Predicting Weather Conditions

Machine Learning Analysis – ClimateWins

Kyle Stanford

09/09/2024

Agenda

1. Objectives & Hypotheses
2. Data Sets & Data Bias
3. Data Optimisation
4. Supervised Machine Learning
5. Scaling and Model Performance
6. Conclusion & Future Steps

Objectives

ClimateWins wants to know if machine learning can be used to predict picnic suitability based on weather data from various stations.

Hypotheses

1. ANN models will outperform KNNs and Decision Trees because they're more complex.
2. Some weather station data may be unsuitable for predictions due to data quality restraints.
3. Scaling the data will significantly improve the performance of machine learning models.

Data Set

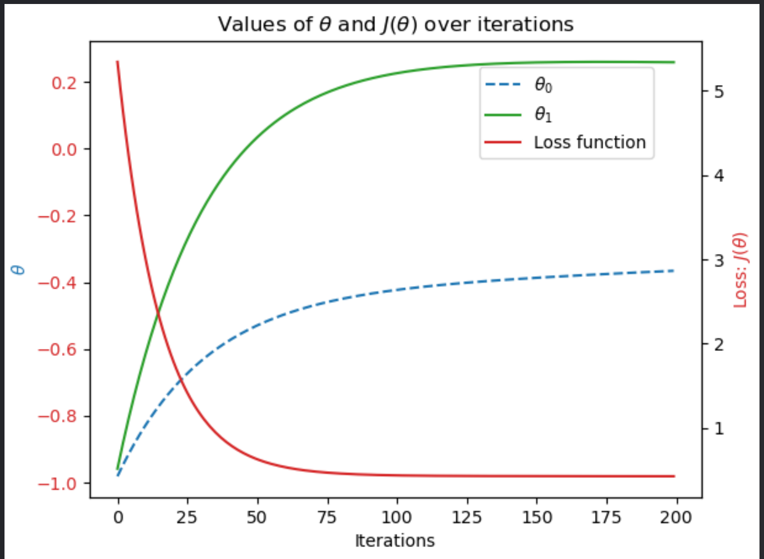
- Data was collected by the European Climate Assessment & Data Set Project.
- Daily weather metrics (temperature, precipitation, wind speed, etc) from 18 weather stations across Europe.
- Data ranges from 1960 to 2022.
- Data Link
- Secondary data set – picnic suitability data.

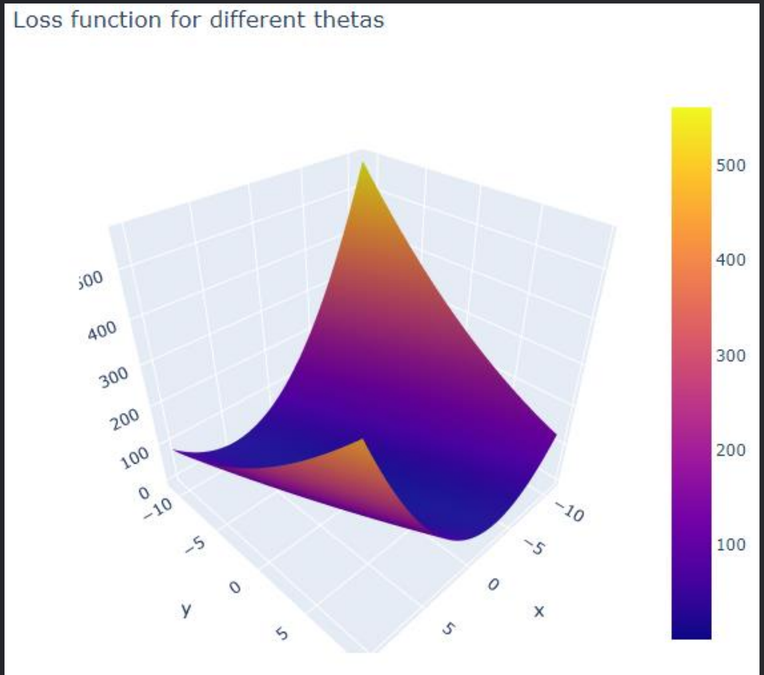
Data Bias

- Sampling Bias: Only 18 of the 23755 weather stations throughout Europe and the Mediterranean have been included. This could misrepresent climate patterns, thus impacting any analyses.
- Measurement Bias:
  - Changes to measurement tools/methods
  - Human subjectivity in determining "pleasant weather"

Data Optimisation

- **Gradient descent** was applied to temperature data from various weather stations across different years.
- Finds parameters for the best-fit line or curve that minimises the cost function (error between predicted and actual values).
- Cost function approached zero (approx. 0.5) in all cases, meaning the data could be approximated accurately.

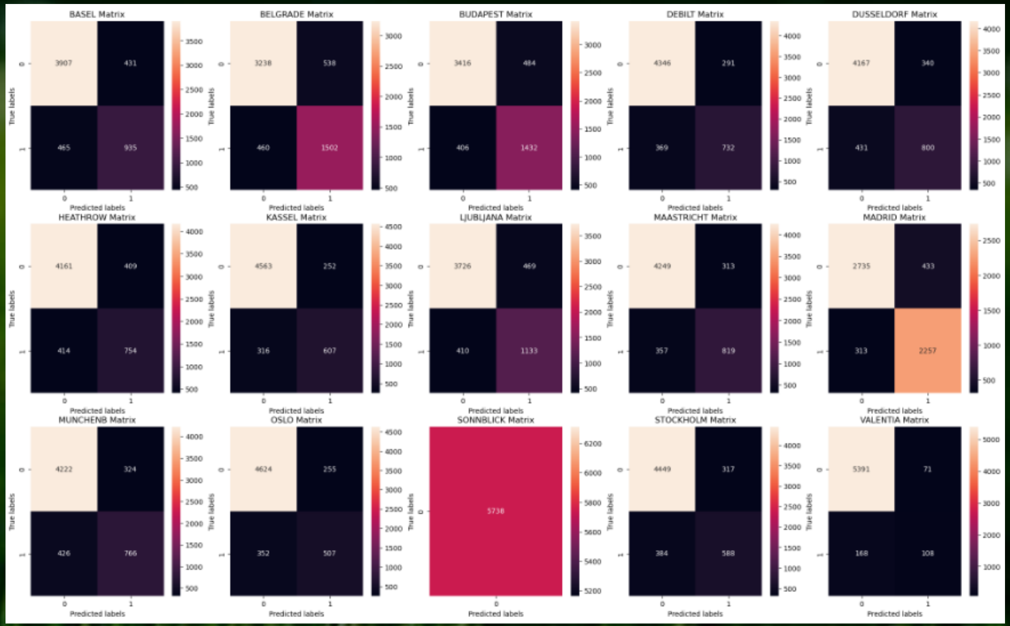




Supervised Machine Learning

K-Nearest Neighbours

KNN was used to classify a given day's weather data as pleasant or unpleasant by comparing data points to their closest neighbouring points and the category to which they belonged.



Confusion Matrix: KNN Testing Data

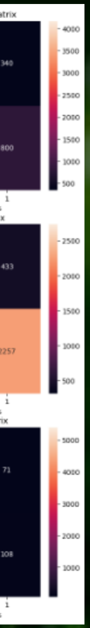
Average Accuracy Scores

Training Data: 93.91 %

Testing Data: 88.15 %

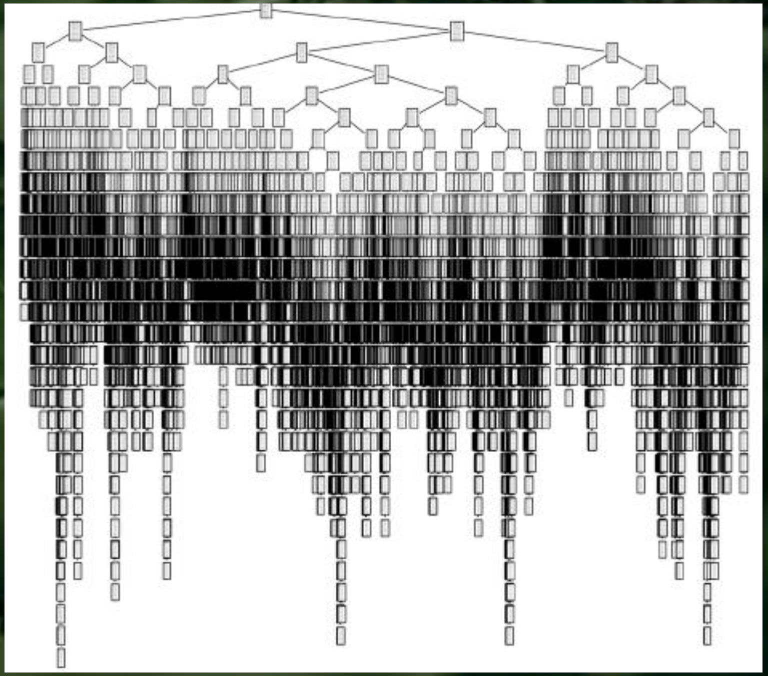
Supervised Machine Learning

Decision Tree Model



Decision Trees make predictions of data points by asking multiple questions – sorting the data like a flowchart.

Our decision tree overfit the training data.



Decision Tree made during project

Average Accuracy Scores

Training Data: 100 %

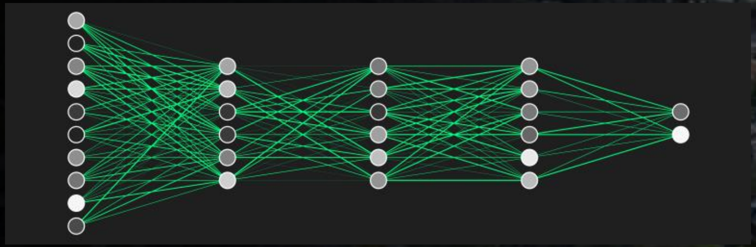
Testing Data: 87.14 %

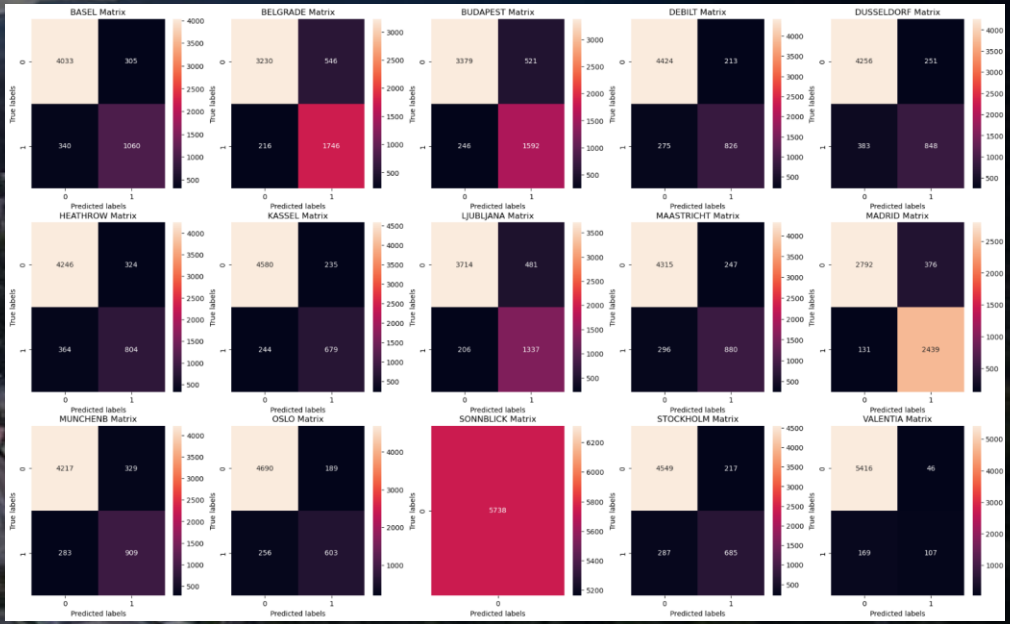
Supervised Machine Learning

Artificial Neural Network

ANNs were used to classify our weather data by processing data points through layers of neurons.

Multiple combinations of neuron layers and neuron totals were trialled.





Confusion Matrix: ANN Testing Data

Average Accuracy Scores

Training Data: 91.37 %

Testing Data: 91.04 %

Did Scaling Matter?

	Scaled Data	Unscaled Data
KNN:	Training Accuracy: 93.91% Testing Accuracy: 88.15%	Training Accuracy: 94.02% Testing Accuracy: 88.45%
Decision Tree:	Training Accuracy: 100% Testing Accuracy: 87.14%	Training Accuracy: 100% Testing Accuracy: 87.23%
ANN:	Training Accuracy: 91.37% Testing Accuracy: 91.04%	Training Accuracy: 91.32% Testing Accuracy: 91.04%

- Scaling is **extremely important.**
- In this case, scaling had minimal impact since the columns were all related to temperature.
- The models' performances, using unscaled data, would likely be diminished by training each with additional weather metrics.

Model Evaluation?

KNN:

Training Accuracy: 93.91%

Testing Accuracy: 88.15%

- The KNN model performs generally well.
- Was relatively easy to set up and implement due to having only a few hyperparameters.
- May be difficult to scale to larger data sets.

Decision Tree:

Training Accuracy: 100%

Testing Accuracy: 87.14%

- Clear case of overfitting.
- There may be a case for using this type of model, but branches will need to be pruned so it can generalise to new data effectively.

ANN:

Training Accuracy: 91.37%

Testing Accuracy: 91.04%

- The best performing model.
- Required the most fine-tuning and experimentation.
- Efficient once trained and extremely useful for complex and high-dimensional data.

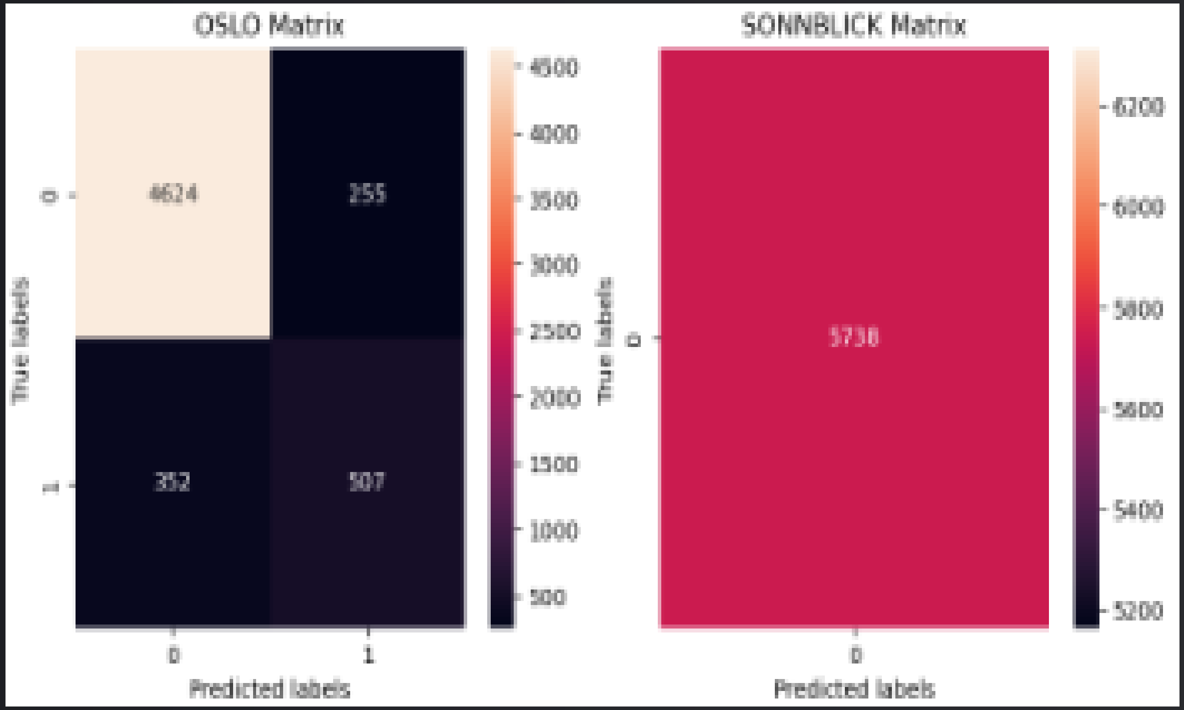
Limitations

1. All models have been trained on an unbalanced data (picnic suitability).

   - Bias towards the majority class.
   - Difficulty in learning minority class patterns.

2. Sonnblick stations weather data was completely unbalanced.

   - Inflates overall accuracy score.
   - Model is completely ungeneralisable to new data.



Conclusions & Future Steps

Overall, machine learning models show promise in predicting climate patterns but should be further explored and experimented with.

KNN:

KNN results were promising and would work well with data of similar size and dimensionality.

However, much larger datasets or those with many extra features should be avoided.

Decision Tree:

Decision trees could be great if steps are taken to reduce the overfitting issue:

- Prune branches from the model to simplify categorisation.
- Use ensemble methods like Random Forests.

ANN:

Best performing model.

Despite its “black-box” nature, it is scalable and may work even better in more complex data with more features.

Avoid use in small datasets due to risk of overfitting.

Questions?