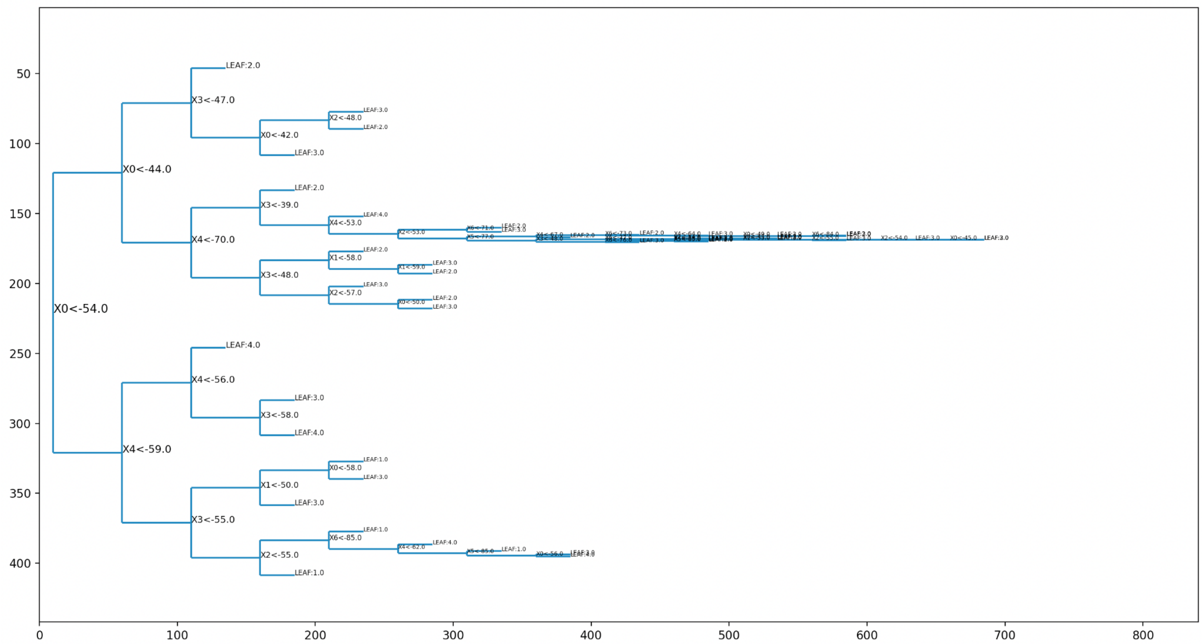
ML Coursework 1

Team: Alexander Goncharov, Jack Qiu, Karan Obhrai, Leo Unoki

Step 2: Bonus Part

Decision Tree for Entire Clean Dataset



Step 3: Evaluation

Dataset Results:
Note: each confusion matrix shown below is the average of all the confusion matrices obtained through one 10-fold cross-validation.

Clean dataset results with no pruning:

Confusion Matrix:

	Room 1 Predicted	**Room 2 Predicted**	**Room 3 Predicted**	**Room 4 Predicted**
Room 1 Actual	49.3	0	0.5	0.2
Room 2 Actual	0	48.4	1.6	0
Room 3 Actual	0.3	1.8	47.7	0.2
Room 4 Actual	0.5	0	0.1	49.4

Accuracy: 0.974
Recall:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.986	0.968	0.954	0.988

Precision:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.98403194	0.96414343	0.95591182	0.99196787

F1-Score:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.98501499	0.96606786	0.95495495	0.98997996

Noisy dataset results with no pruning:

Confusion Matrix:

	Room 1 Predicted	**Room 2 Predicted**	**Room 3 Predicted**	**Room 4 Predicted**
Room 1 Actual	38.6	2.5	3.5	4.4
Room 2 Actual	3.2	39.4	3.9	3.2
Room 3 Actual	2.4	3.9	42.1	3.1
Room 4 Actual	4.8	2.7	3.2	39.1

Accuracy: 0.796

Recall:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.7877551	0.79275654	0.81747573	0.78514056

Precision:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.7877551	0.81237113	0.79886148	0.78514056

F1-Score:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.7877551	0.80244399	0.80806142	0.78514056

Comment for both datasets which rooms are recognized with high/low accuracy, and which rooms are confused:

For the clean dataset, accuracy per class and precision/recall per class were all fairly high (> 95%), so all rooms were recognized with high accuracy and high precision/recall. If anything, rooms 2 and 3 were confused between each other at a higher incidence rate than others

For the noisy dataset, accuracy per class and precision/recall per class were all reasonably high (~ 80%). We see from the confusion matrix that the tree at times confuses Room 1 and 4 with each other at a higher incidence rate than others.

Is there any difference in the performance when using the clean and noisy datasets? If yes/no explain why:

Noisy dataset across all metrics: accuracy and precision/recall per class, was worse than the clean dataset. We reason that because the dataset is noisy, it contains many outliers that both creates a tree that doesn't generalize well and overfits to the training dataset, and data instances/labels that are outliers and are hard for the tree to accurately predict.

Step 4 - Pruning (and evaluation again)

Clean dataset results with pruning:

Confusion Matrix:

	Room 1 Predicted	**Room 2 Predicted**	**Room 3 Predicted**	**Room 4 Predicted**
Room 1 Actual	49.61111111	0	0.27777778	0.11111111
Room 2 Actual	0	47.14444444	2.85555556	0
Room 3 Actual	0.5	1.95555556	47.11111111	0.43333333
Room 4 Actual	0.5	0	0.27777778	49.22222222

Accuracy: 0.9654444444444445

Recall:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.99222222	0.94288889	0.94222222	0.98444444

Precision:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.98024149	0.96017198	0.93248296	0.98906006

F1-Score:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.98619547	0.95145196	0.93732729	0.98674685

Noisy dataset results with pruning:

Confusion Matrix:

	Room 1 Predicted	**Room 2 Predicted**	**Room 3 Predicted**	**Room 4 Predicted**
Room 1 Actual	44.24444444	1.2	1.43333333	2.12222222
Room 2 Actual	1.84444444	43.16666667	3.28888889	1.4
Room 3 Actual	2.32222222	3.12222222	44.16666667	1.88888889
Room 4 Actual	2.33333333	1.38888889	1.7	44.37777778

Accuracy: 0.8797777777777778

Recall:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.90294785	0.8685446	0.85760518	0.89112004

Precision:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.87190716	0.88315526	0.87305074	0.8913189

F1-Score:

Room 1	**Room 2**	**Room 3**	**Room 4**
0.88715607	0.875789	0.86525903	0.89121946

Comment the difference of performance before and after pruning for both datasets. Briefly explain these performance differences:

A decision tree without pruning most often overfits to the training dataset, meaning that many later branches in the tree that only exist to classify outliers in the training dataset. By pruning the dataset, we remove these branches that only exist to classify outliers and thus help the tree generalize better to unseen datasets.

For the noisy dataset, we see a sizable increase in average classification accuracy on the test dataset with pruning compared to that of no pruning. We reason that because the dataset is noisy, it contains a lot of outliers and branches that exist only to classify these outliers, and by pruning, we remove these invaluable outlier branches.

For the clean dataset, we see a slight decrease in average classification accuracy on the test dataset with pruning compared to that of no pruning. We reason that because the dataset is clean, it contains very little to no outliers, and by pruning, we remove valuable classification branches that can generalize past outliers.

Comment on the average depth of the trees that you generated for both datasets, before and after pruning. What can you tell about the relationship between maximal depth and prediction accuracy?:

For the unpruned trees for the clean dataset, it had an average depth of 11, whereas the tree for the noisy dataset had an average depth of 20 giving an mean average depth of 15.5. After pruning, the tree for the clean dataset had a maximum depth of 5 and the noisy dataset had an average depth of 9. This gave a mean average depth of 7, considerably less than the depth for the unpruned tree.

For the clean dataset we see the average slightly from 97.4% to 96.5% after pruning, however for the noisy dataset the accuracy increases from 79.6% to 88.0%. This is a considerable increase for the noisy dataset. Therefore it looks like by reducing the maximal depth for the clean dataset slightly reducing the accuracy however for the noisy dataset, this substantially increases, which we hypothesise is because there is less overfitting. This relationship probably isn't linear, as if we pruned the trees by much more, we'd probably see a reduction in accuracy for both datasets.