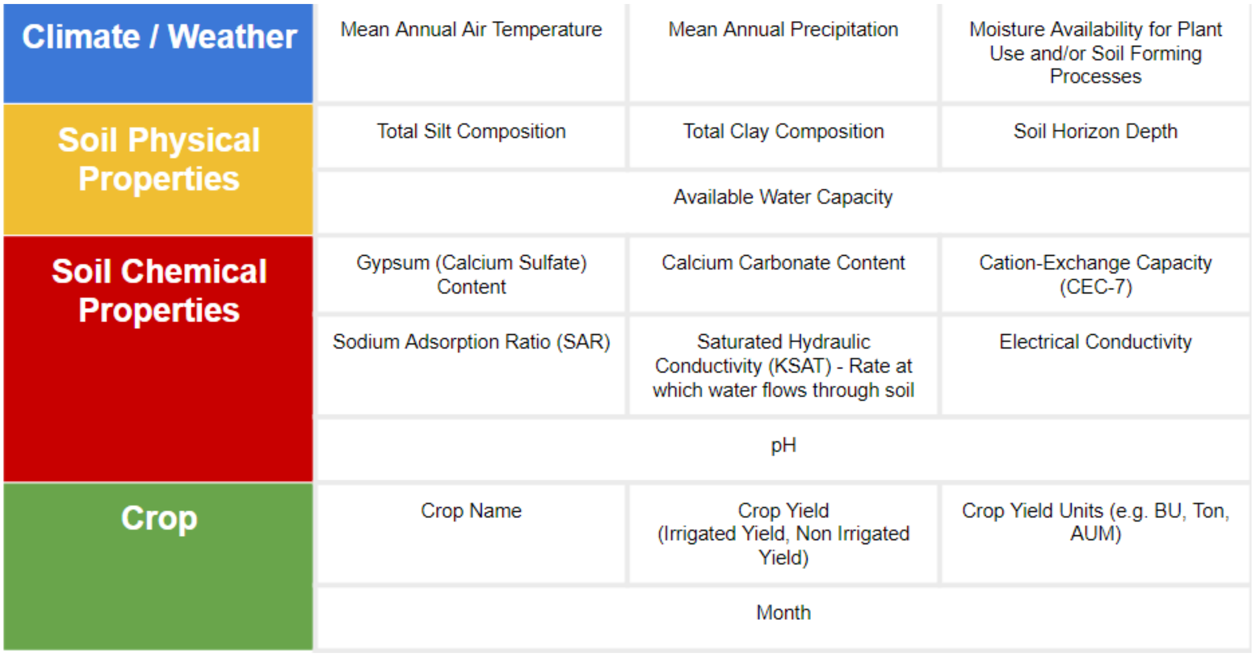
Crop Yield Prediction

Aziza Kurbonova, Florjon Haxhija

We built an ML model that provides a recommendation for farmers on the most suitable crops to grow on a particular farm based on various parameters, such as soil composition and climate data. There is existing research literature that works on this problem as well as a Kaggle problem dataset, defining the most suitable crop as the one that returns the greatest yield. The data that all of this literature works on are limited to India. We intend to replicate the methods used in the literature, with some modifications to the problem scope. First, we intend to curate a similar dataset with the same parameters as found in the literature based in the United States, using publicly available datasets from the U.S. Department of Agriculture. Second, we want to explore adding additional parameters to what constitutes the most suitable crop.

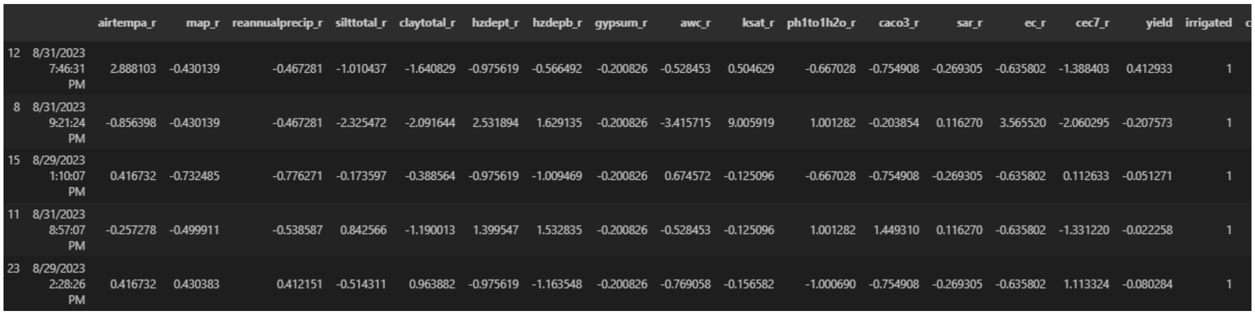
We intend to curate recommendations for the most suitable crop that gives the highest crop yield to farmers in their respective regions. These regions will include the entirety of the United States, specifically demonstrating our results state by county level. We're also intending to give the highest accuracy report on the most suitable crop with a given specific data region.

Before delving into the machine learning model, we must make sure that we're using a suitable dataset for our model. We chose to use the U.S. Department of Agriculture Natural Resource Conservation Service to pull data, roughly 1 million randomly selected data points, and collapse certain features that weren't needed. Here is the contents of the Soil Database that we pulled our data from:



Climate / Weather	Mean Annual Air Temperature	Mean Annual Precipitation	Moisture Availability for Plant Use and/or Soil Forming Processes
Soil Physical Properties	Total Silt Composition	Total Clay Composition	Soil Horizon Depth
	Available Water Capacity		
Soil Chemical Properties	Gypsum (Calcium Sulfate) Content	Calcium Carbonate Content	Cation-Exchange Capacity (CEC-7)
	Sodium Adsorption Ratio (SAR)	Saturated Hydraulic Conductivity (KSAT) - Rate at which water flows through soil	Electrical Conductivity
	pH		
Crop	Crop Name	Crop Yield (Irrigated Yield, Non Irrigated Yield)	Crop Yield Units (e.g. BU, Ton, AUM)
	Month		

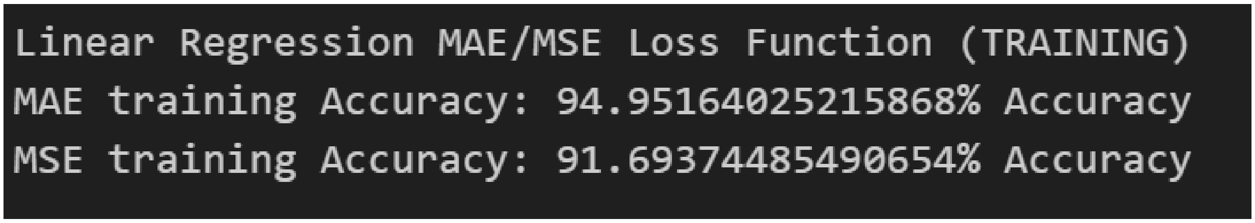
Due to the Natural Resource Conservation Service having empty columns (features), we decided to strip those from our finalized dataset as well with any null values in other categorical columns. The dataset also presented us with two types of yield data, yield with irrigation of the soil, and yield without irrigation. To combine these two types of data into our dataset, we chose to concatenate the yield of the two different states of soil. We also changed the units of all crop yields to be of metric tons, standardized numeric features in our dataset, and created variables in our dataset for our crop name and month. Through all of these efforts, we dwindled our amount of instances in the dataset to 291840 points. A sample of our dataset is shown here:



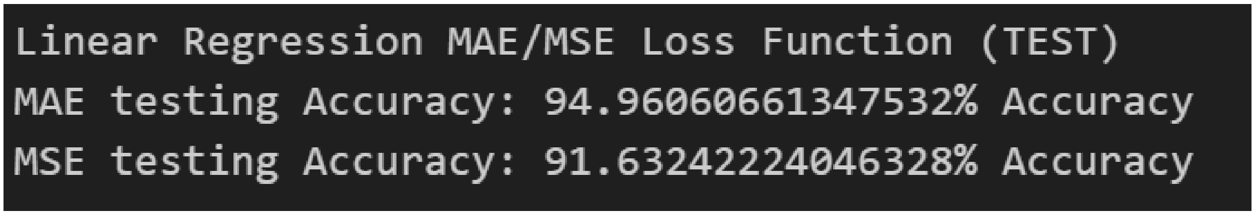
		airtempa_r	map_r	reannualprecip_r	silttotal_r	claytotal_r	hzdept_r	hzdepb_r	gypsum_r	awc_r	ksat_r	ph1to1h2o_r	caco3_r	sar_r	ec_r	cec7_r	yield	irrigated	c
12	8/31/2023 7:46:31 PM	2.888103	-0.430139	-0.467281	-1.010437	-1.640829	-0.975619	-0.566492	-0.200826	-0.528453	0.504629	-0.667028	-0.754908	-0.269305	-0.635802	-1.388403	0.412933	1	
8	8/31/2023 9:21:24 PM	-0.856398	-0.430139	-0.467281	-2.325472	-2.091644	2.531894	1.629135	-0.200826	-3.415715	9.005919	1.001282	-0.203854	0.116270	3.565520	-2.060295	-0.207573	1	
15	8/29/2023 1:10:07 PM	0.416732	-0.732485	-0.776271	-0.173597	-0.388564	-0.975619	-1.009469	-0.200826	0.674572	-0.125096	-0.667028	-0.754908	-0.269305	-0.635802	0.112633	-0.051271	1	
11	8/31/2023 8:57:07 PM	-0.257278	-0.499911	-0.538587	0.842566	-1.190013	1.399547	1.532835	-0.200826	-0.528453	-0.125096	1.001282	1.449310	0.116270	-0.635802	-1.331220	-0.022258	1	
23	8/29/2023 2:28:26 PM	0.416732	0.430383	0.412151	-0.514311	0.963882	-0.975619	-1.163548	-0.200826	-0.769058	-0.156582	-1.000690	-0.754908	-0.269305	-0.635802	1.113324	-0.080284	1	

Moving on to the technical details of the machine learning models, we created a train test split of 2/3 for training and 1/3 for testing. We decided to have a total of 3 different models tested with our dataset, a Linear Regression model with Ridge + Lasso Regularization & MAE/MSE Loss Functions, a Decision Tree Regression model w/ MAE, and a Random Forest Regressor model. To get the lowest possible error between our train and test split, we had to do hyperparameter tuning for each model. For our decision tree model, we chose the lowest error that occurred when the max depth was 20 with a minimum sample split of 5 which also had the lowest MAE at 0.01; Furthermore, for our random forest regressor, we chose our n estimators to be 70, max depth to be 12 and a minimum sample split of 2 which had the lowest test MSE at 0.045.

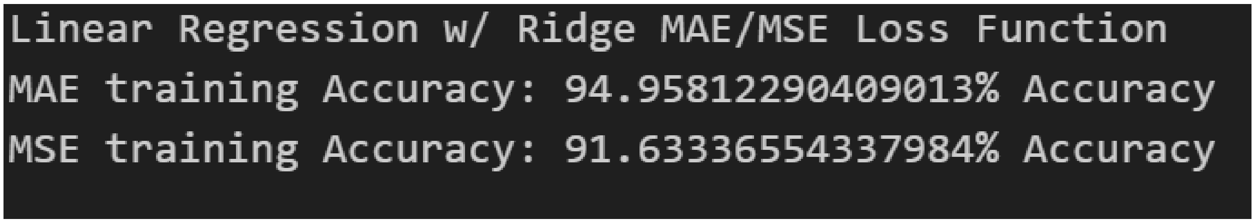
Here are our values of the MAE and MSE for the training and test sets for each model:



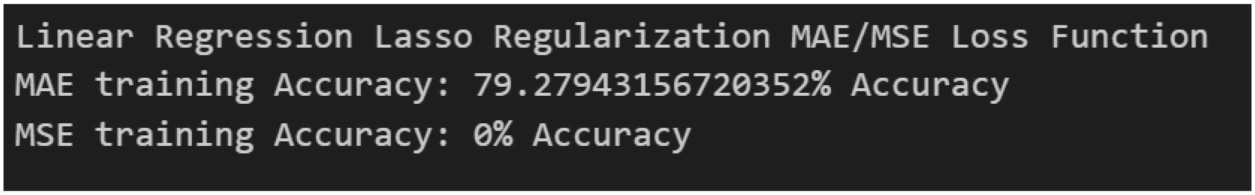
```
Linear Regression MAE/MSE Loss Function (TRAINING)
MAE training Accuracy: 94.95164025215868% Accuracy
MSE training Accuracy: 91.69374485490654% Accuracy
```



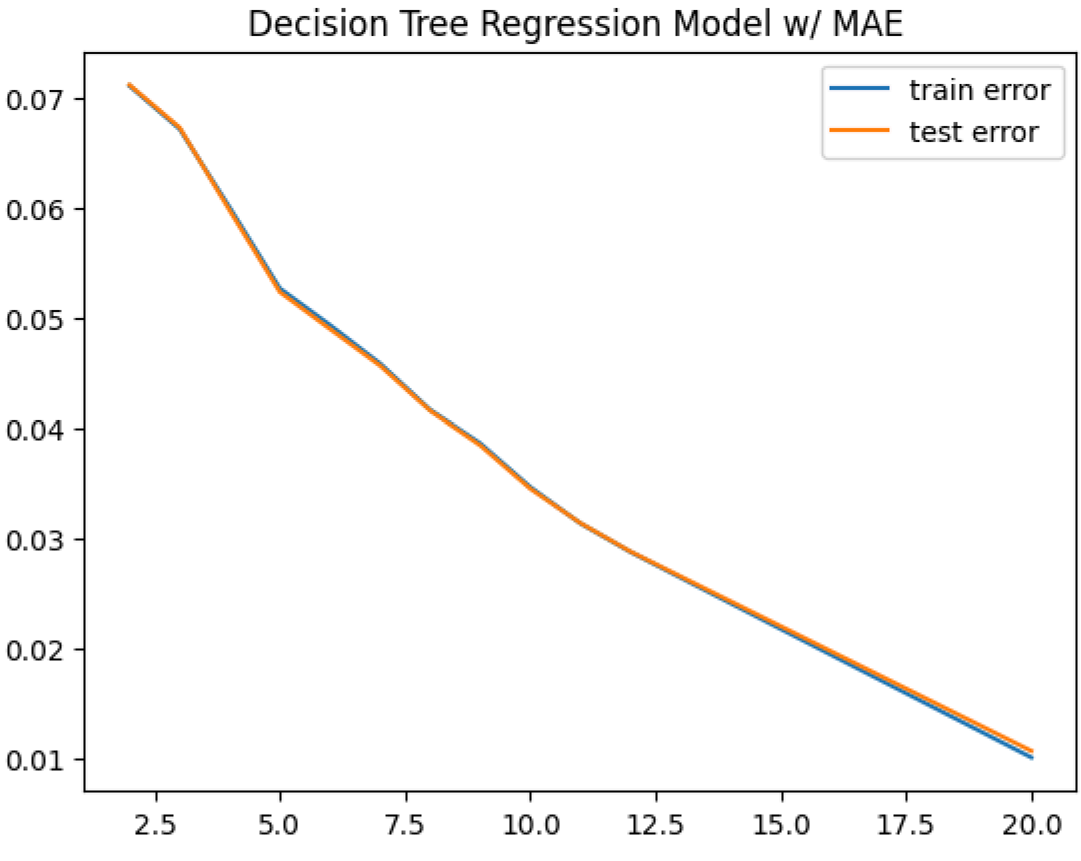
```
Linear Regression MAE/MSE Loss Function (TEST)
MAE testing Accuracy: 94.96060661347532% Accuracy
MSE testing Accuracy: 91.6324224046328% Accuracy
```



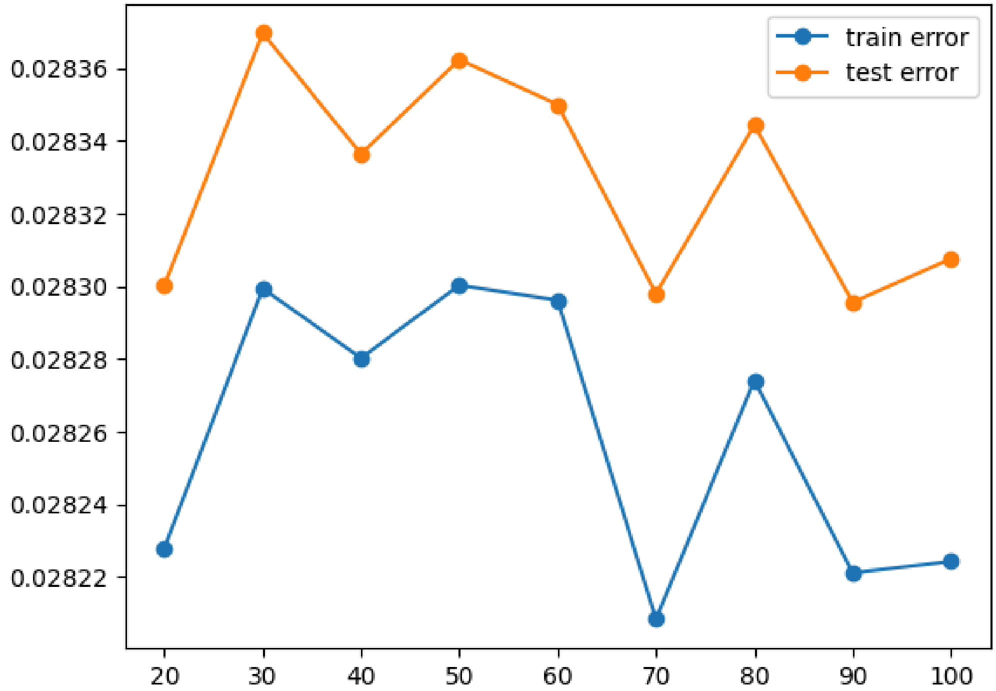
```
Linear Regression w/ Ridge MAE/MSE Loss Function
MAE training Accuracy: 94.95812290409013% Accuracy
MSE training Accuracy: 91.63336554337984% Accuracy
```

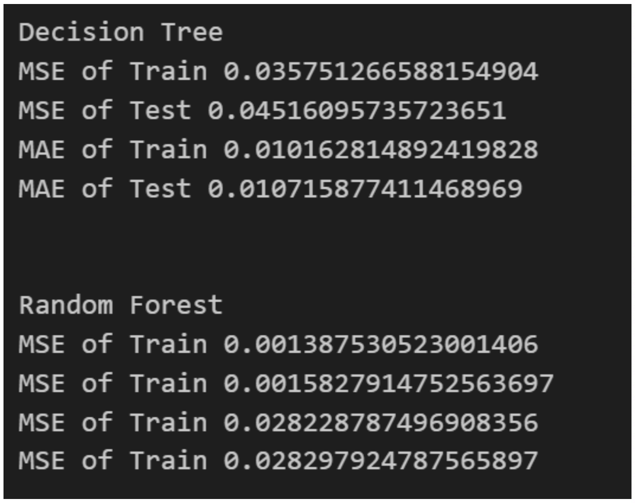


```
Linear Regression Lasso Regularization MAE/MSE Loss Function
MAE training Accuracy: 79.27943156720352% Accuracy
MSE training Accuracy: 0% Accuracy
```



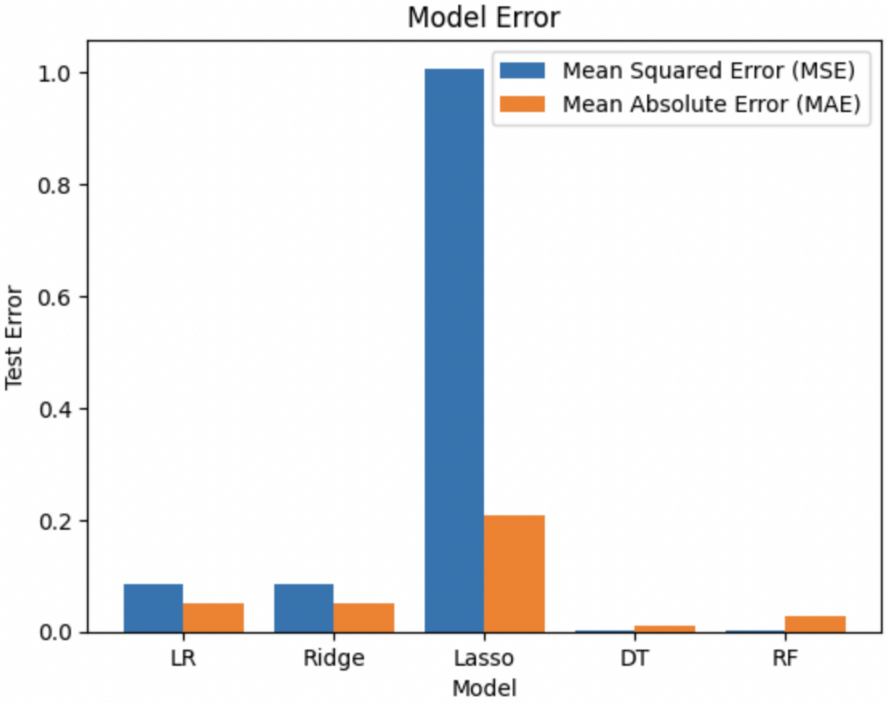
And for the Random Forest Regressor:





```
Decision Tree
MSE of Train 0.035751266588154904
MSE of Test 0.04516095735723651
MAE of Train 0.010162814892419828
MAE of Test 0.010715877411468969


Random Forest
MSE of Train 0.001387530523001406
MSE of Train 0.0015827914752563697
MSE of Train 0.028228787496908356
MSE of Train 0.028297924787565897
```



Overlap between train and test error, implying that the model is not overfitting and is well generalized.

[Notebook]()
[Video Presentation]()

Works Cited:

Dahiphale, Devendra; Shinde, Pratik; Patil, Koninika; Dahiphale, Vijay (2023). Smart Farming: Crop Recommendation using Machine Learning with Challenges and Future Ideas. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.23504496.v1

Hasan M, Marjan MA, Uddin MP, Afjal MI, Kardy S, Ma S and Nam Y (2023) Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. Front. Plant Sci. 14:1234555. doi: 10.3389/fpls.2023.1234555

S. M. PANDE, P. K. RAMESH, A. ANMOL, B. R. AISHWARYA, K. ROHILLA and K. SHAURYA, "Crop Recommender System Using Machine Learning Approach," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1066-1071, doi: 10.1109/ICCMC51019.2021.9418351.