Machine Learning Specialisation

ClimateWins

Weather Conditions and Climate Change

Prepared by Ana Lazarevska, October 2024

Project Objective

To utilise machine learning techniques to predict the potential consequences of climate change, specifically focusing on extreme weather events in Europe.

Hypotheses

- By analysing historical weather data, a machine learning model can accurately predict future weather patterns in Europe, including extreme weather events.
- A supervised learning model, trained on historical weather data, will outperform an unsupervised learning model in predicting future weather patterns, as it can directly learn from past patterns and make more accurate forecasts.
- Machine learning model can predict if a weather will be favourable on a certain day.

Data Info

- The data used in this project is sourced from the European Climate Assessment & Data Set project.
- This dataset contains weather observations from 18 different weather stations across Europe, spanning from the late 1800s to 2022.
- The records are done almost daily with values such as temperature, wind speed, snow, global radiation, etc.
- Data set link

Data Accuracy

- The quality of the data, including the precision of measurements and the completeness of the records, can affect the accuracy of the predictions.
- The complexity of the machine learning model used to analyse the data can impact the accuracy of the predictions. More complex models can potentially capture more nuanced patterns but may also be more prone to overfitting.
- A larger dataset can improve the accuracy of the predictions, as it provides more information for the model to learn from.
- Because machine learning models rely on historical data to make predictions, feeding them incorrect data will cause the algorithm to make incorrect decisions about the future weather events in certain areas. The machine output will always depend on the input, so incorrect output will be result of incorrect input.

Potential Biases

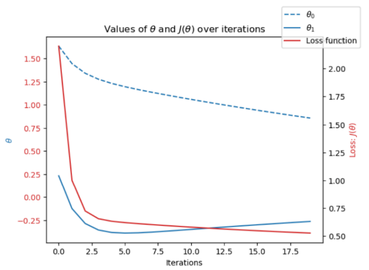
- **Stationary Bias:** The dataset assumes that the historical climate patterns will continue into the future. However, climate change can introduce non-stationary patterns, which might limit the accuracy of the predictions.
- **Selection Bias:** The choice of weather stations might not be entirely random, potentially introducing bias in the data.
- **Spatial Bias:** The data is collected from specific weather stations. It might not accurately represent the climate conditions in regions with fewer stations or unique microclimates.
- **Measurement Bias:** Historical weather instruments might have different calibration standards or accuracies compared to modern ones, leading to potential inaccuracies in the data.

Optimisation

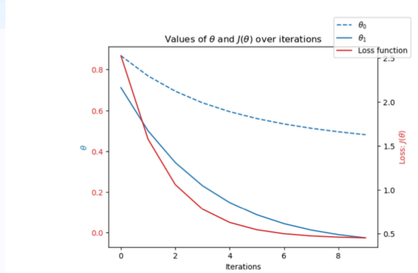
- The gradient descent optimisation was used for the project. Gradient descent is an optimisation algorithm that iteratively adjusts model parameters to minimise the error between predicted and actual outputs.
- In all three chosen weather stations, the mean minimum temperature drastically increased in the last 30 years, while it was more stable in the previous 60 years. Ljubljana experienced biggest increase from the three.
- There is an increase in the mean maximum temperature for all three stations as well. However, this is slight increase and almost steady for the 90 years period.
- From the three weather stations, in the 1960 Ljubljana had the lowest mean minimum temperature, but in 2020 it has almost the same mean minimum temperature as the other two stations.

Weather Station	Year	Mean min temperature	Mean max temperature	Starting theta0	Ending theta0	Starting theta1	Ending theta1	Iterations	Step size
Budapest	2020	-1.58146269926474	1.93579485587908	2	0.85632062	1	-0.26313477	20	0.1
Budapest	1990	-2.35784731009917	1.87697783990677	2	0.54838777	2	-0.21166859	30	0.08
Budapest	1960	-2.41666432607147	1.80639742074001	2	0.72948031	1	-0.28305741	20	0.09
Belgrade	2020	-1.6297300562574	1.94525117152894	1	0.48096044	1	-0.02568489	10	0.05
Belgrade	1990	-2.53766179664758	1.74096652994115	-1	-0.40045976	-1	0.20975368	10	0.1
Belgrade	1960	-2.40147203558906	1.84310885073505	1	0.28880914	1	-0.07187827	10	0.1
Ljubljana	2020	-1.62511313319335	1.93493371135419	0	-0.27340999	1	0.213963	10	0.09
Ljubljana	1990	-2.26689028544172	1.80173467975547	0	-0.27666575	1	0.13672154	30	0.08
Ljubljana	1960	-2.71492339172831	1.66853564815676	-5	-2.25354809	-1	0.98248255	30	0.1

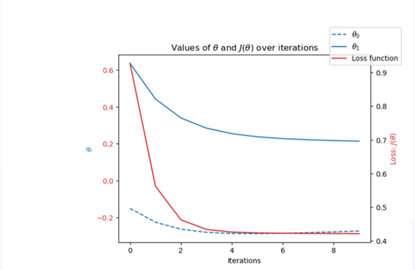
Budapest 2020



Belgrade 2020



Ljubljana 2020



Supervised Learning

Due to the historical nature of the weather data and the goal of the project being to predict future weather patterns, supervised learning was the most suitable approach.

In supervised learning, the algorithm is trained using a dataset where each data point has both input features (like temperature, humidity, and wind speed) and a corresponding output (whether it rained or not).

Algorithm Types

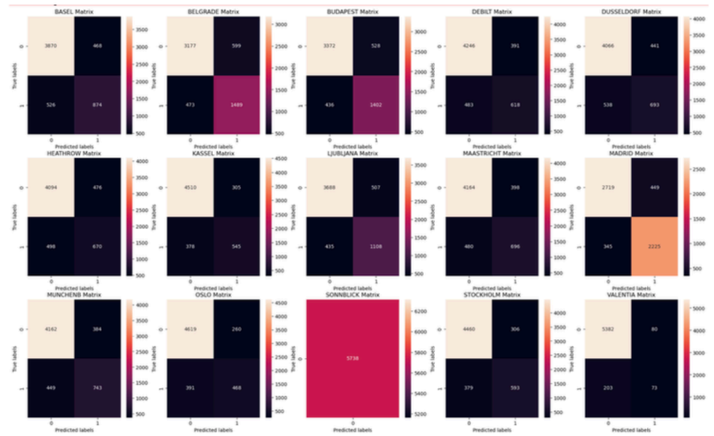
Three algorithm types were used for the project:

- **K-Nearest Neighbours (KNN):** A simple algorithm that classifies data points based on the majority class of their nearest neighbours.
- **Decision Tree:** A tree-like model of decisions and their possible consequences, used to classify or predict outcomes.
- **Artificial Neural Networks (ANN):** Inspired by the human brain, ANNs are complex models with multiple layers of interconnected nodes that learn patterns from data.

KNN

- Given 4 neighbours, most locations achieve accuracy rates between 80% and 90%, indicating generally good performance of the model.
- Sonnblick has a perfect 100% accuracy, but this might be due to limited data, as there are no predictions for "pleasant" days.
- Basel, Belgrade, Budapest, Dusseldorf, Heathrow, Ljubljana, Maastricht, Munchenb, and Stockholm have relatively high numbers of correct predictions for "pleasant" days.
- Debilt, Kassel, Oslo, Sonnblick, and Valentia have particularly high numbers of correct predictions.
- Madrid has a relatively lower number of correct predictions for "unpleasant" days compared to other locations.
- Belgrade, Dusseldorf, Heathrow, Ljubljana, Maastricht, Munchenb have relatively higher numbers of false positives.
- Basel, Belgrade, Budapest, Dusseldorf, Heathrow, Ljubljana, Maastricht, Munchenb have relatively higher numbers of false negatives.

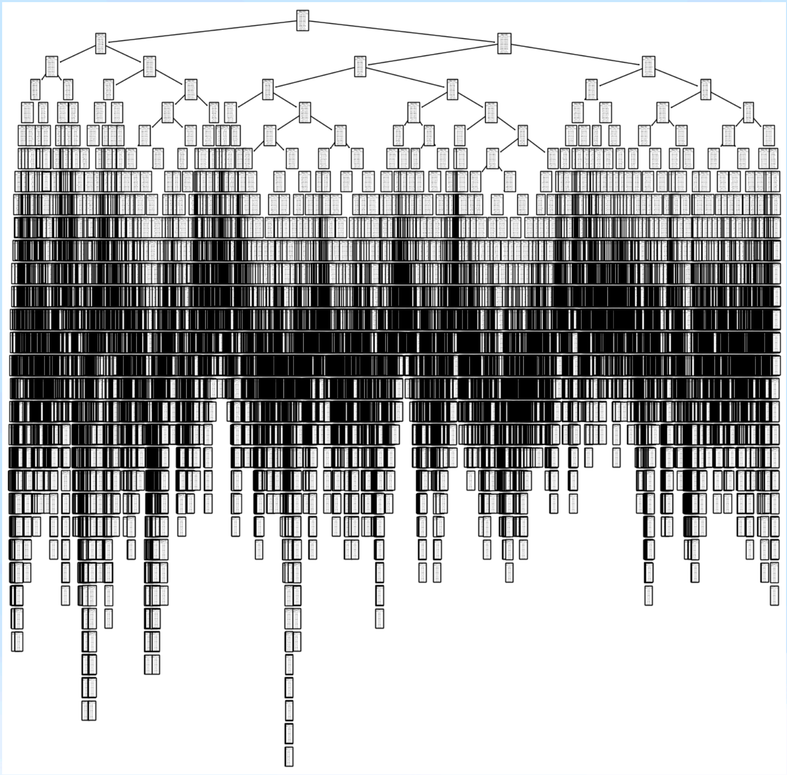
Confusion matrix for the final testing scenario for the scaled dataset

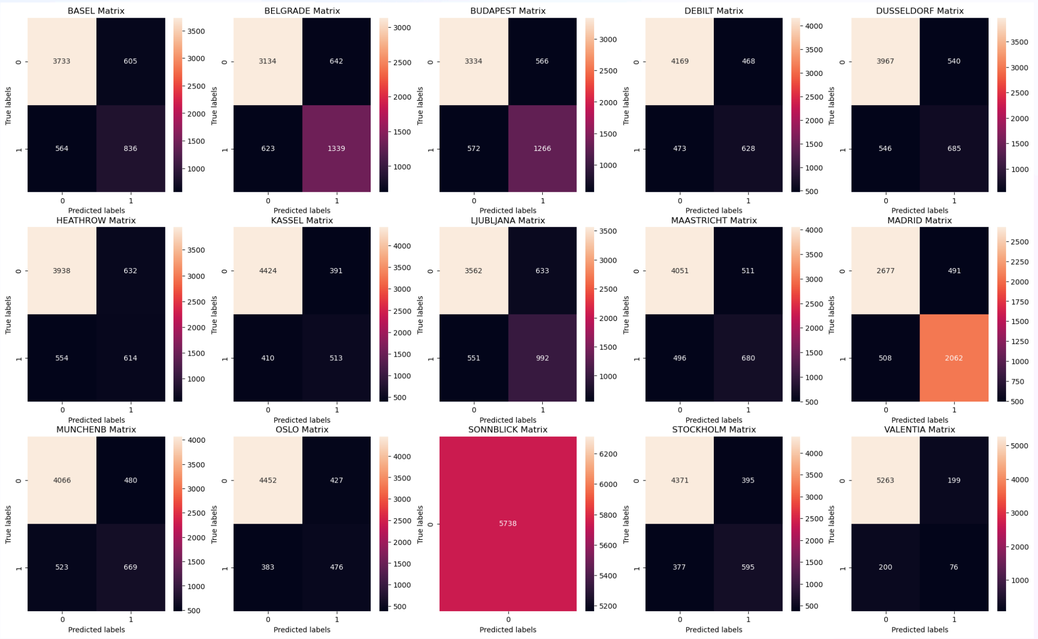


Weather Station	Accurate predictions for 1 (pleasant days)	Accurate predictions for 0 (unpleasant days)	False positives (incorrectly predicted as pleasant)	False negatives (incorrectly predicted as unpleasant)	Accuracy rate
Basel	874	3870	468	526	83%
Belgrade	1489	3177	599	473	81%
Budapest	1402	3372	528	436	83%
Debilt	618	4246	391	483	85%
Dusseldorf	693	4066	441	538	83%
Heathrow	670	4094	476	498	83%
Kassel	545	4510	305	378	88%
Ljubljana	1108	3688	507	435	84%
Maastricht	696	4164	398	480	85%
Madrid	2225	2719	449	345	86%
Munchenb	743	4162	384	449	85%
Oslo	468	4619	260	391	89%
Sonnblick	0	5738	0	0	100%
Stockholm	539	4460	306	379	88%
Valentia	73	5382	80	203	95%

Decision Tree

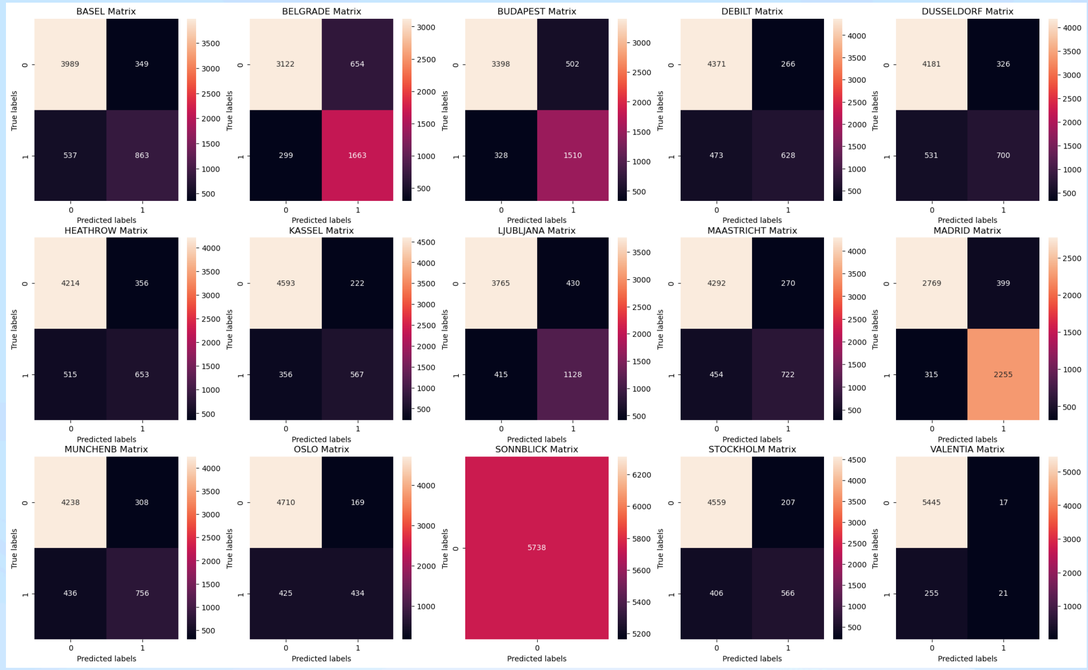
- The tree is very deep and complex, with multiple levels and many branches, and this suggests that it might be overfitting the training data (the model learns the training data too well but struggles to generalise to new, unseen data).
- The training accuracy score of 39.8% suggests that the model is able to correctly predict the target variable for only about 40% of the training samples. This is relatively low and indicates that the model is not fitting the training data well.
- This testing accuracy score of 40.4% is slightly higher than the training accuracy, but still quite low, meaning that the model's ability to predict accurately on data it hasn't seen before is limited.
- The results indicate that the decision tree model has limited predictive power for both training and testing data.

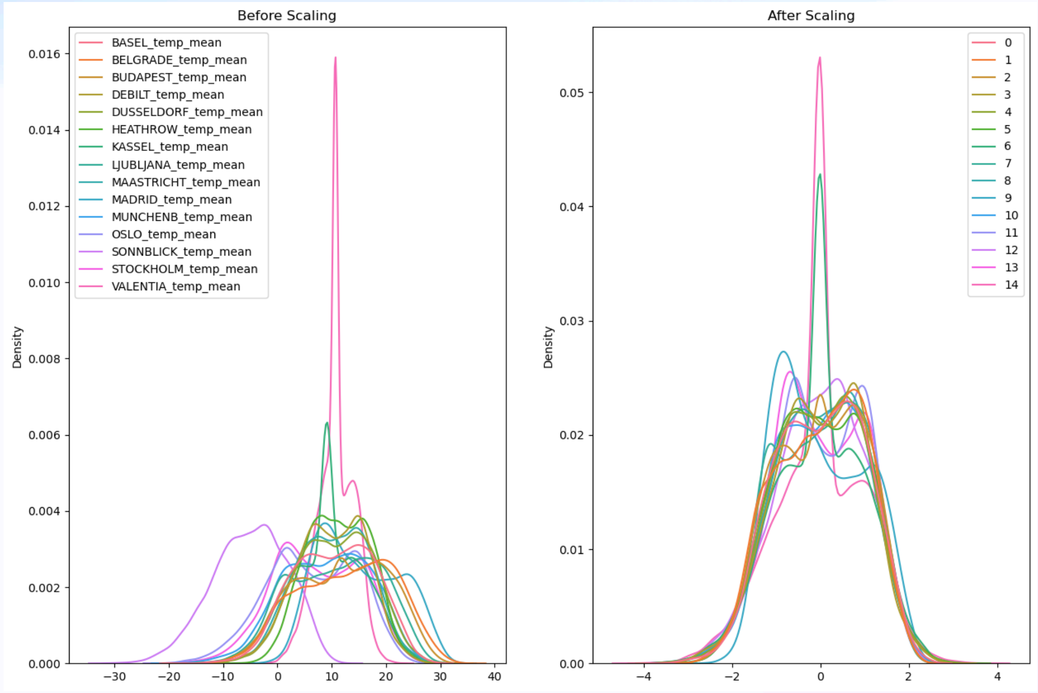




ANN

- Three scenarios were chosen to test the accuracy of the ANN model.
- The best scenario had 4 hidden layers with 50, 20, 50, and 50 nodes respectively, 500 iterations and tolerance of 0.00000001. These choices slightly improved the accuracy of the model to 46.6% for the training data and 45.4% for the test data.
- Scaling the data made slight difference. In the unscaled data the network will give more weight to the larger values, like Valentia. In the scaled model however, everything has more even weigh.





Conclusions

- The machine learning models can accurately predict future weather patterns, thus confirming the initial hypothesis.
- The decision tree model had limited predictive power for both training and testing data.
- I recommend that ClimateWins use the KNN model which has the highest test accuracy.

Next Steps

- Pruning of the decision tree to improve its accuracy.
- Combine supervised and supervised learning methods. While supervised learning is typically the go-to method for weather prediction, unsupervised learning can uncover hidden patterns or anomalies in weather data, and by identifying outliers in weather data, it can help detect extreme weather events or climate change trends that might not be captured by supervised models.
- Deploy the best-performing model into a production environment, such as a web application or API.

Questions?

Contact: ana.lazarevska@dataanalyst.com

Thank you for your attention!

Links:

GitHub: https://github.com/AnaLazarevska/climatewins_machine_learning