A picture containing scale, device

Description automatically generatedDiagram

Description automatically generated1 - Output of Tree Visualisation Function

**Figure 1: Overview of decision tree model**

**Figure 2: Close up view of decision tree root**

2 - Step 3: Evaluation

**Chart

Description automatically generated**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Room 1** | **Room 2** | **Room 3** | **Room 4** |
| **Accuracy** | 0.972 | | | |
| **F1-score** | 0.987 | 0.959 | 0.949 | 0.989 |
| **Precision** | 0.985 | 0.954 | 0.957 | 0.989 |
| **Recall** | 0.990 | 0.964 | 0.942 | 0.988 |

**Figure 3: Clean dataset 10-fold cross validation metrics** (‘default\_rng’ seed set to 1)

Chart

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Room 1** | **Room 2** | **Room 3** | **Room 4** |
| **Accuracy** | 0.804 | | | |
| **F1-score** | 0.778 | 0.826 | 0.804 | 0.797 |
| **Precision** | 0.786 | 0.818 | 0.805 | 0.808 |
| **Recall** | 0.773 | 0.837 | 0.808 | 0.791 |

**Figure 4: Noisy dataset 10-fold cross validation metrics** (‘default\_rng’ seed set to 1)

**Result analysis:**

* **Clean dataset**: Room 3 is the least accurately predicted room with the largest number of false positives, Room 2 being the most mislabeled as Room 3. Room 4 was the most accurately predicted room.
* **Noisy dataset**: Room 1 is the least accurately predicted room with the largest number of false positives, Room 4 being the most mislabeled as Room 1. Room 2 was the most accurately predicted room.

**Dataset differences:** Yes, the noisy dataset resulted in the trained model making significantly more false positives and false negatives. These resulted in an accuracy decreased of 16.8% compared to clean dataset.

3 - Step 4: Pruning

Chart

Description automatically generated

**Figure 5: Clean dataset 10-fold nested cross validation metrics** **after pruning** (‘default\_rng’ seed set to 1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Room 1** | **Room 2** | **Room 3** | **Room 4** |
| **Accuracy** | 0.969 | | | |
| **F1-score** | 0.986 | 0.958 | 0.943 | 0.989 |
| **Precision** | 0.978 | 0.962 | 0.947 | 0.992 |
| **Recall** | 0.994 | 0.954 | 0.941 | 0.987 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Room 1** | **Room 2** | **Room 3** | **Room 4** |
| **Accuracy** | 0.875 | | | |
| **F1-score** | 0.877 | 0.878 | 0.862 | 0.879 |
| **Precision** | 0.863 | 0.880 | 0.872 | 0.885 |
| **Recall** | 0.892 | 0.877 | 0.854 | 0.875 |

Chart

Description automatically generated

**Figure 6: Noisy dataset 10-fold nested cross validation metrics** **after pruning** (‘default\_rng’ seed set to 1)

**Result analysis:**

* **Clean dataset**: Slight performance decrease in accuracy after pruning due to higher bias.
* **Noisy dataset**: Performance increase after pruning due to reduced variance resulting in reduced error from small fluctuations of noise in training set.

**Depth analysis:**

|  |  |  |
| --- | --- | --- |
|  | **Average tree depth before pruning** | **Average tree depth after pruning** |
| **Clean dataset** | 13.30 | 11.24 |
| **Noisy dataset** | 19.90 | 17.42 |

Pruning the decision tree reduces the depth of the decision tree which reduces the variance but increases the bias. In the clean dataset, there is no noise, so the effect of reduced variance is lower, but the higher bias decreases the accuracy. In the noisy dataset, the effect of lower variance in reducing error from noise is much more significant than the higher bias thus leading to greater accuracy.