Introduction to Machine Learning

Gian-Luca Fenocchi, Szymon Kubica, Karim Selih, Devin Thomas

November 3, 2022

Contents

Bonus - Tree Visualisation **2**

Unpruned Clean Dataset . . . 2

Pruned Clean Dataset . . . 2

Unpruned Noisy Dataset . . . 3

Pruned Noisy Dataset . . . 3

Step 3 - Evaluation **4**

Cross Validation Classification Metrics . . . 4

Result Analysis . . . 5

Dataset Differences **5**

Step 4 - Pruning **5**

Cross validation classification metrics after pruning . . . 5

Result Analysis . . . 7

Depth Analysis . . . 7

Bonus - Tree Visualisation

Unpruned Clean Dataset

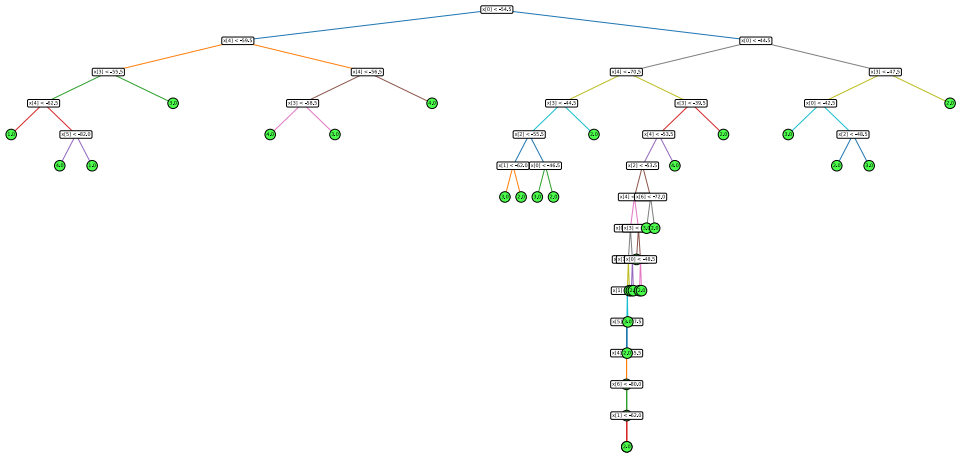

Figure 1: Unpruned tree visualisation function with clean dataset

Pruned Clean Dataset

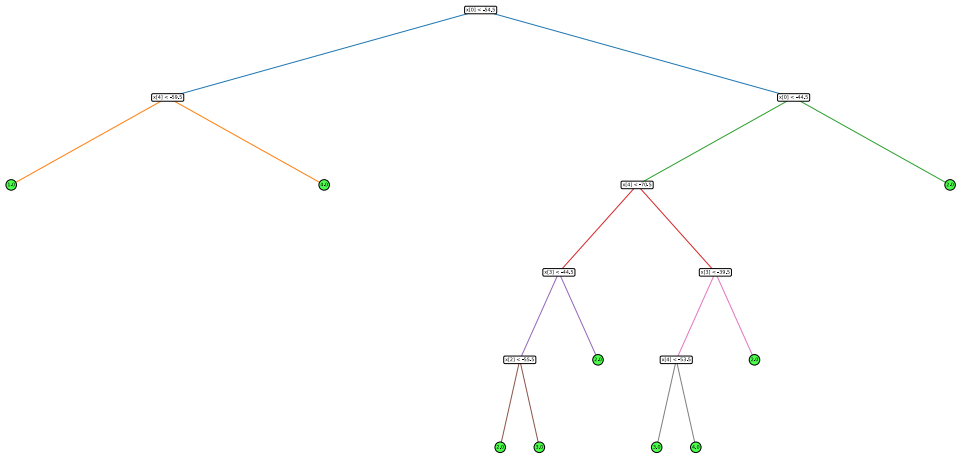

Figure 2: Pruned tree visualisation function with clean dataset

Unpruned Noisy Dataset

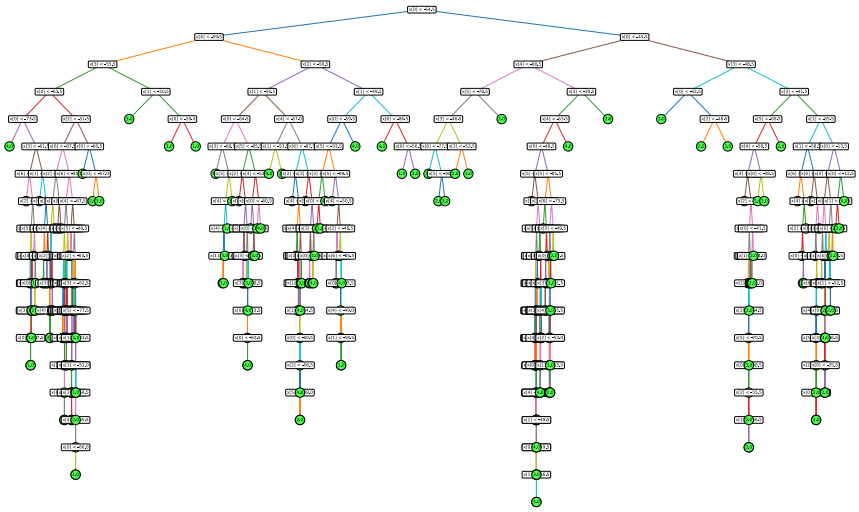

Figure 3: Unpruned tree visualisation function with noisy dataset

Pruned Noisy Dataset

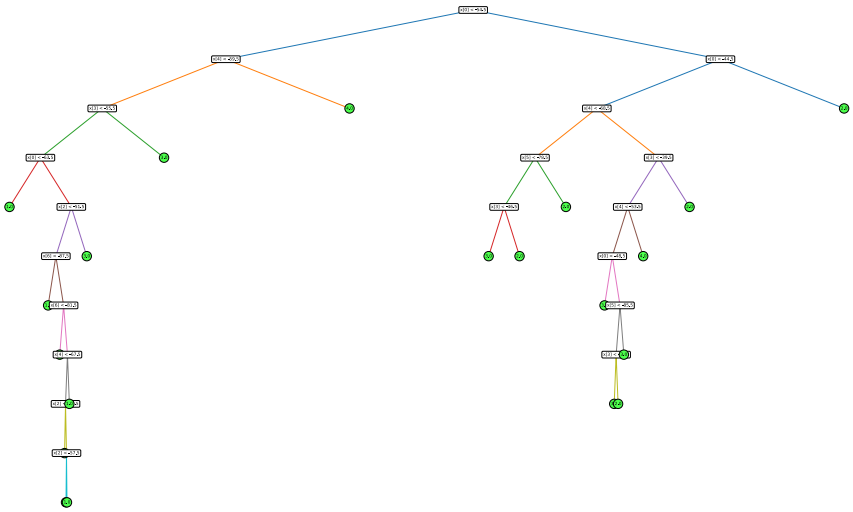

Figure 4: Pruned tree visualisation function with noisy dataset

Step 3 - Evaluation

Cross Validation Classification Metrics

Clean Dataset

The overall confusion matrix obtained by rotating the testing fold 10 times can be seen below. It is a sum of the 10 confusion matrices that we obtained for each tree generated for each fold.

- Confusion matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	496	0	2	2
Room 2 Actual	0	482	18	0
Room 3 Actual	2	19	478	1
Room 4 Actual	4	0	3	493

Table 1: Confusion matrix computed on the clean dataset

- The accuracy: **0.9745**

- Recall, Precision and F1 Per Class

Room	Precision	Recall	F1
1	0.988	0.992	0.990
2	0.962	0.964	0.963
3	0.954	0.956	0.955
4	0.994	0.986	0.990

Table 2: Evaluation metrics computed on the clean dataset

Noisy Dataset

- Confusion matrix

	Room 1 Prediction	Room 2 Prediction	Room 3 Prediction	Room 4 Prediction
Room 1 Actual	403	39	27	29
Room 2 Actual	38	387	31	34
Room 3 Actual	26	23	412	36
Room 4 Actual	37	30	39	409

Table 3: Confusion matrix computed on the noisy dataset

- The accuracy: **0.8055**
- Recall, Precision and F1 Per Class

Room	Precision	Recall	F1
1	0.800	0.809	0.804
2	0.808	0.790	0.799
3	0.809	0.829	0.819
4	0.805	0.794	0.800

Table 4: Evaluation metrics computed on the noisy dataset

Result Analysis

Given the confusion matrix and the evaluation metrics above, we could observe that rooms 1 and 4 tend to get recognised with a very high accuracy, whereas the rooms 2 and 3 noticeably less accurate. It can also be observed that our model tends to misclassify room 3 as room 4 or vice versa.

Dataset Differences

The average accuracy on the noisy dataset is significantly lower than that of the clean detaset. A possible explanation of that behaviour could be that we didn't introduce any cap on the maximum depth of the generated tree. The noisy dataset likely contains outliers from the underlying trend that we want to model with our tree. In that case, our generated trees will likely be over-fitted because they only stop generating new branches when all training data is correctly partitioned.

Step 4 - Pruning

Cross validation classification metrics after pruning

Clean Dataset

- Confusion matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	4485	0	11	4
Room 2 Actual	0	4310	190	0
Room 3 Actual	46	199	4230	25
Room 4 Actual	44	0	30	4426

Table 5: Confusion matrix computed on the clean dataset

- Accuracy
  - Before Pruning: **0.9725**
  - After Pruning: **0.9695**
- Recall, Precision and F1 Per Class

Room	Precision	Recall	F1
1	0.980	0.997	0.988
2	0.956	0.958	0.957
3	0.948	0.940	0.944
4	0.993	0.984	0.988

Table 6: Evaluation metrics computed on the clean dataset with pruning

Noisy Dataset

- Confusion matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	3927	231	150	174
Room 2 Actual	217	3900	126	167
Room 3 Actual	131	175	3919	248
Room 4 Actual	186	179	310	3960

Table 7: Confusion matrix computed on the noisy dataset

- Accuracy
  - Before Pruning: **0.8031**
  - After Pruning: **0.8726**
- Recall, Precision and F1 Per Class

Room	Precision	Recall	F1
1	0.880	0.876	0.880
2	0.870	0.884	0.870
3	0.870	0.876	0.870
4	0.871	0.854	0.871

Table 8: Evaluation metrics computed on the noisy dataset with pruning

Result Analysis

The accuracy of the clean dataset does not change significantly and actually decreases slightly by 0.3%. This would be due to the shuffling and pruning being performed using the validation dataset and thus accuracy is not guaranteed to improve for the test dataset. However, there is a considerable improvement for the noisy dataset by roughly 7%. This is statistically significant to show that the pruned tree does not perform the same as the unpruned tree.

Depth Analysis

For the clean dataset, the average depth for the trees produced was 13.16. However, after pruning, the trees' average depth reduced to 9.8. For the noisy dataset, the average depth for the trees was 19.07. After pruning, this reduced to 14.81. Noise increases the random spread of datapoints, resulting to more decision rules. Both datasets resulted in about a depth average difference of about 5 layers. By reducing the depth of the tree, we begin to increase the accuracy of the model as the tree becomes less overfit to the training data.