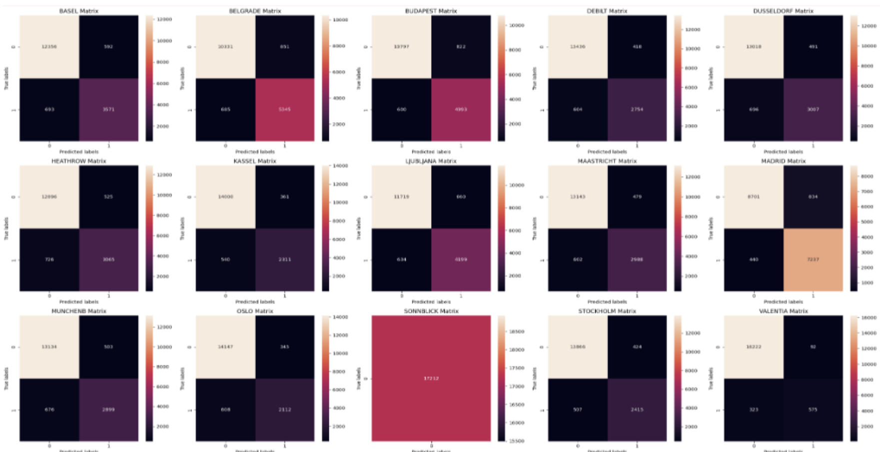
Basics of Machine Learning

1.4 Supervised Learning Algorithms

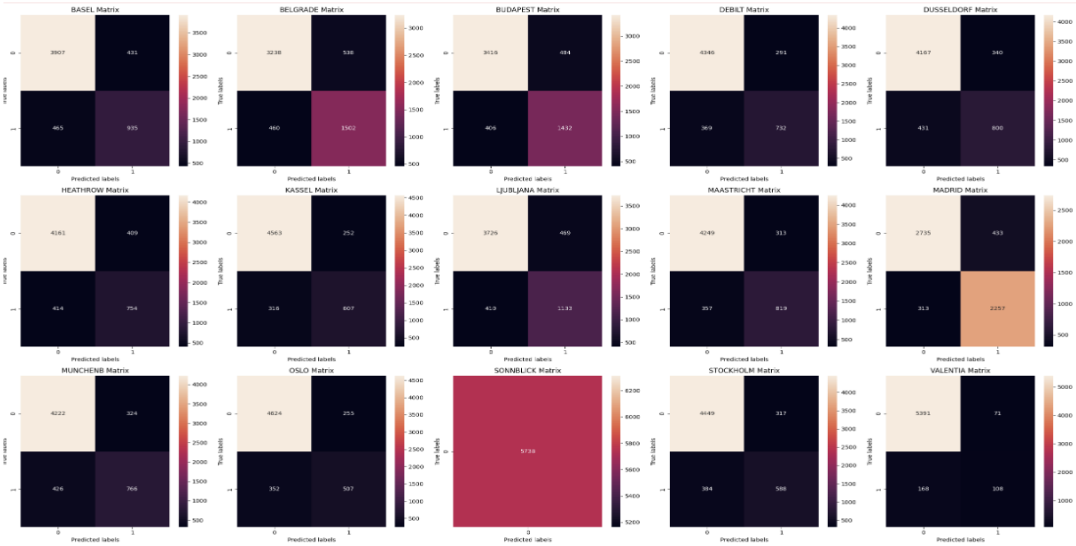
Kendra Jackson

KNN Scaled Data (Train Set)



Weather Station	*Accurate Predictions True Negatives/True Positives*		*False Positive*	*False Negative*	*Accuracy Rate (Overall Percentage of Correct Predictions) (TP+TN)/Total*
Basel	12356	3571	592	693	92.5%
Belgrade	10331	5345	851	685	91.1%
Budapest	10797	4993	822	600	91.7%
Debilt	13436	2754	418	604	94.1%
Dusseldorf	13018	3007	491	696	93.1%
Heathrow	12896	3065	525	726	92.7%
Kassel	14000	2311	361	540	94.8%
Ljubljana	11719	4199	660	634	92.5%
Maastricht	13143	2988	479	602	93.7%
Madrid	8701	7237	834	440	92.6%
Munchenb	13134	2899	503	676	93.2%
Oslo	14147	2112	345	608	94.5%
Sonnblick	17212				100%
Stockholm	13866	2415	424	507	94.6%
Valentia	16222	575	92	323	97.6%
				Average	**93.9%**

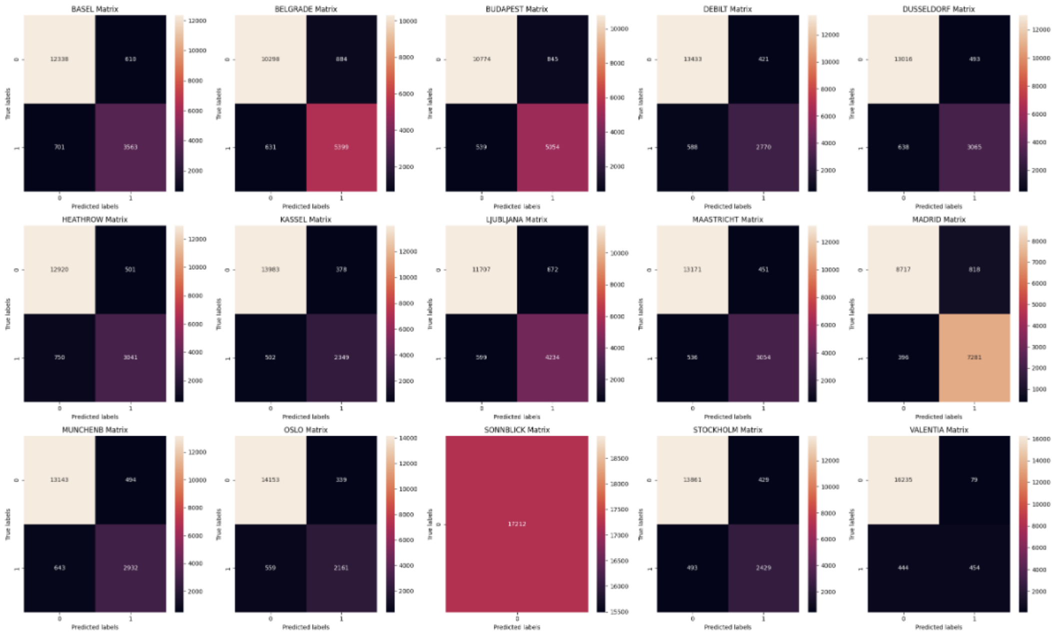
KNN Scaled Data (Test Set)



Weather Station	*Accurate Predictions True Negatives/True Positives*		*False Positive*	*False Negative*	*Accuracy Rate (Overall Percentage of Correct Predictions) (TP+TN)/Total*
Basel	3907	935	431	465	84.3%
Belgrade	3238	1502	538	460	82.6%
Budapest	3416	1432	484	406	84.4%
Debilt	4346	732	291	369	88.4%
Dusseldorf	4167	800	340	431	86.6%
Heathrow	4161	754	409	414	85.7%
Kassel	4563	607	252	316	90.1%
Ljubljana	3726	1133	469	410	84.7%
Maastricht	4249	819	313	357	88.3%
Madrid	2735	2257	433	313	87%
Munchenb	4222	766	324	426	87%
Oslo	4624	507	255	352	89.4%
Sonnblick	5738				100%
Stockholm	4449	588	317	384	87.8%
Valentia	5391	108	71	168	95.8%
				Average	**88.1%**

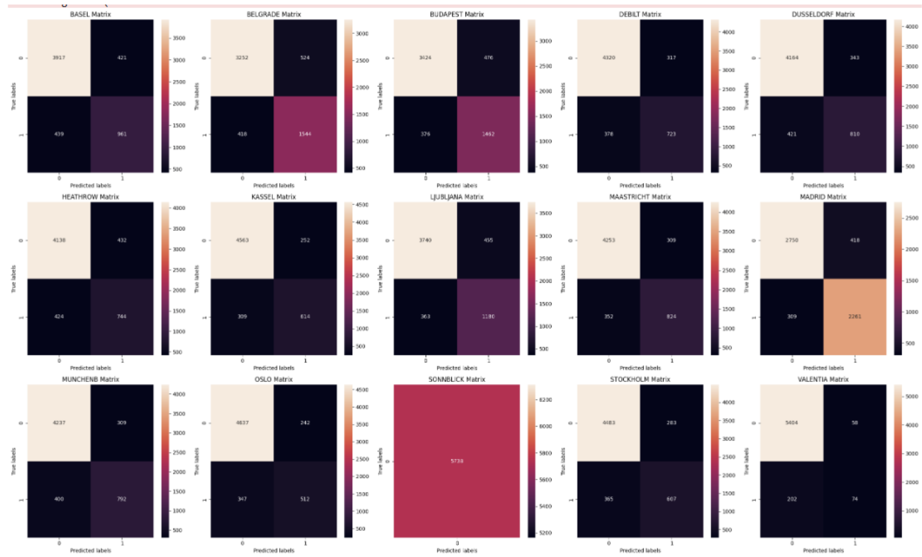
Accuracy decreases by 5.8% between the train and test sets for the scaled data.

KNN Original Prepared Data (Unscaled) – Train Set



Weather Station	*Accurate Predictions True Negatives/True Positives*		*False Positive*	*False Negative*	*Accuracy Rate (Overall Percentage of Correct Predictions)*
Basel	12338	3563	610	701	92.4%
Belgrade	10298	5399	884	631	91.2%
Budapest	10774	5054	845	539	92%
Debilt	13433	2770	421	588	94.1%
Dusseldorf	13016	3065	493	638	93.4%
Heathrow	12920	3041	501	750	92.7%
Kassel	13983	2349	378	502	94.9%
Ljubljana	11707	4234	672	599	92.6%
Maastricht	13171	3054	451	536	94.3%
Madrid	8718	7281	818	396	93%
Munchenb	13143	2932	494	643	93.4%
Oslo	14153	2161	339	559	94.8%
Sonnblick	17212				100%
Stockholm	13861	2429	429	493	94.6%
Valentia	16235	454	79	444	97%
				Average	**94%**

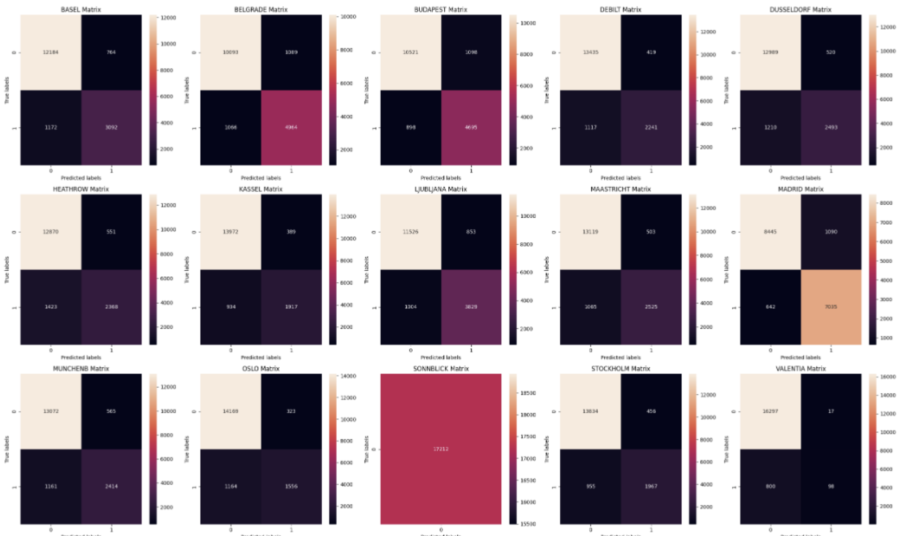
KNN Original Prepared Data (Unscaled) – Test Set



Weather Station	*Accurate Predictions True Negatives/True Positives*		*False Positive*	*False Negative*	*Accuracy Rate (Overall Percentage of Correct Predictions)*
Basel	3917	961	421	439	85%
Belgrade	3252	1544	524	418	83.6%
Budapest	3424	1462	476	376	85.2%
Debilt	4320	723	317	378	87.9%
Dusseldorf	4164	810	343	321	86.7%
Heathrow	4138	744	432	424	85.1%
Kassel	4563	614	252	309	90%
Ljubljana	3740	1180	455	363	85.7%
Maastricht	4253	824	309	352	88.5%
Madrid	2750	2261	418	309	87.3%
Munchenb	4237	792	309	400	87.6%
Oslo	4637	512	242	347	89.7%
Sonnblick	5738				100%
Stockholm	4483	607	283	365	88.7%
Valentia	5404	74	58	202	95.5%
				Average	**88.4%**

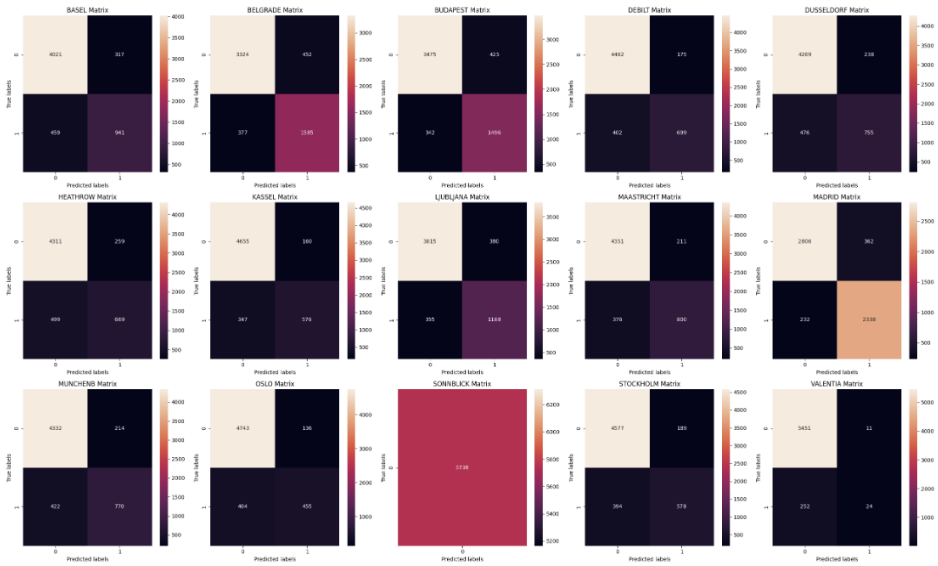
There is a 5.6% decrease in accuracy between train and test sets for the unscaled data. This set has slightly better accuracy so we will look at if the accuracy changes when a higher number of k neighbours is introduced.

KNN Original Prepared Data (Unscaled) – Train Set (parameter for ‘k’ 1-15)



Weather Station	*Accurate Predictions True Negatives/True Positives*		*False Positive*	*False Negative*	*Accuracy Rate (Overall Percentage of Correct Predictions)*
Basel	12184	3092	764	1172	88.8%
Belgrade	10093	4964	1089	1066	87.5%
Budapest	10521	4695	1098	898	88.4%
Debilt	13435	2241	419	1117	91.1%
Dusseldorf	12989	2493	520	1210	89.9%
Heathrow	12870	2368	551	1423	88.5%
Kassel	13972	1917	389	934	92.3%
Ljubljana	11526	3829	853	1004	89.2%
Maastricht	13119	2525	503	1065	90.9%
Madrid	8445	7035	1090	642	89.8%
Munchenb	13072	2414	565	1161	90%
Oslo	14169	1556	323	1164	91.4%
Sonnblick	17212				100%
Stockholm	13834	1967	456	955	91.8%
Valentia	16297	98	17	800	95.3%
				Average	**91%**

KNN Original Prepared Data (Unscaled) – Test Set (parameter for ‘k’ 1-15)



Weather Station	*Accurate Predictions True Negatives/True Positives*		*False Positive*	*False Negative*	*Accuracy Rate (Overall Percentage of Correct Predictions)*
Basel	4021	941	317	459	86.5%
Belgrade	3324	1585	452	377	85.5%
Budapest	3475	1496	425	342	86.6%
Debilt	4462	699	175	402	89.9%
Dusseldorf	4269	755	238	476	87.5%
Heathrow	4311	669	259	499	86.8%
Kassel	4655	576	160	347	91.2%
Ljubljana	3815	1188	380	355	87.2%
Maastricht	4351	800	211	376	89.8%
Madrid	2806	2338	362	232	89.6%
Munchenb	4332	770	214	422	88.9%
Oslo	4743	455	136	404	90.6%
Sonnblick	5738				100%
Stockholm	4577	578	189	394	89.8%
Valentia	5451	24	11	252	95.4%
				Average	**89.6%**

When the parameters are changed to 1-15, the train data set becomes less accurate with predictions compared to lower parameters (1-4). However, the test set has slightly more accuracy (about 1%) with a higher number of neighbours uses.

Parameter values:

Starting: 1-4

Trialed: 1-15

Final: 1-4

A scaled data set and the original prepared data set were used to see how it may affect the accuracy of the KNN model. The KNN weather prediction models show different accuracy levels for each of the 15 stations, with Sonnblick achieving perfect accuracy (100%) at predicting unpleasant weather with both versions of data. This suggests the model is highly accurate at predicting unpleasant weather when faced with data patterns like those of Sonnblick during training. However, this could indicate overfitting and leads to the concern of the adaptability of the model. The overall accuracy rate for the scaled data was 88.1% and the non-scaled data 88.4%. Each version of the model had larger accurate predictions for true negatives (unpleasant weather) suggesting the model may work better for this type of prediction.

Key Takeaways:

- **Accuracy Differences**
  - Sonnblick has 100% accuracy and Valentia 95% accuracy while other areas show lower accuracy like Belgrade or Heathrow. Comparing the numbers between false positives and false negatives, the model seems to struggle slightly more with accurate prediction of “positives” (pleasant weather). This means that the model has some disparities and that its performance varies depending on weather patterns from various geographical locations. This could therefore mean that it may not be a good fit for predicting weather patterns in certain locations.
- **Overfitting Risk**
  - The models 100% accuracy causes suspicion of overfitting. A model can be described as overfit when it appears to have learned the data “too well” and has incorporated noise and outliers during training. This means that its performance will likely be worse on new data. In this regard, the data set used in this training seems to have not properly exposed the model to a diverse range of weather conditions. It could be that our data is slightly biased with many of our data points demonstrating unpleasant weather vs pleasant weather.
- **Generalization Issue**
  - Most weather stations seem to be predicting accurately at around an 86% level without the skew from Valentia and Sonnblick. That being said, there is still a

concerning amount of variability between the accuracy predictions for all stations which suggests the training data may not fully represent real-world conditions. In order for a model to predict well, and across different areas, a model should be able to generalize.

- **Improving Evaluation/Future Improvement**
    - Ideally, this model should be tested with a larger and more varied data test. Incorporating more robust data, covering the above noted discrepancies (not enough pleasant weather data), should help the model distinguish different weather types better. Ultimately, a more accurate model would likely be the outcome if the data we have was expanded on.

In conclusion, the current model demonstrates strong performance with “unpleasant weather” predictions in certain context. However, the overall effectiveness of the model is questionable and its accuracy, if applied to real-world conditions, does not appear to be ideal. In order to improve upon these issues, a more comprehensive approach that incorporates data that is robust and diverse should be used. In doing so, the model would hopefully achieve a more accurate prediction rate improving upon its capabilities to predict weather conditions in real world scenarios.