Development of Machine Learning model for Drought Prediction

By: Anika Pallapothu

ABSTRACT

This research develops a machine learning model for drought prediction using the Random Forest algorithm, employing historical meteorological and soil data to deliver precise drought forecasts. Accurate drought prediction is essential for alleviating negative impacts on agriculture and water resources; however, conventional methods frequently lack precision. This study employs extensive data preprocessing, feature selection, and model training to develop a stable and interpretable predictive model. The algorithm, integrated with an interactive Streamlit application, allows stakeholders to submit data and receive real-time drought predictions. The evaluation criteria, such as accuracy, precision, and recall, demonstrate that the model successfully identifies the links between environmental variables and drought severity. The Random Forest model has robustness and interpretability, making it a significant asset for policymakers, agricultural planners, and researchers. This study also provides a user-friendly yet scientifically robust instrument for proactive drought management and highlights potential avenues for improved model precision and scalability.

Author Keywords. Drought Prediction, Machine Learning, Random Forest Model, Meteorological Data, Soil data

1. Introduction

1.1 Background and Motivation

Droughts profoundly affect agriculture, water supplies, and ecosystems, frequently leading to economic losses and environmental deterioration (Y. Song et al., 2024). Conventional prediction approaches lack precision and adaptability due to the intricacy of drought patterns and the multitude of environmental elements at play (Rezaiy & Shabri, 2024). Recent breakthroughs in machine learning have facilitated enhanced drought predictions by utilizing historical meteorological and soil data to develop efficient data-driven models (Ayinla & Abdulsalam, 2024a). In the present investigation advanced machine learning algorithm is utilized to create a drought prediction system that delivers high accuracy, thereby assisting stakeholders in making decisions on water management, crop planning, and risk mitigation.

1.2 Problem Definition

Drought prediction is the examination of extensive historical meteorological and soil data to forecast instances of water deficiency (Koutroulis et al., 2024). The main objective of this research is to develop a predictive model that can precisely evaluate drought scores, thereby offering early warnings to alleviate negative impacts on agricultural and water resources. Drought prediction is difficult because to the intrinsic complexity and variety of meteorological patterns, soil conditions, and geographical disparities (Zhang et al., 2024). This study seeks to tackle these difficulties by utilizing a Random Forest model, capable of managing extensive datasets with multiple attributes while ensuring interpretability and robustness. This issue is delineated as a supervised learning work, utilizing historical data to train the model for precise prediction of drought scores, ultimately aimed at facilitating proactive decision-making in drought-affected areas.

1.3 Objectives and Scope

The main goal is to create a precise machine learning model employing the Random Forest technique to forecast drought scores using historical weather and soil data. This project addresses both the technical sides of model development and the practical application by incorporating the model into an accessible interface using a Streamlit web application. The project includes data preparation, feature selection, model training, performance evaluation, and the development of an interactive platform allowing users to input data and receive real-time drought predictions. The project seeks to provide a scientifically rigorous and accessible tool by explicitly delineating its objectives and scope, so providing a significant resource for researchers, agricultural stakeholders, and policymakers in addressing and alleviating drought consequences.

2. Literature review

Various researchers have investigated on utilization of machine learning models for drought prediction (Ayinla & Abdulsalam, 2024b; Katipoğlu et al., 2024; Magallanes-Quintanar et al., 2024; TIWARI & Manthankumar P. Brahmbhatt, 2024; Tuğrul & Hinis, 2024). Katipoğlu et al. (2024) formulated a machine learning model for drought prediction by combining Artificial Neural Networks (ANN) with metaheuristic optimization methods, including Firefly Algorithm (FFA), Genetic Algorithm (GA), and Particle Swarm Optimization (PSO). The developed model predicted the Streamflow Drought Index (SDI) during hydrological droughts with a 1-month lead time in the Konya closed basin, employing lag values derived from partial autocorrelation function graphs. The PSO-ANN and FFA-ANN hybrid models exhibited the greatest accuracy, with Coefficient of Determination (R^2) values between 0.443 and 0.931.

Ayinla & Abdulsalam (2024) investigated an innovative method that integrates K-means clustering and the Gradient Boosting Algorithm (KGBA) with Principal Component Analysis (PCA) for predicting droughts. The KGBA model, employing a dataset of 2,756,796 US Drought Monitor records from 2000 to 2016, exhibited elevated precision and recall rates, especially in predicting extreme and exceptional droughts, with an overall accuracy of 46%. Their model surpassed conventional methods, underscoring its potential to improve drought mitigation strategies. Tiwari & Brahmbhatt (2024) created machine learning models, specifically ANN and M5 model trees, to forecast drought indices, specifically the one-month timescale standardized precipitation index (SPI-1) and standardized precipitation evapotranspiration index (SPEI-1) for the central Gujarat region of India. The models employed a 30-year dataset (1986-2015) and were trained with the Levenberg-Marquardt technique. They demonstrated that ANN models surpassed M5 models, especially in predicting SPI-1, hence illustrated their efficacy in drought forecasting. Magallanes-Quintanar et al. (2024) established an Auto-Machine-Learning methodology employing ANN models to forecast the SPI across four regions in Zacatecas, Mexico. Climatological time-series data from 1979 to 2020 functioned as prediction factors. Their models exhibited robust predictive skills, with performance indicators such as Mean Squared Error between 0.0296 and 0.0388, Mean Absolute Error ranging from 0.1214 to 0.1355, and R^2 spanning 0.9342 to 0.9584, hence facilitating drought prediction.

Tuğrul & Hinis (2024) developed a machine learning model for drought prediction using the SPI and various algorithms, including support vector machines (SVM), ANNs, random forest, and decision tree. The M04 model, which incorporated SPI, time steps, and delayed data, yielded the best performance. SVM demonstrated the highest accuracy both before and after applying wavelet transformation, achieving a Nash–Sutcliffe efficiency (NSE) of 0.9942, root mean square error (RMSE) of 0.0764, and R^2 of 0.9971. Kang & Byun (2024) introduced a multi-scale groundwater drought prediction model that employs deep learning, particularly long short-term memory (LSTM) networks. It forecasted both zonal average and point-scale values for the standardized groundwater level index (SGI) utilizing hydrometeorological variables, such as temperature, precipitation, and vapor pressure deficit. The model was evaluated on Jeju Island, with high accuracy and a Nash–Sutcliffe efficiency coefficient more than 0.9 and a RMSE below 0.3, thereby enabling efficient groundwater management procedures. Liu et al. (2024) created a hybrid hydrological-deep learning model to forecast future bivariate hydrological drought features in 179 catchments in China. They employed five bias-corrected global climate model outputs, three shared socioeconomic pathways, and a random forest model to assess meteorological influences on daily streamflow. Their model attained a Kling–Gupta efficiency over 0.8 in 161 catchments, illustrating its efficacy in forecasting drought length and severity, hence tackling the intricacies of hydrological drought projections amid climate change.

Xu et al. (2024) introduced a machine learning stacking ensemble method for drought prediction, employing Precipitation Estimation from Remotely Sensed Information using Artificial Neural Network-Climate Data Record (PERSIANN-CDR), MODIS remote sensing products, and climate zoning data to assess the SPEI-3 across nine sub-regions in China. They determined that the CatBoost Regressor is the most proficient meta-model, with R^2 values of 0.9065 and 0.8218 in the eastern and western regions, respectively, indicating strong seasonal drought monitoring efficacy. Duong et al. (2024) constructed Machine Learning models, namely Gradient Boosting and Extreme Gradient Boosting (XGBoost), to forecast the SPEI in the Mekong Delta. Their models integrated diverse climate elements using data from 11 meteorological sites spanning 1990 to 2022. The findings demonstrated that XGBoost much surpasses conventional forecasting techniques, attaining R^2 values ranging from 0.90 to 0.94 for 1-month predictions, thus improving drought prediction precision and facilitating superior drought management approaches.

Thus, the reviewed literature emphasizes the progression of drought prediction techniques, transitioning from conventional statistical methods to sophisticated machine learning models. Conventional methods offer basic insights into drought patterns but frequently lack the adaptability and precision required to tackle the intricacies of contemporary climate data. Tree-based machine learning models, such as Random Forest, have shown significant enhancements in predicted accuracy and robustness by effectively handling non-linear relationships and numerous input features. Nonetheless, deficiencies persist in model interpretability, scalability, and real-time application, which the present research seeks to rectify. This research also utilizes lessons from prior studies to enhance existing information and create a comprehensive drought prediction system that combines predictive efficacy with practical applicability, paving the way for more accurate and effective drought forecasts.

3. Methodology

3.1 Data Collection

The methodology for acquiring pertinent data to train the drought prediction model is discussed in this section. Precise drought forecasting depends on extensive datasets that encompass the several elements affecting drought conditions, including precipitation, temperature, soil moisture, and humidity. This study utilized historical meteorological and soil data obtained from publicly accessible datasets, including the U.S. drought and meteorological dataset available on Kaggle.

Metadata of the meteorological dataset is Minimum Wind Speed at 10 Meters (m/s) is WS10M_MIN, Specific Humidity at 2 Meters (g/kg) is QV2M, Temperature Range at 2 Meters (C) is T2M_RANGE, Wind Speed at 10 Meters (m/s) is WS10M, Temperature at 2 Meters (C) is T2M, Minimum Wind Speed at 50 Meters (m/s) is WS50M_MIN, Maximum Temperature at 2 Meters (C) is T2M_MAX, Wind Speed at 50 Meters (m/s) is WS50M, Earth Skin Temperature (C) is TS, Wind Speed Range at 50 Meters (m/s) is WS50M_RANGE, Maximum Wind Speed at 50 Meters (m/s) is WS50M_MAX, Maximum Wind Speed at 10 Meters (m/s) is WS10M_MAX, Wind Speed Range at 10 Meters (m/s) is WS10M_RANGE, Surface Pressure (kPa) is PS, Dew/Frost Point at 2 Meters (C) is T2MDEW, Minimum Temperature at 2 Meters (C) is T2M_MIN, Wet Bulb Temperature at 2 Meters (C) is T2MWET, and Precipitation (mm day-1) is PRECTOT. Metadata for soil data is US county FIPS code is fips, Latitude is lat, Longitude is lon, Median elevation (meters) is elevation, 0 % ≤ slope ≤ 0.5 % is slope1, 0.5 % ≤ slope ≤ 2 % is slope 2, 2 % ≤ slope ≤ 5 % is slope3, 5 % ≤ slope ≤ 10 % is slope 4, 10 % ≤ slope ≤ 15 % is slope5, 15 % ≤ slope ≤ 30 % is slope6.

This dataset was chosen for its comprehensive coverage of pertinent features and its temporal depth, facilitating a thorough investigation of drought changes throughout time. The data collection method entailed filtering and structuring the raw data into a format appropriate for machine learning applications. The data quality and relevance directly influence the model's predicted accuracy and efficacy in practical applications.

3.2 Data Preprocessing

Data preparation is crucial in machine learning, as it ensures that the dataset is clean, consistent, and organized to improve the model's learning efficacy. This section outlines the essential procedures done to prepare the raw data for efficient model training. Initially, metrological and soil datasets are merged. The project commenced with pre-processing, which involved addressing missing values and eliminating extraneous features to minimize data noise. Subsequently, data normalization and scaling were implemented to standardize variables with disparate ranges, such as temperature and precipitation, ensuring comparability and preventing any one feature from unduly affecting the model using MinMaxScaler.

Furthermore, category data was encoded, converting qualitative information into a numerical format appropriate for the Random Forest method. The pre-processing processes optimized the dataset for performance, enabling the model to concentrate on the most significant patterns in the data, hence enhancing accuracy and reliability in drought prediction.

3.3 Model Development

Choosing an appropriate model is essential for obtaining trustworthy predictions, particularly due to the complexity of drought patterns affected by multiple environmental factors. Random Forest was selected for its resilience, capacity to manage extensive datasets with numerous features, and its efficacy in identifying non-linear relationships within the data. This ensemble technique generates many decision trees during training, with each tree casting a vote on the result, so improving overall model stability and mitigating the risk of overfitting. Furthermore, Random Forest offers feature relevance scores, enabling the identification of variables that most significantly influence drought conditions. This interpretability aids in enhancing the model and offers insights into environmental factors that influence drought severity. Consequently, the use of Random Forest corresponds with the project's aims of precision, resilience, and clarity in forecasting drought ratings.

To train the model, 'score', 'fips', 'date' columns are dropped. The target column is ‘score’. The model training process entailed partitioning the dataset into training and validation subsets to facilitate an impartial evaluation of the model's predicted performance. The ‘test_size’ is kept as 0.2 and ‘random_state’ as 0. The Random Forest model was subsequently trained on the training set, where it acquired the ability to discern patterns in historical meteorological and soil data linked to drought events. Hyperparameter tuning was conducted to refine the model's parameters, including the number of trees and maximum depth, in order to attain an equilibrium between model complexity and predictive accuracy. Random forest hyper-parameter ‘n_estimators’ is 10. The Random Forest machine learning model is built using these settings and saved in .pkl file format.

The feature importance plot, derived from Gini importance (or mean decrease in impurity), as shown in Figure 1 illustrates the relative significance of different features utilized by the Random Forest model for drought prediction. Features with elevated Gini relevance exert a more substantial influence on the model's decisions. In this diagram:

1. W500M_MAX is the most significant feature, suggesting that this variable, presumably associated with wind speed at a 500 m elevation, is vital for predicting drought conditions in the dataset.
2. Additional parameters, including T2M_MAX, T2M_MIN, W510M_MAX, and W510M_MIN, are also significantly relevant. The temperature and wind-related variables indicate that weather conditions, specifically temperature and wind patterns, are crucial predictors in the model.
3. Elevation and latitude are fairly significant, suggesting that geographic features affect drought forecasts.
4. Land cover classifications, including URB_LAND and FOR_LAND, also influence the forecasts, but to a diminished degree. This indicates that land use and vegetation type could influence drought susceptibility.
5. Numerous more characteristics, such as slopes, particular land cover types, and specific soil properties (SQ1, SQ2, etc.), exhibit diminished relevance scores. These influences the model but possess diminished individual significance.

The predominance of W500M_MAX indicates a significant correlation between drought conditions and high-altitude wind patterns, which may be beneficial in comprehending environmental impacts on drought. This feature importance analysis aids in enhancing the model by concentrating on critical predictors to augment model interpretability and performance potential.

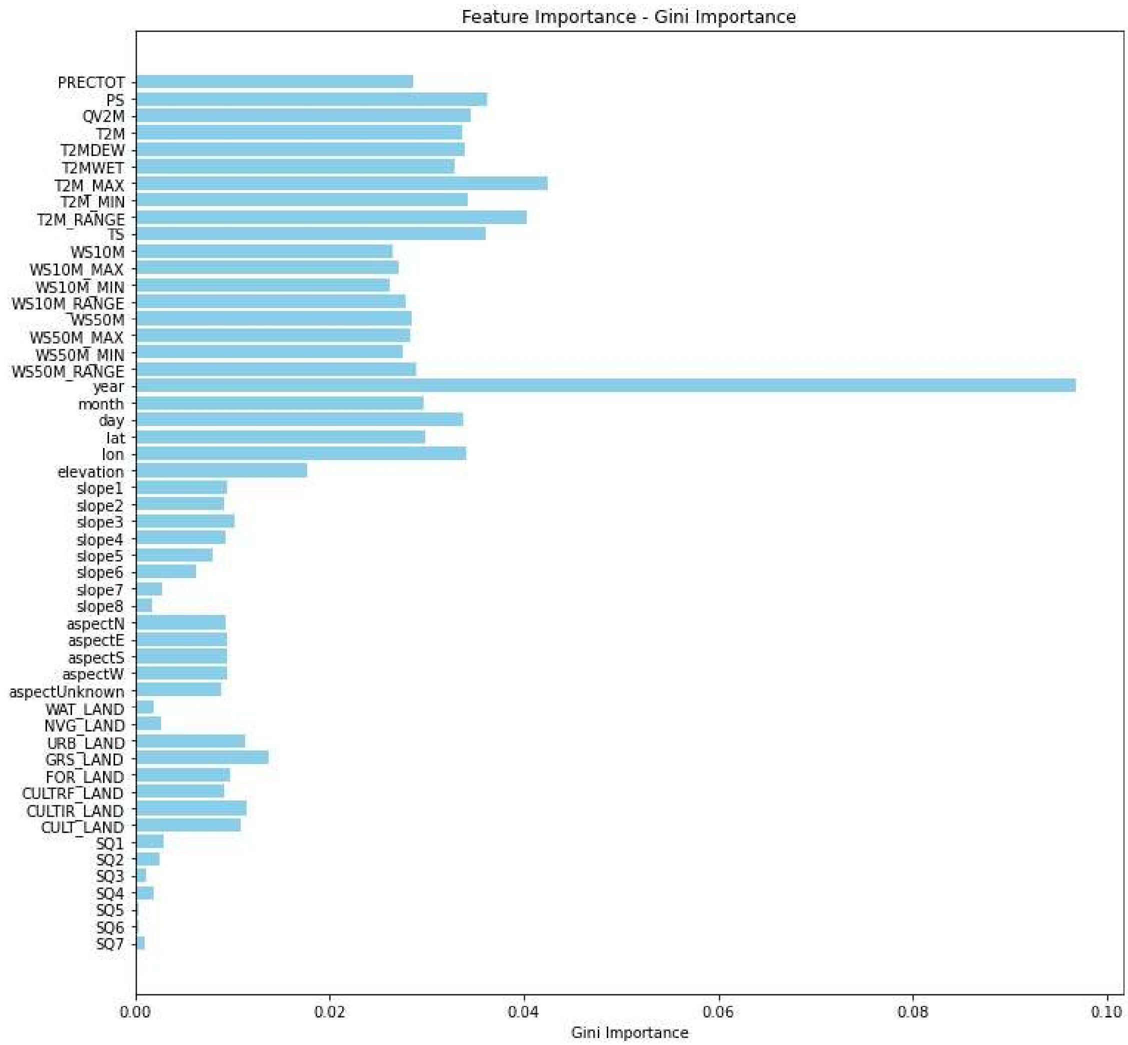

Figure 1. Feature importance plot

3.4 Evaluation Metrics

The predictions are saved in ‘predictions.csv’ file. The ‘score’ column consists of 6 classes, i.e. 0, 1, 2, 3, 4, 5 for six levels of drought. The model underwent validation with the hold-out validation set, employing metrics including accuracy, precision, and recall to evaluate its performance upon completion of training. The validation measures offer insight into the model's generalizability, ensuring it works effectively on historical data while demonstrating robust predicting capability for future drought circumstances. In the present study, accuracy score of 0.768, precision score of 0.75, and recall of 0.768 was obtained for test dataset.

Figure 2 displays a confusion matrix that compares the true labels (actual values) with the expected labels (predictions) generated by the model. Each cell in the matrix denotes the frequency with which a specific true label was predicted as a particular label. Here is an analysis of the matrix:

- The diagonal cells (from top-left to bottom-right) indicate the number of accurate predictions, wherein the predicted label corresponds with the true label. Elevated values in these cells signify superior performance. The model accurately predicted 134,073 instances for label "0" and 16,650 instances for label "1", among others.

- Off-diagonal cells indicate misclassifications, wherein the projected label diverges from the actual label. Specifically, there were 3,383 occurrences in which the actual label was "0" while the model predicted "1", and 18,819 occurrences where the actual label was "1" while the model predicted "0".
- Class-specific performance can be deduced by analyzing each row. For example:
    - Label "0" exhibits a substantial correct count (134,073), indicating strong predictive accuracy for this category.
    - Label "1" exhibits several misclassifications, with numerous instances erroneously projected as "0" (18,819), suggesting potential challenges in differentiating between these two categories.
    - The color coding indicates the magnitude of each cell's value. Darker hues indicate lesser quantities, whilst brighter hues signify bigger quantities. This visual depiction facilitates the rapid identification of the model's strengths and flaws across several labels.
- The confusion matrix elucidates the model's performance across each class and identifies areas of misclassification.

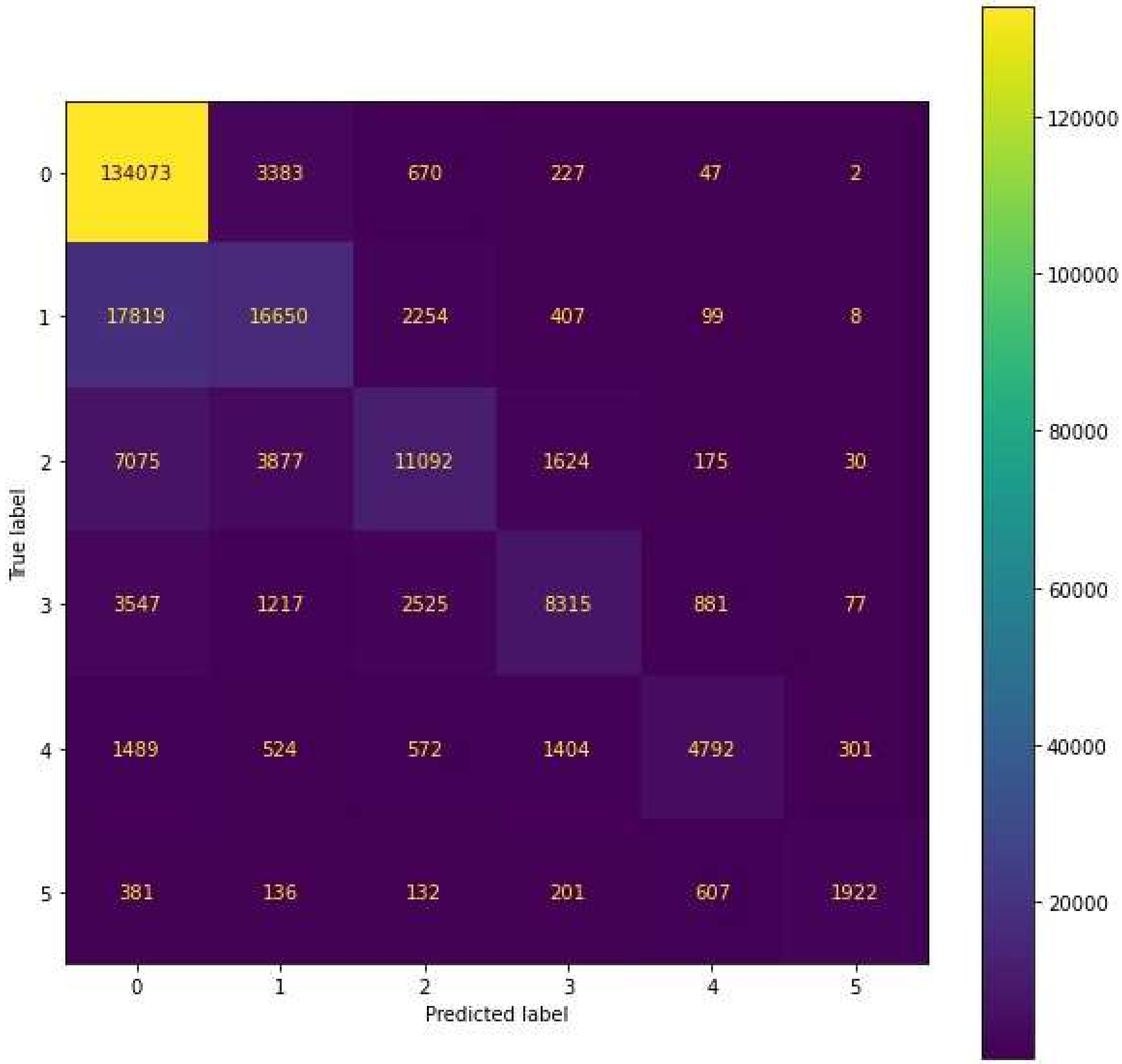

Figure 2. Confusion Matrix

Figure 3 illustrates a predicted score distribution plot, depicting the frequency of expected scores across various classes (0 to 5) derived from a model's output. This style of figure elucidates the distribution of model predictions, aiding in the comprehension of which classes the model predominantly predicts. The following are the principal observations:

- Class 0 Dominance: The predominant number of forecasts is classified as class 0, reaching 300,000 in total. This indicates that the model predominantly favors class 0, likely due to class 0 being the most abundant in the dataset or the model's tendency to overfit to it. This may suggest a skewed dataset in which class 0 is disproportionately represented in the training data.
- Classes 1, 2, and 3 exhibit markedly lower counts relative to class 0, signifying that these predictions are less prevalent. Classes 4 and 5 exhibit an even lower number of predictions, with class 5 demonstrating the minimal count overall. This distribution indicates that the model may have difficulty predicting these higher classes or that these classes are inadequately represented.
- The smooth curve superimposed on the bars illustrates the density estimation, so elucidating the distribution pattern more distinctly. It demonstrates a bias towards lower socioeconomic strata, diminishing as we ascend to higher classes.
- This plot reveals a significant concentration of predictions in the lower class, especially class 0, with comparatively few occurrences in the higher classes. This distribution may result from class imbalance within the dataset or model bias. Resolving this issue may require the application of strategies such as class weighting, resampling, or model optimization to enhance prediction distribution throughout all classes.

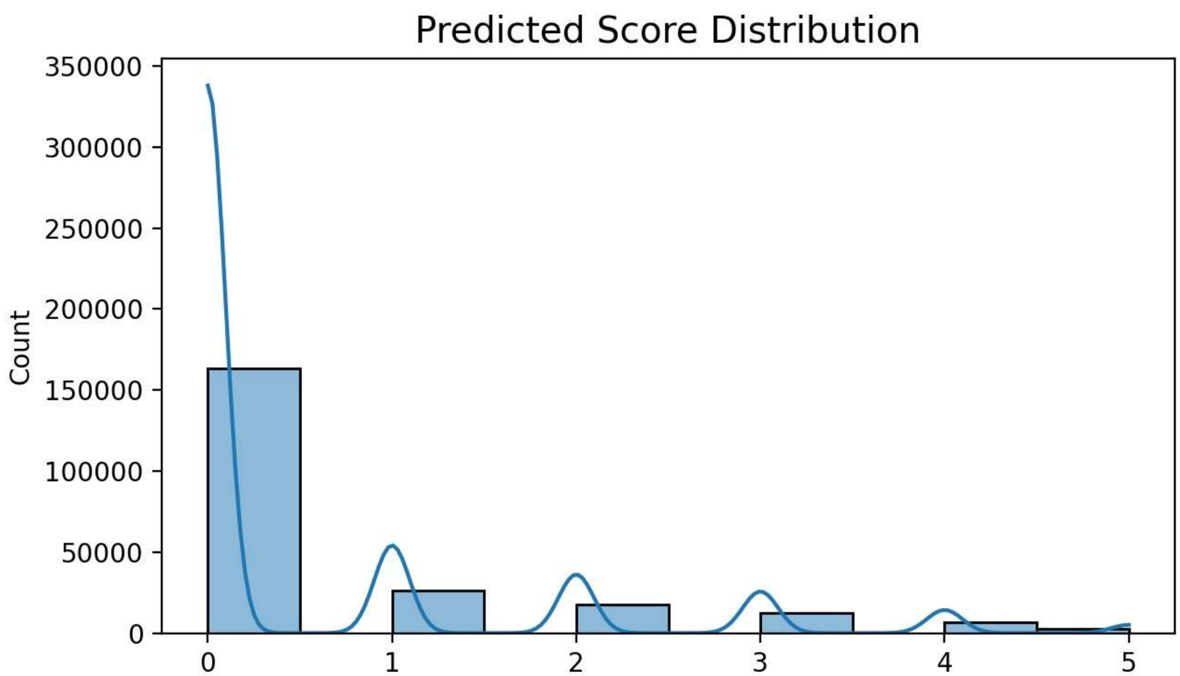

Figure 3. Predicted score distribution

4. System Implementation

4.1 Data Preprocessing script

This section details the creation and operation of a custom Python script designed to automate data preparation for the drought prediction model. This script was developed to manage substantial quantities of raw meteorological and soil data through the consistent and efficient execution of fundamental preprocessing procedures. The script commences by loading data files, thereafter cleaning and filtering the data by rectifying missing values and eliminating superfluous characteristics that may introduce noise. It also executes data normalization and scaling, ensuring that numerical features such as temperature, precipitation, and soil moisture are standardized to a uniform scale. Categorical variables are encoded to ensure interoperability with the machine learning model. The processed data is subsequently stored in a structured format, prepared for model training and evaluation. The automation of these stages by the preprocessing script optimizes the data preparation pipeline, minimizes human error, and improves the reproducibility of results, which is essential for the system's reliability and applicability in future endeavors.

4.2 Model Training process

This section outlines the procedures for training the Random Forest model using the preprocessed drought dataset. The procedure commences with the ingestion of the sanitized and organized data, subsequently partitioning it into training and testing subsets to enable a rigorous assessment. To enhance the model's predictive performance, hyperparameter tuning is conducted by modifying parameters such the number of trees, maximum tree depth, and minimum samples necessary for node splitting. This tuning seeks to equilibrate model complexity with performance, mitigating overfitting and improving generalizability. During training, the Random Forest algorithm constructs several decision trees, each trained on random subsets of the data, so forming an ensemble that generates predictions via a majority voting process. This ensemble method capitalizes on the advantages of individual trees while mitigating vulnerability to noise and volatility in the data. Upon completion of training, the model is preserved in a serialized manner, rendering it suitable for deployment in the drought prediction application. This systematic training procedure guarantees the model's precision, efficacy, and adaptability for practical prediction assignments.

4.3 Streamlit App for user interaction and predictions

This section outlines the creation of an interactive web application enabling users to interact with the drought prediction model in real time. This application, developed with the Streamlit framework, offers a user-friendly interface for uploading test datasets and obtaining prompt drought predictions. The application is engineered for accessibility, allowing even non-technical users to utilize the model's predictive functionalities. The application preprocesses the data for compatibility, subsequently inputting it into the trained Random Forest model, which generates drought scores and visual representations of forecast patterns and essential performance indicators upon uploading a dataset. These representations offer consumers a lucid comprehension of the model's predictions and the fundamental elements influencing them. This Streamlit program functions as a conduit between the intricate predictive model and a practical, user-friendly interface, enabling users such as farmers, academics, and policymakers to make informed decisions based on real-time drought forecasts.

1. Result and Analysis

5.1 Model evaluation results

This section presents the performance metrics and analysis of the Random Forest model following its training and testing on the drought dataset. The model's efficacy was assessed by critical metrics including accuracy, precision, and recall, offering a thorough perspective on its predictive quality. The elevated value of evaluation metrics demonstrate that the model effectively elucidates the correlations between meteorological and edaphic factors and drought intensity. Furthermore, the F1-score was employed to evaluate the equilibrium between precision and recall, guaranteeing that the model exhibits constant performance throughout varying degrees of drought severity. The results confirm the model's robustness and accuracy, establishing it as a dependable instrument for drought prediction. This assessment underscores the model's strengths while offering insights into future enhancements, including the refinement of features or the exploration of alternative models to further increase predictive accuracy.

5.2 Visualizations of predictions and metrics as displayed in the Streamlit app

The Streamlit application offers many charts and graphs that illustrate projected drought scores, facilitating users' comprehension of the data at a glance. Line graphs depict drought patterns over time, demonstrating the variation of expected scores in relation to changing weather and soil conditions. Bar charts and heatmaps are incorporated to illustrate essential performance measures, including accuracy, precision, and recall, providing users with insights into the model's dependability. These visualizations enable users to compare actual and expected values and analyze factors affecting drought severity, providing a clear insight into the model's decision-making process. The application enables users—be they researchers, agricultural planners, or policymakers—to make informed, data-driven decisions with assurance by rendering predictions and measurements clearly available.

6. Discussion

6.1 Analysis of the model's strengths and weaknesses

The Random Forest model's principal strength is in its robustness in managing extensive datasets with numerous features, allowing it to discern intricate correlations between environmental variables and drought severity. The ensemble characteristic of Random Forest mitigates the danger of overfitting, enhancing the model's generalizability to novel data. Moreover, its feature importance functionality provides transparency by identifying the variables that most significantly affect drought conditions, serving as a valuable resource for those seeking insights into environmental factors. Nevertheless, the model possesses many limits, particularly its substantial computational expense, which may result in diminished performance with big datasets. This constraint may impact its efficacy in real-time applications. Moreover, although Random Forest demonstrates commendable performance on the existing dataset, its accuracy may be enhanced by exploring alternative algorithms, such as gradient boosting or deep learning methods, particularly in scenarios where intricate variable interactions are critical. The present research highlights the model's dependability and clarity while pinpointing opportunities for improvement to further augment its predicted precision and efficacy.

6.2 Comparison with other potential models

Gradient boosting machines, such as XGBoost, and deep learning methodologies present viable alternatives owing to their ability to identify intricate, non-linear correlations in extensive datasets (Niazkar et al., 2024; C. E. Song et al., 2024). Gradient boosting models frequently yield superior accuracy compared to Random Forest by systematically rectifying errors; however, they necessitate more meticulous tuning and exhibit increased susceptibility to overfitting. Deep learning models, including neural networks, may improve predictive accuracy by identifying complex patterns and connections among variables; nevertheless, they require extensive datasets and significant processing resources, which may limit their applicability for real-time predictions (Arash Tashakkori et al., 2024; Reddy et al., 2024). Furthermore, support vector machines (SVM) and k-nearest neighbors (k-NN) are evaluated, however their efficacy may diminish with high-dimensional data characteristic of drought datasets (Choesang et al., 2023; Simarmata et al., 2024). Although Random Forest provides a compromise between interpretability and accuracy, its comparison with alternative methods highlights the trade-offs in model complexity, computational efficiency, and interpretability, informing future investigations into models most appropriate for implementing drought prediction across various contexts.

6.3 Limitations of the current approach and possible improvements

A constraint is the Random Forest model's computing requirements, especially during training, which may impede scalability for extensive datasets or real-time applications. Furthermore, although Random Forest achieves excellent accuracy, its dependence on historical data may hinder its ability to adjust to swift changes in climatic patterns induced by global warming, thus compromising prediction reliability. The model also lacks the capability for spatiotemporal analysis, which would enable it to consider spatial changes in drought patterns. Future research could include sophisticated approaches, such as deep learning, particularly recurrent neural networks (RNNs) or convolutional neural networks (CNNs), which are adept at processing sequential and spatial data to address these constraints. Incorporating supplementary real-time data sources, including satellite photography and remote sensing data, may improve the model's adaptability to climatic fluctuations. Finally, establishing an automated pipeline for hyperparameter optimization and model updates will enhance performance consistency and flexibility, hence rendering the system more resilient for prolonged deployment in varied environmental settings.

7. Conclusion

The Random Forest-based machine learning model developed in the present study demonstrates efficacy in drought prediction, providing significant accuracy and user-friendliness for stakeholders in drought-impacted regions. The model delivers timely drought forecasts by analyzing historical meteorological and soil data, so facilitating data-driven decisions for resource management and risk mitigation. While Random Forest provides interpretability and resilience, it has problems like high computational requirements and susceptibility to data imbalance. Subsequent study ought to investigate sophisticated methods such as gradient boosting and deep learning to augment model accuracy and scalability. Furthermore, incorporating real-time data sources, such as satellite imagery, along with spatiotemporal modeling tools, might enhance adaptation to changing climatic patterns. This model establishes a basis for effective and precise drought prediction, overcoming significant limitations in current methodologies and emphasizing opportunities for ongoing advancement.

Conflict of Interest

There is no conflict of interests.

Acknowledgement

I gratefully acknowledge the University of San Francisco, California, for organizing the DSCO23 datathon, providing essential guidance on the dataset, and offering the opportunity to engage in atmospheric research projects focused on air pollution and drought prediction.

References

Arash Tashakkori, Niloufar Erfanibehrouz, Shahin Mirshekari, Abolfazl Sodagartojgi, & Vatsal Gupta. (2024). Enhancing stock market prediction accuracy with recurrent deep learning models: A case study on the CAC40 index. *World Journal of Advanced Research and Reviews*, *23*(1), 2309–2321. https://doi.org/10.30574/wjarr.2024.23.1.2156

Ayinla, B., & Abdulsalam, R. (2024a). Exploring a Novel Approach of K-mean Gradient Boosting Algorithm with PCA for Drought Prediction. *American Journal of Data Mining and Knowledge Discovery*, *9*(1), 1–19. https://doi.org/10.11648/j.ajdmkd.20240901.11

Ayinla, B., & Abdulsalam, R. (2024b). Exploring a Novel Approach of K-mean Gradient Boosting Algorithm with PCA for Drought Prediction. *American Journal of Data Mining and Knowledge Discovery*, *9*(1), 1–19. https://doi.org/10.11648/j.ajdmkd.20240901.11

Choesang, T., Ryntathiang, S., Jacob, B. A., Krishnan, B., & Kokatnoor, S. A. (2023). *Drought Prediction—A Comparative Analysis of Supervised Machine Learning Techniques* (pp. 295–307). https://doi.org/10.1007/978-981-99-2468-4_23

Duong, H. H., Phong, N. D., Ha, T. L., Tang, T. D., Trinh, T. N., Nguyen, T. M., & Nguyen, T. M. (2024). *Application of Machine Learning to Forecast Drought Index for the Mekong Delta*. https://doi.org/10.20944/preprints202405.1693.v1

Kang, D., & Byun, K. (2024). Development of a Multi-Scale Groundwater Drought Prediction Model Using Deep Learning and Hydrometeorological Data. *Water*, *16*(14), 2036. https://doi.org/10.3390/w16142036

Katipoğlu, O. M., Ertugay, N., Elshaboury, N., Aktürk, G., Kartal, V., & Pande, C. B. (2024). A novel metaheuristic optimization and soft computing techniques for improved hydrological drought forecasting. *Physics and Chemistry of the Earth, Parts A/B/C*, *135*, 103646. https://doi.org/10.1016/j.pce.2024.103646

Koutroulis, A., Grillakis, M., Gosling, S., Schmied, H. M., Burek, P., Kou-Giesbrecht, S., Qi, W., Pokhrel, Y., Satoh, Y., Tsanis, I., Stein, L., & Thiery, W. (2024). *Examining the contribution of climate change on global soil moisture drought characteristics*. https://doi.org/10.5194/egusphere-plinius18-46

Liu, R., Yin, J., Slater, L., Kang, S., Yang, Y., Liu, P., Guo, J., Gu, X., Zhang, X., & Volchak, A. (2024). Machine-learning-constrained projection of bivariate hydrological drought magnitudes and socioeconomic risks over China. *Hydrology and Earth System Sciences*, *28*(14), 3305–3326. https://doi.org/10.5194/hess-28-3305-2024

Magallanes-Quintanar, R., Galván-Tejada, C. E., Galván-Tejada, J. I., Gamboa-Rosales, H., Méndez-Gallegos, S. de J., & García-Domínguez, A. (2024). Auto-Machine-Learning Models for Standardized Precipitation Index Prediction in North–Central Mexico. *Climate*, *12*(7), 102. https://doi.org/10.3390/cli12070102

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environmental Modelling & Software*, *174*, 105971. https://doi.org/10.1016/j.envsoft.2024.105971

Reddy, C. V., Reddy, N. V. U., Sharma, S., Kumar, J. P., Lakhanpal, S., & Albawi, A. (2024). Advanced Deep Learning Architectures for Real-Time Human Emotion Recognition and Behavioral Prediction. *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, 1306–1311. https://doi.org/10.1109/IC3SE62002.2024.10592940

Rezaiy, R., & Shabri, A. (2024). Enhancing drought prediction precision with EEMD-ARIMA modeling based on standardized precipitation index. *Water Science & Technology*, *89*(3), 745–770. https://doi.org/10.2166/wst.2024.028

Simarmata, J. E., Weber, G.-W., & Chrisinta, D. (2024). Performance Evaluation of Classification Methods on Big Data: Decision Trees, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines. *Jurnal Matematika, Statistika Dan Komputasi*, *20*(3), 623–638. https://doi.org/10.20956/j.v20i3.32970

Song, C. E., Li, Y., Ramnani, A., Agrawal, P., Agrawal, P., Jang, S.-J., Lee, S.-S., Rosing, T., & Kang, M. (2024). 52.5 TOPS/W 1.7GHz Reconfigurable XGBoost Inference Accelerator Based on Modular-Unit-Tree with Dynamic Data and Compute Gating. *2024 IEEE Custom Integrated Circuits Conference (CICC)*, 1–2. https://doi.org/10.1109/CICC60959.2024.10529017

Song, Y., Joo, J., Kim, H., & Park, M. (2024). Development and Applicability Evaluation of Damage Scale Analysis Techniques for Agricultural Drought. *Water*, *16*(10), 1342. https://doi.org/10.3390/w16101342

TIWARI, M., & Manthankumar P. Brahmbhatt. (2024). Forecasting Drought Indices using Artificial Neural Network and M5 Model Tree Techniques in Middle Gujarat Region of India. *Journal of Agricultural Engineering (India)*, *61*(3), 413–431. https://doi.org/10.52151/jae2024613.1856

Tuğrul, T., & Hinis, M. A. (2024). Improvement of drought forecasting by means of various machine learning algorithms and wavelet transformation. *Acta Geophysica*. https://doi.org/10.1007/s11600-024-01399-z

Xu, X., Chen, F., Wang, B., Harrison, M. T., Chen, Y., Liu, K., Zhang, C., Zhang, M., Zhang, X., Feng, P., & Hu, K. (2024). Unleashing the power of machine learning and remote sensing for robust seasonal drought monitoring: A stacking ensemble approach. *Journal of Hydrology*, *634*, 131102. https://doi.org/10.1016/j.jhydrol.2024.131102

Zhang, J.-L., Huang, X.-M., & Sun, Y.-Z. (2024). Multiscale spatiotemporal meteorological drought prediction: A deep learning approach. *Advances in Climate Change Research*, *15*(2), 211–221. https://doi.org/10.1016/j.accre.2024.04.003