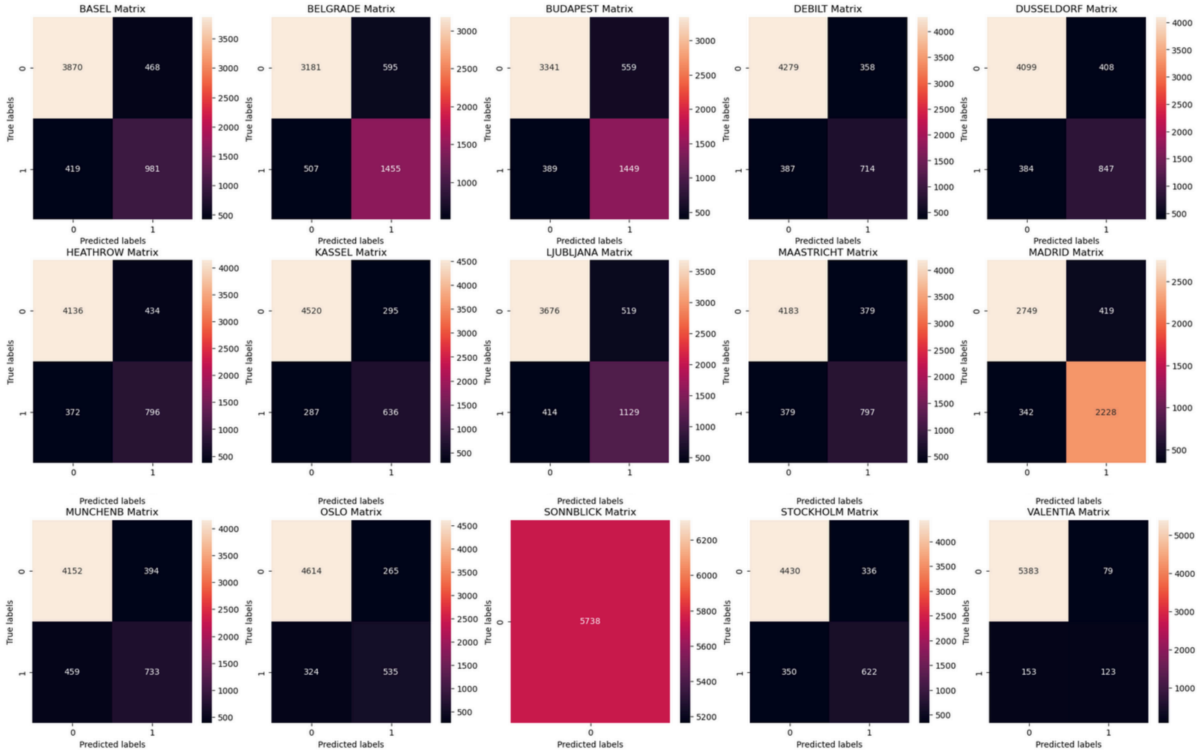
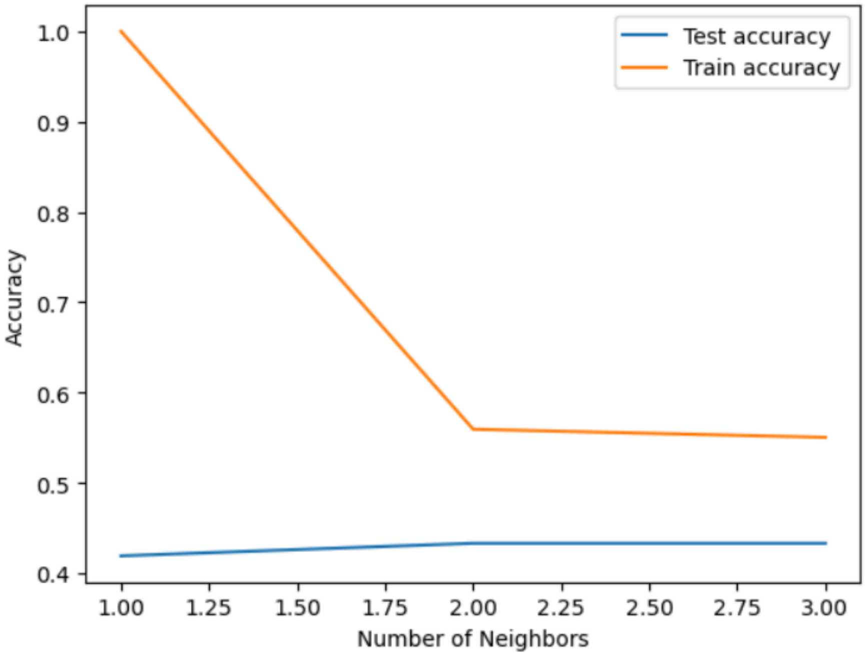
Screenshots





Accuracy

Weather Station	Accurate Predictions (TP + TN)	False Positives	False Negatives	Accuracy Rate
BASEL	3870 + 981 = 4851	468	419	84.5%
BELGRADE	3181 + 1455 = 4636	595	507	81.5%
BUDAPEST	3341 + 1449 = 4790	559	389	83.5%
DEBILT	4279 + 714 = 4993	358	387	86.9%
DUSSELDORF	4099 + 847 = 4946	408	384	86.2%
HEATHROW	4136 + 796 = 4932	434	372	86.0%
KASSEL	4520 + 636 = 5156	295	287	89.9%
LJUBLJANA	3676 + 1129 = 4805	519	414	83.7%
MAASTRICHT	4183 + 797 = 4980	379	379	86.8%
MADRID	2749 + 2228 = 4977	419	342	86.7%
MUNCHENBG	4152 + 733 = 4885	394	459	85.1%
OSLO	4614 + 535 = 5149	265	324	89.7%
SONNBLICK	5738	0	0	100%
STOCKHOLM	4430 + 622 = 5052	336	350	87.9%
VALENTIA	5383 + 123 = 5506	79	153	95.9%

Questions

- How well does this algorithm predict the current data?

  The overall accuracy is 86.71%, which is a good value. The accuracy varies between stations.

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

  Sonnblick is fully accurate, as there seems to be only one value, which the model always predicts. There is overfitting happening.

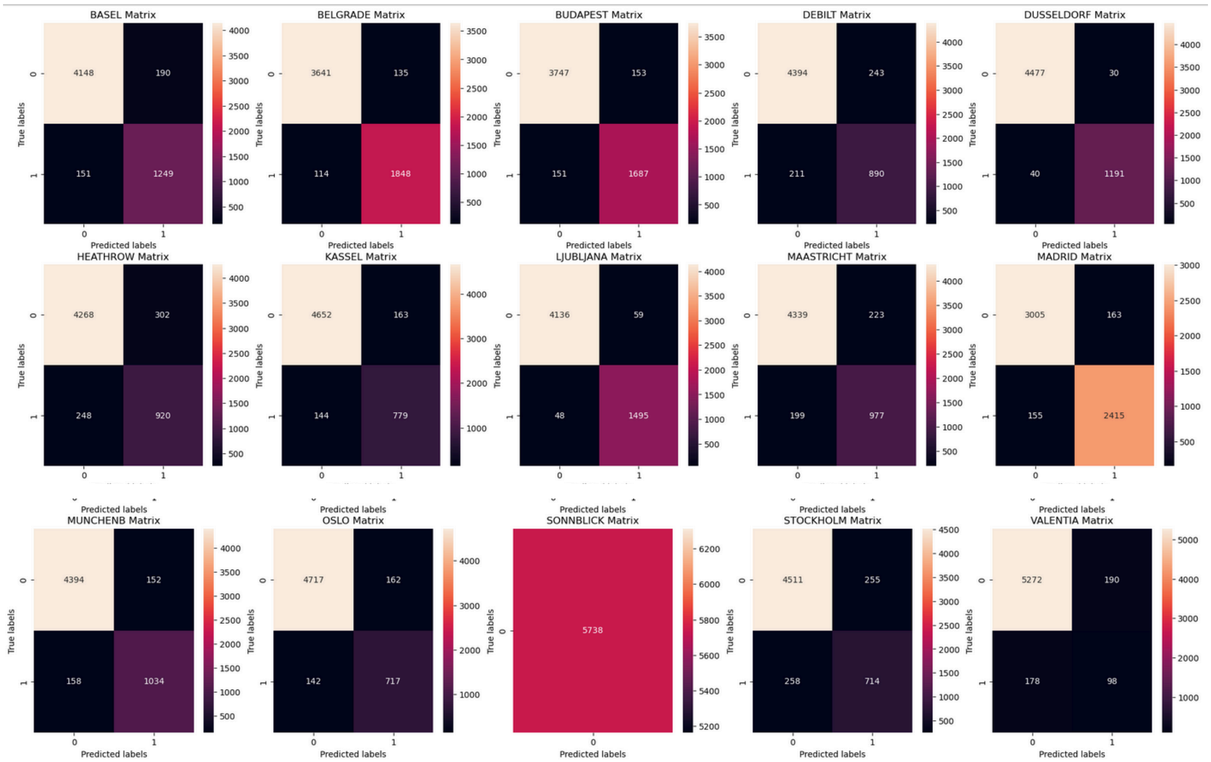
- Are there certain features of the data set (such as particular weather stations) that might contribute to the overall accuracy or inaccuracy?

  Sonnblick might artificially inflate the accuracy.

Task 1.5

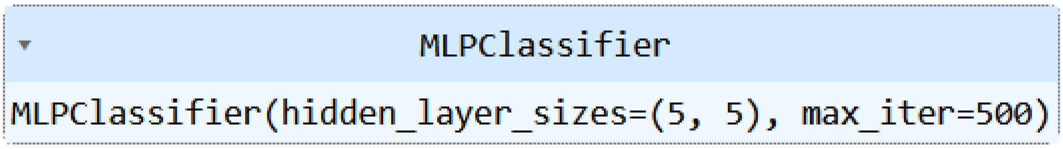
Decision Trees

- The decision tree should be pruned, leaving the main generated branches and leaving out those that are unimportant to the decision. Right now it has too many branches which makes it hard to read, take long to run and most importantly at risk of being overfitted.
- Screenshot of the confusion matrix decision tree:

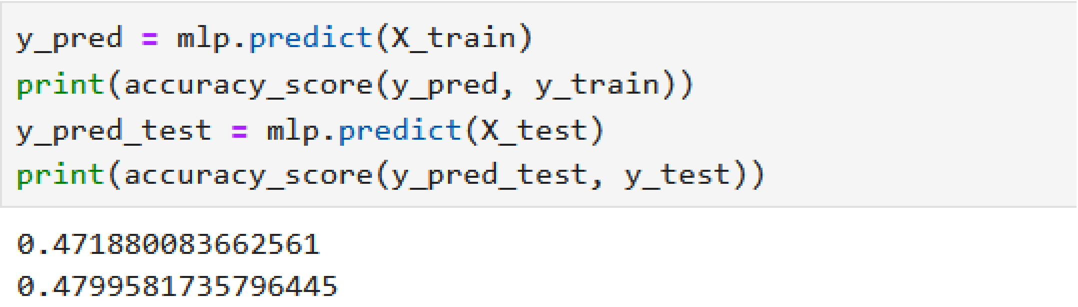


Artificial neural network

Scenario 1:

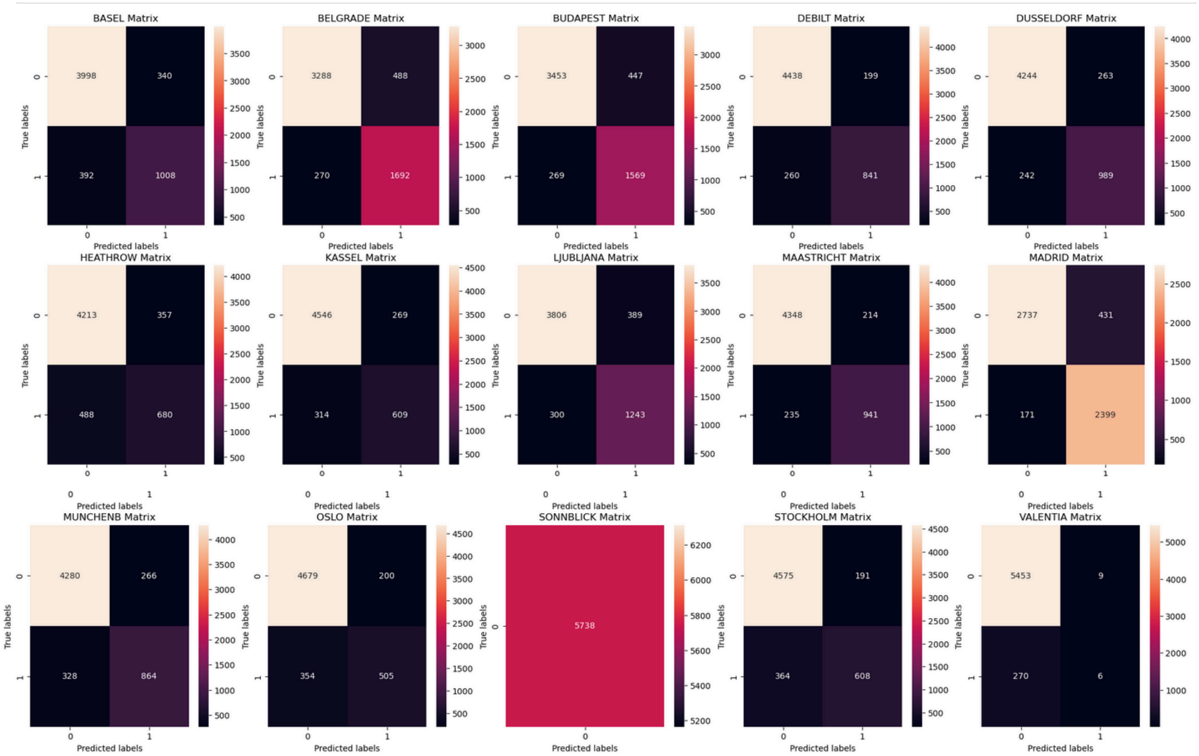


```
MLPClassifier
MLPClassifier(hidden_layer_sizes=(5, 5), max_iter=500)
```

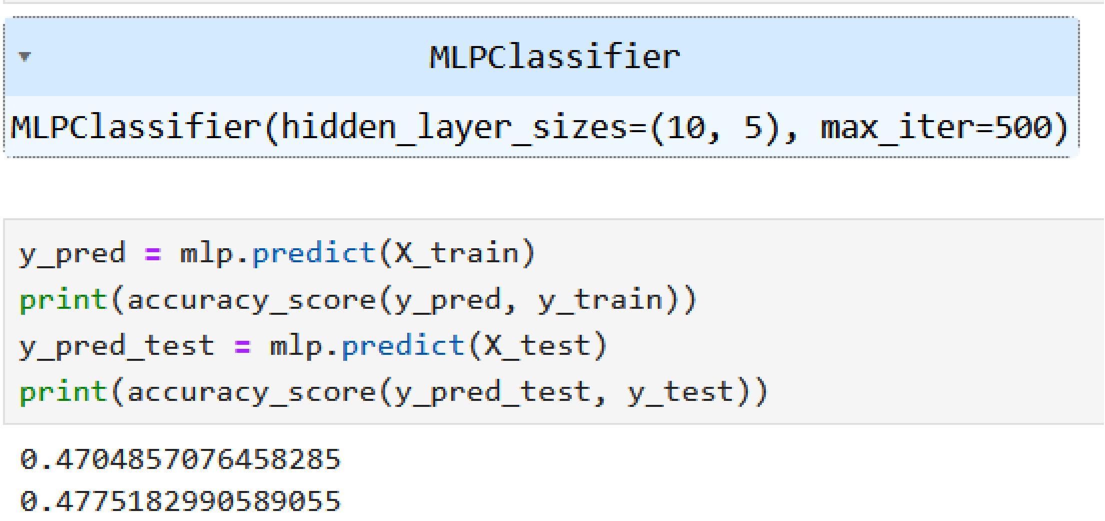


```
y_pred = mlp.predict(X_train)
print(accuracy_score(y_pred, y_train))
y_pred_test = mlp.predict(X_test)
print(accuracy_score(y_pred_test, y_test))
```

```
0.471880083662561
0.4799581735796445
```


BASEL Matrix
3998
340
392
1008
BELGRADE Matrix
3288
488
270
1692
BUDAPEST Matrix
3453
447
269
1569
DEBILT Matrix
4438
199
260
841
DUSSELDORF Matrix
4244
263
242
989
HEATHROW Matrix
4213
357
488
680
KASSEL Matrix
4546
269
314
609
LJUBLJANA Matrix
3806
389
300
1243
MAASTRICHT Matrix
4348
214
235
941
MADRID Matrix
2737
431
171
2399
MUNCHENB Matrix
4280
266
328
864
OSLO Matrix
4679
200
354
505
SONNBLICK Matrix
5738
STOCKHOLM Matrix
4575
191
364
608
VALENTIA Matrix
5453
9
270
6
True labels
Predicted labels

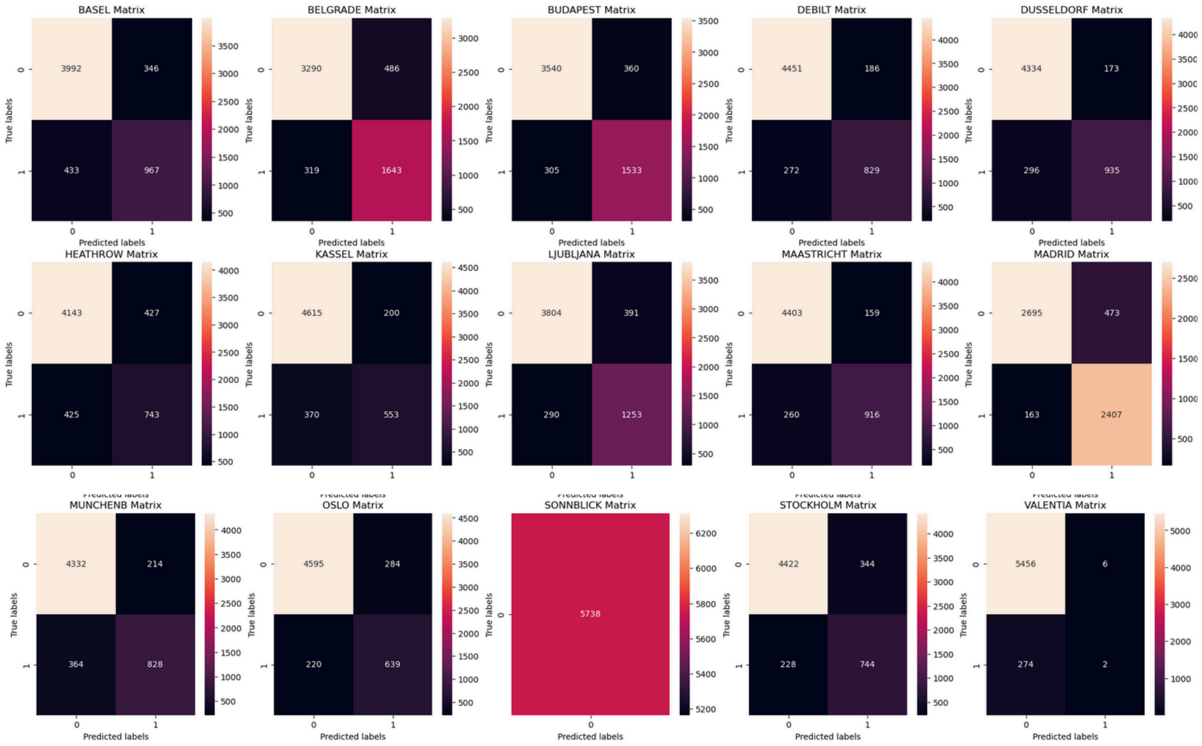
Scenario 2:



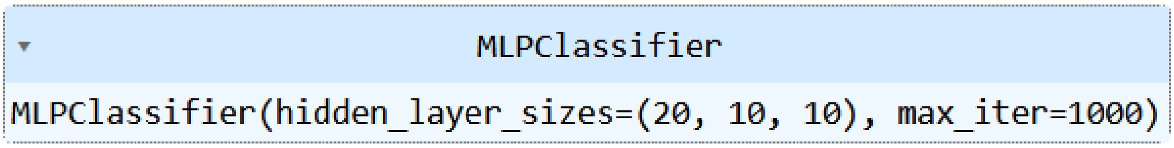
```
MLPClassifier
MLPClassifier(hidden_layer_sizes=(10, 5), max_iter=500)
```

```
y_pred = mlp.predict(X_train)
print(accuracy_score(y_pred, y_train))
y_pred_test = mlp.predict(X_test)
print(accuracy_score(y_pred_test, y_test))
```

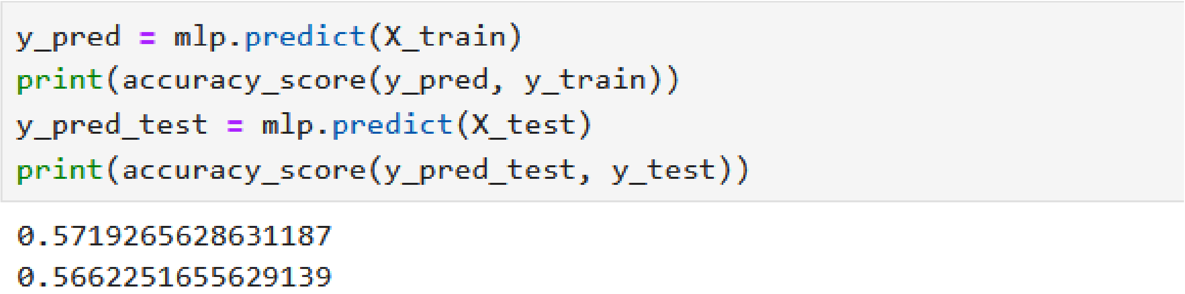
```
0.4704857076458285
0.4775182990589055
```



Scenario 3:

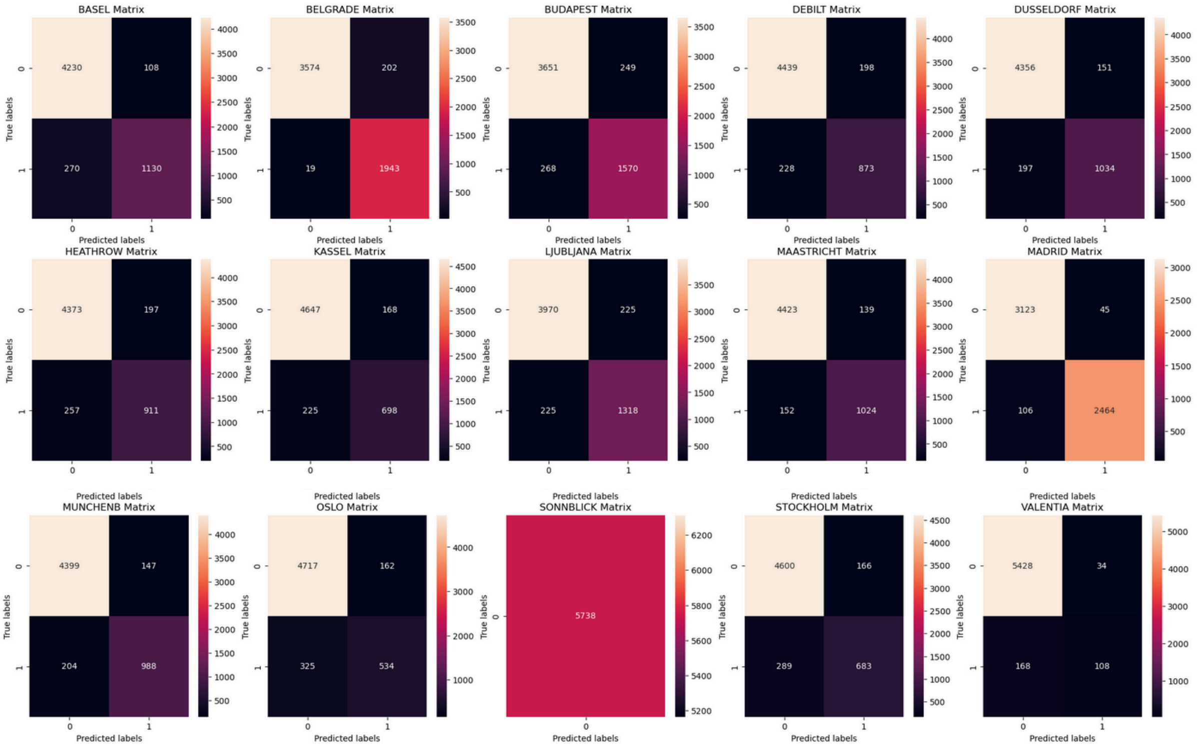


```
MLPClassifier
MLPClassifier(hidden_layer_sizes=(20, 10, 10), max_iter=1000)
```



```
y_pred = mlp.predict(X_train)
print(accuracy_score(y_pred, y_train))
y_pred_test = mlp.predict(X_test)
print(accuracy_score(y_pred_test, y_test))
```

```
0.5719265628631187
0.5662251655629139
```



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

Based on the accuracy metrics, the initial model (KNN from exercise 1.4) with 86.71% accuracy appears to significantly outperform the neural network models and the KNN model in terms of overall prediction accuracy.

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

Sonnblick is fully accurate and overfitted in all models, this is because there is only one value, which the models always predict for it.

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

The wide range in accuracy between stations (81.5% to 95.9%) suggests geographic or climate-specific features are influencing predictability.

Which model would you recommend that ClimateWins use?

I would recommend our initial KNN model from task 1.4 but remove Sonnblick from the data.