# 2023 MSAI 349 Machine Learning Assignment 1 Report

JeongYoon Lee, Hannah Simmons, Rohan Gupta

1. (2.0points) Did you alter the Node data structure? If so, how and why?
    a. We did not change the Node data structure but we did add 2 class functions that 1- is_leaf() : indicate whether a node is a leaf and 2- make_leaf() : create a leaf from an intermediate node. We added these functions to aid in the pruning process.

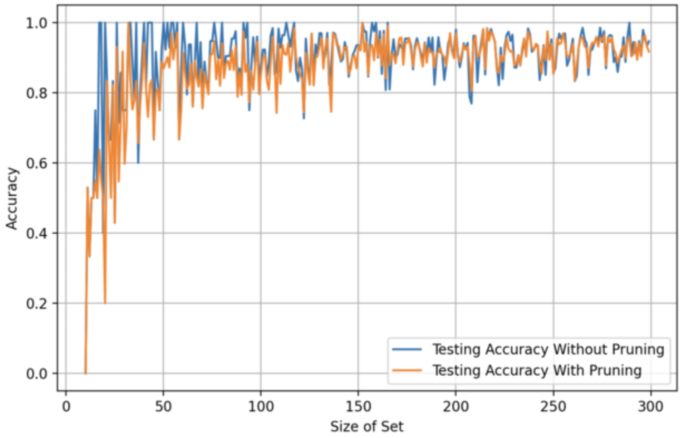
2. (2.0points) How did you handle missing attributes, and why did you choose this strategy?
    a. Instead of removing examples if they had a missing attribute, we substituted in the most common value in the dataset for that attribute. We chose this method because the percentage of missing examples was not too high and there is a high chance that the missing data will look like the majority of the data. Additionally, this technique is cost efficient in terms of time and compute power.

3. (2.0points) How did you perform pruning, and why did you choose this strategy?
    a. We used the reduced error pruning method because it seemed to be more intuitive than the critical value method and more directed than error complexity pruning.

4. (4.0points) Now you will try your learner on the house_votes_84.data, and plot learning curves. Specifically, you should experiment under two settings: with pruning, and without pruning. Use training set sizes ranging between 10 and 300 examples. For each training size you choose, perform 100 random runs, for each run testing on all examples not used for training (see testPruningOnHouseData from unit_tests.py for one example of this). Plot the average accuracy of the 100 runs as one point on a learning curve (x-axis = number of training examples, y-axis = accuracy on test data). Connect the points to show one line representing accuracy *with* pruning, the other *without*. Include your plot in your pdf, and answer two questions:
    a. What is the general trend of both lines as training set size increases, and why does this make sense?

   i. As the training set size increases, both lines converge asymptotically towards 0.95.
   ii. This makes sense because we would expect the models to be more accurate when they are exposed to more examples in the training phase.
   iii. The example distributions of smaller training sets run an increased risk of not being completely representative of the population distribution.
   iv. This would limit the generalizability of the model as it exits the training phase.
   v. Increasing the training dataset size counteracts this and increases the likelihood of the model gaining a better understanding of the population it's trying to model.

b. How does the advantage of pruning change as the dataset size increases? Does this make sense, and why or why not?
   i. As the dataset size increases, we find that the advantage of pruning decreases over using the full model.
   ii. Since we are using the validation dataset to compare the accuracy of each partial tree, each measurement will be more robust as the size of the validation dataset increases.
   iii. As there are only ~400 examples in the house_votes_84.data and we are using 25% of that data as validation dataset, we cannot confidently say that pruning specific nodes is better for the whole model, especially when the size of the training dataset is small.



[Graph 1. Test Accuracy With Pruning vs Without Pruning]

*Note: depending on your particular approach, pruning may not improve accuracy consistently or may decrease it (especially for small data set sizes). You can still receive full credit for this as long as your approach is reasonable and correctly implemented.*

5. (optional 2.0 points) Use your ID3 code to construct a Random Forest classifier using the candy.data dataset. You can construct any number of random trees using methods of your choosing. Justify your design choices and compare results to a single decision tree constructed using the ID3 algorithm.

N/A