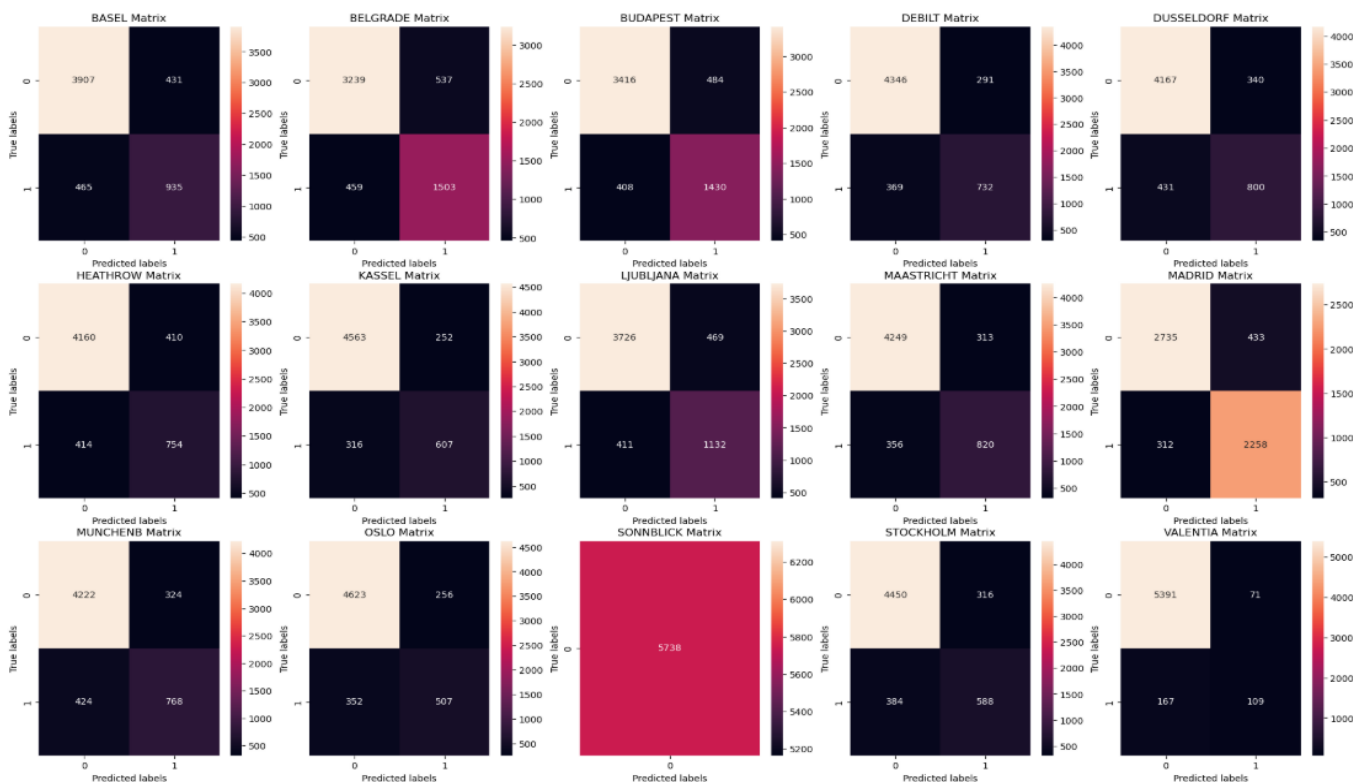


1.4: Supervised Learning Algorithms Part 1

Paul Maden.
Last Update: 08/03/2025

Create a confusion matrix for your final training and testing scenarios. Save a screenshot:



How well does this algorithm predict the current data?

Below is the confusion matrix summarizing the model's classification performance across different weather stations:

Weather Station	True Negatives (TN)	True Positives (TP)	False Positive (FP)	False Negative (FN)	Accuracy Rate
Basel	3907	935	431	465	85%
Belgrade	3239	1303	537	439	84%
Budapest	3416	1410	484	408	85%
Debilt	4346	732	291	389	88%
Dusseldorf	4167	800	380	800	87%
Heathrow	4160	754	332	424	85%
Kassel	4635	607	252	316	90%
Ljubljana	3726	1132	469	411	86%
Maastricht	4249	820	313	356	88%
Madrid	2735	2250	433	322	87%
Munchenb	4222	768	324	400	88%
Oslo	4623	507	256	352	90%
Sonnblick	5738	0	0	0	100%
Stockholm	4450	588	316	308	89%
Valentia	5391	309	71	167	96%
Total					88.53%

The model achieves an overall total accuracy of 88.53%, hence illustrates a strong performance in predicting negatives but does misclassify some positive cases, as shown by the false negatives.

The table below highlights the false negative rate (FNR) for each station, indicating the proportion of actual positive cases that were misclassified as negatives:

(ref: [Confusion matrix - Wikipedia](#))

Weather Station	True Positives (TP)	False Negatives (FN)	False Negative Rate (FNR)
Basel	935	465	33.21%
Belgrade	1303	439	25.20%
Budapest	1410	408	22.44%
Debilt	732	389	34.70%
Dusseldorf	800	800	50.00%
Heathrow	754	424	35.99%
Kassel	607	316	34.24%
Ljubljana	1132	411	26.64%
Maastricht	820	356	30.27%
Madrid	2250	322	12.52%
Munchenb	768	400	34.25%
Oslo	507	352	40.98%
Sonnblick	0	0	nan%
Stockholm	588	308	34.38%
Valentia	309	167	35.08%

Dusseldorf and Oslo have the highest false negative rates, meaning the model is struggling to identify positive cases correctly in these locations.

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

Sonnblick performs best at a 100% accuracy rate! This suggests that it is a strong candidate for overfitting to specific patterns in the dataset rather than learning generalisable trends, since it has no misclassified cases. When compared to the other cities this stands out.

Madrid also has a very low FNR (12.5%) which suggests some overfitting in its positive predictions.

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

Dusseldorf and Oslo have the highest false negative counts, indicating that the model's accuracy is impacted by its failure to correctly identify actual positive cases. These stations potentially have higher variability in weather conditions (i.e. compared to the snowy Sonnblick, up high in the Swiss mountains) causing more misclassifications.

The sample sizes differ in the table, e.g. Belgrade has 5518 samples compared to Dusseldorf which has 6147 samples. This could influence the model as some stations have more training samples than others, where the model learns better from those.

To improve the model's accuracy and reliability, future enhancements should focus on addressing overfitting in Sonnblick, reducing false negatives in Dusseldorf and Oslo, and standardising data collection across weather stations. Retraining with new data will ensure better generalisation and adaptability to changing weather patterns.