70050 ExerciseTypes.CW1

Decision Trees

Submitters

---

cf23	Chen Fan
ky523	Kangle Yuan
vm23	Vinayak Modi
glg23	Gordian Gruentuch

96/100

Excellent work. When a question says 'Analyze' try to answer why and how you see these results rather than simply stating the results.

Code: 18

You should be able to run the scripts for a single dataset.

Emarking

Report – Decision Tree Coursework COMP70050*

1. Output of the tree visualisation function

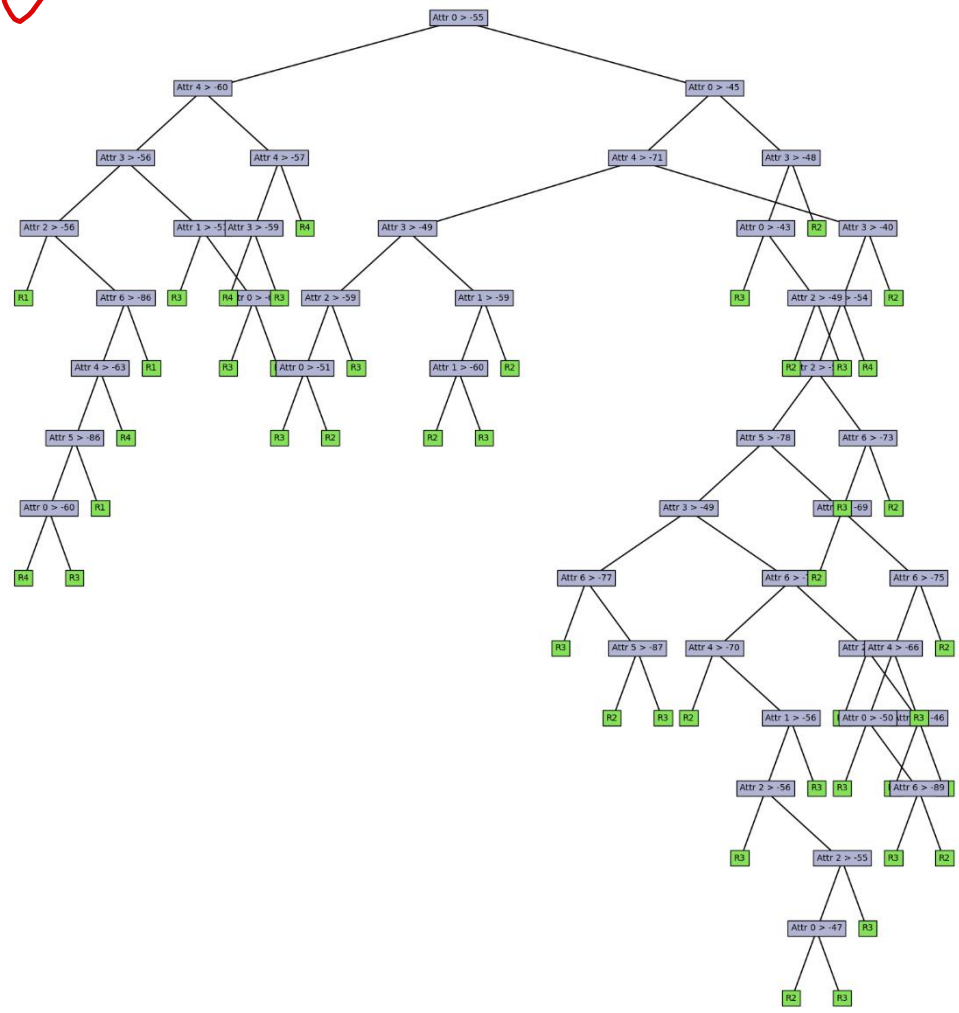


Figure 1 Tree Visualization of the Entire Clean Dataset

* The values and metrics generated in this report are subject to statistical fluctuations resulting from the sampling of the 10 folds.

2. Step 3 - Evaluation +14

1) Cross validation classification metrics

For a visual representation of these metrics, readers are referred to Figures 2 through 6.

- Confusion Matrix ✓ ✓

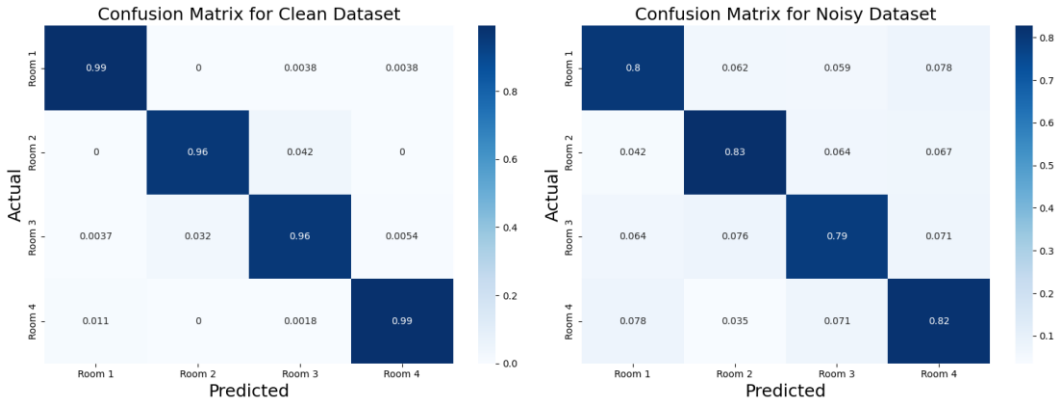


Figure 2 Confusion Matrices of Clean and Noisy Dataset

- The Accuracy

Metric	Clean Dataset	Noisy Dataset
Accuracy	0.9740760747988152	0.8079478125754997

too many dec. points!

Figure 3 Accuracy Comparison Between Clean and Noisy Dataset

- The Recall and Precision Rates Per Class & Macro

Metric	Clean Dataset	Noisy Dataset
Precision (Room 1)	0.98505944	0.81327046
Precision (Room 2)	0.967879	0.82610969
Precision (Room 3)	0.9526365	0.80248224
Precision (Room 4)	0.99079542	0.79044156
Macro-Precision	0.9740925908441165	0.8080759894358382

✓

Figure 4 Precision Comparison Between Clean and Noisy Dataset

* The values and metrics generated in this report are subject to statistical fluctuations resulting from the sampling of the 10 folds.

Metric	Clean Dataset	Noisy Dataset
Recall (Room 1)	0.99241894	0.80041692
Recall (Room 2)	0.95790638	0.82772293
Recall (Room 3)	0.9591107	0.78771608
Recall (Room 4)	0.98686827	0.81593531
Macro-Recall	0.9740760747988151	0.8079478125754997

Figure 5 Recall Comparison Between Clean and Noisy Dataset

- The F1-Measures Derived from The Recall and Precision Rates of the Previous Step

Metric	Clean Dataset	Noisy Dataset
F1 Score (Room 1)	0.98872549	0.8067925
F1 Score (Room 2)	0.96286687	0.82691553
F1 Score (Room 3)	0.95586264	0.79503061
F1 Score (Room 4)	0.98882795	0.80298614
Macro-F1-Score	0.9740707383329147	0.807931192849796

Figure 6 The F1-Score Comparison Between Clean and Noisy Dataset

2) Result analysis

For the clean dataset, the model can most correctly classify rooms 1 and 4, with F1 scores over 98%. However, rooms 2 and 3 are occasionally confused with each other, with rates of 4.3% and 3.2%, respectively.

For the noisy dataset, the model classifies the rooms with an F1 score of about 80% for all classes. The confusion is fairly distributed across all classes, with errors ranging from 3.5% to 7.8%.

3) Dataset differences

Yes, there is a performance difference. The clean dataset, with an accuracy of 97.41%, performs significantly better than the noisy dataset, with an accuracy of 80.79%. This makes the clusters overlap and less distinguishable, interfering with the model's ability to make precise classifications. The training and test data differ more in the case of the noisy compared to the clean data. Therefore, the inevitable noise reduces the classifier's accuracy in noisy dataset.

* The values and metrics generated in this report are subject to statistical fluctuations resulting from the sampling of the 10 folds.

3. Step 4 - Pruning (and evaluation again)

1) Cross validation classification metrics after pruning

- The accuracy

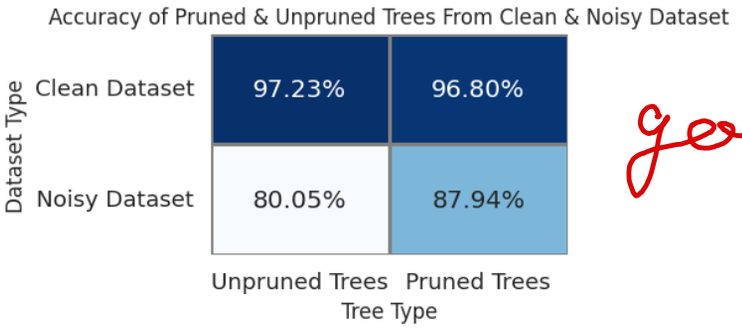


Figure 7 Accuracy Heatmap of Pruned and Unpruned Trees from Clean & Noisy Datasets

- Confusion matrix

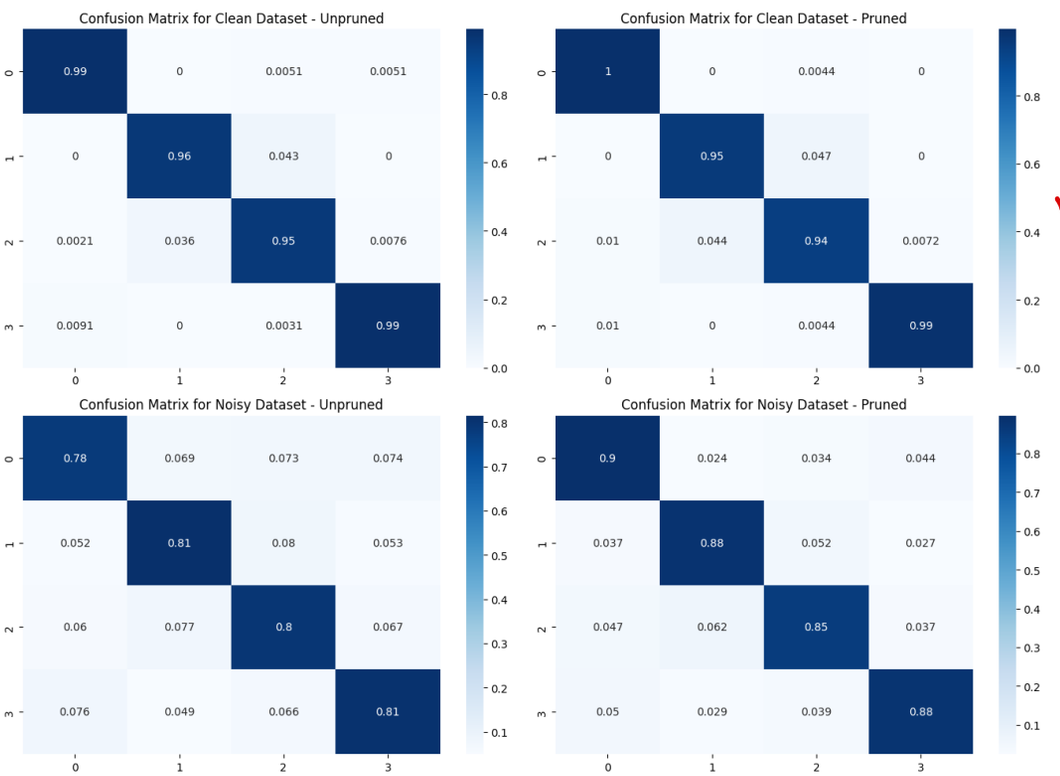
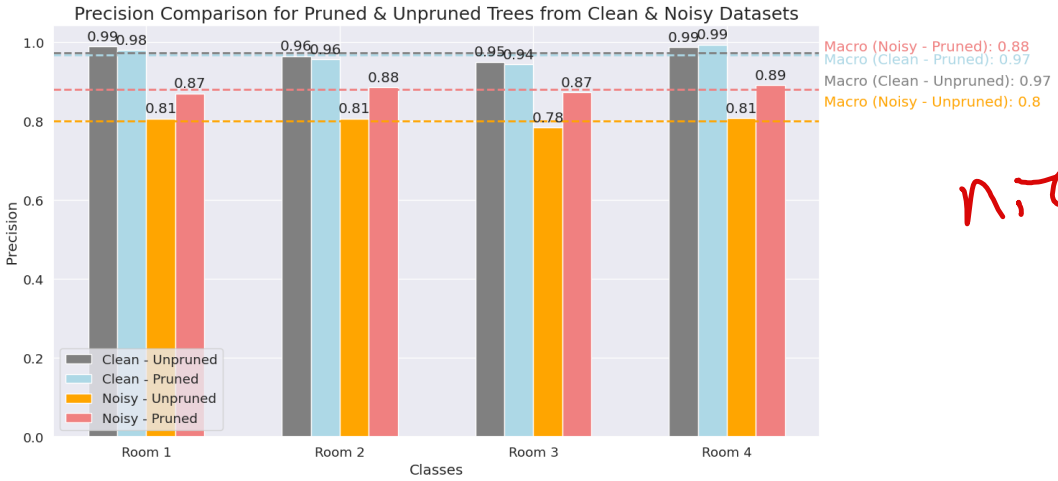


Figure 8 Confusion Matrices for Pruned and Unpruned Trees Generated by Clean and Noisy Datasets

* The values and metrics generated in this report are subject to statistical fluctuations resulting from the sampling of the 10 folds.

- The recall and precision per class



nice!

Figure 7 Precision Comparison of Pruned and Unpruned Trees from Clean & Noisy Datasets

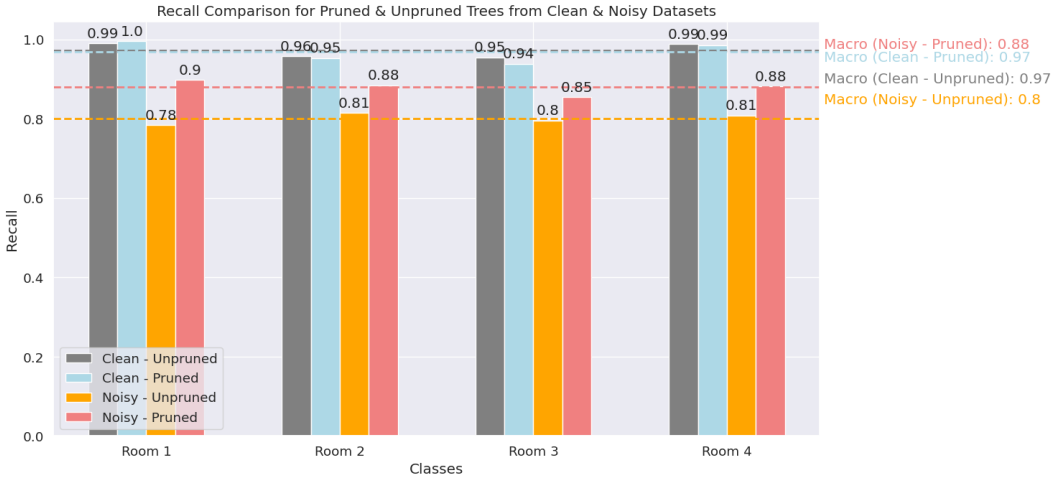


Figure 8 Recall Comparison of Pruned and Unpruned Trees from Clean & Noisy Datasets

- The F1-measures

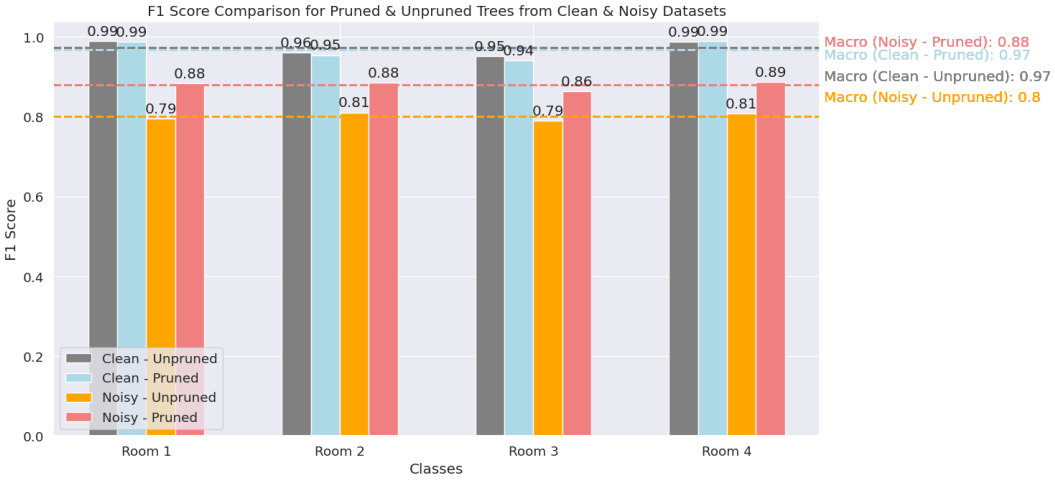


Figure 9 F1-Score Comparison of Pruned and Unpruned Trees from Clean & Noisy Datasets

* The values and metrics generated in this report are subject to statistical fluctuations resulting from the sampling of the 10 folds.

2) Result analysis after pruning +8.

For the clean data, the accuracy dropped slightly after pruning from 97.23% to 96.8%; for the noisy dataset, the accuracy increased greatly from 80.05% to 87.94%. For the noisy dataset, the unpruned tree was overfitted to the training data. Pruning adjustments enhanced its adaptability and performance on test data. Performance differences for the clean data can be attributed to the inherent randomness when partitioning caused by the relatively small fold size.

Cwhy? how?

3) Depth analysis +10

Pruning significantly reduced the average tree depth for both datasets: 12.87 to 9.49 for the clean set and 19.49 to 14.65 for the noisy set. Trees trained without depth limitations or pruning are prone to overfitting. Especially for noisy data, deeper decision nodes can become disproportionately influenced by the data's inherent noise, leading to poor model generalisation. Hence, unrestrained tree growth, especially in the presence of noise, negatively affects prediction accuracy.

Report Quality: +9. Inconsistent and long decimal places makes it difficult to follow.