Amini Soil Prediction Challenge

Team Name: GCAmini.
Members: Chigozie Nkwocha (@Gozie) and Caleb Emelike (@CalebEmelike)

Competition Objective

To build a machine learning model that predicts the availability of 11 essential soil nutrients and calculates the nutrient gaps required for maize crops to achieve a target yield of 4 tons per hectare.

Evaluation Metric: Root Mean Squared Error (RMSE)

Instructions on how to reproduce results

The folder contains Python scripts, Jupyter notebooks and a requirements.txt file that contains a list of Python packages and Libraries used and their versions.

Folders

- Create a `data` folder which contains the given train and test datasets (soil indicators and gaps) alongside the Earth observation data.
- Every other preprocessed dataset will be stored in this folder also.

Code Order

Run the following codes in order of sequence

- Earth_data_preparation.py
- data_prep.py
- soil_prediction_model_RMSE_1066.py or its Jupyter notebook equivalent

Approach

Datasets used

- Soil indicators (Provided train and test data)
- Sentinel-1
- Landsat-8
- Surface temperature (MODIS_MOD11A1_data)
- Evapotranspiration (MODIS_MOD16A2_data)

These datasets are stored in the data folder.

Preparation of Earth Observation data

- Spectral bands were normalised and capped to 0-1 range
- Daily data for each of PIDs in the Earth observation data were aggregated by their monthly mean values across each year
- Spectral band indices such as EVI, NDVI, SAVI, etc, were calculated from the LandSat-8 data.
- Next, we extracted Earth observation data for each of the PIDs. PID coordinates in train and test data were not matching with those in the Earth observation datasets. As a result, we grouped the coordinates of each PID and took their mean values. These centriod values were used to get the PIDs in the Earth observation data closest to the PIDs in the train and test datasets. However, to prevent getting PIDs that are afar off, a distance threshold of 500 m was set.
- After that, we merged closest PIDs with those in Earth observation data and saved on disk in parquet formats in the data directory.

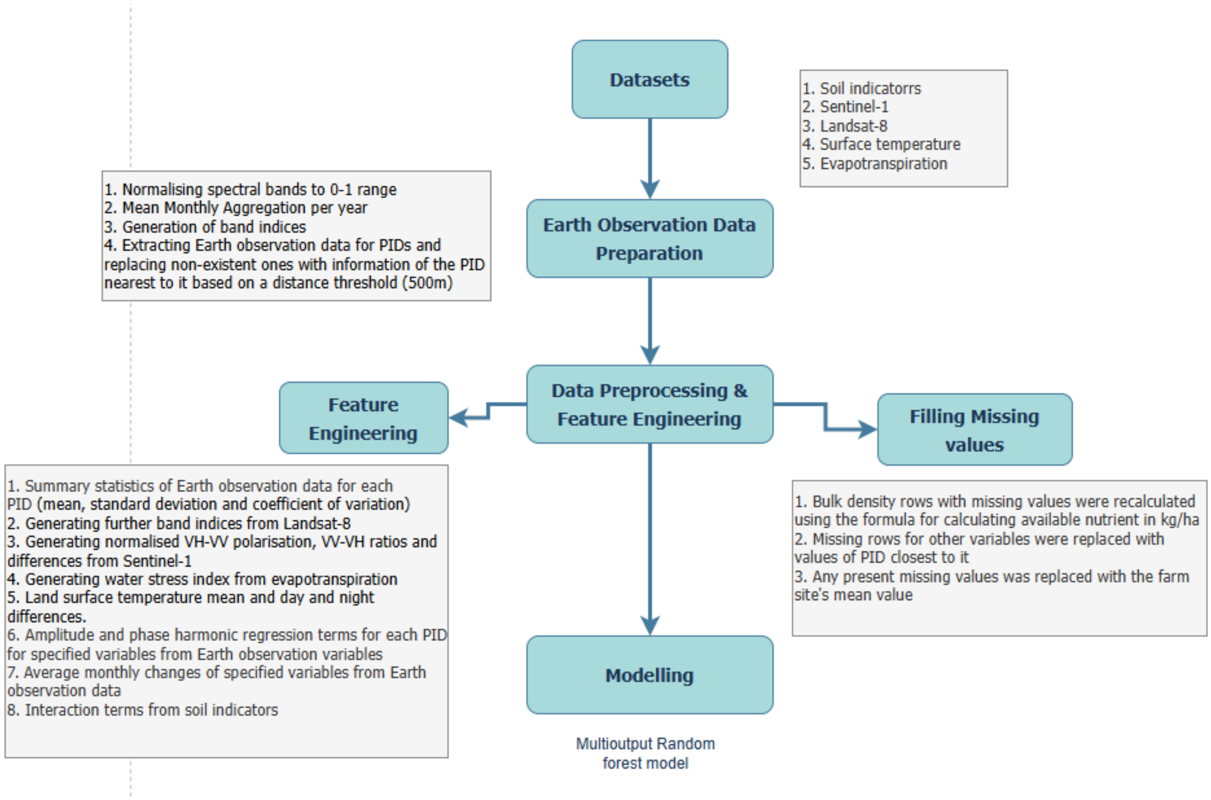
Datasets
1. Soil indicatorrs
2. Sentinel-1
3. Landsat-8
4. Surface temperature
5. Evapotranspiration
Earth Observation Data Preparation
1. Normalising spectral bands to 0-1 range
2. Mean Monthly Aggregation per year
3. Generation of band indices
4. Extracting Earth observation data for PIDs and replacing non-existent ones with information of the PID nearest to it based on a distance threshold (500m)
Data Preprocessing & Feature Engineering
Feature Engineering
1. Summary statistics of Earth observation data for each PID (mean, standard deviation and coefficient of variation)
2. Generating further band indices from Landsat-8
3. Generating normalised VH-VV polarisation, VV-VH ratios and differences from Sentinel-1
4. Generating water stress index from evapotranspiration
5. Land surface temperature mean and day and night differences.
6. Amplitude and phase harmonic regression terms for each PID for specified variables from Earth observation variables
7. Average monthly changes of specified variables from Earth observation data
8. Interaction terms from soil indicators
Filling Missing values
1. Bulk density rows with missing values were recalculated using the formula for calculating available nutrient in kg/ha
2. Missing rows for other variables were replaced with values of PID closest to it
3. Any present missing values was replaced with the farm site's mean value
Modelling
Multioutput Random forest model

Figure 1: flowchart

Data Preprocessing and Feature Engineering

Filling Missing values

- Bulk density rows with missing values were recalculated using the formula for calculating available nutrient in kg/ha.

$$\text{Available (kg/ha)} = \text{Available (ppm)} \times \text{soil depth (cm)} \times \text{bulk density}(g/cm^3) \times 0.1$$

- Missing rows for other variables were replaced with values of PID closest to it.
- Any present missing values was replaced with the mean value at the farm-site level.

Feature Engineering

The following features were generated:

- Summary statistics of Earth observation data for each PID: Summary statistics such as mean, standard deviation and coefficient of variation for each Earth observation were aggregated.
- Generating further band indices from Landsat-8: Further band indices features were generated. These are mainly ratios of existing raw spectral bands
- Generating normalised VH-VV polarisation, VV-VH ratios and differences from Sentinel-1
- Generating water stress index from evapotranspiration, which is the ratio of potential and actual evapotranspiration
- Land surface temperature mean and day and night differences.

- Amplitude and phase harmonic regression terms for each PID for specified variables from Earth observation variables. Amplitude describes how the specified variables fluctuate around the mean annual values. Phase describes its peak time. These features were created because it was discovered that these Earth observation data had seasonal patterns, hence fitting a sinusoidal curve.
- Average monthly changes of specified variables from Earth observation data.
- Aggregating spectral band indices into composite indices: This was done to understand the overall contributions of vegetative, burn-ratio, and soil moisture-related band indices, similar and non-similar band indices. Band indices to be combined were first normalised in a 0-1 range and weights were applied. Weights could be equal weighting, weights obtained from their contributions on the first principal component or by dividing their individual scaled values with the total of their combined scaled values.
- Interaction terms from soil indicators. These include ratios and products of their individual values.

Feature selection

Irrelevant features were dropped. These include the site, PID, longitude and latitudes, MODIS bands (Mbs 1, 3, 5, & 7). These MODIS bands were dropped since LandSat-8, a more accurate spectral imagery, captures this information. In total, about 203 features were used for modelling.

Modelling

Private and Public Leaderboard scores: 1066.884678 & 995.3992514

A multioutput regressor with random forest models was used. To reduce overfitting, a 5-fold cross validation was used to predict the test set at each fold iteration and then, the predictions were averaged. Predictions are saved in the `preds` folder while the feature importance scores is saved in the working environment.

Unfortunately, no hyperparameter tuning was done and also, it was surprising that gradient-boosting models performed poorly on the leaderboard even when they performed better when evaluating model performing on each fold's holdout data. Probably, it is due to overfitting.

One thing worthy of note is that, the pH is one variable that has a strong effect on soil nutrients.

Run time

About 2 hours, due to cross-validation. However, this could be sped up by increasing the `n_jobs` parameter in random forest and multioutput regressor classes.

> Side note: We have a better model (public and private leaderboards: 978.xx, 1047.xx, respectively) which applies mutual information feature-selection method for each of the target variables. It can be found in the model_with_feature_selection notebook.