ClimateWins

Can machine learning be used to predict if weather conditions will be pleasant on a certain day?

- A comparison study of supervised classification algorithms -

Image generated by AI, all rights reserved

Matthieu DA COL

November 2024

Data analyst for ClimateWins

Data: Source & Bias

First data set:

- The data set is owned and collected by the European Climate Assessment & Data Set Project.
- It holds daily weather metrics (temperatures, wind speed, humidity, precipitation, etc.) for a selection of 18 weather stations in Europe.
- The span is a sample from January 1960, to October 2022.

Second data set:

- The data set is corresponding to the same time period and for 15 weather stations, with tags indicating whether the weather is labeled as pleasant or not.
- It was generated by ClimateWins for the purpose of training the classification models.

Potential Bias / limitations:

- ⚠ Evolutions of used instrumentation or metrics over time can introduce inconsistencies.
- ⚠ Any general assumption based on the study must be taken lightly since the sample of weather stations isn't representative of Europe as a whole.
- ⚠ The labeling of pleasant days is mostly subjective, so it can differ for different persons or for different cultures.
- ⚠ The algorithms used are for classification only, and not predictions or forecast, so there is a limited scope to the study.

Machine learning preliminary step

⚠ **OVERFITTING* is a big issue if the models learn from all weather stations at the same time for the same day:**

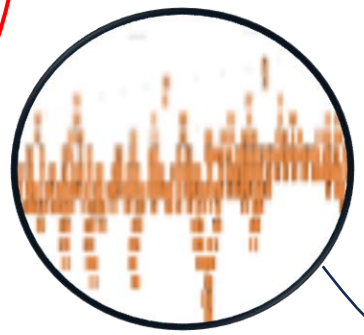
Full comparison :
Overfitted model

VS

Weather in Madrid has an impact on the perceived weather in Stockholm ?

YES

NO



Discriminated comparison :
best fitted models

⬇ **Example of an overfitted <u>decision tree</u> based on all weather stations :**

**The model is too complex and loses accuracy by getting lost in the noise and outliers from the data*

1st model : K-Nearest Neighbor (KNN)

Accuracy score for each weather station:

Station	Accuracy score
Basel	0,9322
Belgrade	0,9178
Budapest	0,937
Debilt	0,9348
Dusseldorf	0,9332
Heathrow	0,9336
Kassel	0,9495
Ljubljana	0,9129
Maastricht	0,9387
Madrid	0,937
Munchenb	0,9393
Oslo	0,9442
Sonnblick	1
Stockholm	0,959
Valentia	0,976

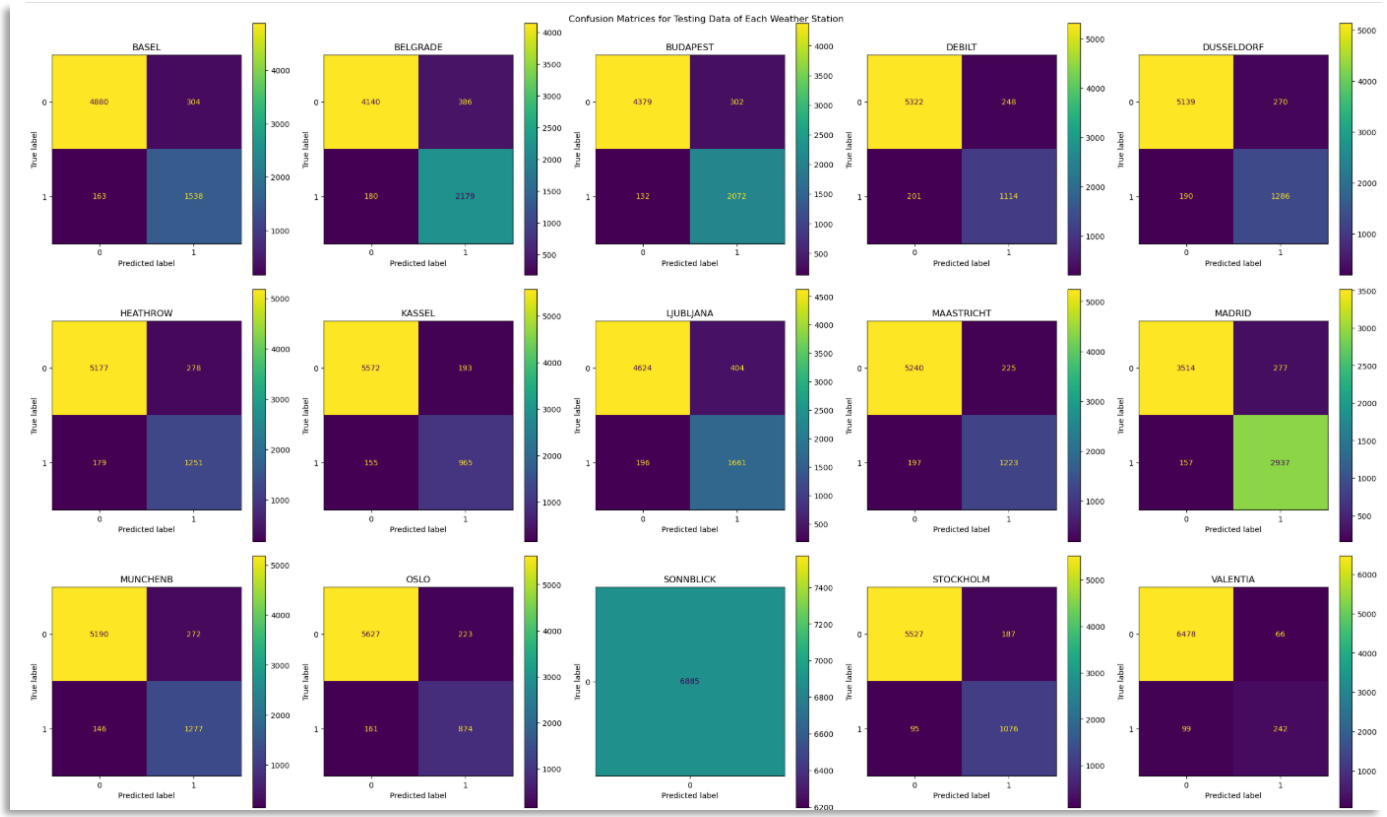
Global avg. accuracy:

94,3%

KNN algorithm classifies data points by comparing them to the closest neighbors and assigning the most common class among those neighbors.

⚠ *Sonnblick has only labels of unpleasant days !*

Confusion matrix (testing data):



The confusion matrix nuances the accuracy, showing that the tags for well performing weather stations like Valentia are not very well balanced (ratio of tags), having more false negatives.

2nd model : Decision Tree

Accuracy score for each weather station:

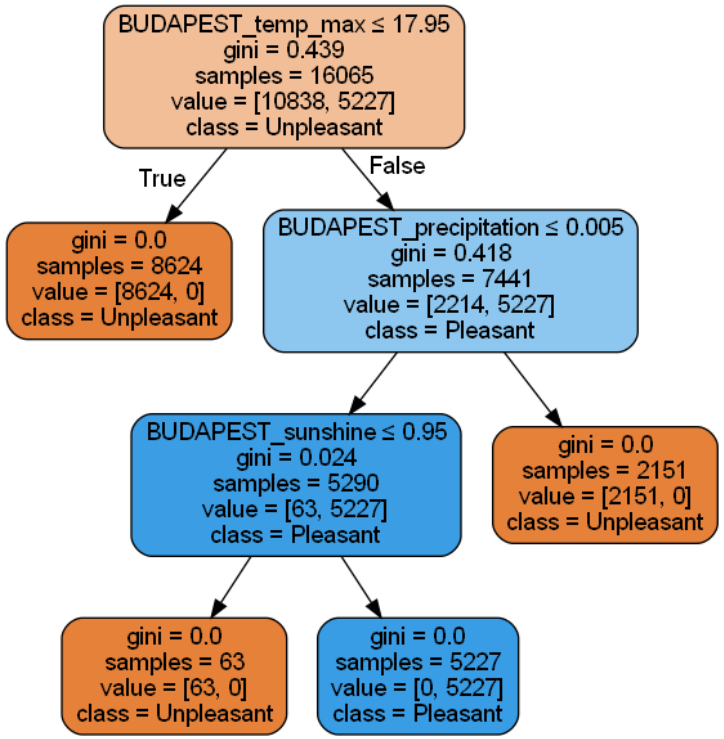
Station	Accuracy score
Basel	1,00
Belgrade	1,00
Budapest	1,00
Debilt	1,00
Dusseldorf	1,00
Heathrow	1,00
Kassel	1,00
Ljubljana	1,00
Maastricht	1,00
Madrid	1,00
Munchenb	1,00
Oslo	1,00
Sonnblick	1,00
Stockholm	1,00
Valentia	1,00

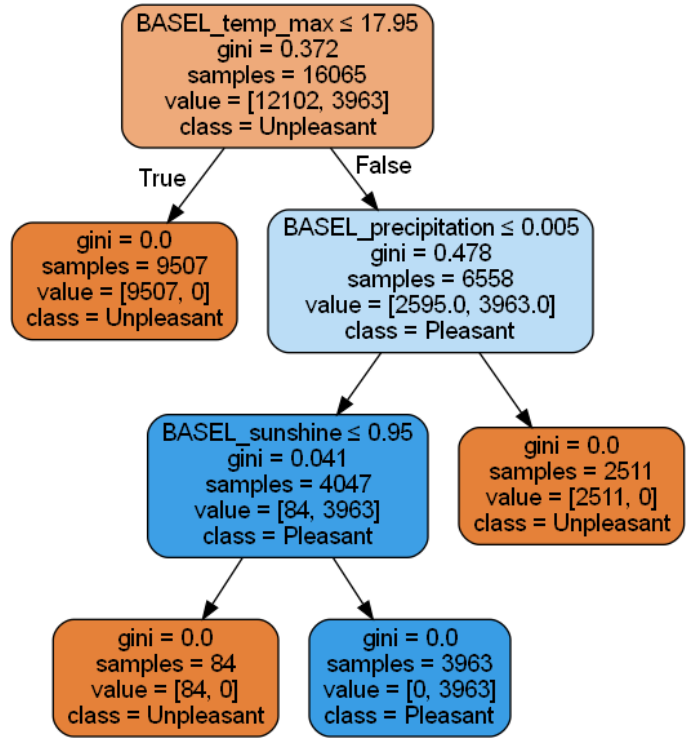
Global avg. accuracy:

100%

Decision tree algorithms split data into branches based on feature values, creating a tree structure to make predictions.

Selection of decision trees:





The decision tree algorithms reached a convergence in **3 steps**, with a starting **threshold temperature of 17,95 °C**.

Is temperature lower or equal to 17,95 °C? → Are precipitations lower or equal to 5 ml ? → Is sunshine lower or equal to 0,95 ?

3rd model : Artificial Neural Network (ANN)

Accuracy score for each weather station best scenario (3rd iteration):

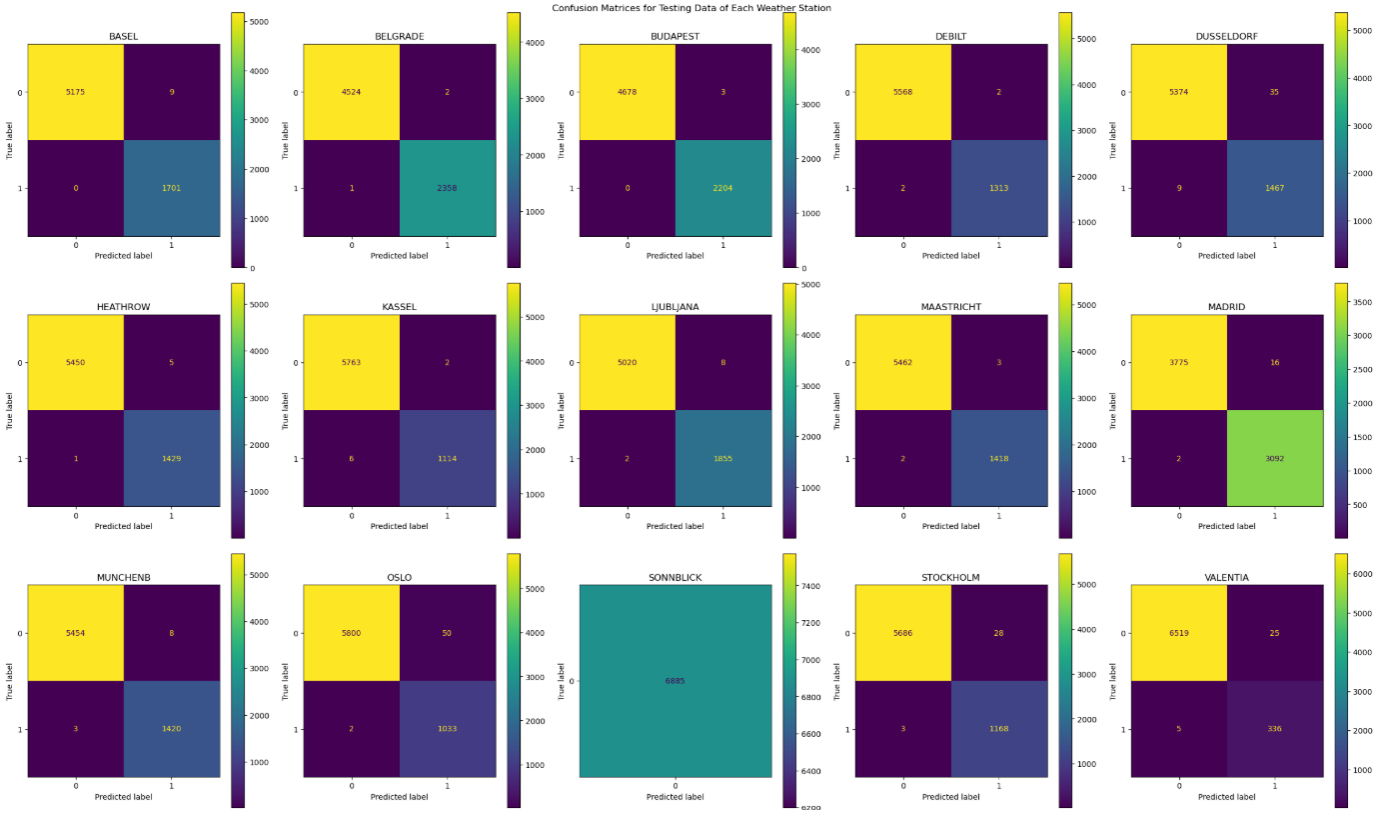
Station	Accuracy score
Basel	0,9987
Belgrade	0,9996
Budapest	0,9996
Debilt	0,9994
Dusseldorf	0,9936
Heathrow	0,9991
Kassel	0,9988
Ljubljana	0,9985
Maastricht	0,9993
Madrid	0,9974
Munchenb	0,9984
Oslo	0,9924
Sonnblick	1
Stockholm	0,9955
Valentia	0,9956

Global avg. accuracy:

99,77%

Artificial Neural Networks (ANN) mimic the human brain by using interconnected nodes (neurons) to process data and make predictions.

Confusion matrix (testing data):



The confusion matrix shows that the model best performs on the stations were there are balanced pleasant and unpleasant days, like Belgrade and Madrid.

Further considerations

Ethical issue:

- ⚠ Climate is evolving, but characteristics of what a pleasant day looks like is not moving as fast: if ClimateWins were to make wrong predictions on where to live in the next years, it might have strong impacts on society, like real estate speculation or increasing migrations.

Derived hypothesis from this study to be verified:

- Will machine learning models be capable to predict climate changes over the next years?
- Are decision trees also capable to adapt to more stations over the world?
- If climate changes, people will tend to move before adapting to their new weather conditions of living?

Recommendations:

- For classifying pleasant days based on weather parameters, decision trees are the most accurate algorithms. However, if more data were to be crossed, the ANN models should not be disregarded.
- Applying these trained models to larger data sets (larger time period more weather stations) could be beneficial in order to test their consistency overtime.
- Enhancing the ANN algorithm with more iterations over new parameters, or helping it by adding an optimization like a gradient descent, could help create a very performing model.
- Another study could involve a combination of these algorithms with time series analyses and forecasting in order to foresee the next most pleasant places to live in Europe.

ClimateWins

Thank you for your kind attention !

For any question, please don’t hesitate to contact me:

All scripts are available on my dedicated GitHub repository: