Predicting Soybean Crop Yield Based on Soil Nutrients and Microbiome
Team Members: Afaf Abdelrahim, Kayla Cross, Stephanie Majernik, Kimberley Ndlovu
Github: https://github.com/kbcross/Erdos_Project_Predict_Crop_Yield.git

Overview

Soybeans are an important crop in the US with 70% of soybeans used for livestock feed and 15% used for human consumption. There is even a growing field of using soybeans for biodiesel. Therefore, high soybean crop yield is essential to keep up with livestock and human consumption. Research has shown soil nutrients and bacteria can influence the growth and therefore the yield of soybeans. Hence in this project we ask: can soil nutrients and soil microbiome be used to predict soybean crop yield? The importance of this is related to the agricultural field as farmers must be able to have efficient crop yield to (1) maintain their income and (2) feed both humans and livestock.

Stakeholders

Stakeholders for this project include individual farmers and the US Department of Agriculture. These stakeholders would be able to use our model to not only predict their soybean crop yield based on the soil nutrients and microbiome, but also identify key features (i.e. nutrients or microbes) that they could supplement their soil with. The data needed for this model is relatively cheap and easy to obtain. Farmers will be able to collect soil in their fields, send it to a laboratory where they will measure the concentration of nutrients and sequence bacterial DNA in the soil, and lastly upload their data to the model. The US Department of Agriculture would be interested in our crop yield prediction tool to determine yield for the month/year. This would allow them to properly allocate soybeans to various categories (livestock, human, biodiesel) and determine whether or not more soybean fields are needed.

Key Performance Indicators

The main outcome we want to predict is soybean yield based on soil nutrients and root microbiome. Additionally, we want to identify key individual nutrients or microbes (potential soil probiotics, we still need to decide on which phylogenetic level) that can predict high or low yield of the crop. On a broader outlook, this will suggest soil health and allow for farmers to determine how to improve soil health for increased crop yield (i.e. addition of probiotic, crop rotation, etc).

Approach

We first tested three different regression models, namely, linear regression (LR), random forest regression (RFR) and support vector regression (SVR). The models are trained and tested on our dataset (with and without soil microbiome as features), and their predictive accuracy is assessed using mean squared error. We then identified important variables by computing feature importance using the random forest algorithm.

Results

We used data found in Niraula, Rose and Chang (2022) where they sampled soil nutrients and soil microbiota in 12 different soybean crop fields across Nebraska and Oklahoma. The data included 16 different soil nutrients, soybean crop yield, and over 4,000 operational taxonomic units (OTUs), which is a way to classify different microbes. For this project, we collapsed the OTUs into the taxonomic hierarchy level Order and removed temperature, rainfall, and electrical current from the soil nutrients.

Potassium (K) and Calcium (Cat) were the most important soil nutrient features in predicting crop yield. The bacterium *Chthoniobacter* , was the most important soil microbiome feature. This bacterium is a general heterotroph that is involved in the transformation of organic carbon compounds in the soil.

We were not able to accurately predict soybean crop yield based on soil nutrients alone or soil nutrients plus microbiome data. However, this could be due to low sample sizes (12 total samples, 6 from each state) as well as the data manipulation we performed on the microbiome data. We collapsed the microbiome data into the order level, however, there are multiple ways in which the microbiome data could be collapsed and should be explored.

Future Directions

Current research suggests that modeling bacterial communities based on relative abundance of the members instead of metabolic pathways misses key differences in communities metabolic output. Our results showed that soil nutrients instead of community members were key predictors of crop yield. In the future, it would be better to use metabolic pathways instead of community members so we can elucidate how the microbiome directly affects the soil nutrient levels to create a more accurate model.