# MSAI 349 Problem Set 1

**Amit Adate**

**October 18, 2018**

## Introduction

This is a brief report that elaborates upon the work done towards submission for the programming assignment 1 for the course EECS349 - machine learning taught by Prof. Douglas Downey at Northwestern University. The code is hyper - Linked Here

## Which other student, if any, is in your group?

The code is a collaboration between Mayank Malik ( NetID - MME3023 ) and Amit Adate ( NetID - ASA5078 )

## Did you alter the Node data structure? If so, how and why?

Yes, the node data structure was altered by adding 4 data types, **list of categories** - in order to store the distinct labels for every feature in the input list to the node, **mode** - in order to calculate the class label that is the most dominant, **mode_perc** - proportion of the mode among all the class labels, **isleaf** - this is a boolean data type that checks whether the node is a leaf or not.

## How did you handle missing attributes, and why did you choose this strategy?

We started with processing missing values into the dominant branch, the accuracy we obtained was in the range of 92.5\% - 93.5\%. Presently we are processing missing values as a separate class label. Presently, the accuracy we obtain is in the range of 93.5\% -95.5\%. Based on the uptick in accuracy, we have chosen the later

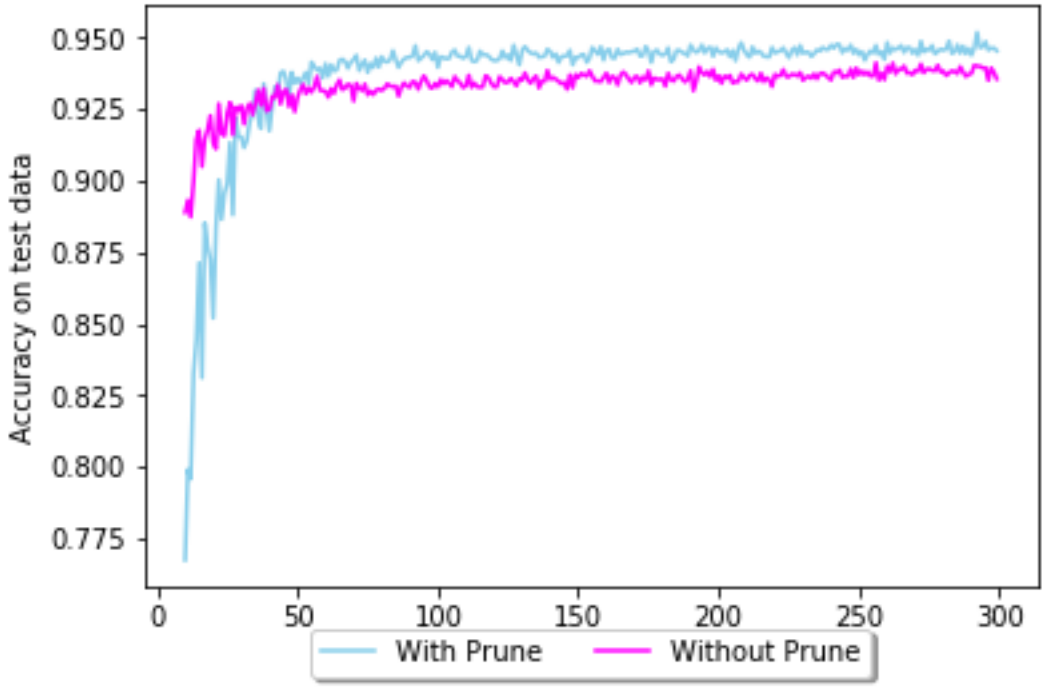
## How did you perform pruning, and why did you choose this strategy?

We traversed through the leaf nodes created after training and evaluated how the accuracy was impacted by removing them, if the accuracy is improved, we pruned

the node. Further, we continued looking through other leaf nodes. Our strategy is based on reduced error pruning - Linked Here.

We chose reduced error pruning because it is a popular pruning technique and easier to implement to its counterparts. Also, a version of reduced error pruning was elaborated in the slides by Prof. Downey and discussed briefly during the lectures. - Linked Here

## Learning Curve




**In about a sentence, what is the general trend of both lines as training set size increases, and why does this make sense?**

As the training set increases, the accuracy of the decision tree with pruning overtakes the accuracy of the tree without pruning.

**how does the advantage of pruning change as the data set size increases? Does this make sense, and why or why not?**

As the tree gets pruned, the nodes which provide little to no information about classifying the target are removed. Yes, this strategy makes sense as during pruning, model is becoming less complex and hence there is an observed improvement in accuracy. Pruning tries to reduce overfitting.