Machine Learning Coursework 1
Decision Trees

Indraneel Dulange, Omar Zeidan
Shaheen Amin, Shaanuka Gunaratne

November 2022

1. Cross Validation Classification Metrics

Clean Dataset

Confusion Matrix

Predicted / Actual	Room 1	Room 2	Room 3	Room 4
Room 1	49.4	0.0	0.5	0.1
Room 2	0.0	47.3	2.7	0.0
Room 3	0.7	1.9	46.9	0.5
Room 4	0.4	0.0	0.5	49.1

Metrics

	Room 1	Room 2	Room 3	Room 4
Precision	0.9772	0.9600	0.9282	0.9871
Recall	0.9873	0.9469	0.9388	0.9833
F1	0.9821	0.9529	0.9329	0.9850
Accuracy	0.9635			

Noisy Dataset

Confusion Matrix

Predicted / Actual	Room 1	Room 2	Room 3	Room 4
Room 1	39.4	2.6	2.6	4.4
Room 2	2.1	41.4	3.0	3.2
Room 3	2.9	4.6	40.9	3.1
Room 4	4.0	3.0	2.8	40.0

Metrics

	Room 1	Room 2	Room 3	Room 4
Precision	0.8142	0.8053	0.8241	0.7891
Recall	0.8034	0.8362	0.7947	0.8067
F1	0.8077	0.8184	0.8053	0.7963
Accuracy	0.8085			

2. Result Analysis

Room 1 has the most accurate readings for the clean dataset and Room 2 for the noisy dataset. While Room 3 has the lowest accuracy for the clean dataset and Room 1 for the noisy dataset. Rooms 2 and 3 are often confused between each other in the clean dataset and Rooms 1 and 4 confuse each other in the noisy dataset.

3. Dataset Differences

The accuracy and other metrics decrease by an average of 16% when using the noisy dataset as compared to the clean dataset – the model overfits to the noisy training set as there is more variance in the data, leading to poorer performance in unseen data tests. In Rooms 1 and 4 performed worse on average.

4. Cross Validation Classification Metrics After Pruning

Clean Dataset

Confusion Matrix

Predicted Actual	Room 1	Room 2	Room 3	Room 4
Room 1	49.9	0.0	0.1	0.0
Room 2	0.0	44.7	5.3	0.0
Room 3	0.7	1.3	47.6	0.4
Room 4	0.4	0.0	1.0	48.6

Metrics

	Room 1	Room 2	Room 3	Room 4
Precision	0.9775	0.9722	0.8846	0.9908
Recall	0.9968	0.8988	0.9514	09744
F1	0.9869	0.9327	0.9152	0.9823
Accuracy	0.9540			

Noisy Dataset

Confusion Matrix

Predicted Actual	Room 1	Room 2	Room 3	Room 4
Room 1	44.1	0.9	1.6	2.4
Room 2	1.9	41.4	5.2	1.2
Room 3	2.1	4.5	42.0	2.9
Room 4	2.5	1.4	2.3	43.6

Metrics

	Room 1	Room 2	Room 3	Room 4
Precision	0.8721	0.8679	0.8219	0.8711
Recall	0.8992	0.8401	0.8215	0.8778
F1	0.8837	0.8469	0.8158	0.8740
Accuracy	0.8555			

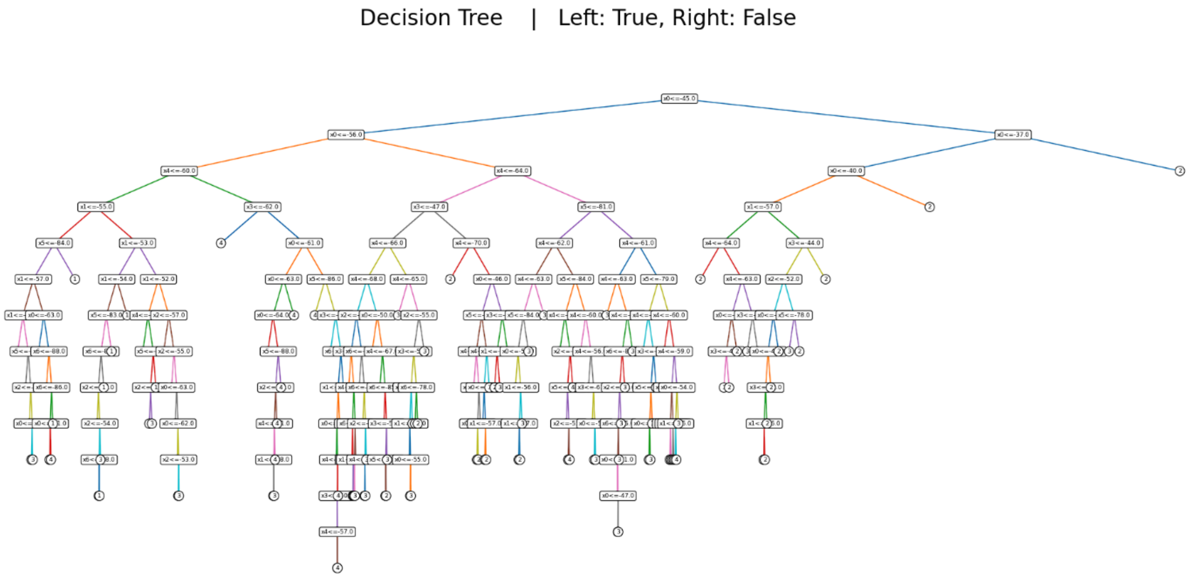
5. Result Analysis After Pruning

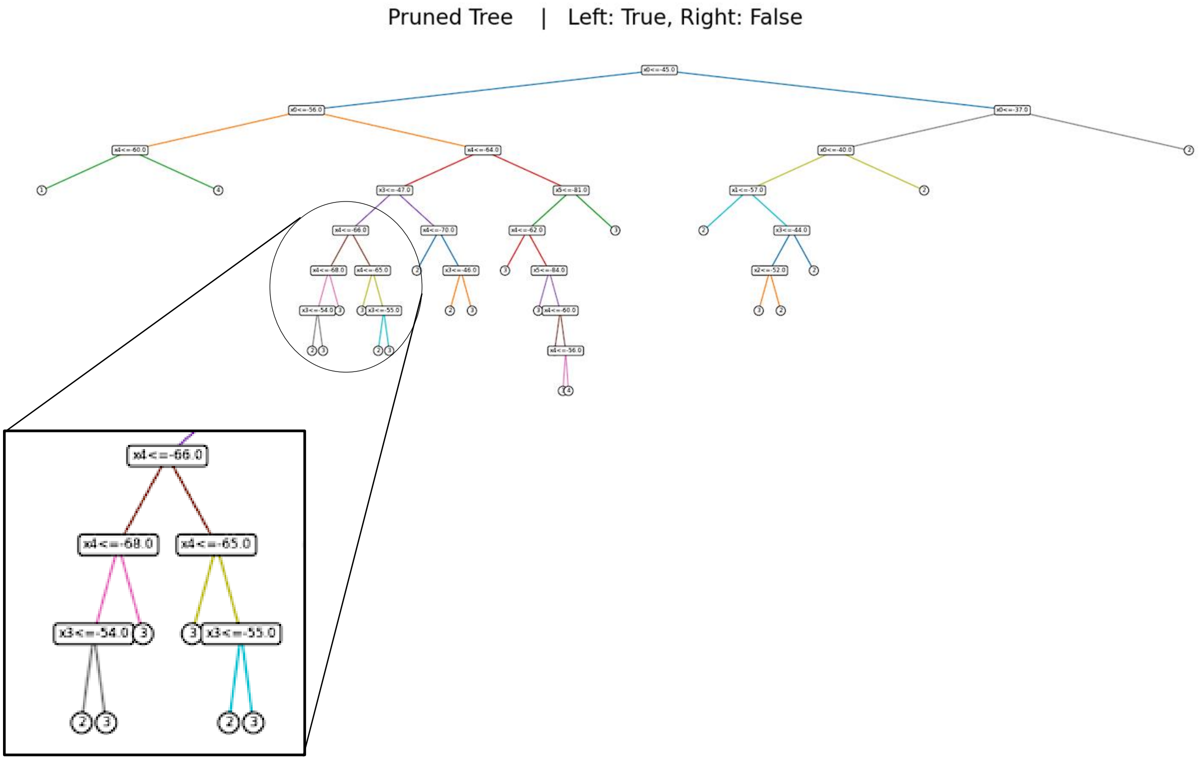
The metrics for the clean dataset do not change significantly after pruning but they do have a strong increase for the noisy dataset by an average of 5% across all metrics. This is because the decision tree is overfitted for the noisy dataset but pruning removes the specific decision nodes which helps to generalize the tree to work better with noisy data.

6. Depth Analysis

For the clean dataset the average depth for unpruned was 13 and pruned was 9.9, while for the noisy dataset the average depth for unpruned was 15 and pruned was 10.2. Pruning trees trained on the clean dataset show that the maximal depth is in itself not a precise indicator for the prediction accuracy.

7. Output of the Tree Visualisation Function





Zoomed in branch of pruned tree