**Prediction of soil fertility and crop productivity through machine learning algorithms**

Mustafa Kamal Shahadat and Dr. Md. Khairul Alam

**Justification:**

Crop productivity is a function of many factors. Among them soil fertility, physiography, crop management, weather parameters etc. are major factors. However, these factors are highly variable both spatially and temporally. A number of machine learning algorithms have been evaluated by many scientists for predicting soil fertility and crop production. The present study focuses on evaluation on various supervised and unsupervised machine learning techniques for predicting soil fertility and crop production.

Objective:

* Prediction of soil fertility status according to agroecological zone.
* Prediction of crop production scenario based upon environmental features.

**Methodology**

Step 1: Source of data and about dataset

We have two datasets at our disposal. One is historical crop yield and production data collected from secondary sources. The dataset contains crop specific yield and production according to locations. The dataset also contains environmental variables related to crop production. There are 11 major crops grown in different rotations of the respective locations. We will apply machine learning algorithms such as multiple regression, random forest, support vector machine etc. in order to predict crop production.

Our second dataset consist some 260 samples of soil test data collected from the farmers field from various locations of coastal region of Bangladesh. Each data point is spatially referenced. So far the dataset contains 15 features viz. location, sample id, latitude, longitude, pH, electrical conductivity (EC), organic matter (OM), organic carbon (OC), total nitrogen (N%), phosphorus (P µm/g), potassium (K meq/100 g), Sulphur (µm/g), zinc (µm/g), boron (µm/g), fertility status etc. We will add other useful features. The main target feature is the fertility status.

Step 2: Data pre-processing and exploratory data analysis

Missing values will be corrected through either dropping or suitable imputation method. Necessary data transformation, encoding (level or one-hot encoding), summary, correlation matrix, visualization, features selection etc. steps will be followed.

Finally, the dataset will be spilted into training and test set, where training set will cover 80% of the data and 20% will be used for testing.

Step 3: Model selection and algorithm setting

Both supervised and unsupervised machine learning approach will be followed. Such as KNN, DTs, SVM, RF, ANN etc. For model tuning necessary hyperparameters will be optimized through grid search.

Step 4: Evaluation of the model: model score, confusion matrix etc.

**Future study:**

* Soil salinity and fertility mapping.
* Preparation of crop suitability map based upon soil fertility and remote sensing data.