Machine Learning
Assignment 2 Part 1 Report

Name - Jatin Mahawar
Roll. No. - 22CS30032

Non-Noisy Data (cardio_noise.csv)

- **Before Pruning**
  0. Number of Nodes - 1367
  1. Accuracy - 0.64
  2. Macro Precision - 0.6403552200321855
  3. Macro Recall Before Pruning - 0.640697454078057

- **After Pruning**
  1. Accuracy - 0.7025
  2. Macro Precision - 0.7022652265226523
  3. Macro Recall After Pruning - 0.702772411041118

Noisy Data (cardio.csv)

- **Before Pruning**
  1. Number of Nodes - 1865
  2. Accuracy - 0.52375
  3. Macro Precision - 0.5237458949453528
  4. Macro Recall Before Pruning - 0.5237257603493343

- **After Pruning**
  1. Accuracy - 0.6208333333333333
  2. Macro Precision - 0.620965409049945
  3. Macro Recall After Pruning - 0.6208031133419816

Techniques used for minimize teh difference between noisy and non-noisy accuracies:

- Nodes are not begin splitted when having information gain less than 0.001.
- Number of Data point in a leaf node restricted to 20.
- Max Height of the Tree is set to be 100.

Comparison

- **Performace Comparison -** Value of Accuracy, precision and recall metrics shows that model perform better on non-noisy data than noisy data.

Impact of Noise

- **Overfit Decision Tree -** Noisy dataset increase the overfitting in model during training, that result in less accuracy of the model.
- **Complexity -** Complexity of Decision tree increase for noisy data. As non-noisy data require 1367 nodes, and noisy data required 1865 nodes.

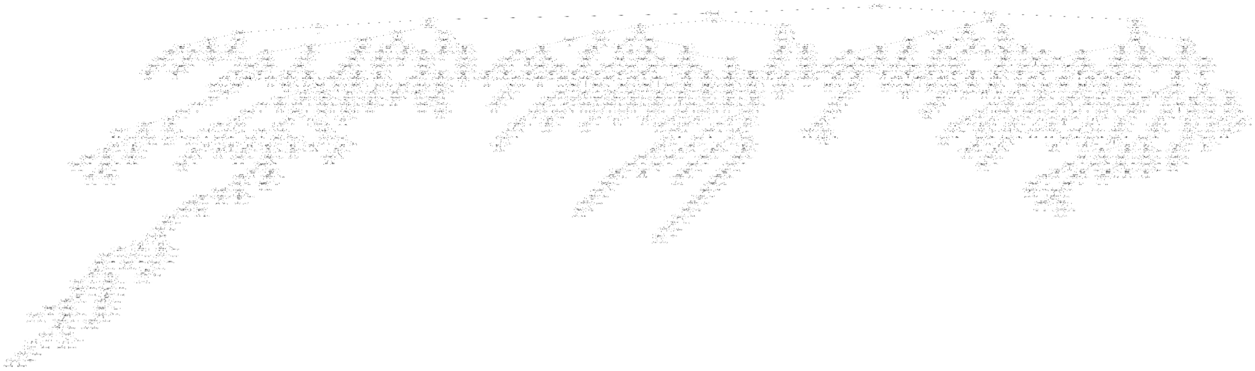
Key Finding

- After applying pruning model improves on both datasets.
- Non-noisy dataset result in more complex decision tree.

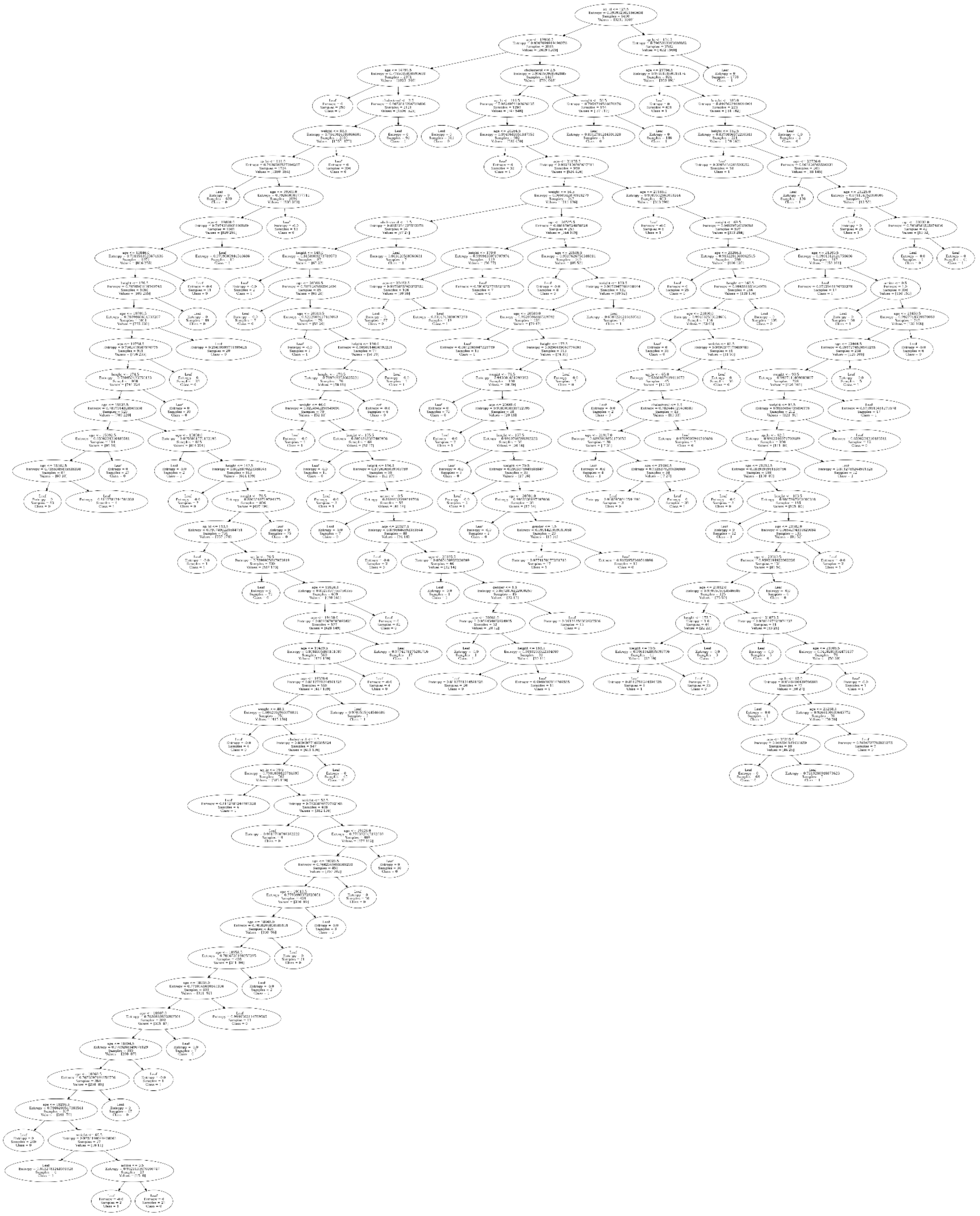
Implications

Applying a good noise removing strategy can result in better accuracy of model, but that could be difficult to apply.

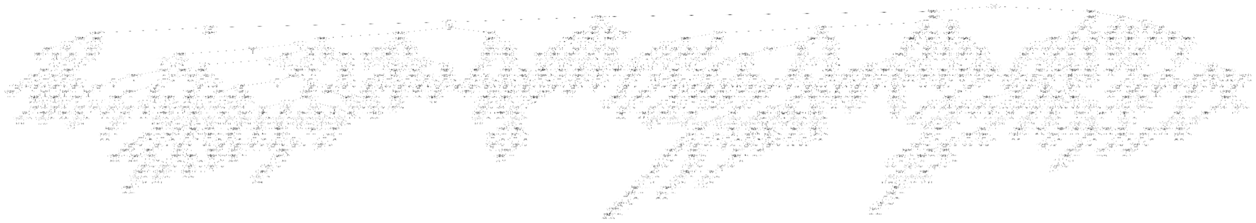
Decision Tree Before Pruning (cardio.csv)



Decision Tree After Pruning (cardio.csv)



Decision Tree Before Pruning (cardio_noise.csv)



Decision Tree After Pruning (cardio_noise.csv)