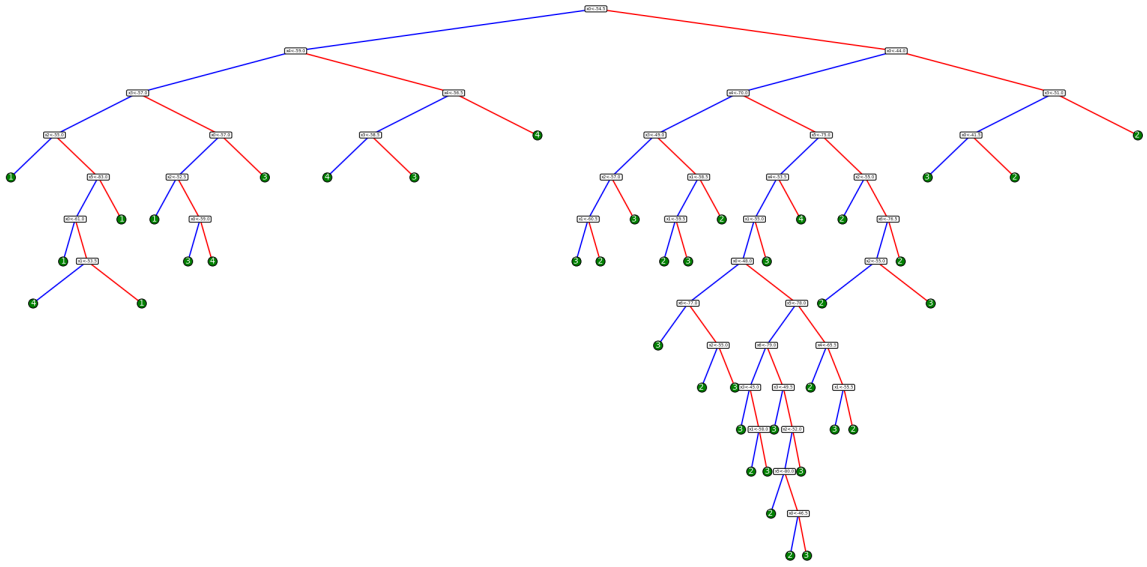
ML coursework one - Decision Trees report

Yasser Hassan, James Taylor (Waisan), Ahmed Elkouny, Yazan Ayyoub

October 2023

1 Tree Visualisation Output



2 Evaluation

2.1 Clean Data Confusion Matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 predicted
Room 1 Actual	493	0	4	3
Room 2 Actual	0	477	23	0
Room 3 Actual	2	30	465	3
Room 4 Actual	3	0	4	493

2.2 Clean Data Set Evaluation Metrics

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.989	0.941	0.9375	0.988
Recall Rates	0.986	0.954	0.930	0.986
F1 Scores	0.988	0.947	0.934	0.987

Accuracy of clean Dataset: 96.39

2.3 Result Analysis Clean Dataset

Room 1 and Room 4 have the highest precision, recall and F1 rates which means that most of the Room 1 and 4 attributes are correctly recognised and there aren't may false positives. Room 2 and 3 have lower precision, recall and F1 scores so have more false positives and miss more attribute recognitions. Rooms 2 and 3 actually often are confused with each other.

2.4 Noisy Data Confusion Matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 predicted
Room 1 Actual	379	36	38	37
Room 2 Actual	36	392	44	25
Room 3 Actual	29	39	408	39
Room 4 Actual	42	27	27	402

2.5 Noise Data Set Evaluation Metrics

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.779	0.794	0.789	0.799
Recall Rates	0.773	0.789	0.792	0.807
F1 Scores	0.776	0.791	0.791	0.803

Accuracy of Noisy Dataset: 79.05

2.6 Result Analysis Noisy Dataset

Here Room 1 has the lowest precision, recall and F1 rates which means that most of the Room 1 attributes are incorrectly recognised and there are many false positives. This class is more often confused than other classes when using noisy data. Room 4 has the highest precision, recall and F1 rates so is confused the least and correctly recognized the most.

2.7 Data Differences

There is around a 17% difference in accuracy between the clean and noisy data sets. This is because there is overfitting on both datasets. This is not a major problem with the clean dataset as most attributes and classifiers follow the same (predictable) split pattern however when noisy data is used the decision tree overfits to noisy (non-typical/non-predictable) attribute values.