Classifying severity of drought using meteorological data and ML models

Ayush Sharma
ayush20042@iiitd.ac.in

Gautam Reddy
gautam20445@iiitd.ac.in

Ujjwal Rastogi
ujjwal20546@iiitd.ac.in

Yash Agrawal
yash20551@iiitd.ac.in

1. Motivation

Since there has been an increase in concern about climate change over the past few decades, we have seen several dramatic events like the Australian Bushfires of 2020 [1] that have entirely changed the terrain. It makes sense to anticipate that as global temperatures rise, so does the likelihood of drought in many places of the world. Therefore, we questioned whether we would be able to accurately forecast such a catastrophe in the future and take the necessary precautions to prevent or deal with the calamity.

2. Problem Statement

Given the meteorological and climatic data for the past years of many locations spanning all over the United States, we are trying to predict the possibility of drought in future in a particular location. Moreover, we need to classify droughts on the basis of their severity and learn how climatic conditions such as rainfall, humidity, etc., along with geographical location and features such as latitude, longitude, temperature, etc., affect the possibility and severity of drought upcoming.

	Lisbon				Munich			
	Train		Test		Train		Test	
λ	Acc	F1	Acc	F1	Acc	F1	Acc	F1
0	0.961	0.861	0.733	0.206	0.959	0.865	0.787	0.176
0.1	0.495	0.233	0.373	0.294	0.506	0.241	0.536	0.215
0.01	0.517	0.245	0.460	0.269	0.519	0.268	0.431	0.275
0.001	0.572	0.261	0.540	0.288	0.490	0.288	0.563	0.266
0.0001	0.765	0.472	0.627	0.259	0.823	0.557	0.719	0.189

	Lisbon				Munich			
Loss-Function	Acc nd	Acc d	Acc	F1	Acc nd	Acc d	Acc	F1
mean absolute error	0.511	0.516	0.540	0.288	0.500	0.582	0.512	0.276
mean squared error	0.440	0.655	0.479	0.312	0.562	0.509	0.553	0.267
binary crossentropy	0.436	0.610	0.467	0.292	0.589	0.440	0.565	0.245
hinge	0.229	0.753	0.323	0.287	0.568	0.486	0.555	0.259
squared hinge	0.486	0.501	0.489	0.261	1.000	0.000	0.840	0.000

		Lisbon				Munich			
Neurons	Architecture	Acc nd	Acc d	Acc	F1	Acc nd	Acc d	Acc	F1
De*Dr*De*Dr*De*De*De	4000*0.5*1000*0.5*500*100*5	0.511	0.516	0.540	0.288	0.562	0.509	0.553	0.267
De*Dr*De*Dr*De*De*De	5000*0.5*1000*0.5*500*100*5	0.581	0.496	0.566	0.292	0.378	0.693	0.428	0.279
De*Dr*De*Dr*De*De*De	5000*0.5*4000*0.5*500*100*5	0.457	0.602	0.483	0.296	0.725	0.338	0.663	0.243
De*Dr*De*Dr*De*De*De	5000*0.5*4000*0.5*1000*100*5	0.570	0.501	0.558	0.290	0.527	0.514	0.525	0.257
De*Dr*De*Dr*De*De*De	5000*0.5*4000*0.5*1000*500*5	0.402	0.635	0.444	0.292	0.683	0.409	0.640	0.266
De*Dr*De*Dr*De*De*De	5000*0.5*4000*0.5*1000*500*100	0.575	0.526	0.566	0.305	0.420	0.619	0.452	0.266

3. Literature Survey

3.1 Applying machine learning for drought prediction using data from a large ensemble of climate simulations

This paper uses meteorological data to predict droughts with a lead time of 1 month using ANN's. They use the meteorological data from the ClimEx Project (Leduc et al,2019) which consists of data for North America and Europe. They are solving a binary classification problem using Standardized Precipitation Index (SPI) as their output class. The paper goes into greater detail regarding where the data came from with specific sources for some of the features that they have used. They explain the reasoning behind how an ANN works and the metrics that they have used to evaluate their models. They ran their model in two different climate regions of Lisbon and Munich The authors have used L2 regularization in their ANN and investigated the effect of different loss functions and architectures, the results of which have been summarized below

3.2 Forecasting standardized precipitation index using data intelligence models: regional investigation of Bangladesh

While the previous paper focused on ANN as the classification model, this paper instead focuses on models such as Random Forests, minimum probability machine regression (MPMR), M5 Tree (M5tree), extreme learning machine (ELM) and online sequential-ELM (OSELM). They too use the SPI as the drought indicator like the previous paper. The paper goes into detail of how all these models work as well as the metrics that they have used to evaluate these models. the metrics they have used are for every base station as the metric and according to them. The SPI was calculated for 1,3,6 and 12 months and the results were reported for each case. According to the authors RF was the best for SP1, and ELM was the best for SP3, SP6 and SP12.

4. Dataset

- The dataset comprises of 18 weather indicators along with 29 soil indicators.
- We have split the dataset into Train-Test-Validation splits as per the following information:

Train	2000-2016	**80%**	**19.3 million**
Validation	2017-2018	**10%**	**2.27 million**
Test	2019-2020	**10%**	**2.27 million**

Meteorological Indicators

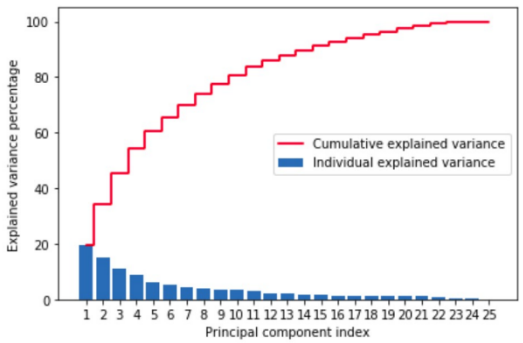
Indicator	Description
WS10M_MIN	Minimum Wind Speed at 10 Meters (m/s)
QV2M	Specific Humidity at 2 Meters (g/kg)
T2M_RANGE	Temperature Range at 2 Meters (C)
WS10M	Wind Speed at 10 Meters (m/s)
T2M	Temperature at 2 Meters (C)
WS50M_MIN	Minimum Wind Speed at 50 Meters (m/s)
T2M_MAX	Maximum Temperature at 2 Meters (C)
WS50M	Wind Speed at 50 Meters (m/s)
TS	Earth Skin Temperature (C)
WS50M_RANG	Wind Speed Range at 50 Meters (m/s)
WS50M_MAX	Maximum Wind Speed at 50 Meters (m/s)
WS10M_MAX	Maximum Wind Speed at 10 Meters (m/s)
WS10M_RANG	Wind Speed Range at 10 Meters (m/s)
PS	Surface Pressure (kPa)
T2MDEW	Dew/Frost Point at 2 Meters (C)
T2M_MIN	Minimum Temperature at 2 Meters (C)
T2MWET	Wet Bulb Temperature at 2 Meters (C)
PRECTOT	Precipitation (mm day-1)

This dataset classifies six stages of drought, ranging from 0 (no drought) to 5. (Exceptional drought). Each entry includes the last 90 days' worth of data from 18 weather indicators and represents the drought level at a certain time point in a specific US region. The target variable is score, for which only weekly data is available.

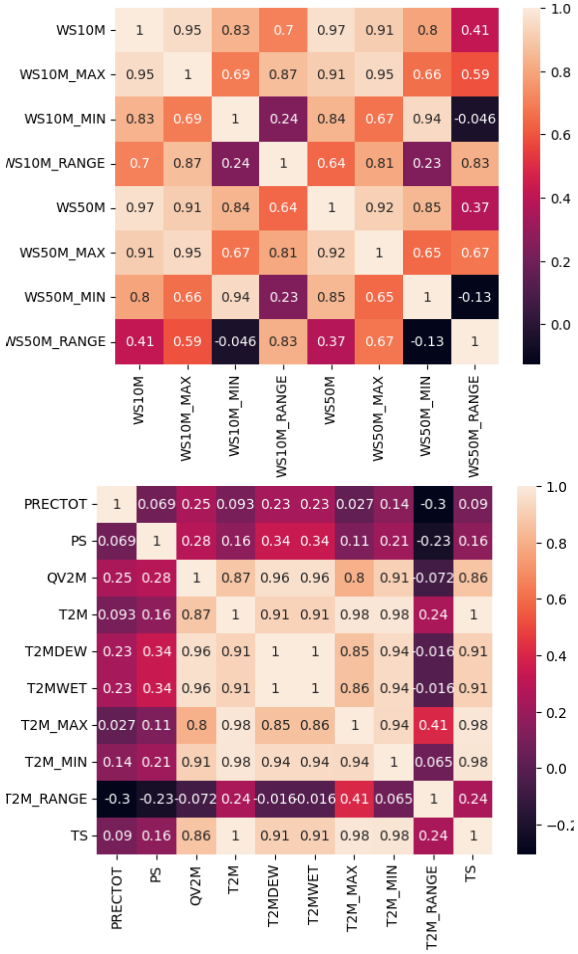
D0	Abnormally Dry
D1	Moderate Drought
D2	Severe Drought
D3	Extreme Drought
D4	Exceptional Drought

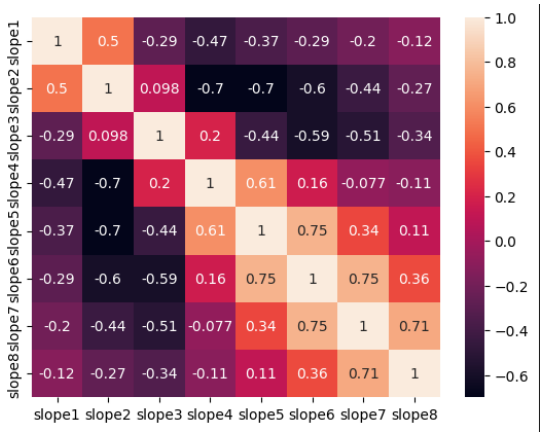
→ Data Pre-Processing Techniques

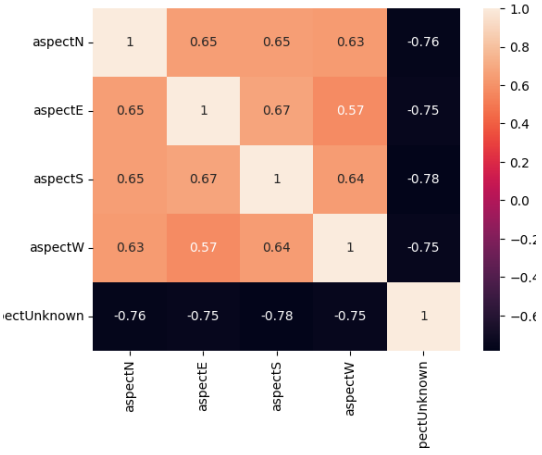
- Merged two different datasets (meteorological data and soil data)
- Left joined soil data on meteorological table based on FIPS id
- Removed all the null value entries
- Reduced the dataset based on date to gauge how older data would affect predictions
- Applied PCA to further reduce the size of the dataset

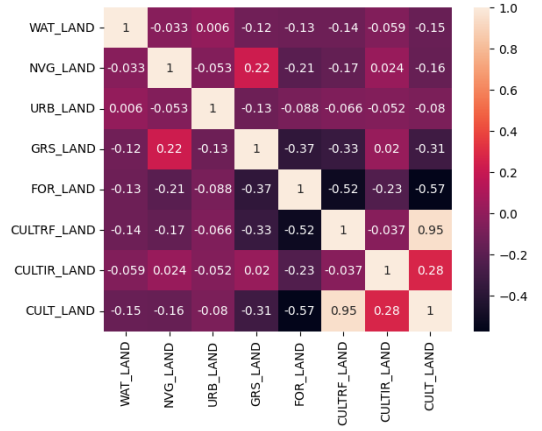


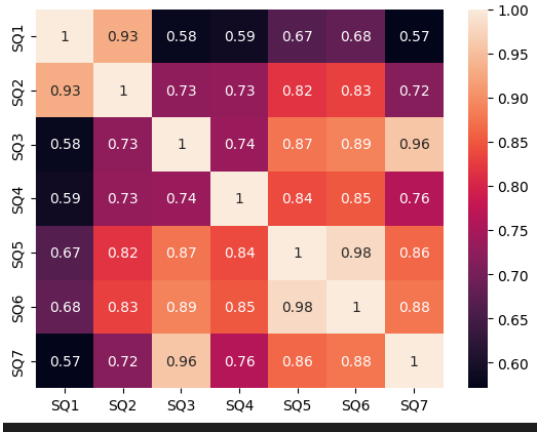
→ Data Visualisation









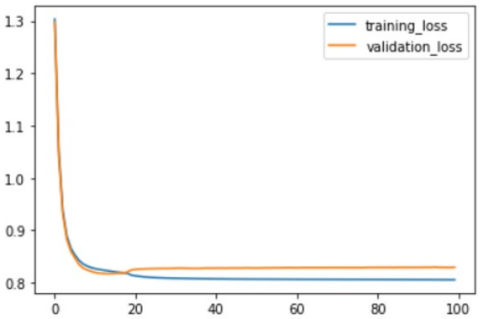


5. Methodology and Model Details

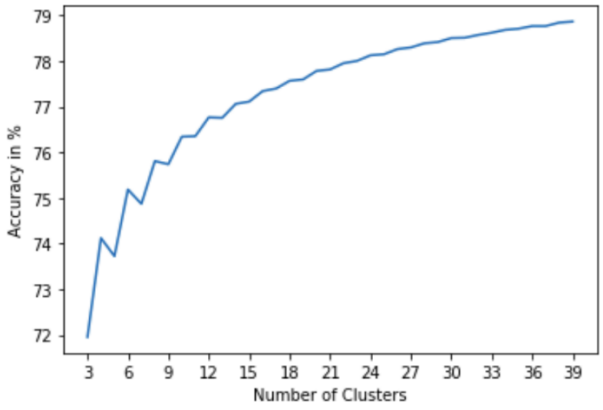
- As this is a multi-class classification problem, we applied Logistic Regression and Random Forests at the beginning
- To fit the target classification categories with actual decimal values, data had to be rounded down or floored.
- Feature selection must be done carefully for better results. As a starting point, 44 features were taken out of 53 for training.
- After unsatisfactory results, feature selection was done using domain knowledge and correlation heatmap. 25 features were selected.

Models Used:

- **Logistic regression with OVR classification:**
  - Since Logistic Regression is one of the simplest classification models, we began our analysis with this model
  - Tuned hyper parameters like max_iters, loss_penalty and class_weights
- **Random Forest Classification:**
  - Ensembled several decision trees to build a classification model.
  - Hyper Parameters tuned: n_bins, max_depth, n_estimators, max_samples
  - Tuning n_bins can improve the performance on dataset.
- **Naïve Bayes Classifier:**
  - Gaussian Naïve Bayes Classifier was used since that was the only model that fits our dataset.
  - Since all our features are continuous in nature there are no categorical features.
- **SVM Classifier:**
  - We used the linear kernel to train our model since a SVM model using the rbf or polynomial kernel does not scale well to larger datasets.
  - There is a tradeoff here between training times and accuracy because large dataset is unlikely to be linearly separable.
  - The hyper parameter that was mainly tuned was tolerance.
- **MLP Classifier:**
  - We tried out various MLP architectures and activation functions.
  - The hidden layer sizes that were used are (10,5), (10,5,3) and (6,4,2).



- **K Neighbor Classification:**
  - K-Neighbor Classifier is sklearn implementation of k-nearest neighbors.
  - We tried this algorithm for various values of k from 3 to 40.
  - As we increased the value of k, our model accuracy increased as shown in the plot given below:



6. Results & Analysis

We present here the results of all the models with their best parameters that were found after hyper-parameter tuning on the test dataset.

Model	Accuracy	Recall	Precision	f1-score
Logistic Regression	0.73	0.73	0.56	0.62
SVM	0.66	0.66	0.59	0.62
Random Forest	0.86	0.87	0.89	0.86
Naive Bayes	0.73	0.73	0.62	0.63
MLP	0.74	0.74	0.55	0.63
K Neighbor Classification	0.77	0.77	0.72	0.71

Model	Accuracy	Recall	Precision	F-Score
Logistic Regression	0.81	0.67	0.81	0.73
Random Forest	0.79	0.71	0.76	0.73
Naïve Bayes Classifier	0.81	0.81	0.71	0.72
SVM	0.72	0.71	0.72	0.71
MLP	0.82	0.82	0.71	0.74
K Neighbour	0.78	0.78	0.71	0.73

Based off the results that we had gotten previously, we hypothesized that perhaps soil and meteorological data that is much older than our test dataset will not be a good indicator for drought prediction because of the various climatic changes that have taken place over two decades. The results that we have shown here confirmed our hypothesis where we noticed that increasing the dataset to include older dates caused the accuracy scores to fall.

SVM was the worst performing model for our problem and Random Forest was the best performing model on the train set and one of the best on the test set. In the end we found that Random Forest classifier and MLP are the best performing models which was shown to be the case in literature as well

7. Conclusion

7.1 Learning

Through this project we learned that feature selection and properly doing EDA is a key component of building good ML models. We learned how to handle big datasets and how to work with them. We learned how the input data being skewed can cause models to not train properly. We also learned how to go about doing hyperparameter tuning. We found out how time-series data can affect predictions and model learning

7.2 Contribution

All of us have contributed equally to the entire project done until now. Almost all the work was done in meetings together, helping each other out and working as a team.

- Gautam: Domain Knowledge, Random Forest, Report and PPT making, EDA, Result Analysis, SVM
- Yash: Domain Knowledge, Logistic Regression, Report and PPT making, Naïve Bayes
- Ayush: Data Preprocessing, Random Forest, Report and PPT making, Data Visualisation, SVM.
- Ujjwal: Data Preprocessing, Logistic Regression, Report and PPT making, EDA, MLP.

References

[1] Trnka, M, Hlavinka, P, Možný, M, *et al.* Czech Drought Monitor System for monitoring and forecasting agricultural drought and drought impacts. *Int J Climatol*. 2020; 40: 5941– 5958. https://doi.org/10.1002/joc.6557

[2] Belayneh, Anteneh & Adamowski, Jan. (2013). Drought forecasting using new machine learning methods. Journal of Water and Land Development. 18. 3-12. 10.2478/jwld-2013.

[3] Aishwarya M Iyengar ,Deepika K ,Kanthi Utkarsha Bharat ,Mitaigar Divya ,Vaidehi M , (2019) " Drought Prediction using Machine Learning Algorithm " , *International Journal of Advances in Computer Science and Cloud Computing (IJACSCC)* , pp. 1-6, Volume-7,Issue-1

[4] Yaseen, Z.M., Ali, M., Sharafati, A. et al. Forecasting standardized precipitation index using data intelligence models: regional investigation of Bangladesh. Sci Rep 11, 3435 (2021). https://doi.org/10.1038/s41598-021-82977-9

[5] Felsche, E., Ludwig, R., 2021. Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations. NHESS - Applying machine learning for drought prediction in a perfect model framework using data from a large ensemble of climate simulations. https://doi.org/https://doi.org/10.5194/nhess-21-3679-2021