Introduction to Machine Learning - Decision Tree Coursework

Jan Marczak (01777601), Alfredo Musumeci (01771313), Stefanos Ioannou (02292458), Oussama Rakye Abouelksim (01731878)

December 26, 2022

1 Tree Visualization

Figure 1 & Figure 2 show an example tree trained on the clean dataset before and after pruning. "A" indicates the splitting attribute and "V" indicates the split value. "C" stands for the class the leaf node belongs to. The tree visualisation was done using Matplotlib library. Whereas node values are hard to see, the tree can be zoomed in while running the code.

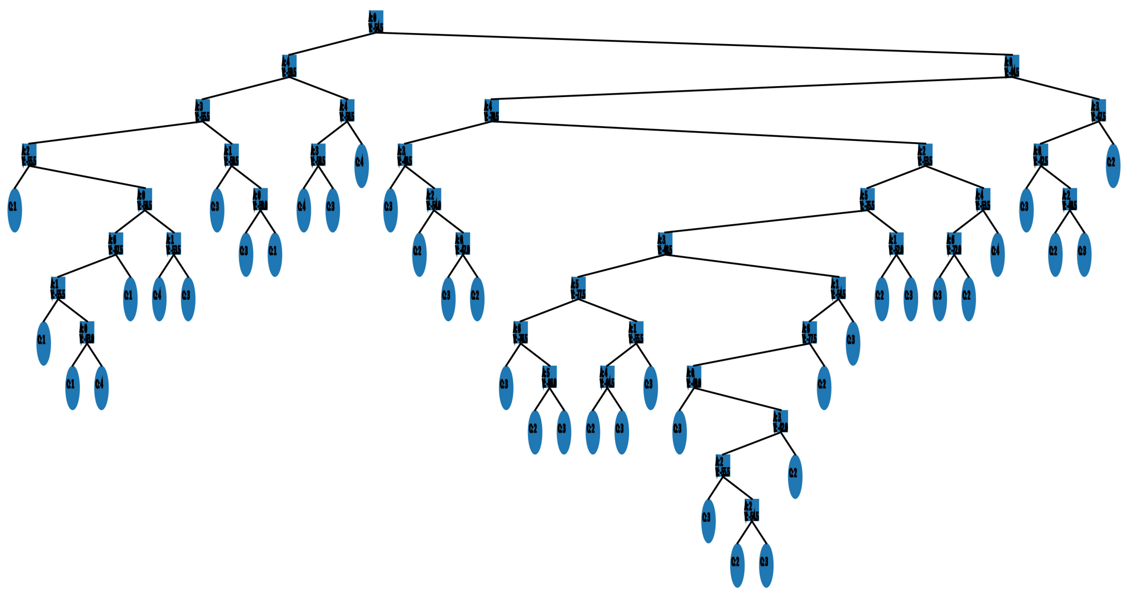

Figure 1: Unpruned tree on clean dataset

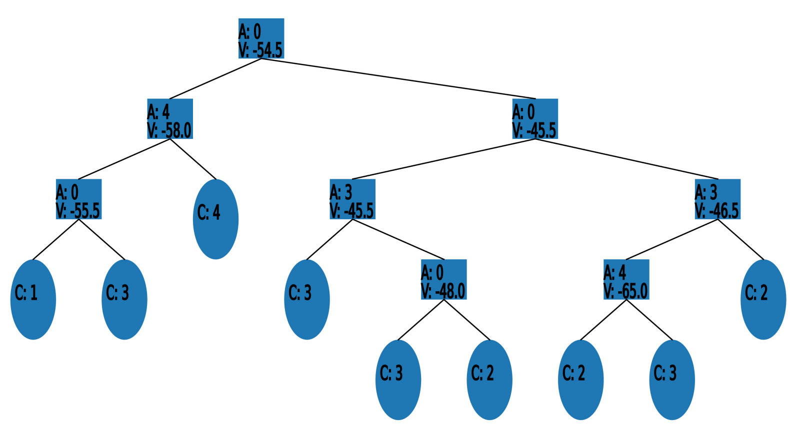

Figure 2: Pruned tree on clean dataset

2 Evaluation

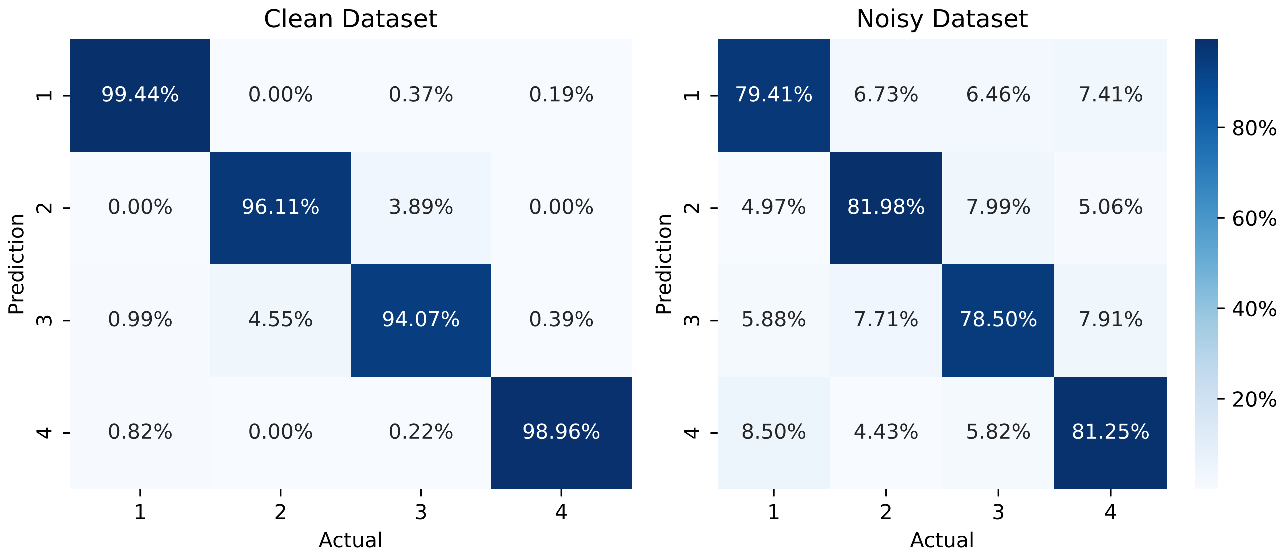

Figure 3: The normalised confusion matrix when training and evaluating the model on the clean dataset (left) and noisy dataset (right), using 10-fold cross-validation.

	Clean Dataset					**Noisy Dataset**				
	1	**2**	**3**	**4**	**Mean**	**1**	**2**	**3**	**4**	**Mean**
Precision	0.988	0.961	0.952	0.992	0.973	0.803	0.816	0.771	0.805	0.799
Recall	0.989	0.958	0.955	0.989	0.973	0.760	0.823	0.802	0.811	0.799
F1	0.988	0.959	0.953	0.990	0.973	0.779	0.817	0.784	0.807	0.797
Mean Accuracy					0.971	**Mean Accuracy**				0.798
Mean Max Depth					13.5	**Mean Max Depth**				19

Table 1: The accuracy, precision, recall, F1 score and maximum depth for the model trained and evaluated on the clean dataset (left) and noisy dataset (right), using 10-fold cross-validation.

2.1 Result Analysis

For the clean dataset, according to the F1 score, the model is better at classifying instances of room 4; the worst average performance is obtained when classifying room 3 instances (see Table 1). According to the F1 score, models trained on the noisy dataset perform best on room 2 and worst on room 1 instances (see Table 1). As can be seen in Figure 3, the error rate for the clean dataset results mainly from miss-classification of rooms 2 and 3 instances.

2.2 Dataset Differences

As shown in Table 1, models trained on the noisy dataset are worse compared to models trained on the clean dataset on all classes. This may be due to the Decision Tree being a high-variance model, thus it tends to overfit to noisy datasets, resulting in poor generalisation (1).

3 Pruning

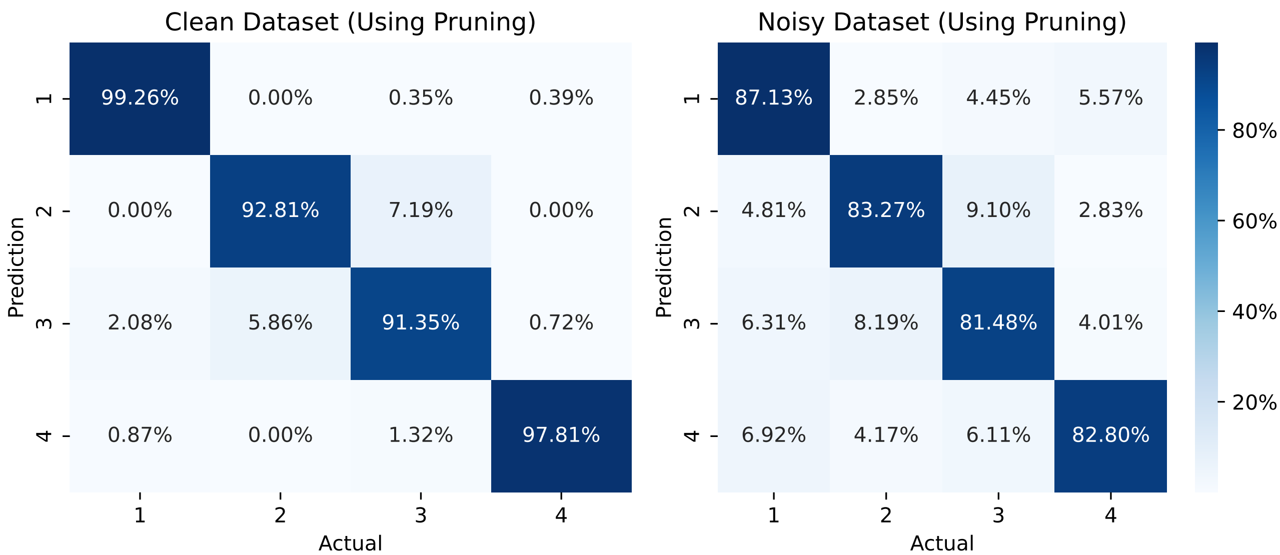

Figure 4: The normalised confusion matrix when training and evaluating the model on the clean dataset (left) and noisy dataset (right), using nested 10-fold cross-validation and pruning.

	Clean Dataset (Using Pruning)					**Noisy Dataset (Using Pruning)**				
	1	**2**	**3**	**4**	**Mean**	**1**	**2**	**3**	**4**	**Mean**
Precision	0.974	0.945	0.919	0.988	0.957	0.832	0.841	0.821	0.867	0.840
Recall	0.992	0.930	0.918	0.983	0.956	0.865	0.833	0.811	0.842	0.838
F1	0.983	0.936	0.917	0.985	0.955	0.847	0.834	0.814	0.852	0.837
Mean Accuracy					0.956	**Mean Accuracy**				0.838
Mean Max Depth					4.60	**Mean Max Depth**				7.43

Table 2: The accuracy, precision, recall, F1 score and maximum depth for the model trained and evaluated on the clean dataset (left) and noisy dataset (right), using nested-10-fold-cross-validation and pruning.

3.1 Result Analysis After Pruning

On the noisy dataset, models trained with pruning perform better on all of the classes compared to unpruned models evaluated on the same dataset. The models' average performance on the clean dataset is still better than that of the noisy dataset, even after pruning. Pruning seems to decrease model variance and increase model bias which may explain the improvement in model performance on the noisy dataset. However, this might have caused the model to under-train on the clean-dataset (1).

3.2 Depth Analysis

As anticipated, the pruned Decision Trees have a lower average depth than the unpruned (see Table 1, Table 2). A lower maximum depth may have decreased model variance, so the model generalises better (i.e. higher accuracy) on the noisy dataset, but might have caused some under-training (i.e. lower accuracy) for the clean dataset.

References

[1] Dietterich, T.G., Kong, E.B.: Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Tech. rep., Technical report, Department of Computer Science, Oregon State University (1995)