# ClimateWins: Assessment of Machine Learning Tools

Sydney Johnson

December 2024

# Objective

- Use machine learning to help predict the consequences of climate change for a European nonprofit organization.
- ClimateWins is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world. This project includes determining which areas can use machine learning, looking at data that already includes answers so that machine learning can be trained to recognize similar answers (supervised learning), and communicating what was found and how it will affect ClimateWins' strategy.

# Hypotheses

The historical weather data from the past 10 to 20 years can be used to build a model that can predict conditions for the future.

Machine learning can be used to predict whether conditions will be favorable on a given day.

Optimization algorithms can be used to determine local and global temperature minimums and maximums,

# Data Source

The data used for this project is from the European Climate Assessment & Data Set Project and has been collected daily from 18 different weather stations across Europe, spanning from the late 1800s to 2022. These stations have recorded data on temperature, wind speed, snow, radiation, cloud cover, and more.

# Data Accuracy

Like all data, there is a potential for bias in this set, but that bias is relatively minimal. The values recorded here are all objective measures that have been consistently recorded. Any potential bias comes from the technology used to measure these variables and the way they have changed since the late 1800s.

There are also several weather stations whose data does not span the entire time at which we are interested in looking (Gdansk, Roma, and Tours), which we'll be excluding from our analysis.

# Classification-Based Machine Learning Models

Three types of models were examined to see how they work with this data: K-Nearest Neighbor (KNN), Decision Tree, and Artificial Neural Network (ANN)

# Algorithms and Success Metrics

- KNN is a good option for this project because we have two clearly labelled groups (pleasant and unpleasant days) into which we are trying to classify our data
- A decision tree may be a good option because it is easy to interpret and helpful in understanding the importance of different features
- ANN is a great option because of its flexibility and ability to manage a large amount of non-linear data

Two metrics will be used to determine the success of the test models:

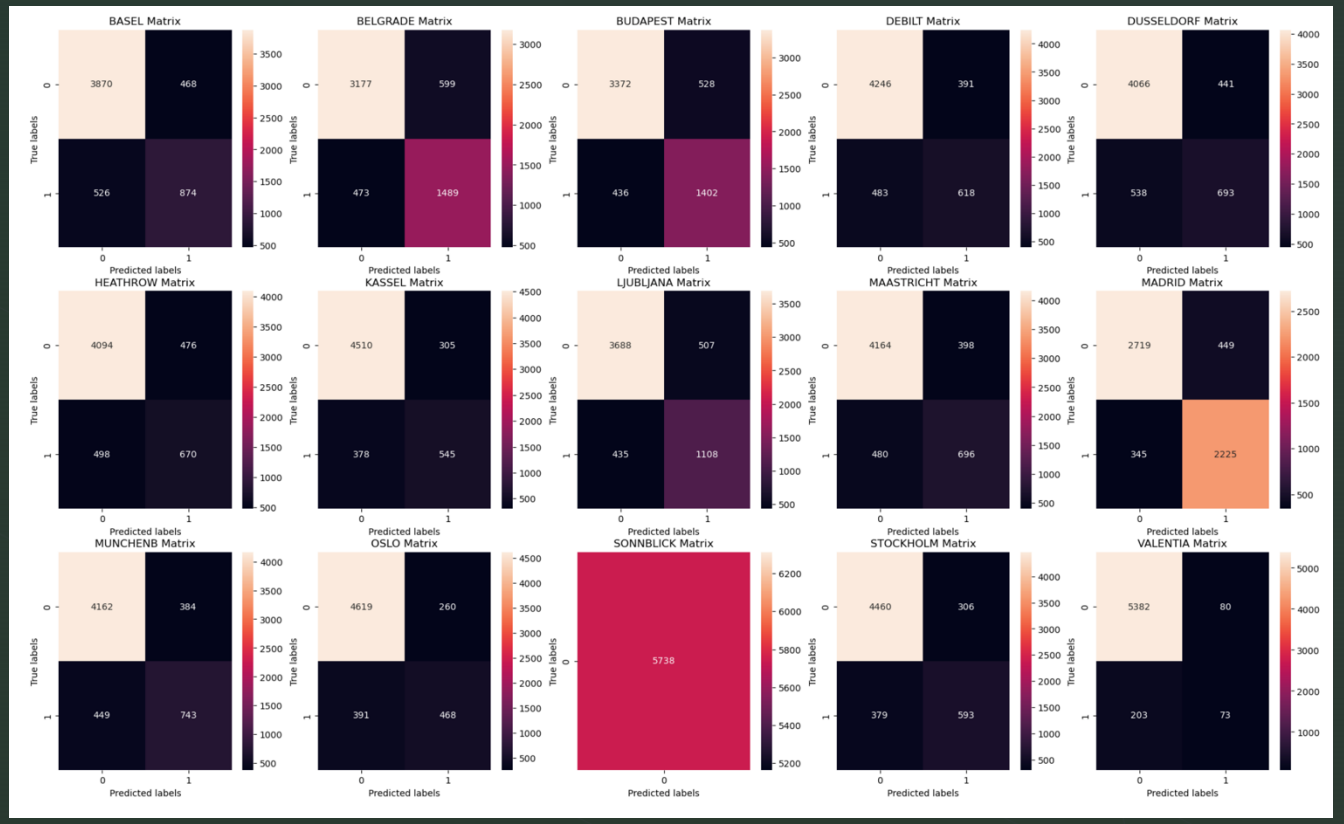
- Test Accuracy
- F1 Score (a metric based on precision and recall)

# K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is a classification-based algorithm in which the user parameter adjusts the number of neighbors to which it compares each data point.

Accuracy: 0.426

F1 Score: 0.678



# Decision Tree

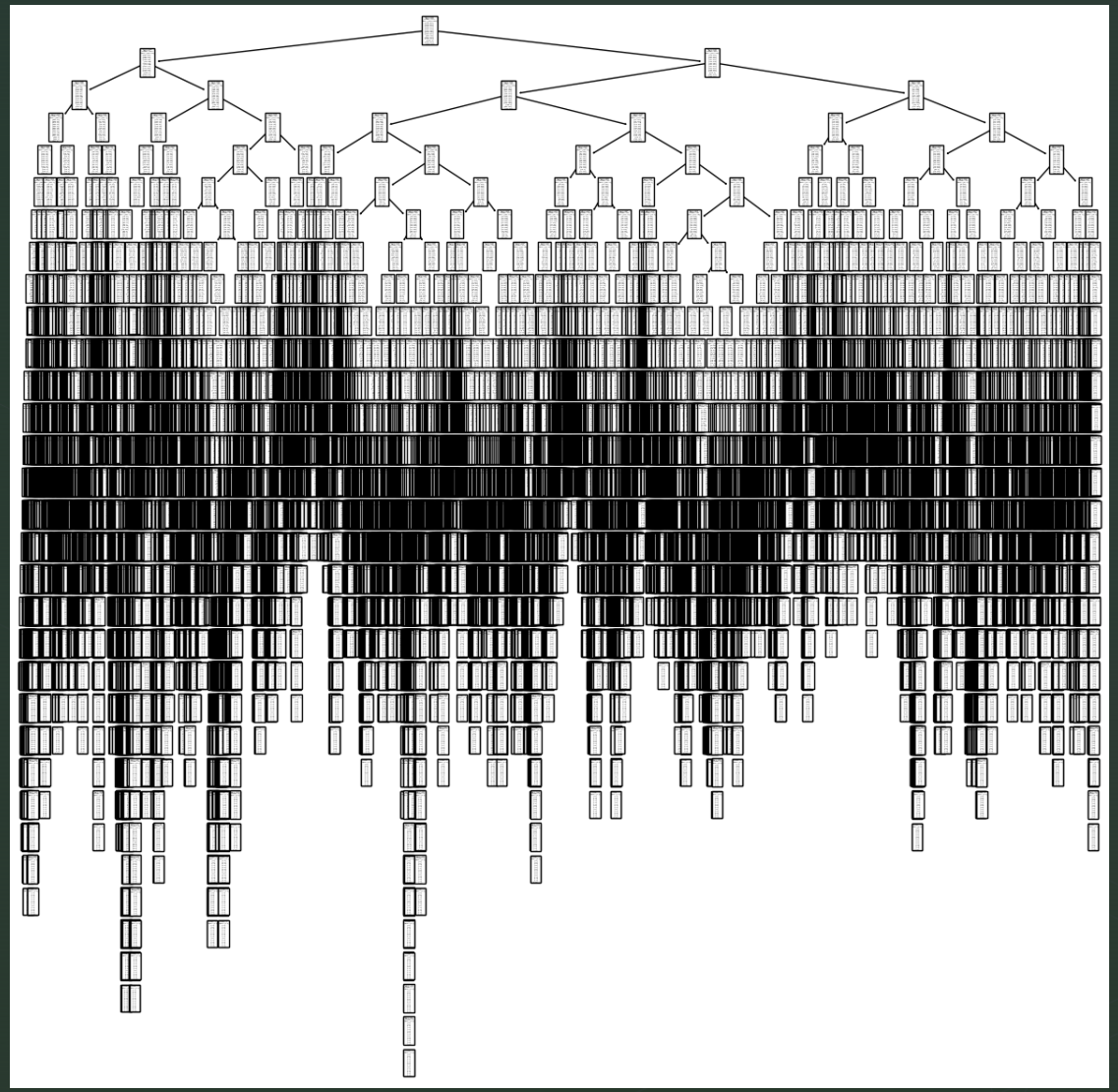
A decision tree is a classification-based algorithm that consists of branches that split a certain number of times and end in leaves.

If the number of branches is too high, the model tends to overfit to the testing data.

This can be solved by pruning the tree to make the algorithm leaner.

Accuracy: .406

F1 Score: .627



# Artificial Neural Network

This confusion matrix is for the most successful of the ANN models that were run

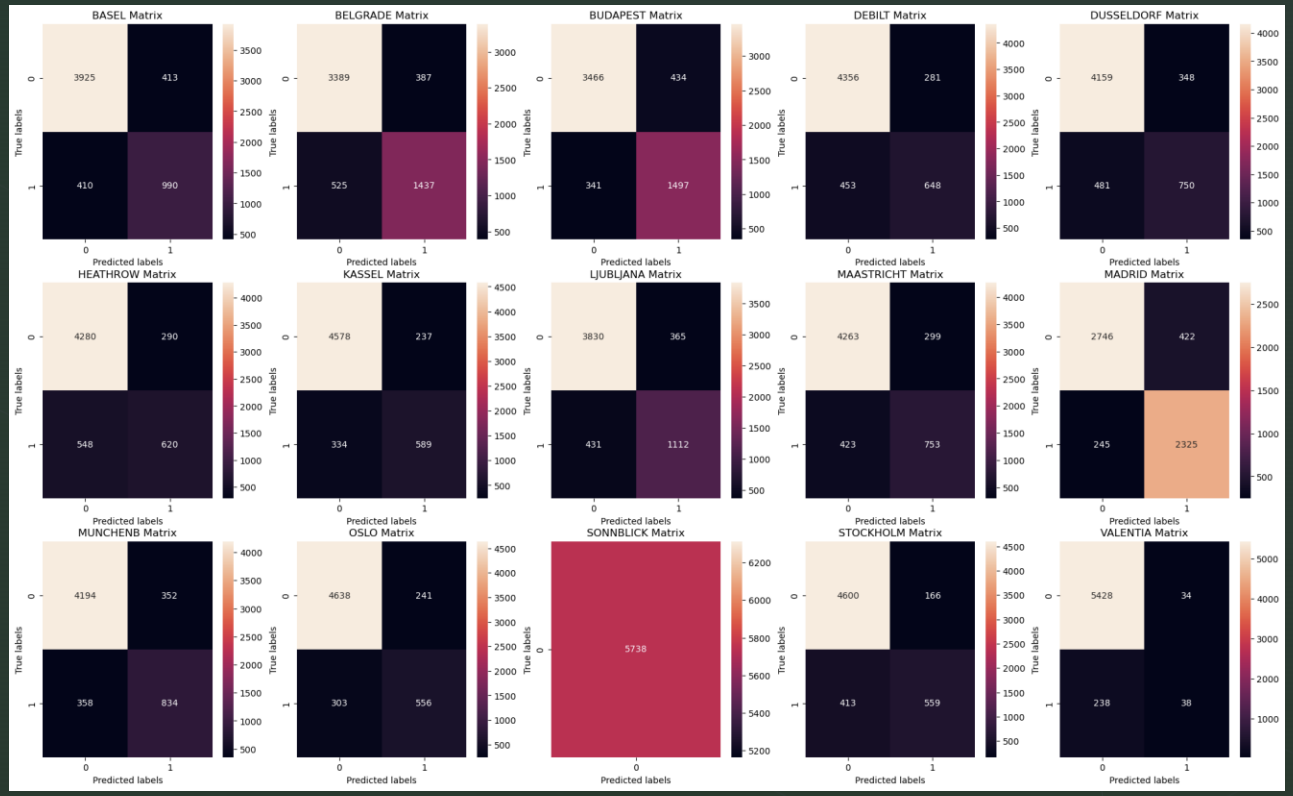
Artificial Neural Network (ANN) algorithms give intuitive results but use a lot of processing power.

Highly customizable in terms of learning rate, maximum epochs, error percentage, and number of nodes and layers.

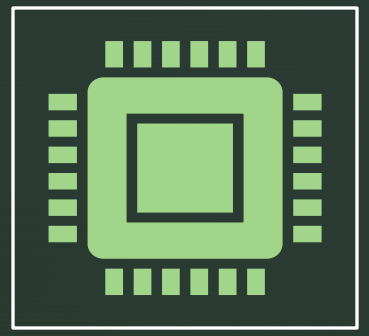
Accuracy: .450

F1 Score: .735

MLPClassifier(hidden_layer_sizes=(100, 50, 25), max_iter=500)



# Summary



The best option for a model moving forward will be an Artificial Neural Network model. Out of 4 models that were tested, the accuracy scores (up to .453) and F1 Scores (up to .735) were significantly higher than other options. With some work, these models could probably be significantly improved.

The decision tree will have to be reevaluated after pruning

It should be possible to predict conditions using machine learning based on the data and algorithms we have access to.

- Thank you!

  Please see my GitHub repository for details on the code and reach out with any questions.

sydneyruthj@gmail.com