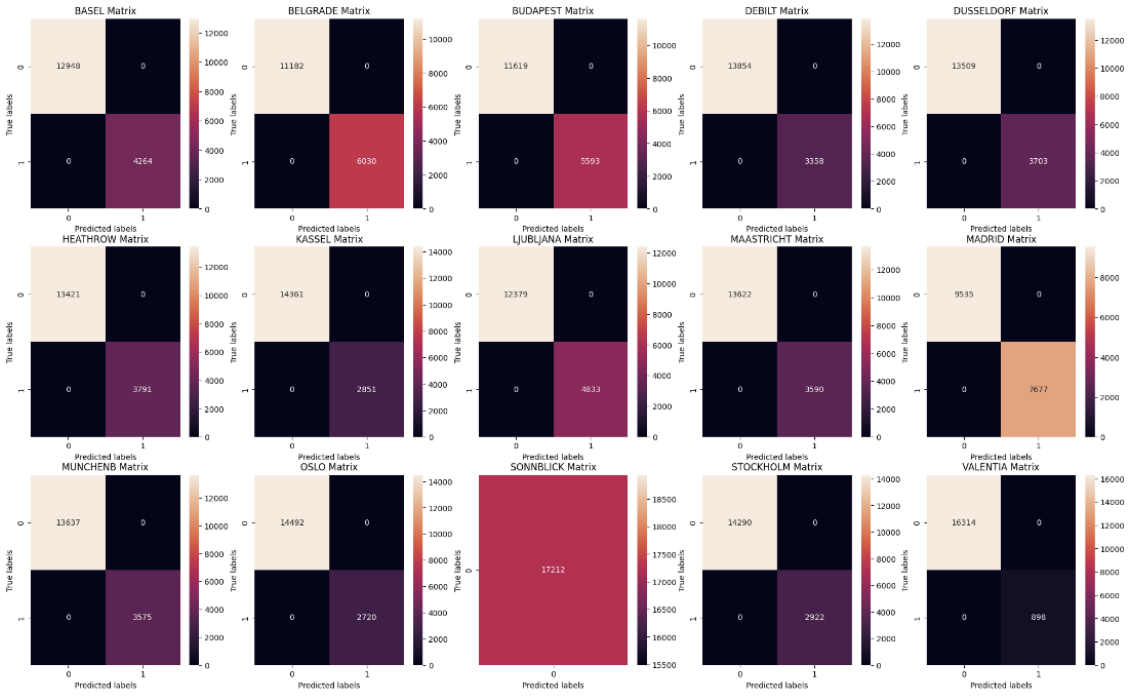
Basics of Machine Learning

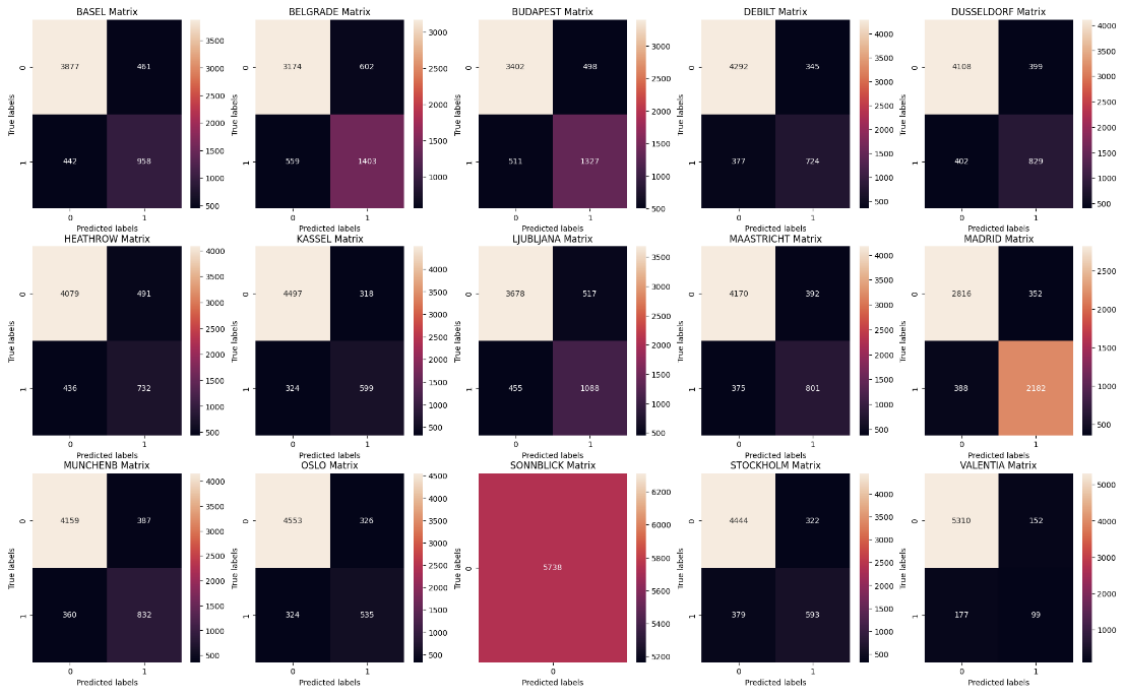
1.5 Supervised Learning Algorithms Part 2

Kendra Jackson

Decision Tree Scaled Data (Train Set)



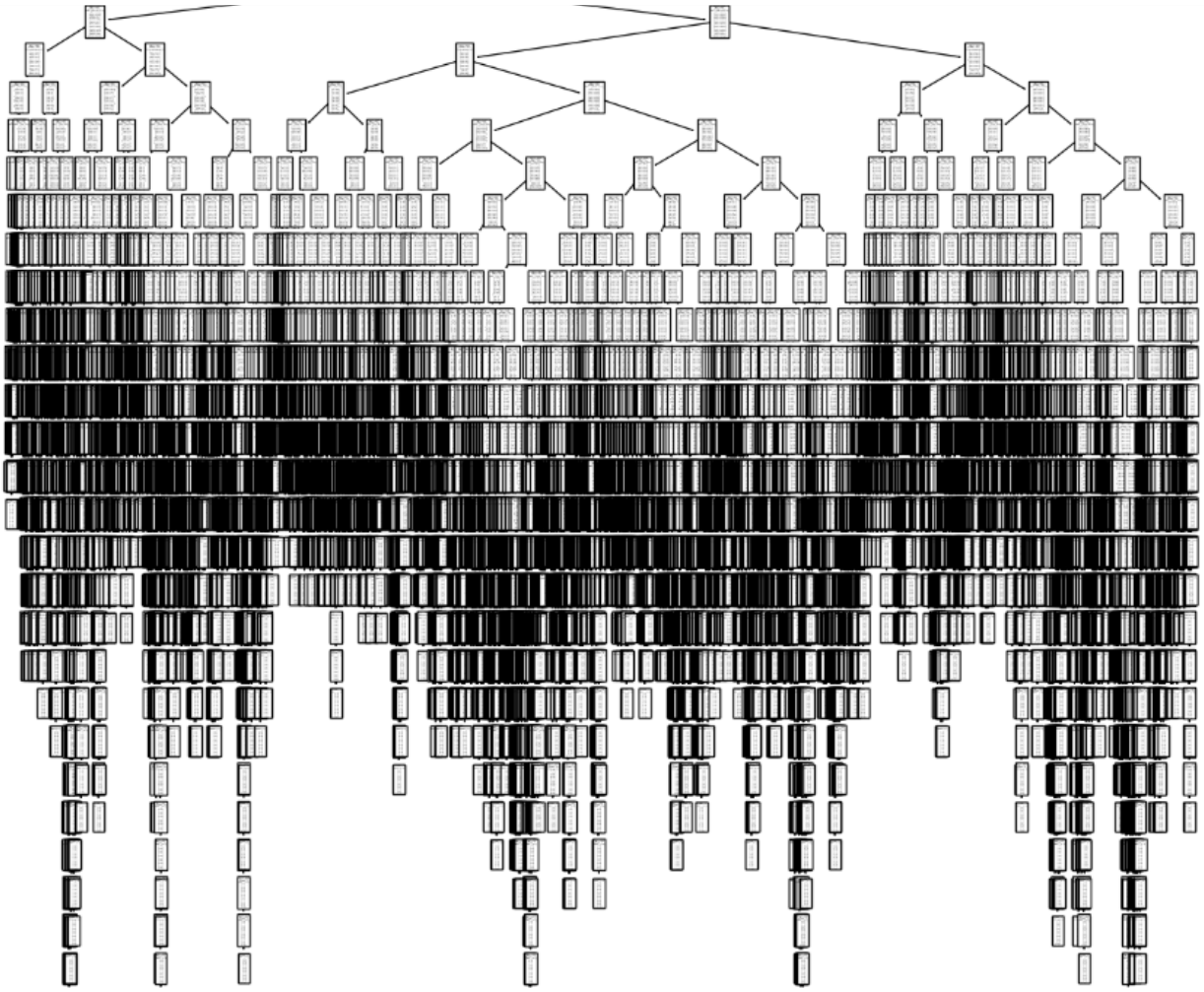
Decision Tree Scaled Data (Test Set)



Train Accuracy: 0.461 → 46.1%

Test Accuracy: 0.471 → 47.1%

The decision tree is very deep/long and complex. It would be better to prune the tree to alleviate some of the complexity and help make the tree more efficient. It is likely that the decision is overfit. We can tell this due to its excessive depth and complexity; it is likely asking too many specific questions. It is also difficult to interpret.



Neural Network Model

Best Model (Determined by highest accuracy)

Layer Count: 2

Layer Size: 70, 50 (larger yields better results)

Maximum iteration: 1000 (smaller (500) yields worse results)

Tolerance: 0.0001 (increasing tolerance does not have much change to accuracy)

Accuracy: 52.4%/49.8%

Trial 1

Layer Count: 2

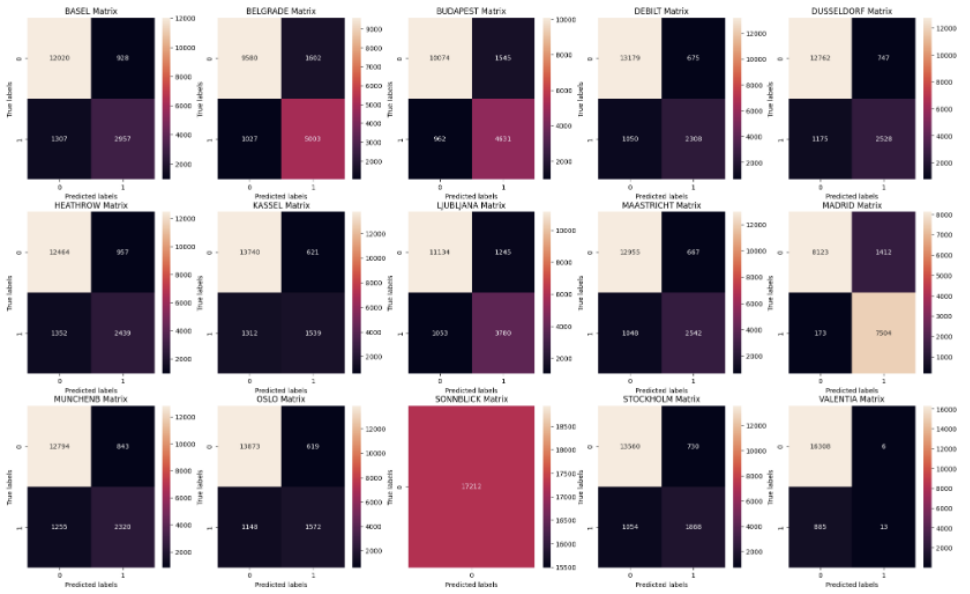
Layer Size: 5, 5

Maximum Iterations: 500

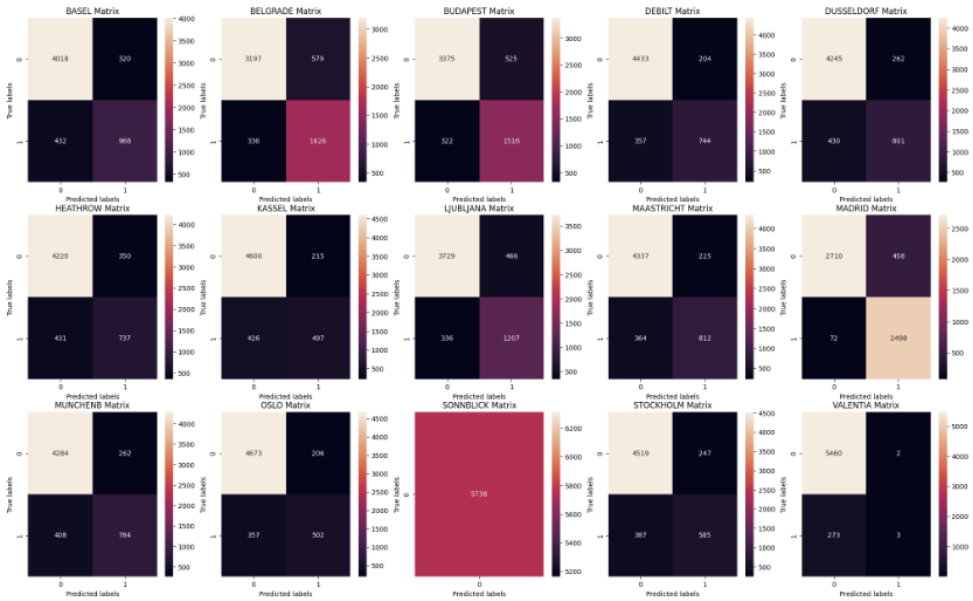
Tolerance 0.0001

Accuracy: 45.6%/46.1%

Training Set



Testing Set



Trial 2

Layer Count: 2

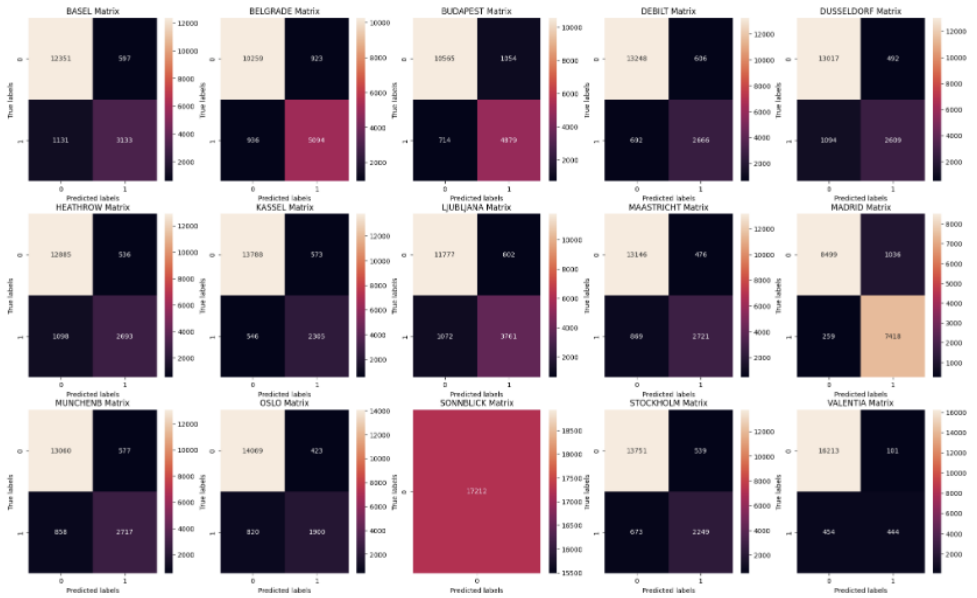
Layer Size: 70, 50

Maximum Iterations: 1000

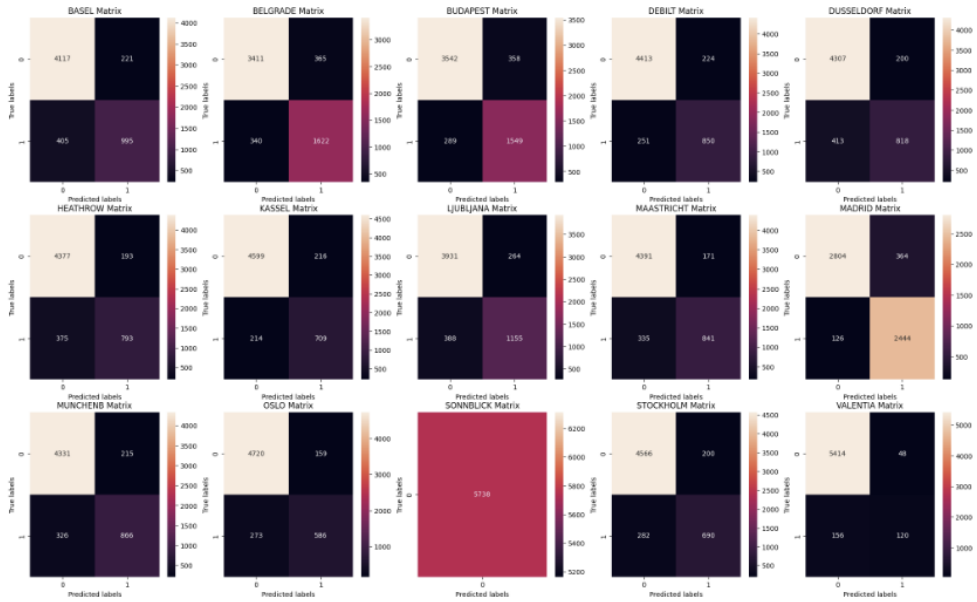
Tolerance 0.0001

Accuracy: 52.4%/49.8%

Training Set



Testing Set



Trial 3

Layer Count: 3

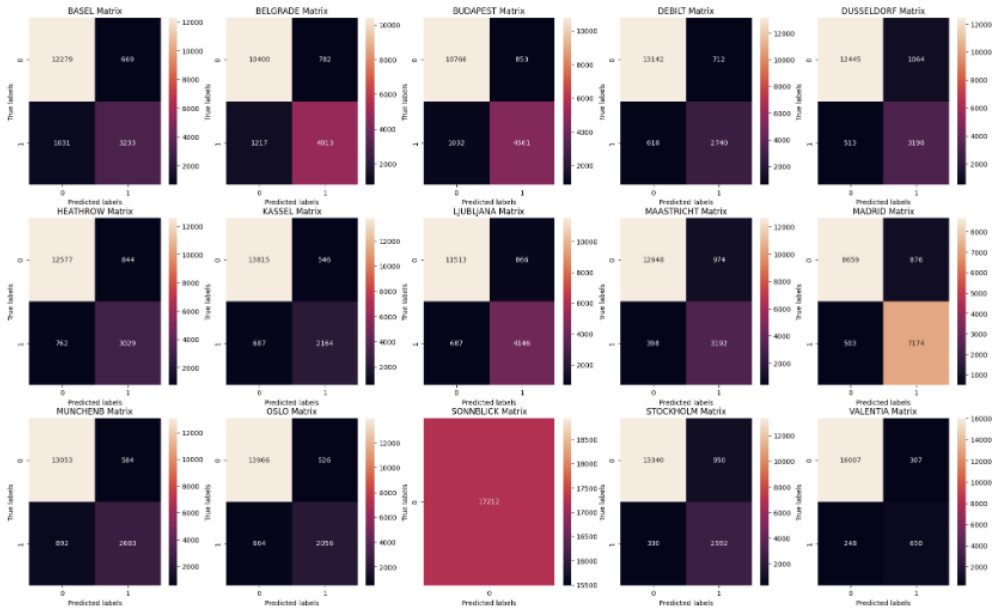
Layer Size: 70,50,50

Maximum Iterations: 1000

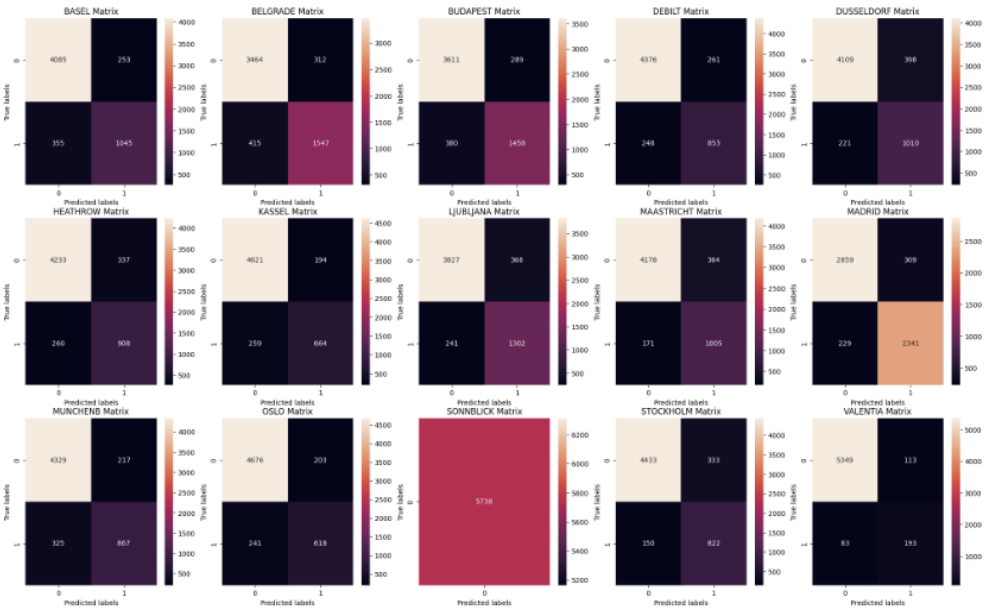
Tolerance 0.0003

Accuracy: 51.9%/49.7%

Training Set



Testing Set



Summary:

After trialing different types of predictive models, it appears that the KNN model from the previous exercise is the model ClimateWins should use. Of the 3 models tested, the KNN model yields the highest accuracy rate and predicts the data the most accurately with average accuracy rates sitting between 88-94%. An accurate model with the ability to predict pleasant daily weather conditions aligns with ClimateWins' goals making the KNN model the appropriate selection. Comparatively, the models trialed within this exercise yield low accuracy on their prediction scores sitting in the high 40's and low 50's.

While the models show 100% accuracy within prediction for Sonnblick, caution must be used as this is indicative of overfitting. While we are seeing this consistent 100% accuracy rate for Sonnblick's prediction of unpleasant weather, there is no variation to the data for this weather station, such as pleasant weather, for the models to train on. This means that for this station, our model is not well-rounded and under different circumstances, like different data, the model may fail. In order to heighten the accuracy of our models, it is imperative that varied data be used to ensure there is no overfitting and that our model should be able to perform well under different conditions.