**Introduction to ML COMP60012**

**Decision Tree Coursework Report**

Runshen Lu 02033159

Han Jiao 02030048

Mo Lin 02027654



https://www.vecteezy.com/vector-art/1234042-decision-tree-design

This report shows results from the decision-tree machine learning program. It includes the visualization of the generated tree and evaluation of the learning algorithm.

**Visualization of binary tree for clean data**

A diagram of a structure

Description automatically generated

**Cross-validation classification metrics**

**Confusion Matrix (clean)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Room 1 (actual) | Room 2 (actual) | Room 3 (actual) | Room 4 (actual) |
| Room 1 (predicted) | 499 | 0 | 3 | 5 |
| Room 2 (predicted) | 0 | 472 | 11 | 0 |
| Room 3 (predicted) | 1 | 28 | 485 | 4 |
| Room 4 (predicted) | 0 | 0 | 1 | 491 |

Accuracy: 97.35%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Room 1 | Room 2 | Room 3 | Room 4 |
| Precision | 98.42% | 97.72% | 93.63% | 99.80% |
| Recall | 99.80% | 94.40% | 97.00% | 98.20% |
| F1 | 0.9911 | 0.9603 | 0.9529 | 0.9899 |

**Confusion Matrix (noisy)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Room 1 (actual) | Room 2 (actual) | Room 3 (actual) | Room 4 (actual) |
| Room 1 (predicted) | 446 | 18 | 20 | 22 |
| Room 2 (predicted) | 10 | 436 | 24 | 13 |
| Room 3 (predicted) | 13 | 32 | 454 | 17 |
| Room 4 (predicted) | 21 | 11 | 17 | 446 |

Accuracy: 89.10%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Room 1 | Room 2 | Room 3 | Room 4 |
| Precision | 88.14% | 90.27% | 87.98% | 90.10% |
| Recall | 91.02% | 87.73% | 88.16% | 89.56% |
| F1 | 0.8956 | 0.8898 | 0.8807 | 0.8983 |

**Result Analysis**

For both clean and noisy dataset, the classifications are quite accurate for the maximum depth=4 we used. The accuracy peaks when max depth equals 3 and 4 and decreases as max depth increases because the overfit phenomenon occurs. Our confusion matrices for both cases show that room 1 and room 4 are predicted most accurately with the highest F1 scores. Room 2 and room 3 are more likely to be confused. The decision tree helps to explain this phenomenon as most of splitting between Room 2 and 3 happens deeper in the tree.

**Datasets Difference**

It is apparent that the model trained by clean dataset has better performance shown in all the evaluation metrics. The difference between the F1 scores for each room in the noisy dataset becomes less obvious. This is because noisy datasets introduce errors like noise, outliers, wrong labels, which could pollute the learning process and introduce confusion to the process of differentiating between different classes, thus lowering the evaluation metrics.