Using Machine Learning with Climate Data

Prepared by Sam Abrams
June 2, 2025

Project Goals

1. **Identify the optimal machine learning model** for predicting climate consequences in Europe using historical weather data spanning over a century
2. **Test three machine learning algorithms** - K-Nearest Neighbor, Neural Networks, and Decision Trees - to determine the most effective approach for weather prediction
3. **Apply findings to extreme weather event prediction** across Europe using the European Climate Assessment & Data Set project data (future step)

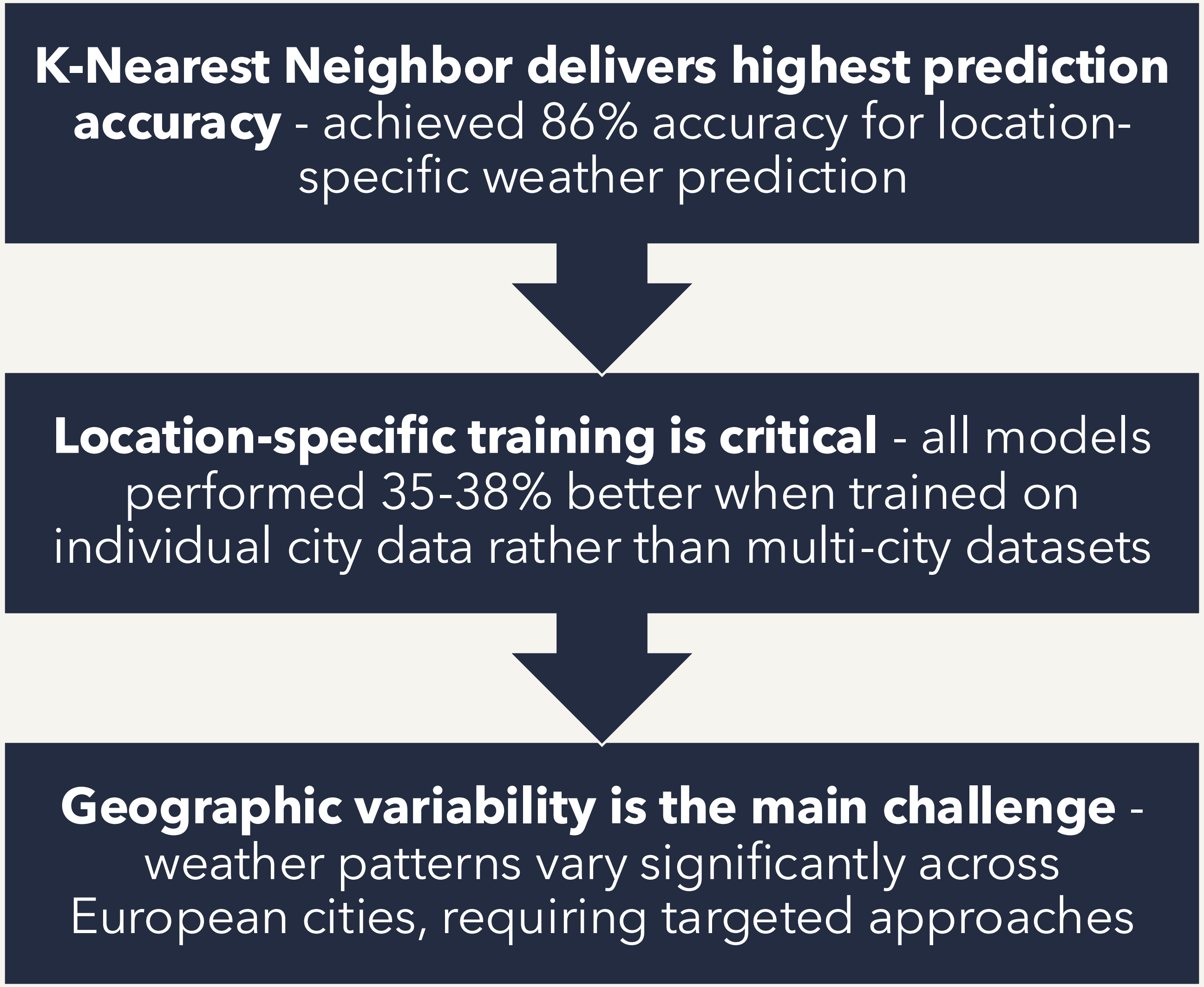
Hypotheses

1. Machine Learning can **accurately predict extreme weather events** in *specific* regions of Europe using historical weather data.

2. There is a **statistically significant increase** in the frequency and intensity of extreme weather events in Europe **over the past 10-20 years** compared to the previous century.

3. **Supervised** learning algorithms, particularly regression models, will be **more effective** at predicting continuous variables like temperature changes compared to classification models.

Key Findings & Recommendations



K-Nearest Neighbor delivers highest prediction accuracy - achieved 86% accuracy for location-specific weather prediction

Location-specific training is critical - all models performed 35-38% better when trained on individual city data rather than multi-city datasets

Geographic variability is the main challenge - weather patterns vary significantly across European cities, requiring targeted approaches

Methodology

Data Sourcing

- The data for this project comes from the **European Climate Assessment & Data Set** project.

- It comprises weather observations from **18 different weather stations** across Europe, spanning from the late 1800s to 2022.

- This includes daily recordings of **temperature**, **wind speed, snow, global radiation**, and other variables.

Potential Data Limitations

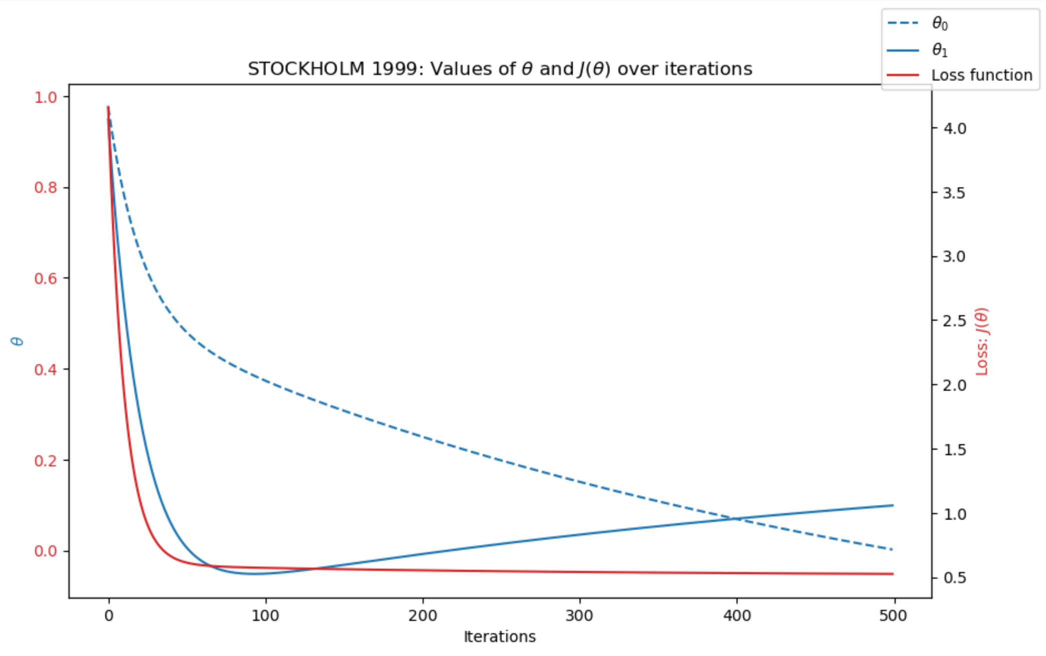
- **Station Placement**: The 18 stations may not be representative of **all** microclimates in Europe, potentially skewing overall trends.
- **Data Collection Methods**: Changes in data collection technology (recording, instruments, etc.) over time might obfuscate long-term trend analyses - early data may be less precise than, or complete as, recent data.
- **Data Completeness**: There might be gaps or missing data for certain periods or stations, which could affect the accuracy of long-term analyses.
- **Geopolitical Issues:** Different institutions collect data differently, have different climate policies, and use that data differently.

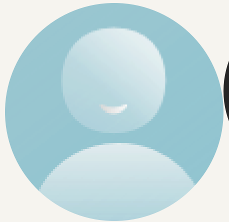
Optimization

- **Gradient descent optimization** was applied to find the best way to describe temperature trends in our weather station data.

- The algorithm iteratively tested different approaches until it discovered the mathematical formula that most accurately captured each city's temperature behavior.

- This process revealed consistent patterns across all locations - showing similar baseline temperatures and seasonal trends, **proving that machine learning can identify universal climate relationships from seemingly different local weather datasets**.





Model Results

Testing Approach

- **Three models tested:**
  1. *K-Nearest Neighbor*
  2. *Decision Trees*
  3. *Artificial Neural Network*

- **Two training scenarios:**
  1. *All cities data*
  2. *Belgrade-only data*

- Evaluation Metric: **Prediction accuracy**

Performance Overview

- **The KNN model gave the most accurate results** across the three models tested.

- When the models were trained on Belgrade’s data only, **they performed between 35-38% better** than the models trained on all cities data.

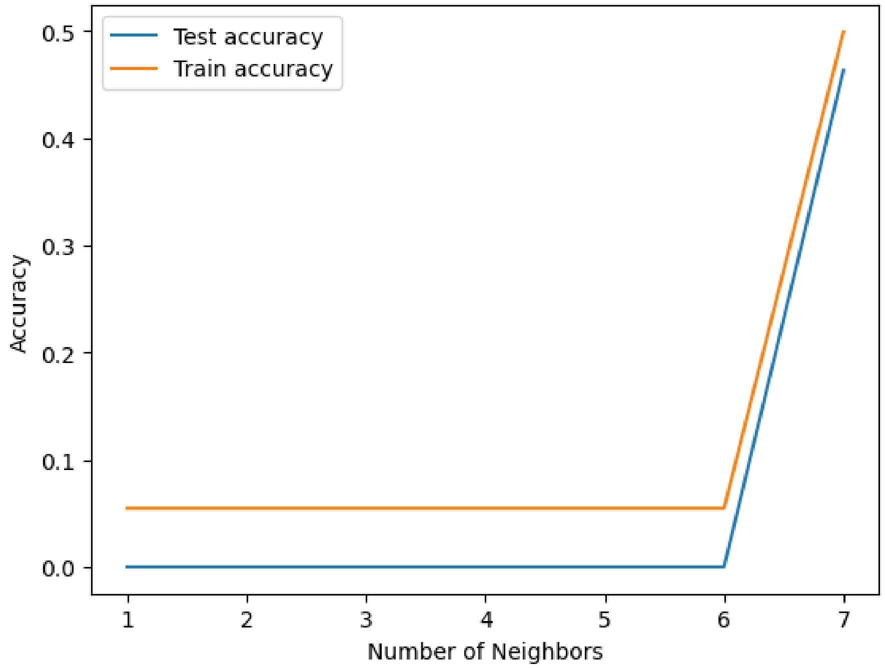
Learning Model	Trained on All Cities	Trained on Belgrade Only
K-Nearest Neighbor	50%	86%
Decision Tree	46%	84%
Artificial Neural Network	46%	81%

K-Nearest Neighbor

- A model that predicts new outcomes by finding the **most similar historical examples** - like predicting tomorrow's temperature by looking at the 5 most similar weather days in your dataset and averaging their temperatures.

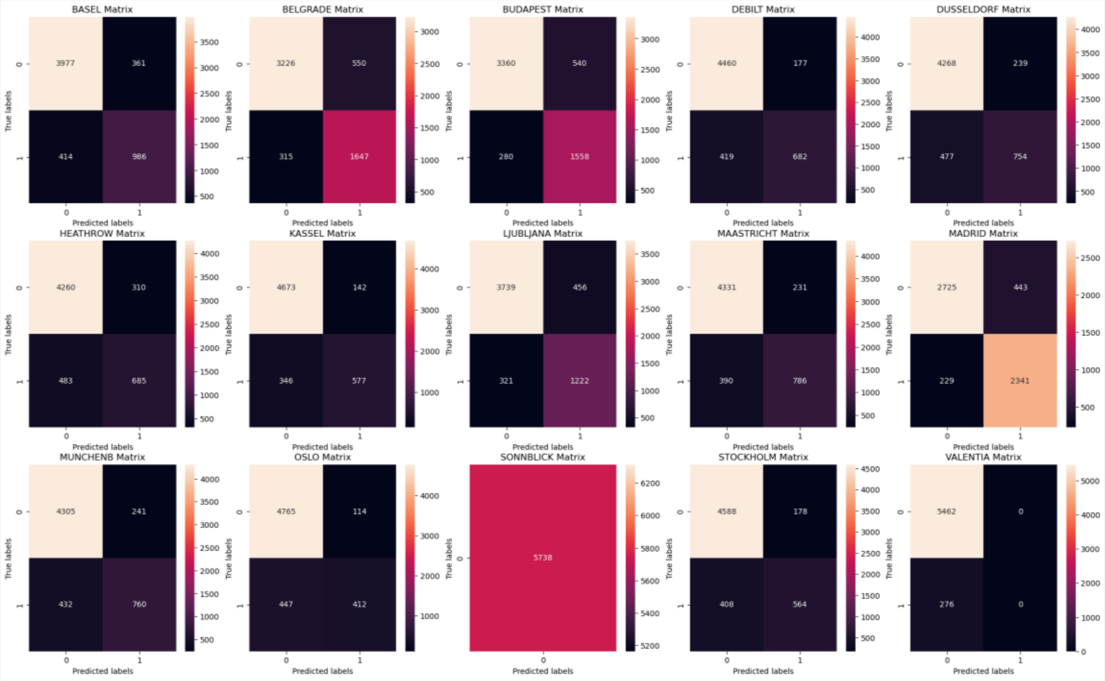
- Using seven “most similar historical examples”, the model was able to predict pleasant weather with about 50% accuracy.

- 50% isn’t great, but let’s take a look at why that number is so low…



K-Nearest Neighbor: Prediction Analysis

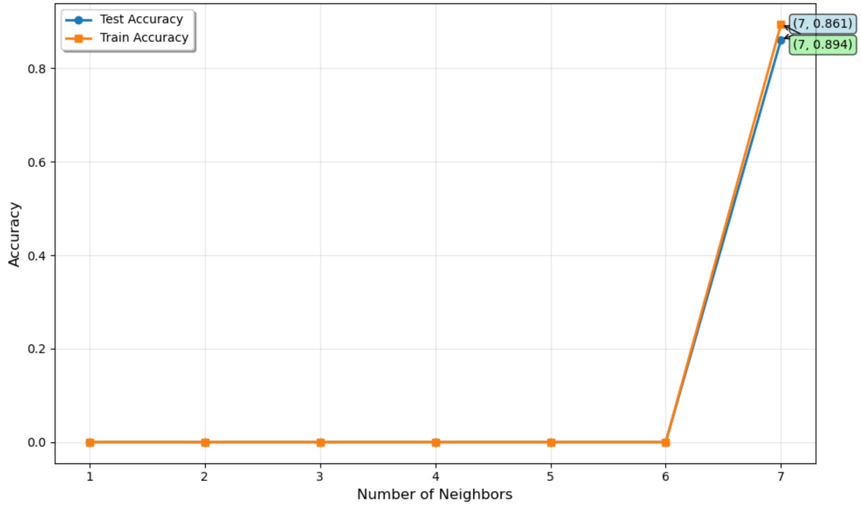
- **Confusion Matrix Overview:** These matrices show prediction accuracy for each city – the key on the right shows what ideal predictions should look like, however, the KNN results don’t look like that. But there’s a reason for that…
- **Geographic and Climate Variability:** The model struggles with accurate predictions because weather patterns vary significantly across different geographic locations and climates, making it difficult to learn consistent patterns that work well for all cities simultaneously.

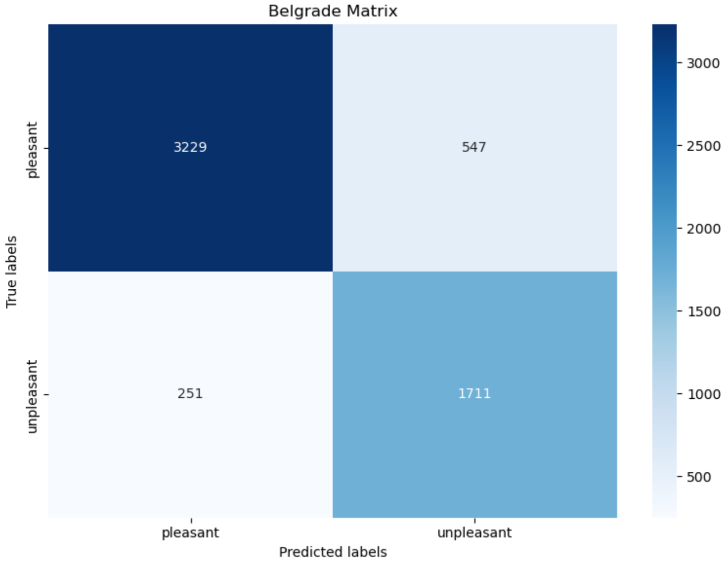


Accurate Unpleasant	Inaccurate Unpleasant
Inaccurate Pleasant	Accurate Pleasant

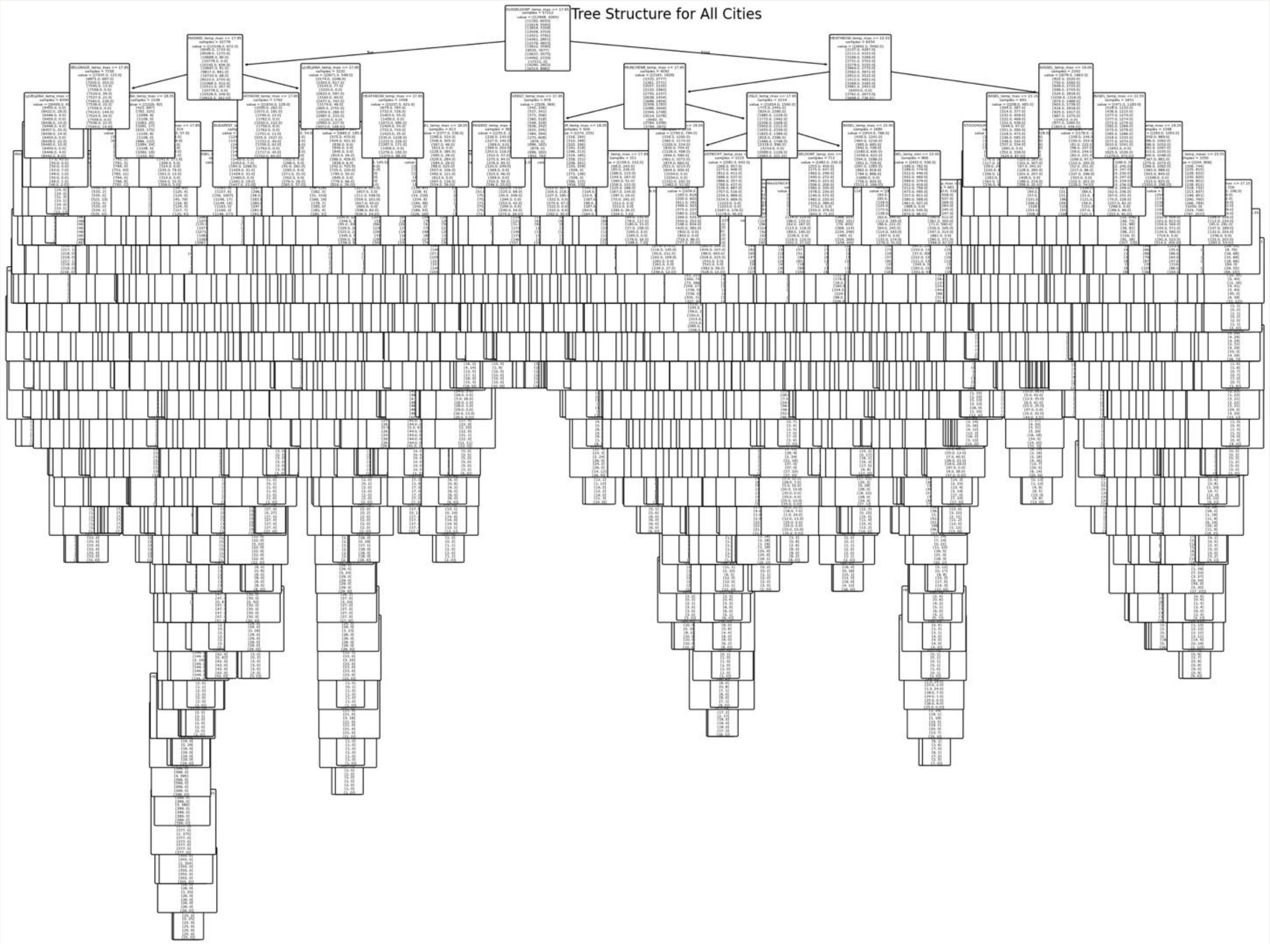
Training KNN Model on a Specific Location

- **Significant Accuracy Improvement:** After retraining the KNN model using only Belgrade's weather data, accuracy increased to **86%** - a substantial improvement over the multi-city approach
- **Location-Specific Training Benefits:** This dramatic improvement demonstrates that weather patterns are highly location-dependent and require targeted training data for optimal predictions
- **Key Takeaway:** Models trained on geographically-focused datasets consistently outperform those trained on diverse, multi-location data due to reduced pattern variability





Decision Tree



- **Decision trees** are models that predicts outcomes by asking simple questions about your data - like "Is temperature above 70°F? If yes, is humidity below 50%?" - until it can classify whether it will rain or not.

- The decision tree created for this data is **extremely overfit**, meaning the predictions are too specific to accurately predict new data.

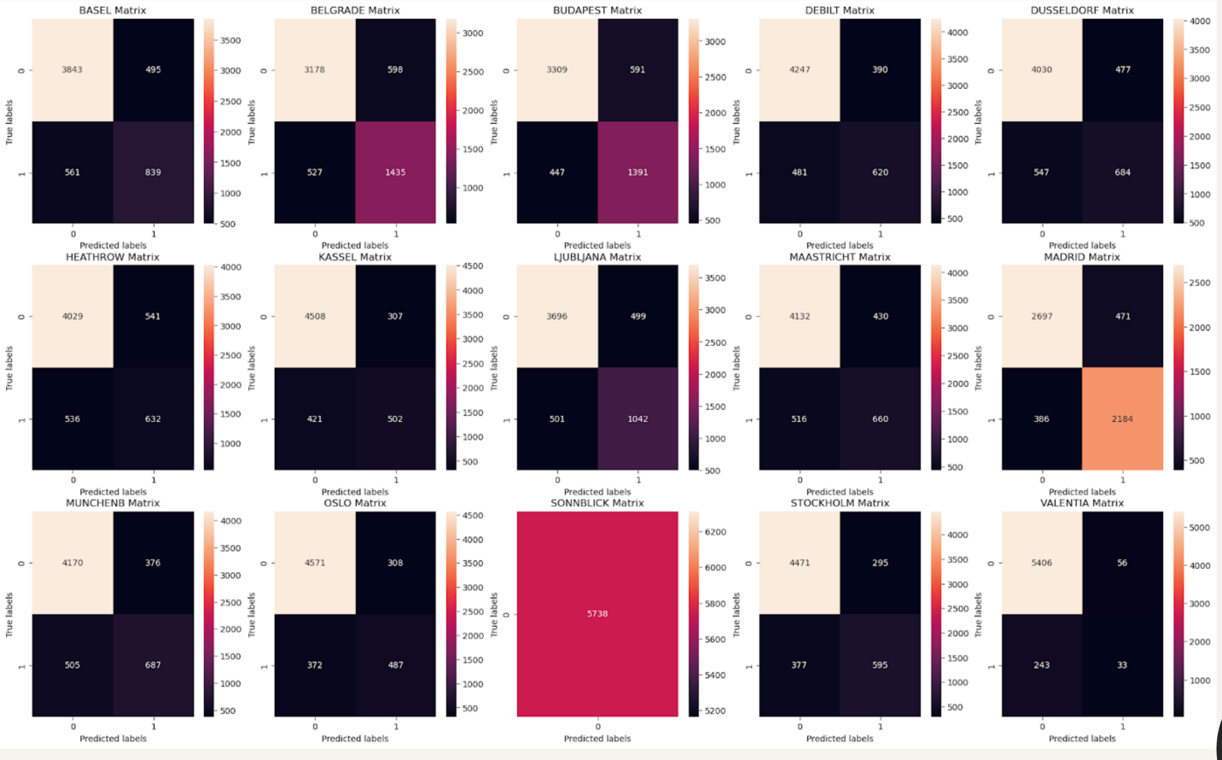
- **Prediction Accuracy**: Overfitting usually leads to a high accuracy score, but this model reported only 46% prediction accuracy. This is due to the variability in the training data.

Decision Tree Confusion Matrix

Accurate Unpleasant	Inaccurate Unpleasant
Inaccurate Pleasant	Accurate Pleasant

- **Variable Training Data Strikes Again**: Similar to the KNN model, the low accuracy is due to the variability in the training data.

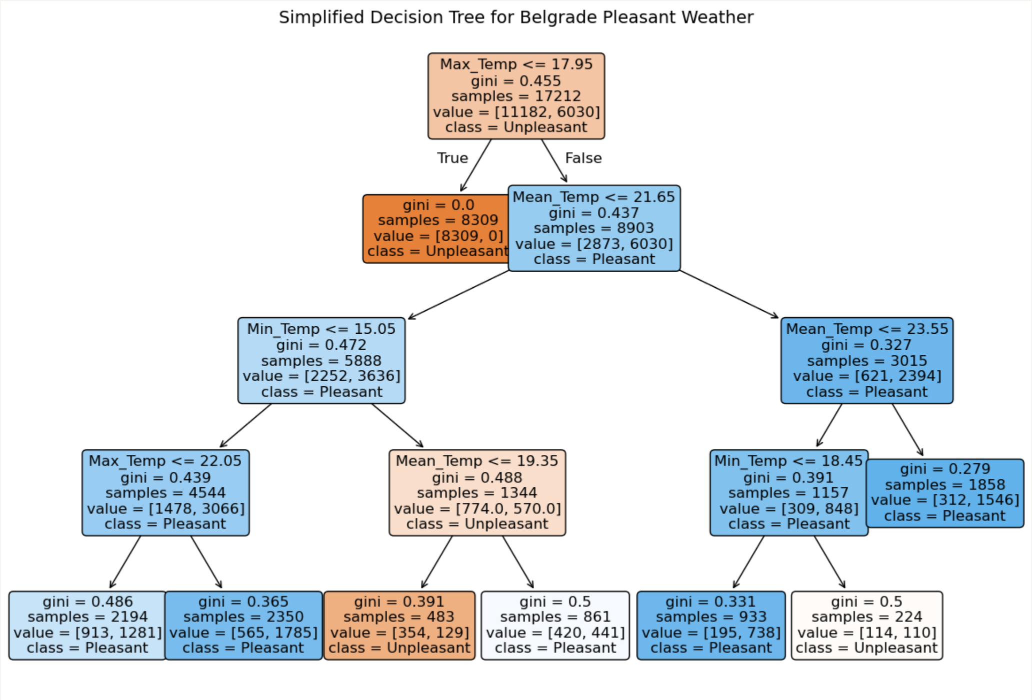
- **To test this theory,** I created another decision tree that was trained only on **Belgrade's** weather data…

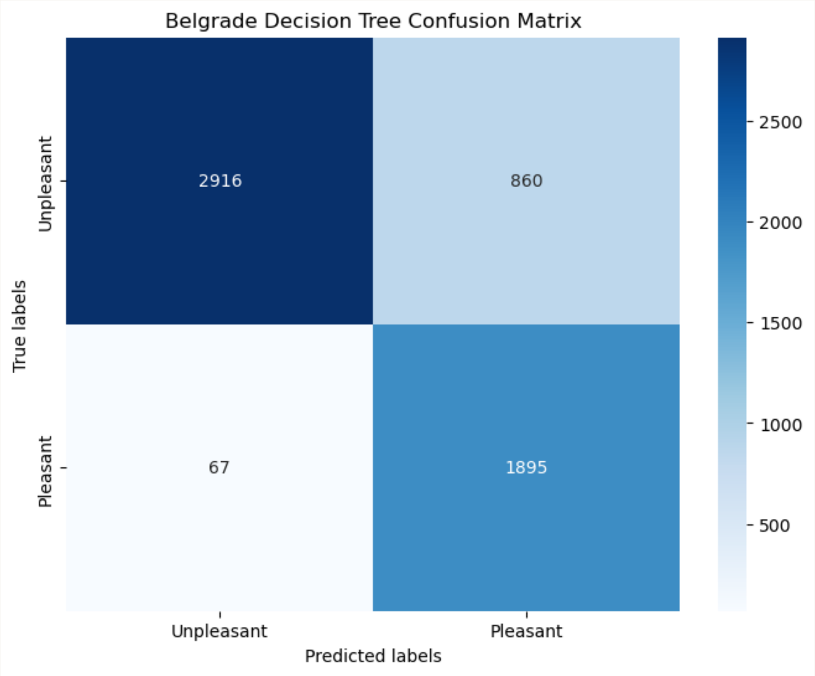


Belgrade Decision Tree

- **Decision trees follow the same pattern as KNN models** - performance improves with location-specific training. The decision tree shown right was created with the same parameters as the previous tree but creates a much more structured set of predictions.

- **Belgrade-focused model achieved 84% accuracy**, again demonstrating the value of targeted data.


Simplified Decision Tree for Belgrade Pleasant Weather

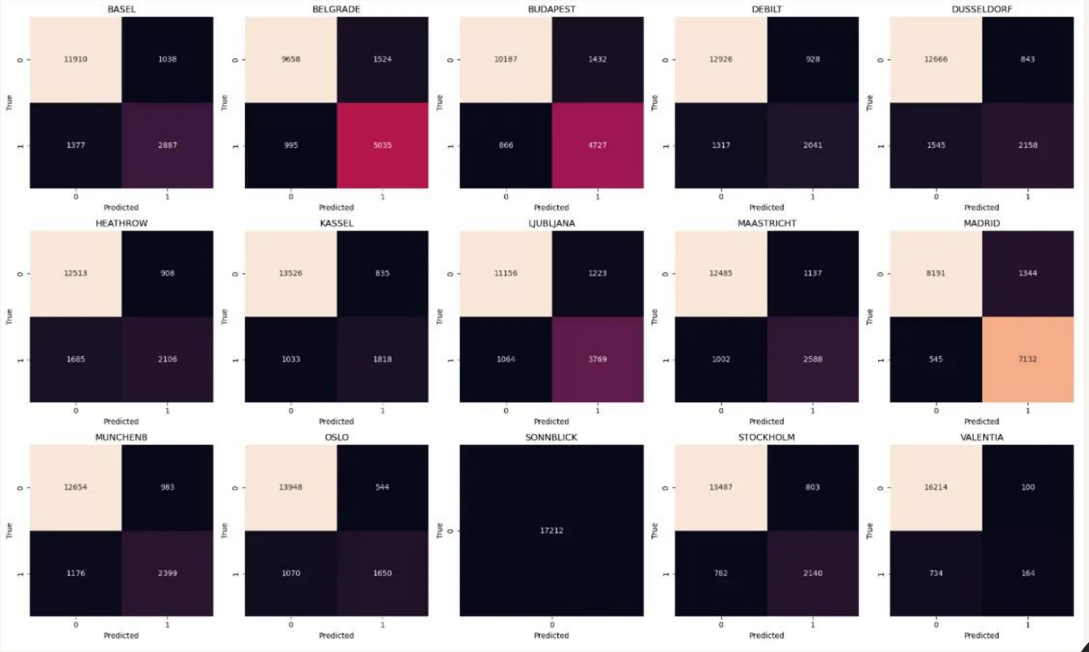

Belgrade Decision Tree Confusion Matrix

Artificial Neural Network

Accurate Unpleasant	Inaccurate Unpleasant
Inaccurate Pleasant	Accurate Pleasant

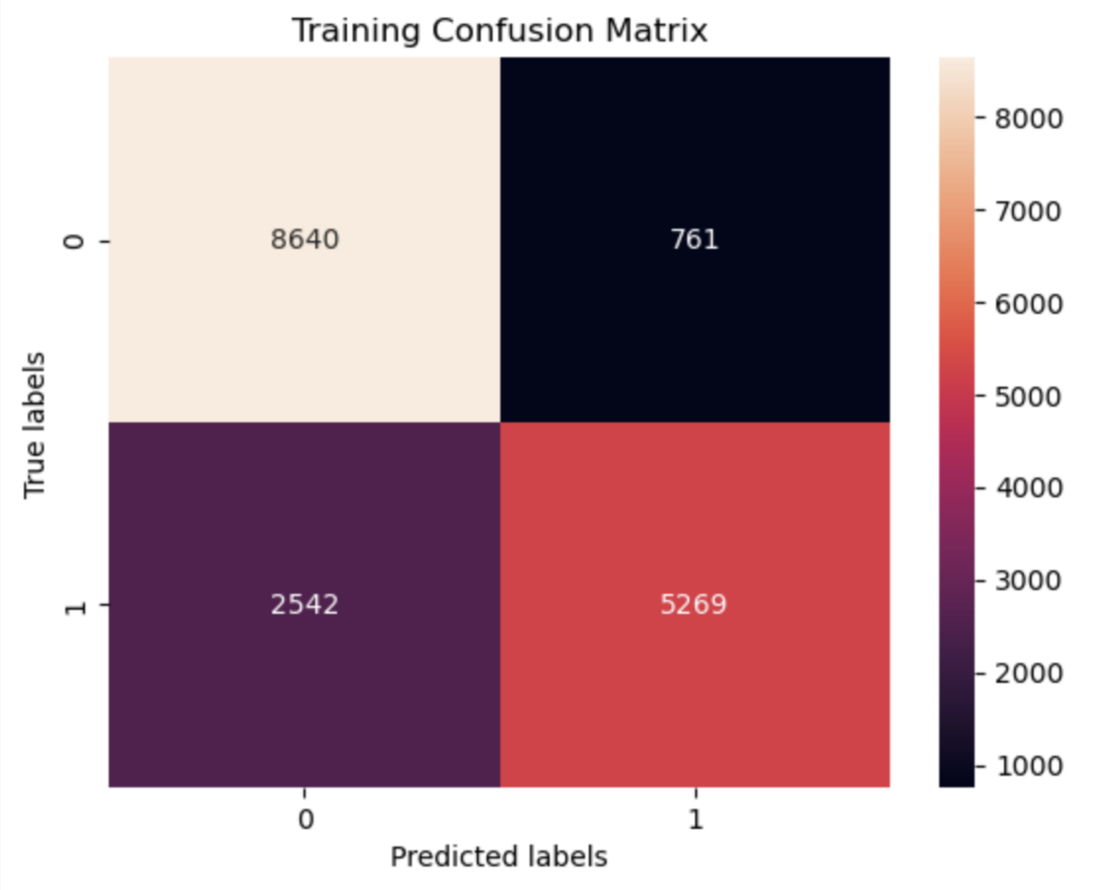
- A model that learns complex relationships in data by processing multiple features simultaneously - like analyzing temperature, humidity, pressure, and wind speed together to predict weather patterns more accurately than simple rules.

- Training the network on all cities gave an accuracy score of 46%, highest among all models



ANN: Belgrade

- The artificial neural network was able to attain **81% accuracy** predicting pleasant weather for Belgrade alone, again confirming the necessity for geographical specificity in the training data.



Recommendations

- **The K-Nearest Neighbor model gave the most accurate results, so that appears to be the best model for making weather predictions**.

- **ONE IMPORTANT CAVEAT**: The **artificial neural network** might be useful the once more variables are considered, **as it tends to handle multivariate training data more efficiently** than the KNN model.

- **Reminder**: these predictions were run using ONLY **average temperatures** to predict "pleasant weather" on any given day. The dataset has more variables we can/will use to train the algorithm, and that's where ANNs typically outshine KNN.

Questions?

Reach out via e-mail: sabrams15@gmail.com

Contact me on my website: www.sam-abrams.com

Thank you for your time!