Journal of Machine Learning Research 3 (2022) 12 Submitted 3/2022; Published

Implementation of ID3 and CART for Classification and Regression

William S. Ventura WVENTUR1@JHU.EDU
Department of Computer Science
Johns Hopkins University
Baltimore, VA 21218-2682, USA

Editor: William S. Ventura

Abstract

This paper looks a look at both the ID3 and CART algorithm. These decision tree algorithms belong to the family of supervised learning algorithms. In this paper, the Iterative Dichotomiser 3 (ID3) and Classification and Regression Tree (CART) will be implemented to handle classification and regression tasks.The ID3 algorithm will be using gain-ratio as the splitting criterion, while the CART algorithm will be using the mean square error and root mean square error as the splitting criterion. They will then be tested on six data sets from the UCI Machine Learning Repository.

Keywords: ID3, CART, Decision Trees

1. Introduction

Decision trees are non-parametric supervised learning algorithms. This paper will take a look at the Iterative Dichotomiser (ID3) and the Classification and Regression Tree (CART). The ID3 decision tree was developed in Ross Quinlan in 1986. It is a top-down greedy approach and it is mainly used for only classification problems(Quinlan, 1986). While the ID3 can only handle classification problems, the CART algorithm is able to do classification and regression tasks. Having been officially proposed in 1984 by Breiman, Stone, Friedman and Olshen, it became one of the most widely used methods for decision tree analytics. Breiman et al. (1984) The main difference between the two algorithms is that while most tree algorithms are binary trees the ID3 became the first algorithm to propose more than two questions for each answer. It however was not the most ideal as ID3 had some disadvantages such as multi-value bias and overfitting. This paper will cover the various methods behind the ID3 and CART algorithms for classification and regression. It is in hopes that by implementing these methods it is possible to build a decision tree algorithm with accurate predicting capacity. Further more comparing the accuracy results of a pruned decision tree versus a fully grown decision tree. Similarly to the *Condensed Nearest Neighbor* Algorithm, pruning is a data compression technique hopes to reduce the size of the decision tree without affecting the accuracy and decreasing the processing speed. However decision trees, are prone to overfitting against the training data and accuracy decreases against unseen data. This paper hypothesizes that the decision trees will perform well on seen data, but on larger unseen data the decision trees will perform worse.

2. Methods

2.1 Entropy

The ID3 algorithm works by selecting the best feature at each step in building the tree. Entropy is the measure of disorder and the entropy of a dataset is the measure of disorder in the target feature of the dataset. Entropy is defined as:

$$H(S) = -\sum_{x \in X} p(x) log_2 p(x)$$

The dataset is denoted as S, where the entropy of the dataset would be the measure of disorder in the target feature of the dataset. Entropy is very important especially in the ID3 algorithm, where entropy is used to calculate the information gain. If H(S) = 0 then the dataset S is pure and perfectly classified.

2.2 Information Gain

Information gain is used by the ID3 algorithm to select the best feature. Information gain measures how well a given attribute classifies the target class and calculates the reduction in the entropy. Information Gain is defined as:

$$IG(S,A) = H(S) - \sum_{t \in T} p(t) H(t)$$

$$IG(S,A) = H(S) - H(S—A)$$

Where *H(S)* is the entropy of the dataset S, T is the subsets created from splitting the data set S at the attribute A , and *H(t)* is the entropy of the subset. The attribute with the highest information gain is selected as the best one.

2.3 Mean Square Error

In addition to having entropy as the splitting criterion for the CART algorithm. For the regression tasks the Mean Square Error (MSE), and Root Mean Square Error (RSME) will be used. They are defined as such:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y - \hat{Y})^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y - \hat{Y})^2}$$

2.4 ID3

The Iterative Dichotomiser 3 (ID3) algorithm was developed by Ross Quinlan in 1986 and is a top down greedy approach. The ID3 Algorithm classifies the data using the attributes

and the tree consists of decision nodes and leafs. What separates the ID3 from the CART Algorithm is that Nodes can have two or more branches.The ID3 algorithm follows principles of Occam's Razor and such its inductive bias lies in forming a generalization beyond the training instances. Like many other decision trees it is prone to overfitting and performance decreases when tested against unseen data.

Using the original dataset prior to any splits S, as the root node, the ID3 algorithm will iterate throughout the attributes of S. ID3 first calculates the system entropy $H(S)$. It then proceeds to calculate the entropy given an attribute $H(S|A)$, and final the information gained given that attribute $IG(S|A)$. The algorithm then selects the attribute with the largest $IG(S|A)$, and splits the dataset S, on that attribute to produce subsets. This continues recursively on each subset until there is no longer any remaining attributes, A.

Aside from tending to overfit the training data, ID3 has other limitations such as possibly arriving at a local optima and not guaranteeing the optimal solution. This can be attributed to the greedy strategy that ID3 implements when selecting the best attribute to split on.

2.5 CART

The Classification and Regression Tree (CART), was proposed in 1984 by Breiman, Stone, Friedman and Olshen, and rapidly became one of the most widely used methods for decision tree analytics. Where the ID3 algorithm is able to nodes with two or more branches, the CART algorithm builds a binary tree by repeatedly splitting the node into two child nodes.

Similarly to the ID3 the CART algorithm finds the best split based on the splitting criterion. This paper will mainly cover using Mean Square Error, MSE as the splitting criterion. This is continued until the binary tree is complete. Breiman et al. (1984)

2.6 Cross Validation and Hyper-Parameter Tuning

2.6.1 ID3 Classification Tasks

Some assumptions made beforehand about the data was that all the instances in the data had a random distribution and that these feature variables, x'_1, ... x'_n were independent of each other and identically distributed (IID). With those assumptions, the data was separated into two randomized subsets for extra precaution. The first of the two subsets, contained a randomized 20% of the original data set. This subset is defined as S_{tuning}, and will be used to tune hyper-parameters. The second subset, which contains the other 80% of the original data set, is defined as $S_{80\%}$. This subset is then saved to be tested on after finding the optimal parameters. Both the Tuning data set and the Testing data set underwent 5-fold cross validation testing for varying values of $MaxDepth$, $MinimumSplit$, and $MinimumGain$ depending on the UCI dataset. The combination of $MaxDepth$, $MinimumSplit$, and $MinimumGain$ values that resulted in lowest MSE in the training data were then used as the values when performing a 5-fold Cross Validation on the

testing data. These 5-fold cross validation were performed on unpruned and pruned trees. The evaluation metric for the ID3 Algorithm was a classification accuracy function.

2.6.2 CART Regression Task

The regression tasks went through the same basic procedures of preprocessing the data and assumptions as the classification data. Just like the classification tasks, the regression data sets were also split up into two subsets for tuning and testing. However unlike the classification task, the regression tasks only had two hyper parameters for tuning, $MaxDepth$, and $MinimumCriterion$. For the regression task, the $MaxDepth$ and $MinimumCriterion$ were tuned to find the combination that gave greatest performance on the tuning dataset, in this case the lowest MSE. Depending on the UCI dataset varying value combinations $MaxDepth$, and $MinimumCriterion$ were attempted to find the optimal parameters. The regression tasks also were evaluated through a 5-fold Cross Validation for both the pruned and unpruned tuning and testing decision trees. The respective hyper parameter values for the combination resulting in the minimal Mean Squared Error (MSE) was then selected to be used with the evaluation subset.

3. Results

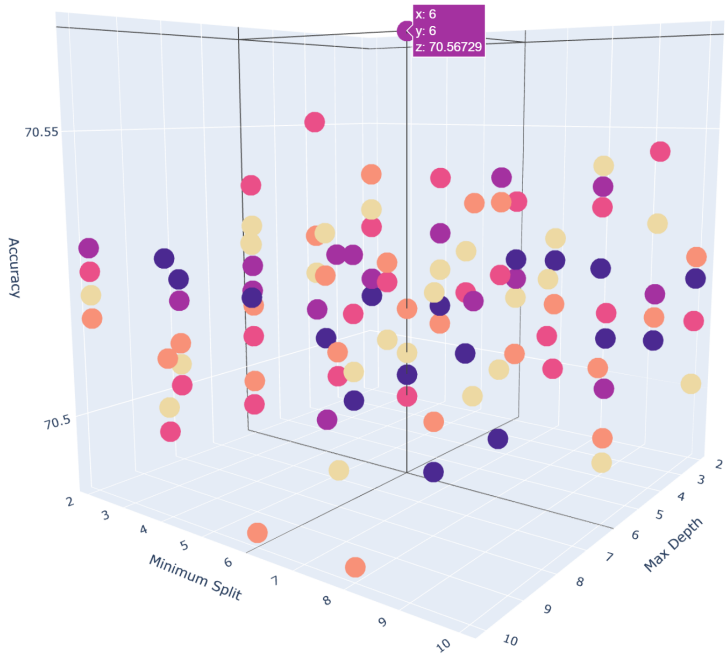
3.1 Car Classification via ID3

Table 1: Car Tuning Subset Pruning Results

Max Depth	Minimum Split	Minimum Gain	Accuracy
2	2	0.2	70.50931677018635
2	2	0.4	70.51345755693582
2	2	0.6	70.52587991718426
2	2	0.8	70.53830227743272
2	2	1	70.53002070393374
2	4	0.2	70.50931677018635
2	4	0.4	70.52587991718426
2	4	0.6	70.53830227743272
2	4	0.8	70.50517598343686
2	4	1	70.51759834368531
2	6	0.2	70.52173913043478
2	6	0.4	70.51759834368531
2	6	0.6	70.53416149068322
2	6	0.8	70.50103519668737
2	6	1	70.51345755693582
2	8	0.2	70.52173913043478
2	8	0.4	70.53830227743272
2	8	0.6	70.53416149068323
2	8	0.8	70.50103519668735
2	8	1	70.5424430641822
2	10	0.2	70.52173913043478
2	10	0.4	70.50103519668737
2	10	0.6	70.51345755693582
2	10	0.8	70.52587991718426
2	10	1	70.50103519668737
4	2	0.2	70.51759834368531
4	2	0.4	70.50931677018635
4	2	0.6	70.55072463768116
4	2	0.8	70.52587991718426
4	2	1	70.51759834368531
4	4	0.2	70.50517598343683
4	4	0.4	70.50517598343684
4	4	0.6	70.51759834368531
4	4	0.8	70.52173913043478
4	4	1	70.50517598343684
4	6	0.2	70.50517598343684
4	6	0.4	70.52587991718426
4	6	0.6	70.51759834368531

Please refer to the output file for complete Table 1 values.

Figure 1: Graph: Car Tuning Dataset Pruning Results (Max Depth vs. Minimum Split vs. Accuracy)

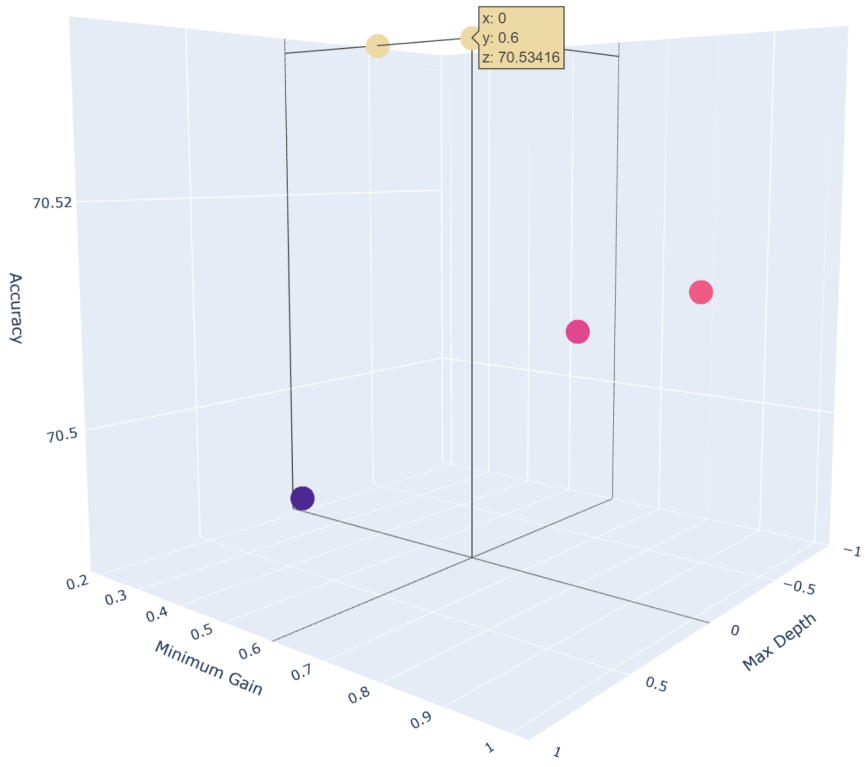


Highest Accuracy of 70.57%, when Max Depth and Minimum Split = 6 and Minimum Gain = 0.4

Table 2: Car Tuning Subset Non-Pruning Results

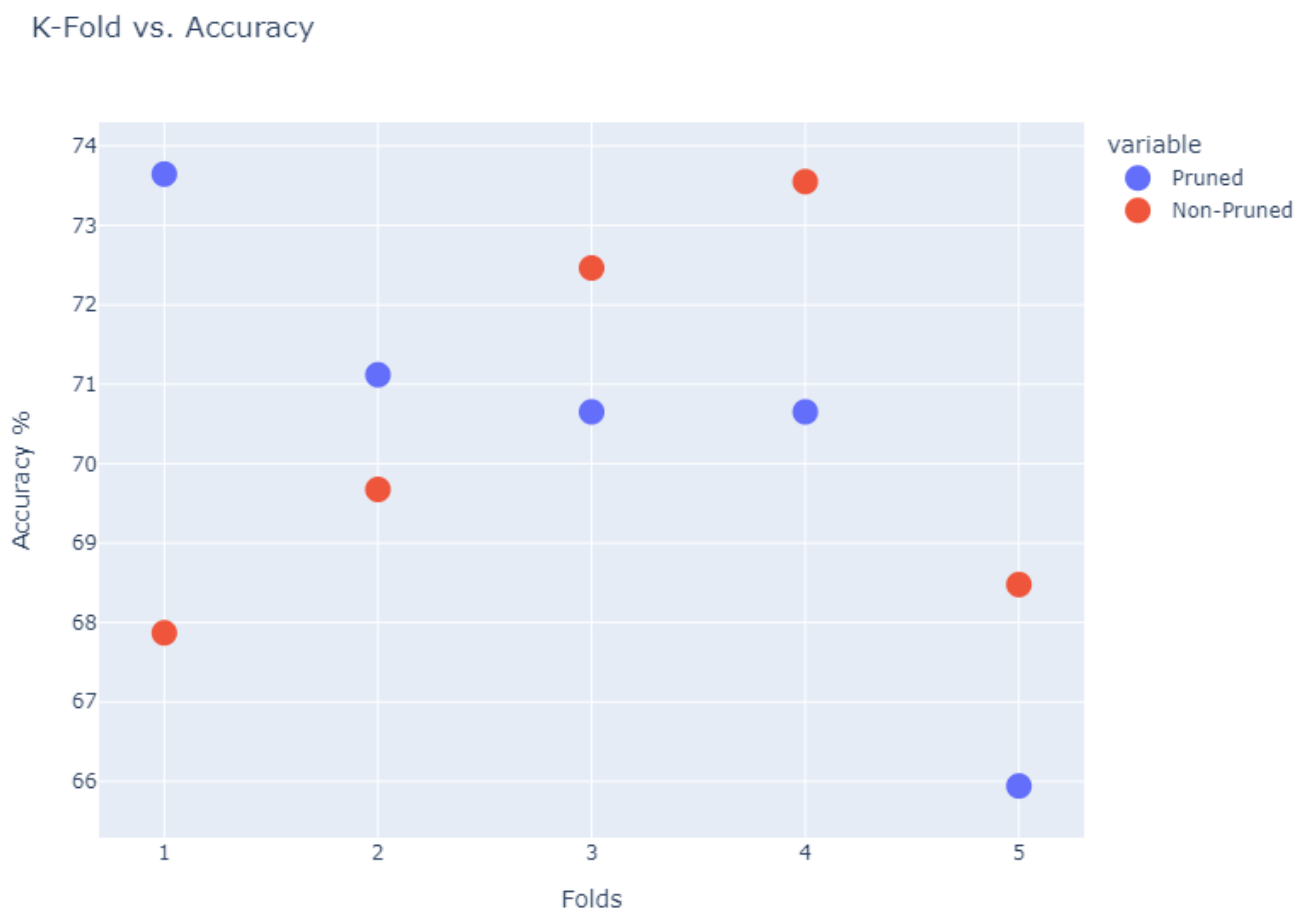
Max Depth	Minimum Split	Minimum Gain	Accuracy
0	0	0.2	70.48861283643892
0	0	0.4	70.53416149068323
0	0	0.6	70.53416149068323
0	0	0.8	70.50931677018635
0	0	1	70.51345755693583

Figure 2: Graph: Car Tuning Dataset Non-Pruning Results (Max Depth vs. Minimum Split vs. Accuracy)



Highest Accuracy of 70.534% achieved when Minimum Gain = 0.4 and 0.6

Figure 3: Graph: Car Testing Dataset Pruned and Unpruned Results (Fold vs. Accuracy)



Pruned Tree 5 Fold Cross Validation Average Accuracy : 70.40234395437662%

Unpruned Tree 5 Fold Cross Validation Average Accuracy : 70.40757599539582%

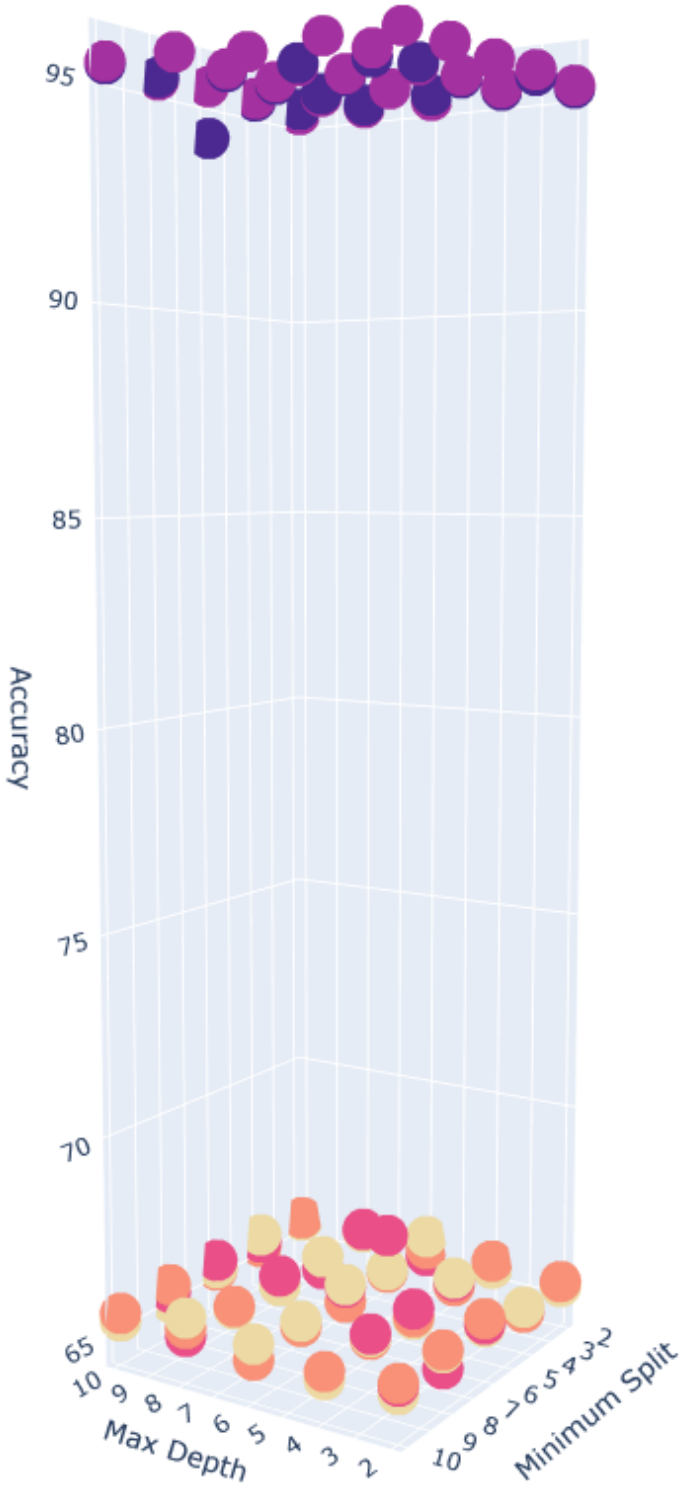
3.2 Voting Classification via ID3

Table 3: House Voting Tuning Subset Pruning Results

Max Depth	Minimum Split	Minimum Gain	Accuracy
2	2	0.2	95.29411764705883
2	2	0.4	95.35947712418302
2	2	0.6	65.42483660130719
2	2	0.8	65.42483660130718
2	2	1	65.29411764705881
2	4	0.2	95.29411764705883
2	4	0.4	95.42483660130719
2	4	0.6	65.42483660130719
2	4	0.8	65.29411764705883
2	4	1	65.42483660130718
2	6	0.2	95.35947712418302
2	6	0.4	95.35947712418302
2	6	0.6	65.68627450980392
2	6	0.8	65.81699346405229
2	6	1	65.62091503267973
2	8	0.2	95.42483660130719
2	8	0.4	95.42483660130719
2	8	0.6	65.29411764705881
2	8	0.8	65.75163398692811
2	8	1	65.68627450980392
2	10	0.2	95.42483660130719
2	10	0.4	95.42483660130719
2	10	0.6	65.62091503267973
2	10	0.8	65.68627450980392
2	10	1	65.42483660130719
4	2	0.2	95.42483660130719
4	2	0.4	95.49019607843137
4	2	0.6	65.55555555555556
4	2	0.8	65.55555555555556
4	2	1	65.359477124183
4	4	0.2	95.42483660130719
4	4	0.4	95.49019607843137
4	4	0.6	65.68627450980394
4	4	0.8	65.55555555555557
4	4	1	65.75163398692811
4	6	0.2	95.49019607843137
4	6	0.4	95.42483660130719
4	6	0.6	65.62091503267975
4	6	0.8	65.42483660130718
4	6	1	65.35947712418302
4	8	0.2	95.29411764705883
4	8	0.4	95.49019607843137
4	8	0.6	65.68627450980392
4	8	0.8	65.55555555555556
4	8	1	65.62091503267973
4	10	0.2	95.42483660130719
4	10	0.4	95.42483660130719
4	10	0.6	65.29411764705883
4	10	0.8	65.49019607843135
4	10	1	65.29411764705883
6	2	0.2	95.49019607843138
6	2	0.4	95.35947712418302
6	2	0.6	65.29411764705881
6	2	0.8	65.49019607843137
6	2	1	65.81699346405229
6	4	0.2	95.35947712418302
6	4	0.4	95.35947712418302

Please refer to the output file for complete Table 3 values.

Figure 4: Graph: House Voting Tuning Dataset Pruning Results (Max Depth vs. Minimum Split vs. Accuracy)

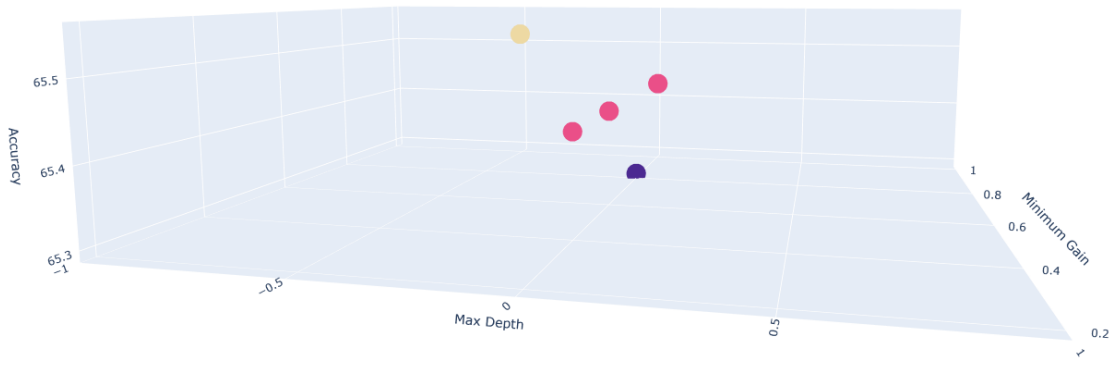


Highest Accuracy of 95.49%, when Max Depth= 6, Minimum Split = 2 and Minimum Gain = 0.2

Table 4: House Voting Tuning Subset Non-Pruning Results

Max Depth	Minimum Split	Minimum Gain	Accuracy
0	0	0.2	65.55555555555556
0	0	0.4	65.42483660130719
0	0	0.6	65.42483660130718
0	0	0.8	65.29411764705883
0	0	1	65.42483660130719

Figure 5: Graph: House Voting Tuning Dataset Non-Pruning Results (Max Depth vs. Minimum Split vs. Accuracy)



Highest Accuracy of 65.556% achieved when Minimum Gain = 0.2

Figure 6: Graph: House Voting Testing Dataset Pruned and Unpruned Results (Fold vs. Accuracy)

Pruned Tree 5 Fold Cross Validation Average Accuracy : 95.12215320910974%

Unpruned Tree 5 Fold Cross Validation Average Accuracy : 69.2497253178465%

3.3 Breast Cancer Classification via ID3

Table 5: Breast Cancer Tuning Subset Pruning Results

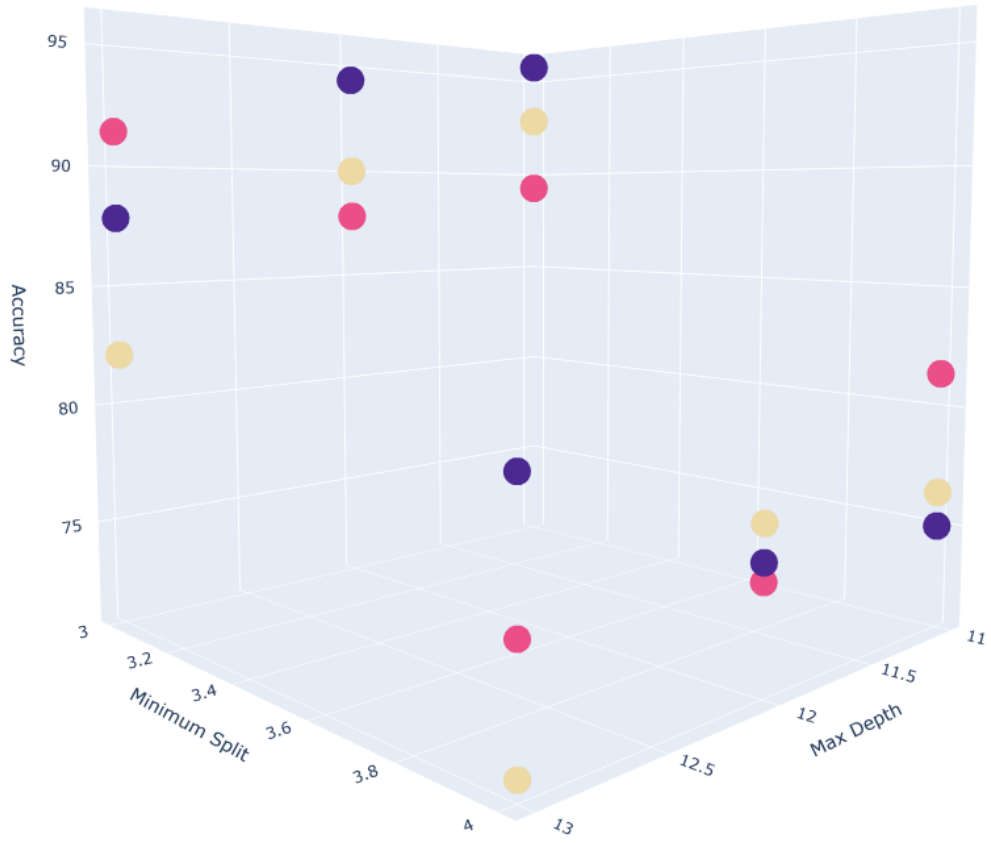
Max Depth	Minimum Split	Minimum Gain	Accuracy
11	3	0.2	95.71428571428572
11	3	0.3	89.28571428571429
11	3	0.4	92.85714285714285
11	4	0.2	75
11	4	0.3	81.42857142857142
11	4	0.4	76.42857142857144
12	3	0.2	94.28571428571429
12	3	0.3	87.85714285714286
12	3	0.4	90.00000000000001
12	4	0.2	75.71428571428571
12	4	0.3	75
12	4	0.4	77.14285714285714
13	3	0.2	87.85714285714286
13	3	0.3	91.42857142857144
13	3	0.4	82.14285714285715
13	4	0.2	80.71428571428571
13	4	0.3	75.71428571428571
13	4	0.4	71.42857142857143

Table 6: Breast Cancer Tuning Subset Non-Pruning Results

Max Depth	Minimum Split	Minimum Gain	Accuracy
0	0	0.2	70.71428571428572
0	0	0.4	70.71428571428572
0	0	0.6	70.71428571428571
0	0	0.8	70.71428571428571
0	0	1	70.71428571428571

In the Unpruned Tree, Minimum Information hyper-parameter did not affect change in accuracy

Figure 7: Graph: Breast Cancer Tuning Dataset Pruning Results (Max Depth vs. Minimum Split vs. Accuracy)



Highest Accuracy of 95.7143%, when Max Depth= 11, Minimum Split = 3 and Minimum Gain = 0.2

Figure 8: Graph: Breast Cancer Testing Dataset Pruned and Unpruned Results (Fold vs. Accuracy)

Pruned Tree 5 Fold Cross Validation Average Accuracy : 91.23552123552125%

Unpruned Tree 5 Fold Cross Validation Average Accuracy : 69.90111442473709%

3.4 Abalone Regression via CART

Table 7: Abalone Tuning Subset Pruning Results

Max Depth	Minimum Criterion	MSE
1	0.05	8.46132335191726
1	0.1	9.67807976078908
1	0.15	7.283756713467984
1	0.2	11.4438557660621
1	0.25	7.312925860348029
3	0.05	7.995196890067388
3	0.1	8.279577006652222
3	0.15	7.4519792135054175
3	0.2	6.437908537462682
3	0.25	6.07849167812727
5	0.05	6.585459486122514
5	0.1	8.867494697756523
5	0.15	8.695643246747997
5	0.2	6.141032984851419
5	0.25	6.9990435601255925
7	0.05	11.15862673947987
7	0.1	7.288892762778763
7	0.15	9.412721250026475
7	0.2	8.635022393255012
7	0.25	7.574471573188331
9	0.05	8.06721280552366
9	0.1	10.368695769617894
9	0.15	7.950878173417158
9	0.2	8.531907766231779
9	0.25	8.799846485804684

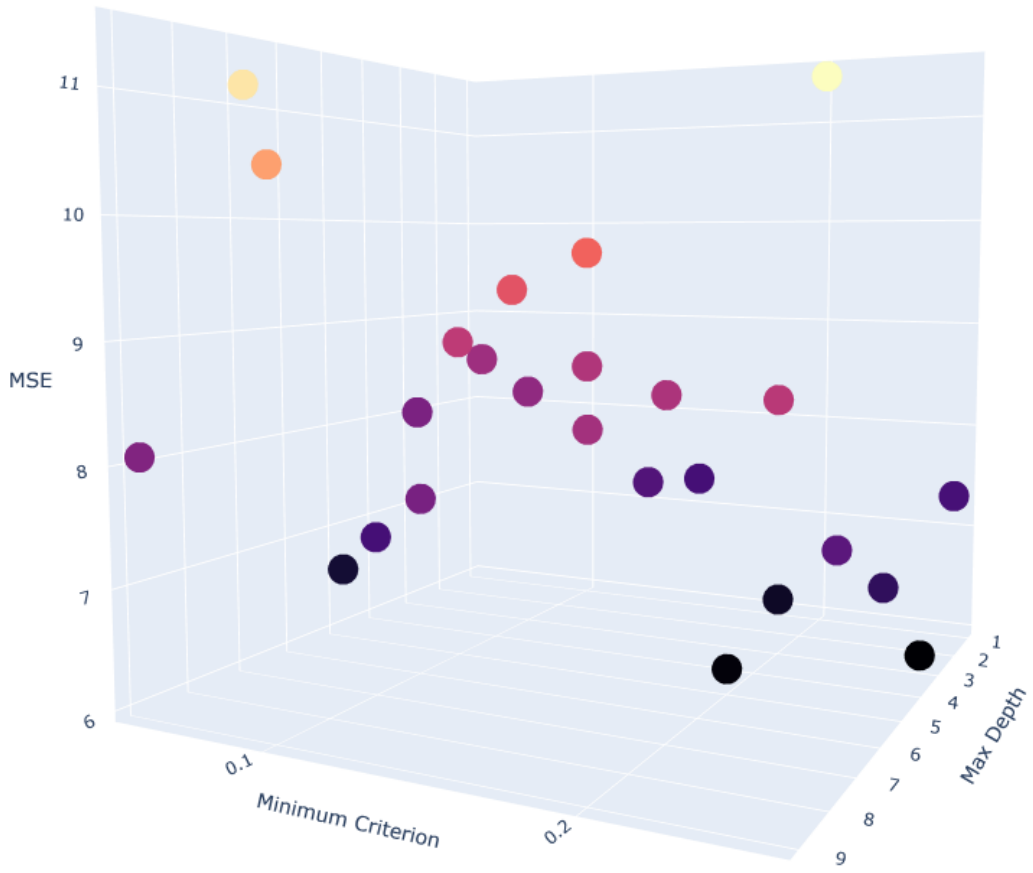
Lowest MSE is 6.078 when Max Depth = 3 and Minimum Splitting Criterion = 0.25

Table 8: Abalone Tuning Subset Non-Pruning Results

Max Depth	Minimum Criterion	MSE
0	0.05	12.44891672032336
0	0.1	7.746440618677559
0	0.15	8.31660751661586
0	0.2	5.572067327817604
0	0.25	8.228201127396948

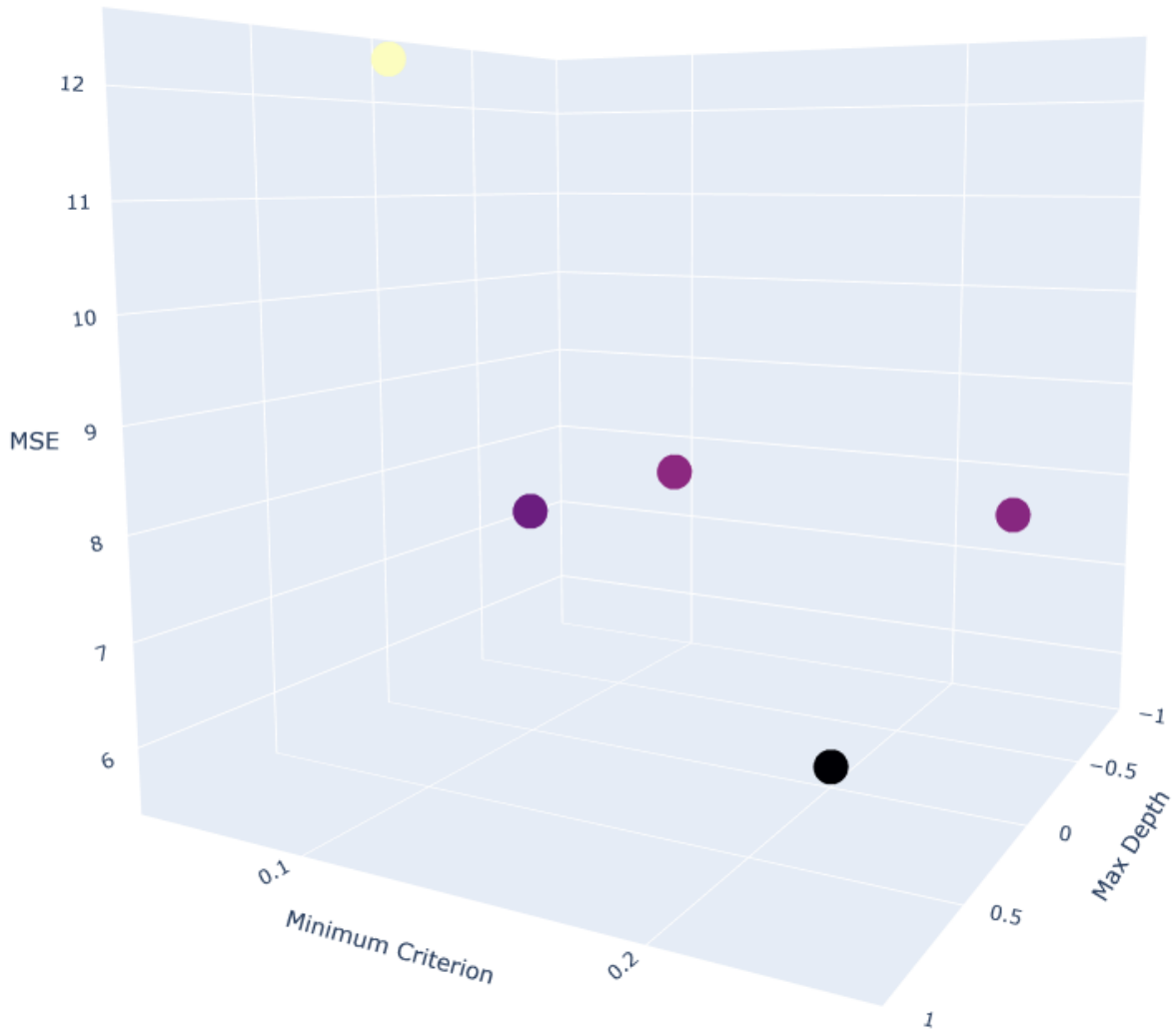
Lowest MSE is 5.572 when Minimum Splitting Criterion = 0.2

Figure 9: Graph: Abalone Tuning Pruned Tree Results (Max Depth vs. Minimum Criterion vs. MSE)



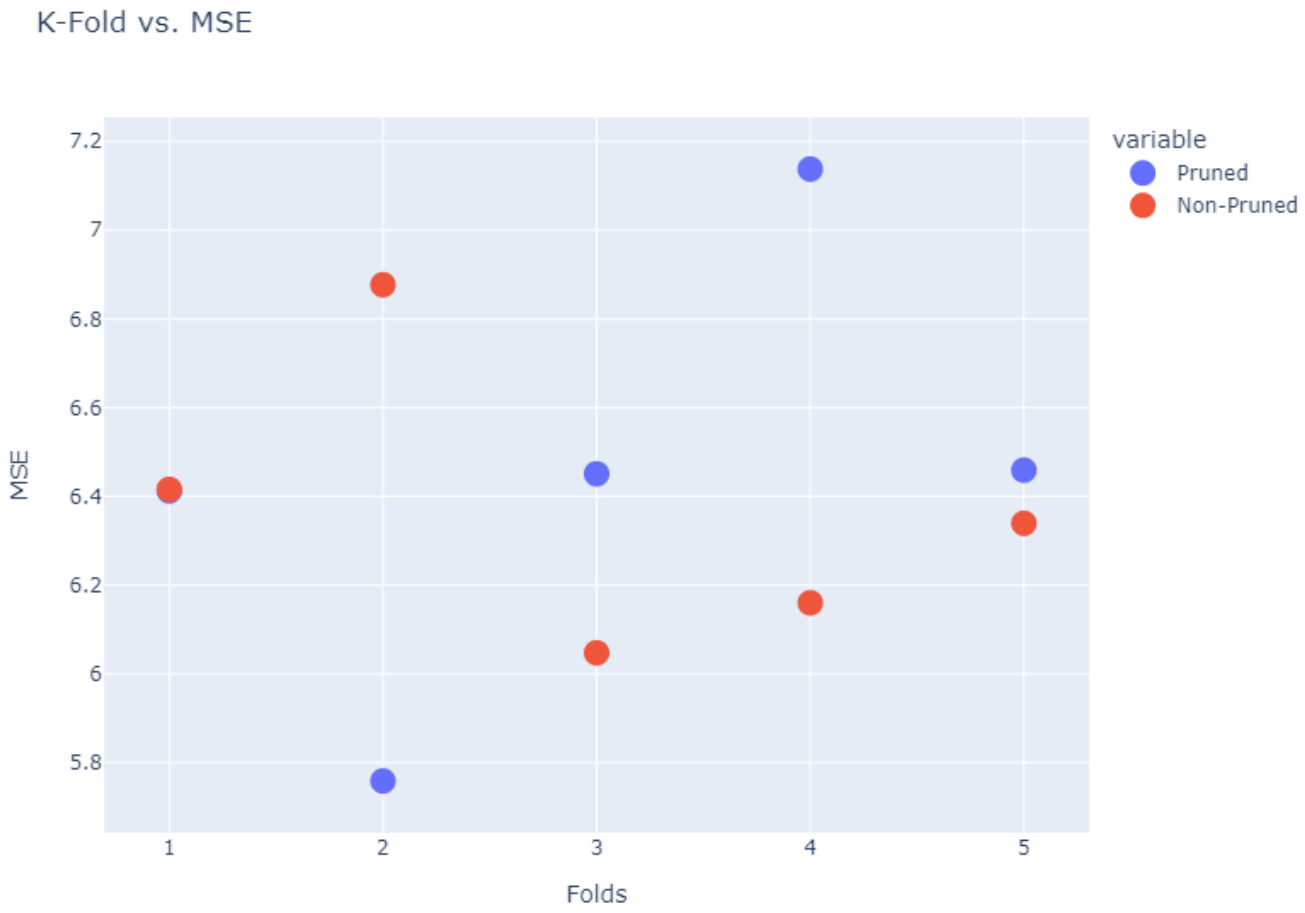
MSE of 6.078, when Max Depth= 3, Minimum Criterion = 0.25

Figure 10: Graph: Abalone Tuning Unpruned Tree Results (Max Depth vs. Minimum Criterion vs. MSE))



Lowest MSE is 5.572 when Minimum Splitting Criterion = 0.2

Figure 11: Graph: Abalone Testing Dataset Pruned and Unpruned Results (Fold vs. MSE)



Pruned Tree 5 Fold Cross Validation Average MSE : 6.4437

Unpruned Tree 5 Fold Cross Validation Average MSE : 6.3679

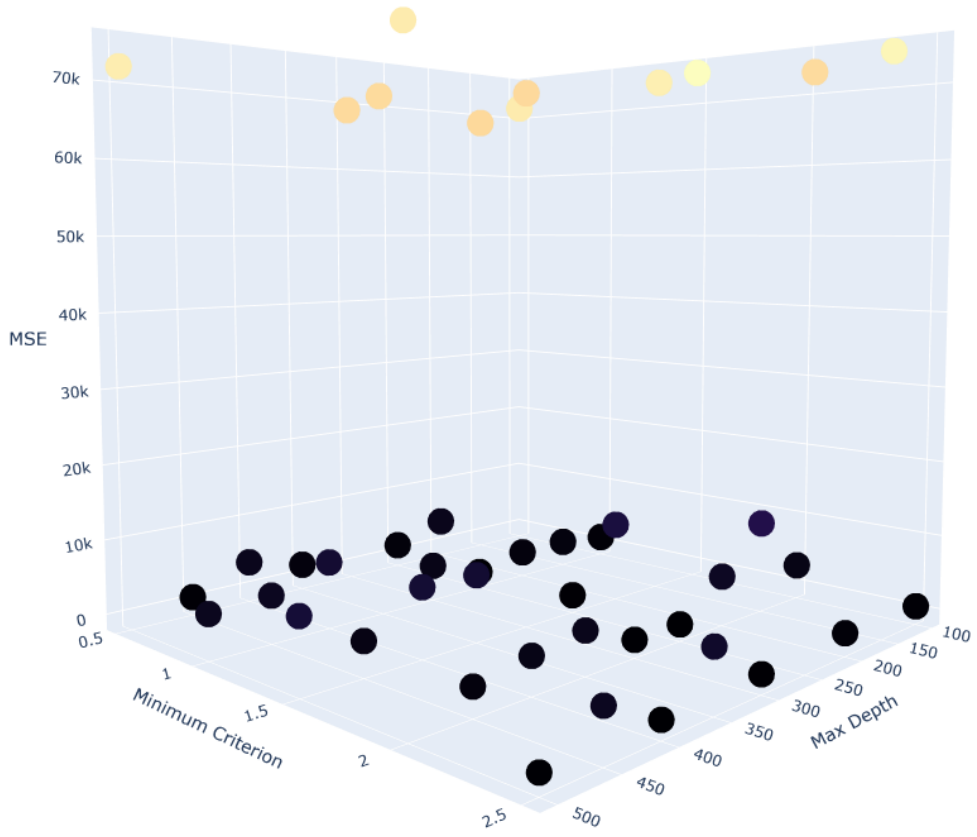
3.5 Computer Regression via CART

Table 9: Computer Tuning Subset Pruning Results

Max Depth	Minimum Criterion	MSE
100	0.5	71496.90625
100	1	740.8090277777777
100	1.5	74346.40625
100	2	2680.183159722222
100	2.5	410.21614583333337
150	0.5	68491.05902777778
150	1	1557.295138888889
150	1.5	72133.61722222222
150	2	10329.461805555555
150	2.5	73145.08420138889
200	0.5	3476.6944444444443
200	1	1536.765625
200	1.5	8739.862777777777
200	2	4626.40625
200	2.5	920.25
250	0.5	1156.57
250	1	232.73958333333331
250	1.5	360.51649305555554
250	2	163.38888888888886
250	2.5	68908.54781250001
300	0.5	68714.96527777778
300	1	2917.2256944444443
300	1.5	68515.125
300	2	272.81944444444446
300	2.5	362.36805555555554
350	0.5	1497.3472222222222
350	1	68782.15625
350	1.5	6783.435
350	2	3540.0138888888887
350	2.5	5764.847222222223
400	0.5	3583.6875
400	1	6862.697916666667
400	1.5	7063.6015625
400	2	2702.8803125000004
400	2.5	175.40625
450	0.5	421.08767361111114
450	1	4281.446180555556
450	1.5	2454.0138888888887
450	2	1525.3715277777778
450	2.5	4182.6849999999995
500	0.5	71803.69
500	1	3888.0720486111113
500	1.5	7479.378472222223
500	2	71485.03125
500	2.5	276.5277777777778

Lowest MSE is 163.388 when Max Depth = 250 and Minimum Splitting Criterion = 2.0

Figure 12: Graph: Computer Tuning Pruned Tree Results (Max Depth vs. Minimum Criterion vs. MSE)



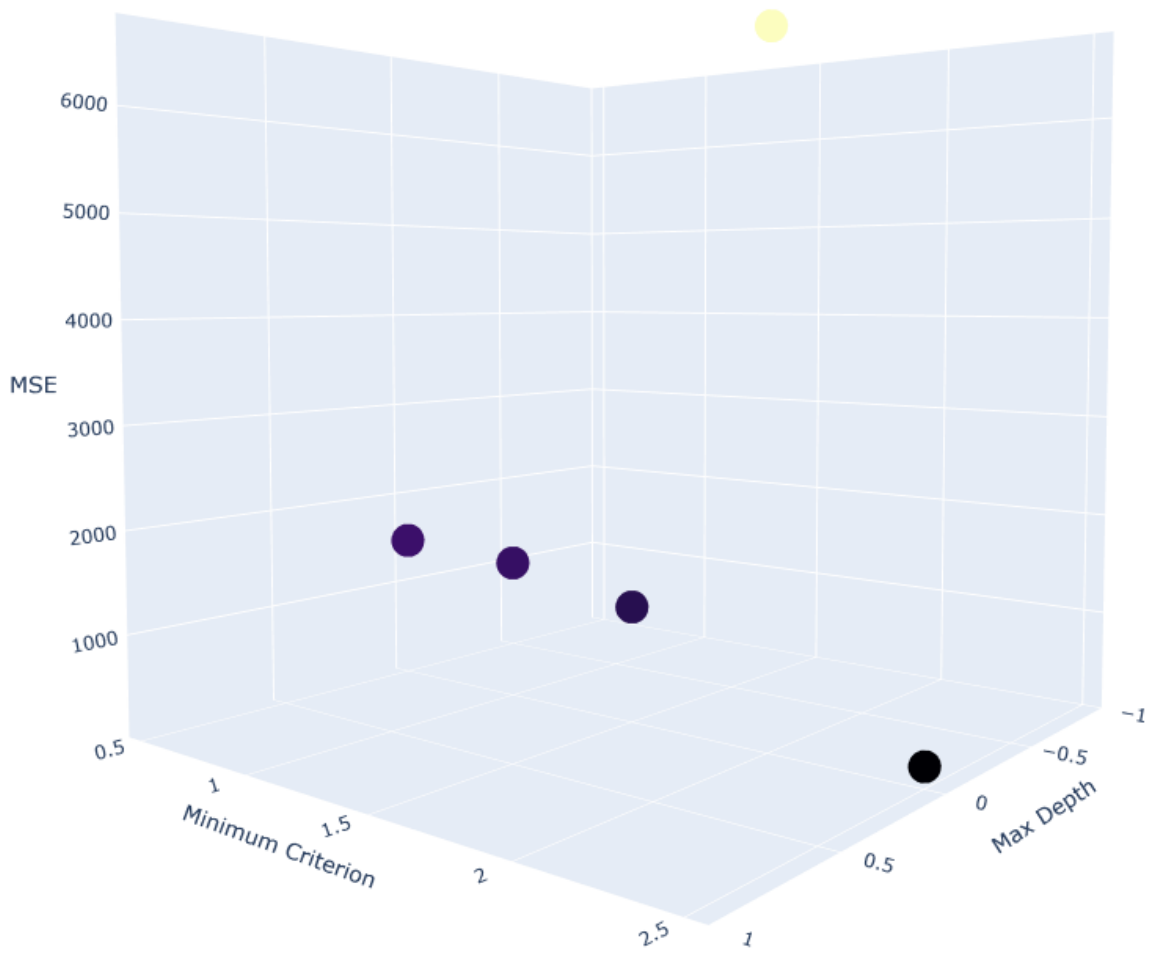
Lowest MSE is 163.388 when Max Depth = 250 and Minimum Splitting = 2.0

Table 10: Computer Tuning Subset Non-Pruning Results

Max Depth	Minimum Criterion	MSE
0	0.5	1494.740451388889
0	1	1422.90625
0	1.5	1167.9444444444446
0	2	6653.473125
0	2.5	213.1865625

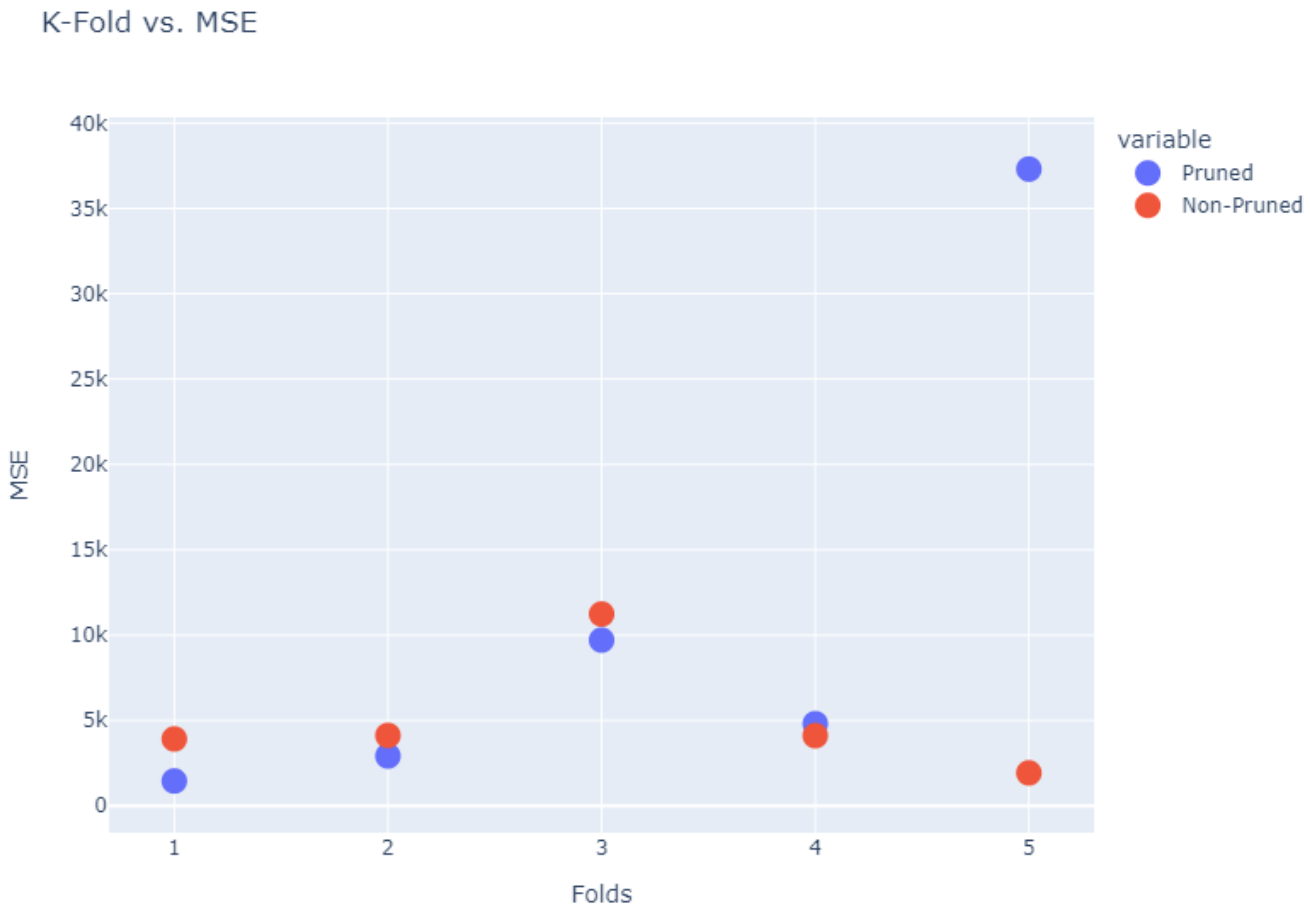
Lowest MSE is 213.187 when Minimum Splitting Criterion = 2.5

Figure 13: Graph: Computer Tuning Unpruned Tree Results (Max Depth vs. Minimum Criterion vs. MSE))



Lowest MSE is 213.187 when Minimum Splitting Criterion = 2.5

Figure 14: Graph: Computer Testing Dataset Pruned and Unpruned Results (Fold vs. MSE)



Pruned Tree 5 Fold Cross Validation Average MSE : 11241.0579

Unpruned Tree 5 Fold Cross Validation Average MSE : 5065.4900

3.6 Forest Regression via CART

Table 11: Forest Tuning Subset Pruning Results

Max Depth	Minimum Criterion	MSE
10	0.5	384.31982119712376
10	1	192.44110183061227
10	1.5	613.7758485000977
10	2	129.97360599619992
10	2.5	68.68169002220343
30	0.5	617.477249902408
30	1	50.51945445766721
30	1.5	144.663495212275
30	2	197.75415567590449
30	2.5	73.22418328075042
50	0.5	164.31760249753793
50	1	696.5052827889108
50	1.5	645.8187709787516
50	2	69.35806976112846
50	2.5	101.73113455972222
70	0.5	121.99350770699762
70	1	549.9794525051652
70	1.5	124.57710683823204
70	2	499.74978718759496
70	2.5	147.54533499026144
90	0.5	1115.2011291843853
90	1	84.39170997980145
90	1.5	414.15978399999995
90	2	545.0992736156925
90	2.5	29.254880028117917

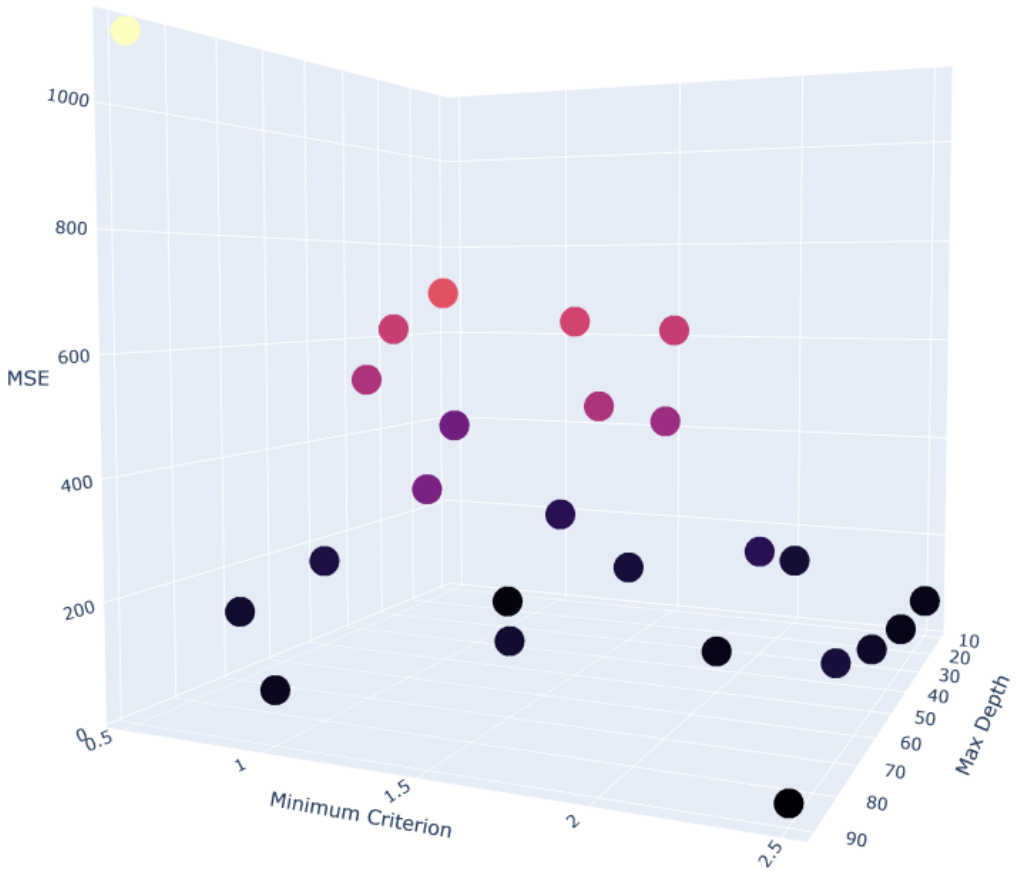
Lowest MSE is 29.2549 when Max Depth = 90 and Minimum Splitting Criterion = 2.5

Table 12: Forest Tuning Subset Non-Pruning Results

Max Depth	Minimum Criterion	MSE
0	0.5	201.79151771117606
0	1	1053.9052190965278
0	1.5	49.2940889616864
0	2	51.60587189424789
0	2.5	21.79443433479819

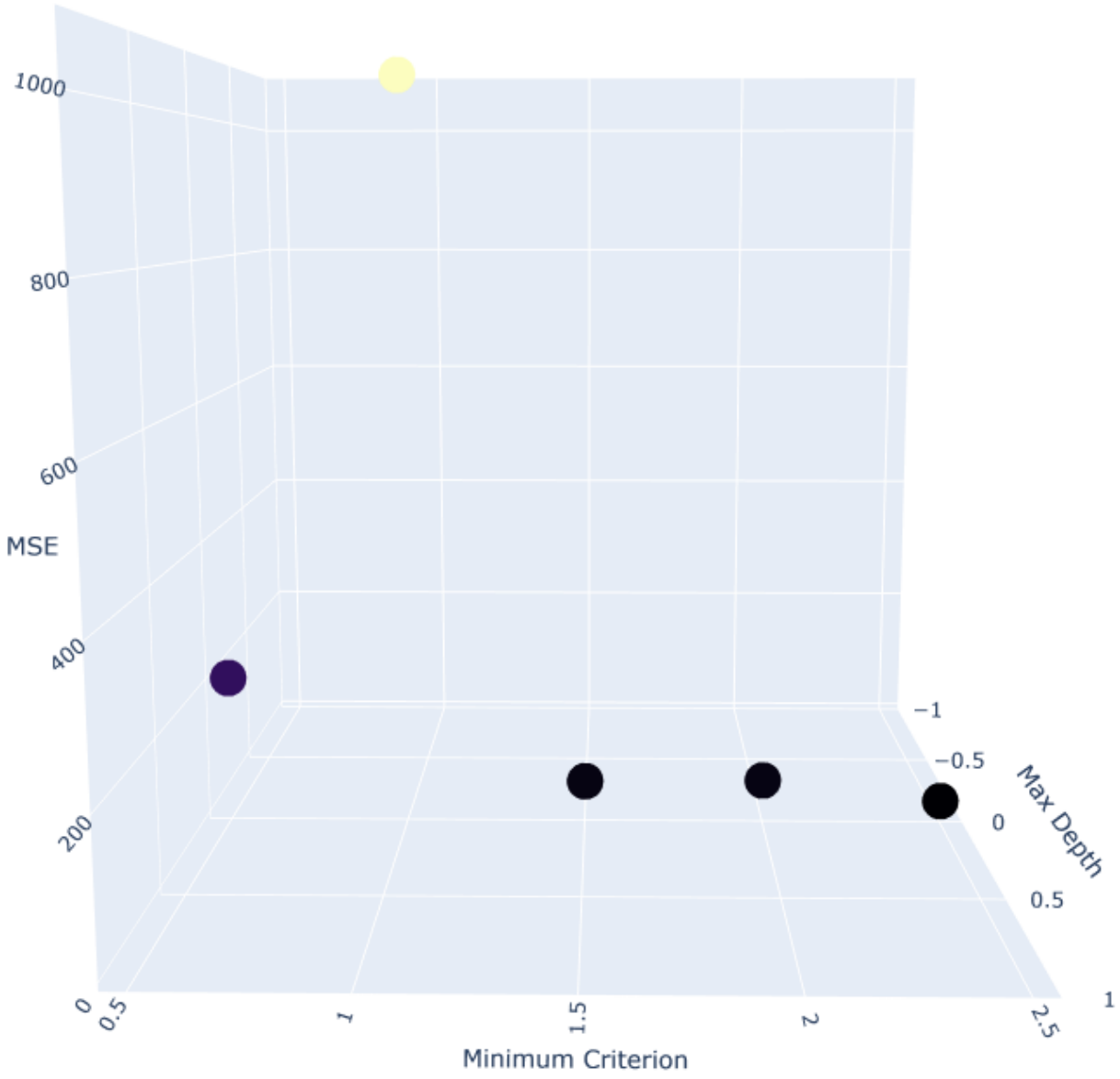
Lowest MSE is 21.7944 when Minimum Splitting Criterion = 2.5

Figure 15: Graph: Forest Tuning Pruned Tree Results (Max Depth vs. Minimum Criterion vs. MSE)



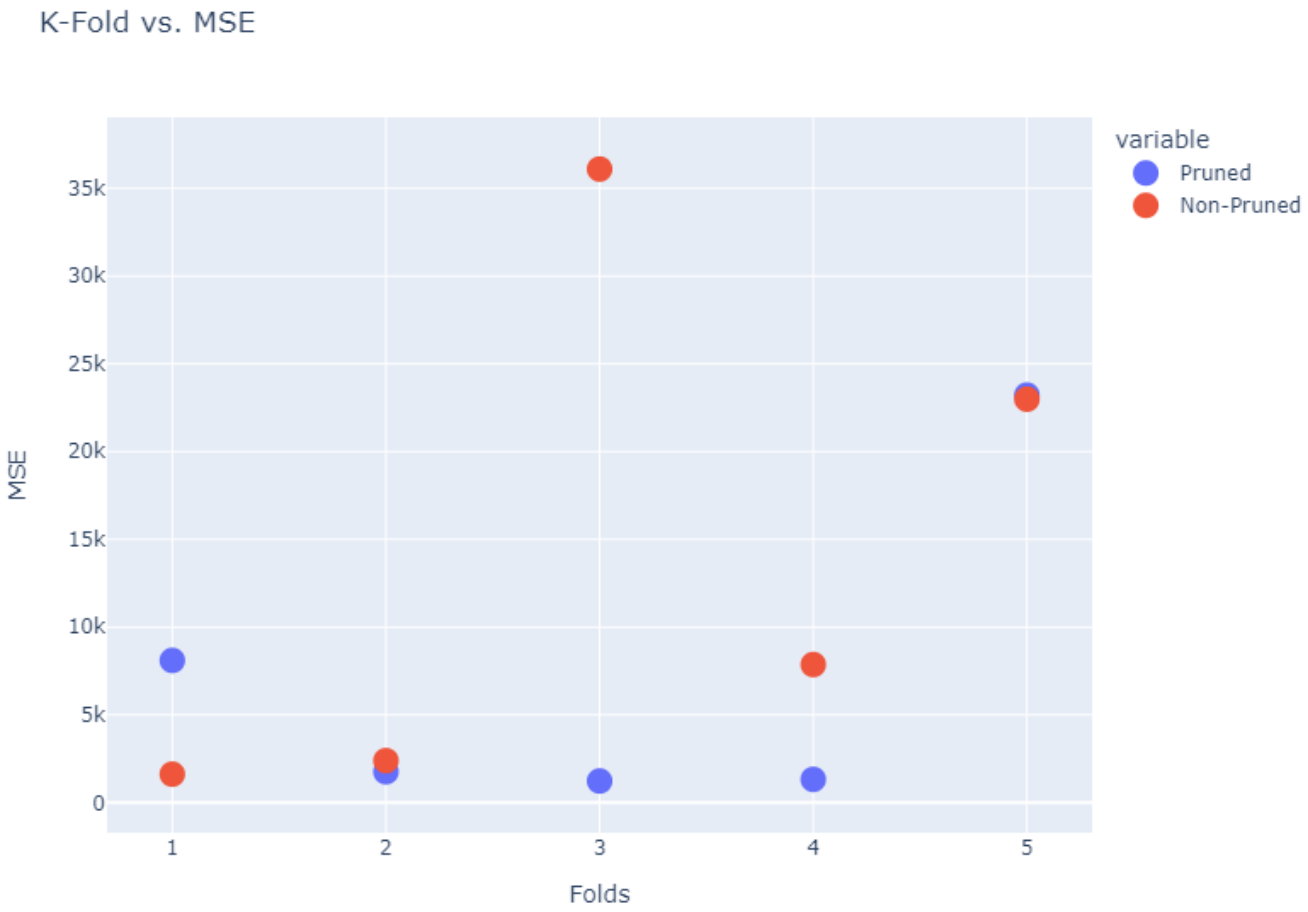
MSE of 29.2549 when Max Depth = 90 and Minimum Splitting Criterion = 2.5

Figure 16: Graph: Forest Tuning Unpruned Tree Results (Max Depth vs. Minimum Criterion vs. MSE))



Lowest MSE is 21.7944 when Minimum Splitting Criterion = 2.5

Figure 17: Graph: Forest Fire Testing Dataset Pruned and Unpruned Results (Fold vs. MSE)



Pruned Tree 5 Fold Cross Validation Average MSE : 7125.7629

Unpruned Tree 5 Fold Cross Validation Average MSE : 14191.774

4. Discussion

4.1 Classification Tasks

4.1.1 Car Evaluation Data

When looking at the results from implementing the ID3 Algorithm on the Car Evaluation Data, for the Unpruned Decision Tree from the tuning data the optimal combination of hyper parameters were $MaxDepth = 0$, $MinimumSplit = 0$ and $MinimumGain = 0.4$ or 0.6 resulting in 70.534% accuracy. When looking at the *car.name* file, it is important to note that the class distribution among the class was skewed more to the unacceptable class, *unacc*. This skewed distribution is often attributed to under sampling. To reduce overfitting pruning was applied to reduce the size of the decision tree. Implementing the ID3 Algorithm

on the Car Tuning Data, for a Pruned Decision Tree, the resulting optimal combination of hyper parameters were $MaxDepth = 6$, $MinimumSplit = 6$ and $MinimumGain = 0.4$ resulting in 70.57% accuracy.
When applying the combination of hyper parameters that resulted in a Pruned Decision Tree to the testing data, the average accuracy achieved was 70.4023%. When applying that combination of hyper parameters that resulted in an Unpruned Decision Tree from the testing data, the average accuracy achieved was 70.4076%. In this case the predictive performance had a insignificantly increase.

4.1.2 House Voting Data

Through the same process of tuning as the previous classification problem, the hyper-parameter combination that resulted in the highest accuracy for an Unpruned Decision Tree from the tuning data was $MaxDepth = 0$, $MinimumSplit = 0$ and $MinimumGain = 0.2$, with the accuracy being 65.556%. When pruning the tuning data to in an attempt to reduce the size of the decision tree such that it removes non-critical sections needed to make a classification, the optimal combination of hyper parameters were $MaxDepth = 6$, $MinimumSplit = 6$ and $MinimumGain = 0.2$, with the accuracy being 95.49%. Applying the respective hyper parameters on the testing data resulting in 69.250% accuracy for the Unpruned Tree and 95.122% accuracy for the Pruned Decision Tree. Pruning the decision tree in this case helped reduce the overfitting and increase the predictive capacity. While the achieved accuracy on the testing data was 95.122% for the Pruned Tree the *house-votes-84.names* file states that this predictive accuracy could be STAGGER's asymptote and could be observed in Figure 4.

4.1.3 Breast Cancer Data

The hyper-parameter combination that resulted in the highest accuracy for an Unpruned Decision Tree was $MaxDepth = 0$, $MinimumSplit = 0$ and any value of $MinimumGain$ as the values tested for that did not change the maximum accuracy of 70.7143%. When looking at the results from tuning the Pruned Decision Tree, the accuracy of the ID3 algorithm performed well above expected. The combination of optimal hyper-parameters were $MaxDepth = 11$, $MinimumSplit = 3$, and $MinimumGain = 0.2$, with an accuracy of 95.7134%.

Applying the respective combinations onto the testing data, the Unpruned Decision Tree had an accuracy of 69.9011% while the Pruned Decision Tree had an accuracy of 91.2356%.

4.2 Regression Tasks

4.2.1 Abalone Data

Unlike the ID3 algorithm, the CART algorithm only had two hyper-parameters that needed tuning, $maxDepth$ and $minCriterion$. The criterion function in these regression cases was Mean Square Error (MSE). Through tuning, it was attempted to find an optimal combination of hyper-parameters that minimizes the MSE of the decision trees. By setting $maxDepth = 0$, when tuning the Unpruned Decision Tree, the lowest $MSE = 6.078$ occurred

when $minCriterion = 0.2$. When tuning the Pruned Decision Tree, the lowest $MSE = 6.078$ occurred when $maxDepth = 3$ and $minCriterion = 0.25$. Using the respective combination of hyper parameters, for the testing data the Unpruned Tree resulted in an average $MSE = 6.3679$ and Pruned Tree resulted in an average $MSE = 6.4437$.

4.2.2 Computer Data

When implement the CART algorithm on the computer Data, there were issues with finding the best range of values for each hyper parameter. As a result from tuning, the lowest $MSE = 213.187$ occurred when $minCriterion = 2.5$, for the Unpruned Decision Tree. When tuning the Pruned Decision Tree, the lowest $MSE = 163.388$ occurred when $maxDepth = 250$ and $minCriterion = 2.0$. Using the respective combination of hyper parameters, for the testing data the Unpruned Tree resulted in an average MSE =5065.49 and Pruned Tree resulted in an average $MSE = 11241.0579$. As shown in Figure 14, the $MSEs$ were severely greater than expected. It is possible that the tuning values selected arrived at a local optima where as there is yet to find global optimum tuning values to reduce the MSE.

4.2.3 Forest Fire Data

From tuning, the lowest $MSE = 21.7944$ occurred when $minCriterion = 2.5$, for the Unpruned Decision Tree. When tuning the Pruned Decision Tree, the lowest $MSE = 29.2549$ occurred when $maxDepth = 90$ and $minCriterion = 2.5$. From the *forestfires.name* file it was stated that this is a very difficult regression task as it is not clear how many outliers there are and the number of examples with a large burned area is relatively small. When using the respective combination of hyper parameters, for the testing data the Unpruned Tree resulted in an average MSE =5065.49 and Pruned Tree resulted in an average $MSE = 11241.0579$. As shown in Figure 17, the $MSEs$ were severely greater than expected. This and the computer data results show how these models are fitting against the training data. With the inclusion of a small sample of large burned areas, this model thus is unable to generalize and performs poorly against unseen data.

5. Conclusion

In the process of implementing the ID3 and the CART algorithm, it has shown that while a decision trees are relatively good at making predictions on seen data, it is a challenge to tackle the concepts of overfitting, local optima, and generalization. A decision that is too large, risks overfitting the data and having poor generalization of new samples and small trees may not grasp all the critical information of the search space. Through this project, it has shown how arriving at local optima hyper parameter values, can lead to overfitting and a limited search space for the decision tree. Though overfitting can be tackled through pruning, that is only is the global optimum hyper-parameter values are chosen. Lastly, while the classification task produced good results for the Car Evaluation Data, for the other two it shows a concerning increase in accuracy between the pruned and unpruned trees this low error rate is a good indicator of overfitting. Besides the Abalone Data, the regression tasks may have not produced the best results, but have shown the power of size reduction and how that can greatly affect processing. This paper finds that decision trees

did perform well on seen data but had difficulty when it came to generalization and unseen data.

References

L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984. ISBN 9780412048418. URL `https://books.google.com/books?id=JwQx-WOmSyQC`.

J. R. Quinlan. Induction of decision trees. *MACH. LEARN*, 1:81–106, 1986.