# MLCourse-LU

s2538334

Assignment 2

1. Splitting for "airborne" on the place where "feathers" is now would lead to two subtrees, as birds still need to be split off from bugs, which would lead to extra uncertainty and thereby making the algorithm less accurate. Having only one class split off increases the accuracy as there is not any confusion in the split-off leaf and therefore makes the model better.

2. ...By checking the actual classes of the instances against the leaf in which they ended up being classified

3. ... The training and test sets are split different every time, and as the model needs the training set to do predictions, this also leads to different predictions and therefore different trees every time

4. ... See below for the tree. Apparently this particular training set has led the model to believe that half of the mammals dont give milk, for example.

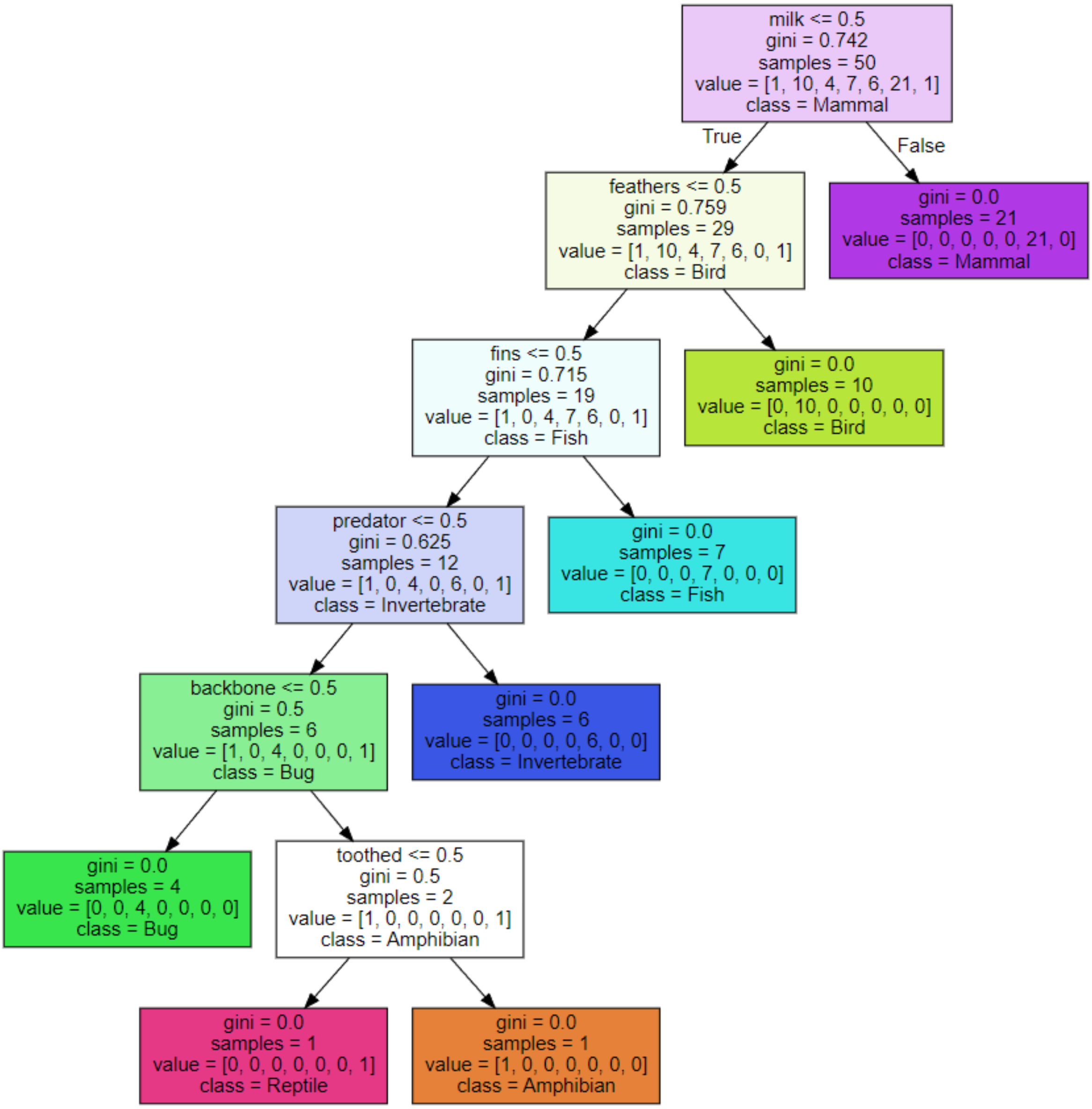

Figure 1: An example of a bad tree (accuracyscore = 0.80)

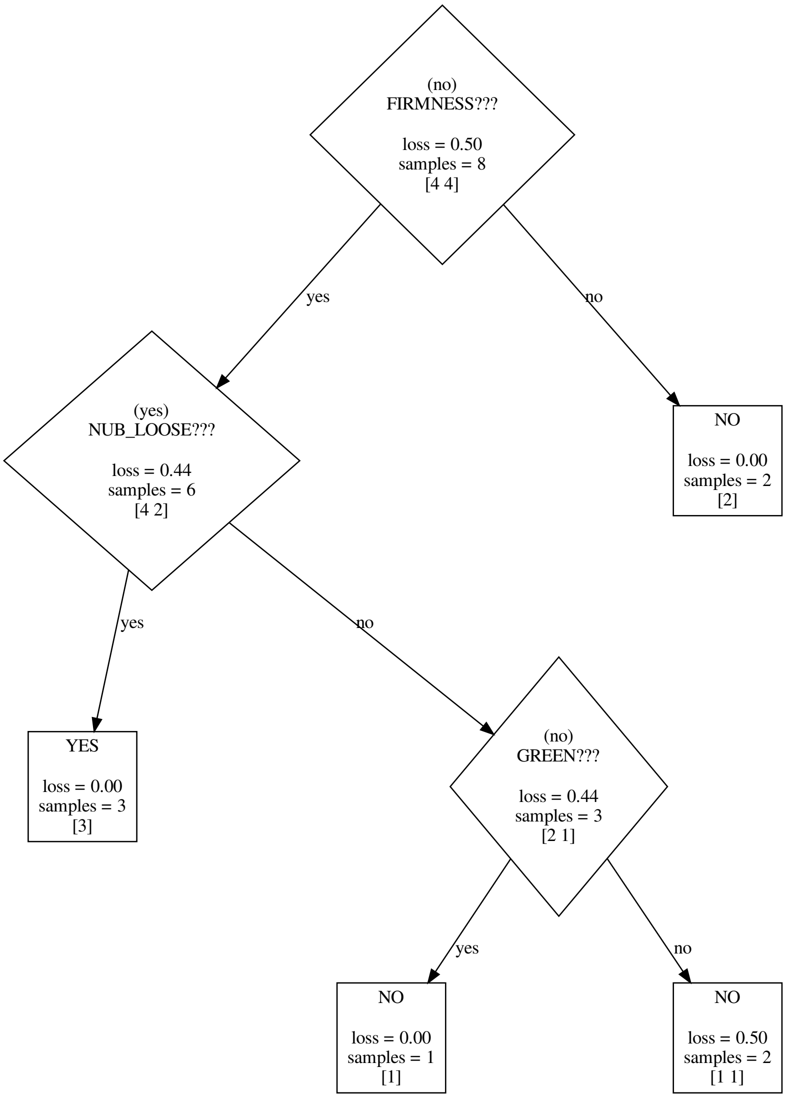

Figure 2: This figure represents a part of the decision tree for avocados. First it is checked for firmness; if yes, samples go into the NUBLOOSE node; if no, samples go into the NO node. Samples represent how many avocados there are in a node, while loss is a measure for the impurity in a node.

5. ...

| **Model + Loss** | **Mean** | **Standard Deviation** |
|---|---|---|
| Scikit-Learn + Gini | 0.93 | 0.04 |
| Scikit-Learn + Entropy | 0.93 | 0.05 |
| ? | ? | ? |
| ? | ? | ? |
| ? | ? | ? |

Table 1: Which of these/your setups works best? It seems like the Scikit-Learn set works slightly better, but both are really close in performance.

6. ... I would rather have a somewhat lower accuracy model with lower standard deviation, as such a model is more reliable and less unpredictable than a high-accuracy model with a large standard deviation, making a lower accuracy model more suitable to do predictions with.

7. ...

   For larger (up to k=100) values of k the mean converges to 0.95, while the standard deviation increases too, to about 0.23 for entropy and 0.22 for Gini. The increase in standard deviation is because the training size decreases with higher k, which means there are a number of really bad models in the set with a low accuracy number, dragging the standard deviation down. The mean becomes somewhat higher, because most of the models are still quite good.

   For smaller values of k (k=3) the accuracy swings between about 0.93 and 0.97, while the standard deviation swings between 0.02 and 0.07. The training sizes are quite large, but as there are only three sets, the mean and standard deviation are quite dependent on chance.