**Model Description**

This unsupervised learning model segments geographic zones into **agro-clusters** based on rainfall, temperature, and pesticide usage. These clusters serve as an environmental fingerprint, guiding further prediction and analysis of crop yield in a climate-resilient way.

It can be used to:

* Discover spatial patterns and environmental typologies
* Tailor local farming strategies and input usage
* Feed agro-zone identifiers into supervised prediction models for enhanced accuracy

**Input:**

* Cleaned environmental data (rainfall, temperature, pesticide levels)
* Scaled and preprocessed to reduce bias

**Output:**

* Cluster assignments per region
* Centroids representing typical environmental conditions

**Model Architecture:**

* K-Means clustering with Euclidean distance
* Optimal cluster count determined via **Elbow Method**
* Supplemented by **Hierarchical Clustering** for validation

**Performance**

* **Silhouette Score:** 0.63
* Clusters meaningfully separated and interpreted
* Improved downstream model performance by adding agro\_zone as a feature

**Limitations**

* Static clustering doesn’t account for seasonality
* Sensitive to outliers and scale, requiring careful preprocessing
* Sensitive to outliers and scale, requiring careful preprocessing

**⚙️ Model Card**

*Model Name:* agri\_yield\_model.ridge *Version:* 1.0 *Date:* 06th Aug 2025 *Author:* [Madhusmita Mishra]

**Model Description**

This Ridge Regression model predicts **crop yield per hectare**, balancing simplicity with regularization to control multicollinearity and reduce overfitting.

Suited for:

* Baseline modeling and benchmarking
* Scenarios where interpretability is key

**Input:**

* Environmental features, agro-zones, and interactions
* Preprocessed for scale and encoding

**Output:**

* Continuous predicted yield (tons/hectare)

**Model Architecture:**

* Ridge Regression with α tuned via cross-validation
* Penalizes large coefficients, supports multicollinear data

**Performance**

| **Metric** | **Value** |
| --- | --- |
| R² | 0.72 |
| MAE | 0.19 |
| RMSE | 0.27 |

**Limitations**

* May underperform if nonlinear relationships dominate
* Less flexible compared to tree-based models

**🌲 Model Card**

*Model Name:* agri\_yield\_model.Random Forest *Version:* 1.0 *Date:* 06th Aug 2025 *Author:* [Madhusmita Mishra]

**Model Description**

The Random Forest model uses an ensemble of decision trees to predict crop yields. It handles high-dimensional, nonlinear features effectively and is robust to noise.

Ideal for:

* Situations with complex feature interactions
* Data with outliers or skew

**Input:**

* Environmental, chemical, and zone-level features
* Bootstrapped training samples

**Output:**

* Yield prediction (continuous)
* Feature importance scores

**Model Architecture:**

* 300 trees, max depth=15, min\_samples\_leaf=5
* No pruning, bagging applied

**Performance**

| **Metric** | **Value** |
| --- | --- |
| R² | 0.81 |
| MAE | 0.16 |
| RMSE | 0.21 |

**Limitations**

* Slower inference time
* Less interpretable than linear models
* Can overfit with noisy features

**🚀 Model Card**

*Model Name:* agri\_yield\_model.gbm *Version:* 1.0 *Date:* 06th Aug 2025 *Author:* [MAdhu]

**Model Description**

This Gradient Boosting model is optimized using **Bayesian Search** to maximize predictive power while remaining generalizable. It captures nonlinear patterns across climate and chemical variables effectively.

Recommended for:

* High-performance needs
* Real-world prediction with uncertainty estimation

**Input:**

* Same preprocessed feature set including agro-zones
* With added interaction and polynomial features

**Output:**

* Predicted yield
* SHAP values to explain individual predictions

**Model Architecture:**

* Learning rate: 0.03
* Estimators: 500
* Max depth: 10
* Subsample: 0.75
* Optimized via 5-fold cross-validated BayesSearchCV

**Performance**

| **Metric** | **Value** |
| --- | --- |
| R² | 0.87 |
| MAE | 0.13 |
| RMSE | 0.18 |

**Limitations**

* Requires tuning for new geographies
* Computationally more expensive to retrain
* Interpretability depends on SHAP or feature analysis