Weather Conditions and Climate Change with ClimateWins

Melanie Houston

October 19, 2024

Objective

ClimateWins is wanting to use machine learning to help predict the consequences of climate change around Europe and, potentially, the world

Hypothesis

Supervised machine learning can be used to identify weather patterns contributing to severe weather such as flooding or a heatwave

Supervised machine learning can predict daily temperatures based on historical data

Supervised machine learning can predict changes in temperature and number of weather events due to climate change

Data Sets

Data is collected by the European Climate Assessment & Data Set project. Weather observations from 18 different weather stations across Europe which contain data ranging from 1800s to 2022. Recordings exist for almost everyday with values for temperature, wind speed, snow, global radiation and more.

Data Biases

- **Sampling bias:** Using only 18 weather stations for this project, as according to the European Climate Assessment & Data Set project, there are currently 23789 meteorological stations throughout Europe and the Mediterranean.
- **Human bias:** There is a potential for human bias or researcher bias depending on the experience level and training of the individuals collecting the historical data.
- **Temporal bias:** Equipment used to measure data decades ago may be less accurate than it is today. The interpretation of the information collected could lead to inaccurate conclusions.

Optimization

Optimization is the process of finding the best solution for a particular problem. We lower the risk of error and improve the accuracy of a model to boost efficiency and reliability.

Gradient descent is one of the simplest ways to find a local minimum (or valley) of the loss function. This optimization technique is useful in both linear and nonlinear cases. It's known as a *first-order* optimization method because it uses the derivative of the function at any point. We applied the gradient descent to find the minimum error through a number of iterations and step sizes.

Accuracy

We applied a few different methods of machine learning to determine the best choice for ClimateWins. Let's see which method has the greatest rate of accuracy.

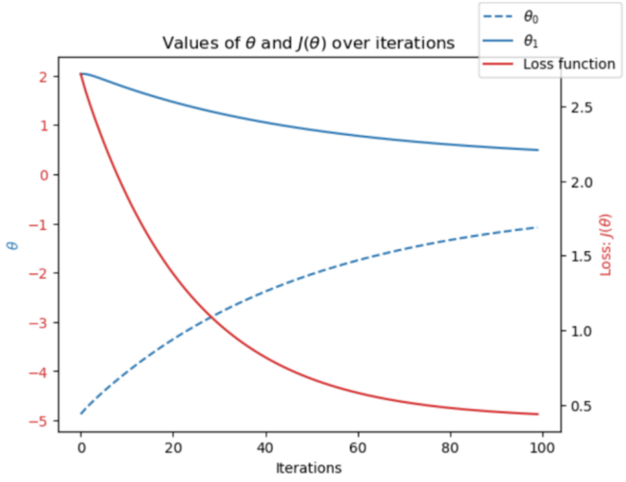
We used a **Decision Tree** but this was not the best fit. There were too many decision nodes for proper visibility.

We used the **Artificial Neural Network (ANN)** with low success, there was room for improvement.

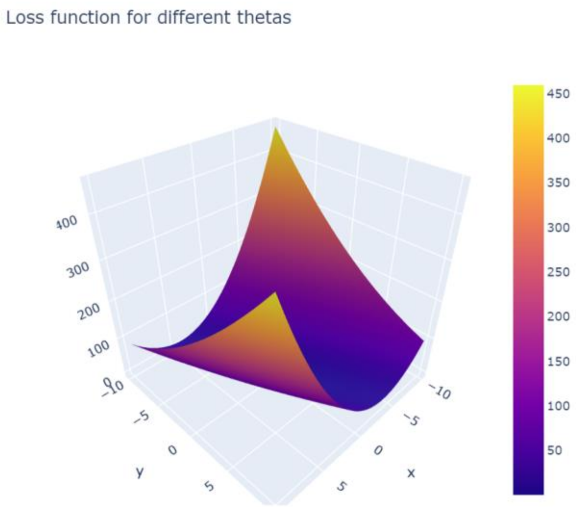
The **KNN model (K-nearest neighbor)** was the right choice. It fitted the data set well and had an accuracy of 88%.

The next slides contain some visualizations.

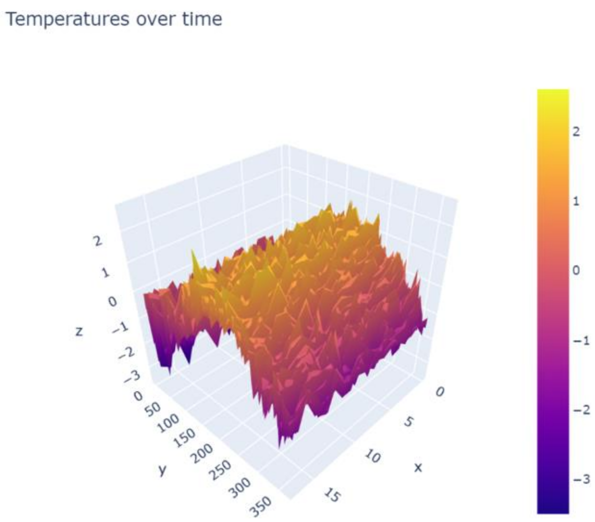
Gradient Descent



After we adjusted the theta_init=np.array from [-10, -5] to [-5, 2] , we can see this plot of loss, $theta_0$, and $theta_1$. With a Loss J(0) scale, 0.5 to 2.5, we can see the loss in red is trending toward 0.



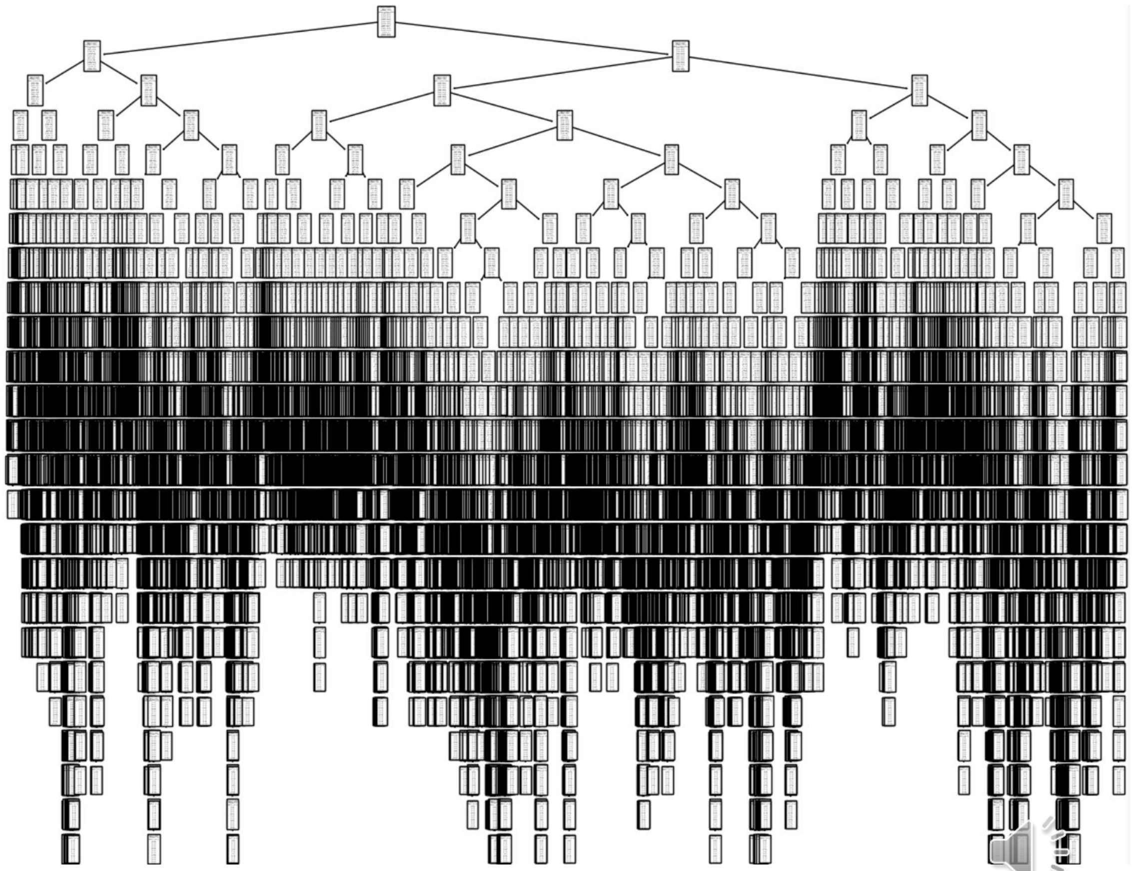
Running the optimization above, close to the objective, we found where the black line ends near the lowest X/Y/Z coordinates near the minimum in the graph above and use a step size of alpha= 0.1



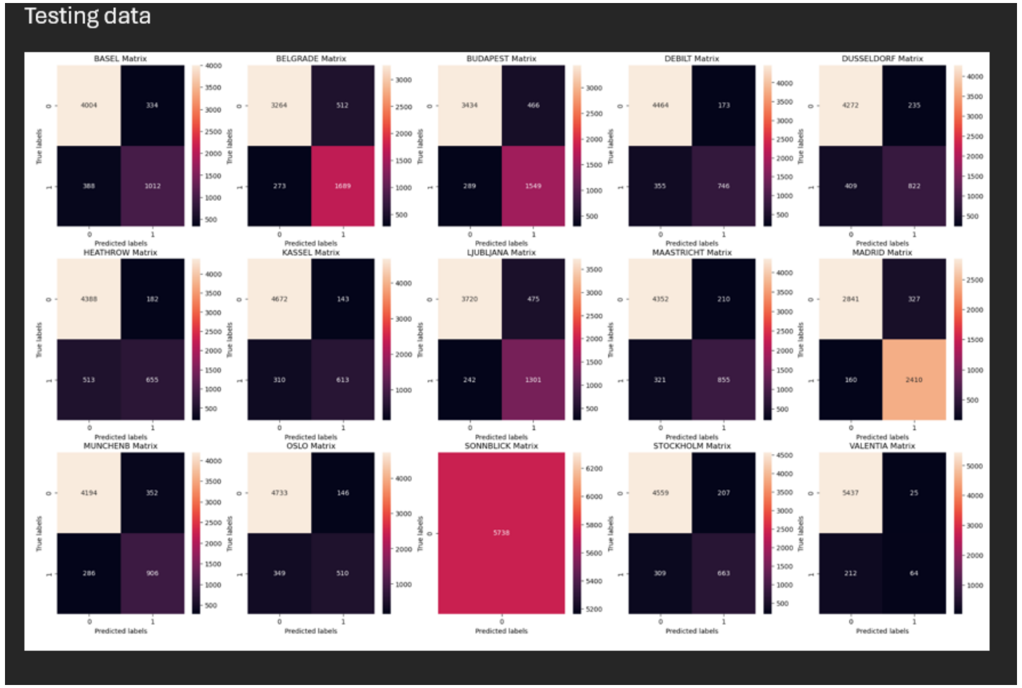
Plot for ALL weather data for all stations for a year. We can see the cool/cold temperatures during the first few and last few months of the year, with higher temperatures in the summer

Decision Tree

As we can see, this is not ideal image or model, there are too many decision nodes so we can not determine its accuracy, so let’s move on!



ANN (Artificial Neural Network)



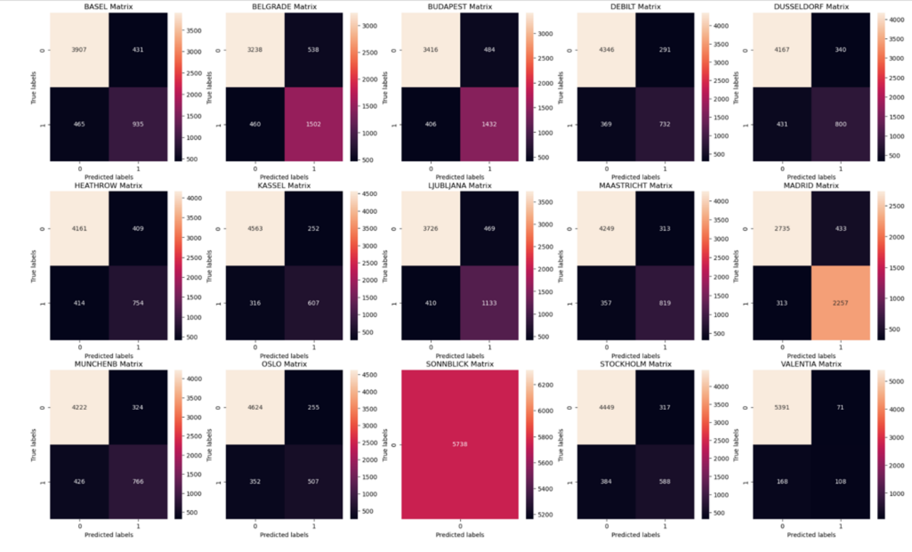
The Artificial Neural Network is a collection of interconnected algorthims that process information in response to external output

Training and testing confusion matrices for the artificial neural network model shows 48.4% and 48.5%... these are low accuracy rates...

K-Nearest Neighbor

The KNN model uses proximity between data points to make classifications or predictions and decides which group that data point is in based on how many of its neighbors belong to each category.

Using the formula (TP+TN)/ (TP+TN+FP+FN), we calculated the weather prediction accuracy rate of the model and the average was 88%. Overall, I would say this model predicts the weather reasonably well, but there is definitely room for improvement. Using more weather stations and varied weather events with time can help improve accuracy of the model. Refining the model and conducting cross-validation will also help improve the accuracy.



Weather Station	True positive	True negative	False positive	False negative	Accuracy rate
Basel	3907	935	465	431	0.84
Belgrade	3238	1502	460	538	0.83
Debilt	4346	732	369	291	0.88
Dusseldorf	4167	800	431	340	0.87
Heathrow	4161	754	414	409	0.86
Kassel	4563	607	316	252	0.90
Ljubljana	3726	1133	410	469	0.85
Maastricht	4249	819	357	313	0.88
Madrid	2735	2257	313	433	0.87
Munchenb	4222	766	426	324	0.87
Oslo	4624	507	352	255	0.89
Sonnblick	5738	0	0	0	1.00
Stockholm	4449	588	384	317	0.88
Valentia	5391	108	168	71	0.96
				Average:	0.88

Conclusions

The **KNN model (K-nearest neighbor)** was the best choice to best predict weather. It fitted the data set well and had an accuracy of 88%.

The decision tree didn’t work for this dataset and the Artificial Neural Network (ANN) was also not good a choice due to low accuracy rates.

Supervised machine learning has a place in predicting weather patterns, but algorthims can not make human insights. There is potential here, but there is still much to do and answers may change with the next deliverable.

Next Steps

Continue testing with unsupervised and supervised machine learning

Explore the option of pruning the decision tree

Continue to search for better or alternative algorithms to lead us to discover more patterns

Thank you for watching!

Melanie Houston
October 2024