Decision Tree Coursework

Riya Chard, Vamika Gupta, Anish Narain, Elsa Polo-Laube

November 2, 2023

Contents

1 Creating Decision Trees **2**

2 Evaluation **3**

2.1 Cross Validation Classification Metrics 3

2.1.1 Confusion Matrix 3

2.1.2 Accuracy 3

2.1.3 Precision 3

2.1.4 Recall 3

2.1.5 F1 4

2.2 Result Analysis 4

2.3 Dataset Differences 4

1 Creating Decision Trees

Using NumPy and Matplotlib, the following decision tree was trained on the entire clean dataset Fig. 1.

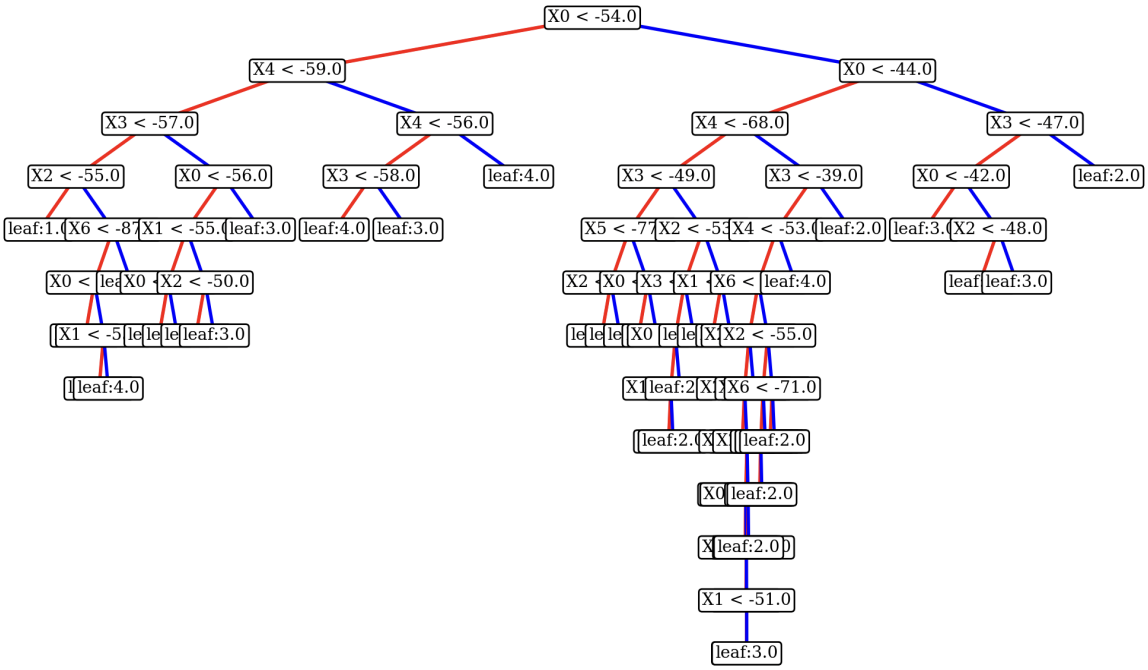

Figure 1: Visual representation of the full tree

Here it is zoomed in to the first left sub-tree Fig. 2.

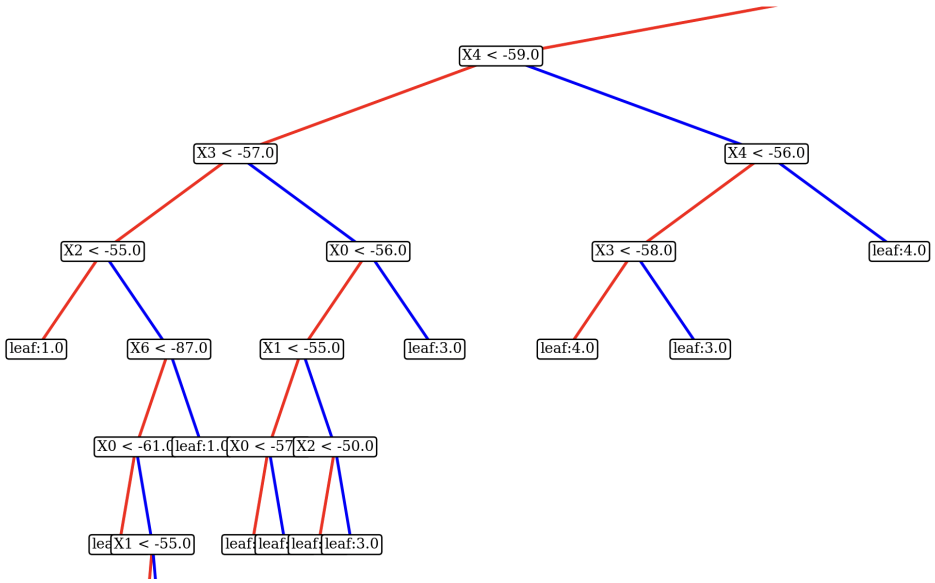

Figure 2: Zoomed-in section of the decision tree

2 Evaluation

Below are all the evaluation metrics calculated from the clean and noisy data sets. The metrics were calculated by finding the average over all 10 folds. The clean dataset was balanced so neither normalising, upsampling or downsampling was necessary.

2.1 Cross Validation Classification Metrics

2.1.1 Confusion Matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	49.4	0	0.1	0.5
Room 2 Actual	0	47.6	2.4	0
Room 3 Actual	0.5	1.8	47.6	0.1
Room 4 Actual	0.4	0	0.2	49.4

Table 1: Confusion Matrix for Clean Dataset

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	39.0	3.9	2.8	3.3
Room 2 Actual	3.1	39.9	4.3	2.4
Room 3 Actual	2.6	3.5	41.6	3.8
Room 4 Actual	4.7	2.2	2.7	40.2

Table 2: Confusion Matrix for Noisy Dataset

2.1.2 Accuracy

	Average Accuracy
Clean Dataset	0.9700
Noisy Dataset	0.8035

Table 3: Average Accuracy Across 10 Folds

2.1.3 Precision

	Room 1	Room 2	Room 3	Room 4	Macro-Averaged
Clean Dataset	0.9820	0.9643	0.9473	0.9873	0.9702
Noisy Dataset	0.7868	0.8032	0.8116	0.8062	0.8020

Table 4: Average Precision

2.1.4 Recall

	Room 1	Room 2	Room 3	Room 4	Macro-Averaged
Clean Dataset	0.9884	0.9524	0.9520	0.9874	0.9701
Noisy Dataset	0.8002	0.8021	0.8115	0.8073	0.8053

Table 5: Average Recall

2.1.5 F1

	Room 1	Room 2	Room 3	Room 4	Macro-Averaged
Clean Dataset	0.9850	0.9581	0.9494	0.9872	0.9699
Noisy Dataset	0.7904	0.8013	0.8099	0.8051	0.8017

Table 6: Average F1

2.2 Result Analysis

For the clean dataset, Room 1 and Room 4 are correctly recognized with high precision, recall, and F1 measures, indicating that the model performs well for these rooms. However, Room 2 and Room 3 have some confusion with others as shown in the confusion matrix: Room 2 has a small number of false positives, and Room 3 has some false negatives, leading to slightly lower precision, recall, and F1 measures for these rooms.

In the noisy dataset, Room 4 is correctly recognized with the highest precision, recall, and F1 measures. Room 2 also performs relatively well, but there is some confusion with Rooms 1 and 3. Room 1 and Room 3 exhibit more confusion with each other and Room 2, as evidenced by the non-zero values in their respective rows and columns in the confusion matrix.

2.3 Dataset Differences

The evaluation of the clean data shows high performance it includes high outputs for Accuracy, Precision and Recall, suggesting the majority of the time the prediction is the true value. The noisy data set shows reduced performance in all evaluation metrics. This can be explained by overfitting. The decision tree closely mimics the training set, which includes noise, resulting in poor generalization. As a result, the model trained on the clean dataset performs better on similarly clean test data, while the model trained on the noisy dataset struggles due to overfitting.