Data Specialization

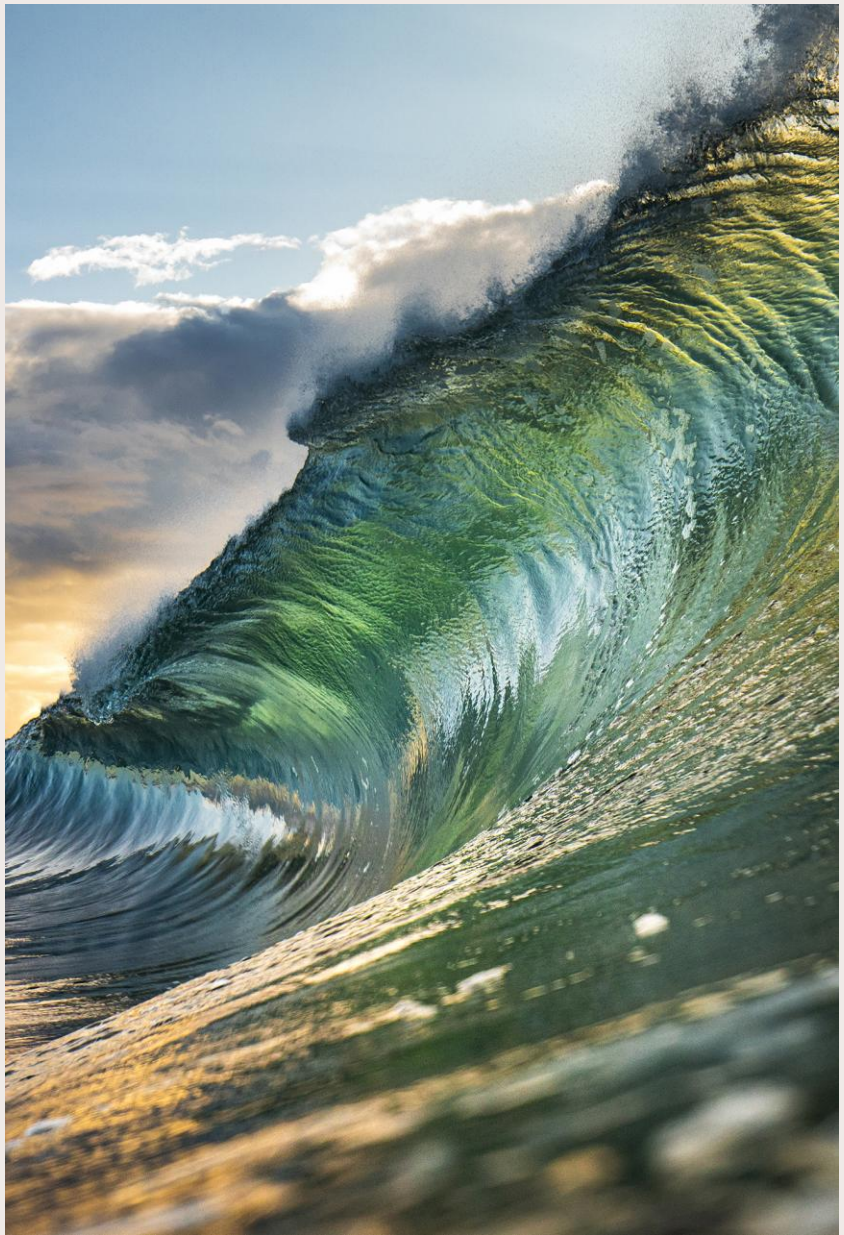
Weather Conditions and Climate Change with Climate Wins

November 20, 2024

Kendra Jackson

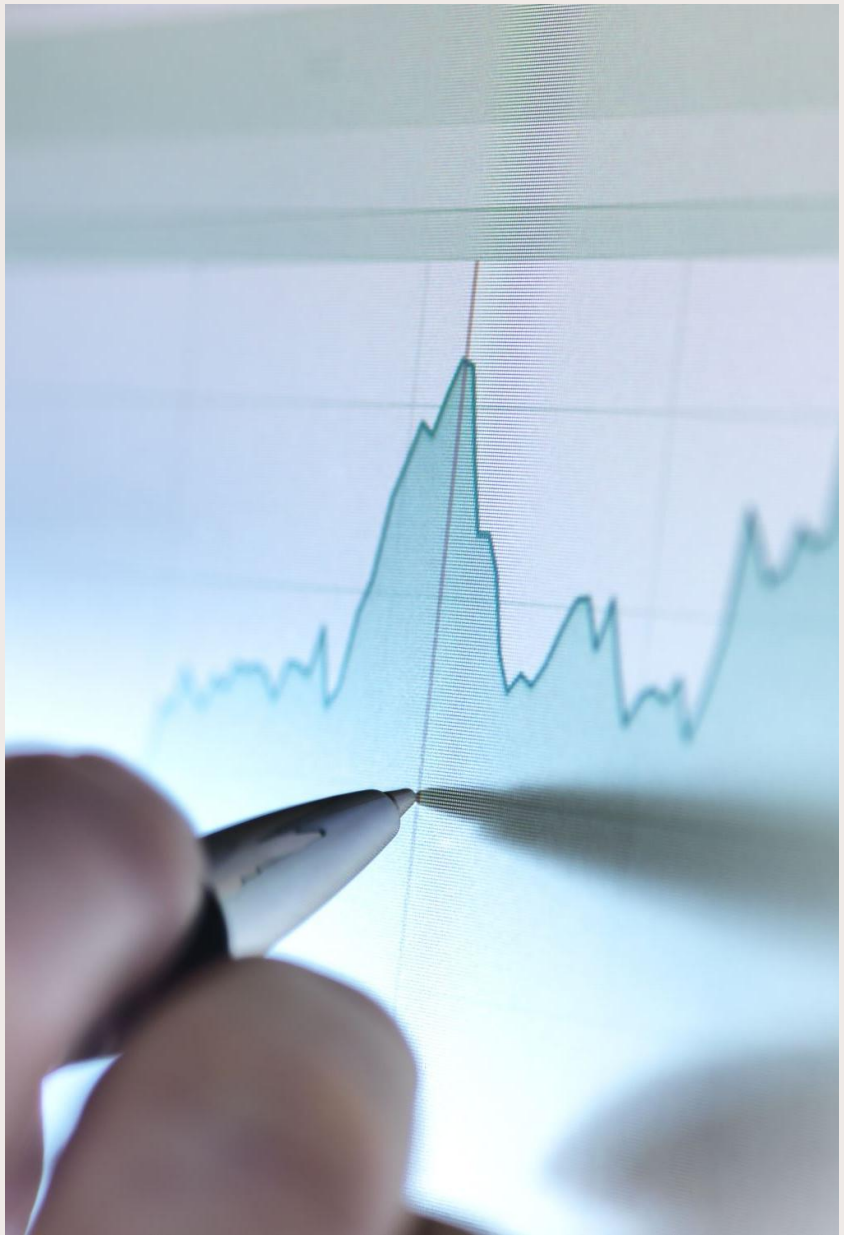
Objective

The company, ClimateWins, is a nonprofit European organization. Their main concern is an increase in extreme weather events, especially those within the previous 10–20 years. However, ClimateWins wants to incorporate machine learning into their analysis of climate data. Their hope is to utilize historical data and machine learning to predict the consequences of climate change, including extreme weather events, within Europe and, potentially, the world.



Hypothesis

1. Machine learning models trained on historical climate data can accurately predict an increase in average daily temperatures across Europe within the next 5 years.
2. Machine learning algorithms can identify patterns and correlations between different extreme weather conditions (Ex. Heatwaves and wildfires, heavy rainfall and flooding) and predict the likelihood of one event occurring given the presence of another
3. Machine learning models utilizing real-time and historical weather data can effectively predict the occurrence, intensity, and location of extreme weather events (ex. Floods, typhoons, droughts) with a higher accuracy than traditional forecasting methods
4. Warmer temperatures correlate positively with occurrence of pleasant weather days while colder temperatures, winds, snow, rain, and the like correlate to unpleasant weather days

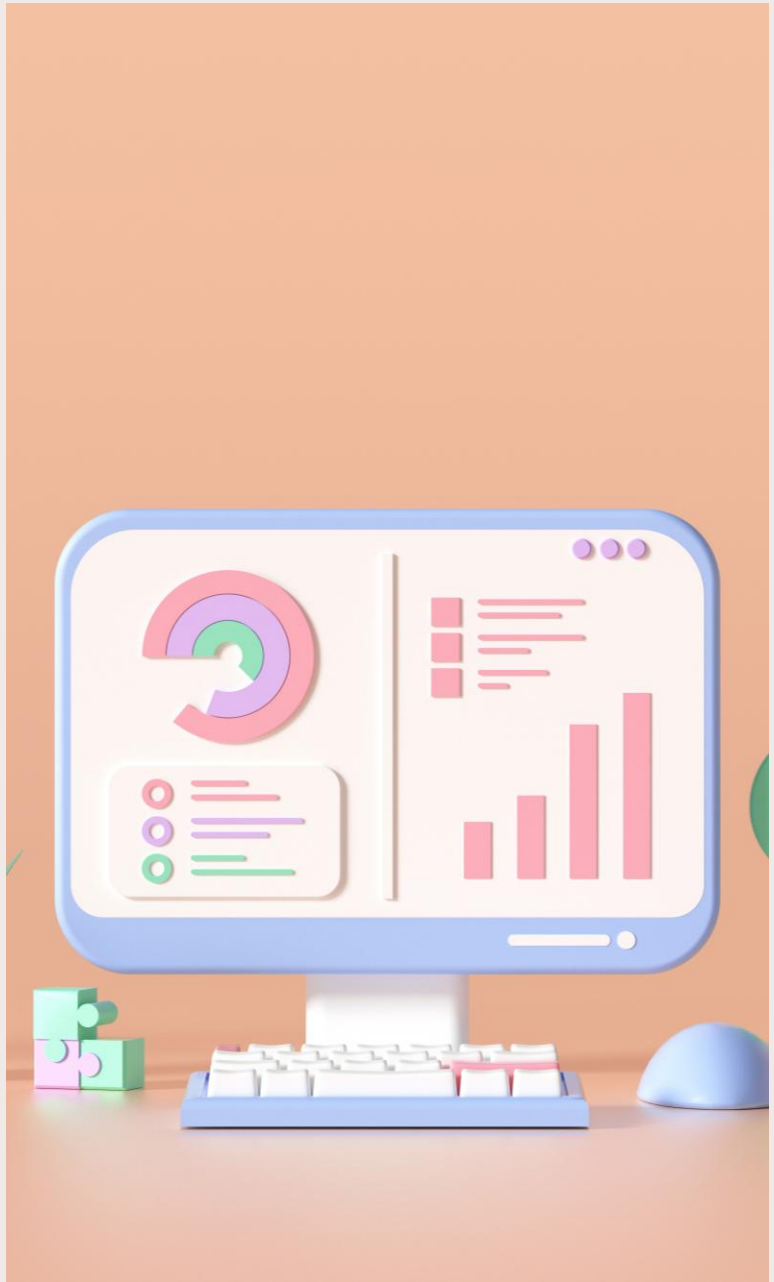


The Data

Collected through hurricane predictions from the The National Oceanic and Atmospheric Administration (NOAA), an American company, and typhoon data from The Japan Meteorological Agency (JMA) in Japan, world temperatures, and a “great deal of other data”.

The Dataset

- Based on weather collected from 18 weather stations across Europe
- Collected information across a period ranging from the late 1800s to 2022.
- Records were made almost daily
- Values recorded included temperature, wind speed, snow, global radiation, and more
- Collected by the European Climate Assessment & Data Project
- Downloadable Dataset Link



Data Bias

1. Collection Bias
    - Over 26,000 weather stations exist across Europe, data was collected from only 18. The sampled weather stations may not accurately represent the diverse climates across Europe.
2. Location Bias
    - Data is centralized to Europe; predictions may not be capable of generalizing and therefore predict accurately in other regions like Russia, Canada, or Chile.
3. Sampling Bias
    - Selection of limited stations can skew the results due to inaccurate representation of regional data
4. Temporal Bias
    - Data range spans a period of over 150+ years, historical conditions may no longer represent current conditions and may mislead the algorithm to determine more milder weather

Data Prediction Accuracy

- Accuracy depends on Machine Learning Algorithm applied to historical data
    - KNN (Best Fit) ~ 88% (Test Data)
    - Decision Tree ~ 47% (Test Data)
        - Requires pruning
    - Artificial Neural Network ~ 50% (Test Data)

Data Optimization

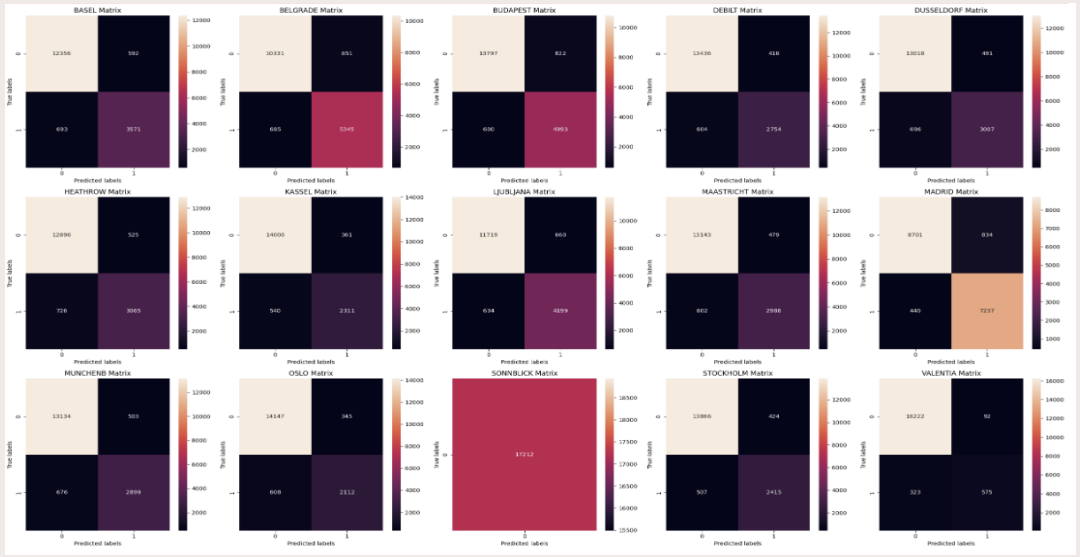
- Data optimized through **Gradient Descent**
- Gradient Descent was utilized to find a local minimum or “valley” within the data, this function represents error within a model.
- For this data, gradient descent was applied to find the minimum error, this was completed by adjusting iteration, Theta values, and step size.
- Results near to 0 were achieved
    - When lower error is achieved, the model fits the data better and can therefore make more accurate predictions

Machine Learning Algorithms

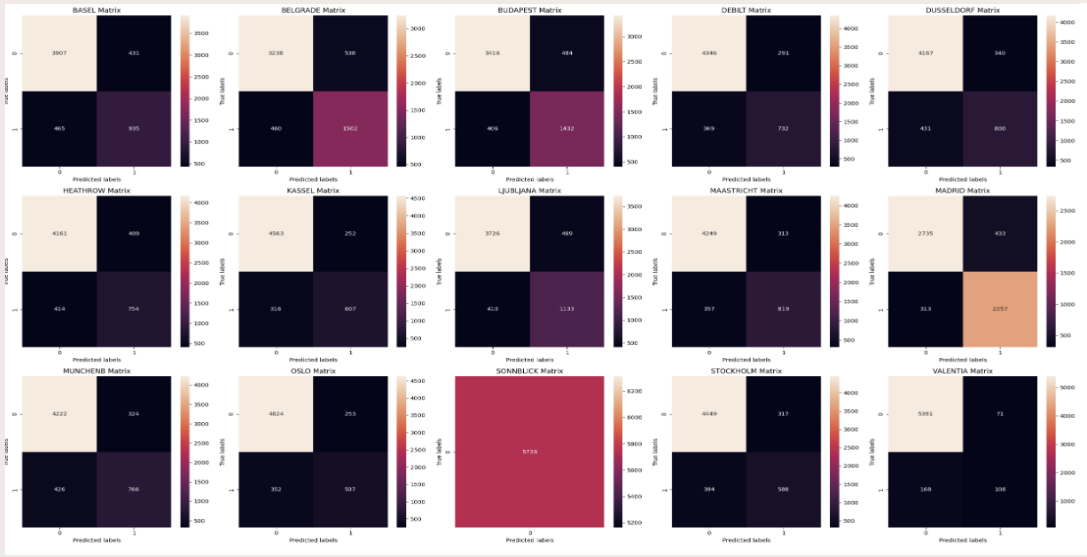
1. K-Nearest Neighbour (KNN)
2. Decision Tree
3. Artificial Neural Network (ANN)

K-Nearest Neighbour (KNN)

Training Dataset (Accuracy)



Testing Dataset (Accuracy)



- Assigns distance to other data points from a selected data point
- Looks at an assigned value of nearest data points (Ex. Looks at 3 closest “neighbours”)
- Makes a prediction based on the “neighbours” or data points closest to it

K-Nearest Neighbour (KNN)

Training Dataset (Accuracy)

Weather Station	Accurate Predictions True Negatives/True Positives		False Positive	False Negative	Accuracy Rate (Overall Percentage of Correct Predictions) (TP+TN)/Total
Basel	12356	3571	592	693	92.5%
Belgrade	10331	5345	851	685	91.1%
Budapest	10797	4993	822	600	91.7%
Debilt	13436	2754	418	604	94.1%
Dusseldorf	13018	3007	491	696	93.1%
Heathrow	12896	3065	525	726	92.7%
Kassel	14000	2311	361	540	94.8%
Ljubljana	11719	4199	660	634	92.5%
Maastricht	13143	2988	479	602	93.7%
Madrid	8701	7237	834	440	92.6%
Munchenb	13134	2899	503	676	93.2%
Oslo	14147	2112	345	608	94.5%
Sonnblick	17212				100%
Stockholm	13866	2415	424	507	94.6%
Valentia	16222	575	92	323	97.6%
				Average	**93.9%**

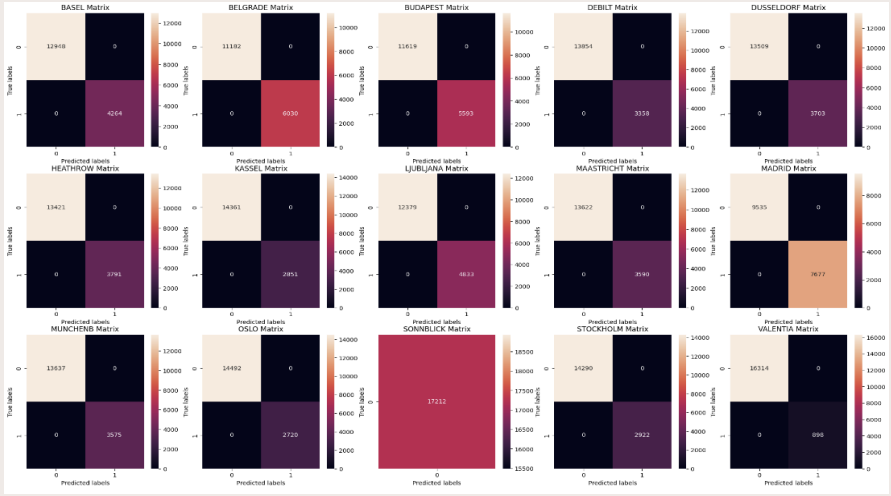
Testing Dataset (Accuracy)

Weather Station	Accurate Predictions True Negatives/True Positives		False Positive	False Negative	Accuracy Rate (Overall Percentage of Correct Predictions) (TP+TN)/Total
Basel	3907	935	431	465	84.3%
Belgrade	3238	1502	538	460	82.6%
Budapest	3416	1432	484	406	84.4%
Debilt	4346	732	291	369	88.4%
Dusseldorf	4167	800	340	431	86.6%
Heathrow	4161	754	409	414	85.7%
Kassel	4563	607	252	316	90.1%
Ljubljana	3726	1133	469	410	84.7%
Maastricht	4249	819	313	357	88.3%
Madrid	2735	2257	433	313	87%
Munchenb	4222	766	324	426	87%
Oslo	4624	507	255	352	89.4%
Sonnblick	5738				100%
Stockholm	4449	588	317	384	87.8%
Valentia	5391	108	71	168	95.8%
				Average	**88.1%**

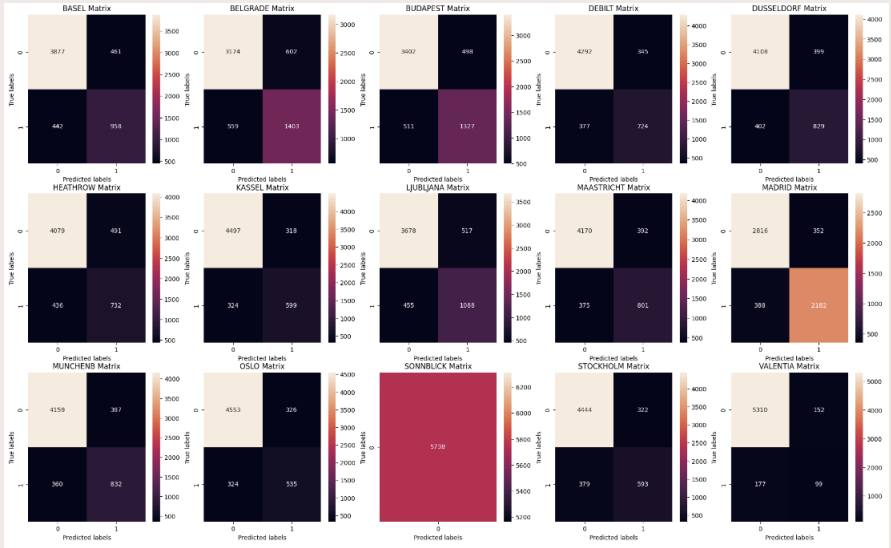
- The algorithm is better at predicting true negatives, or unpleasant weather, suggesting it may work better for this type of prediction
- The model had slightly more difficulty predicting true positives, or pleasant weather
- Overall testing accuracy sits around 88%
- Sonnblick had 100% accuracy, this leads to the suspicion that the model may be overfit and has learned this specific data too well. This means its performance may worsen on new data.

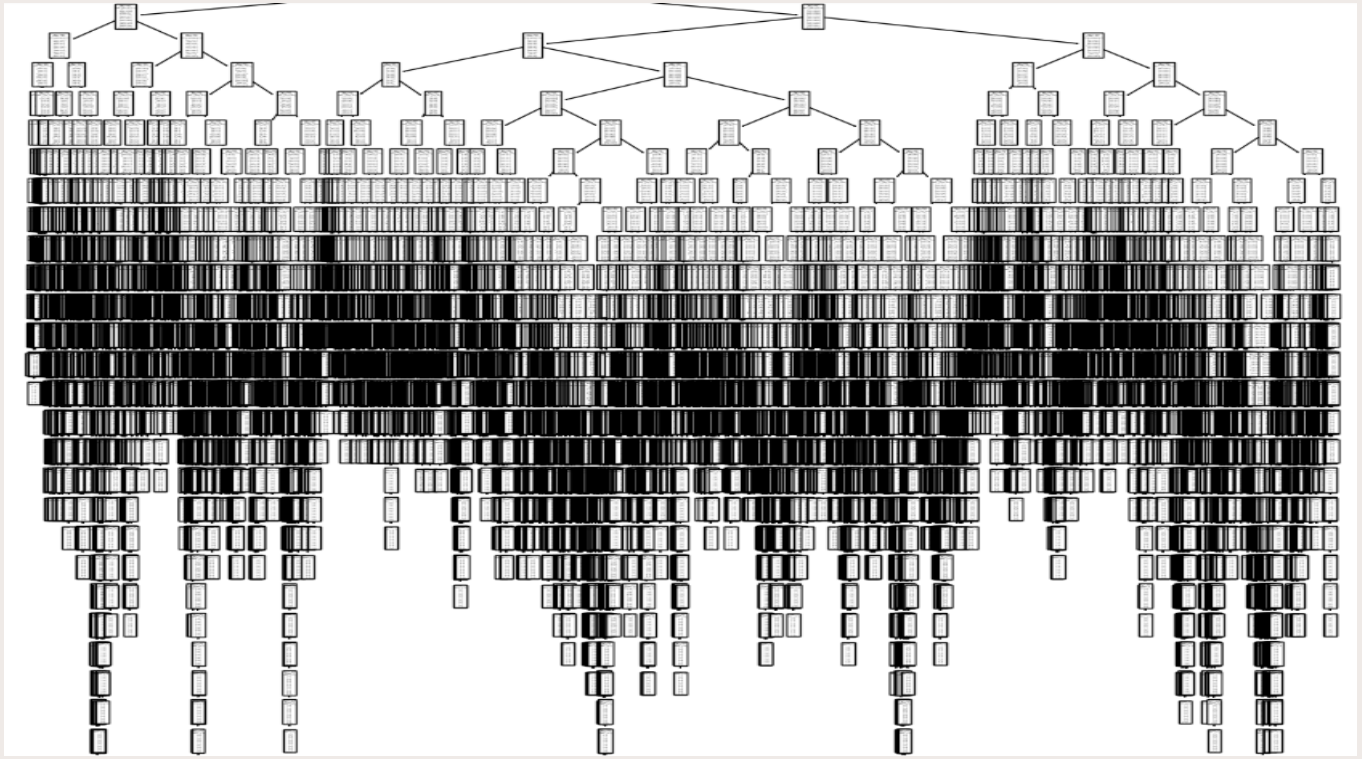
Decision Tree

Training Dataset (Accuracy)



Testing Dataset (Accuracy)

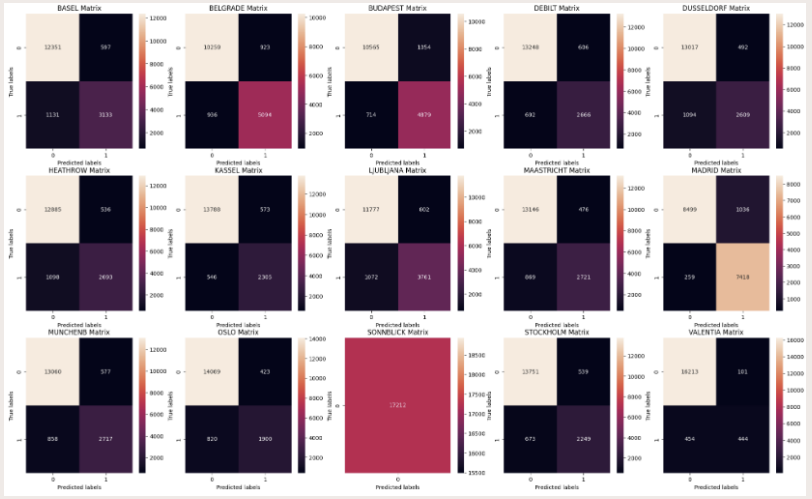




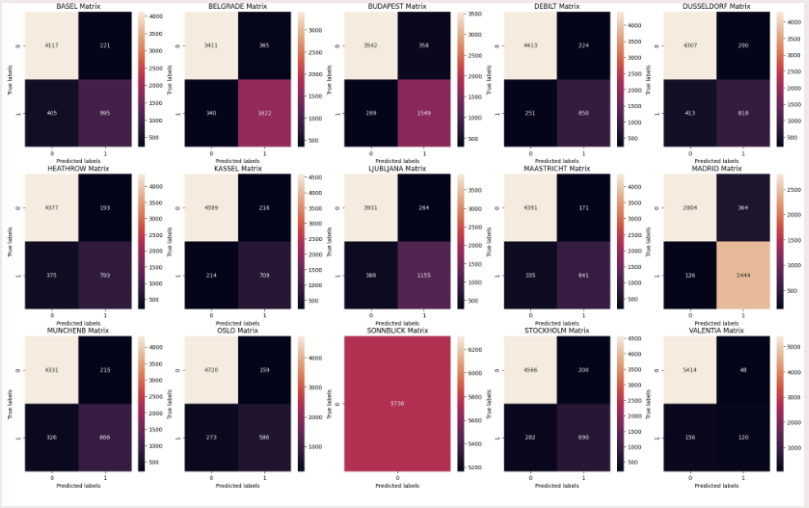
- The accuracy score was ~47% for testing data
- The decision tree is likely overfit due to its excessive depth and complexity
- The decision tree is likely asking too many specific questions causing it to be inefficient
- The decision tree needs to be pruned and then its accuracy reassessed, for now it is not a good choice of a model

Artificial Neural Network (ANN)

Training Dataset (Accuracy)



Testing Dataset (Accuracy)



- Performed multiple scenarios to determine how changing hyperparameters of this machine learning algorithm may affect model accuracy
- The most accurate trial yielded an accuracy for testing data of ~50%
- Similar to other algorithms, Sonnblick achieved an accuracy of 100%

Conclusion and Next Steps

- Findings thus far demonstrate the ability of machine learning to accurately predict certain weather events, this indicates that machine learning could be able to predict adverse and extreme events as well, aligning with our hypothesis.
- The KNN model appears to work the best for this data as it has the highest test accuracy.
- The decision tree model is likely overfit and requires pruning which may lead to improvement

Prune	Prune the Decision Tree Model to improve accuracy and reduce overfitting.
Explore	Trial methods that combine models to improve accuracy. Models can include both supervised and unsupervised as well as previously explored models.
Discover	Utilize unsupervised machine learning models to identify previously unfound patterns and differences in weather data.

Thank you

Any Questions?

Please Contact me below:

Kendra Jackson

Or Visit: