

Weather Conditions and Climate Change with ClimateWins

Dirk Davis

8/13/25

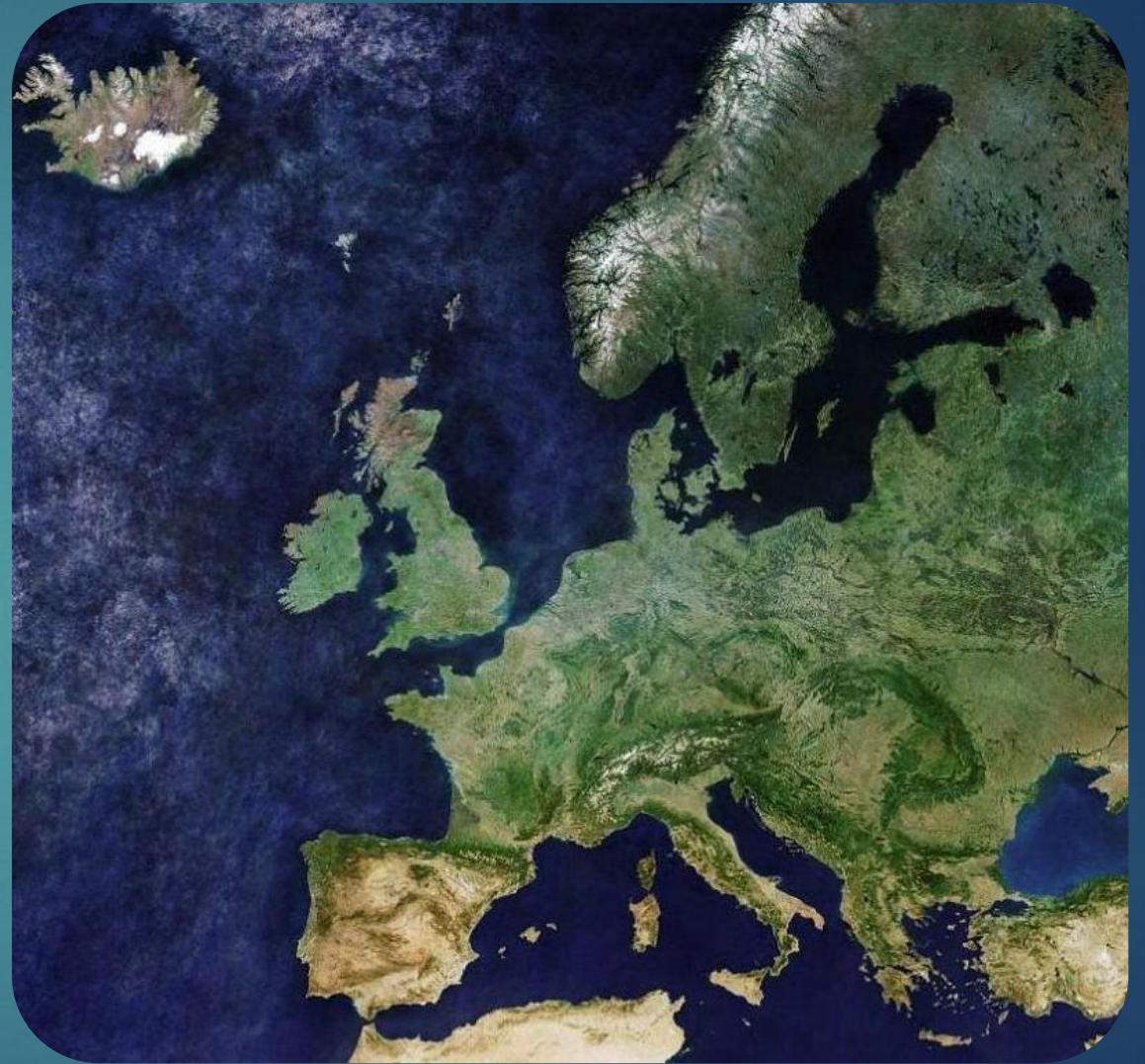


Photo by the Copernicus Sentinel-3A weather satellite of the European Space Agency (ESA)

ClimateWins

Project: Climate Forecasting

Objective

ClimateWins, a non-profit organization, seeks to apply machine learning (ML) to predict the consequences of climate change in Europe and beyond. It is concerned with the acceleration of extreme weather events in recent decades, believing that such fat-tailed risk can be viably modeled and predicted with ML.

Data Sources

- Weather data: European Climate Assessment & Dataset project
- Pleasant Weather answers dataset: Undisclosed origin

Tools

Python, Excel, ML-friendly libraries including tensorflow/keras. ChatGPT for ML model customization, fine-tuning and regularization Python code.



AI-rendered image by OpenAI's ChatGPT, GPT-4o model

Datasets

Weather dataset: Daily observations (1960-2022) at 18 European weather stations, measuring temperature (mean, max, min), sunshine, cloud cover, wind speed, humidity, air pressure, precipitation, snow depth & global radiation.

Pleasant Weather answers dataset: Daily boolean "1" (Pleasant) or "0" (Unpleasant) weather outcome (1960-2022) at each of the 18 European weather stations.

Data Biases: Potentially include measurement location & temporal period selection, measurement errors, non-uniform scales & standards, significant features (variables) omission, collection equipment malfunction and aging, undisclosed interpolation or extrapolation of missing data, day length is latitude-dependent thus varies.

Red Flags on Pleasant Weather answers dataset

Undisclosed origin, undefined criteria, subjective, accuracy issues (entire month of October 2022 had "Unpleasant" daily weather at all 18 weather stations: would you believe that verdict?), oversimplified into a binary classification outcome, missing data at some stations.

Weather dataset completeness: Valentia & Kassel weather stations' three temperature features time series data flatlines for 8-10+ years at the end of the 60+ year study period, e.g. > 5% of observations. These two weather stations were thus removed from the analysis. The unconfirmed rumour is that the teams went out for Guinness and Lowenbrau, never to return to their posts.

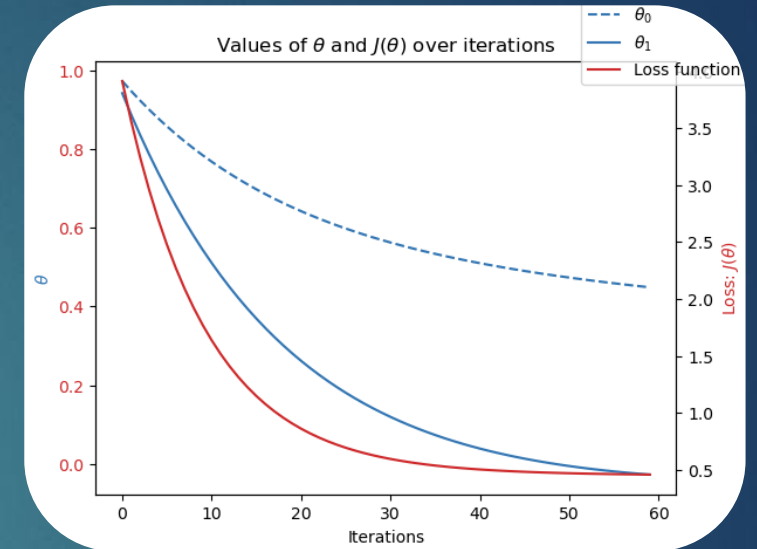
Hypotheses

- (1) Causality of extreme weather events (drought, floods as defined by the EU-backed Copernicus Climate Change Service, C3S) cannot be determined solely from temperature variation.
- (2) Weather forecasting accuracy is inversely correlated with forecast period length.
- (3) Leveraging unsupervised non-linear t-SNE or UMAP dimensionality reduction techniques to optimize feature selection for the "Pleasant Weather" classification exercise via the ANN (Artificial Neural Network) model will increase the model "Balanced Accuracy (Test)" score by at least 10%, compared to the constraint of using only temperature features as model inputs.

Optimization for problem-solving

Gradient Descent: The simplest version of this optimization method was explored to introduce loss minimization as a concept. The exercise highlighted the relationship between parameters (the thetas in the line plot, to the right) and model (algorithm) performance, where "training" a model involves finding the best parameters to minimize loss.

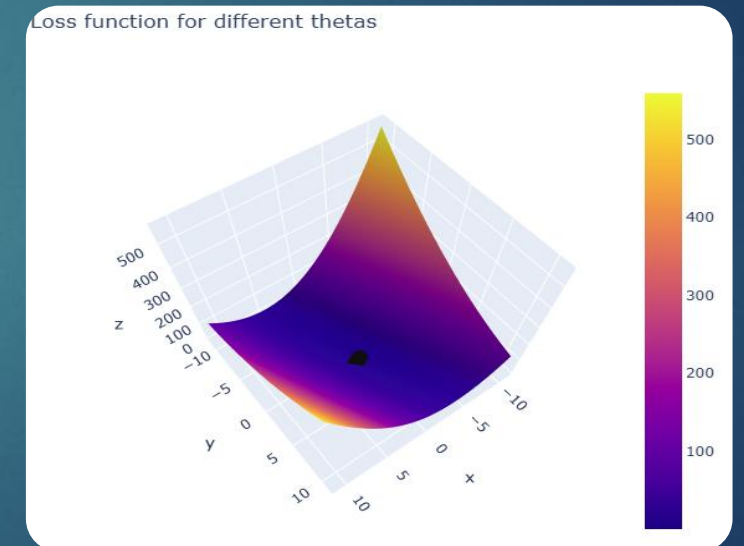
Adaptive Moment Estimation ("Adam"): This powerful optimization method was later used to significantly improve ANN Balanced Accuracy (Test) model performance on a Pleasant/Unpleasant weather day classification exercise (in the next slides), in conjunction with regularization (L2, dropout) techniques which minimized model overfit on training data.



Supervised learning

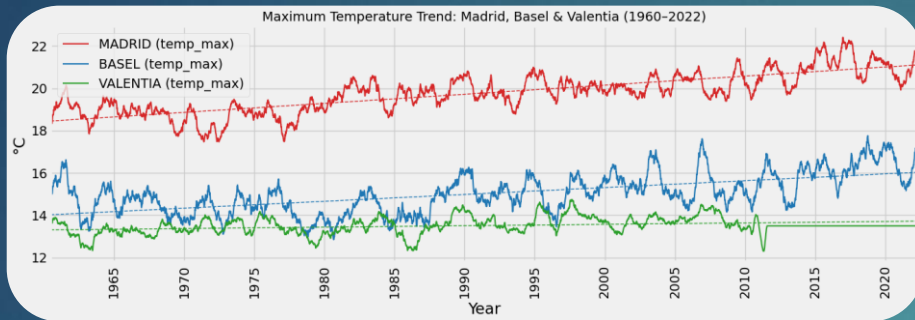
K-Nearest Neighbors (KNN) model

This algorithm uses Euclidean, Manhattan or Minkowski distance (closer=better) between data points to classify or predict the value of a new data point in the vicinity of the training data points (neighbors). KNN was used to classify Pleasant/Unpleasant weather days based on the weather dataset. However, subpar KNN model results confirmed that it does not work well with complex features such as weather data.

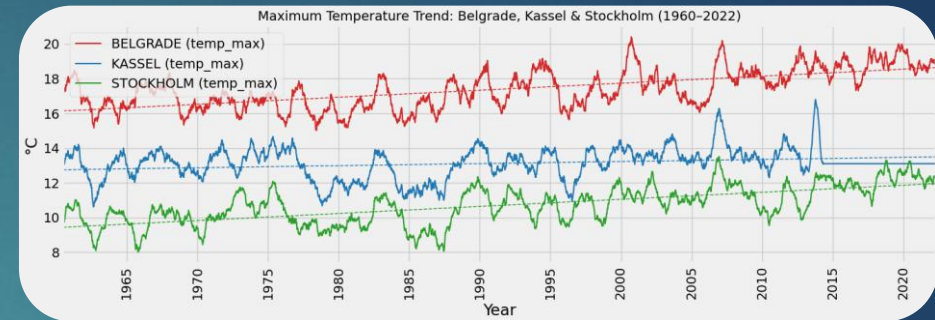


Gradient descent followed a path along the surface to find the x, y coordinates (thetas) associated with the loss function minimum (black dot).

Data & Model adjustments for binary Pleasant/Unpleasant weather classification prediction



Which data series would you reject?



Weather data: Valencia & Kassel weather station data series were removed (flatlining for >5% of observations); Roma, Gdansk & Tours were removed for insufficient answers dataset values; temperature raw data series were 2-step transformed first by differencing then by z-score rescaling with stationarization as the goal (confirmed by decomposition plots, ADF & KPSS tests and critically, ACF plots) to mitigate ANN model sensitivity to non-stationary data. Three temperature features (mean, max, min) were specified in the Exercise as the model feature-selection "constraint", to handicap model predictive accuracy and explore comparative model weaknesses. Nonetheless, the minimum temperature feature was removed based on a feature importance ranking by the .feature_importances_ attribute, to reduce data "noise" and improve model performance (which it did). Two features remained: mean, max temperatures. No other features (precipitation, cloud cover, wind speed, humidity, etc.) were allowed for use in the binary classification prediction Exercise.

Decision Tree ("DT") model: GridSearchCV was run to automate discovery of the best hyperparameters by testing all combinations in a grid using 3-fold cross-validation. Metadata includes: Criterion=entropy, max depth=7, min samples split=2, min samples leaf=20, **class weight=balanced** (a critical argument, based on the imbalanced class answers data profile), random state=42.

Artificial Neural Network ("ANN") model: GridSearchCV was run again to discover hyperparameter best configuration. Metadata includes: tensorflow/keras library/API to build model, hidden units=(40,) as network architecture, optimization (algorithm=Adam, lr=0.001, activation='relu', loss=binary_crossentropy), regularization (L2 alpha=0.001, dropout=0.10) to tame the overfit problem, random state=42, training control & evaluation (epochs=300, validation split=0.10, batch size=256).

Decision Tree (DT) versus Artificial Neural Network (ANN) model comparative results

Results were handicapped by: (1) class imbalance, (2) features constraint and (3) answers dataset red flags

Weather Station	Model	BASELINE	
		Unpleasant (0) %	Pleasant (1) %
Basel	DT	75.32	24.68
Basel	ANN	75.32	24.68
Oslo	DT	84.40	15.60
Oslo	ANN	84.40	15.60
Stockholm	DT	83.03	16.97
Stockholm	ANN	83.03	16.97
Budapest	DT	67.62	32.38
Budapest	ANN	67.62	32.38
Debilt	DT	80.57	19.43
Debilt	ANN	80.57	19.43
Dusseldorf	DT	78.50	21.50
Dusseldorf	ANN	78.50	21.50
MunchenB	DT	79.23	20.77
MunchenB	ANN	79.23	20.77
Madrid	DT	55.35	44.65
Madrid	ANN	55.35	44.65
Belgrade	DT	65.18	34.83
Belgrade	ANN	65.18	34.83
Ljubljana	DT	72.22	27.78
Ljubljana	ANN	72.22	27.78
Heathrow	DT	78.39	21.61
Heathrow	ANN	78.39	21.61
Maastricht	DT	79.23	20.77
Maastricht	ANN	79.23	20.77
Group - Mean	DT	74.92	25.08
Group - Mean	ANN	74.92	25.08
Group - Sdev	DT	8.11	8.11
Group - Sdev	ANN	8.11	8.11
Group - Max	DT	84.40	44.65
Group - Max	ANN	84.40	44.65
Group - Min	DT	55.35	15.60
Group - Min	ANN	55.35	15.60

Weather Station	Model	CLASS-SPECIFIC SCORES (TEST)								
		Recall (Pleasant)	Lift	Precision	Gap (Recall-	Specificity	Precision	Gap (Recall-	Balanced	F1 Score
		Test %	(Recall) %	(Pleasant) Test %	Precision) %	(Unpleasant) Test %	(Unpleasant) Test %	Precision) %	Accuracy (Test) %	(Pleasant) Test %
Basel	DT	66.69	42.01	37.85	28.84	63.37	85.05	-21.68	65.03	48.29
Basel	ANN	49.93	25.25	55.88	-5.95	86.81	83.83	2.98	68.37	52.74
Oslo	DT	72.34	56.74	23.29	49.05	55.04	91.34	-36.30	63.69	35.24
Oslo	ANN	14.05	-1.55	50.39	-36.34	97.39	85.72	11.67	55.72	21.97
Stockholm	DT	57.90	40.93	31.50	26.40	73.43	89.21	-15.78	65.66	40.80
Stockholm	ANN	18.10	1.13	53.71	-35.61	96.71	84.84	11.87	57.40	27.08
Budapest	DT	68.17	35.79	46.18	21.99	62.38	80.54	-18.16	65.27	55.06
Budapest	ANN	58.57	26.19	55.21	3.36	77.50	79.80	-2.30	68.04	56.84
Debilt	DT	67.96	48.53	30.52	37.44	62.07	88.76	-26.69	65.02	42.13
Debilt	ANN	34.25	14.82	53.90	-19.65	92.82	85.20	7.62	63.53	41.88
Dusseldorf	DT	74.25	52.75	31.15	43.10	53.73	88.10	-34.37	63.99	43.89
Dusseldorf	ANN	36.85	15.35	53.26	-16.41	90.88	83.62	7.26	63.87	43.56
MunchenB	DT	75.58	54.81	32.97	42.61	59.20	90.13	-30.93	67.39	45.91
MunchenB	ANN	39.45	18.68	53.73	-14.28	90.98	84.98	6.00	65.22	45.50
Madrid	DT	72.02	27.37	53.99	18.03	50.35	68.98	-18.63	61.18	61.71
Madrid	ANN	80.32	35.67	58.15	22.17	53.25	76.98	-23.73	66.78	67.46
Belgrade	DT	57.83	23.01	53.57	4.26	73.20	76.45	-3.25	65.52	55.62
Belgrade	ANN	66.13	31.31	57.60	8.53	73.98	80.34	-6.36	70.06	61.57
Ljubljana	DT	80.21	52.43	40.65	39.56	55.22	87.95	-32.73	67.71	53.95
Ljubljana	ANN	60.62	32.84	57.99	2.63	83.21	84.68	-1.47	71.91	59.27
Heathrow	DT	68.97	47.36	27.59	41.38	48.23	84.46	-36.23	58.60	39.41
Heathrow	ANN	24.06	2.45	48.81	-24.75	92.78	81.03	11.75	58.42	32.23
Maastricht	DT	61.27	40.50	34.10	27.17	68.16	86.74	-18.58	64.71	43.81
Maastricht	ANN	40.38	19.61	54.43	-14.05	90.91	85.01	5.90	65.64	46.36
Group - Mean	DT	68.60	43.52	36.95	31.65	60.37	84.81	-24.44	64.48	47.15
Group - Mean	ANN	43.56	18.48	54.42	-10.86	85.60	83.00	2.60	64.58	46.37
Group - Sdev	DT	6.64	10.27	9.44	12.35	7.96	6.25	9.72	2.40	7.55
Group - Sdev	ANN	19.31	12.09	2.74	16.98	11.92	2.66	9.73	4.88	13.51
Group - Max	DT	80.21	56.74	53.99	49.05	73.43	91.34	-3.25	67.71	61.71
Group - Max	ANN	80.32	35.67	58.15	22.17	97.39	85.72	11.87	71.91	67.46
Group - Min	DT	57.83	23.01	23.29	4.26	48.23	68.98	-36.30	58.60	35.24
Group - Min	ANN	14.05	-1.55	48.81	-36.34	53.25	76.98	-23.73	55.72	21.97

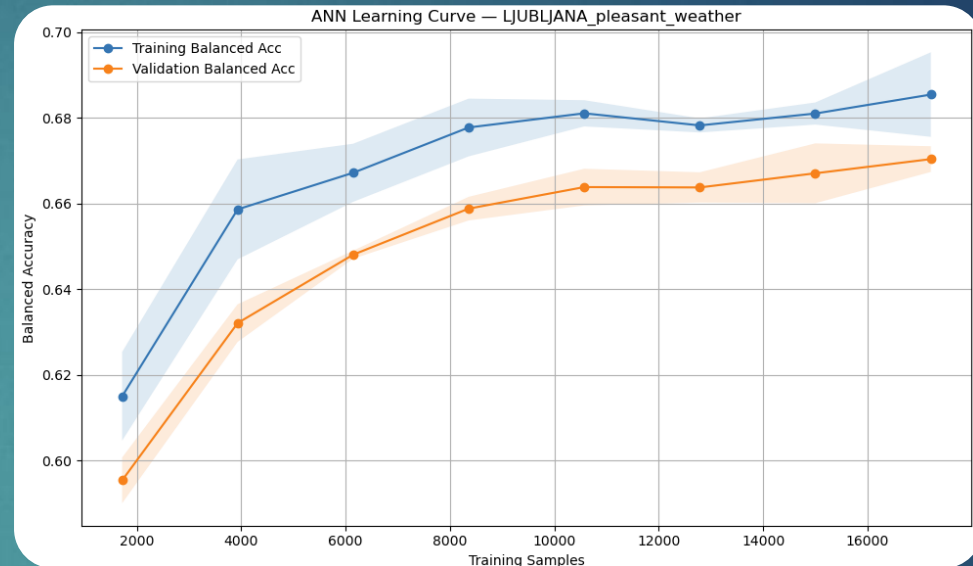
Baseline presents the answers dataset's proportions of Pleasant / Unpleasant weather day outcomes across 60+ years of daily data.
Do you believe these outcomes?

ANN outperforms DT on overall "balance" of various metrics, if evaluation goals include minimizing the gap between Recall and Precision scores. Lift is calculated against Baseline, quantifying model value-add on the relevant metric. ANN Lift for Balanced Accuracy (Test) % in a different table was 14.58%, confirming outperformance against Baseline. Sonnblick (Austria) weather station, which had 100% Unpleasant days and 100% prediction accuracy (both models), was excluded from group evaluation to limit results distortion, although it was included in model training.

Model overfit mitigation: how did the DT and ANN models compare?

Overfitting training data negatively impacts a model's ability to generalize to (unseen) test data.

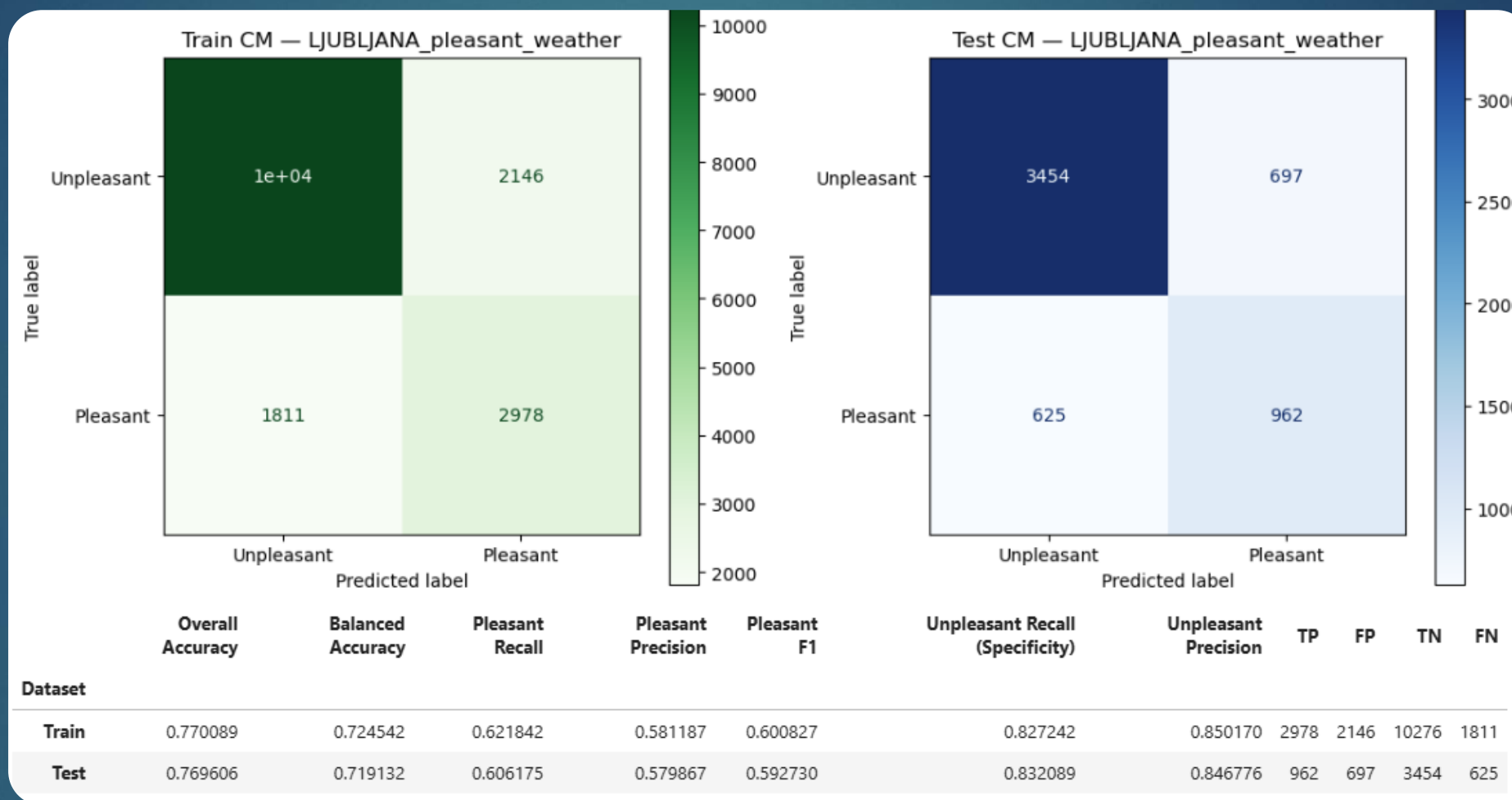
Weather Station	Model	LEARNING CURVE INSIGHTS				Plateau Reached?
		Final Train Score %	Final Validation Score %	Gap %	Curve Shape	
Basel	DT	69.00	65.00	4.00	Converge then moderate gap-parallel	Y
Basel	ANN	68.80	67.30	1.50	Rising curve tapers, small gap - parallel	Y
Oslo	DT	67.80	63.00	4.80	Converge then moderate gap-parallel	Y
Oslo	ANN	69.00	67.60	1.40	Rising curve tapers, small gap - parallel	Y
Stockholm	DT	67.00	63.00	4.00	Converge then moderate gap-parallel	Y
Stockholm	ANN	68.70	67.20	1.50	Rising curve tapers, small gap - parallel	Y
Budapest	DT	68.10	65.20	2.90	Converge then moderate gap-parallel	Y
Budapest	ANN	68.30	66.80	1.50	Rising curve tapers, small gap - parallel	Y
Debilt	DT	68.10	65.10	3.00	Converge then moderate gap-parallel	Y
Debilt	ANN	68.30	67.10	1.20	Rising curve tapers, small gap - parallel	Y
Dusseldorf	DT	67.00	64.00	3.00	Converge to a gap	Y
Dusseldorf	ANN	68.20	67.10	1.10	Rising curve tapers, small gap - parallel	Y
MunchenB	DT	71.00	67.40	3.60	Converge then moderate gap-parallel	Y
MunchenB	ANN	68.50	67.20	1.30	Rising curve tapers, small gap - parallel	Y
Madrid	DT	63.40	60.80	2.60	Converge then moderate gap-parallel	Y
Madrid	ANN	67.80	66.70	1.10	Rising curve tapers, small gap - parallel	Y
Belgrade	DT	67.90	65.10	2.80	Converge to a gap	Y
Belgrade	ANN	68.10	67.10	1.00	Rising curve tapers, small gap - parallel	Y
Ljubljana	DT	70.50	67.40	3.10	Converge then moderate gap-parallel	Y
Ljubljana	ANN	68.70	67.10	1.60	Rising curve tapers, small gap - parallel	Y
Heathrow	DT	63.30	59.90	3.40	Converge then moderate gap-parallel	Y
Heathrow	ANN	68.05	66.90	1.15	Rising curve tapers, small gap - parallel	Y
Maastricht	DT	68.00	65.10	2.90	Converge then moderate gap-parallel	Y
Maastricht	ANN	68.80	67.20	1.60	Rising curve tapers, small gap - parallel	Y
Group - Mean	DT	67.59	64.25	3.34		
Group - Mean	ANN	68.44	67.11	1.33		
Group - Sdev	DT	2.23	2.19	0.62		
Group - Sdev	ANN	0.35	0.23	0.20		
Group - Max	DT	71.00	67.40	4.80		
Group - Max	ANN	69.00	67.60	1.60		
Group - Min	DT	63.30	59.90	2.60		
Group - Min	ANN	67.80	66.70	1.00		



ANN model regularization (L2, dropout) helped significantly reduce overfit on training data. Ljubljana weather station learning curves (above) confirm taming of the overfit tiger for that particular station. The curves confirm model learning took place, revealed through the Balanced Accuracy score (decimals, not %) increase as training sample size grows. The parallel curves end in a modest gap ~ 1.6%, considered a *small* degree of overfit. (Smaller=better)

One of the goals in model fine-tuning and regularization (when needed) is to minimize overfit on training data. The smaller the differential between the train-vs-test (validation) Balanced Accuracy % scores (table above), the lesser the degree of model overfit. ANN model for the win, again: ANN's weather station group mean train-vs-test scores gap of 1.33% bested DT's 3.34%.

ANN model best individual station results: Ljubljana weather station



Scores in decimals, confusion matrices in day counts. Ljubljana weather station had the best overall results using the ANN model for classification prediction. TP = True Positives (predicted Pleasant days that turned out Pleasant), FP = False Positives (predicted Pleasant days that turned out Unpleasant), TN = True Negatives (predicted Unpleasant days that turned out Unpleasant), FN = False Negatives (predicted Unpleasant days that turned out Pleasant). Balanced Accuracy is a better performance measure than Overall Accuracy, which is misleading on imbalanced classes (where the Baseline was ~75% Unpleasant, 25% Pleasant days). Small Train-vs-Test differentials on various performance metrics. Recall is the % of actual outcomes that were predicted correctly for the specified class, while Precision is the % of predictions that were correct for the specified class.

Algorithm Lookback

- Daily weather classification prediction: The ANN model outperformed the DT and KNN models overall, in terms of balancing various performance metrics (reducing differentials between these) and minimizing model overfit on training data. **It pushed the envelope or "ceiling" of obtainable results despite the handicaps of class imbalance, features constraint to temperature-only and a red-flagged answers dataset.**
- Weather data is chaotic (see Edward Lorenz famous 1972 allegory of the "Butterfly Effect") and non-linear. The ANN model is more powerful, subtle and adaptable to complex data patterns and tasks, versus the other models cited. It can handle noise, outliers and imbalanced classes better than the other models examined here.
- One disadvantage is that ANN is difficult to interpret or indecipherable "under the hood", in the hidden layers, versus DT (gini or entropy criterion) or KNN (proximity-based). However, it finds widespread use in the real world, implying that legions of data scientists have found a comfort level despite its "Black Box" aspect.

Algorithms for the three Hypotheses

Hypothesis #1: Unsupervised, non-linear t-SNE or UMAP to find features, then ARCH/GARCH models for extreme weather event forecasting. ARCH/GARCH handle volatility and conditional heteroskedasticity well.

Hypothesis #2: ARCH/GARCH on any weather feature of interest.

Hypothesis #3: ANN.

Summary

Hypothesis #1: Causality of extreme weather events (drought, floods as defined by the EU-backed Copernicus Climate Change Service, C3S) cannot be determined solely from temperature variation.

Methods: t-SNE/UMAP for feature discovery, ARCH/GARCH for extreme weather event forecasting.

Hypothesis #2: Weather forecasting accuracy is inversely correlated with forecast period length.

Methods: ARCH/GARCH on any weather feature of interest.

Hypothesis #3: The Pleasant/Unpleasant weather classification prediction exercise's ANN model-based weather station group Balanced Accuracy (Test) score can be increased > 10% by expanding the model input features beyond temperature only.

Methods: t-SNE/UMAP for feature discovery, ANN for classification prediction.

Next Steps


- Explore dimensionality reduction techniques and more algorithms (unsupervised, supervised), including multivariate forecasting models and Random Forests.
- Assess European weather domain experts' definitions of extreme weather events and available datasets.
- If available and sufficiently robust: add warmer-latitude weather stations to the weather dataset mix for spatial diversification. Consider omitting Sonnblick for binary classification prediction training purposes.
- Make more predictions, evaluate results, find insights.
- Identify the Pleasant/Unpleasant weather answers dataset provenance, methodology if any. Evaluate accuracy if possible.

Thank You

GitHub repository for the Python scripts and datasets used for this project:



Contact

 ddav16psy@gmail.com