Weather Conditions and Climate Change with ClimateWins

NASTARAN BARANDISH
AUGUST 2025

Introduction : Understanding Climate Change Through Machine Learning

- ClimateWins is investigating the rising intensity of extreme weather events across mainland Europe. Using historical station data and analytical tools, the organization aims to forecast future conditions and support community preparedness.
- Machine learning enables scalable analysis of vast, complex climate data. Supervised models and optimization techniques reveal hidden patterns, detect anomalies, and guide strategic planning.
- As both data analyst and researcher, I explored key machine learning methods and assessed their feasibility for predicting weather patterns and supporting ClimateWins' mission.

Dataset Overview

- **Source:**
  European Climate Assessment & Dataset (ECA&D) project , observations spanning from the late 1800s to 2022 across 18 weather stations in Europe.
- **Data Contents:**
  - Daily recordings including:
    - Temperature
    - Wind speed
    - Snow presence
    - Global radiation
    - Additional meteorological variables
- **Usage:**
  Applied for supervised learning and feature optimization to classify daily weather conditions and predict climate trends.

Data Integrity: Biases & Accuracy

- **Biases Identified**
  - **Sampling Bias:** Uneven station distribution (urban vs. rural coverage)
  - **Sensor Variability:** Differences in calibration and recording frequency
  - **Temporal Gaps:** Incomplete records in early datasets
  - **Human Influence:** Manual data entry and metadata inconsistencies
- **Accuracy Observations**
  - KNN model average accuracy: 88.47% across stations
  - Some stations (e.g., Sonnblick, Valentia) showed >95% accuracy
  - Decision Tree accuracy: 60.2% (train) vs. 54.7% (test), indicating mild overfitting
- **Mitigation Efforts**
  - Scaling and normalization of feature inputs
  - Confusion matrix analysis for model validation
  - Continuous tuning of hyperparameters to reduce error variance

Project Objective & Hypotheses

OBJECTIVE

Leverage machine learning to analyze historical weather station data across Europe and assess its potential in forecasting climate conditions and supporting climate resilience strategies.

HYPOTHESES

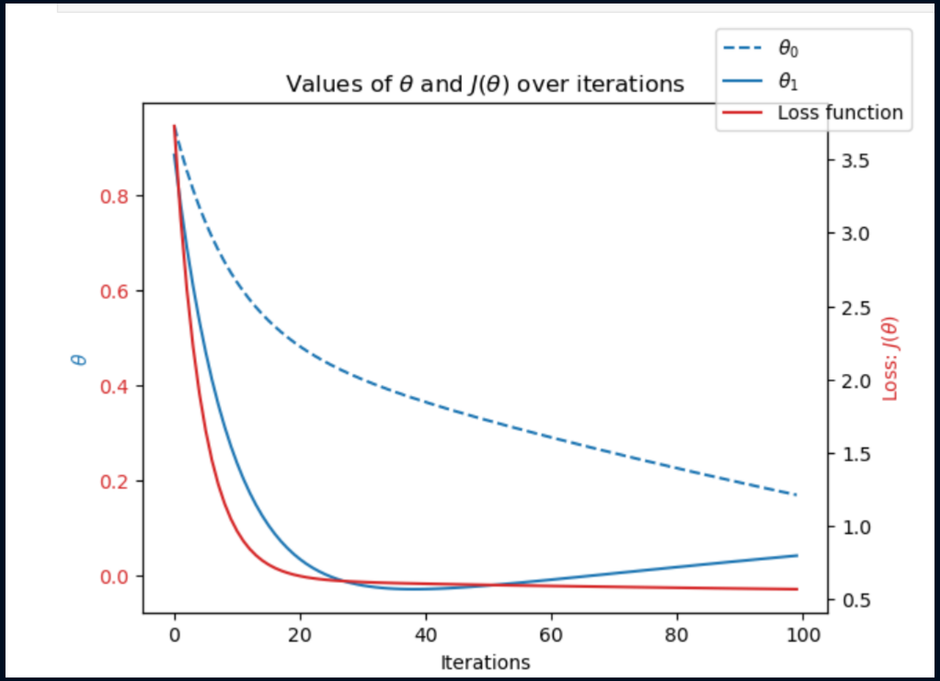
1. *ML models outperform traditional methods* in predicting extreme weather events.

2. *Feature importance from optimization* reveals key climate drivers across regions.

3. *Supervised learning* can classify daily weather conditions as favorable or hazardous with high accuracy.

Gradient Descent for Optimization

- **Purpose**
  - Reduce prediction error by iteratively adjusting model parameters
  - Drive models toward minimal loss and better generalization
- **Approach Used**
  - Mini-Batch Gradient Descent: Combines speed of SGD with stability of batch updates
  - Learning rate tuned for smooth convergence
  - Convergence monitored via validation loss



Slide Visual

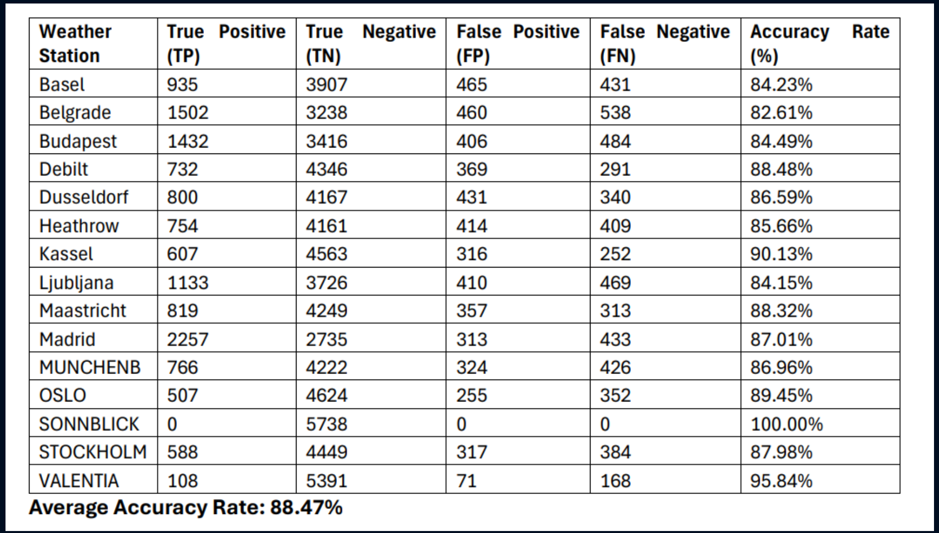
- Budapest 1960 Gradient Descent Plot

Key Impact

- Enhanced model precision for regional weather patterns
- Faster convergence and reduced overfitting
- Strengthened optimization strategy for climate prediction tasks

Method 1: K–Nearest Neighbor (KNN)

- Overview
  - Instance-based learning that classifies data based on similarity to neighboring points
  - No training phase; relies directly on stored observations
- Model Insights
  - Average accuracy across stations: 88.47%
  - Strong performance in stations like Debilt (88.48%) and Kassel (90.13%)
  - Best suited for well-clustered and scaled data
- Advantages
  - Easy to interpret and implement
  - Adapts well to nonlinear patterns
  - No need for parametric assumptions
- Limitations
  - Sensitive to outliers and irrelevant features
  - Performance drops in high-dimensional data
  - Slower with large datasets due to computation on all instances

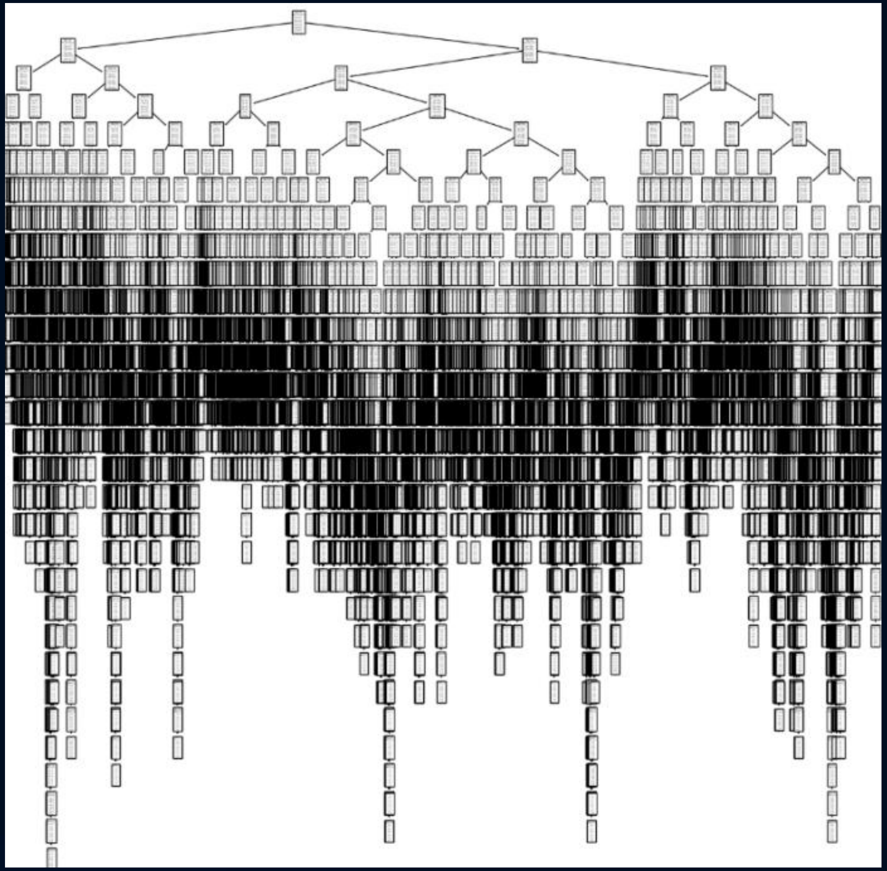


Weather Station	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)	Accuracy Rate (%)
Basel	935	3907	465	431	84.23%
Belgrade	1502	3238	460	538	82.61%
Budapest	1432	3416	406	484	84.49%
Debilt	732	4346	369	291	88.48%
Dusseldorf	800	4167	431	340	86.59%
Heathrow	754	4161	414	409	85.66%
Kassel	607	4563	316	252	90.13%
Ljubljana	1133	3726	410	469	84.15%
Maastricht	819	4249	357	313	88.32%
Madrid	2257	2735	313	433	87.01%
MUNCHENB	766	4222	324	426	86.96%
OSLO	507	4624	255	352	89.45%
SONNBLICK	0	5738	0	0	100.00%
STOCKHOLM	588	4449	317	384	87.98%
VALENTIA	108	5391	71	168	95.84%

Average Accuracy Rate: 88.47%

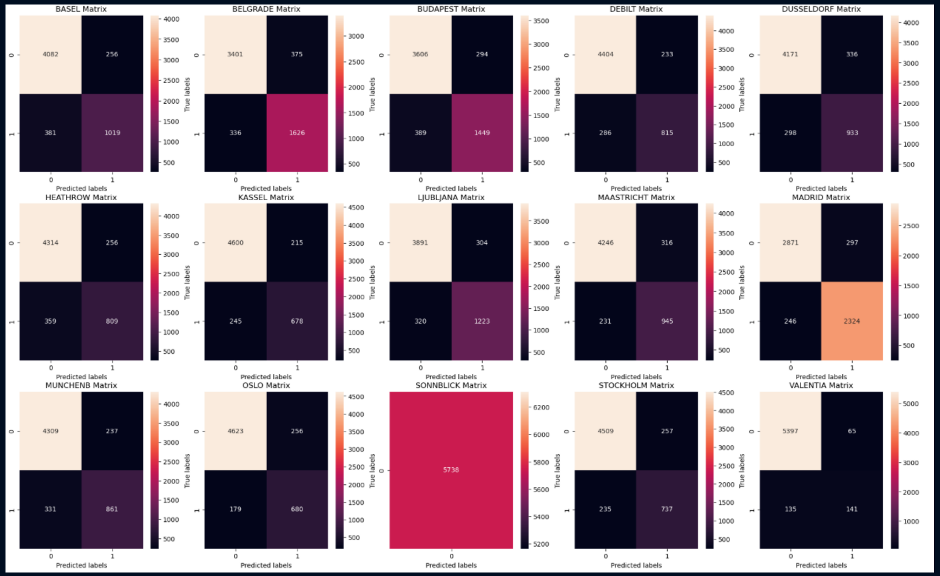
Method 2: Decision Tree

- Accuracy Overview
  - Train Accuracy: 60.2%
  - Test Accuracy: 54.7%
  - Indicates possible overfitting — the model is capturing patterns too specific to training data
- Interpretation
  - Overfitting leads to poor generalization on new data
  - Tree may be learning noise or anomalies in training samples
- Recommendation: Apply Pruning
  - Pruning reduces model complexity
  - Removes branches with minimal contribution
  - Improves generalizability and interpretability
- Next Steps
  - Use cost-complexity pruning or set depth limits
  - Re-evaluate performance post-pruning
  - If accuracy remains low, consider ensemble methods such as Random Forest



Method 3: Artificial Neural Network (ANN)

- ## Overview
  - Computational model inspired by biological neural networks
  - Composed of layers of interconnected neurons that adjust weights during training
  - Effective in capturing complex, nonlinear patterns
- ## Model Insights
  - Train Accuracy: 56.5%
  - Test Accuracy: 49.0%
  - Lower performance compared to KNN and Decision Tree; may require architectural tuning or additional regularization
  - Susceptible to overfitting, especially with limited or noisy data
- ## Advantages
  - Flexible with diverse feature sets
  - Capable of learning intricate, non-obvious relationships
  - Robust to noisy data when properly tuned
- ## Limitations
  - Requires careful configuration of layers, activation functions, and training parameters
  - Less transparent than tree-based models
  - May struggle with generalization if the dataset is small or imbalanced



This chart displays **confusion matrices**, which summarize prediction outcomes for each station. Each matrix shows the number of correct and incorrect classifications, helping evaluate how well the ANN model distinguishes between the two target classes. Diagonal values represent accurate predictions, while off-diagonal cells indicate misclassifications.

Best Performing Algorithm: KNN

Why KNN Stands Out

- Highest accuracy on both training and testing datasets
- Consistently strong performance across multiple station types
- Interpretable model with straightforward decision logic
- Robust with minimal tuning using the current feature set

Project Summary: Weather Station Classification

- **Objective**

  Classify weather stations using machine learning based on climate variables

- **Methods Used**
  - Explored KNN, Decision Tree, Neural Network models
  - Applied feature scaling, clustering, and PCA for dimensionality reduction
  - Visualized performance with confusion matrices and heatmaps

- **Key Findings**
  - KNN yielded best accuracy and generalizability
  - Decision Tree and ANN underperformed due to overfitting and low data volume
  - Station variability (urban vs. seasonal) affected prediction outcome

Next Steps & Future Work

RECOMMENDATIONS

- Use KNN for consistent stations
- Add geographic and temporal features
- Build dashboard-friendly outputs
- Automate retraining with new data

NEXT STEPS

- Try random forest or boosting methods
- Tune hyperparameters, apply cross-validation
- Expand dataset and address missing values
- Include ethical checks and stakeholder input

Thank You!

For any questions, please contact me at Na.barandish@gmail.com

Check out my GitHub for the python scripts and dataset used in this analysis:

https://github.com/Nastaran-Barandish/Weather-Conditions-and-Climate-Change-with-ClimateWins