**Problem Statement**

Crop yield is influenced by a variety of environmental and chemical factors — temperature swings, rainfall inconsistencies, and pesticide usage — all of which vary wildly in the real world. Predicting yield reliably across these changing landscapes is a challenge.

That's where this project comes in: to develop a **climate-smart crop yield prediction model** that adapts to these dynamics while staying accurate, understandable, and sustainable. I have used following dataset for this.

<https://www.kaggle.com/code/patelris/crop-yield-eda-viz/notebook>

**Acknowledgements** This project uses publicly available datasets sourced from:

* Food and Agriculture Organization (FAO) <http://www.fao.org/home/en/>
* World Bank Open Data <https://data.worldbank.org/>

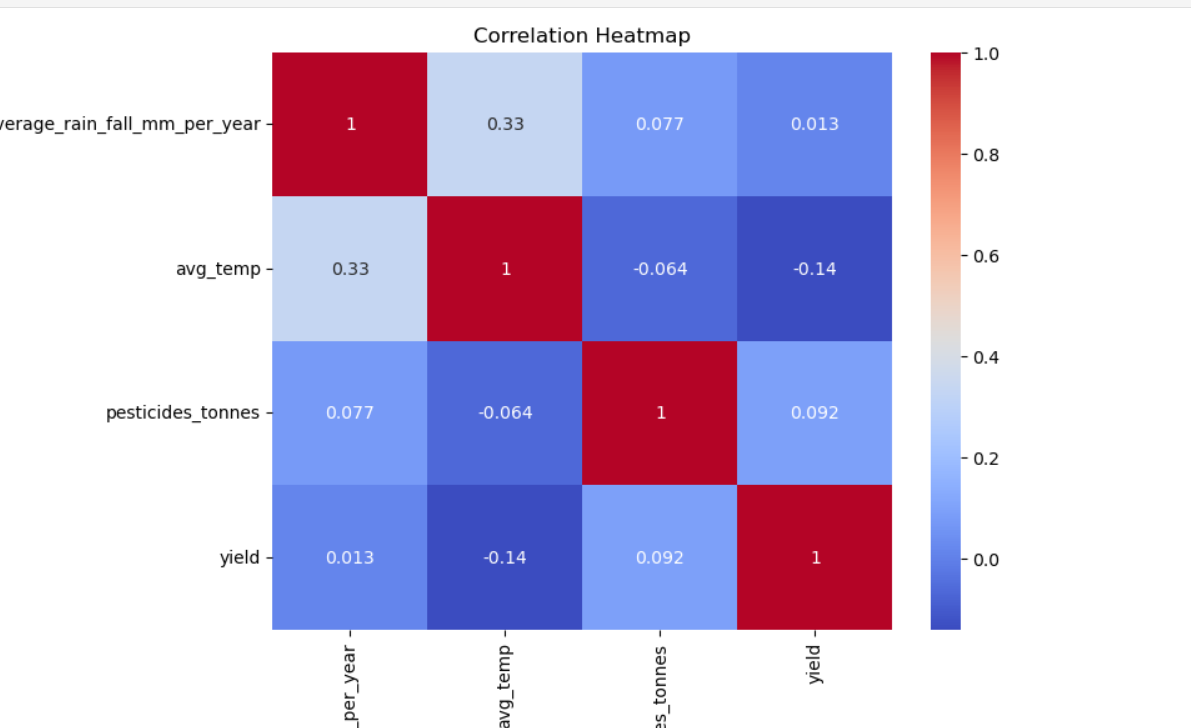
The dataset used for crop yield prediction was originally compiled and shared by patelris on Kaggle, and includes data derived from FAO and World Bank sources.  
[Crop Yield Prediction Dataset](https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset)

**1. Data Cleaning & Integration**

* Merge yield.csv, rainfall.csv, temp.csv, pesticides.csv
* Standardize column names
* Handle missing values with bootstrapping and imputation
* Detect and treat outliers using IQR and z-score

**2. Exploratory Data Analysis**

* Mean, median, mode, std dev, quartiles
* Visuals: boxplots, histograms, scatter plots, heatmaps
* Apply **Central Limit Theorem** via sampling distributions
* Study **normal distribution shifts** by manipulating variables



**3. Unsupervised Learning**

* Apply **K-Means** and **Hierarchical Clustering** to group regions/crops
* Add cluster labels as new features
* Use **PCA** for dimensionality reduction and visualization

**4. Feature Engineering**

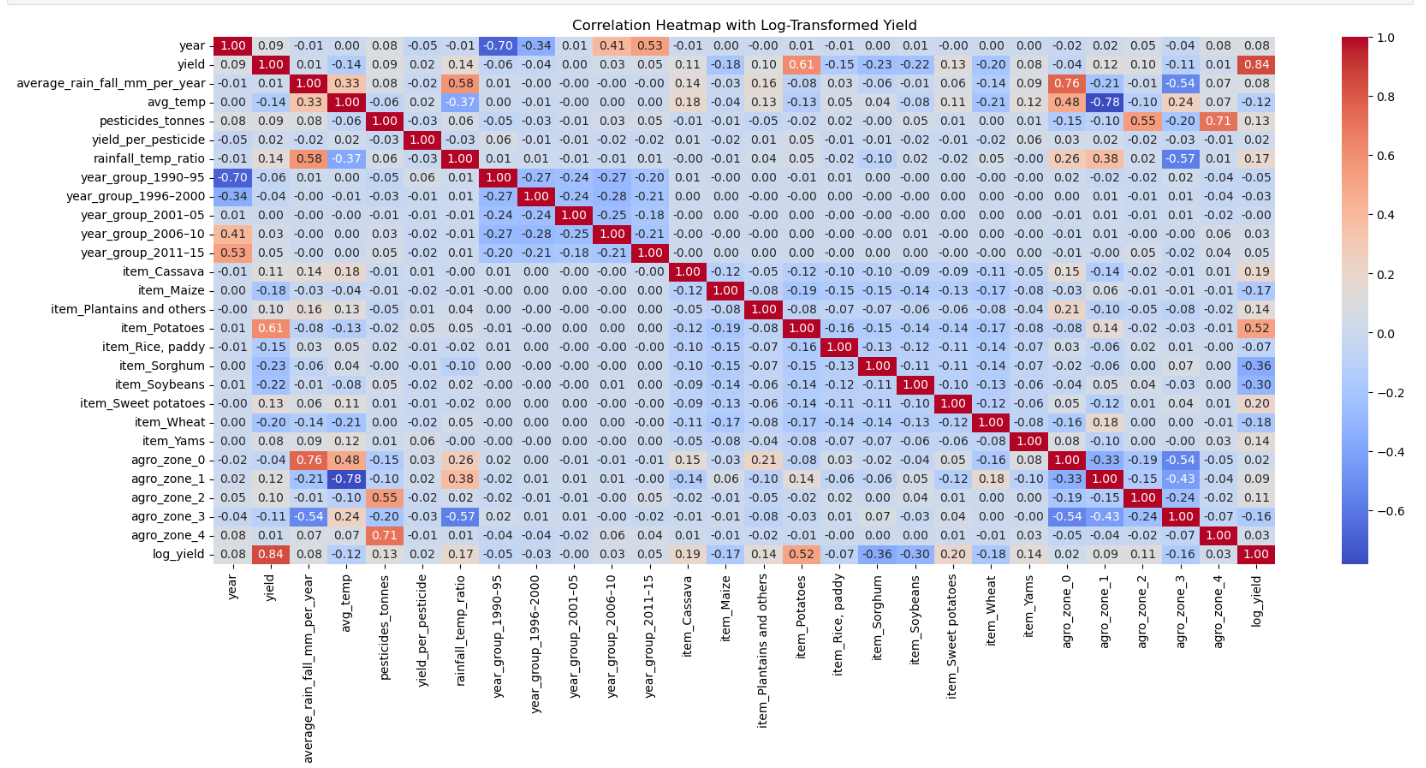
* One-hot and label encoding
* Interaction terms (e.g. rainfall × pesticide)
* Add bootstrapped confidence intervals
* Correlation coefficients and multicollinearity checks

**5. Modeling & Evaluation**

* Models: Ridge Regression, Decision Tree, Gradient Boost Regression
* Use **Bayesian Optimization** for hyperparameter tuning
* Apply **Maximum Likelihood Estimation** for parameter fitting
* Evaluate with R², RMSE, MAE, confusion matrix, lift chart
* Use **K-Fold Cross Validation** and **Stratified K-Fold** for robustness

**6. Bias–Variance Tradeoff & Generalisation**

* Compare models on training vs validation
* Plot learning curves
* Discuss overfitting vs underfitting



**Strategic Implications of Correlation**

* **Zone 0** is both warm and wet — ideal for crops like rice, yams, or cassava. Low pesticide use might indicate sustainable practices.
* **Zone 4** is warm and pesticide-intensive — possibly high-value crops or pest-prone areas. Could benefit from integrated pest management.
* **Zone 3** has moderate rainfall correlation — may be transitional or mixed farming zone.
* **Zone 1 and 2** are temperature-driven — crop choices here may depend more on heat tolerance than water availability.

**Final Descision regarding Model choice**

**Model Selection Breakdown**

| **Criteria** | **Ridge Regression** | **Random Forest** | **Gradient Boosting** |
| --- | --- | --- | --- |
| Accuracy | 📉 Low | ✅ Highest | 👍 Good |
| Interpretability | ✅ High (linear model) | ⚠️ Moderate (trees ensemble) | ⚠️ Moderate |
| Handling Outliers | ❌ Poor | ✅ Strong | ✅ Strong |
| Nonlinear Relationships | ❌ Poor | ✅ Excellent | ✅ Excellent |
| Uncertainty Estimation | ⚠️ Basic | ✅ Can use bootstrapping | ✅ Can use SHAP & intervals |
| Generalizability | ✅ High | ✅ High | ✅ Medium-High |
| Sustainability Impact | ❌ Static assumptions | ✅ Feature insights inform action | ✅ Good but sensitive to tuning |

**Interpretation of CV R² Scores with the help of Cross Validation using KFold, StratifiedKFold, cross\_val\_score**

| **Model** | **Mean CV R²** | **Std. Dev ±** | **What It Means** |
| --- | --- | --- | --- |
| **Ridge Regression** | 0.659 | ± 0.011 | Moderate fit. Linear assumptions limit its ability to model complex relationships. |
| **Random Forest** | 0.983 | ± 0.001 | Excellent fit. Extremely high performance and very stable across folds. |
| **Gradient Boosting** | 0.934 | ± 0.002 | Strong fit. Also quite stable, though slightly less than RF. Can improve with tuning. |

To build a **robust, interpretable, and sustainable system**   
  
This model pipeline is best because  
  
**This Is Powerful**

* I am **customizing predictions** to agro-ecological realities
* I am **adding structure** to a flat dataset of data fetched from different files
* I am **making it actionable** for policy, farming decisions, and sustainability

**Limitations & Future Work**

**While promising, the model works best with static regional data and short-term forecasts. Future work may include:**

* **Satellite image integration**
* **Long-term climate trend adaptation**
* **Soil quality data enhancement**