Predicting soybean crop yield based on soil nutrients and microbiome

Afaf Abdelrahim, Kayla Cross, Stephanie Majernik, Kimberley Ndlovu

Introduction

- Soybeans are an important US crop
  - 70% of soybeans being used for livestock feed
  - 15% used for human consumption
- Not only are soil nutrients important for plant growth, but also the soil microbiome.
  - Soil microbes promote growth by controlling the availability and acquisition of nutrients by plants.
  - Meaning, if the soil microbiome is not at an optimal state, the plant has higher risk of plant disease, reduced yield, higher greenhouse gas emissions, and more.

Research Question & Key Stakeholders

Question: Can soil nutrients and root microbiome be used to predict soybean crop yield?

Key Stakeholders

1. On a small scale, individual farmers and small farming companies to determine their expected yield (and therefore profit) based on their soil nutrients and microbiome.
2. On a large scale, the US Department of Agriculture to determine yield for the month or year to allow for proper allocation of soybeans to various categories and determine if more soybean fields will be needed in the future.

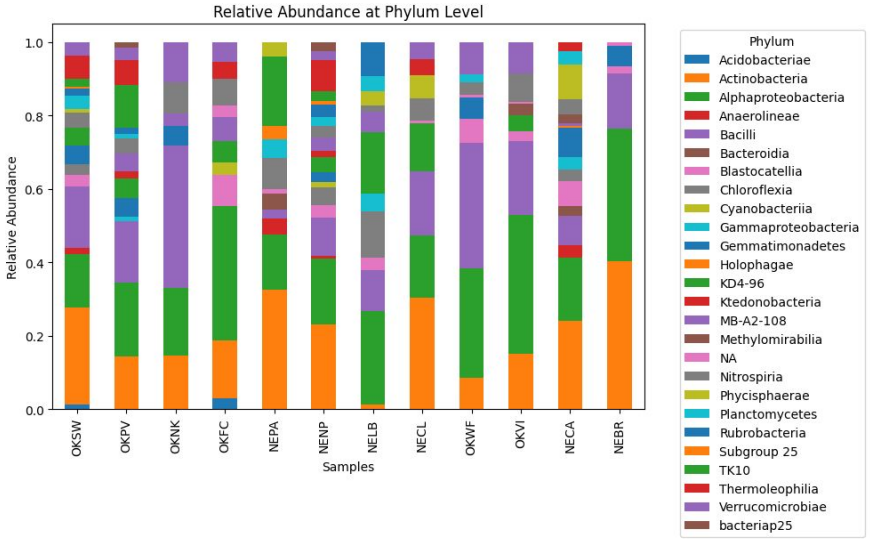
Data Structure

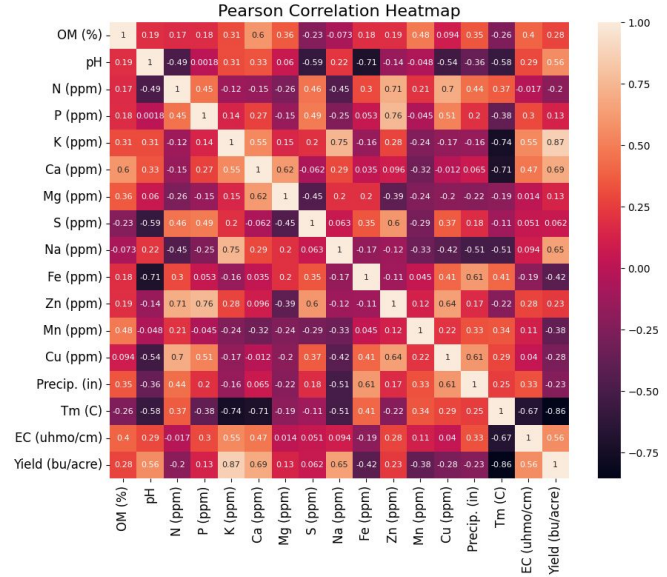
- Collected data provided information on soybean yield, soil nutrients, and bacterial 16S rRNA gene sequencing reads
  - Bacterial 16S rRNA reads can be processed to provide information on bacterial abundance at the genus level
  - Soil nutrients provides information on factors known to affect crop yield

	Yield (bu/acre)	OM (%)	pH	N (ppm)	P (ppm)	K (ppm)	Ca (ppm)	Mg (ppm)	S (ppm)	Na (ppm)	...	Rokubacteriales	Rubrobacterales	SBR1031	Sphingomonadales	Streptomycetales
0	59.2	2.46	7.0	1	40	752	1932	312	11	21	...	0	0	0	317	0
1	53.9	3.52	7.0	9	92	413	3513	422	10	12	...	78	0	35	238	228
2	58.6	3.98	6.8	18	55	504	2474	270	13	8	...	0	0	0	83	0
3	43.6	3.31	6.8	0	60	291	2990	425	9	8	...	0	129	0	0	0
4	54.4	2.63	7.0	14	227	419	2687	275	16	10	...	0	174	0	191	0

Primary Features

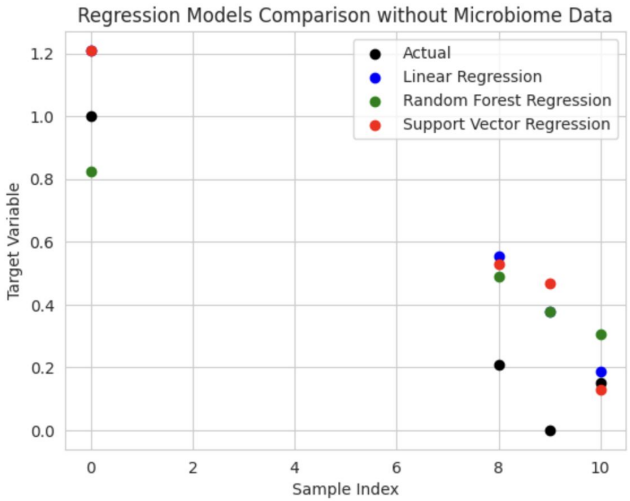
- Primary features of interest were bacterial abundance and soil nutrients
  - Temperature, rainfall, and electric current were removed due to being uncontrollable variables





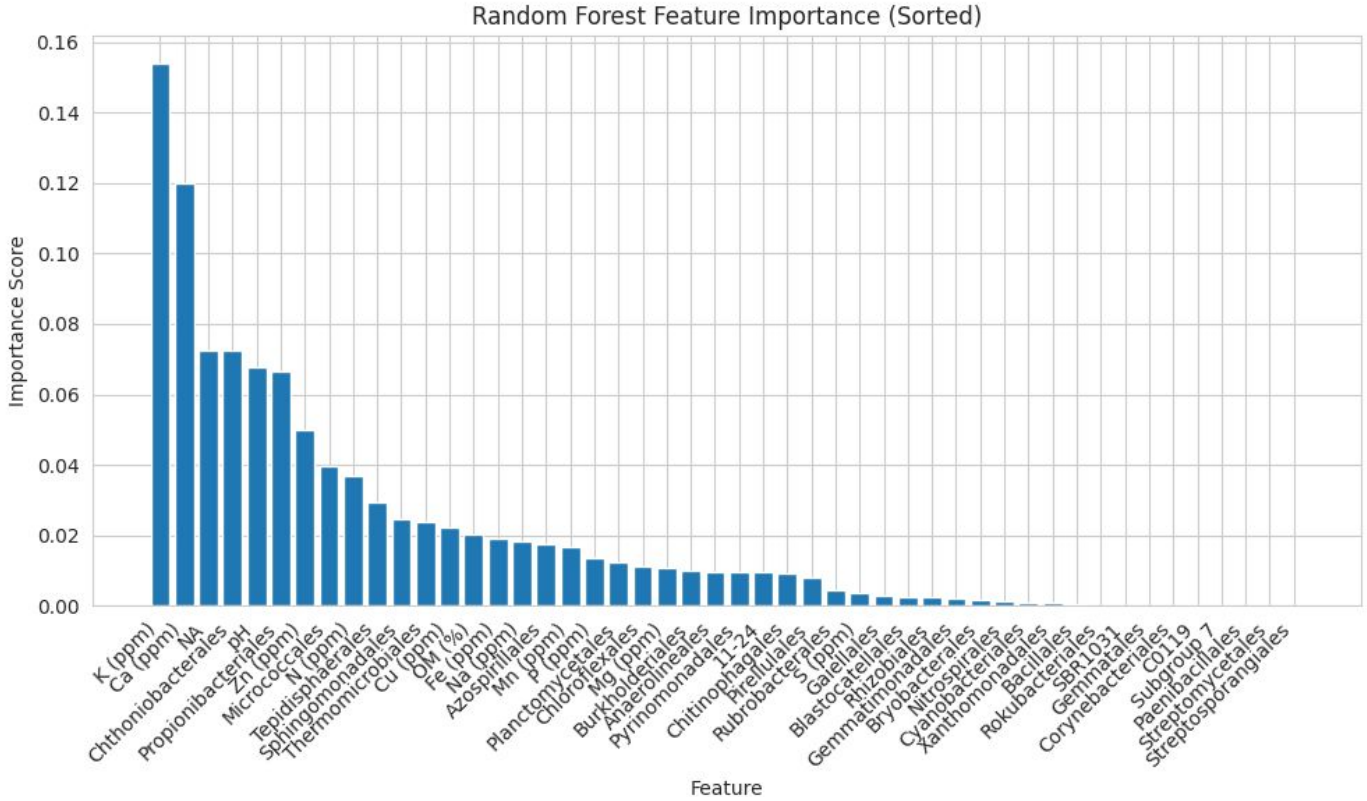
Regression Model: A Comparative Analysis

- Linear, random forest, and support vector regression was tested with and without the bacterial abundance.
- The model with the best performance only included soil nutrients with random forest regression



With or without microbe data?	Linear Regression MSE	RF Regression MSE	Support Vector Regression MSE
With	26.708	209.662	219.714
Without	0.076	0.069	0.091

Random Forest Feature Importance Analysis



Results & Conclusions

Strengths:

- Soil nutrient data, especially nutrients Potassium and Calcium, is able to predict soybean crop yield.
- We tested multiple models and compared performance.

Weaknesses:

- Sample size was low with only 6 samples from Nebraska and 6 samples from Oklahoma. A greater sample size that included more states would be more ideal.
- Including soil microbiome data in the model reduced the performance. This should be explored more to determine why and if the data could be manipulated to provide a better model

Was our approach accurate? No, we were not able to accurately predict, but this is most likely due to weaknesses stated above.

Stakeholders and Future Directions

- Stakeholders, individual farmers & USDA, will be able to use soil nutrient and microbiome data to predict their crop yield.

- Bacteria are much more complex than their genus. Many share metabolic genes and contribute to metabolic networks.

- Future models would take these networks into account to create a model that would better connect microbial nutrient production to prediction of crop yield.

Links/References/Resources

Github: https://github.com/kbcross/Erdos_Project_Predict_Crop_Yield.git

Executive Summary

Extra Slides if needed

Results - K Nearest Neighbors Regression

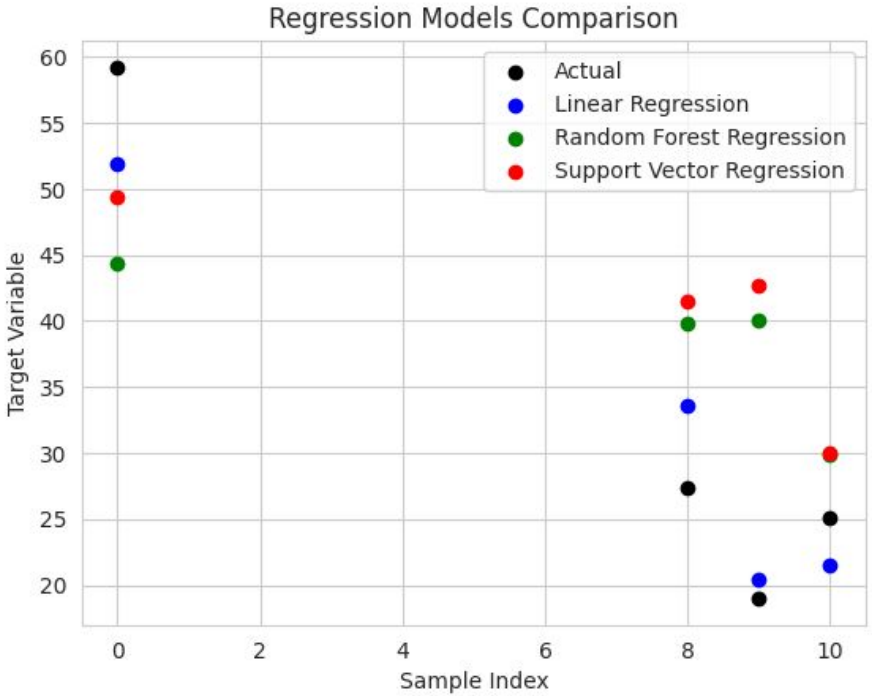
KNN was not talked about in our presentation, but we did do it

Main Conclusion on KNN:

Soil nutrient data alone was better at predicting actual data

K = 3 with weight = “distance” had the lowest mean squared error (0.24)

With microbe data?	K = ?	weight	MSE
No	3	uniform	0.256
No	3	distance	0.24
No	5	uniform	0.267
No	5	distance	0.25
Yes	2	distance	255.07
Yes	3	distance	141.26
Yes	3	uniform	138.89
Yes	4	distance	241.55



With or without microbe data?	Linear Regression MSE	RF Regression MSE	Support Vector Regression MSE
With	26.708	209.662	219.714
Without	0.076	0.069	0.091

Feature importance computed using shap values