Introduction to ML - Decision Tree Coursework

Yilun Cheng, Pengfei Xiao, Qiancheng Fan

November 2, 2023

1 Tree visualization

Figure 4 is the big overview of the tree and the figure 1 is a zoomed-in version near the root.

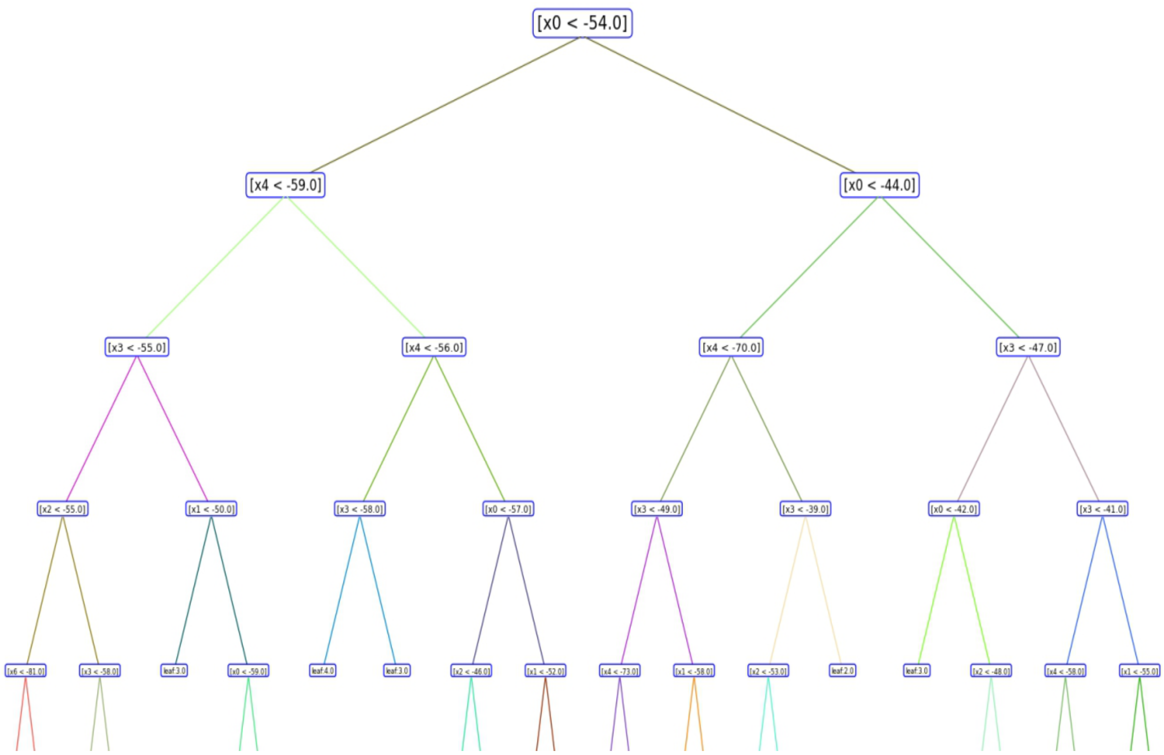
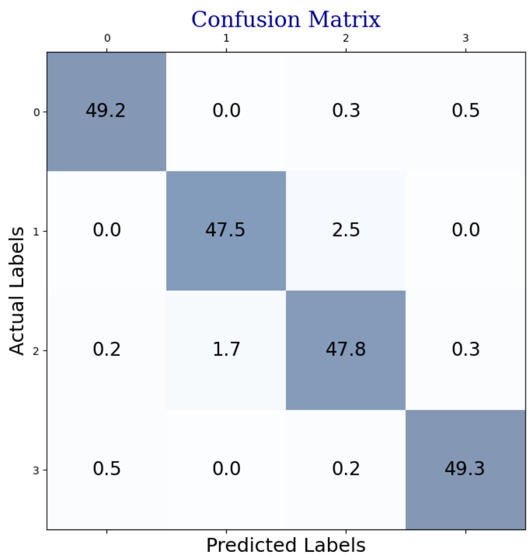


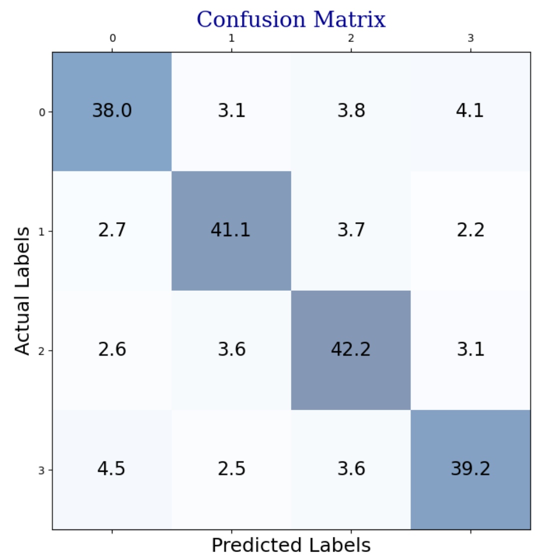
Figure 1: Visualization of the head of output tree trained on the entire clean dataset.

2 Evaluation on the trained trees

2.1 Cross validation classification metrics



(a) Confusion matrix of the tree trained on clean set.



(b) Confusion matrix of the tree trained on noisy set.

Figure 2: The Confusion Matrices of the trees trained on both datasets.

Metric	Clean set	Noisy set
Accuracy Average	0.969	0.802
Recall		
Class 1	0.9842450530863879	0.7759208544095565
Class 2	0.9483285437732877	0.8268539975991661
Class 3	0.9555807214189739	0.8201763449805309
Class 4	0.9862896348060741	0.7897798340117661
Precision		
Class 1	0.9859528734570749	0.7965205115839527
Class 2	0.9661271605408172	0.8173745170301501
Class 3	0.9391069757820033	0.7951296455199124
Class 4	0.9839185394775457	0.8077187206775602
F1-Score		
Class 1	0.9667747871490636	0.8114038571473374
Class 2	0.9571451156491971	0.8220869312614182
Class 3	0.9436952325666926	0.8106815735126127
Class 4	0.9657957761431425	0.8171743546925613

Table 1: Performance of the trees trained on Clean set and Noisy set

2.2 Result analysis

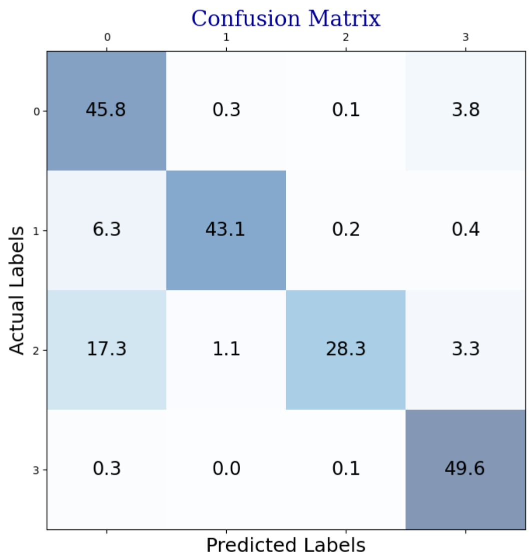
The decision trees trained on both datasets show high accuracy for all four labels. For the clean set, the decision tree holds a high average accuracy of 96.9%. Room 1 and 4 are most likely to be correctly recognized, while instances from Room 2 and 3 are more likely to be confused with others. For the noisy set, Class 1 and 4 hold relatively low Recall values, which means instances in Room 1 and 4 are less likely to be correctly recognized. All four class hold similar precision values and F1 scores.

2.3 Dataset differences

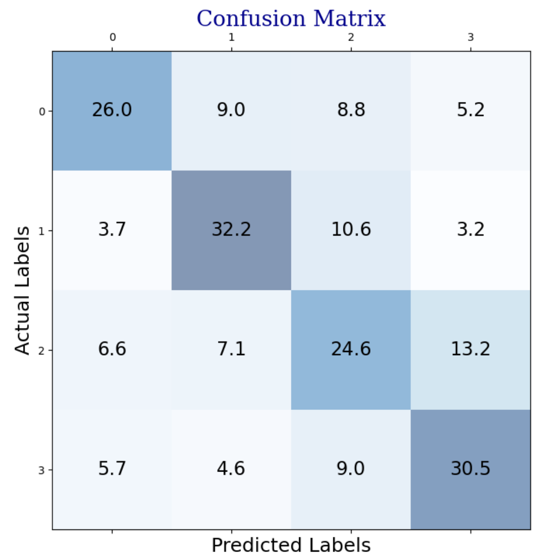
Comparing with the clean set, the overall accuracy of the tree for noisy dataset reduced significantly from 0.969 to 0.802. The F1-score for all rooms dropped. This is because that the noisy dataset contains errors and outliers that can result in wrong thresholds. Noise can introduce incorrect splits in the tree and make it harder to extract meaningful patterns. Also, noisy data can lead to overfitting where noise are fitted, resulting in complex boundaries, making it harder to recognize test cases correctly.

3 Pruning (and evaluation again)

3.1 Cross validation classification metrics after pruning



(a) Confusion matrix after pruning trained on clean set.



(b) Confusion matrix after pruning trained on noisy set.

Figure 3: The Confusion Matrices after pruning trained on both datasets.

Metric	Clean set	Noisy set
Accuracy Average	0.834	0.5665
Recall		
Class 1	0.9184175299421846	0.5347685619748843
Class 2	0.8651642379359771	0.6500926532192915
Class 3	0.5627069337135867	0.47562488502951894
Class 4	0.992112561375371	0.6143696939156706
Precision		
Class 1	0.6647133077235937	0.6324435102014889
Class 2	0.9698408362679634	0.6186547780577631
Class 3	0.982197657152285	0.4621477263526103
Class 4	0.878204047131878	0.5816512654543049
F1-Score		
Class 1	0.7518068148057572	0.6411466456611985
Class 2	0.9145169349386876	0.6339842212559941
Class 3	0.919973817595777	0.5402408456335125
Class 4	0.8716353758374923	0.6139704993465426

Table 2: Performance after pruning on Clean set and Noisy set

3.2 Result analysis after pruning

Post-pruning resulted in decreased performance in both datasets, with accuracy decreasing to 83.4% in the clean dataset and dropping to 56.7% in the noisy dataset. This performance difference after pruning is due to the less distinct features in the noisy dataset compared to the clean dataset. Consequently, the more significant decline in performance after pruning in the noisy dataset means reduced overfitting risk following pruning, contributing to enhanced generalization ability across different datasets.

3.3 Depth analysis

On the clean-set, the average depth decreased from 5.27 to 4.34 after pruning, while on the noisy-set, it reduced from 7.06 to 6.64, with more complex pruning conditions on the noisy-set.The relationship between maximal depth and prediction accuracy is that increasing the maximal depth of decision trees enhances accuracy but can lead to overfitting. Conversely, reducing tree depth through pruning lowers prediction accuracy, especially in complex datasets, yet it improves generalization ability.

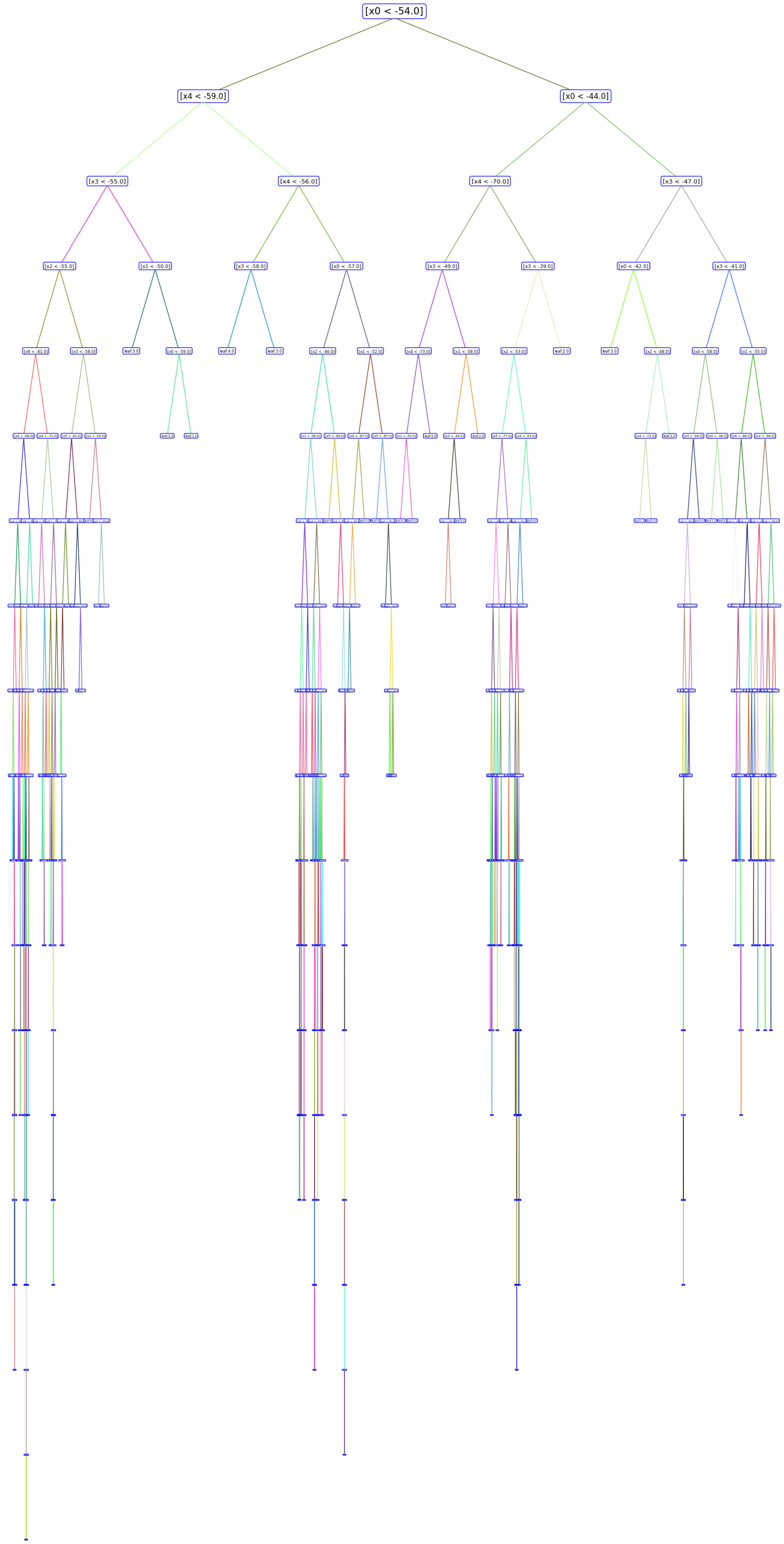


Figure 4: Visualization of the entire output tree trained on the entire clean dataset.