California’s Changing Climate: Predicting Future Drought

Andres, Anvesh, Christopher, Owen

June 13, 2024

Abstract

Droughts are common in California due to their frequent swings between extremely dry and wet periods, making climatic patterns unpredictable. California's changing climate complicates predictions about its future. Recurring droughts in California pose significant challenges as they take a serious toll on water resources, agriculture, economy, society, and the environment. Motivated by this imperative, our aim for this project is to predict California’s future drought occurrence by gathering and combining multiple datasets from more than the past ten years from different locations based on their features. Through multiple data representations, our findings suggest that increased precipitation, higher snowpack levels, lower temperatures, and greater percentage capacities and averages of reservoirs result in a lower likelihood of drought development. Furthermore, our analysis quantifies the observed correlation among different combinations of features which conveys how these features play a role in assessing future droughts in advance and mitigating their impacts.

1 Introduction

Droughts in California profoundly impact its environment, economy, and society. California's climate tends to swing between extreme dry and wet periods, making climatic patterns highly unpredictable. Predicting future droughts in advance would be highly beneficial for efficiently managing water resources and mitigating adverse effects promptly. In this project, we have collected and merged multiple datasets from 2010-2023 from different locations based on their features to predict drought conditions more effectively. In this project, we have collected and merged multiple datasets from 2010-2023 from various locations by focusing on their features to predict drought conditions effectively. Snowpack data, including snow water content and snow depth, is crucial for assessing water supply availability since snowmelt replenishes many reservoirs and groundwater resources, especially during the summer months. Reservoir data

provides insights into current reservoir levels (percentage of capacity), maximum capacity, historical averages, and percentage of average, allowing us to compare and correlate with historical trends. This information highlights the availability of water resources and informs water management techniques. Precipitation data displays information on long-term rainfall and snowfall trends while temperature data displays information on weather trends from the past several years wherein they both show how much the historical values deviate from the average. Overall, these datasets are well interconnected as they give us an overview of determining how snowpack levels in the winter and spring correlate with water levels in reservoirs, rivers, and groundwater during summer which helps us understand how winter precipitation impacts summer water supplies. The California Drought Monitor is a website that updates the state's drought status every week. It shows how much of California is in a drought, which can be seen by looking at a certain intensity or cumulatively across multiple categories within a designated range.

In this paper, we aim to address a fundamental question: How could we predict the drought classification of a month in California as early as one year prior using various machine learning models? Section 2 outlines the underlying model and details the experimental methodology, Section 3 presents the results, and Section 4 discusses the caveats and future works.

2 Model & Method

We model the interaction between drought status and various features using different forms of data representations. The datasets consist of observations collected at an interval of one month over thirteen years from October 2010 to September 2023. These dates were chosen as they include the biggest drought years in California.

For the daily and weekly data, we take the average of all the values, convert it to monthly data, and create a pivot table to have the month as a unique identifier for each sample in the dataset. We join the different datasets on the month index. We also grouped the snowpack data from various mountains, counties (for precipitation and temperatures), and reservoirs in the regions of Northern, Central, and Southern California to make the respective features to study how these regions affect the drought status.

The first column in the dataset represents the date in 'MonthYear' format on a monthly basis while the rest of the columns represent features consisting of numerical data (e.g., snow water content (inches), snow depth (inches), temperature, precipitation (inches), etc.). Our target variable is the drought index for a specific period in the future.

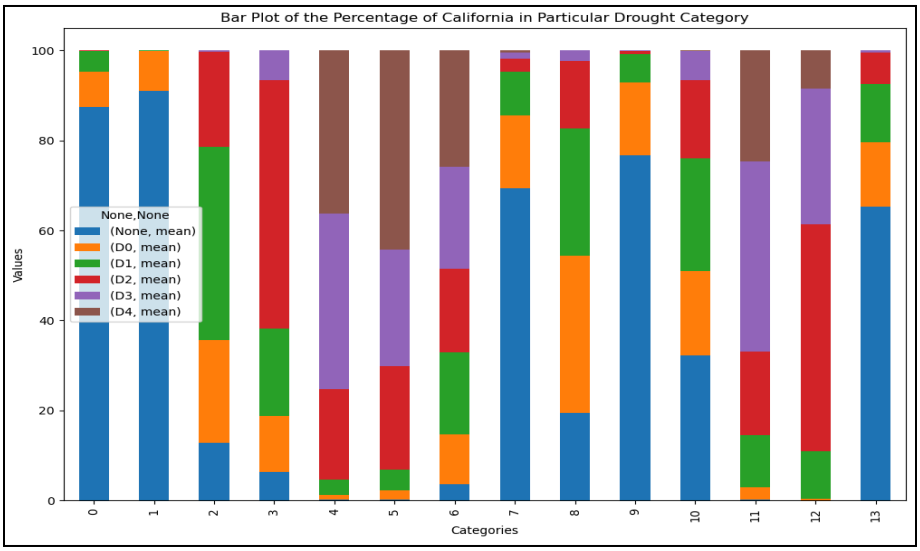

Figure 1. The stacked bar plot shows the years from 2010-2023 are depicted in different drought intensities measured in percentages. None (blue) represents no drought, D0 (orange) represents abnormally dry, D1 (green) represents moderate drought, D2 (red) represents severe drought, D3 (purple) represents extreme drought, and D4 (brown) represents exceptional drought.

Figure 1 above shows the different intensities of drought California was facing throughout the years from 2010-2023. Looking at the bar plot above, we see that 2010-2011, 2017, 2019, and 2023 were the non-drought years whereas 2012-2016 and 2021-2022 were the biggest drought years.

We make a second target variable, a boolean variable by putting a threshold value of at least 35% of California being in some drought condition to classify the month as a drought month. We then get the average of the same month over the two previous years for each feature and align it with the current month's drought status. This makes our 156-month dataset from October 2010 to September 2023 a 132-month dataset from October 2012 to September 2023 as we lose the first 24 months with insufficient historical data. In the process of data cleaning, we replace cells with

missing values represented by ‘M’ with ‘NaN’ for null, replace ‘T’ which represents very low values with 0, and omit unnecessary columns.

2.1 Assumptions

To maintain simplicity in our model, we only focus on California rather than the entire United States and make the following assumptions: Our model assumes a random occurrence of drought because California tends to swing a lot between extreme wet and extreme dry conditions. Our model also assumes that California’s drought status depends on snowpack levels, amount of precipitation, reservoir levels, and temperature from three regions of the state. We also assume that the various locations had equal effects in their respective regions: snowpack data, precipitation, reservoir data, and temperature when we averaged (non-weighted) them.

2.2 Method

We construct a correlation square matrix among all the features (drought indicators) and based on that, we perform correlation analysis to study which combination of features have the most in common as well as which features exert the strongest influence on the drought status.

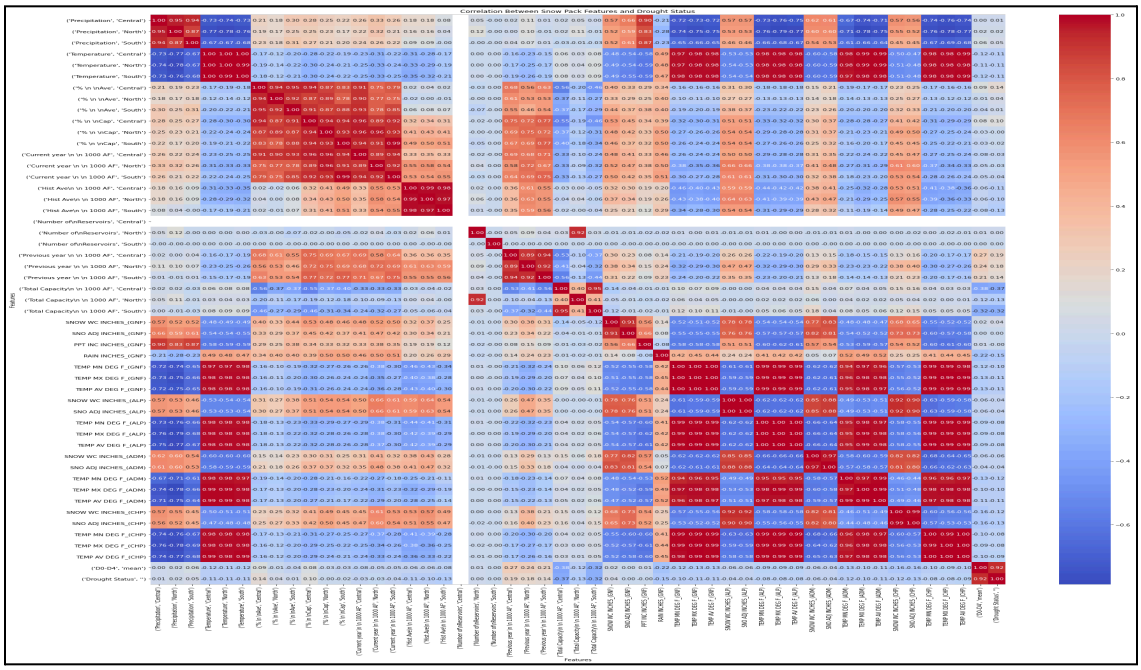

Figure 2. The plot above shows a combined correlation matrix measuring the strength of the association of all the features from the reservoir, snowpack, temperature, and precipitation data, and drought status data.

As shown in Figure 2, there is the expected correlation of precipitation and temperature data for the different regions of California as well as the snowpack data. To our surprise, we find that none of the features individually has a correlation greater than 0.5 or lower than -0.5 with the drought status. This is unideal as typically the high correlation of features with the target variables represents an importance in predicting the target variable; moreover, what astonishes us the most is the correlation between 'drought status' and 'current capacity of reservoirs (1,000 AF)' since the matrix depicts almost no correlation between these features. In fact, we expect there to be a strong correlation between these features as a higher capacity of reservoir levels would imply a less likelihood of drought occurrence.

We proceed to create a machine learning model that takes in this dataset to predict the drought status. To maximize the efficiency of our model we attempt dimensionality reduction using PCA and using the first five Principal Components as our input features for the methods. This results in low model accuracy and unclear clusters from the PC analysis of any pairs of the five principal components. This means the relevance of the dimensions of the data dropped when PCA is done. We chose to use the dataset in its entirety without any dimensionality reduction which we see success using classification models such as decision trees and random forest classifiers. As it would be ideal to have an exact percentage of the state that is in drought as our prediction, we attempt a few regression models targeting the drought index D0-D4 shown in Figure 1. This shows less success in terms of accuracy than the classification models.

3 Analysis of Results

This section presents the performance evaluation of various machine learning models applied to predict drought occurrences. Initially, we used regression models because of their simplicity and effectiveness in handling continuous data. The regression models that we tested include Random Forest Regression, Lasso Regression, and Ridge Regression. The metrics used to evaluate the models are Mean Squared Error (MSE) and R-squared (R^2) score.

Model	MSE	R^2
Random Forest Regression	0.0439	0.7144
Lasso Regression	0.1447	0.0408
Ridge Regression	0.1401	0.0716

Figure 3. The table shows the performance metrics of the regression models. The Random Forest Regression, Ridge Regression, and Lasso Regression were evaluated using MSE and R^2 scores.

In Figure 3, Random Forest Regression achieved an MSE of 0.0439 and an R^2 score of 0.7144, Lasso Regression achieved an MSE of 0.1447 and an R^2 score of 0.0408, and Ridge Regression achieved an MSE of 0.1401 and an R^2 score of 0.0716. The regression models showed moderate predictive accuracy but were insufficient for predicting future drought occurrences. The R^2 scores for Ridge and Lasso Regression models were low, indicating that these models were not effective at capturing the relationship between the features and the target variable. Although the Random Forest Regression model performed better, it did not meet the desired accuracy and precision for our predictive goals.

Given these unsatisfactory results, regression models weren't good enough for accurately predicting drought occurrences. This led us to explore more complex models and consider classification approaches, which might better handle the nuances of our dataset and provide more reliable predictions. The classification models we tested are Decision Tree Classifier, Random Forest Classifier, Logistic Regression, and Support Vector Machine. The metrics used to evaluate the models are accuracy, precision, recall, and f1-score.

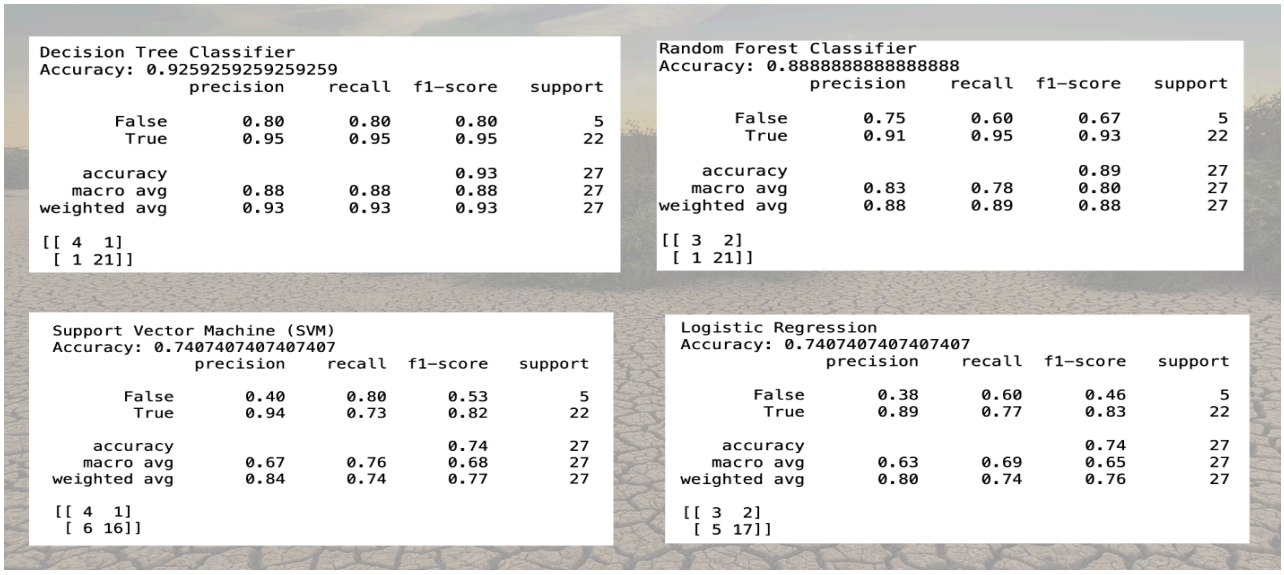

Figure 4. The table shows the performance metrics of the four different classification models. The Decision Tree Classifier, Random Forest Classifier, Logistic Regression, and Support Vector Machine were evaluated using accuracy, precision, recall, and f1-score.

Figure 4 above shows how the different classification models performed on our dataset. The best-performing classification model is the Decision Tree Classifier with an accuracy score of 0.9259 meaning that the model correctly predicted the outcome for 92.59% of the cases in the test set. The second-best model was the Random Forest Classifier with an accuracy of 0.889 followed by Support Vector Machine and Logistic Regression with an accuracy of 0.7407. From the results of the classification models, we see that there was a lower precision, recall, and f1-score for predicting non-droughts compared to predicting droughts. This is because our training and testing target variable was dominated by “true”. This might affect our model performance when predicting an upcoming year's lack of drought. While this is unideal, it is less impactful than a false negative on an upcoming drought year. We do a feature importance curve to analyze the features affecting our best-performing models: Random Forest Classifier and Decision Tree. They both indicate that the reservoir features are most relevant followed by the snowpack features then finally precipitation and temperature.

4 Discussion

4.1 Caveats

Our model simplifies the data by averaging it at multiple points; this is likely to bias the model due to potential outliers. Our model is also only successful when predicting the drought or no drought classification but not the actual percentage of the state in drought. Using classification limits the model's ability to predict the severity or specific characteristics of droughts. Our training data is heavily skewed to drought years which is causing a higher performance when predicting the drought months than the non-drought months. This is not as big of a caveat as it is safer, however, it could mislead entities such as governments to allocate funds to drought preparation when it is not necessary.

4.2 Future Work

We began with an approach to devise a predictive model for future drought utilizing the data from 2010-2023 consisting of a variety of different features. To make our results more accurate, we must consider taking more drought years and non-drought years from over the past decades into account and include them as part of our dataset; moreover, we would also consider implementing different principle component combinations like clustering and logistic regression to classify ongoing and/or predict future drought or devise a model that could predict future drought in advance or devise a probability model based on the considered features that would predict the likelihood of drought occurrence in a particular year by identifying early warning signs and thresholds to improve upon drought forecasts in advance. Consequently, we would have more evidence to back up our model with reliable and accurate results.

5 Conclusion

In conclusion, this project aimed to enhance California's prediction of drought occurrences by using various machine learning models and extensive datasets. We achieved more accurate predictions by shifting from regression to classification models and utilizing the full dataset without dimensionality reduction. Our findings highlight the critical role of drought features such as precipitation, snowpack levels, temperature, and reservoir capacities in forecasting droughts.

This improved our understanding and predictive capability, and contributed significantly to the field of climate science and resource management. By providing early warnings and identifying key indicators, our work supports better preparedness and mitigation strategies that ultimately helps manage water resources more efficiently and reduce the adverse effects of droughts on the environment, economy, and society.

6 Contributions

Owen contributed to data collection by collecting snowpack data and reservoir data. He contributed to the data cleaning making the datasets have one index of month by using pivots to make the other features columns of the dataset. Made the code to get the two previous year averages to align with the current year drought status. Contributed to predictive model selection and optimization. Contributed to writing the final paper.

Christopher contributed to data collection by collecting temperature, precipitation, and reservoir data. He contributed to data cleaning by adding columns to each dataset to identify the region of each data point and by separating the dates. He also decided to try different machine-learning models to see which one worked the best. Finally, he also contributed to creating the slides and final paper.

Anvesh contributed to preliminary research by focusing on the availability of snowpack, rainfall, drought status, and precipitation data from credible online sources such as the California Department of Water Resources and how to use them. Additionally, he collaborated with group members on coding tasks and contributed to the analysis of various data representations. Anvesh further contributed to the presentation slides and final paper where he worked on multiple sections of the paper.

Andres contributed to data selection and collection, collaborating on which regions would be representative of the status of California drought concerning snowpack, reservoir, precipitation, and average temperature. He also assisted with data cleaning and finalization on the structure of the dataset. Collaborated on changing drought data to binary by creating a threshold which assigned a true or false value concerning the presence of a drought.