Problem Statement : Data Analytics

Plant health highly depends upon the availability of nutrients in soil that can be estimated through several advanced techniques such as spectrometry, spectroradiometry and PXRF(Portable X-ray Fluorescence) spectroscopy etc.

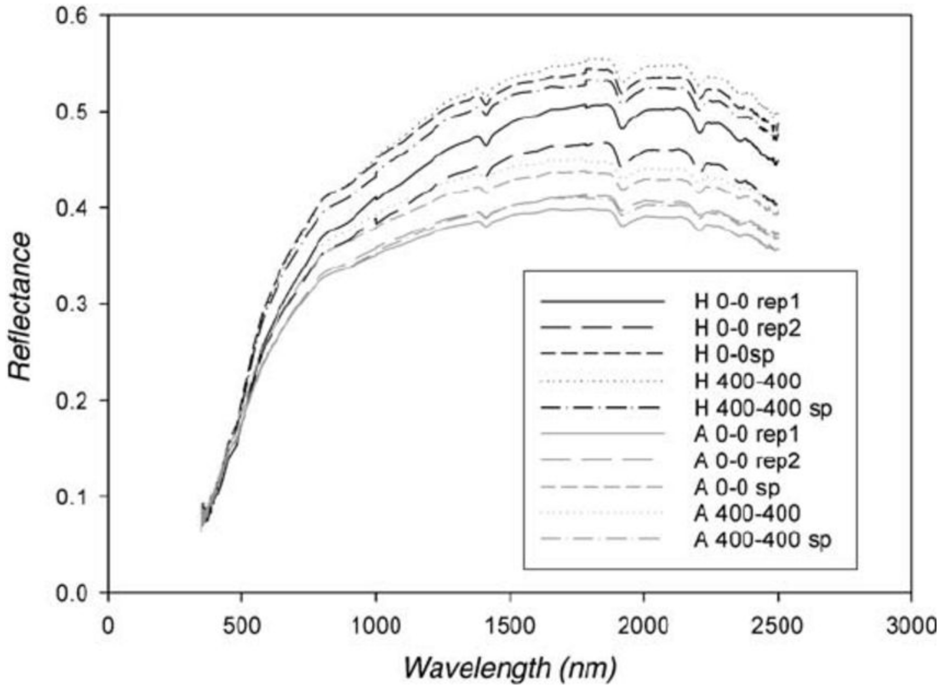
There is a lot of research going in the field of Spectrometry. A spectrometer is a scientific instrument used to separate and measure spectral components of a physical phenomenon. Another instrument, spectroradiometer, is designed to measure spectral radiance or irradiance across various spectral ranges. These techniques are rapid, portable, and provides multi-elemental analysis with results generally comparable to traditional laboratory-based techniques.

The attached dataset contains the data of **137** soil samples collected from four different locations. **TC** represents **Total Soil carbon**, **TN** represents **Total Soil Nitrogen**, green highlighted part represents the **elemental data** (ppm) measured with a specific soil sensor and **reflectance values** from 350 nm to 2500 nm measured using a spectroradiometer. The spectroradiometer data (spectra) is available with only one pretreatment ie. the first derivative of original absorbance data.

Elemental Data(columns highlighted in green):- These columns shows an estimate of the concentration of elements (like Zinc, Sulphur, Potassium, Calcium, Titanium, Manganese, Iron, Rubidium, Strontium, Aluminium and Silicon) in ppm, present in the soil sample.

Spectral Data(columns not highlighted with any colour):- These columns give the value of the first-order derivative of reflectance of different soil samples at different wavelengths ranging from 350 to 2500 nm.

Below given graph is a plot of reflectance versus wavelength for a given soil sample to show their relationship and the data given is of the first-order derivative of reflectance at different wavelengths.



A. The problem statement is to use multivariate machine learning methods to predict TC and TN (the yellow highlighted columns) using

1. Only elemental data(highlighted in green)
2. Only Spectral data(not highlighted)
3. Combination of elemental and spectral data.

B. Also, compare the model's accuracy in the three different sets of data(ie. only spectral, only elemental and a combination of both).

NOTE: While comparing the model, this thing should be kept in mind that the same model or a combination of models can be used for comparison.

C. Also, select the influential variables used for each method used. In other words, study the correlation of all the independent variables with the dependent variables.

Judging for the model comparison(**part A and B**) would be done on the basis of RMSE score(Root Mean Squared Error)