Project Part 2.2: Modelling Plan

Students: Bahadir Bagci, Hassan Majeed, Zoheb Rahman

Team Name: Spectral Data Soft Sensor (B)

1. Introduction

Following the the last part, which included initial data processing and number of principal component selection steps, we now can shift our focus to modeling. From now on, we will be working on developing a model which predicts selected five plant traits: carbon, chlorophyll, equivalent water thickness, leaf mass per area, and nitrogen. These were chosen because they have fewer missing values compared to the others. We'll build a separate regression model for each trait. The input for these models will be the spectral band values collected for each sample.

2. Methodology

Last time, we used PCA earlier to reduce the number of features by creating new ones that are combinations of the original spectral bands. Since these new features are linear, we chose to use linear regression to model the relationship between them and the selected plant traits. Although other options like Ridge or Lasso could be used, we went with basic linear regression, because it's easier to interpret and fits our goal at this stage.

2.1.Tools and Path

Model calibration will be performed in Python, using the libraries such as scikit-learn(for Linear Regression) and matplotlib(for visualization).

2.2.Model Calibration Strategy

Based on the nature of this dataset, missing trait values do not imply that the corresponding traits are absent or zero. Rather, they indicate that those traits were not measured for that particular sample. Since even the least missing trait has over 30% missing values, we did not apply imputation. Filling in such a large amount of missing data without clear patterns could distort the relationships the model tries to learn. Instead, for each selected trait, we filtered out the rows where the target value is missing. We made sure to use only the real, available values for each trait, instead of filling in missing ones.

First of all, we split the dataset into training and test sets.(%70-%30) Then, for generalization, we checked how well the models perform during training. In order to do that, we used 5-fold cross validation on the training set. In brief, this method splits the data into five parts, trains the model on four parts, and validates it on the remaining part—repeating this process five times so as to help us understand the consistency of the models.