Writing Decision Tree (CART) From Scratch

Aeon Williams, CS398F2020

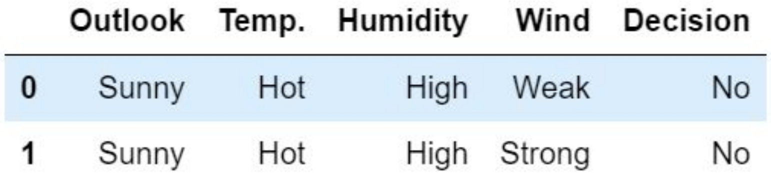
The Code

My decision tree takes a training dataset, and starts by determining the root node by finding the best split. Best split is quantified as having the the smallest Gini Index. Next, nodes are created recursively by evaluating the best split, and determining if there are terminal nodes based on the stopping conditions.

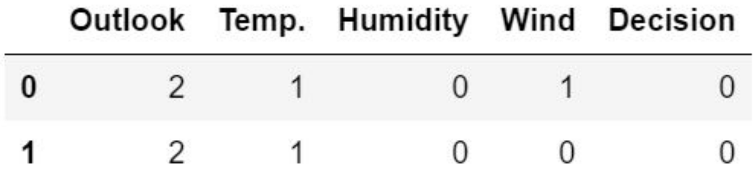
Predictions are made by navigating the tree, and finding the value of the terminal node that is hit.

Tennis Dataset

The first dataset evaluated with the model is an encoded tennis dataset, predicting Decision.



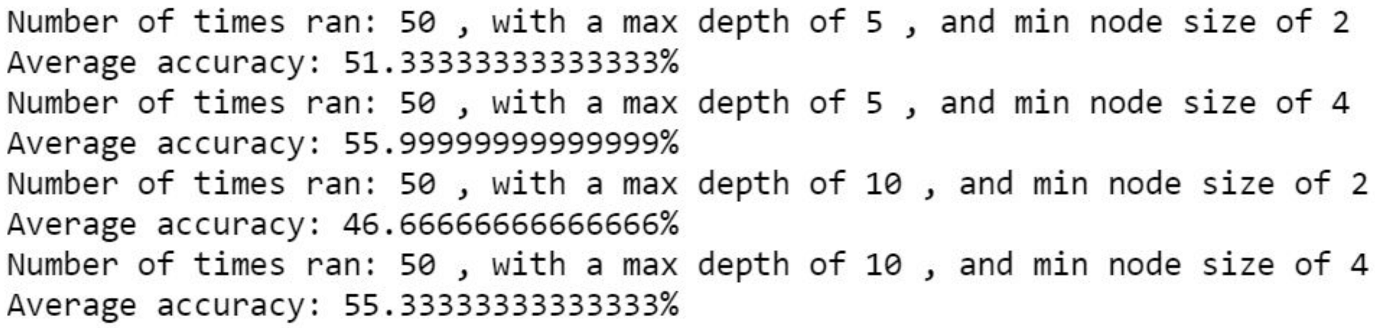
	Outlook	Temp.	Humidity	Wind	Decision
0	Sunny	Hot	High	Weak	No
1	Sunny	Hot	High	Strong	No



	Outlook	Temp.	Humidity	Wind	Decision
0	2	1	0	1	0
1	2	1	0	0	0

Accuracy

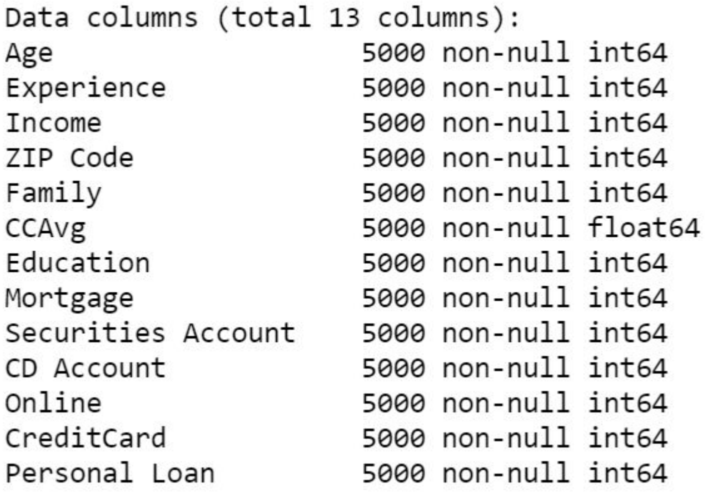
This model was not very accurate. I suspect this might be because of low sample size for training data and/or improper encoding. I would not trust it's predictions.



```
Number of times ran: 50 , with a max depth of 5 , and min node size of 2
Average accuracy: 51.33333333333333%
Number of times ran: 50 , with a max depth of 5 , and min node size of 4
Average accuracy: 55.99999999999999%
Number of times ran: 50 , with a max depth of 10 , and min node size of 2
Average accuracy: 46.66666666666666%
Number of times ran: 50 , with a max depth of 10 , and min node size of 4
Average accuracy: 55.33333333333333%
```

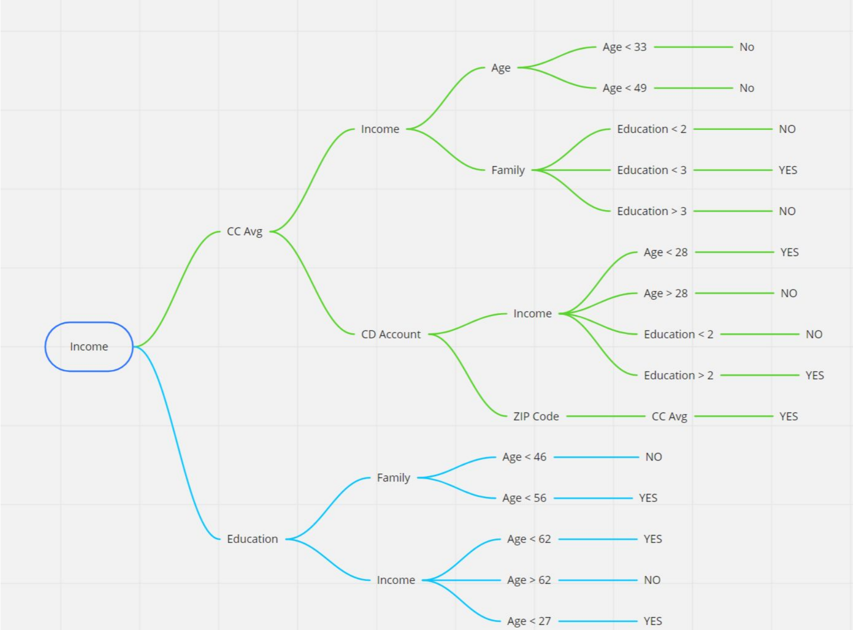
Personal Loan Dataset

This was the next dataset evaluated, to predict Personal Loan.



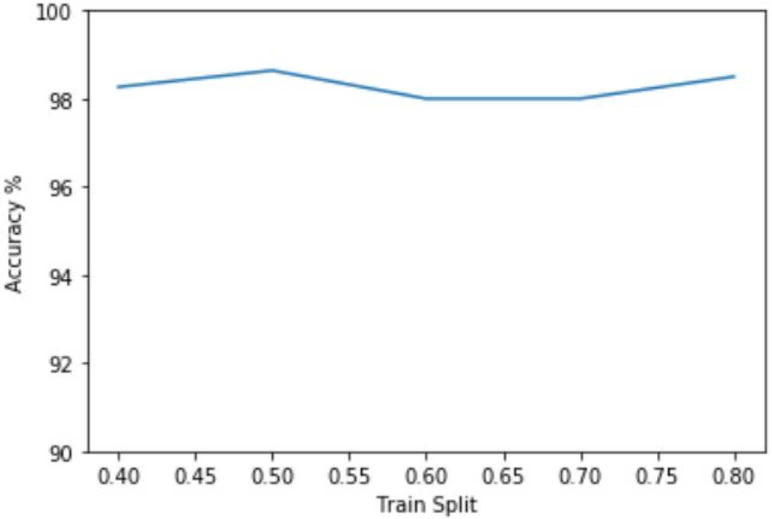
```
Data columns (total 13 columns):
Age                      5000 non-null int64
Experience               5000 non-null int64
Income                   5000 non-null int64
ZIP Code                 5000 non-null int64
Family                   5000 non-null int64
CCAvg                    5000 non-null float64
Education                5000 non-null int64
Mortgage                 5000 non-null int64
Securities Account       5000 non-null int64
CD Account               5000 non-null int64
Online                   5000 non-null int64
CreditCard               5000 non-null int64
Personal Loan            5000 non-null int64
```

Sample Graph of Personal Loan Tree



Accuracy

Overall, this model was very accurate, averaging around 97%.

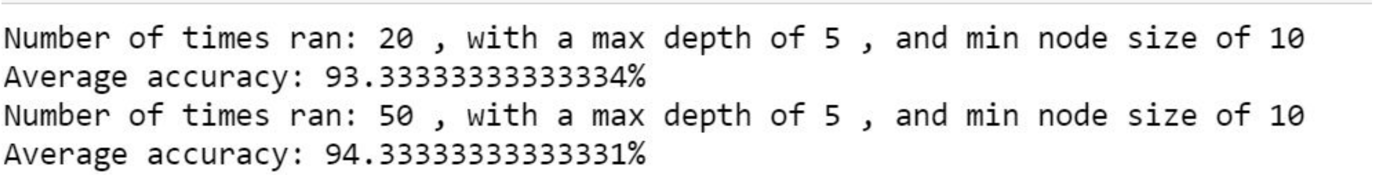


Next Steps

The only “pre-pruning” methods used are limiting the maximum depths, and ensuring a minimum number of samples per node. Additional pre and post pruning techniques would remove some of the unnecessary splitting happening in the model currently.

Bonus: Iris Dataset

To determine if the personal loan tree was actually accurate, and not overfit, I tested the model on an additional numerical dataset. The accuracy was still above 90%, so I consider the model to be legitimately good.



```
Number of times ran: 20 , with a max depth of 5 , and min node size of 10
Average accuracy: 93.33333333333334%
Number of times ran: 50 , with a max depth of 5 , and min node size of 10
Average accuracy: 94.33333333333331%
```