MACHINE LEARNING AND PATTERN RECOGNITION

State-wise Agricultural Prediction Model

A Data-Driven Framework for Production estimation

Aksh Jhawar, Adyant Jha, Krissh Modi

Plaksha University

Problem Statement

Problem

Crop production depends heavily on weather conditions such as temperature, rainfall, and humidity during the growing season. Current production estimates mainly rely on surveys, which can often be inaccurate. This project develops a machine learning model to predict total crop production before harvest using weather data from the crop growing months.

Applications

The model can help governments and policymakers estimate crop production in advance for better agricultural planning, storage management, procurement, irrigation planning, and relief measures for farmers.

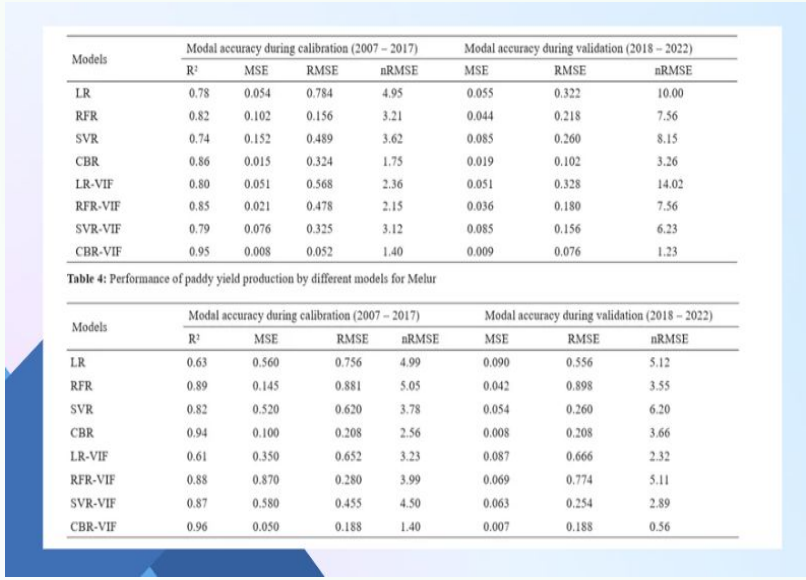
Impact

Early prediction of crop production can reduce dependence on inaccurate surveys, improve policy decisions, support farmers through timely interventions, and strengthen food security and resource management.

Literature Review

1. Weather based paddy yield prediction using machine learning regression algorithms- Journal of Agrometeorology

ISSN: 0972-1665 Issue: September 2024



Models	Modal accuracy during calibration (2007 – 2017)				Modal accuracy during validation (2018 – 2022)		
	R^2	MSE	RMSE	nRMSE	MSE	RMSE	nRMSE
LR	0.78	0.054	0.784	4.95	0.055	0.322	10.00
RFR	0.82	0.102	0.156	3.21	0.044	0.218	7.56
SVR	0.74	0.152	0.489	3.62	0.085	0.260	8.15
CBR	0.86	0.015	0.324	1.75	0.019	0.102	3.26
LR-VIF	0.80	0.051	0.568	2.36	0.051	0.328	14.02
RFR-VIF	0.85	0.021	0.478	2.15	0.036	0.180	7.56
SVR-VIF	0.79	0.076	0.325	3.12	0.085	0.156	6.23
CBR-VIF	0.95	0.008	0.052	1.40	0.009	0.076	1.23

Table 4: Performance of paddy yield production by different models for Melur

Models	Modal accuracy during calibration (2007 – 2017)				Modal accuracy during validation (2018 – 2022)		
	R^2	MSE	RMSE	nRMSE	MSE	RMSE	nRMSE
LR	0.63	0.560	0.756	4.99	0.090	0.556	5.12
RFR	0.89	0.145	0.881	5.05	0.042	0.898	3.55
SVR	0.82	0.520	0.620	3.78	0.054	0.260	6.20
CBR	0.94	0.100	0.208	2.56	0.008	0.208	3.66
LR-VIF	0.61	0.350	0.652	3.23	0.087	0.666	2.32
RFR-VIF	0.88	0.870	0.280	3.99	0.069	0.774	5.11
SVR-VIF	0.87	0.580	0.455	4.50	0.063	0.254	2.89
CBR-VIF	0.96	0.050	0.188	1.40	0.007	0.188	0.56

This research paper predicts agricultural Production for the Paddy crop in Madurai taking into account the following features:

Variable name	Variable ID	Variable type	Description
Maximum temperature	MAXT	Predictor	Maximum temperature for various division
Minimum temperature	MINT	Predictor	Minimum temperature for various division
Rainfall normal	RN	Predictor	Rainfall normal value
Actual rainfall	AR	Predictor	Rainfall actual value
Starting month	SM	Predictor	Starting month for various season
Ending month	EM	Predictor	Ending month for various season
Division name	DN	Predictor	District list in Madurai district
Duration	DUR	Predictor	Duration based on no.of. days
Production	PRD	Target	Production ratio
Crop year	CY	Predictor	Year of crop production
Seed name	SN	Predictor	Collection of paddy name in Madurai districts

Literature Review

2. Crop Selection and Yield Prediction us
Machine Learning- Current Agriculture
Research Journal

Approach ISSN: 2398-4985 Issue:
November 2023

This research paper predicts agricultural yield for a specific crop taking into account only 4 features, City, Crop, Annual Rainfall (in mm), Season.

Uses Kaggle Dataset, which may explain high R 2

Figure 8: Regression Results for Yield Prediction

Literature Review

3. Machine Learning Techniques for Weather based Crop Yield Prediction.-Third International Conference on Artificial Intelligence and Smart Energy (ICAIS 2023)

ISBN: 978-1-6654-6216-7

This research paper predicts agricultural yield for a crop taking into account only 4 features:- Season, District, yield type and Area.

Also uses Kaggle Dataset which may explain higer R 2

Machine learning *Techniques*	Statistics of Performance			
	Accuracy	*RSD*	*R2*	*Adj. R2*
DTR	92.43	5.76	96.82	94.89
RFR	96.67	7.43	97.33	96.39

Dataset & Features I

1. Agronomic & Structural Features

Source: ICRISAT Database

- Irrigation Coverage
- HYV Area (High Yielding Variety)
- Soil Type

2. Core Environmental Features

Source: NASA Weather Database

- Total Rainfall
- Average Temperature
- Humidity

Dataset & Features II

3. Secondary Environmental Features

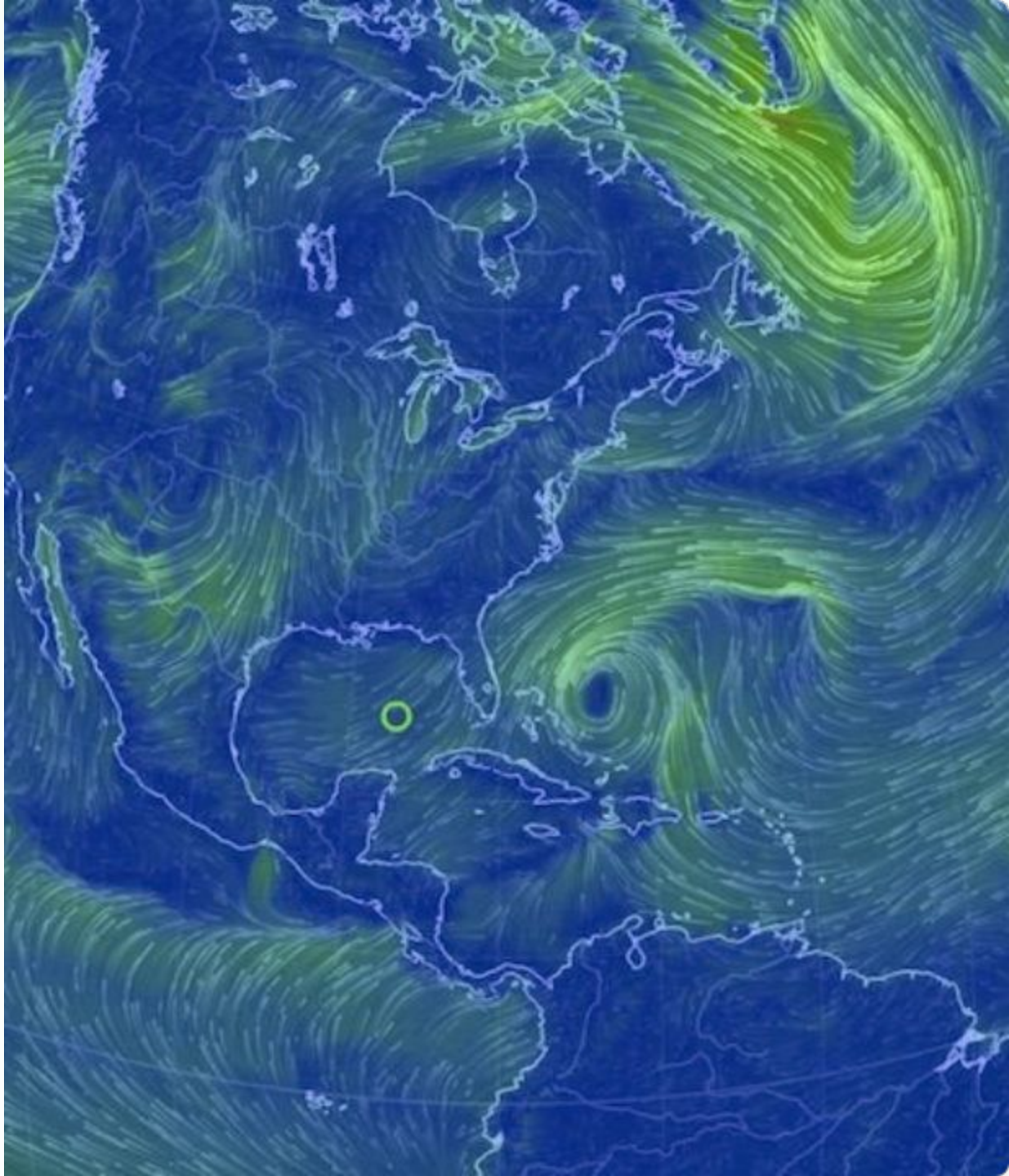
Source: NASA Weather Database

- Sunshine Hours / Solar Radiation
- Wind Speed

4. Derived / Biological Features

Calculated Internally

- Growing Degree Days (GDD)
- Critical Growth Stage Rainfall



Dual-Track Preprocessing Strategy

Shared Initial Steps

Domain-specific feature selection followed by **Standardization** to ensure equal scaling across all features.

> **Temporal Flattening:** Monthly data was reduced to yearly format using sum or mean calculations.

Data Imputation

Addressed missing data for **Solar Radiation (1981-1983)** by employing district-wise mean imputation.

Dataset Bifurcation

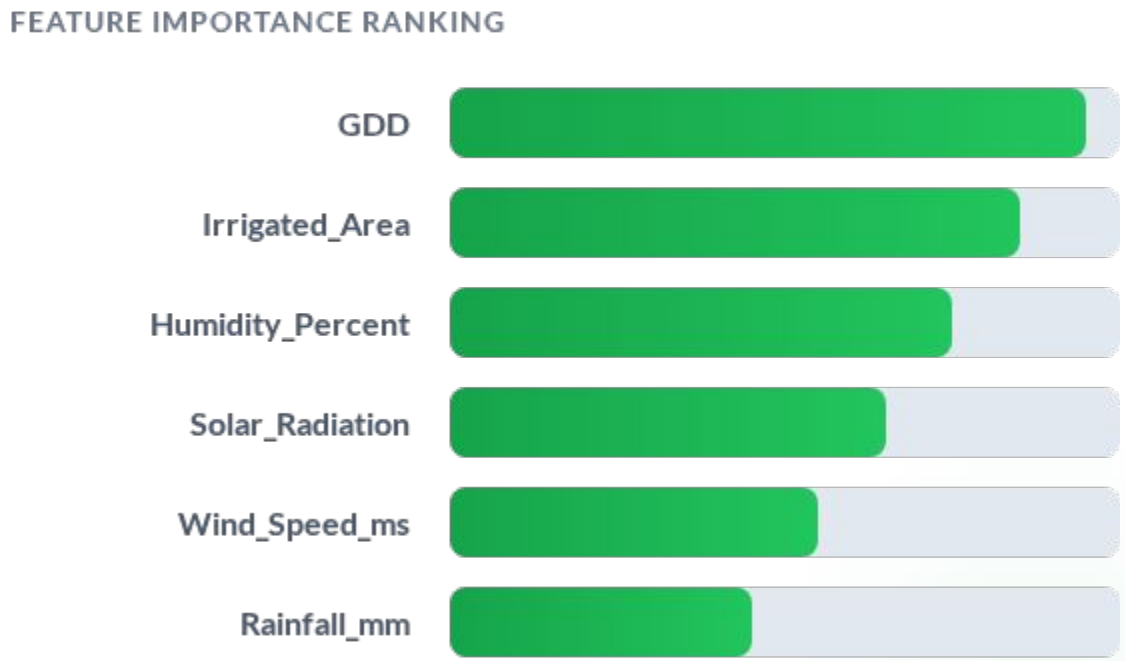
- **Track 1:** Advanced models (NN, XGBoost, SVR, Random Forest).
- **Track 2:** Regularized models (Ridge, Elastic Net).

Dataset 1: Feature Selection & Importance

Refinement Process

- **Correlation Filter:** Dropped redundant variables from highly correlated pairs.
- **Random Forest Test:** Ranked features by objective importance.
- **SHAP Algorithm:** Verified results to ensure biological and logical consistency.

**Dataset 2 utilized original features as selection is inherently handled by Ridge/Elastic Net.*



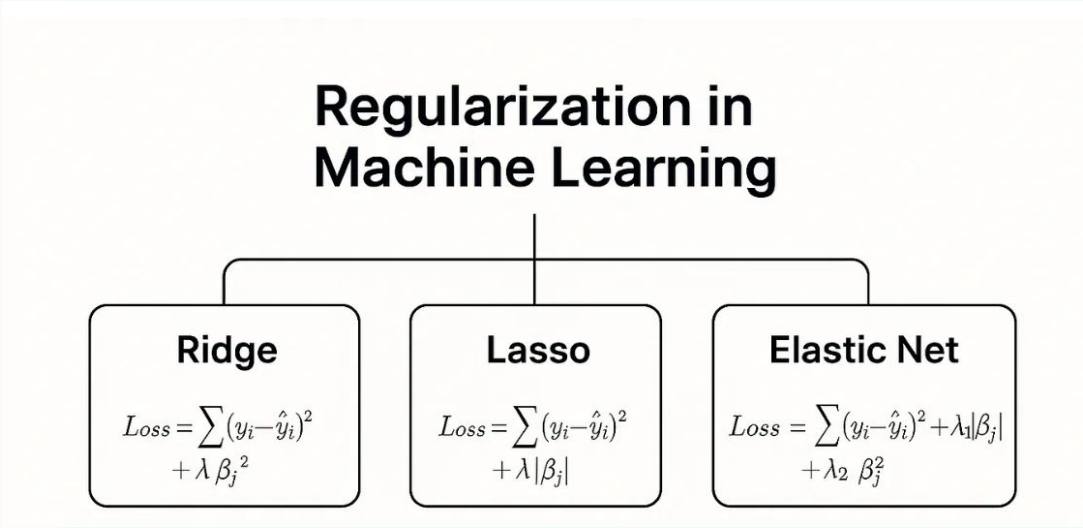
DATASET 2: FEATURE INTEGRATION

Automated Selection

For the second dataset, manual feature selection was deliberately avoided to preserve the full variance of the input space.

Algorithmic Handling: Techniques like Ridge Regression and Elastic Net possess inherent regularization properties.

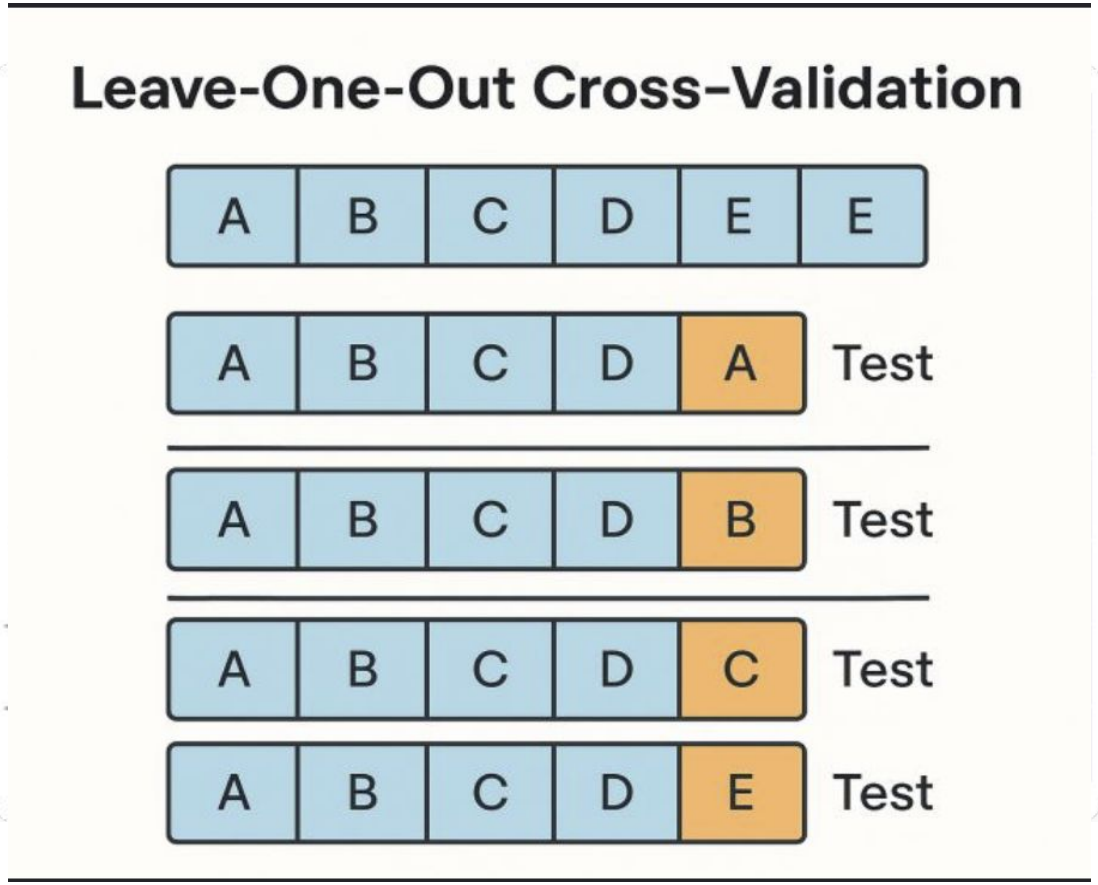
Integrity: We kept all original features, allowing the models to perform internal coefficient shrinkage and selection.



Strategic Decision:

Selection is managed by L1 and L2 penalties during model training.

Validation & Data Augmentation



Handling Small Datasets

After temporal flattening, the dataset size was restricted to approximately **~400 rows**, requiring robust validation and augmentation strategies.

- **LOOCV:** Leave One Out Cross Validation used for train-test splitting.
- **Augmentation:** Applied to the training set only to enhance model generalization.

Augmentation Techniques Used:

Gaussian Jitter, Block Bootstrap, Mixup, etc.

MODELS USED FOR CROP YIELD PREDICTION

To improve prediction accuracy, we used multiple machine learning models because different algorithms capture different patterns in agricultural data such as temperature, rainfall, humidity, soil conditions, and growing degree days (GDD).

1. Ridge Regression

MACHINE LEARNING

Why we used it:

Ridge Regression works well for datasets where input features are highly correlated, which is common in climate and crop data. It also helps reduce overfitting.

How it works (briefly):

It is a linear regression model that adds a penalty term to keep the model weights small. This improves stability and generalization.

2. Elastic Net

MACHINE LEARNING

Why we used it:

Elastic Net is useful when there are many features and some may be less important. It performs both feature selection and regularization.

How it works (briefly):

It combines the properties of Ridge and Lasso Regression by applying two types of penalties, helping the model select important variables while avoiding overfitting.

MODELS USED FOR CROP YIELD PREDICTION

3. Random Forest

MACHINE LEARNING

Why we used it:

Random Forest handles nonlinear relationships very well and performs strongly on agricultural datasets with complex environmental interactions.

How it works (briefly):

It creates many decision trees using random subsets of data and features. The final prediction is obtained by averaging the outputs of all trees.

4. Gradient Boosting

MACHINE LEARNING

Why we used it:

Gradient Boosting often provides high prediction accuracy by learning from previous errors step by step.

How it works (briefly):

It builds decision trees sequentially, where each new tree tries to correct the mistakes made by earlier trees.

MODELS USED FOR CROP YIELD PREDICTION

5. Support Vector Regression (SVR – RBF)

MACHINE LEARNING

Why we used it:

SVR is effective for capturing nonlinear patterns in crop and climate data, especially when relationships between variables are complex.

How it works (briefly):

Using the RBF (Radial Basis Function) kernel, SVR maps data into a higher-dimensional space and finds the best curve that fits the data within an acceptable error range.

6. Neural Network (MLP)

MACHINE LEARNING

Why we used it:

Neural Networks are powerful for learning highly complex and nonlinear relationships in large agricultural datasets.

How it works (briefly):

An MLP (Multi-Layer Perceptron) consists of interconnected layers of neurons that learn patterns by adjusting weights during training using backpropagation.

OVERALL APPROACH

Using multiple models allows comparison of performance metrics such as:

- MAE (Mean Absolute Error)
- √× RMSE (Root Mean Square Error)
- % R^2 Score

Model	RMSE	R2	MAPE
Ridge Regression	164.28	0.5969	29.07%
Elastic Net	156.79	0.5905	28.45%
Random Forest	165.84	0.4410	31.20%
Gradient Boosting	182.96	0.3993	30.18%
SVR (RBF)	189.90	0.4017	27.90%
Neural Network (MLP)	185.94	0.5490	30.66%

This helps identify the most suitable model for accurate crop yield prediction under varying environmental conditions.

CHALLENGES FACED DURING THE PROJECT

Dataset Availability

One of the major challenges we faced was dataset availability and limited data points. The crop production data was available only on a yearly basis, which resulted in a very small dataset for training machine learning models.
Since ML models generally perform better with larger datasets, this became a significant limitation.

Data Augmentation

To address this issue, we used several data augmentation techniques to synthetically increase the dataset size and improve model generalization:

- **Gaussian Jittering** – adding small random noise to numerical features to create slightly varied samples.
- **Mixup** – combining multiple samples to generate new synthetic data points.
- **Bootstrap Sampling** – creating multiple resampled datasets from the original data.

CHALLENGES FACED DURING THE PROJECT

Feature Engineering

We divided each year into different crop growth stages (such as **sowing, vegetative growth, flowering, and harvesting**) so that weather and environmental conditions during each stage could be analyzed separately. This improved the biological relevance of the data but also increased the number of features significantly.

Feature Selection Dilemma

For feature selection, we initially used Pearson correlation analysis. However, we observed that Pearson correlation sometimes selected features that were not agriculturally meaningful, such as:

SELECTED (NOT MEANINGFUL)

-SOIL MOISTURE AFTER HARVEST
- WIND SPEED AFTER HARVEST

IGNORED (IMPORTANT)

-RAINFALL DURING GROWING PERIOD
- SOIL MOISTURE DURING SOWING

FUTURE CHALLENGES

- One challenge when scaling the solution is handling large amounts of weather and crop data efficiently, which may increase computational and storage requirements.

- Another challenge is that crops and regions react differently to changing weather conditions, so the model may require regular retraining and updates to deliver good values of performance metrics.

Thank You

Any questions?