Assignment 2 Part 1 Report

Sharanya Chakraborty

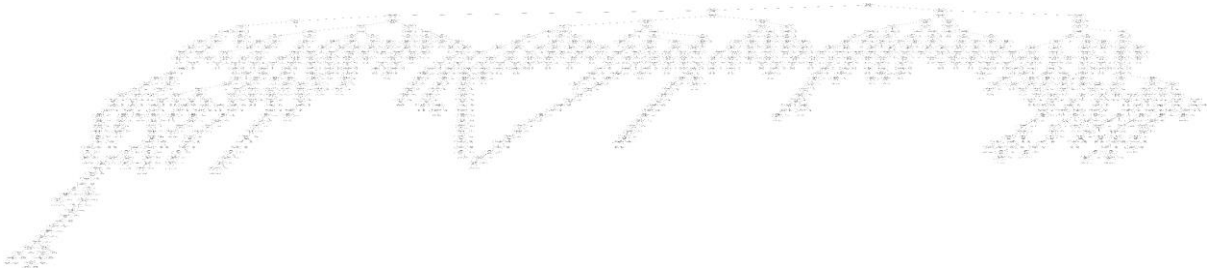
22CS10088

Summary of results from non-noisy and noisy datasets:

- Evaluation results for non-noisy dataset cardio.csv:
  1) Before pruning:

     ```
     Accuracy,Macro Precision, Macro Recall Before Pruning
     (0.6545, 0.655693821878793, 0.6558744293908456)
     ```

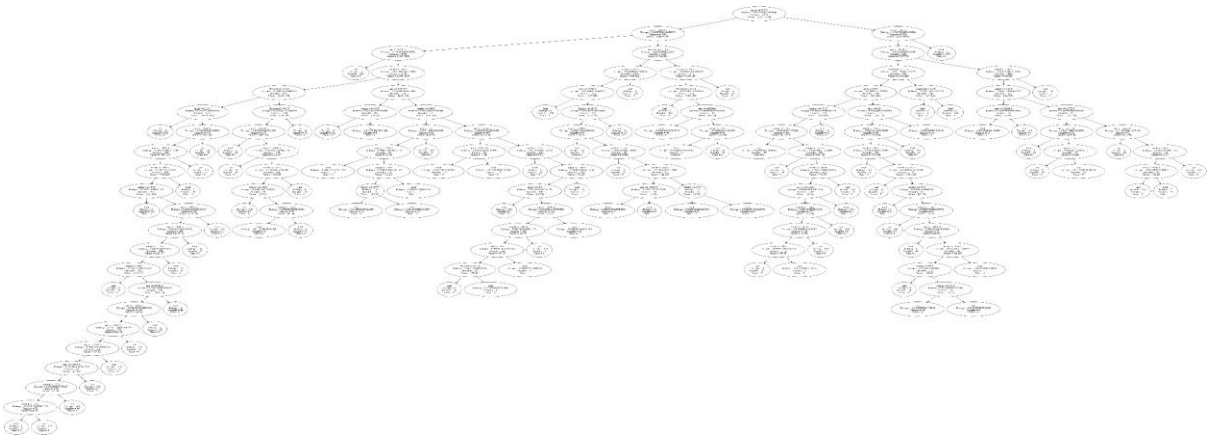
     Tree:

  2) After post-pruning

     ```
     Accuracy,Macro Precision, Macro Recall After Post Pruning
     (0.7045, 0.714059934318555, 0.7091239313454294)
     ```

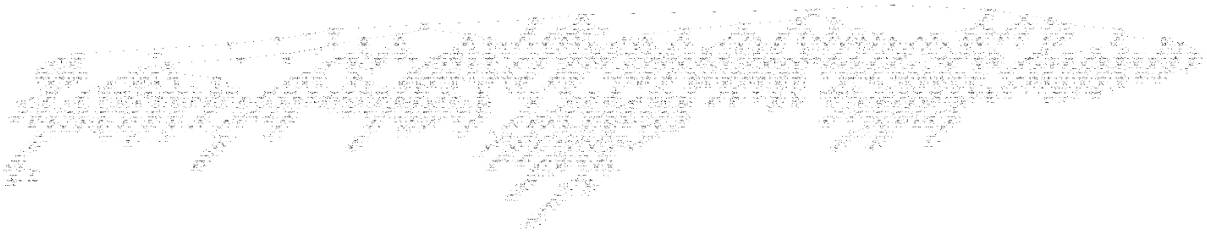
     Tree:

- Evaluation results for noisy dataset cardio_noise.csv
  1) Before pruning

```
Accuracy,Macro Precision, Macro Recall Before Pruning
(0.5120833333333333, 0.512072447186951, 0.5120687835243987)
```

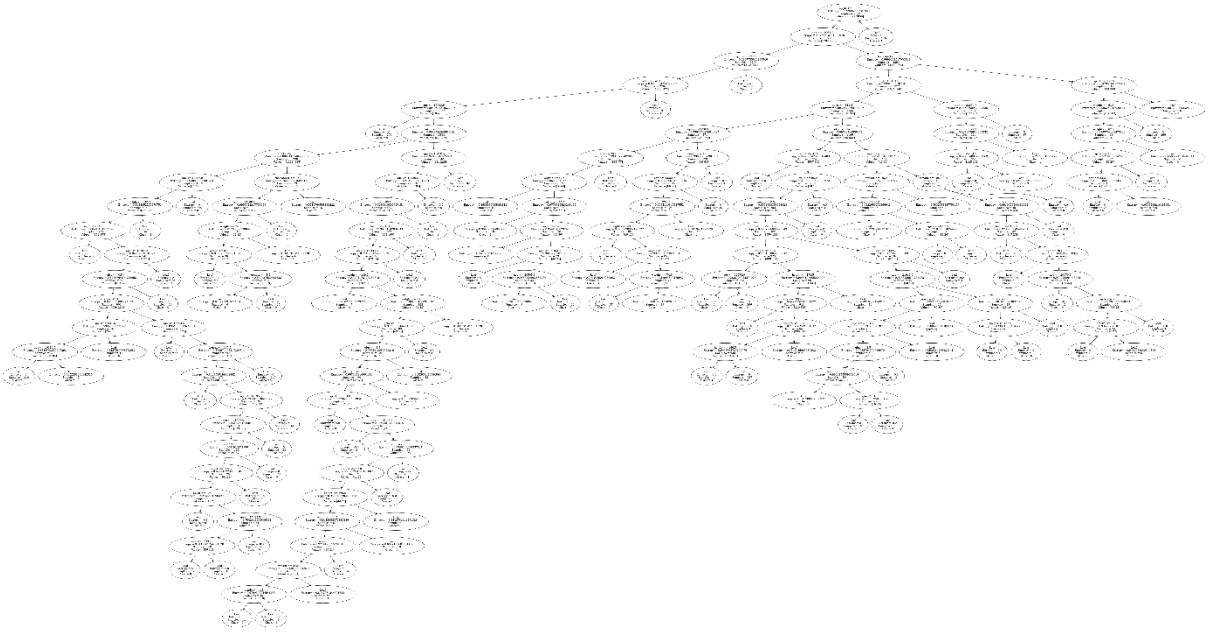
Tree:



  2) After post-pruning

```
Accuracy,Macro Precision, Macro Recall After Post Pruning
(0.615, 0.6151001112347052, 0.6149725415903933)
```

Tree:



<u>Techniques used to minimize difference between non-noisy and noisy accuracies:</u>

1) Pre-pruning implemented while building the decision tree so as to increase accuracy before post-pruning on noisy dataset (as this tries to prevent overfitting).
2) Minimum number of data points in leaf nodes restricted to 20 while pre-pruning.
3) Nodes were not split if information gain was less than 0.001.
4) Max height of tree was restricted to 100 by default.

<u>Comparison and Impact of Noise:</u>

- **Performance Comparison:** The accuracy, precision, and recall metrics clearly show that the model performs significantly better on the non-noisy dataset compared to the noisy dataset, both before and after pruning. Noise introduced variability that affected the model's ability to generalize from the training data to test data.
- **Impact of Noise:**

1) Noise in the dataset caused the decision trees to overfit during training, resulting in poor performance on test data. This is evident from the lower accuracy and less stable precision and recall scores on the noisy dataset compared to the non-noisy dataset.
2) It also increased the complexity of the decision tree, given the fact that the tree made using the non-noisy dataset had 1393 nodes, while the one made using the noisy dataset had 1971 nodes (this is before post-pruning).

Key Findings and Implications:

- **Findings:** Post-pruning significantly improves model performance on both datasets, but noise reduces the effectiveness of pruning, particularly evident in the accuracy drop between noisy and non-noisy datasets.
- **Implications:** It's crucial to preprocess datasets to minimize noise and choose appropriate pruning strategies to enhance model robustness.