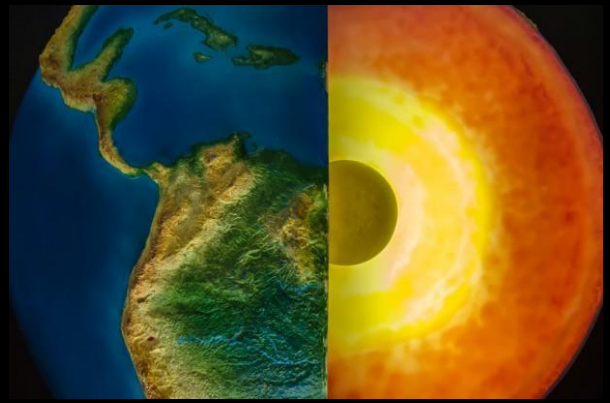
Predicting the Consequences of Climate Change with Machine Learning



Kimberly Mizrahi

December 2024

Introduction

ClimateWins is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world.

The data used in this presentation was collected through hurricane predictions from The National Oceanic and Atmospheric Administration (NOAA) in the U.S., typhoon data from The Japan Meteorological Agency (JMA) in Japan, world temperatures, and a great deal of other data.

Hypotheses

1. Can machine learning predict a significant increase in average daily temperatures in Europe in the next years?
2. Can machine learning effectively predict extreme weather conditions?
3. Can machine learning identify a correlation or pattern between extreme weather conditions with other variables, such as heatwave?

Data Sets

- The Dataset used comes from European Climate Assessment & Data Set project
- The weather dataset includes observations from 18 different weather stations across Europe, containing data ranging from the late 1800s to 2022.
- Recordings exist for almost every day with values such as temperature, wind speed, snow, global radiation, and more.

Data Bias

Collection Bias: The data was collected around 18 different weather stations across Europe, while, according to the European Climate Assessment and Dataset there are a total of 23755 weather stations across Europe.

Location Bias: Considering that the data set was collected by European Assessment & Dataset and contain information about Europe and the Mediterranean, this data may not be able to predict correctly the weather from places such as Brazil or Canada, since it’s different from the sample origin.

Temporal Bias: Since the range is quite large, from late 1800s to 2022, it is possible that some data may not represent the reality anymore and could mislead the machine learning algorithm.

Data Accuracy

- The accuracy of the data depends on the Machine Learning algorithm applied, as shown in the next slides, which highlight significant differences in accuracy across algorithms, as well as their performance on training and test sets.
- The accuracy of the decision tree could not be assessed, as it requires pruning to ensure proper evaluation.
- The Artificial Neural Network (ANN) algorithm achieved an overall accuracy of 40% on the test set, while its accuracy on the training set ranged from 40% to 70%.
- The KNN model demonstrated the best fit for this dataset, achieving an accuracy of 88% on the test set and 92% on the training set.

Data Optimization

- The dataset was optimized using Gradient Descent.
- Gradient Descent is one of the simplest methods to find a local minimum (or valley) and can be applied to both linear and nonlinear cases.
- In this analysis, Gradient Descent was applied to minimize error by adjusting the number of iterations and the step size (alpha), which varied depending on the case.
- By fine-tuning theta0, theta1, the number of iterations, and alpha, it was possible to achieve a result close to 0, as shown in the next slide.

KNN (Accuracy)

Test Set

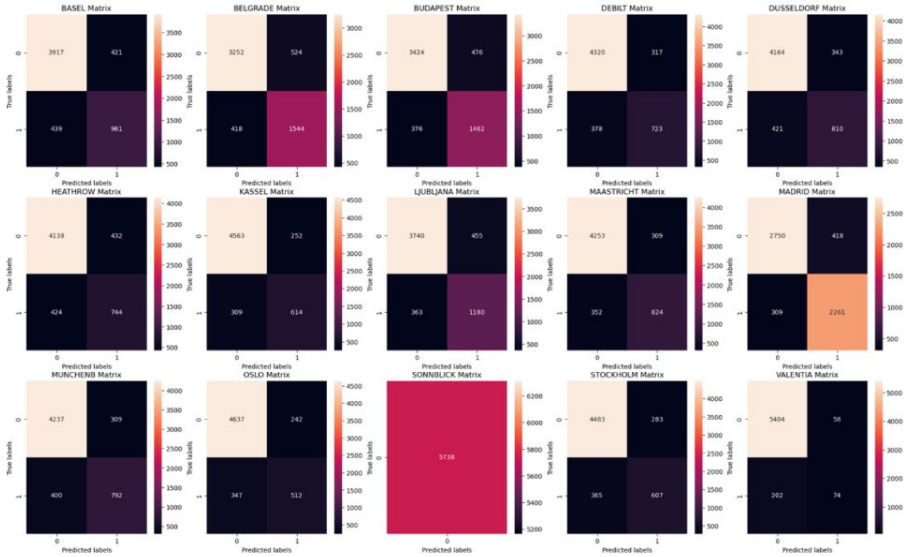
Weather Station	0: True	1: True	0: Mislabel as 1	1: Mislabeled as 0	Total True Positive	Total False Positive	Accuracy
Basel	3917	961	421	439	4878	860	85.01%
Belgrade	3252	1544	524	418	4796	942	83.58%
Budapest	3424	1462	476	376	4886	852	85.15%
Debilt	4320	723	317	378	5043	695	87.89%
Dusseldorf	4164	810	343	421	4974	764	86.69%
Heathrow	4138	744	432	424	4882	856	85.08%
Kassel	4563	614	252	309	5177	561	90.22%
Ljubljana	3740	1180	455	363	4920	818	85.74%
Maastricht	4253	824	309	352	5077	661	88.48%
Madrid	2750	2261	418	309	5011	727	87.33%
Munchenb	4237	792	309	400	5029	709	87.64%
Oslo	4637	512	242	347	5149	589	89.74%
Sonnblick	5738	0	0	0	5738	0	100.00%
Stockholm	4483	607	283	365	5090	648	88.71%
Valentia	5404	74	58	202	5478	260	95.47%
						Average	88.45%

Training Set

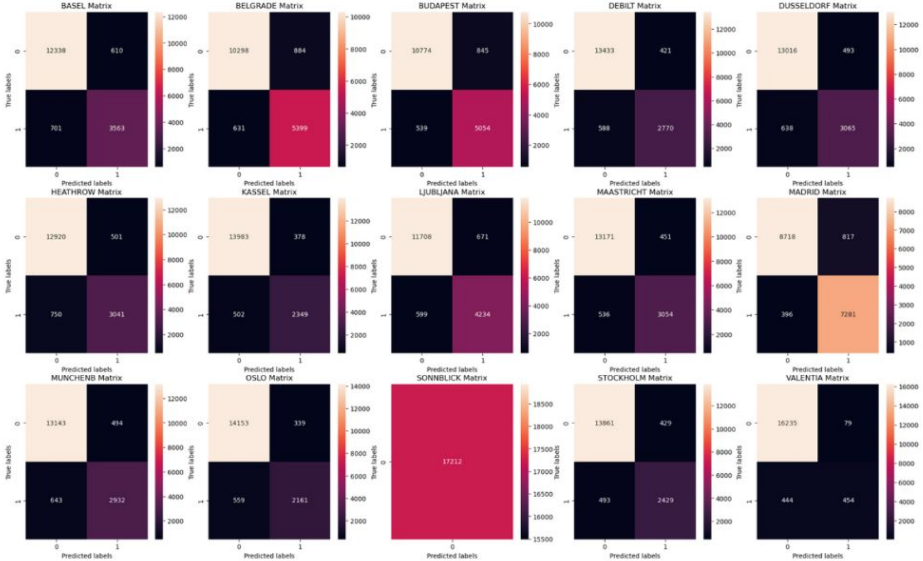
Weather Station	0: True	1: True	0: Mislabel as 1	1: Mislabeled as 0	Total True Positive	Total False Positive	Accuracy
Basel	12338	3563	610	701	15901	1311	92.38%
Belgrade	10298	5399	884	631	15697	1515	91.20%
Budapest	10774	5054	845	539	15828	1384	91.96%
Debilt	13433	2770	421	588	16203	1009	94.14%
Dusseldorf	13016	3065	493	638	16081	1131	93.43%
Heathrow	12920	3041	501	750	15961	1251	92.73%
Kassel	13983	2349	378	502	16332	880	94.89%
Ljubljana	11708	4234	671	599	15942	1270	92.62%
Maastricht	13171	3054	451	536	16225	987	94.27%
Madrid	8718	7281	817	396	15999	1213	92.95%
Munchenb	13143	2932	494	643	16075	1137	93.39%
Oslo	14153	2161	339	559	16314	898	94.78%
Sonnblick	17212	0	0	0	17212	0	100.00%
Stockholm	13861	2429	429	493	16290	922	94.64%
Valentia	16235	454	79	444	16689	523	96.96%
						AVERAGE:	94.02%

KNN (Confusion Matrix)

Test Set

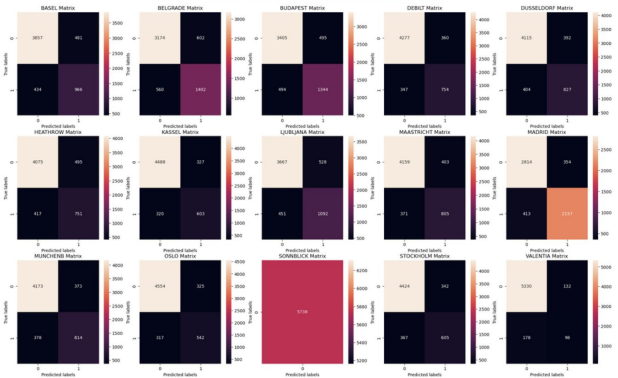


Training Set

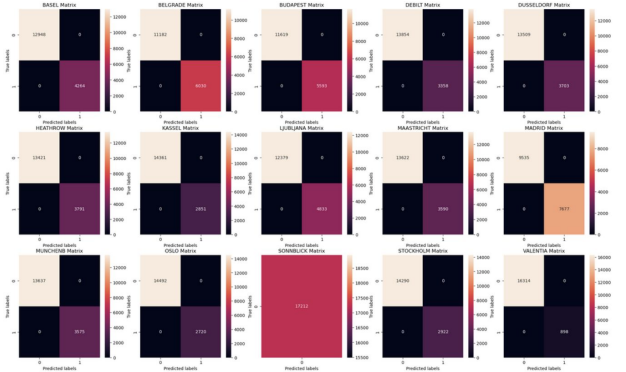


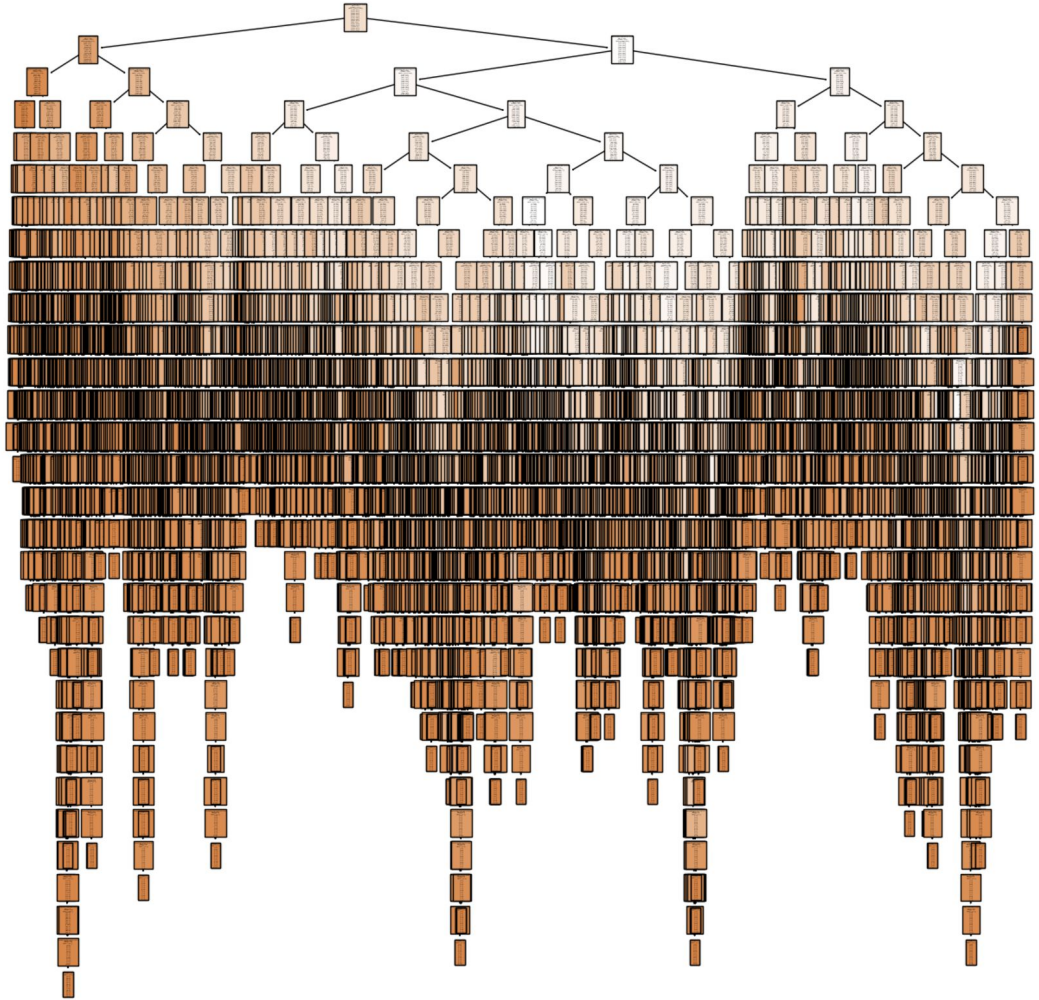
Decision Tree

Test Set



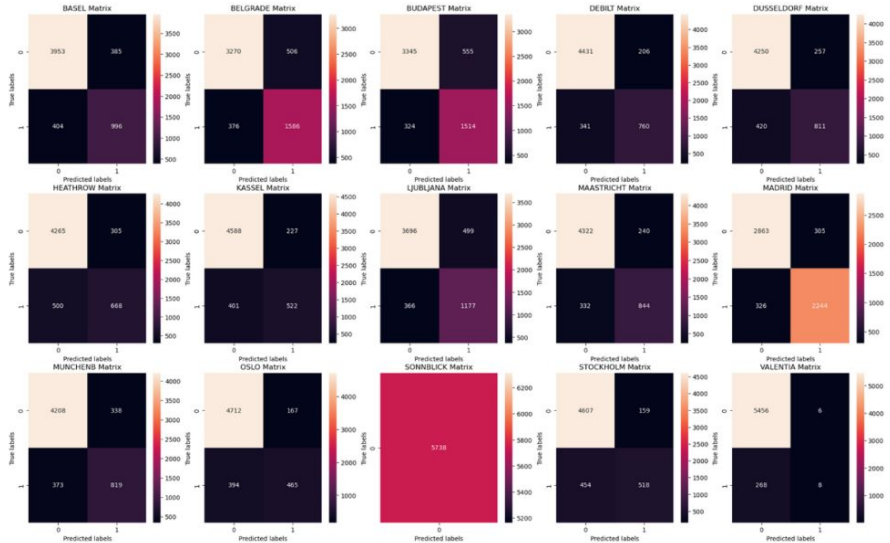
Training Set



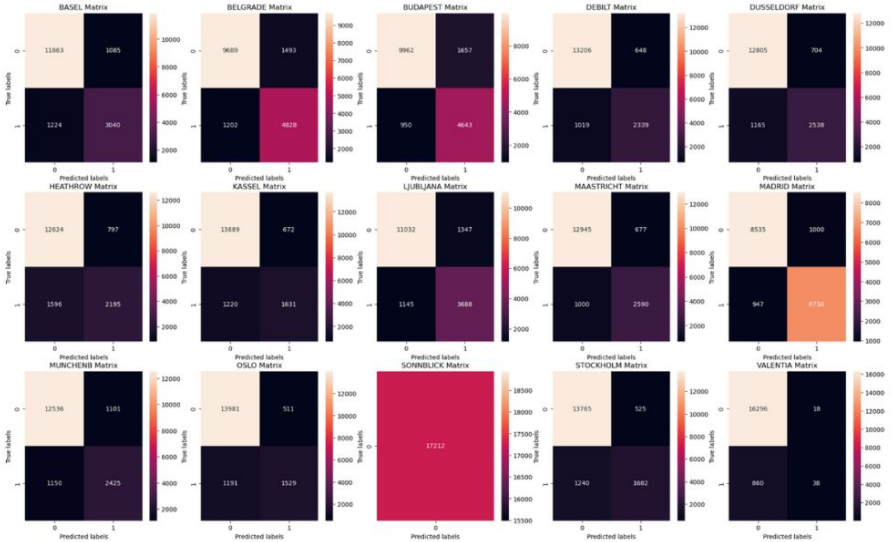


ANN: Scenario 1

Test Set

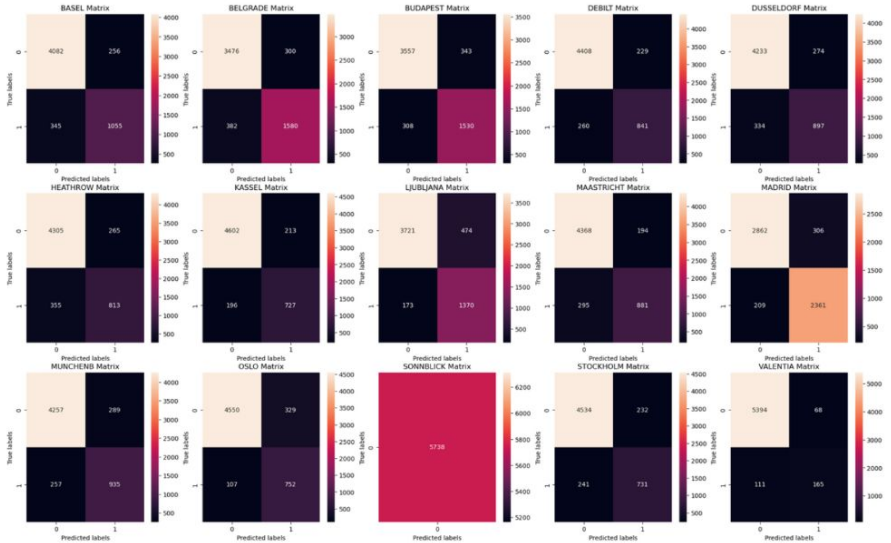


Training Set

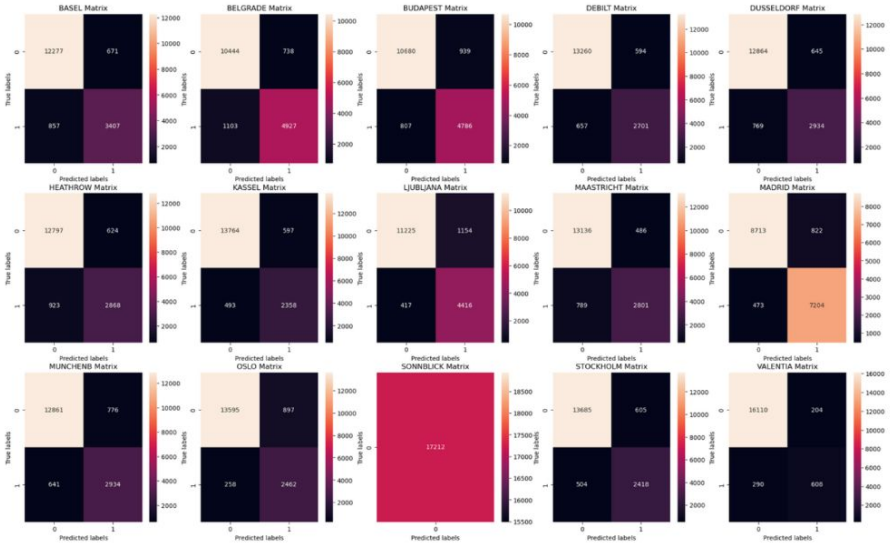


ANN: Scenario 2

Test Set

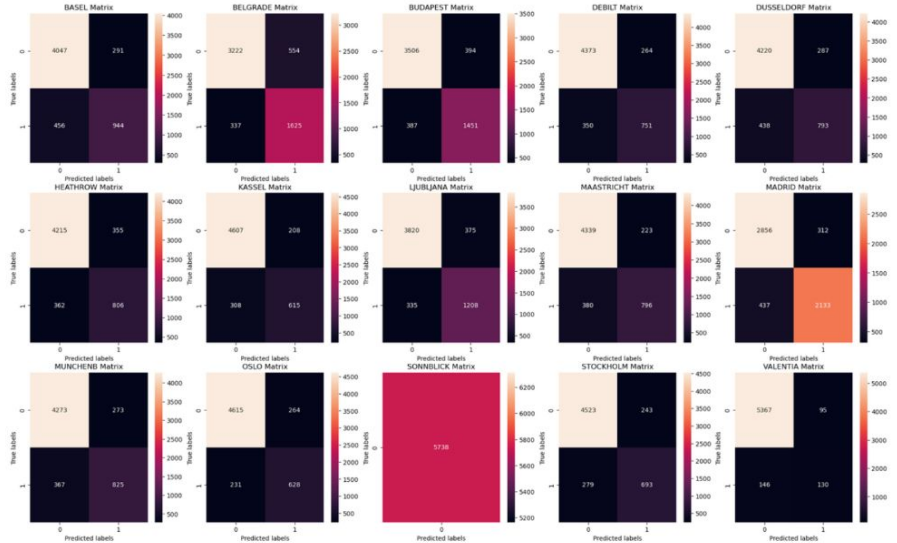


Training Set

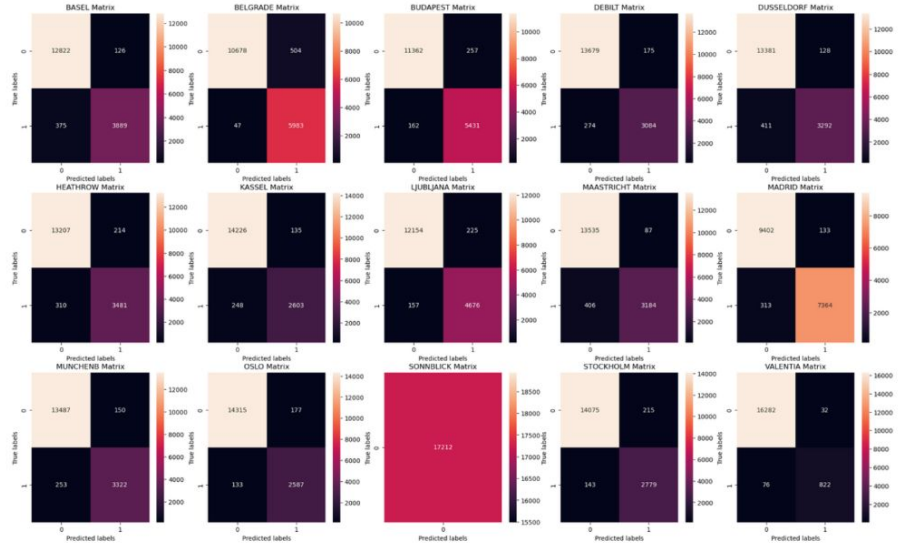


ANN: Scenario 3

Test Set

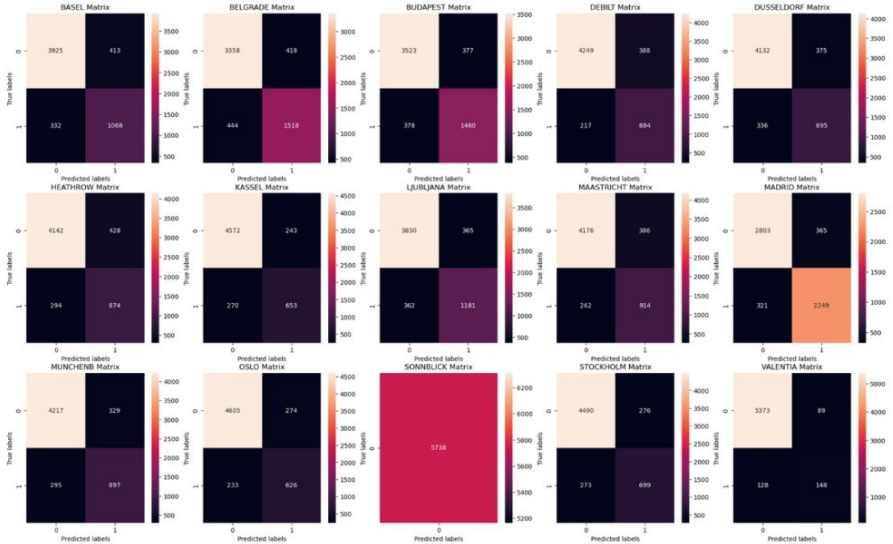


Training Set

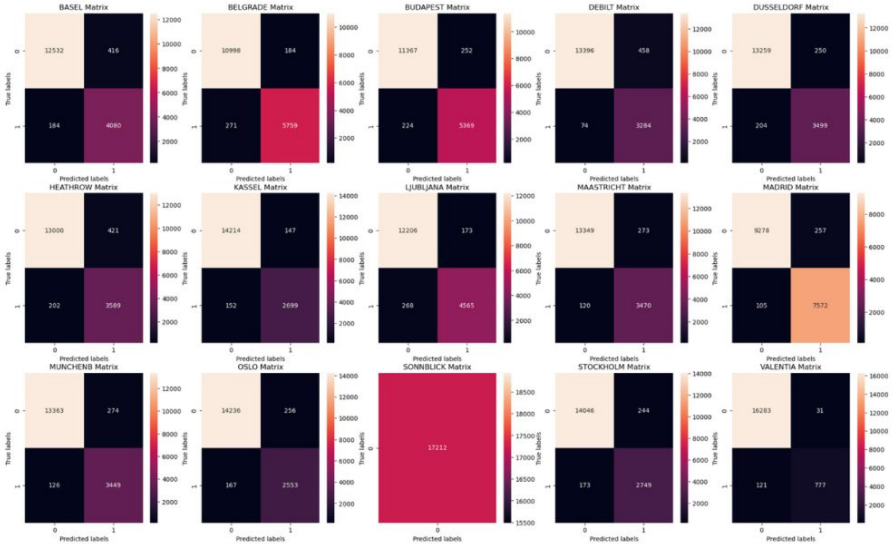


ANN: Scenario 4

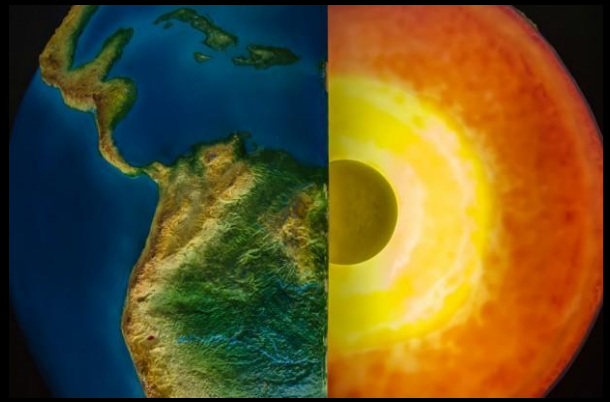
Test Set



Training Set



Project Questions & Answers



Question 1: *How is machine learning used? Is it applicable to weather data?*

Machine Learning algorithms work inductively, identifying patterns to establish rules for solving problems. In this project, the following were used:

- **K-Nearest Neighbor (KNN):** Classifies or predicts group membership based on data point proximity.
- **Decision Tree:** A hierarchical structure of root nodes, branches, and leaf nodes, where each decision leads to a specific outcome.
- **Artificial Neural Network (ANN):** A network of nodes organized in layers, where nodes activate and pass data forward if their output exceeds a threshold.

Question 2: *ClimateWins has heard of ethical concerns surrounding machine learning and AI. Are there any concerns specific to this project?*

Ethical concerns in climate change data include:

- **Privacy issues:** Monitoring private property, land, or activities under the pretext of climate change.
- **Bias**: Political influence may affect data collection and algorithm training based on analysts' personal beliefs.

However, since no personal information is linked to the data in this case, privacy risks are minimized.

Question 3: *Historically, what have the maximums and minimums in temperature been?*

According to the European Climate Assessment & Data Set Project data set, the minimum and maximum temperature registered were:

1. **Sonnblick**: -34.3º C on January 13th,1968.
2. **Belgrade**: 43.6º C on July 24th, 2007.

Question 4: *Can machine learning be used to predict whether weather conditions will be favorable on a certain day?*

The KNN algorithm accurately predicted pleasant weather 88% of the time. With further training and sufficient data, it could predict favorable weather conditions for specific days.

Conclusion

- The current dataset was better predicted by the KNN algorithm, achieving an average accuracy of 88.45% on the test set.
- The decision tree requires pruning to improve its accuracy.
- While the performance of the ANN is unpredictable, it achieved only 40% accuracy on the test set, significantly lower than the KNN algorithm.
- Based on the findings from this project, the KNN algorithm is the best option, achieving 88% accuracy in predicting climate temperature.

Next Steps

- Continue testing using both supervised and unsupervised algorithms after optimizing them.
- Further refine the decision tree by pruning to improve accuracy.
- Explore additional options, such as new algorithms or combinations of previously used algorithms, to uncover previously undefined patterns.

Thank You

Kimberly Mizrahi