COURSEWORK

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Introduction to Machine Learning CW1

Authors:
Jacob Peake
Ryan Meierhofer
Yash Belur
Donavon Clay

Date: November 2, 2023

1 Tree Visualisation

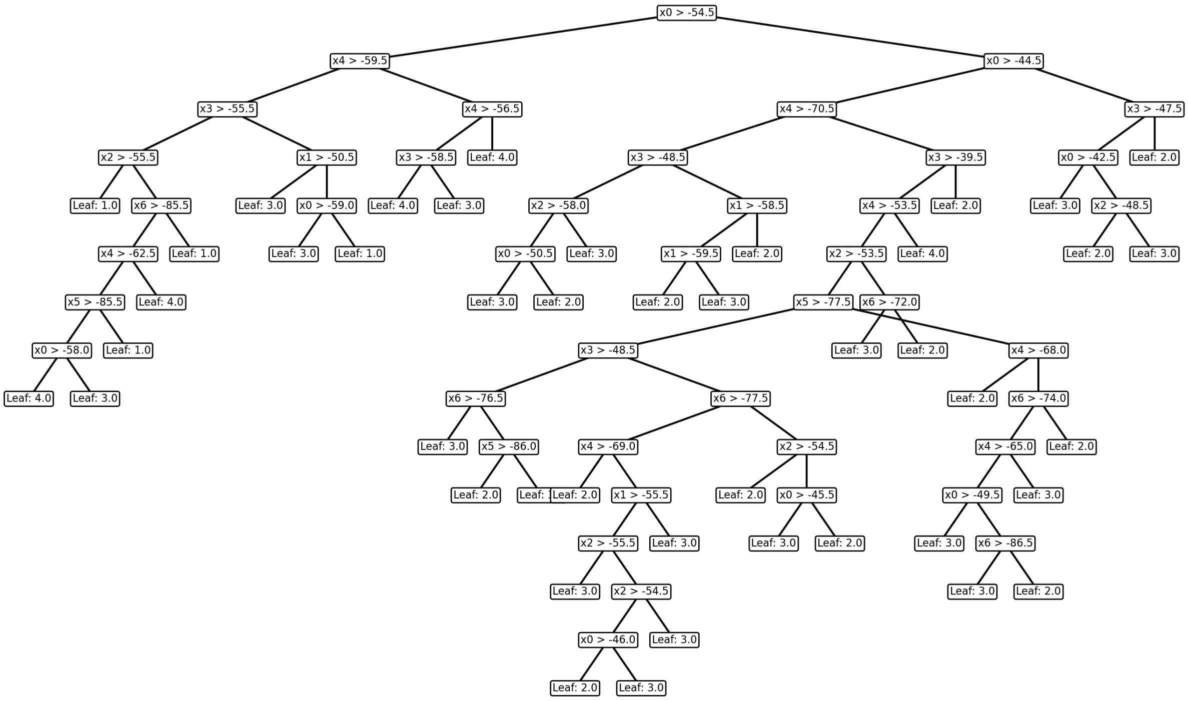

Figure 1: Plot of the decision tree using the clean dataset.

2 Evaluation

2.1 Cross-Validation Classification Metrics

The Evaluation Metrics for each dataset was computed by evaluating the decision tree on each fold, and computing the mean of all metrics.

Confusion Matrix:

Clean Dataset:

	Class 1 Predicted	Class 2 Predicted	Class 3 Predicted	Class 4 Predicted
Class 1 Actual	49.5	0	0.4	0.1
Class 2 Actual	0	48.1	1.9	0
Class 3 Actual	0.2	2.0	47.7	0.1
Class 4 Actual	0.5	0	0.1	49.4

Noisy Dataset:

	Class 1 Predicted	Class 2 Predicted	Class 3 Predicted	Class 4 Predicted
Class 1 Actual	38.6	3.3	3.2	3.9
Class 2 Actual	2.7	40.6	3.2	3.2
Class 3 Actual	3.0	3.2	42.5	2.8
Class 4 Actual	4.7	2.3	2.8	40.0

Accuracy:

Clean Dataset: 0.9735

Noisy Dataset: 0.8085

Precision:

Clean Dataset: [Class 1: 0.9861, Class 2: 0.9601, Class 3: 0.9521, Class 4: 0.996]

Noisy Dataset: [Class 1: 0.7878, Class 2: 0.8219, Class 3: 0.8221, Class 4: 0.8016]

Recall:

Clean Dataset: [Class 1: 0.990, Class 2: 0.962, Class 3: 0.954, Class 4: 0.988]

Noisy Dataset: [Class 1: 0.7878, Class 2: 0.8169, Class 3: 0.8252, Class 4: 0.8032]

F1-Measure:

Clean Dataset: [Class 1: 0.988, Class 2: 0.961, Class 3: 0.953, Class 4: 0.992]

Noisy Dataset: [Class 1: 0.7878, Class 2: 0.8194, Class 3: 0.8236, Class 4: 0.8024]

2.2 Result Analysis

For the clean dataset, rooms 1 & 4 are almost always correctly recognised. Rooms 2 & 3 are sometimes confused with each other. This is reflected in the confusion matrix at $C_{23} \& C_{32}$, & in the lower precision & recall for classes 2 & 3. This may be due to a WiFi Signal that is close to the boundary between rooms 2 & 3. For the noisy dataset, all rooms are sometimes confused with each other (see off-diagonal entries of confusion matrix & precision/recall measures).

2.3 Dataset Differences

The clean dataset shows consistently greater performance than the noisy dataset - reflected in the greater accuracy, and precision & recall for each class. This is due to the clean dataset providing a better representation of the underlying distribution that the decision tree in trying to model. This means that the 'clean data' model will generalise better to new, unseen, data, whereas the 'noisy data' model may overfit, as it learns the noise too.