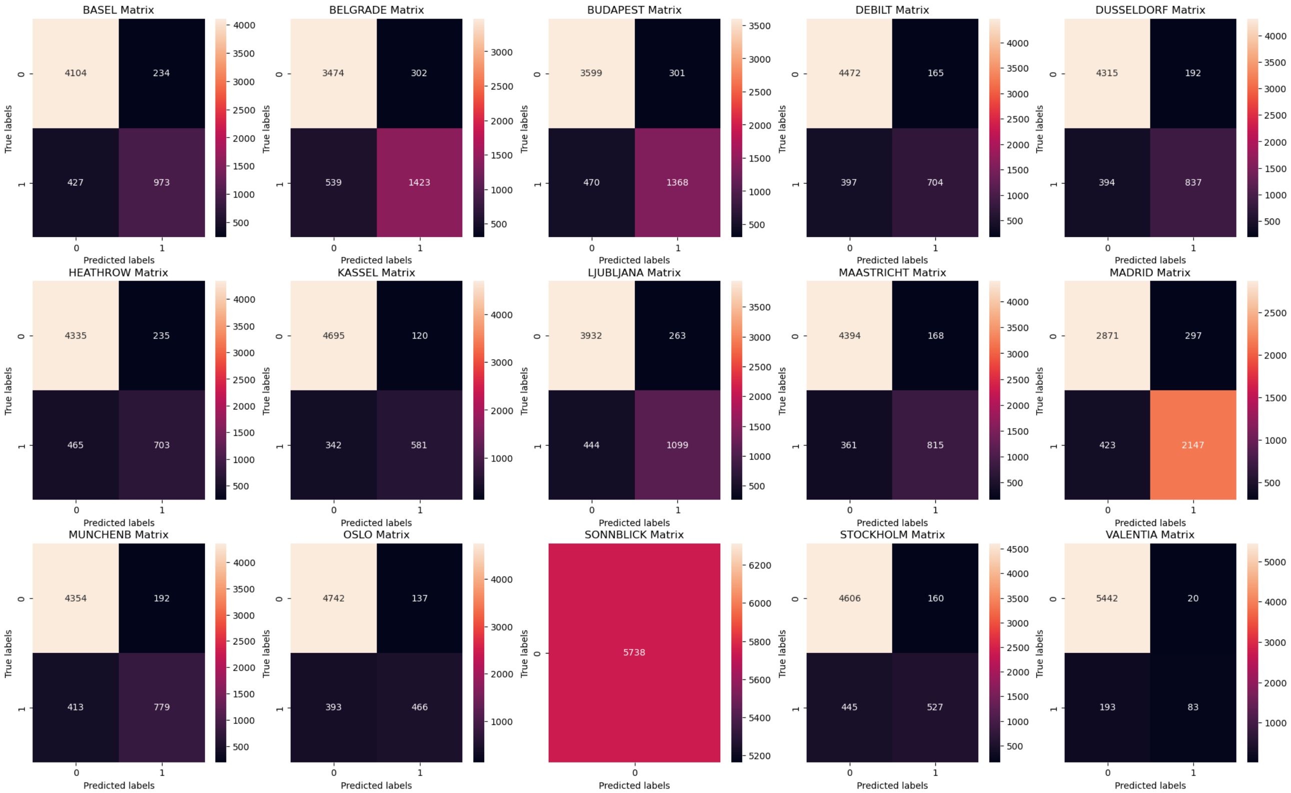
Machine Learning - Supervised Learning Algorithms

KNN Model

Confusion Matrix



Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	4104	973	234	427	**88%**
Belgrade	3474	1423	302	539	**85%**
Budapest	3599	1368	301	470	**87%**
Debilt	4472	704	165	397	**90%**
Dusseldorf	4315	837	192	394	**90%**
Heathrow	4335	703	235	465	**88%**
Kassel	4695	581	120	342	**92%**
Ljubljana	3932	1099	263	444	**88%**
Maastricht	4394	815	168	361	**91%**
Madrid	2871	2147	297	423	**87%**
Munchenb	4354	779	192	413	**89%**
Oslo	4742	466	137	393	**91%**
Sonnblick	5738	0	0	0	**100%**
Stockholm	4606	527	160	445	**89%**
Valentia	5442	83	20	193	**96%**

Record your starting parameters, your final parameters, and the accuracy of the training and testing data. How does the number of neighbors affect the accuracy of the answers?

I started with k-range at 1-4. The model is the most accurate for the training set at 1 number of neighbors and then the accuracy goes steeply down after 2 neighbors. The model works poorly for the test set, predicting the values at less than 50% accuracy. Changing the k-range to 1-5 did not change the results.

How well does this algorithm predict the current data?

From the results of the confusion matrix, this algorithm predicts at 85% to 100% for the various stations (average 90%).

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

Sonnblick has 100% accuracy. If we take a look at the value counts of Sonnblick answers, we see that it is 100% 0 (unpleasant weather). The model is overfitting to Sonnblick.

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

Depending on the location of the weather stations, some features may contribute to overall inaccuracy. Stations in southern Europe, such as Valentia, Roma and Madrid, may not have any snow depth, lower precipitation, and more 'pleasant' days. Northern stations may have more precipitation, less sunshine and lower temperatures. Depending on how the criteria for 'pleasant' weather was formulated, this model will be more accurate for the stations where the features fit the criteria, and inaccurate for stations that don't.

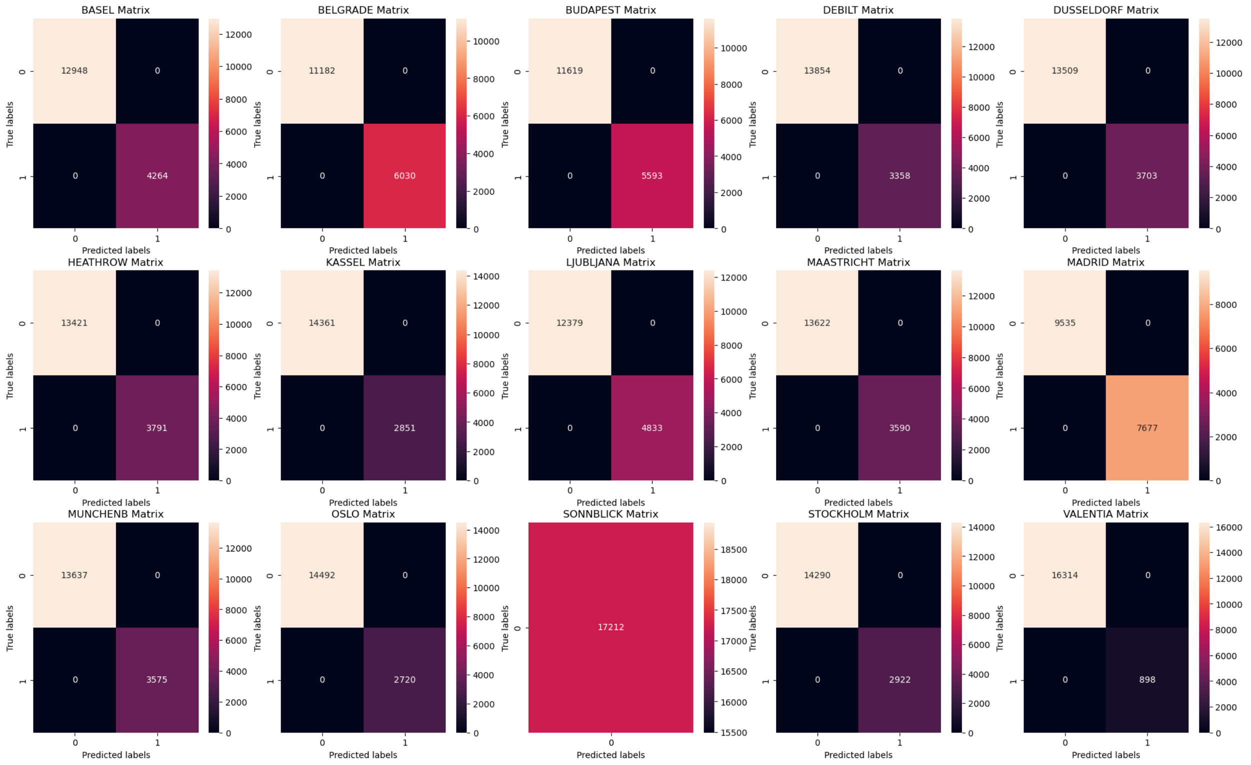
I would argue that predicting pleasant weather should not be based on the entire data set, but rather only the individual station's variables. For instance, Basel predictions should be based only on Basel data.

Running KNN on just one station

For X, I took only the variables pertaining to Basel.
For y, the pleasant weather answers for Basel.
I started with k-range 1-4, then expanded to 1-10. The model is the most accurate for the training set at k = 1. It drops down to 96%, but the testing set starts to rise after. By k=9 they both reach around 94%, which is higher than the accuracy of the model on the entire data set. The confusion matrix confirms that the model predicted about 94% correctly for Basel.

Decision Tree Model

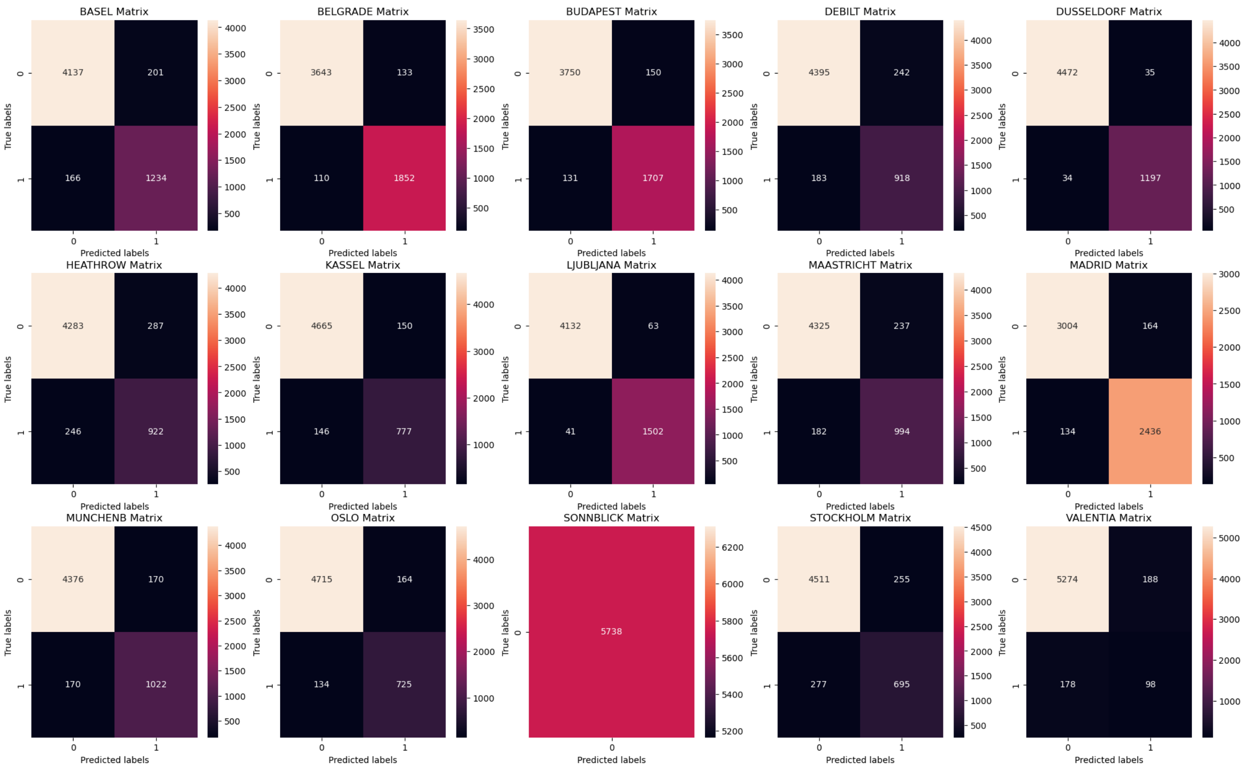
Confusion Matrix for Training Data



Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	12948	4264	0	0	**100%**
Belgrade	11182	6030	0	0	**100%**
Budapest	11619	5593	0	0	**100%**
Debilt	13854	3358	0	0	**100%**
Dusseldorf	13059	3703	0	0	**100%**
Heathrow	13421	3791	0	0	**100%**
Kassel	14361	2851	0	0	**100%**
Ljubljana	12379	4833	0	0	**100%**

Maastricht	13622	3590	0	0	**100%**
Madrid	9535	7677	0	0	**100%**
Munchenb	13637	3575	0	0	**100%**
Oslo	14492	2720	0	0	**100%**
Sonnblick	17212	0	0	0	**100%**
Stockholm	14290	2922	0	0	**100%**
Valentia	16314	898	0	0	**100%**

Confusion Matrix for Testing Data



Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	4137	1234	201	166	**94%**
Belgrade	3643	1852	133	110	**96%**
Budapest	3750	1707	150	131	**95%**
Debilt	4395	918	242	183	**93%**
Dusseldorf	4472	1197	35	34	**99%**
Heathrow	4283	922	287	246	**91%**
Kassel	4665	777	150	146	**95%**
Ljubljana	4132	1502	63	41	**98%**
Maastricht	4325	994	237	182	**93%**
Madrid	3004	2436	164	134	**95%**
Munchenb	4376	1022	170	170	**94%**
Oslo	4715	725	164	134	**95%**
Sonnblick	5738	0	0	0	**100%**
Stockholm	4511	695	255	277	**91%**
Valentia	5274	98	188	178	**94%**

Record the accuracy of the training and testing data.

Train accuracy score: 0.9196301564722617
Test accuracy score: 0.6294876263506448

Confusion matrix: training 100%, test 95%

Do you think the decision tree needs to be pruned? Why?

Yes, it needs to be pruned. There is overfitting, as the model predicts 100% of the training data, and has more errors on the testing data.

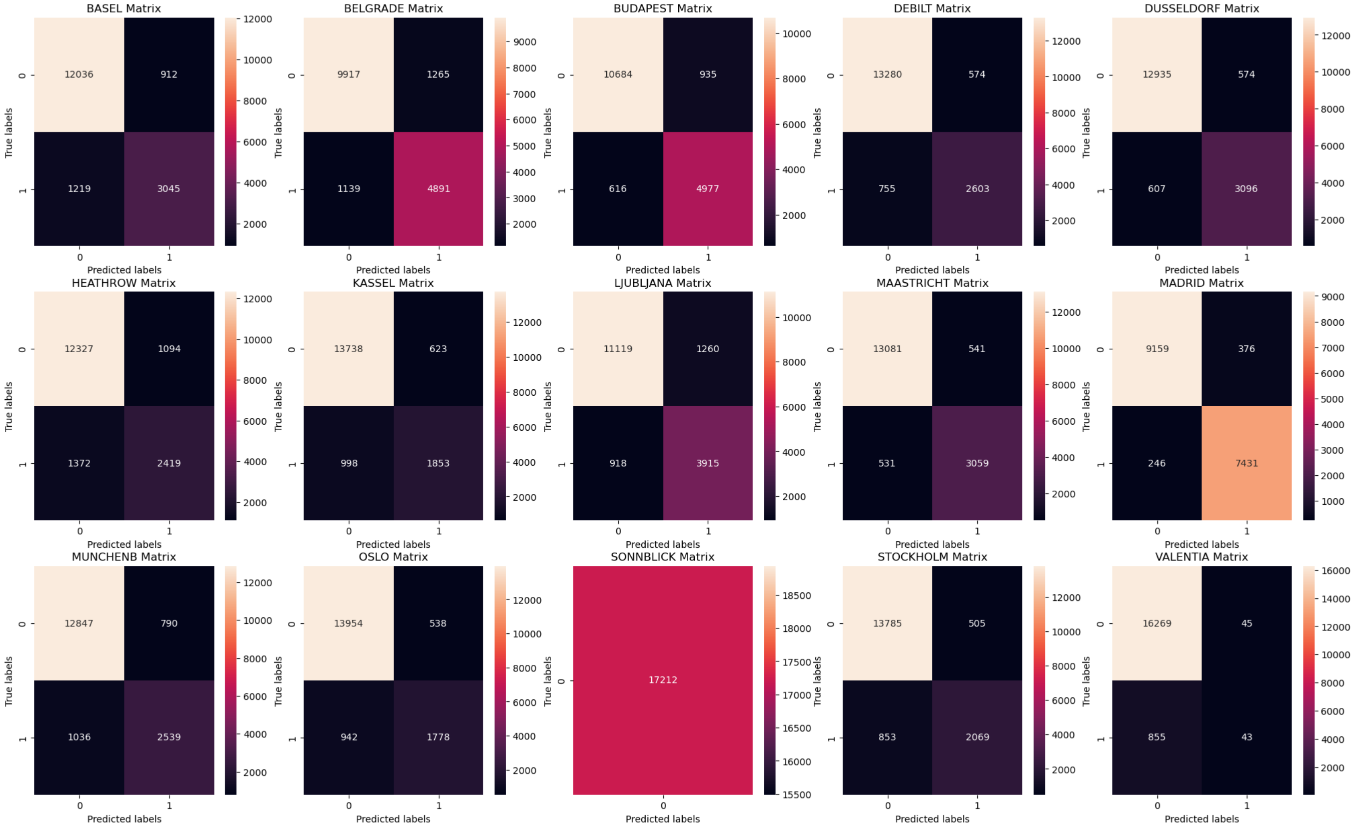
ANN Model

Overview of Scenarios

Scenario	hidden layer sizes =	max iter =	tol =	Train accuracy score	Test accuracy score	Confusion Matrix Accuracy (train/test)
1	(5, 5)	500	0.0001	0.488	0.495	91%/ 91%
2	(20, 10, 5)	1000	0.0001	0.558	0.547	94%/ 93%
3	(50, 20, 10)	1000	0.00001	0.713	0.623	97%/ 95%

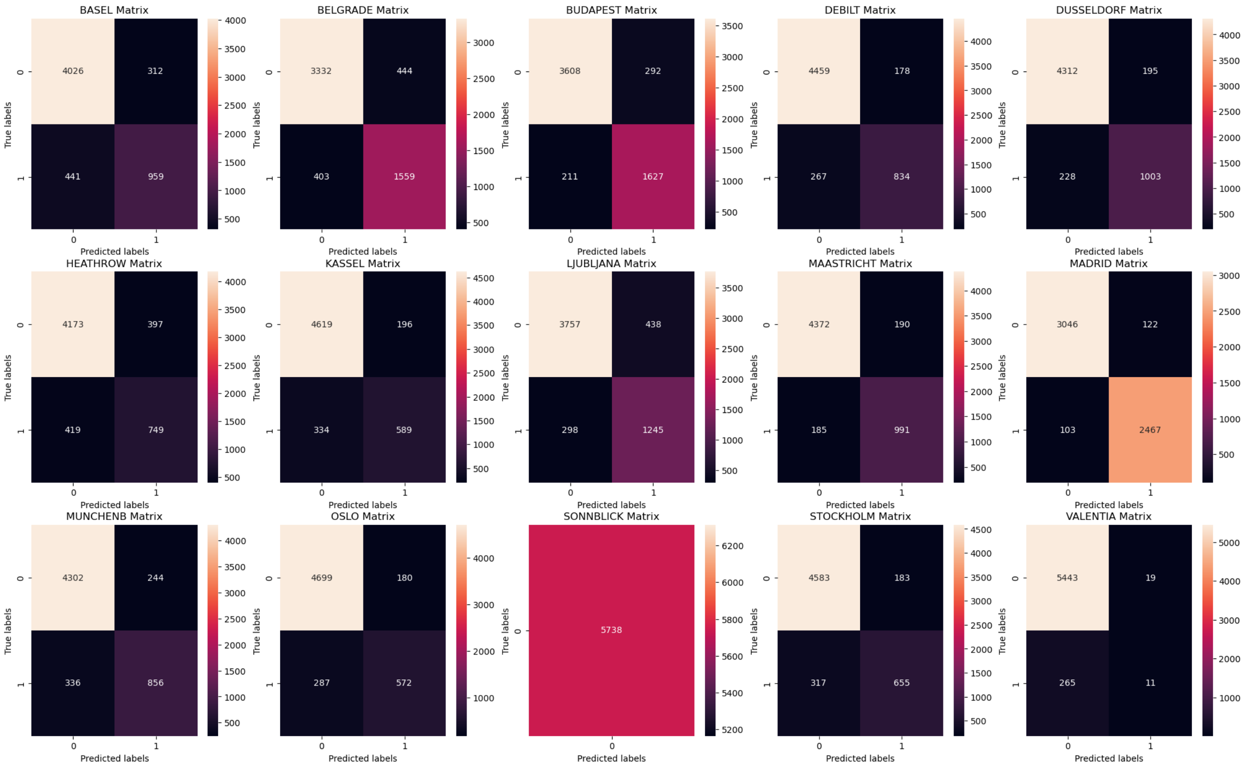
Scenario 1 Confusion Matrices

Training Data



Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	12036	3045	912	1219	**88%**
Belgrade	9917	4891	1265	1139	**86%**
Budapest	10684	4977	935	616	**91%**
Debilt	13280	2603	574	755	**92%**
Dusseldorf	12935	3096	574	607	**93%**
Heathrow	12327	2419	1094	1372	**86%**
Kassel	13738	1853	623	998	**91%**
Ljubljana	11119	3915	1260	918	**87%**
Maastricht	13081	3059	541	531	**94%**
Madrid	9159	7431	376	246	**96%**
Munchenb	12847	2539	790	1036	**89%**
Oslo	13954	1778	538	942	**91%**
Sonnblick	17212	0	0	0	**100%**
Stockholm	13785	2069	505	853	**92%**
Valentia	16269	43	45	855	**95%**

Testing Data

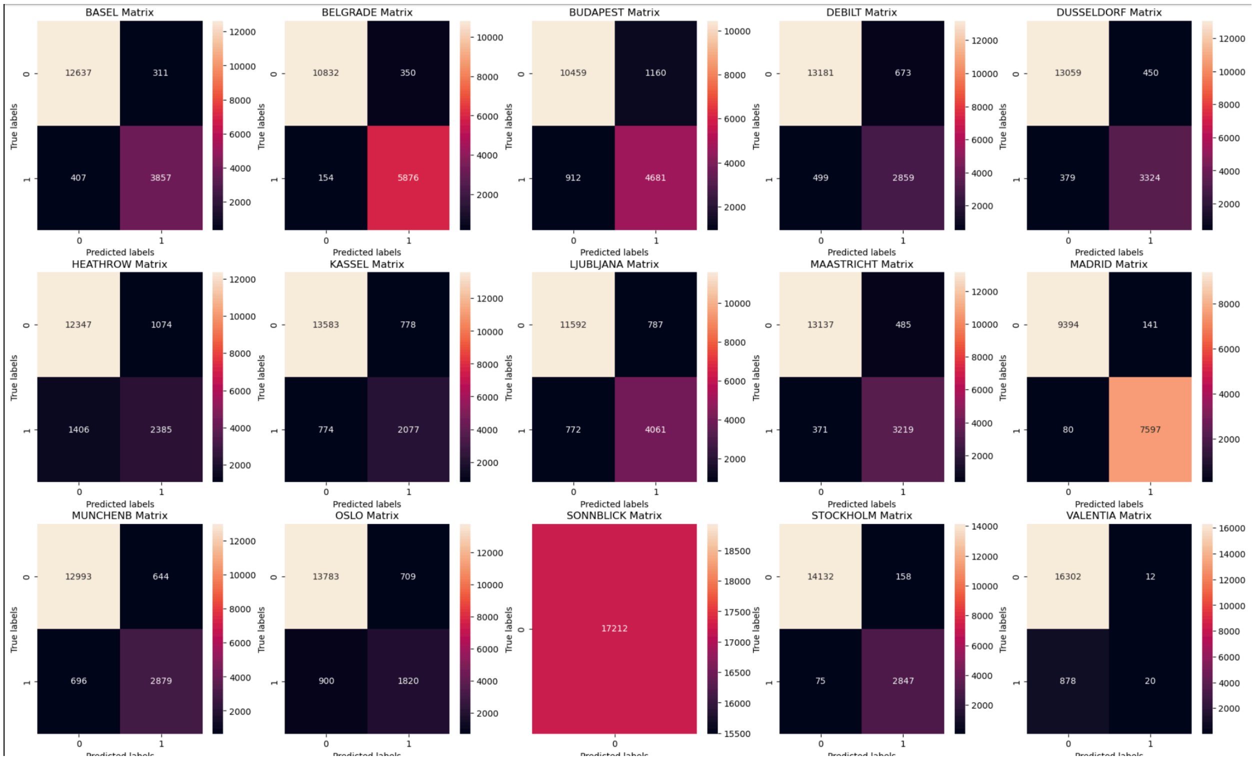


Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	4026	959	312	441	**87%**
Belgrade	3332	1559	444	403	**85%**
Budapest	3608	1627	292	211	**91%**
Debilt	4459	834	178	267	**92%**
Dusseldorf	4212	1003	195	228	**91%**
Heathrow	4173	749	397	419	**86%**
Kassel	4619	589	169	334	**91%**
Ljubljana	3757	1245	438	298	**87%**
Maastricht	4372	991	190	185	**93%**

Madrid	3046	2467	122	103	**96%**
Munchenb	4302	856	244	336	**90%**
Oslo	4699	572	180	287	**92%**
Sonnblick	5738	0	0	0	**100%**
Stockholm	4583	655	183	317	**91%**
Valentia	5443	11	19	265	**95%**

Scenario 2 Confusion Matrices

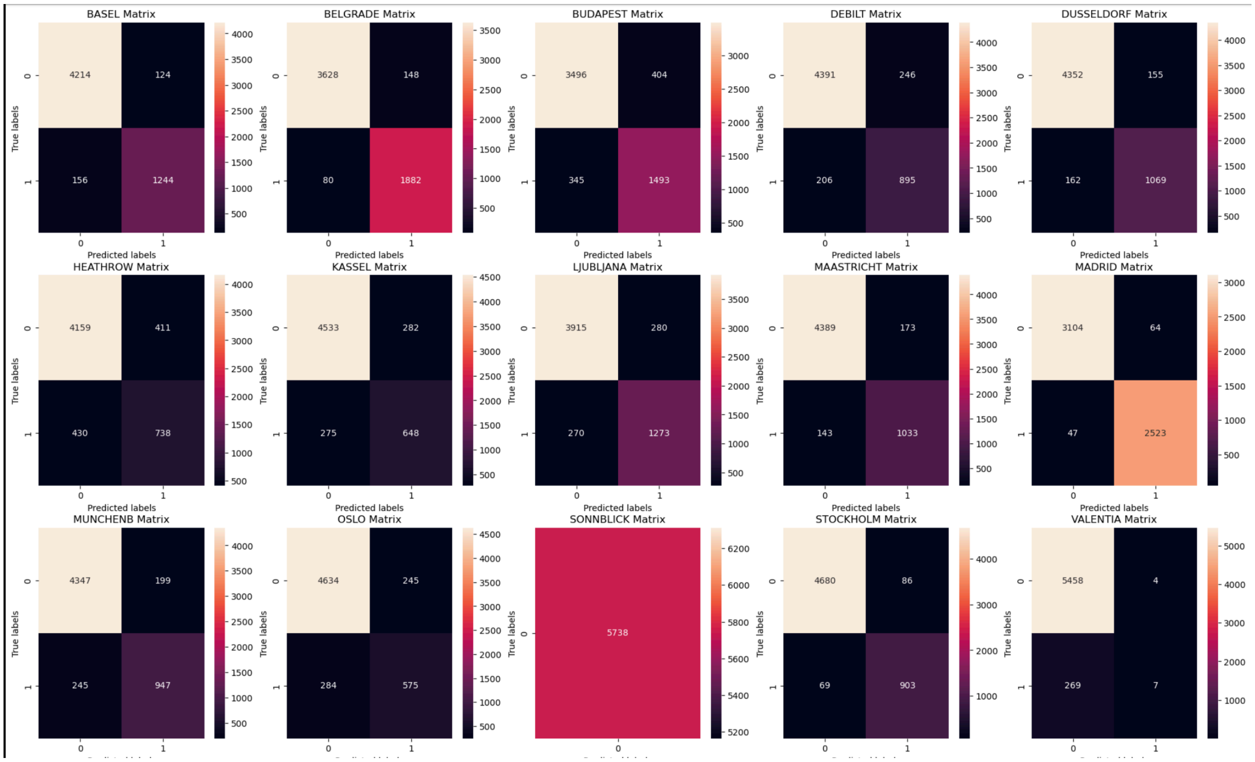
Training Data



Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	12637	3857	311	407	**96%**
Belgrade	10832	5876	350	154	**97%**
Budapest	10459	4681	1160	912	**88%**
Debilt	13181	2859	673	499	**93%**
Dusseldorf	13059	3324	450	379	**95%**
Heathrow	12347	2385	1074	1406	**86%**
Kassel	13583	2077	778	774	**91%**
Ljubljana	11592	4061	787	772	**91%**
Maastricht	13137	3219	485	371	**95%**

Madrid	9394	7597	141	80	**99%**
Munchenb	12993	2879	644	696	**92%**
Oslo	13783	1820	709	900	**91%**
Sonnblick	17212	0	0	0	**100%**
Stockholm	14132	2847	158	75	**99%**
Valentia	16302	20	12	878	**95%**

Testing Data

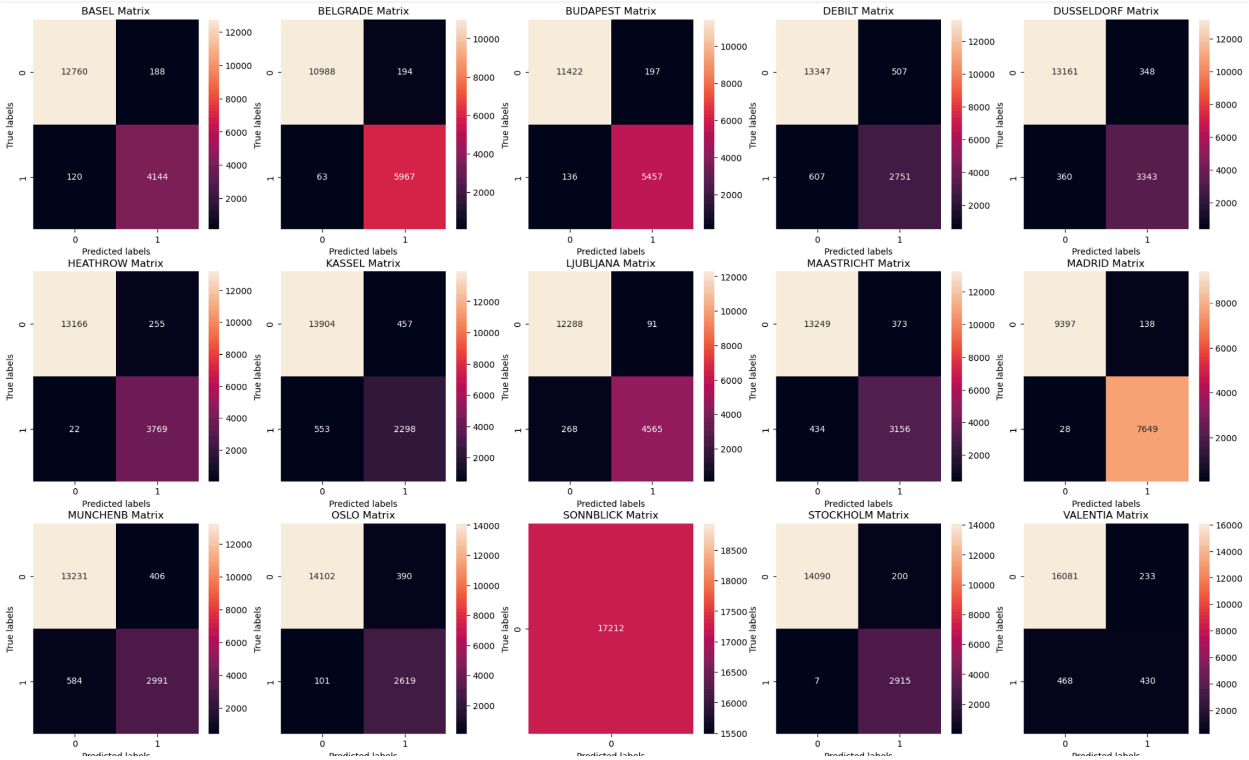


Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	4214	1244	124	156	**95%**
Belgrade	3628	1882	148	80	**96%**
Budapest	3496	1493	404	345	**87%**
Debilt	4391	895	246	206	**92%**
Dusseldorf	4352	1069	155	162	**94%**
Heathrow	4159	738	411	430	**85%**
Kassel	4533	648	282	275	**90%**
Ljubljana	3915	1273	280	270	**90%**
Maastricht	4389	1033	173	143	**94%**

Madrid	3104	2523	64	47	**98%**
Munchenb	4347	947	199	245	**92%**
Oslo	4634	575	245	284	**91%**
Sonnblick	5738	0	0	0	**100%**
Stockholm	4680	903	86	59	**97%**
Valentia	5458	7	4	269	**95%**

Scenario 3 Confusion Matrices

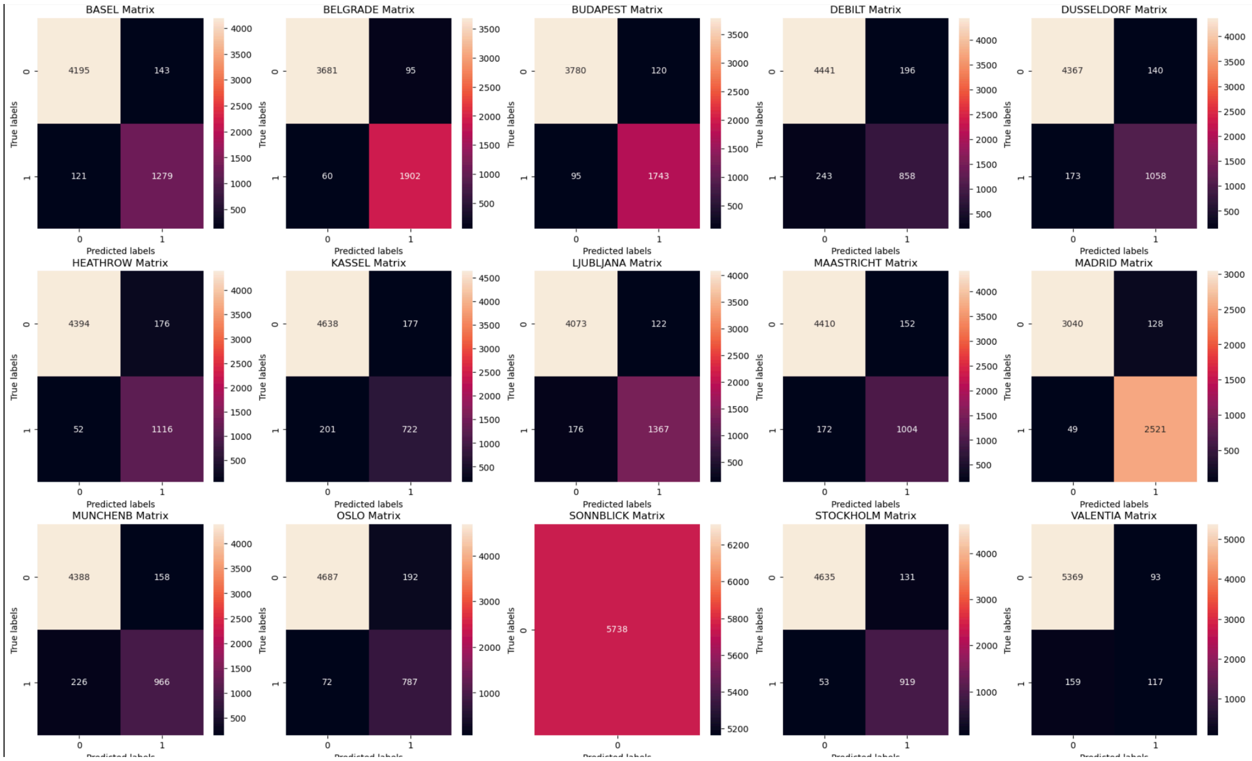
Training Data



Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	12760	4144	188	120	**98%**
Belgrade	10988	5967	194	63	**99%**
Budapest	11422	5457	197	136	**98%**
Debilt	13347	2751	507	607	**94%**
Dusseldorf	13161	3343	348	360	**96%**
Heathrow	13166	3769	255	22	**98%**
Kassel	13904	2298	457	553	**94%**
Ljubljana	12288	4565	91	268	**98%**

Maastricht	13249	3156	373	434	**95%**
Madrid	9397	7649	138	28	**99%**
Munchenb	13231	2991	406	584	**94%**
Oslo	14102	2619	390	101	**97%**
Sonnblick	17212	0	0	0	**100%**
Stockholm	14090	2915	200	7	**99%**
Valentia	16081	430	233	468	**96%**

Testing Data



Station	Accurate 0	Accurate 1	False Pos	False Neg	Accuracy Rate
Basel	4195	1279	143	121	**95%**
Belgrade	3681	1902	95	60	**97%**
Budapest	3780	1743	120	95	**96%**
Debilt	4441	858	196	243	**92%**
Dusseldorf	4367	1058	140	173	**95%**
Heathrow	4394	1116	176	52	**96%**
Kassel	4638	722	177	201	**93%**
Ljubljana	4073	1367	122	176	**95%**
Maastricht	4410	1004	152	172	**94%**
Madrid	3040	2521	128	49	**97%**
Munchenb	4388	966	158	226	**93%**
Oslo	4687	787	192	72	**95%**
Sonnblick	5738	0	0	0	**100%**
Stockholm	4635	919	131	53	**97%**
Valentia	5369	117	93	159	**96%**

Questions:

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

I think the ANN model works the best in predicting current data. It predicts most accurately and with less overfitting than the decision tree model. There is some room for experiment with the parameters, but Scenario 3 already produces quite effective results.

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

Sonnblick is always fully accurate. The training data overfits and produces a higher accuracy rate than the testing data.

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

From observing the distribution of different features, certain variables such as precipitation and snow depth have many outliers which may contribute to overall accuracy of the models.

Which model would you recommend that ClimateWins use?

I would recommend ClimateWins use the ANN model, as it is the most accurate, and there is a potential also for improving on its performance by experimenting more with the parameters.