# Decision Tree Coursework

dgs119 mc620 bet20

October 2022

## 1 Evaluation

### 1.1 Cross Validation Classification Metrics

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Room 1 | Room 2 | Room 3 | Room 4 |
| Actual | Room 1 | 49.7 | 0.0 | 0.1 | 0.2 |
| | Room 2 | 0.0 | 47.9 | 2.1 | 0.0 |
| | Room 3 | 0.2 | 1.8 | 47.8 | 0.2 |
| | Room 4 | 0.4 | 0.0 | 0.2 | 49.4 |

(a) Clean Dataset Confusion Matrix. Accuracy: 97.3%

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Room 1 | Room 2 | Room 3 | Room 4 |
| Actual | Room 1 | 38.7 | 2.8 | 3.4 | 4.1 |
| | Room 2 | 3.0 | 39.7 | 4.1 | 2.9 |
| | Room 3 | 3.2 | 3.4 | 41.2 | 3.7 |
| | Room 4 | 4.7 | 2.5 | 2.9 | 39.7 |

(b) Noisy Dataset Confusion Matrix. Accuracy: 79.7%

Figure 1: Confusion Matrices

| | Recall | Precision | $F_1$ |
|---|---|---|---|
| Room 1 | 0.99 | 0.99 | 0.99 |
| Room 2 | 0.96 | 0.96 | 0.96 |
| Room 3 | 0.96 | 0.95 | 0.95 |
| Room 4 | 0.99 | 0.99 | 0.99 |
| Macros | 0.97 | 0.97 | 0.97 |

(a) Clean Dataset Metrics

| | Recall | Precision | $F_1$ |
|---|---|---|---|
| Room 1 | 0.79 | 0.78 | 0.78 |
| Room 2 | 0.80 | 0.82 | 0.81 |
| Room 3 | 0.80 | 0.80 | 0.80 |
| Room 4 | 0.79 | 0.79 | 0.79 |
| Macros | 0.80 | 0.80 | 0.80 |

(b) Noisy Dataset Metrics

Figure 2: Recall, Precision and $F_1$ Metrics

### 1.2 Result Analysis

Rooms 1 and 4 have the highest recall on the clean dataset (Fig 2a). Rooms 2 and 3 are most confused with each other (Fig 1a). This could be because they share WiFi emitters, causing similar signal strengths to be detected in both rooms thus making them harder to differentiate. The trend is reversed on the noisy dataset: Rooms 2 and 3 have the highest recall, and Rooms 1 and 4 are most likely to be confused with one another (Fig 2b, 1b).

### 1.3 Dataset Differences

The accuracy of the tree trained on the noisy dataset is lower by 17.6% compared to the tree trained on the clean dataset. There is also more confusion across all the rooms for the noisy dataset's tree (Fig 1). The tree trained on the noisy dataset overfits to noise and has a high variance. Hence, it fails to generalize well on unseen data and has a significantly lower accuracy when evaluated on the test dataset.

# 2 Pruning

## 2.1 Cross Validation Classification Metrics

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Room 1 | Room 2 | Room 3 | Room 4 |
| Actual | Room 1 | 49.5 | 0.0 | 0.3 | 0.3 |
| | Room 2 | 0.0 | 47.7 | 2.3 | 0.0 |
| | Room 3 | 0.3 | 2.5 | 47.1 | 0.2 |
| | Room 4 | 0.4 | 0.0 | 0.2 | 49.4 |

(a) Clean Dataset Confusion Matrix. Accuracy: 96.8%

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Room 1 | Room 2 | Room 3 | Room 4 |
| Actual | Room 1 | 38.7 | 2.8 | 3.4 | 4.1 |
| | Room 2 | 3.0 | 39.7 | 4.1 | 2.9 |
| | Room 3 | 3.2 | 3.4 | 41.2 | 3.7 |
| | Room 4 | 4.7 | 2.5 | 2.9 | 39.7 |

(b) Noisy Dataset Confusion Matrix. Accuracy: 82.8%

Figure 3: Confusion Matrices

| | Recall | Precision | $F_1$ |
|---|---|---|---|
| Room 1 | 0.99 | 0.99 | 0.99 |
| Room 2 | 0.95 | 0.95 | 0.95 |
| Room 3 | 0.94 | 0.94 | 0.94 |
| Room 4 | 0.99 | 0.99 | 0.99 |
| Macros | 0.97 | 0.97 | 0.97 |

(a) Clean Dataset Metrics

| | Recall | Precision | $F_1$ |
|---|---|---|---|
| Room 1 | 0.82 | 0.82 | 0.82 |
| Room 2 | 0.83 | 0.84 | 0.83 |
| Room 3 | 0.83 | 0.82 | 0.83 |
| Room 4 | 0.84 | 0.83 | 0.84 |
| Macros | 0.83 | 0.83 | 0.83 |

(b) Noisy Dataset Metrics

Figure 4: Recall, Precision and $F_1$ Metrics

## 2.2 Result Analysis

There is a negligible change in the accuracy, 0.5%, of the tree trained on the clean dataset after pruning. Pruning the tree trained on the noisy dataset, however, improves the accuracy by 3.1%. The tree trained on the noisy dataset fits noise and has a high variance. Pruning reduces the tree's variance whilst maintaining bias, allowing it to generalize to unseen data better.

## 2.3 Depth Analysis

| | Unpruned | Pruned |
|---|---|---|
| Clean Dataset | 13.3 | 11.9 |
| Noisy Dataset | 20.5 | 18.9 |

(a) Depth of Trees

| | Unpruned | Pruned |
|---|---|---|
| Clean Dataset | 0.974 | 0.968 |
| Noisy Dataset | 0.796 | 0.828 |

(b) Accuracy of Trees

Figure 5: Depth and Accuracy of Trees

Noise increases the depth of the tree (Fig 5) as trees are created with more decision rules. Reducing tree depth by pruning has a negligible change in accuracy for the clean dataset and increases the accuracy of the tree trained on the noisy dataset. This shows that a deep tree might have a higher variance and generalizes poorly to unseen data.

# 3 Visualization

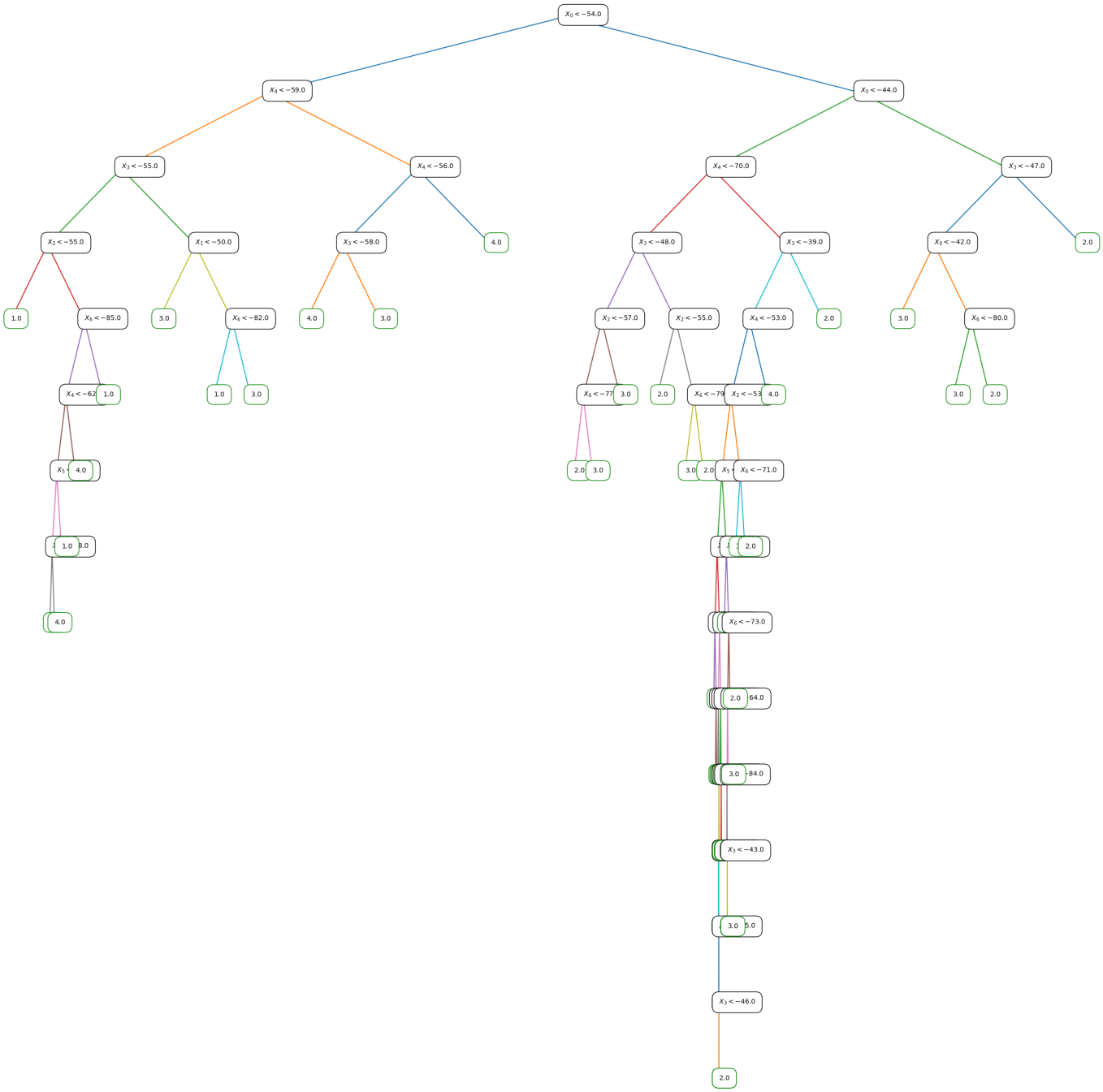


Figure 6: Tree Trained on the Entire Clean Dataset