A Weather Story

By

Thomas Efstathiades
25.11.2025

Objective

The objective of this project is to help predict the consequences of climate change, while using machine learning.

Objective

ClimateWins, an European nonprofit organization, aims to categorize and predict the weather in mainland Europe and iis concerned with the increase in extreme weather events, especially within the past 10 to 20 years.

Research Questions I

1. Clarifying questions
   - 1.1. Has there been a significant change in the amount of “unpleasant” days per year?
   - 1.2. Did mean temperature change during this timeframe?
   - 1.3. Have range and variance of temperatures changed?
   - 1.4. Did snow depth change over time?
   - 1.5. Did humidity change?
2. Adjoining questions
   - 2.1. Did the amount of “unpleasant days” per year increase or decrease?
   - 2.2. Did mean temperature change geographically by region?
   - 2.3. Did mean temperature increase or decrease over the years?
   - 2.4. Have range and variance of temperatures changed geographically by region?
   - 2.5. Have range and variance of temperatures increased or decreased?
   - 2.6. Did snow depth increase or decrease over time?
   - 2.7. Did humidity increase or decrease?

Research Questions II

1. Funneling questions
   1.1. Is there a trend in the amount of “unpleasant” days per year?
   1.2. Is there a certain trend in mean temperature during this timeframe?
   1.3. Is there a certain trend in the range and variance of temperatures?
   1.4. Is there a certain trend in snow depth?
   1.5. Was a substantial increase in CO2-emissions recorded?
   1.6. Is there a certain trend in humidity?
2. Elevating questions
   2.1. Where there any significant environmental or industrial changes?
   2.2. Any significant changes in environmental or industrial policies?
   2.3. Any natural disasters with potential influence on the weather?

Research Hypothesis

If climate change continues to impede European weather patterns, then the number of unpleasant days per year will increase significantly over time.

Null - Hypotheses (H0)

1) H0: The number of “unpleasant” days per year is not significantly changing over the years.
2) H0: Mean temperatures are not significantly changing over the years, while range and variance of temperatures are keeping the same.
3) H0: Mean snow depth is not significantly changing throughout the years, while mean humidity is remaining the same.

Data Set

➢ The data is collected by the *European Climate Assessment & Data Set project (ECA&D)*.

➢ *ClimateWins* been sorting through hurricane predictions from The National Oceanic and Atmospheric Administration (NOAA) in the U.S., typhoon data from The Japan Meteorological Agency (JMA) in Japan, world temperatures and a great deal of other data.

Data Set

➢ The data set for this project is based on weather observations from 18 different weather stations across Europe, which contain data ranging from the late 1800s to 2022.

➢ Recordings exist for almost every day with values such as temperature, wind speed, snow, global radiation and more.

Potential Bias I

Collection bias ->

- data collection methods
- personal information

Representation or sample bias ->

- localisation of weather stations -> *Diversity*, in this sense, may mean more representative, and therefore, more accurate data, and further, weather predictions being accessible to all.

Potential Bias II

Human or cultural bias ->

- certain level of training and experience in validating prototype models
- different believes of validators regarding climate change may also skew the outcome.

Regional bias ->

- uneven distribution of weather stations across Europe as well as hurricane data streamed from the U.S. and typhoon data from Japan.

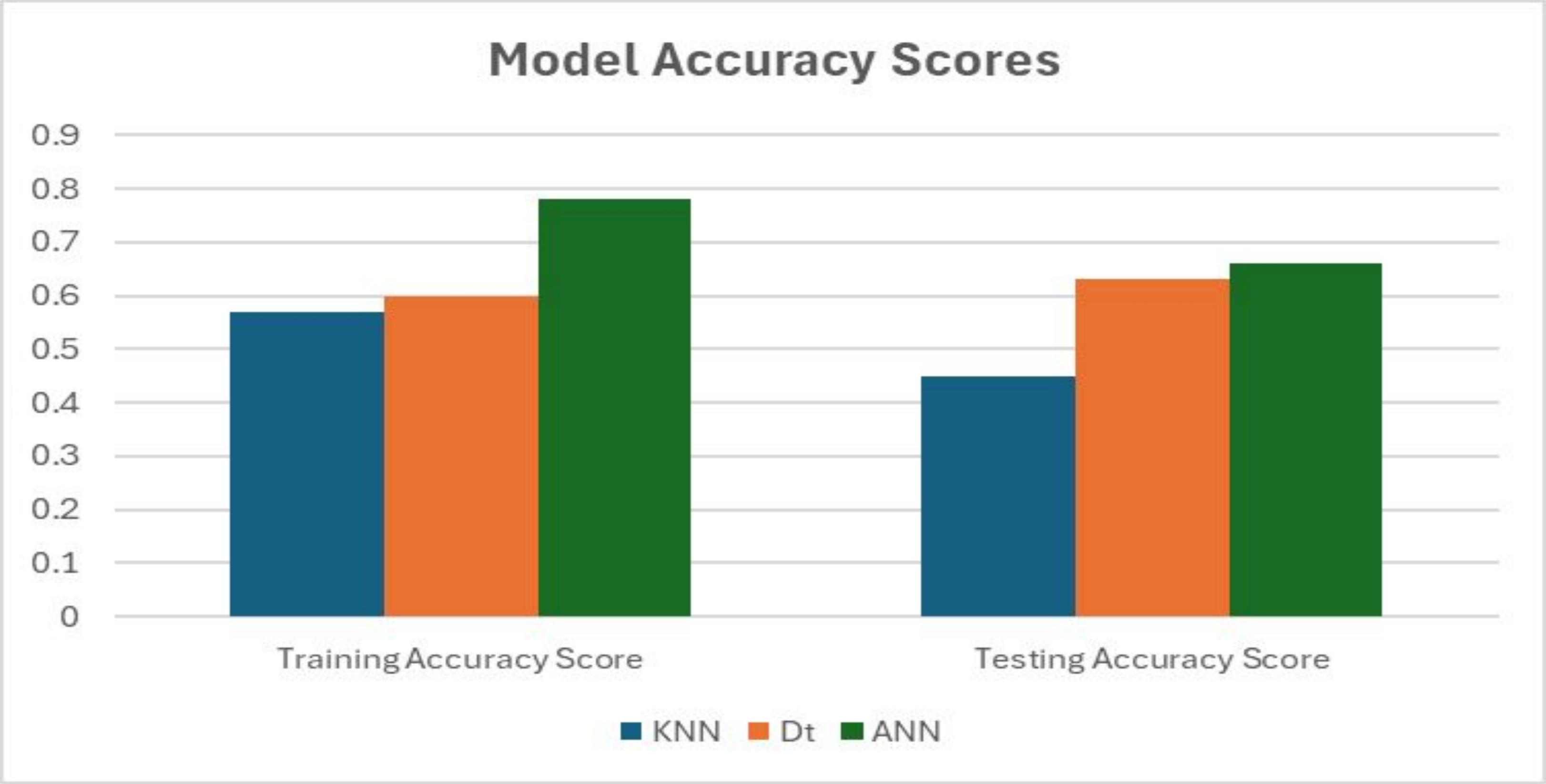
Validity & Accuracy Of The Data

- The European Climate Assessment & Data Set project (ECA&D) was initiated by the ECSN in 1998 and has received financial support from the EUMETNET and the European Commission. EUMETNET is a collaborative network comprising 33 European National Meteorological Services.

Validity & Accuracy Of The Data

- Can be considered internal data from an official European website, this data can be assigned as trustworthy. Although, technically, it’s external data provided by third-parties.

Accuracy Scores Of The Different Models



Optimization Technique

- The *Gradient Descent method* has been used for optimizing the predictability of mean temperatures by supervised machine learning.
- As such the *linear regression model* has been used.
- The *Gradient Descent algorithm* converged towards ‘0’, and thus led to optimization of the linear regression model, throughout all of the data sets of 3 different weather stations and over 3 different years each.

Supervised Learning

- Various *supervised learning algorithms* have been assessed on the weather data, *linear regression, k-Nearest Neighbor (KNN), Decision Tree and Artificial Neural Network (ANN)*.
- After assessing and comparing the accuracy of these models, it’s appearing that the *ANN model* might be the most effective in predicting the weather conditions on the data provided.
- However, *unsupervised learning* is still to be assessed on the data and may potentially provide an even better solution.

Summary

The main H0 for statistical analysis is the following:

> H0: The number of “unpleasant” days per year is not significantly changing over the years.

The ANN model may be used, as it provides the most accurate results so far, in predicting the unseen values.

Next steps: as opposed to supervised learning, unsupervised learning is still to be assessed.

Future analysis: results of this analysis may further be linked to bad weather events, e.g. hurricanes, typhoons, etc., to find connections and improve the predictability.

Summary Table

Supervised ML Model	Accuracy	Context	Strengths	Limitations	Notes
KNN	~ 50%	linear & non-linear data	ease of implementation	size of data set & "curse of dimensionality"	scaling is important
Dt	~ 60%	linear & non-linear data	relative fast/low computational cost	prone to overfitting/jumping to noise	manually adjusting hyperparameters
ANN	~ 70%	linear & non-linear data	best for complex non-linear data	computational resources	very powerful but requires more data

Questions ?

Contact details

thomas.ef@outlook.com

LinkedIn

Thank you !