ClimateWins Data Analysis

KACI ERWIN

MAY 2024

Objective

Leverage machine learning to predict future weather conditions and potential extreme weather events in mainland Europe using historical climate data.

The data provided could be analyzed to test these hypotheses

- Hypothesis: There has been a significant increase in average annual temperatures in mainland Europe over the past century.
- Hypothesis: The frequency of extreme weather events, such as heatwaves and heavy precipitation, has increased over the last century
- Hypothesis: Machine learning models can accurately predict favorable weather conditions and extreme weather events in mainland Europe based on historical weather data.

Climate Data

Source
The data used by ClimateWins is the European Climate Assessment and Dataset (ECA&D), collected and funded by the Royal Netherlands Meteorological Institute (KNMI).

Contents
The ECA&D dataset includes a wide range of weather information, such as temperature, precipitation, wind speed, and humidity, from 18 weather stations across Europe. It spans multiple decades, enabling detailed analysis of historical weather patterns and trends.

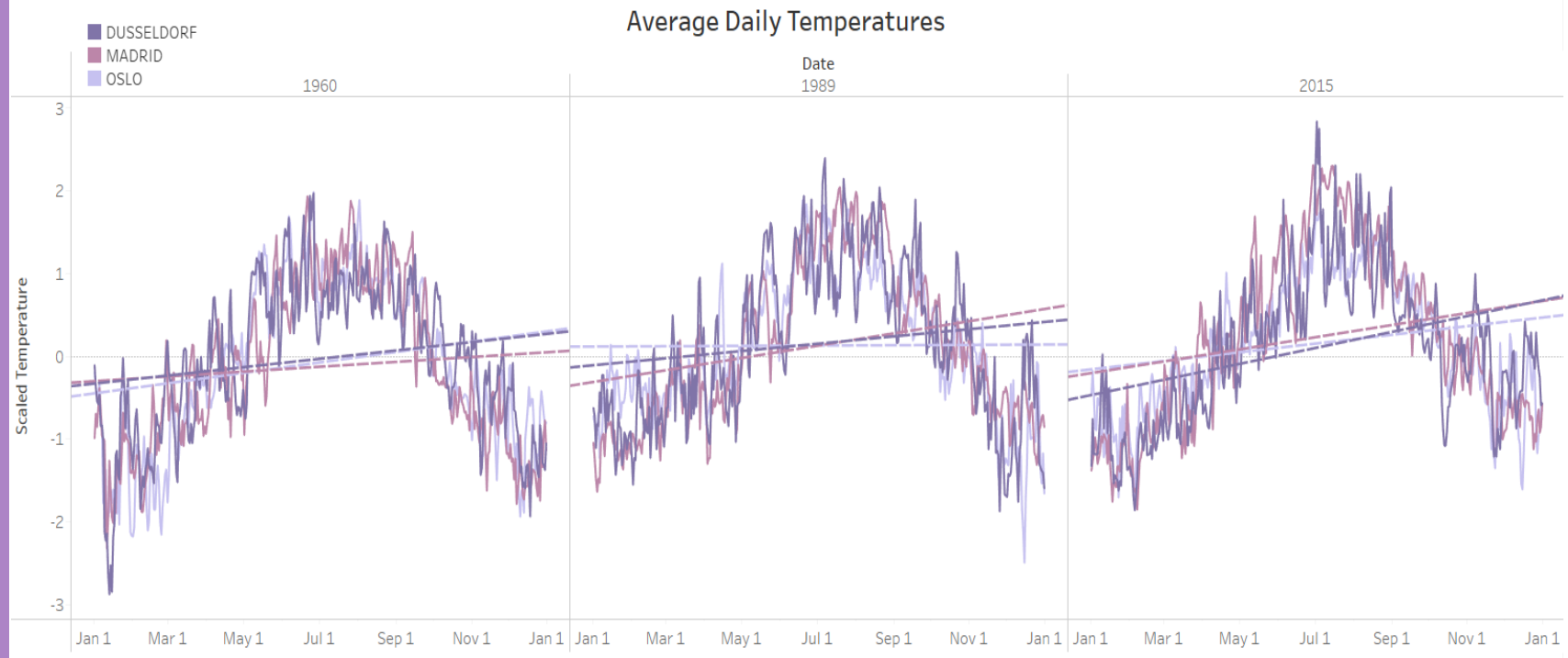
Accuracy
The data is meticulously collected and validated by the KNMI to ensure high accuracy and reliability. The weather stations are strategically located across Europe, providing a representative sample of the continent's diverse climate conditions.

Bias
Potential biases may arise from the geographic distribution of weather stations, which might not fully capture local variations in climate. Also, historical data collection methods and changes in measurement technology over time can introduce inconsistencies.

Gradient Descent Optimization

Gradient descent optimization was used to determine the features of the dataset by fitting a linear model to scaled daily average temperatures from 3 locations in 3 different years. The resulting slopes for each fit are shown.

Slopes (theta1) from gradient descent optimization			
Location	1960	1989	2015
Madrid	0.095	0.24	0.24
Oslo	0.2	0.0065	0.17
Dusseldorf	0.16	0.14	0.31



Gradient Descent Optimization

Results

- All three weather stations have positive slopes for all 3 years analyzed. These slopes indicate an upward trend in average temperatures each year suggesting that warming has occurred over the past several decades.

- The magnitude of the slopes do not follow a chronological trend. The largest magnitude is found in different years for each location. This suggests that while warming is occurring it is not consistent.

- The magnitude of the slopes do not follow a geographical trend. The largest magnitude is found at a different location for each year analyzed, indicating that the rate of temperature increase varies spatially across Europe. This suggests that different regions may be experiencing the impacts of climate change at different rates and intensities.

The analysis so far is limited. The analysis should be repeated for more years and more stations to provide more confidence in the results. The analysis should also be expanded to incorporate more climate data variables to provide more holistic results.

Supervised Machine Learning

A dataset that showed pleasantness ratings (as either 0 or 1) for each day at each station was added to the climate dataset. Three algorithms were used to model the data to predict pleasantness based on the climate variables.

K-Nearest Neighbors: This algorithm classifies a data point based on the most common class among its nearest neighbors in the feature space, using a distance metric to determine closeness.

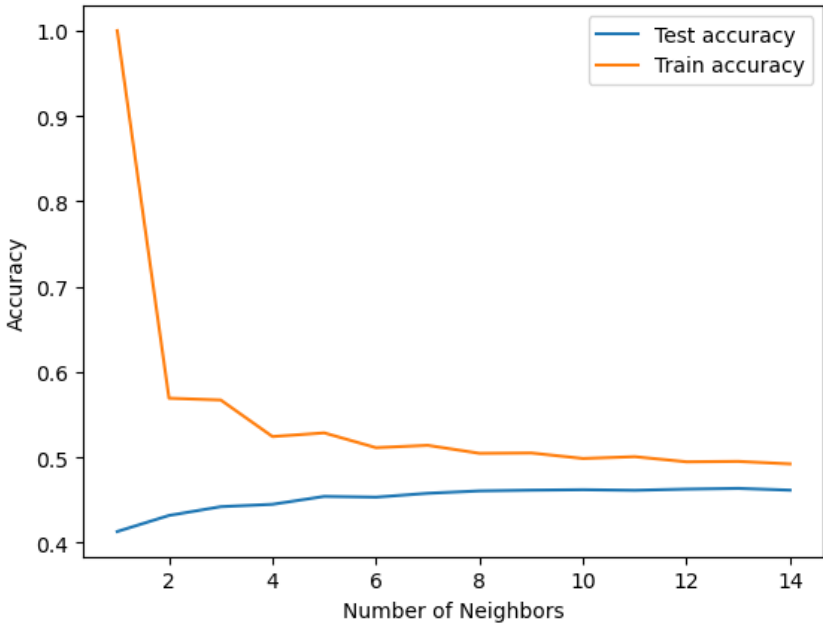
Decision Tree: This model predicts outcomes by splitting the dataset into subsets based on feature values, creating a tree where each node represents a decision point, leading to a prediction at the leaf nodes.

Artificial Neural Network (MPClassifier): This algorithm uses interconnected layers of neurons that adjust their weights through training to learn complex patterns and make predictions.

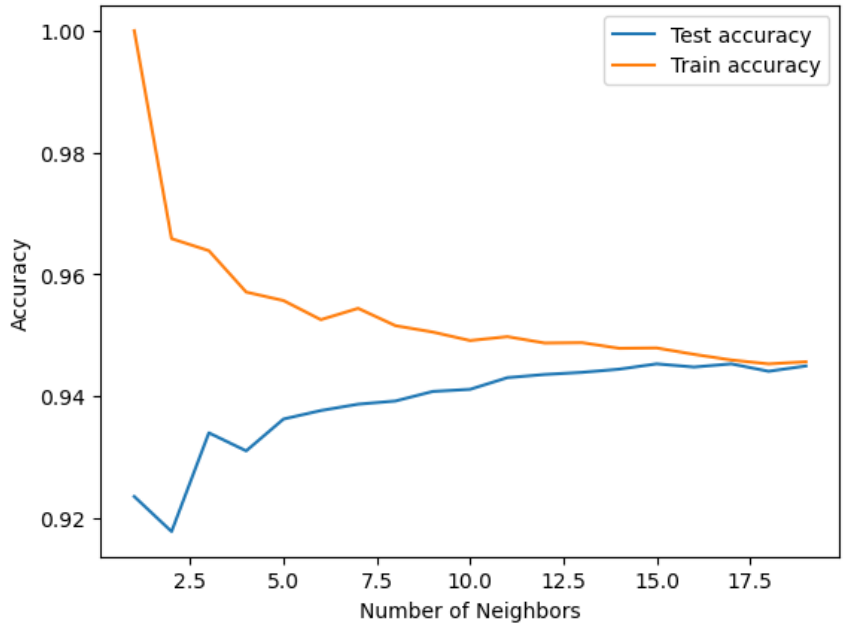
K- Nearest Neighbors: Analysis

Full Dataset:
A KNN model fit to the full climate dataset converges around 6 neighbors with a training accuracy of 0.49 and a testing accuracy of 0.46.

Accuracy vs Neighbors for Full Dataset



Accuracy vs Neighbors for Heathrow Dataset



Limited Dataset:
A KNN model fit to climate data for only the Heathrow station converged around 13 neighbors with a training accuracy of 0.95 and a testing accuracy of 0.94.

K- Nearest Neighbors: Conclusion

Results

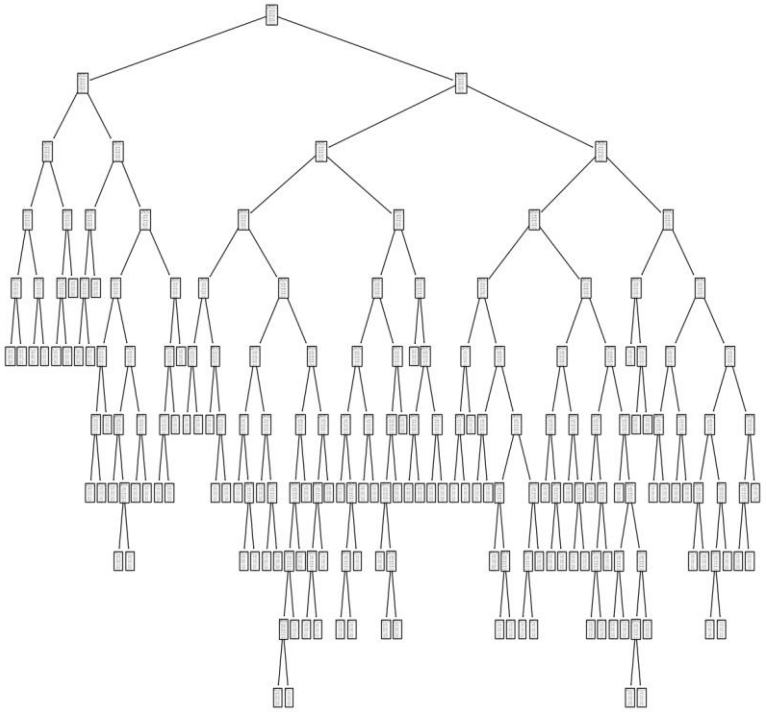
While the KNN model converges when fitted to the full dataset, the low accuracies indicate that the data's complexity leads to **underfitting**. In contrast, the high accuracies achieved by the KNN model on the simpler Heathrow dataset suggest that the model is **well-suited and appropriate for this simpler data.**

Conclusions

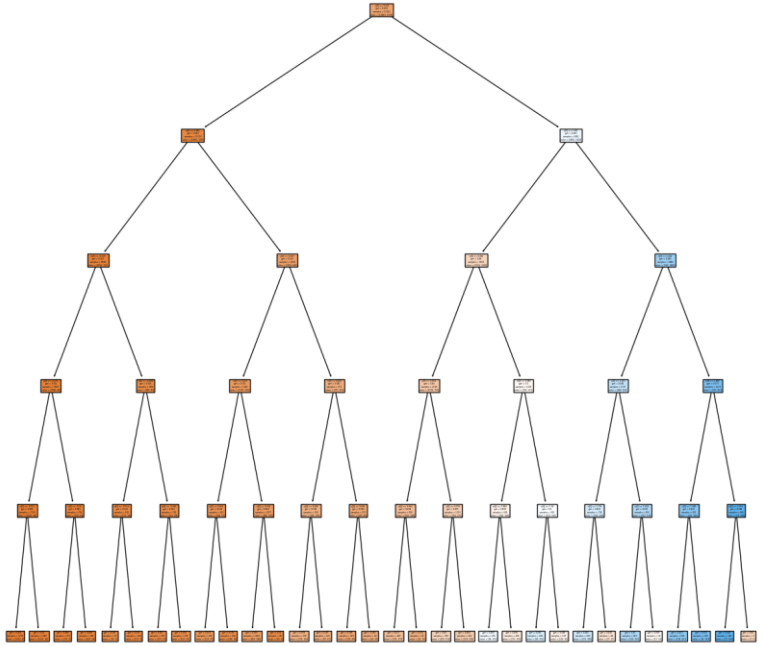
The KNN model shows promise for predicting weather pleasantness, particularly with simpler datasets. However, the low accuracies when applied to the full dataset indicate that the model may struggle with the complexity and variability of larger, more diverse data. This suggests that KNN, while appropriate in certain contexts, may benefit from **additional preprocessing** to improve its performance on complex datasets. Future steps could include **feature engineering** and possibly **integrating more advanced modeling techniques** to handle the complexity of the full dataset.

Decision Tree: Analysis

After parameter optimization, the decision tree created to fit the full dataset was complex with many nodes and long branches. The accuracy of this model fit to the training data is 0.83. The accuracy of this model fit to the testing data is 0.44.



Decision Tree for Full Dataset



Decision Tree for Heathrow Dataset

A decision tree was created to fit a dataset that contained only data for the Heathrow location. The accuracy of this model fit to the training data is 0.83. The accuracy of this model fit to the testing data is 0.84. The model shows bias in predictions with the “pleasant” outcome being mislabeled about half of the time.

Decision Tree: Conclusions

Results

The parameter-optimized decision tree model for the full dataset was complex and **overfit** the training data. The decision tree model achieved higher accuracy for a simpler dataset, but the model showed **bias** in the pleasantness predictions.

Conclusions

These results suggest that a decision tree can effectively model simpler, more homogeneous datasets but may struggle with more complex, heterogeneous data. If a decision tree model is chosen, the data may need to be **segmented into simpler subsets** for higher accuracy or **ensemble methods (Random Forests)** or **feature engineering techniques** could be implemented to improve the model performance for the full dataset.

Artificial Neural Network (MPCLassifier): Analysis

Parameters for an ANN model were adjusted while the model was fit to the full dataset. As adjustments were made, the training accuracies increased while the testing accuracies remained constant.

	Run 1	Run 2	Run 3	Run 4	Run 5
Layers	2	3	3	3	3
Nodes	5, 5	20, 10, 10	100, 50, 20	200, 100, 100	200, 200, 200
iter	500	1000	1000	1000	1000
tolerance	0.0001	0.0001	0.0001	0.000001	0.000001
converged?	yes	yes	yes	yes	yes
Test Accuracy	0.44	0.45	0.44	0.44	0.42
Train Accuracy	0.44	0.46	0.48	0.58	0.78

Artificial Neural Network (MPCLassifier): Conclusions

Results

While the ANN models converge when fit to the full dataset, the resulting accuracies indicate that as the models become more complex, they begin to **overfit** the data. This overfitting suggests that the models are capturing noise and specific patterns in the training data rather than generalizable trends.

Conclusion

An ANN model that fits the data appropriately has not yet been identified. Despite this, ANN models have many parameters that can be adjusted to help the models fit complex data more effectively. These include **network architecture, activation functions, regularization techniques, training epochs**, and **early stopping**. Further model refinement could identify the optimal configuration for this dataset.

Conclusions: which model should we use?

While all models show potential with appropriate data segmentation and/or further model tuning, the more complex models (Decision Tree and ANN) have the highest potential to fit the dataset. A hybrid model approach is recommended.

Recommendations

Tuning and hybrid approach: Experiment with parameter tuning, ensemble methods, and hybrid models to balance the complexity of the data with the complexity of the models.

Data Segmentations: If the models prove difficult to appropriately tune, consider segmenting data by region to reduce variables and improve model accuracy.

Regularization: Apply regularization methods in complex models to mitigate overfitting and improve generalization.

Thank you!