Imperial College London

INTRODUCTION TO MACHINE LEARNING

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Coursework 1-Decision Tree

Author:
Xuchun Hu, Zuoou Li, Yaqi Zha, Jiayi Xie
CID:
06011335, 06008185, 06013006, 06009199

Date: October 31, 2024

Contents

1 Introduction

1.1 Background

As the potential applications in smart homes continue to expand, the technology of indoor positioning based on WIFI signal strength has emerged as a significant and practical issue. Decision trees, as an efficient machine learning algorithm, are particularly well-suited for addressing this classification problem. This report will explore how to utilize decision tree algorithms to perform indoor location recognition based on the collected WIFI signal strength data.

1.2 Dataset

There are two datasets being used – a clean dataset and a noisy dataset (specific distinctions will be discussed in subsequent sections of the report). Each dataset comprises 2,000 samples, where each sample consists of 7 features representing WIFI signal strengths and one label indicating the room number. For each dataset, samples are divided into training and testing sets in an 8:2 ratio to facilitate model training and performance evaluation.

2 Result

2.1 Visualization on the clean dataset

Before pruning, the decision tree has approximately 5 to 7 layers, and the decision nodes in the lower layers are more complex. After pruning, although the visualization seems to show more nodes, this is actually a display artifact. In the unpruned tree, the dense and overlapping nodes were often hidden or rendered as leaf nodes due to space limitations. The pruned tree, despite having fewer actual nodes, displays them more clearly because of better spacing. Indeed, the tree structure is simplified after pruning, which helps reduce overfitting despite what the visualization might imply.

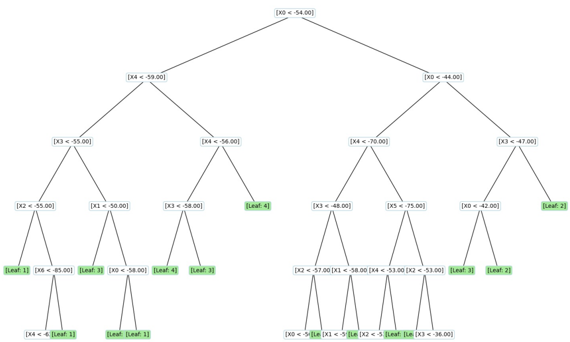

Figure 1: Original tree of clean dataset

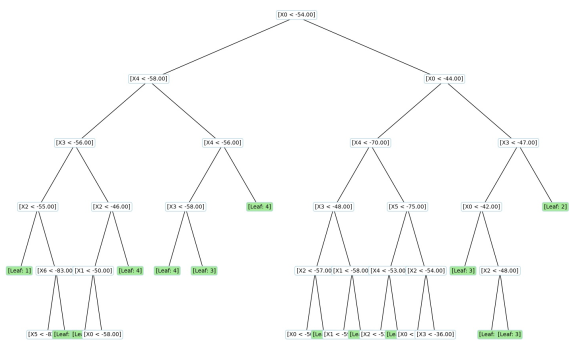

Figure 2: Pruned tree of clean dataset

2.2 Visualization on the noisy dataset

The decision tree trained on the noisy dataset has more than 7 layers before pruning, and the decision path is complex, especially in the bottom layer. After pruning, while the visualization might show more visible nodes due to better spacing (previously, dense nodes often overlapped or were hidden), the actual tree structure is simplified. The decision boundary is clearer, and some leaf nodes are moved to higher layers. The total number of layers is around 6, indicating that the complexity and noise sensitivity are reduced.

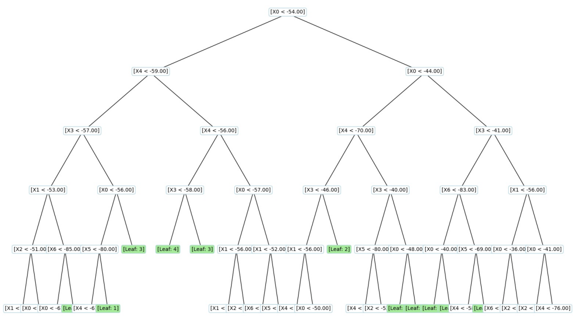

Figure 3: Original tree of noisy dataset

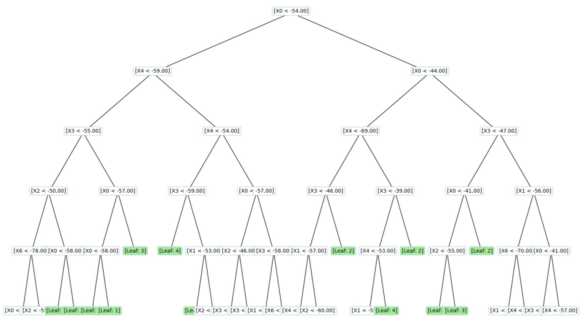

Figure 4: Pruned tree of noisy dataset

3 Evaluation

3.1 Cross-validation

Accuracy measures overall prediction correctness, while Precision indicates the accuracy of positive predictions, and Recall shows the proportion of actual positives captured. The F1-score evaluates model performance, particularly in multi-class scenarios. Macro-averaging computes metrics for each class independently and averages them, ensuring equal weight for all classes, which mitigates the influence of larger classes in imbalanced datasets.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{Trace(M)}{\sum(M)}$$

$$Recall = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$F1-Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Macro-average = \frac{1}{N}\sum_{i=1}^{N} Metric_i$$

3.1.1 Clean Dataset

Confusion matrix:

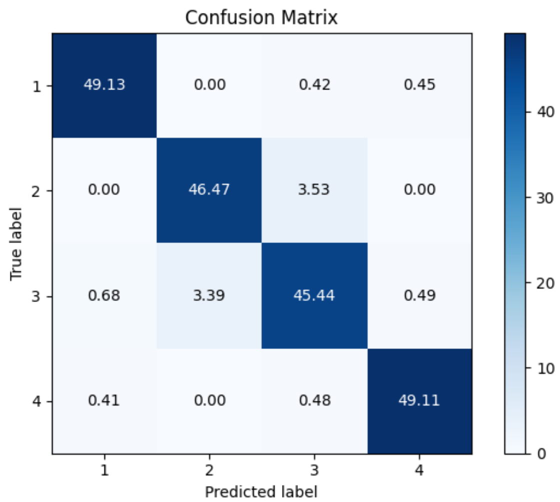

Figure 5: Original confusion matrix on clean dataset

Accuracy:

$$\frac{(49.13 + 46.47 + 45.44 + 49.11)}{\text{sum of all elements}} = \frac{190.15}{200} = 95.1\%$$

Per Class Metrics:
Class 1:

$$\text{Precision} = \frac{49.13}{49.13 + 0.00 + 0.68 + 0.41} = \frac{49.13}{50.22} = 97.8\%$$
$$\text{Recall} = \frac{49.13}{49.13 + 0.00 + 0.42 + 0.45} = \frac{49.13}{50} = 98.3\%$$
$$\text{F1} = 2 \cdot \frac{0.978 \cdot 0.983}{0.978 + 0.983} = 98.0\%$$

Class 2:

$$\text{Precision} = \frac{46.47}{0.00 + 46.47 + 3.39 + 0.00} = \frac{46.47}{49.86} = 93.2\%$$
$$\text{Recall} = \frac{46.47}{0.00 + 46.47 + 3.53 + 0.00} = \frac{46.47}{50} = 92.9\%$$
$$\text{F1} = 2 \cdot \frac{0.932 \cdot 0.929}{0.932 + 0.929} = 93.1\%$$

Class 3:

$$\text{Precision} = \frac{45.44}{0.42 + 3.53 + 45.44 + 0.48} = \frac{45.44}{49.87} = 91.1\%$$
$$\text{Recall} = \frac{45.44}{0.68 + 3.39 + 45.44 + 0.49} = \frac{45.44}{50} = 90.9\%$$

$$\text{F1} = 2 \cdot \frac{0.911 \cdot 0.909}{0.911 + 0.909} = 91.0\%$$

Class 4:

$$\text{Precision} = \frac{49.11}{0.45 + 0.00 + 0.49 + 49.11} = \frac{49.11}{50.05} = 98.1\%$$

$$\text{Recall} = \frac{49.11}{0.41 + 0.00 + 0.48 + 49.11} = \frac{49.11}{50} = 98.2\%$$

$$\text{F1} = 2 \cdot \frac{0.981 \cdot 0.982}{0.981 + 0.982} = 98.2\%$$

Macro-average:

$$\text{Macro-average precision} = \frac{0.978 + 0.932 + 0.911 + 0.981}{4} = 95.1\%$$

$$\text{Macro-average Recall} = \frac{0.983 + 0.929 + 0.909 + 0.982}{4} = 95.1\%$$

$$\text{Macro-average F1} = \frac{0.980 + 0.931 + 0.910 + 0.982}{4} = 95.1\%$$

Table 1: Cross validation classification metrics: Clean Dataset

Class	Recall(%)	Precision(%)	F1(%)
1	98.3	97.8	98.0
2	92.9	93.2	93.1
3	90.9	91.1	91.0
4	98.2	98.1	98.2
Macro-avg	95.1	95.1	95.1

3.1.2 Noisy Dataset

Confusion matrix:

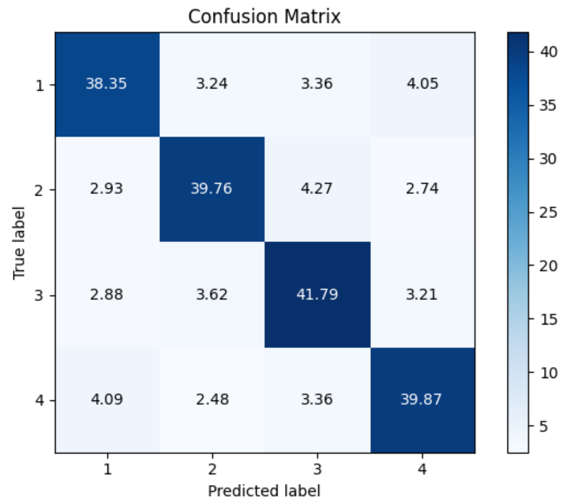

Figure 6: Original confusion matrix on noisy dataset

Accuracy:

$$\frac{(38.35+39.76+41.79+39.87)}{\text{sum of all elements}}=\frac{159.77}{200}=79.9\%$$

Per Class Metrics:

Class 1:

$$\text{Precision}=\frac{38.35}{38.35+2.93+2.88+4.09}=\frac{38.35}{48.25}=79.5\%$$
$$\text{Recall}=\frac{38.35}{38.35+3.24+3.36+4.05}=\frac{38.35}{49}=78.3\%$$
$$\text{F1}=2\cdot\frac{0.795\cdot 0.783}{0.795+0.783}=78.9\%$$

Class 2:

$$\text{Precision}=\frac{39.76}{3.24+39.76+3.62+2.48}=\frac{39.76}{49.1}=81.0\%$$
$$\text{Recall}=\frac{39.76}{2.93+39.76+4.27+2.74}=\frac{39.76}{49.7}=80.0\%$$
$$\text{F1}=2\cdot\frac{0.810\cdot 0.800}{0.810+0.800}=80.5\%$$

Class 3:

$$\text{Precision}=\frac{41.79}{3.36+4.27+41.79+3.36}=\frac{41.79}{52.78}=79.2\%$$
$$\text{Recall}=\frac{41.79}{2.88+3.62+41.79+3.21}=\frac{41.79}{51.5}=81.1\%$$
$$\text{F1}=2\cdot\frac{0.792\cdot 0.811}{0.792+0.811}=80.1\%$$

Class 4:

$$\text{Precision}=\frac{39.87}{4.05+2.74+3.21+39.87}=\frac{39.87}{49.87}=79.9\%$$
$$\text{Recall}=\frac{39.87}{4.09+2.48+3.36+39.87}=\frac{39.87}{49.8}=80.1\%$$
$$\text{F1}=2\cdot\frac{0.799\cdot 0.801}{0.799+0.801}=80.0\%$$

Macro-average:

$$\text{Macro-average precision}=\frac{0.795+0.810+0.792+0.799}{4}=79.9\%$$
$$\text{Macro-average Recall}=\frac{0.783+0.800+0.811+0.801}{4}=79.9\%$$
$$\text{Macro-average F1}=\frac{0.789+0.805+0.801+0.800}{4}=79.9\%$$

Table 2: Cross validation classification metrics: Noisy Dataset

Class	Recall(%)	Precision(%)	F1(%)
1	78.3	79.5	78.9
2	80.0	81.0	80.5
3	81.1	79.2	80.1
4	80.1	79.9	80.0
Macro-avg	79.9	79.9	79.9

3.2 Result Analysis

In the clean dataset, Classes 1 and 4 achieve high accuracy (95.1%) and F1 scores (98%) with minimal confusion. Class 3 shows minor misclassification with Class 2, likely due to signal similarity between adjacent rooms. In the noisy dataset, both accuracy and F1 scores drop to around 80%, with more uniform confusion, indicating that noise significantly disrupts WiFi signal-based location discrimination.

3.3 Dataset Differences

Yes, there is a significant performance gap between datasets. The clean dataset achieves 95.1% macro-average F1-score, while the noisy dataset drops to 79.9%. This substantial difference arises from noise in WiFi measurements making room discrimination more challenging, resulting in more uniform confusion patterns across rooms compared to the clean dataset's localized confusion between adjacent rooms.

4 Pruning and Re-evaluation

4.1 Cross-validation

4.1.1 Clean Dataset

Confusion matrix:

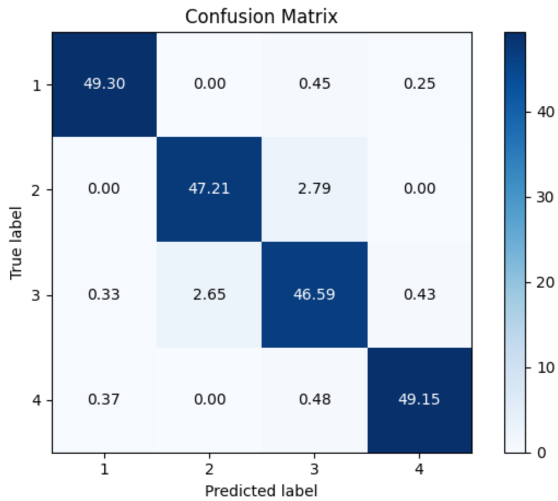

Figure 7: Confusion matrix after pruning on clean dataset

Accuracy:

$$\frac{(49.3 + 47.21 + 46.59 + 49.15)}{\text{sum of all elements}} = \frac{192.25}{200} = 96.1\%$$

Per Class Metrics:

Class 1:

$$\text{Precision} = \frac{49.3}{49.3 + 0.7} = \frac{49.3}{50.4} = 98.6\%$$

$$\text{Recall} = \frac{49.3}{49.3 + 0.7} = \frac{49.3}{50} = 98.6\%$$

$$\text{F1} = 2 \cdot \frac{0.986 \cdot 0.986}{0.986 + 0.986} = 98.6\%$$

Class 2:

$$\text{Precision} = \frac{47.21}{47.21 + 2.65} = \frac{47.21}{49.86} = 94.7\%$$

$$\text{Recall} = \frac{47.21}{47.21 + 2.79} = \frac{47.21}{50} = 94.4\%$$

$$\text{F1} = 2 \cdot \frac{0.947 \cdot 0.944}{0.947 + 0.944} = 94.5\%$$

Class 3:

$$\text{Precision} = \frac{46.59}{46.59 + 3.72} = \frac{46.59}{50.31} = 92.6\%$$

$$\text{Recall} = \frac{46.59}{46.59 + 3.41} = \frac{46.59}{50} = 93.2\%$$

$$\text{F1} = 2 \cdot \frac{0.926 \cdot 0.932}{0.926 + 0.932} = 92.9\%$$

Class 4:

$$\text{Precision} = \frac{49.15}{49.15 + 0.68} = \frac{49.15}{49.83} = 98.6\%$$

$$\text{Recall} = \frac{49.15}{49.15 + 0.85} = \frac{49.15}{50} = 98.3\%$$

$$\text{F1} = 2 \cdot \frac{0.983 \cdot 0.986}{0.983 + 0.986} = 98.4\%$$

Macro-average:

$$\text{Macro-average precision} = \frac{0.986 + 0.947 + 0.926 + 0.986}{4} = 96.1\%$$

$$\text{Macro-average Recall} = \frac{0.986 + 0.944 + 0.932 + 0.983}{4} = 96.2\%$$

$$\text{Macro-average F1} = \frac{0.986 + 0.945 + 0.929 + 0.984}{4} = 96.1\%$$

Table 3: Cross validation classification metrics: Clean Dataset

Class	Recall(%)	Precision(%)	F1(%)
1	98.6	98.6	98.6
2	94.4	94.7	94.5
3	93.2	92.6	92.9
4	98.3	98.6	98.4
Macro-avg	96.2	96.1	96.1

4.1.2 Noisy Dataset

Confusion matrix:

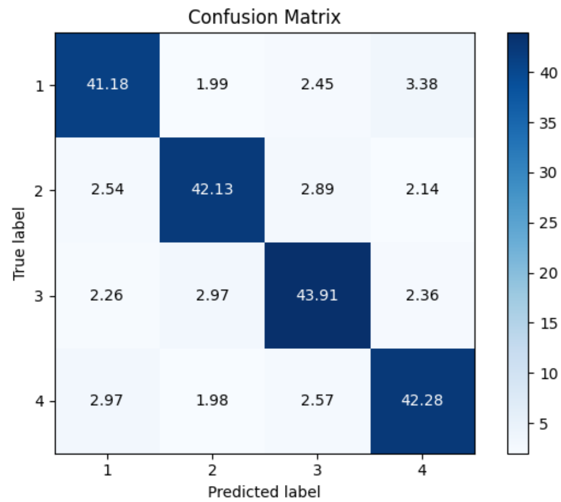

Figure 8: Confusion matrix after pruning on noisy dataset

Accuracy:

$$\frac{(41.18 + 42.13 + 43.91 + 42.28)}{\text{sum of all elements}} = \frac{169.5}{200} = 84.8\%$$

Per Class Metrics:

Class 1:

$$\text{Precision} = \frac{41.18}{41.18 + 7.77} = \frac{41.18}{48.95} = 84.1\%$$
$$\text{Recall} = \frac{41.18}{41.18 + 7.82} = \frac{41.18}{49.0} = 84.1\%$$
$$\text{F1} = 2 \cdot \frac{0.841 \cdot 0.841}{0.841 + 0.841} = 84.1\%$$

Class 2:

$$\text{Precision} = \frac{42.13}{42.13 + 6.94} = \frac{42.13}{49.07} = 85.9\%$$
$$\text{Recall} = \frac{42.13}{42.13 + 7.57} = \frac{42.13}{49.7} = 84.8\%$$
$$\text{F1} = 2 \cdot \frac{0.859 \cdot 0.848}{0.859 + 0.848} = 85.3\%$$

Class 3:

$$\text{Precision} = \frac{43.91}{43.91 + 7.91} = \frac{43.91}{51.82} = 84.7\%$$
$$\text{Recall} = \frac{43.91}{43.91 + 7.59} = \frac{43.91}{51.5} = 85.3\%$$

$$F1 = 2 \cdot \frac{0.847 \cdot 0.853}{0.847 + 0.853} = 85.0\%$$

Class 4:

$$\text{Precision} = \frac{42.28}{42.28 + 7.88} = \frac{42.28}{50.16} = 84.3\%$$
$$\text{Recall} = \frac{42.28}{42.28 + 7.52} = \frac{42.28}{49.8} = 84.9\%$$
$$F1 = 2 \cdot \frac{0.843 \cdot 0.849}{0.843 + 0.849} = 84.6\%$$

Macro-average:

$$\text{Macro-average precision} = \frac{0.841 + 0.859 + 0.847 + 0.843}{4} = 84.8\%$$
$$\text{Macro-average Recall} = \frac{0.841 + 0.848 + 0.853 + 0.849}{4} = 84.8\%$$
$$\text{Macro-average F1 score} = \frac{0.841 + 0.853 + 0.850 + 0.846}{4} = 84.8\%$$

Table 4: Cross validation classification metrics: Noisy Dataset

Class	Recall(%)	Precision(%)	F1(%)
1	84.1	84.1	84.1
2	84.8	85.9	85.3
3	85.3	84.7	85.0
4	84.9	84.3	84.6
Macro-avg	84.8	84.8	84.8

4.2 Result Analysis

After pruning, the noisy dataset's F1-score improves from 0.80 to 0.85, as pruning reduces overfitting by removing noise-fitting branches. The clean dataset shows minimal improvement (F1: 0.95→0.96) since its original tree was already well-suited to clean patterns. This highlights pruning's greater effectiveness in noisy data by reducing the influence of noise, leading to a more generalized model.

Table 5: The difference in performance before and after pruning for both datasets.

Prune	Clean Dataset		Noisy Dataset	
	Accuracy	F1-Score	Accuracy	F1-Score
✓	0.96↑	0.96↑	0.85↑	0.85↑
✗	0.95	0.95	0.80	0.80
#Pruned Nodes	39↓		57↓	
Depth	4↓		5↓	

4.3 Depth Analysis

The noisy dataset generates deeper, more complex trees due to the influence of noisy data, while the clean dataset produces simpler ones. After pruning, both trees reduce in depth while maintaining or improving accuracy, showing that deeper trees don't guarantee better performance. Pruned trees highlight that simpler models would generalize better, especially with noisy data.