CareerFoundry ML with Python
November 25, 2025
Charlotte Lin

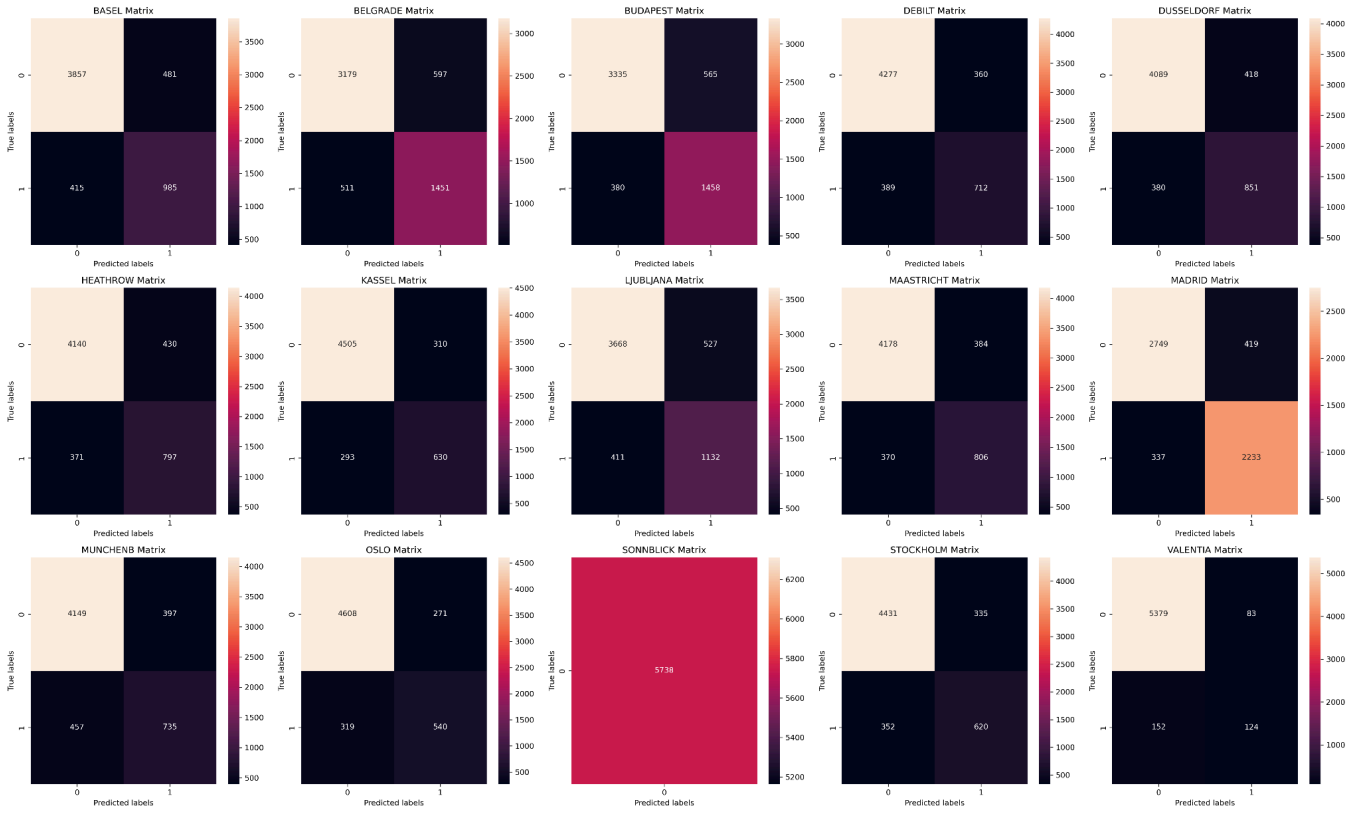
Achievement 1.5:
Supervised Learning Algorithms Part 2

Pleasant Weather Prediction

This exercise evaluates how well supervised learning models can predict “pleasant” weather across multiple European stations. Continuing from the KNN model programmed in Exercise 1.4, I proceeded with Decision Trees, and Artificial Neural Networks (ANN). Performance was assessed primarily on the test set through macro-F1 scores and station-level confusion matrices, since the data are imbalanced and vary by location. The aim was to identify the strongest model while understanding how algorithm design, pruning, and hyperparameter choices influence predictive accuracy.

KNN Model

k	Train F1	Test F1
1	1.00	0.81
2	0.90	0.77
3	0.91	0.82



	Station	True Negative	False Positive	False Negative	True Positive	Accuracy (%)
0	BASEL	3857	481	415	985	84.38
1	BELGRADE	3179	597	511	1451	80.69
2	BUDAPEST	3335	565	380	1458	83.53
3	DEBILT	4277	360	389	712	86.95
4	DUSSELDORF	4089	418	380	851	86.09
5	HEATHROW	4140	430	371	797	86.04
6	KASSEL	4505	310	293	630	89.49
7	LJUBLJANA	3668	527	411	1132	83.65
8	MAASTRICHT	4178	384	370	806	86.86
9	MADRID	2749	419	337	2233	86.82
10	MUNCHENB	4149	397	457	735	85.12
11	OSLO	4608	271	319	540	89.72
12	SONNBLICK	5738	0	0	0	100.00
13	STOCKHOLM	4431	335	352	620	88.03
14	VALENTIA	5379	83	152	124	95.90

Across all 15 stations, the **average accuracy rate is 87.55%**, which shows that the KNN model performs reasonably well overall at distinguishing pleasant from unpleasant weather conditions. Most stations show balanced performance, with accuracy rates clustered between **83% and 90%**.

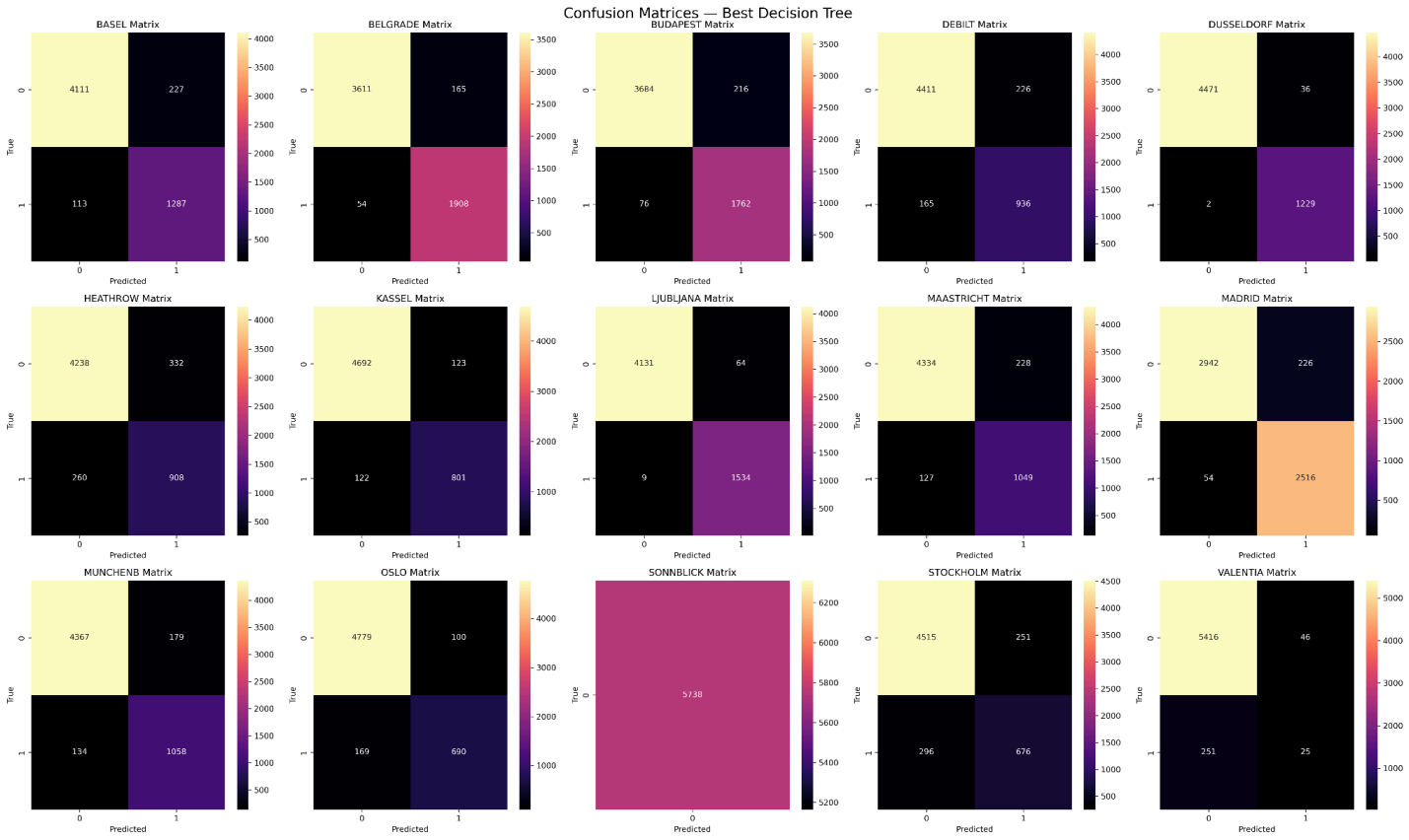
A few stations stand out. **Kassel** and **Oslo** achieve among the highest balanced accuracies (≈89–90%), suggesting their local weather patterns align more closely with the training data. In contrast, **Sonnblick** shows a perfect 100% accuracy but predicts *only* unpleasant days, which reflects extreme label imbalance rather than true predictive skill.

The KNN model provides a reasonable first attempt at predicting pleasant-weather suitability across European stations. The next exercises using Decision Trees and Artificial Neural Networks will allow for deeper comparison and may help address limitations identified here.

Decision Tree Model

Confusion Matrices for the Best Decision Tree

To make sure that my evaluation reflects the most reliable and generalizable decision-tree model, I tested multiple tree configurations rather than relying on a single default model. Because decision trees naturally grow very deep and tend to overfit, I included pruned variations in my analysis and compared their performance using macro-F1, which was a more appropriate metric for this imbalanced, multi-label dataset. This allowed me to identify the **"best" decision tree (max_depth = 10)**, which achieved the strongest test-set macro-F1 and the smallest train–test performance gap. Therefore, all further evaluation uses the best pruned model, as it best represents true predictive performance on unseen data.



	Station	Accurate 0 (TN)	Accurate 1 (TP)	False Pos (FP)	False Neg (FN)	Accuracy Rate (%)
0	BASEL	4111	1287	227	113	94.07
1	BELGRADE	3611	1908	165	54	96.18
2	BUDAPEST	3684	1762	216	76	94.91
3	DEBILT	4411	936	226	165	93.19
4	DUSSELDORF	4471	1229	36	2	99.34
5	HEATHROW	4238	908	332	260	89.68
6	KASSEL	4692	801	123	122	95.73
7	LJUBLJANA	4131	1534	64	9	98.73
8	MAASTRICHT	4334	1049	228	127	93.81
9	MADRID	2942	2516	226	54	95.12
10	MUNCHENB	4367	1058	179	134	94.55
11	OSLO	4779	690	100	169	95.31
12	SONNBLICK	5738	0	0	0	100.00
13	STOCKHOLM	4515	676	251	296	90.47
14	VALENTIA	5416	25	46	251	94.82

Accuracy of the Training and Testing Data (Best Model: Depth = 10)

Based on the selected best model:

- Training macro-F1: 0.9502
- Testing macro-F1: 0.9268

These values indicate high performance on both training and testing sets, a small gap between training and testing F1 (≈0.02), and minimal overfitting. The model generalizes well while still capturing meaningful structure in the data.

Does the Decision Tree Need to Be Pruned? Why?

Pruning was needed as described earlier. It was a necessary step to achieve a reliable model. The unpruned decision tree showed clear signs of overfitting: it produced an extremely high training macro-F1 but a noticeably lower test macro-F1, indicating that the model was memorizing training patterns rather than learning generalizable structure. By introducing pruning through max_depth constraints, I was able to compare several versions of the model and evaluate their generalization using macro-F1 on the test set.

ANN Model

Scenario Details

Here is the direct output in Python regarding all 3 scenarios with the macro-FI scores for the training and testing set.

Scenario	hidden_layers	max_iter	tol	Train macro-F1	Test macro-F1
Scenario 1	(5, 5)	500	0.0001	0.8718	0.8662
Scenario 2	(25, 10)	800	0.0001	0.9511	0.9413
Scenario 3	(50, 25, 10)	800	0.0001	0.9617	0.9355

In evaluating the artificial neural network (ANN) models, I focused on comparing confusion matrices using the **testing dataset only** rather than plotting both training and testing matrices. This choice reflects that testing results provide the most meaningful assessment of generalization, while training confusion matrices tend to be overly optimistic (close to 100% in this case) because the model has already seen that data. In addition, I relied on macro-F1 scores, a more appropriate metric for this imbalanced, multi-station dataset, to guide model selection across the three ANN scenarios. Since macro-F1 measures how well the model balances false positives and false negatives across all labels, the testing set offers the clearest picture of real performance. For this reason, all confusion matrices and accuracy tables in this section are derived from the best ANN models evaluated on the testing dataset, which makes sure the analysis reflects how the ANN would perform on unseen weather conditions.

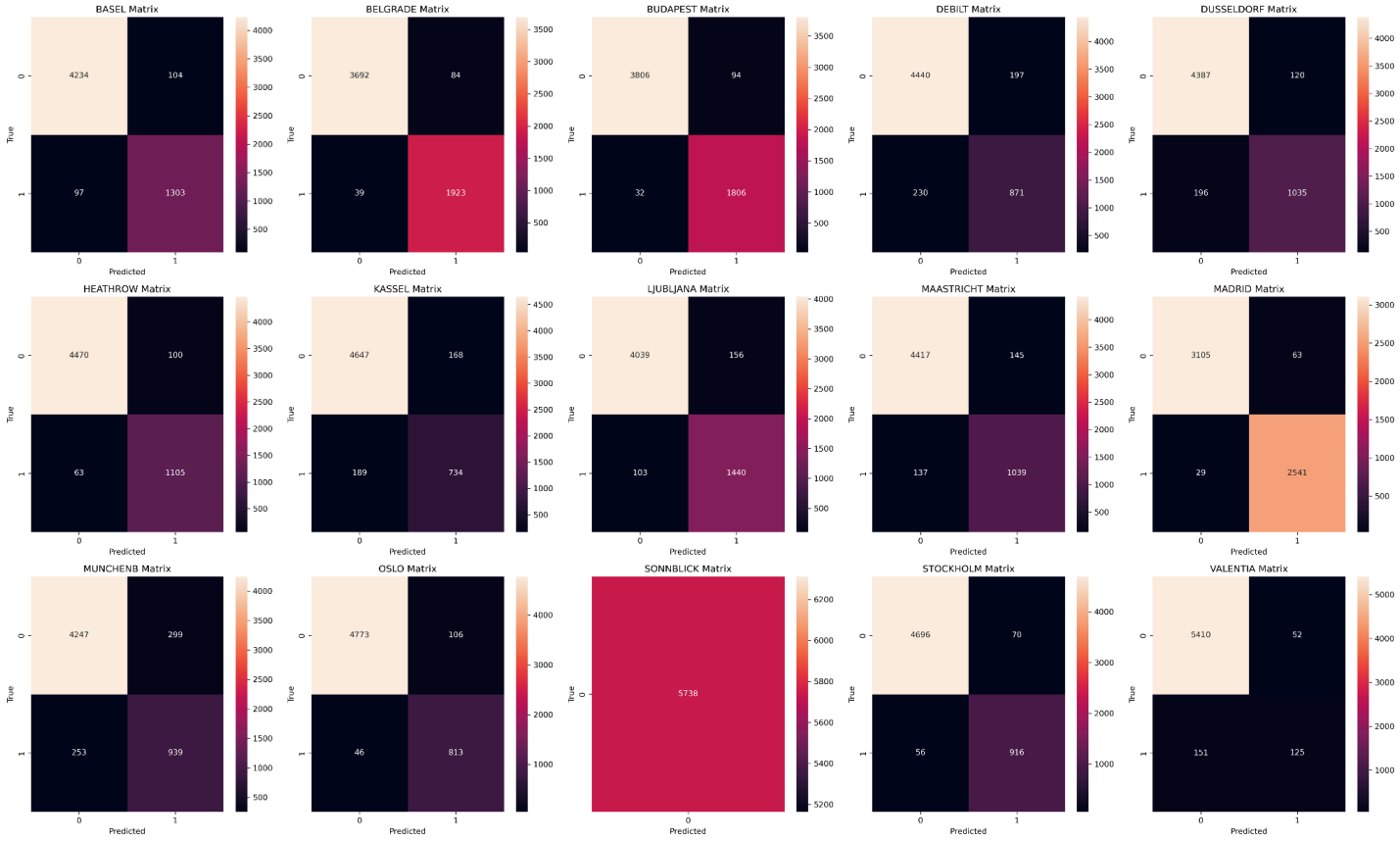
With this set up in mind, I can then provide additional **average accuracy rates** for all 3 scenarios' testing set:

- Scenario 1: 95.69%
- Scenario 2: 93.08%
- Scenario 3: 95.78%

Note that full accuracy rate tables are provided in the pages below.

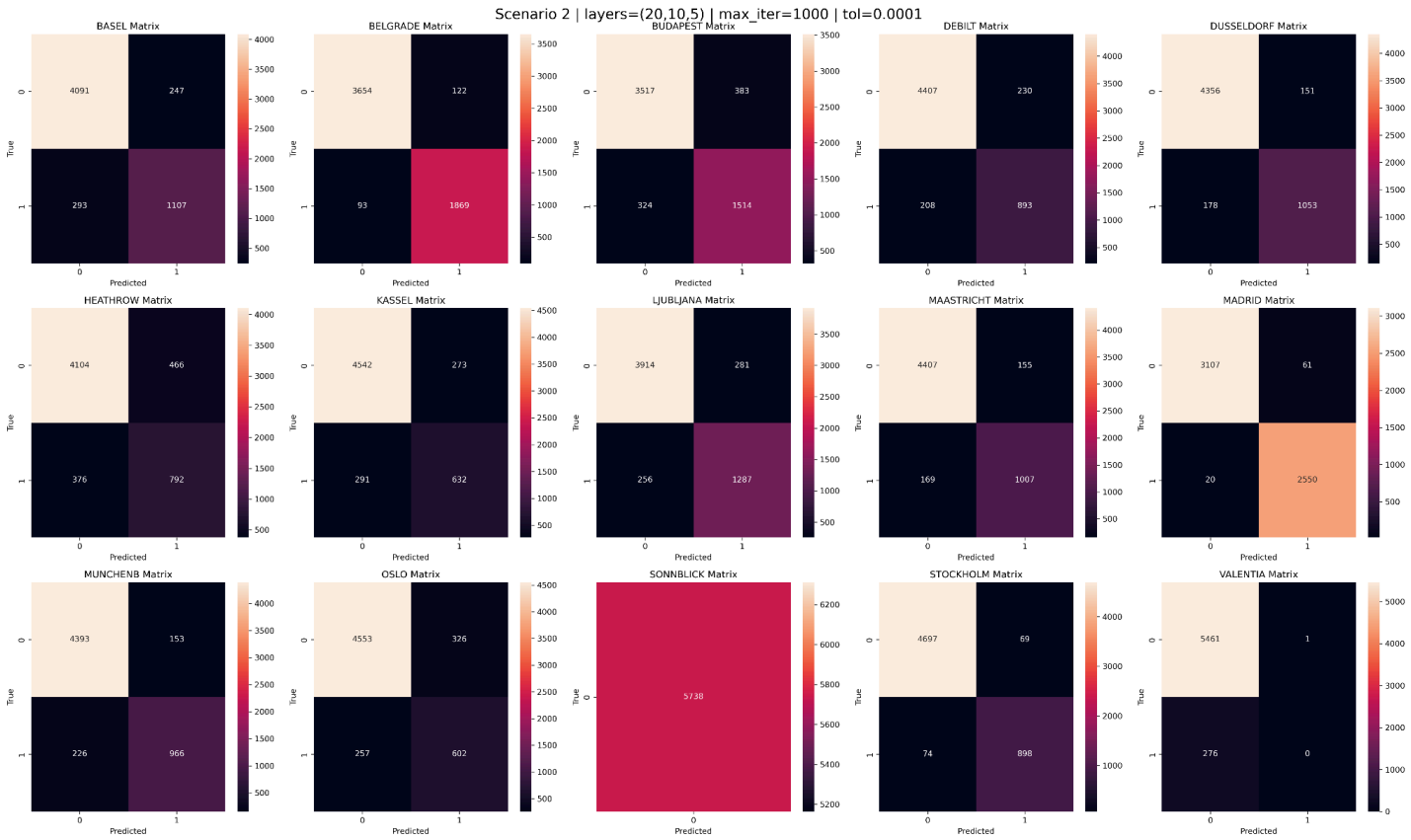
The macro-FI scores and the accuracy rates reveal that **Scenario 2 is likely the best ANN model** because (1) it has the highest test macro-F1 score, (2) its train-test gap indicates great generalization and (3) it beats scenarios 1 and 3 on the metric that matters the most for imbalance multi-label data.

Scenario 1 Confusion Matrices - Testing Set



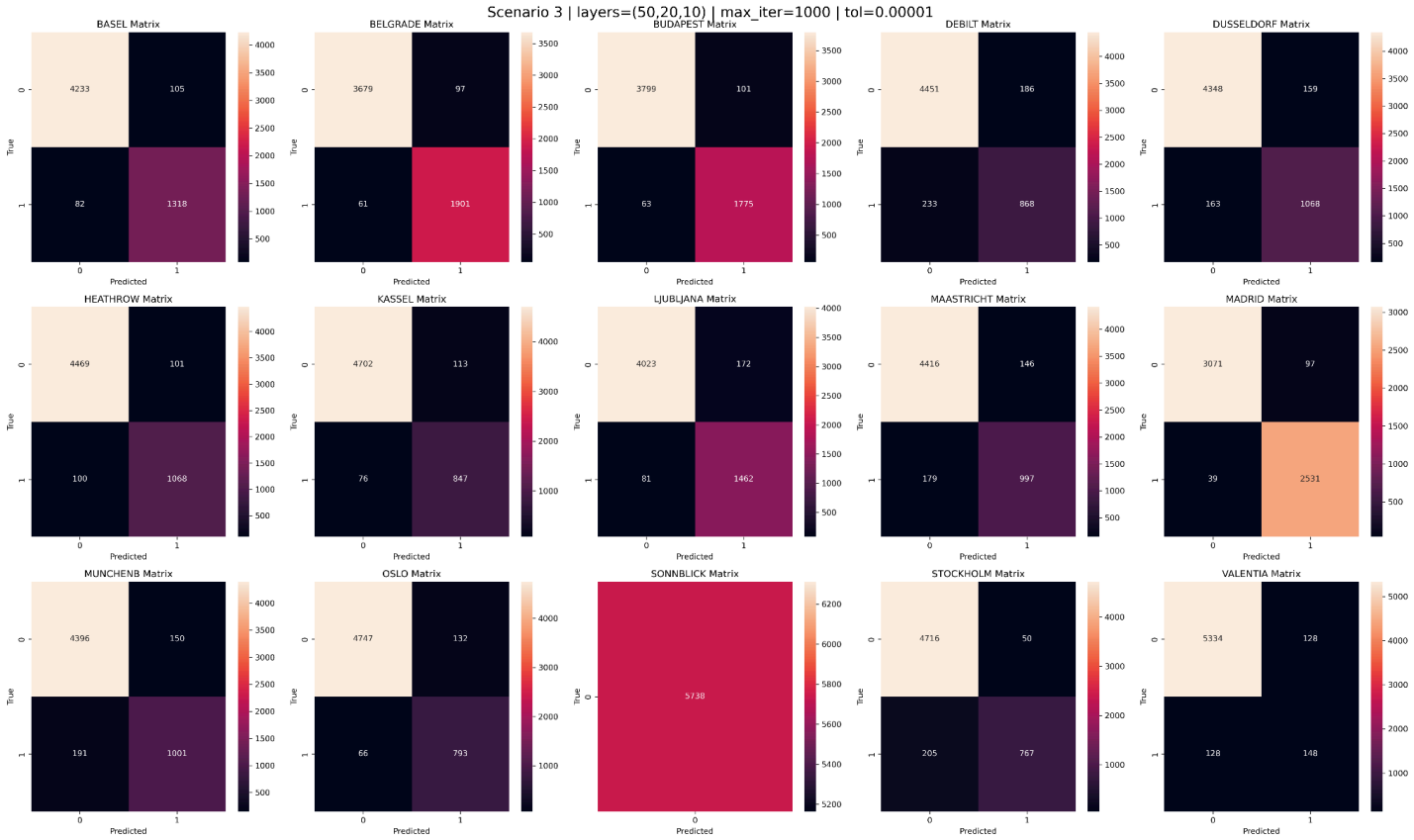
	Station	Accurate 0 (TN)	Accurate 1 (TP)	False Pos (FP)	False Neg (FN)	Accuracy Rate (%)
0	BASEL	4183	1297	155	103	95.50
1	BELGRADE	3670	1866	106	96	96.48
2	BUDAPEST	3814	1701	86	137	96.11
3	DEBILT	4364	922	273	179	92.12
4	DUSSELDORF	4336	1064	171	167	94.11
5	HEATHROW	4427	1017	143	151	94.88
6	KASSEL	4699	784	116	139	95.56
7	LJUBLJANA	4023	1411	172	132	94.70
8	MAASTRICHT	4383	1019	179	157	94.14
9	MADRID	3081	2490	87	80	97.09
10	MUNCHENB	4395	1109	151	83	95.92
11	OSLO	4727	749	152	110	95.43
12	SONNBLICK	5738	0	0	0	100.00
13	STOCKHOLM	4732	834	34	138	97.00
14	VALENTIA	5386	140	76	136	96.31

Scenario 2 Confusion Matrices - Testing Set



	Station	Accurate 0 (TN)	Accurate 1 (TP)	False Pos (FP)	False Neg (FN)	Accuracy Rate (%)
0	BASEL	4091	1107	247	293	90.59
1	BELGRADE	3654	1869	122	93	96.25
2	BUDAPEST	3517	1514	383	324	87.68
3	DEBILT	4407	893	230	208	92.37
4	DUSSELDORF	4356	1053	151	178	94.27
5	HEATHROW	4104	792	466	376	85.33
6	KASSEL	4542	632	273	291	90.17
7	LJUBLJANA	3914	1287	281	256	90.64
8	MAASTRICHT	4407	1007	155	169	94.35
9	MADRID	3107	2550	61	20	98.59
10	MUNCHENB	4393	966	153	226	93.39
11	OSLO	4553	602	326	257	89.84
12	SONNBLICK	5738	0	0	0	100.00
13	STOCKHOLM	4697	898	69	74	97.51
14	VALENTIA	5461	0	1	276	95.17

Scenario 3 Confusion Matrices - Testing Set



	Station	Accurate 0 (TN)	Accurate 1 (TP)	False Pos (FP)	False Neg (FN)	Accuracy Rate (%)
0	BASEL	4233	1318	105	82	96.74
1	BELGRADE	3679	1901	97	61	97.25
2	BUDAPEST	3799	1775	101	63	97.14
3	DEBILT	4451	868	186	233	92.70
4	DUSSELDORF	4348	1068	159	163	94.39
5	HEATHROW	4469	1068	101	100	96.50
6	KASSEL	4702	847	113	76	96.71
7	LJUBLJANA	4023	1462	172	81	95.59
8	MAASTRICHT	4416	997	146	179	94.34
9	MADRID	3071	2531	97	39	97.63
10	MUNCHENB	4396	1001	150	191	94.06
11	OSLO	4747	793	132	66	96.55
12	SONNBLICK	5738	0	0	0	100.00
13	STOCKHOLM	4716	767	50	205	95.56
14	VALENTIA	5334	148	128	128	95.54

Questions

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

Across all three algorithms, the **Artificial Neural Network (ANN)**, specifically **Scenario 2**, provides the strongest overall performance for predicting pleasant weather. The pruned Decision Tree also performs well, achieving a test macro-F1 of 0.9268, while KNN delivers lower generalization performance with test macro-F1 values between 0.77 and 0.82. Because macro-F1 measures how well a model balances false positives and false negatives across all 15 stations, the ANN demonstrates the most consistent and accurate ability to generalize to unseen weather conditions.

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

Yes, Sonnblick consistently achieves 100% accuracy, but this does not imply perfect predictive ability. Sonnblick's labels are extremely imbalanced, and the station almost always reports the same "unpleasant" outcome. Regarding overfitting:

- KNN k=1 clearly overfits with a perfect training F1 of 1.00 but dropping sharply on the test set to 0.81.
- The unpruned Decision Tree also overfits, which is why pruning (max_depth = 10) was necessary.
- By contrast, the ANN models do not show much overfitting; the gaps between training and testing macro-F1 scores remain small across all scenarios. This suggests strong generalization and well-regularized training.

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

Variables such as precipitation and snow depth contain many outliers and long sequences of zeros, making them harder for the models to learn meaningful patterns from. In addition, the multi-label structure means the models must learn many small sub-problems rather than one large one. This naturally produces uneven accuracy across stations, since some locations have more stable weather patterns than others.

Which model would you recommend that ClimateWins use?

Based on all results, the **Artificial Neural Network — Scenario 2** is the best choice for ClimateWins. It offers:

- The highest test macro-F1 (0.9413), the clearest indicator of real predictive performance
- Strong generalization with minimal train–test gap
- Flexible learning capacity, necessary for multi-label, multi-station weather prediction
- Consistent station-level performance, even in stations with more balanced labels

While Decision Trees provide interpretability and perform reasonably well, their macro-F1 is slightly lower. KNN, although simple, underperforms on generalization and is sensitive to high dimensionality.

Therefore, for both predictive accuracy and real-world applicability, ANN Scenario 2 is the strongest and most dependable model for ClimateWins' pleasant-weather forecasting use case.