Amy Zhang

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

Analyzing all the confusion matrices simultaneously proves quite challenging. It's immediately clear, however, that KNN struggles significantly, unsurprising given that weather data is inherently non-linear and multi-dimensional. While Decision Trees, with further pruning, seem poised to be an optimal solution—their binary decision-making perfectly aligns with classifying 'pleasant' versus 'unpleasant' conditions—the intricate interplay of numerous weather features presents a compelling case for Artificial Neural Networks. An ANN could potentially capture the complex relationships between variables like temperature, pressure, and humidity in ways a decision tree might not.

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

One station—Sonnblick—shows what looks like perfect accuracy. But upon closer inspection, this is misleading. The confusion matrix reveals that the model predicts only one class (unpleasant) for that station, which likely reflects class imbalance rather than genuine predictive accuracy. So in reality, no station is fully accurate, and yes, there are definite signs of overfitting, especially with unpruned decision trees.

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

Yes. Several factors likely impact the models' ability to predict accurately. **Temperature features** (mean, min, max) are continuous and rich; these likely contribute significantly. **No explicit time or seasonal features** are present, which limits the models' ability to capture weather patterns that follow cyclical trends (e.g., warmer summers). **Class imbalance** (stations with mostly unpleasant days) can bias the models toward always predicting one class, skewing results. Including additional variables like **humidity and pressure**, could further boost performance.

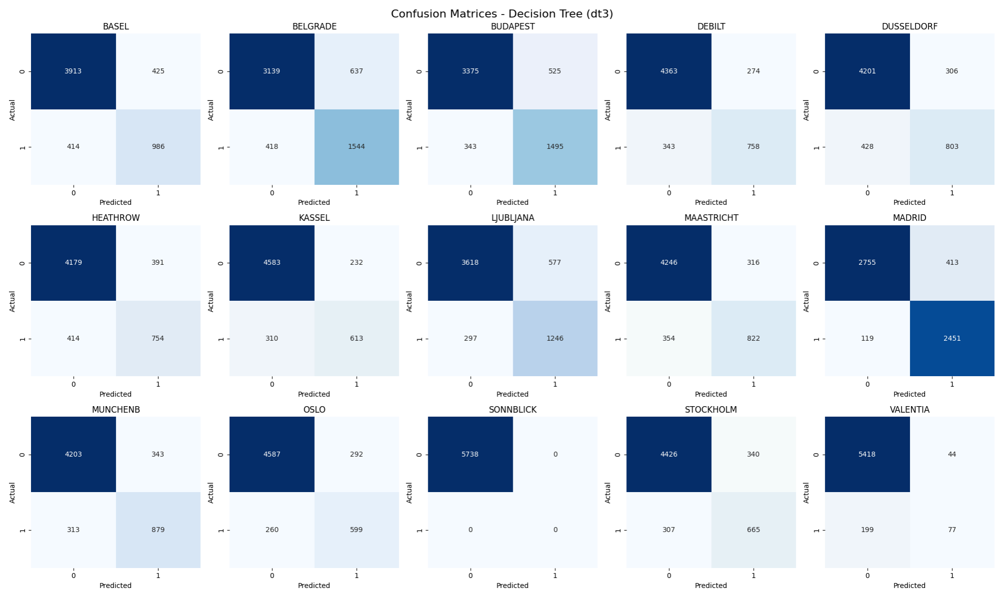
Which model would you recommend that ClimateWins use?

For the current stage of development, I would recommend:

- **ANN (MLP)**: for best overall accuracy and ability to model complex, nonlinear patterns in weather data.
- **Pruned Decision Tree**: as a strong backup option when interpretability is important—especially if explaining predictions to stakeholders is a goal.
- **Random Forest or Ensemble Models**: could also be explored to combine robustness with moderate interpretability.

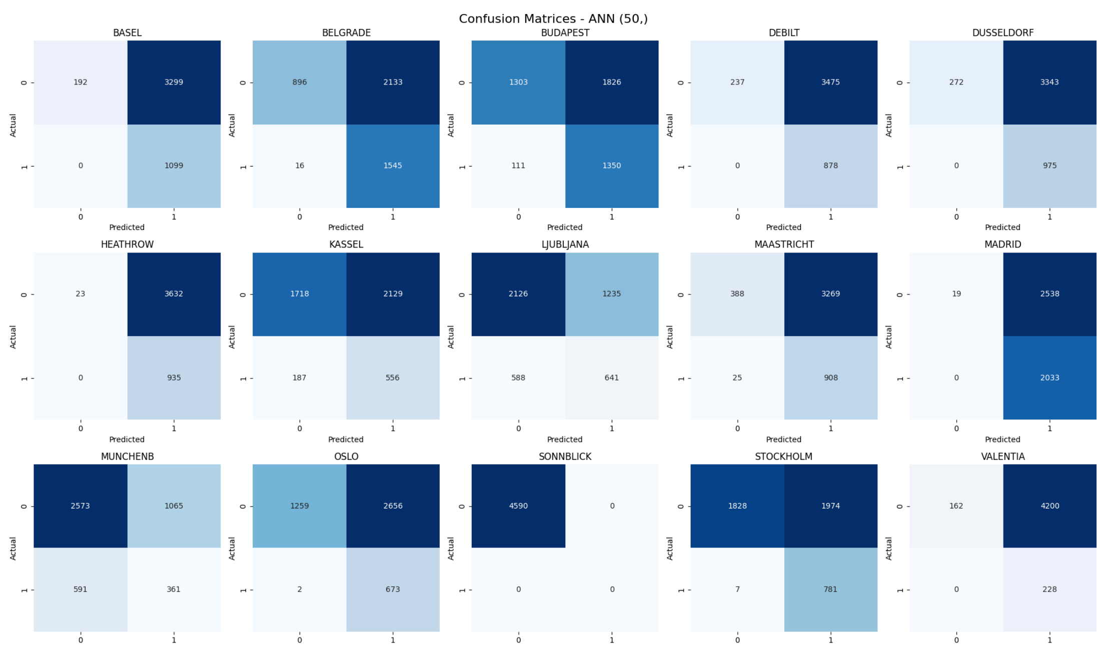
Amy Zhang

1.5 Decision Tree Model



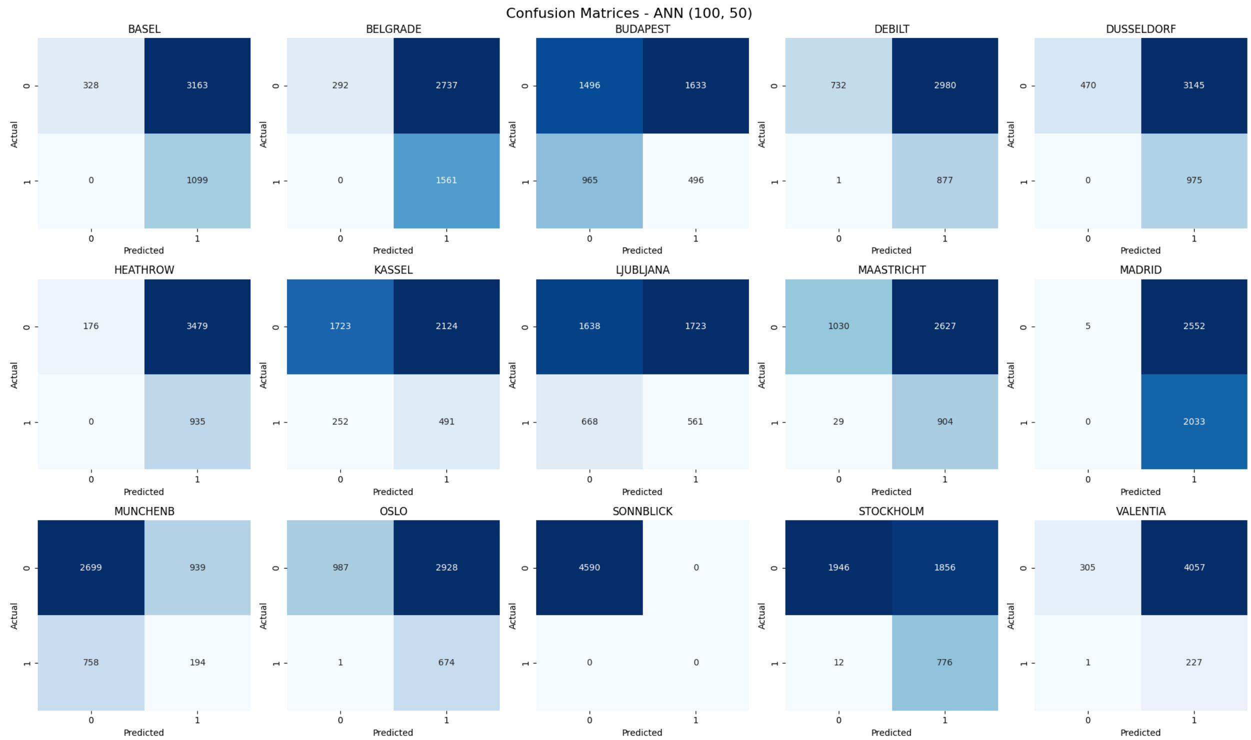
1.5 Artificial Neural Network

(50,)

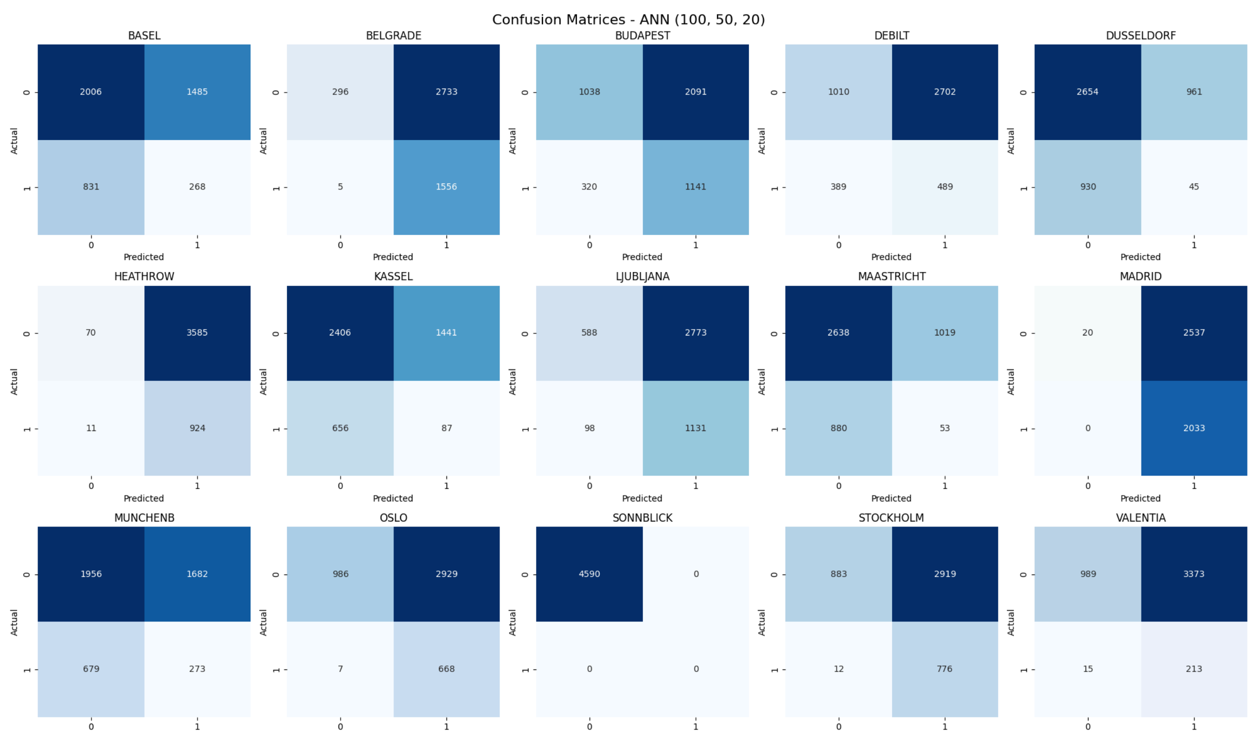


Amy Zhang

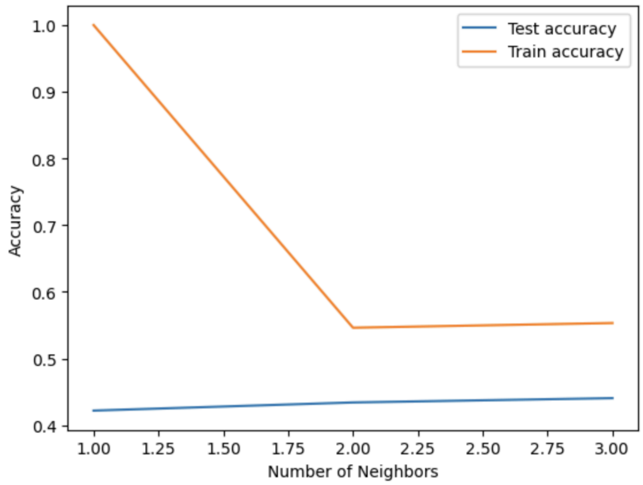
(100, 50)



(100, 50, 20)



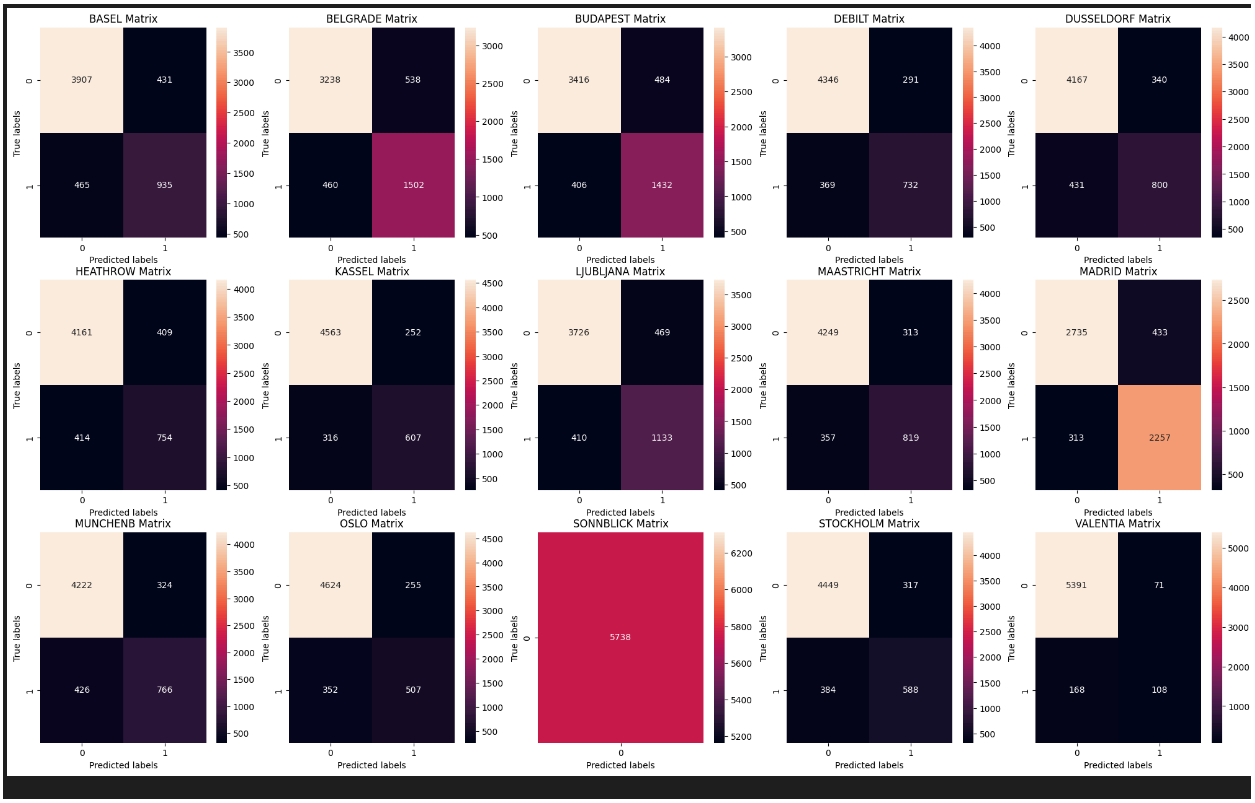
1.4 K-Nearest Neighbors Model



Conclusions:

It appears the model consistently performs better at identifying unpleasant days than pleasant ones. Across all stations, there are notable counts of both false positives and false negatives — no weather station achieves high accuracy. SONNBLICK stands out as potentially problematic, predicting only a single class, which suggests either data quality issues or extreme class imbalance. While overfitting may be a concern, especially given the steep drop in accuracy between 1 and 2 neighbors, the final model used k = 3, which already showed signs of degraded performance. The unreliable predictions from SONNBLICK likely skew the overall accuracy and may be masking the true behavior of the model on more balanced stations.

Amy Zhang


BASEL Matrix
3907 431 465 935
BELGRADE Matrix
3238 538 460 1502
BUDAPEST Matrix
3416 484 406 1432
DEBILT Matrix
4346 291 369 732
DUSSELDORF Matrix
4167 340 431 800
HEATHROW Matrix
4161 409 414 754
KASSEL Matrix
4563 252 316 607
LJUBLJANA Matrix
3726 469 410 1133
MAASTRICHT Matrix
4249 313 357 819
MADRID Matrix
2735 433 313 2257
MUNCHENB Matrix
4222 324 426 766
OSLO Matrix
4624 255 352 507
SONNBLICK Matrix
5738
STOCKHOLM Matrix
4449 317 384 588
VALENTIA Matrix
5391 71 168 108
True labels
Predicted labels