Machine Learning Solutions for Weather Prediction

Prepared By Vicky Hung | Data Analyst | July, 2024

PROJECT OVERVIEW

ClimateWins, a fictional European nonprofit organization, leverages machine learning to predict the consequences of climate change.

As an analyst for ClimateWins, it is crucial to rely on data-driven insights to evaluate tools that categorize and predict weather patterns across mainland Europe.

The organization is concerned with the increase in extreme weather events, particularly over the past 10 to 20 years. Utilizing data from the past century, **we aim to use machine learning to find new patterns in weather changes, predict the consequences of climate change around Europe, and identify safe places to live in Europe.**

FINAL REPORT
OVERVIEW

- Machine Learning Options
- Data Source & Additional Data Needs
- Algorithm Exploration: Key Outcomes
- Thought Experiments
- Summary & Next Steps

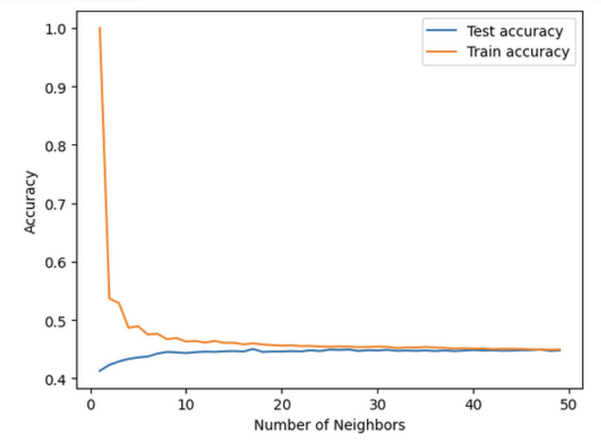
ALGORITHMS EXPLORED

- In the interim report, three supervised models were tested to determine which was most effective for predicting "pleasant weather." Below is a summary of the results:

K-Nearest Neighbors (KNN):

The model fitted to the entire dataset converges with approximately 25 neighbors achieving a **training accuracy of 0.450 and a testing accuracy of 0.448**

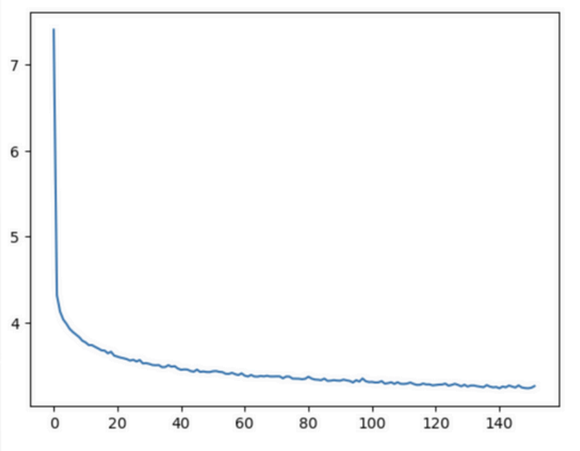
K_range = (1,50)



Artificial Neural Network (ANN):

With 2 layer (80, 80), 500 max iterations and tolerance = 0.0001 drives the best **accuracy score of the training (0.468) and testing data (0.457)**

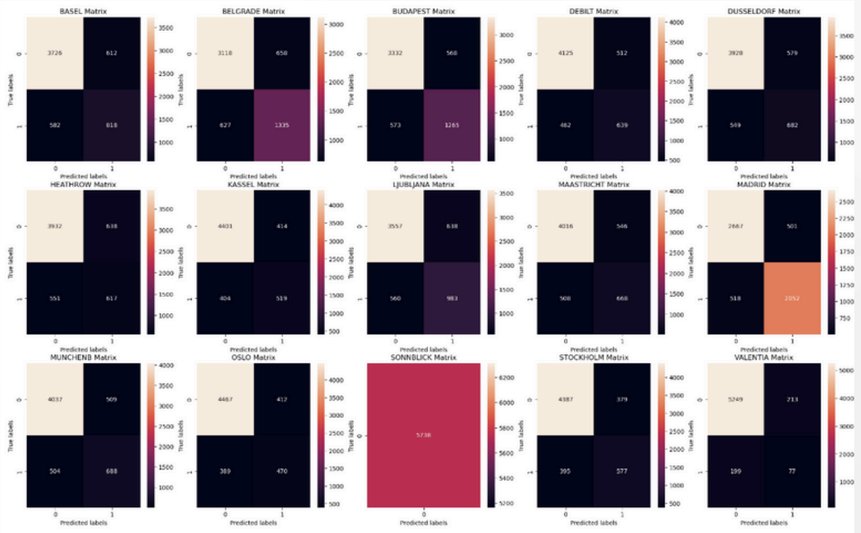
Loss Curve



Decision Tree:

The decision tree yielded an **accuracy score of 0.40** for both training and testing.

Testing Data Accuracy 0.40170791216451723



ALGORITHMS EXPLORED

- Following the interim report, advanced machine learning and deep learning techniques were applied to uncover data patterns and structures, enhancing prediction accuracy.

Hierarchical Clustering
Hierarchical Clustering creates nested clusters by merging or splitting based on distance, producing a dendrogram to show cluster arrangement.

Principal Component Analysis (PCA)
PCA reduces dimensionality by transforming high-dimensional data into principal components, maximizing variance and minimizing information loss.

Convolutional Neural Network (CNN)
CNNs are deep learning models for structured grid data like images, using convolutional layers to learn spatial hierarchies of features automatically.

Recurrent Neural Networks (RNN)
RNNs are deep learning models for sequential data, maintaining a hidden state to capture past information, ideal for time series prediction and natural language processing.

ALGORITHMS EXPLORED

- Following the interim report, advanced machine learning and deep learning techniques were applied to uncover data patterns and structures, enhancing prediction accuracy.

Random Forests
Random Forests are ensemble methods that build multiple decision trees and output the mode or mean prediction, enhancing accuracy and reducing overfitting.

Grid Search
Grid Search exhaustively searches a specified parameter grid to find optimal model parameters, enhancing algorithm performance.

Bayesian Optimization
Bayesian Optimization is a probabilistic method for optimizing hyperparameters, using a surrogate model and utility function to find the global minimum efficiently.

Generative Adversarial Networks (GANs)
GANs are deep learning models with a generator and discriminator that compete, producing realistic synthetic data by evaluating and improving data authenticity.

DATA

Data Set

Europe Temperature Data Set - This data is collected by the European Climate Assessment & Data Set project (www.ecad.eu)

The dataset includes weather observations from 18 European stations, spanning the late 1800s to 2022, covering temperature, wind speed, snowfall, and other variables.

Data Accuracy & Reliability

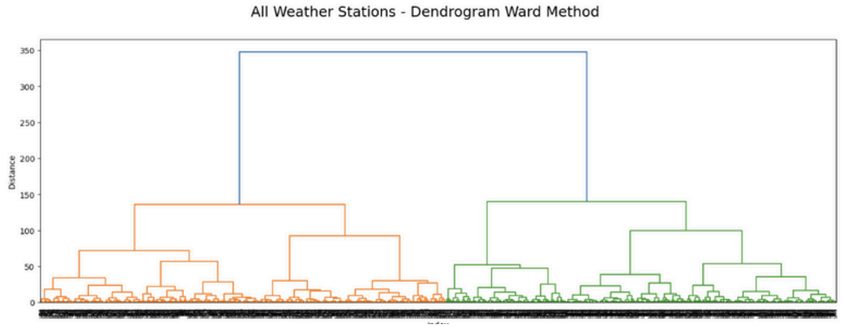
- KNMI meticulously gathers and validates the data, ensuring top-tier accuracy and reliability
- Weather stations across Europe are strategically placed to represent diverse climates
- The data is thoroughly checked for missing values, duplicates, and inaccuracies

Additional Data Source Consideration

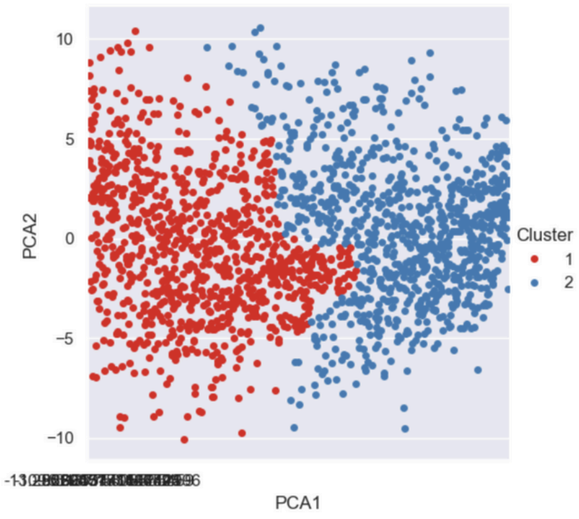
- Extreme Weather Event Records: Historical data on natural disasters and extreme weather events.
- Satellite Data & Ground-based Measurements: including UV index forecasts and historical UV radiation data
- Environmental and Geographical Data: including land use, air quality, and water resources.
- Sea Level Rise and Coastal Data: including sea level rise, ocean temperatures, and marine conditions.
- Remote Sensing and Satellite Data: Satellite data on land, ocean, and atmospheric conditions.
- Emissions Data: Data on greenhouse gas emissions.

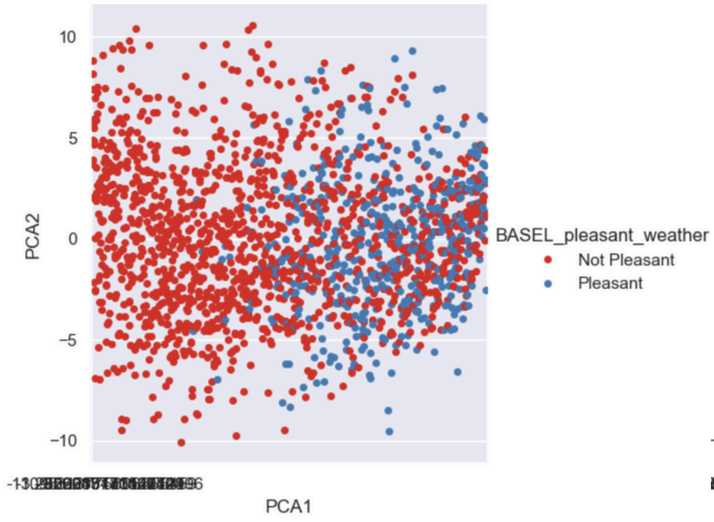
KEY FINDINGS FROM ALGORITHM EXPLORATION

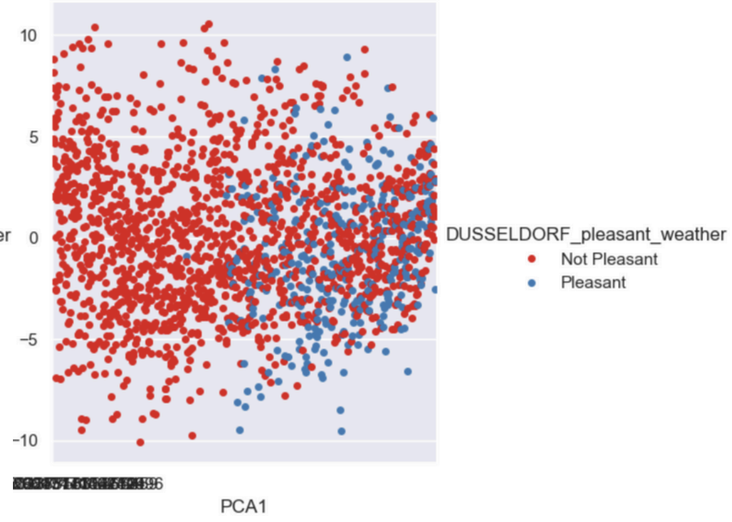
- Using the **Dendrogram Ward method, combined with PCA** to reduce variables from 147 to 2, results in more well-defined clusters.



- Two clusters were created with a threshold value of 150 and plotted on the chart to compare with data from two weather stations. **The results show that the clusters effectively represent Pleasant and Not Pleasant weather conditions.**

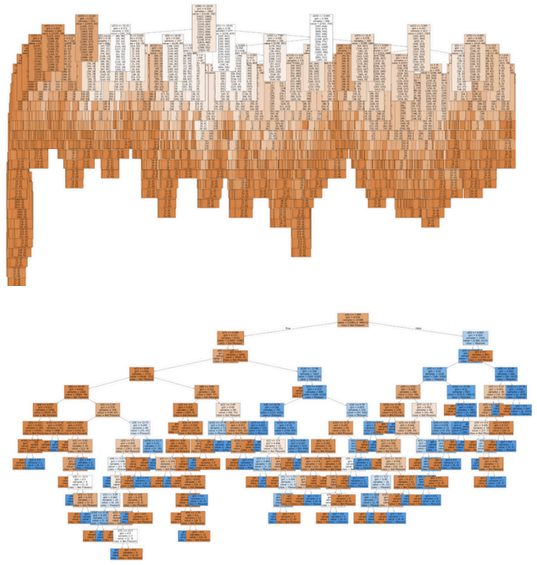




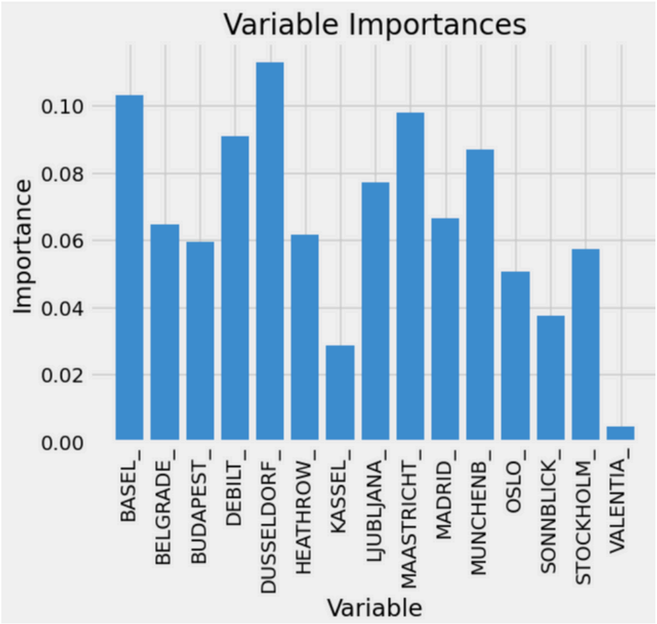


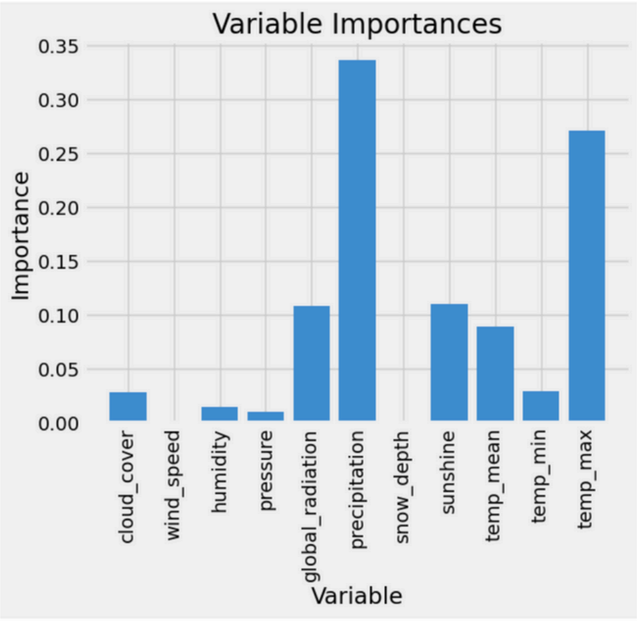
KEY FINDINGS FROM ALGORITHM EXPLORATION

- **The random forest model** yielded an **accuracy score of 0.60** for all weather stations and **1.0** for individual weather stations.



- The **variable importance** charts identify the top influential weather stations (Düsseldorf, Maastricht, and Basel) and the most influential weather observations (precipitation, maximum temperature, and global radiation).





KEY FINDINGS FROM ALGORITHM EXPLORATION

- **The CNN** with **Bayesian Optimization** has significantly improve the **accuracy score to 0.7954** and reduce loss to 0.5482

Key Parameters of the CNN Model

```
epochs = 100
batch_size = 64
n_hidden = 64
optimizer = Adam(learning_rate=0.000001)

timesteps = len(X_train[0])
input_dim = len(X_train[0][0])
n_classes = len(y_train[0])

# Implement complex layers
model = Sequential()
model.add(Conv1D(n_hidden, kernel_size=2, activation='relu', input_shape=(timesteps, input_dim)))
model.add(Dense(64, activation='relu'))
model.add(MaxPooling1D())
model.add(Dropout(0.4))
model.add(Flatten())
model.add(Dense(n_classes, activation='softmax'))
```

```
Epoch 100/100
287/287 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.0735 - loss: 53.5616
```

Key Parameters of the CNN Model with Bayesian Optimization

```
# Define optimizer with the specified learning rate
optimizer = Adadelta(learning_rate=0.6051038616257767)

# Verify the label values in y_train are within the valid range [0, n_classes-1]
y_train = np.array(y_train)
n_classes = 6
y_train = np.clip(y_train, 0, n_classes - 1)

# Parameters
epochs = 32
batch_size = 961
timesteps = len(X_train[0])
input_dim = len(X_train[0][0])
layers1 = 1
layers2 = 3
activation = 'softsign'
kernel = 1
neurons = 74
normalization = 0.020584494295802447
dropout = 0.7319939418114051
dropout_rate = 0.17959754525911098
```

```
Epoch 32/32
20/20 - 0s - 19ms/step - accuracy: 0.7954 - loss: 0.5482
```

THOUGHT EXPERIMENTS

Thought Experiment 1: Identifying weather patterns outside the regional norm in Europe

Thought Experiment 2: Finding new patterns in weather changes over the last 60 years

Thought Experiment 3: Determining the Safest Places to Live in Europe Within the Next 25 to 50 Years

Thought Experiment 1

IDENTIFYING WEATHER PATTERNS OUTSIDE THE REGIONAL NORM IN EUROPE.

Objective:

- Determine weather patterns that deviate significantly from regional norms to understand anomalies.

Data Needed:

- Historical weather observations
- Extreme Weather Event Records.
- Satellite Data & Ground-based Measurements

1. **Data Preprocessing:** Clean and standardize weather observations, extreme event records, and satellite data to ensure consistency and accuracy before applying machine learning models.
2. **KNN:** Use KNN to classify weather events based on regional norms, identifying outliers that represent significant deviations from typical patterns in the dataset.
3. **ANN:** Train an ANN to model complex relationships within the weather data, enhancing the detection of subtle anomalies and patterns that KNN might miss.
4. **Anomaly Analysis and Validation:** Validate identified anomalies by cross-referencing with extreme weather event records and satellite data, ensuring accurate detection of significant deviations from regional weather norms.

Thought Experiment 2

FINDING NEW PATTERNS IN WEATHER CHANGES OVER THE LAST 60 YEARS

Objective:

- Identify new patterns in weather changes in Europe over the past 60 years using historical weather data

Data Needed:

- Historical weather observations
- Extreme Weather Event Records
- Environmental and Geographical Data

1. **PCA:** To reduce the dimensionality of the dataset from multiple variables to a few principal components, highlighting the main patterns and trends in weather changes.
2. **Hierarchical Clustering:** Group similar weather patterns by creating clusters based on the principal components. This helps in identifying and visualizing new trends and changes over time.
3. **Time-Series Analysis:** Apply time-series analysis techniques to the clustered data to detect temporal patterns and shifts, examining how weather changes evolve over the 60-year period.
4. **Validation and Interpretation:** Validate the identified patterns using Extreme Weather Event Records and Environmental Data. Interpret the results to understand the implications of these new patterns on climate change.

Thought Experiment 3

DETERMINING THE SAFEST PLACES TO LIVE IN EUROPE WITHIN THE NEXT 25 TO 50 YEARS

Objective:
Identify the safest areas to live in Europe based on projected climate change impacts and weather patterns.

Data Needed:

- Historical weather observations
- Sea Level Rise and Coastal Data
- Environmental and Geographical Data
- Satellite Data & Ground-based Measurements
- Emissions Data

1. **Random Forests:** To assess and rank the safety of different regions based on multiple criteria such as historical weather data, environmental risks, and Sea Level Rise
2. **Grid Search:** To optimize the parameters of the predictive models used, ensuring the most accurate predictions.
3. **RNN:** Analyze sequential climate data to predict future trends and impacts, offering a dynamic and time-aware safety perspective.
4. **Bayesian Optimization:** To refine the model by incorporating probabilistic approaches to parameter selection, ensuring the best possible predictive performance and reliability in identifying safe regions.

SUMMARY

Thought Experiment 1:
Identifying weather patterns outside the regional norm in Europe

This Has the Most Potential:

- **Resource Efficiency:** Leverages existing data and tools, making it cost-effective for ClimateWins with limited resources.
- **Actionable Insights:** Provides immediate insights without needing to develop new products, allowing quick implementation of findings.
- **High Impact:** Focuses on detecting anomalies that can help us take quick actions to boost climate resilience and preparedness.

Thought Experiment 2:
Finding New Patterns in Weather Changes Over the Last 60 Years

Potential Challenges:

- **Dimensionality Reduction:** Reducing the dimensionality of the dataset without losing critical information can be challenging.
- **Complex Pattern Detection:** Hierarchical clustering and time-series analysis must be carefully adjusted to spot and show these patterns accurately.
- **Scalability and Generalization:** Making sure the models work well with other regions or datasets can be challenging. The results should be reliable and useful across different climates and locations.

Thought Experiment 3:
Determining the Safest Places to Live in Europe Within the Next 25 to 50 Years

Potential Challenges:

- **Data Integration:** Integrating diverse datasets, such as historical weather observations, sea level rise data, environmental data, and emissions data, can be complex.
- **Algorithm Complexity:** Implementing advanced algorithms like RNNs and Bayesian Optimization requires significant computational resources and expertise.
- **Validation and Reliability:** Validating predictions for future conditions is inherently challenging due to the lack of real-world future data.

NEXT STEPS

Data Collection and Preprocessing:

- Gather and clean historical weather data, extreme weather event records, and satellite measurements. Ensure data consistency and accuracy.

Data Management

- Set up robust data storage and management systems to handle large volumes of weather data and ensure data security and integrity.

Resource Allocation:

- Ensure that skilled data scientists, analysts, and IT staff are available to manage the data collection, preprocessing, and model development.
- Invest in high-performance computing resources to handle the computational demands of advanced algorithms like ANN and KNN.

Training and Skill Development

- Organize workshops to keep the team updated on recent advancements in climate science and machine learning applications.

THANK YOU

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

vicky@climatewins.com