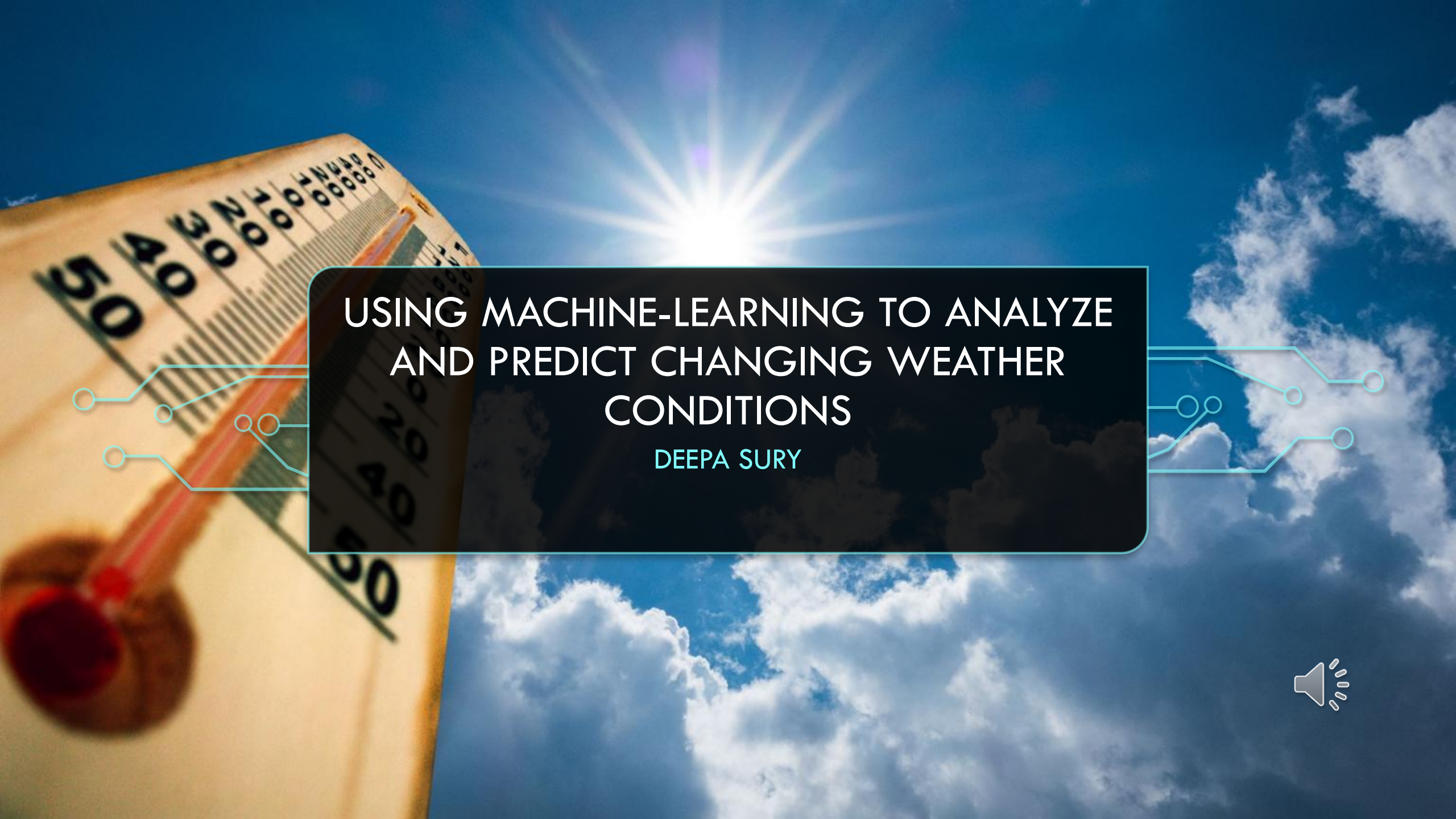


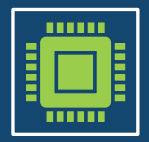
USING MACHINE-LEARNING TO ANALYZE AND PREDICT CHANGING WEATHER CONDITIONS

DEEPA SURY

OVERVIEW

- Weather conditions have been dramatically changing over the last few decades.
- **Extreme weather events** with severe consequences have become more common and are widely predicted to increase in frequency.
- This is why ClimateWins aims to improve our ability to **detect and predict changing weather conditions with the help of machine-learning.**

HYPOTHESES



A single machine-learning algorithm will not be sufficient to accurately predict weather conditions. This will likely require a combination of algorithms.

Data from certain locations may be more complete or may reflect certain patterns more strongly than others. This may lead to machine-learning algorithms overfitting the data from these locations.

Extreme weather conditions (e.g. record high or low temperatures, wind speeds, precipitation, storms) have increased in frequency during the past decade and will likely continue to do so in the future.

THE DATA

The data we are using was collected by the European Climate Assessment & Data Set project.

This data contains weather observations, such as minimum and maximum temperatures; humidity and wind speed, taken from 18 different European weather stations from 1960 until 2022.

The data has been thoroughly checked for quality and completeness; it contains neither null nor duplicate values.

POSSIBLE LIMITATIONS AND BIASES OF THE DATA

This data has been measured and collected by human specialists. If a specialist is not adequately trained to identify certain patterns or events, it could lead to **collection bias** appearing in the data.

Uneven levels and rates of economic development across regions over time could lead to varying levels of accuracy, particularly in historical data.

This data comes from a selection of 18 out of a total of 25,007 European weather stations. This poses the risk of **sampling bias** in the data, as the data from these locations may not reflect all weather developments in the region.

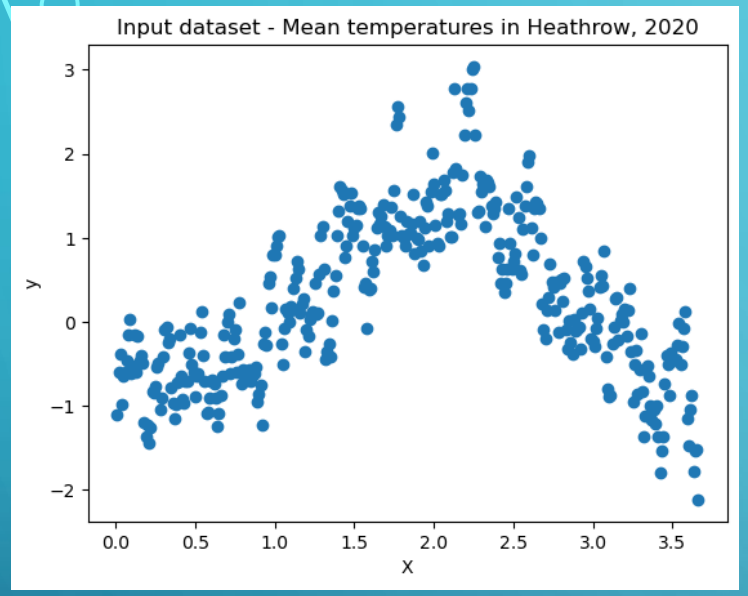
The observations and insights gathered from this data would **primarily reflect climate developments in the European and Mediterranean regions** and would therefore not be particularly useful in making predictions for the rest of the world.

DATA OPTIMIZATION

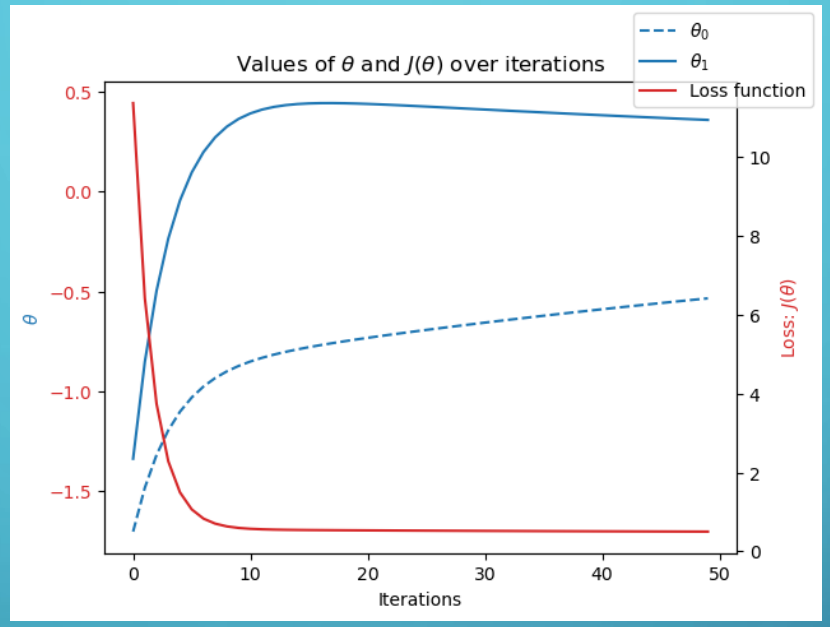
- Data optimization is a necessary step to minimize the amount of **loss** (wrong answers) produced by a machine-learning algorithm.
- The data was optimized using a **Gradient Descent** algorithm.
- This method determines the minimum loss level for the data landscape by adjusting step-size and the number of iterations.
- The following table shows the optimized data from 3 selected weather stations during 3 selected years.

Year	Location	Mean temp. min	Mean temp. max	Mean temp. range	Theta0	Theta1	Step size	Nr. of iterations
1965	Basel	-2.24	1.88	4.12	-0.5	0.5	0.1	100
1965	Heathrow	-2.43	1.64	4.07	-1	0	0.1	100
1965	Madrid	-2.08	2.02	4.1	-0.5	0	0.01	100
1992	Basel	-2.19	2.24	4.43	-1	0	0.05	500
1992	Heathrow	-2.19	2.09	4.28	0	0	0.1	500
1992	Madrid	-2.01	2.22	4.23	-1	0	0.1	500
2020	Basel	-1.68	2.31	3.99	0	0	0.01	100
2020	Heathrow	-2.12	3.04	5.16	-2	-2	0.05	50
2020	Madrid	-1.64	2.33	3.97	-2	-2	0.05	50

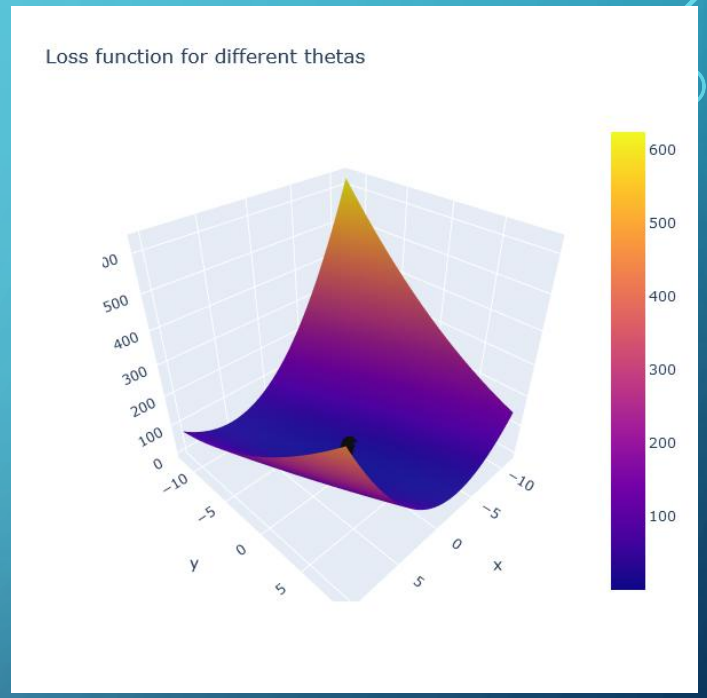
DATA OPTIMIZATION USING GRADIENT DESCENT



Plotting the mean daily temperatures recorded at Heathrow in 2020.



Gradient descent applied to mean temperature data from Heathrow in 2020. Note the loss value trending towards zero through the iterations.



A visualization of the loss function for different model parameter values.

THE ALGORITHMS

k-Nearest Neighbours (kNN)

- An instance-based, supervised machine-learning model.
- Classifies data points based on their proximity to data points it has been trained on.
- The user can adjust the number of nearest-neighbours (k) as a parameter.

Decision Tree

- A supervised classification model.
- Approaches a solution (the "leaf") through a series of increasingly-specific questions (the "branches") – similar to a game of "20 Questions".

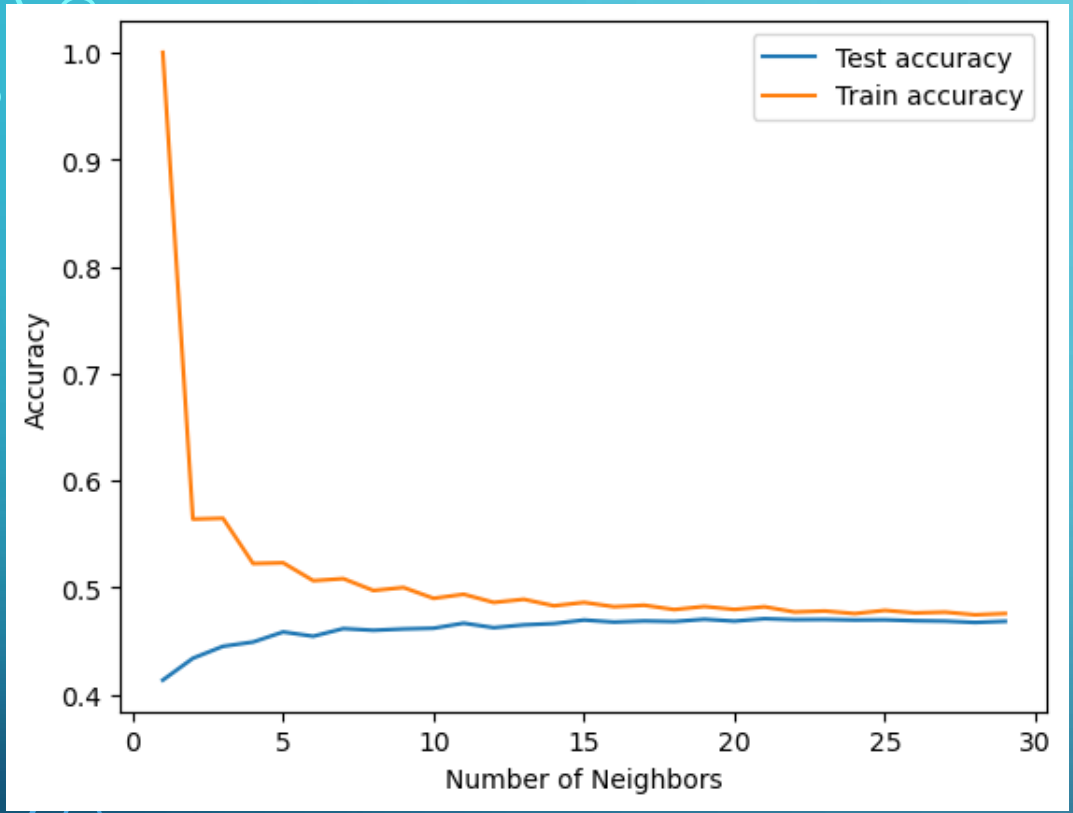
Artificial Neural Network (ANN)

- Based on the Perceptron, one of the oldest examples of machine-learning, in which inputs are multiplied by their weights to compute an answer.
- An ANN consists of up to 3 layers, each containing a certain number of inputs (nodes) multiplied by their weights.
- The user can adjust the number of layers as well as the number of nodes contained in each layer.

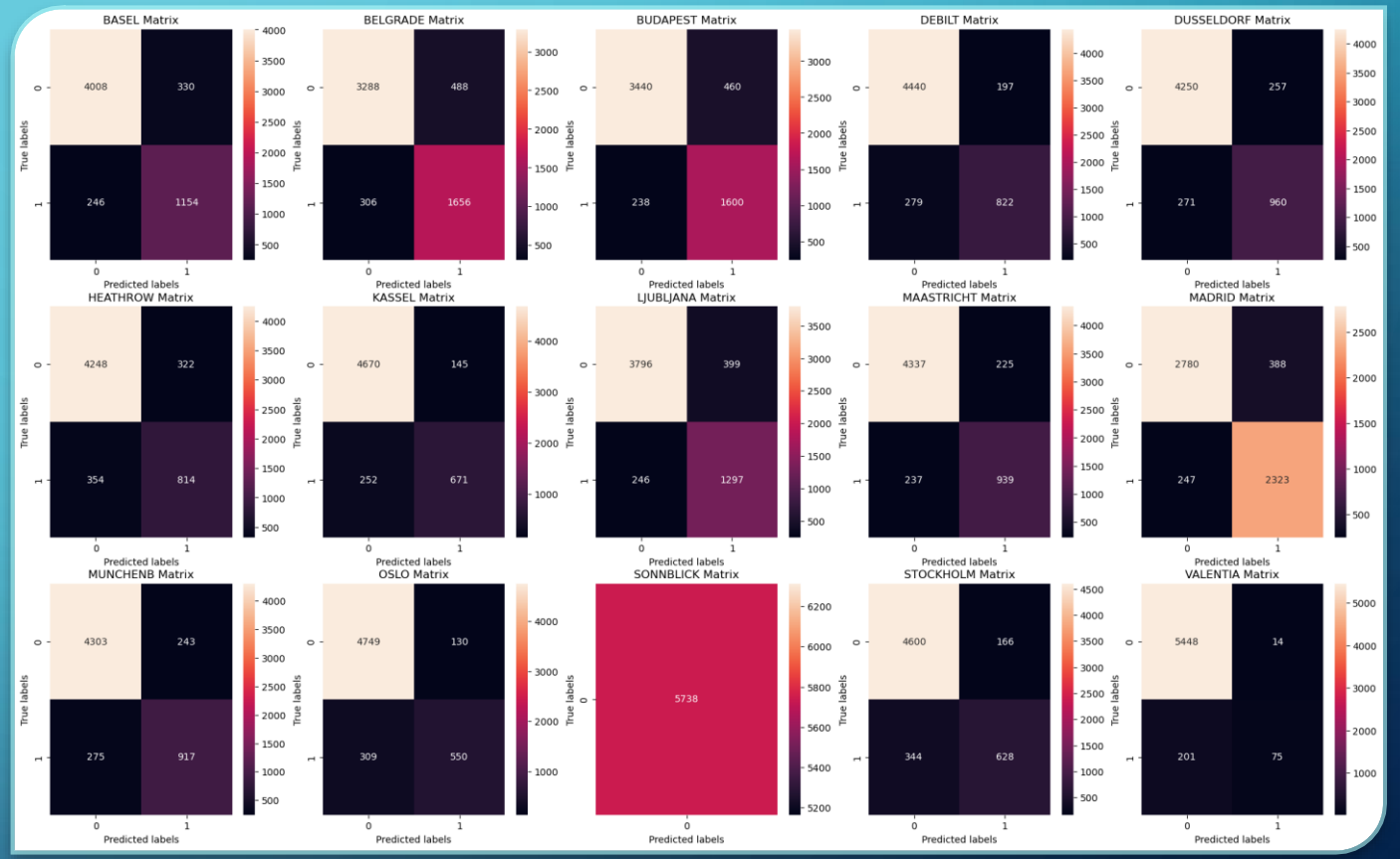
PREDICTING PLEASANT WEATHER

- A survey has been conducted daily over the past years across Europe asking the public if the weather that day is suitable for a picnic or similar outdoor activities.
- Using this data along with our ECAD dataset, we will apply different machine-learning algorithms to predict whether the weather on a particular day is pleasant enough for outdoor activities.

USING KNN ALGORITHMS TO PREDICT PLEASANT WEATHER



The accuracy of the KNN algorithm with train and test sets of data as the number of nearest-neighbours increases.



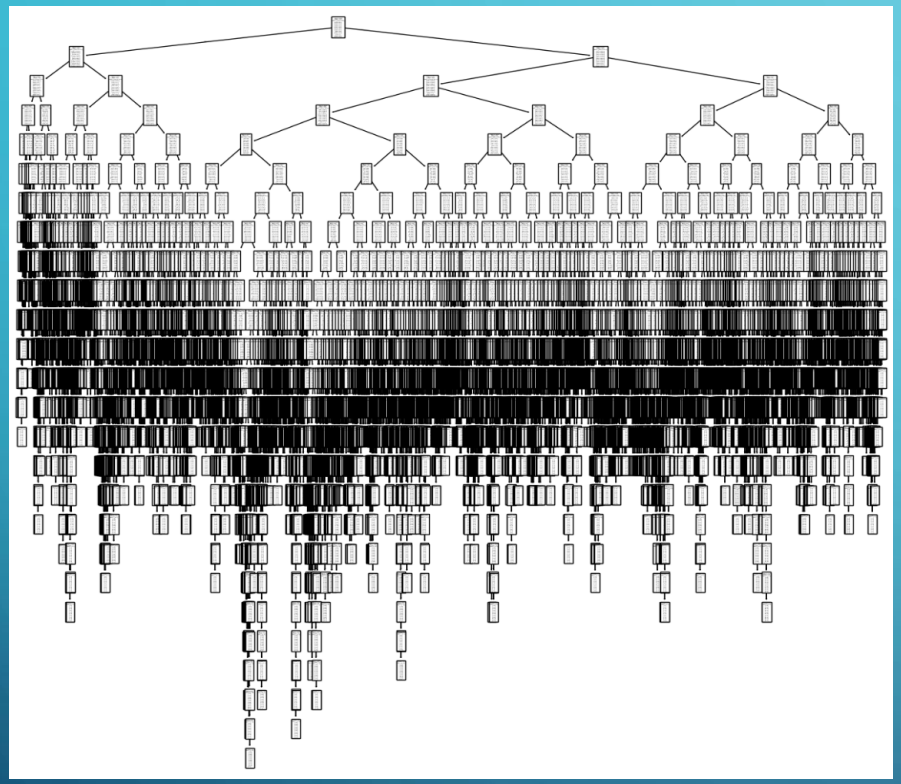
This confusion matrix shows the accuracy of the KNN algorithm in predicting pleasant weather for each European weather station.

- The KNN algorithm displayed the **highest level of accuracy for Valentia** and the **lowest for Belgrade.**
- Further analysis of the Pleasant Weather dataset showed that **Sonnblick and Valentia recorded the highest number of unpleasant weather days,** with the **Sonnblick data containing only negative reports.** The KNN algorithm displayed the highest accuracy for both locations, with Sonnblick producing 100% accuracy. **This suggests that the model works best on data with a higher number of negative answers.**

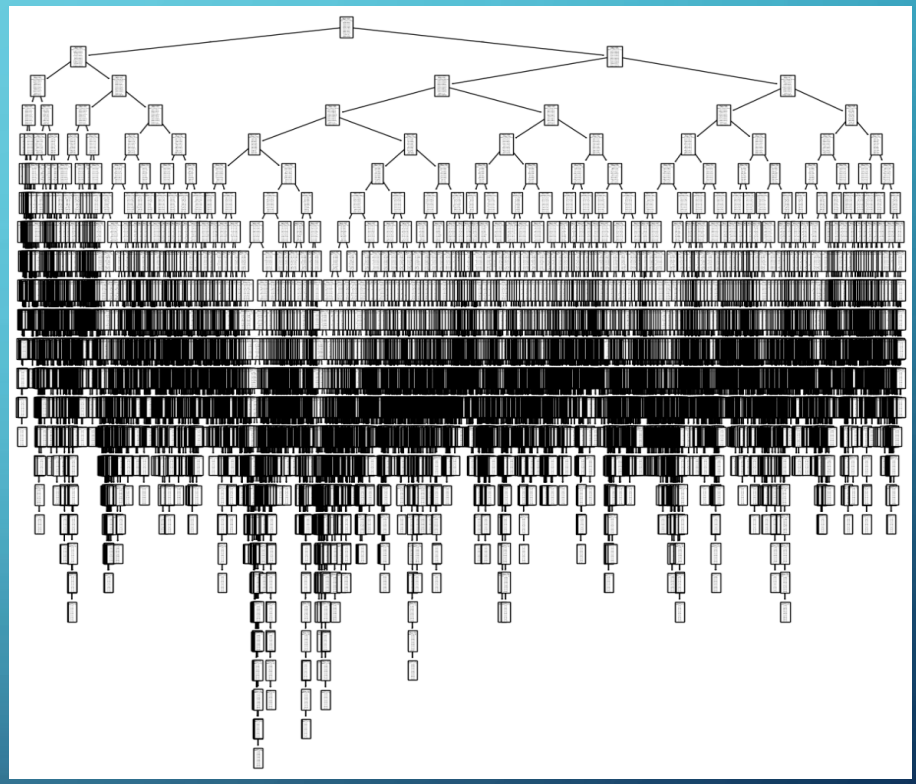
Weather Station	Accurate Prediction 0	Accurate Prediction 1	False Positive	False Negative	Total Accuracy
Basel	4008	1154	246	330	90%
Belgrade	3288	1656	**488**	306	86%
Budapest	3440	1600	460	238	88%
Debilt	4440	822	197	279	92%
Dusseldorf	4250	960	257	271	91%
Heathrow	4248	814	322	**354**	88%
Kassel	4670	671	145	252	93%
Ljubljana	3796	1297	399	246	89%
Maastricht	4337	939	225	237	92%
Madrid	2780	**2323**	388	247	89%
Munchen	4303	917	243	275	91%
Oslo	4749	550	130	309	92%
Sonnblick	**5738**	0	0	0	**100%**
Stockholm	4600	628	166	344	91%
Valentia	**5448**	75	14	201	**96%**
				Average	**91%**

PREDICTING PLEASANT WEATHER WITH DECISION TREES

Unscaled Data



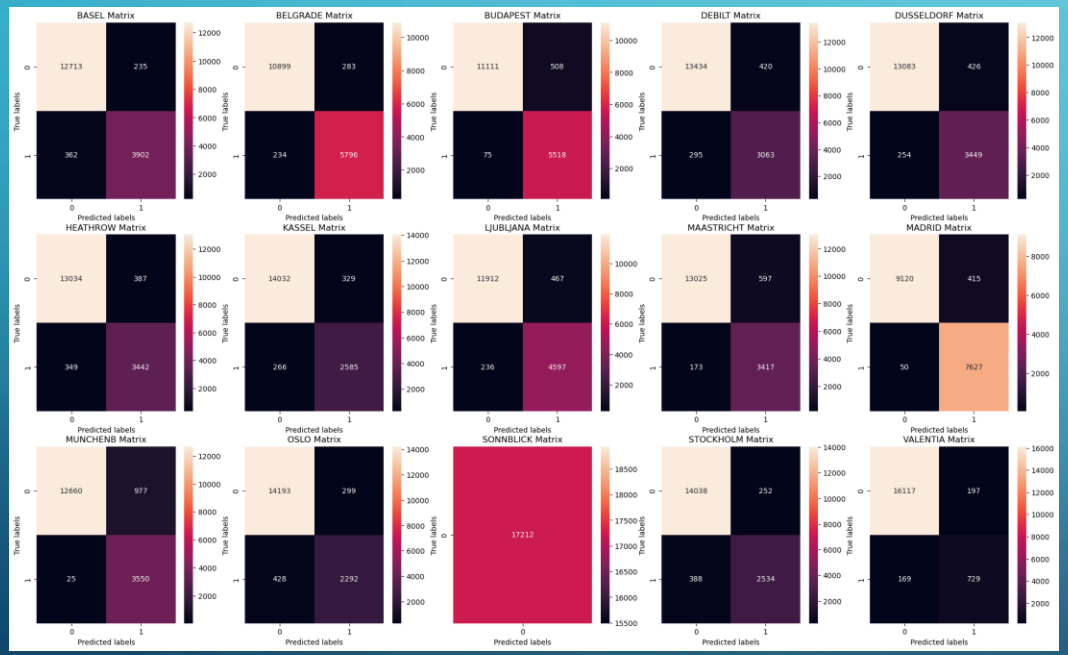
Scaled Data

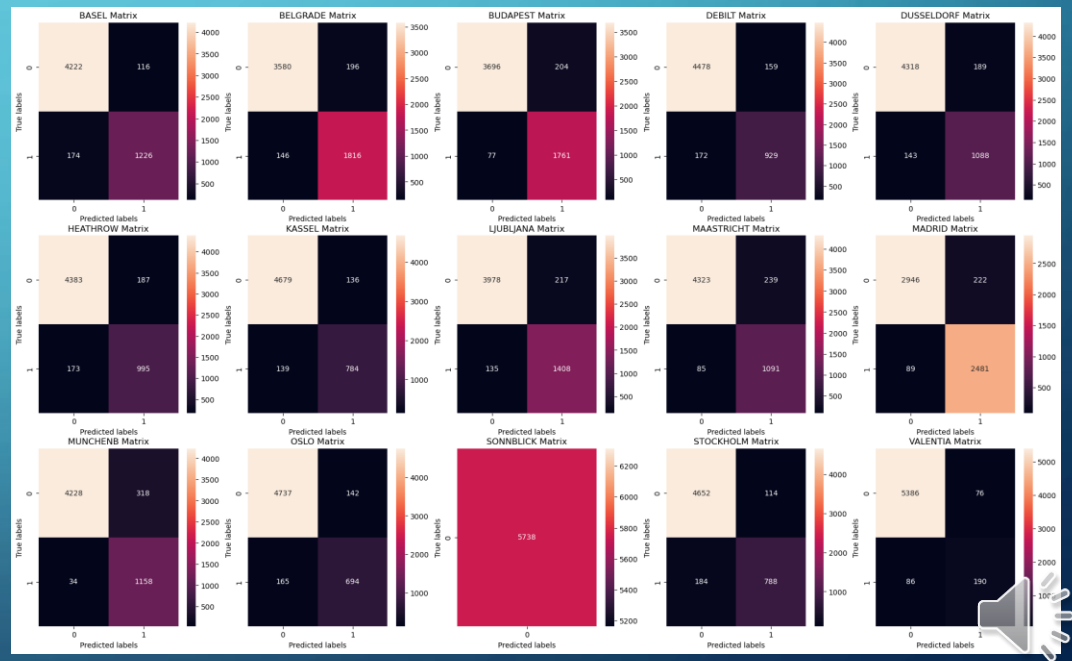


Both decision trees displayed between 56 and 61% accuracy.
Both decision trees would require pruning (removing overly-specific branches) to avoid overfitting the data.
At this point, it is unclear whether scaling the data affects the accuracy of the decision tree model.

USING ARTIFICIAL NEURAL NETWORK (ANN) MODELLING TO PREDICT PLEASANT WEATHER

- The most accurate ANN model using test data had **3 layers of 100, 50 and 50 nodes**.
- The ANN with the largest number of nodes per layer showed almost perfect accuracy with the training data, yet much lower accuracy with the test data. This means that these layer sizes resulted in overfitting the training data.





Confusion matrices for the most accurate ANN model using train (left) and test (right) data. The accuracy of this model was 64.2% (train) and 56.4% (test).

OBSERVATIONS

The **KNN Algorithm** showed the highest level of accuracy with data from locations that recorded a high number of unpleasant weather days.

The **ANN Model** with 3 layers of 100, 50 and 50 nodes, displayed the most accuracy with the test set of data. Higher numbers of nodes resulted in the model overfitting the train set of the data while showing lower accuracy for the test set.

Finally, the **Decision Tree** model would require pruning (removing overly-specific branches) to avoid overfitting.

CONCLUSION

- The **KNN algorithm** works best for predicting pleasant and unpleasant weather out of the 3 models tested, however, there is a risk of the model overfitting certain types of data.
- Due to the complex and nonlinear nature of this data, it is likely that **a combination of these models** would more accurately predict certain weather conditions.

NEXT STEPS

- Continue testing the KNN model with the data provided and, if possible, find ways to prevent or reduce the risk of overfitting certain types of data.
- Re-examine the quality of the Pleasant Weather dataset, particularly for the data collected from Sonnblick station.
- Do further research into the accuracy of data recorded in recent years compared to older historical data. Determine how much this difference in accuracy affects predictions using this data.
- Test the accuracy of other machine-learning models or combinations thereof with this data.

CONTACT

E-Mail: suryd@protonmail.com
GitHub: github.com/sryds