COL341 Assignment - 3

3.1 Binary Classification

a) Decision Tree from scratch

Splitting criterion as 'GINI'

Training time = 110 seconds

For Train set

PRECISION = 0.82250607728215
RECALL = 0.8460000000000001
ACCURACY = 0.87

For Validation set

PRECISION = 0.7548689324129427
RECALL = 0.8307142857142857
ACCURACY = 0.89125

Splitting criterion as 'ENTROPY'

Training time = 129 seconds

For Train set

PRECISION = 0.872630407914148
RECALL = 0.8903333333333333
ACCURACY = 0.9085

For Validation set

PRECISION = 0.8167538034006876
RECALL = 0.8842857142857143
ACCURACY = 0.925

b) Decision Tree sklearn

Splitting criterion = 'GINI'

Training time = 5 seconds

For Train set

PRECISION = 0.9877300613496932
RECALL = 0.966
ACCURACY = 0.9885

For Validation set

PRECISION = 0.9090909090909091
RECALL = 0.8
ACCURACY = 0.965

Splitting criterion = 'ENTROPY'

Training time = 5 seconds

For Train set

PRECISION = 0.9960159362549801
RECALL = 1.0
ACCURACY = 0.999

For Validation set

PRECISION = 0.9090909090909091
RECALL = 0.9
ACCURACY = 0.97625

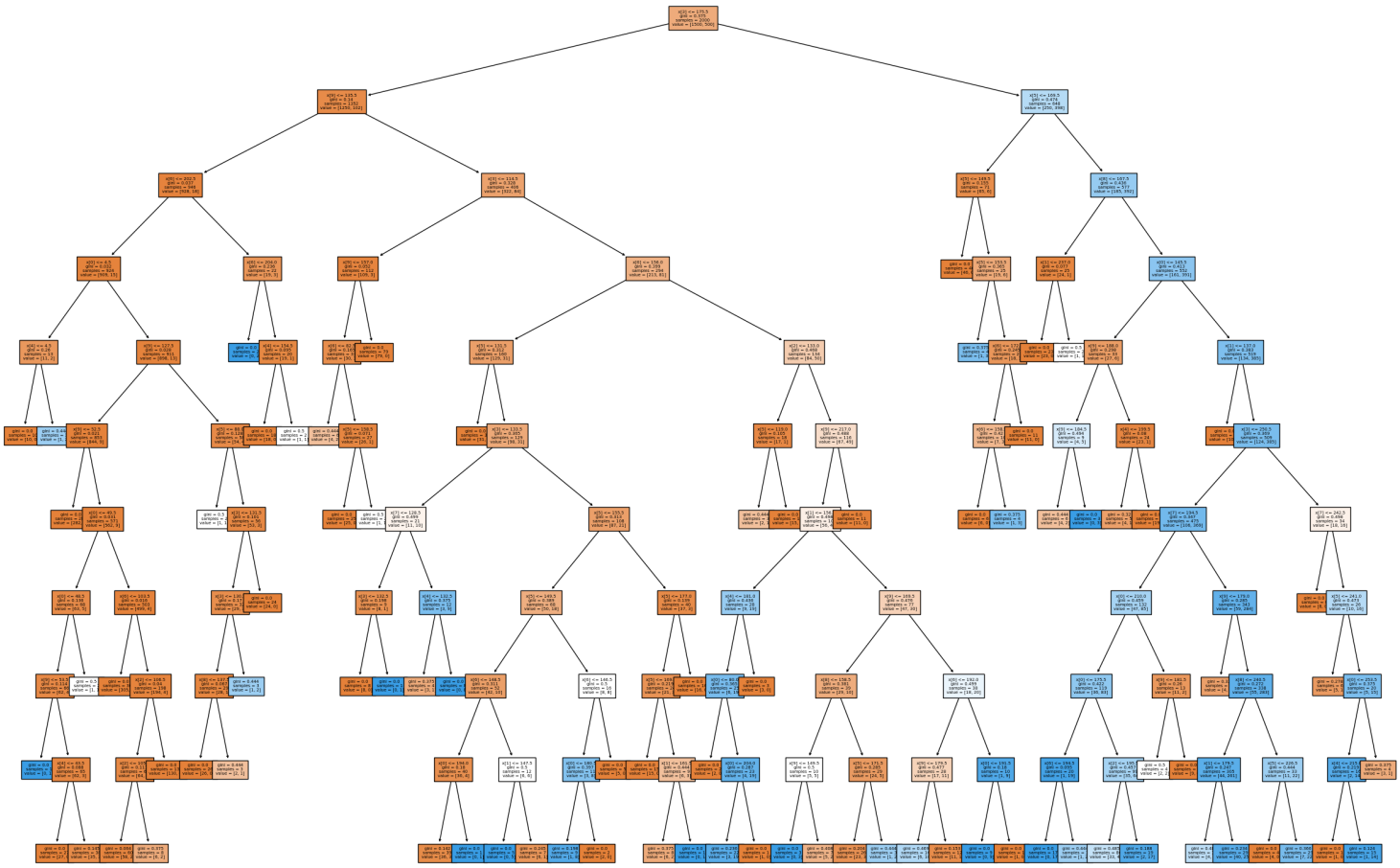
Comparison with my model->

Clearly the accuracy for the sklearn model is far more better than my model (part a) because it uses advanced mathematical and statistical techniques to make decisions based on the input data. It is also well optimised since it ran in just 5 seconds and my model took almost 2 mins to train.

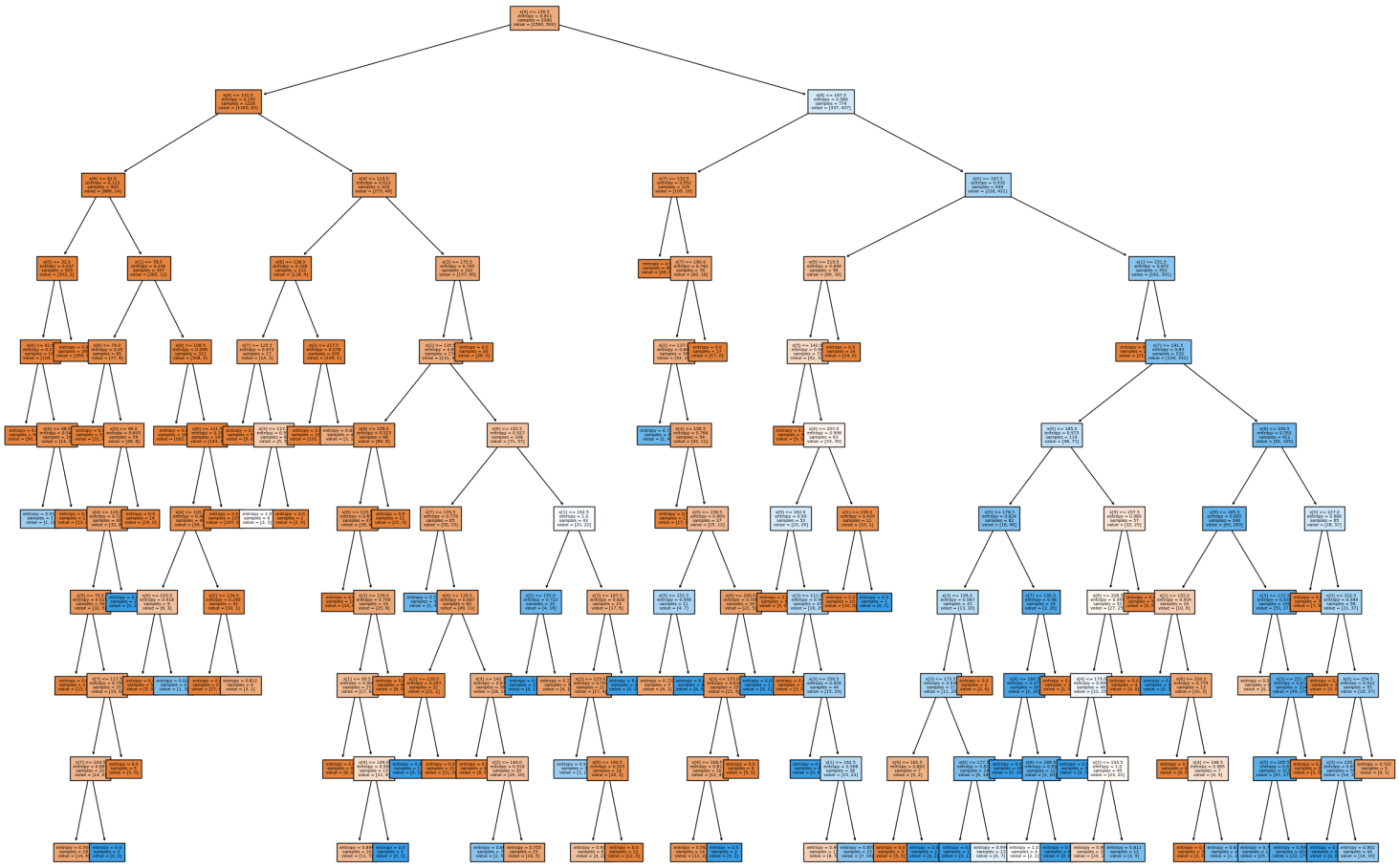
c) Decision Tree Grid-Search and Visualisation

Visualisation

For criterion : Gini



For criterion : Entropy



Optimal set of parameters obtained are

- Criterion : entropy
- Maxdepth : None
- min_sample_split : 2

Training time : less than a second

For Train Set

PRECISION = 1.0
RECALL = 1.0
ACCURACY = 1.0

For Validation Set

PRECISION = 0.77
RECALL = 0.77
ACCURACY = 0.9425

Comparison with A part

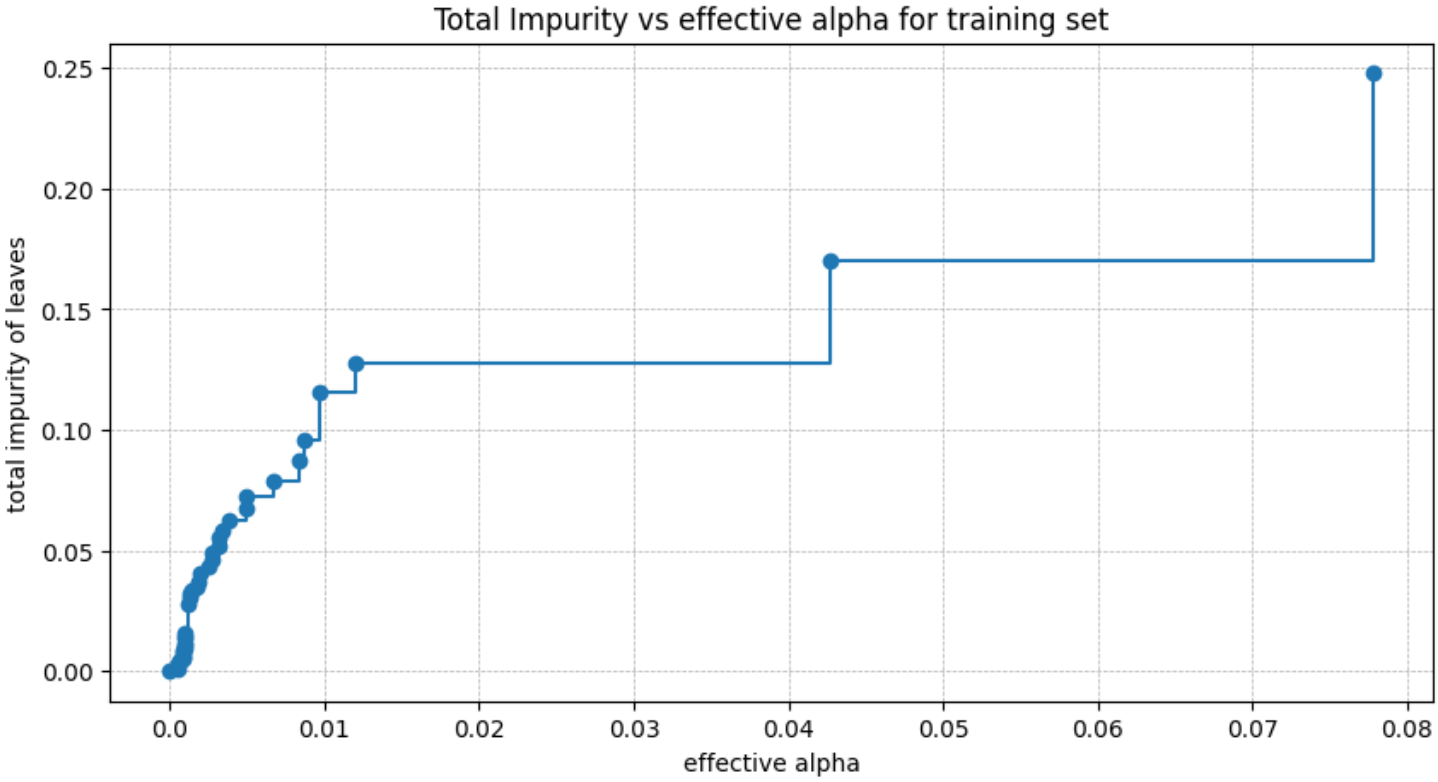
Since we are using just 10 features, the training time reduced significantly. On comparison with A, the training time of A is approx 2 mins and the training time of C is just a few milli seconds now.
If we talk about the accuracies, the accuracy obtained in this part is still higher than the accuracy obtained in part A (for validation set)

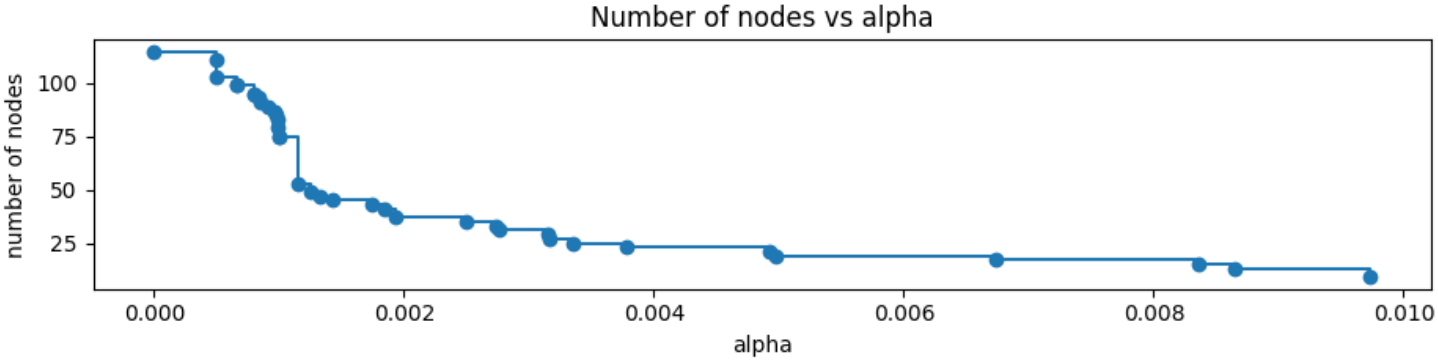
Comparison with B part

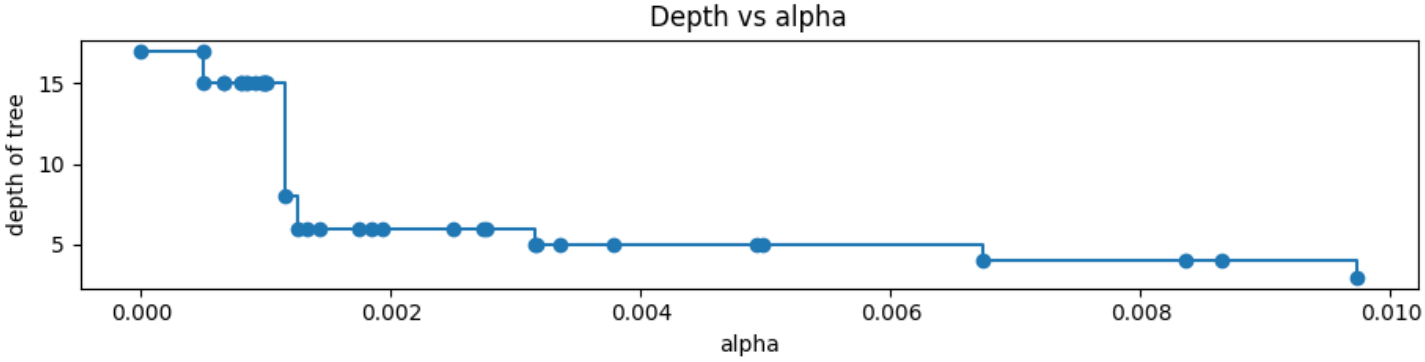
We can see that the training time has dropped significantly. From 5 seconds to a few milliseconds. But the validation set accuracies have dropped significantly from 97% to 94% (although the training accuracy has increased from 99% to 100%)

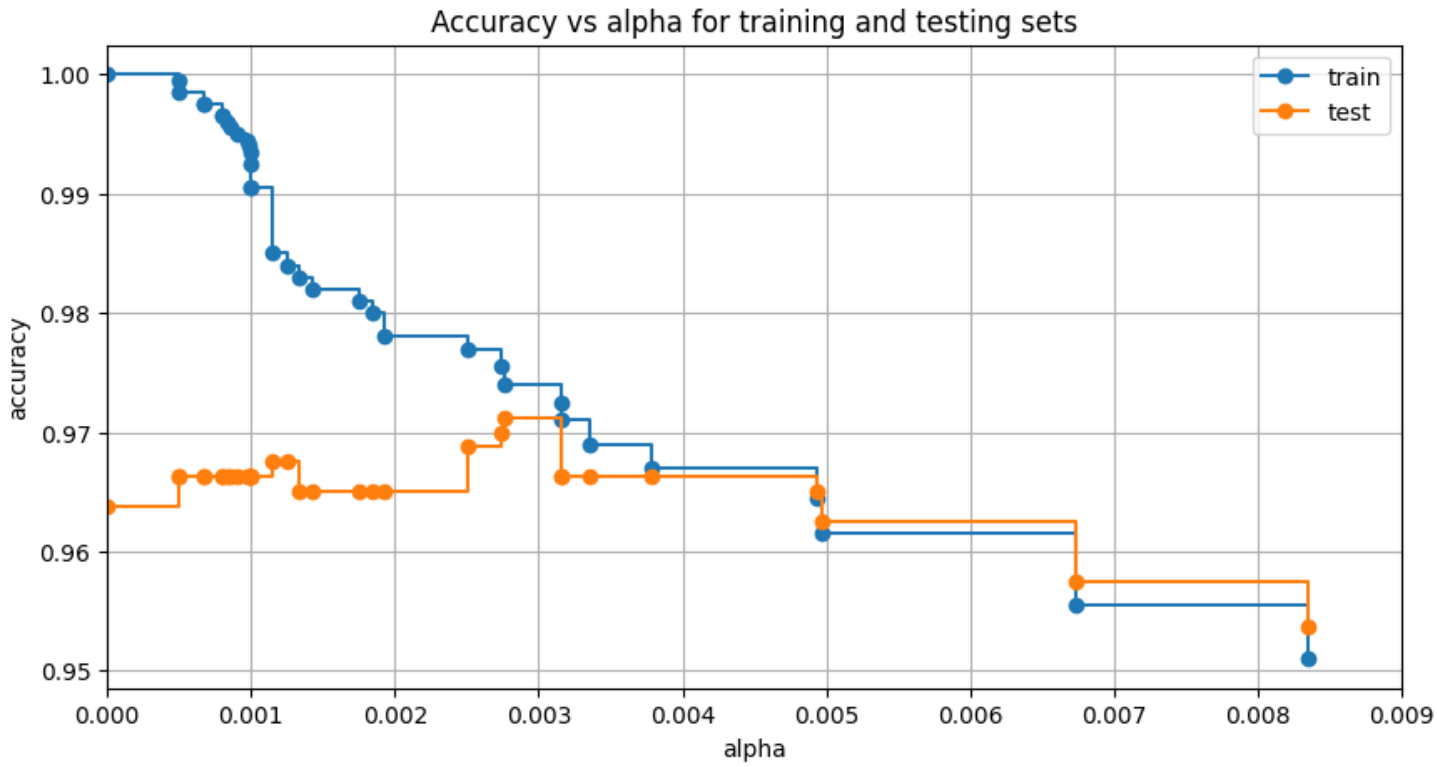
d) Decision Tree Post Pruning with Cost Complexity Pruning

PLOTS

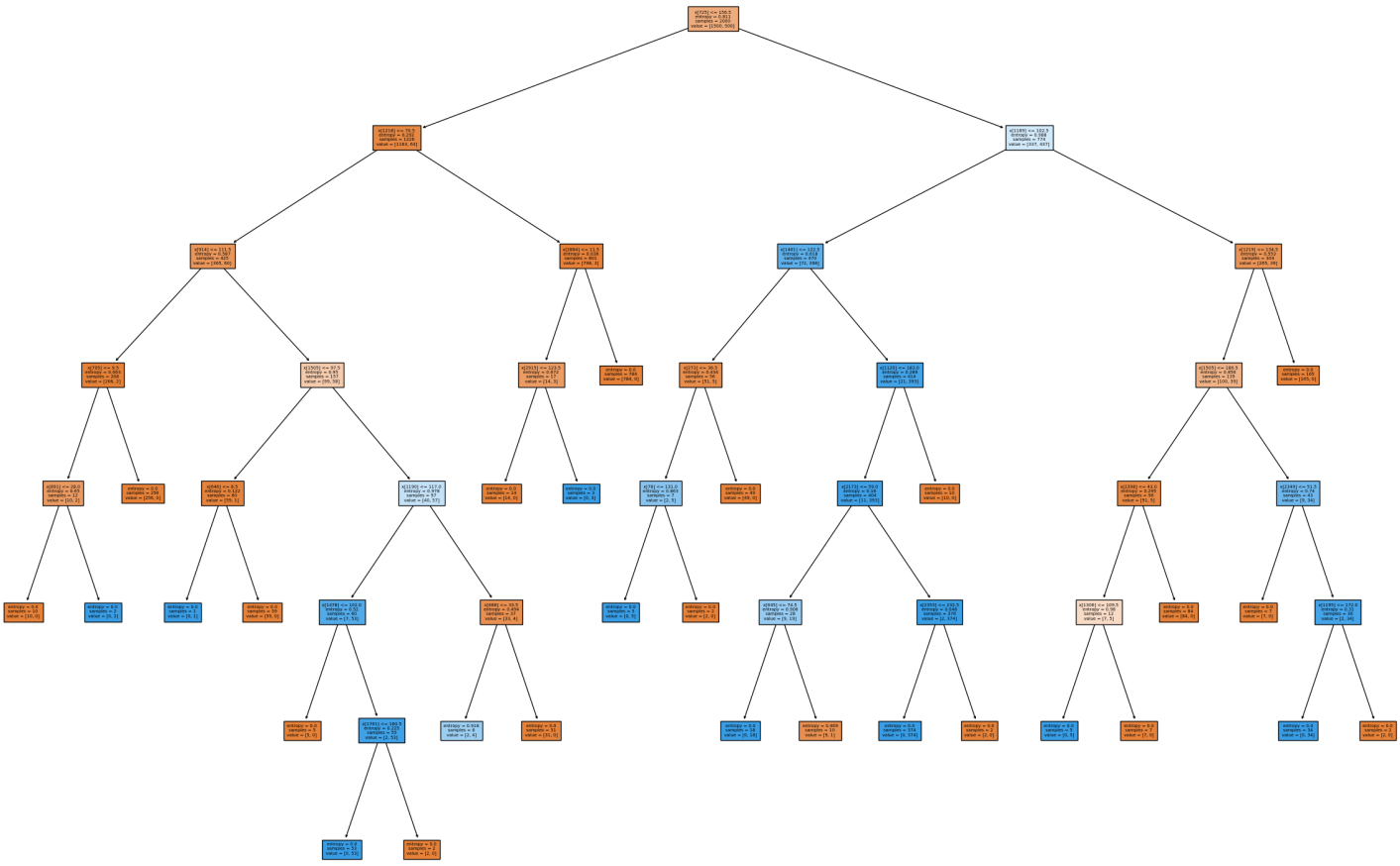

Total Impurity vs effective alpha for training set
total impurity of leaves
effective alpha
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08


Number of nodes vs alpha
number of nodes
alpha
100
75
50
25
0.000
0.002
0.004
0.006
0.008
0.010


Depth vs alpha
depth of tree
alpha
15
10
5
0.000
0.002
0.004
0.006
0.008
0.010



Visualization



Observations

The impurity of a node is a measure of how mixed the labels of the training examples are at that node.

We can see as the magnitude of alpha increase the total impurity decreases (which is sum of impurities of all nodes). By penalising the impurity of its nodes, increasing the value of alpha forces the tree to be simpler and smaller.

As we can see the total impurity decrease with increasing alpha →since the number of nodes are decreasing

We can also see that the number of nodes and the depth of the tree both decrease with increasing value of alphas.

Validation and training accuracy plot:
We can see that the training accuracy keep on decreasing with increasing alpha consistently. This is because we have lesser nodes for the split not and the tree do lesser overfitting.
For the validation accuracy : we can see that it first increases(not regularly tho) then reaches a peak at alpha = 0.003 (approx) and then decreases

Visualization : The tree generated has much lesser number of nodes than the one generated in part c

Training time for best alpha -> 4.4 seconds

Optimal set of parameters obtained are

- Criterion : entropy
- ccp_alpha :0.0027625497690715057

Test set

PRECISION = 0.9960079840319361
RECALL = 0.998
ACCURACY = 0.9985

Validation set

PRECISION = 0.9081632653061225
RECALL = 0.89
ACCURACY = 0.975

e) Random Forest

Optimal set of parameters obtained are

- Criterion : entropy
- max_depth : None

- min_samples_split : 5
- n_estimators : 100

Training time -> 4 seconds

For training set

PRECISION = 1.0
RECALL = 1.0
ACCURACY = 1.0

For Validation set

PRECISION = 1.0
RECALL = 0.9
ACCURACY = 0.9875

e) Gradient Boost, XG Boost

Gradient Boost

Optimal set of parameters obtained are

- max_depth : 6
- n_estimators : 50
- subsample : 0.6

Time to train → 1min 3.4sec

Training set

PRECISION = 1.0
RECALL = 1.0
ACCURACY = 1.0

Validation set

PRECISION = 0.9888888888888889
RECALL = 0.89
ACCURACY = 0.985

XGradBosst

Optimal set of parameters obtained are

- max_depth : 5
- n_estimators : 30
- subsample : 0.5

Time to train → 2.8 seconds

Training set

PRECISION = 1.0
RECALL = 1.0
ACCURACY = 1.0

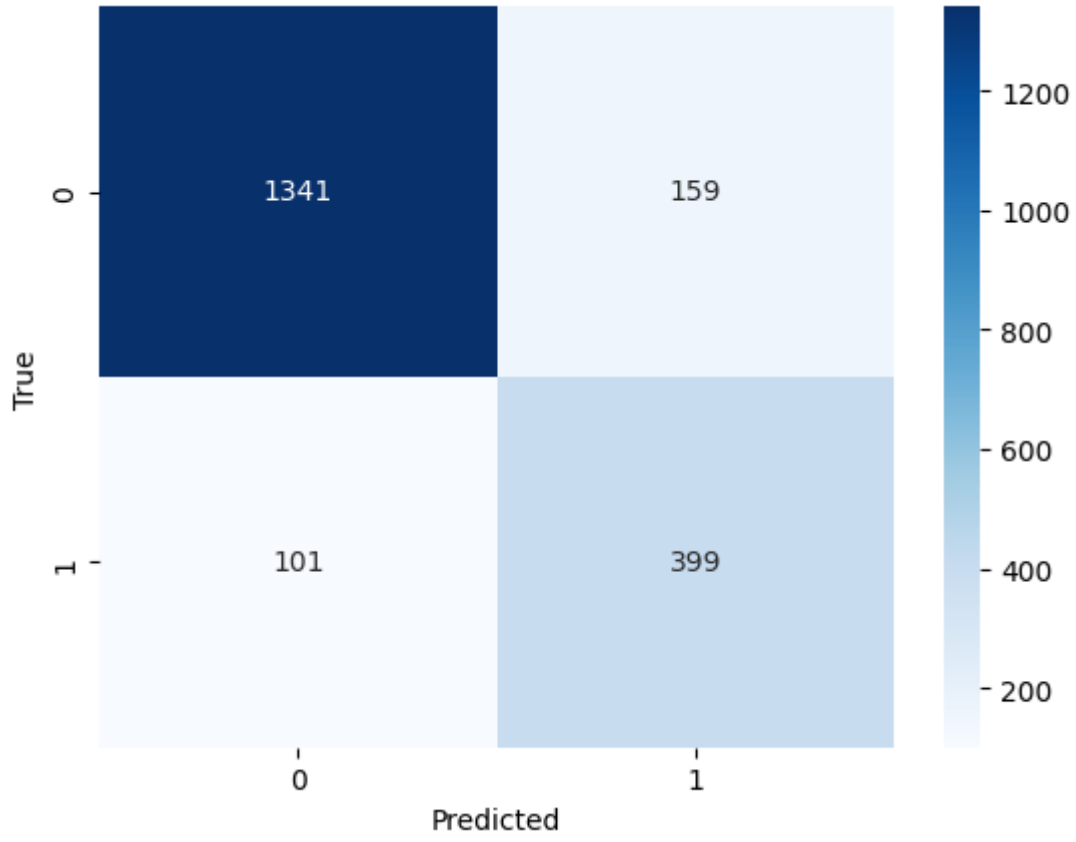
Validation set

PRECISION = 0.9888888888888889
RECALL = 0.89
ACCURACY = 0.985

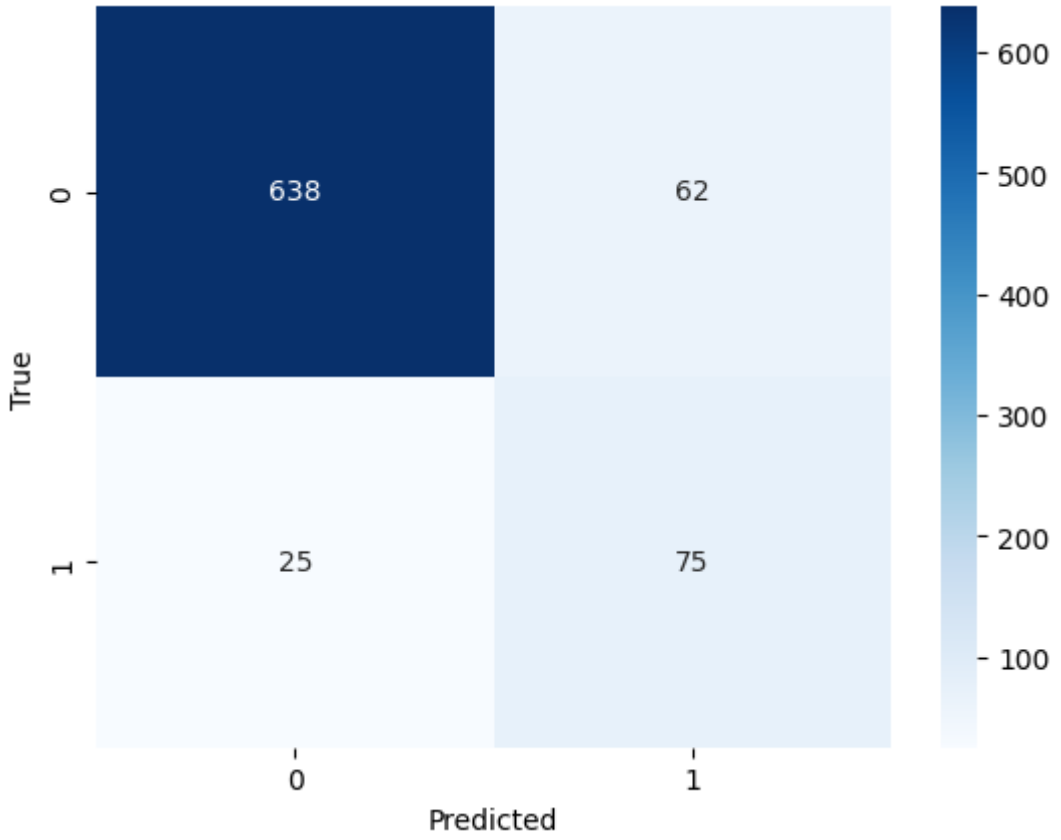
f) Confusion matrix

Part A (Gini)

Training set

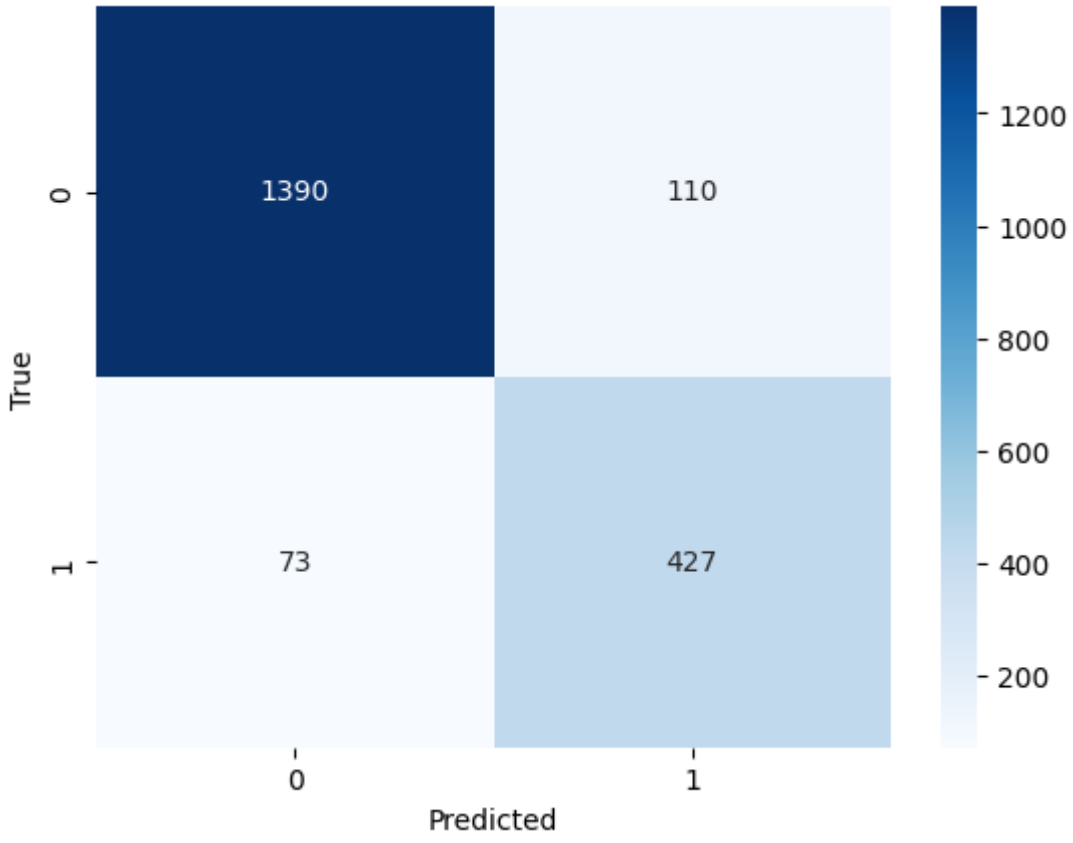


Validation set

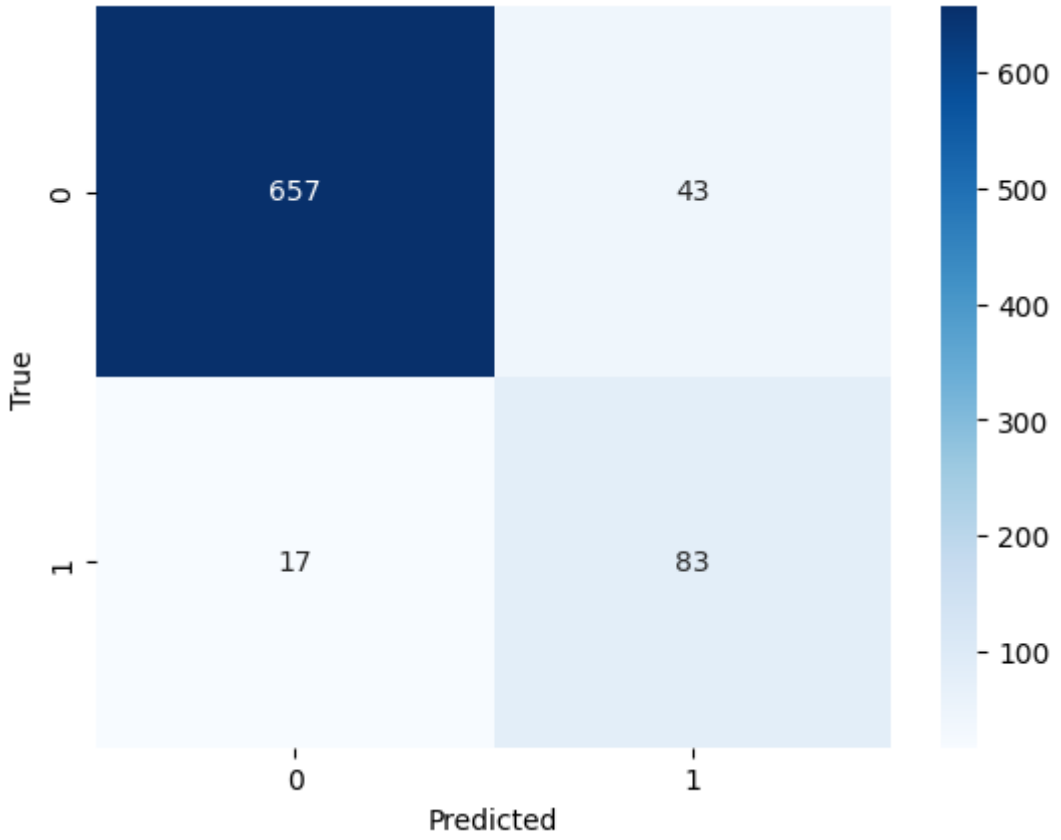


Part A (entropy)

Training set

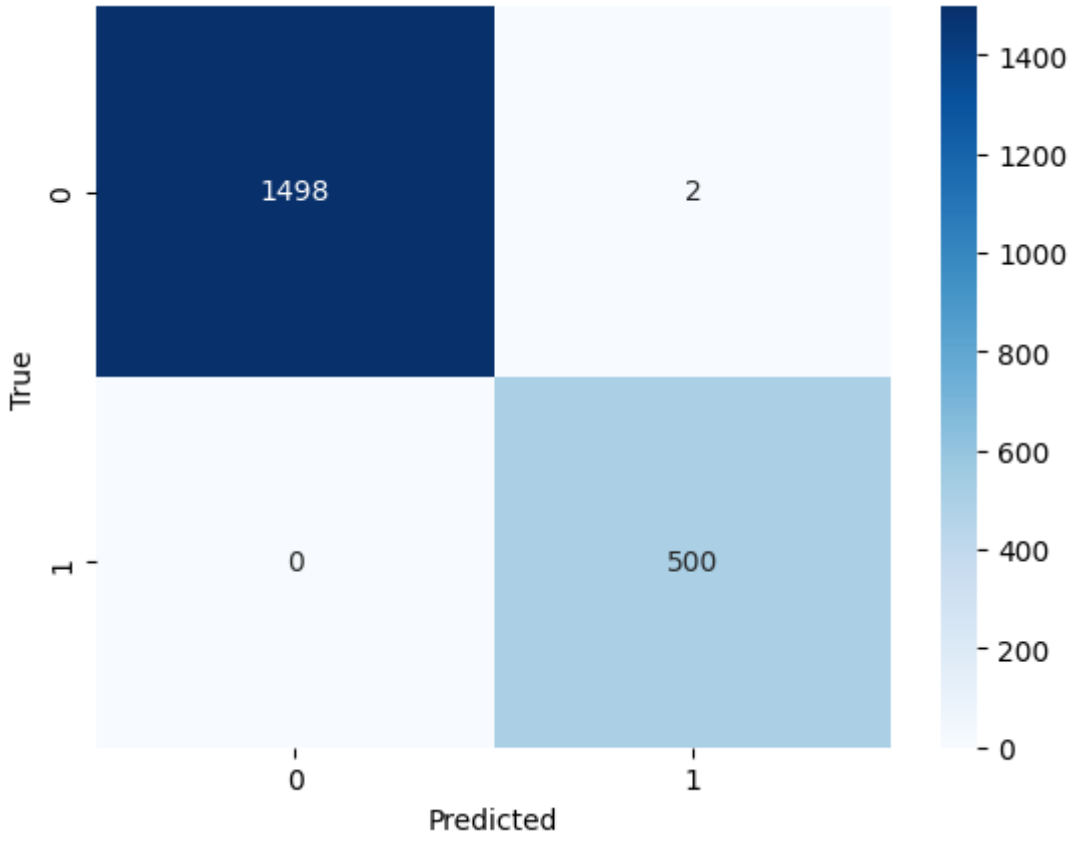


Validation set

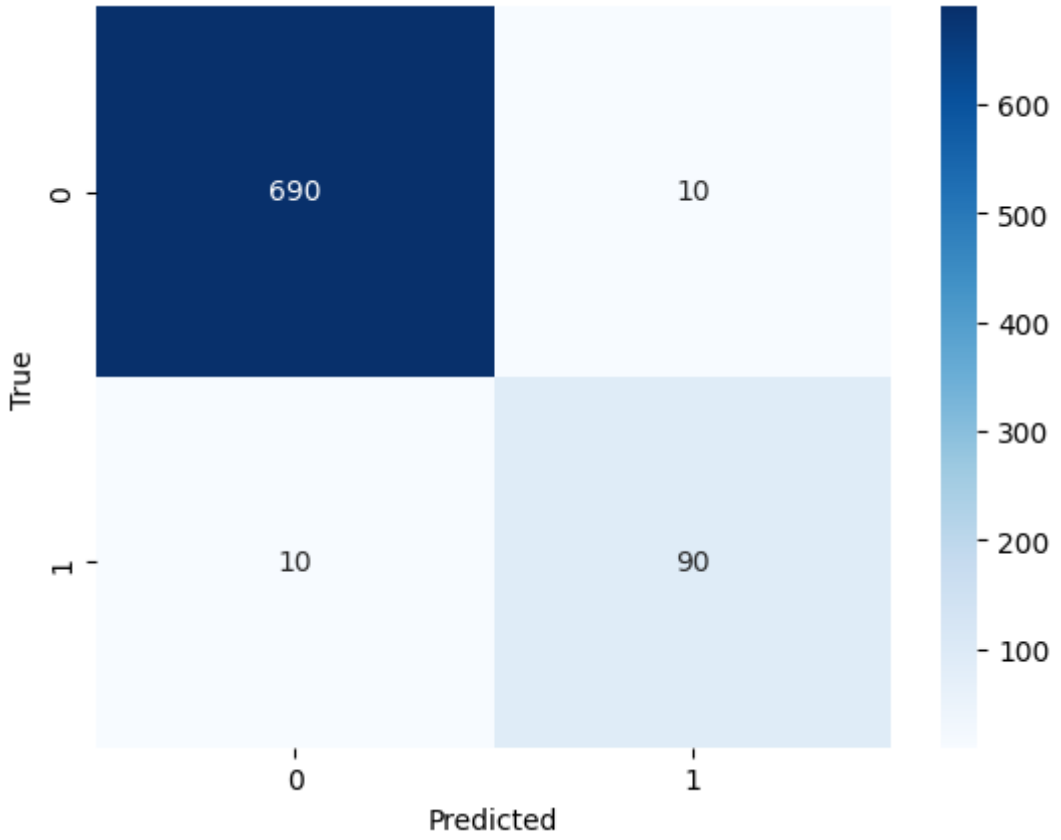


Part B

Training set

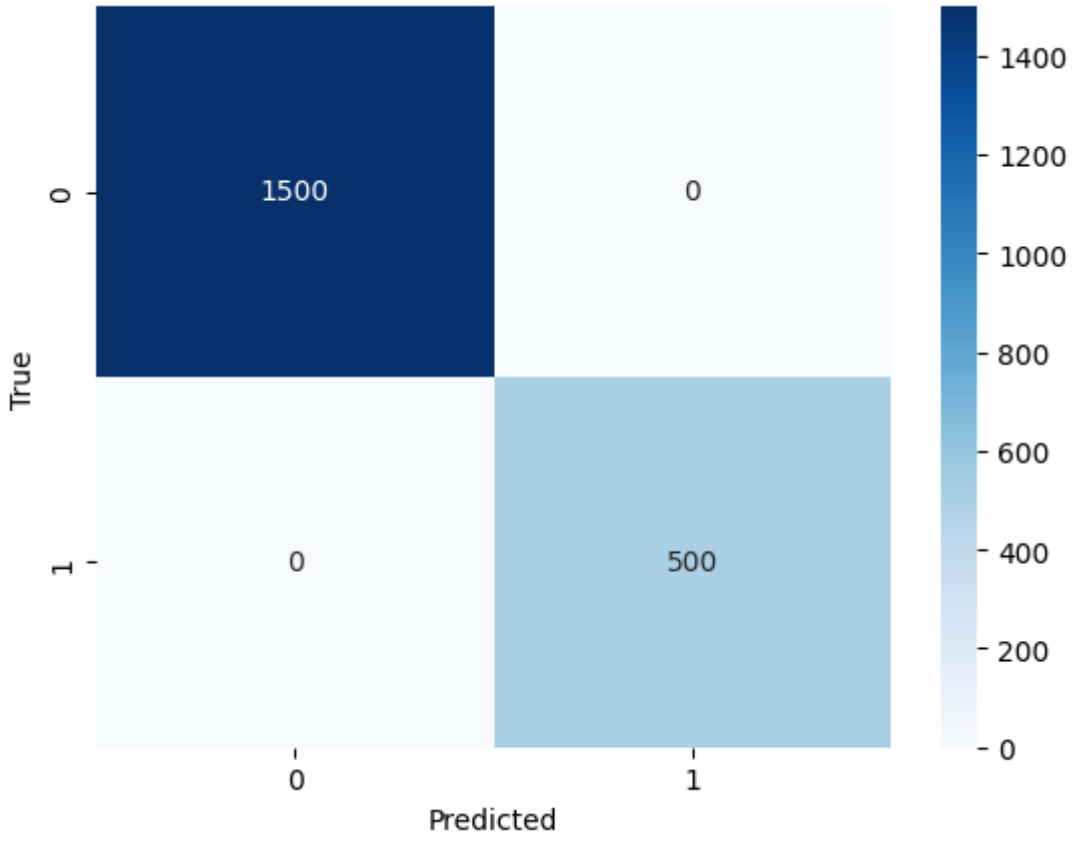


Validation set

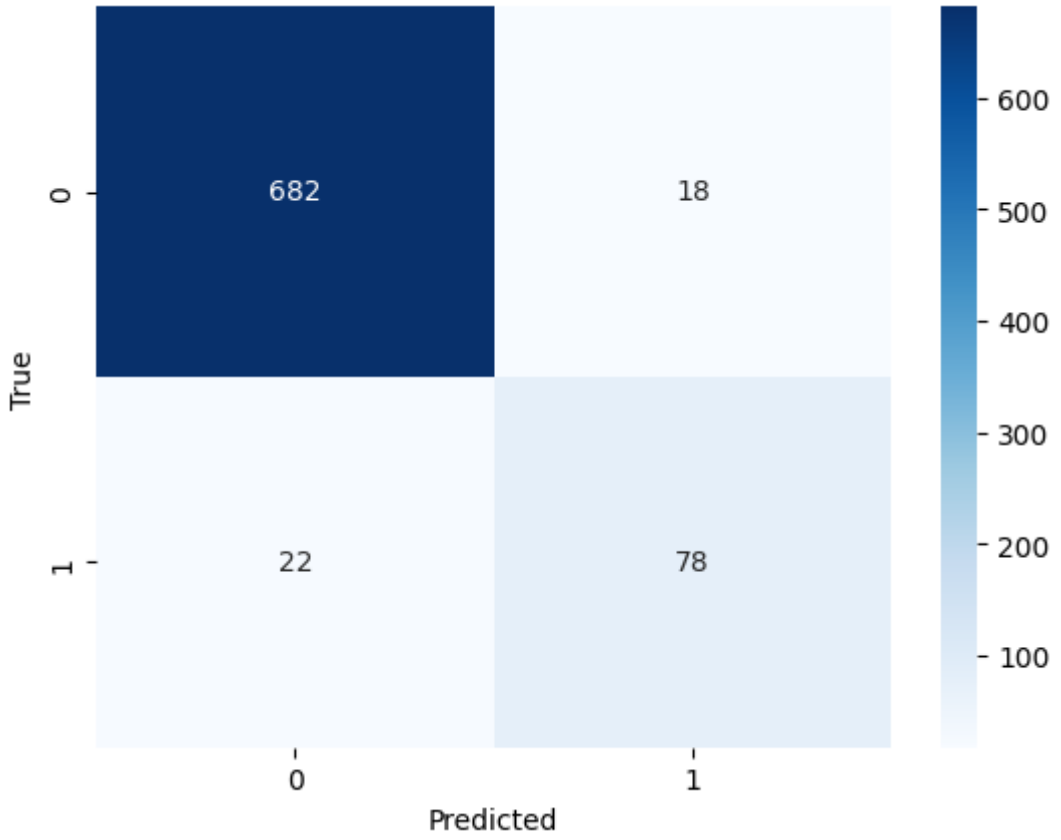


Part C

Training set

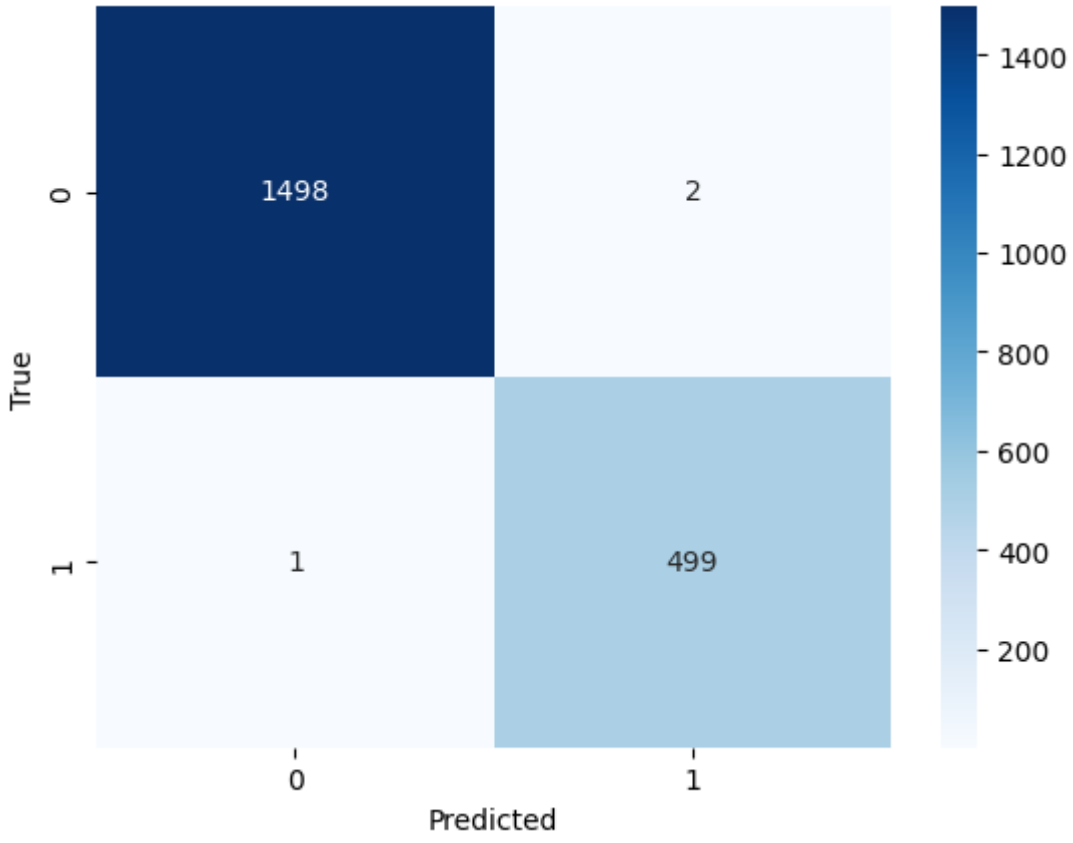


Validation set

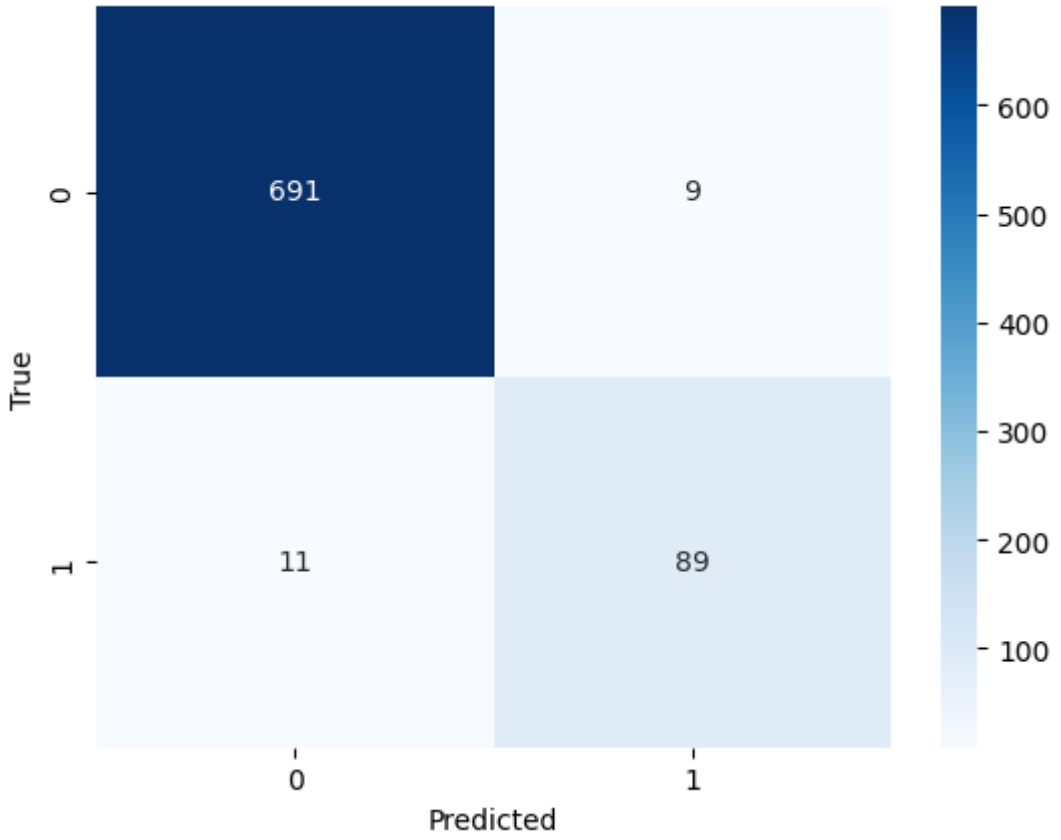


Part D

Training set

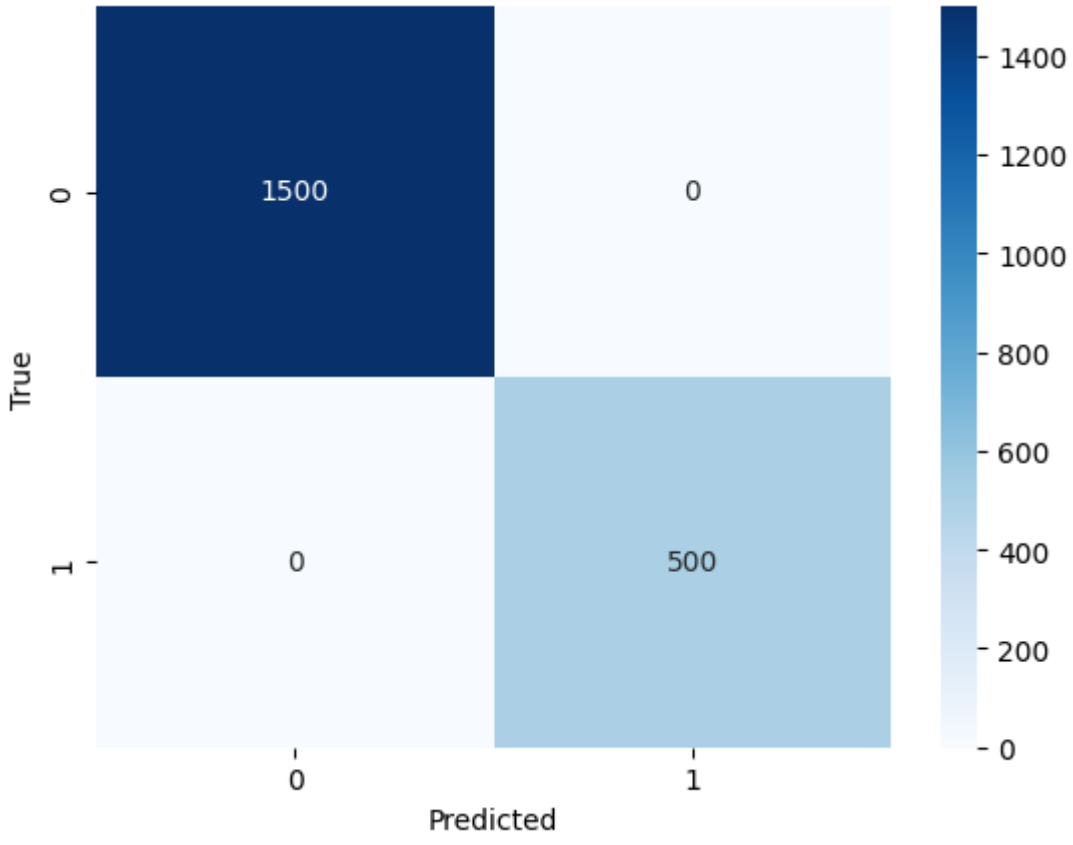


Validation set

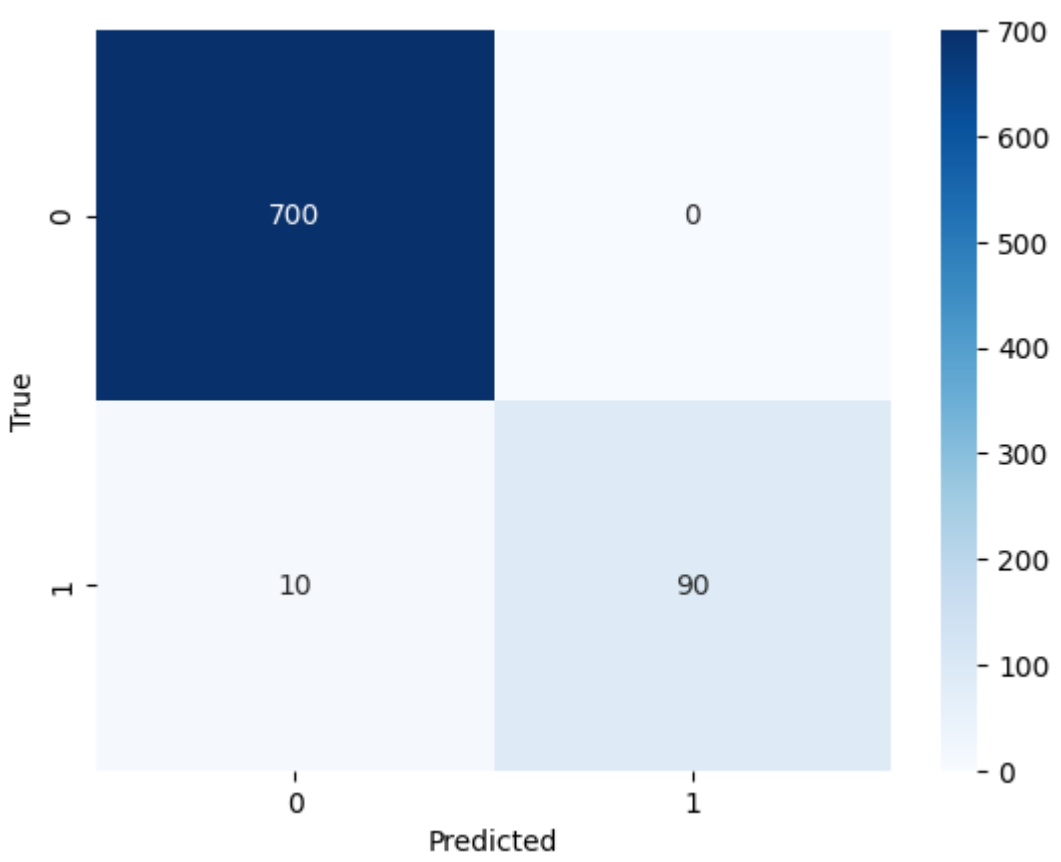


Part E

Training set

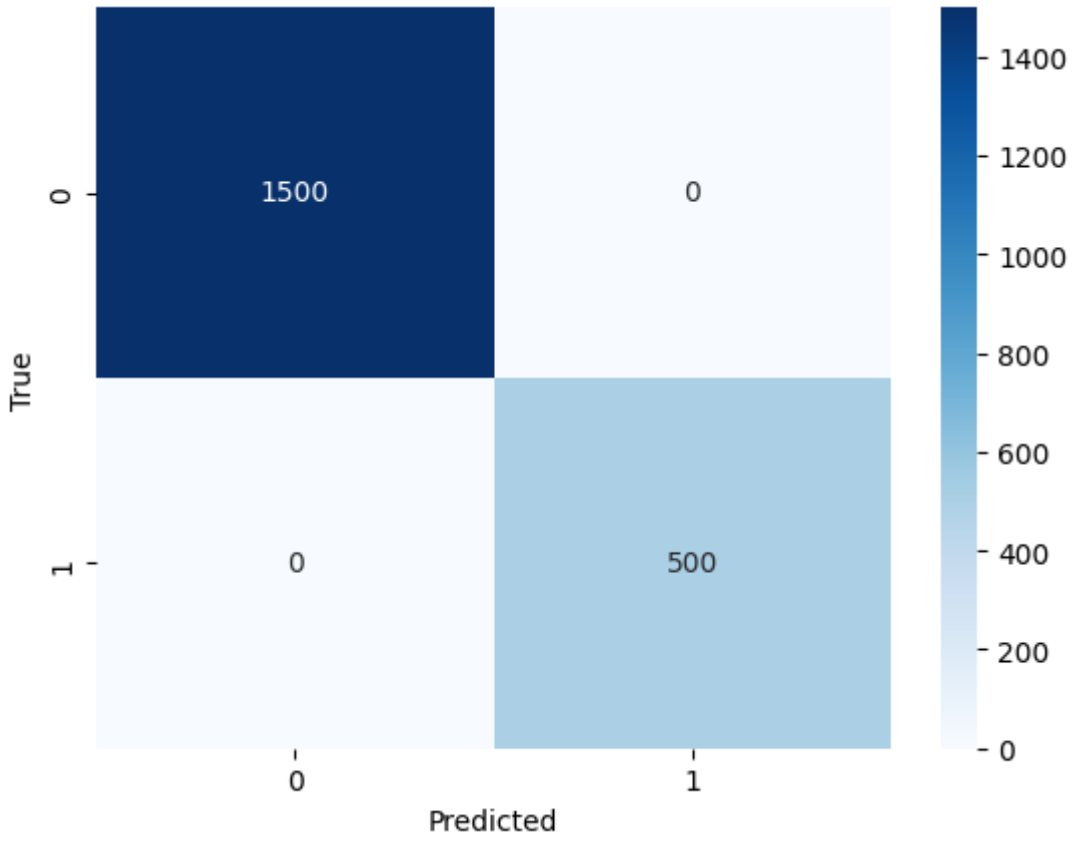


Validation set

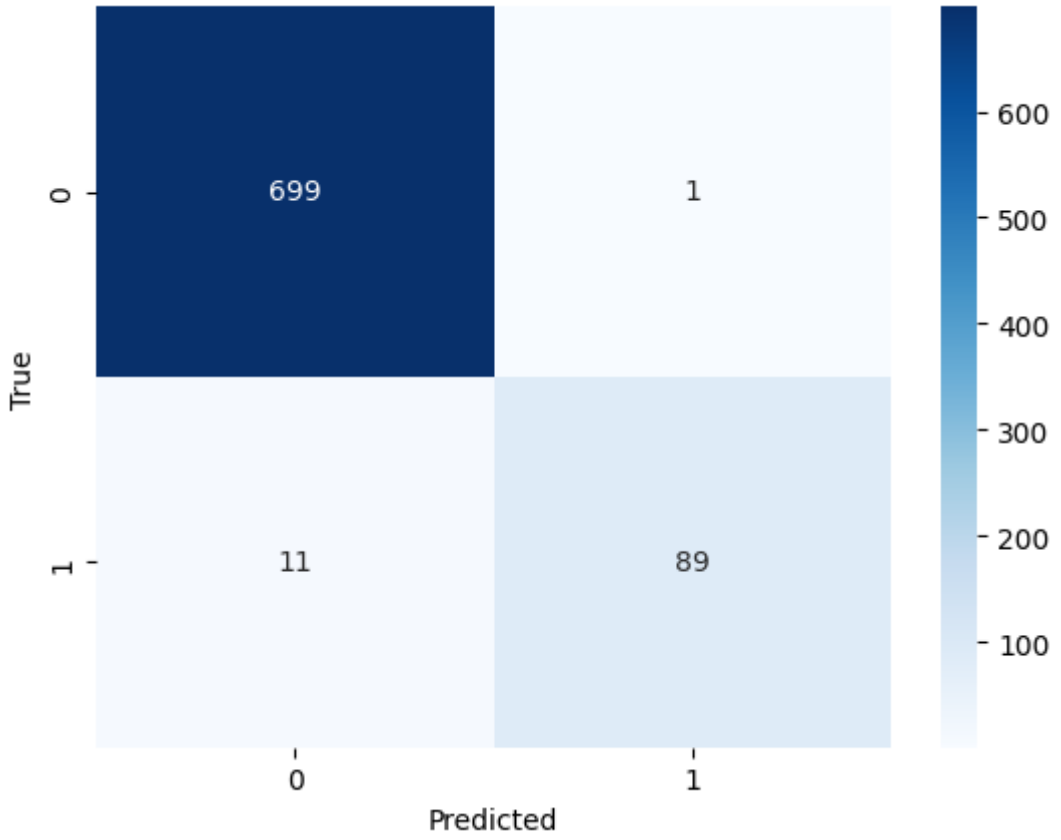


Part F (Gradient boost)

Training set

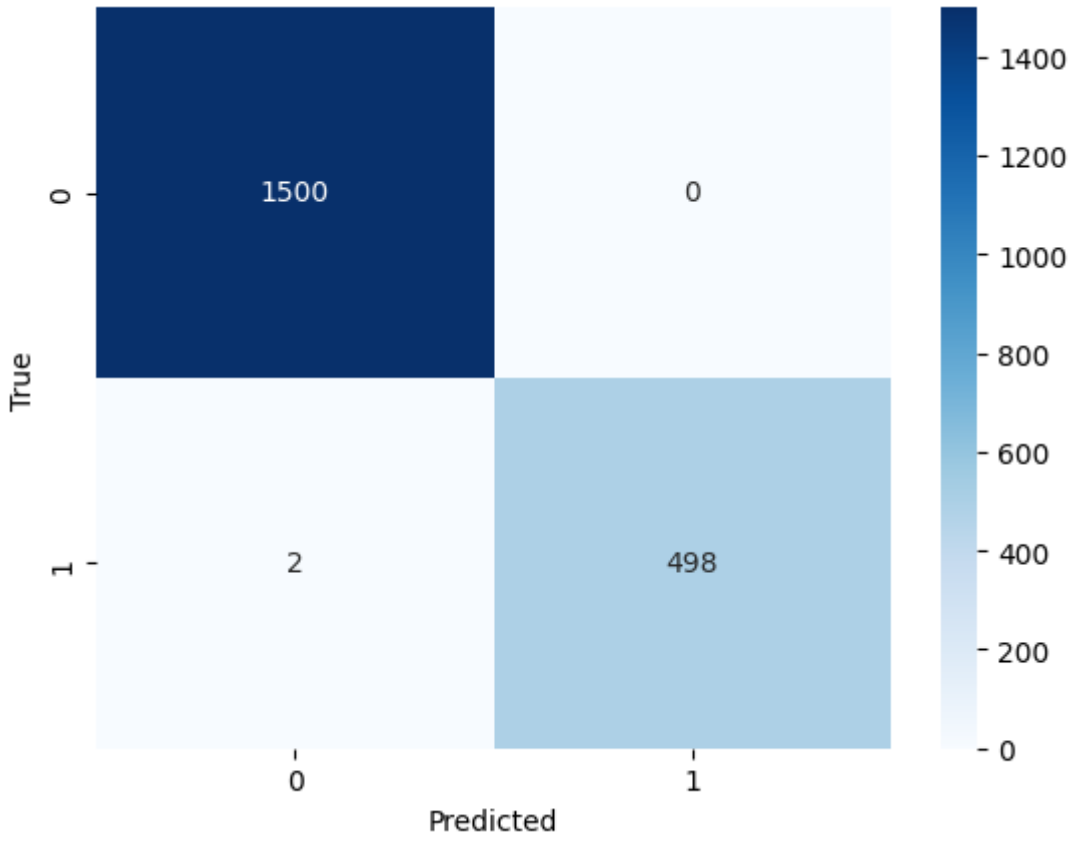


Validation set

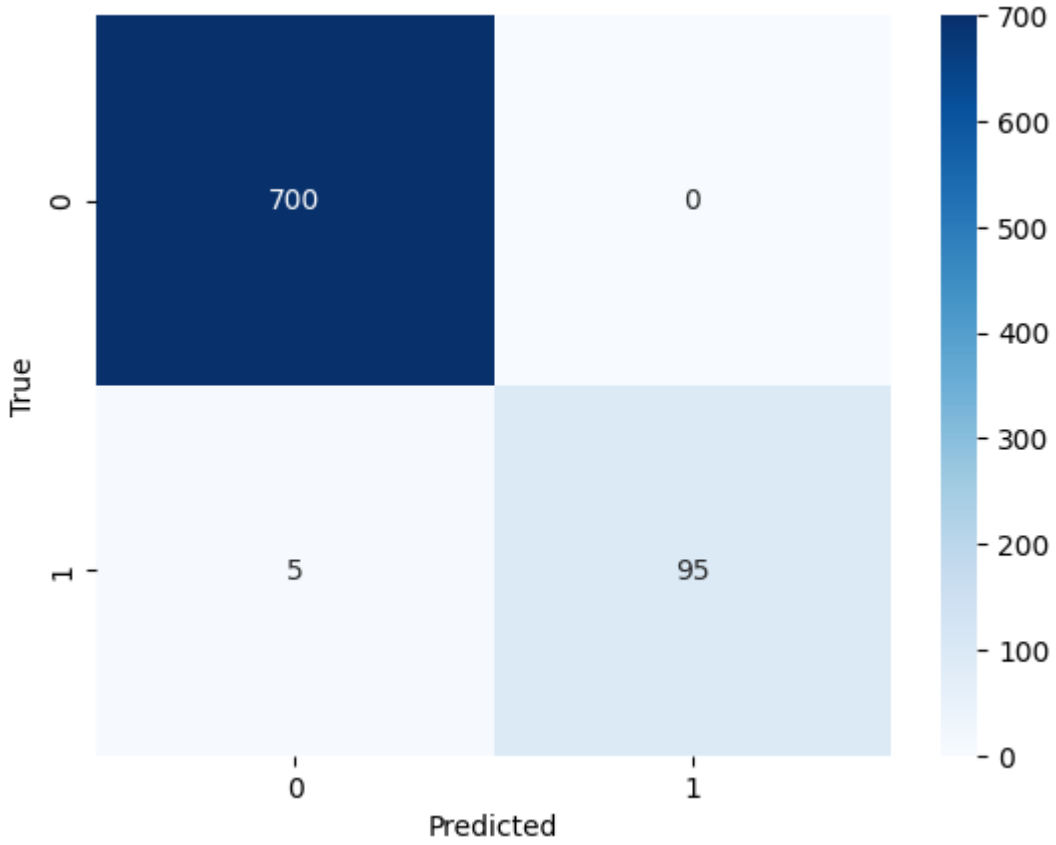


Part F (XGradient boost)

Training set



Validation set



Multi-Class Classification

a) Decision Tree sklearn

Criterion = Gini

Train time = 3.2 seconds

Training set

PRECISION = 0.969
RECALL = 0.969
ACCURACY = 0.969

Validation set

PRECISION = 0.74
RECALL = 0.74
ACCURACY = 0.74

Criterion = Entropy

Train time = 4 seconds

Training set

PRECISION = 0.971
RECALL = 0.971
ACCURACY = 0.971

Validation set

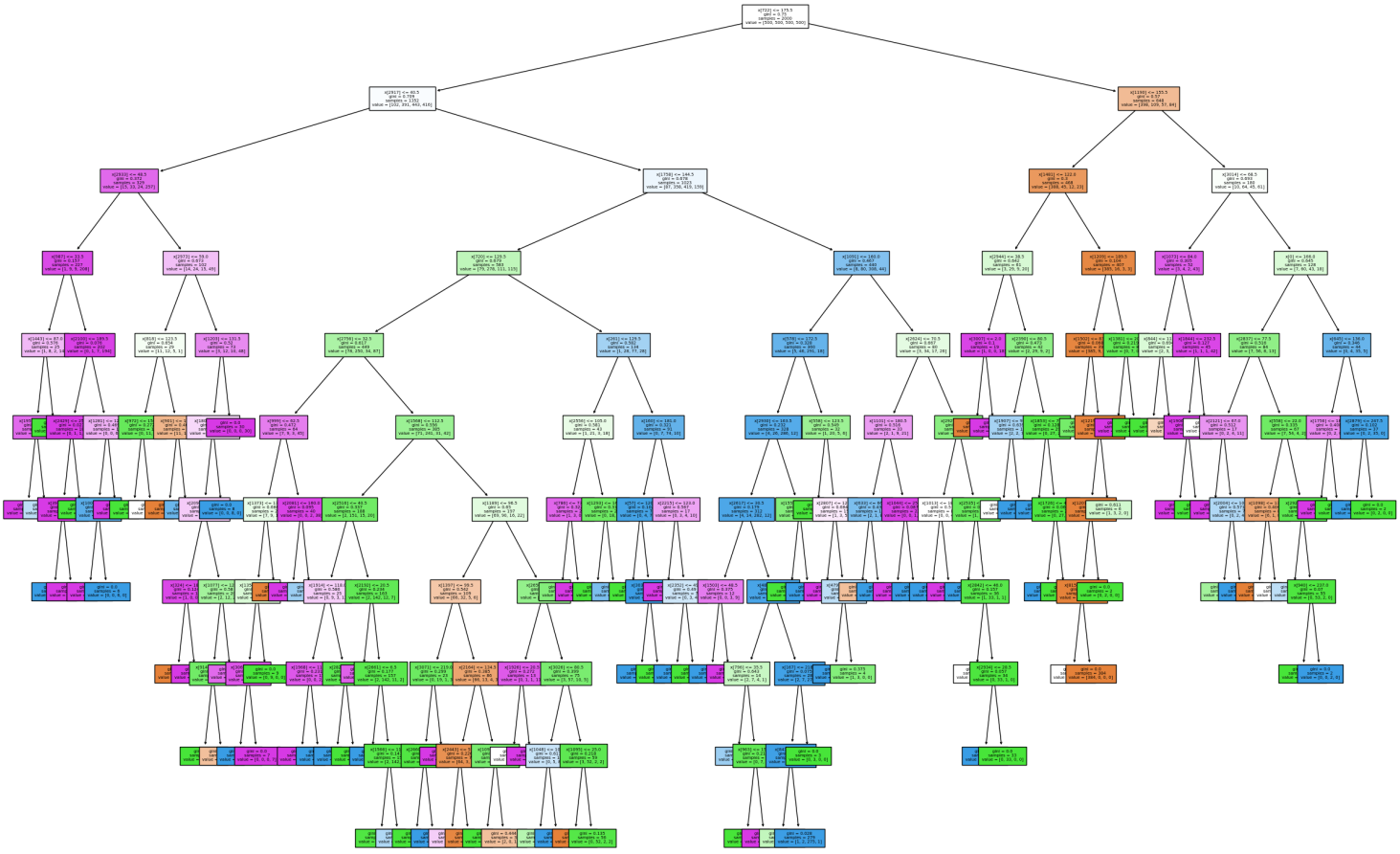
PRECISION = 0.72
RECALL = 0.72
ACCURACY = 0.72

b) Decision Tree Grid Search and visualisation selecting 10 best features

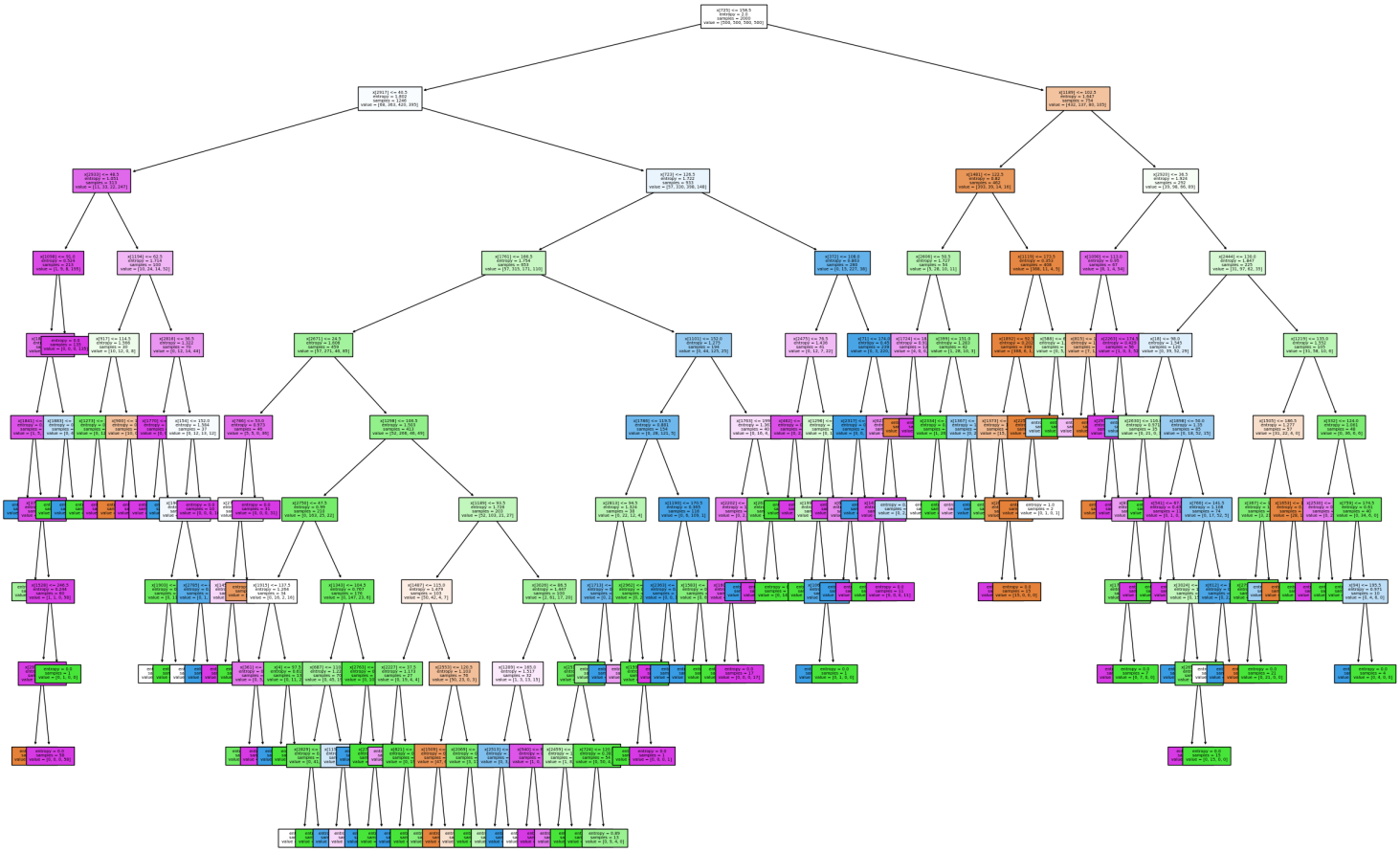
Training time : less than a second

Visualization

For criterion : Gini



For criterion : Entropy



Optimal set of parameters obtained are

- Criterion : entropy
- Max_depth : 5
- min_sample_split : 2

Training time : 3.8 seconds

Training set

PRECISION = 0.669
RECALL = 0.669
ACCURACY = 0.669

Validation set

PRECISION = 0.5975
RECALL = 0.5975
ACCURACY = 0.5975

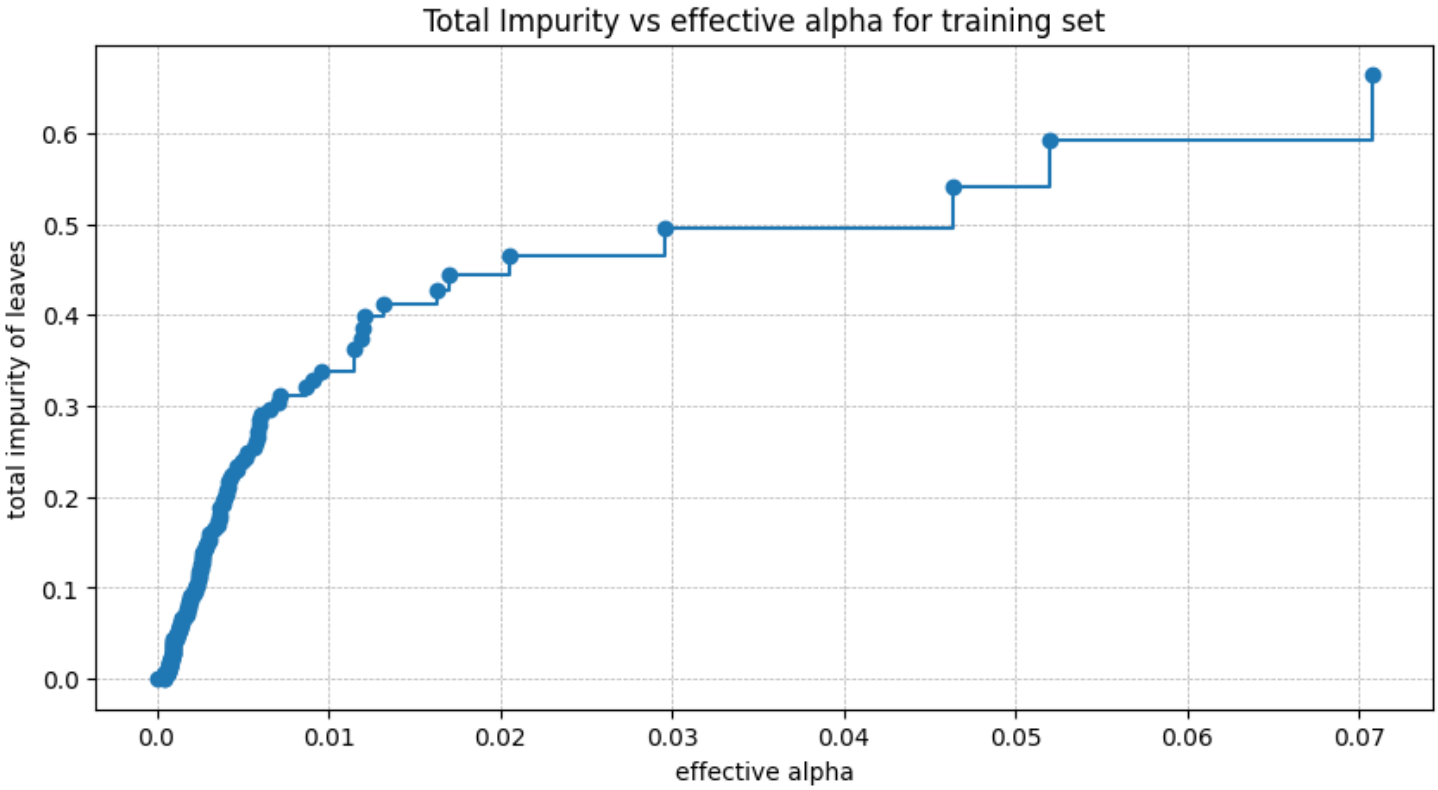
Comparison with A part

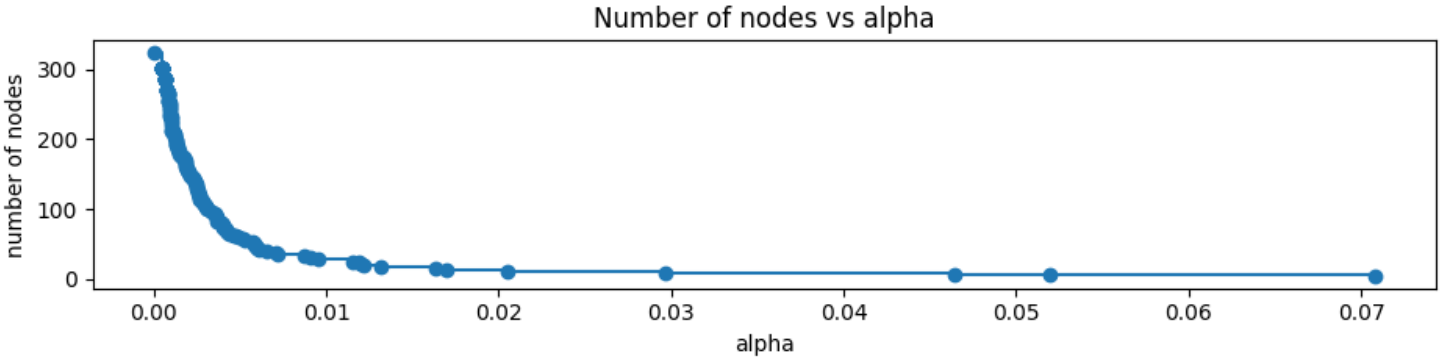
Since we are using just 10 features, the training time reduced significantly. On comparison with A,

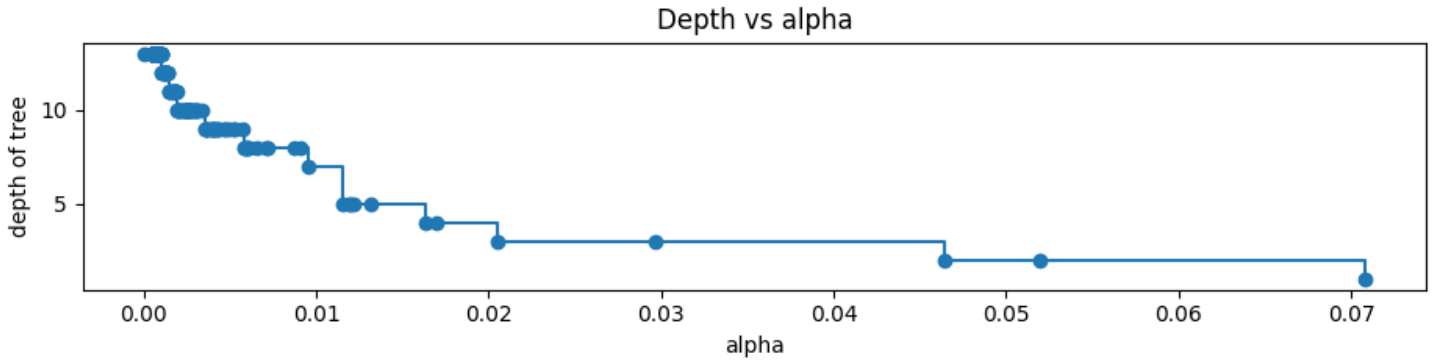
If we talk about the accuracies, the accuracy obtained in this part has also reduced significantly since we are now using just top 10 features instead of 32 times 32 times 3 features

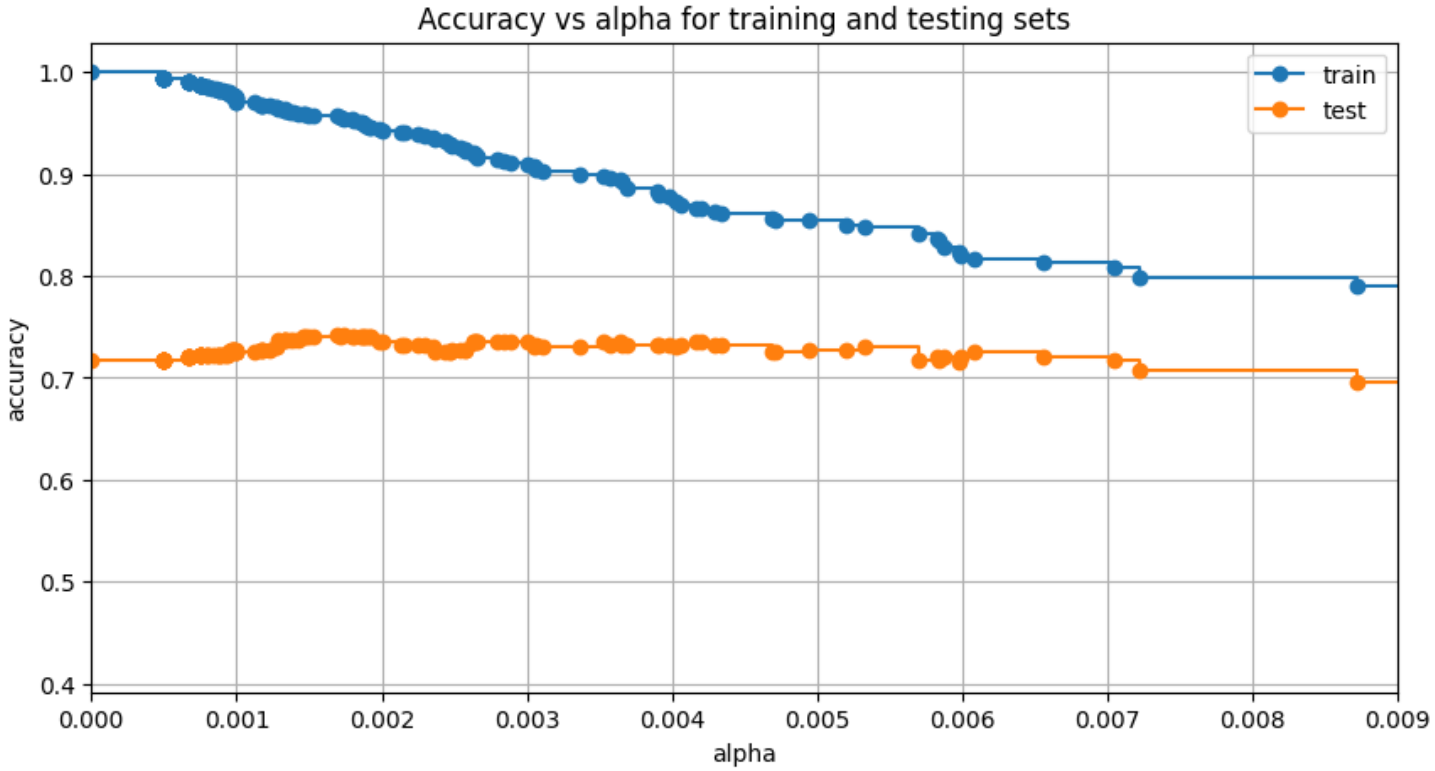
c) Decision Tree Post Pruning with Cost Complexity Pruning

Plots

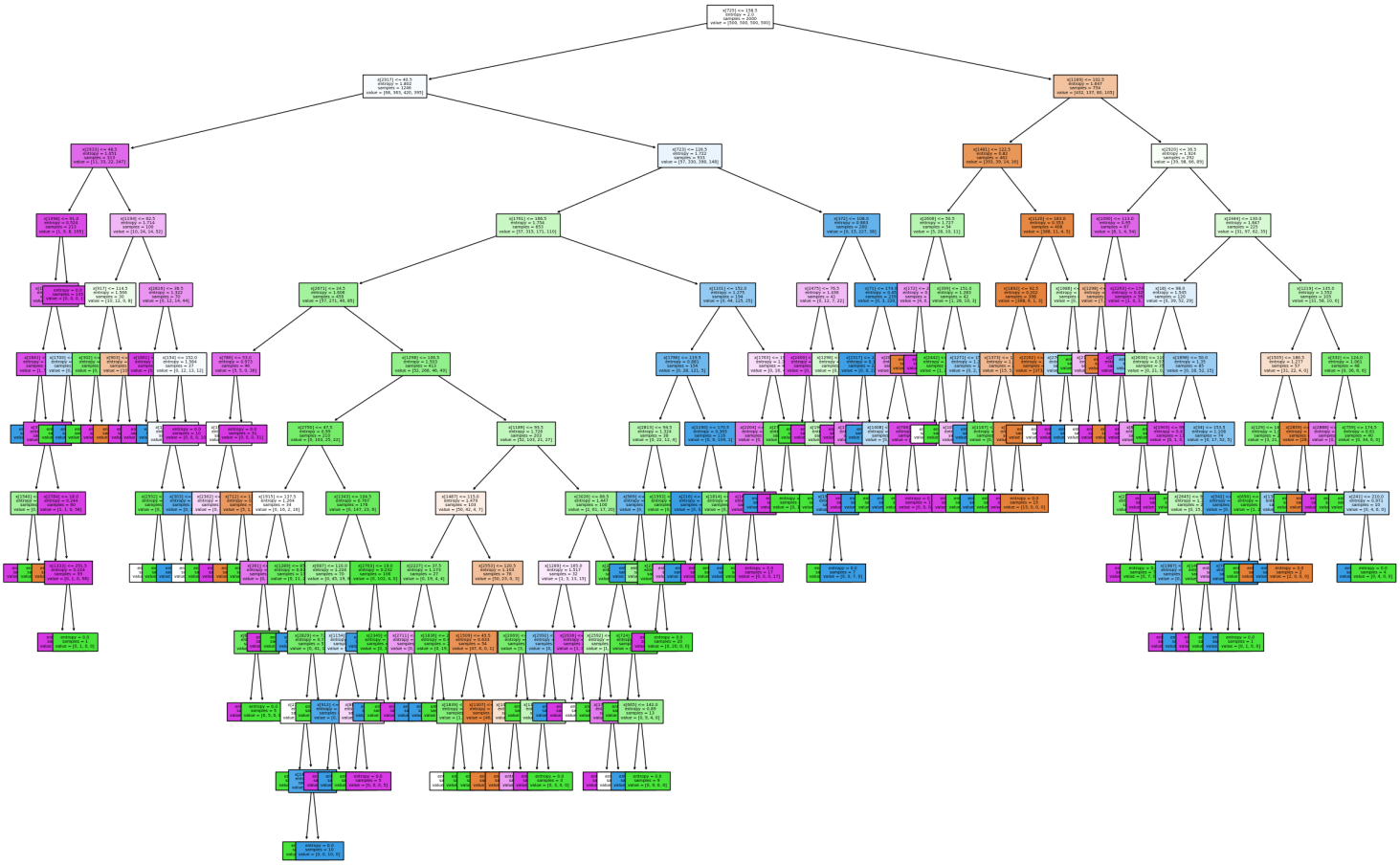








Visualization



Observations

The impurity of a node is a measure of how mixed the labels of the training examples are at that node.

We can see as the magnitude of alpha increase the total impurity decreases (which is sum of impurities of all nodes). By penalising the impurity of its nodes, increasing the value of alpha forces the tree to be simpler and smaller.

As we can see the total impurity decrease with increasing alpha →since the number of nodes are decreasing

We can also see that the number of nodes and the depth of the tree both decrease with increasing value of alphas.

Validation and training accuracy plot:
We can see that the training accuracy keep on decreasing with increasing alpha consistently. This is because we have lesser nodes for the split not and the tree do lesser overfitting.
For the validation accuracy : we can see that validaiton accuracy remains almost constant. I reaches maximum in the region between 0.001 and 0.002

Visualization : The tree generated has much lesser number of nodes than the one generated in part b

Train time -> 3.8 seconds

Train set

PRECISION = 0.9935
RECALL = 0.9935
ACCURACY = 0.9935

Validation set

PRECISION = 0.7275
RECALL = 0.7275
ACCURACY = 0.7275

d) Random Forest

Training time - > 12.9seconds

Optimal set of parameters obtained are

- Criterion : entropy
- Max_depth : None
- min_sample_split : 10
- n_estimators : 200

Train set

PRECISION = 1.0

RECALL = 1.0
ACCURACY = 1.0

Validation set

PRECISION = 0.88
RECALL = 0.88
ACCURACY = 0.88

e) Gradient Boosted Trees and XGBoost

Gradient boost

Train time -> 5min 34 sec

Optimal set of parameters obtained are

- max_depth : 9
- n_estimators : 50
- subsample : 0.5

Train set

PRECISION = 1.0
RECALL = 1.0
ACCURACY = 1.0

Validation set

PRECISION = 0.6725
RECALL = 0.6725
ACCURACY = 0.6725

XGradient Boost

Train time -> 35 sec

Optimal set of parameters obtained are

- max_depth : 9
- n_estimators : 50

- subsample : 0.6

Train set

PRECISION = 1.0
RECALL = 1.0
ACCURACY = 1.0

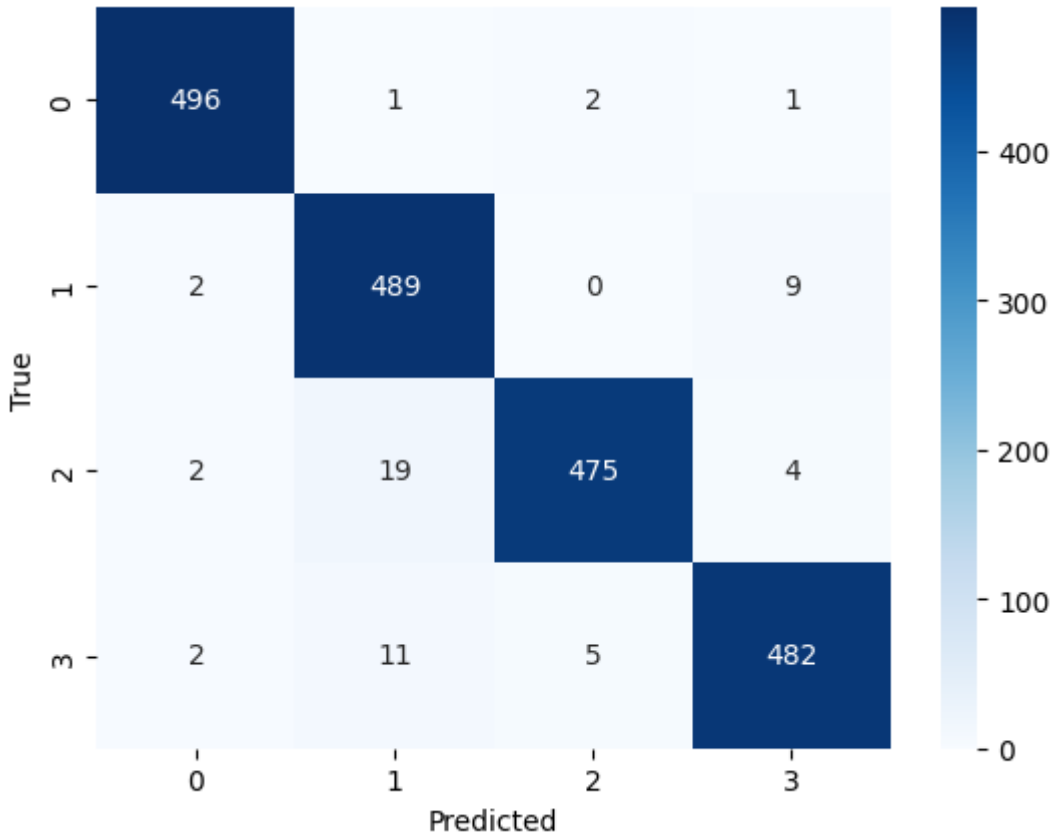
Validation set

PRECISION = 0.68
RECALL = 0.68
ACCURACY = 0.68

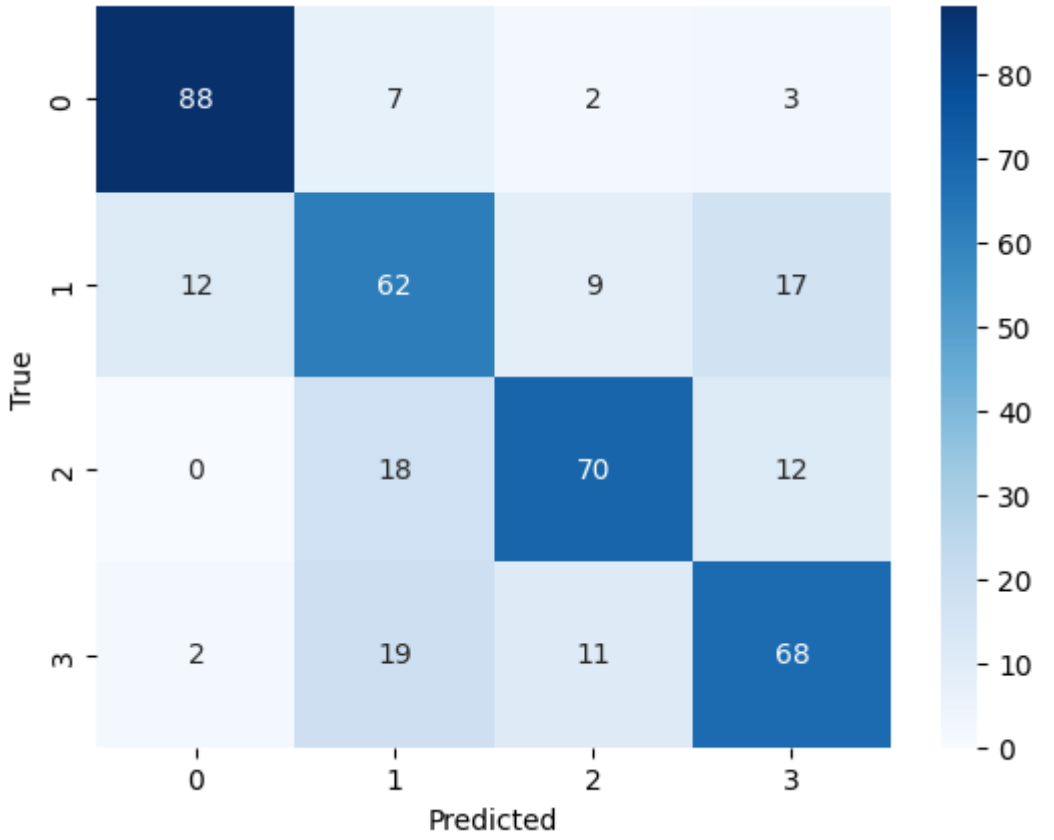
f) Confusion matrix

Part A (Entropy)

Train set

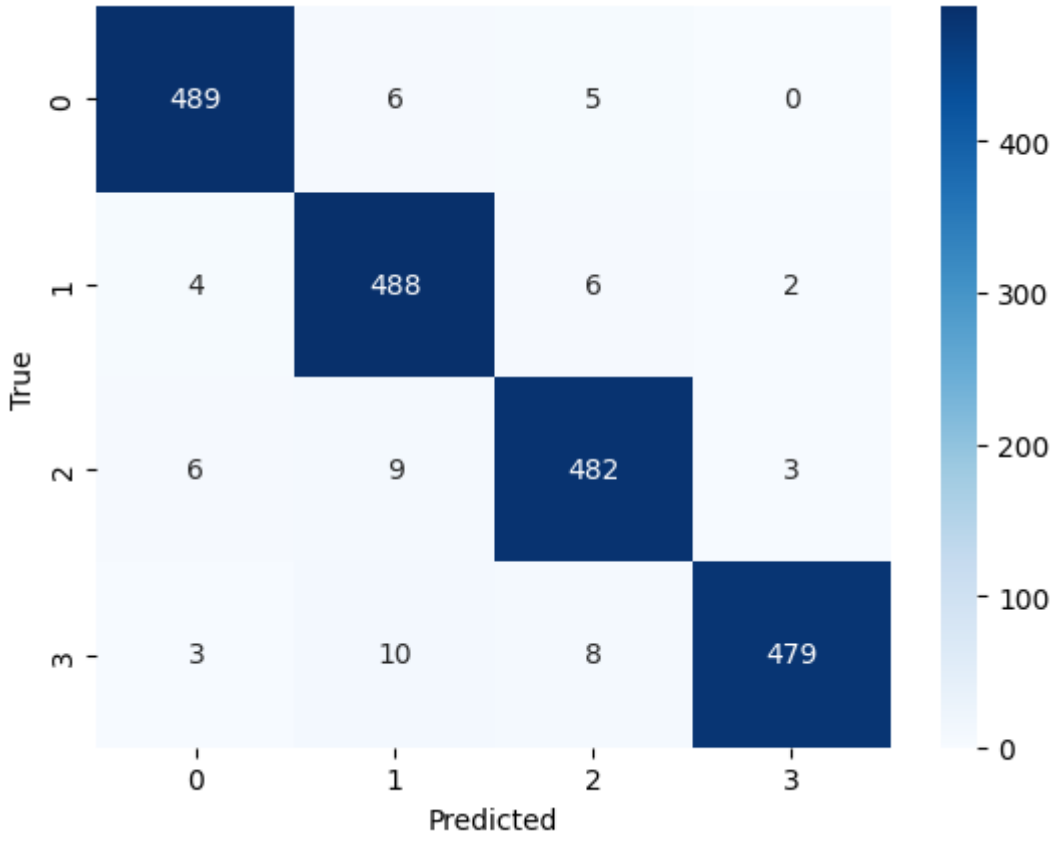


Validation set

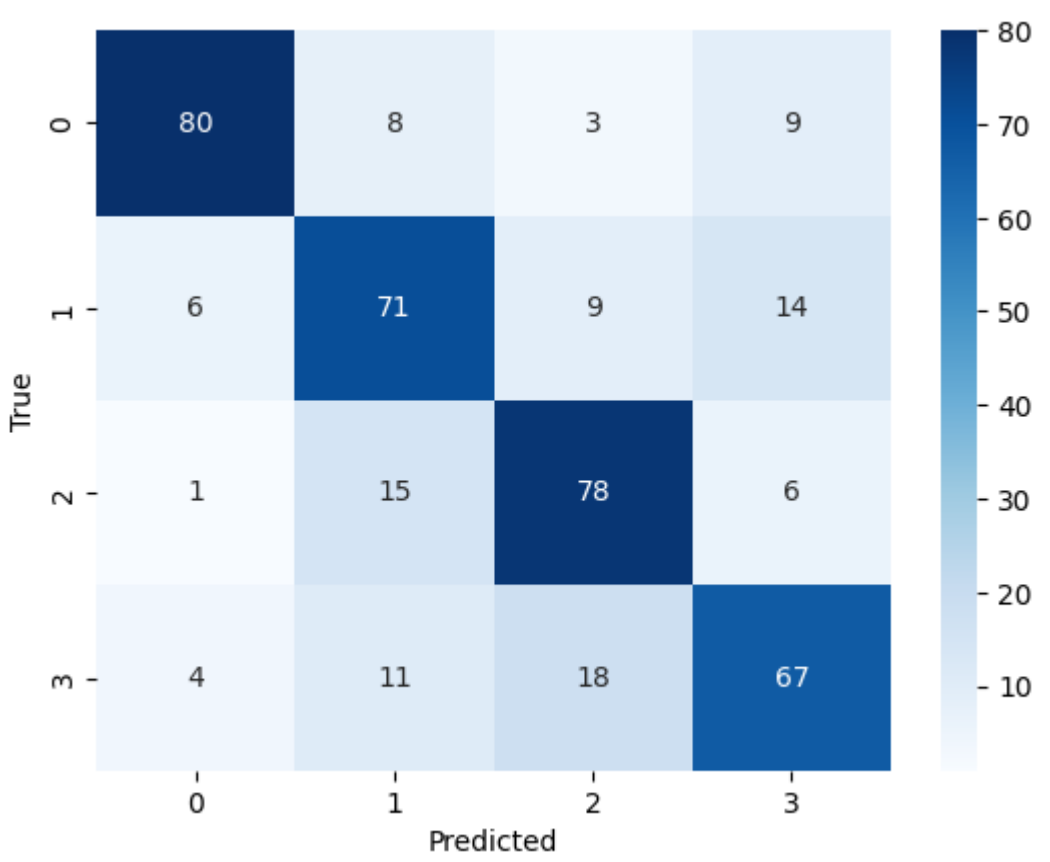


Part A (Gini)

Train set

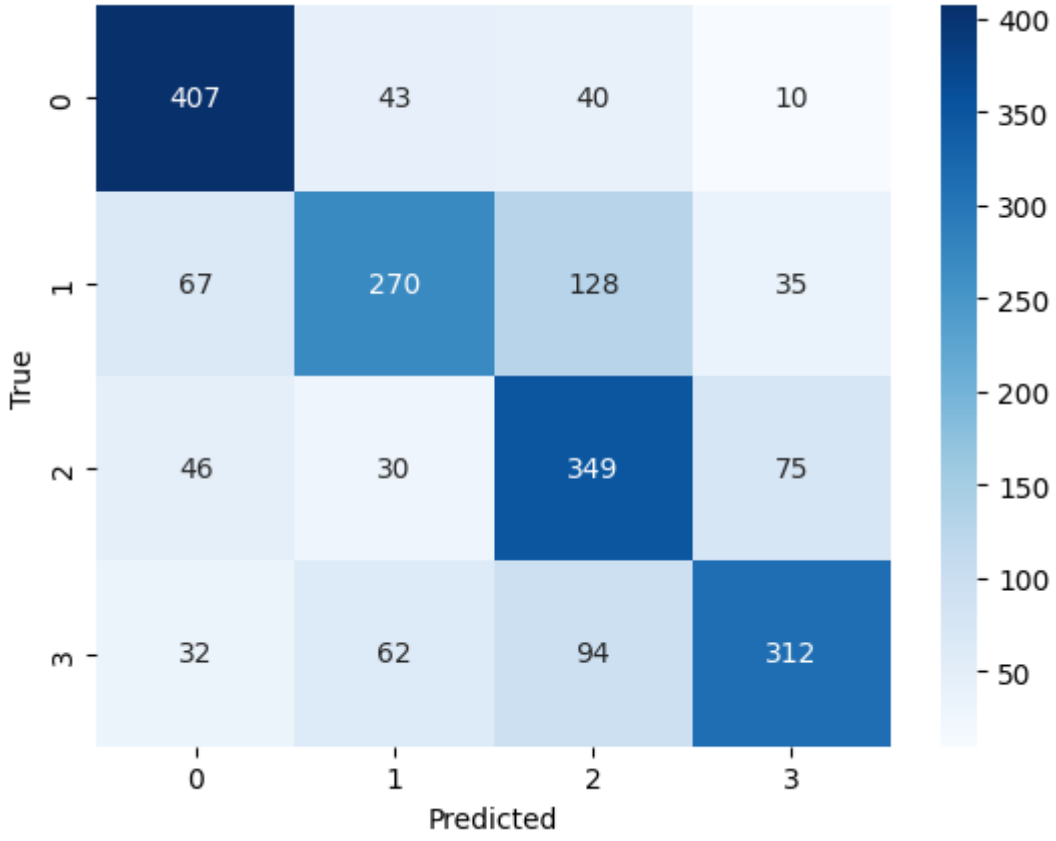


Validation set

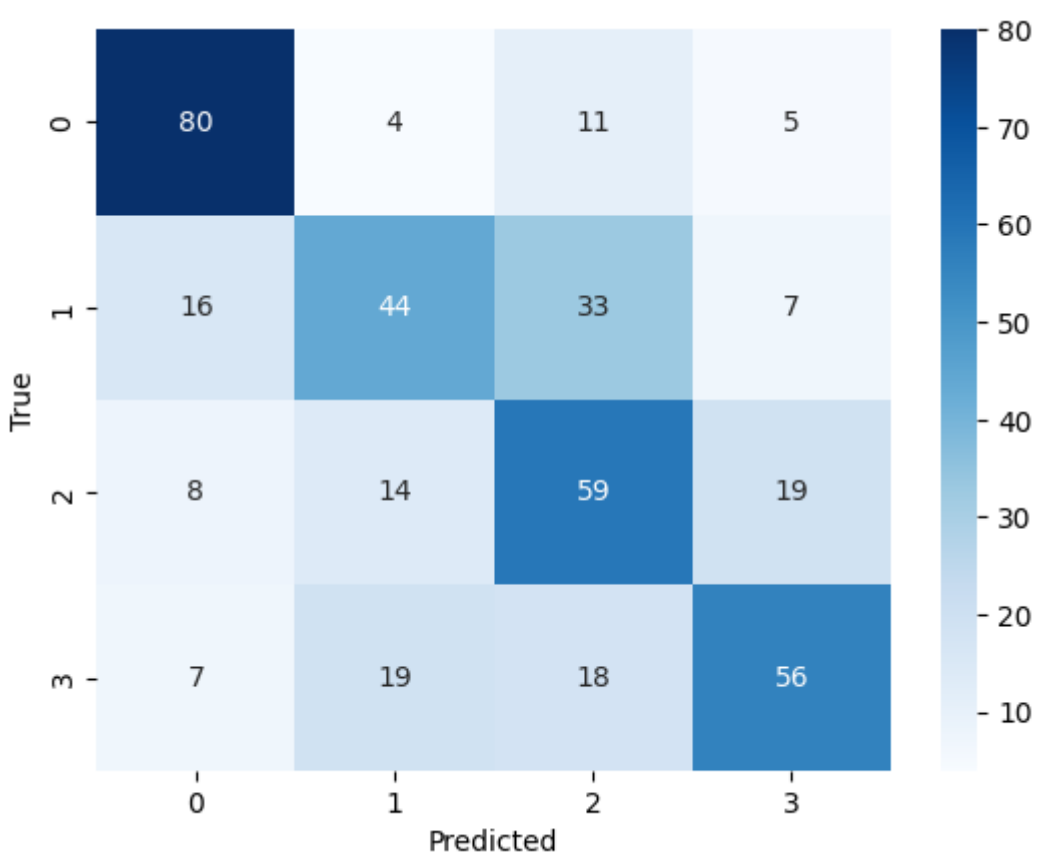


Part B

Train set

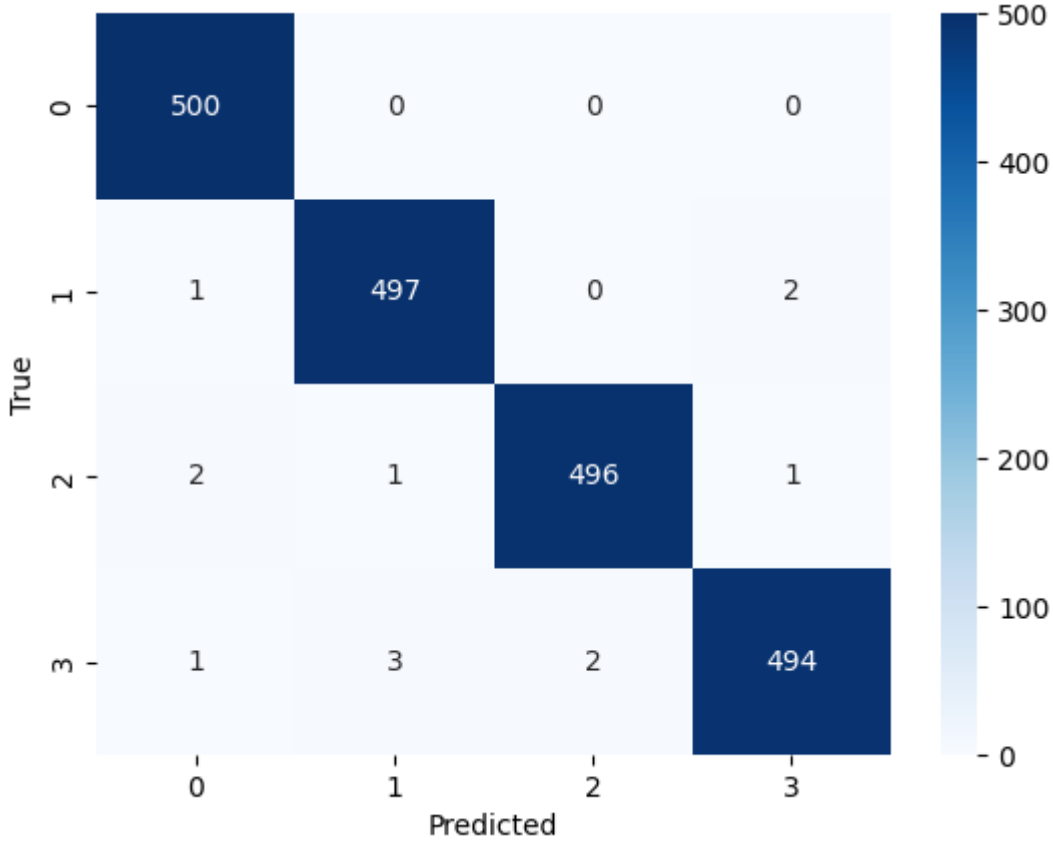


Validation set

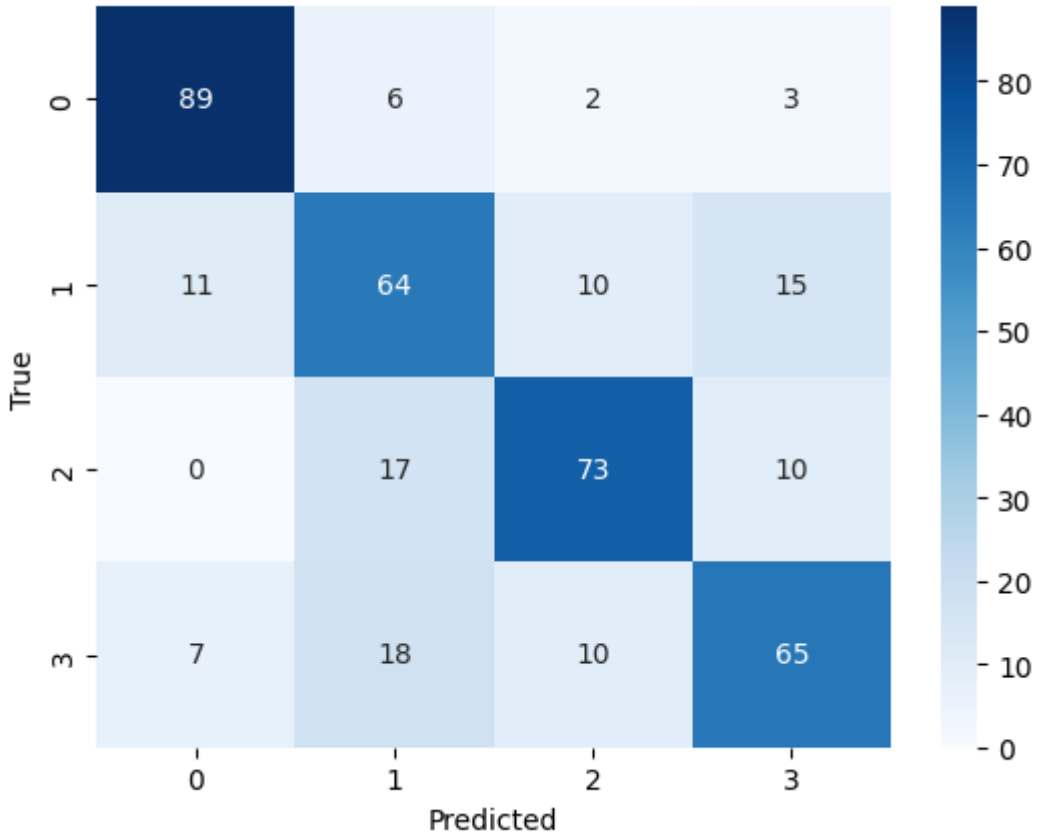


Part C

Train set

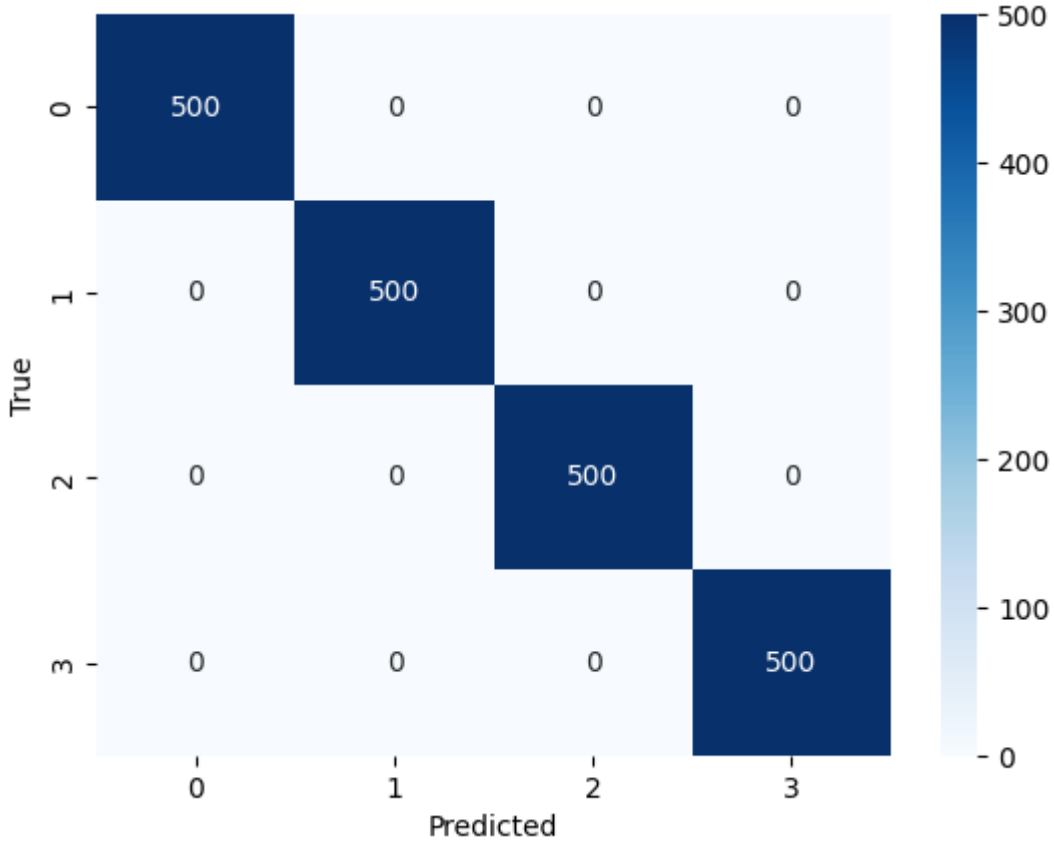


Validation set

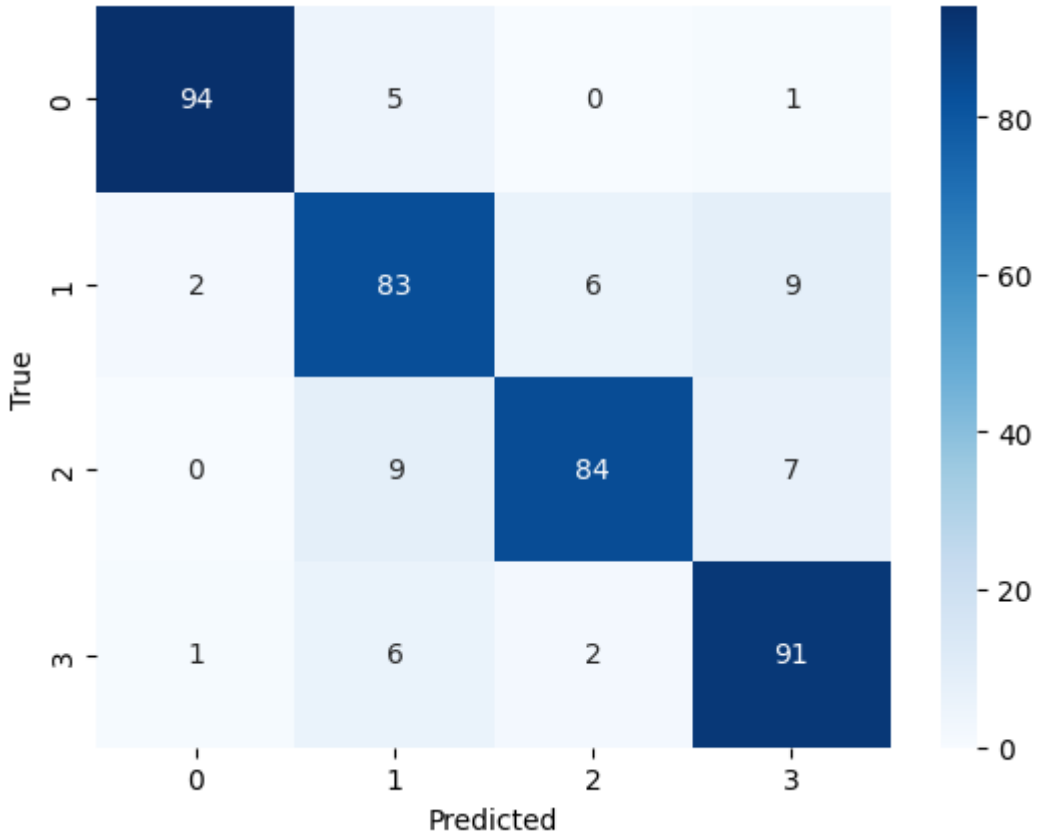


Part D

Train set

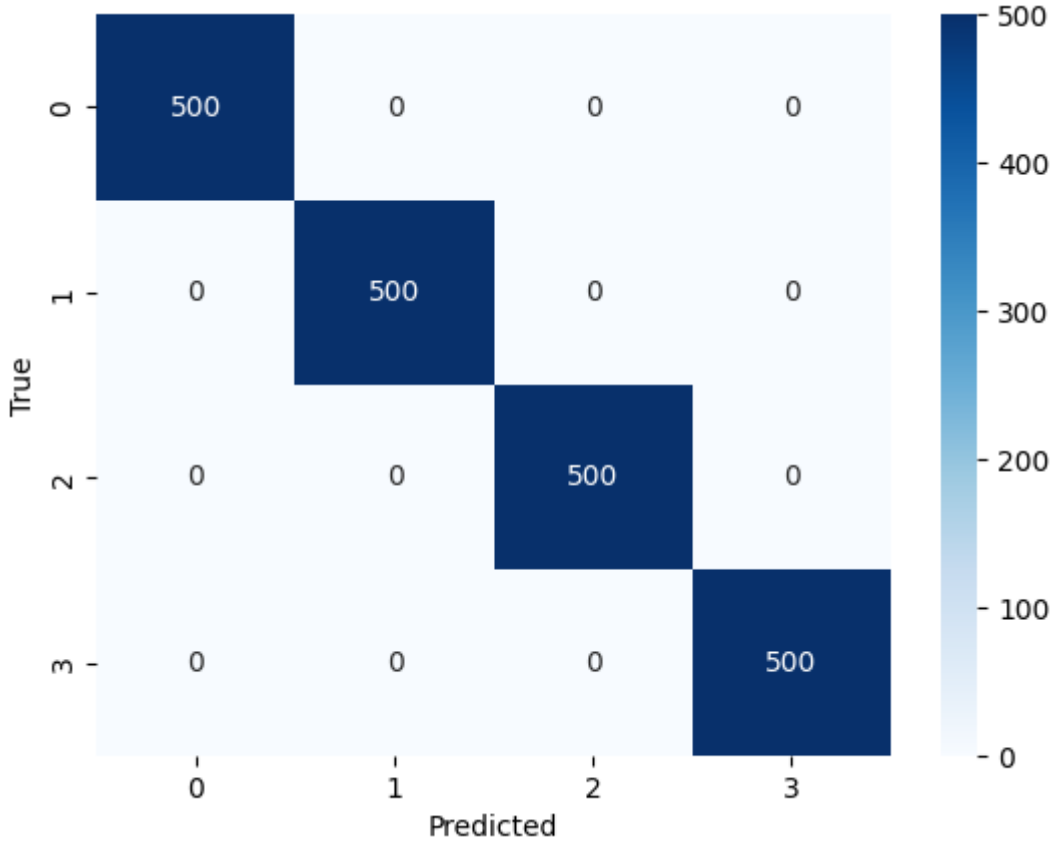


Validation set

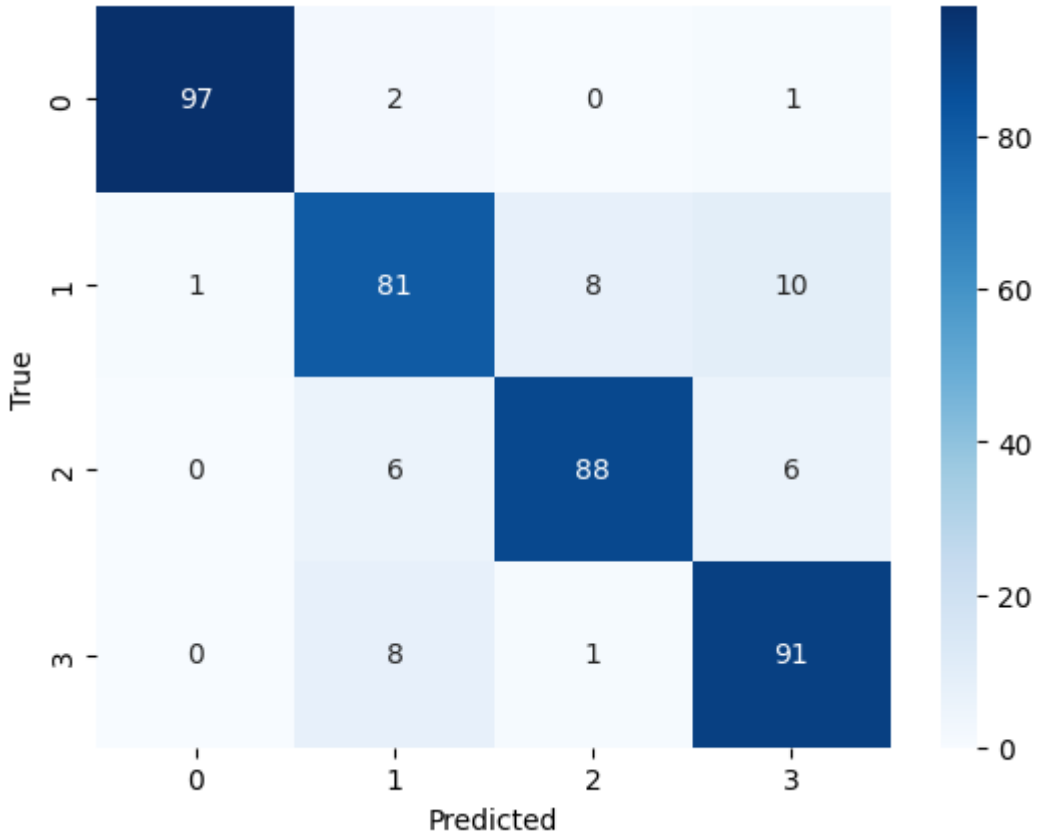


Part E (gradient Boost)

Train set

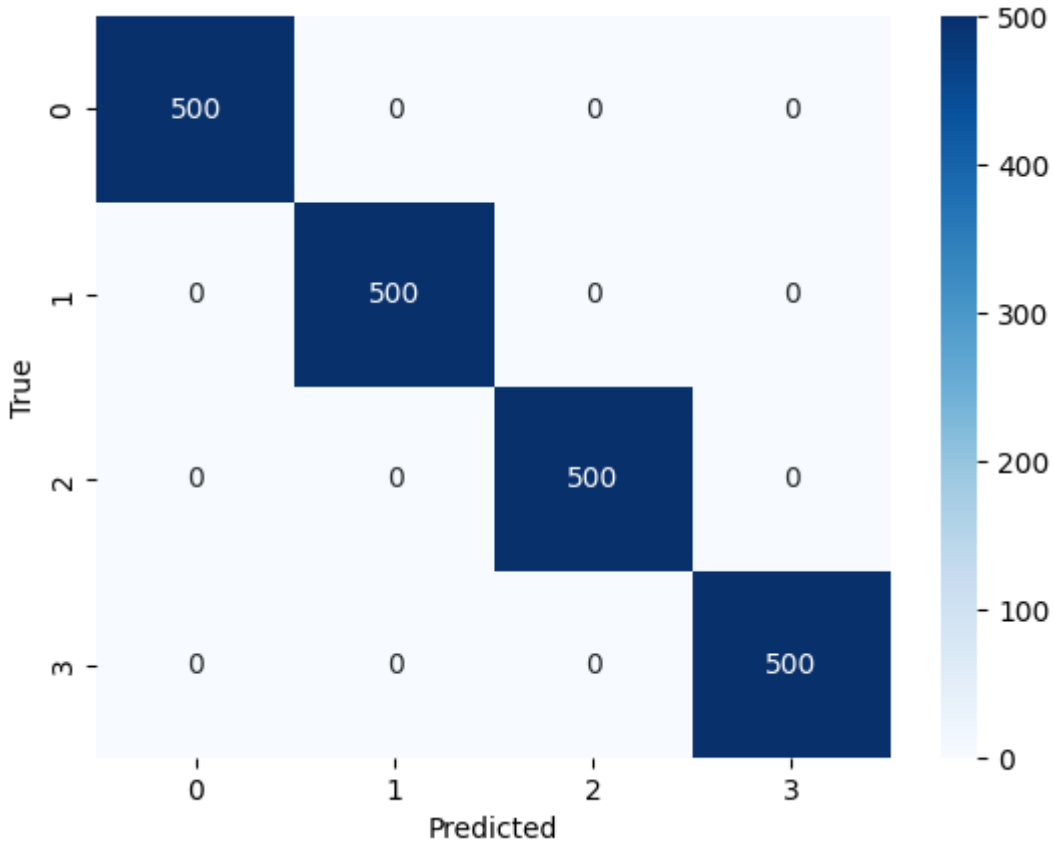


Validation set

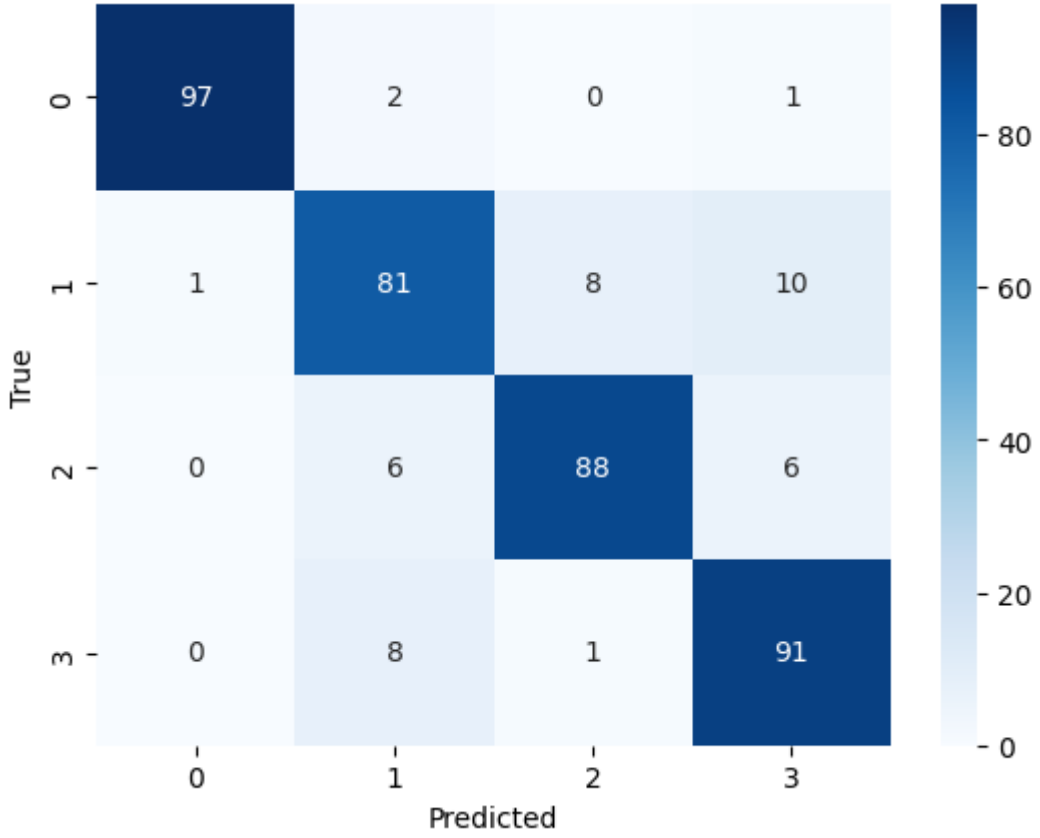


Part E (Xgradient Boost)

Train set



Validation set



g) Real-time Application

Decision Tree sklearn

outputs generated
2.png,1.0
7.png,3.0
9.png,1.0
5.png,3.0
10.png,1.0
3.png,3.0
6.png,3.0
1.png,2.0
8.png,1.0
4.png,3.0
accuracy = 40%

Decision Tree Grid Search and visualisation

outputs generated
2.png,3.0
7.png,3.0
9.png,3.0
5.png,3.0
10.png,3.0

3.png,3.0
6.png,2.0
1.png,1.0
8.png,1.0
4.png,3.0
accuracy = 20%

Decision Tree Post Pruning with Cost Complexity Pruning

2.png,1.0
7.png,3.0
9.png,1.0
5.png,3.0
10.png,1.0
3.png,3.0
6.png,3.0
1.png,3.0
8.png,1.0
4.png,0.0
accuracy = 40%

Random Forest

2.png,3.0
7.png,3.0
9.png,3.0
5.png,3.0
10.png,3.0
3.png,3.0
6.png,3.0
1.png,2.0
8.png,3.0
4.png,3.0
accuracy=0%

XGBoost

2.png,3
7.png,3
9.png,3
5.png,3
10.png,3
3.png,3

6.png,3
1.png,3
8.png,3
4.png,3
accuracy=0%