Report for Coursework 1

Output of the Tree Visualisation Function

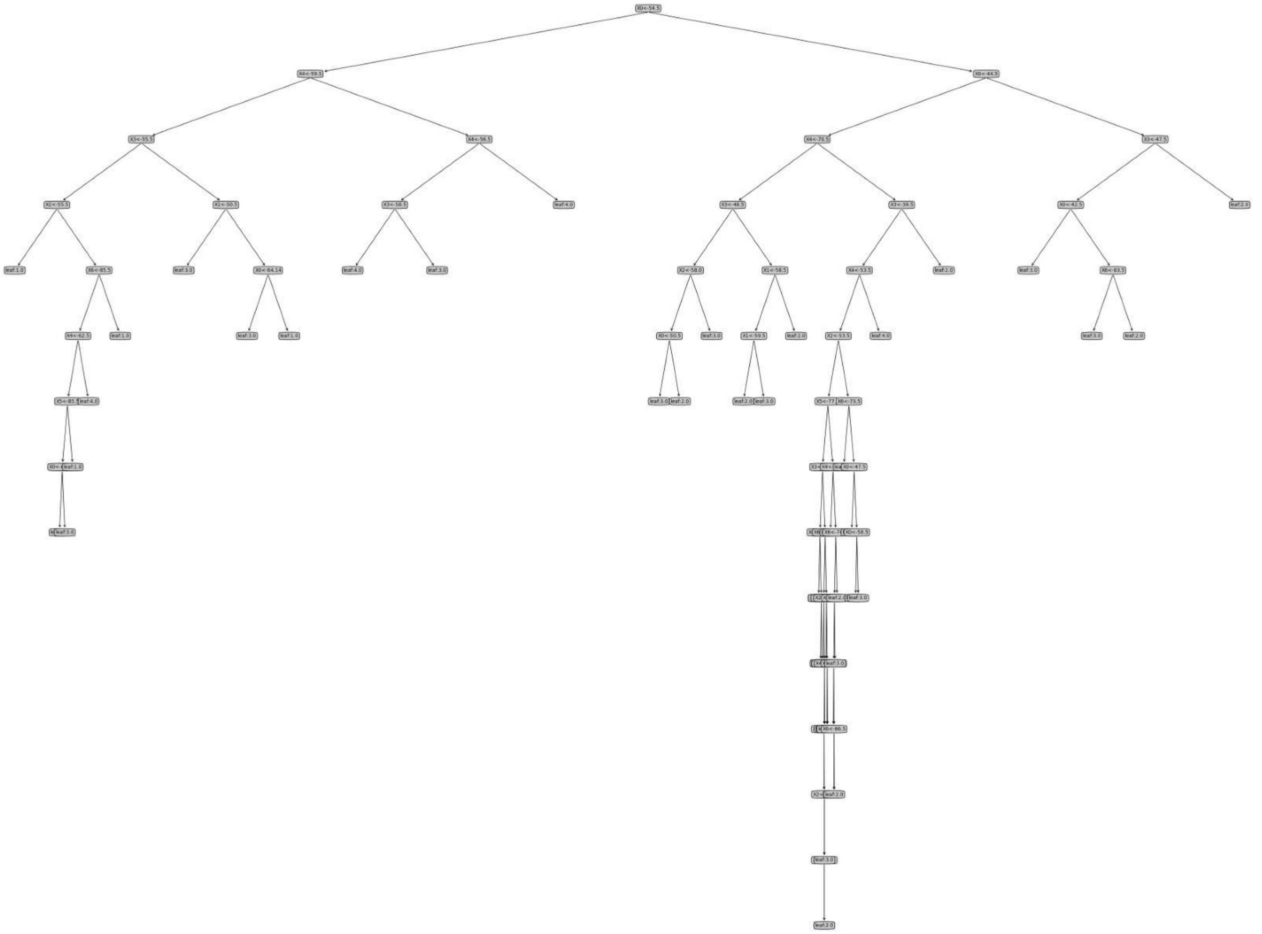


Figure 1: Overview of the decision tree model trained on entire clean dataset

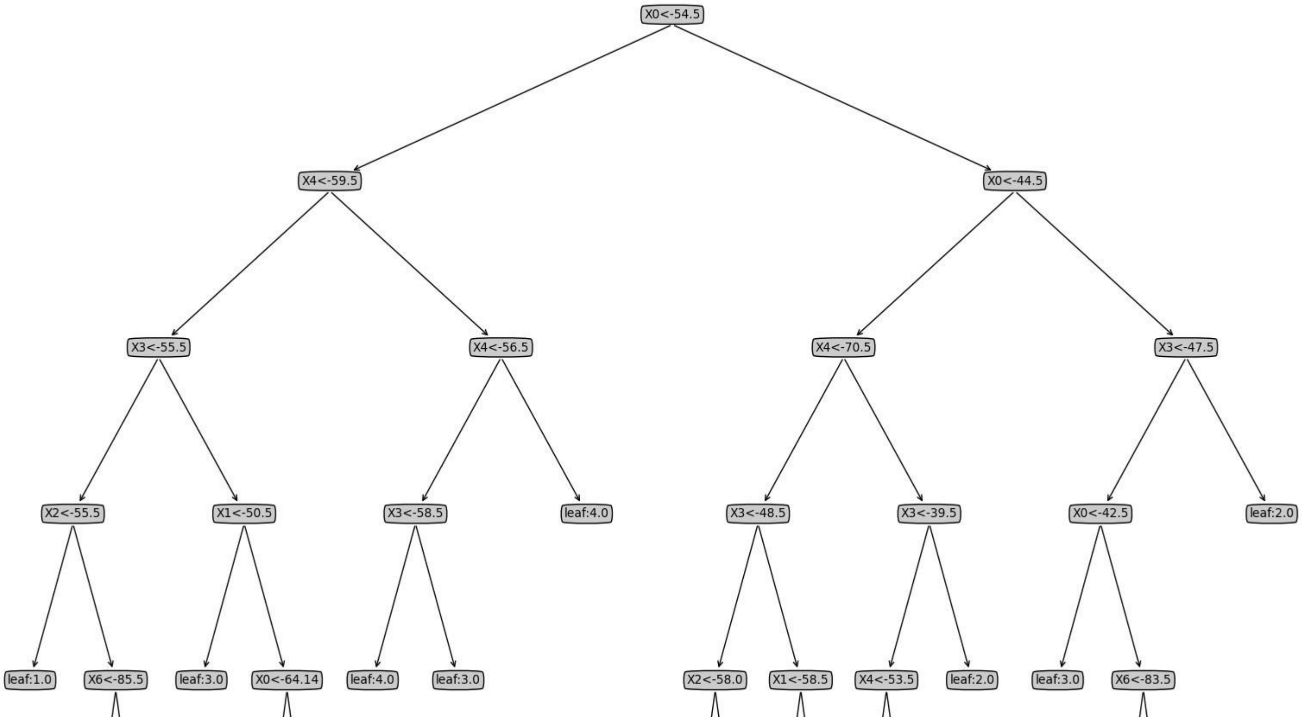
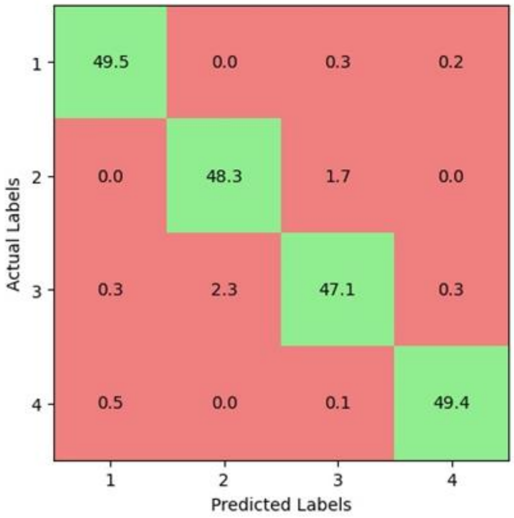


Figure 2: Close up view of the decision tree

Step 3 – Evaluation

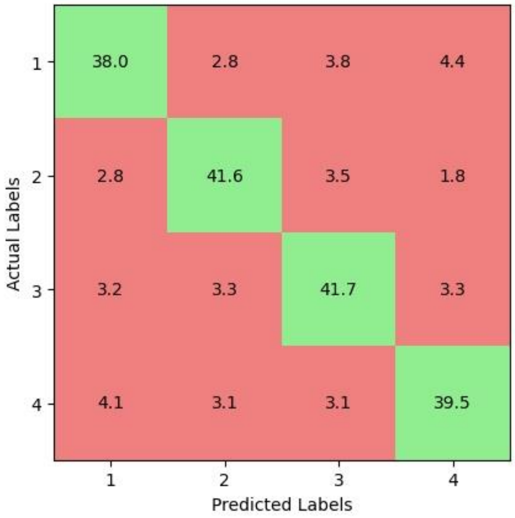
Cross validation classification metrics (seed = 1)

- Clean dataset



	Room 1	**Room 2**	**Room 3**	**Room 4**
Accuracy	0.972			
Precision	0.985	0.954	0.957	0.989
Recall	0.990	0.964	0.942	0.988
F1-score	0.987	0.959	0.949	0.989

- Noisy dataset



	Room 1	**Room 2**	**Room 3**	**Room 4**
Accuracy	0.804			
Precision	0.786	0.818	0.805	0.808
Recall	0.773	0.837	0.808	0.791
F1-score	0.778	0.826	0.804	0.797

Result analysis

For clean dataset, Room 1 and 4 are the most accurately predicted rooms with the highest overall recall and precision. Room 3 and Room 2 are confused as they are commonly mislabelled as each other.

For noisy dataset, Room 2 is the most accurately predicted room. Room 1 is the least accurately predicted room with the largest number of false positives. Room 4 is the most confused one and is commonly mislabelled as Room 1.

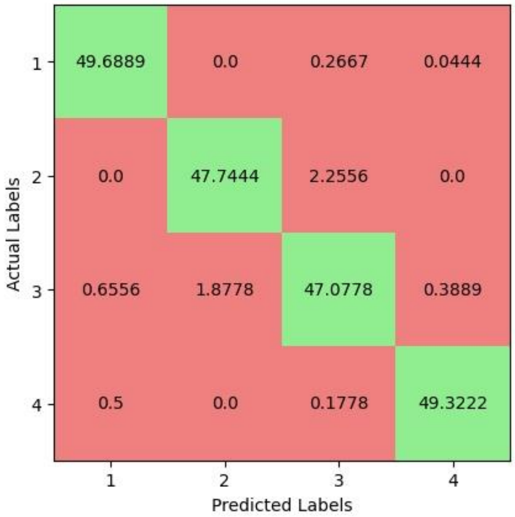
Dataset differences

Yes, the clean dataset has better performance than the noisy dataset. There are significantly more false positives and negatives in the model trained by the noisy dataset and result in an accuracy decrease of 16.8%. That is because the noisy dataset contains a lot more noises causing the trained model to be overfitted, hence the model has worse performance when testing with unknown dataset.

Step 4 - Pruning

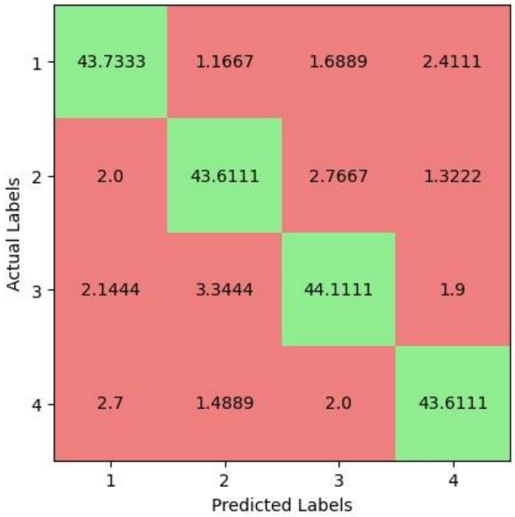
Cross validation classification metrics after pruning (seed = 1)

- Clean dataset



	Room 1	Room 2	Room 3	Room 4
Accuracy	0.969			
Precision	0.978	0.962	0.947	0.992
Recall	0.994	0.954	0.941	0.987
F1-score	0.986	0.958	0.943	0.989

- Noisy dataset



	Room 1	Room 2	Room 3	Room 4
Accuracy	0.875			
Precision	0.863	0.880	0.872	0.885
Recall	0.892	0.877	0.854	0.875
F1-score	0.877	0.878	0.862	0.879

Result analysis

For clean dataset, as the trained tree before pruning already provides accurate predictions, the overall performance decreases slightly after pruning due to higher bias.

For noisy dataset, accuracy increases from 80.4% to 87.5% after pruning. This performance increase is due to the reduced variance which fixes the overfitting problem of the unpruned tree, and such simpler model is less influenced by the fluctuations of noise in the training dataset.

Depth analysis

For the clean dataset, the average tree depth before and after pruning is 13.30 and 11.24 respectively. While for the noisy dataset, the average tree depth after pruning decreases from 19.90 to 17.42. Hence, pruning would reduce the average depth of the decision tree, which reduces the variance and increases the bias. For an overfitted model, lower tree depth would give higher accuracy. However, when the tree depth becomes too low, the accuracy would decrease instead as model becomes too simple and hence has low variance and high bias.