PRAKRITI 2020
Data Analytics

Team Intangible

- Saurav Kumar Nishant
- Tushar Mohandass
- Saw Sachin Azad
- Saurabh Kumar Pandey

Estimation of Yield and Nutrient Concentration

“ *Aim:*

Predict the Total Soil Carbon & Total Soil Nitrogen by using multivariate machine learning models from elemental & spectral data

DATA SETS : ELEMENTAL DATA

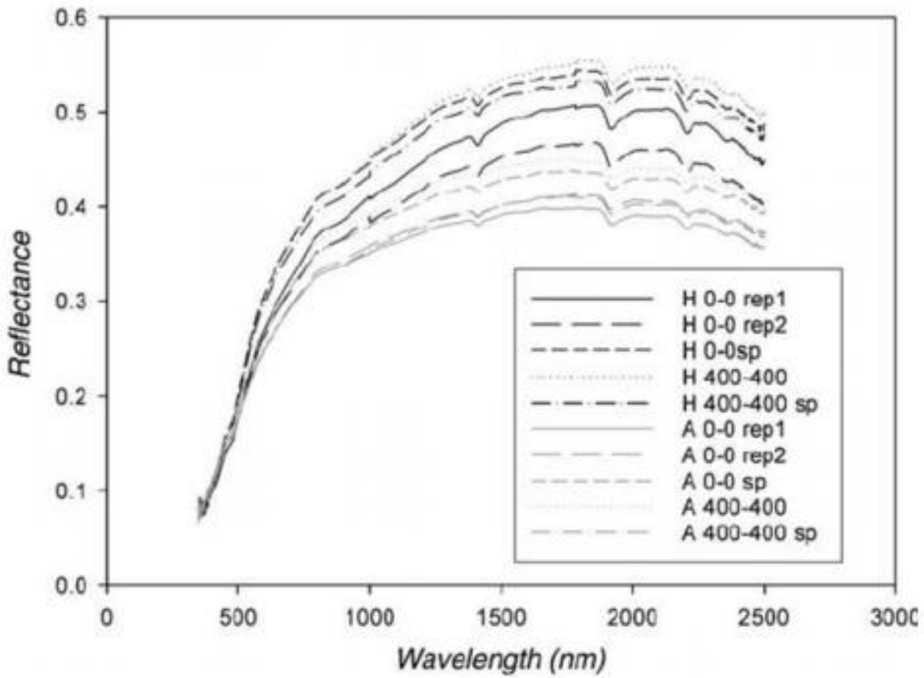
Concentration in ppm

- Zinc
- Sulphur
- Potassium
- Calcium
- Titanium
- Manganese
- Iron
- Rubidium
- Strontium
- Aluminium
- Silicon

DATA SETS: SPECTRAL DATA

- **Reflectance Data** from 350 nm to 2500 nm measured using a **spectroradiometer**.
- The data is available with one pre-treatment i.e. the **first derivative of original absorbance data**
- Therefore, each column give the value of the first-order derivative of reflectance of different **soil samples** at different **wavelengths ranging from 350-2500 nm**

RELATIONSHIP BETWEEN REFLECTANCE AND THE WAVELENGTH FOR A SOIL SAMPLE



As given in the Problem Statement. The actual data gives the first order derivative value of the reflectance which can show a linear relationship with TC and TN values

OBJECTIVES

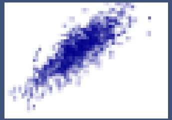
- Use **Multivariate Machine Learning** models to predict TC and TN using
  a. Only elemental data
  b. Only Spectral Data
  c. Combination of Elemental and Spectral data
- Compare **Model Accuracy** for the 3 cases
- Identifying influential variables and **correlation** study
- **RMSE** Reporting

Model Building Steps & Feature Selection

- Data has been imported in the Google Colab.
- Null value has been checked (there aren't any NULL values) and data types of each feature has been checked.
- Two columns (**TC (%) and TN (%)**) which were supposed to be of **float type** were in **object type** so we checked the string value and removed it.
- Finally when **string (n.a.)** was removed from these two columns then type was converted from object to float by typecasting.
- Data set was split into **X (Input) and y1 and y2 (outputs)**.
- Then Input data X was modelled into three categories **X1 (Only elemental data(highlighted in green)), X2 (Only Spectral data(not highlighted)) and X3 (Combination of elemental and spectral data).**

Model Building Steps & Feature Selection

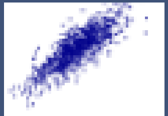
- **Pearson's correlations coefficient** were checked for each of the three situations and if two features were found to be correlated one from those two were removed.
- **Min-Max Scaling** was done on both Input and Output in all the three situations.
- Then on all the three given sets of data three models were trained i.e. **Linear Regression model, Decision Tree Regressor and Gradient Boosting Regressor** on X_scaled_corr.
- **X_scaled_corr** is the **best set of features** obtained from Pearson's Correlation Coefficient after it has been scaled.
- **Root mean squared errors** were calculated in each case to select the best Model.



Pearson’s Correlation Coefficient

- **Correlation Coefficient** is calculated among the attributes of elemental/ spectroradiometer data
- A benchmark of either **0.85 and 0.9** are set based on the **correlation value distribution** and the respective attributes are combined to avoid multicollinearity.
- It is also referred to as the **Pearson’s R test**.
- It can take any value from **+1 to -1**. If the value is greater than zero then the variables are positively dependant and vice versa.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

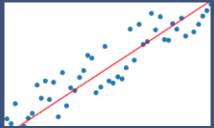


Correlated Features

Results:

Number of pairs of correlated features (based on Pearson's Correlation coefficient) :

X1-Only elemental data(highlighted in green) >0.85	X2-Only Spectral data(not highlighted) >0.9	X3-Combination of elemental and spectral data. >0.9
4	1886	1887



Linear Regression Model

- Linear Regression is **a linear approach** that models the relationship between a **scalar** response and one or **more explanatory variables**.
- We perform a multiple linear regression by using **Ordinary least square regression** technique (**unbiased estimator**) to model TC and TN as a function of the elemental or **spectrometric data** or both.

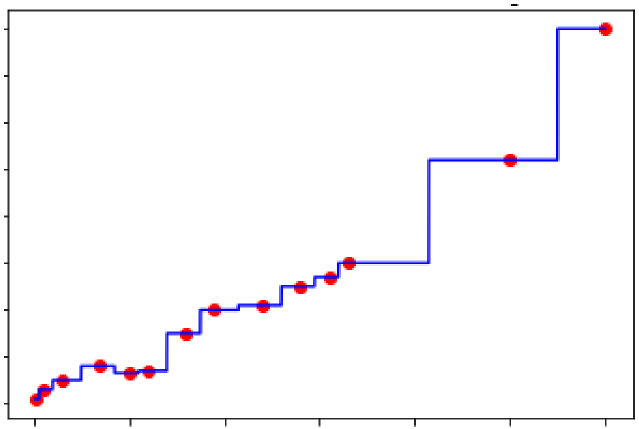
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

$$E = \sum_{i=1}^{m} (Y_i - \hat{Y}_i)^2$$

- In OLS regression**, error function** i.e. the sum of squares of the difference of predicted values and actual values; is **minimized** using a **gradient descent approach**

Decision Tree Regression

- **Decision tree** is a **supervised algorithm** that builds regression models in the form of a tree structure.
- It breaks down a **dataset into smaller and smaller** subsets while at the same time an associated decision tree is **incrementally developed**.
- **Decision tree regression** observes features of an object and trains a model in the **structure of a tree** to predict data in the future to produce meaningful continuous output.



Gradient Boosting Regression

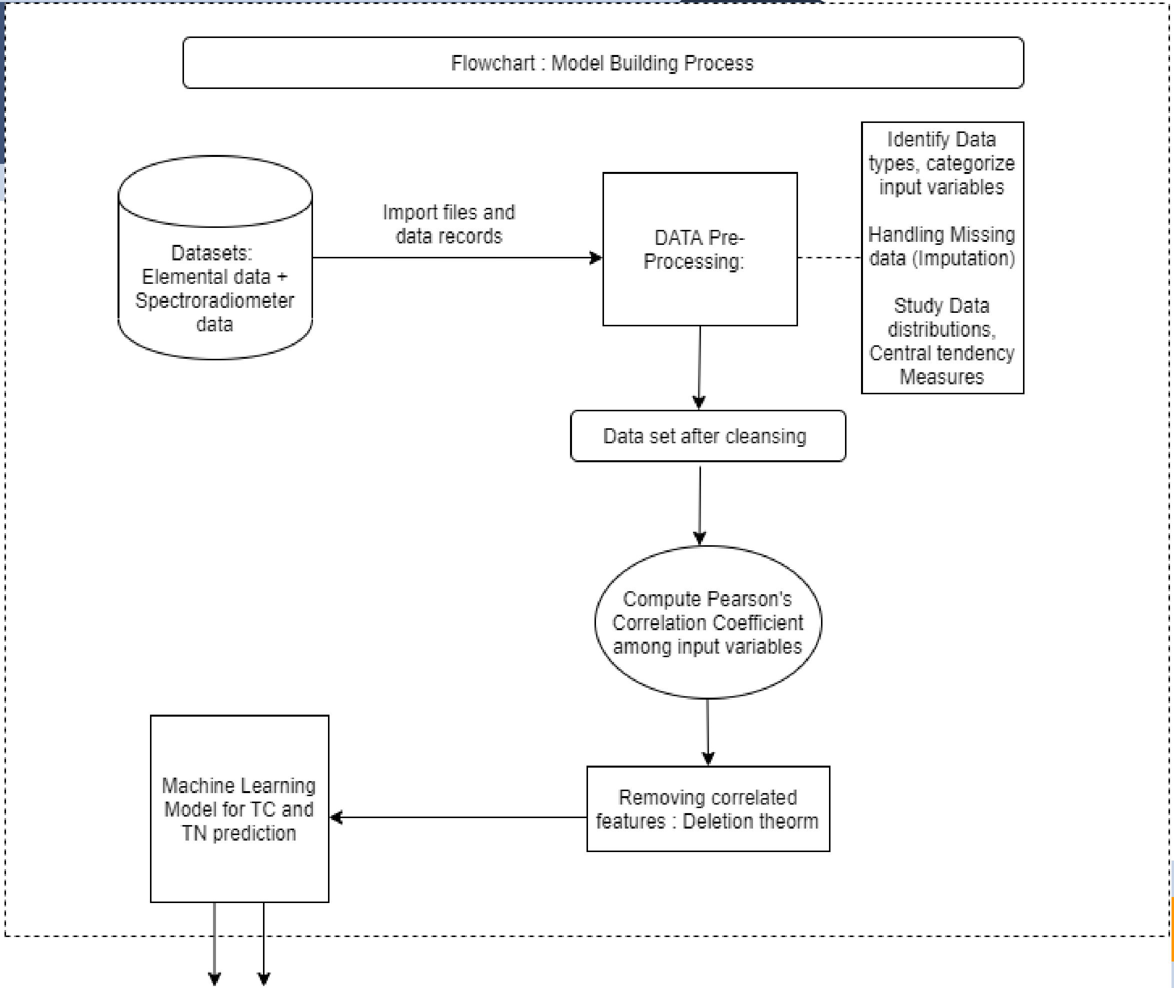
- **Gradient boosting** is a **machine learning technique** used for regression which produce prediction model in the form of an **ensemble of weak prediction models**.
- It builds the model in a stage wise fashion like other boosting methods do and generalizes them by allowing optimization of a **arbitrary differential loss function**.

$$F_m(x) = F_{m-1}(x) + \underset{h_m \in \mathcal{H}}{\arg\min} \left[\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right].$$

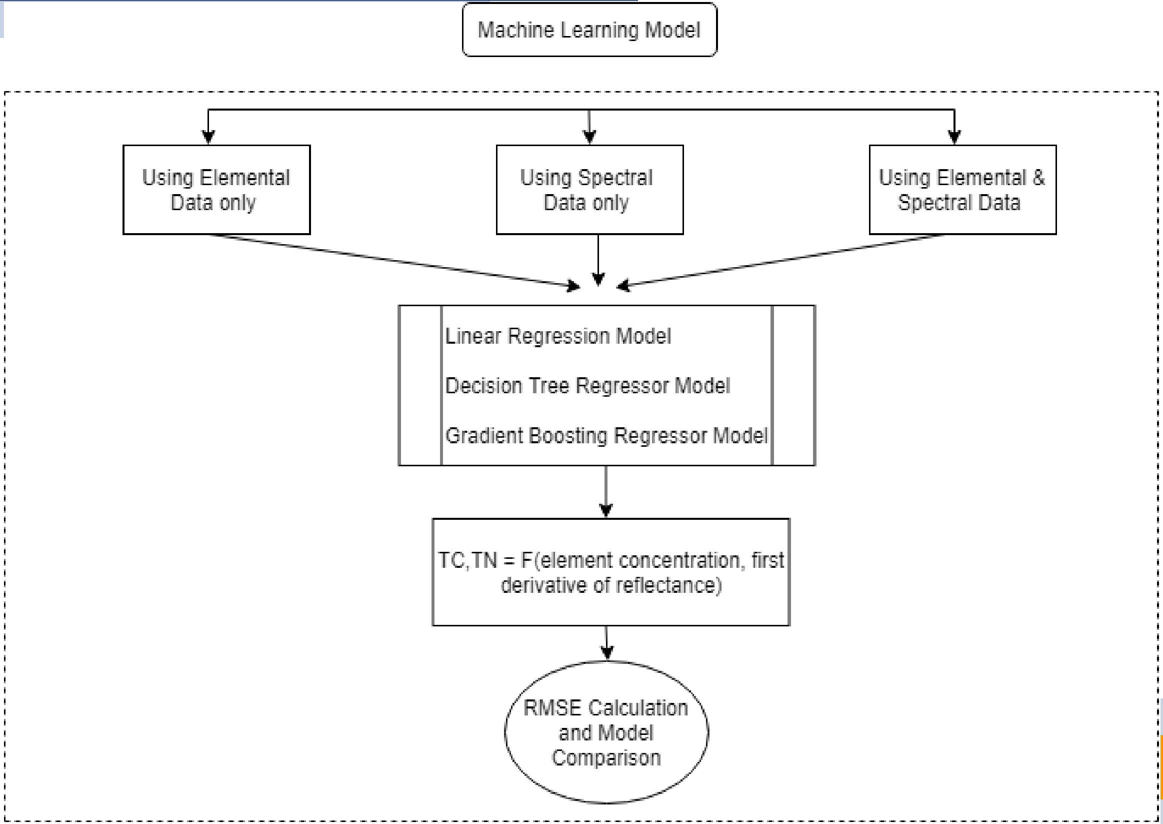
- h is the **base learner's function**
- L is the **Loss function**
- **Steepest Descent** is apply to solve the minimization problem

(1) Model Building Process

- Data pre-processing is carried out & input variables are categorized
- Handling Missing data: Data Imputation
- Exploratory Data Analysis



(2) Model Building Process



- 3 different ML models are used to predict the **TC, TN** values.
- **Linear Regression, Decision tree regression** and **gradient boosting regression model** are used

Root Mean Square Error

- The **RMSE** is evaluated for all the test cases based on the model built for predicting **TC and TN**.
- The **root mean square error** is a measure of the differences between values predicted by a model or an estimator and then the values observed.
- The RMSE serves to aggregate the magnitudes of the errors in prediction for various times into a **single measure of predictive power**.
- It is a measure of accuracy to **compare forecasting errors** of different models for a particular dataset

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

RMSE Results Tabulation for Total Soil Carbon (TC)

Root Mean Squared Error (RMSE) for TC(%) (y1):

	X1-Only elemental data(highlighted in green)	X2-Only Spectral data(not highlighted)	X3-Combination of elemental and spectral data.
Linear Regression	**0.21422**	**0.09985**	**0.09792**
Decision Tree Regressor	**0.09777**	**0.11456**	**0.09259**
Gradient Boosting Regressor	**0.08214**	**0.13700**	**0.09728**

- **Among the 3 model building process:**
- ✓ RMSE is the least for **Gradient Boosting Regressor** in case of estimation with only **elemental data**
- ✓ In case of **spectral data** based estimation, **linear regression** performs the based regressor based estimation
- ✓ In case of **both the data attributes** the best model is **decision tree regressor** which has the least RMSE

RMSE Results Tabulation for Total Soil Nitrogen (TN)

Root Mean Squared Error (RMSE) for TN(%) (y2):

	X1-Only elemental data(highlighted in green)	X2-Only Spectral data(not highlighted)	X3-Combination of elemental and spectral data.
Linear Regression	**0.00783**	**0.00446**	**0.00434**
Decision Tree Regressor	**0.00442**	**0.00373**	**0.00456**
Gradient Boosting Regressor	**0.00468**	**0.00470**	**0.00528**

- **Among the 3 model building process:**
- ✓ RMSE is the least for **Decision Tree Regressor** in case of estimation with only **elemental data**
- ✓ In case of **spectral data** based estimation, **decision tree** performs the based regressor based estimation
- ✓ In case of **both the data attributes** the best model is **linear regression** which has the least RMSE

THANK YOU