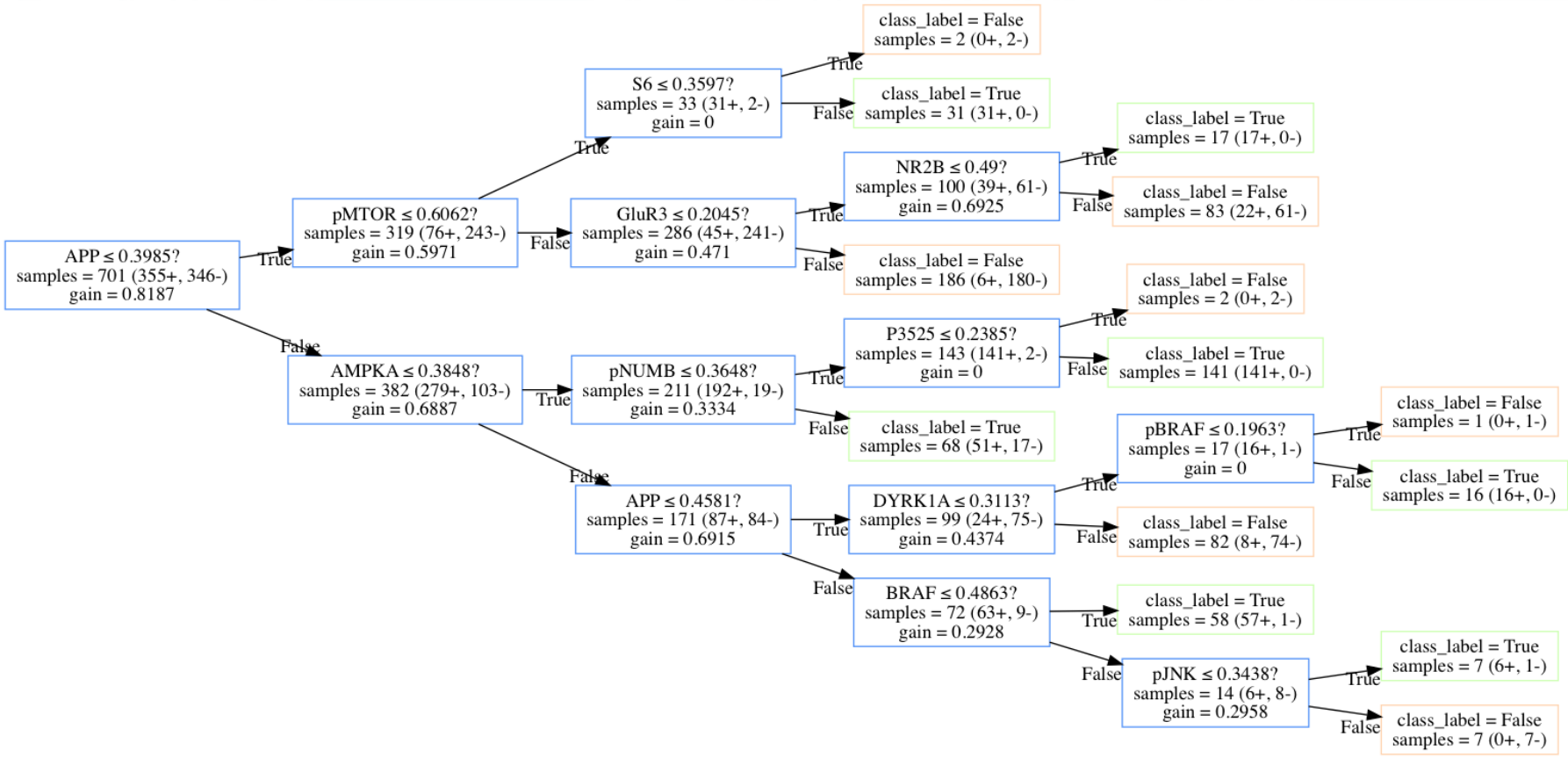
Maximum depth 3 pruned



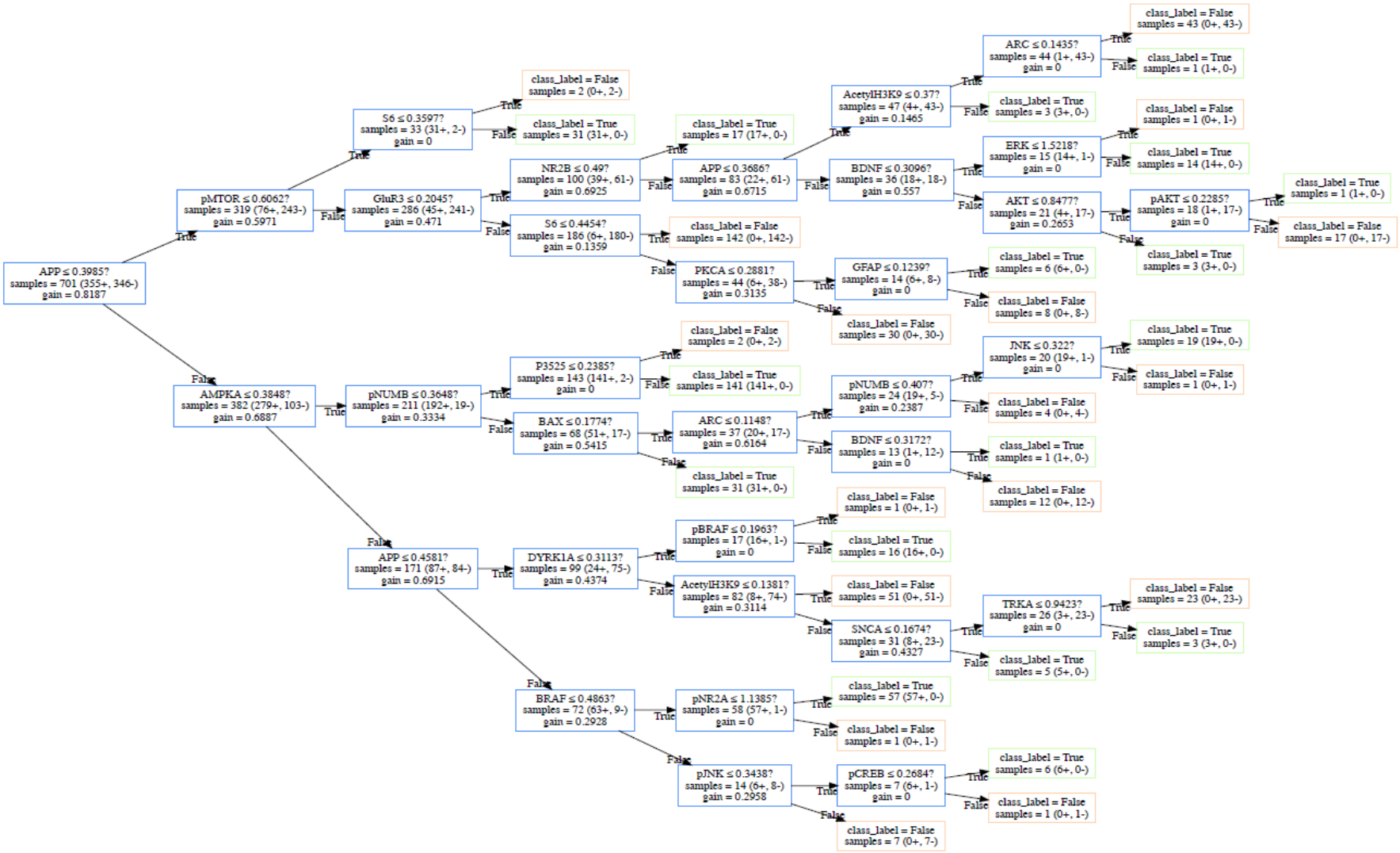
Maximum depth 5



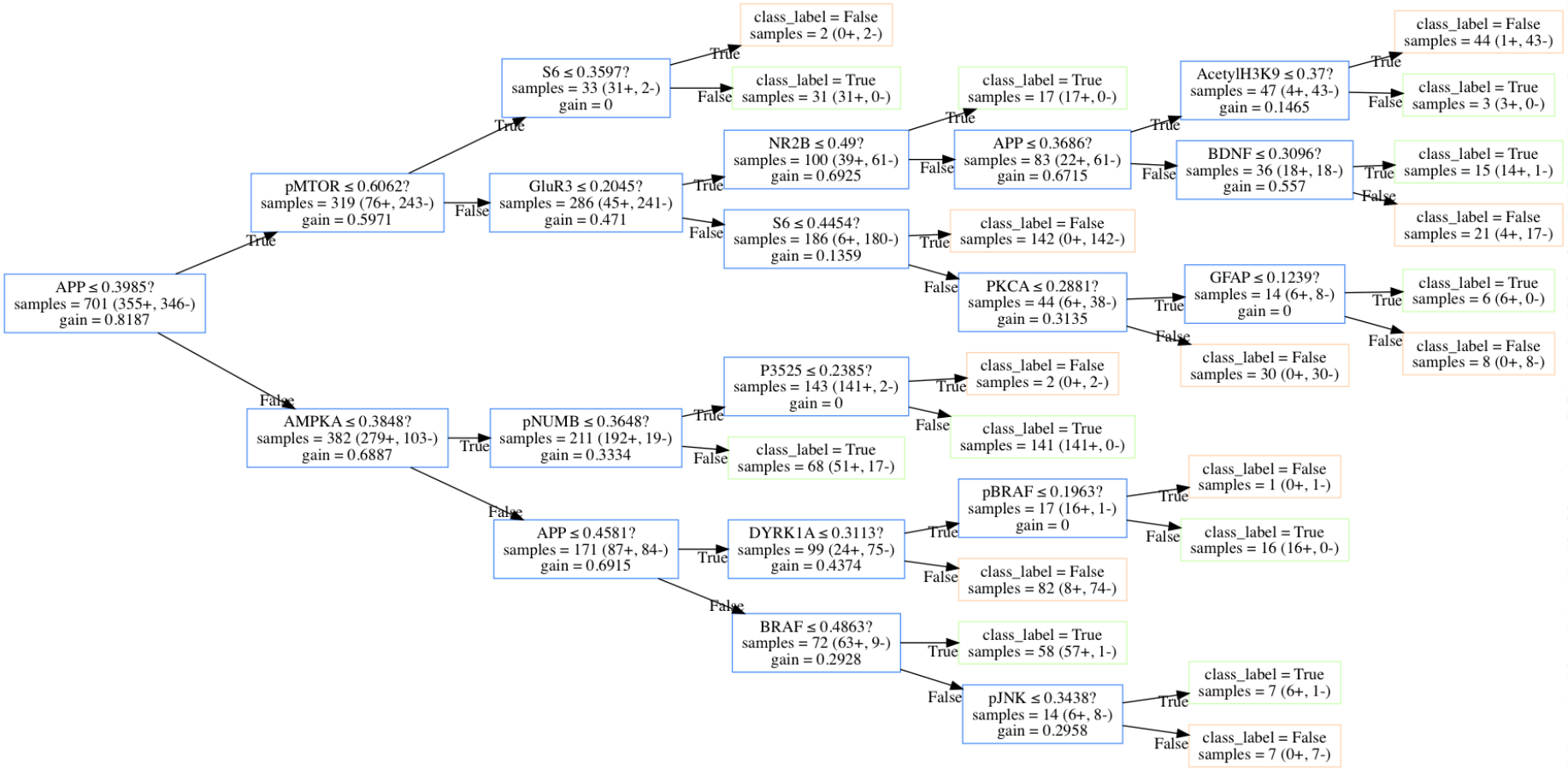
Maximum depth 5 pruned



Maximum depth 10



Maximum depth 10 pruned



Thank you!