Climate Wins Weather Analysis

Predicting Weather Conditions and Climate Change

Rob Thomas Schassler

May 2015

Introduction and Background

Climate Wins is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world. However, it's hard to grasp how everything is changing in the world at once, so ClimateWins utilizes machine learning towards this end.

It is concerning that there has been an increase in extreme weather events in mainland Europe--especially in the past 10 to 20 years—but with historical data on weather patterns, machine learning can be used to predict and plan for adverse weather.

Hypotheses

1. Machine learning algorithms can be utilized to accurately predict weather conditions across mainland Europe.

2. Accuracy of predictions will vary between geographic locations, due to climatological variables outside the scope of this analysis.

3. Machine learning results can support correlative factors in climate change, such as temperatures rising leading to greater adverse weather conditions

Data Sourcing and Profile

Data was collected by the European Climate Assessment & Data Set Project by compiling yearly weather observations from weather stations across mainland Europe.

Dates in the total dataset range from the 1800s up to 2022, with daily observations from 1960 to 2022.

Weather observation variables from each station includes temperature, wind speed, precipitation, humidity, snowfall, and global radiation.

Data Bias and Accuracy

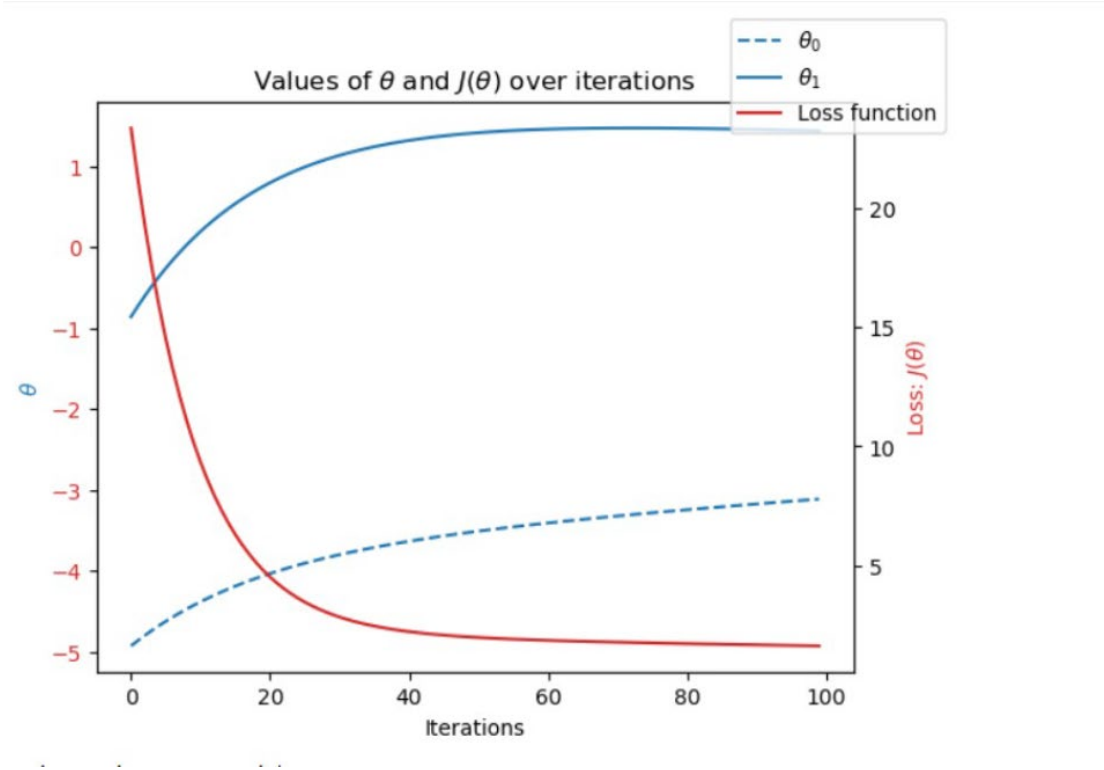
There is potential for temporal bias; data from earlier years may be less accurate due to outdated collection methodology.

There is a potential for historical bias; with the large range in years collected, recent or more acute changes may be overlooked in the algorithm.

There is a potential for selection bias; data is collected from weather stations spread across a wide geographic range and may be over representative of high population regions.

Optimization

- First the dataset was scaled to diminish the effects of outlying data
- Gradient Descent function was used to optimize the data
  - Specifically, gradient descent method adjusts parameters by testing them towards converging on 0. The adjusting of the parameters allows for a more accurate model because it reduces the loss function over consecutive iterations



Supervised Learning Algorithms

K-Nearest Neighbor:

- Makes predictions by calculating the distance between new data points and all other data points in the training set. Nearby data points are grouped based on the number of neighbors in each group.

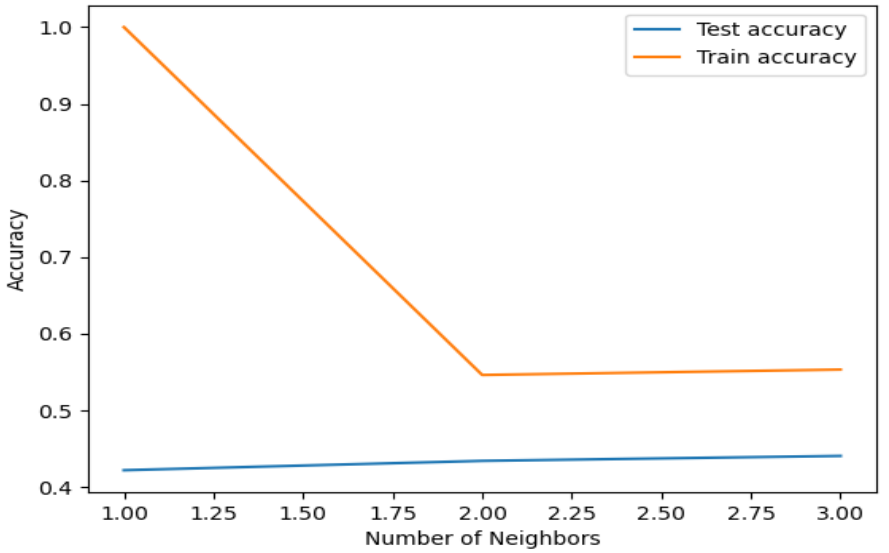
Decision Tree:

- Makes predictions by answering a series of questions about the features of the data. The tree consists of nodes and branches in which new data points follow the flow of the tree based on its best fit.

Artificial Neural Network (ANN):

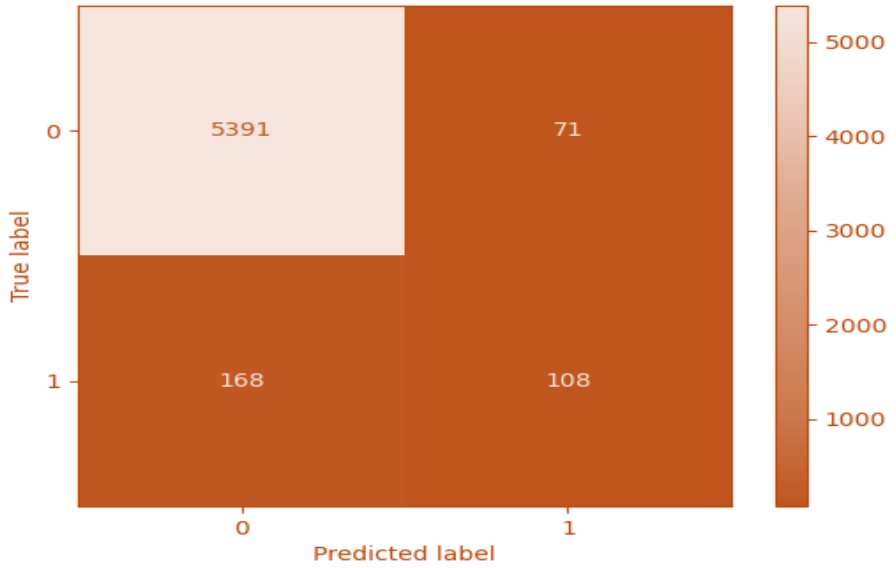
- Mimics the way the human brain processes information. It consists of layers of interconnected "neurons" that learn patterns from data through training

K-Nearest Neighbor Method



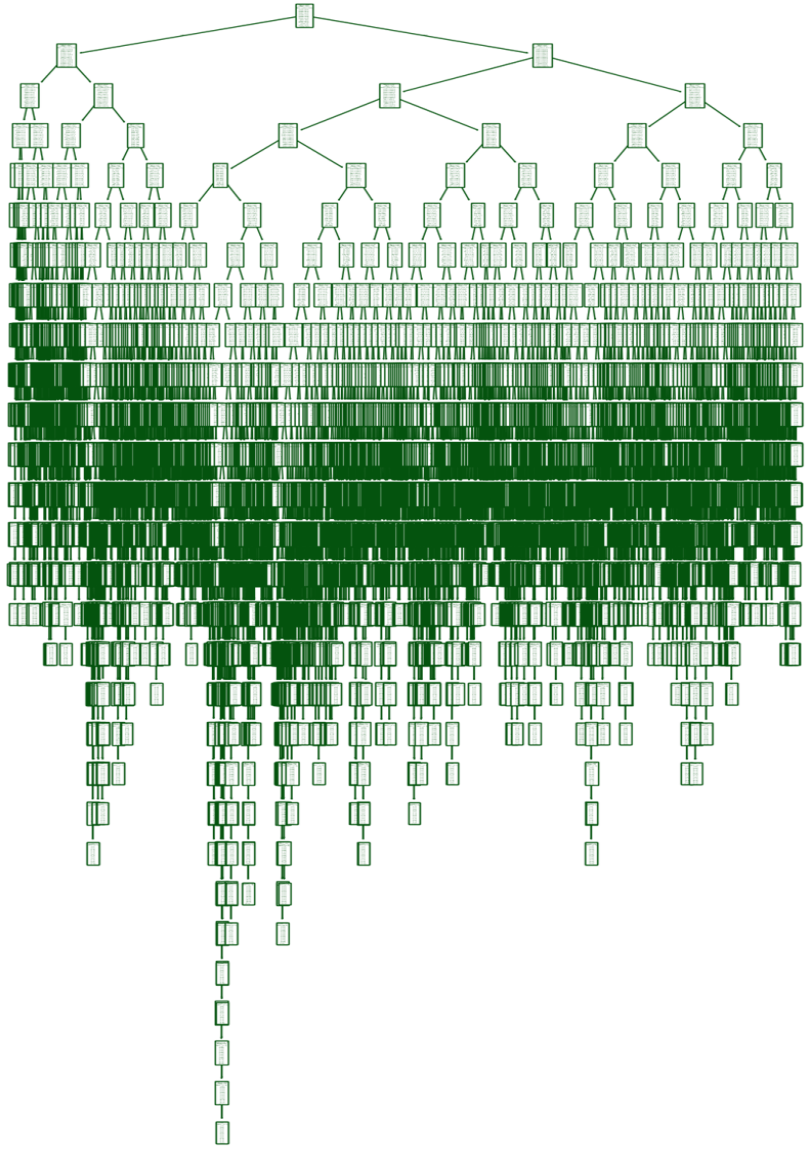
- The model has varied accuracy at different stations, however across all stations, the accuracy rate--number of accurate predictions vs. wrong predictions—was never below 80%.

- This suggests the algorithm was successful, as it is predicting correctly most of the time. The confusion matrix shown demonstrates the algorithm's predictions—with '0' being an 'unpleasant day' and ' being a 'pleasant day'



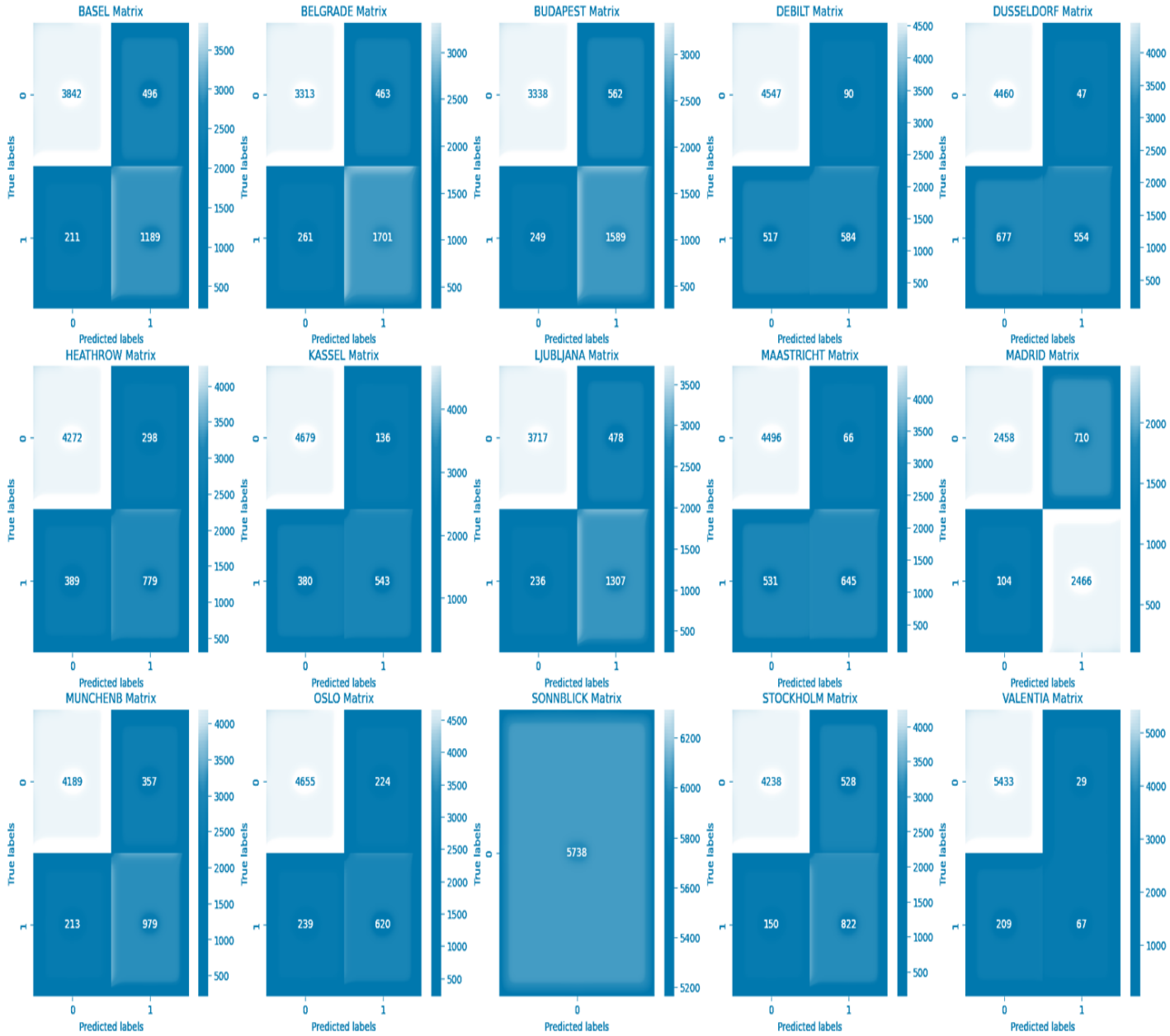
Decision Tree Method

- The decision tree model predicted an accuracy of about 47%.
- In its current format, the decision tree is a poor method of accurately predicting the weather data, as it is too complex to interpret.
- Steps *could* be taken to “prune” the decision tree. Now it is technically overfitting and classifying too rigidly. Pruning would result in a leaner tree by only allowing a certain number of splits.



ANN Method

- The ANN model achieved around 50% accuracy.
    - Though, this was at a max hidden layer size of 40 and number of iterations capped at 2,800.
    - Raising the hidden layer input would probably train higher accuracy results for the ANN model
- The confusion matrix shows the number of accurate and inaccurate predictions at each location for unpleasant days (0) and pleasant days (1).
- Correct guesses for unpleasant days are recorded in the top left of each matrix and correct guesses for pleasant days is in the bottom right. The other diagonals are incorrect guesses.

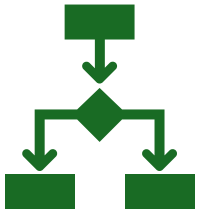


Algorithm Results Recap

The KNN model at 80% accuracy, performed the best in predicting days with pleasant vs. unpleasant weather.

However, an accuracy of 80% is probably not high enough to serve the purpose of Climate Wins' mission claim to accurately predict unpleasant weather.

The data set's other features besides temperature are more unpredictable, affecting the accuracy in way not accounted for by the algorithms.



The ANN models were achieving around 50% and the decision tree model was around the same (about 47%).

However, more substantial changes of the control parameters would almost certainly train higher accuracy results for the ANN model.

Though it will require more testing of the control parameters than was done in the comparison thus far, the ANN model can generally be trained up to around 95% - 100%

The controls of the ANN model should be further tested to train the accuracy to at least 90%.

Although the KNN model was highest in predicting pleasant or unpleasant weather thus far, the ANN model is generally better suited to complex datasets like this with more "noise" included.

Summary

01

MACHINE LEARNING ALGORITHMS CAN INDEED BE UTILIZED TO PREDICT WEATHER TO AT THE VERY LEAST 80% (WITH PROBABLE POTENTIAL TO ACHIEVE A HIGHER ACCURACY IN THE 90^{TH} PERCENTILE.

02

ACCURACY DOES IN FACT VARY BASED ON GEOGRAPHIC LOCATION AND WEATHER VARIABLES BESIDES TEMPERATURE LIKE WIND VELOCITY OR HUMIDITY.

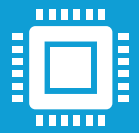
03

MACHINE LEARNING CAN DEMONSTRATE CLIMATE CHANGE BY IDENTIFYING SUMMATIVE CHANGING WEATHER CONDITIONS OVER TIME.

Next Steps

Build in more datapoints by including more of the weather variables besides temperature into the analysis

Continue working with the control parameters of the ANN algorithm to achieve accuracy results above 90%



Employ unsupervised machine learning algorithms (and specifically deep learning) to observe further potential towards deploying a system that will continue to learn and train towards higher weather prediction accuracy as climates continue to change in real time

Thank You!

For further questions, please contact me at:
robschassler@gmail.com

To view the scripts and datasets used, please visit my github repository for this analysis at:

https://github.com/rschassler/ClimateWins-Machine-Learning