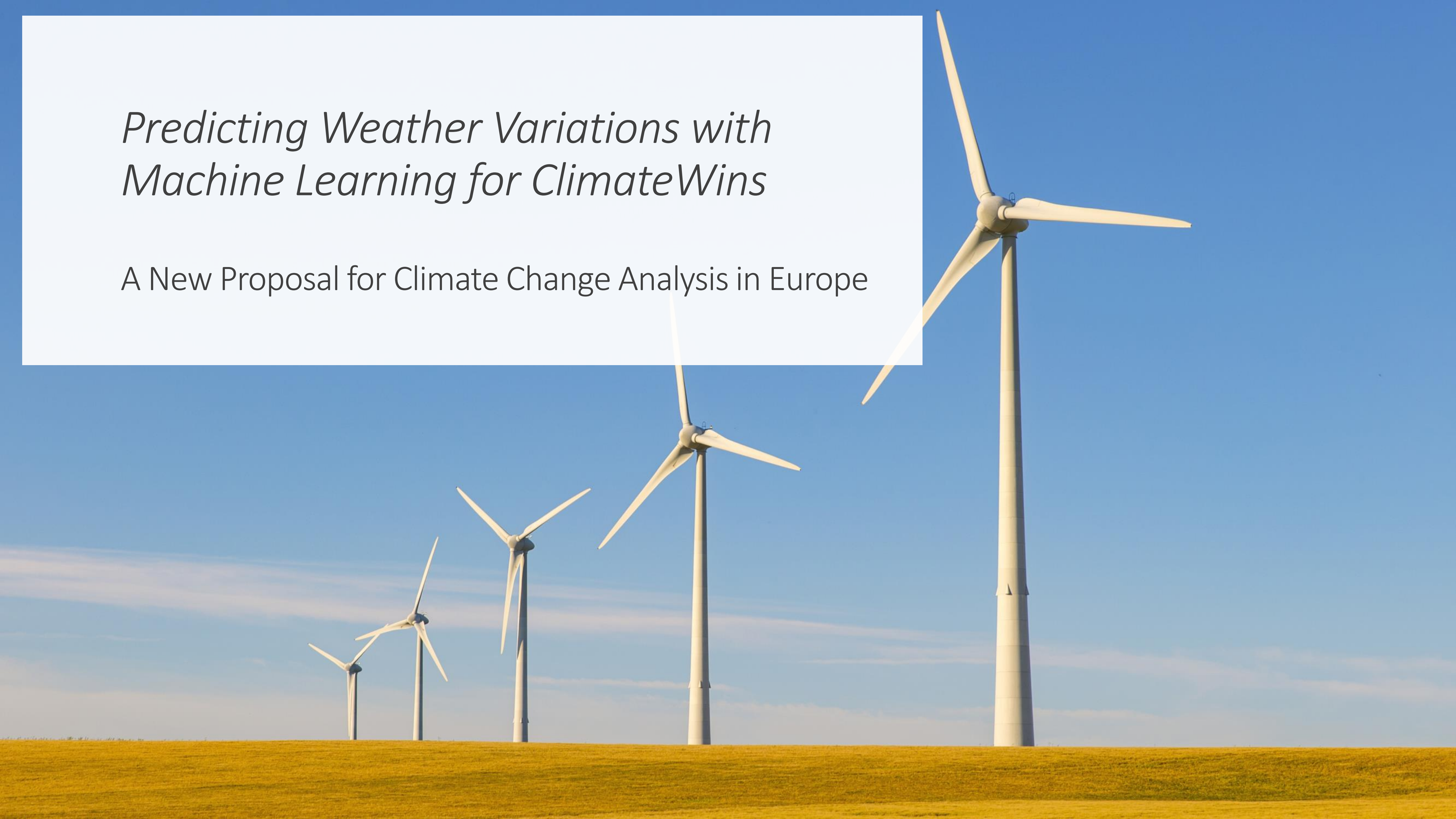
Predicting Weather Variations with Machine Learning for ClimateWins

A New Proposal for Climate Change Analysis in Europe



Introduction & Objectives

Who is ClimateWins?

Nonprofit focusing on extreme weather & climate change with limited resources, looking for data-driven machine learning solutions.

Project Objectives

- Identify weather patterns outside the regional norm in Europe.
- Determine if unusual weather patterns are increasing.
- Generate future weather condition possibilities (25–50 years).
- Determine the safest European regions for habitation (25–50 years).

Overview of Proposed Thought Experiments

Hierarchical Clustering & PCA

- **Goal**: Classify and understand weather extremes vs. “normal days.”
- **Method**: Using dendrograms (complete, single, average, Ward’s), possibly combining with PCA for dimensionality reduction.

CNN & GAN Approach

- **Goal**: Generate and predict realistic future weather patterns
- **Method**: Use CNN for pattern detection in complex data; use GANs to create “synthetic” weather scenarios.

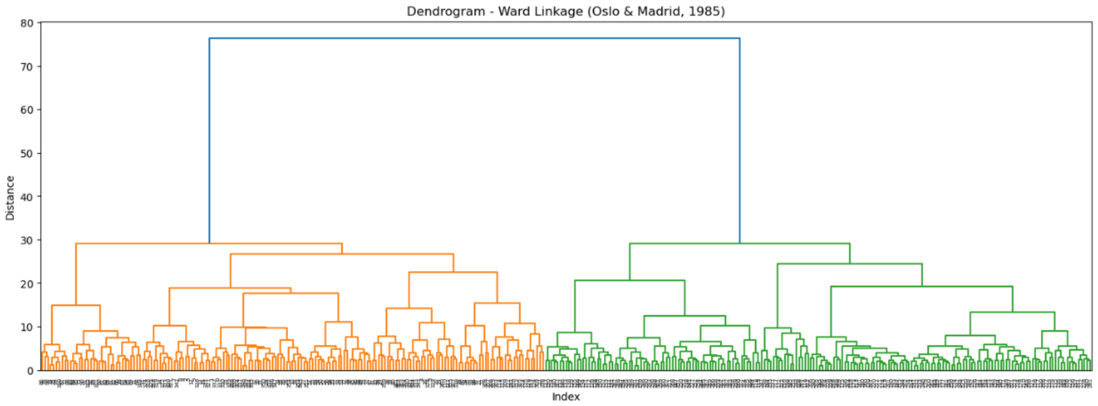
Random Forest & Optimization

- **Goal**: Identify critical features (temp, precipitation, etc.) and find the safest places to live.
- **Method**: Train a RandomForestClassifier, optimize hyperparameters, interpret feature importances.

Machine Learning Options & Their Value

Hierarchical Clustering + PCA:

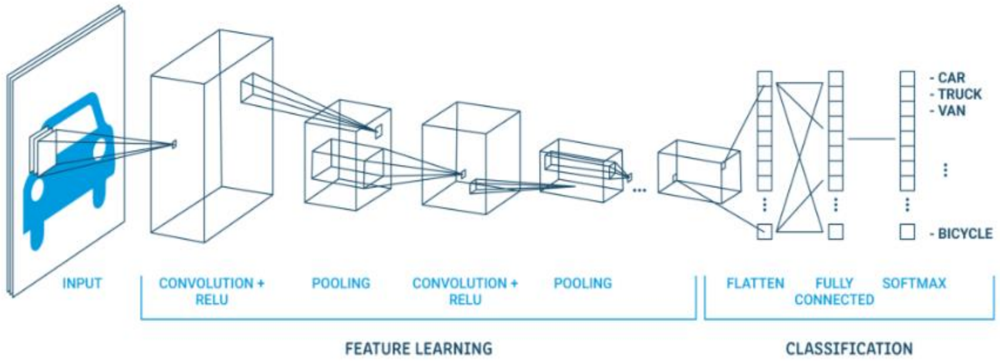
Unsupervised, can uncover hidden structures. PCA reduces dimensionality to highlight essential patterns (e.G., “Pleasant” vs. “Unpleasant” days).



Example: Dendrogram analyses of Oslo & Madrid (1985) revealed better cluster clarity after applying PCA.

CNN (Convolutional Neural Network):

Strong for spatial or image-like data (e.G., Radar/satellite). Can detect local patterns quickly.



Example Observations: Achieved up to **~19%** with unoptimized attempts on multiple stations; saw improvements up to **81%** after hyperparameter tuning in other contexts.

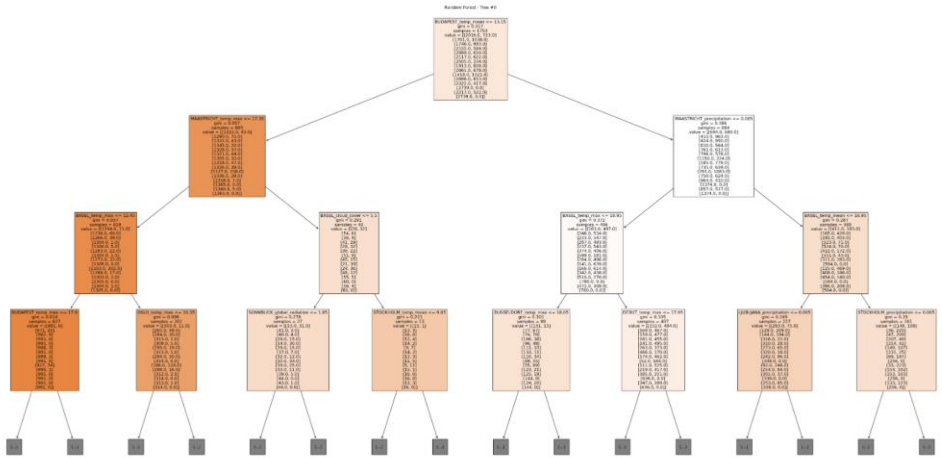
Machine Learning Options & Their Value

GAN (Generative Adversarial Network)

Generates synthetic yet realistic data, allowing us to simulate future climate scenarios beyond the historical record.

Random Forest

High interpretability for feature importances; can handle multi-label classification with ease.



Example Observations: Achieved 59.1% accuracy for a decade of multi-station data, 100% in single-station cases (like Maastricht).

Data Needed Beyond Historical Weather

Radar and Satellite Imagery:

For CNN to detect weather fronts, cloud structures, storms, etc.

Extreme Weather Events:

Hurricanes, storms, droughts, and heatwaves data to see if extremes are rising.

Geospatial & Socioeconomic Data:

Population density, infrastructure resilience—helps determine “safest places.”

Health Impact Data:

Illness/injury and mortality rates tied to weather extremes (validating what “dangerous” means).

Thought Experiment 1 — Hierarchical Clustering & PCA

Objective

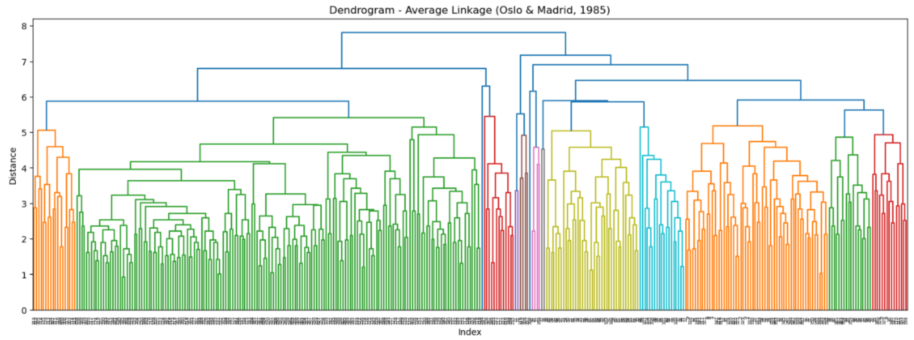
- Identify unusual vs. typical weather patterns by grouping daily observations.

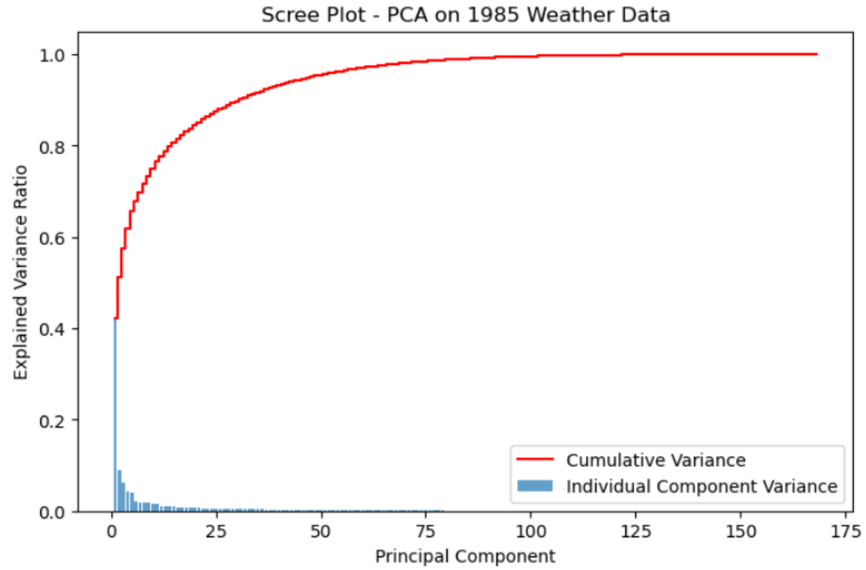
Method

- Reduce dimensionality with PCA (e.g., from 50+ features down to ~20 principal components).
- Apply different linkage methods (complete, ward, etc.) to see how clusters form.

Key Findings

- After PCA, dendrograms are “sharper”; single linkage still “chains,” but complete linkage finds more nuanced subgroups.
- Allows us to see patterns like “pleasant days” vs. “extreme days.”





Thought Experiment 2 — CNN & GAN

Objective

- Simulate and predict future weather conditions using advanced neural networks.

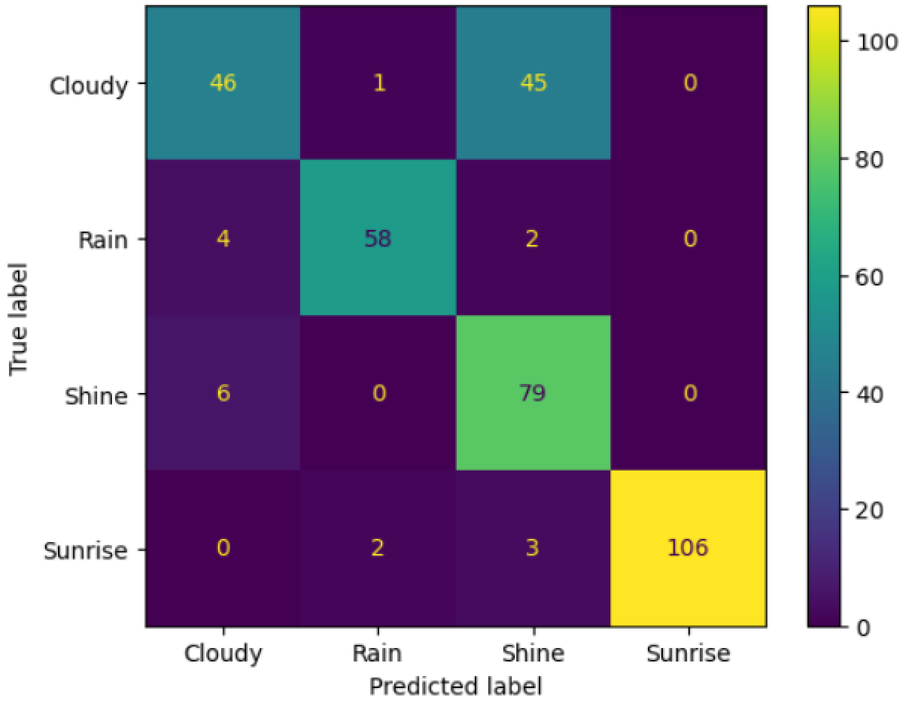
CNN

- Trial runs had issues with unscaled data & class imbalance, leading to ~7–19% accuracy.
- With hyperparameter tuning (filters, kernel size, dropout), accuracy can jump to 70–80%.

GAN

- Could create synthetic “future” weather data, adding variety and volume to the training set.
- Realistic scenarios for years 2050, 2070, etc.

The model performed well in classifying "Sunrise" and "Shine" images but struggled to distinguish between "Cloudy" and "Shine" or "Cloudy" and "Rain." This highlights areas for improvement, such as refining the dataset or further tuning hyperparameters



The plots of training and validation accuracy and loss over epochs showed steady improvements in accuracy but some fluctuation in validation loss, reflecting the challenges of generalizing to unseen data.

Thought Experiment 3 — Random Forest & Optimization

Objective

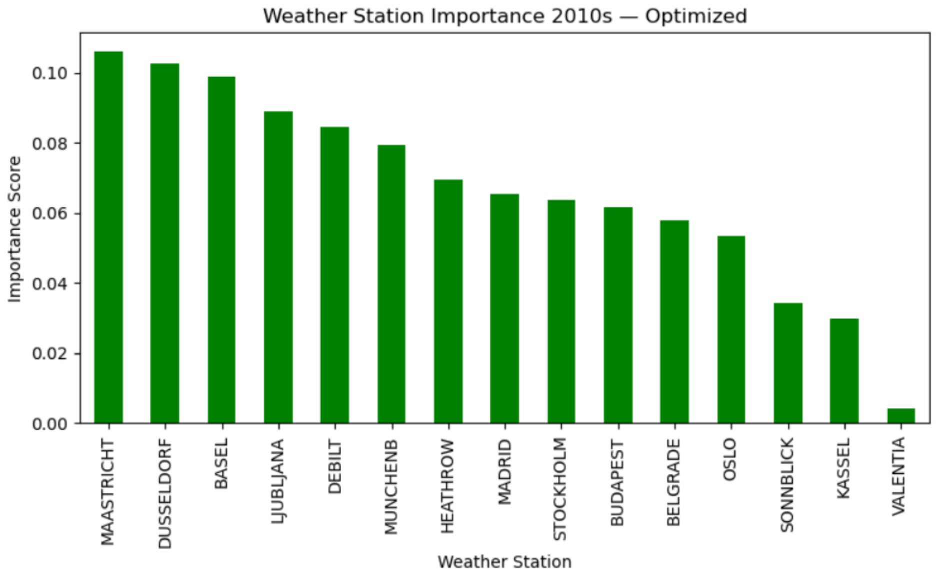
- Determine safe living regions & key climate features.

Method

- Train a RandomForestClassifier on multi-label station data.
- Hyperparameter optimization (e.g., n_estimators, max_depth).
- Examine feature importances (temp_max, precipitation, etc.).

Key Findings

- ~59.1% accuracy across 15 stations (2010–2019).
- Single stations (like Maastricht) can yield 100% accuracy but might be artificially "easy."
- Consistent top features: precipitation + temperature variables (temp_max, temp_mean).



The optimization primarily affected the model's structure rather than its predictive performance.

Summary & Recommendations

Which Experiment Shines?

- **GAN + CNN** has the highest potential for future-focused predictions if we properly tune the network and gather robust training data.
- **Random Forest** is excellent for interpretability and simpler “where is it safe to live” questions.
- **Hierarchical Clustering + PCA** helps classify and find subtle patterns but doesn’t directly predict the future—more of an exploratory tool.

Next Steps

- **Data Gathering**: Integrate satellite/radar imagery, health stats, and extreme-event records.
- **Optimization**: Scale inputs, handle class imbalance (e.g., weighted classes for CNN).
- **Deploymen**t: Start with simpler Random Forest models for near-term insights; then implement CNN+GAN for broader, long-range forecasting.

Thank you

Questions?
Let’s Discuss!

Isaac Contreras

isaac.contreras.ko@gmail.com

LinkedIn

GitHub Repository

Tableau Interactive Vizs