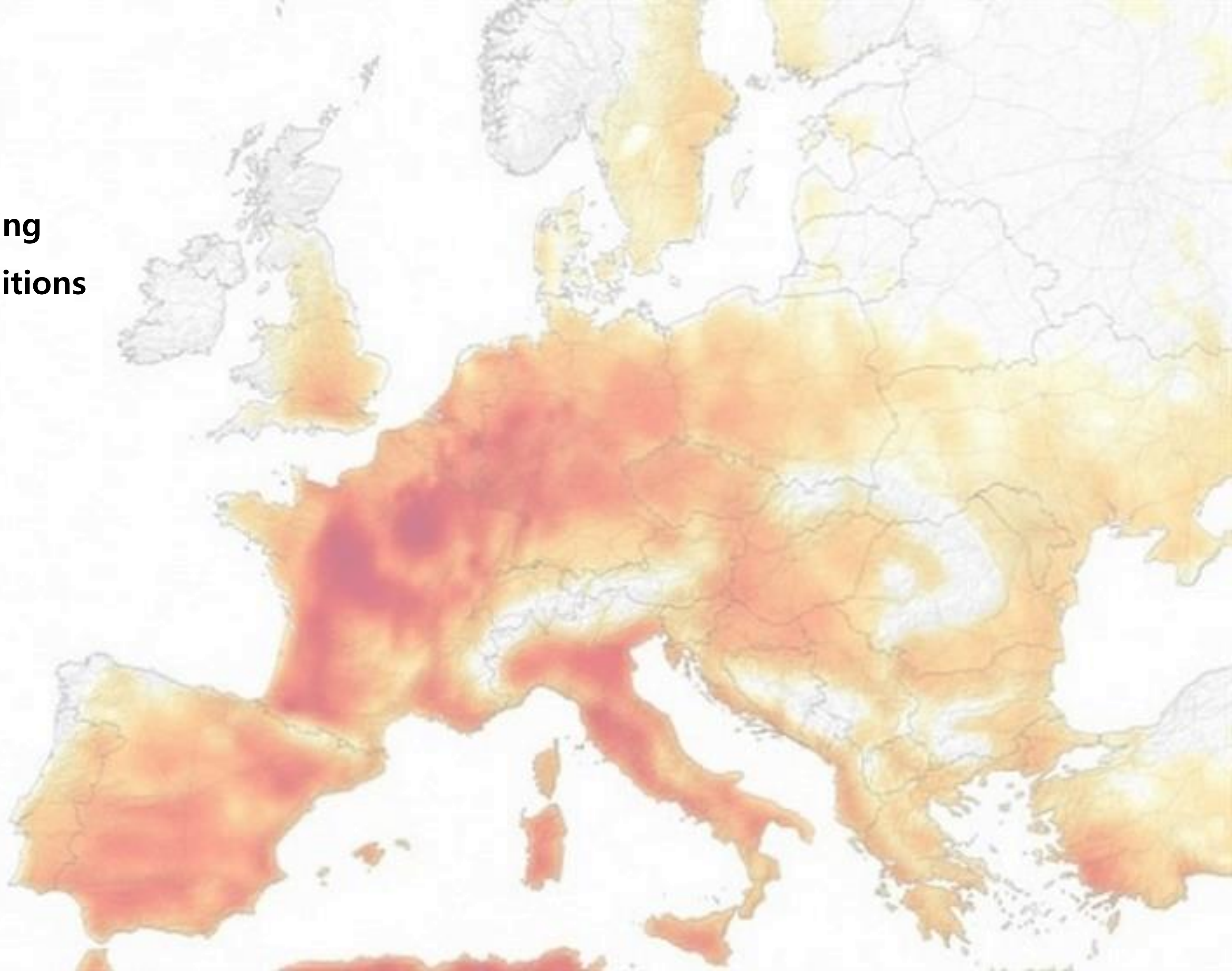
Applying Machine Learning to Predict Weather Conditions



Jinwoo Chung (Data Analyst)
ClimateWins
2024. 4. 28

Table of Contents

1. Background and Objectives

ClimateWins is a non-profit organization dedicated to understanding and mitigating the impacts of climate change, particularly focusing on the weather patterns across Europe. In recent decades, Europe has experienced a marked increase in extreme weather events, including intense storms, prolonged droughts, and devastating floods. These events have significant implications not only on the environment but also on socio-economic stability across the continent.

Recognizing the urgent need for better preparedness and adaptive strategies, ClimateWins has embarked on a project to harness the power of machine learning to predict and analyze these weather changes. The organization aims to utilize historical weather data spanning over a century to develop predictive models that can forecast future weather conditions and identify potential safe havens in the face of escalating climate impacts.

Objectives

- Finding new patterns in weather changes over the last 60 years.
- Identifying weather patterns outside the regional norm in Europe.
- Determining whether unusual weather patterns are increasing.
- Generating possibilities for future weather conditions over the next 25 to 50 years based on current trends.
- Determining the safest places for people to live in Europe within the next 25 to 50 years..

2. Data & Machine Learning Tools

- Scope: Records from 1960 to 2022 with daily weather observations across 18 locations in Europe.
- Variables: Includes metrics like cloud cover, wind speed, humidity, pressure, global radiation, precipitation, snow depth, sunshine, and temperatures & Pleasant Day records
- Data Significance: Enables the identification of long-term climate patterns and extreme weather events critical for predictive modeling.

Machine Learning Tools & Libraries

- Python: The primary programming language for data analysis and machine learning tasks.
- TensorFlow & Keras: Advanced libraries for building neural network models to analyze complex patterns within the data.
- Scikit-Learn: Used for implementing classical machine learning algorithms for comparative analysis and model selection.

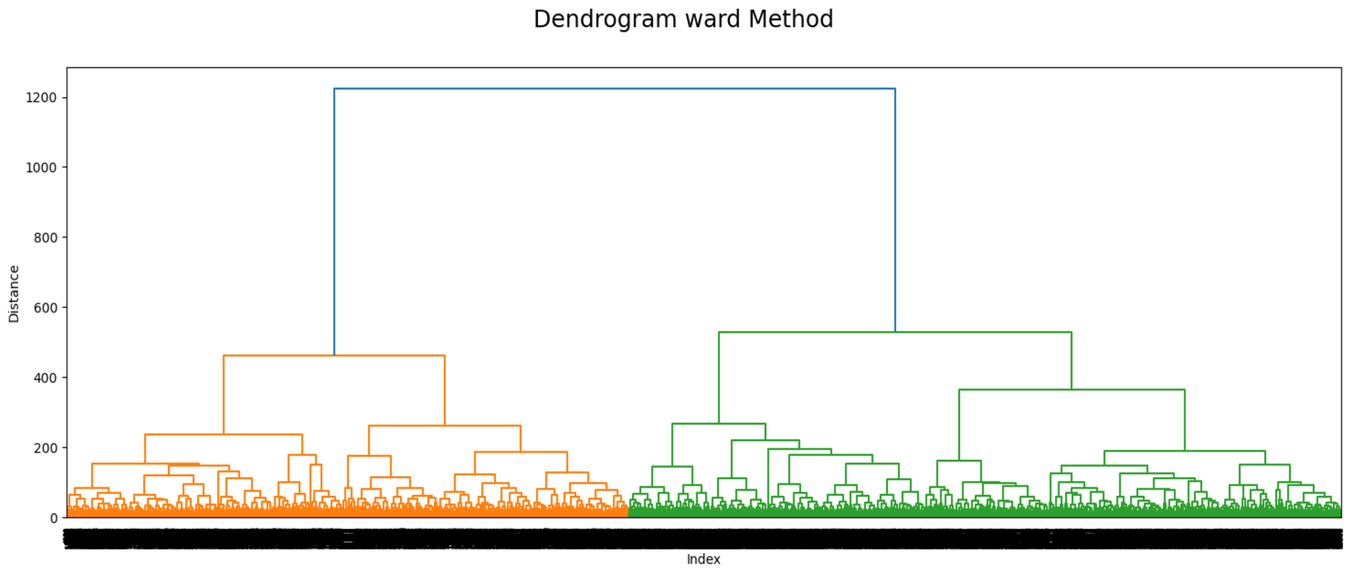
3. Why Pleasant Day?

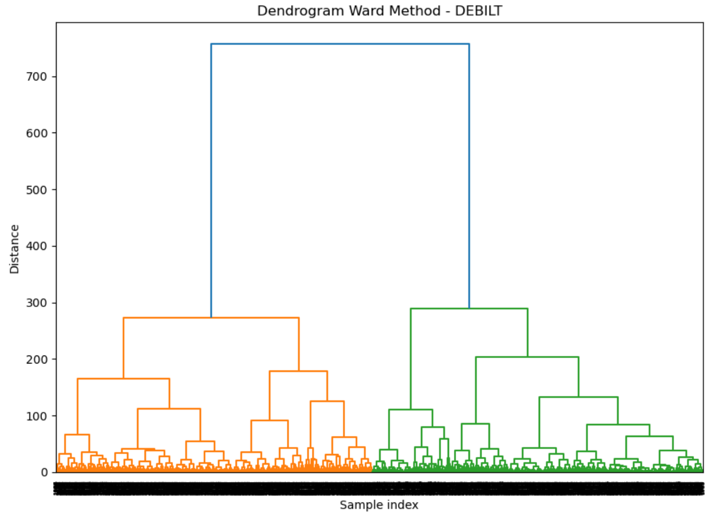
The concept of a "***pleasant day***" is often characterized by weather conditions that are generally agreeable and comfortable for the majority of people. This might include moderate temperatures, low humidity, clear skies, or a light breeze. In the context of future weather prediction, the identification and frequency of pleasant days can be important for several reasons like public health, economic planning, and climate change monitoring, etc.

A pleasant day can be a ***target variable for machine learning*** that recognizes patterns in the weather and predicts the future.

4. Dendrogram Clustering

A dendrogram is a ***tree-shaped diagram*** commonly used in hierarchical clustering. This diagram visually represents the grouping based on the similarity between objects and shows how the objects are connected in stages to form clusters. By illustrating the connections between objects at each level, dendrograms are ***useful in identifying meaningful patterns in complex datasets and understanding the relationships between data points.***





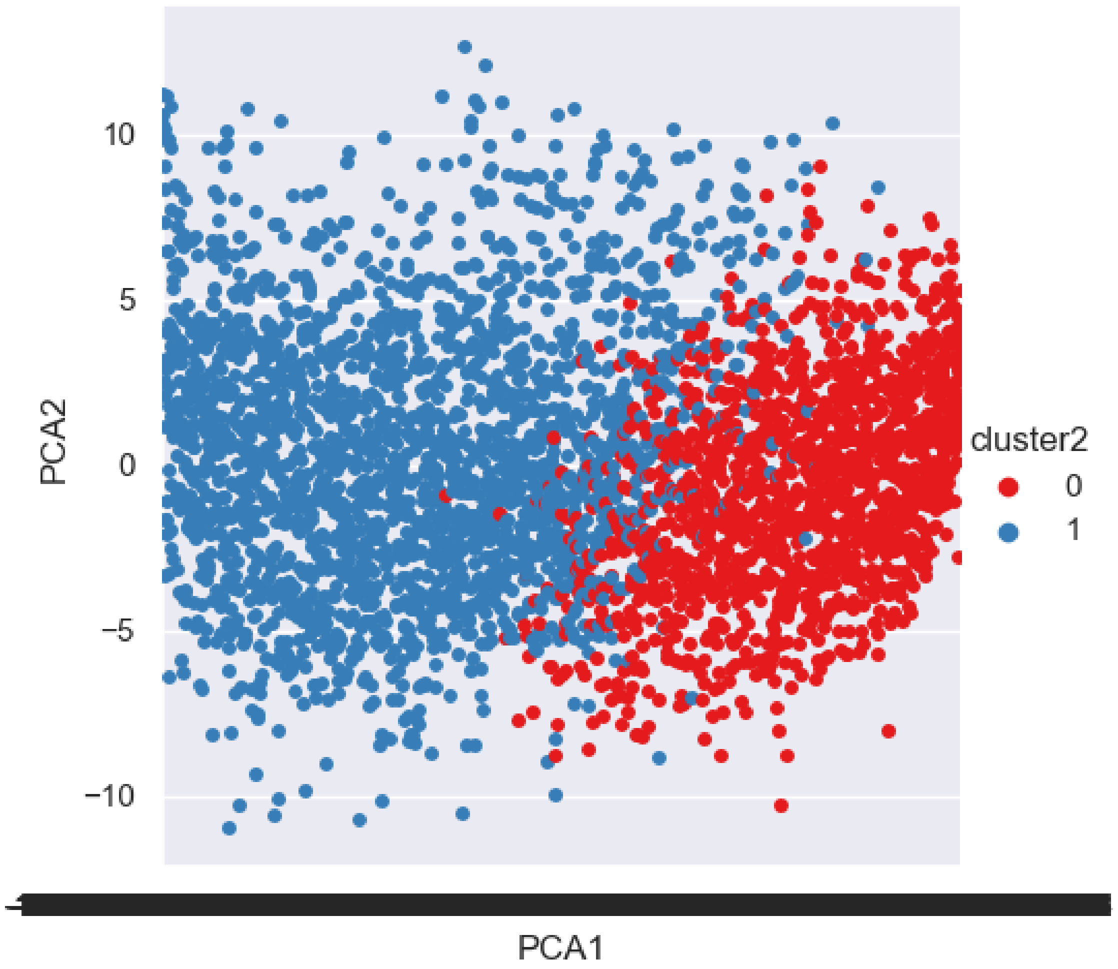
4. Dendrogram Clustering

Our dataset can be separated into two clusters, as shown in the dendrogram.

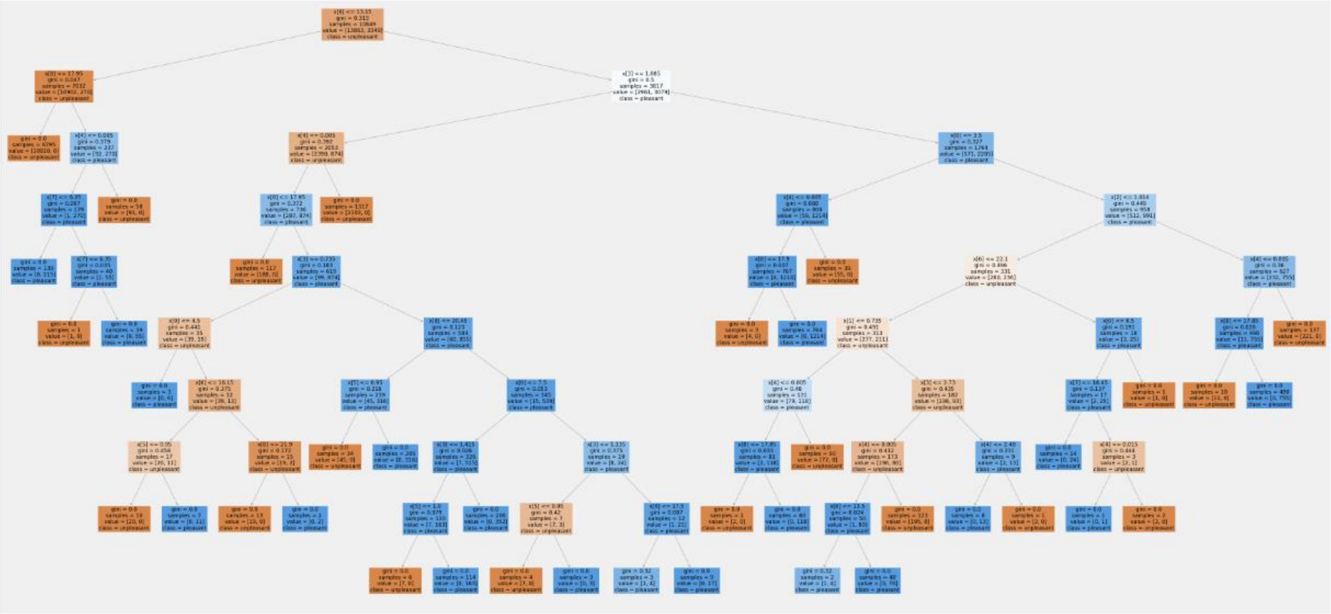
In order to improve the ***resource utilization efficiency*** of the non-profit organization ClimateWins, we attempted dimensionality reduction using PCA.

The graph on the right shows a scatter plot of a dataset of 135+ variables reduced to 2 variables by PCA, and how the 2 clusters in the dendrogram are separated.

The result is that while there is some overlap, the two clusters are mostly distinct, confirming the feasibility of data dimensionality reduction.



5. Important Cities and Factors – RandomForest

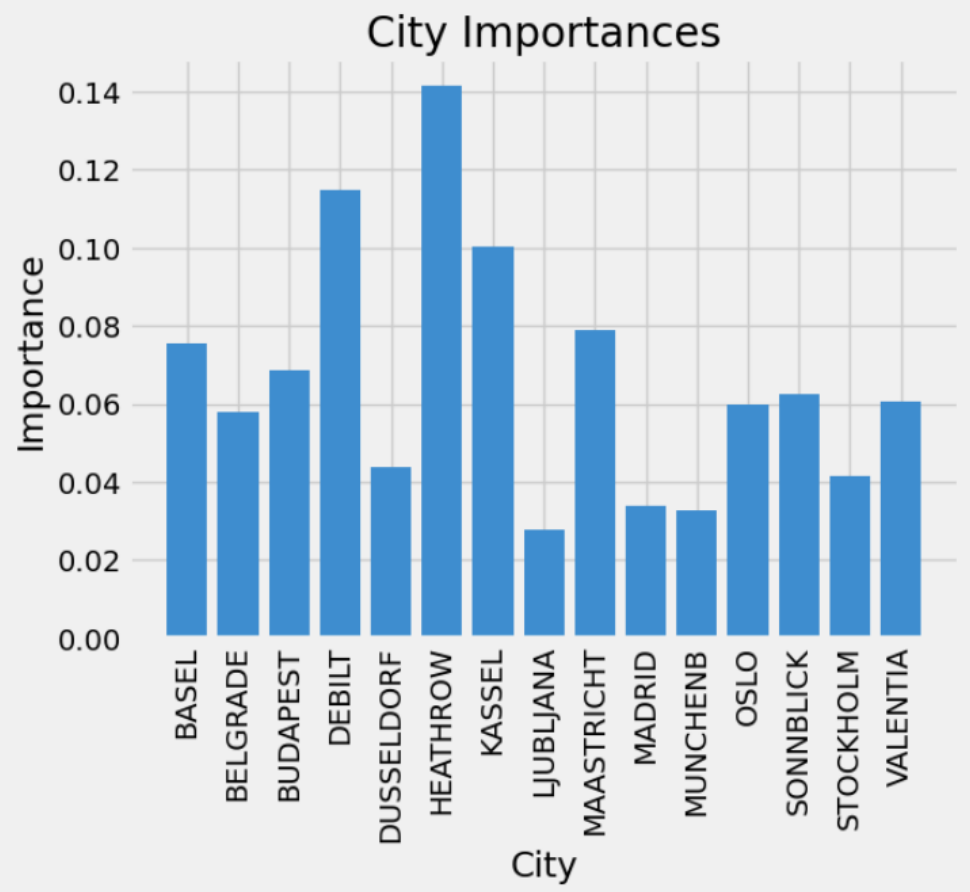


Unsupervised Learning Models - Random Forest, etc.

A random forest is an ensemble learning method that combines multiple decision trees. Each tree is trained on a random subset of the data, and the final prediction is made by aggregating the predictions from all the individual trees. This approach reduces overfitting and improves the generalization ability of the model by capturing a wide variety of data variations.

5. Important Cities and Factors – RandomForest

The Random Forest model is an optimization of the GridSearch and RandomSearch hyperparameters. The model allows us to see that different stations and climate factors have different importance.

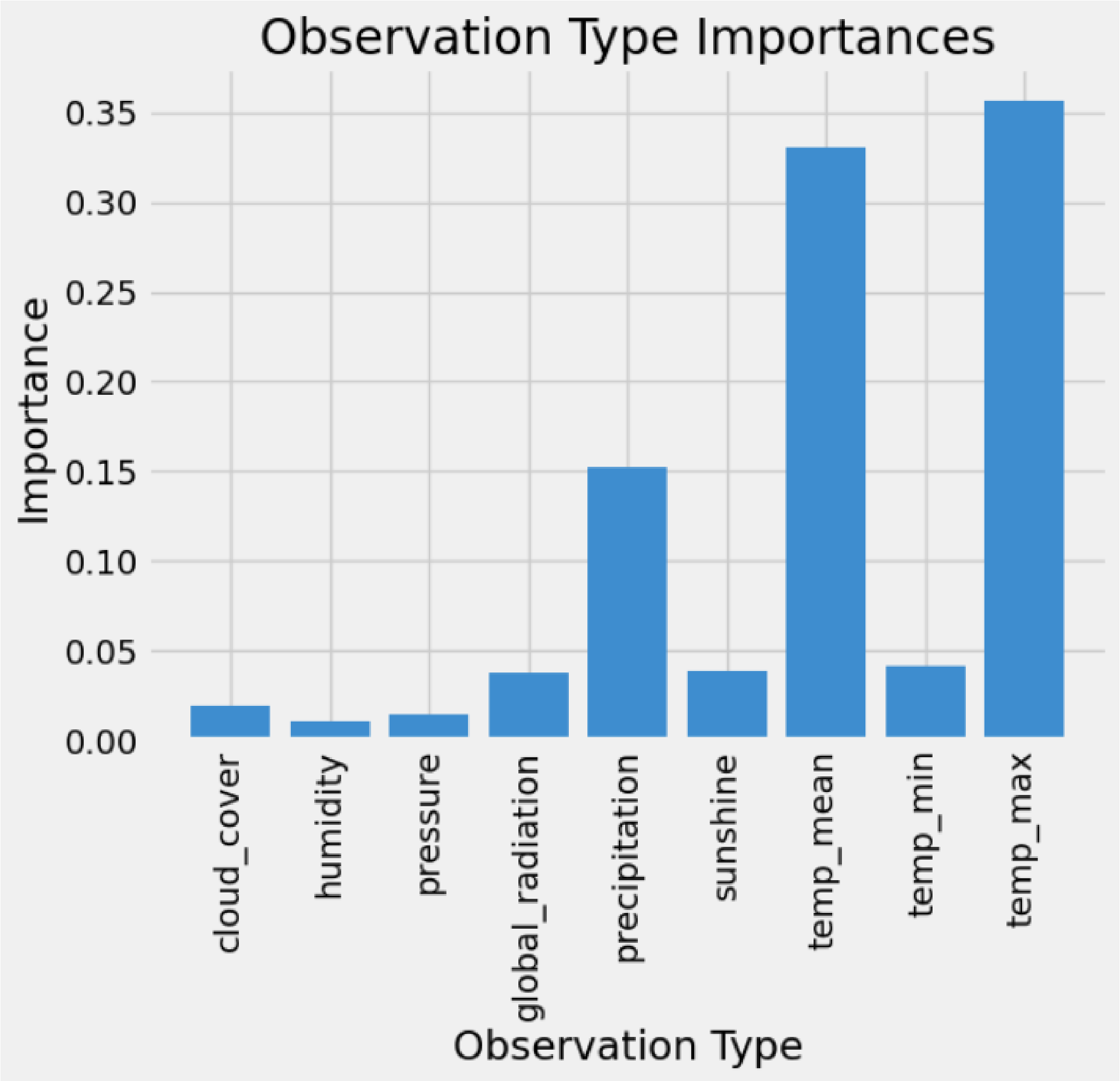


City	Importance
BASEL	0.075445
BELGRADE	0.057926
BUDAPEST	0.068572
DEBILT	0.114975
DUSSELDORF	0.043738
HEATHROW	0.141364
KASSEL	0.100108
LJUBLJANA	0.02789
MAASTRICHT	0.07882
MADRID	0.033717
MUNCHENB	0.032741
OSLO	0.059939
SONNBLICK	0.062624
STOCKHOLM	0.041431
VALENTIA	0.060709

Most Important Cities

1. Heathrow
2. DeBlit
3. Kassel

5. Important Cities and Factors – RandomForest



By climate component, **maximum temperature, minimum temperature, and precipitation** were identified as the most important climate factors.

The findings from Random Forest provide important implications for what ClimateWins should focus on to capture climate change in Europe.

6. CNN Model

```
180/180 [==============================] - 1s 2ms/step
Pred        BASEL  BELGRADE  BUDAPEST  DEBILT  DUSSELDORF  HEATHROW  KASSEL  \
True
BASEL        2628        28       426     117           9        53       8
BELGRADE      550         0       410      50           7        29       1
BUDAPEST       96         0        85      15           1         8       0
DEBILT         47         0        41       6           0         1       0
DUSSELDORF     18         0        10       1           1         2       0
HEATHROW       47         0        34       3           0         7       0
KASSEL          3         0         9       2           0         1       0
LJUBLJANA      31         0        25       3           0         1       0
MAASTRICHT      6         0         0       0           0         0       0
MADRID        218         7        84      26           3        17       3
MUNCHENB        0         0         2       0           0         0       0
OSLO            0         0         1       2           0         0       0
STOCKHOLM       2         0         2       0           0         0       0
VALENTIA        0         0         0       0           1         0       0

Pred        LJUBLJANA  MAASTRICHT  MADRID  MUNCHENB  OSLO  SONNBLICK  \
True
BASEL               8           0     360         1    11          2
BELGRADE            6           0      47         0    11          0
BUDAPEST            2           0       5         0     0          0
DEBILT              0           0       1         0     0          0
DUSSELDORF          0           0       1         0     0          0
HEATHROW            2           0       3         0     0          0
KASSEL              0           0       0         0     0          0
LJUBLJANA           0           0       1         0     0          0
MAASTRICHT          0           0       0         0     0          0
MADRID              5           1      62         0     0          0
MUNCHENB            1           0       0         0     0          0
OSLO                0           0       0         0     0          0
STOCKHOLM           0           0       0         0     0          0
VALENTIA            0           0       1         0     0          0

Pred        STOCKHOLM  VALENTIA
True
BASEL              11         1
BELGRADE            0         2
BUDAPEST            0         2
DEBILT              0         0
DUSSELDORF          0         1
HEATHROW            0         0
KASSEL              0         0
LJUBLJANA           0         1
MAASTRICHT          0         0
MADRID              0         1
MUNCHENB            0         0
OSLO                0         0
STOCKHOLM           0         0
VALENTIA            0         0
```

CNN, or Convolutional Neural Network, is a type of deep learning model that's especially well-suited for analyzing visual imagery. CNNs are composed of layers that process and transform an input image to produce a prediction output.

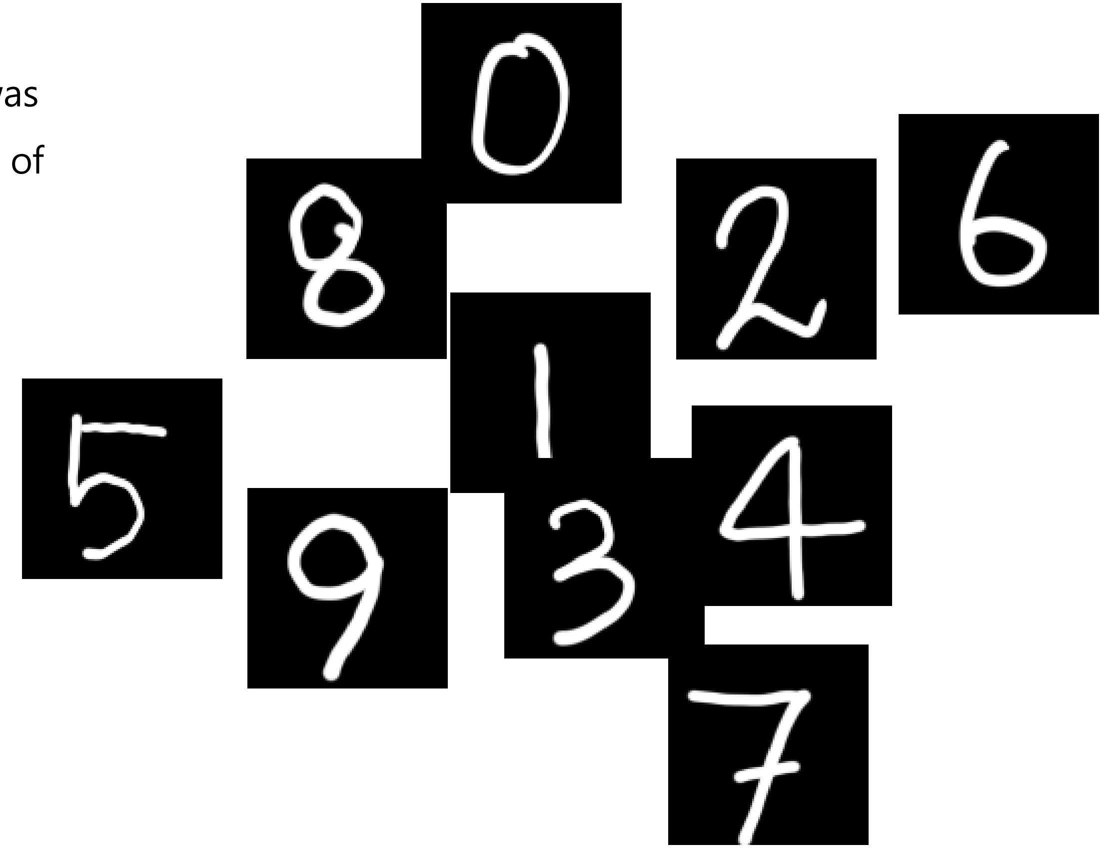
- Convolutional Layers: These layers apply filters to the input image to create feature maps that highlight specific features like edges, textures, or shapes.
- Activation Functions: After convolution, an activation function like ReLU (Rectified Linear Unit) is applied to introduce non-linearity, helping the network learn complex patterns.
- Pooling Layers: These layers reduce the spatial dimensions (width and height) of the input volume for the next convolutional layer. It’s also a form of down-sampling that retains the most important information.
- Fully Connected Layers: Towards the end, the network includes one or more fully connected layers, which take the high-level features from the output of the convolutional layers and use them to perform classification.
- Output Layer: The last layer gives the final prediction. In a classification task, it often uses the softmax activation function to output probabilities of different classes.

6. CNN Model

CNNs are particularly **powerful in tasks like image recognition**, object detection, and even areas outside of image processing like time series analysis or natural language processing, where the hierarchical pattern recognition capabilities of CNNs can be leveraged.

The CNN model trained through Bayesian optimization was able to distinguish between images. Below is an example of how the CNN model recognizes handwriting, showing a recognition rate of around 99%.

```
313/313 [==============================] - 3s 9ms/step
[[ 978    0    1    0    0    0    0    1    0    0]
 [    0 1128    1    1    0    1    3    1    0    0]
 [    1    0 1026    1    0    0    0    3    1    0]
 [    0    0    2 1004    0    3    0    0    1    0]
 [    1    0    1    0  961    0    4    0    3   12]
 [    0    0    0    3    0  887    1    0    1    0]
 [    5    1    0    1    1    4  946    0    0    0]
 [    0    0    7    1    0    0    0 1016    1    3]
 [    1    0    1    1    0    0    0    1  968    2]
 [    3    0    0    2    2    6    0    2    2  992]]
```



7. GAN Model

Another deep learning model, the Generative Adversarial Network (GAN), trains itself on images to create new, similar images. .

A GAN is a sophisticated AI model composed of two neural networks: ***a generator that creates data, and a discriminator that evaluates it***. These networks compete, improving their abilities over time. GANs are particularly adept at image-related tasks. The discriminator enhances its image recognition capabilities by distinguishing real from generated images, while the generator improves at creating visually convincing images, useful for photo-realistic image generation and style transfer.




7. GAN Model

What can GAN do for future weather prediction?

- Identify the path of a typhoon

The extent of damage caused by meteorological natural disasters such as typhoons varies greatly depending on the size of the typhoon and whether or not the country is prepared for it. The likely path and intensity of a typhoon can be predicted if a GAN is trained on continuous images of typhoons.

- Localized precipitation forecasting

Currently, weather forecasting, especially precipitation probability and precipitation amount, is done on a city-by-city basis. However, as observational systems improve in accuracy, and models such as GANs improve in training and accuracy, it is expected that weather forecasting will become more efficient (and less expensive) to provide more granular localized weather information.

- Predicting extreme weather

Extreme weather is no longer a surprise: every year, extreme heat waves, cold waves, and heavy snowfall cause loss of life and property in many parts of the world. If deep learning models such as GANs are further developed, it is expected that we can identify the causes of these extreme weather events or detect the signs of their occurrence in advance. If the trained weather data detects weather events that are different from the predicted weather pictures, it will be possible to identify them as signs of extreme weather and prepare for them in advance.

8. Conclusion

In this project, we have embarked on a comprehensive journey with the goal of harnessing machine learning to understand and predict weather changes across Europe, a continent that has seen a significant increase in extreme weather events in recent decades. Our approach integrated various machine learning techniques, including unsupervised learning models like Random Forests, and advanced neural networks such as CNNs and GANs, to analyze and predict weather patterns from historical data spanning over sixty years.

Our findings have not only enhanced our understanding of the climate dynamics in Europe but also equipped ClimateWins with predictive models that can forecast future weather conditions with considerable accuracy. This capability is critical for preparing for and mitigating the impacts of climate change and extreme weather events. Through the use of machine learning, we have been able to identify key weather patterns and determine the safest areas for future habitation.

Next Steps & Future Analysis

1. **Model Refinement and Validation**: Continue to refine and validate the predictive models through extended testing with additional data sets
2. **Integration of Real-time Data**: Develop capabilities to integrate real-time weather data into the models, allowing for dynamic predictions and adjustments as new data becomes available.

Thank You For Your Attention.

Further analysis will be updated and reported soon.

Feel free to reach out with any further questions or inquiries.

jinwoochung@outlook.com

http://www.jinwoochung.com

Linkedin.com/in/ChungJinwoo

Jinwoo Chung (Data Analyst)
ClimateWins
2024. 4. 28