# COL341 - Assignment 3

**Eshan Jain**
**2020CS50424**

## 3.1 Binary Classification

### a) Decision Tree from Scratch

Time to build the information gain decision tree:
6201.594185352325 seconds

Information Gain Decision Tree:

Training Accuracy: 0.9805

Training Precision: 0.9600798403193613

Training Recall: 0.962

Training Confusion Matrix:

[[481, 20], [19, 1480]]

Validation Accuracy: 0.76

Validation Precision: 0.5416666666666666

Validation Recall: 0.26

Validation Confusion Matrix:

[[26, 22], [74, 278]]

**b) Decision Tree sklearn**

Training Time: 4.319708347320557 seconds

Training Accuracy: 0.9885

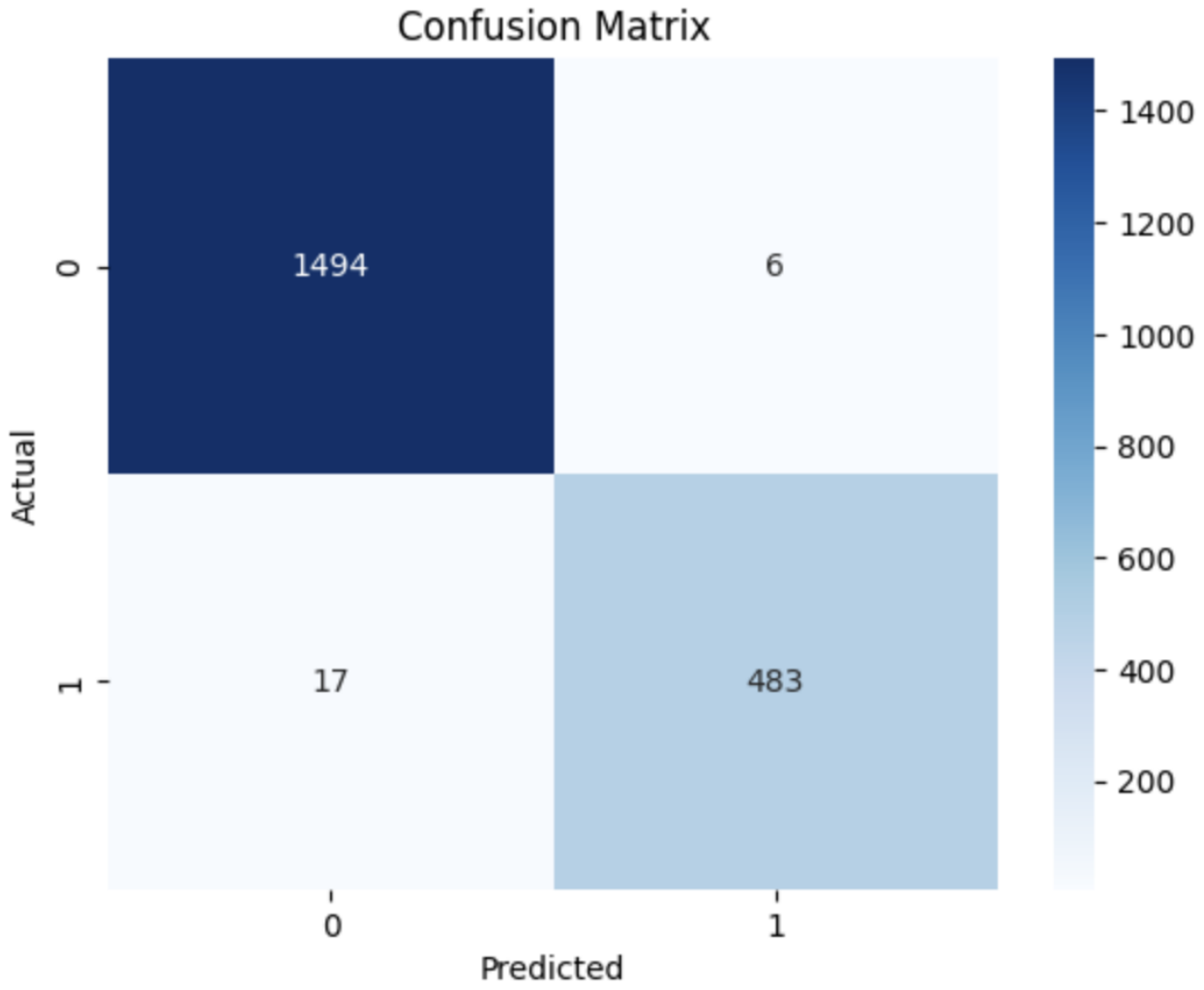
Training Precision: 0.9877300613496932

Training Recall: 0.966

Training Confusion Matrix:

[[1494 6]

[ 17 483]]




Validation Accuracy: 0.93

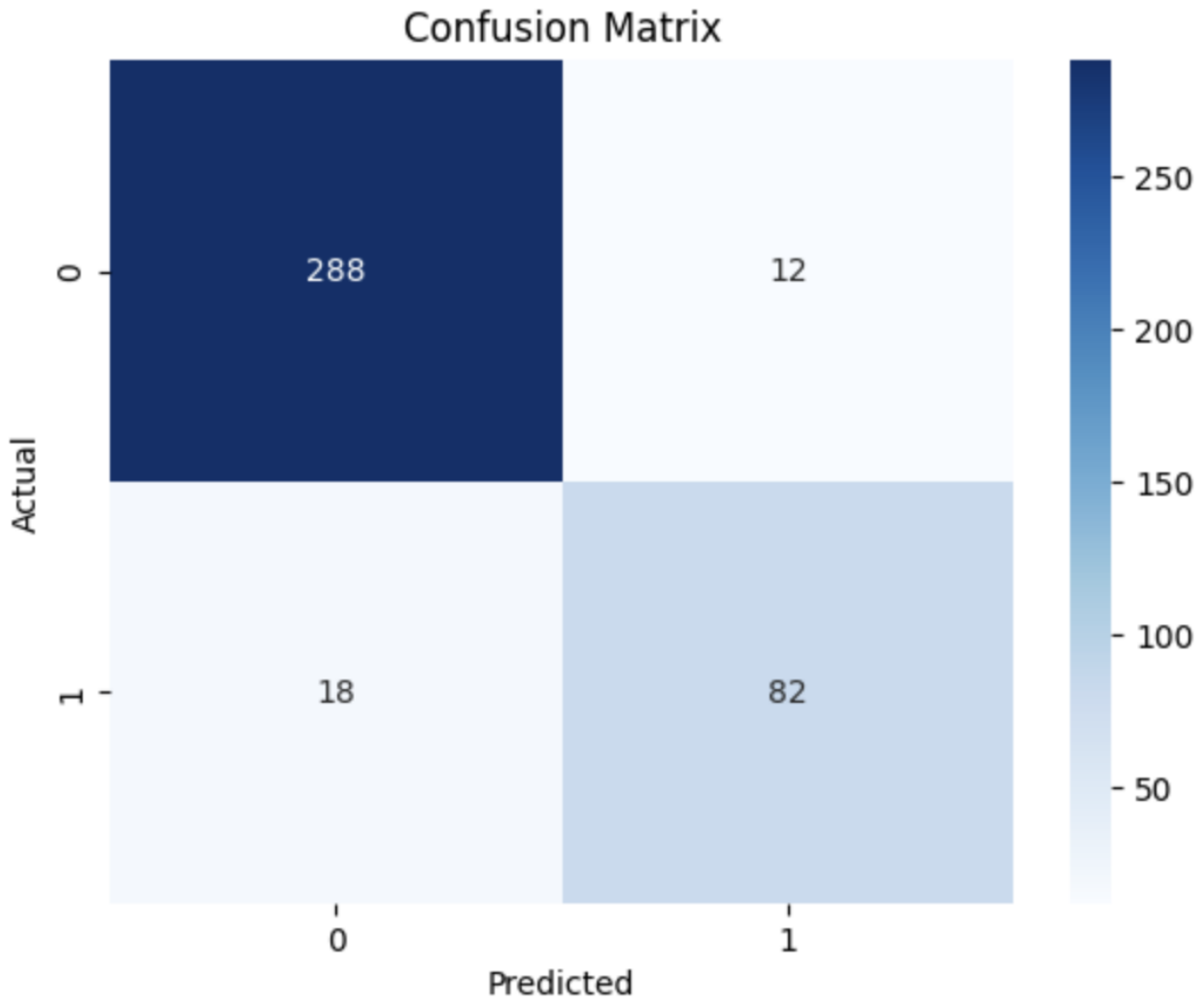
Validation Precision: 0.9

Validation Recall: 0.81

Validation Confusion Matrix:

[[291 9]

[ 19 81]]




**c) Decision Tree Grid Search and Visualisation**

**Selecting top 10 features:**

Training Time in case of top 10 features: 0.030390024185180664

Training Accuracy for top 10 features: 0.9235

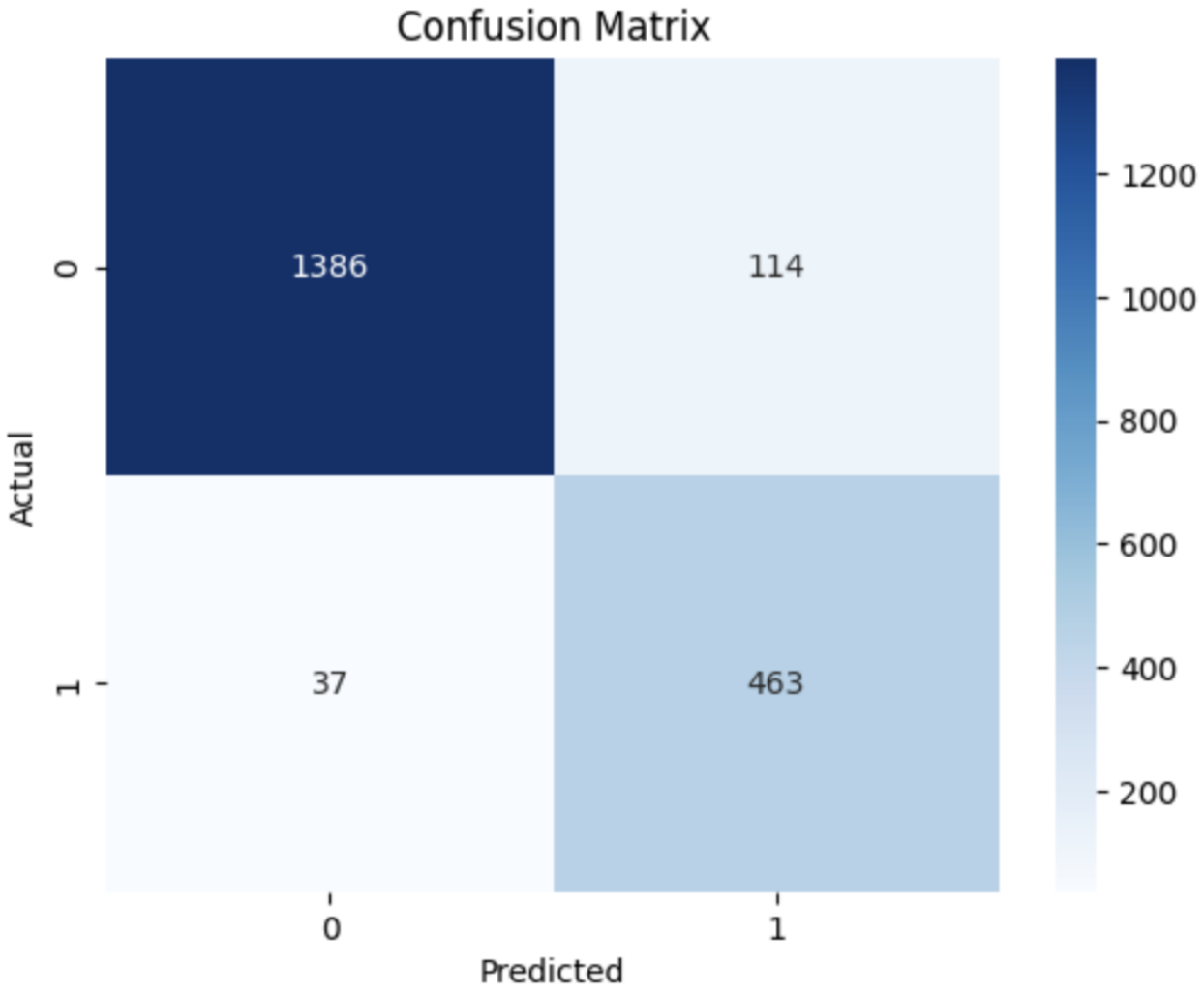
Training Precision for top 10 features: 0.8126126126126126

Training Recall for top 10 features: 0.902

Training Confusion Matrix for top 10 features:

[[1396 104]

[ 49 451]]




Validation Accuracy for top 10 features: 0.865

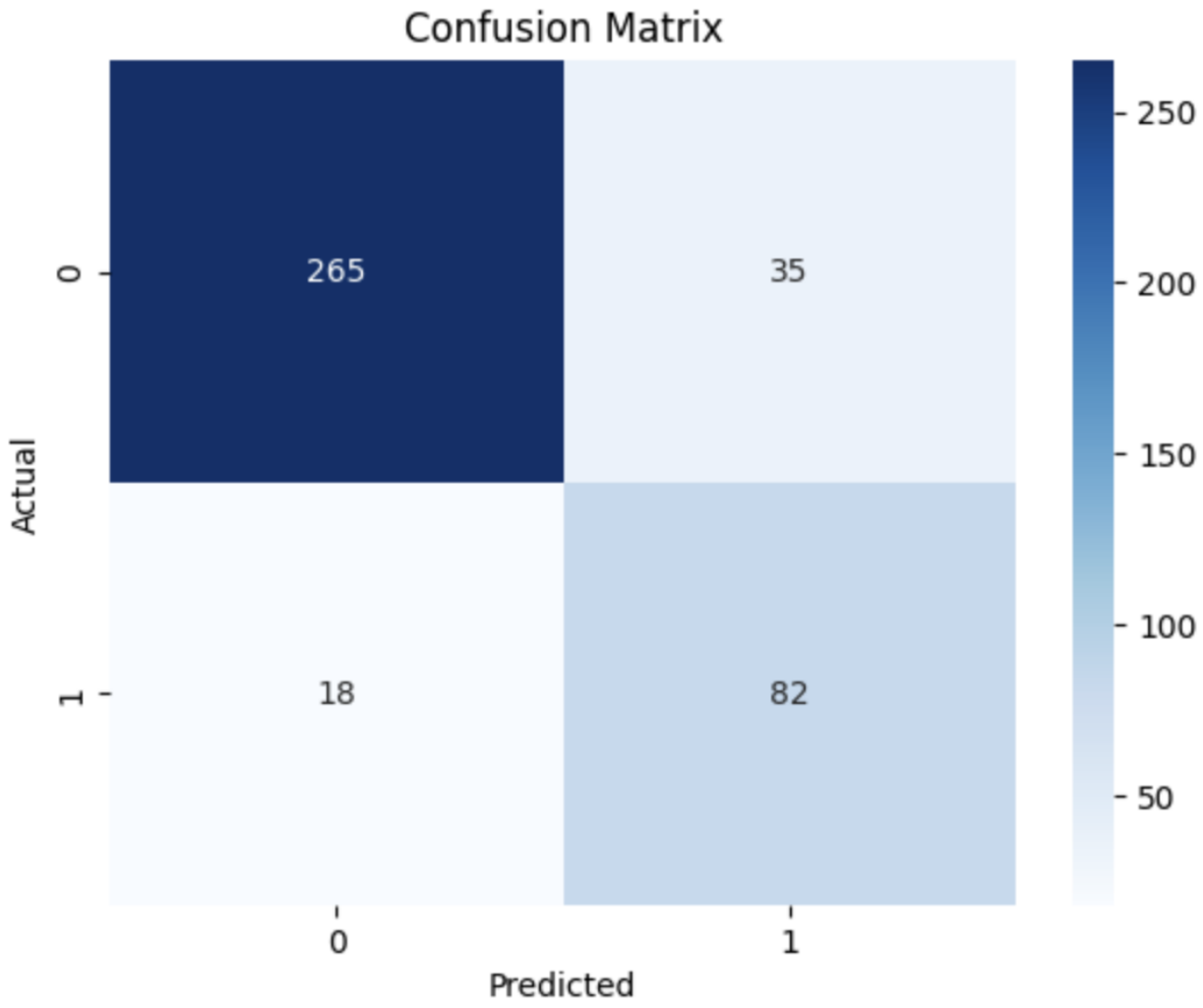
Validation Precision for top 10 features: 0.7017543859649122

Validation Recall for top 10 features: 0.8

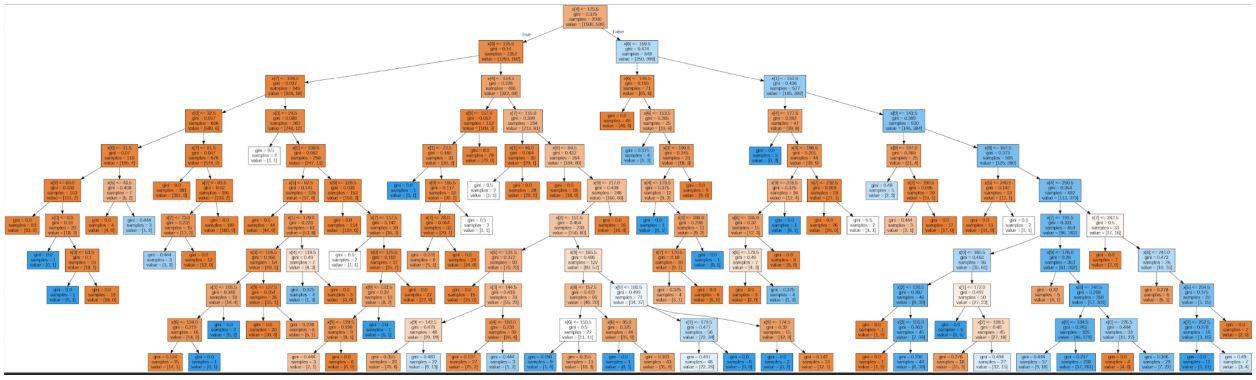
Validation Confusion Matrix for top 10 features:

[[266 34]

[ 20 80]]



**Visualising the tree:**



**Grid Search over the top 10 features:**

Best Hyperparameters: {'criterion': 'gini', 'max_depth': 5, 'min_samples_split': 4}

Training Accuracy after performing grid search: 0.887

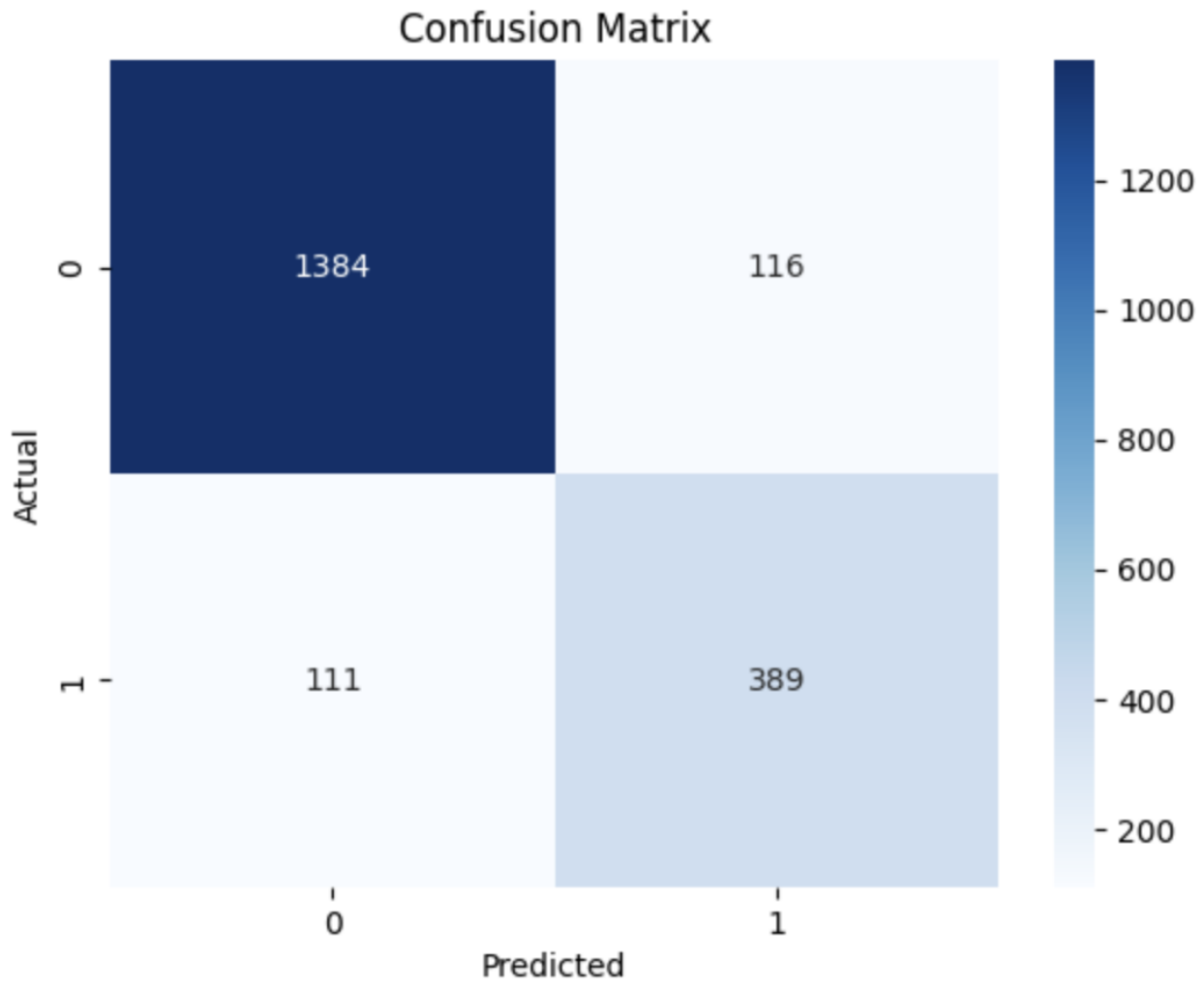
Training Precision after performing grid search: 0.7718253968253969

Training Recall after performing grid search: 0.778

Training Confusion Matrix after performing grid search:

[[1385 115]

[ 111 389]]

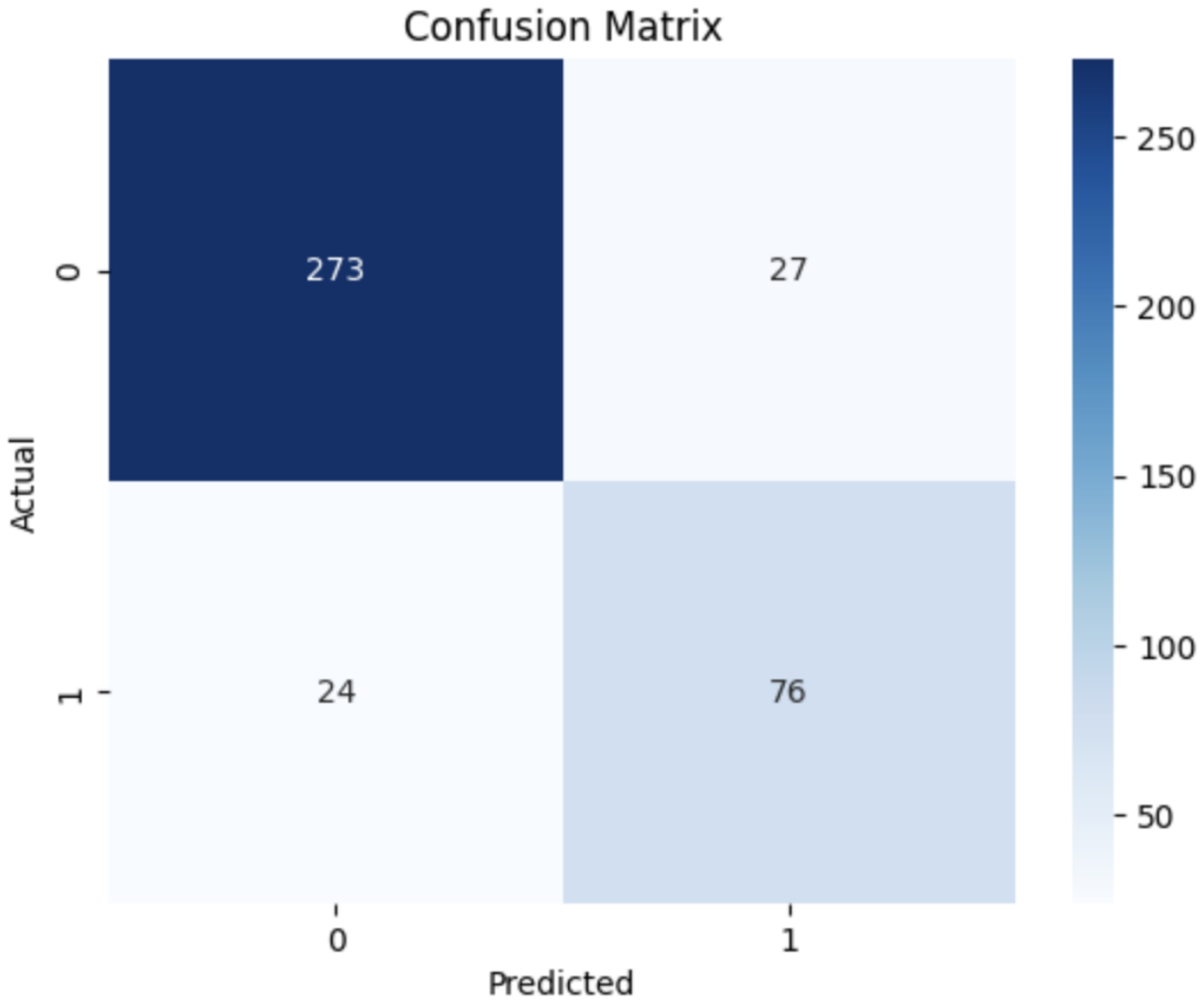



Validation Accuracy after performing grid search: 0.8725

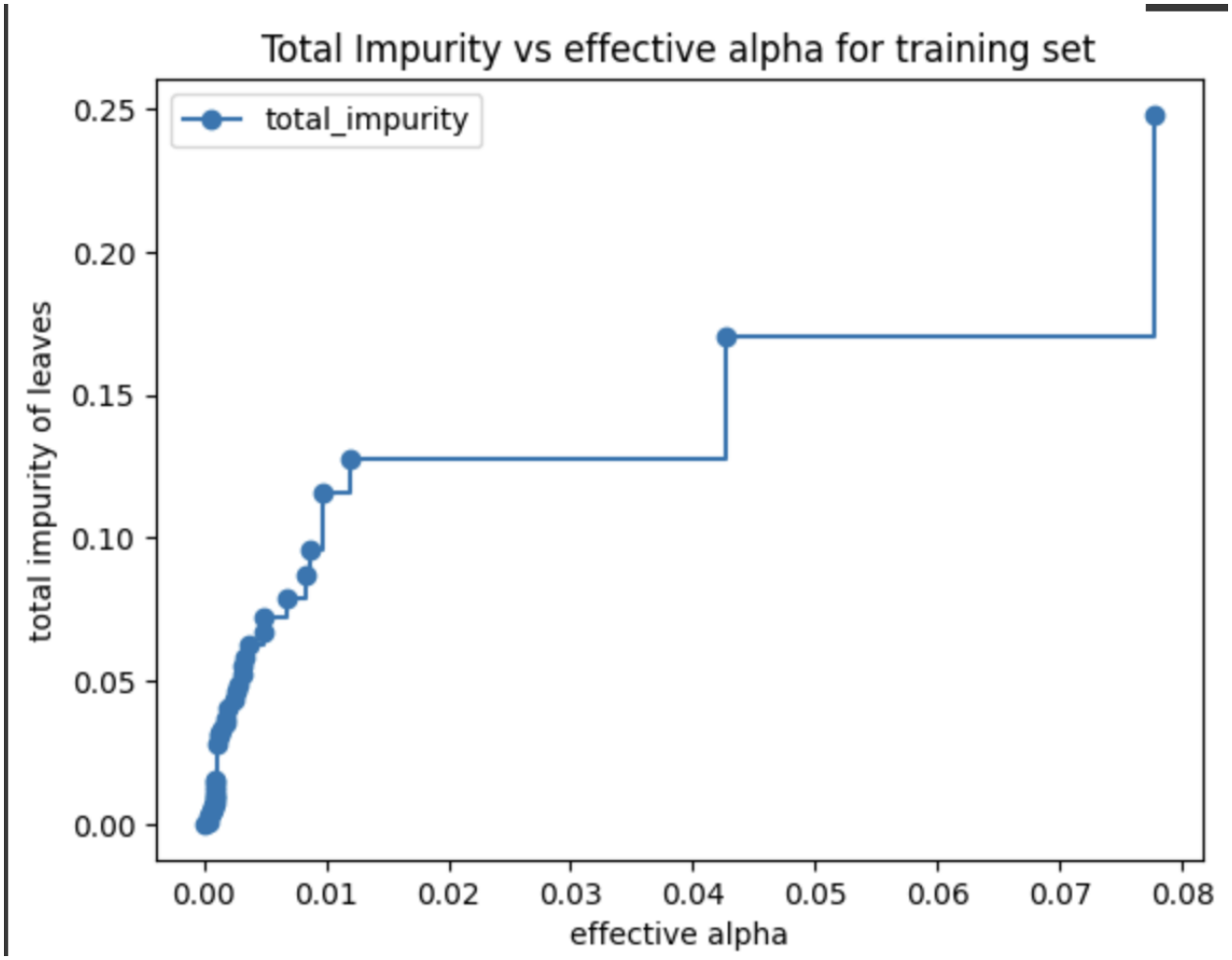
Validation Precision after performing grid search: 0.7378640776699029

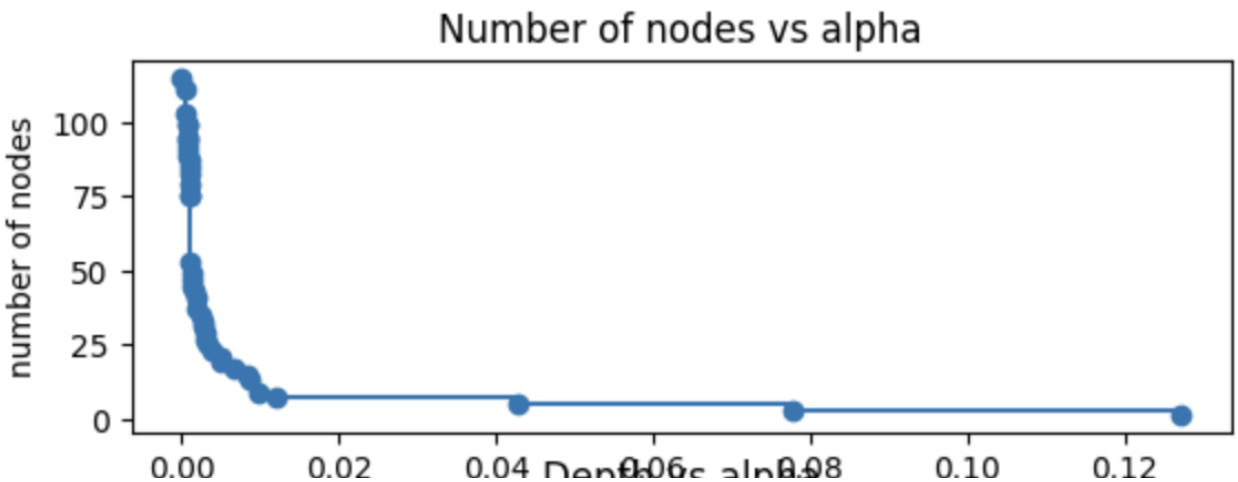
Validation Recall after performing grid search: 0.76

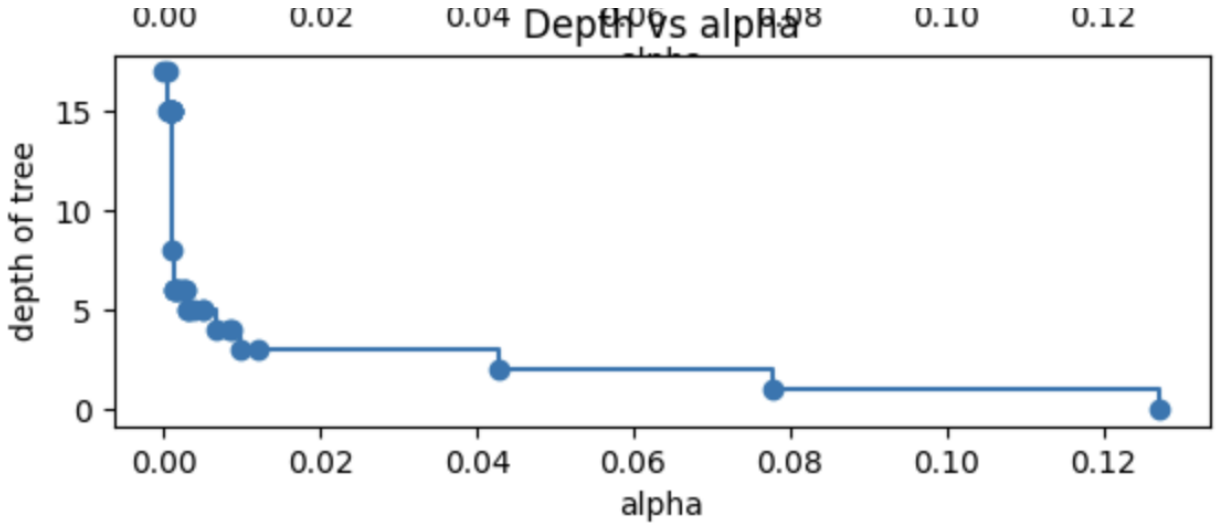
Validation Confusion Matrix after performing grid search:

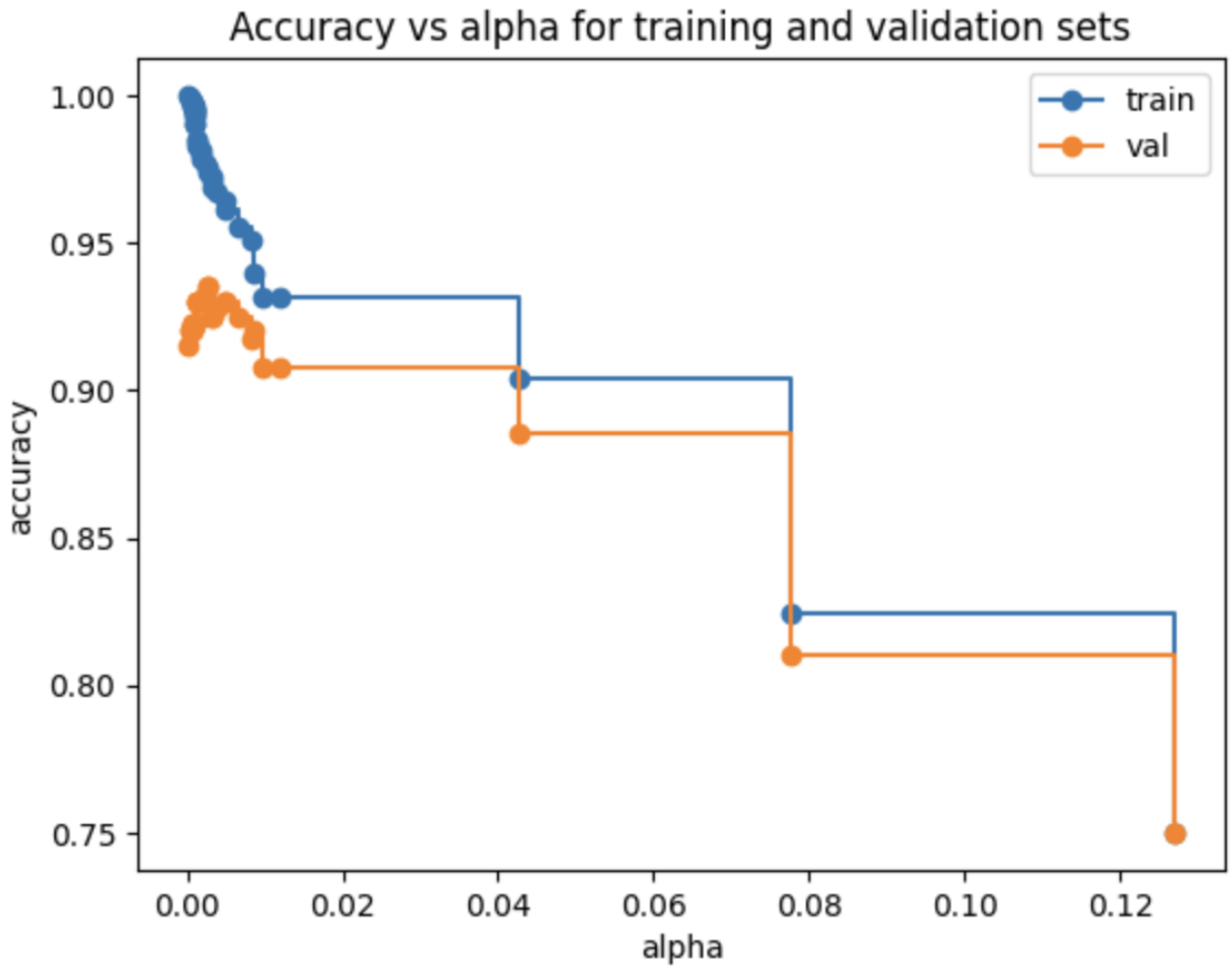


## d) Decision Tree Post Pruning with Cost Complexity Pruning




Number of nodes vs alpha
number of nodes
100
75
50
25
0
0.00
0.02
0.04
0.06
0.08
0.10
0.12


Depth vs alpha
depth of tree
15
10
5
0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
alpha




Best alpha: 0.0027352941176470593

**Best performing tree statistics:**

Best training accuracy: 0.9755

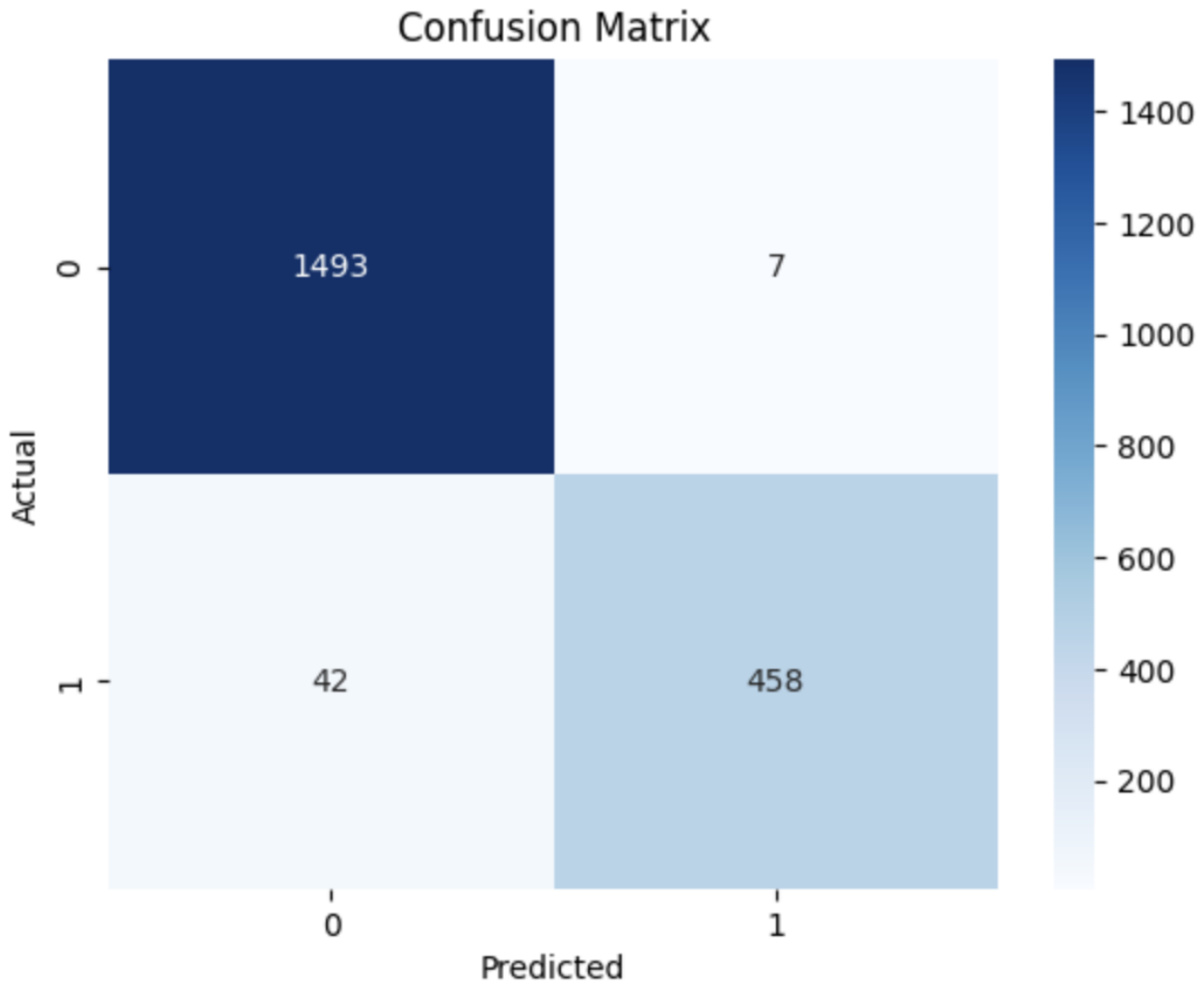
Best training precision: 0.9849462365591398

Best training recall: 0.916

Training Confusion Matrix:

[[1493 7]

[ 42 458]]




Best validation accuracy: 0.935

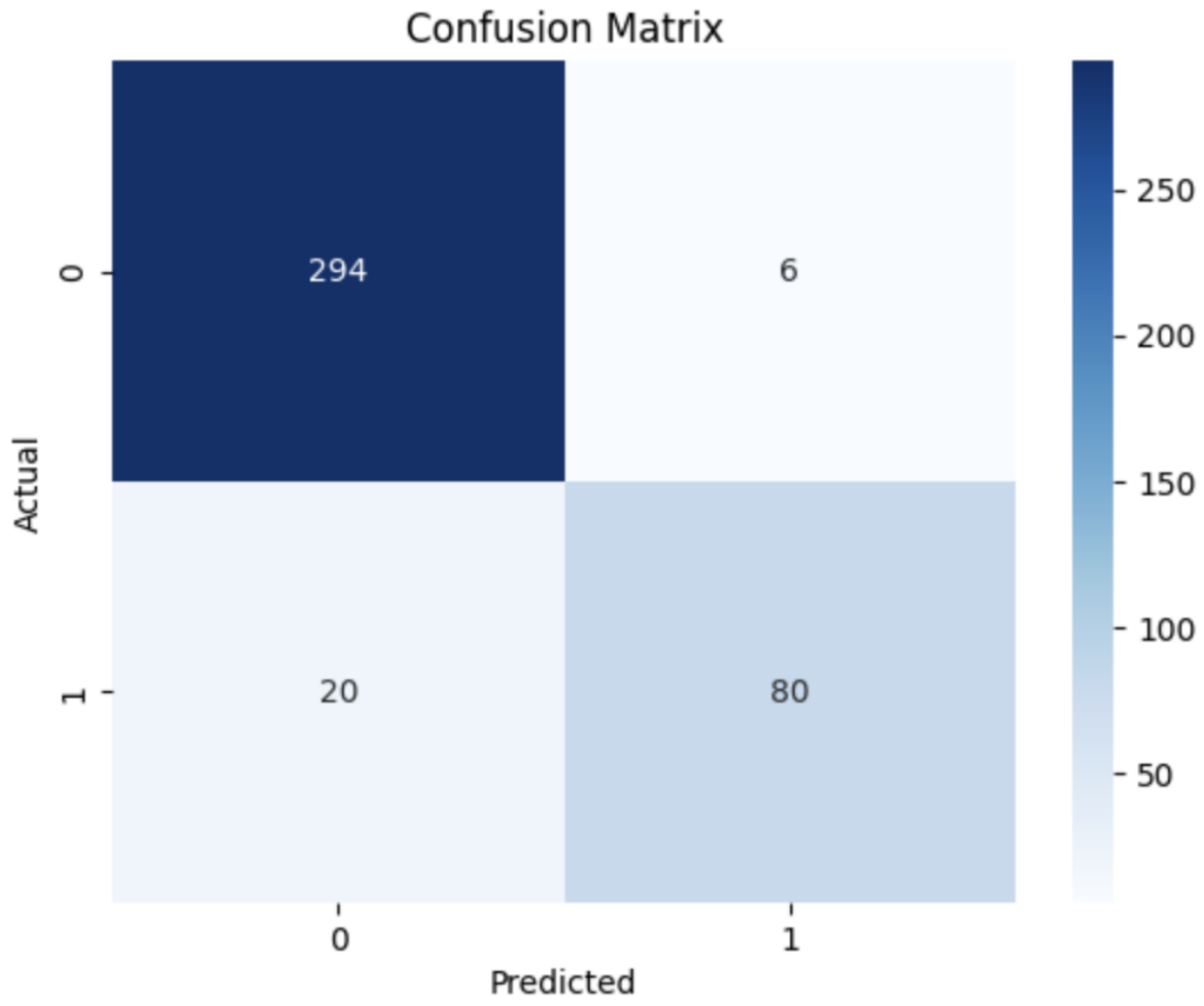
Best validation precision: 0.9302325581395349

Best validation recall: 0.8

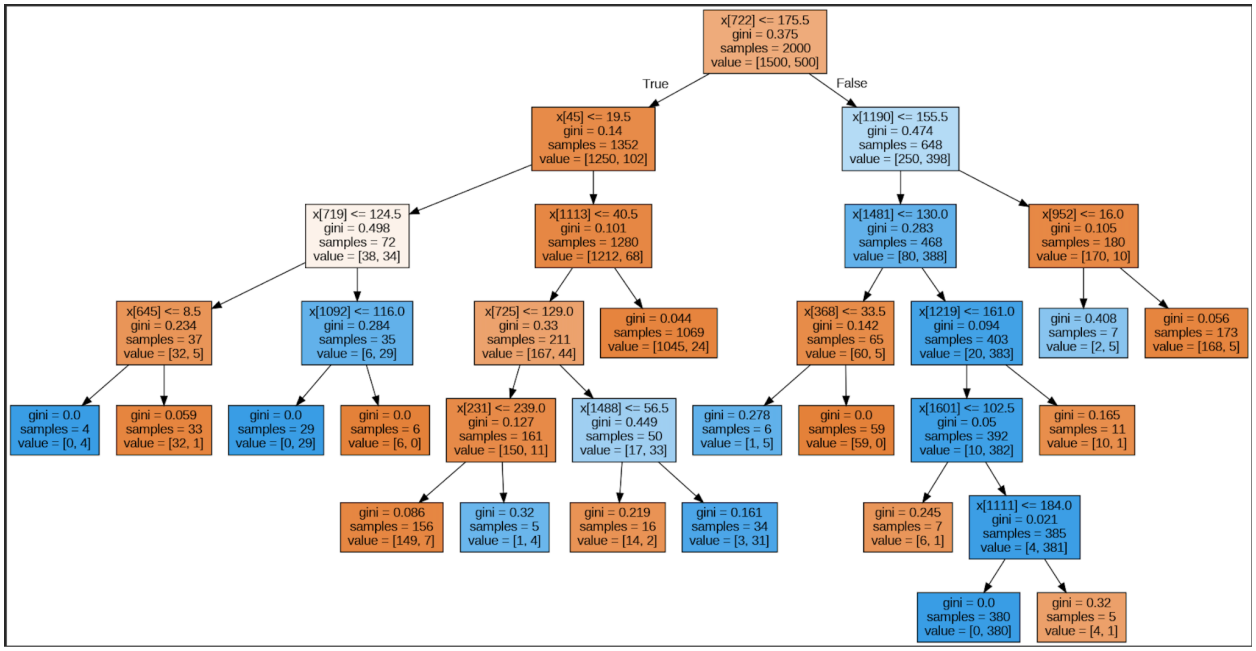
Validation Confusion Matrix:

[[294 6]

[ 20 80]]



## Best Pruned Tree:



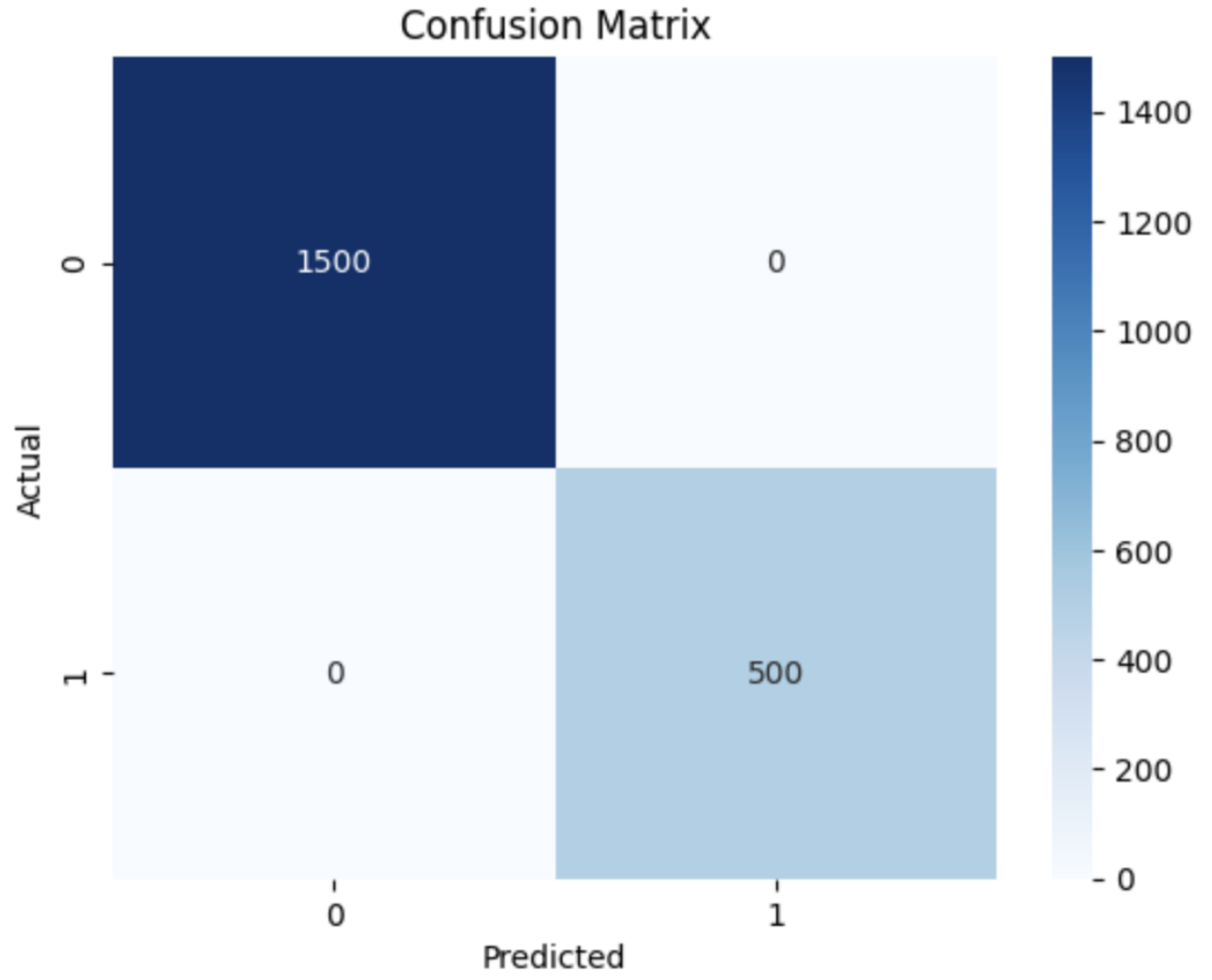

## e) Random Forests

Training Accuracy for random forest: 1.0

Training Precision for random forest: 1.0

Training Recall for random forest: 1.0

Training Confusion Matrix for random forest:

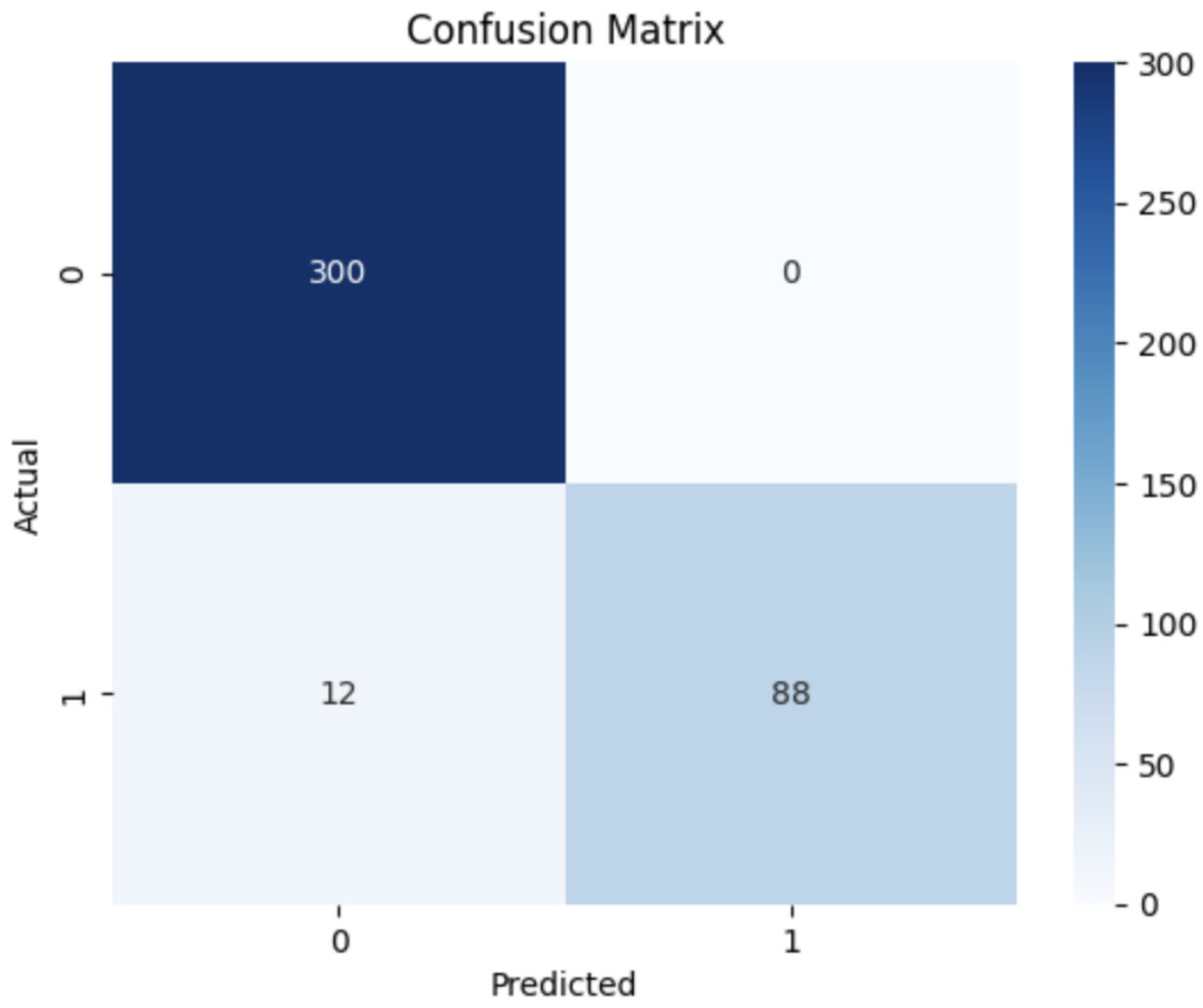



Validation Accuracy for random forest:  0.97

Validation Precision for random forest:  0.9807692307692308

Validation Recall for random forest:  0.94

Validation Confusion Matrix for random forest:


Confusion Matrix
300
0
12
88
0
1
Actual
Predicted
300
250
200
150
100
50
0

**Grid Search to find best hyperparameters:**

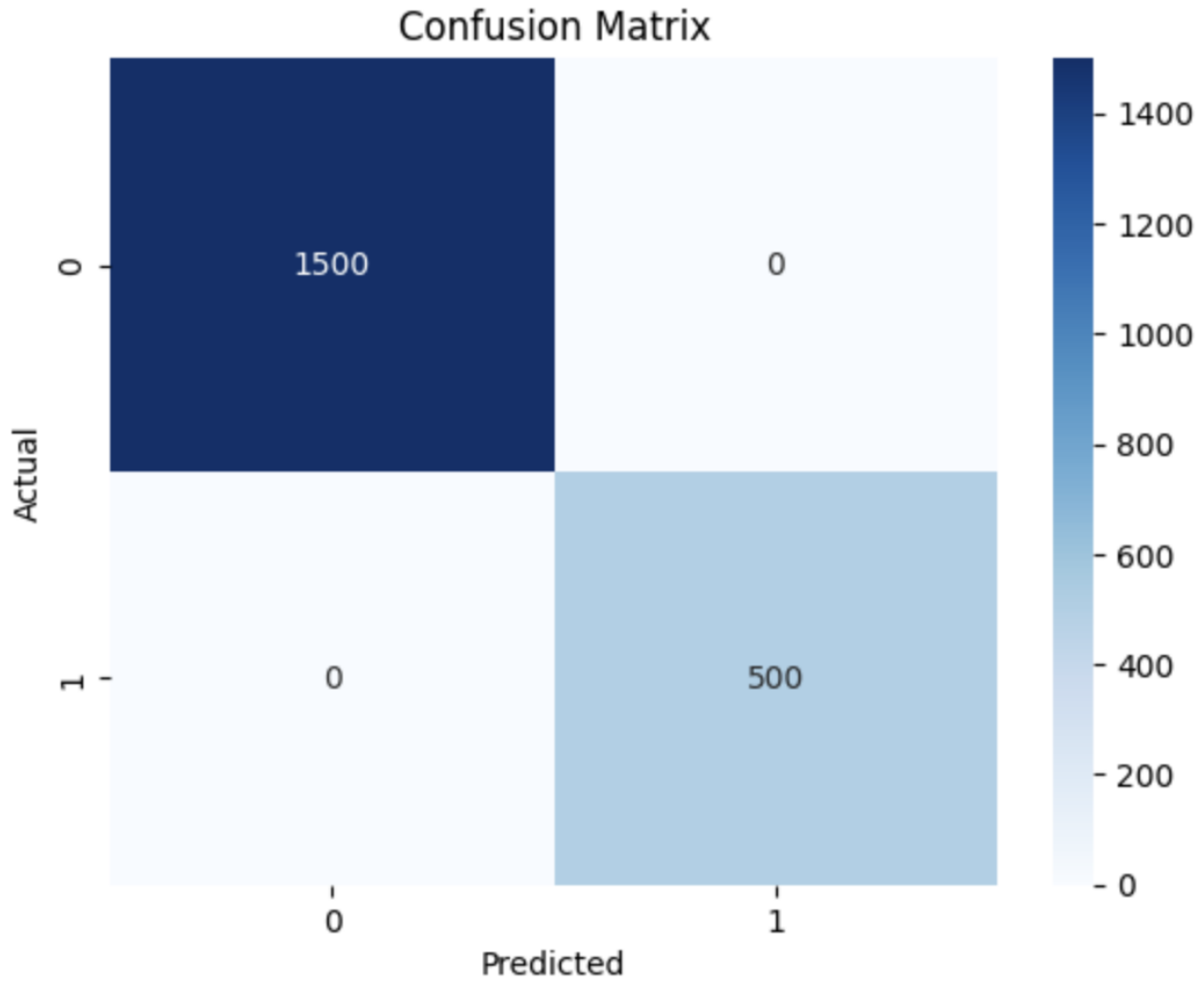
Best Hyperparameters: {'criterion': 'entropy', 'max_depth': 7, 'min_samples_split': 10, 'n_estimators': 100}

Training Accuracy: 0.982

Training Precision: 1.0

Training Recall: 1.0

Training Confusion Matrix:



Validation Accuracy: 0.9775

Validation Precision: 0.989247311827957

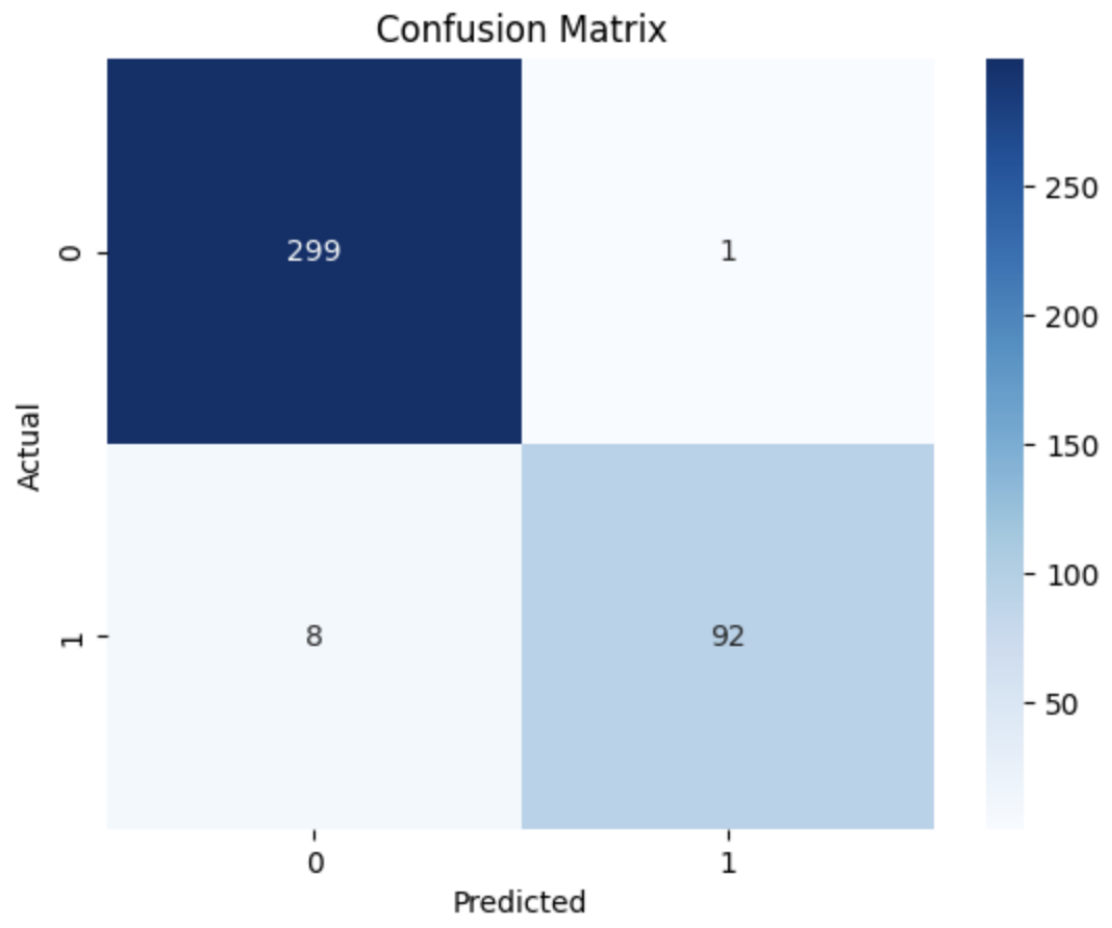
Validation Recall: 0.92

Validation Confusion Matrix:

[[299 1]

[ 8 92]]

---



**f) Gradient Boosted Trees and XGBoost**

Gradient Boosting Training Accuracy: 1.0

Gradient Boosting Training Precision: 1.0

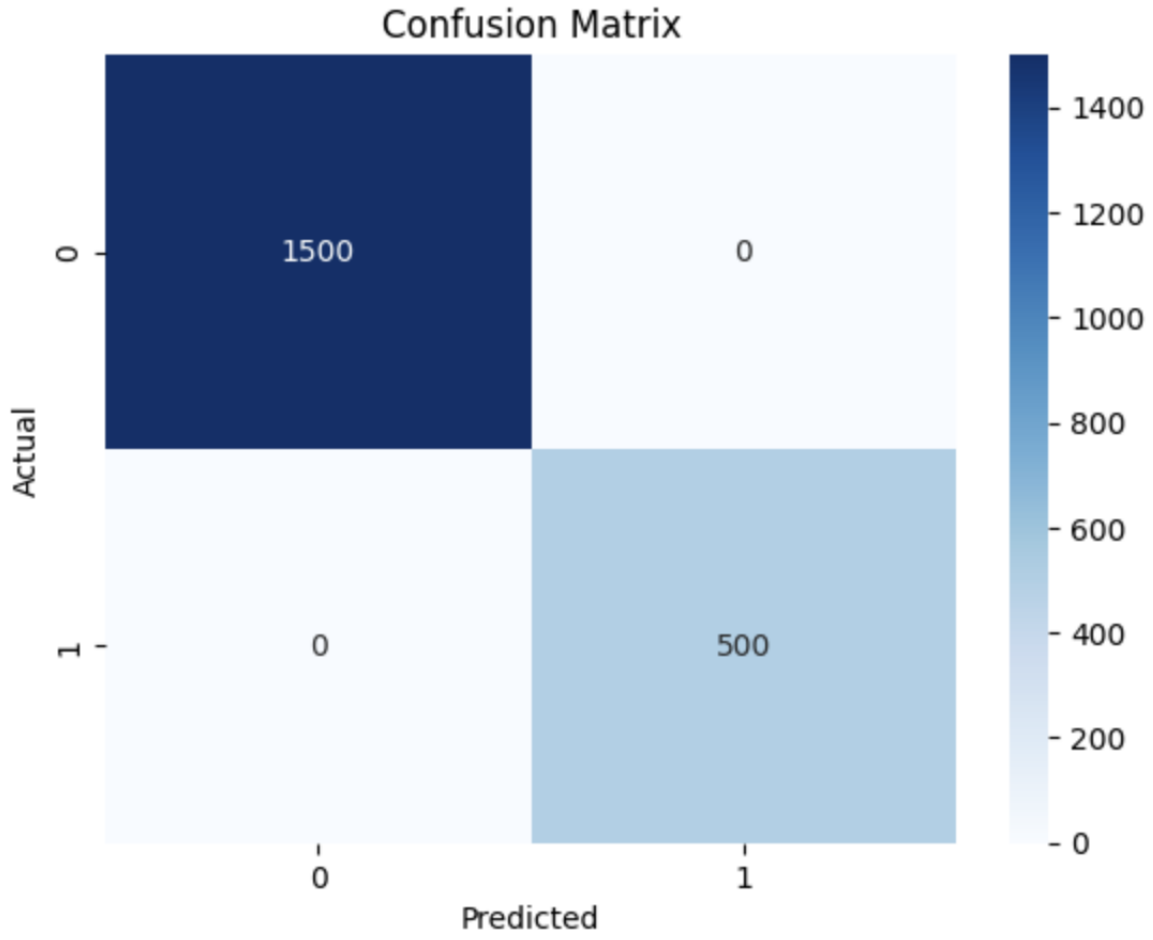
Gradient Boosting Training Recall: 1.0

Gradient Boosting Validation Accuracy: 0.98

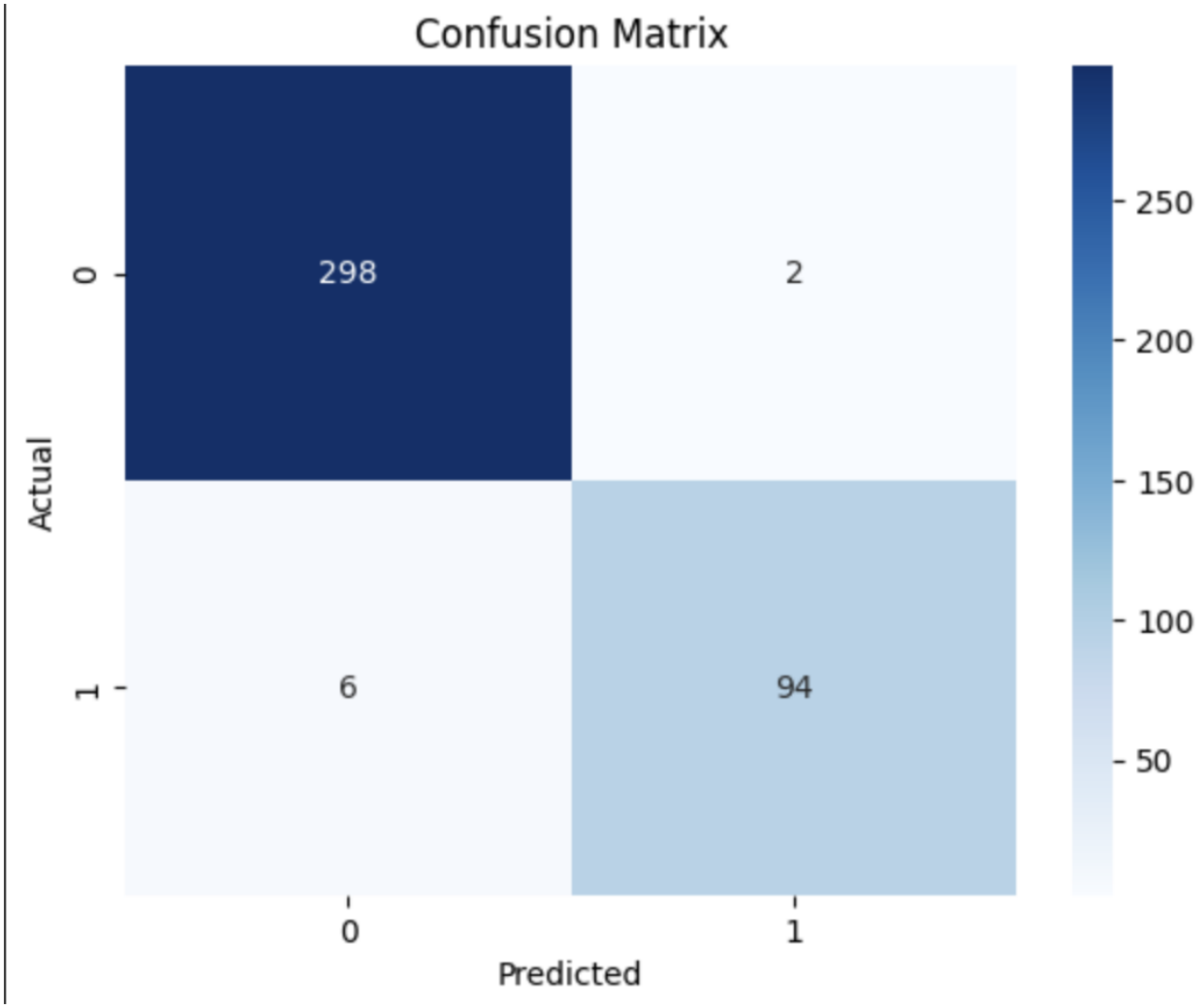
Gradient Boosting Validation Precision: 0.9791666666666666

Gradient Boosting Validation Recall: 0.94

Gradient Boosting Training Confusion Matrix:




Gradient Boosting Validation Confusion Matrix:

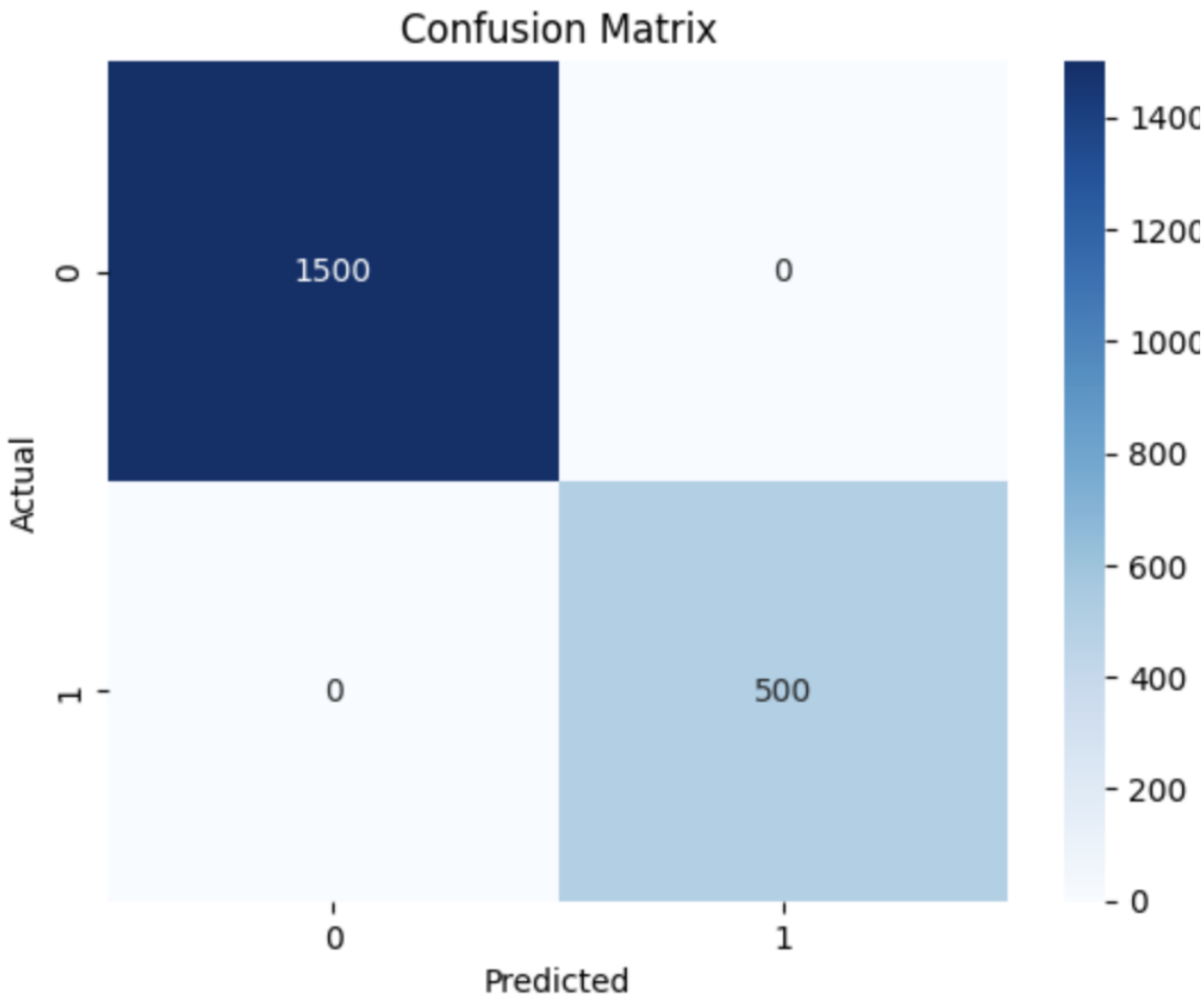



**Grid search to find best hyperparameters for Gradient Boost:**

Best Hyperparameters:  {'max_depth': 8, 'n_estimators': 50, 'subsample': 0.6}

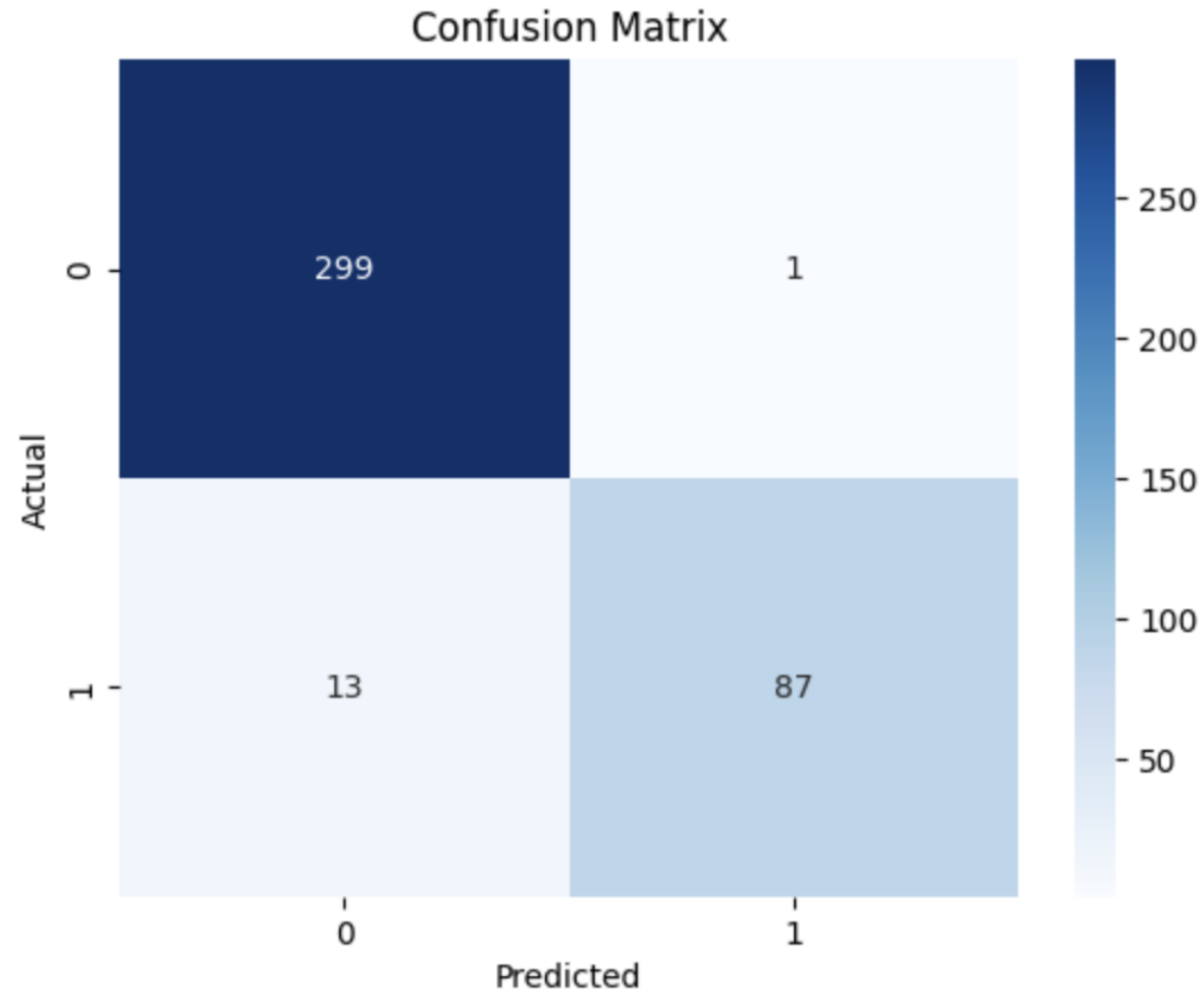
Gradient Boosting Training Accuracy: 1.0

Gradient Boosting Training Confusion Matrix:




Gradient Boosting Validation Accuracy: 0.965

Gradient Boosting Validation Confusion Matrix:



**Extreme Gradient Boosting XGBoost:**

XGBoost Training Accuracy: 1.0

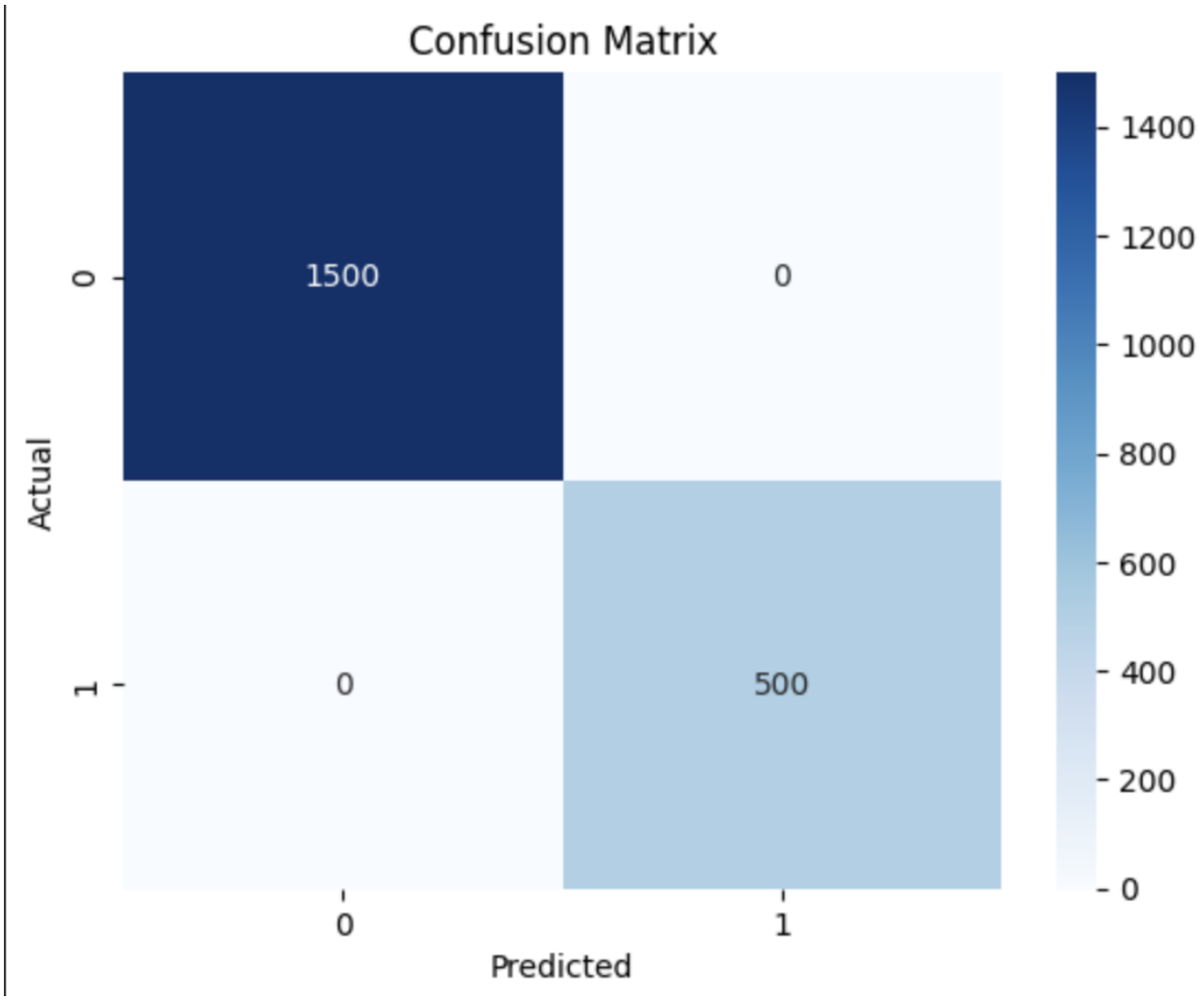
XGBoost Training Precision: 1.0

XGBoost Training Recall: 1.0

XGBoost Training Confusion Matrix:

[[1500 0]

[ 0 500]]

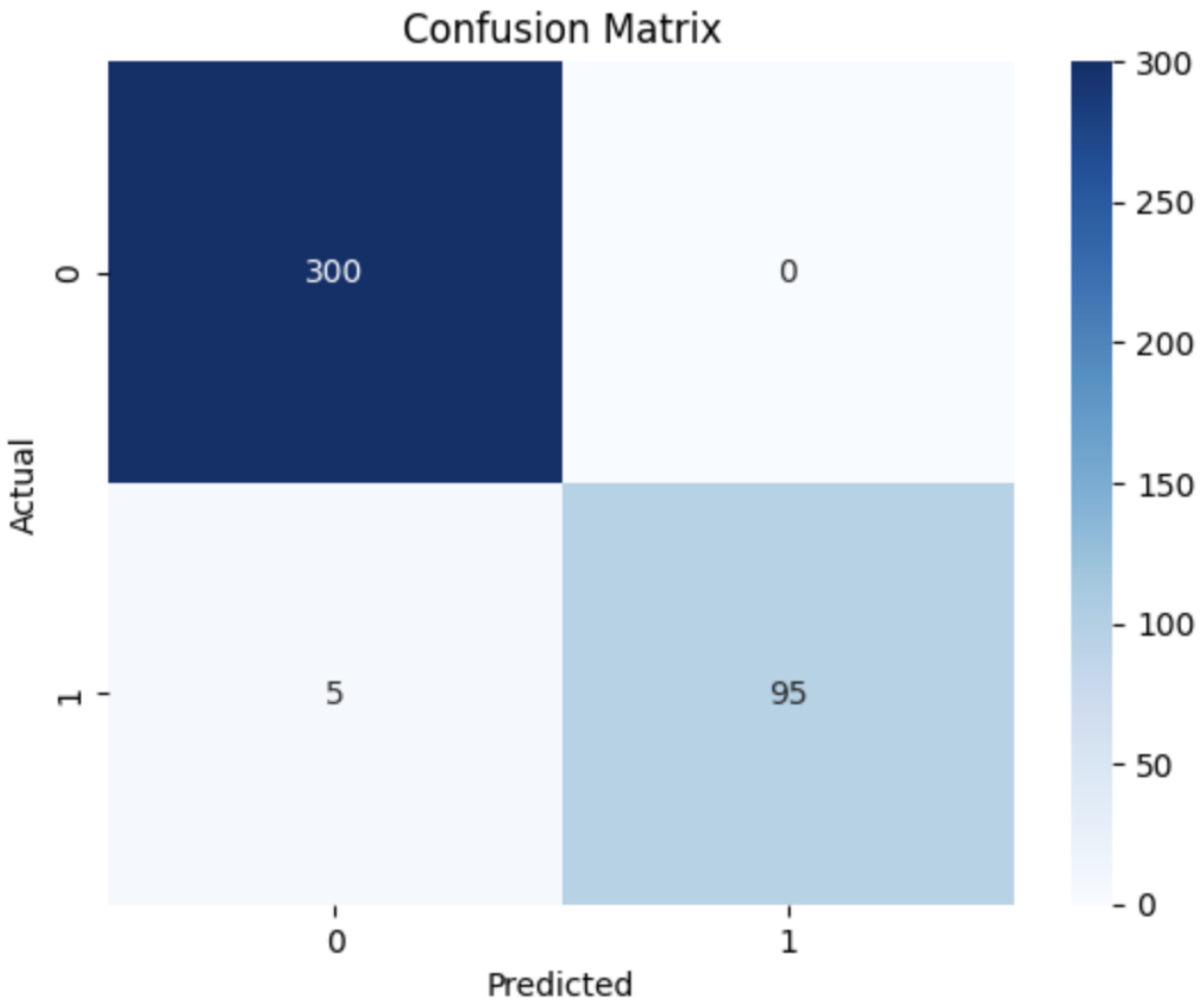



XGBoost Validation Accuracy: 0.9875

XGBoost Validation Precision: 1.0

XGBoost Validation Recall: 0.95

XGBoost Validation Confusion Matrix:




**Grid search to find best hyperparameters for XGBoost**

XGBoost Training Time:  6018.1490960121155 seconds

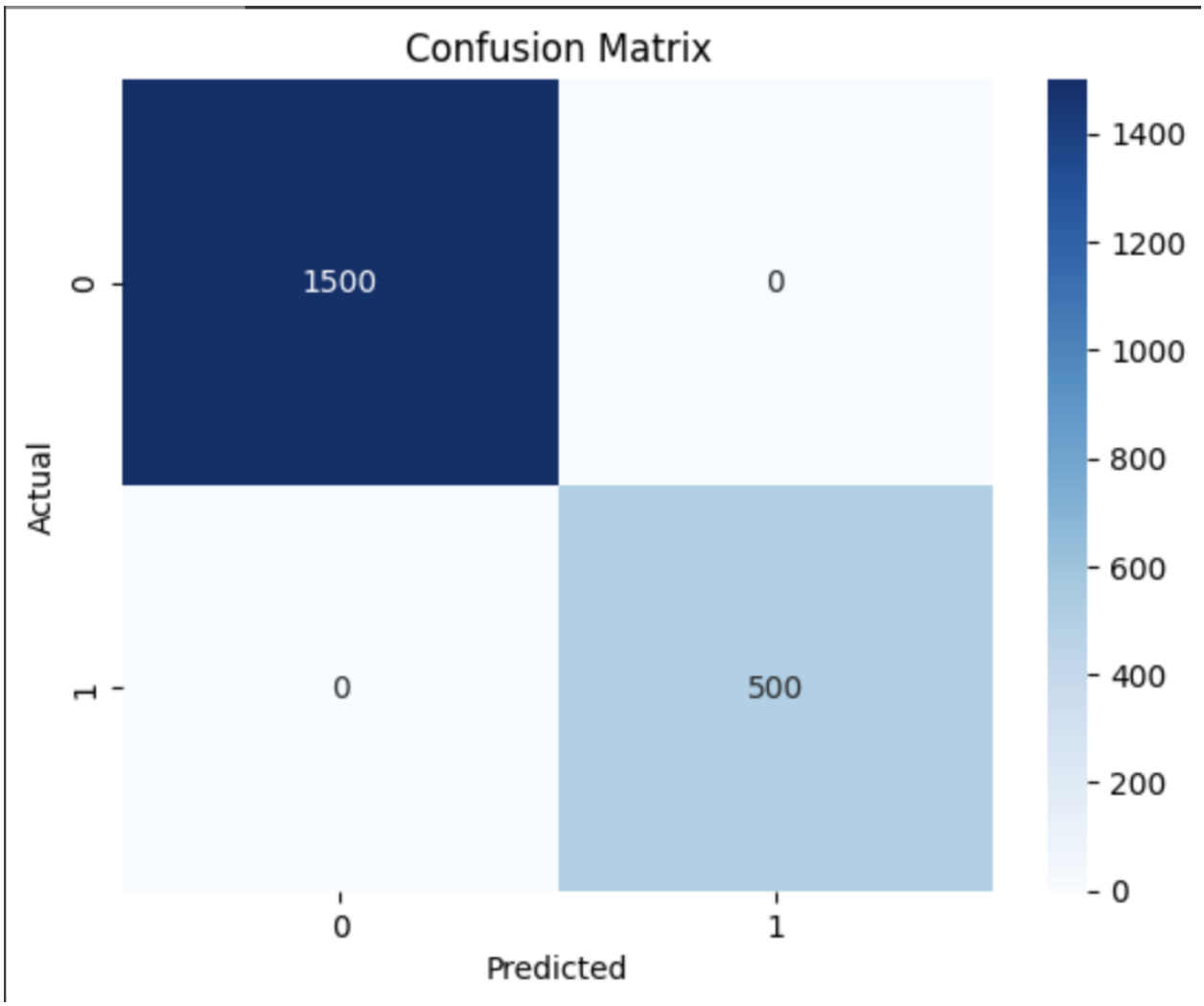
XGBoost Best Hyperparameters:  {'max_depth': 8, 'n_estimators': 50, 'subsample': 0.6}

XGBoost Training Accuracy:  0.9824999999999999

XGBoost Training Precision:  1.0

XGBoost Training Recall:  1.0

XGBoost Training Confusion Matrix:

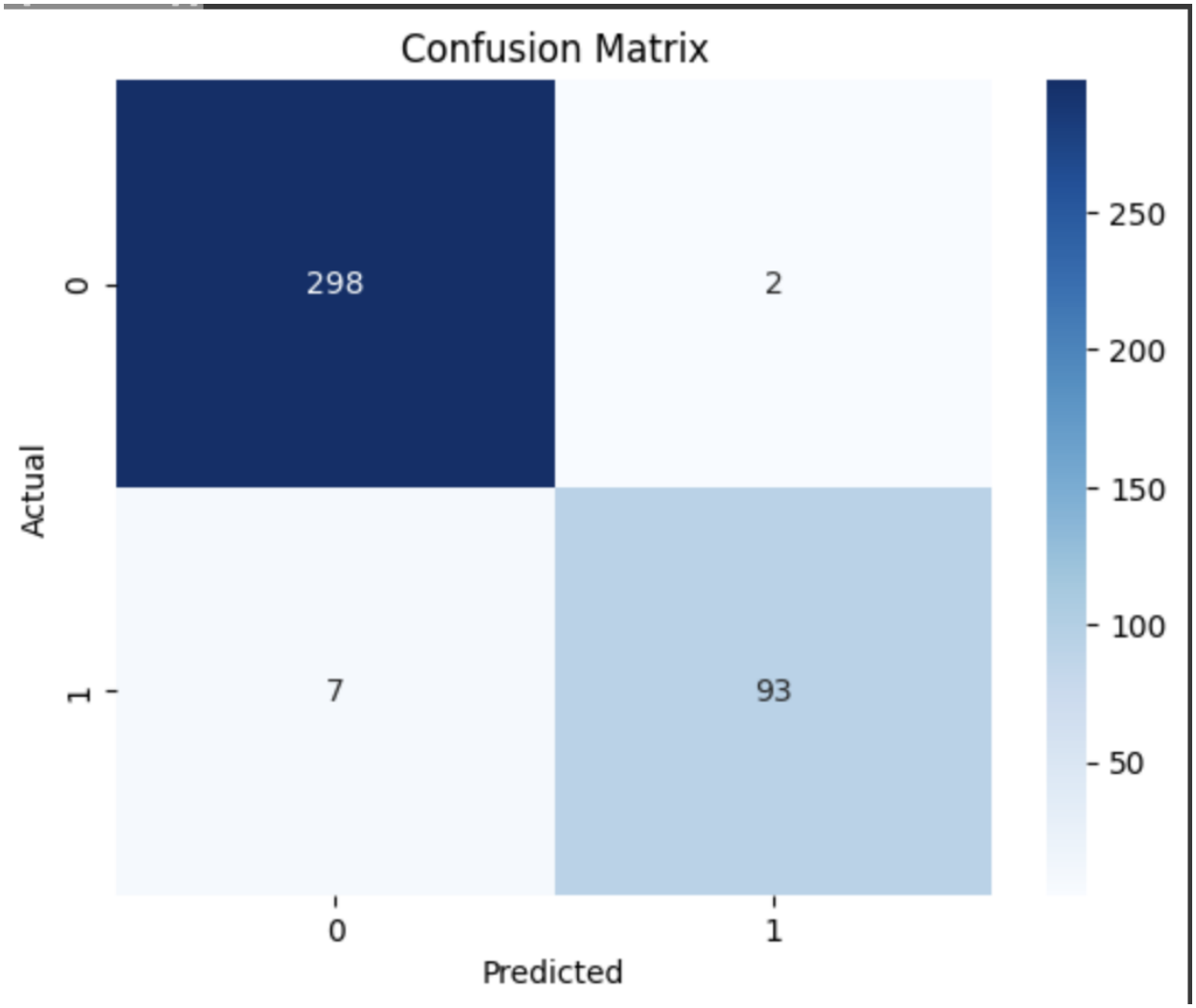


XGBoost Validation Accuracy: 0.9775

XGBoost Validation Precision: 0.9789473684210527

XGBoost Validation Recall: 0.93

XGBoost Validation Confusion Matrix:



g) Confusion matrices for all parts have been written in the respective parts.

## 3.2 Multi-Class Classification

### a) Decision Tree sklearn

Training time: 4.3754870891571045

Training Accuracy: 0.7425

Training Precision: 0.7450347943147482

Training Recall: 0.7425

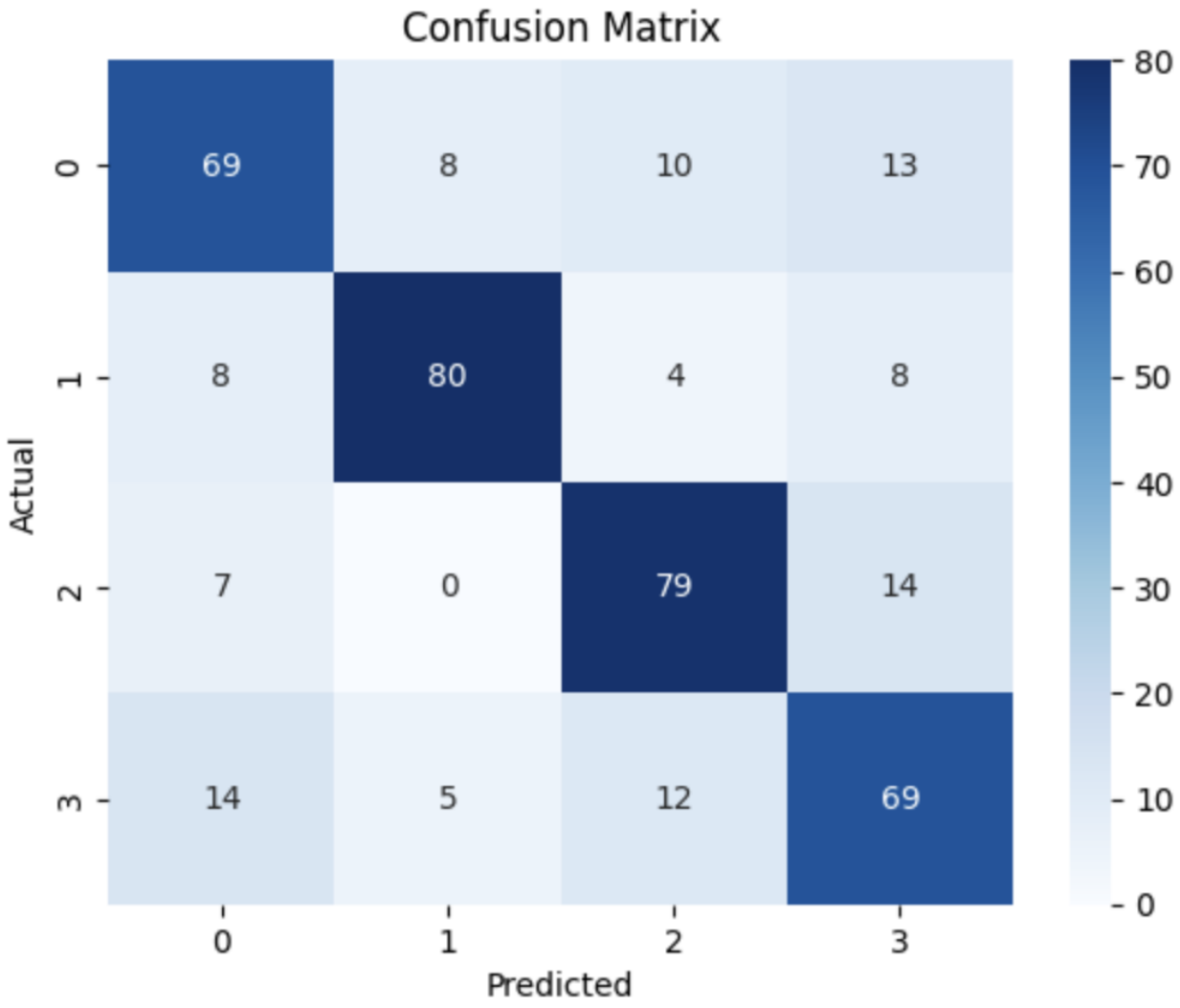
Training Confusion Matrix:

[[69 8 10 13]

[ 8 80 4 8]

[ 7 0 79 14]

[14 5 12 69]]



Validation Accuracy: 0.7425

Validation Precision: 0.7450347943147482

Validation Recall: 0.7425

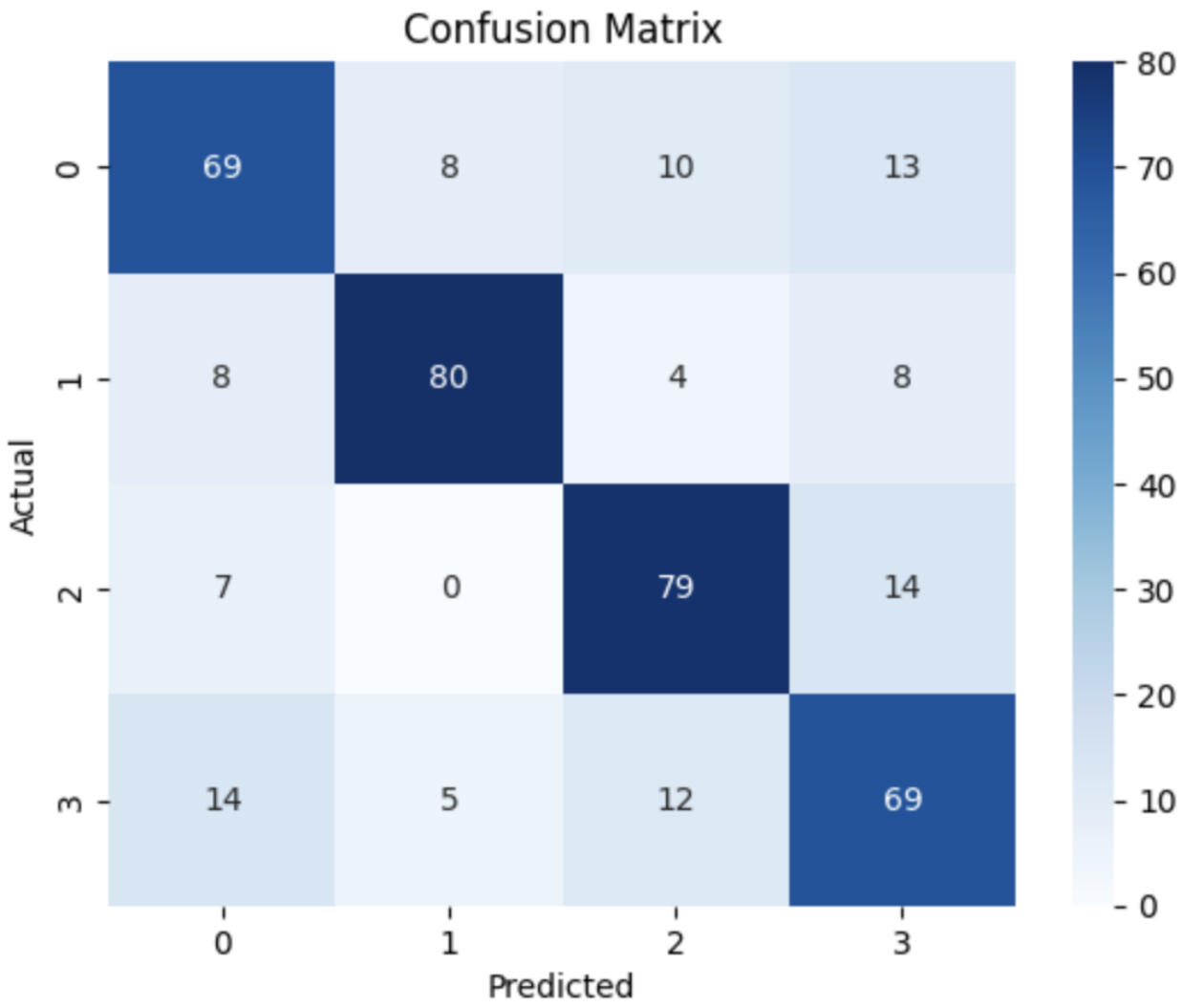
Validation Confusion Matrix:

[[69 8 10 13]

[ 8 80 4 8]

[ 7 0 79 14]

[14 5 12 69]]




**b) Decision Tree Grid Search and Visualisation**

**Selecting top 10 features:**

Feature Selection Time: 0.0641489028930664

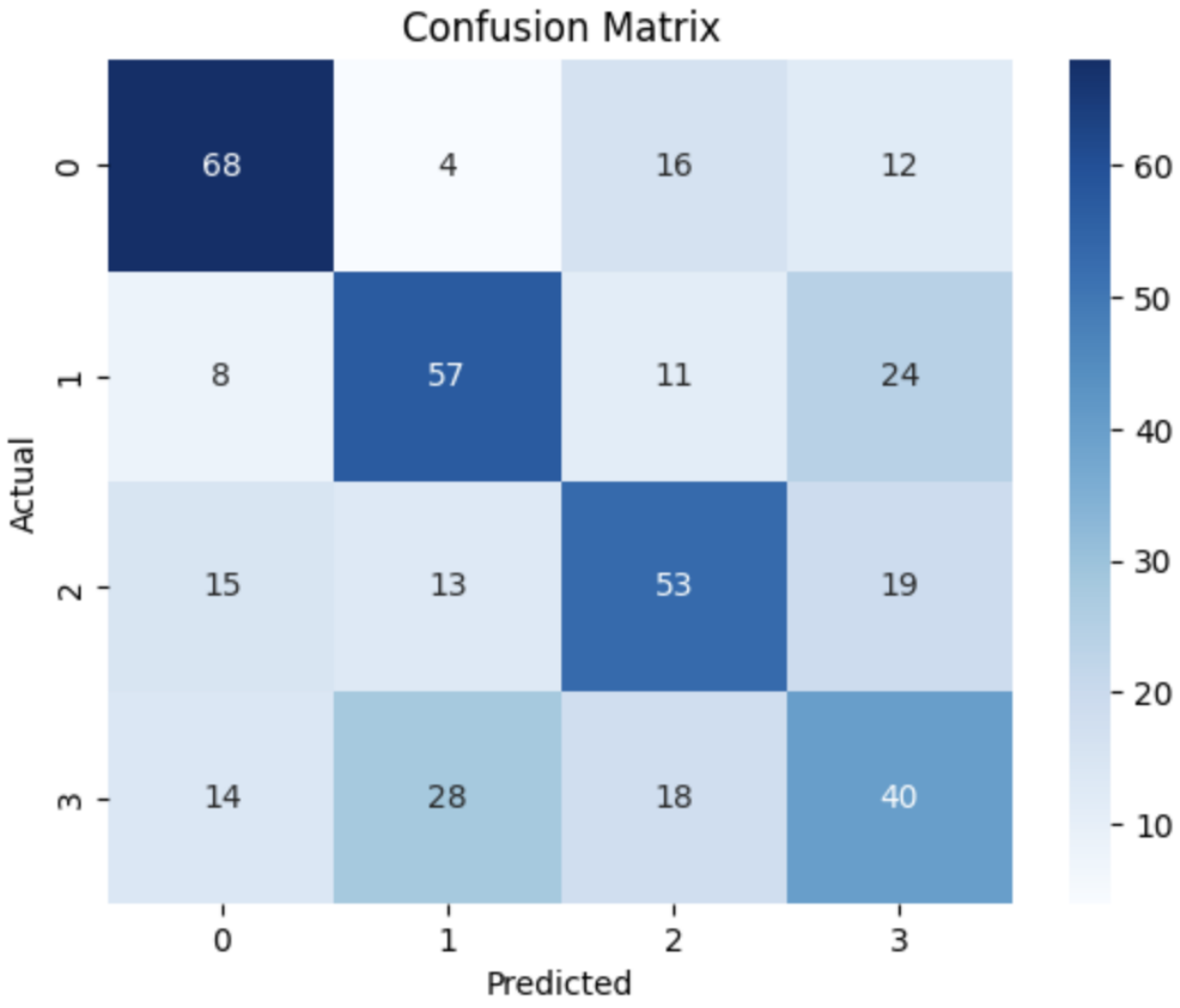
Training Time for top 10 features: 0.017633438110351562

Training Accuracy for top 10 features: 0.545

Training Precision for top 10 features: 0.542077883785093

Training Recall for top 10 features: 0.545

Training Confusion Matrix for top 10 features:

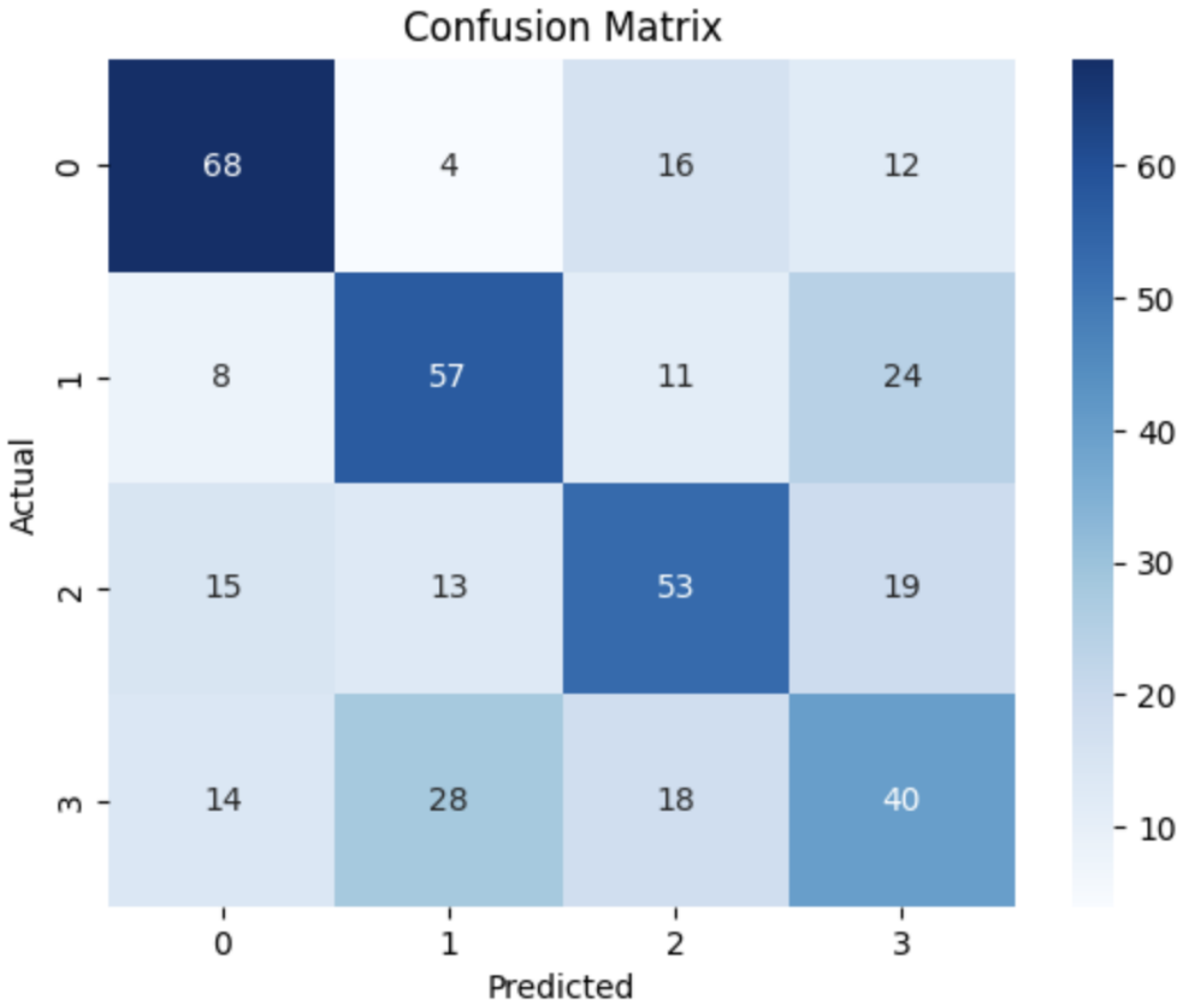


Validation Accuracy for top 10 features: 0.545

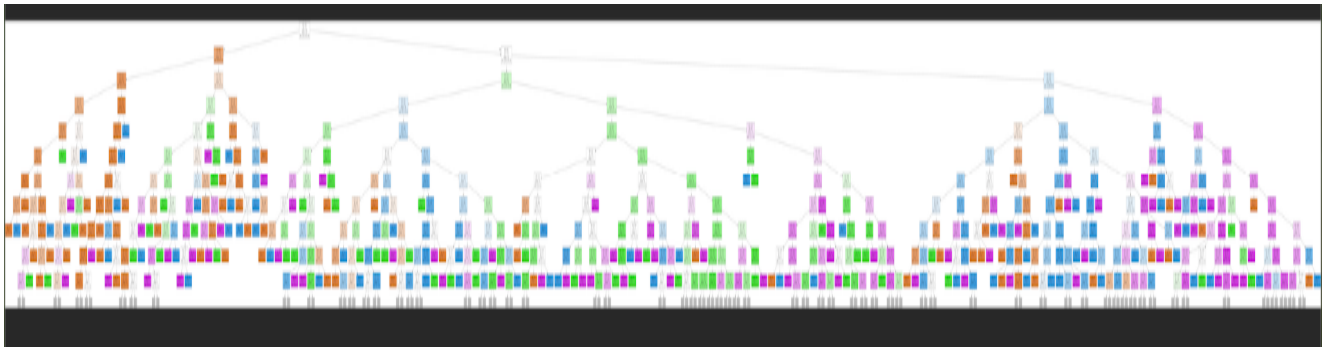
Validation Precision for top 10 features: 0.542077883785093

Validation Recall for top 10 features: 0.545

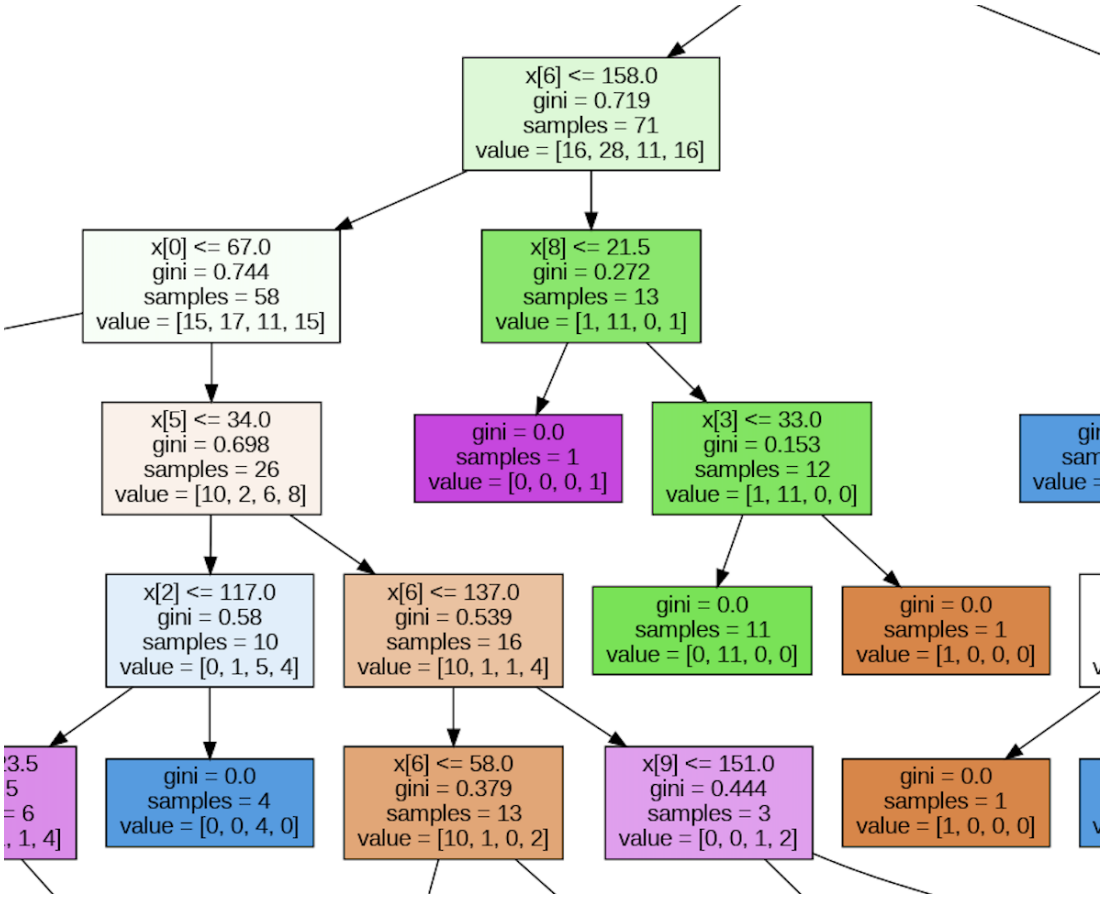
Validation Confusion Matrix for top 10 features:



## Visualising the decision tree:



Close up view of a few nodes:



**Grid Search over the top 10 features:**

Training Time for top 10 features with Grid Search: 2.7509918212890625

Best Parameters: {'criterion': 'gini', 'max_depth': 7, 'min_samples_split': 7}

Training Accuracy for top 10 features with Grid Search: 0.5325

Training Precision for top 10 features with Grid Search: 0.5286157160884155

Training Recall for top 10 features with Grid Search: 0.5325

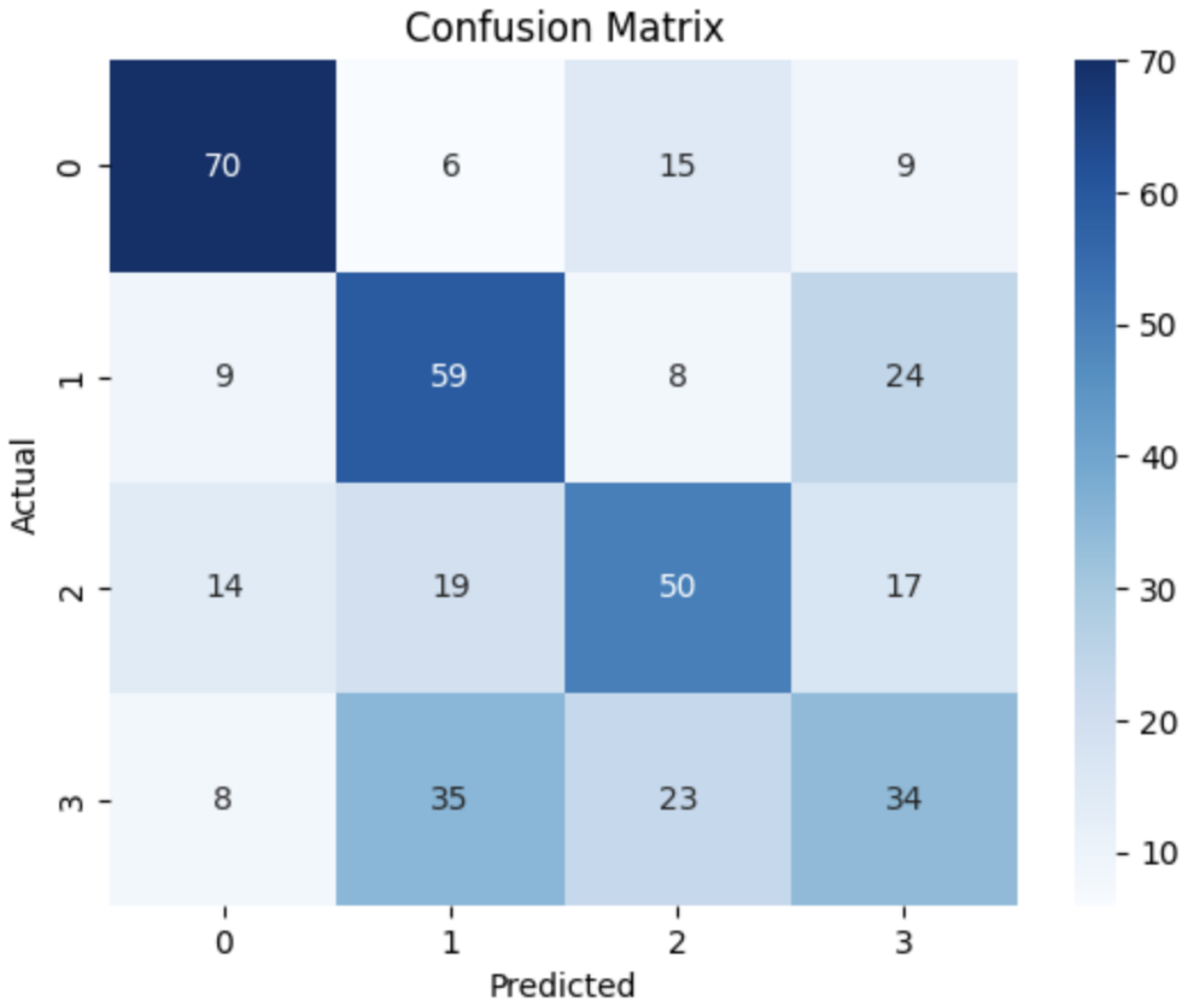
Training Confusion Matrix for top 10 features with Grid Search:

[[70 6 15 9]

[ 9 59 8 24]

[14 19 50 17]

[ 8 35 23 34]]




Validation Accuracy for top 10 features with Grid Search: 0.5325

Validation Precision for top 10 features with Grid Search: 0.5286157160884155

Validation Recall for top 10 features with Grid Search: 0.5325

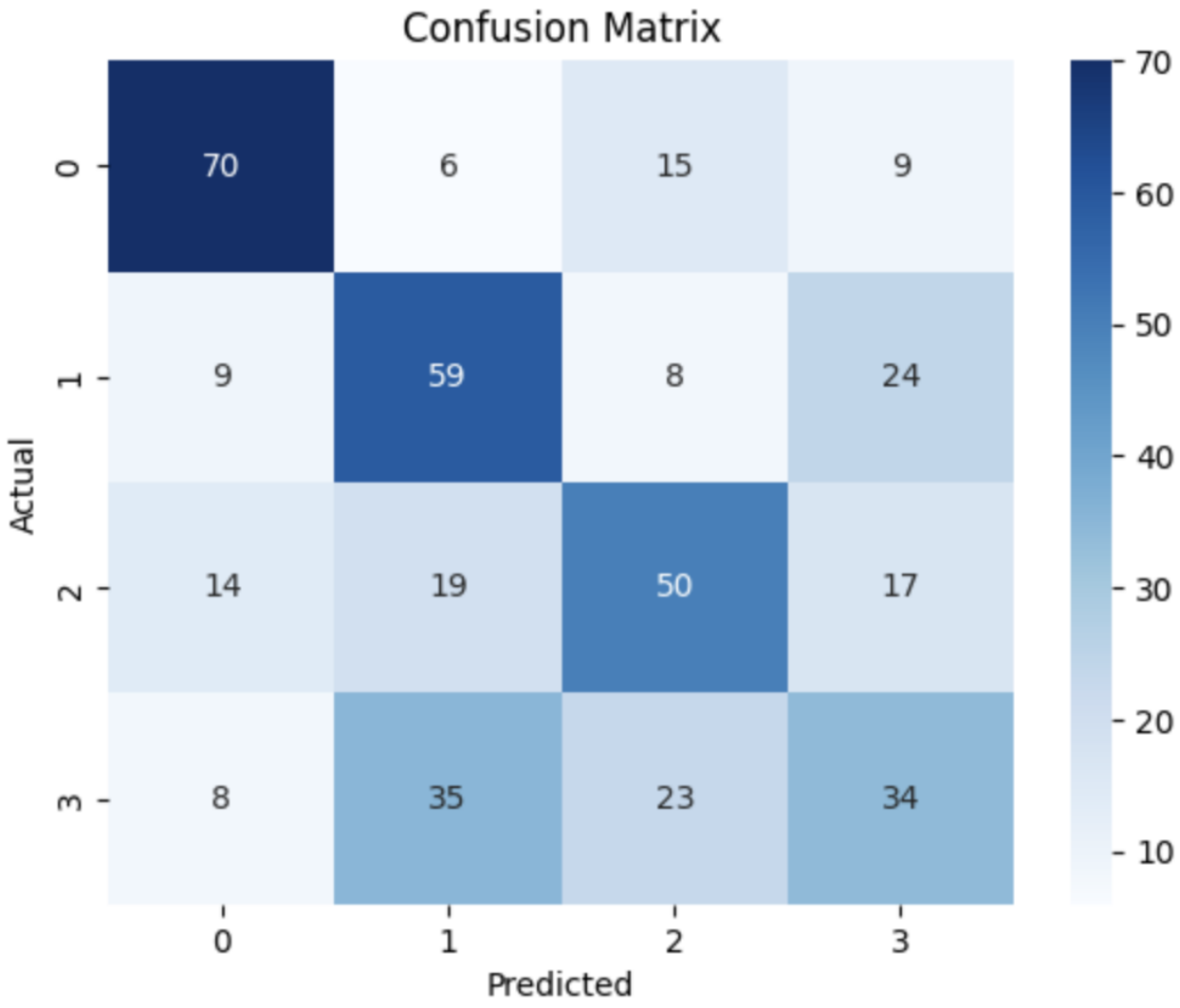
Validation Confusion Matrix for top 10 features with Grid Search:

[[70 6 15 9]

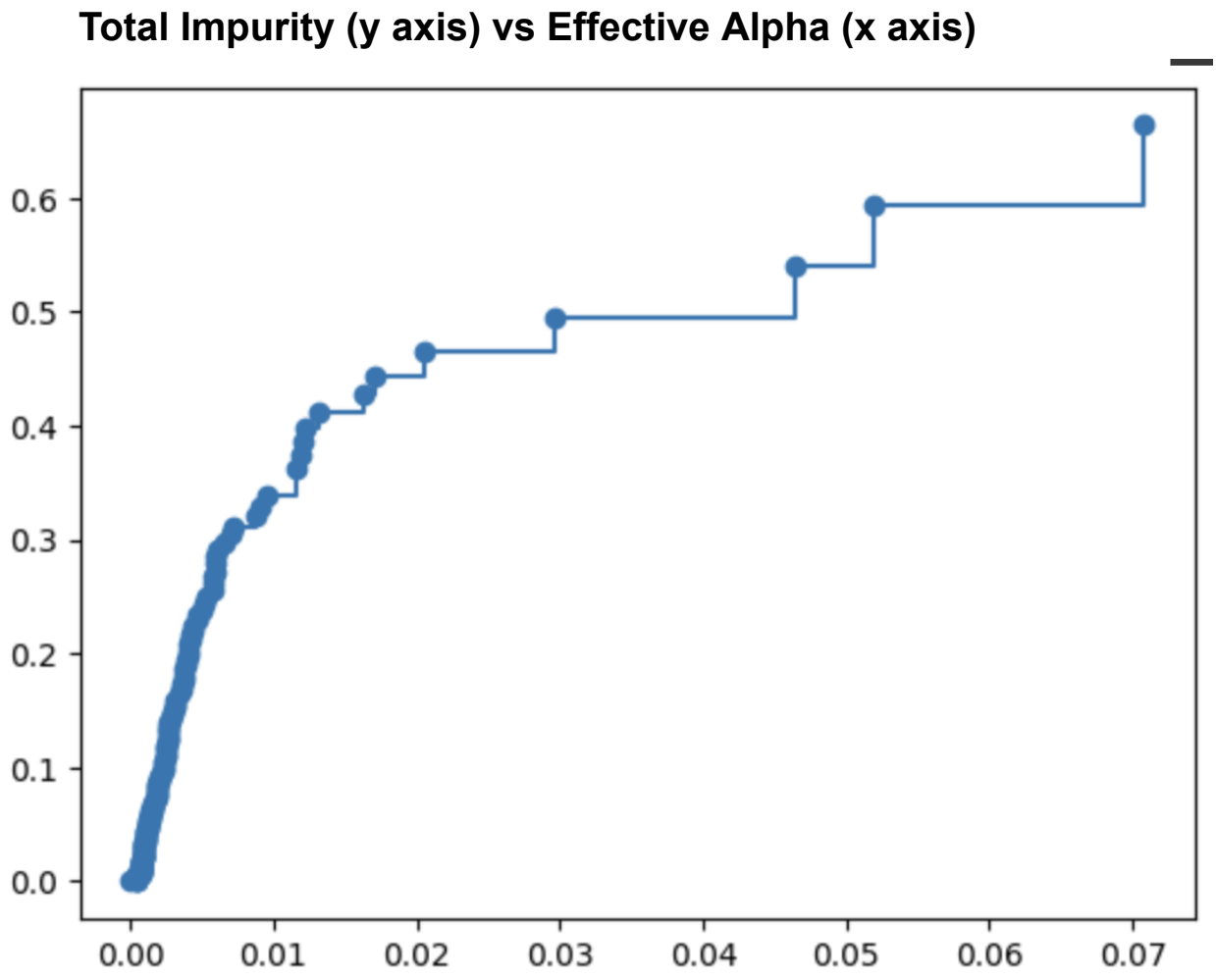
[ 9 59 8 24]

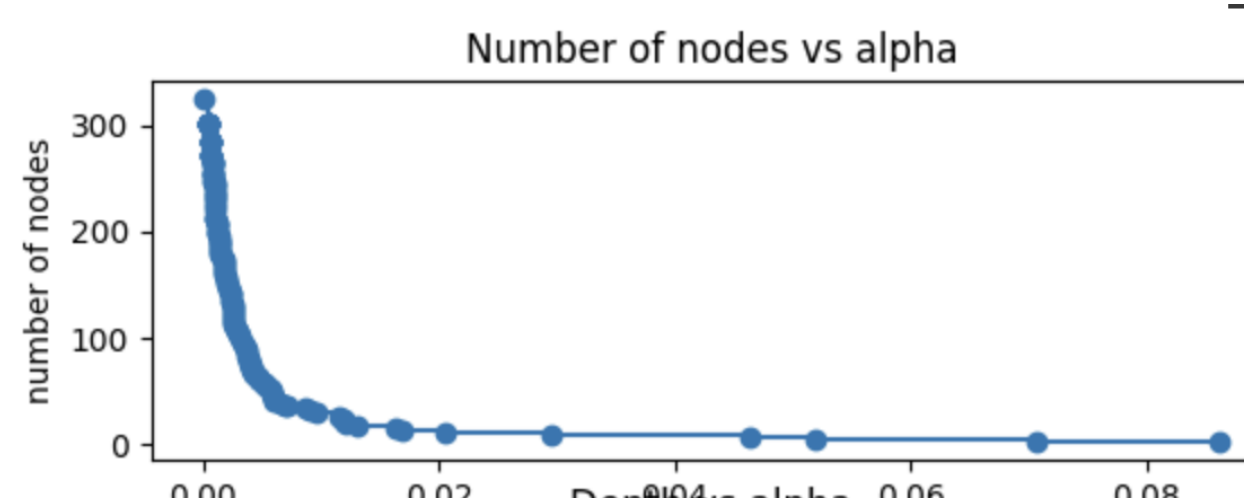
[14 19 50 17]

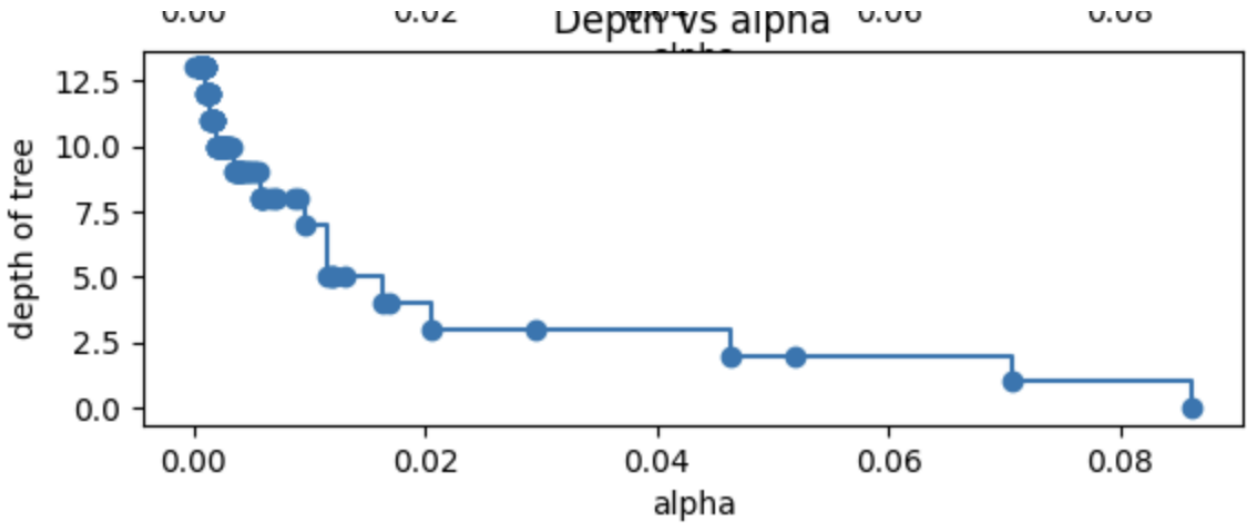
[ 8 35 23 34]]

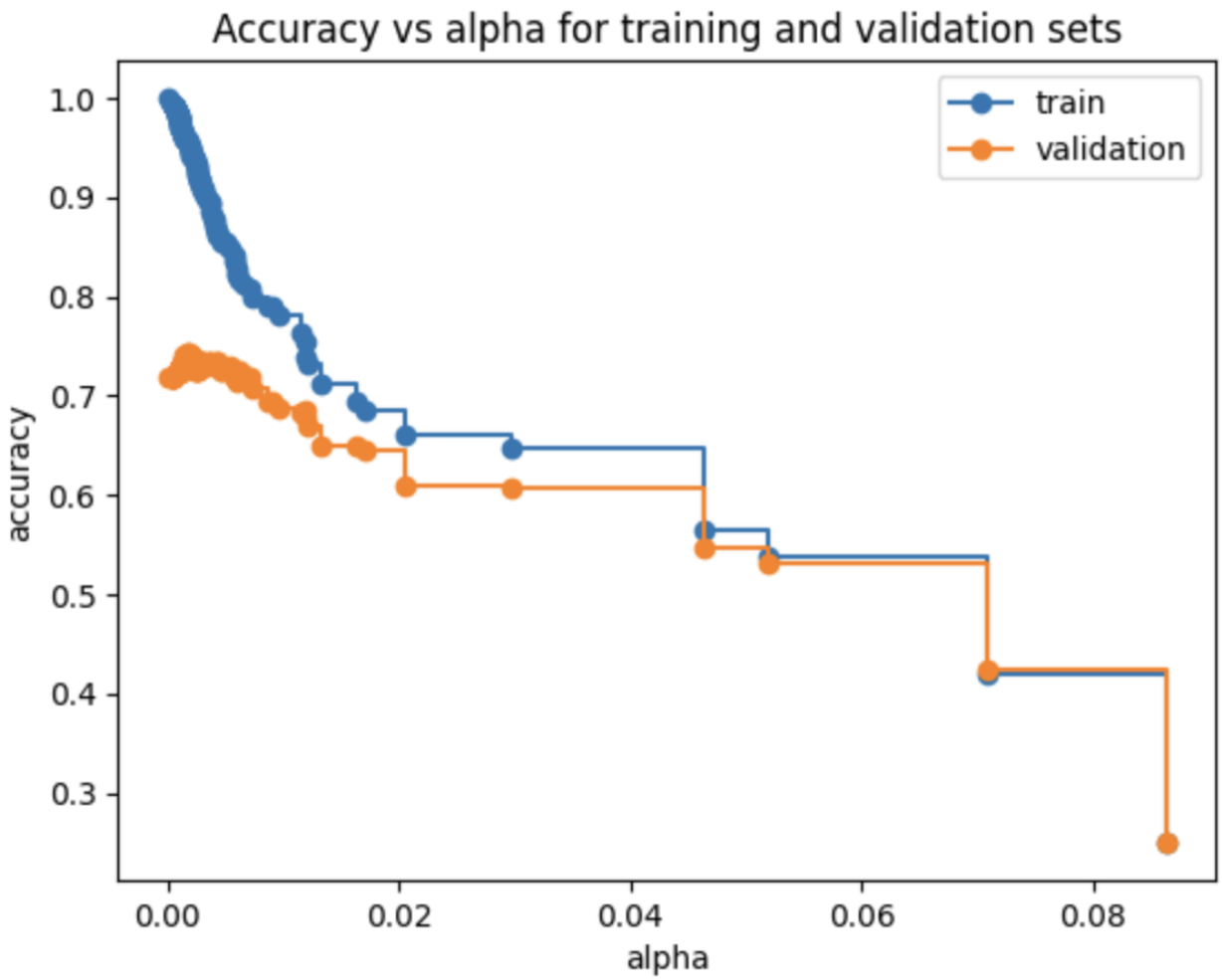


## c) Decision Tree Post Pruning with Cost Complexity Pruning




Number of nodes vs alpha
number of nodes
300
200
100
0
0.00
0.02
0.06
0.08


Depth vs alpha
depth of tree
12.5
10.0
7.5
5.0
2.5
0.0
0.00
0.02
0.04
0.06
0.08
alpha


Accuracy vs alpha for training and validation sets
train
validation
accuracy
alpha
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.00
0.02
0.04
0.06
0.08

Best alpha: 0.0016969696969696972

Best performing tree statistics:

Training Accuracy for best performing tree: 0.957

Training Precision for best performing tree: 0.9572802620240429

Training Recall for best performing tree: 0.9570000000000001

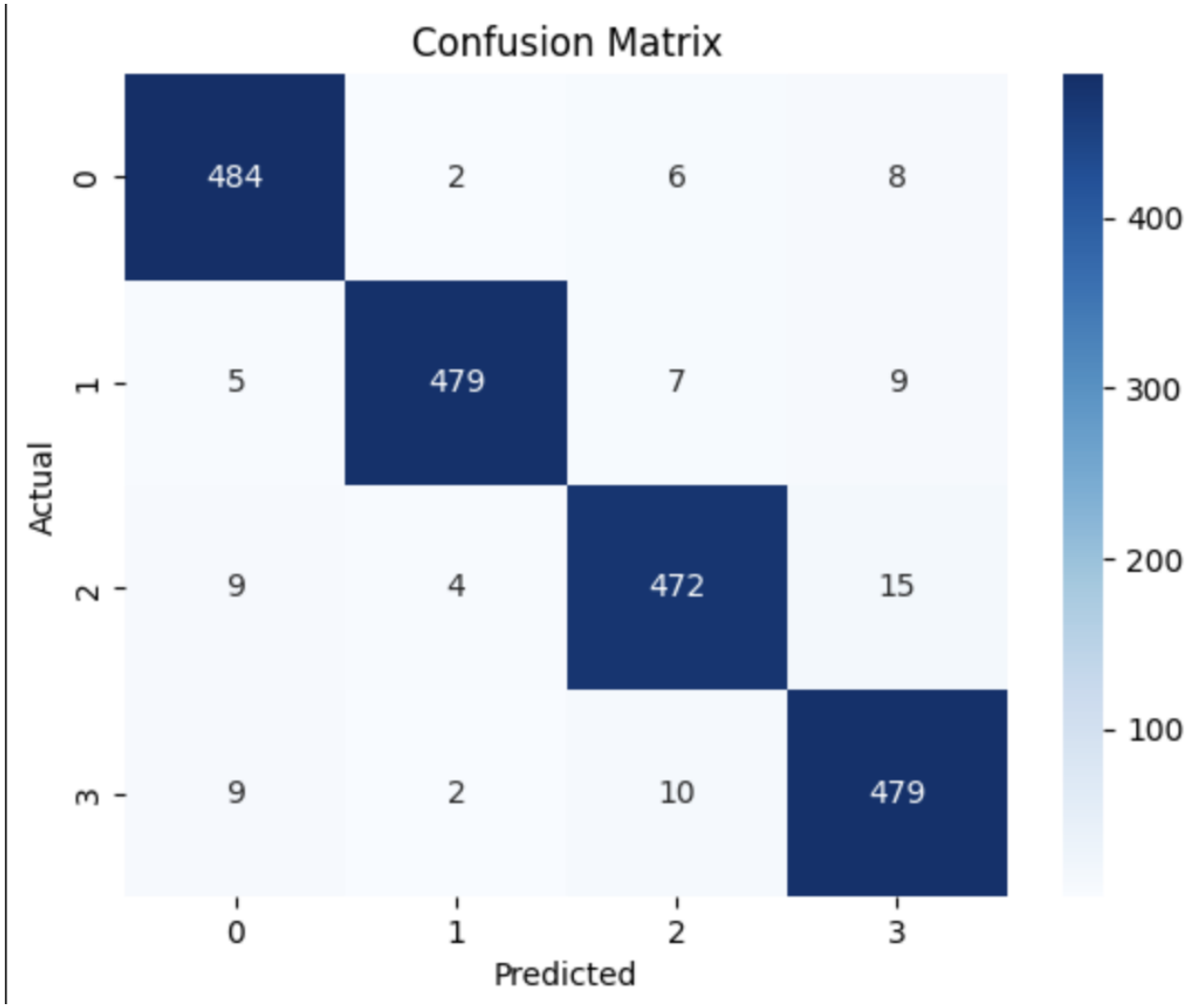
Training Confusion Matrix for best performing tree:

[[484 2 6 8]

[ 5 479 7 9]

[ 9 4 472 15]

[ 9 2 10 479]]




Validation Accuracy for best performing tree: 0.7425

Validation Precision for best performing tree: 0.7479564886279222

Validation Recall for best performing tree: 0.7425

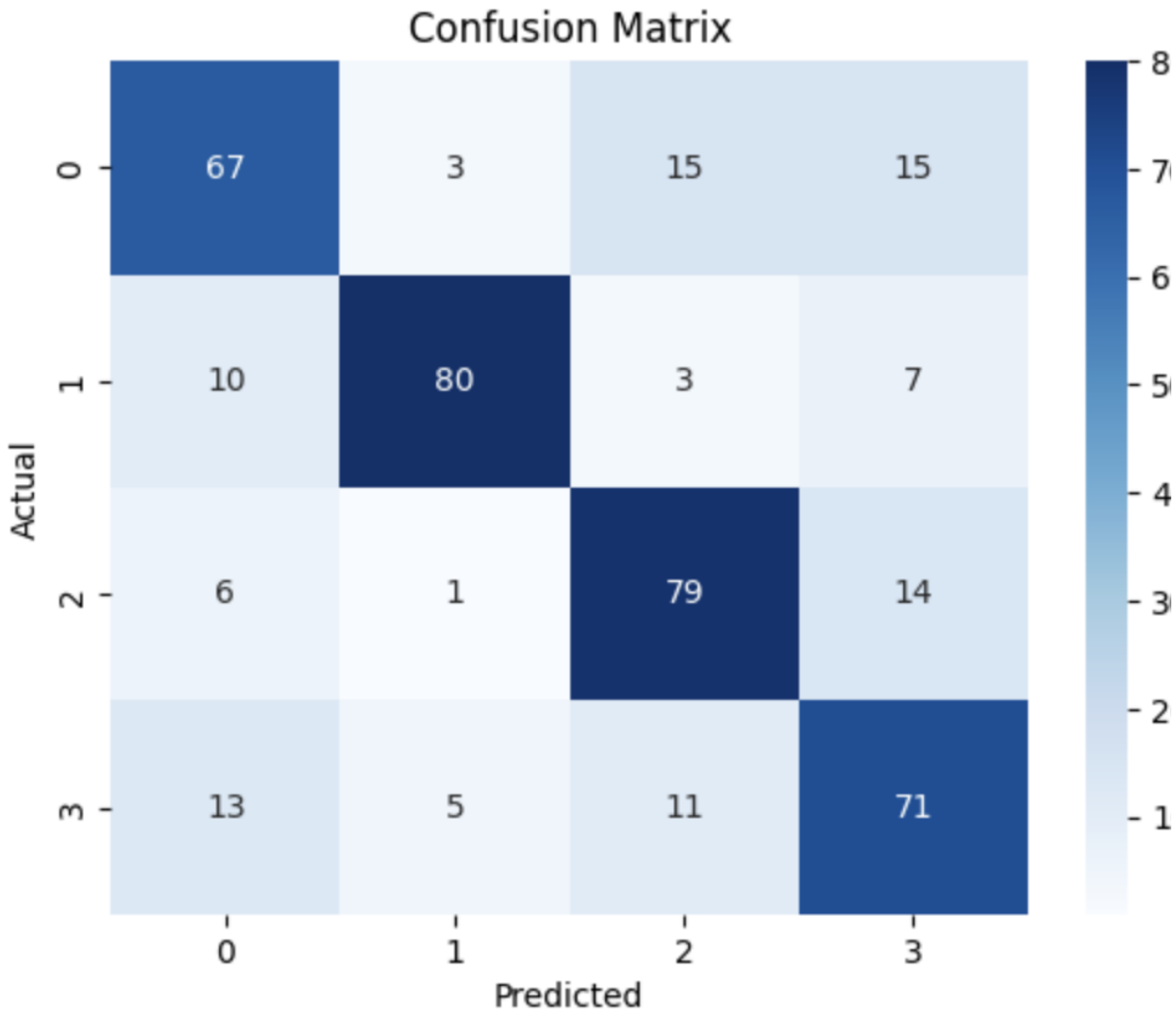
Validation Confusion Matrix for best performing tree:

[[67 3 15 15]

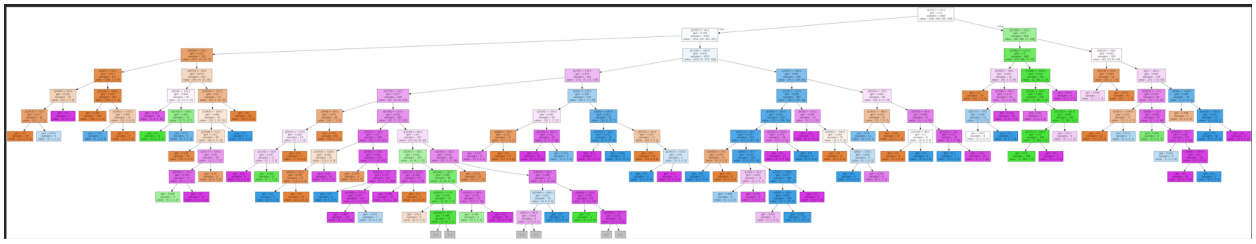
[10 80 3 7]

[ 6 1 79 14]

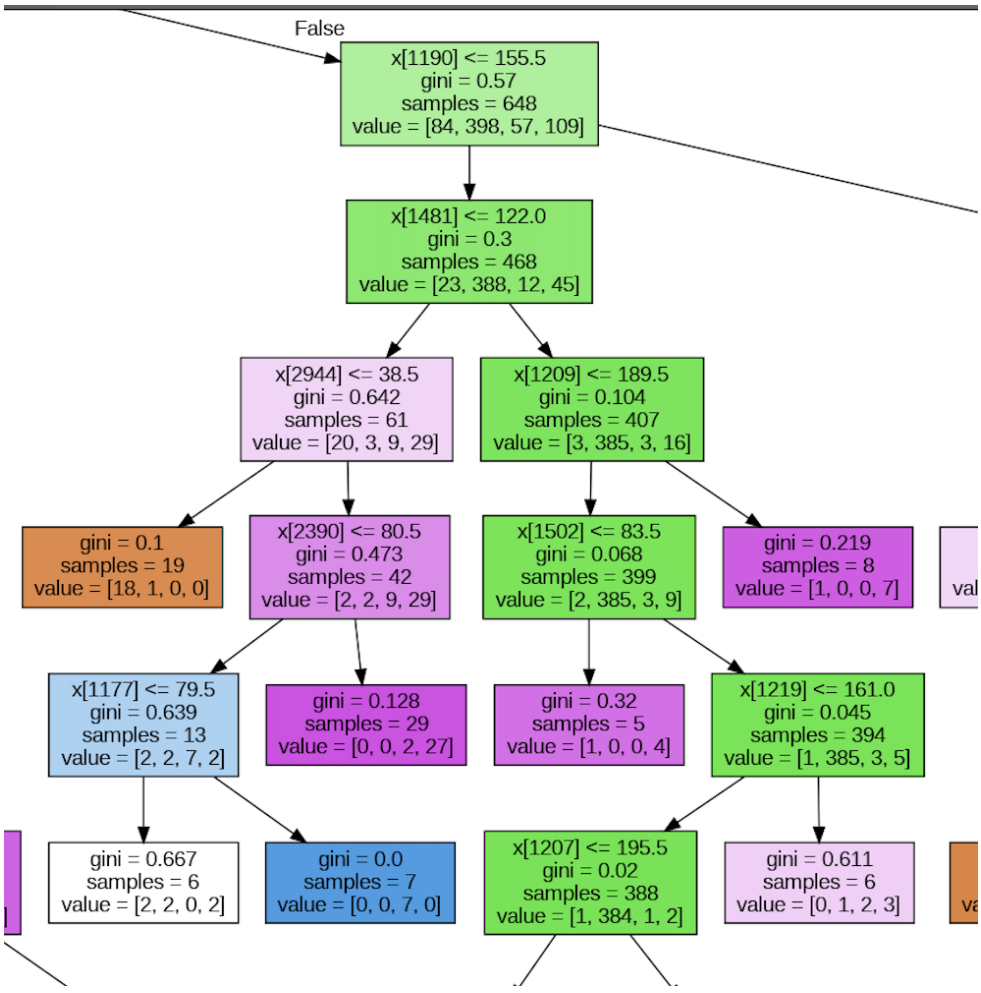
[13 5 11 71]]



## Best Pruned Tree:



This is a close up view of the tree (such that individual nodes are visible):



## d) Random Forest

### Trained on default parameters

Training Accuracy for random forest: 1.0

Training Precision for random forest: 1.0

Training Recall for random forest: 1.0

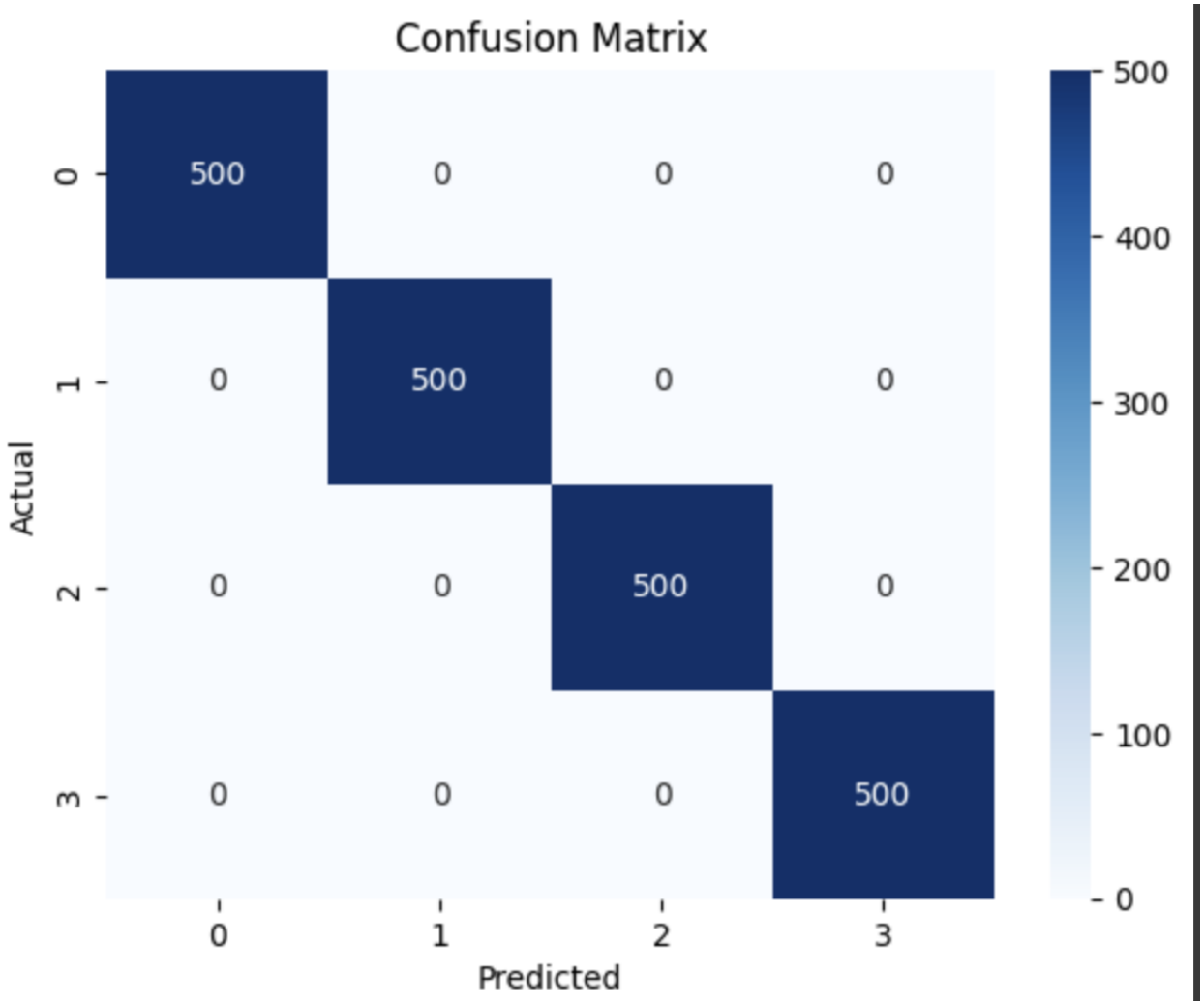
Training Confusion Matrix for random forest:

[[500 0 0 0]

[ 0 500 0 0]

[ 0 0 500 0]

[ 0 0 0 500]]




Validation Accuracy for random forest: 0.8725

Validation Precision for random forest: 0.8792709705753184

Validation Recall for random forest: 0.8724999999999999

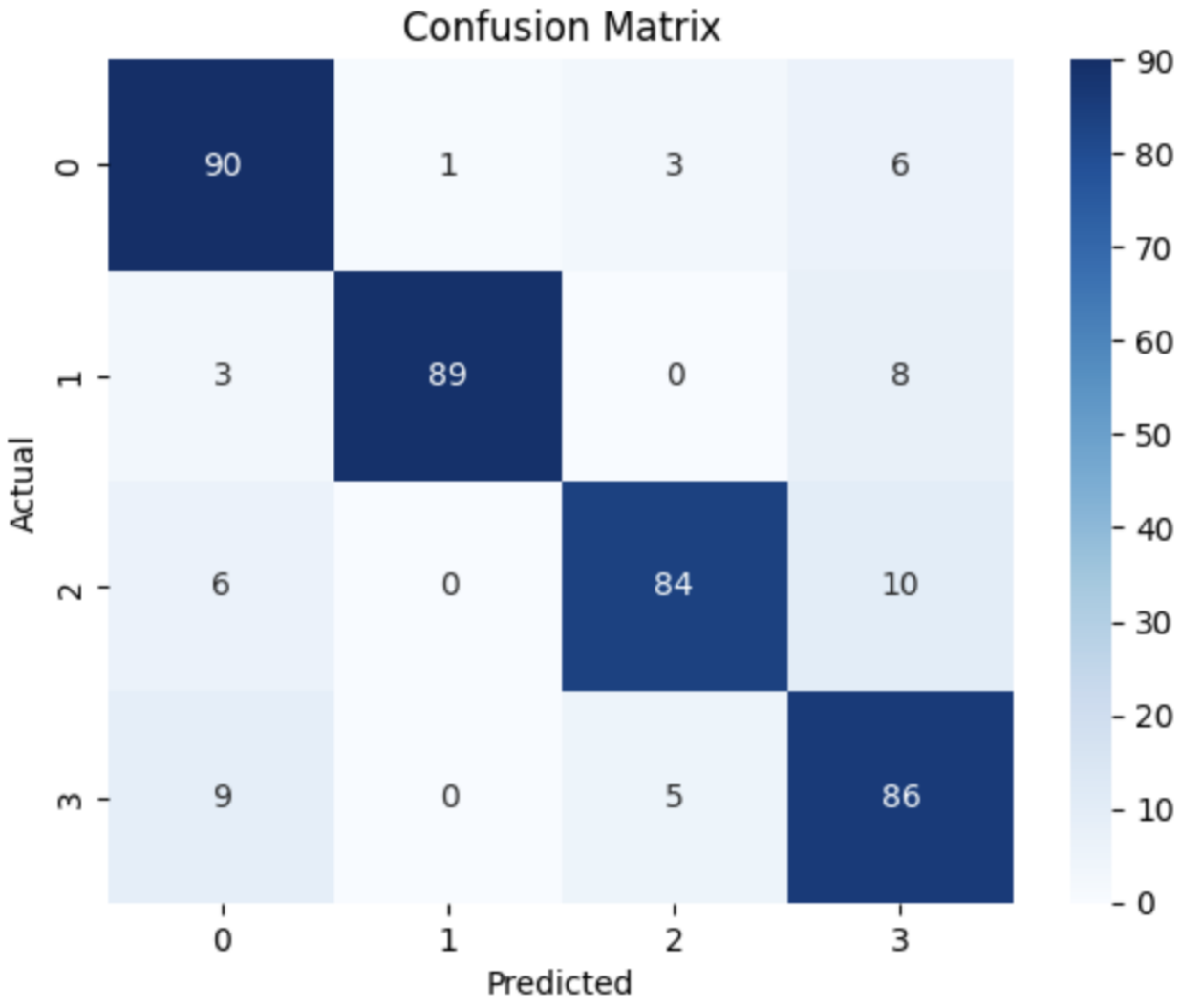
Validation Confusion Matrix for random forest:

[[90 1 3 6]

[ 3 89 0 8]

[ 6 0 84 10]

[ 9 0 5 86]]



Best Parameters: {'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 200}

Training Accuracy for random forest with best hyperparameters: 1.0

Training Precision for random forest with best hyperparameters: 1.0

Training Recall for random forest with best hyperparameters: 1.0

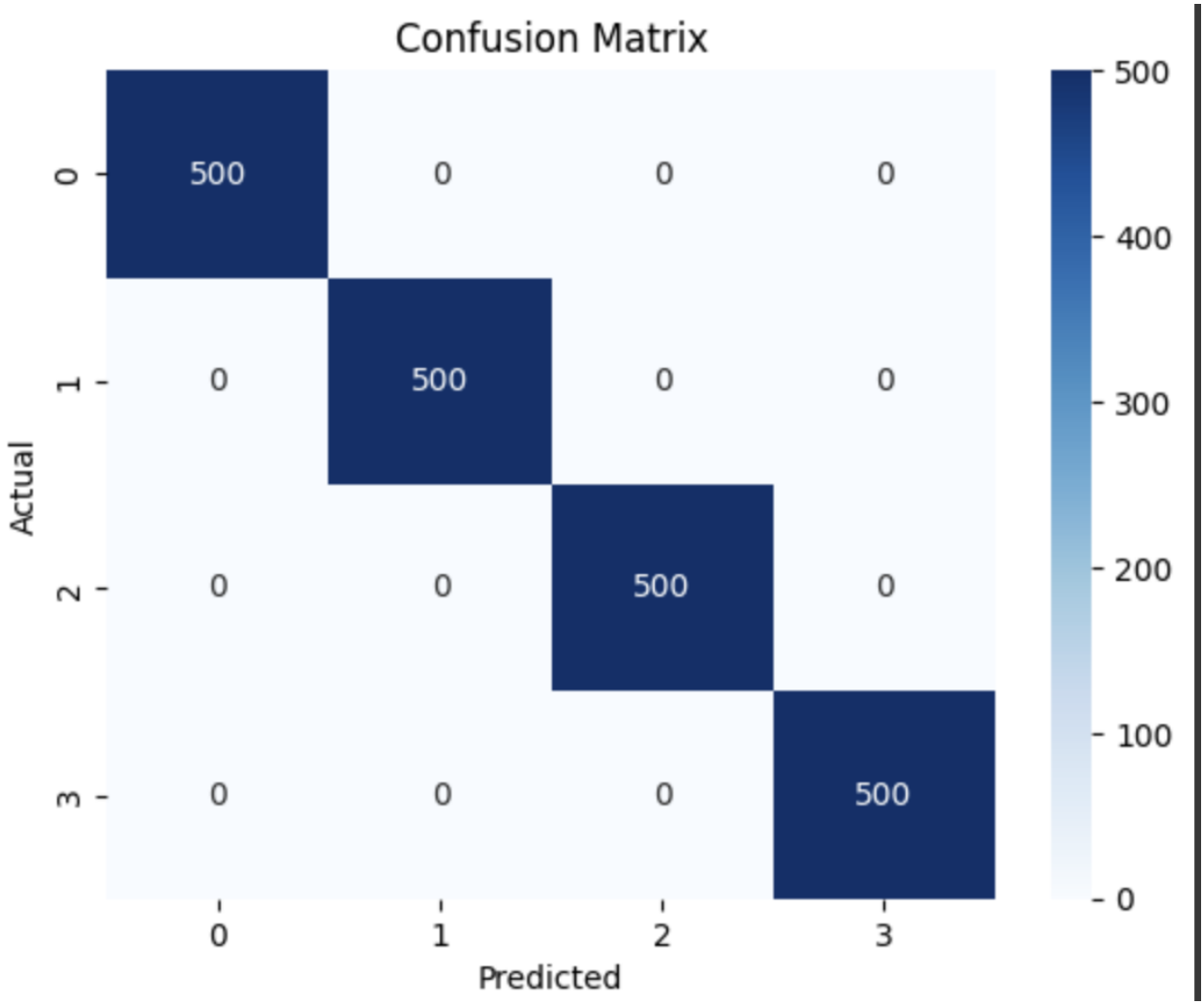
Training Confusion Matrix for random forest with best hyperparameters:

[[500 0 0 0]

[ 0 500 0 0]

[ 0 0 500 0]

[ 0 0 0 500]]




Validation Accuracy for random forest with best hyperparameters: 0.8975

Validation Precision for random forest with best hyperparameters: 0.8998582499486065

Validation Recall for random forest with best hyperparameters: 0.8975

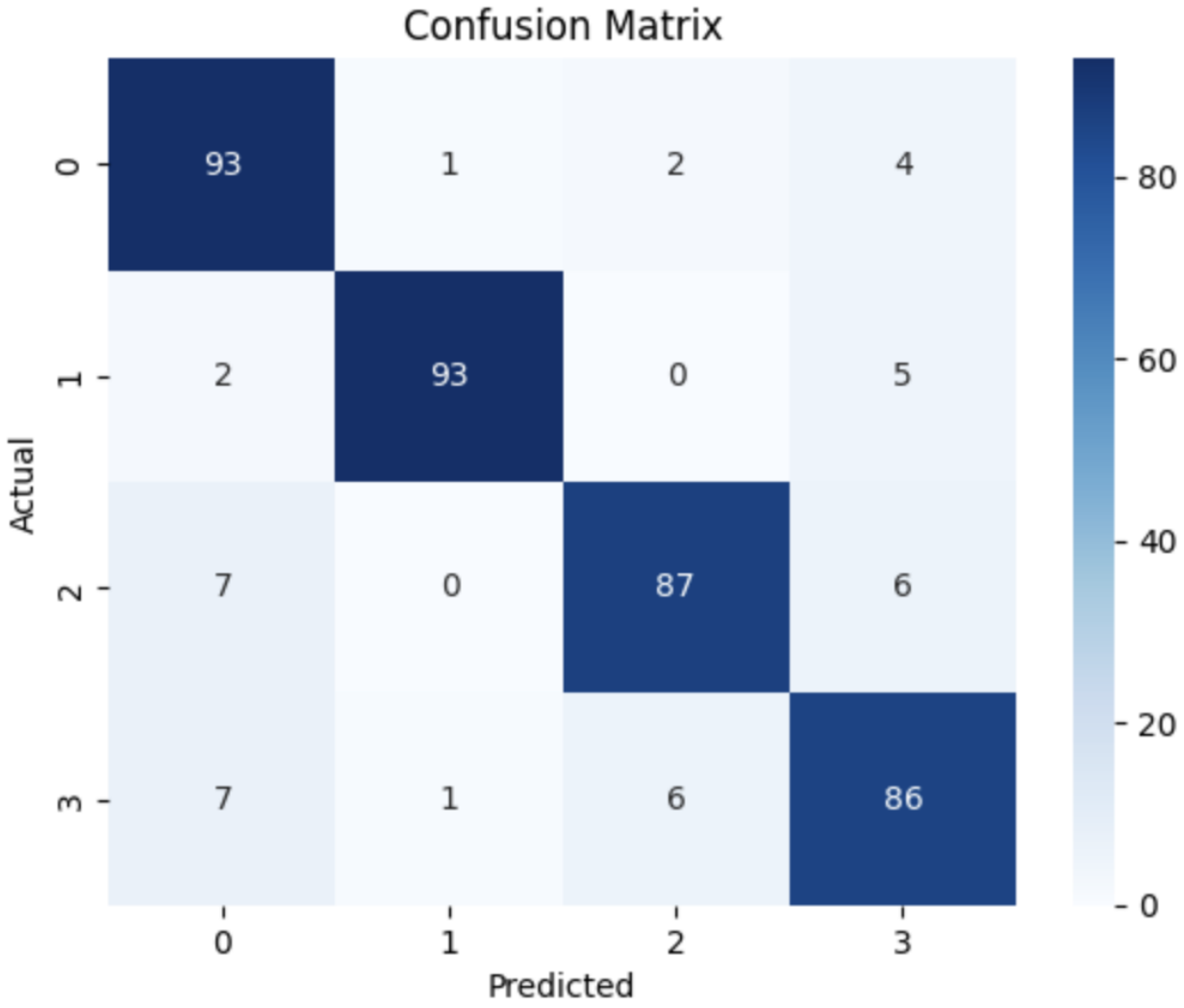
Validation Confusion Matrix for random forest with best hyperparameters:

[[93 1 2 4]

[ 2 93 0 5]

[ 7 0 87 6]

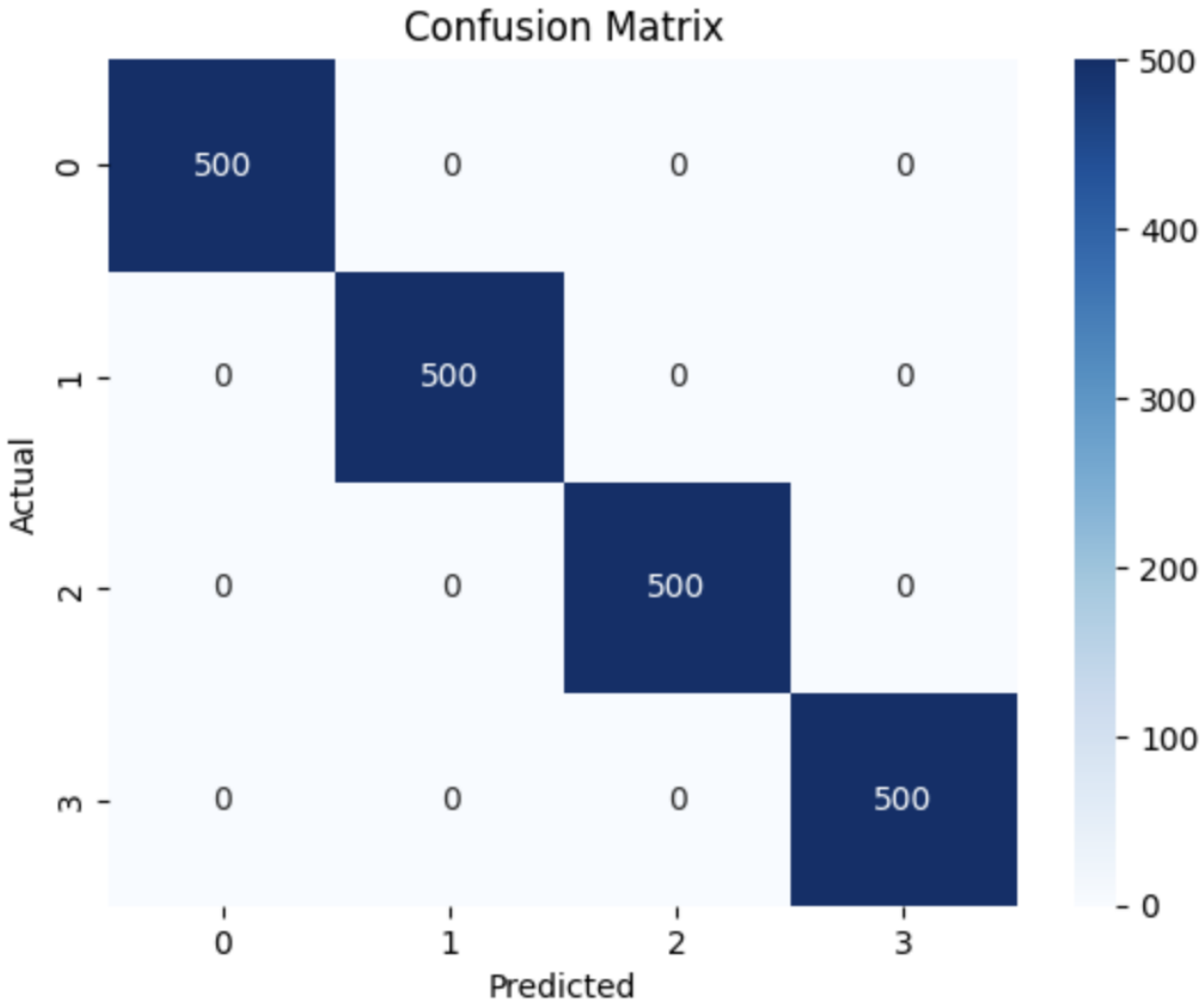
[ 7 1 6 86]]



## e) Gradient Boosted Trees and XGBoost

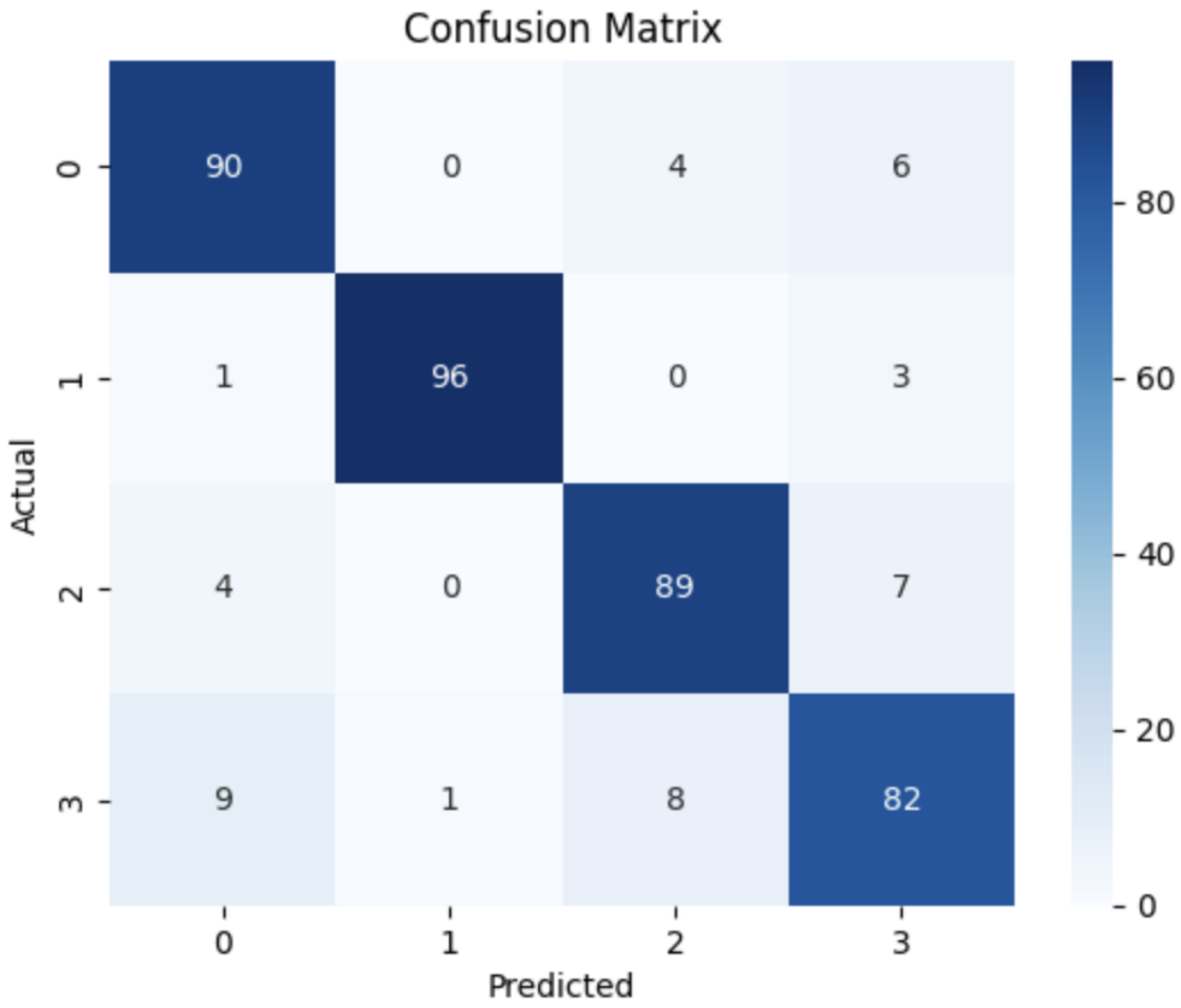
Gradient Boosting Training Accuracy: 1.0

Gradient Boosting Training Confusion Matrix:



Gradient Boosting Validation Accuracy: 0.8925

Gradient Boosting Validation Confusion Matrix:

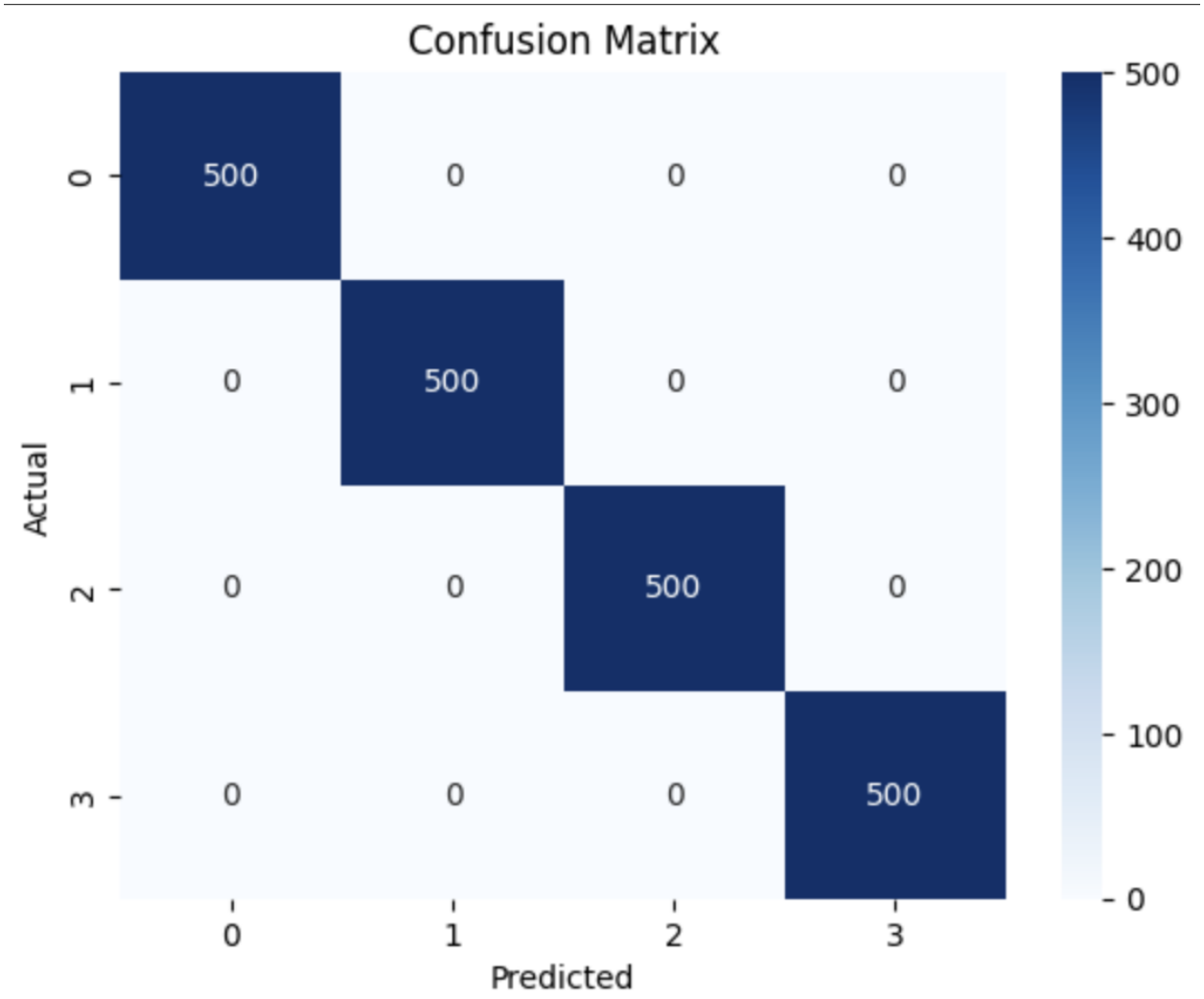


**Grid Search to find the best hyperparameters:**

Best Hyperparameters:  {'max_depth': 8, 'n_estimators': 50, 'subsample': 0.6}

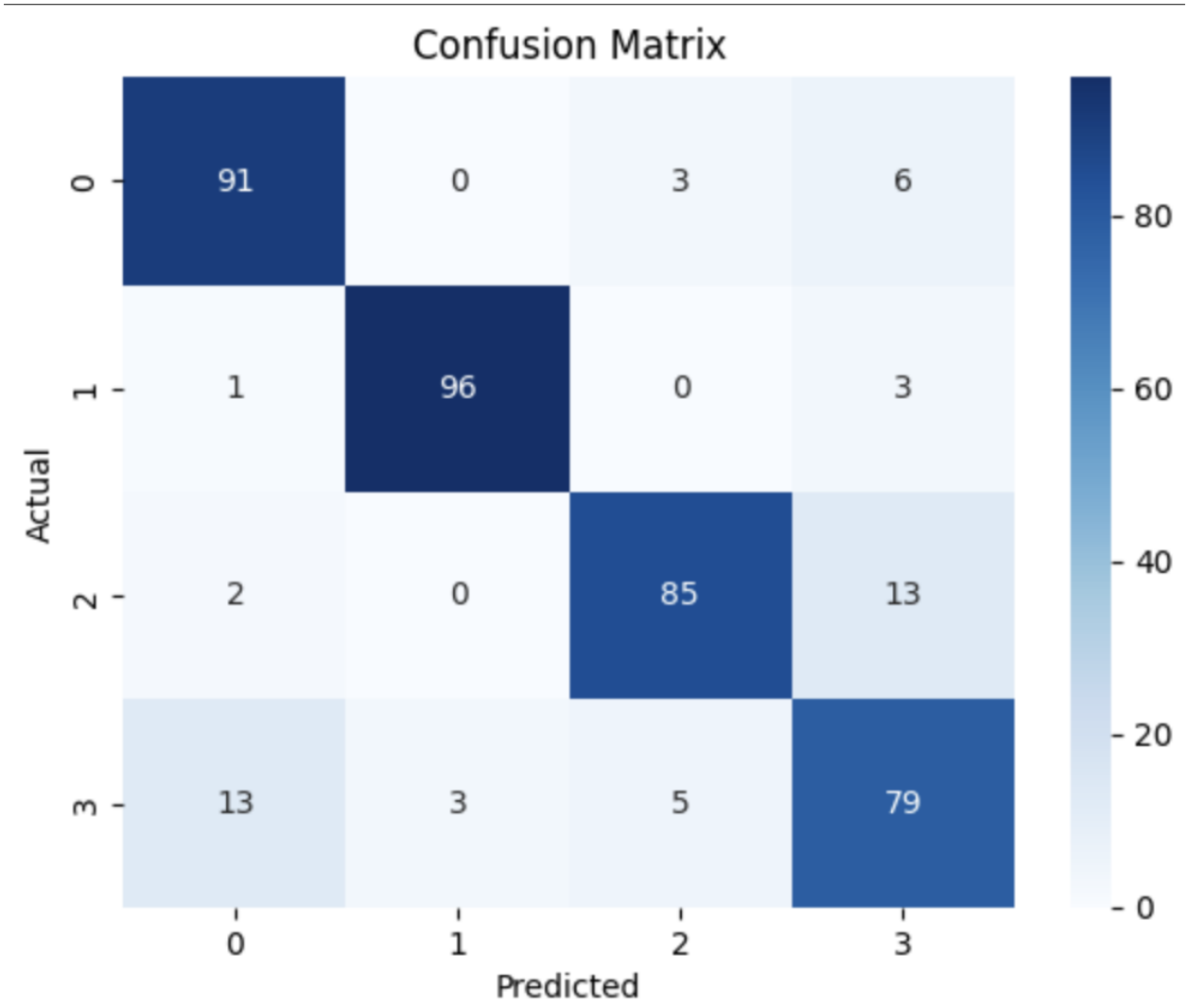
Gradient Boosting Training Accuracy: 1.0

Gradient Boosting Training Confusion Matrix:



Gradient Boosting Validation Accuracy: 0.8775

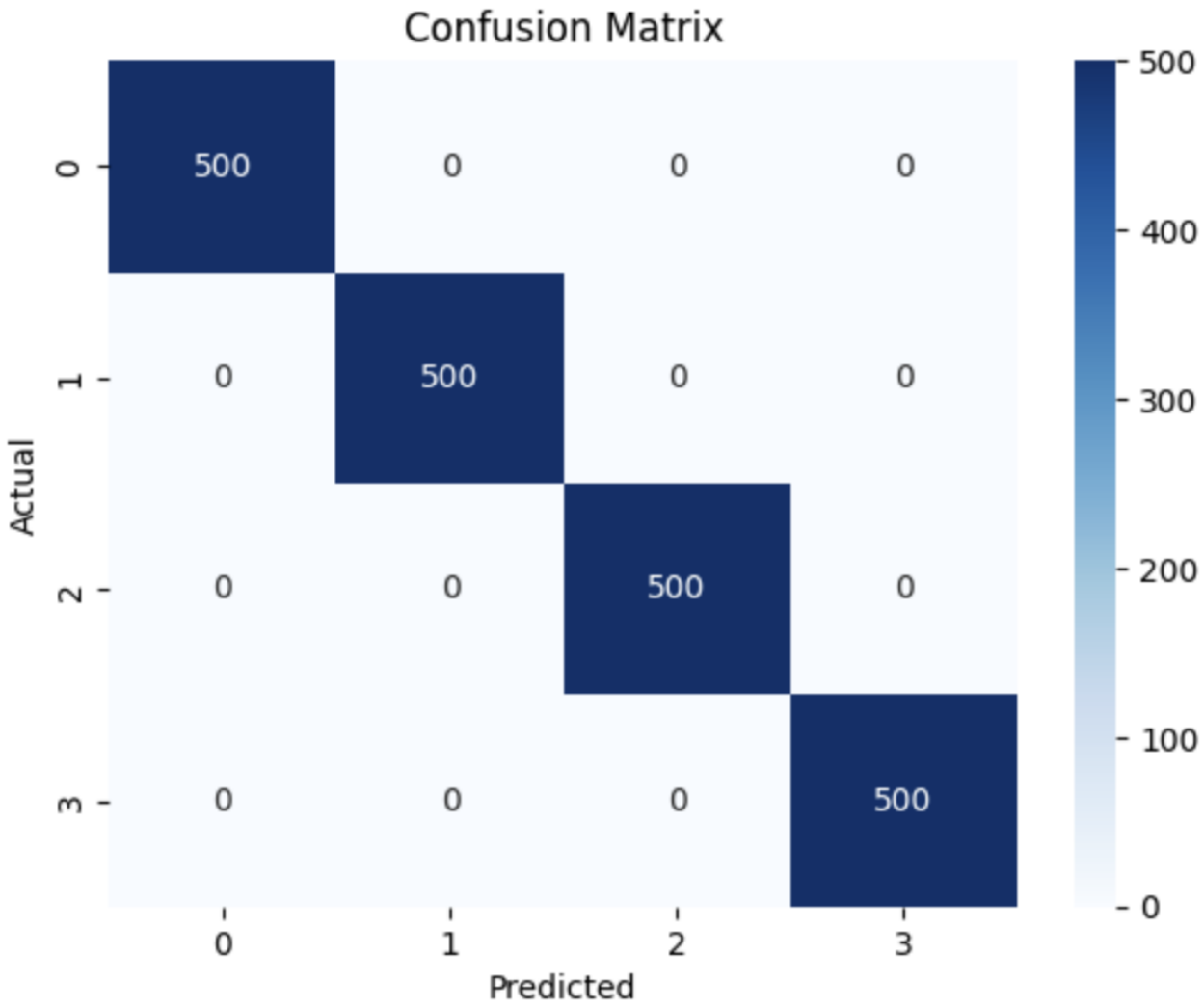
Gradient Boosting Validation Confusion Matrix:



**Extreme Gradient Boosting XGBoost:**

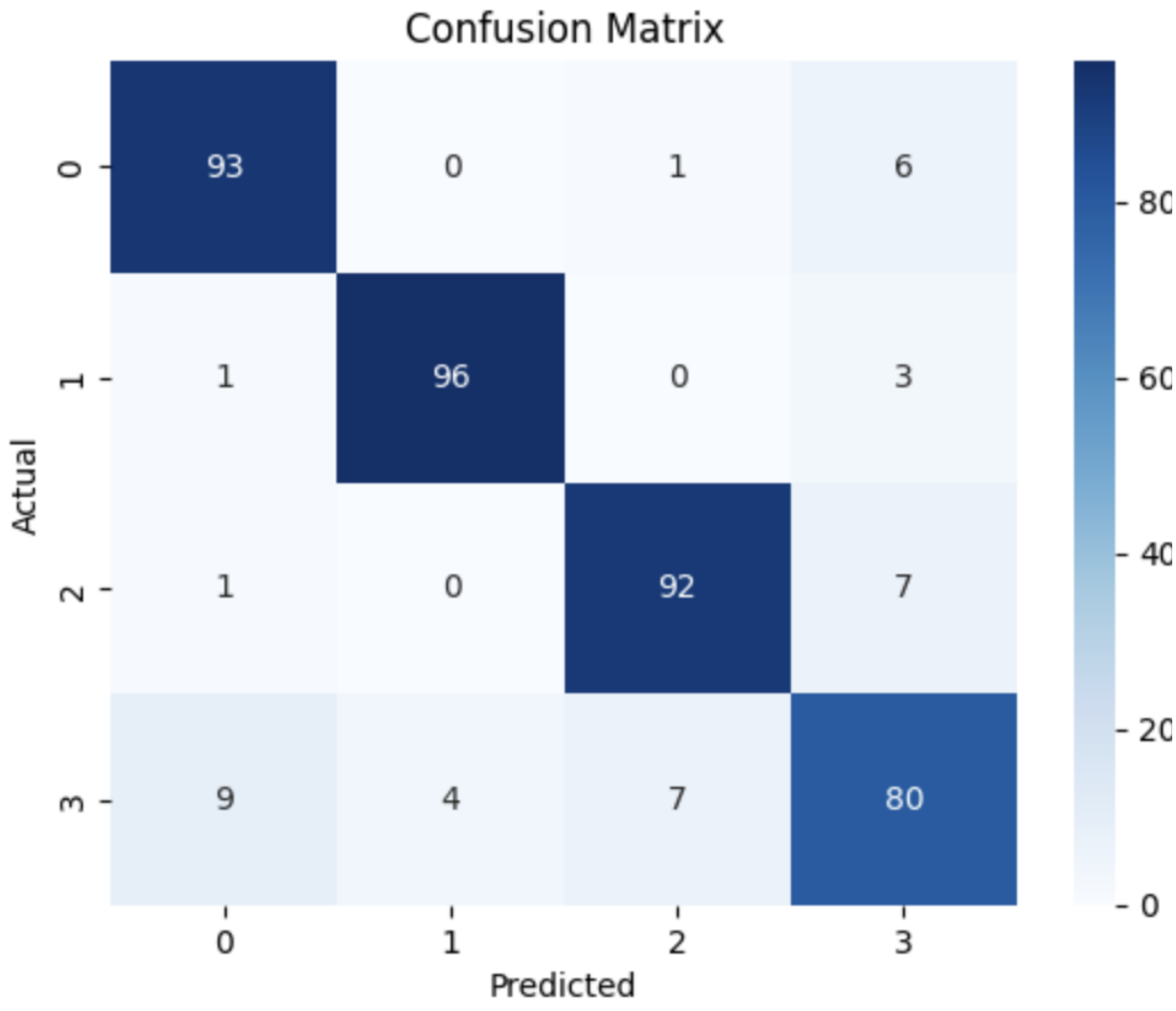
XGBoost Training Accuracy: 1.0

XGBoost Training Confusion Matrix:



XGBoost Validation Accuracy: 0.9025

XGBoost Validation Confusion Matrix:

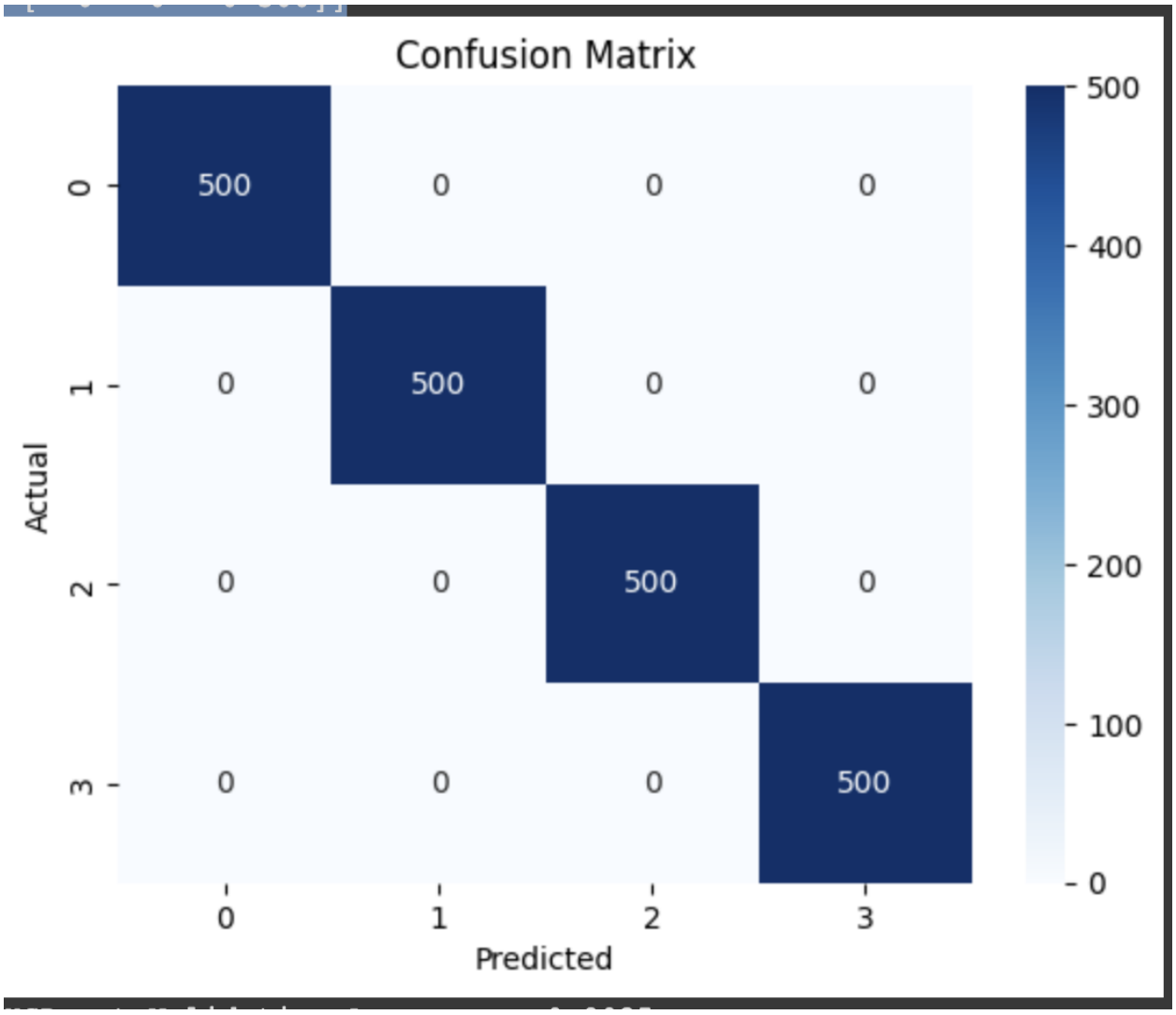


**Grid Search to find the best hyperparameters:**

XGBoost Best Hyperparameters:  {'max_depth': 8, 'n_estimators': 50, 'subsample': 0.6}

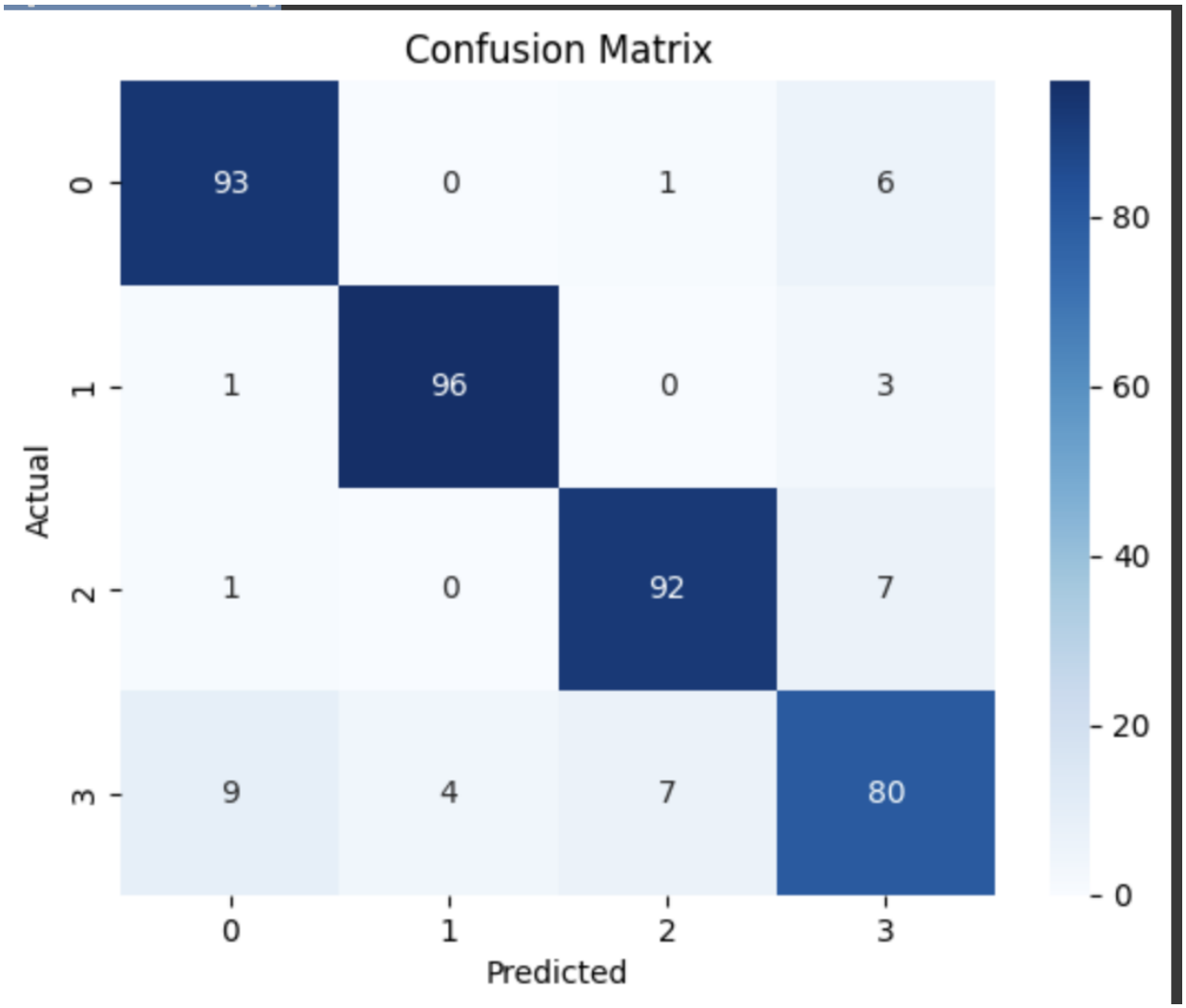
XGBoost Training Accuracy: 1.0

XGBoost Training Confusion Matrix:




XGBoost Validation Accuracy: 0.9025

XGBoost Validation Confusion Matrix:




**f)** Confusion matrices for all parts have been written in the respective parts.

**g) Real-Time Testing**

From 10 images, 7 images were predicted as face and 3 were predicted as dog.

So the accuracy came out to be 70%.