Analyzing the Impact of Environmental Factors on Drought Intensity Using Machine Learning

Aditya Aiyer

James Logan High School, Union City, CA

Abstract

A drought is a period of drier-than-normal conditions. These dry conditions can reduce the quality and quantity of water, raise illnesses and diseases, and, in turn, mortality rates. Drought is a global issue and understanding how different environmental aspects affect drought can allow for preparation and prevention in a time of crisis. The US Drought Monitor categorizes drought on a scale. First, several data visualizations were performed using a real-world environmental dataset to understand how individual factors affected drought intensity. The features were then organized into a graph from highest to lowest contribution to drought prediction. The three top and the three lowest features were plotted on violin plots to analyze how each feature varied across drought categorizations. These results highlight which environmental features should be closely monitored during drought prediction. Five different models, including a neural network, were compared to determine which would produce the highest accuracy results, and the RandomForestClassifier yielded the highest accuracy. A model was then developed that would take multiple environmental factors and predict drought intensity. A time-series split was run on the RandomForestClassifier to compare it to a DummyClassifier, which is a simple model used as a performance baseline. The results revealed that the top 3

drought predictors were surface pressure, temperature range at 2 meters, and the maximum temperature at 2 meters. In addition, the RandomForestClassifier outperformed the baseline model after 10000 lines. This study provides a comprehensive review of important environmental features and effective models for early drought warning systems.

Keywords: environment, drought, permutation, time-series split, cross-validation

Introduction

With the rise in global warming, drought is becoming a significant issue. Over the past 2 decades, drought occurrences have increased by 17% in the Western United States ("New Research Finds Rising Heat Driving Western U.S. Droughts," n.d.). Drought prediction and measures to protect against it are valuable assets for places worldwide. Each place has a different set of environmental conditions, so it is essential to analyze many factors to gain as accurate a representation of potential droughts as possible. Datasets detailing the impact of environmental conditions on droughts are available, and predictive drought models have been created, but there has not yet been a paper that assessed the most impactful environmental conditions, compared various models for drought prediction through different methods, and created visualizations for future use.

In the past, information has been collected and datasets have been organized, such as the global database of droughts from 1951 - 2016 ("A New Global Database of Meteorological Drought Events from 1951 to 2016" 2019) and the one being used in this study, DroughtED (Christo Minixhofer et al. 2021; Naumann et al. 2014), which is a novel dataset for drought-driven forecasting (Christoph Minixhofer 2021).

Furthermore, studies based on climate data and soil moisture using AI models for drought prediction have been made (Oyounalsoud et al. 2024). However, some of these studies use a limited number of models, so there are more models to be tested, such as MLPNN, SVR, ANFIS, and EDT (Azimi et al. 2022). Another example is the LSTM model ("Explainable AI in Drought Forecasting," 2021). This study aims to combine these elements, using prior datasets and features along with more models to develop and train an AI to predict droughts with high levels of accuracy while also assessing the most critical environmental conditions to look for and to create visualizations.

This study investigates the following research question: among various environmental factors, which were the most impactful contributors to drought levels? To answer this, a logistic regression model was used to narrow down the top three. The model was then trained to predict drought by assessing the levels of the most important factors. Various methods for drought prediction were also performed on numerous models to determine which resulted in the highest accuracy using cross-validation. Finally, a time-series split was applied to both a RandomForestClassifier and a DummyClassifier model to assess the size of a dataset required for the RFC to outperform a baseline dummy. Additionally, many plots and visualizations were created to show how different environmental conditions impacted the drought score. Violin plots were also generated for some of the top-ranking features to observe how they individually impacted the score.

Method

Dataset

A dataset called "Predicting Droughts using Weather and Soil Data" was used (Christoph

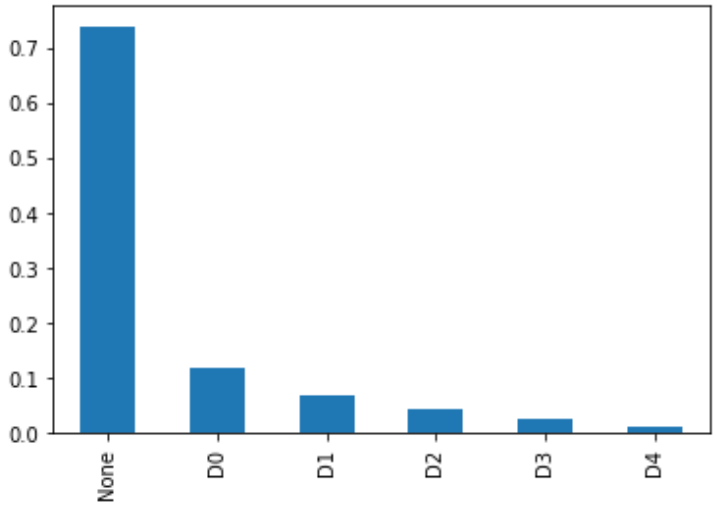
Minixhofer 2021). This dataset, based on real data, recorded different levels of environmental factors and reported the drought intensity for each area. The dataset contains approximately 19.3 million lines across 21 features.

Dataset Distribution

The dataset contained a list of historical drought events occurring throughout various United States counties. Each drought was classified into one of six intensity levels: None, D0, D1, D2, D3, and D4. As the D level increases, the drought intensity also increases, meaning temperatures were hotter and more severe for the residents, etc. The drought monitor provides a full description for each value ("Drought Classification," n.d.).

Figure 1

Distribution of the selected dataset, where each drought intensity is measured by the probability of occurring in the dataset



As indicated in Figure 1, the "None" classification contained the largest percentage of the dataset, indicating an imbalance. This shows that a naive classifier, such as a DummyClassifier, would always predict None. Therefore, the DummyClassifier was used as a baseline comparison against the RandomForestClassifier.

Models

Each factor was plotted against drought intensity using scatterplot graphs. The distribution of various drought intensities was assessed when plotted on the factor 1 vs. factor 2 graph.

Permutation importance ("4.2. Permutation Feature Importance," n.d.) was used, which quantitatively measures a feature's contribution to the performance on a dataset. The RandomForestClassifier model was used to retrieve a list of feature importance values for each feature within the drought dataset and plotted on a bar chart.

Violin plots were generated for the three most important and the three least important features to understand the distribution of the data and the corresponding shape of the violin curves.

A 10-fold cross-validation was performed on five different models to compare their accuracy results. This approach allowed the dataset to be divided into subsets and ensured that only historical data was used to predict the drought conditions of future dates.

Time-series splits were applied for RandomForestClassifier ("RandomForestClassifier," n.d.) and DummyClassifier ("DummyClassifier," n.d.). to break the dataset into parts, and to vary the training and testing percentages. This ensured the model does not use future data to predict past events. To achieve this, the model was organized chronologically.

To test the change in the accuracy of the model, the number of lines included as part of the training was varied. A DummyClassifier was used to evaluate how the accuracy of a naive model compares to the RandomForestClassifier, which uses feature data. Both models were trained on the same train-test split to ensure consistency in the data used for comparison. This method allowed for an understanding of the dataset size required for the RFC model to outperform the naïve model.

Results

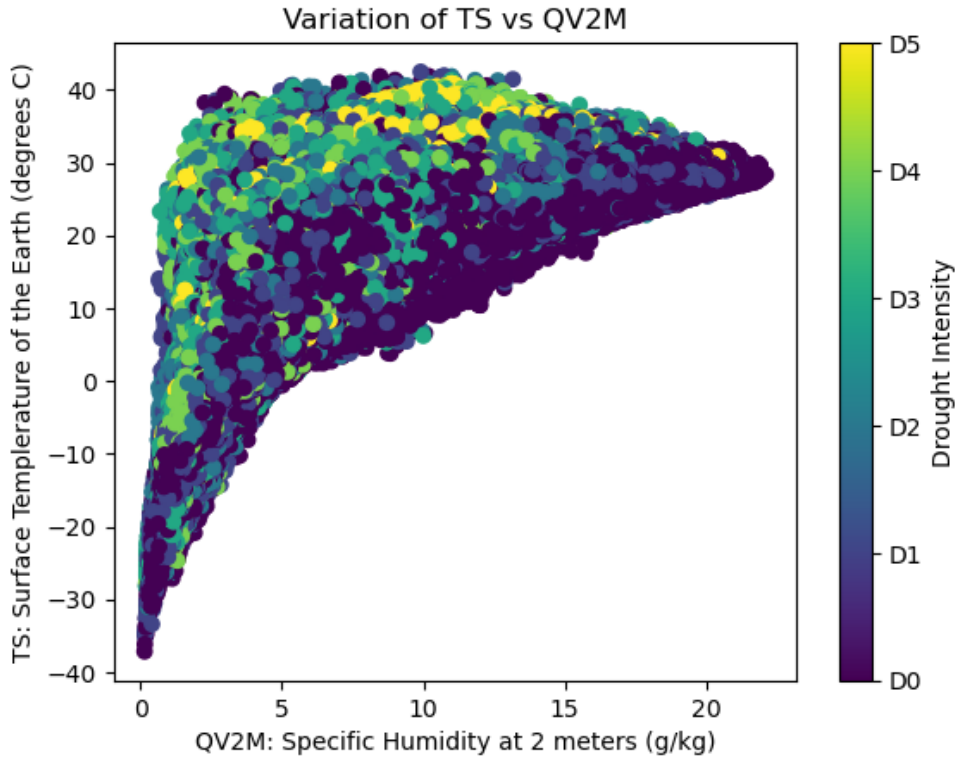
Comparing different environmental features in the dataset to drought intensity

The surface temperature and specific humidity were plotted against drought intensity. As indicated in Figure 2 below, as the surface temperature increased, the drought intensity increased; however, as the specific humidity increased, drought intensity decreased. The greatest

drought intensity, indicated by the lightest colors, occurred where surface temperature was around 35 °C, and specific humidity at 2 meters was 5 to 10 g/kg.

Figure 2

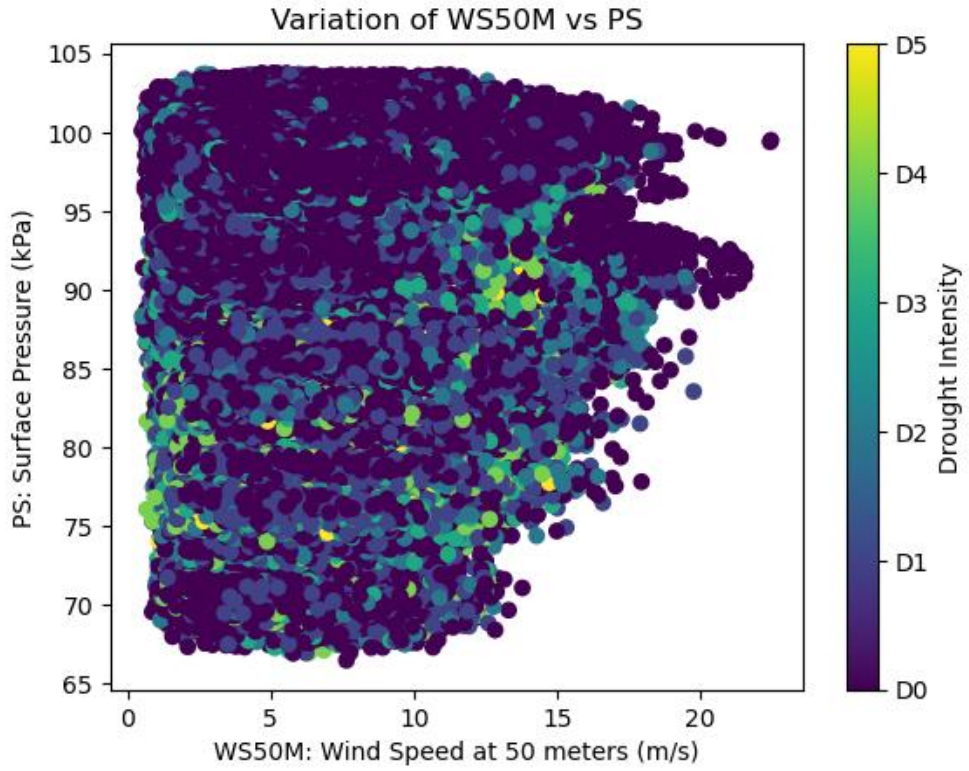
Drought intensity trends based on changes in TS (Surface Temperature of the Earth in degrees C) and QV2M (Specific Humidity at 2 meters in g/kg).



The lighter colors indicate greater drought intensities, with level 5 being the most intense. The drought intensity was then evaluated against wind speed and surface pressure. As observed in Figure 3 below, there was no clear correlation between wind speed at 50 meters and drought intensity. As surface pressure increased, drought intensity somewhat decreased, but there was no clear trend. The lightest colors, or greatest intensity, were not concentrated at any particular wind speed but occurred at surface pressure levels of 75 to 90 kPa.

Figure 3

Drought intensity trends based on changes in WS50M (Wind Speed at 50 meters in m/s) and PS (Surface Pressure in kPa), where lighter colors indicate greater drought intensities, with level 5 being the most intense.

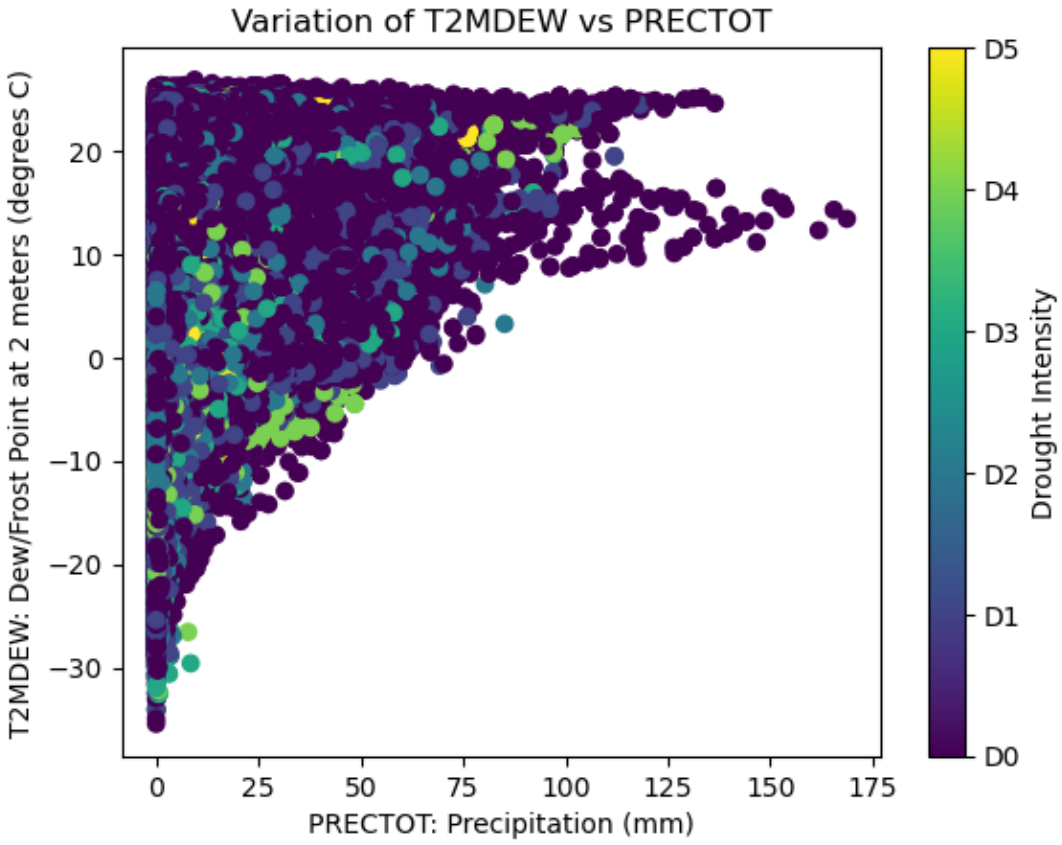


The impact of precipitation and dew/frost was analyzed on drought intensity. As seen in Figure 4 below, there was no clear correlation between dew/frost point and drought intensity, although there seemed to be a concentration of light spots at moderate temperatures. There was no noticeable correlation between precipitation and drought intensity, although drought

occurrences tapered off at around 100 mm. The lightest colors, or greatest intensity, occurred at a dew/frost point of -10 ℃ to 10 ℃, and a precipitation level of 25 to 50 mm.

Figure 4

Drought intensity trends based on changes in T2MDEW (Dew/Frost point at 2 meters in degrees Celsius) and PRECTOT (Precipitation in mm), where lighter colors indicate greater drought intensities, with level 5 being the most intense



Determining important correlations of drought via machine learning

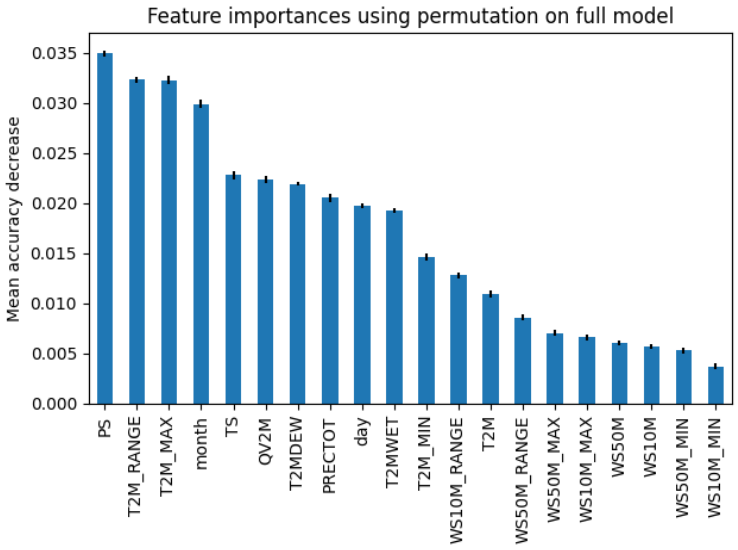
Twenty environmental features in the dataset were compared using permutation to

determine the most important features while predicting drought intensity. The original dataset contained “FIPS code”, “Date”, and “score” as three features, but FIPS code was omitted as it was not relevant to an environmental condition, and the score feature was removed because that was the dependent variable being measured. The date feature was then split into month and day as two separate features, to predict if the month may indicate signs of seasonal drought predictability.

As indicated in Figure 5 below, the top four selected features were PS (Surface Pressure), T2M_Range (Temperature Range at 2 meters), T2M_Max (Maximum Temperature at 2 meters), and month, in descending order. The bottom three selected features were WS10M (Wind Speed at 10 meters), WS50M_MIN (Minimum Wind Speed at 50 meters), and WS10M_MIN (Minimum Wind Speed at 10 meters) in descending order.

Figure 5

Performing feature importance using permutation on a full RandomForestClassifier model, for which the mean accuracy decrease of each dataset feature was measured.



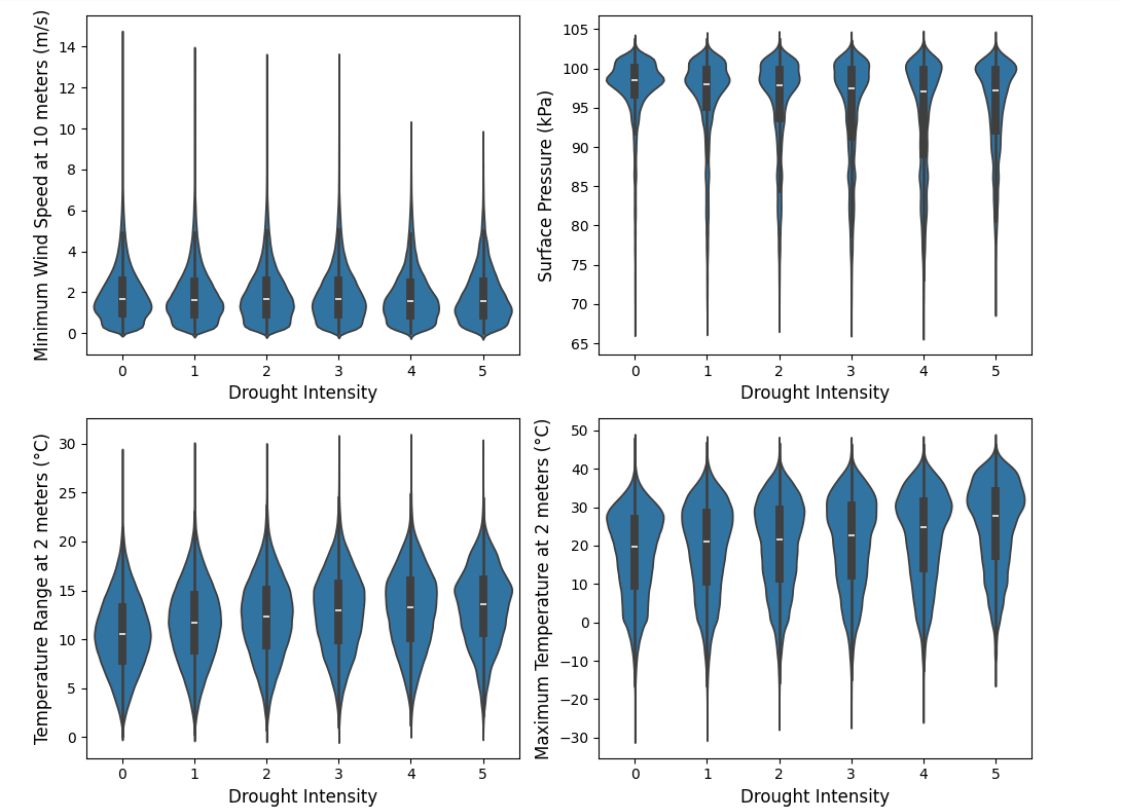
Violin plots of feature metric concentrations for each drought intensity

Based on the results of the feature importance using permutation, violin plots were created using the top three selected features, as well as the features that returned the lowest importance. These plots measured the concentration of various environmental features as drought intensity increased. They revealed visual insights into using Minimum Wind Speed at 10 meters, Surface Pressure, Temperature Range at 2 meters, and Maximum Temperature at 2 meters for predicting drought intensity.

As indicated in Figure 6 below, Minimum Wind Speed at 10 meters, which was the least important feature, remained stable from D2 to D4, and each graph had a similar shape/distribution.

Figure 6

Violin Plots analyzing trends of the least important feature and the top three most important features against drought intensity



Surface pressure, the topmost important feature, was concentrated around 95-105 kPa across all drought intensities. Across intensities, there was a difference in distribution, as each intensity number had a distinct shape.

The temperature range at 2 meters, which was ranked the second most important feature, had varied concentrations across intensities. As intensity increased, the temperature range also gradually increased, with D0 having its greatest concentration at around 10 and D5 having its greatest concentration at almost 15.

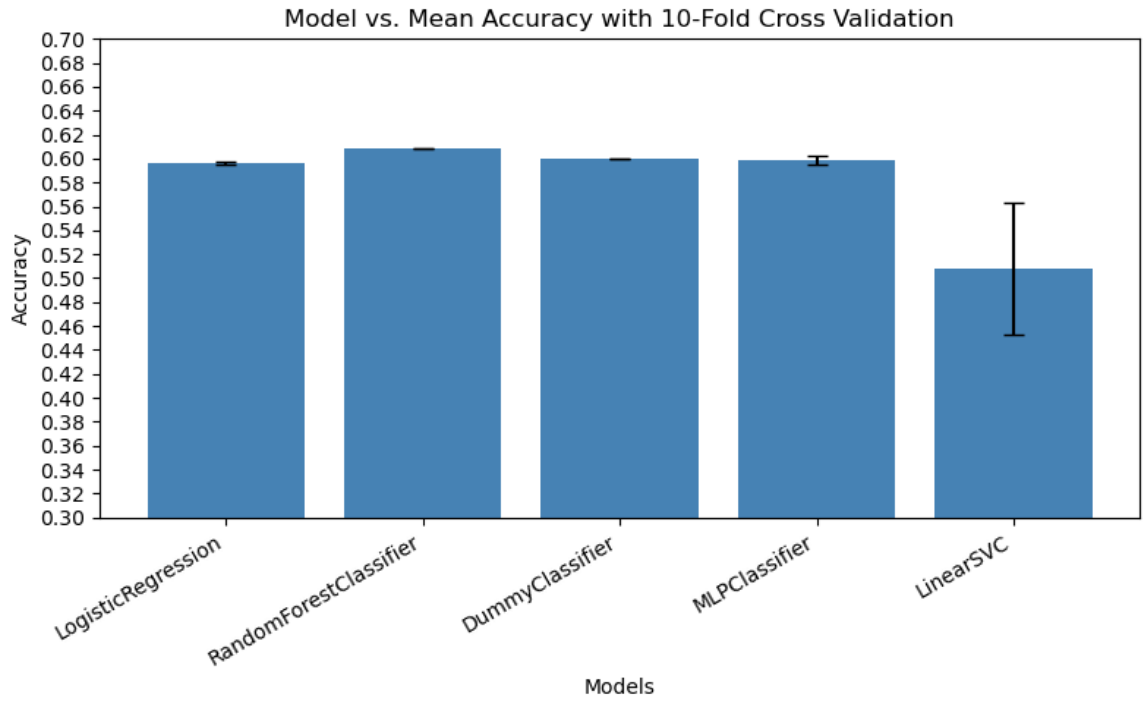
The maximum temperature at 2 meters, which was ranked the third most important feature, also had varying concentrations across intensities. D0 ranged from -10 to 40, while D5 ranged from 0 to 45, and there were slight shape differences among the intensity levels.

Cross-Validation Model Comparison

To select a model to measure accuracy with, five different models were compared using 10-fold cross-validation.

Figure 7

Accuracy of each model on 10-fold Cross-Validation



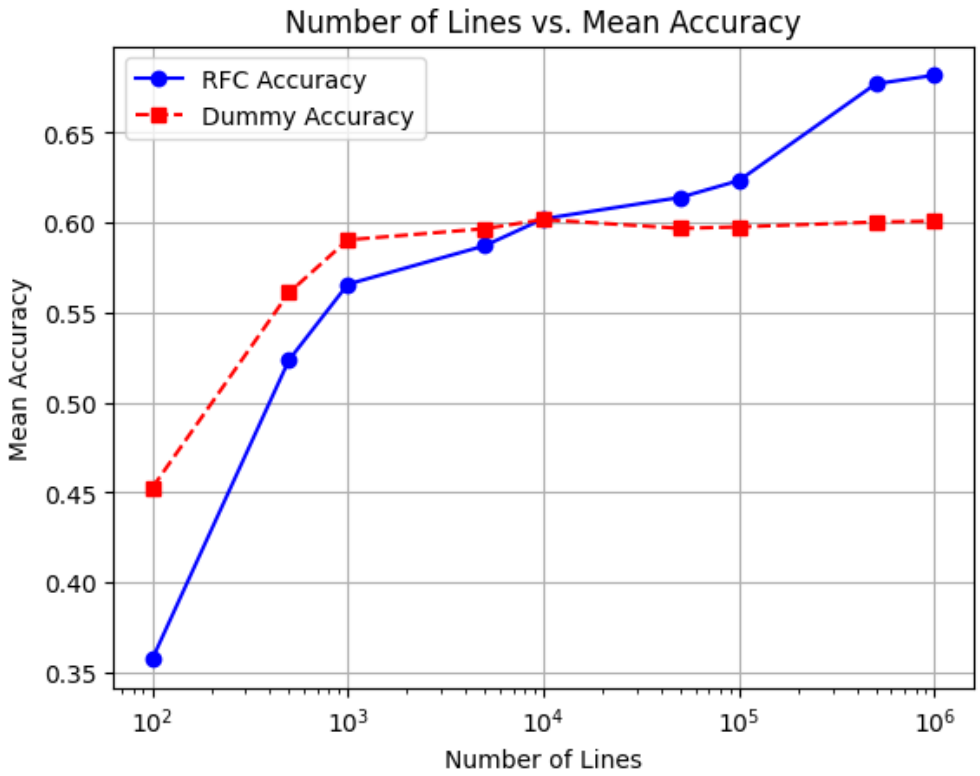
As indicated in Figure 7, LinearSVC performed with the lowest accuracy, at 45%. The other models performed with similar accuracy, with RandomForestClassifier having slightly greater accuracy at 60%. Based on the results, the RandomForestClassifier was chosen to run a set of time-series splits against the DummyClassifier.

Time-Series Split

Time-series splits were then performed with various cross-validation folds on the RandomForestClassifier and the DummyClassifier to establish a point-of-comparison baseline. The DummyClassifier was selected because it provides a naive baseline score, representing the accuracy of consistently predicting the most common drought intensity. Since "None" is the most frequent intensity, the classifier consistently predicted "None."

Figure 8

RandomForestClassifier accuracy comparison against DummyClassifier as lines included in the dataset increases



The models were evaluated using 6 time-series splits, and the number of lines increased exponentially, as indicated by Figure 8. Initially, the DummyClassifier reported higher accuracies than the RandomForestClassifier, but after 1000 lines, the DummyClassifier's performance began to plateau while the RandomForestClassifier continued to improve. The RandomForestClassifier's accuracy exceeds that of the DummyClassifier at 10000 lines.

Discussion

Throughout this paper, multiple tasks were performed using the DroughtED dataset to produce results with scientific applications. Feature importance was assessed using permutation, and a variety of visualizations were generated to examine how drought intensity changed when comparing different feature pairs. Violin plots were employed to observe how intensity values were distributed across different levels of each environmental feature. Time-series splits were applied to compare the accuracy between the DummyClassifier and the RandomForestClassifier.

Permutation results identified surface pressure as the most important feature in determining drought intensity. This finding aligns with the understanding that high-pressure systems suppress rainfall, making already dry regions drier (Bengali, 2023). Interestingly, among the top drought indicators listed by the Center for Climate and Energy Solutions, the only feature that appeared in the top five was temperature (Bengali, 2023). The way temperature is measured also varies. Permutation also indicated that the minimum wind speed at 10 meters was the feature with the least importance in determining drought intensity. Furthermore, the month was identified as the fourth most important feature, suggesting that seasonality may be an important factor to consider while predicting drought conditions.

While performing 10-fold cross-validation across the five model choices, the RandomForestClassifier model returned the greatest accuracy. There are some explanations for the results of the other selected models. For instance, the MultilayerPerceptron classifier (MLPClassifier), a neural network, was not expected to perform as well due to the dataset having already been organized and labeled. Therefore, it underperformed relative to the RandomForestClassifier. Another observation was the large standard deviation of the LinearSVC model. This can be attributed to the model being sensitive to the dataset imbalance and outliers within the samples.

The RandomForestClassifier consistently outperformed the DummyClassifier baseline on datasets with greater than 10000 lines when evaluated using time-series splits. This highlights how the unbalanced nature of the data created substantial performance differences between prediction methods. The model's performance over the baseline demonstrates its practical value for drought prediction in real-world applications.

A key limitation of the study was that the dataset did not include all relevant environmental indicators. Although 21 columns were available, environmental conditions extend far beyond this. Some of the columns—such as day, month, year, and date, as well as FIPS code—were primarily chronological or geographic identifiers and are more difficult to directly associate with drought intensity. While the month could hint at seasonal effects, most of these identifiers had limited application. Additionally, certain environmental features were split into different measurements (e.g., maximum temperature, minimum temperature, and surface temperature), which could complicate feature interpretation. Another limitation involved ensuring that the model relied only on past data, not future data, which was addressed by using time-series splits.

Future applications of this work involve implementing the predictive methods, models, and identified top features in various global regions. The visualizations provided here may serve as tools for local governments to track critical environmental indicators and anticipate increases in drought intensity.

Conclusion

This study successfully identified the most impactful environmental factors contributing to drought intensity using the DroughtED dataset. The feature importance analysis revealed that surface pressure, temperature range at 2 meters, and maximum temperature at 2 meters were the top predictors of drought intensity. By comparing multiple prediction methods, the RandomForestClassifier outperformed other tested models, as well as the DummyClassifier baseline, demonstrating the effectiveness of using this model in forecasting drought conditions.

The findings of this research extend beyond the ability of this data. By defining minimum environmental causes of drought and testing a resilient predictive model, the study provides a basis for early drought warning systems. This data can be applied worldwide to facilitate anticipatory policy planning, resource allocation, and strategic mitigation planning in water scarcity risk zones. Last, the study offers valuable tools for facilitating sustainable environmental planning with increased climatic variability.

References

4.2. Permutation feature importance. (n.d.). Scikit-learn. Retrieved May 27, 2025, from https://scikit-learn.org/stable/modules/permutation_importance.html

A new global database of meteorological drought events from 1951 to 2016. (2019). Journal of Hydrology: Regional Studies, 22, 100593. https://doi.org/10.1016/j.ejrh.2019.100593

Azimi, S. M. E., Sadatinejad, S. J., Malekian, A., & Jahangir, M. H. (2022). Application of artificial intelligence hybrid models for meteorological drought prediction. Natural Hazards, 116(2), 2565–2589. https://doi.org/10.1007/s11069-022-05476-0

Dikshit, A., & Pradhan, B. (2021). Interpretable and explainable AI (XAI) model for spatial drought prediction. Science of the Total Environment, 801, 149797. https://doi.org/10.1016/j.scitotenv.2021.149797

Drought classification. (n.d.). U.S. Drought Monitor. Retrieved April 27, 2025, from https://droughtmonitor.unl.edu/About/AbouttheData/DroughtClassification.aspx

DummyClassifier. (n.d.). Scikit-learn. Retrieved May 27, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

Explainable AI in drought forecasting. (2021). Machine Learning with Applications, 6, 100192. https://doi.org/10.1016/j.mlwa.2021.100192

Kikon, A., & Deka, P. C. (2022). Artificial intelligence application in drought assessment, monitoring and forecasting: A review. Stochastic Environmental Research and Risk Assessment, 36, 1197–1214. https://doi.org/10.1007/s00477-021-02129-3

LogisticRegression. (n.d.). Scikit-learn. Retrieved May 27, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Minixhofer, C. (2021). Predict droughts using weather & soil data. Kaggle. https://www.kaggle.com/cdminix/us-drought-meteorological-data

Minixhofer, C., Swan, M., McMeekin, C., & Andreadis, P. (2021). DroughtED: A dataset and methodology for drought forecasting spanning multiple climate zones. In ICML 2021 Workshop on Tackling Climate Change with Machine Learning. Climate Change AI. https://www.climatechange.ai/papers/icml2021/22/paper.pdf

Naumann, G., Dutra, E., Barbosa, P., Pappenberger, F., Wetterhall, F., & Vogt, J. V. (2014). Comparison of drought indicators derived from multiple data sets over Africa. Hydrology and Earth System Sciences, 18(5), 1625–1640. https://doi.org/10.5194/hess-18-1625-2014

New research finds rising heat driving western U.S. droughts. (n.d.). Drought.gov. Retrieved April 27, 2025, from https://www.drought.gov/news/new-research-finds-rising-heat-driving-western-us-droughts-2024-11-08

Oyounalsoud, M. S., Yilmaz, A. G., Abdallah, M., & Abdeljaber, A. (2024). Drought prediction using artificial intelligence models based on climate data and soil moisture. Scientific Reports, 14(1), 19700. https://doi.org/10.1038/s41598-024-46661-6

RandomForestClassifier. (n.d.). Scikit-learn. Retrieved May 27, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Vij, P., & Tiwari, A. (2025). AI-driven drought monitoring: Advanced machine learning techniques for early prediction. In SHS Web of Conferences (Vol. 216, p. 01024). EDP Sciences. https://doi.org/10.1051/shsconf/202521601024