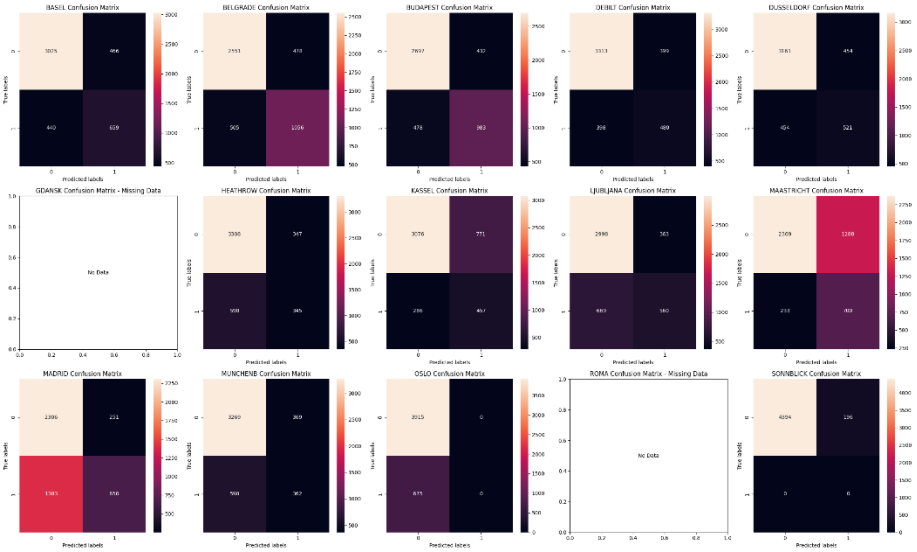
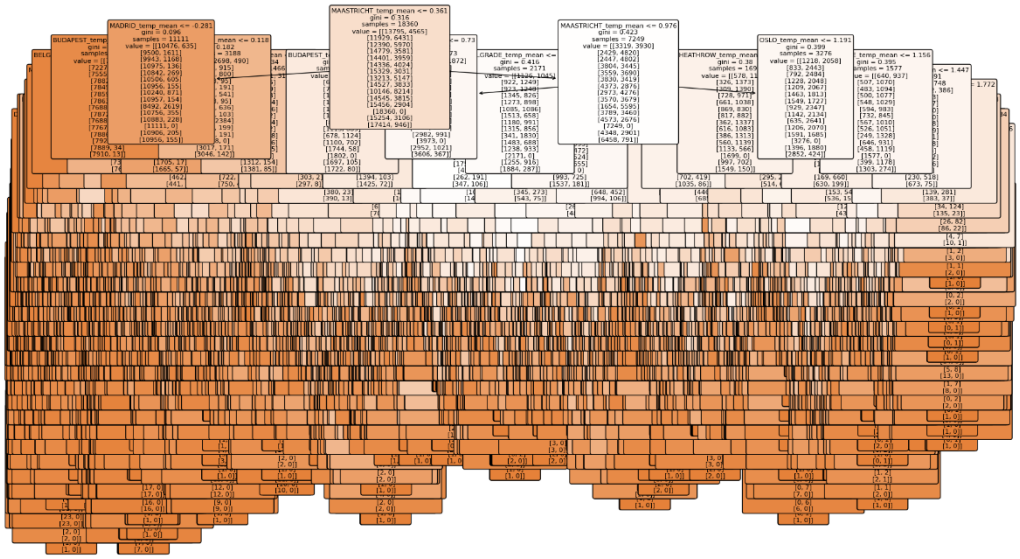
Luis Gil

Decision Tree Model

Confusion Matrix



Decision Tree



Luis Gil

Accuracy:

- **Training Accuracy**: 1.0 (Indicates overfitting as the model perfectly classifies the training data)
- **Testing Accuracy**: 0.41 (Shows the model struggles to generalize well to unseen data)

Decision Tree Pruning:

Yes, the decision tree **definitely needs to be pruned**. The training accuracy of 1.0 is a clear indicator of overfitting, which means the tree has memorized the training data instead of learning generalized patterns. Overfitting results in low testing accuracy (0.41), which shows the model doesn't perform well on unseen data. Pruning can help to simplify the decision tree, removing unnecessary complexity and improving its generalization ability on testing data.

Scaling¶

- After scaling, the mean values of all features become very close to zero, and the standard deviations are normalized to 1. This indicates that scaling has effectively standardized the data, which is often crucial for improving the performance of machine learning algorithms like neural networks.

Initial Neural Network (ANN) Results

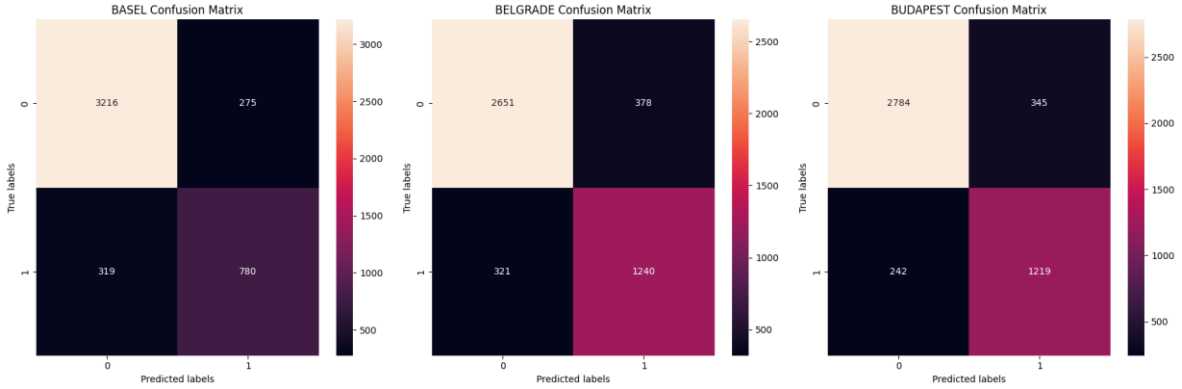
- With hidden_layer_sizes=(100,), max_iter=500, the ANN model achieved:
  - A testing accuracy of approximately **74%**.
  - Some stations have higher precision, such as **station 9** (precision of **0.85**) and **station 2** (precision of **0.75**).
  - Some stations, however, show lower precision and recall, such as **station 14** (precision of **0.55** and recall of **0.19**).
  - **Station 12** has no support, meaning the data might be missing or incomplete for this station, causing it to perform poorly.

Luis Gil

Testing New Parameters

- By adjusting the parameters with hidden_layer_sizes=(100, 50), max_iter=1000, and a lower tolerance of tol=0.0001, there are slight improvements:
    - Overall performance remains similar with a testing accuracy of **74%**.
    - Precision and recall improved for some stations but decreased slightly for others, suggesting that the model is fairly stable with these parameters.

Confusion Matrices for BASEL, BELGRADE, and BUDAPEST:



Key Observations:

- **BASEL:**
    - Predicted "pleasant" weather correctly in **3216** cases, but **275** times incorrectly classified as "unpleasant."
    - Misclassified **319** true "unpleasant" weather cases as "pleasant" and correctly identified **780** cases.
- **BELGRADE:**
    - Predicted "pleasant" weather correctly in **2651** cases, with **378** false negatives.
    - Correctly identified **1240** cases of "unpleasant" weather but misclassified **321** times.

- **BUDAPEST:**
    - **2784** cases of "pleasant" weather were predicted correctly, with **345** misclassifications.
    - **1219** cases of "unpleasant" weather were predicted correctly, while **242** were classified incorrectly.

The ANN model seems to perform relatively well, but as seen from the confusion matrices, there is room for improvement, particularly in reducing the number of misclassifications.

Conclusion

Among the algorithms tested, the Artificial Neural Network (ANN) model appears to best predict the current data with a testing accuracy of 74%, offering better precision for certain weather stations like BASEL, BELGRADE, and BUDAPEST compared to the KNN and Decision Tree models. None of the stations are fully accurate, but some, like SONNBLICK and BUDAPEST, exhibit high precision and recall, while others, like MAASTRICHT and STOCKHOLM, show more misclassifications. Overfitting is most evident in the Decision Tree model, where the training accuracy is 100%, indicating that the model is too complex. Features like temperature variability and precipitation patterns seem to contribute significantly to model accuracy, with clearer patterns leading to better performance for certain stations. Overall, I recommend ClimateWins use the ANN model, as it provides the best balance between accuracy and generalization without significant overfitting.