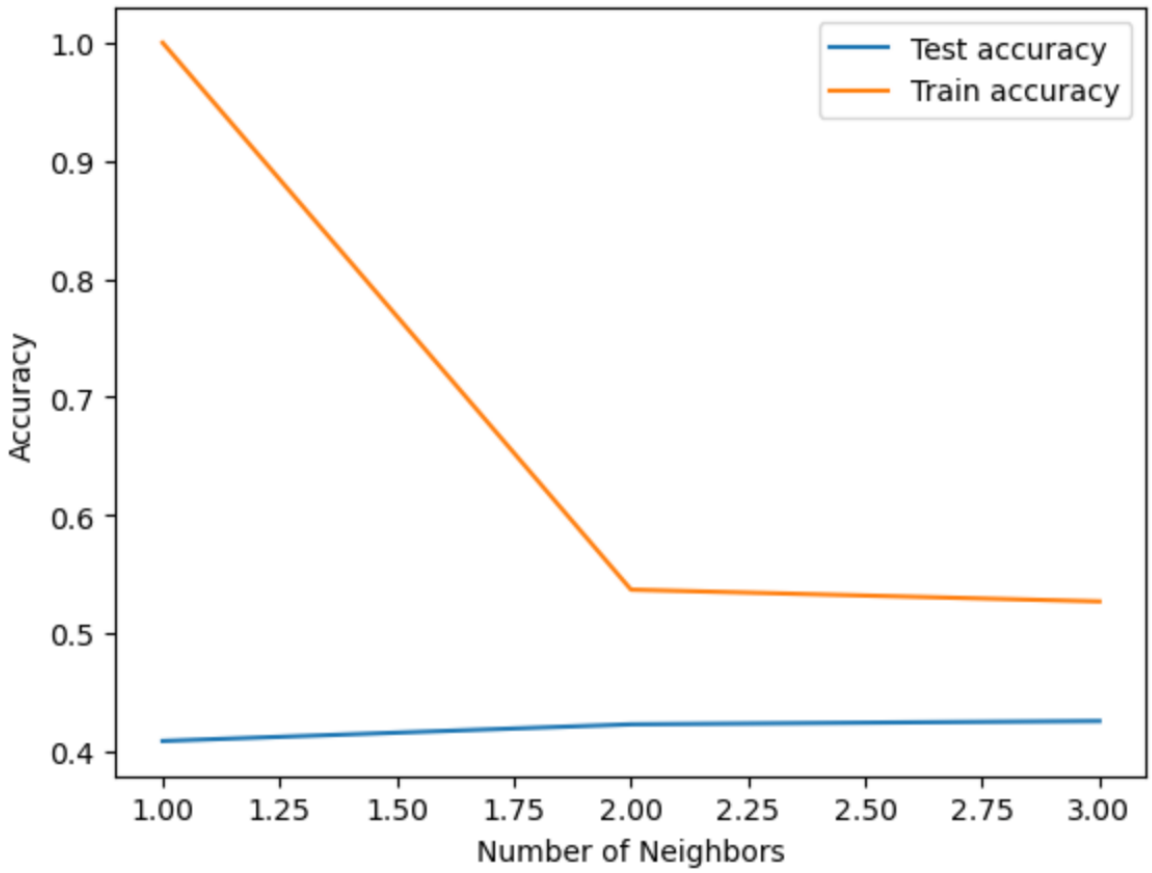
1.4 Answers - KNN

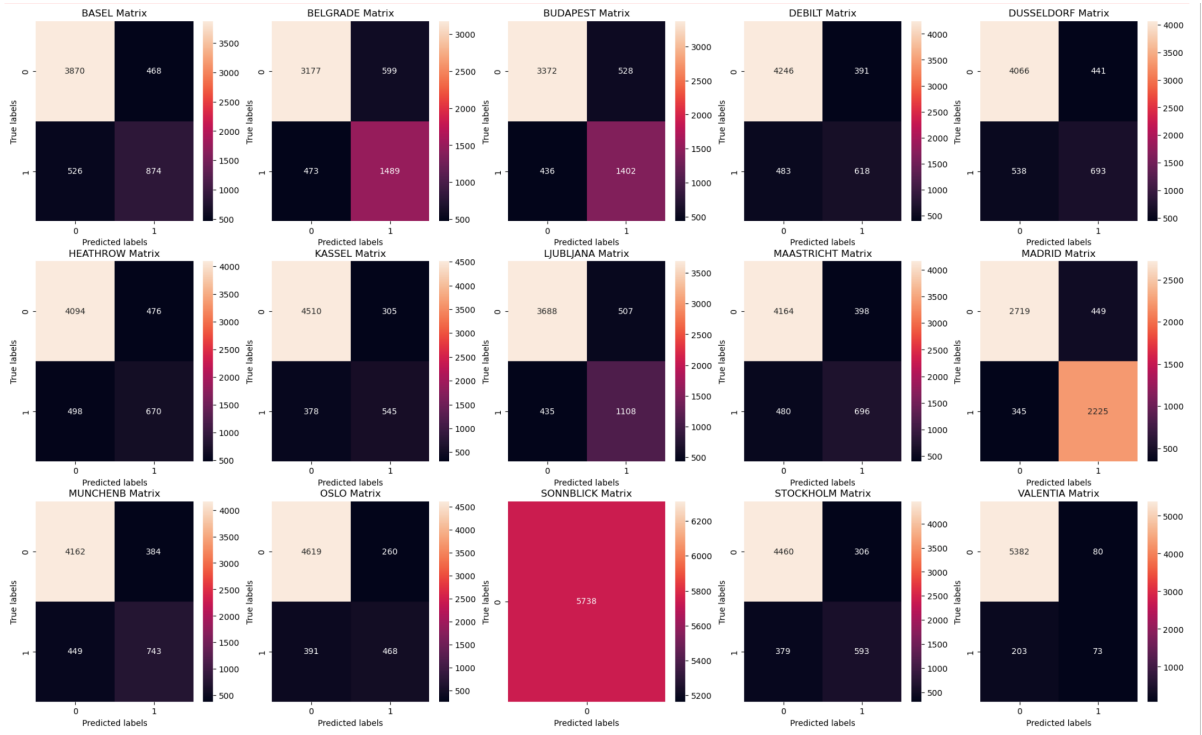
- Record your starting parameters, your final parameters, and the accuracy of the training and testing data. How does the number of neighbours affect the accuracy of the answers?

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



This shows for Training data the accuracy starts at 100 percent but sharply decreases as the neighbours change to 2, then has a very gradual decrease from 2 to 3. For Test data, it starts much lower and then only very gradually rises as the neighbours increase.

- Create a confusion matrix for your final training and testing scenarios. Save a screenshot in a document.



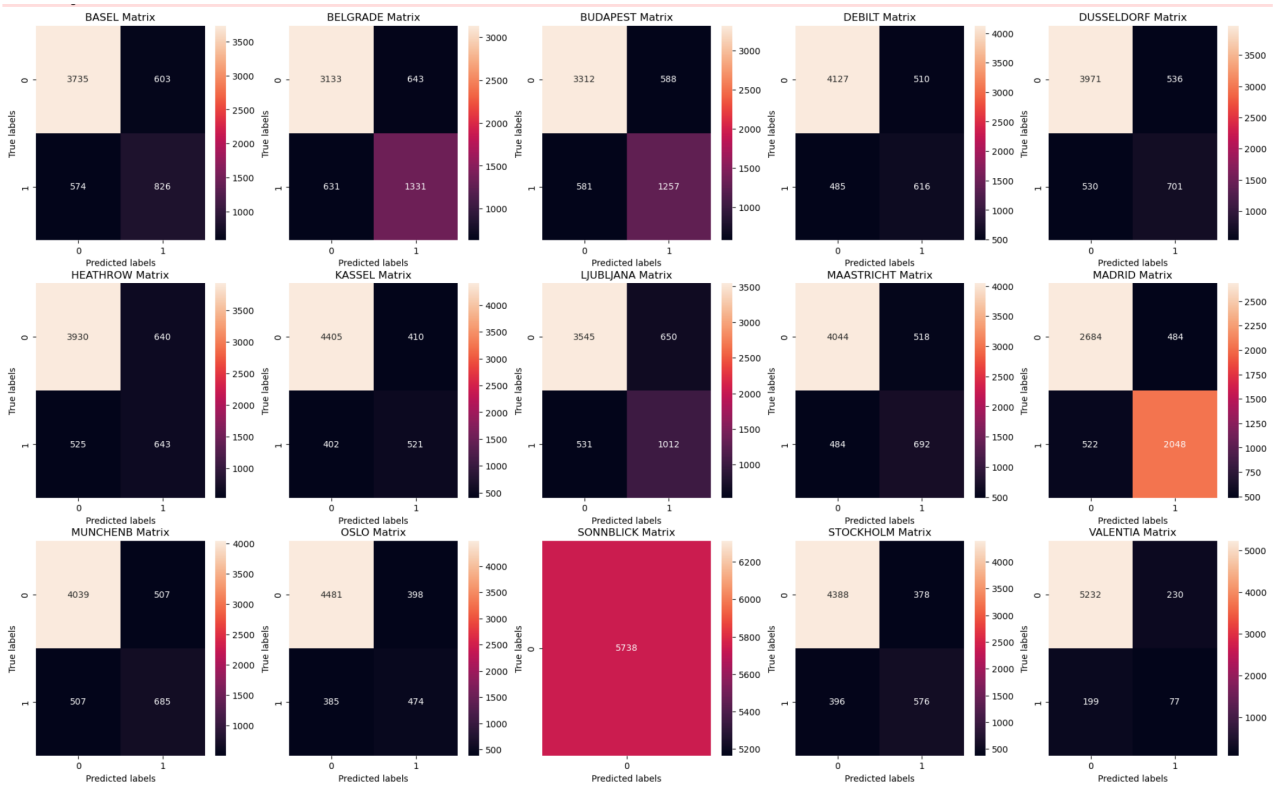
2. Write a paragraph answering the following questions:
    - How well does this algorithm predict the current data?
    - Are any weather stations fully accurate? Is there any overfitting happening?
    - Are there certain features of the data set (such as particular weather stations) that might contribute to the overall accuracy or inaccuracy?

Generally, it works well to predict the current data as the majority (80-90%) are accurate. No stations are fully accurate, although Valentia comes close with 95% accuracy and may skew the overall accuracy to be higher, while Sonnblick data does not work in the matrix as the entirety are for 0 (unpleasant weather) which would mean no-one should go out and is likely an error with the answer data!

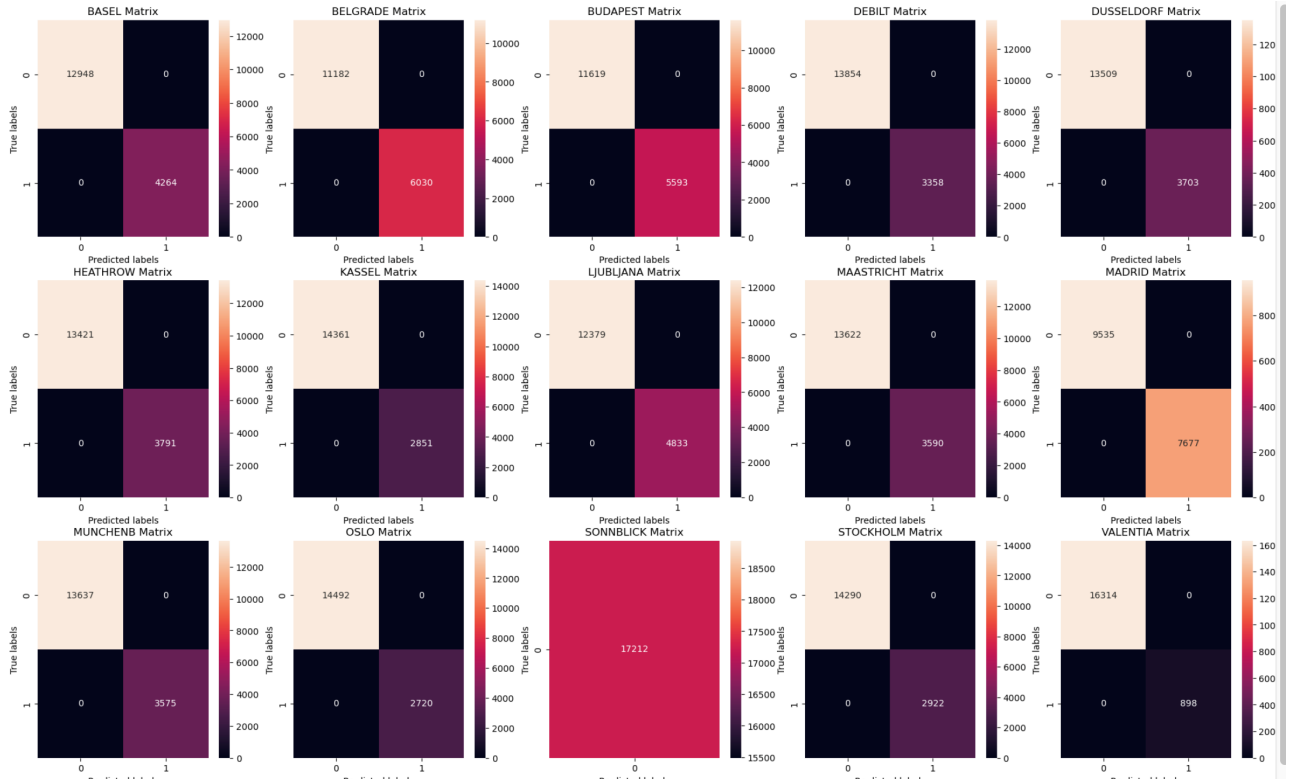
1.5 Answers – Decision Tree

- Accuracy – training is 39.9%, test is marginally higher at 40.5%
- I believe it should be pruned as the accuracy for both training and test data are quite low. Because the data is much denser in contrast to the sparse data of the Iris set, it can be pruned to lower the chance of overfitting and make the model quicker to run

Test Data Confusion Matrix



Training Data Confusion Matrix



1.5 Answers – ANN

Scaling is required as after scaling, the distributions are much closer to normal

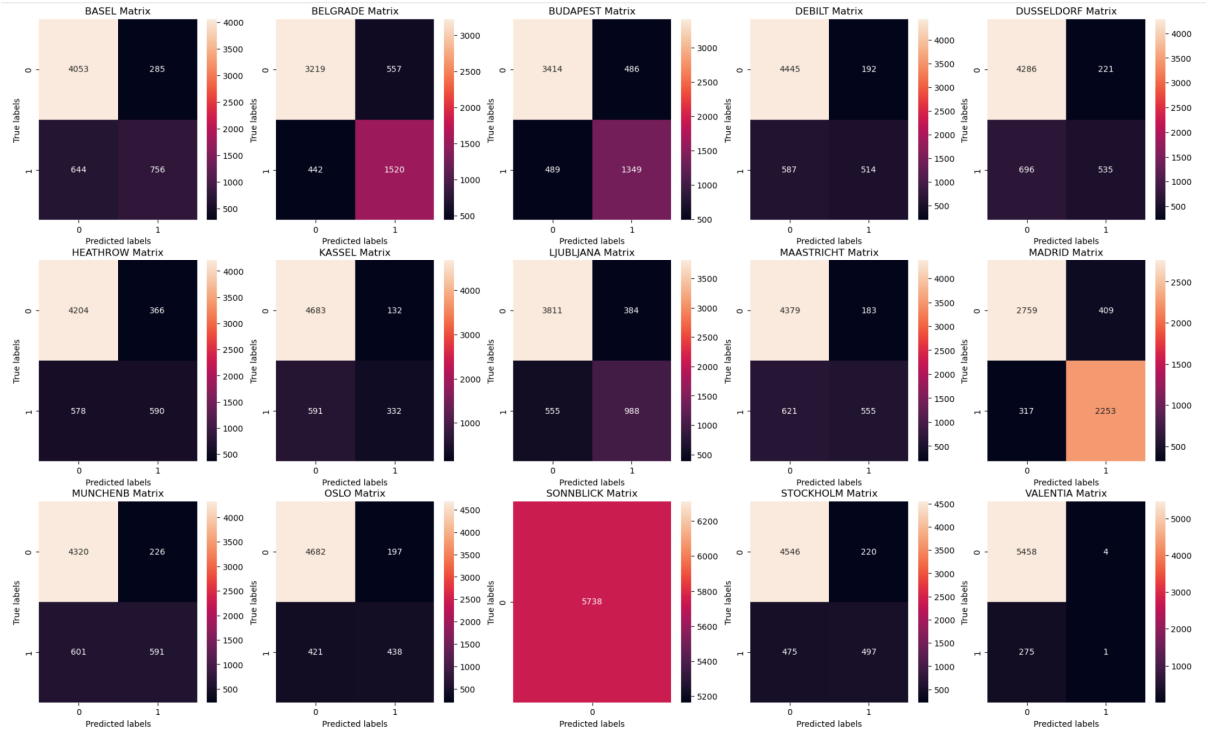
- Test out the number of layers, number of nodes per layer, max iterations, and tolerance. What combination drives the best accuracy of the training and testing data? Record your answers.
- Create a multi-station confusion matrix for at least three of your training and testing scenarios. Save screenshots in the document you created in Exercise 1.4.

Scenario #1

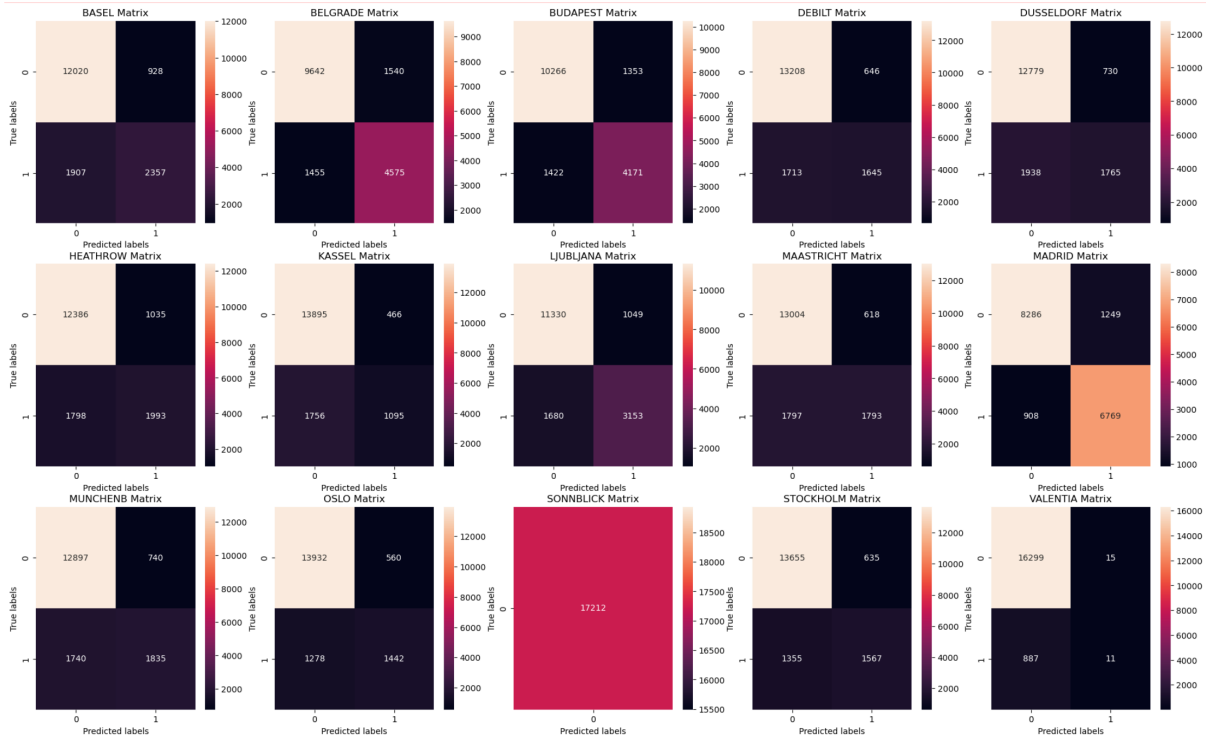
2 hidden layers with 5 nodes each, 500 max iterations, 0.0001 tolerance

Training accuracy is 43.7%, test accuracy is 44%

S1 Test Data Confusion Matrix



S1 Training Data Confusion Matrix

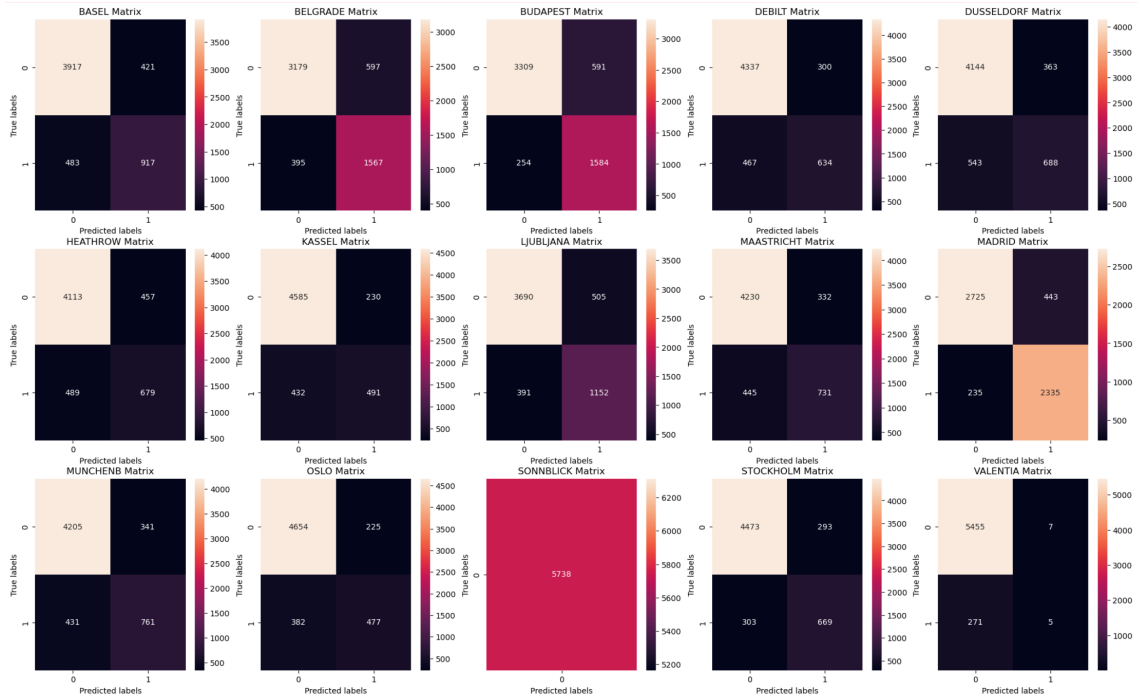


Scenario #2

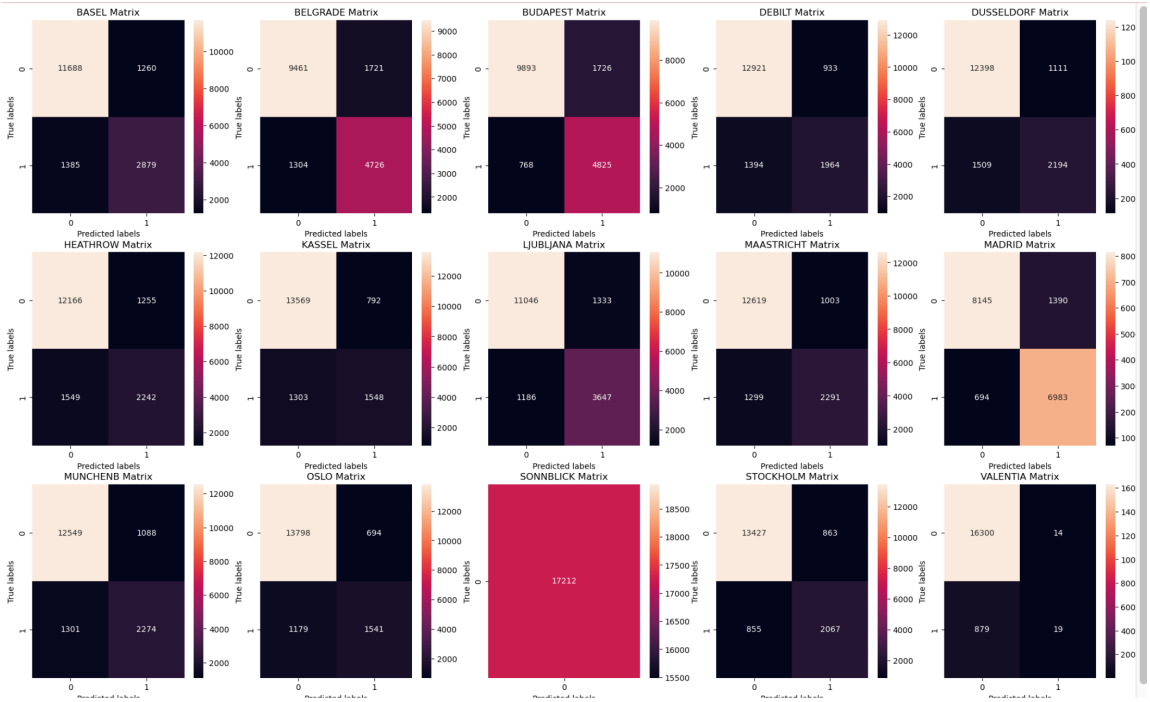
2 hidden layers with one with 10 nodes, one with 5 nodes, 500 max iterations, 0.0001 tolerance

Training accuracy is 44%, test accuracy is 44.8%. Only marginal improvement

S2 Test Data Confusion Matrix



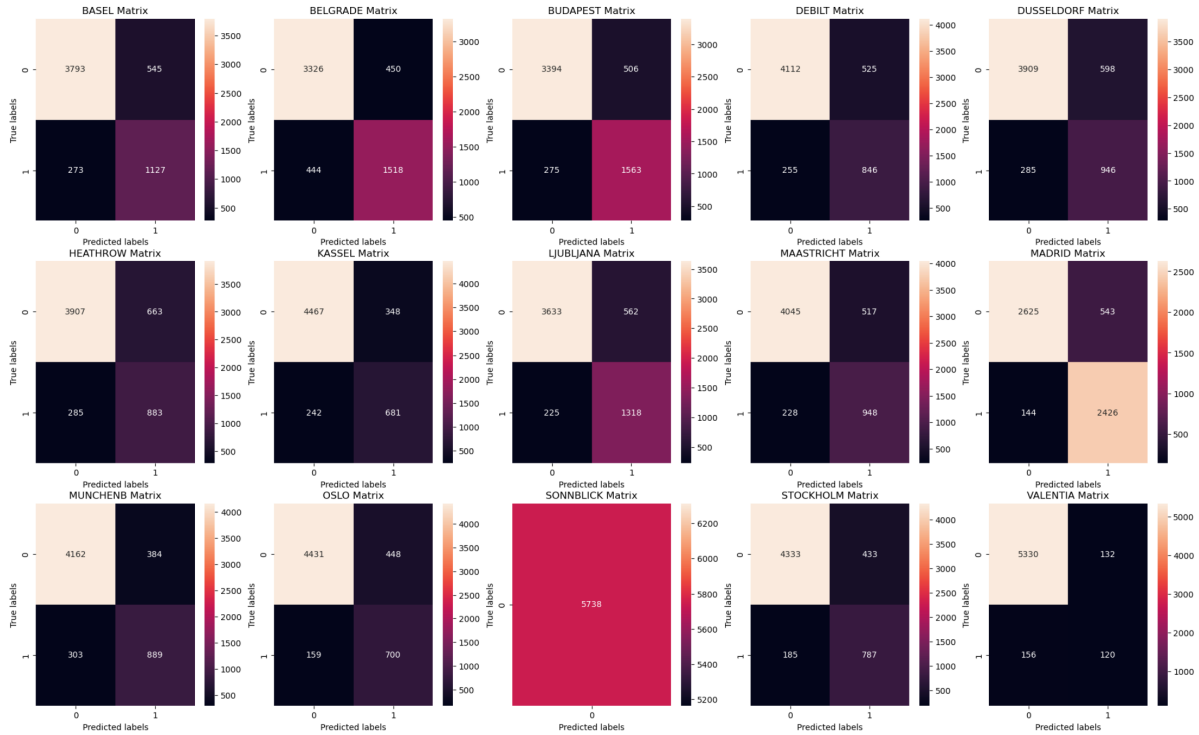
S2 Training Data Confusion Matrix



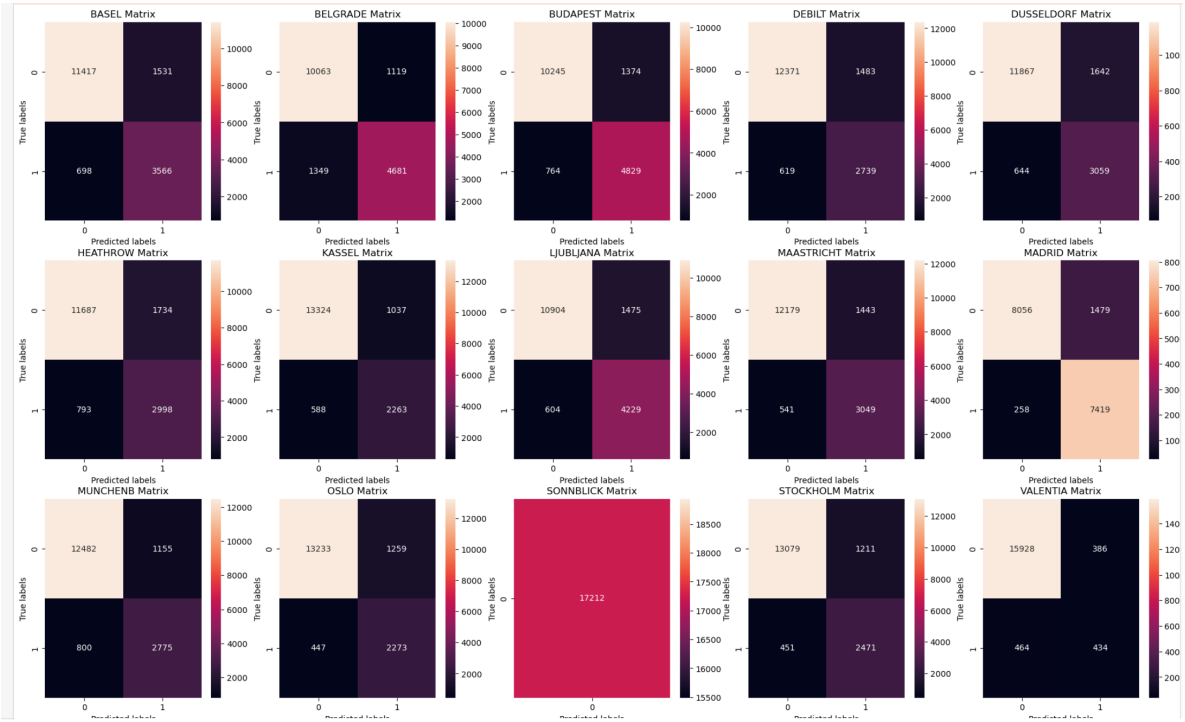
Scenario #3 = 3 hidden layers with one with 100 nodes, one with 50 nodes, one with 25 nodes, 2000 max iterations, 0.0001 tolerance

Training accuracy is 47%, test accuracy is 44.5%. Training data score increases 3% whereas test data slightly decreases, as it should when optimizing training data.

S3 Test Data Confusion Matrix



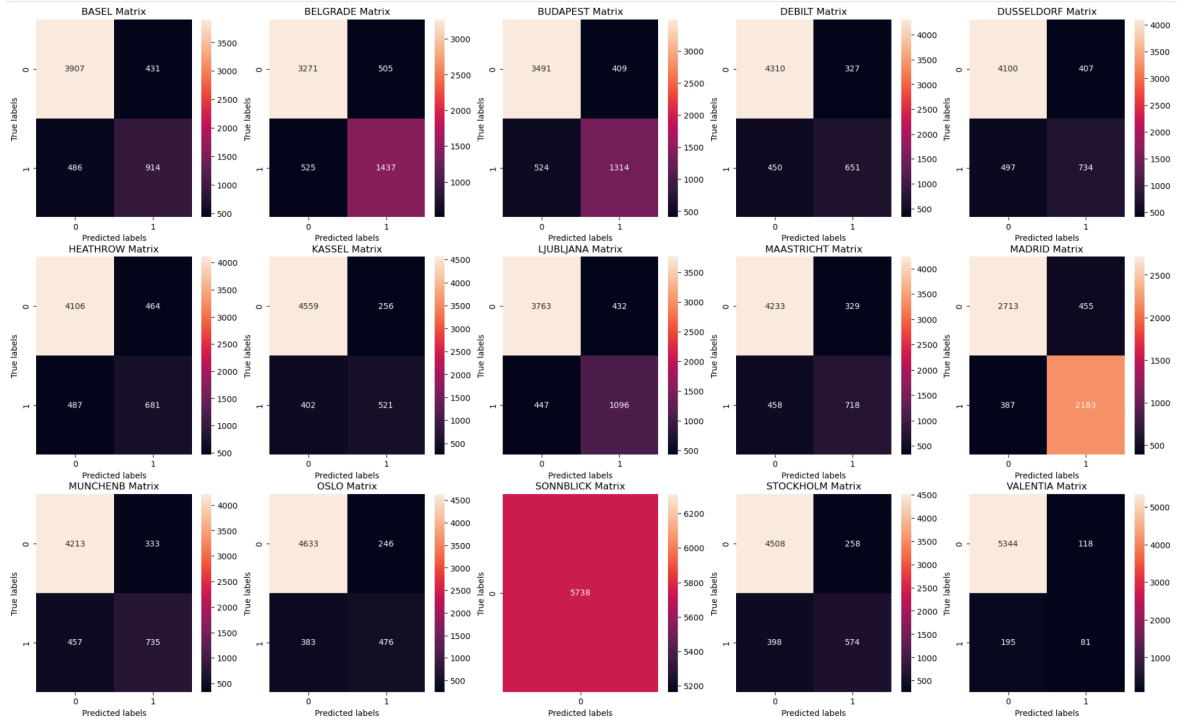
S3 Training Data Confusion Matrix



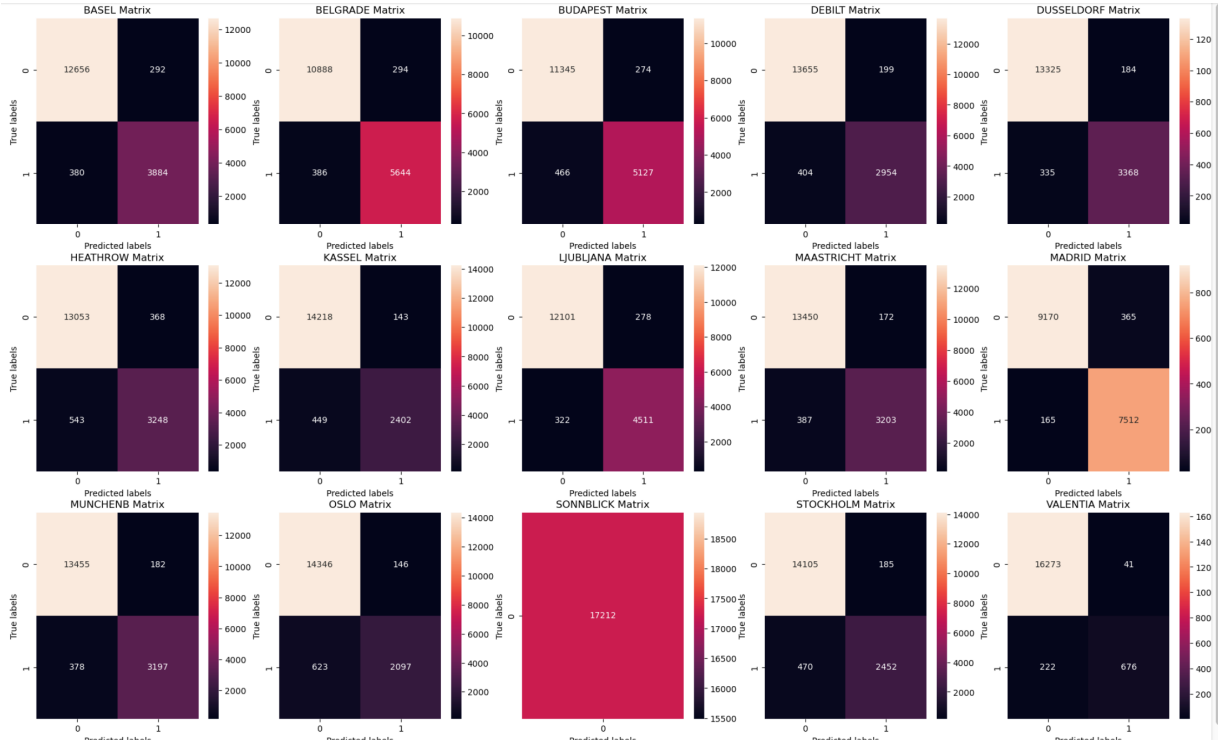
Scenario #4 = 3 hidden layers with one with 500 nodes, one with 250 nodes, one with 125 nodes, 1000 max iterations, 0.0001 tolerance

Training accuracy is 72.5%, test accuracy is 42.4%. Training data score vastly increases 25.5% as test data slightly decreases further

S4 Test Data Confusion Matrix



S4 Training Data Confusion Matrix



1. Write a paragraph that answers the following questions:
    - Which of these algorithms (including the KNN model from Exercise 1.4) do you think best predicts the current data?
    - Are any weather stations fully accurate? Is there any overfitting happening?
    - Are there certain features of the data set that might contribute to the overall accuracy?
    - Which model would you recommend that ClimateWins use?

The KNN model best predicts the current data with 80-90% accuracy compared to 72.5% for ANN and around 40% for the Decision Tree. Therefore, I would recommend ClimateWins uses the KNN algorithm; while no stations are fully accurate, Valentia comes close with 95% accuracy and may skew the overall accuracy to be higher, while Sonnblick data does not work in the matrix as the entirety are for 0 (unpleasant weather) and could be lowering the overall accuracy. I would further recommend acknowledging the limitation with Sonnblick and focusing on the other stations.