

Predictive Artificial Intelligence for Multi-Crop Yield Optimization in Bangladesh under Climate Variability

Onjon Sutradhar

Student ID: 2431963

MSc Data Science and Artificial Intelligence



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Table of contents

01 The Problem

What is the problem statement and motivation behind the project ?

02 Existing Works

What others have done ? What are their techniques ?

03 Methodology

What have we done ? What are the steps followed ?

04 Experimentation

Experimental Setup

05 Results

Implementation and results of the experiments in real world scenarios

06 Conclusions

What is the conclusion ? And What are the future work ideas

01

Problem Statement

Problem Statement

How can we accurately predict multi-crop yield under climate variability using machine learning?

Agriculture is a major part of Bangladesh's economy, employing over 40% of the workforce. However, it is highly vulnerable to climate variability such as irregular rainfall, floods, and temperature changes.

The Real Issue

- **Climate variability disrupts yield patterns**

Climate in Bangladesh is highly unpredictable, with irregular rainfall, floods, droughts, and temperature fluctuations. These changes directly affect planting schedules, crop growth cycles, and final production.

- **Traditional methods rely on experience**

Farmers primarily depend on historical knowledge, seasonal patterns, and intuition passed down over generations.

- **Cannot capture complex, non-linear relationships**

Crop yield depends on multiple interacting factors such as temperature, humidity, soil conditions, and cultivation area.

Motivation

There is a strong need for data-driven agricultural decision systems to improve:

- Crop selection
- Resource allocation
- Yield optimization

Goal

Multi-crop prediction
Climate-aware decision support
Real-world usability

02

Existing Works

Machine Learning for Crop Yield Prediction

- Jeong et al. (2016): Applied Random Forest using weather and remote sensing data → achieved high prediction accuracy ($R^2 > 0.85$) and captured non-linear relationships effectively.
- Khaki & Wang (2019): Used deep neural networks for corn yield prediction → outperformed traditional ML models but required large-scale datasets and high computation.
- Patil et al. (2025): Compared multiple regression models (RF, KNN, XGBoost, AdaBoost) → showed model performance varies across different crops, no single best model.

Machine Learning in Agriculture

- Shan He et al. (2022): Combined feature selection with ML classifiers → improved crop classification accuracy using remote sensing data.
- Prem Kumar et al. (2024): Built soil-based crop recommendation system → enabled real-time crop selection using classification models.
- Bhullar et al. (2023): Used semi-supervised deep learning for land suitability → predicted multi-crop suitability with limited labeled data.

Limitations

- Mostly single-crop studies
- Focus on developed countries
- Poor real-world deployment

Research Gap

- Lack of multi-crop models
- Limited work in developing regions
- No balance between performance & interpretability

03

Methodology

Proposed Methodology

- **Pipeline:**

- Data Collection - First a dataset is picked carefully
- EDA - Exploratory Data Analysis is done
- Feature Engineering - Choose effective Features
- Model Training - Train models from traditional to ensemble
- Evaluation - Evaluate all models and compare among them

Dataset

SPAS-Dataset-BD

Public agricultural dataset

(Mendeley Data)

Peer-reviewed (Data in Brief)

Ensures data quality and reliability

Dataset Coverage

Geographical Scope

64 districts across Bangladesh

Covers all agro-ecological zones

Temporal Scope

Kharif-1 (Apr–Jun)

Kharif-2 (Jul–Oct)

Rabi (Nov–Mar)

Crop Diversity

73 crops

Cereals, pulses,
vegetables, fruits, cash
crops

Dataset

Kharif-1 Crops

- Aus, Jute, Maize 1
- Mango, Jackfruit, Guava
- Lady's Finger, Kakrol, Jhinga
- Green Coconut, Green Papaya

Kharif-2 Crops

- Aman, Arhar, Lentil
- Cabbage, Cauliflower, Carrot
- Garlic, Gram, Beans
- Tobacco, Sugarcane

Rabi Crops

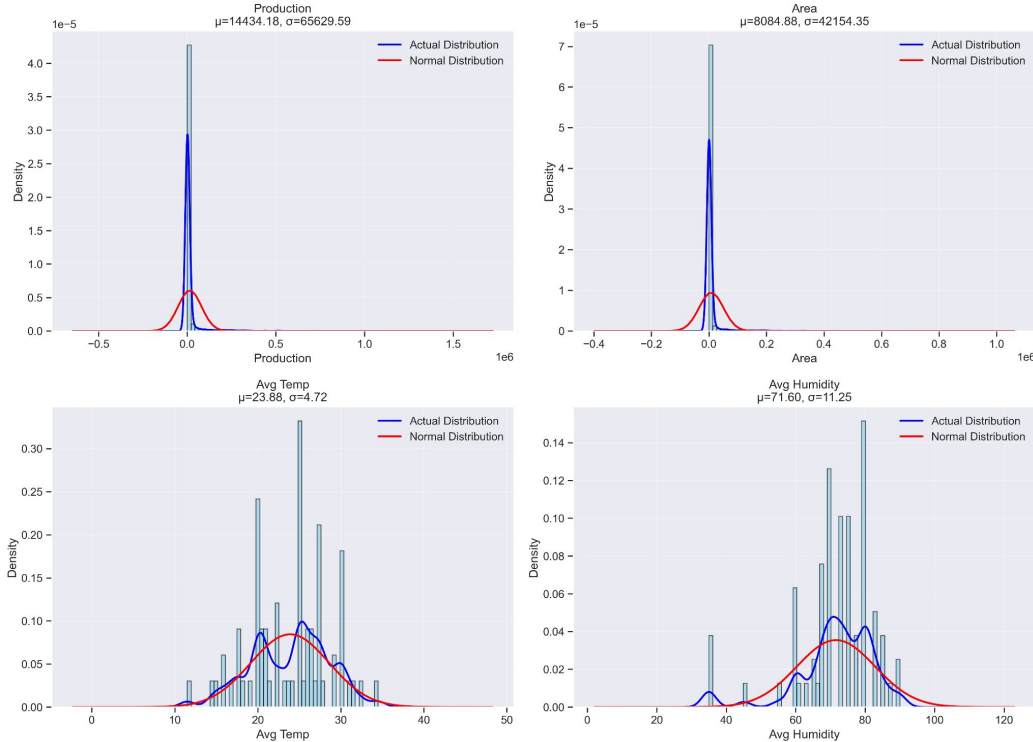
- Boro, Wheat, Maize 2
- Banana, Mango (late), Lemon
- Chili, Ginger, Onion
- Sweet Potato, Sesame

04

Implementation & Experimentation

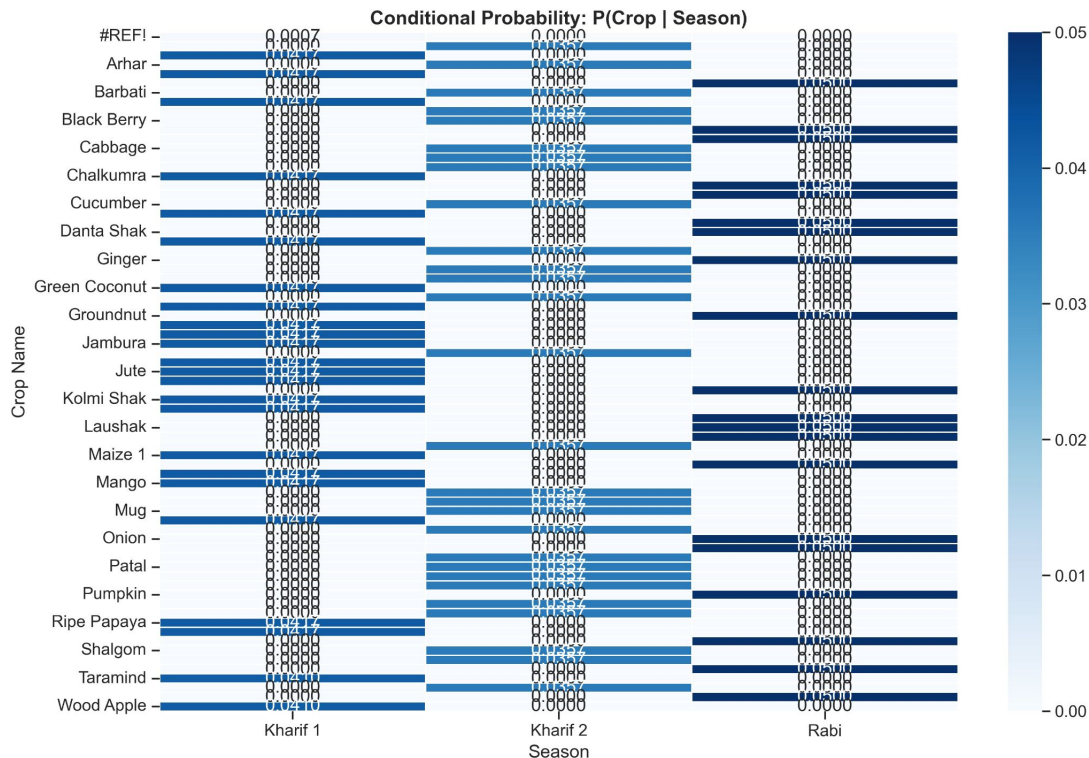
EDA

Distribution Analysis with Normal Curve Overlay



Normal Curve Overlays on Histograms.

EDA



Conditional Probability Heatmap $P(\text{Crop} | \text{Season})$. Strong diagonal and block patterns demonstrate

deterministic crop-season relationships aligned with Bangladesh's monsoon agricultural calendar

EDA

Season Distribution Summary:

Kharif-1 (April-June): 25 unique crops, 1,600 observations

Kharif-2 (July-October): 28 unique crops, 1,792 observations

Rabi (November-March): 20 unique crops, 1,280 observations

Data Preprocessing Step

Algorithm 1 Data Loading and Validation

```
1: Input: CSV file path
2: Output: Validated DataFrame  $D$ 
3:
4:  $D \leftarrow \text{LoadCSV}(\text{filepath})$ 
5:  $D[\text{District}] \leftarrow \text{ConvertToCategory}(D[\text{District}])$ 
6:  $D[\text{Season}] \leftarrow \text{ConvertToCategory}(D[\text{Season}])$ 
7:  $D[\text{Crop Name}] \leftarrow \text{ConvertToCategory}(D[\text{Crop Name}])$ 
8:
9: if  $D.\text{columns} \neq 15$  then
10:   raise  $\text{ValidationError}(\text{"Unexpected column count"})$ 
11: end if
12: if  $D[\text{Season}].\text{unique}().\text{count}() \neq 3$  then
13:   raise  $\text{ValidationError}(\text{"Expected 3 seasons"})$ 
14: end if
15: return  $D$ 
```

Dataset is loaded and
validated

Data Preprocessing Step

Missing values
are treated

Algorithm 2 Missing Value Treatment

```
1: Input: DataFrame  $D$ 
2: Output: Cleaned DataFrame  $D'$ 
3:
4:  $D[\text{AP Ratio}] \leftarrow \text{ConvertToNumeric}(D[\text{AP Ratio}])$ 
5:  $D' \leftarrow \text{DropRows}(D, \text{where Season is null})$ 
6:
7: for each column  $c$  in numeric columns of  $D'$  do
8:   if  $c$  has missing values then
9:      $\text{median}_c \leftarrow \text{Median}(D'[c], \text{ignoring nulls})$ 
10:     $D'[c] \leftarrow \text{FillNA}(D'[c], \text{median}_c)$ 
11:   end if
12: end for
13: return  $D'$ 
```

Data Preprocessing Step

Algorithm 3 Outlier Detection using IQR Method

- 1: **Input:** Dataset D , column c , factor $f = 1.5$
 - 2: **Output:** Set of outlier indices O
 - 3:
 - 4: $Q_1 \leftarrow \text{Quantile}(D[c], 0.25)$
 - 5: $Q_3 \leftarrow \text{Quantile}(D[c], 0.75)$
 - 6: $IQR \leftarrow Q_3 - Q_1$
 - 7: $lower \leftarrow Q_1 - f \times IQR$
 - 8: $upper \leftarrow Q_3 + f \times IQR$
 - 9: $O \leftarrow \{i : D[c][i] < lower \text{ or } D[c][i] > upper\}$
 - 10: **return** $O, lower, upper$
-

Outliers are identified
and bounds are set

Data Preprocessing Step

Algorithm 4 Agricultural Feature Engineering

```
1: Input: Feature matrix  $X$  with raw variables
2: Output: Augmented feature matrix  $X'$  with engineered features
3:
4:  $X' \leftarrow X$                                 ▷ Copy original features
5:
6:                                ▷ Temperature-based features
7:  $X'[\text{Temp\_Range}] \leftarrow X[\text{Max Temp}] - X[\text{Min Temp}]$ 
8:  $X'[\text{Temp\_Deviation}] \leftarrow X[\text{Avg Temp}] - 25$                                 ▷ 25°C baseline
9:
10:                                ▷ Humidity-based features
11:  $X'[\text{Humidity\_Range}] \leftarrow X[\text{Max RH}] - X[\text{Min RH}]$ 
12:  $X'[\text{Humidity\_Deviation}] \leftarrow X[\text{Avg Humidity}] - 70$                                 ▷ 70% baseline
13:
14:                                ▷ Interaction features
15:  $X'[\text{Temp\_Humidity\_Int}] \leftarrow X[\text{Avg Temp}] \times X[\text{Avg Humidity}]$ 
16:  $X'[\text{Area\_AP\_Int}] \leftarrow X[\text{Area}] \times X[\text{AP Ratio}]$ 
17:
18: return  $X'$ 
```

Essential Features are
identified and Scaled

Method 1

Approach V1: Global Quantile-Based Classification

Production values are categorized into balanced low–very high classes using global quantile thresholds.

Algorithm 5 Global Quantile-Based Yield Categorization (V1)

```
1: Input: Dataset  $D$  with Production values
2: Output: Dataset  $D$  with Yield_Class labels
3:
4:  $Q \leftarrow \text{ComputeQuantiles}(D[\text{Production}], [0.25, 0.50, 0.75])$ 
5:
6: for each row  $i$  in  $D$  do
7:   if  $D[\text{Production}][i] \leq Q_{0.25}$  then
8:      $D[\text{Yield\_Class}][i] \leftarrow \text{"Low"}$ 
9:   else if  $D[\text{Production}][i] \leq Q_{0.50}$  then
10:     $D[\text{Yield\_Class}][i] \leftarrow \text{"Medium"}$ 
11:  else if  $D[\text{Production}][i] \leq Q_{0.75}$  then
12:     $D[\text{Yield\_Class}][i] \leftarrow \text{"High"}$ 
13:  else
14:     $D[\text{Yield\_Class}][i] \leftarrow \text{"Very High"}$ 
15:  end if
16: end for
17: return  $D$ 
```

Approach V2: Crop-Aware Percentile Ranking

Yield classes are crop-specific percentile-based, with global quantiles used when data is insufficient.

Method 2

Algorithm 6 Crop-Aware Yield Classification (V2)

```
1: Input: Dataset  $D$  with Production and Crop Name
2: Output: Dataset  $D$  with crop-relative Yield_Class
3:
4: for each unique crop  $c$  in  $D[\text{Crop Name}]$  do
5:    $D_c \leftarrow$  Select rows where Crop Name =  $c$ 
6:   if  $|D_c| \geq 8$  then                                ▷ Sufficient samples for ranking
7:      $\text{rank}_c \leftarrow \text{PercentileRank}(D_c[\text{Production}])$ 
8:     Apply cut:  $\text{rank}_c \in [0, 0.25] \rightarrow \text{"Low"}$ ,
9:                $[0.25, 0.50] \rightarrow \text{"Medium"}$ ,
10:               $[0.50, 0.75] \rightarrow \text{"High"}$ ,
11:               $[0.75, 1.00] \rightarrow \text{"Very High"}$ 
12:   else                                                ▷ Sparse crop: use global quantiles
13:     Apply global quantile approach as in V1
14:   end if
15: end for
16: return  $D$ 
```

Method 3 - Final Regression Pipeline

The regression pipeline preprocesses data, splits it, and trains models evaluated by R^2 , RMSE, and MAE.

Algorithm 12 Final Regression Pipeline with Ensemble Methods

```

1: Input: Cleaned dataset  $D$  with Production as continuous target
2: Output: Best ensemble regressor  $M^*$  with performance metrics
3:
4:                                     ▷ Preprocessing Phase
5: Drop features: {Yield_Per_Area, Transplant, Growth, Harvest}           ▷ Prevent leakage
6:  $X \leftarrow$  Feature matrix from  $D$ ,  $y \leftarrow D[\text{Production}]$ 
7: Apply RobustScaler to all numeric features in  $X$ 
8: Apply LabelEncoder to categorical features
9: Split  $(X_{train}, X_{val}, X_{test}, y_{train}, y_{val}, y_{test}) \leftarrow \text{StratifiedSplit}(X, y)$ 
10:
11:                                     ▷ Base Model Training Phase
12: Initialize gradient boosting models:
13:    $M_1 \leftarrow \text{XGBoostRegressor}(n\_estimators=100, \text{max\_depth}=7, \text{learning\_rate}=0.1)$ 
14:    $M_2 \leftarrow \text{LightGBMRegressor}(n\_estimators=100, \text{num\_leaves}=31)$ 
15:    $M_3 \leftarrow \text{CatBoostRegressor}(n\_estimators=100, \text{depth}=6)$ 
16:    $M_4 \leftarrow \text{RandomForestRegressor}(n\_estimators=100)$ 
17:    $M_5 \leftarrow \text{ExtraTreesRegressor}(n\_estimators=100)$ 
18:
19: for each model  $M_i$  do
20:   Train:  $M_i.\text{fit}(X_{train}, y_{train})$ 
21:    $\hat{y}_{val}^{(i)} \leftarrow M_i.\text{predict}(X_{val})$ 
22:   Compute:  $R_i^2, RMSE_i, MAE_i$  on validation set
23: end for
24:
25:                                     ▷ Ensemble Construction Phase

```

Method 3 - Final Regression Pipeline

Ensembles combine base models via weighting and stacking, then the best is chosen by highest R^2 with metrics and feature importance reported.

```

26:  $M_{vote} \leftarrow \text{VotingRegressor}(\$ 
27:     estimators= $[(M_1, M_2, M_3, M_4)]$ ,
28:     weights= $[2, 2, 2, 1]$ 
29: )
30:
31:  $M_{stack} \leftarrow \text{StackingRegressor}(\$ 
32:     estimators= $[(M_1, M_2, M_3, M_4)]$ ,
33:     final_estimator= $\text{RidgeRegression}(\alpha = 1.0)$ ,
34:     cv=5
35: )
36:
37: Train ensemble models:  $M_{vote}.fit(X_{train}, y_{train})$ 
38:                         $M_{stack}.fit(X_{train}, y_{train})$ 
39:
40:
41: Evaluate all models on  $X_{test}$ :
42: for each model  $M_j \in \{\text{all trained models}\}$  do
43:      $\hat{y}_{test}^{(j)} \leftarrow M_j.predict(X_{test})$ 
44:      $R_j^2 \leftarrow \text{Compute}(y_{test}, \hat{y}_{test}^{(j)})$ 
45:      $RMSE_j \leftarrow \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{test}[i] - \hat{y}_{test}^{(j)}[i])^2}$ 
46:      $MAE_j \leftarrow \frac{1}{n} \sum_{i=1}^n |y_{test}[i] - \hat{y}_{test}^{(j)}[i]|$ 
47: end for
48:
49:  $M^* \leftarrow \arg \max_{M_j} R_j^2$ 
50: return  $M^*$ , test metrics, feature importances

```

▷ Weight boosting models higher
 ▷ Use cross-validation for meta-features
 ▷ Final Evaluation Phase
 ▷ Select best model by R^2 score

05

Results

EDA - Descriptive Statistics Results

Variable	Mean	Median	Std Dev	Min	Max	Skewness
Production (MT)	8,084.88	280.00	42,154.35	0	696,412	8.12
Area (ha)	3,545.32	175.00	18,625.47	0	663,734	15.82
AP Ratio	2.85	1.15	5.91	0.01	27.69	2.47
Avg Temp (°C)	23.88	25.00	4.72	11.5	34.5	-0.23
Avg Humidity (%)	71.36	70.00	5.68	60	85	0.45

EDA - Descriptive Statistics Results

- **Highly skewed output:** Mean (8,085 MT) far exceeds median (280 MT), showing extreme right skew driven by a few dominant crops.
- **High variability:** Std. deviation $>$ mean for production and area, indicating strong heterogeneity across crops and regions.
- **Wide temperature range:** 11.5°C – 34.5°C reflects clear seasonal variation in a tropical monsoon climate.
- **Moderate humidity spread:** Mostly 60–85%, stable but still impactful for crop performance.

EDA - Correlation Analysis

Variable	Correlation with Production
Area	0.9234***
AP Ratio	0.1567***
Avg Temperature	-0.0842*
Avg Humidity	0.0523
Max Temperature	-0.0756*
Min Temperature	-0.0641
Max Relative Humidity	0.0489
Min Relative Humidity	0.0512

*** $p < 0.001$, * $p < 0.05$

- **Area dominates output:** Strong correlation ($r = 0.92$) confirms production is mainly driven by cultivated area and supports data consistency.

- **Efficiency matters somewhat:** AP Ratio has a modest positive link ($r = 0.16$), showing yield per area adds to production beyond scale.

- **Climate link is weak:** Temperature and humidity show low correlation, likely due to aggregation effects, non-linear dynamics, and crop diversity masking relationships.

V1 vs V2

Approach	Best Model	Primary Metric	Val Score	Test Score
V1 (Global Classification)	Logistic Regression	Accuracy	66.45%	65.18%
		F1-Macro	0.6655	0.6507
V2 (Crop-Aware Classification)	Gradient Boosting	Accuracy	38.16%	N/A
		F1-Macro	0.3850	N/A

V1 vs V2

Approach	Best Model	Primary Metric	Val Score	Test Score
V1 (Global Classification)	Logistic Regression	Accuracy	66.45%	65.18%
		F1-Macro	0.6655	0.6507
V2 (Crop-Aware Classification)	Gradient Boosting	Accuracy	38.16%	N/A
		F1-Macro	0.3850	N/A

Verdict

Results became less than expected

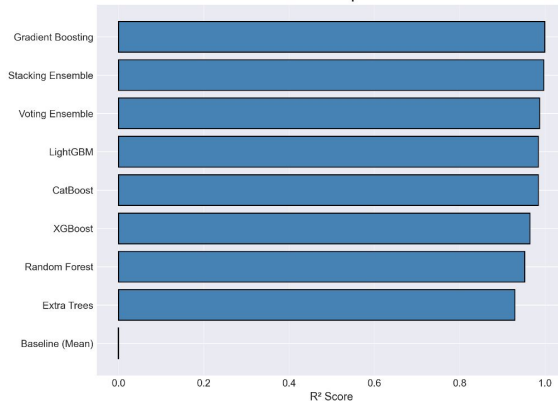
V1 vs V2 vs Final Regression

Approach	Best Model	Primary Metric	Val Score	Test Score
V1 (Global Classification)	Logistic Regression	Accuracy	66.45%	65.18%
		F1-Macro	0.6655	0.6507
V2 (Crop-Aware Classification)	Gradient Boosting	Accuracy	38.16%	N/A
		F1-Macro	0.3850	N/A
Final (Regression)	Stacking Ensemble	R ²	0.9897	0.9974
		RMSE (tons)	7,248	3,793
		MAE (tons)	1,456	472
		MAPE (%)	135.8	109.3

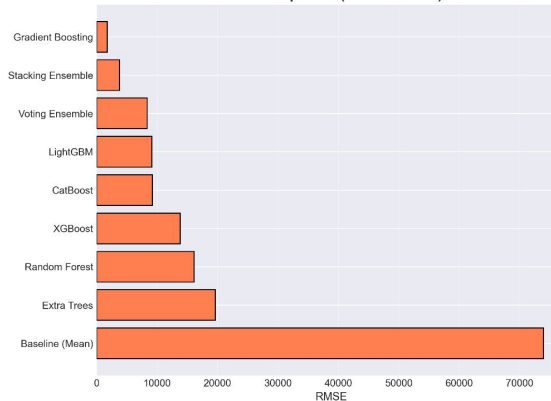
Verdict

Framing the problem from classification to regression made things better

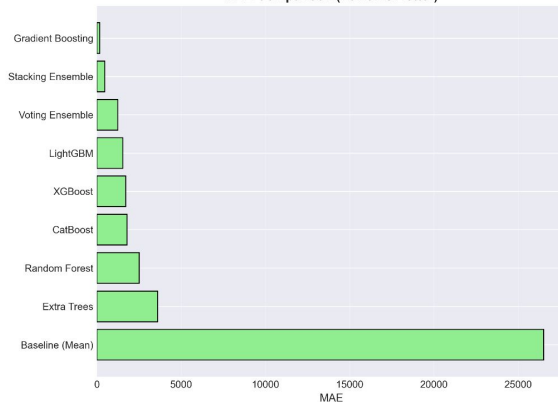
R² Score Comparison



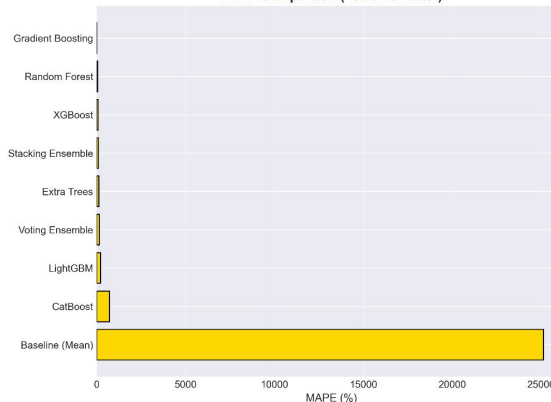
RMSE Comparison (Lower is Better)



MAE Comparison (Lower is Better)



MAPE Comparison (Lower is Better)



Comparison between Regression Approaches

Model	Accuracy	F1-Macro	F1-Weighted	Balanced Acc
Logistic Regression	0.6645	0.6655	0.6656	0.6644
Gradient Boosting	0.6423	0.6458	0.6459	0.6421
K-Nearest Neighbors	0.6312	0.6330	0.6332	0.6310
Random Forest	0.6280	0.6279	0.6280	0.6278
Extra Trees	0.6264	0.6264	0.6266	0.6262
Decision Tree	0.6248	0.6240	0.6241	0.6246
Support Vector Machine	0.6121	0.6115	0.6117	0.6118
Naive Bayes	0.5278	0.4702	0.4705	0.5272

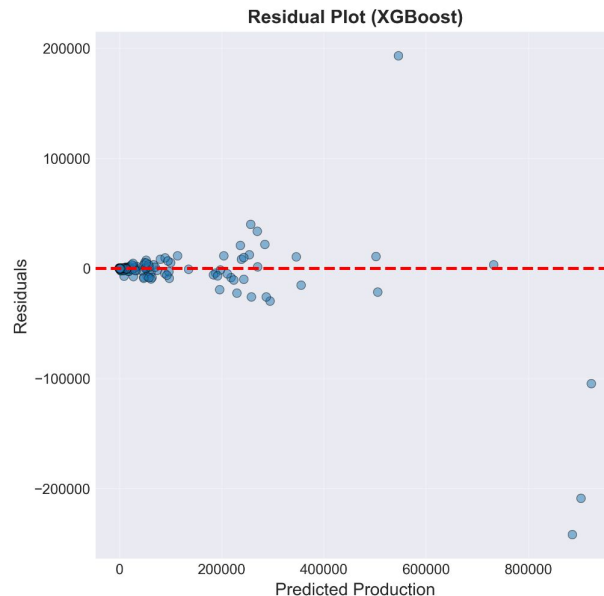
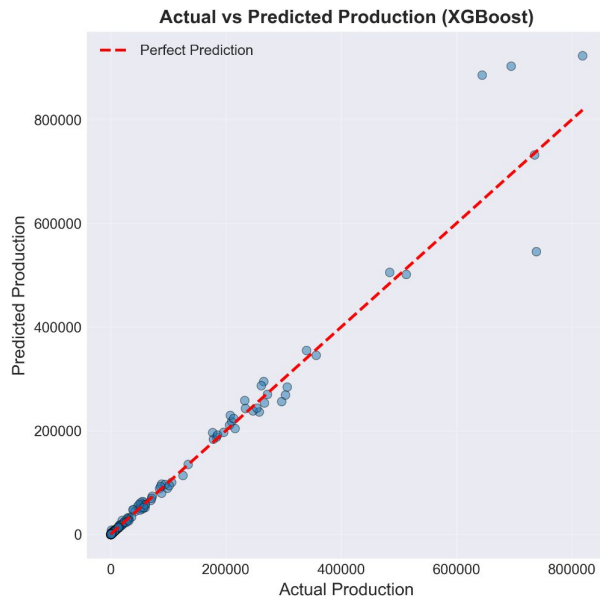
Comparison
between
models of V1

Model	Accuracy	F1-Macro	F1-Weighted	Balanced Acc
Naive Bayes	0.3863	0.3261	0.3262	0.3865
Gradient Boosting	0.3816	0.3850	0.3849	0.3816
Extra Trees	0.3768	0.3695	0.3695	0.3767
Soft Voting (Top 3)	0.3752	0.3713	0.3714	0.3751
Logistic Regression	0.3625	0.3560	0.3559	0.3625
Random Forest	0.3498	0.3461	0.3461	0.3498
Support Vector Machine	0.3498	0.3477	0.3475	0.3499
Decision Tree	0.3132	0.3124	0.3122	0.3133
K-Nearest Neighbors	0.2703	0.2682	0.2681	0.2704

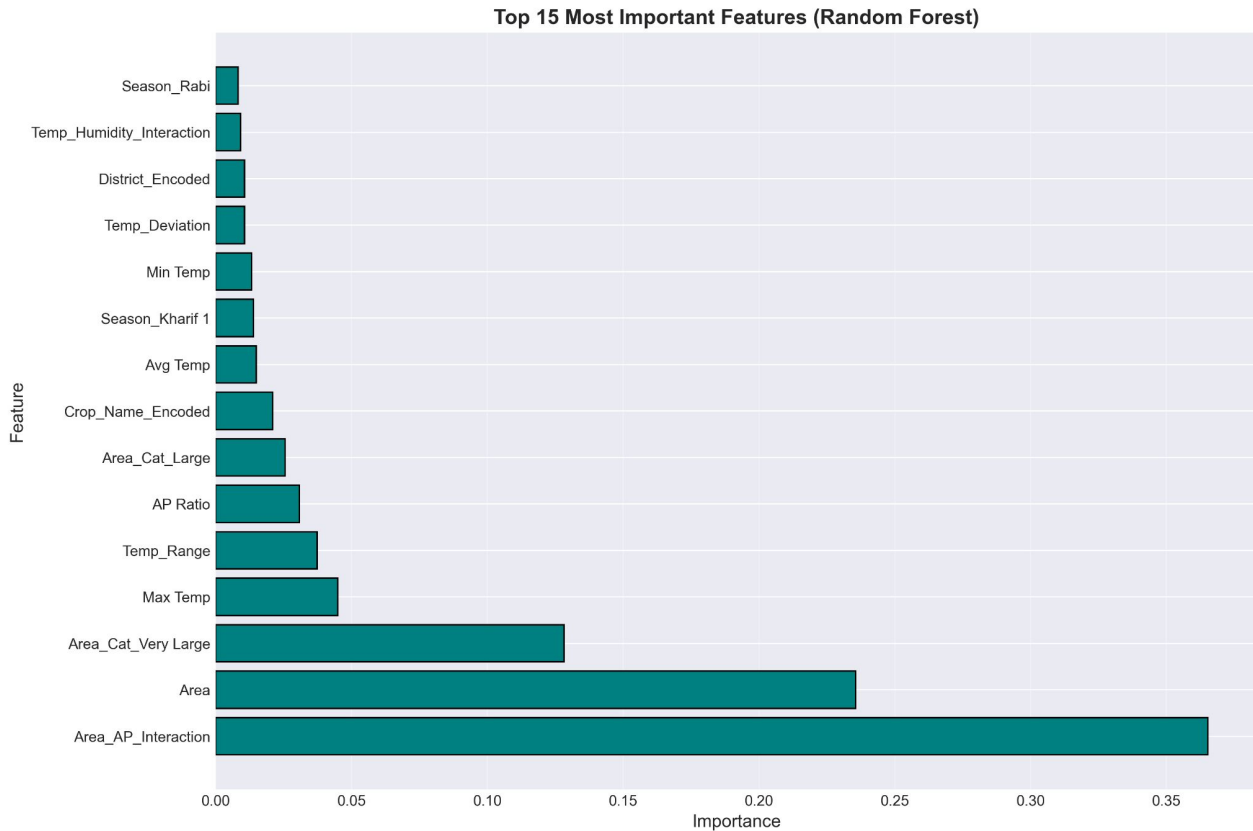
Comparison
between
models of V2

Model	R ² Score	RMSE (tons)	MAE (tons)	MAPE (%)
Stacking Ensemble	0.9974	3,793	472	109.3
Voting Ensemble	0.9872	8,363	1,241	151.8
XGBoost	0.9854	8,921	1,387	159.2
LightGBM	0.9841	9,324	1,452	165.4
CatBoost	0.9838	9,421	1,479	167.1
Random Forest	0.9812	10,145	1,623	178.9

Regression
performance
on test set



**Regression
performance
on test set**



Importance of
features

06

Conclusion

Key Findings & Contributions

Feature Importance

- Area: **Most dominant (~13.7%)**
- Climate factors: ~11%
(temperature + humidity)
- District-level effects also significant

Insight:

→ Production mainly driven by scale + environment

Seasonal Variation

- Kharif-2: **~68.9% accuracy**
- Rabi: **~61.8% accuracy**

Observation:

→ Seasonal differences affect prediction reliability
→ Monsoon season more predictable

Key Findings & Contributions

Finally,

Regression performed much better than classification approaches based on the dataset. It was hard to find a good dataset specially in the domain where country is Bangladesh and multi crop was present. Even after finding such, expectation was that there will be more columns like fertilizer, pesticides used, money spent, seed quality etc. However due to the limitation of available dataset features; it was only possible to do regression and ensemble methods. Deep learning methods would not fit this scenario infact, it would be overkill. Using crop identity for yield also proved subpar.

References

- J. R. Jui, “Agriculture, structural transformation, and economic growth in bangladesh: Trends, challenges, and policy implications,” *Journal of Scientific Reports*, vol. 10, no. 1, pp. 196–209, 2025.
- [2] S. Dasgupta, M. M. Hossain, M. Huq, and D. Wheeler, “Climate change and soil salinity: The case of coastal bangladesh,” *Ambio*, vol. 44, no. 8, pp. 815–826, 2015.
- [3] J. Chen and V. Mueller, “Coastal climate change, soil salinity and human migration in bangladesh,” *Nature climate change*, vol. 8, no. 11, pp. 981–985, 2018.
- [4] J. Molla and A. Rayhan, “Indigenous knowledge and challenges of sustainable paddy cultivation: A cross-cultural study on uncovering traditional knowledge, modern challenges and climate resilience in mymensingh.”
- [5] S. Pokhariyal, N. Patel, and A. Govind, “Machine learning-driven remote sensing applications for agriculture in india—a systematic review,” *Agronomy*, vol. 13, no. 9, p. 2302, 2023.
- [6] S. O. Araujo, R. S. Peres, J. C. Ramalho, F. Lidon, and J. Barata, “Machine learning applications in agriculture: current trends, challenges, and future perspectives,” *Agronomy*, vol. 13, no. 12, p. 2976, 2023.

See the rest on our paper

Thanks!

Do you have any questions?

Onjon Sutradhar