Predictive Model for determining Experimental Site Response to N Fertilizer Addition

John Salako

johnsalako3@gmail.com

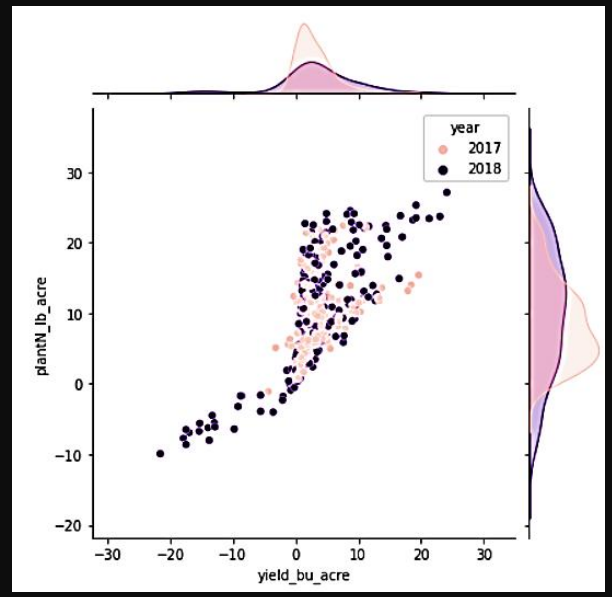
. . . *Consistency Across ML Methods Provides Confidence*

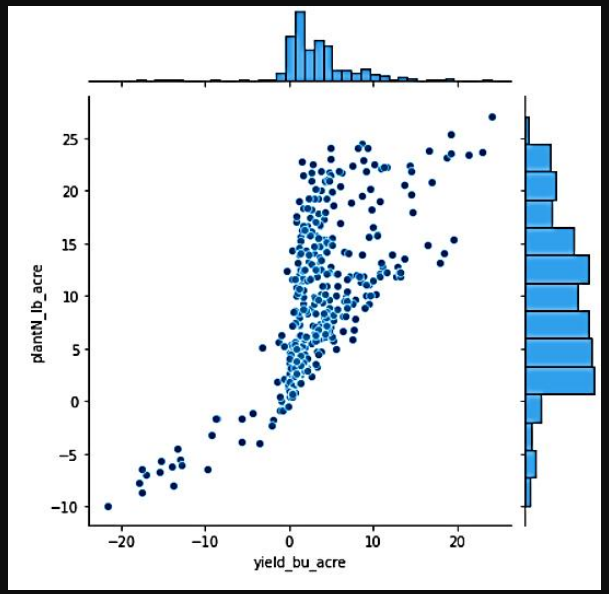
Contents

- Data preparation and Visualizations
- Preprocessing Steps
- Machine Learning Process
- Accuracy and Validation of Model
- Insights from Model
- Conclusion

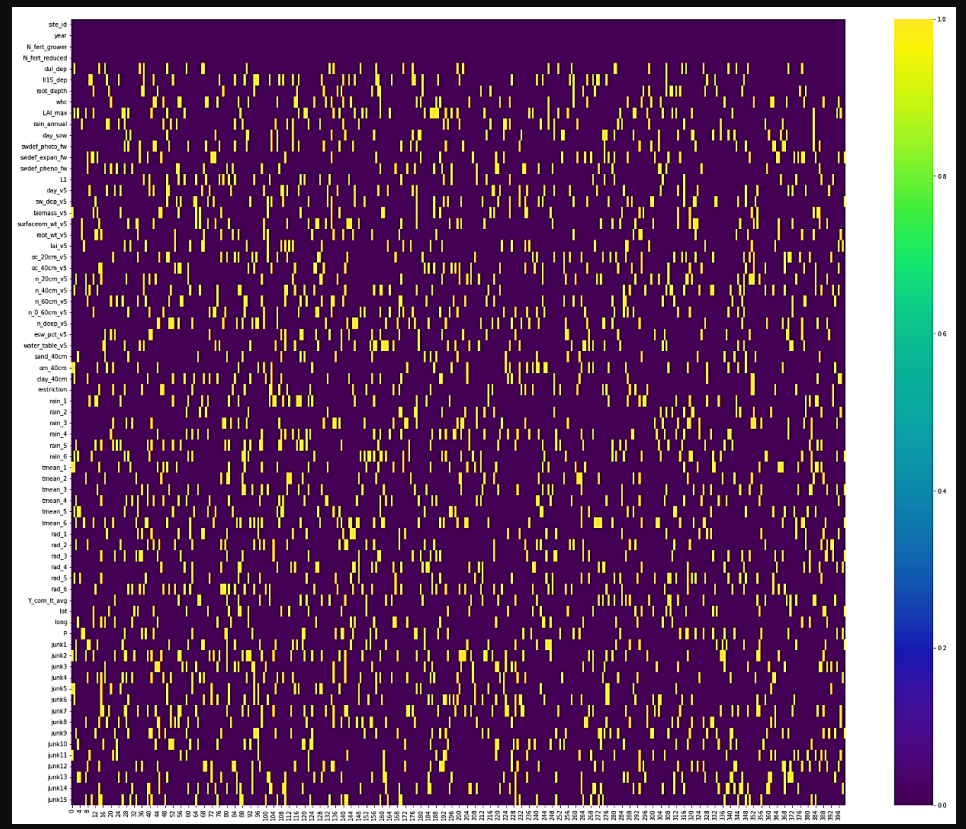
Data preparation and Visualizations

- Reformating the target data
- Assessing targets distribution via Seaborn *jointplots*
- Data Missingness Analysis (MAR)
- Understanding Predictors Correlation through EDA

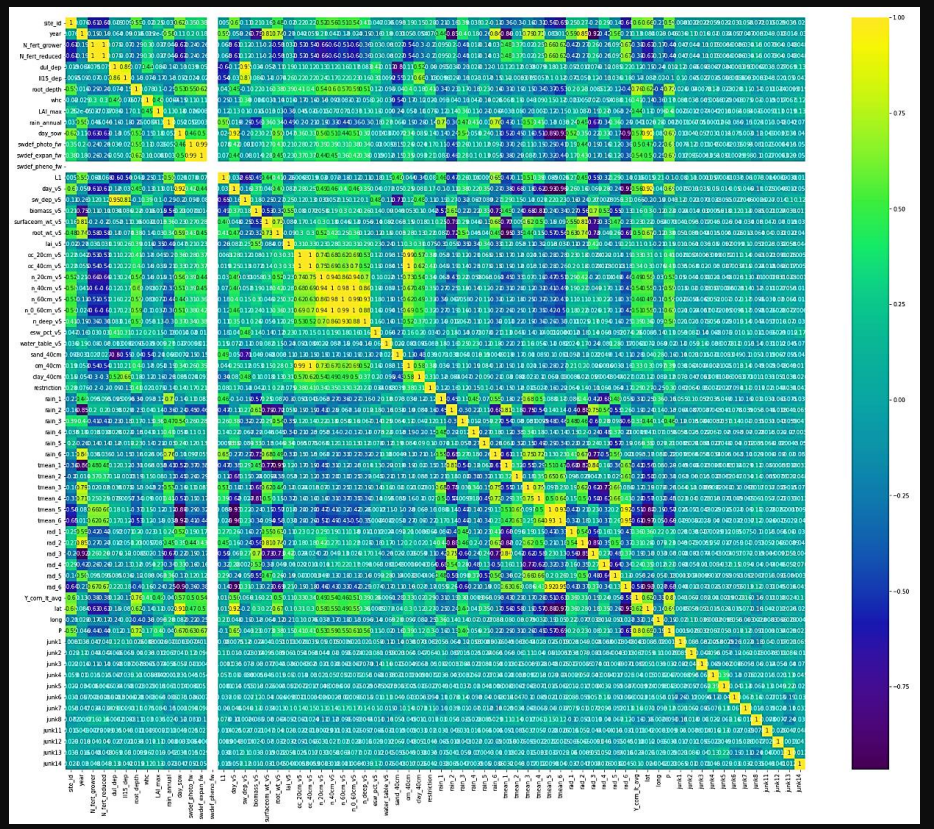




Data Preparation:
Missingness and EDA

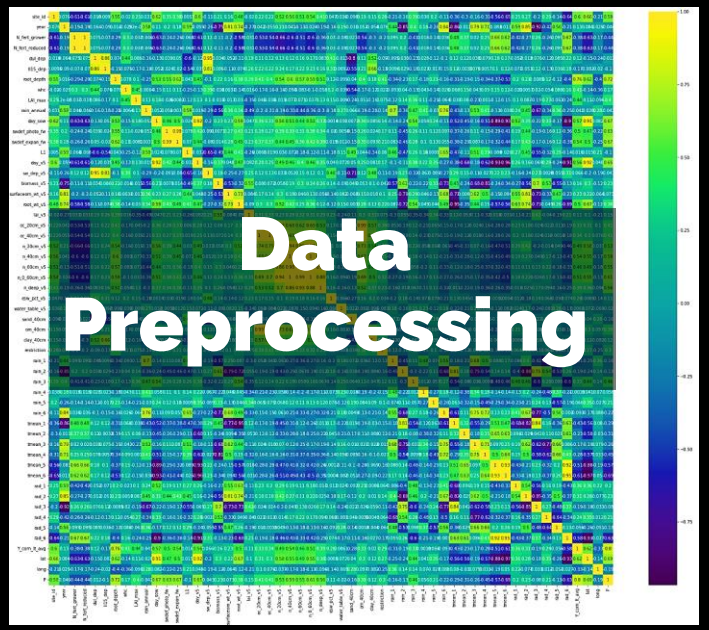


Missingness: **Data missing at random**



EDA: **Visualization revealed predictors' correlation**

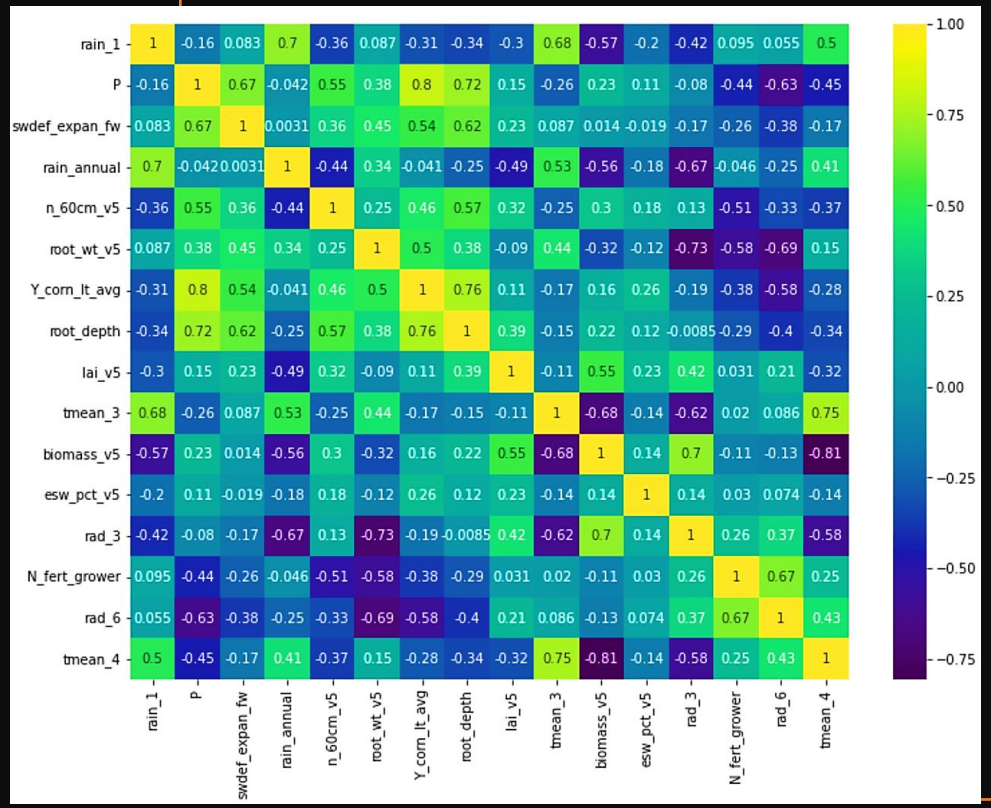
Data Preprocessing



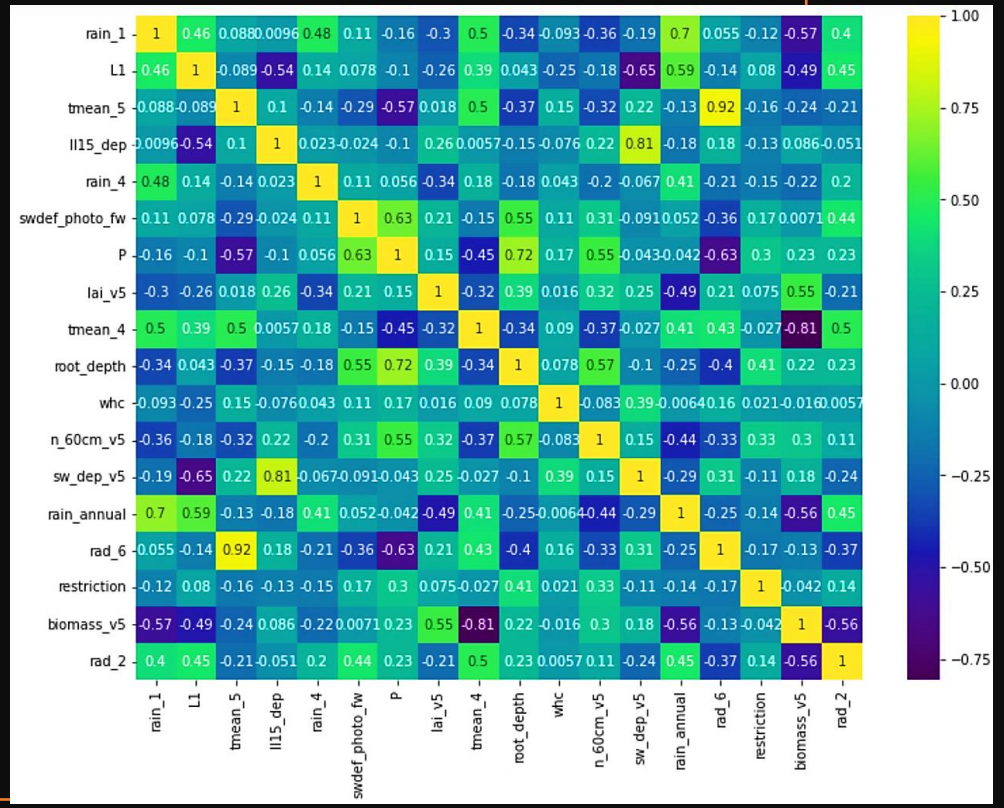
1. EDA aided the removal of redundant features
2. EDA: Removed 16 features
3. Feature Encoding
4. Data Imputation using KNN
5. Lasso and forward Regression step method selected optimal features
6. Scaling using StandardScaler

Dimensionality Reduction

Lasso and Forward Regression in R, was used to select optimal features out of the fifty-five features for both responses



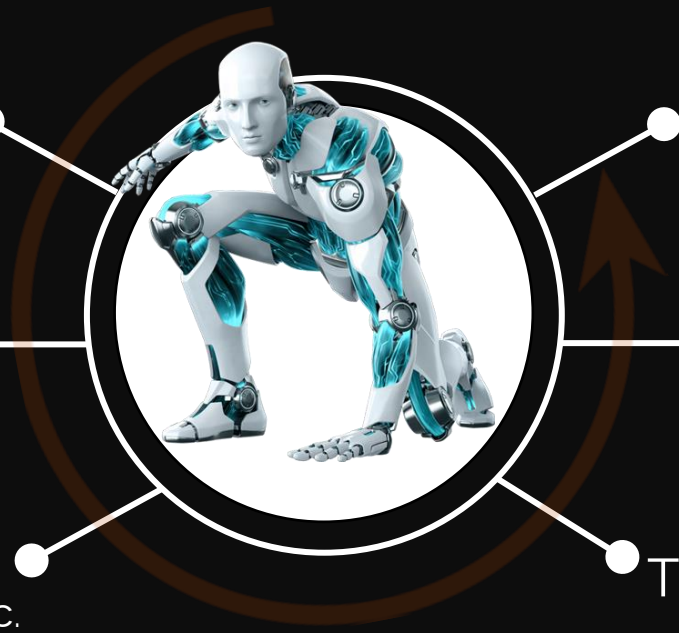
yield_bu_acre Dominant Features



plantN_lb_acre Dominant Feature

Image showing the correlation between the features in each response

Machine Learning Process



Transforming Label to categorical data

Train-Test Splitting

Building Classifiers:
DT, KNN, SVC, RF, NN, etc.

Model Accuracy and Precision using Model scoring and Confusion Matrix

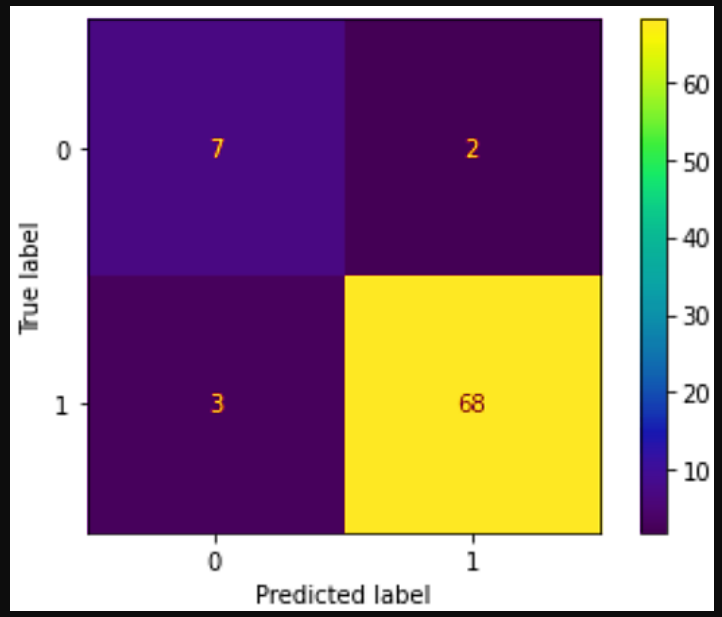
Cross Validation

Training Model

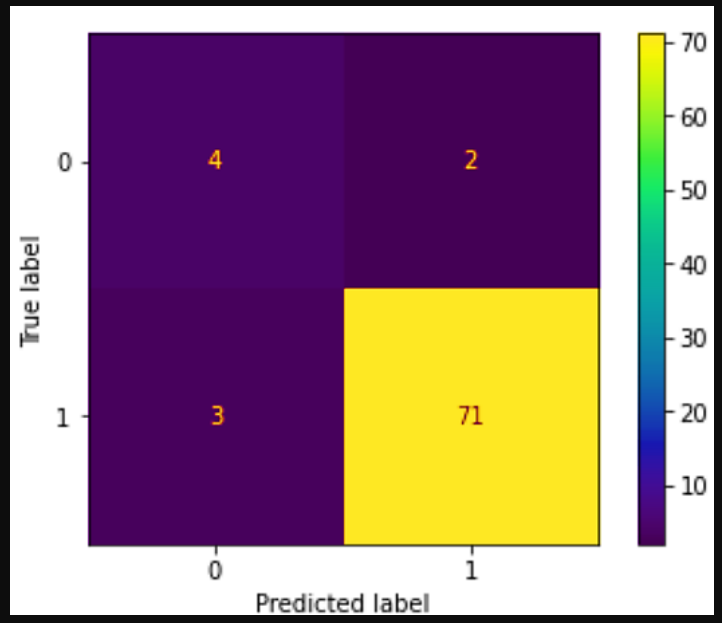
Model Accuracy and Validation

	Classification Algorithm	yield_bu_acre	plantN_lb_acre
0	Decision Tree Classifier	0.9375	0.9375
1	KNeighbors Classifier	0.9375	0.9500
2	SVC(kernel="linear")	0.8875	0.9250
3	SVC(gamma=2)	0.9000	0.9375
4	Gaussian Process Classifier (with RBF)	0.9500	0.9500
5	Decision Tree Classifier (max depth=5)	0.9375	0.9500
6	Random Forest Classifier	0.9250	0.9375
7	MLPClassifier (Neural Network)	0.9250	0.9625
8	AdaBoostClassifier	0.9375	0.9375
9	Gaussian Naive Bayes	0.8625	0.8750
10	Quadratic Discriminant Analysis	0.9375	0.9375

Model Summary Table for the Classification algorithms



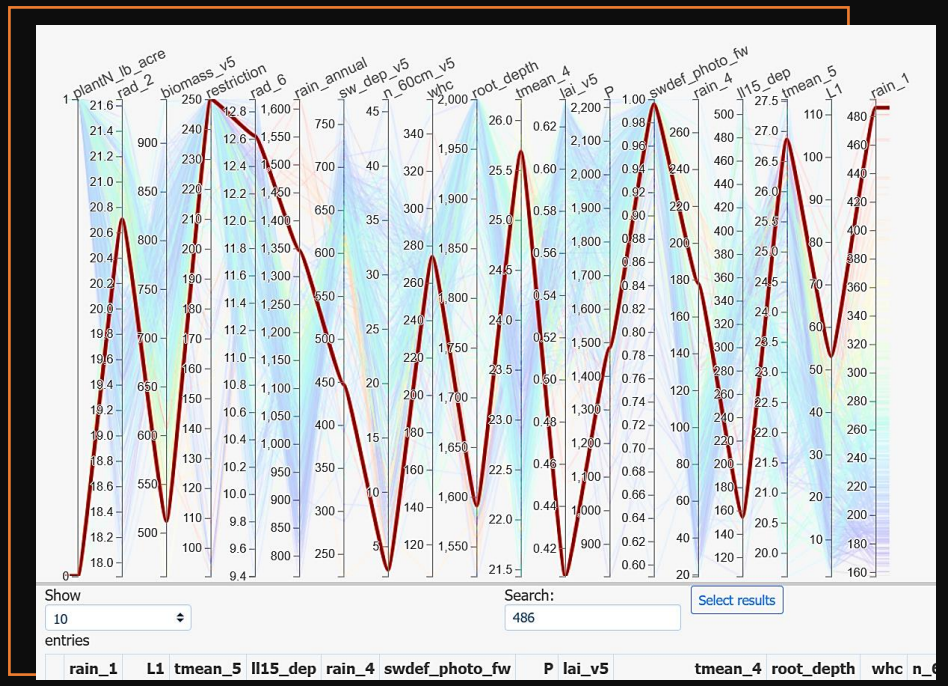
yield_bu_acre Confusion Matrix



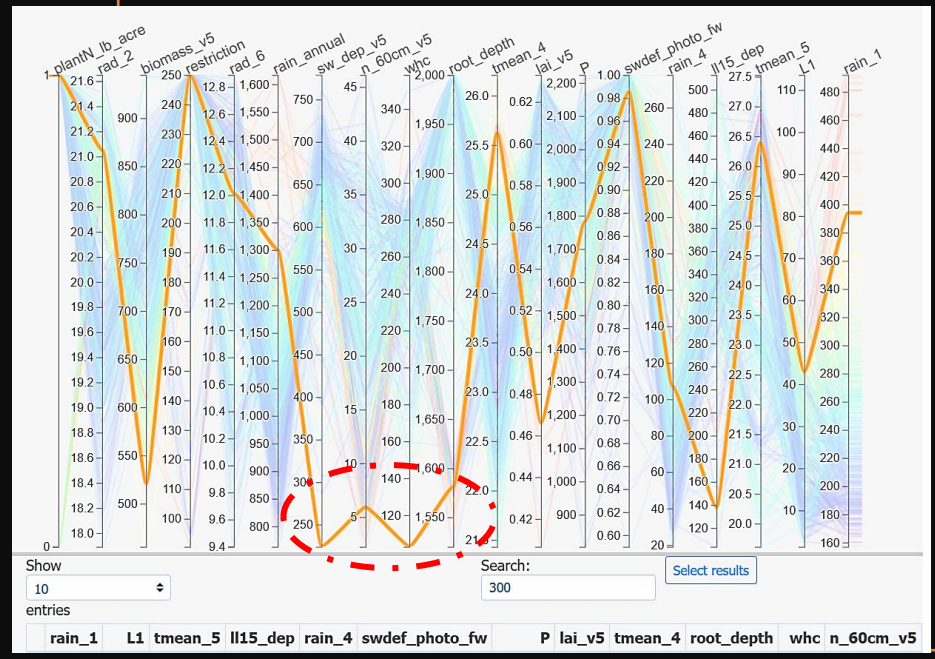
plantN_lb_acre Confusion Matrix

Insights from the Model - plantN_lb_acre

- The High dimensional Interactive Plotting (HiPlot) shows a distinguishable trend between features.
- From the figures, we see a repeatable trend between certain features.
- The Positive response was the only target that had sites with the lowest sw_dep_v5, n_60cm_v5, and whc features



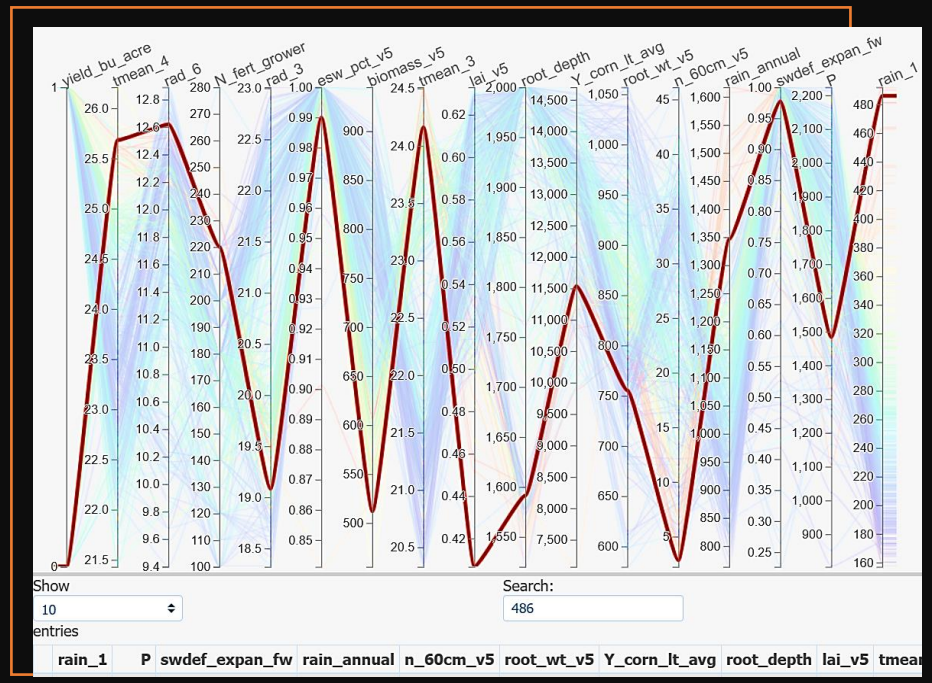
Negative Response



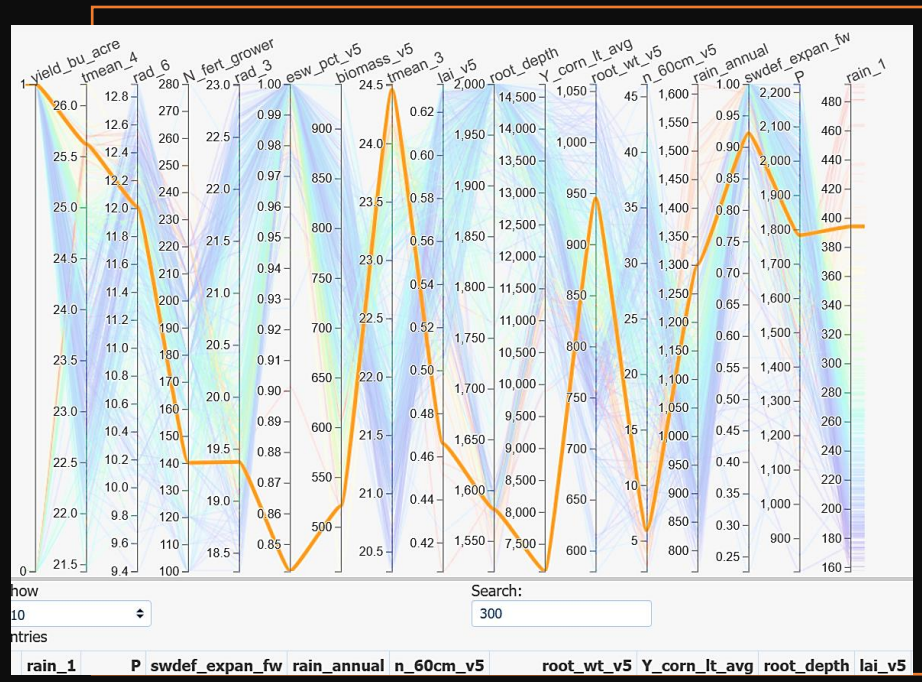
Positive Response

Insights from the Model - yield_bu_acre

- The features used show discernable characteristics for sites that will give respond to N addition.
- Although these are individual instances of the two traits, variances in both scenarios are what the machine learning algorithms learn from.



Negative Response



Positive Response

Conclusion

- The features used for the prediction are key site characteristics as their variations aided accurate predictions.
- The model showed that given relevant features, the response status (1, 0) for both traits could be predicted accurately.
- The ML model developed produced at least 92 % accuracy in identifying experimental sites that are more likely to respond to N fertilizer additions
- Cross-validation using the eleven classification algorithms produced 92% accuracy for the plantN_bu_acre trait and 94% accuracy for plantN_lb_acre
- Therefore, the model can be used to determine future sites that could be used to test new strains of diazotrophic microbes that supply corn with fixed atmospheric Nitrogen (N).