ClimateWins Weather Predictions & Climate Change

Introduction:

ClimateWins is a fictional European nonprofit organization, that is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world. It's concerned with extreme weather events, especially in the past 10-20 years. Through use of machine learning, it wants to see if weather conditions can be predicted by looking historically at the temperature highs and lows, and exploring whether conditions can be predicted to a specific given day and can prevent danger.

Goal:

Utilize machine learning (both supervised and unsupervised) algorithms to make weather predictions the consequences of climate change around Europe and, potentially the world.

Hypothesis:

- ClimateWins can help predict climate change around Europe (and potentially, around the world).
- The weather climate across Europe will gradually increase over time.
- Supervised and Unsupervised Deep Learning algorithms are are optimal tools in in predictive analysis necessary for weather forecasting.

Steps and Skills:

- Data Cleaning
- Gradient Descent Optimization
- Pairplots

- **Supervised Learning Algorithms**
  - K-Nearest Neighbor (KNN)
  - Artificial Neural Network (ANN)
  - Decision Tree

- **Unsupervised Learning Algorithms:**
  - Random Forest
  - K-Means Clustering
  - Dendrograms
  - Dimensionality Reduction
  - Convolution Neural Network (CNN)
  - Recurrent Neural Networks (RNN)
  - Generative Adversarial Networks (GANs)

Tools Used:

jupyter GitHub

ClimateWins Weather Predictions & Climate Change

<u>Optimization</u> lowers the risk of error and improves the accuracy of a model, often used to determine which algorithms to use.
It helps understand valleys and peaks of the local/global landscape of the data.

- **Gradient Descent** (used in both linear and nonlinear data) was used in this study to determine the local minimums and maximums of the data points.
  - Three iterations performed, adjusting step lengths (alpha) in order to get a result as near to 0 as possible

Is Climate Increasing?

- Belgrade has freezing minimum temps getting colder over past 20 years. It has warmed up by about 5 degrees over past 60 years, when looking at the mean per year. The max mean raised 1 degree higher than 60 years ago).
- In general, all means, mins, and max temperatures have increased, with the exception of Valencia (where data is likely skewed due to permanency of 10.7 report).

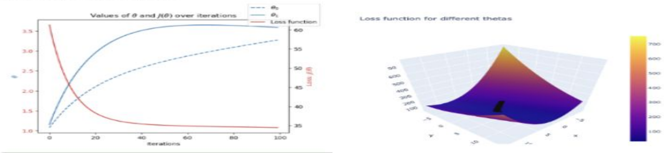
This chart shows data for approximately a 60 years span of temperatures in Madrid, Valentia, and Belgrade in the years: 1980, 2000, & 2018

Weather Station	Year	Theta0		Theta1		Iterations	Step size	Mean Temp	Max Mean	Min Mean
		Start	End	Start	End					
MADRID	1980	1	0	1	0	100	.01	14.19 (57.5F)	30.4 (87F)	-.2 (31F)
MADRID	2000	1	-1	1	0	100	.01	15 (59F)	29.4 9 (84F)	.6 (33F)
MADRID	2018	1	0	1	0	100	.01	15.57 (60F)	32.9 (91F)	1.6 (35F)
VALENTIA	1980	1	0	1	0	100	.01	10.35 (50.6F)	17.9 (64F)	1.5 (34F)
VALENTIA	2000	1	0	1	0	100	.02	10.76 (51F)	19.5 (67F)	1.4 (34.5F)
VALENTIA	2018	1	0	1	0	100	.01	10.7 (51F)	10.7 (51F)	10.7 (51F)
BELGRADE	1980	1	0	1	0	100	.01	10.77 (51F)	27.4 (81F)	-9.3 (15F)
BELGRADE	2000	1	1	1	1	100	.03	14.19 (57F)	32.8 (91)	-9.9 (14F)
BELGRADE	2018	1	.5	1	1	100	.02	14.94 (57.5F)	28.4 (83F)	-5.8 (21.5F)

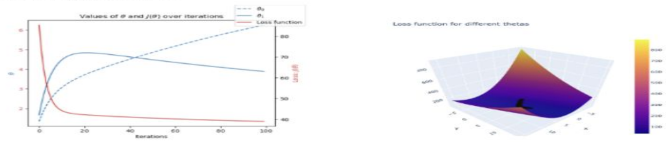
Belgrade:

3 iterations
showing
loss of function
and *loss profile*

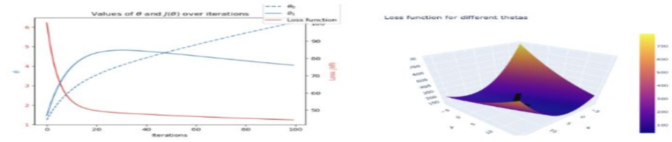
BELGRADE 1980



BELGRADE 2000

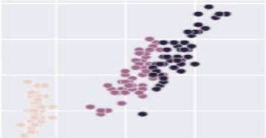


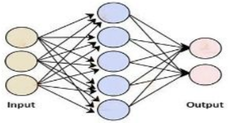
BELGRADE 2018

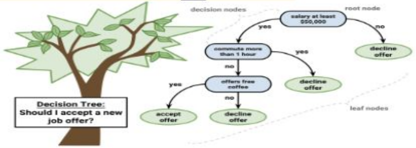


ClimateWins Weather Predictions & Climate Change

Supervised Machine Learning



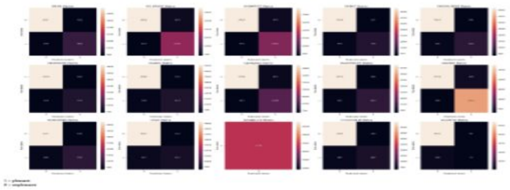




K-Nearest Neighbor (KNN):

classifies data on its proximity to its neighbors

Test Accuracy Scores = 88.46%



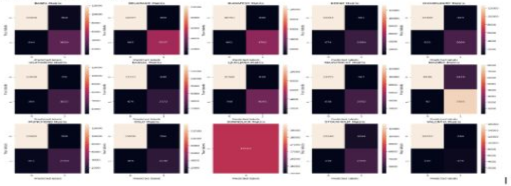
Weather Station	Accurate Predications		False Positive	False Negative	Accuracy:
	0 (unpleasant)	1 (pleasant)			
Basal	3917	961	421	439	4878/5738 – 85%
Belgrade	3252	1544	524	418	4796/5738 – 84%
Budapest	3424	1462	476	376	4886/5738 – 85%
Debilt	4320	723	317	378	5043/5738 – 88%
Desseldorf	4164	810	343	421	4974/5738 – 87%
Heathrow	4138	744	432	424	4882/5738 – 85%
Kassel	4563	614	252	309	5177/5738 – 90%
Ljubljana	3740	1180	455	363	4920/5738 – 86%
Maastricht	4253	824	309	352	5077/5738 – 88%
Madrid	2750	2261	418	309	5011/5738 – 87%
Munchenb	4237	792	309	400	5029/5738 – 88%
Oslo	4637	512	242	347	5149/5738 – 90%
Sonnblick	5738	0	0	0	5738/5738 – 100%
Stockholm	4483	607	283	365	5090/5738 – 89%
Valentia	5404	74	58	202	5478/5738 – 95%

Artificial Neural Network (ANN):

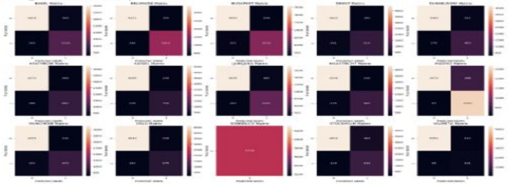
Replicates the human brain, consisting of input and output layers, along with hidden layers by adjusting weights to obtain outcomes

Train accuracy score = 52%
Test Accuracy Score = 49%

Training: 52% accuracy



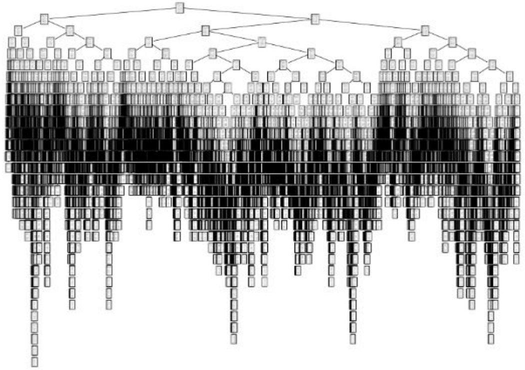
Testing: 49% accuracy



Decision Tree

Decision trees model these kinds of questions for many objects at once. They have roots, branches, and leaves. The *root* is the first question asked, with yes/no answer. Leading to another question/ *branch*. The answer is the *leaf*/stopping point.

Train accuracy score = 46%,
Test Accuracy Score = 47%



ClimateWins Weather Predictions & Climate Change

Algorithm Overview: Machine Learning
(Supervised/Unsupervised)

Random Forest:
(Clustering with Dendrograms)

Identify extreme weather conditions:

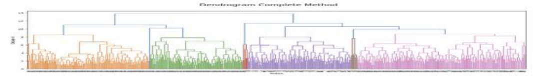
Compare all stations with Complete method (reduced data) to find trends, then narrow down specific locations to find the extreme weather condition patterns using Single method.

Combines multiple decision trees to make the most accurate predictions. Each decision tree trains on a random sample of the total data, with final prediction made by averaging the predictions of all trees.

-----------------Dendrograms comparing Madrid and Belgrade in 2010-----------------

Single method: this looks at the two closet members of each cluster, showing that most days in 2010 are clumped into one category, represented by the orange. This is not providing any meaningful insights. I am guessing that the small group to the left were classified outliers?

Dendrogram Single Method can find outliers, interpreted as extreme weather conditions *when looking at specific weather stations*. It is not as useful when comparing all weather stations



Complete method: This is looking at the distance between the farthest members of each cluster, broken down more evenly into 6 categories, a more even distribution for the year. This may be interesting to look at to gain insight from patterns in these clusters.

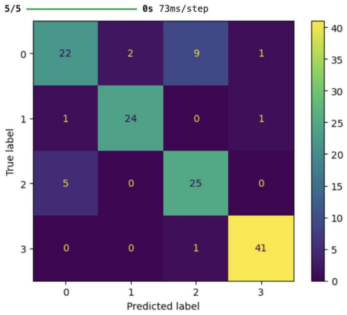
Dendrogram Complete Method gives clear, distinct clusters with less noise to find key patterns in the data (all weather stations per year)

Deep Learning with CNN & RNN:
(Convolutional Neural Network & Recurrent Neural Network)

Determine if unusual weather patterns are increasing:

Use **CNNs** and **GAN**s to analyze spatial data to find trends across these regions

CNNs handle **images & numerical** data as they were inspired by visual cortex processes of the brain. RNNs handle temporal data such as **text, handwriting, and speech**. RNNs also use LSTM (long short-term memory), imitating a brain that forgets unimportant data – updating the significant data as needed)



CNN model Confusion Matrix
showing 4 classes: cloudy, rain, sunshine and sunrise
91% Accuracy

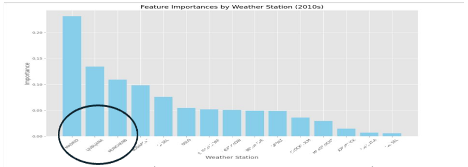
Incorrect prediction:
Class cloudy, predicted shine

GANs:
(Generative Adversarial Network)

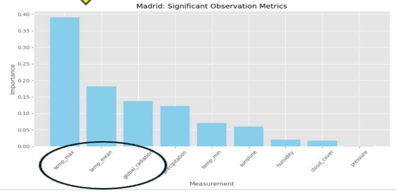
Determine safest places to live in the future:

Use **Random Forest / Decision Trees (for 2010s)** to use predicted weather conditions to determine regional safety. Use the identified features within top locations to gauge future weather predictions.

Use two adversarial neural networks working against each other – generator (creates data) & Discriminator (samples of real & artificial, discriminating which is real). Can create artificial data (text & images)



First, narrow down top locations.
Then, look at their top features.



ClimateWins Weather Predictions & Climate Change

Insights:

- Supervised and Unsupervised Machine Learning models can make accurate weather predictions
    - his study has shown slight temperature increase over time, and has potential to generalize to the rest of the world
- **KNN (**K-Nearest Neighbor) **model was the best choice** for unsupervised method (88% accuracy, much higher when compared with the other two).
    - **ANN (Artificial Neural Network) has potential** to produce a more complete depiction of the data, presuming adjustments are made when eliciting this model as long as there is a good understanding of the model and how to manipulate it for best outcomes.
    - **Decision Tree** was not useful as the data was too deep and complex for a meaningful insight, nor was **RNN** due to its low accuracy rate and lack of insightful information. These algorithm may run better on sections of data.
- **Random Forest Algorithms** are highly accurate and provide useful breakdown of data to pinpoint weather data features and top locations
    - Dendrograms are useful in finding outliers or extreme weather events (single) and for finding season trends year round (complete).

Recommendations:

- **Prune Decision Tree** data to reach a better conclusion and avoid overfitting
- **Run more iterations on ANN model** to find more options for training the model to find more optimal variables.
    - Rule out if Sonnblick's 100% accuracy was due to overfitting or data error
- Make adjustments within the weather data
    - **Further clean/check for error**s: as in the case of Valencia showing no variability in min/max/mean temps for the year chosen in this analysis
    - Incorporate more complete data: **include more descriptions features that encompass 'pleasant' vs. 'unpleasant'**
    - **Separate locations into hot/cold regions** to eliminate cultural bias, particularly in perception of what is pleasant vs. unpleasant
- **Combine supervised and unsupervised algorithms** for best results.
    - Use **Random Forest to narrow down** the data features (either by location or year). **Create subsets** of data on these important variables. Then, **run those through a CNN** to produce higher accuracy. Then, use final results to **plug into a GAN** to generate more realistic artificial results.
- **Increase sample size to other parts of the world** besides Europe and run the algorithms again to compare and see if these methods can be used as a generalization for world data.