Forecasting Drought Area Percentage in California using Machine Learning Algorithms

Rama Krishna Poluru, Dharma Teja Kolluri, Saranya Gondeli and Manish Kumar Sesetti

Department of Applied Data Science, San Jose State University

DATA 270 : Data Analyst Process

Dr. Eduardo Chan

February 17, 2023

1. Proposed Research Problem

1.1 ***Background:-*** According to the U.S. Drought Monitor, it is estimated that there has been a water shortage of 30% in 2022 compared to 2021 in California state. It is said to have state reservoirs standing at a capacity of less than 55% by September 2022 which is the first time that happened in the past 35 years. In September 2022, 74.4% of areas are in exceptional to severe drought levels which had a consequential effect on the state's economy, and agriculture. Furthermore, it led to a raise in water bills for the customers' daily use and drinking. Therefore, we propose to develop a machine learning model to predict the drought area percentage in California using historical weather data, which can help in better planning and management of events in the state for the areas where it is going to be more drought-prone.

1.2 ***Problem Statement:-*** In California, the major environmental issue is the drought, by far it has a lot of effect on agriculture, water supply, and the state economy. It is quite a challenge for government officials to handle water management in drought-prone areas. The change in climatic conditions like lack of rainfall impacts to frequency and severity of droughts, and it is crucial to implement accurate methods for predicting drought-prone area percentage for the California region. This project aims to develop a reliable and accurate machine learning model to predict drought-prone areas percentage for the California region using historical weather data. The Standardized Precipitation Evapotranspiration Index (SPEI), vegetation index, and soil moisture index are used as indicators to predict the drought-prone areas percentage for the California region.

1.3 ***Methodology:-*** To solve our problem, we will start by taking out the important information from the satellite pictures we collected. We will then look at different sources of data to see which things make droughts worse in California. After that, we will divide the data into two groups, one to train with and one to test with. This will help us find out if our predictions are correct. We will use four different machine learning techniques that are good for time series data. After researching, we found that the Decision tree model is easy to understand and can help us learn from the data. We also found that using the Random Forest model helps us avoid overfitting. Artificial Neural Networks (ANN) are also good for time series data because they can capture how the data changes over time. Lastly, we found that Long Short-Term Memory (LSTM) is good for time series data because it can remember short-term trends. So we chose to do these models for our project. After that we train these models with the train data and then, we will test the accuracy of these models using some other data. By comparing their accuracy, we can decide which model is most helpful in predicting the amount of drought in California in different climate conditions. With this knowledge, we can offer useful suggestions for managing drought in the state, reducing its negative impact on people and the economy.

1.4 ***Conclusion:-*** In conclusion, the severe drought conditions in California have led to a negative impact on agriculture, the economy, and the water supply. To manage the situation more efficiently, accurate predictions of the percentage of land that is under drought-prone state is needed. A machine learning model is proposed that uses historical weather data to forecast the percentage of California's land that will be in a drought-prone state. This project uses three datasets from public sources, including the SPEI, NDVI, and SMI values, to train the model accurately. By using these indicators, we can provide information to drought early warning

systems, assess the impacts of drought conditions, and better plan and manage drought events in the state.

1. Background

2.1 Research

2.1.1 ***Motivation:–*** California has been experiencing very dry weather, which we call a drought, and this has caused a lot of problems. It has made it difficult for farmers to grow crops, led to a shortage of water for people to use, and even hurt the economy. Droughts can also increase the chances of big fires that can harm homes and the environment. It is important to be able to predict when droughts might happen so we can be ready and lessen their impact on everyone. In California, droughts can last for many years and cause different issues like dry soil, no water, and harm to plants and animals.

California did not get much rain in 2020-2021, which made the drought very bad. The drought made the water level in Lake Mead, where people get their water, very low. To fix this, California is trying to save water, use it in a smarter way, and make more water by removing salt from ocean water and reusing water. So we can use computers to predict and manage droughts, and warn people so they can prepare for them.

2.1.2 Relevant work reference

2.1.2.a ***Paper title:*** *Drought prediction based on SPI and SPEI with varying timescales using LSTM recurrent neural network*

S. Poornima and M. Pushpalatha, Soft Comput., vol. 23, no. 18, pp. 8399-8412, Sep. 2019

How was the specific research problem tackled or resolved by other research papers? What are the practical implications, and how were they beneficial?

The paper focuses on the use of meteorological drought indices, such as the standardized precipitation index and standardized precipitation evapotranspiration index, to identify drought and assess its severity. Prior research has often relied on statistical methods like Holt-Winters and ARIMA but Holt-Winters and ARIMA are not very effective in predicting drought indices accurately over long periods of time. This paper compares the performance of ARIMA statistical model with LSTM using multivariate input in predicting 1-, 6- and 12-month drought indices, with the hope of improving performance. LSTM, which uses long short-term memory in a recurrent neural network, is particularly effective at handling real-time nonlinear data, and can better help authorities prepare for and mitigate the impact of natural disasters.

Problem statement of their research:

The aim of this paper is to enhance the prediction accuracy of 1-, 6-, and 12-month drought indices by comparing the performance of the ARIMA statistical model with LSTM, which uses multivariate input.

Methodology:

The proposed LSTM RNN model is split into two models. The first model uses SPI and SPEI. The AdaGrad algorithm is used to reduce the root mean square (RMS) loss. It does this by increasing the learning rate for more sparse parameters and decreasing the learning rate for less

sparse ones. The second model is more intricate than the first one, as it takes additional parameters such as temperature and relative humidity into account. This model is trained using the RMSProp algorithm, which is better suited for this case since it addresses the issue of decreasing learning rates. The methodology is to improve LSTM network by reducing the number of training epochs while maintaining accuracy and avoiding the problem of vanishing gradients caused by the use of sigmoid and hyperbolic tangent functions in the LSTM.

Conclusions:

Many variables are the primary focus of the LSTM model, whereas one independent variable will be the primary focus of the ARIMA model. Temperature and humidity are the only two variables that are excluded from the equations used to calculate the SPI and SPEI values. When making predictions over extended periods of time, the LSTM model is more accurate than the ARIMA model. While there was a positive connection between the variables and the SPI and SPEI values, LSTM performed well. Future research can make advantage of a huge dataset and incorporate deep learning techniques that can recognize the many patterns of droughts to get better outcomes.

2.1.2.b ***Paper title:*** *Evapotranspiration evaluation models based on machine learning algorithms—A comparative study*

F. Granata: Agricult. Water Manage., vol. 217, pp. 303-315, May 2019.

How was the specific research problem tackled or resolved by other research papers? What are the practical implications, and how were they beneficial?

This paper aimed to develop the forecasting of actual evapotranspiration(the process of evaporation) using M5P Regression Tree, Random Forest and Support Vector Regression models to find the best accurate model. This paper also showed how hybrid models could be used to predict the evapotranspiration index values which describe the water resource level in the area.

Problem statement of their research:

The paper discusses training the data of temperatures, wind speed and relative humidity to predict the evapotranspiration index using various machine learning algorithms to find the most accurate model.

Methodology:

In this paper, 75% of the data has been taken to train the data and 25% of the data has been used to test the data. Data consists of the min, max, and mean values of temperatures, wind speed, and relative humidity values. This data is trained using M5P Regression Tree, Random Forest and Support Vector Regression models.

Conclusions:

Among all the models used to predict the evapotranspiration index, M5P Regression showed the best accuracy with 95%. Among all the variables, T(min), T, RH, and W had the best input combination which means that this combination indicates a high impact on evaporation. Also, as

per the data taken in the paper, all models significantly overestimated the pan evaporation values. And these significant changes might be due to the dataset taken in the paper.

2.1.2.c ***Paper title:*** *Drought analysis with machine learning methods*

E. E. Başakın, Ö. Ekmekcioğlu and M. Ozger

Pamukkale Univ. J. Eng. Sci., vol. 25, no. 8, pp. 985-991, 2019.

How was the particular research problem addressed or solved by other research papers? What is the practical application, and how was it useful

This study aims to develop a new drought index called CTEI that provides comprehensive observations of droughts over large areas with good spatial and temporal resolution. The study applied five machine learning models to model the new index in the Ganga river basin, and the best-performing model was the eighth model that used the SVM algorithm, followed by the Matern 5/2 Gaussian model. The SVM and Matern 5/2 Gaussian methods were the best-performing machine learning algorithms in predicting CTEI for the Ganga basin.

Problem statement of their research:

The study focuses on developing a new drought index, the Combined Terrestrial Evapotranspiration Index (CTEI), in the Ganga river basin, using hydro-meteorological variables, precipitation (P), potential evapotranspiration (PET), and Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage anomalies (TWSAs). The aim is to model the CTEI using five different machine learning algorithms: Random Forest (RF), Support

Vector Machine (SVM), boosted trees, bagging, and Matern 5/2 Gaussian process regression (GPR), and compare their accuracy and stability in predicting droughts.

Methodology:

The paper outlines the methodology used to derive the Combined Terrestrial Evapotranspiration Index (CTEI) which quantifies drought events at a regional scale. The CTEI is derived from GRACE TWSA data and meteorological variables (precipitation and evapotranspiration). The index was compared to existing drought indices and demonstrated good correlations. The paper further details the machine learning models used to predict the CTEI. The Support Vector Machine model is a supervised learning algorithm used for regression with the aim of minimizing error by individualizing the hyperplane. Decision-tree learning is used to create a model that forecasts the target value based on several independent inputs. Various techniques were often deployed to construct more than one decision tree.

Conclusions:

The study developed a new drought index, called CTEI, for the Ganga river basin by using five different machine learning models, based on meteorological variables and GRACE TWSA. The models were tested on monthly data collected from 2003 to 2016, with data from 2003-2013 used for training and data from 2014-2016 used for testing. The best performing model in predicting CTEI was found to be the SVM algorithm with eight model settings, followed by the Matern GPR algorithm and Model 6, which used data fusion of various hydroclimatic parameter.

2.2 Literature Survey:

Numerous studies and research have been conducted on the identification of drought and its impact on various fields. Various papers have discussed multiple ways to find drought forecasting based on different sets of variables and data. One common variable among most of the research papers is SPEI and SPI values. One popular approach for forecasting and predicting drought is using LSTM, which has demonstrated better results when additional variables, such as relative humidity and temperature, were included in the training data. Along with that, there are other Machine Learning models such as Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Tree Classifier, boosted trees, bagging, Matern 5/2 Gaussian process regression (GPR) and M5P Regression. Accuracies of each model vary due to the datasets used in various papers but the predictions could show better results when accurate datasets are used such as temperatures, soil moisture, vegetation index and evapotranspiration. From the review of existing literature, it is clear that a considerable amount of research has been dedicated to utilizing machine learning algorithms for drought prediction.

2. Data Source and Datasets

3.1 where to find the data set

For this project, we obtained data from three public sources. The data we collected was necessary for calculating the Standardized Precipitation-Evapotranspiration Index (SPEI) value. Additionally, we gathered information on the vegetation index and soil moisture index by analyzing satellite image data. To train our machine learning models for predicting the percentage of California's land classified as drought-prone, we collected data covering the period from 2000 to 2020. For historical drought data, we sourced our information from a public website.

3.2 How to collect data, if applicable

Dataset URLs

3.2.1. https://power.larc.nasa.gov/data-access-viewer/

3.2.2. https://developers.google.com/earth-engine/datasets

3.2.3. https://droughtmonitor.unl.edu/CurrentMap/StateDroughtMonitor.aspx?CA

3.3 Explain what the dataset contains

We received a California environment dataset from the NASA Power website, which includes monthly data on air temperature, earth's temperature, humidity, wind speed, and precipitation. This information can be used to determine the SPEI, which indicates when precipitation and evapotranspiration are in balance. This value can be used to monitor drought conditions, assess how they impact agriculture, water resources, and other businesses, and provide information to drought early warning systems.

We later collected satellite images of the state to retrieve the Normalized Difference Vegetation Index (NDVI) value. This index tells us the percentage of vegetation present on the land, including plants and trees. We also used the images to get the Soil Moisture Index (SMI) value, which provides insight into current soil moisture conditions relative to historical data for the same time of year.

Additionally, we have obtained data from the U.S. Drought Monitor website that shows the percentage of the country's land area that is considered to be drought-prone. This information will help us train the model to forecast the percentage of California's land area that will experience drought in the future.

3.4 Examples of data from your dataset

	Year	Month	Earth Skin Temperature (C)	Temperature at 2 Meters (C)	Specific Humidity at 2 Meters (g/kg)	Wind Speed at 2 Meters (m/s)	Precipitation Corrected (mm/ day)
	1985	January	-0.03	1.3	3.36	0.62	0.82
	1985	February	0.18	2.05	3.6	0.72	2.23
	1985	March	0.46	1.25	3.72	0.69	4.27
	1985	April	8.65	9.6	4.94	0.67	0.41
	1985	May	10.46	11.01	4.33	0.7	0.06
	1985	June	19.05	19.44	4.82	0.73	0.3
	1985	July	21.98	22.25	5.43	0.7	0.42
	1985	August	18.39	19.15	3.85	0.7	0.05
	1985	September	11.95	12.67	5.0	0.62	1.5
	1985	October	9.12	10.3	4.03	0.67	1.71
	1985	November	0.54	2.41	3.6	0.66	5.08

Fig 3.4.1 Snapshot of data showing Weather deatails

Data Tables U.S. Drought Monitor

Week	None	D0-D4	D1-D4	D2-D4	D3-D4	D4	DSCI
2023-02-14	0.64	99.36	84.60	32.62	0.00	0.00	217
2023-02-07	0.64	99.36	84.60	32.62	0.00	0.00	217
2023-01-31	0.64	99.36	89.56	32.57	0.00	0.00	221
2023-01-24	0.64	99.36	89.56	32.57	0.00	0.00	221
2023-01-17	0.64	99.36	92.12	42.84	0.00	0.00	234
2023-01-10	0.00	100.00	95.38	46.00	0.32	0.00	242
2023-01-03	0.00	100.00	97.93	71.14	27.10	0.00	296
2022-12-27	0.00	100.00	97.94	80.56	35.50	7.16	321

Fig 3.4.2 Snapshot of the table showing percentages of drought areas in California

3.4.1 Dataset Details

We have gathered a lot of information for this project, spanning the years 2000 to 2020. The datasets we gathered come in a range of sizes, from daily recordings to monthly records. We will employ a total of 10–12 attributes to train the model with each dataset including 5–6 features. We can make sure that our model precisely and thoroughly forecasts the portion of California's land area that is anticipated to be in a drought-prone state in the future.

3.4.2 Attributes Details

Combined we have 10+ attributes. Below is the information about the main attributes:

A. Earth skin temperature : The Earth's surface has a certain amount of heat, and it needs to maintain a balance to keep things stable. This heat balance is affected by how much heat is exchanged between the surface and the air around it. This value is collected in celsius.
B. Temperature at 2 meters : Temperature at 2 meters refers to the air temperature at a height of 2 meters above the Earth's surface. It is a commonly used measurement in meteorology, atmospheric science, and environmental monitoring.
C. Specific humidity at 2 meters: Specific humidity at 2 meters is a measure of the amount of moisture present in the air at a height of 2 meters above the Earth's surface. It is expressed in grams of water vapor per kilogram of dry air (g/kg), which is a common unit of measurement for atmospheric moisture content.
D. Wind speed at 2 meters: Wind speed at 2 meters refers to the speed at which air is moving at a height of 2 meters above the Earth's surface.
E. Precipitation corrected: Precipitation Corrected (mm/day), which is commonly given in millimeters per day, is a measurement of the amount of precipitation that falls on a specific area over a specific period of time.

F. NDVI: NDVI stands for Normalized Difference Vegetation Index, which is a measure of the amount of healthy vegetation in a particular area.

G. SMI: A measurement of the moisture content in the topsoil layer of the Earth's surface, the SMI abbreviates soil moisture as a percentage. It is frequently stated as a relative index, where higher values denote wetter soil conditions and lower ones denote drier conditions.

The drought data that we got from the U.S drought monitor website shows the droughts are classified in 5 levels

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

In the dataset they have provided the percent area for each day for the california state.

References

Mokhtar.A et al., "Estimation of SPEI Meteorological Drought Using Machine Learning Algorithms," in IEEE Access, vol. 9, pp. 65503-65523, 2021, doi: 10.1109/ACCESS.2021.3074305. https://research-repository.griffith.edu.au/bitstream/handle/10072/404189/Gyasi-Agyei483825-P ublished.pdf?sequence=2

Khan.N, "Prediction of droughts over Pakistan using machine learning algorithms", Adv.Water Resour., vol. 139, May 2020. https://ui.adsabs.harvard.edu/abs/2020AdWR..13903562K/abstract

Vicente-Serrano . S.M, Van der Schrier . G, Beguería .S, et al., "Contribution of precipitation and reference evapotranspiration to drought indices under different climates", J. Hydrol., vol. 526, pp. 42-54, Jul. 2015 https://www.researchgate.net/publication/275836498_Contribution_of_precipitation_and_reference_evapotranspiration_to_drought_indices_under_different_climates

Xu .L, Chen. N, Zhang . X , et al., "An evaluation of statistical NMME and hybrid models for drought prediction in China", J. Hydrol., vol. 566, pp. 235-249, Nov. 2018. https://www.sciencedirect.com/science/article/pii/S0168169922002423

Ganguli .F and Reddy M.J, "Ensemble prediction of regional droughts using climate inputs and the SVM-copula approach", Hydrol. Processes, vol. 28, no. 19, pp. 4989-5009, Sep. 2014.

https://onlinelibrary.wiley.com/doi/10.1002/hyp.9966

Granata. F, "Evapotranspiration evaluation models based on machine learning algorithms —A comparative study", Agricult. Water Manage., vol. 217, pp. 303-315, May 2019.

https://www.researchgate.net/publication/331678180_Evapotranspiration_evaluation_models_ba sed_on_machine_learning_algorithms-A_comparative_study

Masinde. M, "Artificial neural networks models for predicting effective drought index : Factoring effects of rainfall variability", Mitigation Adaptation Strategies Global Change, vol. 19, no. 8, pp. 1139-1162, Dec. 2014.

https://ideas.repec.org/a/spr/masfgc/v19y2014i8p1139-1162.html

Zhang. R, Z.-Y. Chen, L.-J. Xu and C.-Q. Ou, "Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province China", Sci. Total Environ., vol. 665, pp. 338-346, May 2019.

https://www.mdpi.com/2073-445X/11/11/2040

Başakın E.E, Ekmekcioğlu .Ö and Ozger.M, "Drought analysis with machine learning methods", Pamukkale Univ. J. Eng. Sci., vol. 25, no. 8, pp. 985-991, 2019.

https://www.mdpi.com/2073-4441/13/4/547

Shahbazi . A.N, "Seasonal meteorological drought prediction using support vector machine"

World Appl. Sci. J., vol. 13, no. 6, pp. 1387-1397, 2011.

https://onlinelibrary.wiley.com/doi/10.1002/env.2154

Poornima .S and Pushpalatha .M, "Drought prediction based on SPI and SPEI with

varying timescales using LSTM recurrent neural network", Soft Comput., vol. 23, no. 18, pp. 8399-8412, Sep. 2019.

https://www.researchgate.net/publication/333654021_Drought_prediction_based_on_SPI_and_S PEI_with_varying_timescales_using_LSTM_recurrent_neural_network

Breiman .L, "Random forests", Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

https://www.scirp.org/(S(czeh2tfqw2orz553k1w0r45))/reference/referencespapers.aspx?referenc eid=173455