Synopsis on Drought Prediction with Machine Learning using Weather and Soil Data

Title: *Drought Prediction with Machine Learning using Weather and Soil Data*

Introduction and Objective:

Drought is a serious natural disaster that can severely impact agriculture, water resources, and ecosystems. This project aims to create a machine learning model that forecasts the severity of drought by analyzing meteorological and soil data, sorting it into one of six defined levels. The model is designed to aid in early detection, ultimately leading to improved planning and resource management strategies.

Dataset Used:

The dataset used in this study was sourced from Kaggle, built using open data offered by the NASA POWER Project and the authors of the US Drought Monitor. It includes drought severity levels for U.S. countries, classified into six categories-None(no drought), D0(abnormally dry), D1(moderate drought), D2(severe drought), D3(extreme drought), D4(exceptional drought) and previous 90 days of 18 meteorological features.

1. *Weather/meteorological data is stored in train_timeseries.csv and contains daily weather observations along with weekly drought severity scores.*
2. *Soil data is stored in soil_data.csv and contains static soil characteristics of US regions using fips code.*

Methodology:

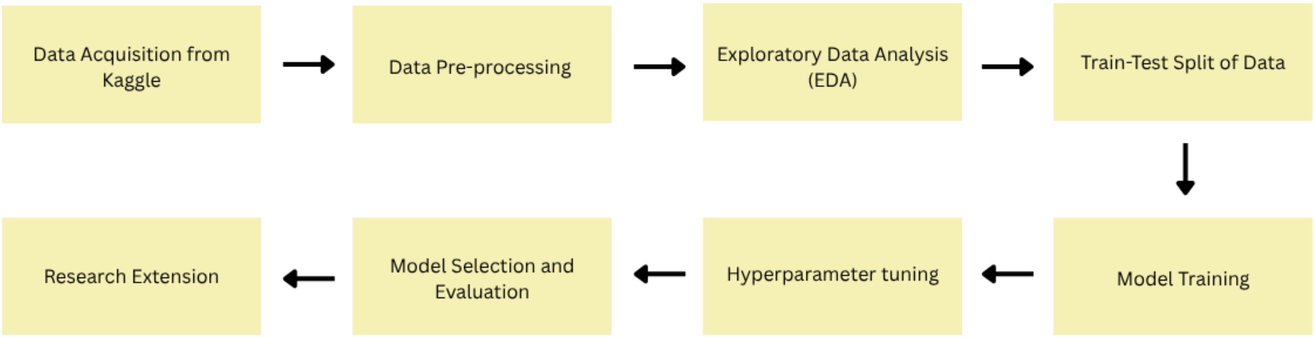

Fig1.0 methodology plan for machine learning based drought predictor

The approach taken in this research is quite methodical, aiming to predict drought severity by leveraging weather and soil data through machine learning techniques. The data was sourced from the US Drought Meteorological Data repository on Kaggle, which offers daily weather observations like precipitation, temperature, and drought scores, along with static soil characteristics at the county level. To create a cohesive dataset that links meteorological and soil information, both datasets were merged using the common fips identifier.

After merging, the data underwent preprocessing to ensure it was of high quality and suitable for modeling. Missing 'NaN' values were addressed, and any irrelevant or overly sparse features were discarded to reduce complexity. Date fields were transformed into numerical components such as year, month, and day, while outliers were identified and managed using statistical methods like the three-sigma rule. To ensure all features were on comparable scales, numerical attributes were standardized. Given that the target variable (drought severity) showed a significant class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was employed to create synthetic samples of the minority classes, helping to prevent model bias.

Once the data was ready, exploratory data analysis (EDA) was conducted. As part of the exploratory data analysis (EDA), we conducted univariate, multivariate, and correlation analyses to get a better grasp of the data and steer our modeling efforts. In the univariate analysis, we took a close look at individual features like temperature, precipitation, and soil properties, using histograms, boxplots, and countplots to evaluate distributions, spot outliers, and uncover any class imbalances in drought severity. The multivariate analysis delved into the relationships between variables, examining how precipitation and soil characteristics changed across different drought severity classes, with boxplots and scatter plots helping us identify predictive patterns. Lastly, we performed correlation analysis by calculating and visualizing a correlation matrix to reveal dependencies among features and pinpoint redundancies, which was crucial for feature selection.

After gaining insights from the data, The dataset was divided into training and validation sets in an 80:20 ratio, ensuring that the distribution of drought severity classes was preserved in both subsets through stratification. A variety of machine learning models were trained on the training set, including Random Forest, Decision Tree, Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost), all aimed at predicting drought severity. We fine-tuned the hyperparameters of each model through grid search with cross-validation, ensuring we found the best configurations to enhance predictive performance while steering clear of overfitting. To evaluate how well the models performed, we used the test set and looked at classification metrics like *accuracy, precision, recall, F1-score, and confusion matrices, which helped us understand their effectiveness across different levels of drought severity. We also took a closer look at feature importance to pinpoint the key predictors of drought conditions. Ultimately, we developed a full machine learning pipeline and saved it with Joblib for easy reproduction and deployment. The trained model was then incorporated into a real-time drought monitoring web app built with Streamlit, allowing users to interactively input weather and soil information and receive predictions about drought severity.*

Tools and Libraries:

This project was implemented using Python and the following libraries:

- *Data Manipulation: pandas, numpy*
- *Data Visualization: matplotlib, seaborn*
- *Machine Learning and Modelling: scikit-learn, xgboost, lightgbm, RandomForestClassifier,DecisionTreeClassifier*
- *Model Tuning and Evaluation: GridSearchCV, imblearn, SMOTE*
- *Other: warnings*

Expected Outcomes:

- *A trained and validated machine learning model capable of predicting drought severity levels with reasonable accuracy.*
- *Comparative analysis of multiple algorithms (XGBoost, LightGBM, Random Forest ,Decision Tree) to identify the best performing model based on evaluation metrics.*
- *Identification of key meteorological and soil features that influence drought conditions.*
- *A reproducible and scalable machine learning pipeline that can be adapted for drought prediction in other geographical areas.*
- *Potential integration of this model into early warning systems to support pro-active decision making.*