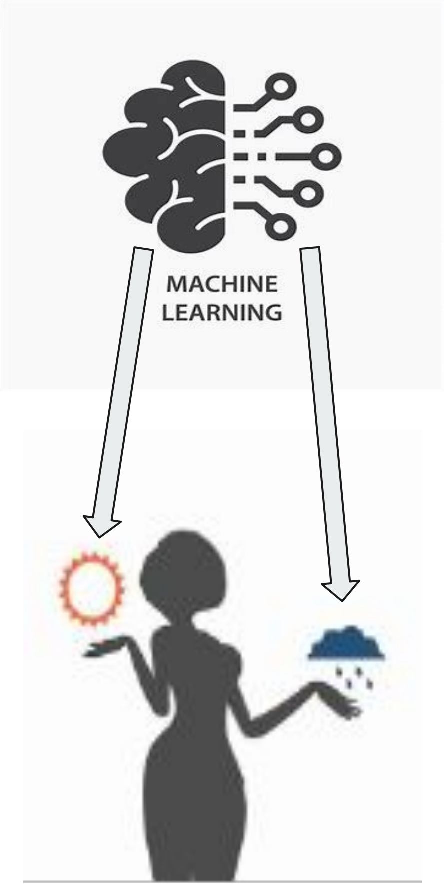
ClimateWins Weather Predictions & Climate Change

Nancy Kolaski
September, 2024



Objective:

ClimateWins is a European nonprofit organization, that is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world.

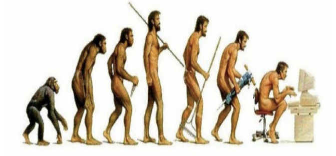
Hypothesis:

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

DATA: Data was collected between 1800s to 2022 by ‘European Climate Assessment and Data Set Project’ found at https://www.ecad.eu/

Bias

- **Location Bias:**
  - Data utilized for this analysis was partially collected from extreme mountain top weather.
  - Much of the weather has been mostly categorized as ‘unpleasant’, or less than ideal for a picnic. This dataset is unique in some it its location, and has little chance of generalizing to its intended location of Europe, and even less chance of generalizing to the rest of world



- **Temporal Bias:**
  - Data was collected across a vast time span (early 1800s to 2022).
  - The collection methods used prior to the technological advancements that exist now are prone to error and would not be as complete as the data collected in more recent years.

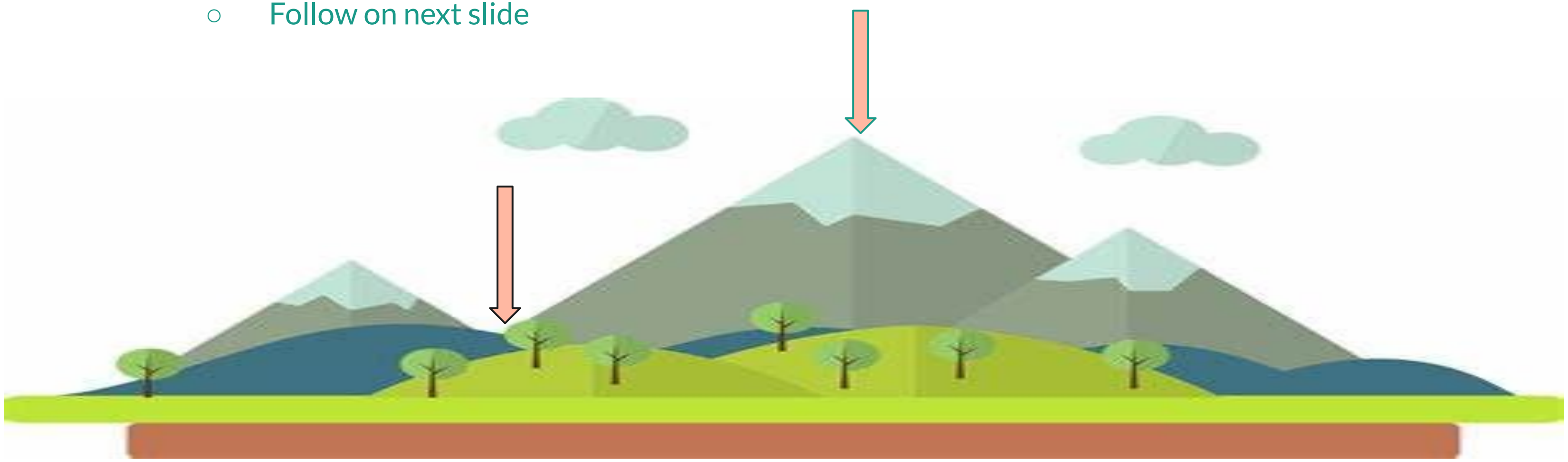
- **Regional or Cultural Bias:**
  - Each region has had a history of exposure to certain climate events. They may have less readiness or action when compared to another. Interpretations may vary from region to region. What is considered hot weather would not tolerate cold and rainy temps. This skews any classification of pleasant vs. unpleasant, as we have in this dataset.

- **Small Sample Size:**
  - Out of the 23,755 weather stations in Europe, this analysis only studied 18 weather stations.

What is optimization & what did it reveal about temperatures over the past 60 years?

Optimization 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
    - Follow on next slide



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).

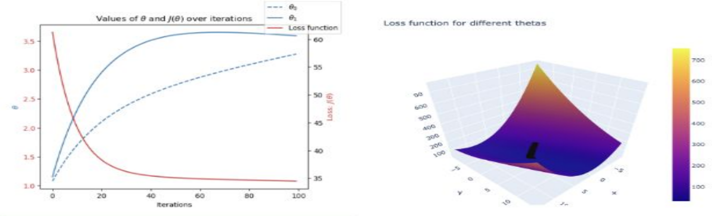
This chart shows data for approximately a 60 years span of temperatures in Madrid, Valentia, and Belgrade in the years: 1980, 2000, & 2018

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)

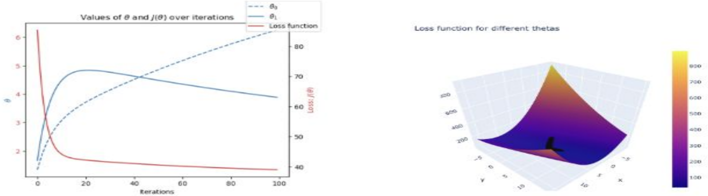
Belgrade:

3 iterations
showing
loss of function
and loss profile

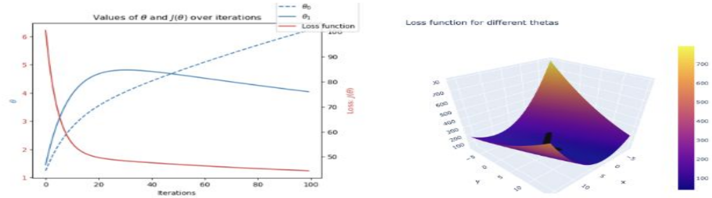
BELGRADE 1980



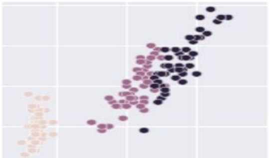
BELGRADE 2000

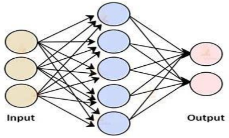


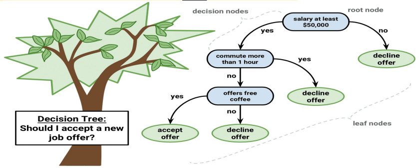
BELGRADE 2018



Supervised Machine Learning



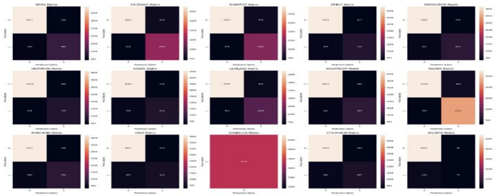




K-Nearest Neighbor (KNN):

classifies data on its proximity to its neighbors

Test Accuracy Scores = 88.46%



1 = pleasant
0 = unpleasant

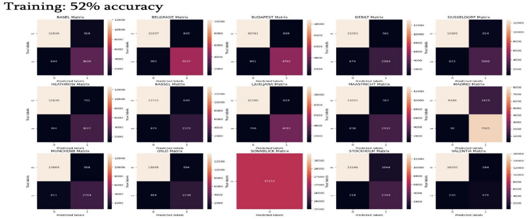
Weather Station	Accurate Predications 0 (unpleasant)	Accurate Predications 1 (pleasant)	False Positive	False Negative	Accuracy: (Accurate Predictions/Total Values)
Basal	3917	961	421	439	4878/5738 = **85%**
Belgrade	3252	1544	524	418	4796/5738 = **84%**
Budapest	3424	1462	476	376	4886/5738 = **85%**
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):

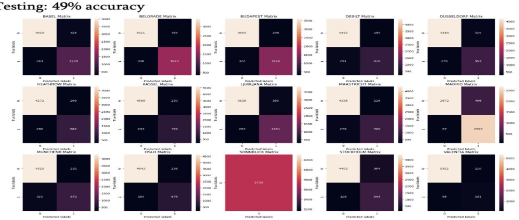
Replicates the human brain, consisting of input and output layers, along with hidden layers by adjusting weights to obtain outcomes

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

Training: 52% accuracy



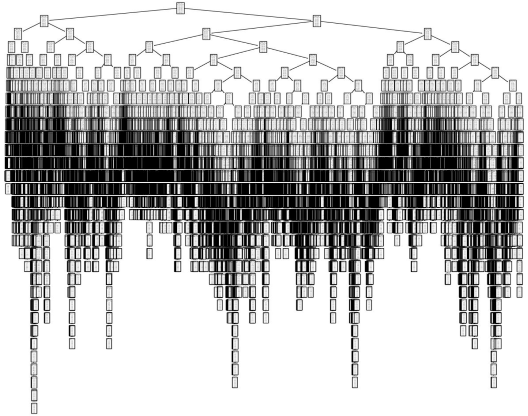
Testing: 49% accuracy



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%



Conclusion:

- **Supervised Machine Learning models can make accurate weather predictions**
  - This study has already shown slight increase over time, and has potential to generalize to the rest of the world
- **KNN (K-Nearest Neighbor) model was the best choice** for this study so far (88% accuracy, much higher when compared with the other two).
  - **ANN (Artifical 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

Next Steps:

1. Prune Decision Tree data to reach a better conclusion and avoid overfitting
2. Run more iterations on ANN model to find more options for training the model to find more optimal variables.
    - Rule out of Sonnblick's 100% accuracy was due to overfitting or data error
3. Make adjustments within the data
    - cleaning/checking for errors: as in the case of Valencia showing no variability in min/max/mean temps for the year chosen in this analysis
    - incorporating more complete data: include more descriptions of data when categorizing into 'pleasant' vs. 'unpleasant'
    - Separate locations into hot/cold regions to eliminate cultural bias, particularly in perception of what is pleasant vs. unpleasant
4. Increase sample size and run the algorithms again to compare
    - more appropriate for generalizing to other locations and ultimately, the world

Thank you for your time and consideration. Questions?

Nancy Kolaski

Nancykolaski@gmail.com

Github.com/Nancy-Kolaski