

CLIMATEWINS  
PROJECT SUMMARY



Keanu Gomes March, 2024  
*CareerFoundry Student*



# I. PROJECT OBJECTIVE AND HYPOTHESES

[Tableau Dashboard](#)

## GOAL OBJECTIVE

Apply optimization algorithms, supervised and unsupervised machine learning techniques to predict the consequences of climate change as a data analyst at the European non profit organization, ClimateWins.



## HYPOTHESES THAT CAN BE PROPOSED FROM THIS DATA:

1. Which algorithm predicts pleasant weather days best?
2. Will warmer temperatures correlate positively with the occurrence of pleasant weather days?
3. Does higher global radiation correspond to increased temperatures in cities?

# II. DATA ETHICS



## DATA SOURCE

<https://www.ecad.eu>



## BIAS TYPES

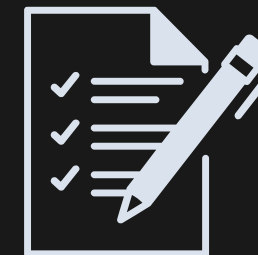
### Selection Bias

Only 18 out of 26321 weather stations were chosen as sample data



## DATA ACCURACY

The data selected for this analysis comes from reliable and trustable sourcing, as is it from 87 participants from verified meteorological stations across Europe totaling 26321 weather stations and 13 characteristics to be analyzed.



## DATA DIMENSIONS

22,951 rows x 170 columns



**18**

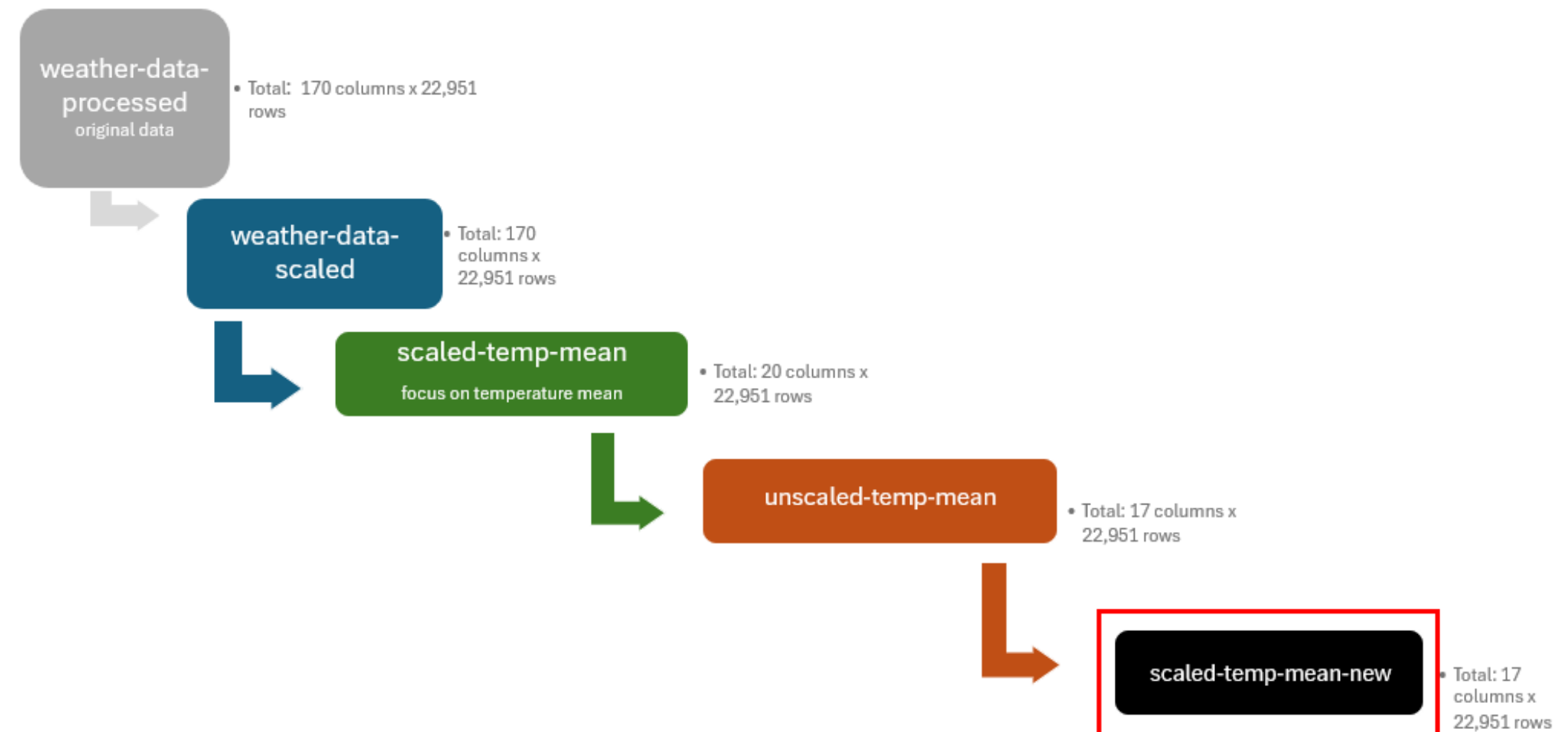
Total weather stations

# III. DATA WORKFLOW

## TEMPERATURE (MEAN) FOCUSED ANALYSIS

- Displayed on the right, is the population flow of my analysis through the (already) processed & cleaned data received from the weather stations.
- In order to feed the data into our supervised learning algorithms, we must removed non-pertinent columns & scale the data in order to normalize it for a more accurate analysis.

### CLIMATEWINS Population flow



# GRADIENT DESCENT

## IV. How was optimization used to determine the features of this data set?

# GRADIENT DESCENT

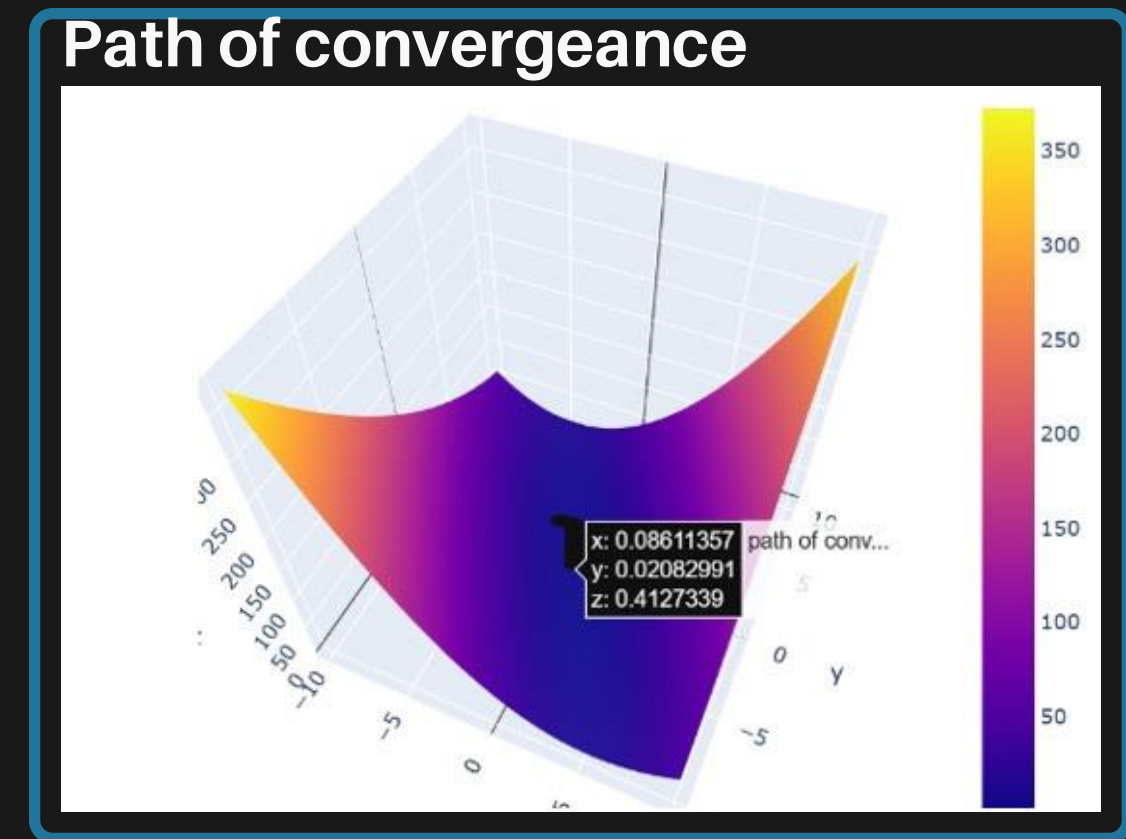
## IV. How was optimization used to determine the features of this data set?

# Path of convergence

A 3D surface plot illustrating the path of convergence for a function. The surface is colored according to a scale from 50 (dark blue) to 350 (yellow). The axes are labeled x, y, and z. A specific point on the surface is highlighted with a black box containing the coordinates:

x: 0.08611357  
y: 0.02082991  
z: 0.4127339

The text "path of conv..." is also visible near the highlighted point.



- Figured below, is a screenshot of the ending parameters I used for optimizing the algorithm.

# Loss function

Values of  $\theta$  and  $J(\theta)$  over iterations\_LJUBLJA

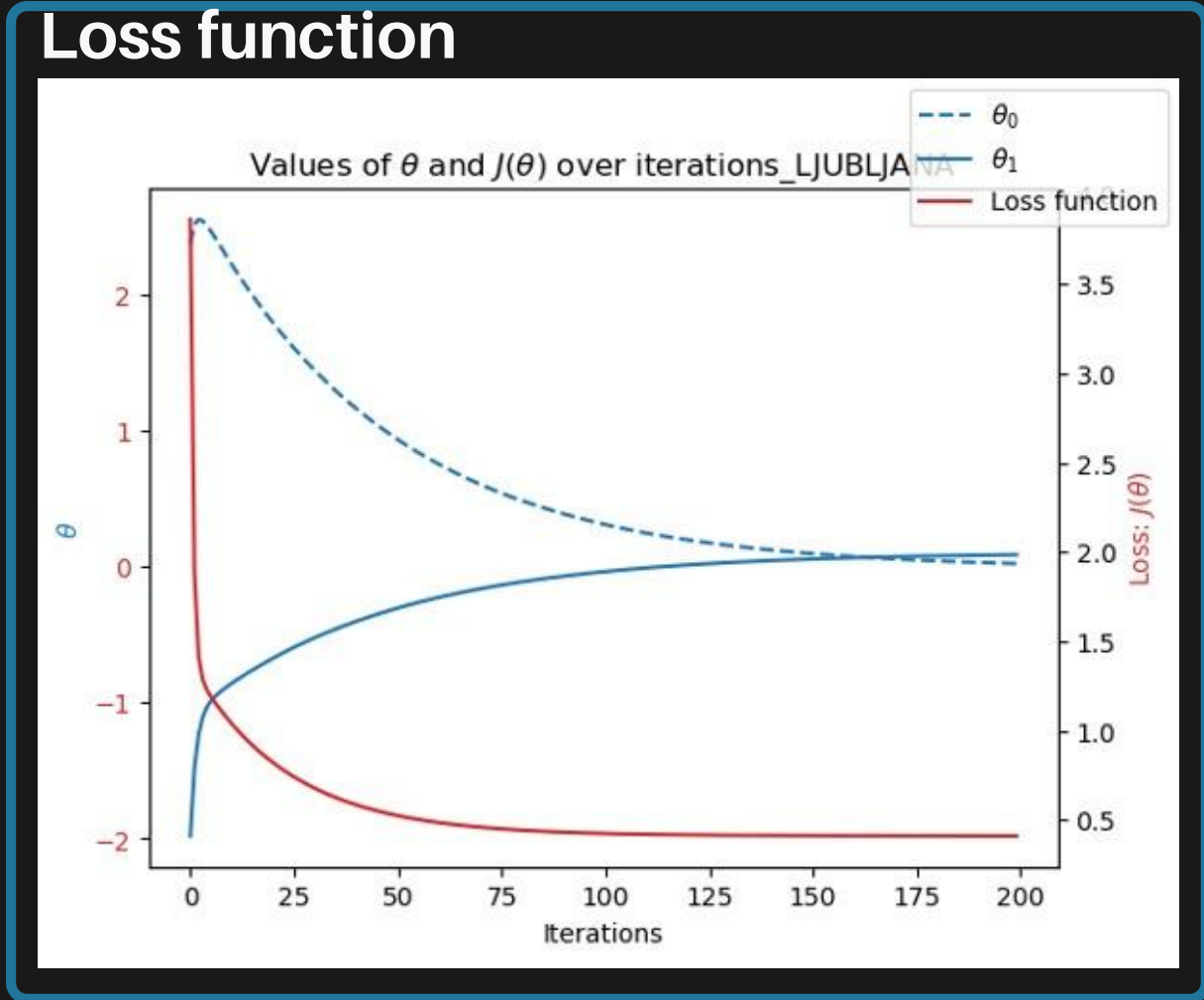
Legend:

- $\theta_0$  (dashed blue line)
- $\theta_1$  (solid blue line)
- Loss function (solid red line)

The graph shows the convergence of the parameters  $\theta_0$  and  $\theta_1$  and the loss function  $J(\theta)$  over 200 iterations. The x-axis represents the number of iterations (0 to 200). The left y-axis represents the values of  $\theta_0$  and  $\theta_1$  (-2 to 2). The right y-axis represents the loss function  $J(\theta)$  (0.5 to 3.5).

Approximate values from the graph:

Iterations	$\theta_0$	$\theta_1$	Loss: $J(\theta)$
0	2.5	-2.0	3.5
25	1.5	-0.5	1.5
50	1.0	-0.2	0.8
75	0.6	-0.1	0.6
100	0.4	-0.05	0.55
125	0.3	-0.02	0.52
150	0.2	-0.01	0.51
175	0.15	-0.005	0.505
200	0.1	0.0	0.5

[illegible][illegible]

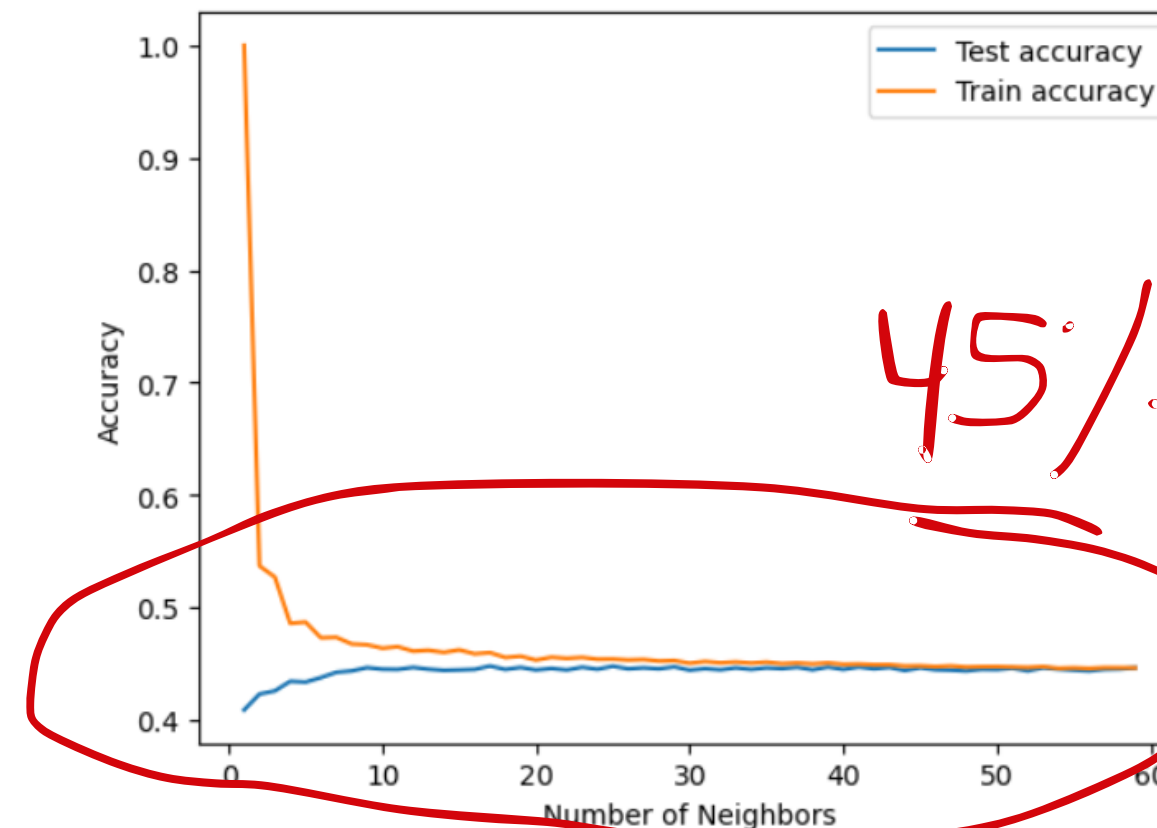
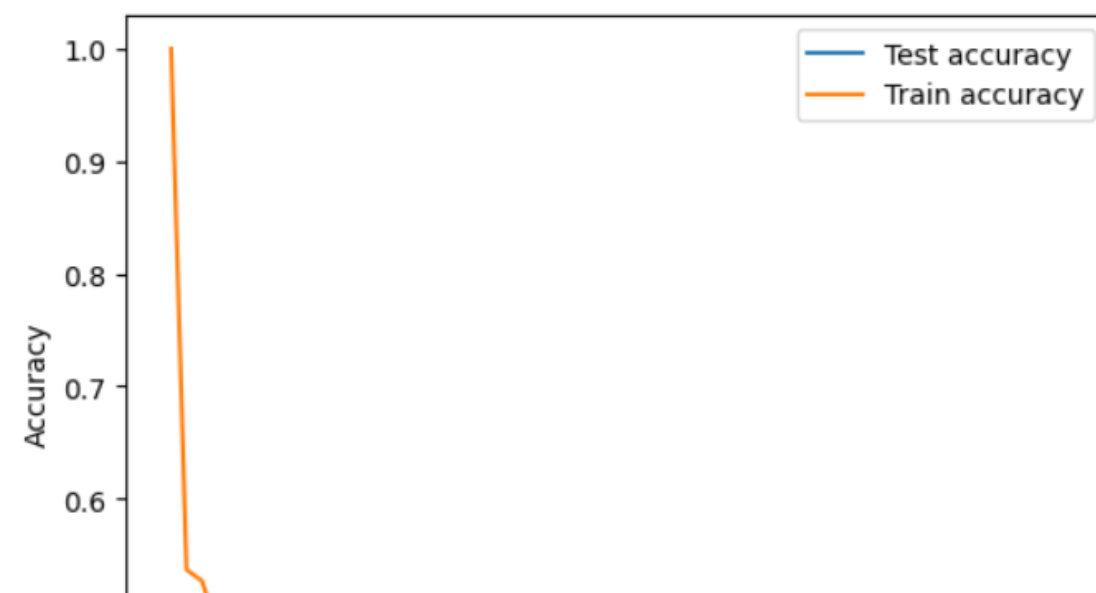
# V. ACCURACY SCORES

KNN

{K-NEAREST NEIGHBORS}

```
In [20]: 1 #plot the relationship between number of neighbors and accuracy
2 plt.plot(k_range, test_acc, label = 'Test accuracy')
3 plt.plot(k_range, train_acc, label = 'Train accuracy')
4 plt.legend()
5 plt.xlabel('Number of Neighbors')
6 plt.ylabel('Accuracy')
```

Out[20]: Text(0, 0.5, 'Accuracy')



- Overall Testing Accuracy: approx. 0.45 or 45%
- Individual Station Accuracy: 0.82 - 0.95 or 82-95%
- Interpretation: Potential overfitting; lower overall accuracy compared to individual station scores.



# V.I. ACCURACY SCORES

{DECISION TREE}

```
In [12]: ▶ 1 #What is the testing accuracy score? Using the cross validation method
          2 y_pred = weather_dt.predict(X_test)
          3 print('Test accuracy score: ', accuracy_score(y_test, y_pred))
          4 multilabel_confusion_matrix(y_test, y_pred)
```

Test accuracy score: 0.4051934471941443

```
Out[12]: array([[3735, 603],
                [ 555, 845]],

              [[3143, 633],
                [ 622, 1340]],
```

- Overall Testing Accuracy: approx. 0.405 or 40%
- Individual Station Accuracy: 0.82 - 0.95 or 82-95%
- Interpretation: Potential overfitting; lower overall accuracy compared to individual station scores.

## V.II. ACCURACY SCORES

ANN

{ARTIFICIAL NEURAL NETWORK}

```
2 mlp = MLPClassifier(hidden_layer_sizes=(20, 10, 10), max_iter=1000, tol=0.0001) #increasing hidden layers
3 #Fit the data to the model
4 mlp.fit(X_train, y_train)
```

Out[31]:

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

```
In [32]: 1 y_pred = mlp.predict(X_train)
2 print(accuracy_score(y_pred, y_train))
3 y_pred_test = mlp.predict(X_test)
4 print(accuracy_score(y_pred_test, y_test))
```

```
0.45044155240529865
0.4527710003485535
```

- Overall Testing Accuracy: approx. 0.452 or 45%
- Individual Station Accuracy: 0.82 - 0.95 or 82-95%
- Interpretation: Potential overfitting; lower overall accuracy compared to individual station scores.



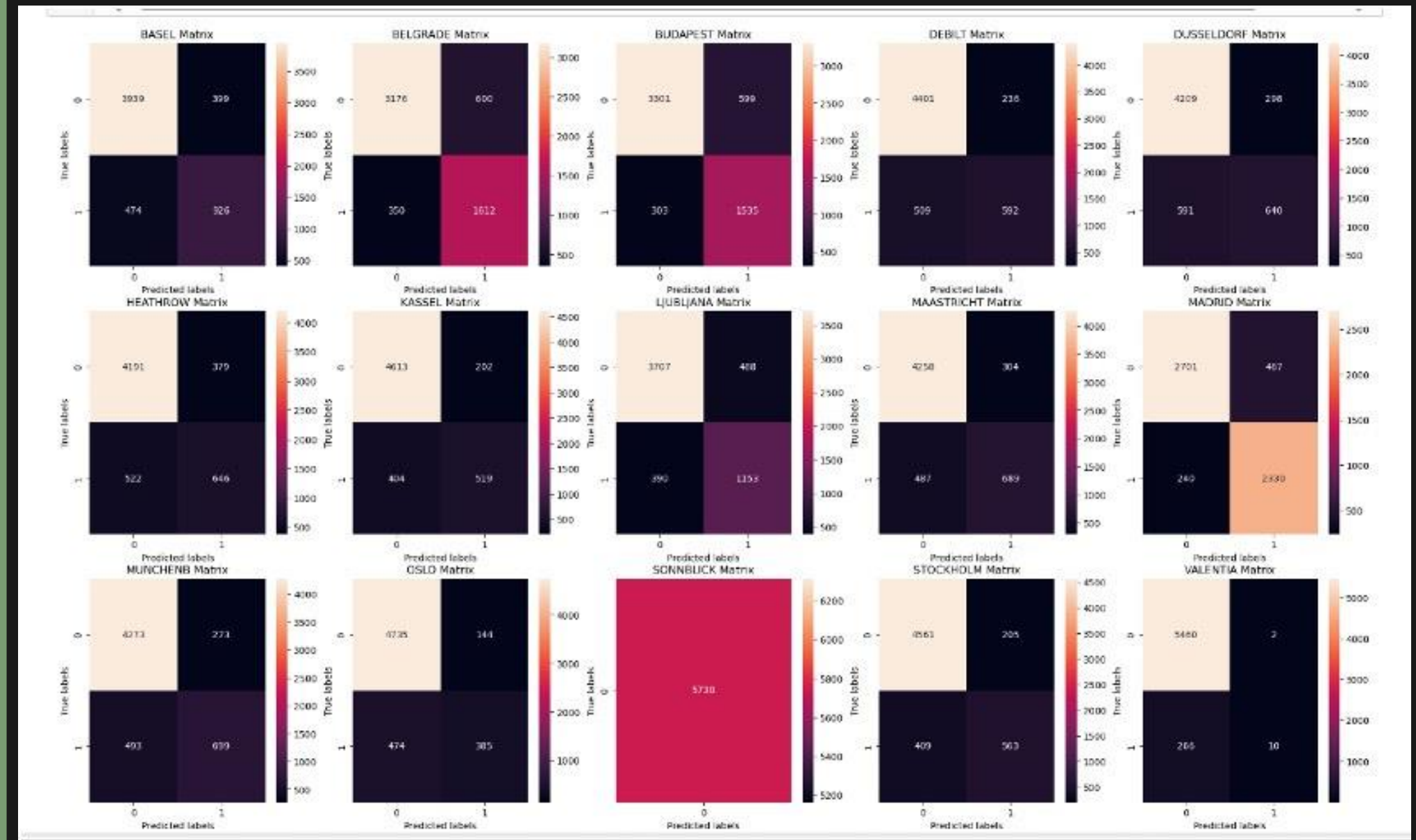
## KNN/ANN/ OR DECISION TREE?

VI. How accurately do the algorithms predict pleasant and non-pleasant days per weather station?

### VALENTIA PREDICTION METRICS for 60 neighbors

1. Accuracy: 95.34%
2. Precision: 99.96%
3. Recall (Sensitivity): 95.37%
4. F1 Score: 97.61%

## Confusion Matrix Scores (pleasant vs non-pleasant weather)



- VALENTIA seems to have the least false positives and negatives, & the highest number of true positives out of every station and algorithm used, this indicates that it may be the most accurate at the individual level.

VII. Which supervised learning algorithm types will be most effective for our hypotheses?



## PRIMARY RECOMMENDATIONS

### CONSISTENT TREND:

**40-45%**

Overall Accuracy

**82-100%**

Individual Station Accuracy

### VALENTIA STANDS OUT:

**95%**

Achieves high accuracy scores consistently around

### RECOMMENDATIONS:

- Investigate data quality and potential biases.
- Conduct feature importance analysis to leverage Valentia's strengths.
- Continue model refinement for improved accuracy.

### SUMMARY:

- Engage stakeholders to discuss implications and actions.
- All models show potential overfitting with lower overall accuracy compared to individual station scores.
- Further analysis and model refinement are necessary to address overfitting and improve generalization performance.



**THANKS FOR FOLLOWING ALONG!**

Any Questions?

Please contact me below at  
keanudatatech@gmail.com

or visit:

<https://keanudatatech.github.io/portfolio>