1.5: Supervised Learning Algorithms Part 2

Paul Maden.

Last Update: 25/03/2025

Decision Tree model - Record the accuracy of the training and testing data.

Training Accuracy: 46.10%

Testing Accuracy: 46.11%

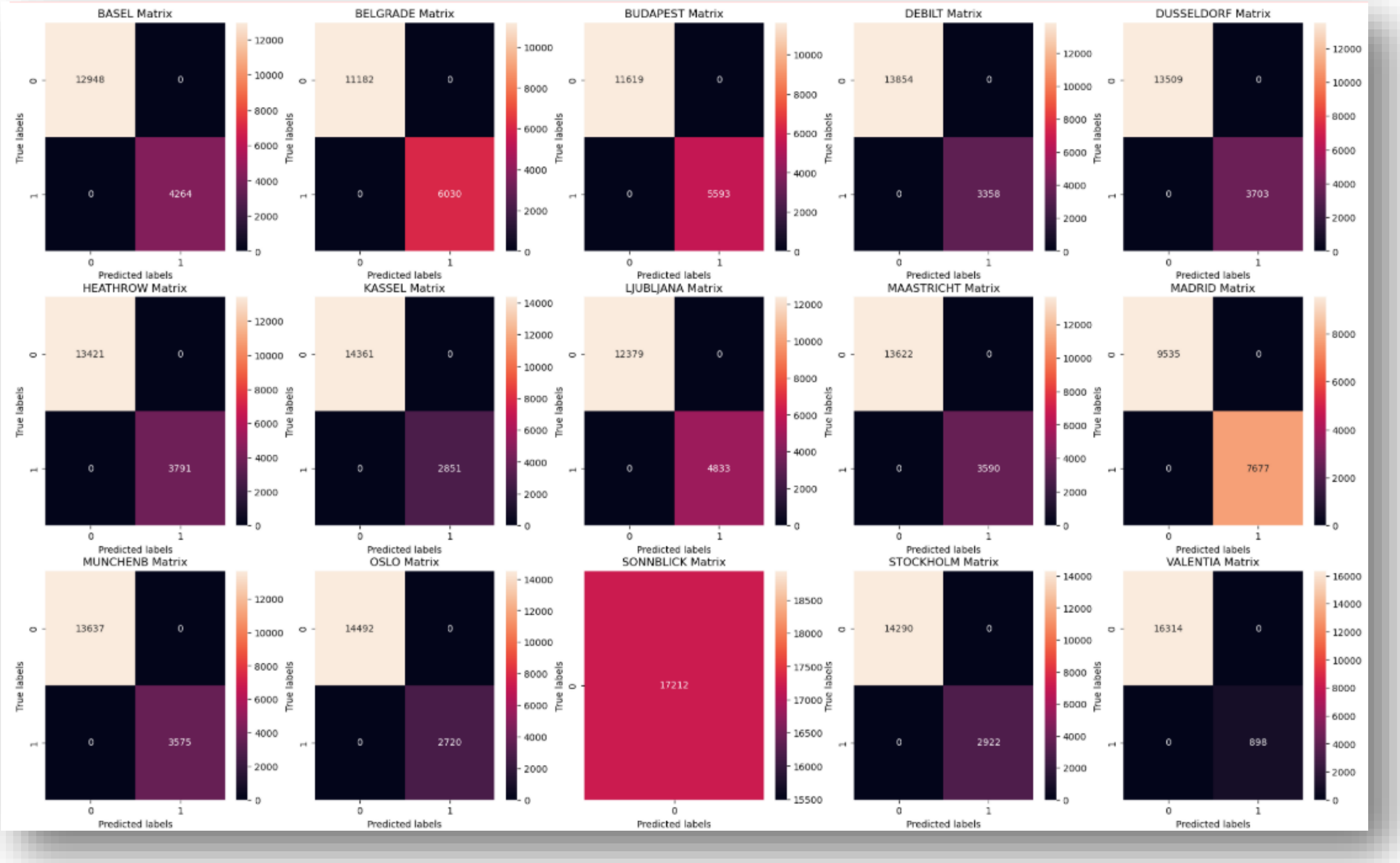
These scores indicate that the Decision Tree is not performing well on the weather dataset. A model with accuracy around 50% on a binary classification problem suggests it is only slightly better than random guessing.

Decision Tree model - Do you think the decision tree needs to be pruned? Why?

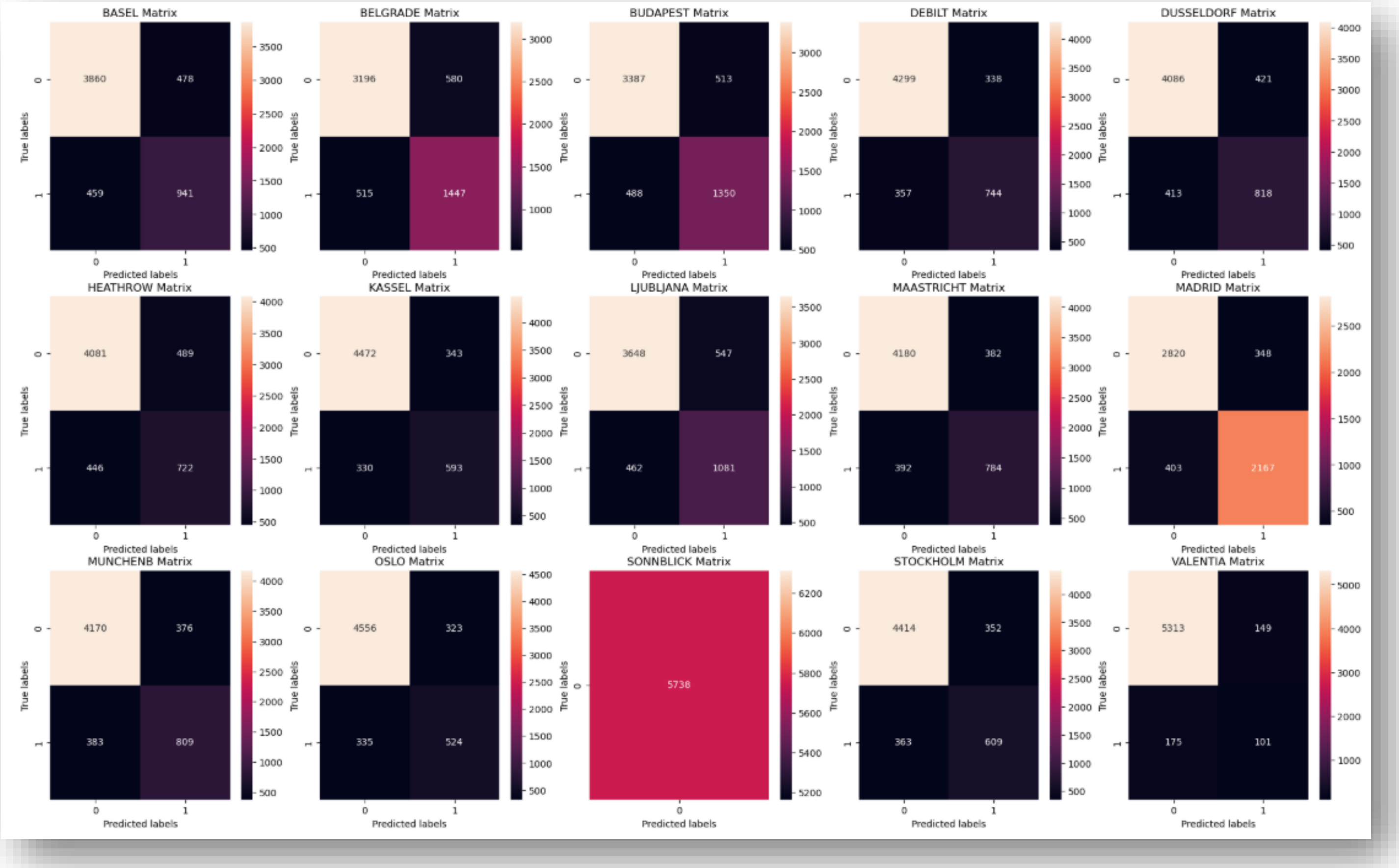
I think the Decision Tree *should* be pruned:

Both training and testing data show large number of misclassifications (i.e. False negatives and False Positives), which suggests overfitting.
Also looking at the training data there are many zeros (i.e. no false positives or false negatives in some cities, e.g. Basel, Belgrade, Budapest etc.), which suggests the tree is memorising pattern instead of learning generalised rules.
The tree may be too complex. Pruning can simplify the tree, making it easier to generalise and focus on the most important decision splits.

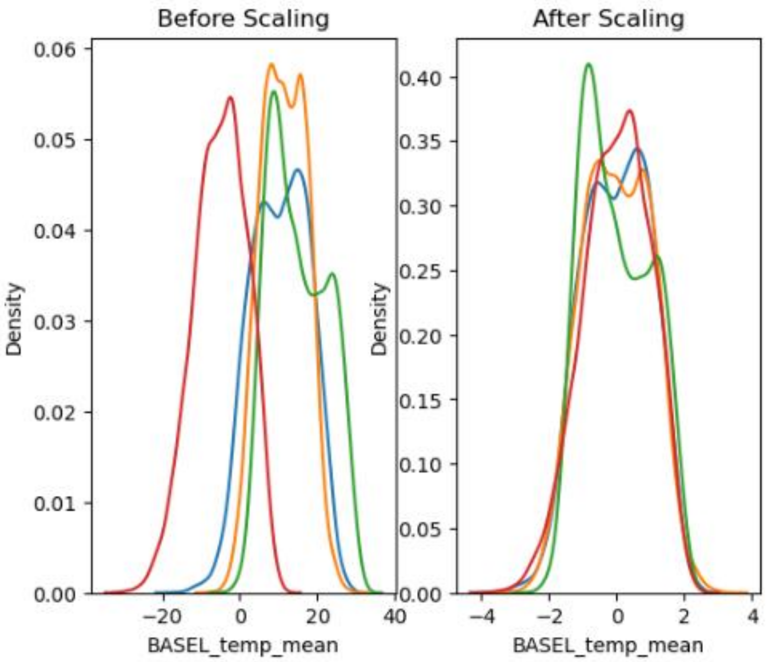
Training data (Decision Tree):



Testing data (Decision Tree):



ANN model - Investigate your unscaled data. Run scaling in this script and decide if scaled data will make a difference. Record your answers



Unscaled Accuracy: 0.4739

Scaled Accuracy: 0.4744

Hence, scaled is very slightly more accurate

ANN model - 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.

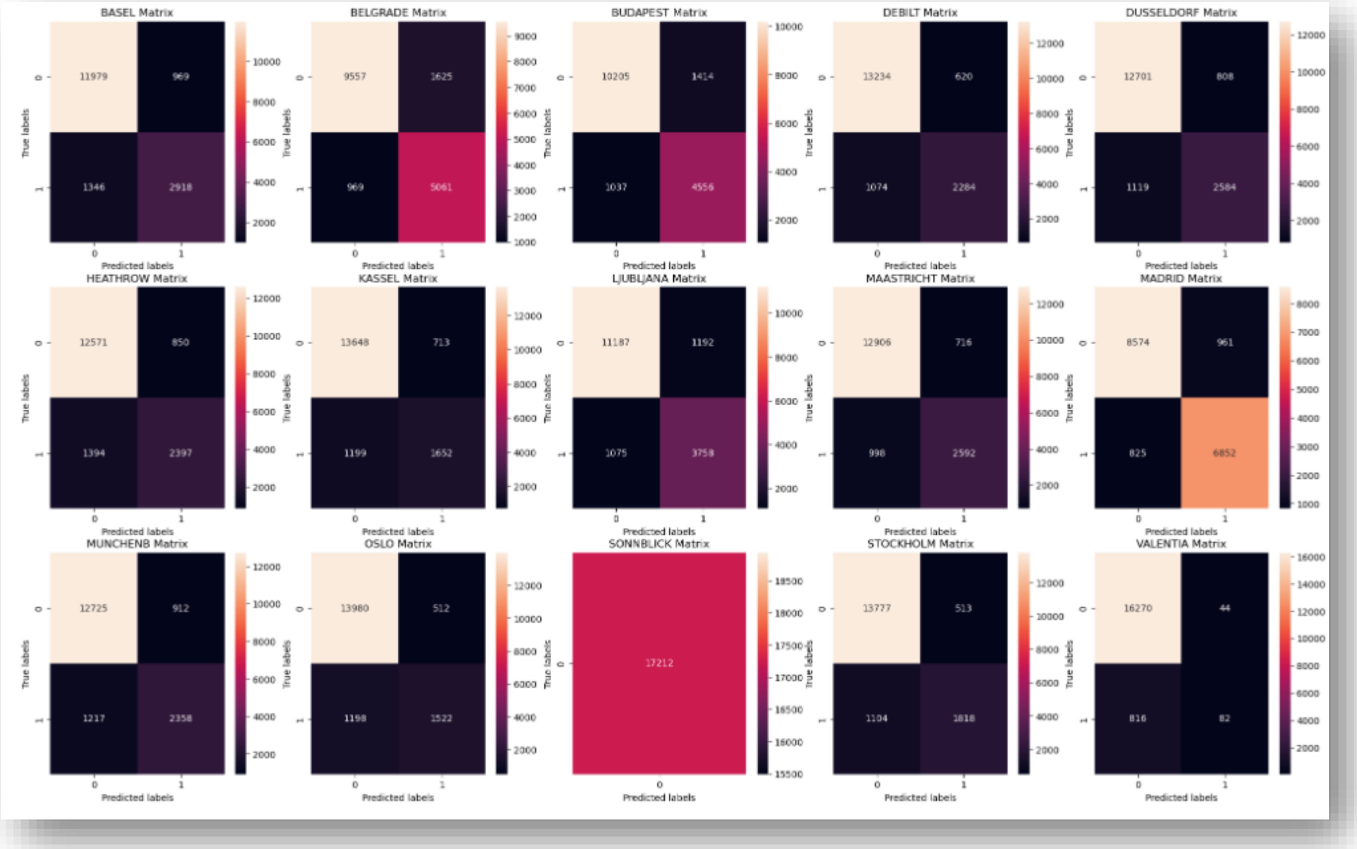
Best accuracy was

- Number of layers: The model with (50,50,60) configuration had the highest accuracy. It achieved 51.2% training accuracy and 49.8% testing accuracy. This suggests that increasing the number of nodes per layer improves accuracy.
- Max iterations: Increasing max iterations to **1000** (as seen in the (5,5) configuration) resulted in **lower accuracy** (41.8% training, 42% testing). The best-performing configuration had **500 iterations**, which suggests that excessive iterations might cause overfitting.
- Tolerance: The model with **tol=0.0001** consistently outperformed models with **tol=0.01**. A lower tolerance value ensures the model continues optimizing longer before stopping, which likely contributed to better performance.

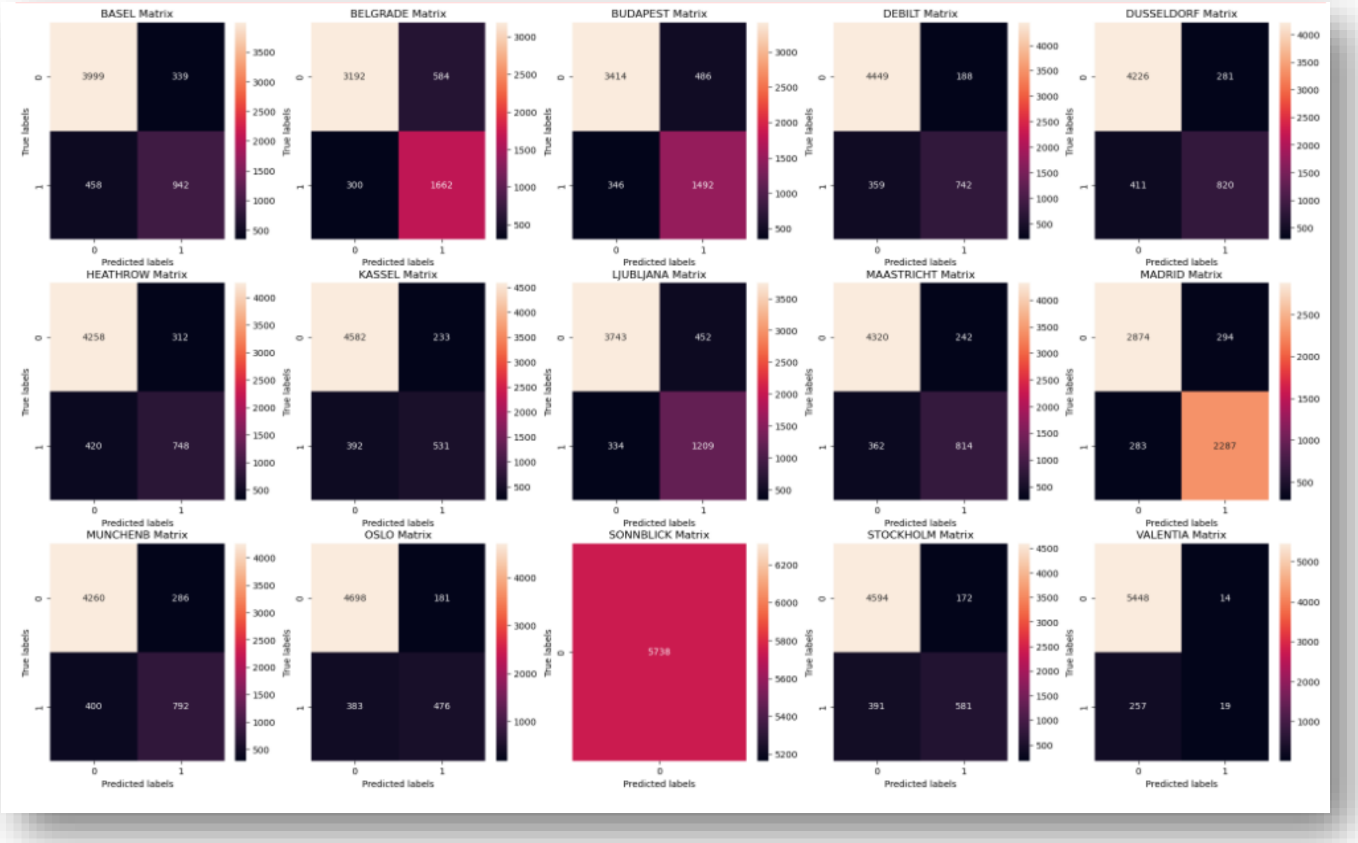
Recording:

Model Configuration	**Training Accuracy**	**Testing Accuracy**
(5,5), max_iter=500, tol=0.0001	**47.3%**	**47.8%**
(10,10), max_iter=500, tol=0.0001	**47.3%**	**47.8%**
(5,5), max_iter=1000, tol=0.0001	**41.8%**	**42.0%**
(5,5), max_iter=500, tol=0.01	**41.0%**	**40.8%**
(50,50,60), max_iter=500, tol=0.0001	**51.2%**	**49.8%**

Testing data (ANN):



Testing data (ANN):



Which of these algorithms (including the KNN model from Exercise 1.4) do you think best predicts the current data?

I would say the KNN model best predicts the current data since it has the highest accuracy (83%). This is significantly higher than both the Decision Tree (46.1% training & 46.11% test accuracy) and ANN models (51.2% training & 49.8% test accuracy).

The ANN is potentially struggling to generalize the weather data, where as the Decision Tree seems to be suffering from overfitting and poor generalization.

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

Decision Tree Model: Sonnblick shows to have a 100% accuracy rate. Some other stations have low false positive or false negative rates.

ANN Model: Sonnblick shows to have a 100% accuracy rate (all predictions are either true positives or true negatives). This suggests overfitting because real-world data is unlikely to have perfect classification without errors.

KNN Model: Sonnblick shows to have a 100% accuracy rate (suggesting overfitting). KNN performs the best overall with an accuracy of 83% across all stations.

Are there certain features of the data set that might contribute to the overall accuracy?

The data set special features such as temperature and daily weather are unpredictable which affects its overall accuracy. Some stations, such as Sonnblick, may have less variable weather patterns (e.g., consistently snowy), making classification easier for the model. Other stations with high weather variability (e.g., Madrid or Oslo) may lead to more misclassifications because patterns are less distinct.

Also, if certain weather stations have more training samples than others, the model might prioritize those stations, leading to biased accuracy. Dusseldorf had 6147 samples, where as Belgrade had 5518.

Which model would you recommend that 'ClimateWins' use?

The KNN model gives the highest accuracy rate, thus I would recommend *ClimateWins* to use KNN model to predict if the weather is suitable for picnic. KNN had the lowest FN and FP rates, making it the best generalizable model and balances accuracy across weather stations.

However, if I spent more time further tuning the ANN model, it could become a viable option in the future. *ClimateWins* could consider using KNN for predicting pleasant weather conditions.