Achievement 2.3: Complex Machine Learning Models and Keras—Part 2 **Olufemi Ige** (02/03/2026)

Overview

This report documents the development and evaluation of Random Forest classification models designed to identify the most influential weather indicators of a "pleasant day." Using historical European weather data, the analysis examines both a multi-station model and focused, station-specific models to understand which locations and environmental variables contribute most strongly to prediction accuracy.

The work is structured in two stages:
1. All-Stations Model: A Random Forest model trained on combined data from multiple European weather stations (2010–2019).
2. Individual Station Models: Separate Random Forest models trained on data from the three most influential stations identified in the first stage: Basel, München, and Maastricht.

The findings support ClimateWins' goal of improving forecasting efficiency and guiding investment in high-value weather stations and sensors.

Methods Summary

Data Preparation

Weather data from multiple European stations were cleaned and standardized. Three stations (Gdansk, Roma, Tours) and two observation types (depth and wind speed) were removed to ensure consistency. The final dataset included: - 15 weather stations - 9 weather variables per station - 135 total features.

Time Filtering

To ensure comparability and manageable scope, the dataset was restricted to the decade 2010–2019, resulting in 3,652 daily observations with corresponding labels indicating whether each day was considered pleasant.

Feature Selection

All weather variables were retained for modeling, while non-predictive fields such as date and month were excluded. This allowed the model to learn relationships directly from environmental conditions.

Modeling Approach

A Random Forest Classifier with 100 decision trees was used. The dataset was split into 75% training and 25% testing subsets. Random Forest was selected for its strong performance on tabular data and its built-in feature importance metrics.

Feature Importance Analysis

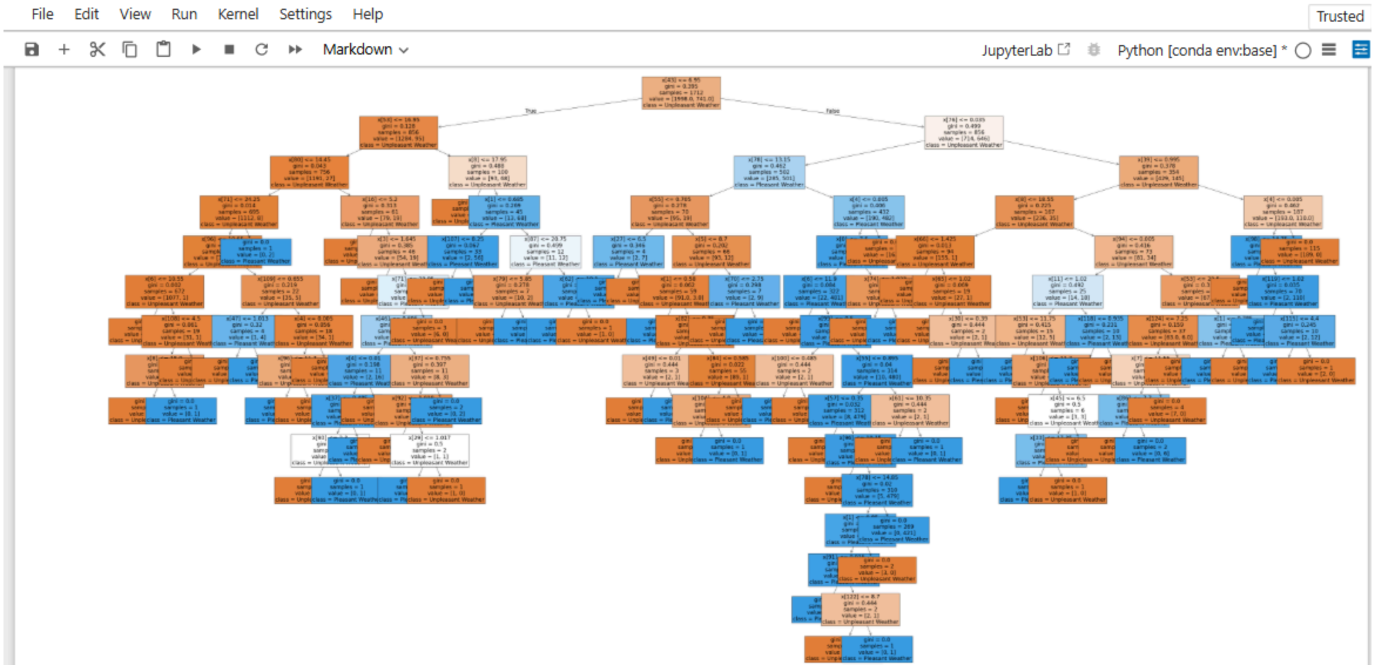
Post-training, feature importance scores were aggregated by station to determine which locations most strongly influenced predictions. This step informed the selection of stations for more in-depth analysis.

Random Forest Model—All Stations (2010s)

The initial model was trained using data from all stations covering the 2010–2019 period.

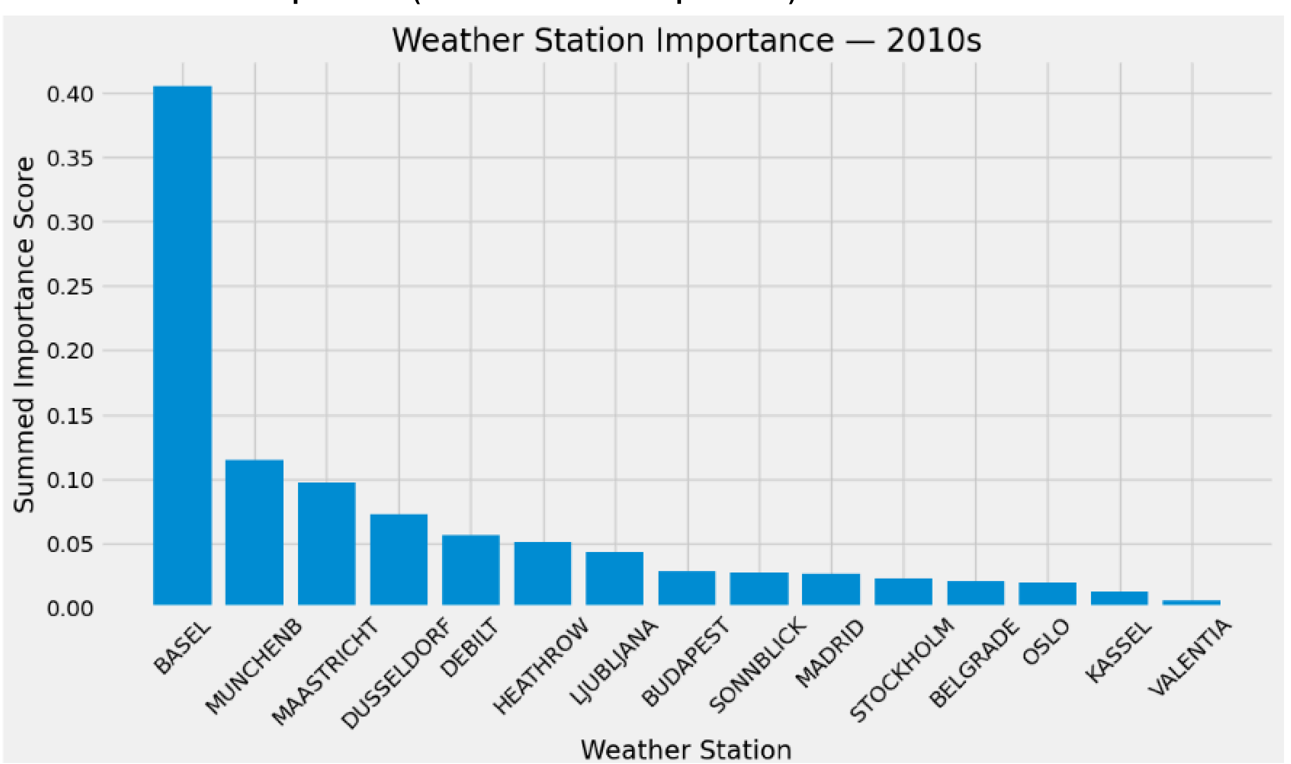
Key Results: - Test accuracy: Approximately 99% - Strong generalization without manual depth tuning.

Figure 1. Decision Tree (All Stations, 2010s)



Random Forest decision tree—all stations, 2010s

Figure 2. Weather Station Importance (Summed Feature Importance)



Weather station importance bar chart – 2010s

Station Importance Ranking: 1. Basel—dominant influence on predictions **2. München –** strong secondary influence **3. Maastricht—**third-highest contribution.

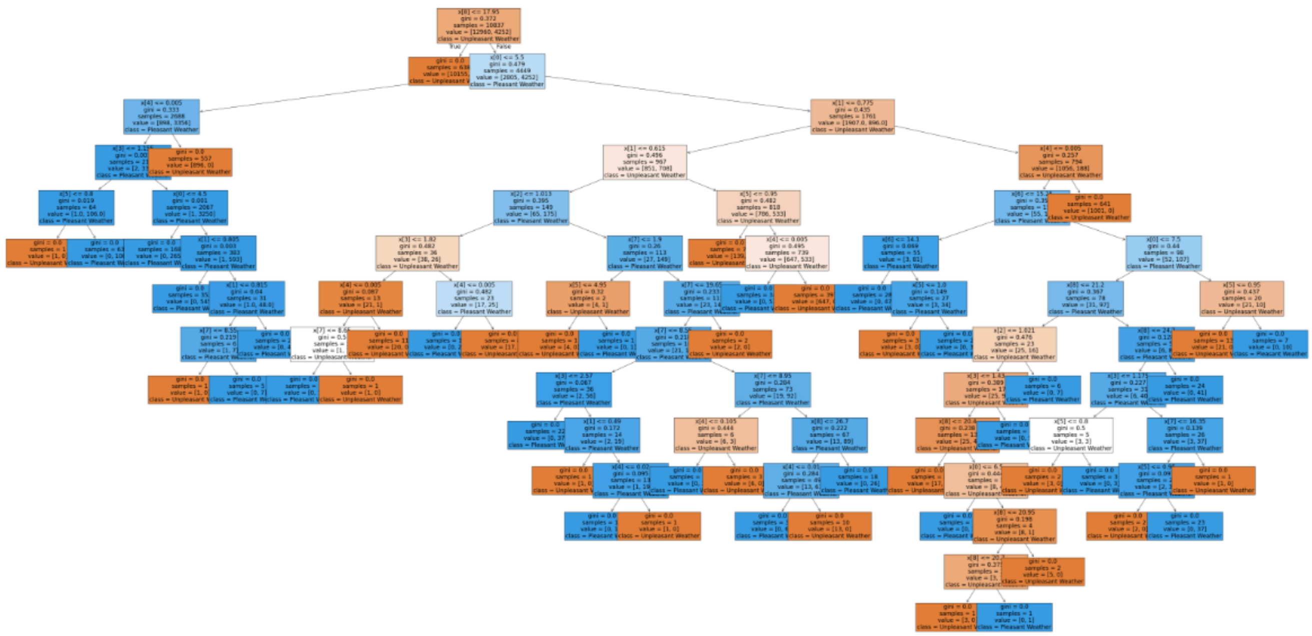
Although all stations contributed information, these three locations accounted for a disproportionately large share of the model's predictive power.

Individual Random Forest Model—Basel (All Years)

A station-specific model was trained using all available Basel data.

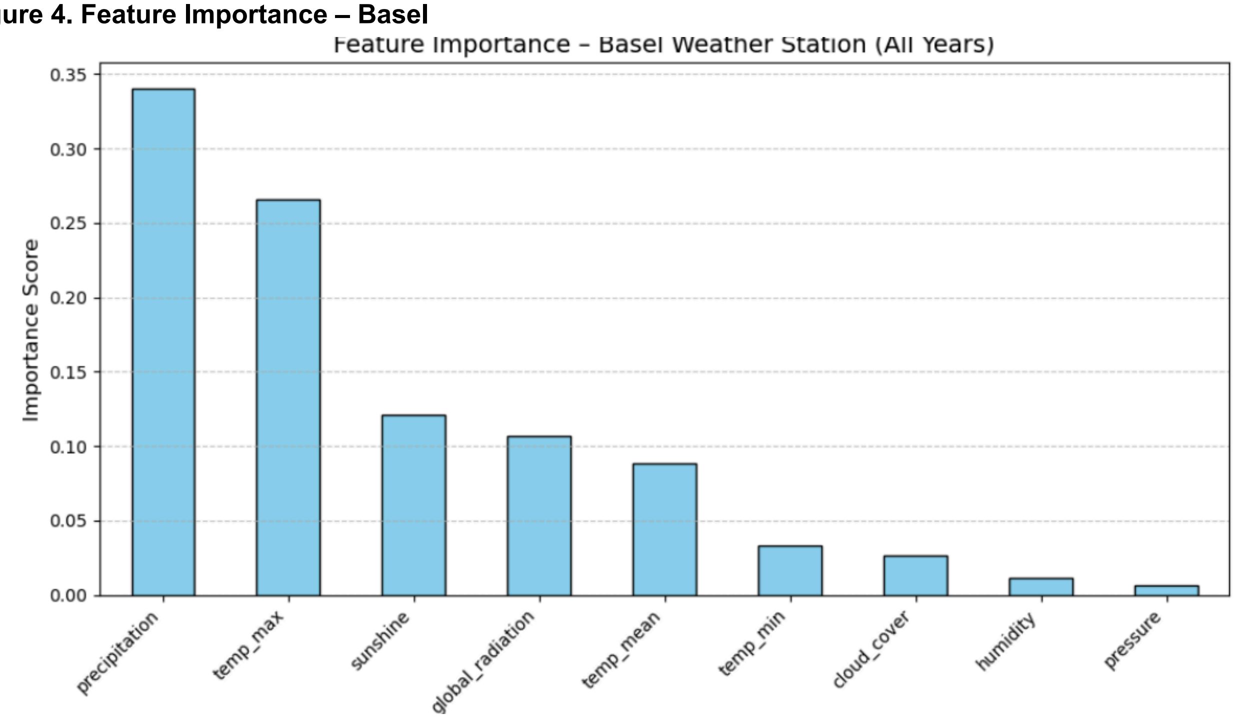
Performance: - Test accuracy: **100%** (excellent fit; potential overfitting considered)

Figure 3. Basel Random Forest Decision Tree



Basel decision tree visualization

Figure 4. Feature Importance – Basel



Basel feature importance bar chart

Top Predictors: - Precipitation (strongest indicator) - Maximum temperature - Sunshine - Global radiation

Lower-Importance Variables: - Humidity - Air pressure - Cloud cover

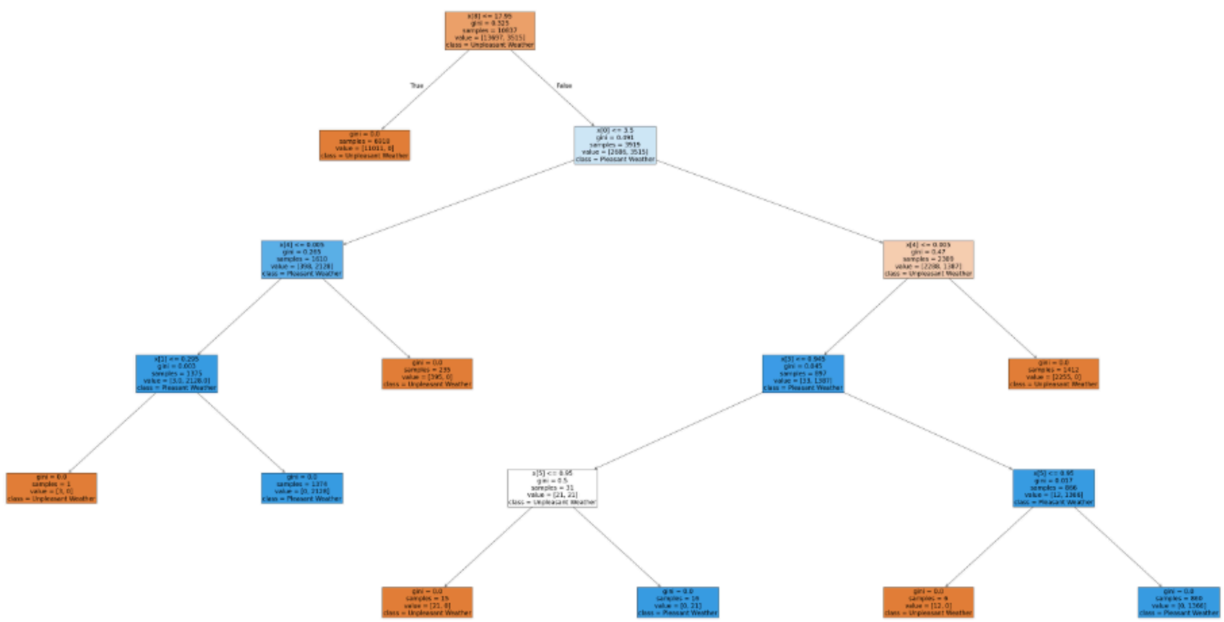
These results show that dry, warm, and sunny conditions are the clearest signals of a pleasant day in Basel.

Individual Random Forest Model—München (All Years)

The München station model demonstrated similarly strong performance.

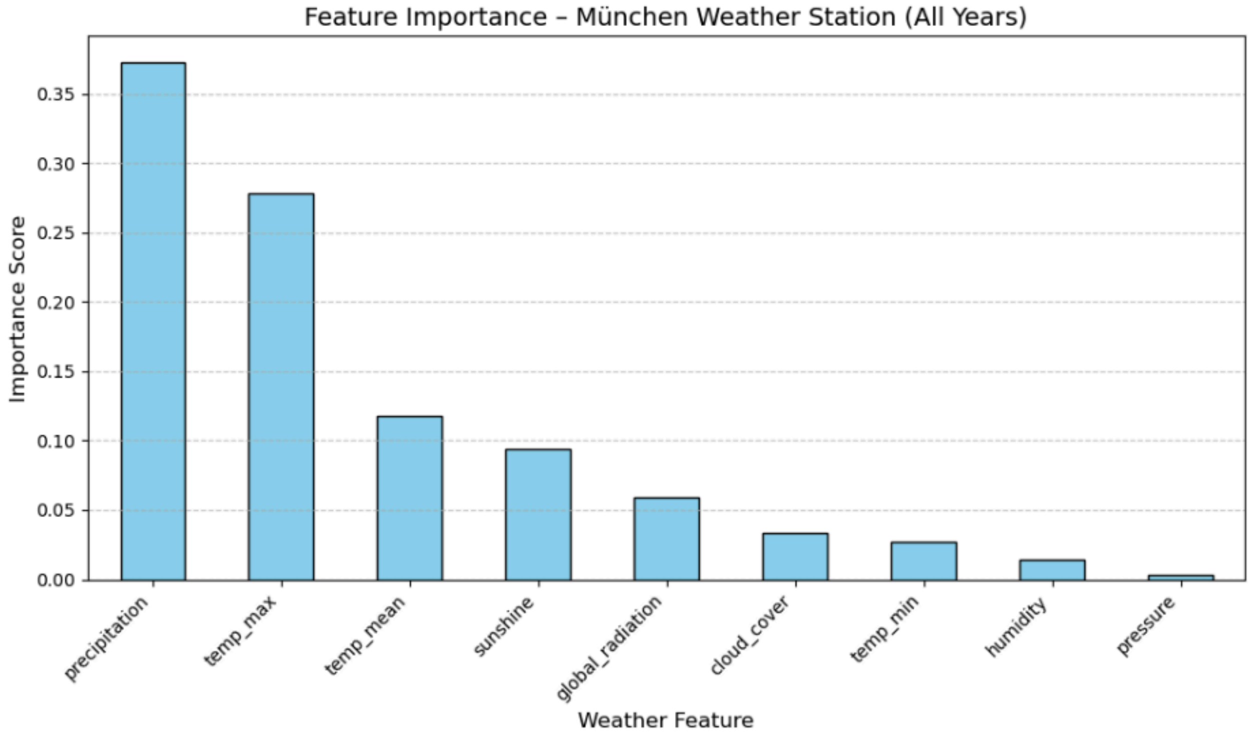
Performance: - Test accuracy: 100%

Figure 5. München Random Forest Decision Tree



München decision tree visualization

Figure 6. Feature Importance – München



München feature importance bar chart

Top Predictors: - Precipitation - Maximum temperature - Sunshine - Mean temperature - Global radiation

Lower-Importance Variables: - Cloud cover - Minimum temperature - Air pressure - Humidity

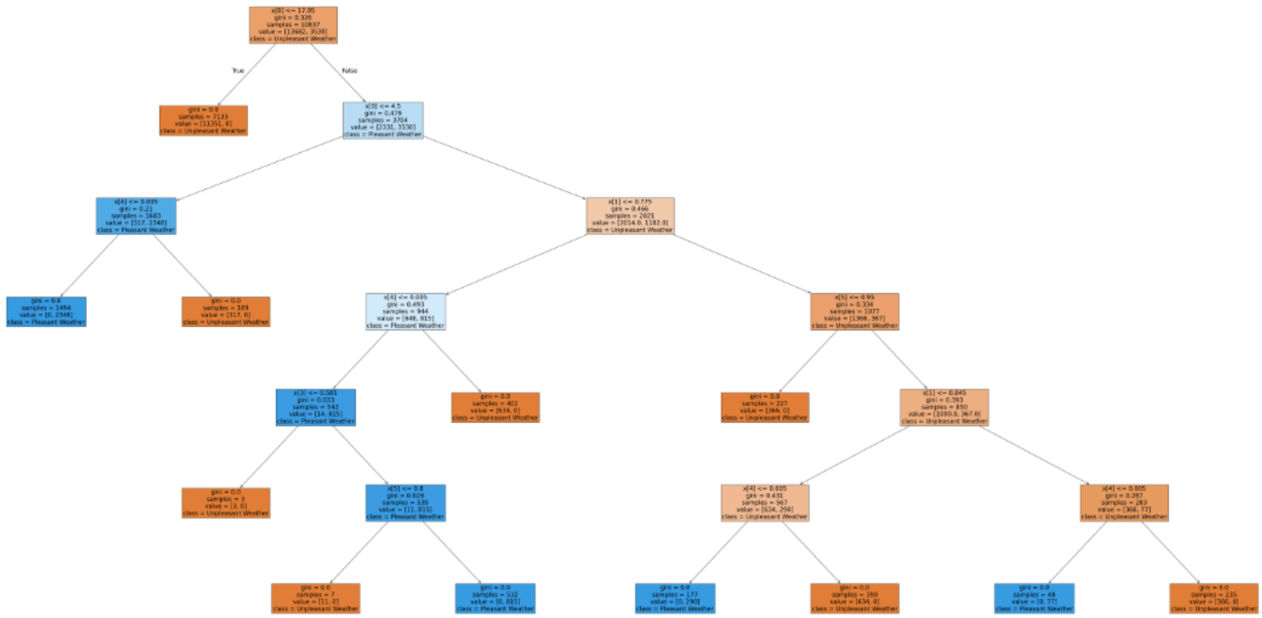
The results closely mirror Basel's patterns, reinforcing the importance of rainfall and warmth in determining pleasant weather.

Individual Random Forest Model—Maastricht (All Years)

The Maastricht model also achieved perfect classification accuracy on the test set.

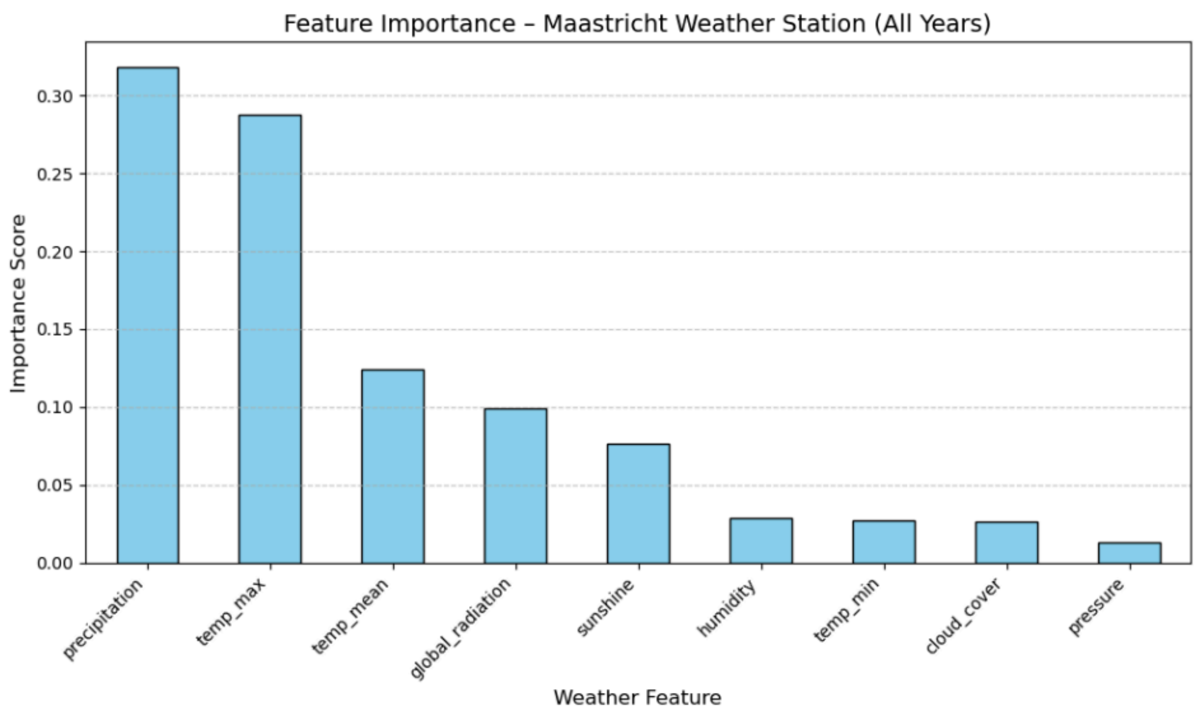
Performance: - Test accuracy: 100%

Figure 7. Maastricht Random Forest Decision Tree



Maastricht decision tree visualization

Figure 8. Feature Importance – Maastricht



Maastricht feature importance bar chart

Top Predictors: - Precipitation - Maximum temperature

Supporting Predictors: - Mean temperature - Sunshine - Solar radiation

Lower-Importance Variables: - Humidity - Minimum temperature - Cloud cover - Air pressure

Across all three stations, the same core weather signals repeatedly drove accurate predictions.

Insights and Conclusions

All-Stations Model:

- Achieved near-perfect accuracy using a wide set of weather features.
- Basel emerged as the most influential station, contributing roughly two-fifths of total station importance.
- München and Maastricht also provided strong, consistent signals.

Individual Station Models

- All three station-specific models achieved 100% test accuracy.
- Precipitation and maximum temperature were the dominant predictors across locations.
- Variables such as humidity and air pressure consistently showed low importance.

These findings suggest that pleasant-day prediction can be driven by a relatively small subset of weather variables.

Recommendations for ClimateWins

1. **Prioritize High-Value Weather Signals:**
   Focus forecasting efforts on rainfall and daily maximum temperature, supported by sunshine and solar radiation.
2. **Invest Strategically in Sensors**
   Allocate resources toward high-impact stations—particularly Basel, München, and Maastricht—and ensure accurate precipitation and temperature measurement.
3. **Adopt Location-Specific Models**
   Tailored station-level models outperform generalized approaches and should be used where sufficient local data exists.
4. **Validate Across Time and Conditions**
   Test models on different years and extreme seasons to confirm robustness and avoid overfitting.

Key Takeaway:
This analysis highlights the effectiveness of Random Forest models for structured climate data and demonstrates the value of feature importance analysis in guiding practical decision-making. While perfect accuracy is encouraging, it also underscores the importance of rigorous validation to ensure real-world reliability.

Overall, the analysis provides ClimateWins with clear, actionable insights into which stations and weather variables matter most when predicting pleasant days.