Can Machine Learning Algorithms Predict a Good Day?

ClimateWins

Rodeesha Simmonds

August 29,2025

Objective & Hypotheses

- **Objective:**
- Use supervised machine learning to predict whether European weather will be pleasant or unpleasant.

- **Hypotheses:**
- 1. If simpler models are used, then they will generalize better to unseen weather data than more complex models.
- 2. If certain weather stations contain stronger or more complete data, then they will contribute more to prediction accuracy than others.
- 3. If weather data is scaled before training, then ANN will perform better than when trained on unscaled data.

Data Source & Biases

- **Data Source:**
- European Climate Assessment & Dataset (ECA&D), 18 weather stations (1800s–2022)

- **Biases:**
- Pleasant/unpleasant labels provided by ClimateWins
- Missing stations (GDANSK, ROMA, TOURS)
- Regional differences in defining pleasant weather
- Limited data coverage

Optimization & Preprocessing

Removed non-predictive columns: DATE, MONTH

Dropped stations with insufficient labels

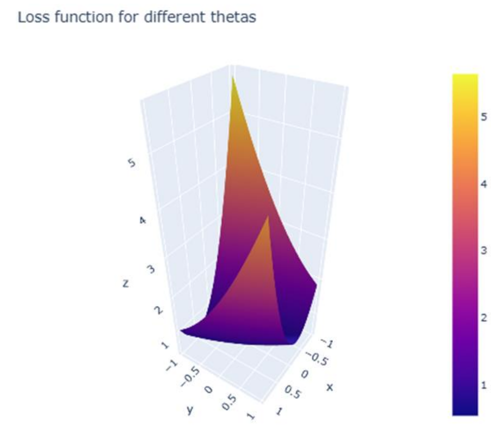
Feature scaling for ANN

Gradient Descent Optimization:

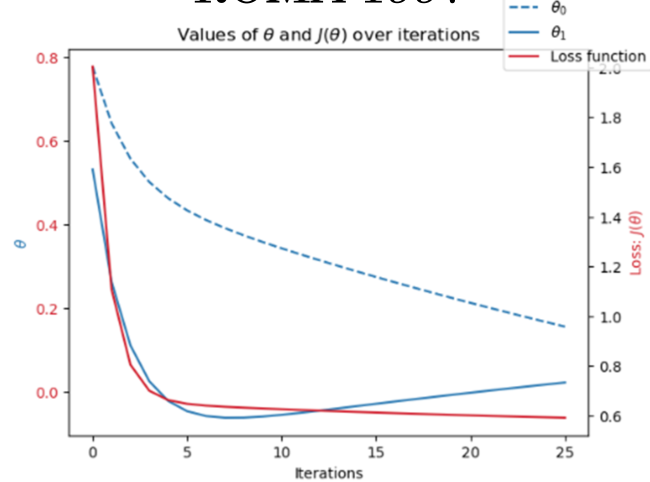
- Used iterative updates to minimize loss
- Tracked θ_0, θ_1 across iterations
- Visualized convergence using loss surfaces & contour plots
- Showed how different learning rates affect speed of convergence

Hyperparameter tuning:

- **Decision Tree**: pruned depth, min samples
- **ANN**: varied layers (20–100 nodes), iterations (500–5000), tolerance (1e-3 to 1e-4)



ROMA 1997



Algorithms Used

- **KNN**
  - Pros: Simple, effective with small feature sets
  - Cons: Slower with large data, sensitive to scaling
- **Decision Tree**
  - Pros: Interpretable, handles nonlinear relationships
  - Cons: Overfits unless pruned
- **ANN**
  - Pros: Captures complex patterns
  - Cons: Requires scaling, tuning; risk of overfitting
- **Result:**
  - KNN generalized best; ANN and Trees struggled

1

- Decision Tree (Unpruned): Train = 100%, Test ≈ 26%

2

- Decision Tree (Pruned): Train ≈ 46%, Test ≈ 31%

3

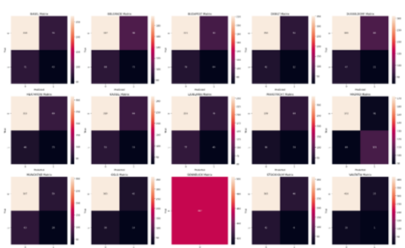
- ANN (50,50): Train = 100%, Test ≈ 26%

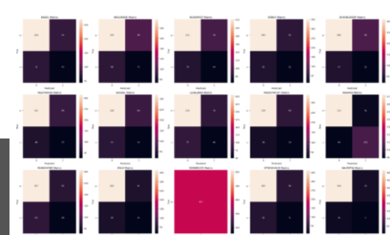
4

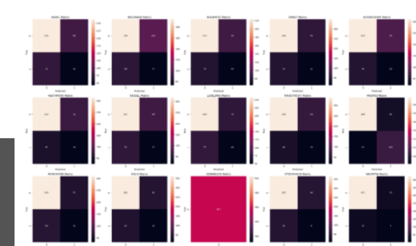
- ANN (20,20, tol=1e-3, max_iter=5000): Train ≈ 78%, Test ≈ 29%

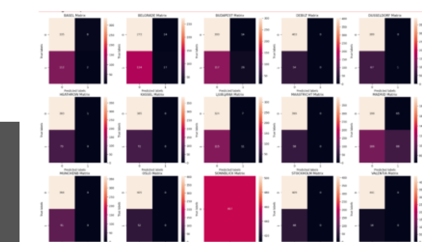
5

- KNN: Test ≈ better balance than Tree/ANN









Accuracy

KNN Accuracy

- Best Test Accuracy: ~30% (around k=7–9)
- Best Training Accuracy: ~100% (at k=1), but declines with larger k
- Trend: Small k: Overfits (perfect train accuracy, poor generalization)
- Medium k: Balanced accuracy — best test performance
- Large k: Underfits (train/test both low)
- General Range:
  - Train: 70–100%
  - Test: 25–30%

Summary of Results

- **KNN**: Best relative test accuracy, balanced results: Train: ≈100%, Test: ≈30%
- **Decision Tree (Unpruned)**: Overfit; perfect training but poor testing
- **Decision Tree (Pruned):** Lower train accuracy, slightly better test accuracy
- **ANN (50,50):** Overfit; failed to generalize
- **ANN (20,20, 5000 iters):** Better balance (Train ≈ 78%, Test ≈ 29%)
- *No station reached full accuracy*
- *Overall accuracy modest (25–35%) → limits in current dataset*

Limitations & Next Steps

- Missing stations reduce geographic balance
- Labels subjective across regions
- Accuracy capped at 25–35%
- Risk of overfitting in complex models

- Add more stations & fill gaps
- Engineer new features (humidity, wind chill)
- Try ensembles (Random Forest, Boosting)
- Apply cross-validation across stations

Thank You

Link: Rodeesha1/Machine-Learning-with-Python-Basics: Machine learning to help predict the consequences of climate change for European nonprofit organization, ClimateWins