Femina

12/09/2024

Predicting Climate Change Consequences with Machine Learning

Objective and Hypotheses

- **Objective:** To assess machine learning models for predicting weather and climate patterns across Europe, identifying critical predictors of climate change.
- **Hypotheses:**

1. *If* machine learning models are trained on historical weather data, *then* they will be able to predict future extreme weather events with a reasonable degree of accuracy.
2. *If* specific weather stations in different regions of Europe are analyzed, *then* machine learning models will identify distinct trends in climate change that vary regionally.
3. *If* machine learning models are applied to weather data, *then* they can accurately classify days with pleasant and unpleasant weather conditions.

Data Source and Biases

- Data from European weather stations, provided by the European Climate Assessment & Data Set project.
- **Biases:**

1. Regional Bias: ML models trained primarily on data from specific regions may not represent global climate conditions leads to biased predictions.
2. Historical Data Bias: Climate data often reflects incomplete or biased historical records.
3. Over Predictions: Errors in predictions could lead to harmful outcomes like misallocation of resources, inadequate preparation for extreme weather, or neglect of vulnerable communities.

Optimization and Features

- Data optimization involves various strategies to improve data management, ensuring better **efficiency, reliability, and accessibility**.
- The goal is to **maximize the utility** of existing resources by refining how data is processed and used.

Gradient Descent:

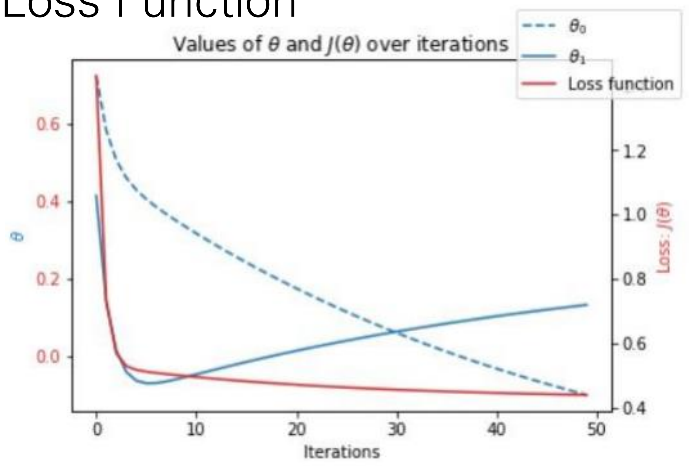
- **Gradient Descent** is a key method for finding the **best-fit parameters** for machine learning models.
- It works by iteratively minimizing the **error or cost function** to get closer to the optimal solution.
- Used to adjust model weights for higher accuracy and better predictions.

Optimization process for Debilt's 2021 weather data using gradient descent.

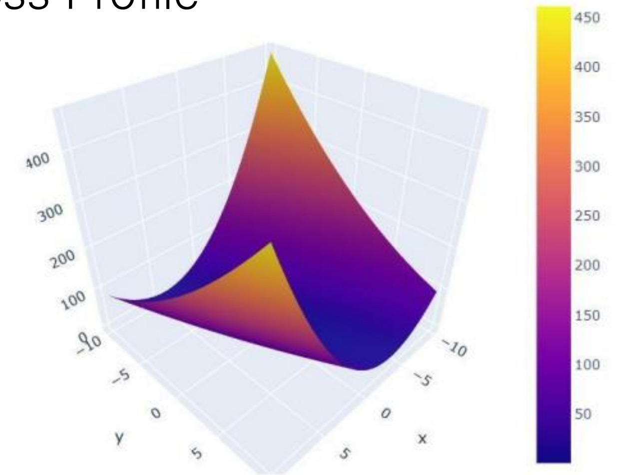
Key Insights:

- The **steep initial drop** in the loss function suggests that the starting values for θ_0 and θ_1 were **far from optimal**.
- This behavior indicates that the **temperature data** for Debilt in 2021 displayed **significant variability or seasonality**.
- The model adjusted rapidly during the **first few iterations** to account for this variability, reflecting a need for further tuning of the learning rate or additional features.

Loss Function



Loss Profile



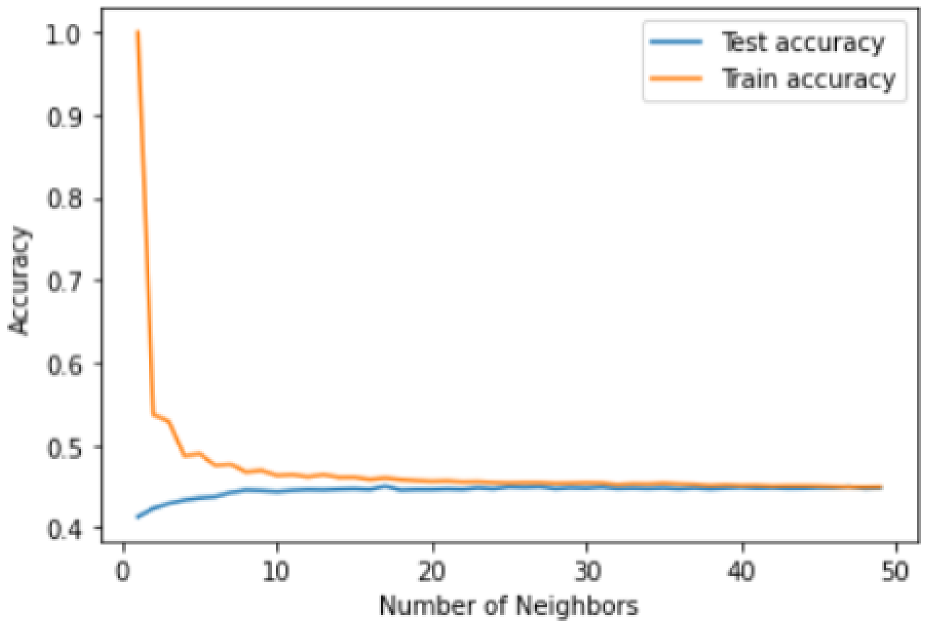
Supervised Learning and Algorithms:K-Nearest Neighbors (KNN) Model Analysis

- What is KNN?
- K-Nearest Neighbors (KNN) is a supervised learning algorithm used for classification and regression.
- It classifies data points by comparing the input to the k nearest neighbors in the feature space and assigning the most common label among those neighbors.
- KNN is simple yet effective for many problems, but its performance depends heavily on the choice of k (the number of neighbors).

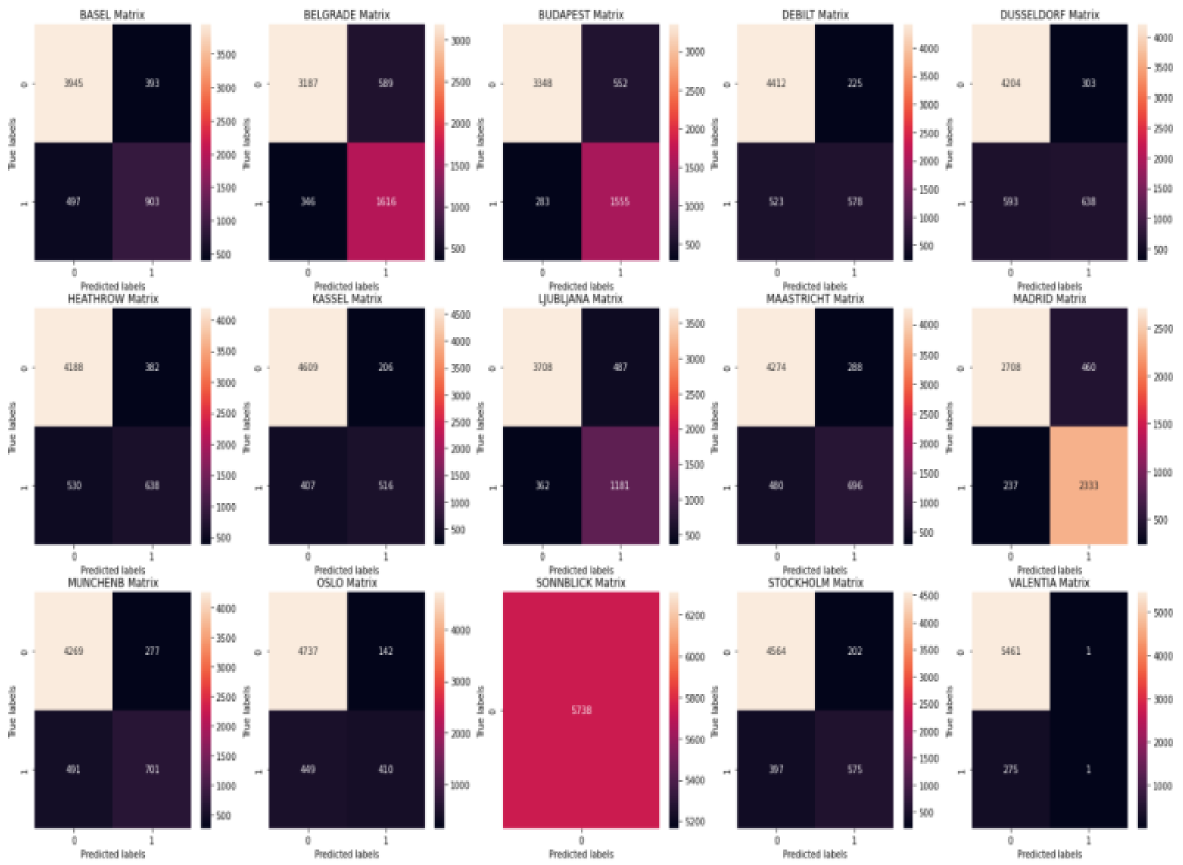
Observations from KNN Results:

- For **small values of k** (e.g., **k=1**), the model has **high train accuracy** (~100%) but **low test accuracy**.
- This indicates **overfitting**—the model memorizes the training data but **fails to generalize** to unseen data.
- When **k increases to 4-5**, both **train and test accuracies stabilize**, suggesting a balance between **overfitting** and **underfitting**.

Accuracy Plot



Performance by Weather Station(Using KNN Nodel):



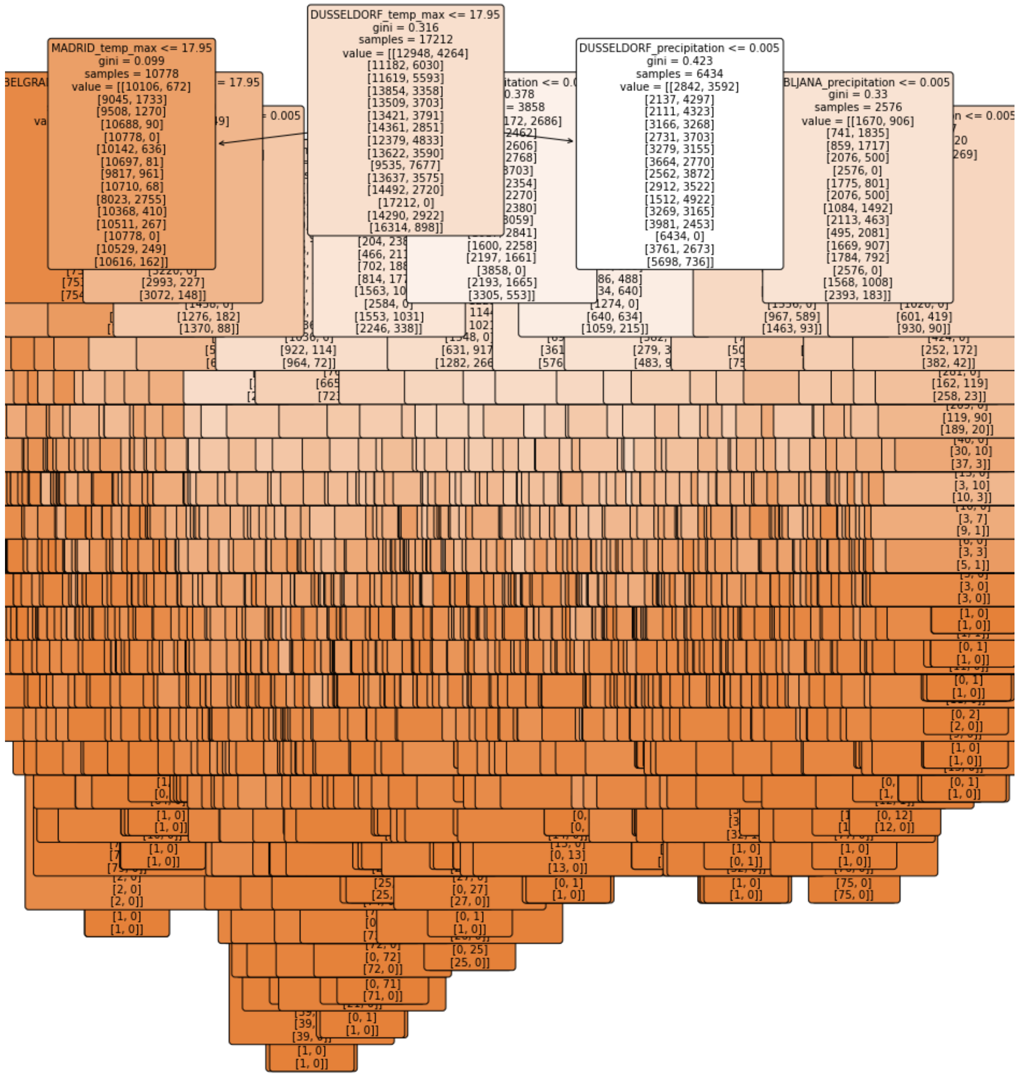
- High Accuracy (Kassel, Oslo, Stockholm, Valentia):
  - These stations show **high prediction accuracy** for both "pleasant" and "unpleasant" weather days.
  - The **diagonal values** in the confusion matrix (true positives and true negatives) indicate effective prediction for both classes.
- Imbalance (Sonnblick):
  - The model **only predicts "unpleasant" days**, with no correct predictions for "pleasant" days.
  - This suggests **data imbalance** or **unique weather conditions** at this station affecting model performance.
- Moderate Misclassifications (Heathrow, Munich, Debilt, Dusseldorf, Ljubljana):
  - These stations show **confusion** between predicting "pleasant" and "unpleasant" days.
  - Moderate misclassifications (off-diagonal values) indicate that the KNN model struggles with these locations.

Decision Tree Model Analysis

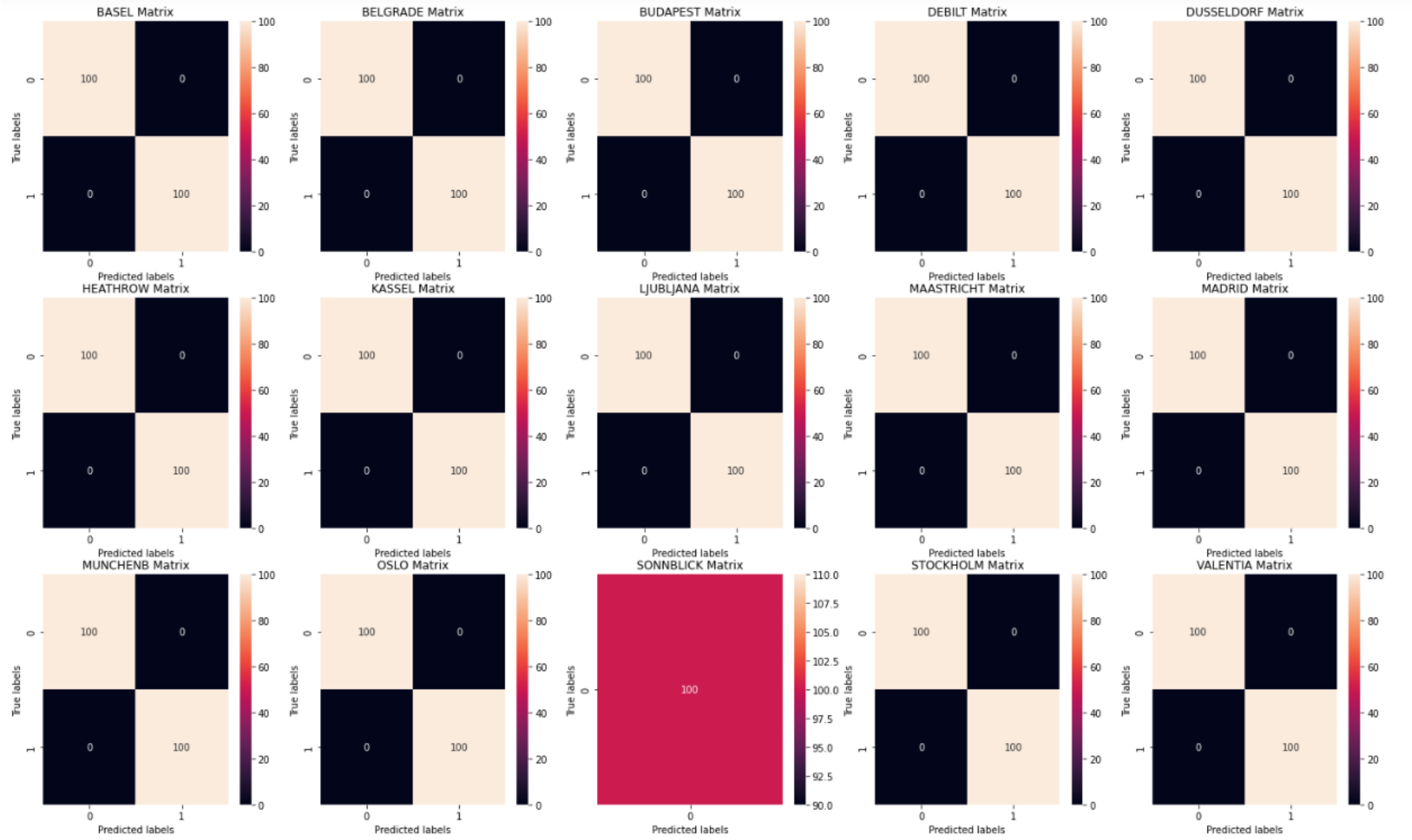
What is a Decision Tree?

- A **Decision Tree** is a supervised learning algorithm used for both **classification** and **regression** tasks.
- It works by **splitting the dataset** into subsets based on feature values, forming a tree structure with **decision nodes** and **leaf nodes**.
- The model recursively chooses the best feature to split the data, aiming to create the most homogeneous subsets.

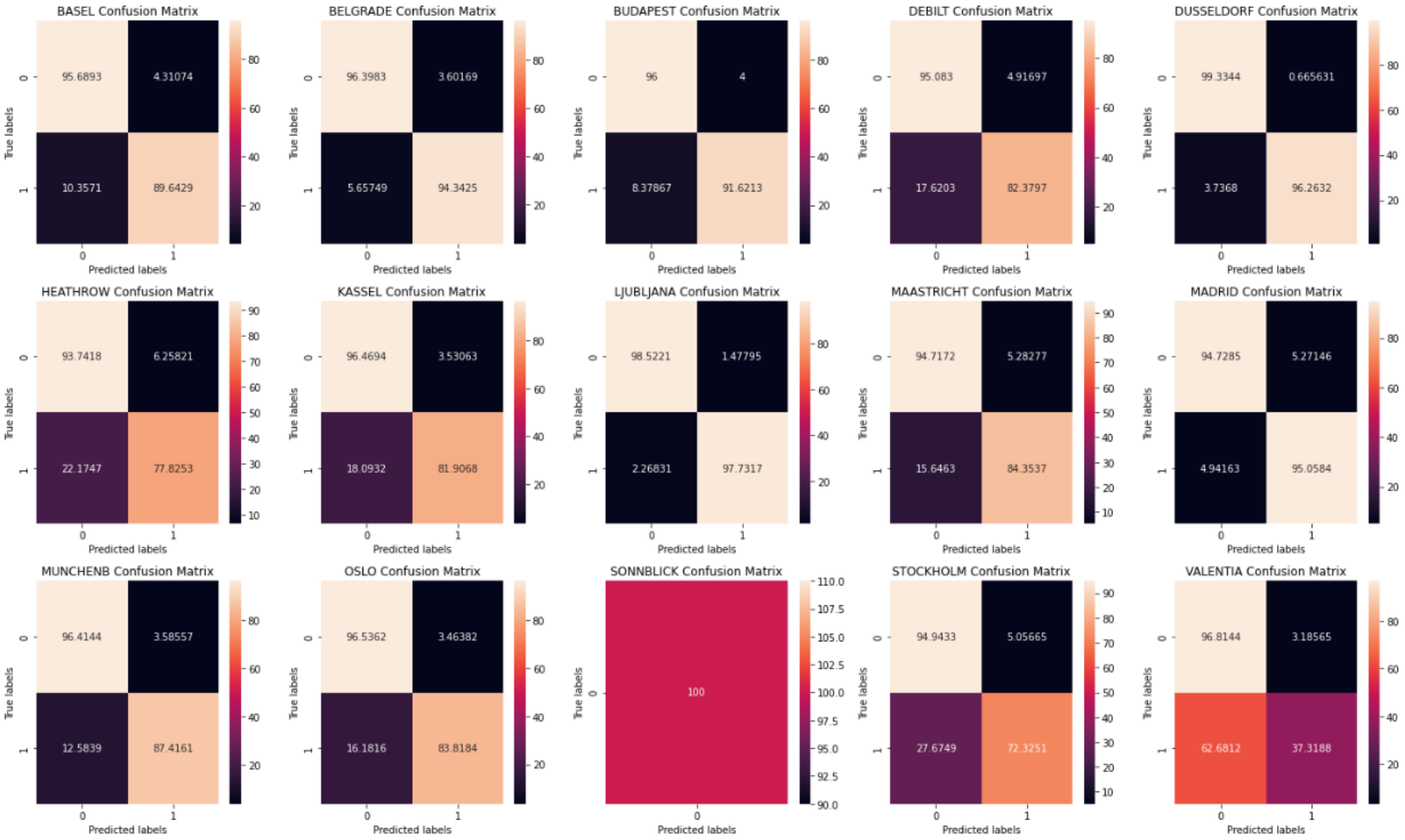
Decision Tree for Climatewins data



Training Data(Confusion Matrix in percentage) with 60% accuracy



Testing Data(Confusion Matrix in percentage) with 63% accuracy



Artificial Neural Network (ANN) Model Analysis

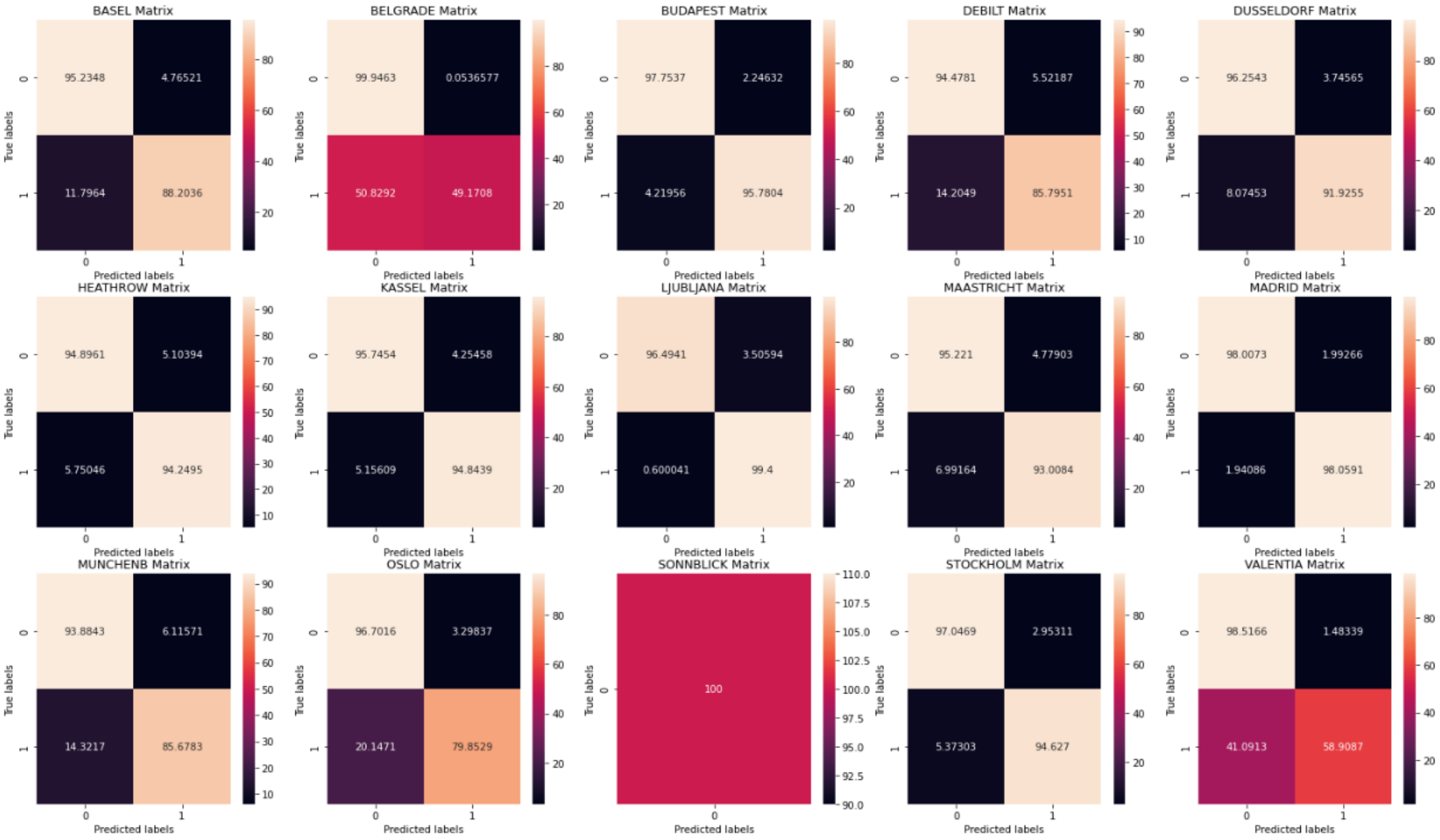
What is an ANN?

- **Artificial Neural Networks (ANN)** are computational models inspired by the human brain. They consist of interconnected **layers of neurons** that process information.
- ANNs are effective at **learning complex patterns** in data by using multiple layers (input, hidden, and output) to **transform inputs** into predictions.

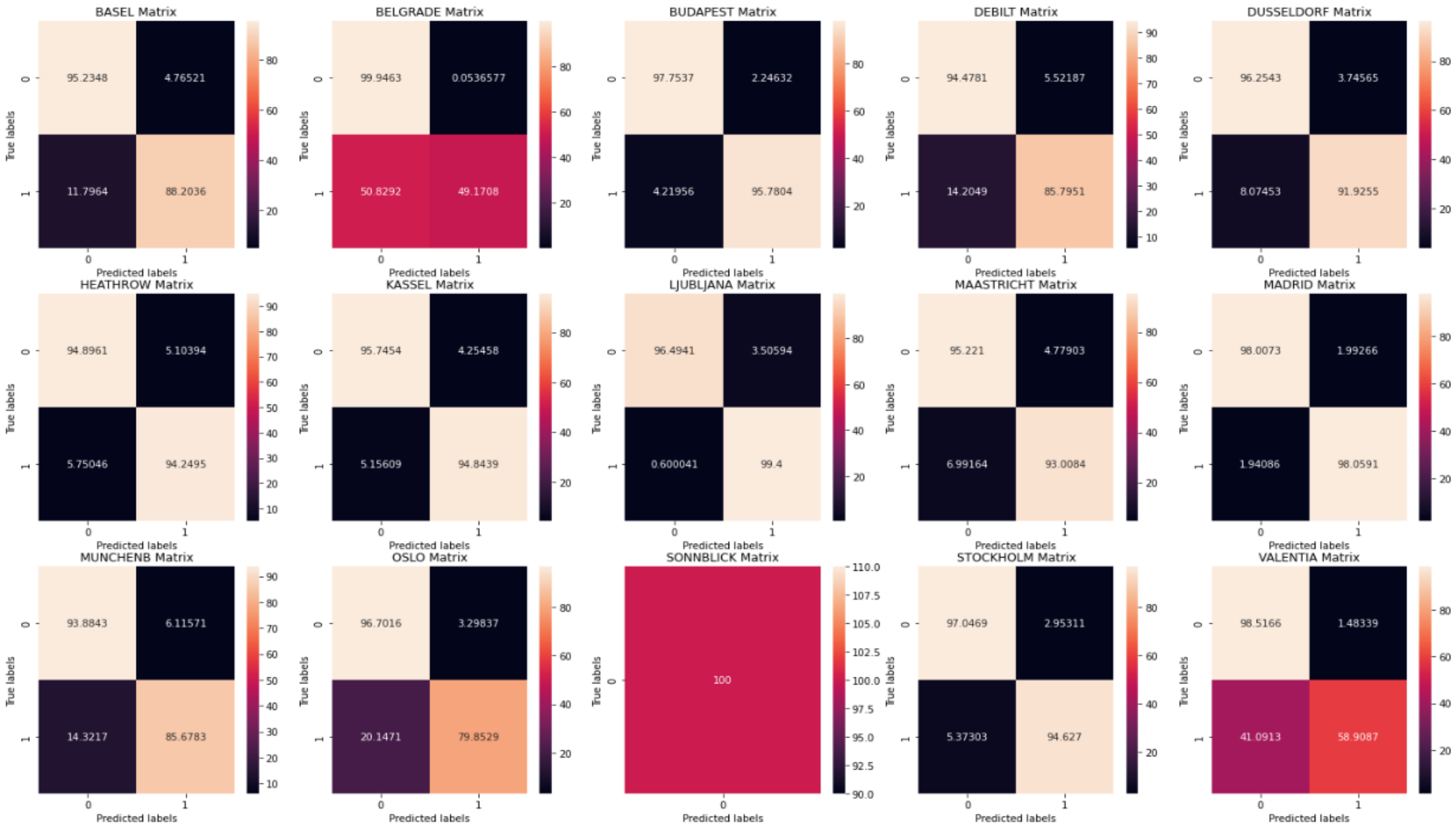
Model Architecture:

- **Hidden Layers:** (35, 25, 15)
  - The model has 3 hidden layers with **35, 25, and 15 neurons**, respectively.
- **Max Iterations:** 2000
  - The model is allowed to train for a maximum of **2000 iterations**.
- **Tolerance (tol):** 0.00000001
  - The model stops when the improvement in the loss function becomes less than this value, ensuring convergence.
- The ANN achieves **63% accuracy on training data** and **60% accuracy on testing data**, indicating moderate success in learning from the dataset.

Training Data(Confusion Matrix in percentage) with 63% accuracy



Testing Data(Confusion Matrix in percentage) with 60% accuracy



Summary and Future Steps

- Hypotheses:

1. Machine learning models can predict **pleasant** and **unpleasant** weather days based on historical data.
2. Some weather stations, due to their unique conditions, may present **imbalanced outcomes**, affecting model performance.
3. Different machine learning models (KNN, ANN, Decision Tree) will perform variably across weather stations, with potential for **overfitting** or **underfitting**.

Summary and Future Steps

- Methods Chosen:
- KNN Model: Demonstrated the **best performance** for predicting weather patterns, with k values between **3-5** optimizing accuracy across most stations.
- ANN Model: Showed **balanced generalization** but slightly lower accuracy (~60%) compared to KNN.
- Decision Tree: The decision tree shows strong performance in classifying **unpleasant** climates across most locations but faces challenges when it comes to **pleasant** weather, especially in cities with higher variability or imbalance in their climate data.

Summary and Future Steps

- Next Steps:

1. **Further optimize the KNN model**, especially focusing on improving accuracy at stations like **Sonnblick** and **Heathrow**.
2. **Enhance data preprocessing**, including addressing data imbalances for stations where prediction struggles.
3. **Tune hyperparameters** in the ANN model to improve overall accuracy and capture more complex weather patterns.
4. **Experiment with alternative models** (e.g., Random Forest, SVM) to explore different approaches for predicting difficult stations.

THANK YOU

Feel free to ask any questions or request further details.

Femina

fjasmin76@gmail.com