Machine Learning for Weather Prediction with ClimateWins

October 2024

Poorya Behnamie

Project Overview

ClimateWins is a European nonprofit focused on using machine learning to help predict the consequences of climate change around Europe and, potentially, the world.

The questions this project aims to answer:

How is machine learning used? Is it applicable to weather data?	Are there any ethical concerns specific to this project?	Can machine learning be used to predict whether weather conditions will be favorable on a certain day?

Objective and Hypotheses

Objective:

The goal is to predict daily weather conditions that are pleasant for outdoor activities across various European weather stations using machine learning.

Key Question:

Can machine learning reliably predict whether the weather will be pleasant based on historical weather data?

Hypotheses:

1. Machine learning can predict daily weather patterns in Europe with high accuracy.
2. Different machine learning algorithms have varying levels of success in predicting pleasant weather.
3. There are certain weather stations or data features that contribute more to the overall accuracy of predictions.

We hypothesize that different algorithms will have varying success in modeling "pleasant" weather, and this presentation aims to explore which model is most effective.

Data Sources and Biases

Data Source:

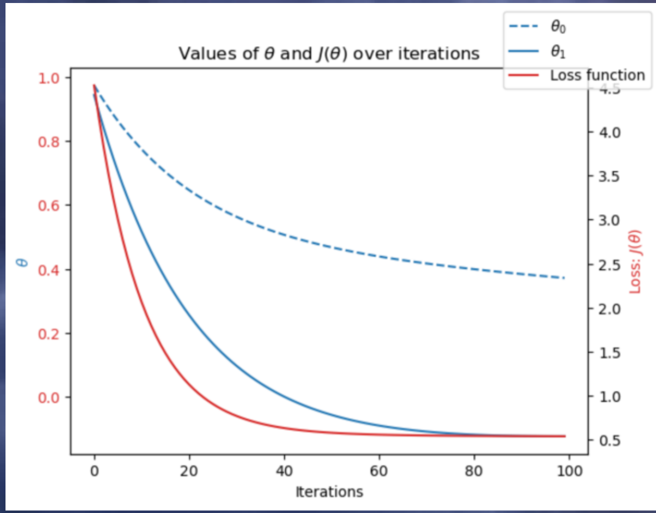
- Weather data from 18 European weather stations ranging from 1800s to 2022.
- Features include average temperature, humidity, wind speed, and more, with labels for "pleasant" and "unpleasant" days.

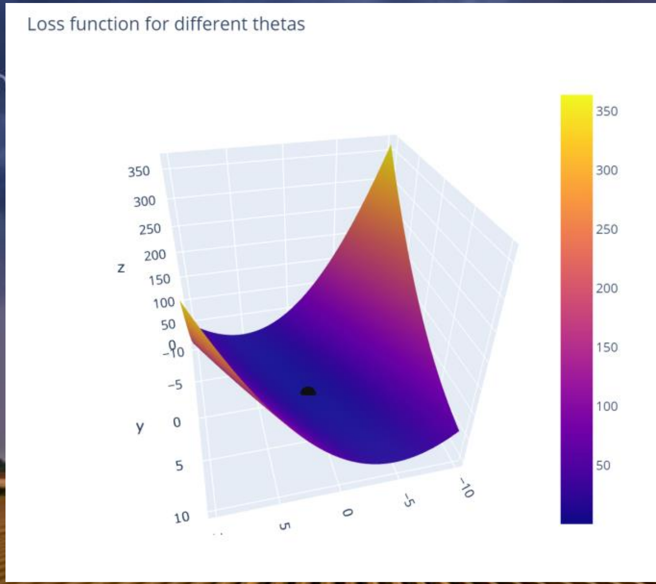
Biases:

- Collection Bias: Data is collected only from European weather stations.
- Location Bias: Weather data may not generalize to regions outside of Europe.
- Temporal Bias: Older data may not accurately represent current weather patterns.

Accuracy:

- The accuracy of machine learning models depends on the algorithm and optimization technique applied.





Supervised Learning Algorithms & Optimization

Algorithms Used:

- **K-Nearest Neighbors (KNN):** Achieved 88.45% accuracy.
- **Decision Tree:** Accuracy of 63.33%, but overfitting is a concern.
- **Artificial Neural Network (ANN):** Train Accuracy of 78%, Test Accuracy of 68%.

Optimization:

Used gradient descent to minimize error in training models and find the best learning rates and parameters.

K-Nearest Neighbors (KNN) Model

Objective:

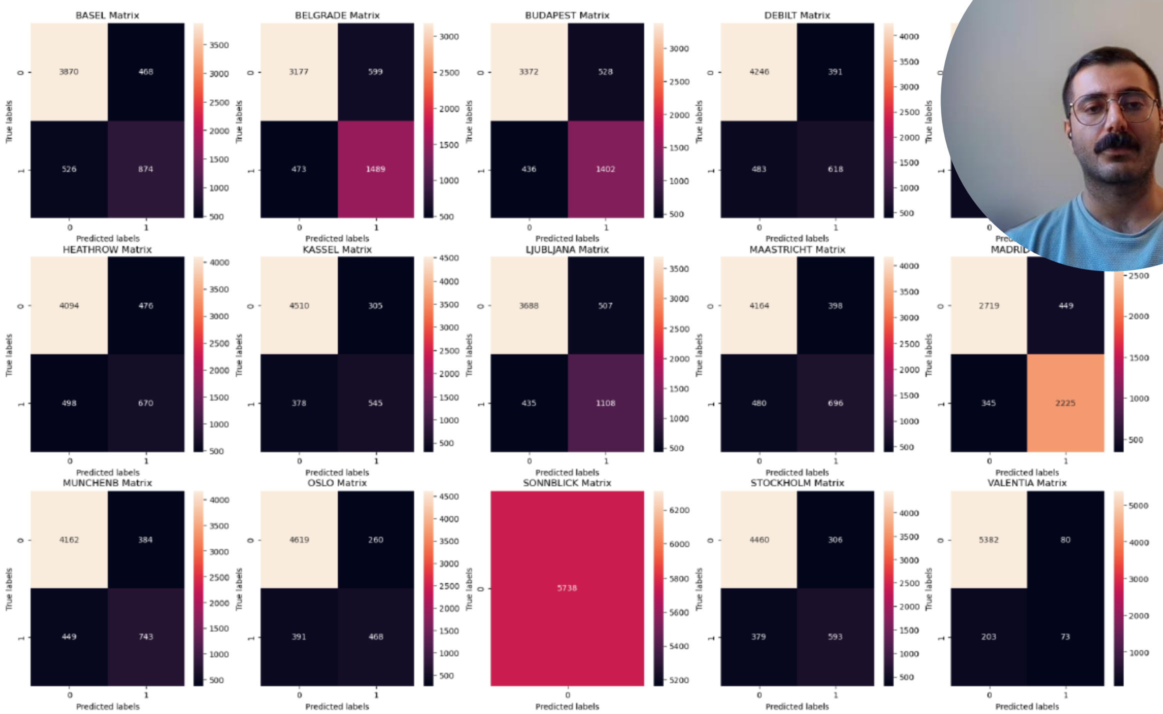
- Predict pleasant vs. unpleasant weather using KNN.

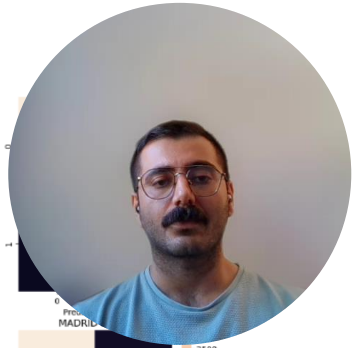
Model Performance:

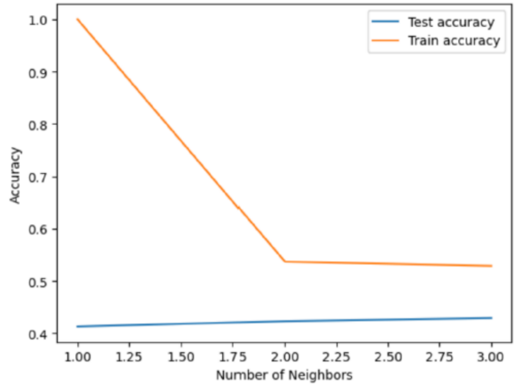
- Train Accuracy: 100%
- Test Accuracy: 87%

Conclusion: KNN showed signs of overfitting, with a significant performance gap between training and test data.

Bias: KNN relies heavily on feature scaling, and the bias variance trade-off is evident.





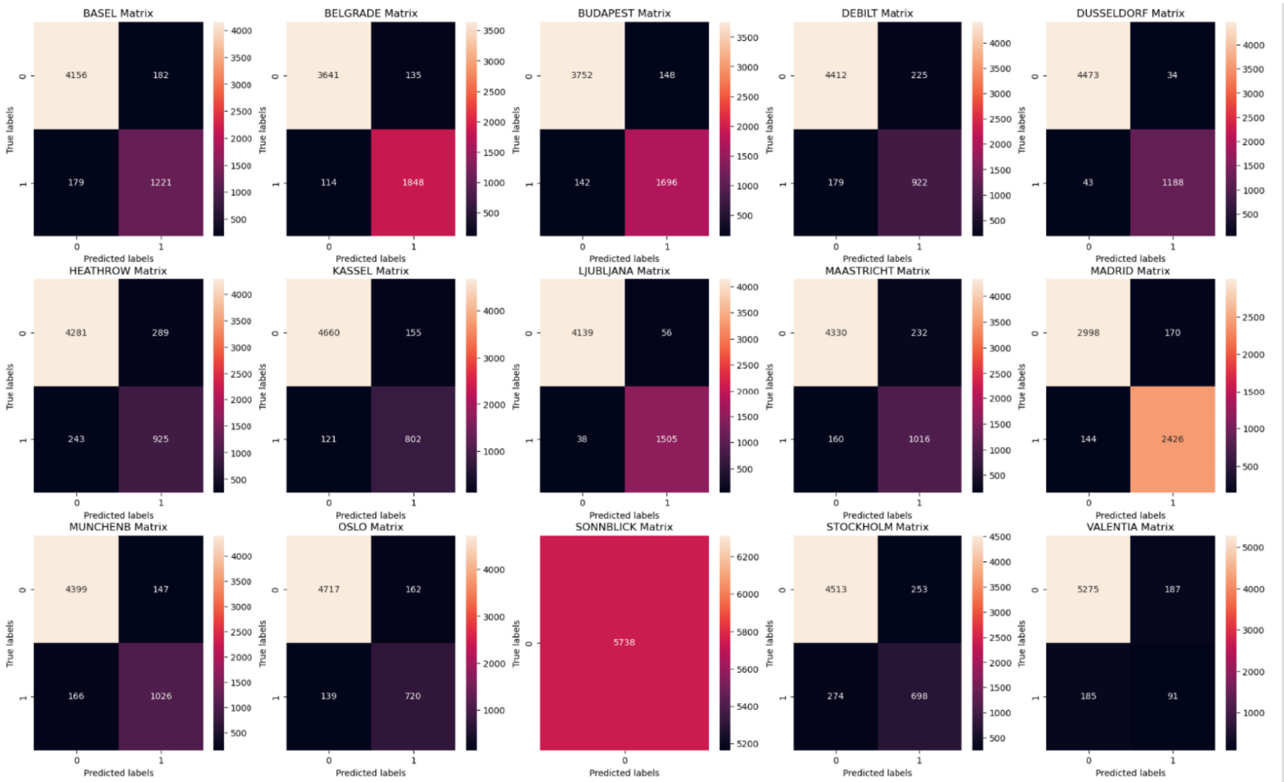



```
Accuracy for BASEL: 0.83
Accuracy for BELGRADE: 0.81
Accuracy for BUDAPEST: 0.83
Accuracy for DEBILT: 0.85
Accuracy for DUSSELDORF: 0.83
Accuracy for HEATHROW: 0.83
Accuracy for KASSEL: 0.88
Accuracy for LJUBLJANA: 0.84
Accuracy for MAASTRICHT: 0.85
Accuracy for MADRID: 0.86
Accuracy for MUNCHENB: 0.85
Accuracy for OSLO: 0.89
Accuracy for SONNBLICK: 1.00
Accuracy for STOCKHOLM: 0.88
Accuracy for VALENTIA: 0.95

Overall Accuracy: 0.87
```

Decision Tree Model

```
Decision Tree Train Accuracy: 1.0
Decision Tree Test Accuracy: 0.6333217148832345
```



Objective:

- Use decision trees to classify weather days based on pleasantness.

Model Performance:

- Train Accuracy: 100%
- Test Accuracy: 63.33%
- **Overfitting**: Clear evidence of overfitting in the decision tree model, with the need for pruning to avoid overly rigid classifications

Optimization: Tree depth tuning didn't improve test accuracy significantly

Bias: The model overfits due to deep trees capturing noise in training data.

Artificial Neural Networks (ANN)

Objective:

- Model nonlinear data relationships using ANN.

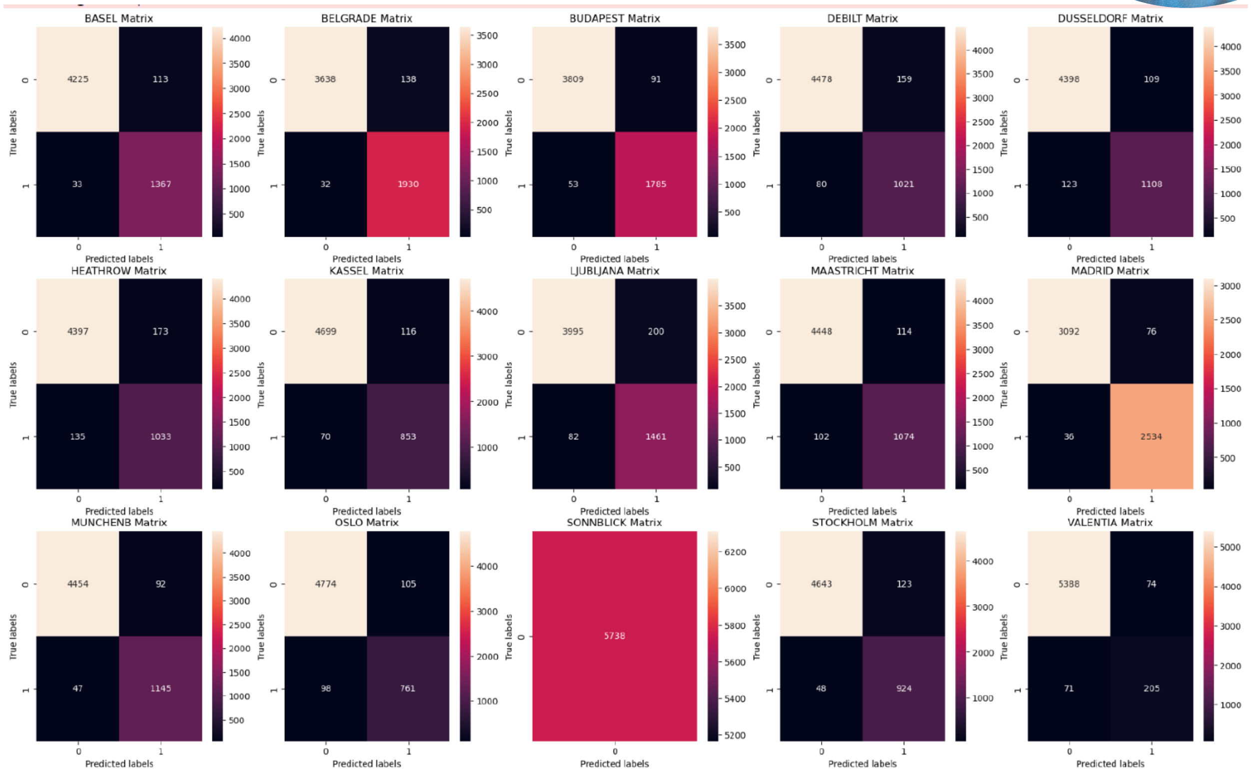
Model Performance:

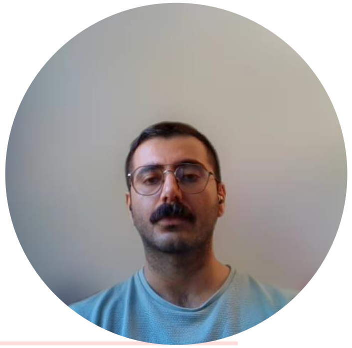
- Train Accuracy: 78.2%
- Test Accuracy: 68.4%

Conclusion: ANN provided more balanced results, showing potential for better generalization but still struggled with the complexity of the data.

Limitations: Despite tuning layer sizes, regularization, and iterations, accuracy improvements plateaued.

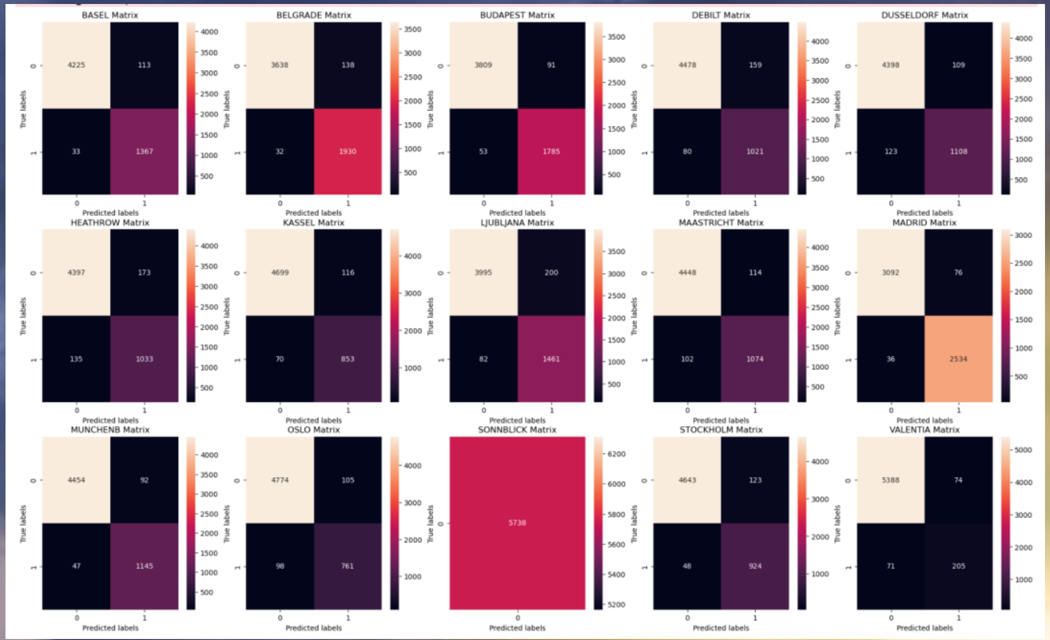
```
ANN Train Accuracy: 0.7819544503834535
ANN Test Accuracy: 0.6840362495643081
```





Confusion Matrix Analysis

KNN Confusion Matrix:	High accuracy, misclassification rates are low.
Decision Tree:	Overfitting led to more misclassifications; training accuracy was perfect, but test accuracy was lower.
ANN:	Mixed results depending on layers and iterations. The test accuracy improved with tuning, but overfitting was reduced compared to the decision tree.



Evaluation of Models

KNN:

Best-performing model, providing 88% accuracy in predicting pleasant weather. Highest training accuracy but overfits heavily.

Decision Tree:

Overfitting is evident; pruning is necessary for better results.

ANN:

Can improve with more layers and iterations but remains less accurate than KNN. Best generalization among the models, but still underperformed compared to expectations.

Scaling:

Data scaling had a minimal effect on improving accuracy.

Recommendations for ClimateWins

Recommendation:

- Based on overall performance, **ANN** with further tuning could offer the best balance between training and testing accuracy.

Next Steps:

Feature Engineering: Add additional weather features like wind speed, radiation, and precipitation.

Data Collection: Improve underperforming stations like Roma and Gdansk by filling data gaps.

Model Refinement: Consider using ensemble methods (e.g., Random Forest or XGBoost) to reduce overfitting.

Ethical Oversight: Regularly audit model outputs to ensure fairness and accountability in predictive outcomes.

Conclusion: While ANN shows promise, model generalization and interpretability must be improved for practical use by ClimateWins.

Questions & Conclusion

Machine learning can successfully predict pleasant weather, with **KNN** being the most reliable model. Further refinement of **ANN** and decision trees could lead to improvements in prediction accuracy.

Thank You for Your Time

If you have any questions,
please feel free to contact me.

Pooryabehnamie@gmail.com