Introduction to Machine Learning

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

November 4, 2022

Contents

Bonus - Tree Visualisation

Unpruned Clean Dataset

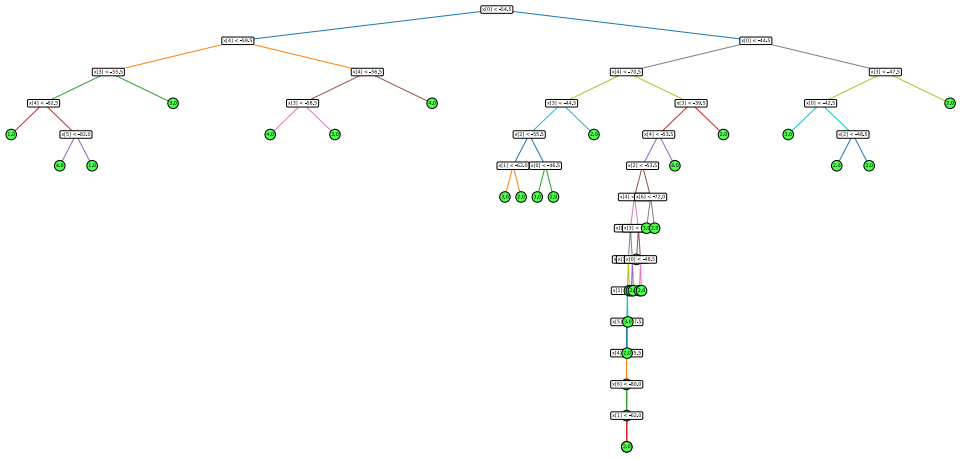

Figure 1: Unpruned tree visualisation function with clean dataset

Pruned Clean Dataset

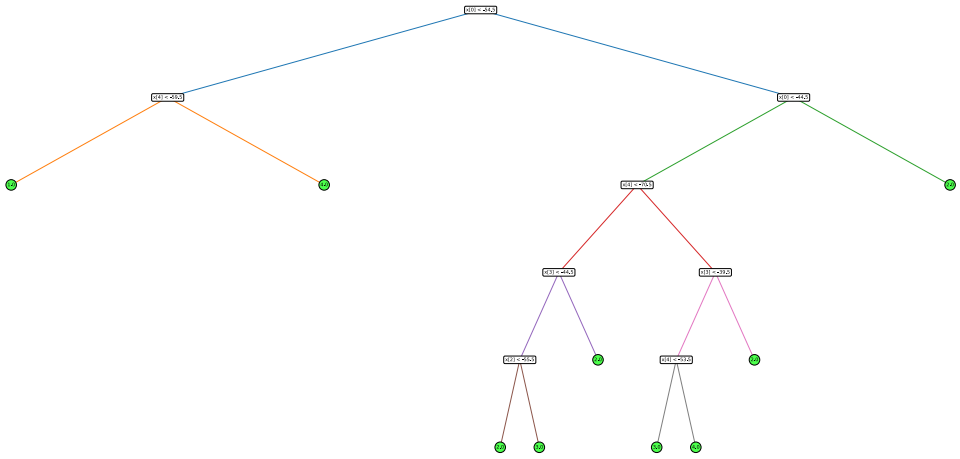

Figure 2: Pruned tree visualisation function with clean dataset

Unpruned Noisy Dataset

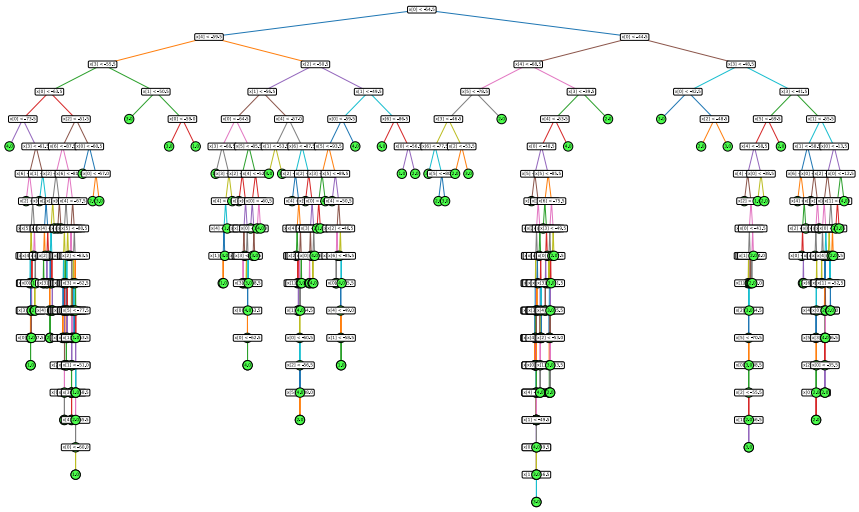

Figure 3: Unpruned tree visualisation function with noisy dataset

Pruned Noisy Dataset

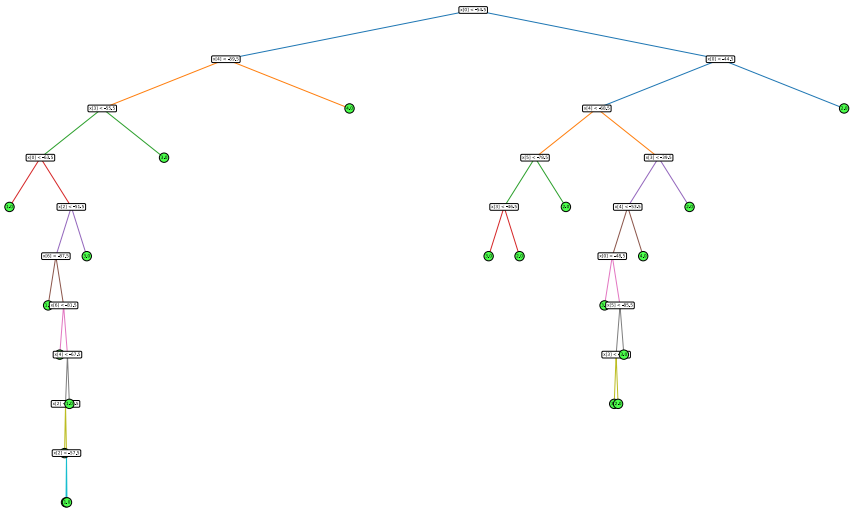

Figure 4: Pruned tree visualisation function with noisy dataset

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	493	0	3	4
Room 2 Actual	0	484	16	0
Room 3 Actual	1	15	481	3
Room 4 Actual	4	0	2	494

Table 1: Confusion matrix computed on the clean dataset

- The accuracy: **0.9760**

- Recall, Precision and F1 Per Class

Room	Precision	Recall	F1
1	0.990	0.986	0.988
2	0.970	0.968	0.969
3	0.958	0.962	0.960
4	0.986	0.988	0.987

Table 2: Evaluation metrics computed on the clean dataset (9.619s)

Noisy Dataset

- Confusion matrix

	Room 1 Prediction	Room 2 Prediction	Room 3 Prediction	Room 4 Prediction
Room 1 Actual	412	35	27	24
Room 2 Actual	41	385	34	30
Room 3 Actual	27	26	407	37
Room 4 Actual	39	23	34	419

Table 3: Confusion matrix computed on the noisy dataset (52.225s)

- The accuracy: **0.8115**
- Recall, Precision and F1 Per Class

Room	Precision	Recall	F1
1	0.794	0.827	0.810
2	0.821	0.786	0.803
3	0.811	0.819	0.815
4	0.822	0.814	0.818

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 and vice versa. Therefore, these rooms are confused.

Dataset Differences

The average accuracy on the noisy dataset is significantly lower than that of the clean dataset. Due to each room having a similar F1 measure in the noisy dataset, it is possible that an even amount of noise was applied to each room. Our model then tries to fit this noisy data causing it to perform less well on unseen data as it can't generalise as well (overfitting). We could combat this by placing a cap on the depth of the tree or by pruning (as seen later).

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	4475	0	19	6
Room 2 Actual	0	4302	198	0
Room 3 Actual	49	210	4216	25
Room 4 Actual	39	0	44	4417

Table 5: Confusion matrix computed on the clean dataset (47.545s)

- Accuracy
  - Before Pruning: **0.9691**
  - After Pruning: **0.9672**
- Recall, Precision and F1 Per Class

Room	Precision	Recall	F1
1	0.981	0.994	0.988
2	0.953	0.956	0.955
3	0.942	0.937	0.939
4	0.993	0.982	0.987

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	3954	219	138	171
Room 2 Actual	227	3907	118	158
Room 3 Actual	113	183	3924	253
Room 4 Actual	177	191	351	3916

Table 7: Confusion matrix computed on the noisy dataset (120.942s)

- Accuracy
  - Before Pruning: **0.8016**
  - After Pruning: **0.8723**
- Recall, Precision and F1 Per Class

Room	Precision	Recall	F1
1	0.884	0.882	0.883
2	0.868	0.886	0.877
3	0.866	0.877	0.872
4	0.871	0.845	0.858

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%. 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. The reason for this is due to pruning reducing the effect of overfitting to the noisy training data allowing the tree to generalise better to unseen data - resulting in an increase in accuracy.

Depth Analysis

For the clean dataset, the average depth for the trees produced was 12.02. However, after pruning, the trees' average depth reduced to 8.44. For the noisy dataset, the average depth for the trees was 17.29. After pruning, this reduced to 13.02. Noise increases the random spread of datapoints, resulting to more decision rules. Both datasets resulted in about a depth average difference of about 4 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. We prevent underfitting by only pruning when the accuracy stays the same or increases.