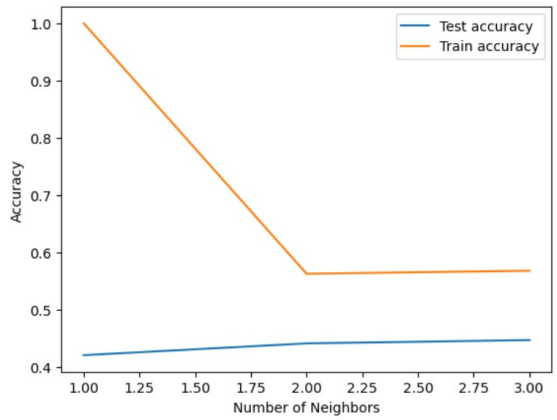
Exercise 4+5

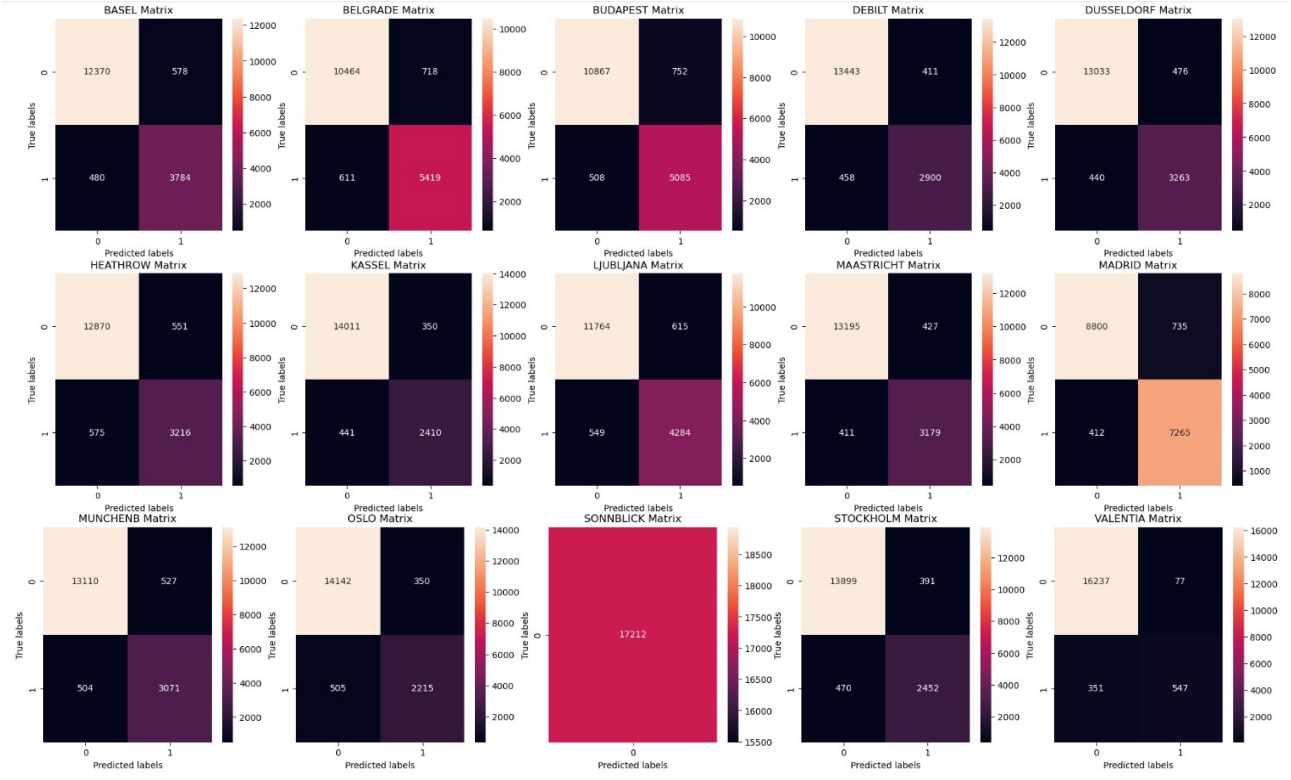
KNN – k-nearest neighbors algorithm

1a) Accuracy of the training & testing data for k = 1 – 3

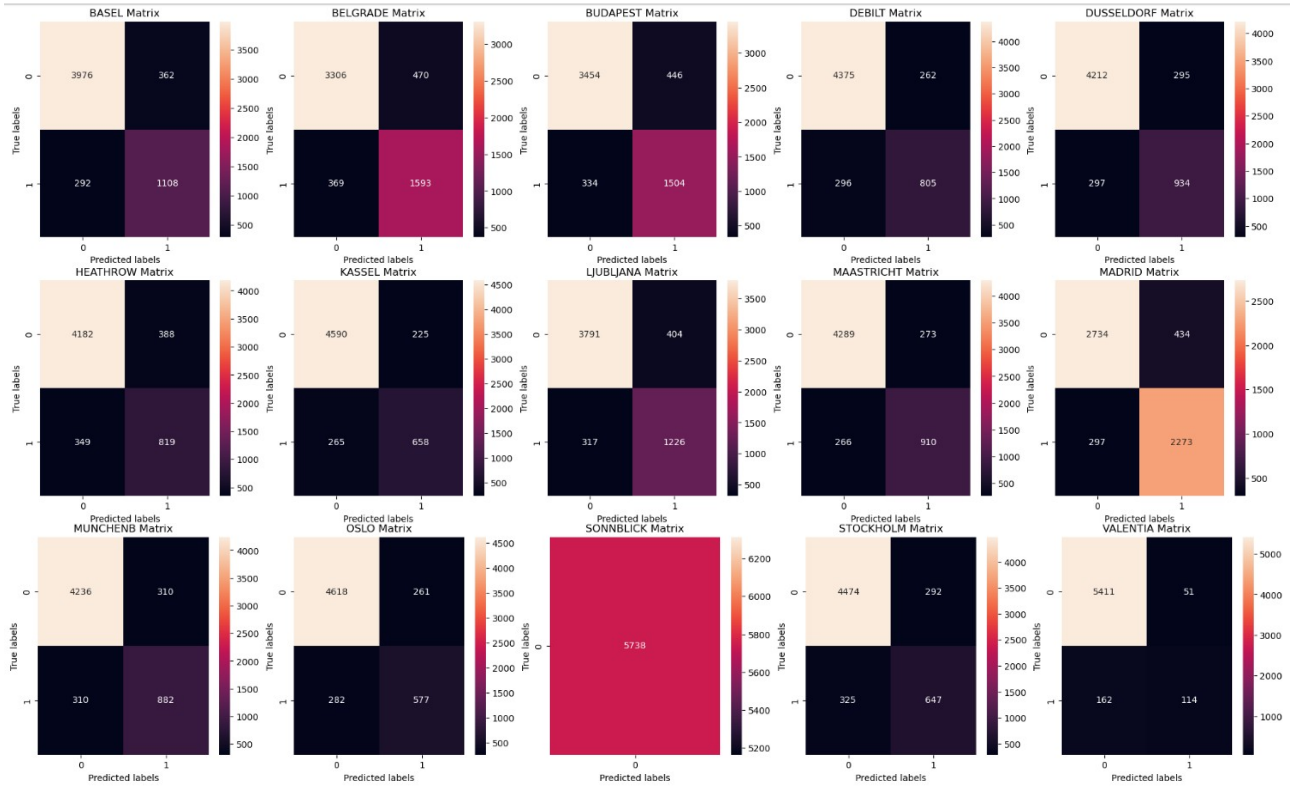


→ When k = 1 training and testing accuracies appear quite apart, indicating overfitting of the data. Using k = 2 or 3 would be a better solution, since testing accuracy keeps slightly improving.

1b) Confusion matrix on the training data



1c) Confusion matrix on the testing data



2) How well does this algorithm predict the current data?

→ The KNN algorithm, in general, is not really doing a great job here, with a rather low recall and precision rate.
→ E.g. the Belgrade weather station is missing 369 (false negative) out of a total of 1962 truly "pleasant" observations (*low recall*) and shows lots of false positive ones (*low precision*) as well, with 470 truly "unpleasant" observations out of a total of 2063 predicted "pleasant" observations.

→ *Train accuracy score*: ***57%***
→ *Test accuracy score*: ***45%***
→ *Recall rate* of about ***81%***
→ *Precision rate* of about ***77%***

→ **Those numbers should be better for the system to be implemented.**

3. Are any weather stations fully accurate? Is there any overfitting happening?

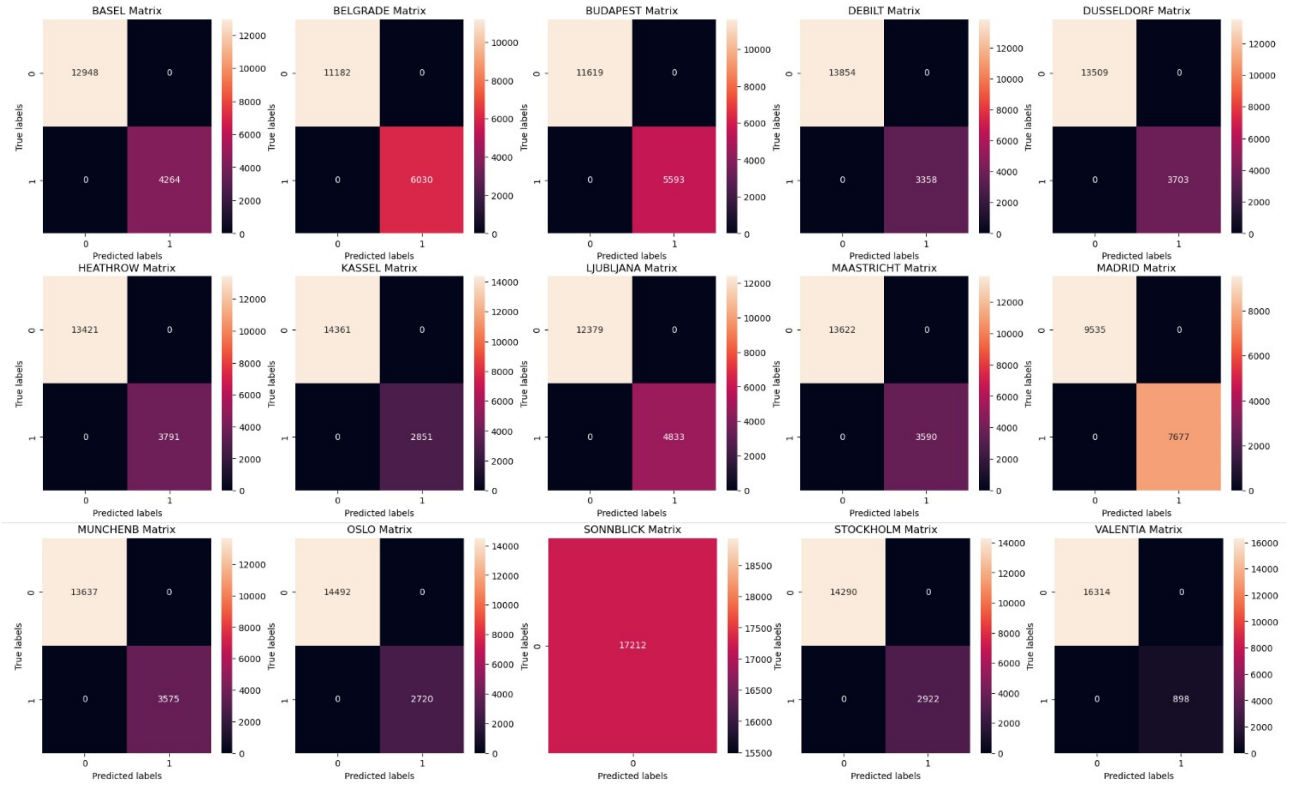
→ The SONNBLICK weather station is showing only "unpleasant" observations, which might lead to 100% accuracy. It would need to be investigated if the data for this weather station is complete or correct.

4. Are there certain features of the data set (such as particular weather stations) that might contribute to the overall accuracy or inaccuracy?

→ The SONNBLICK weather station may be a bias regarding the overall accuracy of the model. Further than that, e.g. the Kassel weather station seems to have relatively few "pleasant" observations, however numbers for precision and recall are similar to the Belgrade weather station.

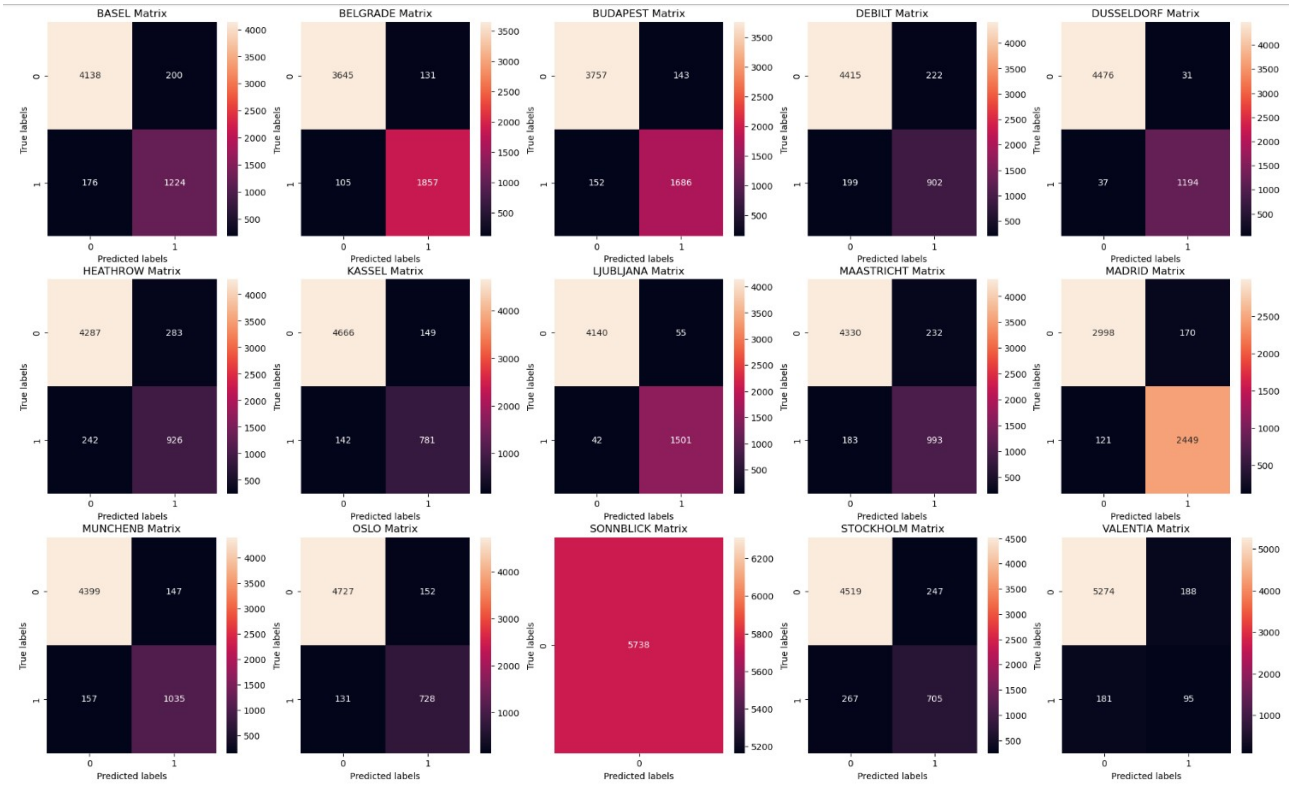
Decision tree algorithm

5a) Confusion matrix on the training data



→ *Train accuracy score*: ***0,60***

5b) Confusion matrix on the testing data



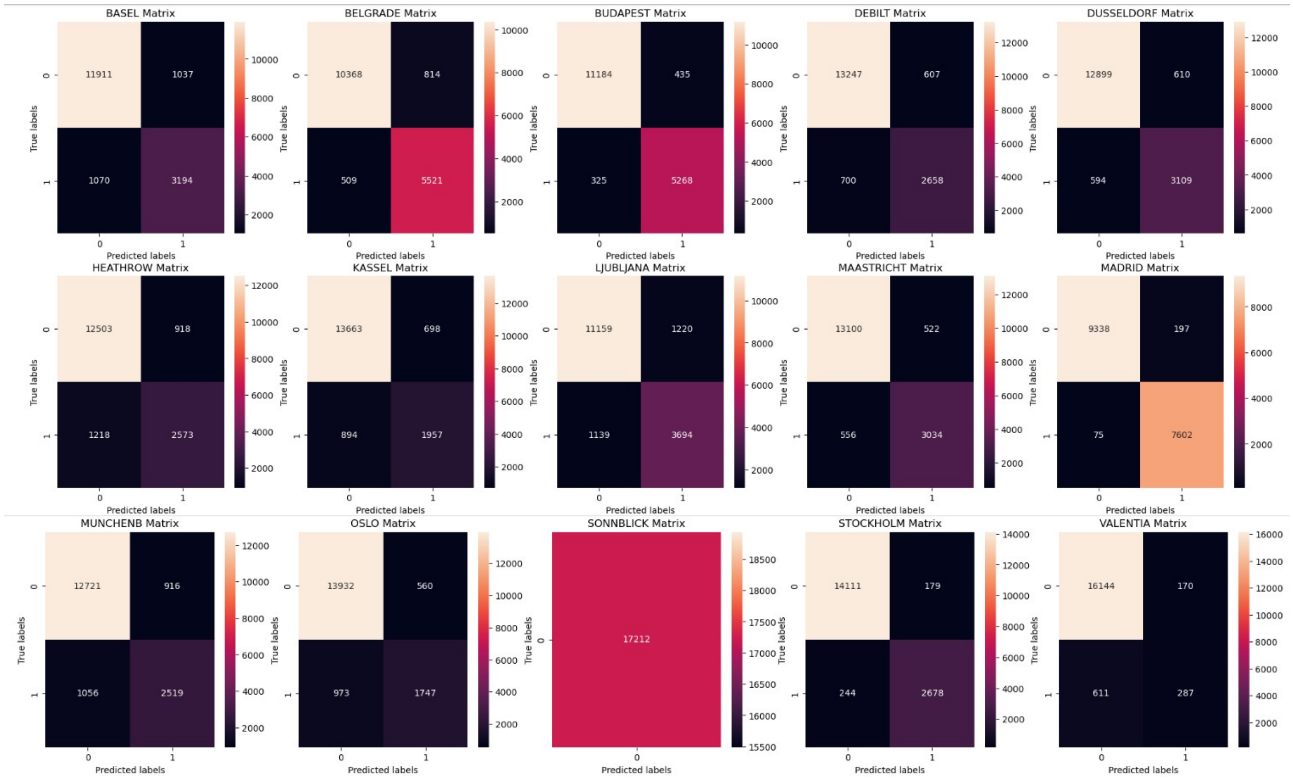
→ *Test accuracy score*: ***0,63***
→ The algorithm is not performing well, with rather low recall and precision rates.

→ e.g. the Stockholm weather station is missing 267 (false negative) out of a total of 972 truly “pleasant” observations (low recall)
→ and is showing lots of false positive ones (low precision),
with 247 truly “unpleasant” observations out of a total of 952 predicted “pleasant” observations

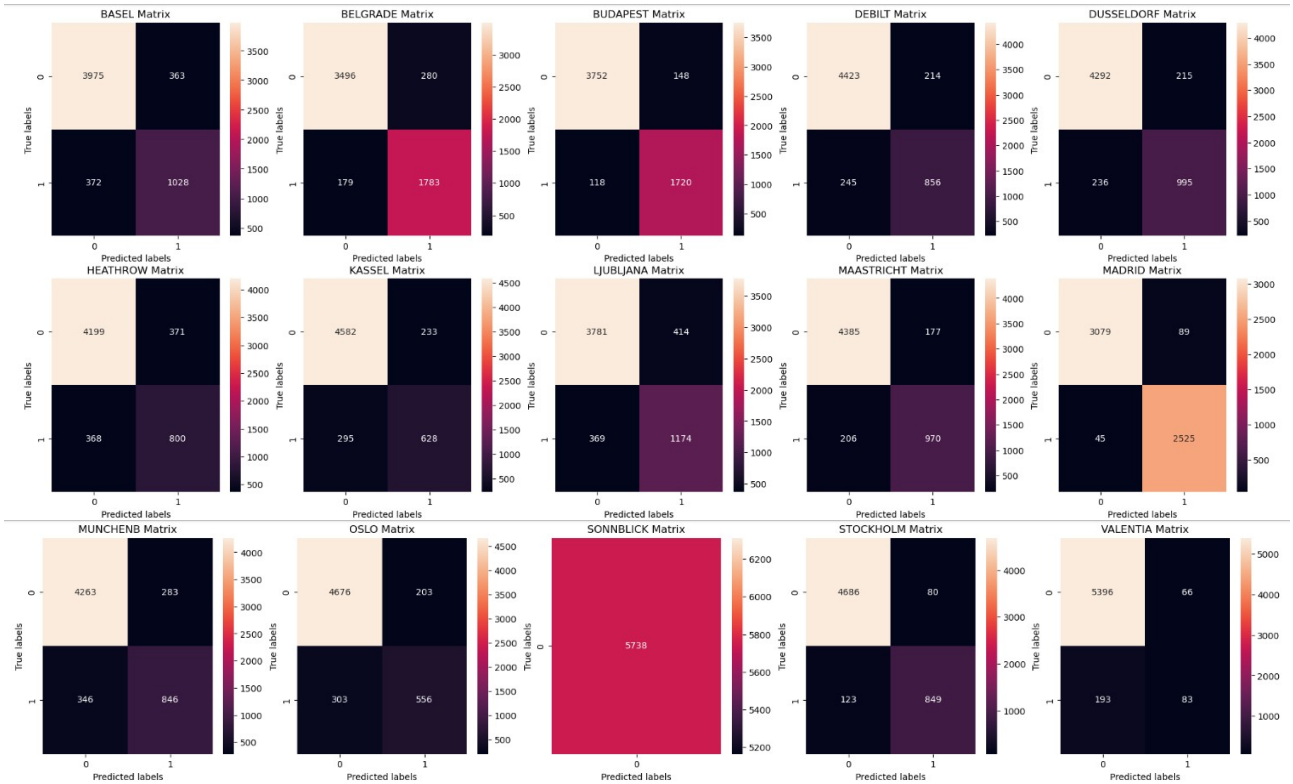
→ *Recall rate* of about ***72%***
→ *Precision rate* of about ***74%***

Artificial Neural Network

6a) Confusion matrix 1.1.: Multi-station confusion matrix on the training data
Model parameters: (*hidden_layer_sizes=(10, 5), max_iter=500, tol=0.0001)*

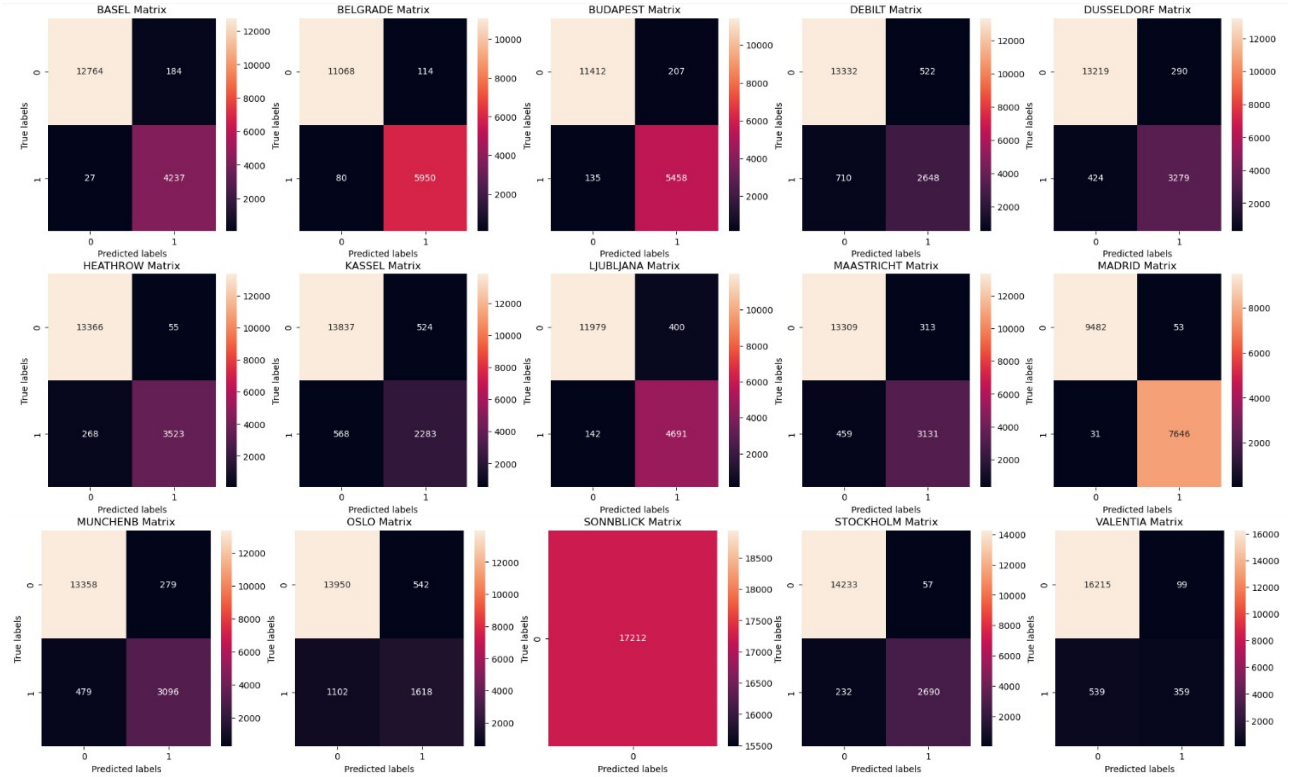


Confusion matrix 1.2.: Multi-station confusion matrix on the testing data
Model parameters: (*hidden_layer_sizes=(10, 5), max_iter=500, tol=0.0001)*

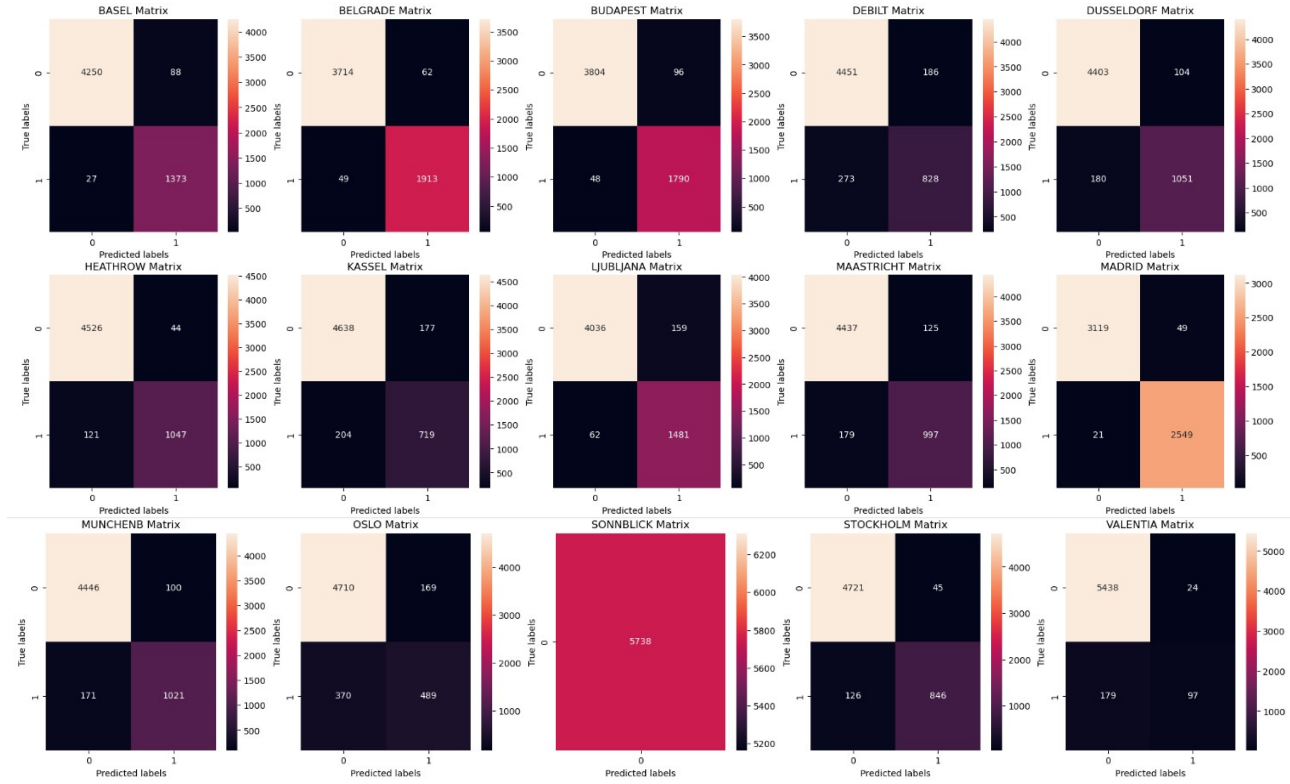


→ *Accuracy scores* on both the *training and testing data* respectively: ***0.53, 0.53***

6b) Confusion matrix 2.1.: Multi-station confusion matrix on the training data
Model parameters: *(hidden_layer_sizes=(20, 10, 10), max_iter=1000, tol=0.0001)*

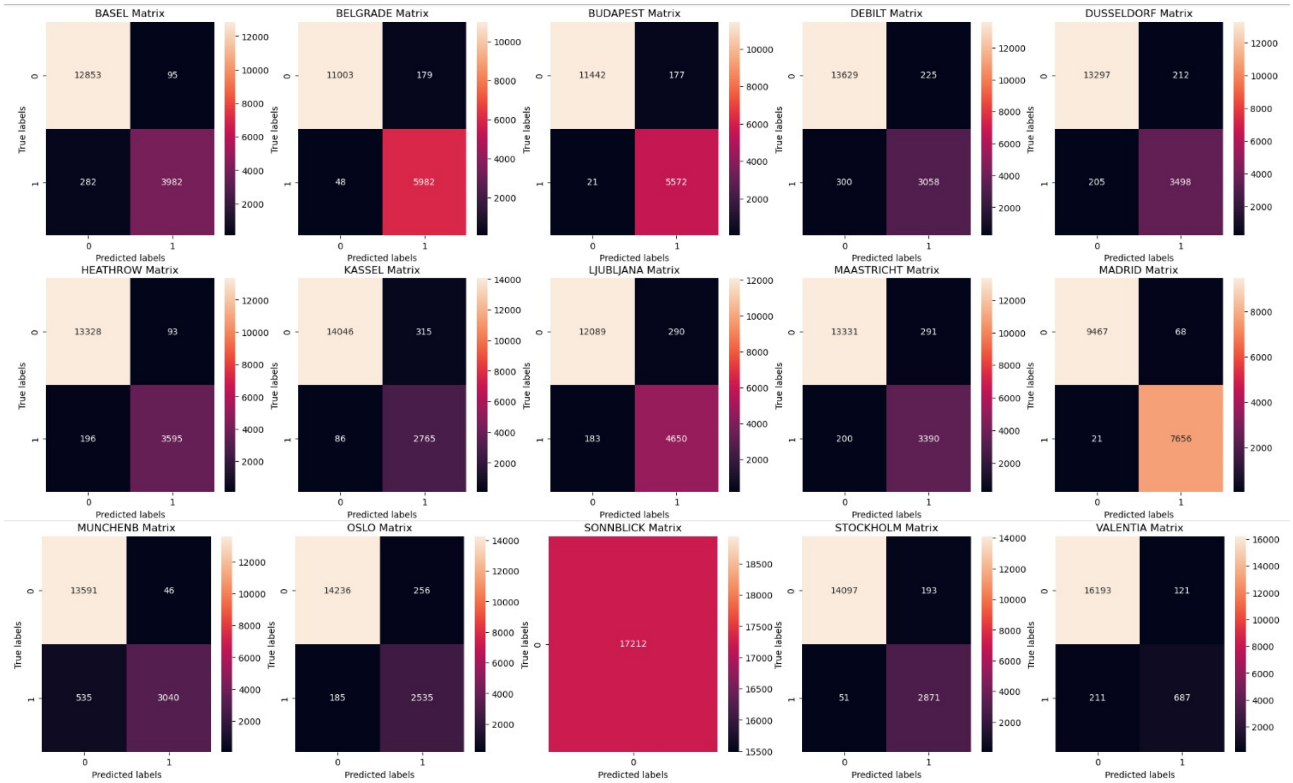


Confusion matrix 2.2.: Multi-station confusion matrix on the testing data
Model parameters: *(hidden_layer_sizes=(20, 10, 10), max_iter=1000, tol=0.0001)*

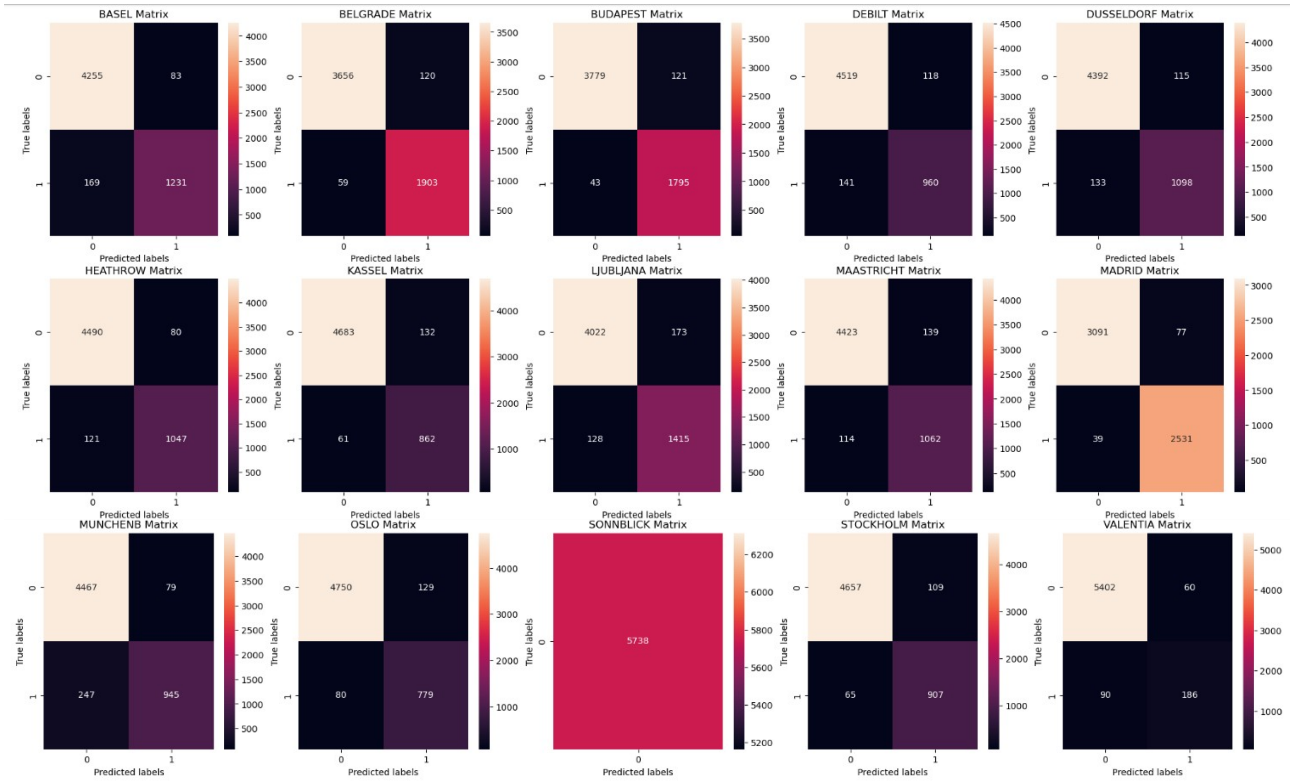


→ *Accuracy scores* on both the *training and testing data* respectively: ***0.68, 0.65***

6c) Confusion matrix 3.1.: Multi-station confusion matrix on the training data
Model parameters: (hidden_layer_sizes=(50, 25, 25), max_iter=1000, tol=0.0001)



Confusion matrix 3.2.: Multi-station confusion matrix on the testing data
Model parameters: (hidden_layer_sizes=(50, 25, 25), max_iter=1000, tol=0.0001)



→ *Accuracy scores* on both the *training and testing data* respectively: ***0.78, 0.66***
→ *Recall rate of about* ***97%***
→ *Precision rate of about* ***94%***

→ ***As the accuracy score on the testing data has shown to decrease with higher node numbers, this is the final model. Further adjustments would only lead to higher accuracy scores on the training data, however, scores on the testing data are dropping, which might be due to overfitting. Moreover, accuracy scores are not high and it should be considered if another approach might be better.***

7) Which of these algorithms best predicts the current data?

→ I'd suggest to use the ANN model, as the last configuration is showing the best overall results (a slightly higher test accuracy score, but also the confusion matrix appearing more accurate, with less errors, and thus a higher recall and precision rate).

8) Any weather stations fully accurate? Is there any overfitting happening?

→ As already mentioned, the SONNBLICK weather station consists only of 'unpleasant' data. Might want to be considered to exclude that weather station from the analysis.

In addition, overfitting appears to happen when further adjusting model parameters. Maybe after dropping the SONNBLICK weather station that would look different.

9) Any features of the data set that contribute to the overall accuracy?

→ One important aspect is the amount of data/data points/observations available.
→ Then, of course, data quality measures like validity, consistency and reliability are important.

10) Which model to recommend ClimateWins to use?

→ I’d recommend the ANN model.