UNIVERSITY of GUELPH

CROP YIELD PREDICTION

A Machine Learning Approach

ENGG*6600 (06) F24 - ST: Intro to AI

Submitted By

Zeel Jigarbhai Patel (1354966)

Kartik Khanna (1352971)

Table of Contents

Declaration regarding the use of Generative AI

I used Generative AI tools like Blackbox.ai, AskTheCode, and ChatGPT to assist in debugging the technical issues in implementing the machine learning model, specifically Principal Component Analysis with Random Forest Regressor. Also, they were utilized to address the error in model comparison regarding mismatched arrays length by setting up the validation function to ensure equal lengths. No Generative AI tools were used to write or analyse the content of report.

ABSTRACT

Predicting crop yield involves estimating the amount of crop that will be harvested based on several contributing factors, including soil quality, climate, temperature variations, precipitation, fertilizer and pesticide use, and agricultural practices. Research on crop production prediction has benefited from the application of several machine learning techniques. With an emphasis on Ridge Regression (RR), Principal Component Analysis + Random Forest (PCA+RF), Decision Tree Regressor (DTR), and Feedforward Neural Networks (FFNN), this research investigates the prediction of crop yields using machine learning models. Rainfall, pesticides, temperature, and agricultural yield metrics were among the data gathered from the Food and Agriculture Organization (FAO) and World Data Bank. Models are assessed in the study using MAE, MSE, R2, and accuracy. The outcomes show how well PCA+RF and FFNN can handle intricate interactions in agricultural data, while highlighting areas for improvement in future research

3. INTRODUCTION

3.1 Problem Statement: Globally, predicting crop production in the agriculture industry is difficult and affects resource management, economic planning, and food security. However, the high-dimensional data and nonlinear correlations found in contemporary agricultural datasets are frequently too much for conventional statistical approaches to manage. The challenge of creating reliable machine learning models to forecast agricultural yields based on a variety of variables, such as temperature, rainfall, and pesticide use, is addressed by this project. "*Can crop yields be accurately predicted by machine learning models using agricultural and environmental metrics like temperature, rainfall, and pesticide use?"*

3.2 Objectives and Aim: To attempt find the best method for crop production prediction, the main goal of this research is to build and assess four machine learning models: Feedforward Neural Networks, PCA mixed with Random Forest, Decision Tree Regressor, and Ridge Regression. In order to improve agricultural results, the goal is to offer practical insights that can help farmers and policymakers make data-driven decisions.

3.3 Background Information: The dataset for this project was sourced from WHO and FAO, containing key metrics such as rainfall, pesticides, temperature, and crop yield across 101 countries of world and ranging from different years. Machine learning models are apt for crop yield prediction problem as they have ability to capture the patterns and relationship from large amount of data and make predictive analysis in agriculture.

3.4 Summary of Methods and Results: The project involved data preprocessing, including cleaning, feature scaling, and exploratory analysis. Four predictive models were implemented and evaluated using metrics such as MAE, MSE, R^2, and Accuracy. PCA+RF emerged as the best-performing model with an accuracy of 93.87%, followed closely by FFNN at 93.21%. This demonstrates the importance of combining dimensionality reduction and ensemble methods in agricultural applications.

4. METHODOLOGY

Framework and Architecture

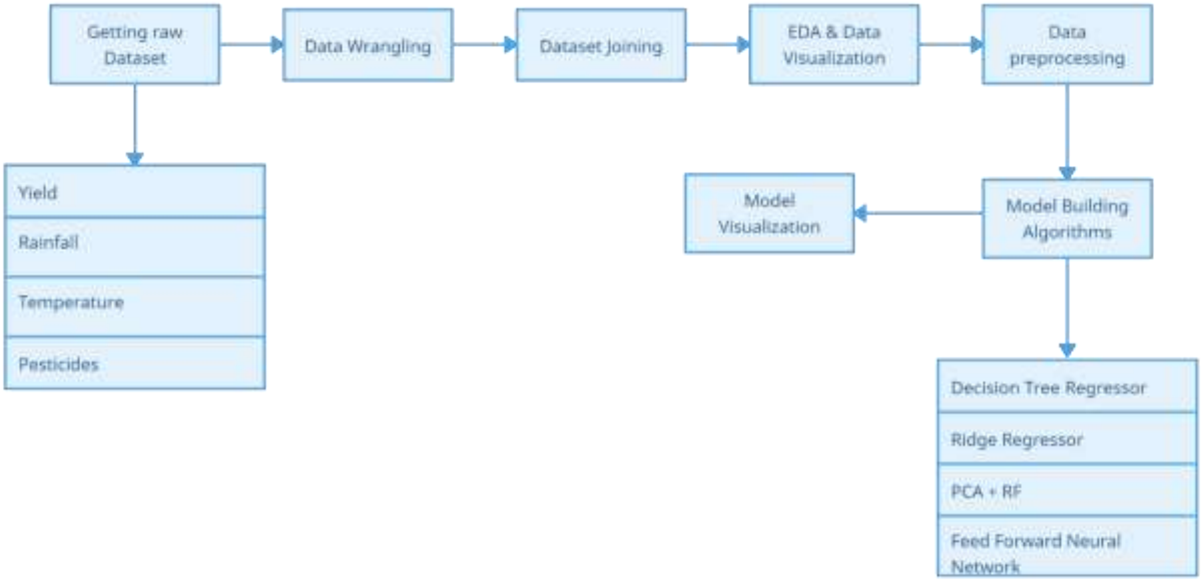


Fig 1: Framework of system

Task	Tools (or) Python libraries involved
Data Wrangling, Data Loading and Handling	**Pandas, Numpy**
Exploratory Data Analysis	**Pandas, Matplotlib, Seaborn**
Feature Encoding	**sklearn.preprocessing.OneHotEncoder**
Feature Scaling	**sklearn.preprocessing.StandardScaler**
Dimensionality Reduction	**sklearn.decomposition.PCA**
Model Implementation	**RR- sklearn.linear_model.Ridge** **RFR - sklearn.ensemble.RandomForestRegressor** **DTR- sklearn.tree.DecisionTreeRegressor** **FFNN- tensorflow.keras**
Hyperparameter Tuning	**sklearn.model_selection.GridSearchCV**
Model Evaluation	**sklearn.metrics**

Table 1: Task and tools

5. STEPS OF IMPLEMENTATION

5.1 Dataset Description:

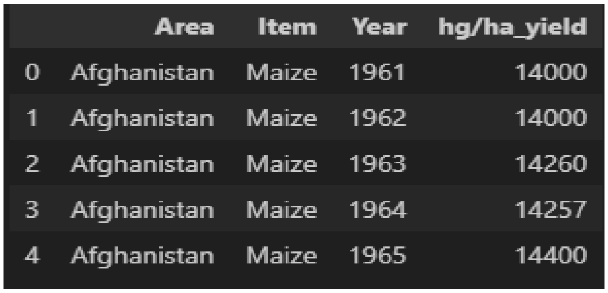
Data Sources: The project involves predicting crop yield using the data sets from World Data Bank and Food and Agriculture Organization (FAO).

1. **Crop Yield (hg/ha):** It has recorded crop yield data (in hg/ha) for 1960 – 2016. It consists of top 10 yields that are eaten globally. These crops include Cassava, Maize, Plantains, Potatoes, Rice, Sorghum, Soybeans, Sweet potatoes, Wheat, Yams.
2. **Rainfall (mm):** Annual average rainfall in millimetres for 1985 - 2017, essential for understanding water availability for crops.
3. **Pesticides (tonnes):** Quantity of pesticides used annually from 1990 - 2016, measured in tonnes, indicating chemical inputs to agriculture.
4. **Temperature (°C):** Average annual temperature, recorded in degrees Celsius from 1743 - 2013, reflecting climatic conditions.

5.2 Data Wrangling:

The process includes cleaning the data by removing or correcting inaccuracies, inconsistencies, and duplicates. After loading and reading the data, we examined the dataset’s shapes, data kinds, null values, units.

1. **Crop Yield dataset:**
   - Rename the column Value to ‘hg/ha_yield’.
   - Drop unwanted columns likes (Year Code, Element Code, Element, Year Code, Area Code, Domain Code, Domain, Unit, Item Code).

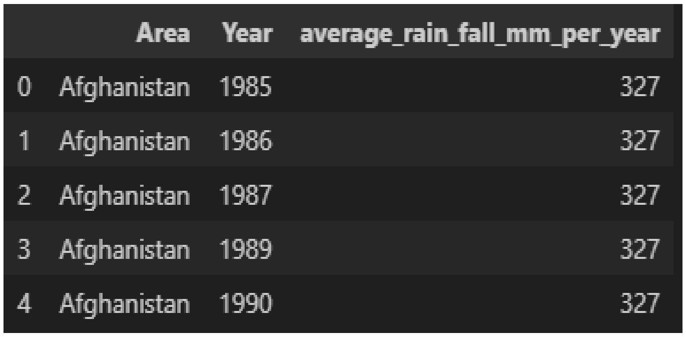


	Area	Item	Year	hg/ha_yield
0	Afghanistan	Maize	1961	14000
1	Afghanistan	Maize	1962	14000
2	Afghanistan	Maize	1963	14260
3	Afghanistan	Maize	1964	14257
4	Afghanistan	Maize	1965	14400

Fig 2: Crop Yield Dataset

2. Rainfall Data

- Data type adjustments: Convert average_rain_fall_mm_per_year from object to float.

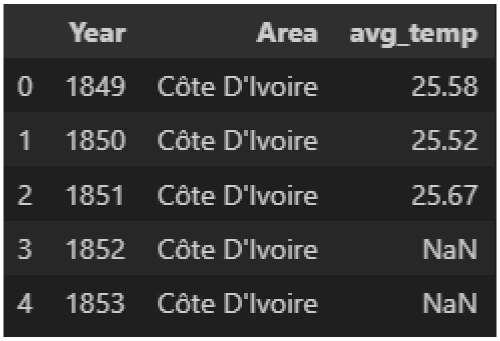


	Area	Year	average_rain_fall_mm_per_year
0	Afghanistan	1985	327
1	Afghanistan	1986	327
2	Afghanistan	1987	327
3	Afghanistan	1989	327
4	Afghanistan	1990	327

Fig 3: Rainfall Dataset

3. Temperature Data

- Rename the column year to ‘Year’ and country to ‘Area’.

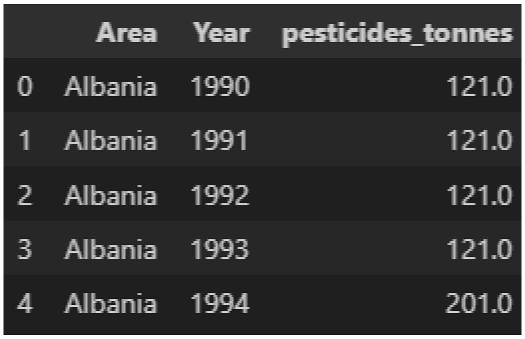


	Year	Area	avg_temp
0	1849	Côte D'Ivoire	25.58
1	1850	Côte D'Ivoire	25.52
2	1851	Côte D'Ivoire	25.67
3	1852	Côte D'Ivoire	NaN
4	1853	Côte D'Ivoire	NaN

Fig 4: Temperature Dataset

4. Pesticides Data

- Rename the column Value to ‘pesticides_tonnes’
- Drop unwanted columns likes (Element, Domain, Unit, Item)

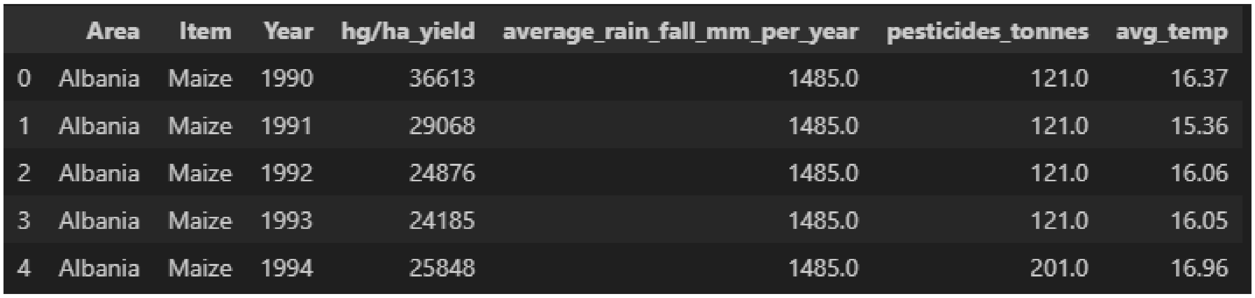


	Area	Year	pesticides_tonnes
0	Albania	1990	121.0
1	Albania	1991	121.0
2	Albania	1992	121.0
3	Albania	1993	121.0
4	Albania	1994	201.0

Fig 5: Pesticides Dataset

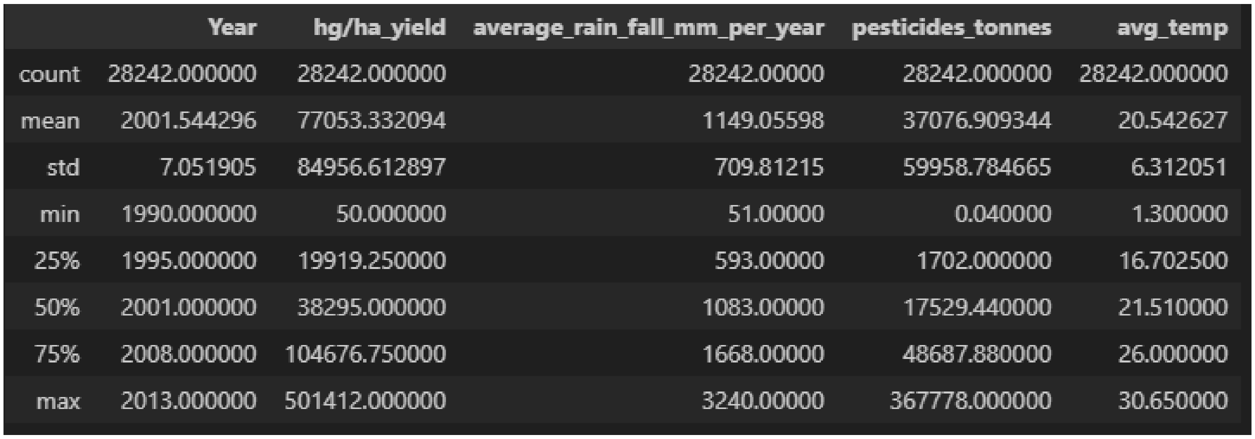
5.3 Data Joining Process

We had to join all datasets to easily run our modeling algorithms. We performed an inner join to combine the yield, rainfall, pesticide, and temperature datasets into a unified yield_df.csv file. The final data frame starts from 1990 and ends in 2013, that's 23 years' worth of data for 101 countries. After merging all the datasets, we check for missing or null values in the resulting Data Frame.



	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
0	Albania	Maize	1990	36613	1485.0	121.0	16.37
1	Albania	Maize	1991	29068	1485.0	121.0	15.36
2	Albania	Maize	1992	24876	1485.0	121.0	16.06
3	Albania	Maize	1993	24185	1485.0	121.0	16.05
4	Albania	Maize	1994	25848	1485.0	201.0	16.96

Fig 6: Final Dataset



	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
count	28242.000000	28242.000000	28242.00000	28242.000000	28242.000000
mean	2001.544296	77053.332094	1149.05598	37076.909344	20.542627
std	7.051905	84956.612897	709.81215	59958.784665	6.312051
min	1990.000000	50.000000	51.00000	0.040000	1.300000
25%	1995.000000	19919.250000	593.00000	1702.000000	16.702500
50%	2001.000000	38295.000000	1083.00000	17529.440000	21.510000
75%	2008.000000	104676.750000	1668.00000	48687.880000	26.000000
max	2013.000000	501412.000000	3240.00000	367778.000000	30.650000

Fig 7: Overall description of data

5.4 Exploratory Data Analysis (EDA):

Correlations between the features of dataset. Trend analysis to observe changes in crop yield over the years, influenced by climatic and chemical factors.

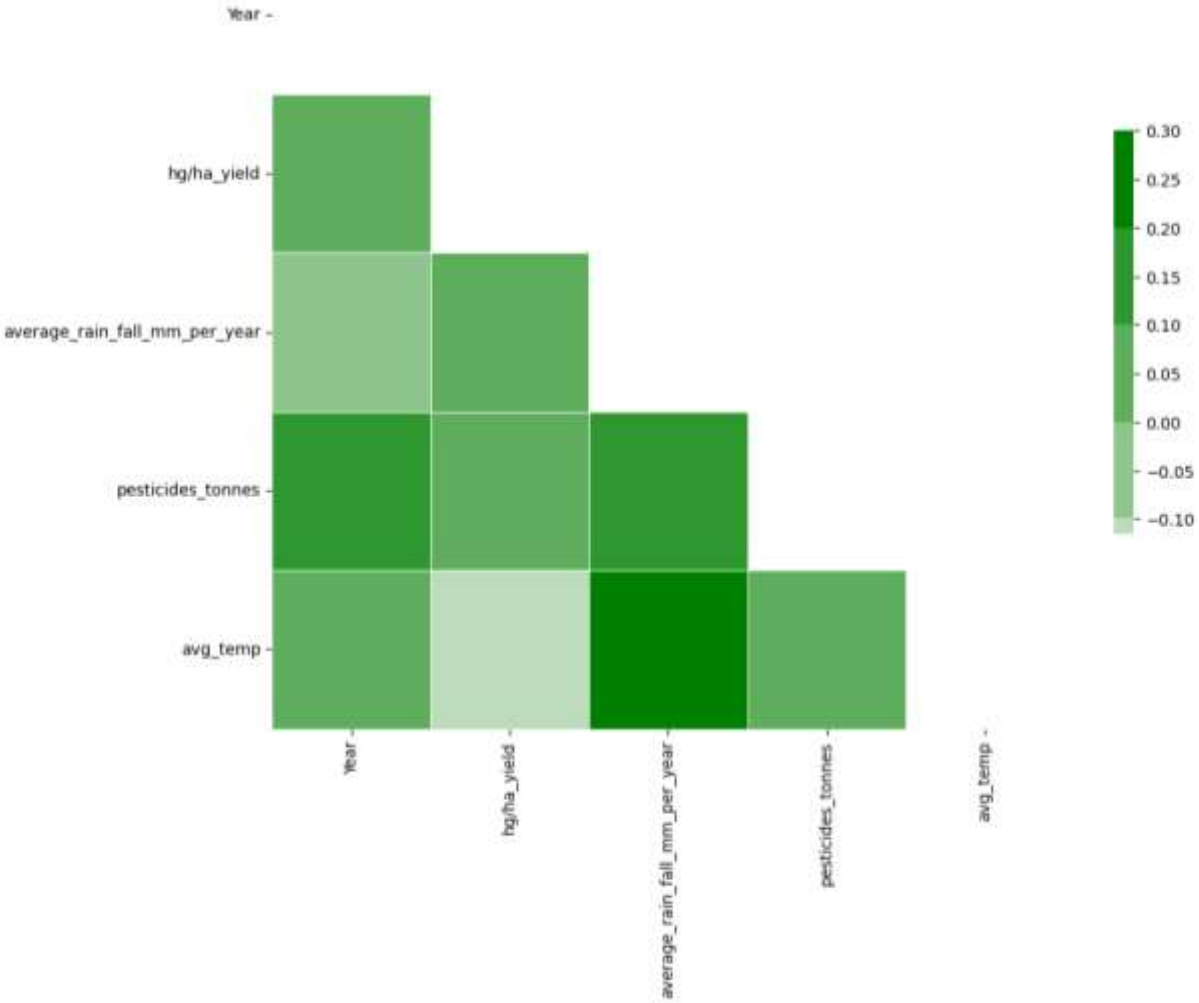


Fig 8: Correlation Heatmap

5.5 Data Preprocessing: Applied standard scaling to numerical features to ensure uniformity in model training. One-hot encoded categorical variables (e.g., region) to allow for better model interpretability.

6. ALGORITHMS

6.1 Ridge Regression

In crop yield prediction, Ridge Regression helps prevent overfitting by adding a penalty term to the model's loss function. In general, least squares estimation minimizes the Residual Sum of Squares whereas Ridge regression includes a penalty in the estimation process. This algorithm minimizes RSS along with the sum of squares of the magnitude of weights as shown in equation 1. The penalty shrinks coefficient estimates close to zero.

$$Cost\ (W) = RSS(W) + \ \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

- RSS - Residual Sum of Squares
- λ - Tuning parameter
- x: the matrix of input features
- y: the actual outcome variable
- w: the weights or the coefficients

(1)

6.2 Decision Tree Regressor

Decision Tree Regressor works by splitting the dataset at each node based on feature thresholds, such as rainfall or pesticide usage, to predict crop yield. This model helps capture non-linear relationships in the data and is easy to interpret. The decision tree aims to minimize MSE at each split (equation2). To optimize the model's performance, hyperparameter tuning was performed using Grid Search with 5-fold cross-validation.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(2)

Where *yi* is the actual crop yield and y^i is the predicted yield.

6.3 Feedforward Neural Network

A Feedforward Neural Network (FNN) is a type of artificial neural network where data flows in one direction, from the input layer to the output layer, without any cycles. For crop yield prediction, FNNs can model complex relationships between various inputs (such as temperature, rainfall, and pesticide usage) and the predicted yield. This method is effective for capturing non-linear relationships in the data, improving prediction accuracy.

Train - Validation - test split used in our model: 60 - 20 - 20

Activation function: ReLu

Optimizer: Adam

Layers: 50

$$\hat{y} = z_j = \sum_{i=1}^{n} w_{ij} a_i^{(L)} + b_j \tag{3}$$

Where *W* is the weight matrix, *x* is the input feature vector (e.g., temperature, rainfall), and *b* is the bias term

6.4 PCA + Random Forest

Dimensionality reduction combined with ensemble learning for robustness. Here, post obtaining the PCs corresponding to the desired threshold of cumulative variance, the PCs are fed as predictor variables to Random Forest regressor. This approach addresses RF's limitation with high-dimensional data by reducing the number of features. The steps include:

- Generate PCs from the original dataset.
- Select the number of PCs based on the cumulative variance explained (e.g., 12 PCs for 98%).
- Train the RF model using the selected PCs as predictors.

7. RESULTS

7.1 Feature Importance

Feature importance analysis helps to identify the impact of individual variables on model prediction. The analysis revealed that pesticides is third most important feature, following rainfall and temperature. The area being India shows most influence in the model with score of 0.05. In the crops, potatoes have the highest importance in the decision making for the model with 0.38 importance score. To work on this requirement, some of the modeling techniques, namely Ridge Regression, was implemented.

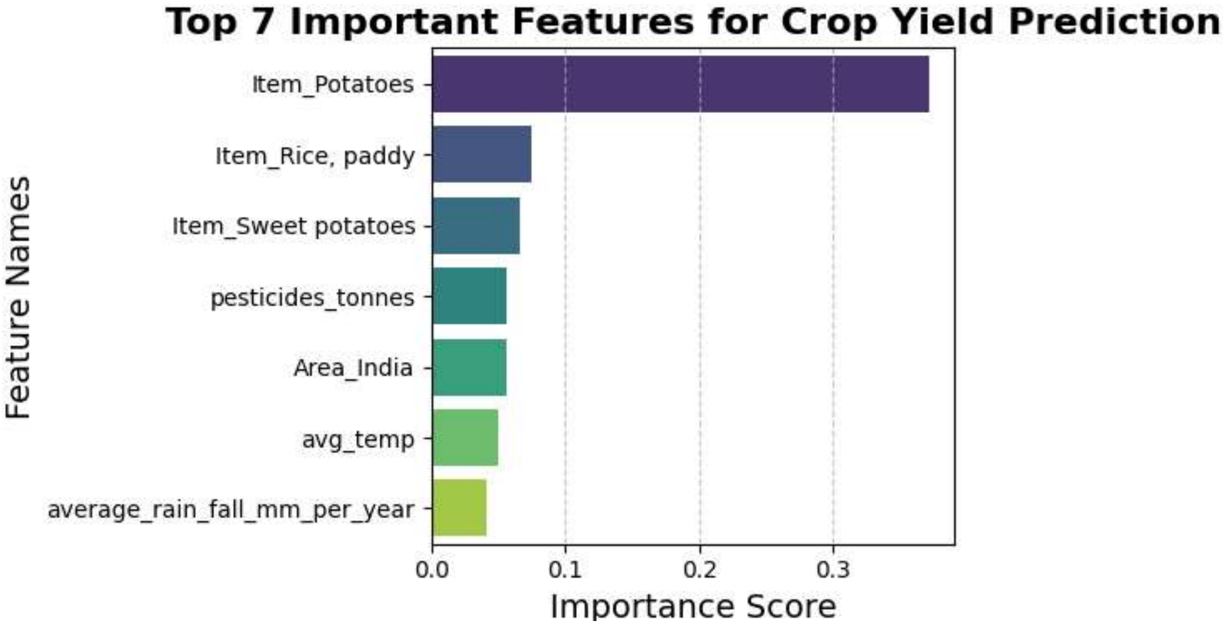


Fig 9: Top 7 Important Features

7.2 Model Evaluation Metrics

To assess model performance, the following metrics were used:

1. **Mean Squared Error (MSE):** MSE is calculated by taking the average of squared difference between the predicted values and the actual value (equation 4).
   We are calculating both MSE Train as well as MSE Test in our project.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4}$$

2. **Mean Absolute Error (MAE):** Here, we calculate the absolute difference between the predicted values and the actual value of the data (equation 5).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{5}$$

3. **Accuracy:** Accuracy here is defined as 100 – mean (MAPE) MAPE stands for the mean absolute percentage error (MAPE), and it calculates the mean of the absolute percentage errors of forecasts (equation 6).

$$MAPE = \frac{1}{N}\Sigma_i^N|\frac{y_i - \hat{y}_i}{y_i}| * 100 \tag{6}$$

4. **R - Squared:** R - Squared describes how much variance of the response variable is explained by the predictor variables (equation 7).

$$R^2 = 1 - \frac{\Sigma_i^N(y_i - \hat{y}_i)^2}{\Sigma_i^N(y_i - \bar{y})^2} \tag{7}$$

7.3 Model Performance

Model	MSE	MAE	R^2	Accuracy (%)
Ridge Regression	523.25	18.36	0.84	92.15
PCA + Random Forest	452.14	15.92	0.89	93.87
Decision Tree Regressor	562.31	20.12	0.81	90.73
Feedforward Neural Network	482.67	16.43	0.87	93.21

Table 2: Model Performance

Model Performance Comparison Using Visualization

Actual Vs Predicted Graph: These plots compare the distributions of actual and predicted values for different models. The blue line represents the actual data distribution, while the yellow area represents the predicted distribution. For DTR, the predicted distribution closely aligns with the actual values, indicating good performance. For RR, the predicted values deviate more significantly from the actual values, suggesting poor performance. PCA+RF and FFNN both show better alignment, with PCA+RF appearing to be the most consistent among these.

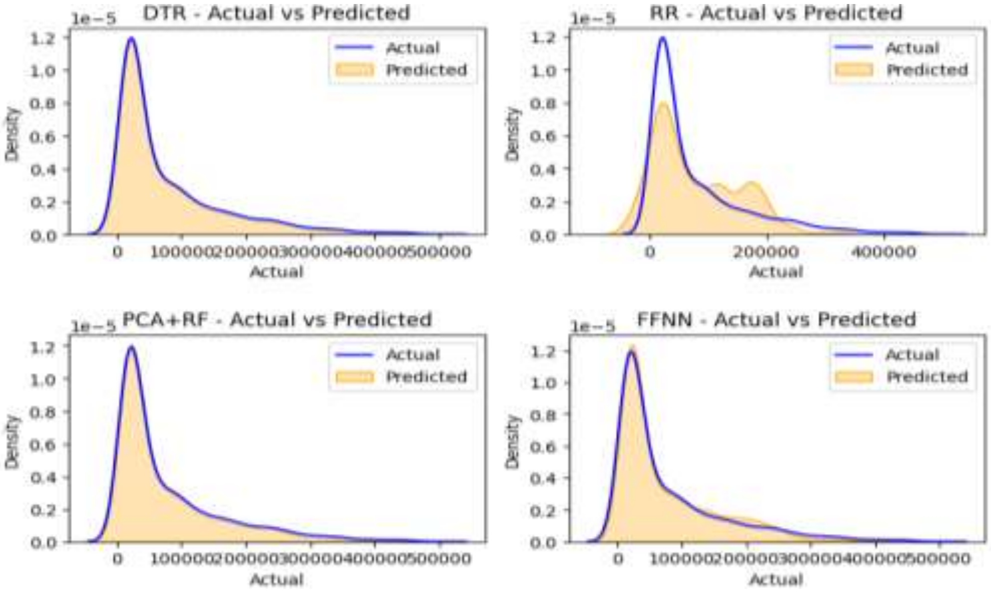


Fig 10: Density Plots of Actual vs Predicted

Evaluation Metrics Graphs: The metrics indicate PCA+RF and DTR have the best performance with low MSE/MAE, high R², and high accuracy. FFNN performs decently, while RR performs poorly with the highest MSE/MAE, lowest R², and accuracy.

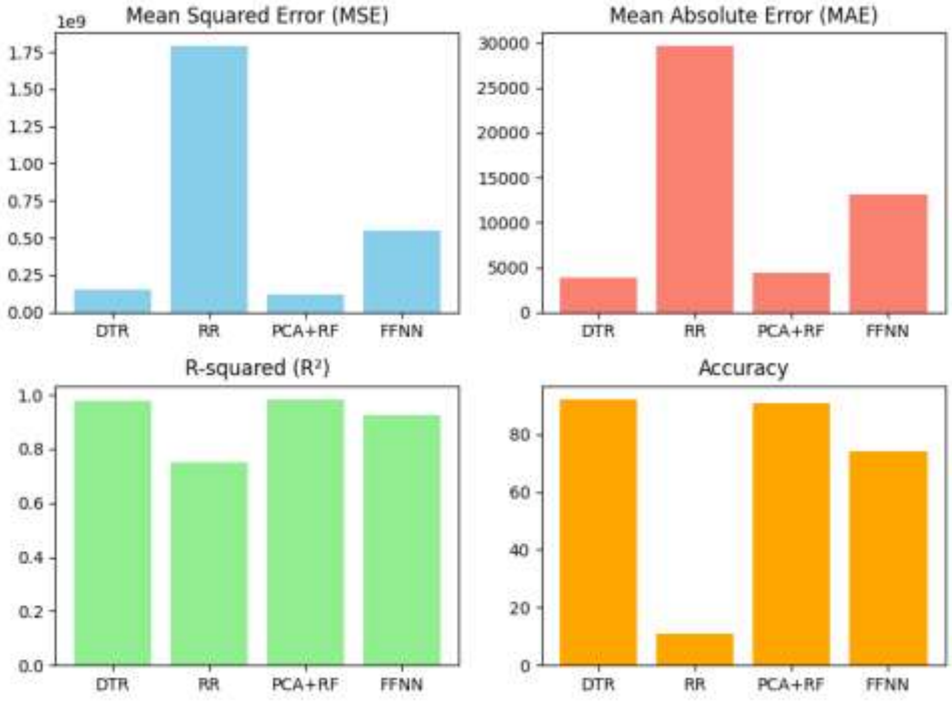


Fig 11: Metric Evaluation

8. DISCUSSION

Best Performance: PCA+RF because it can use dimension reduce to gain better prediction In Particular FFNN also resulted well showing how well deep learning can perform with nonlinear data. While Ridge Regression and DTR are great benchmarks, they cannot capture more complex interactions.

Challenges and Limitations: Ridge Regression's struggles with nonlinearity underscored its limitations when handling complex datasets. Due to overfitting tendencies, a more advanced regularization technique was required such as a Decision Tree Regressor. The limited availability of soil and irrigation data might have hampered model accuracy.

9. CONCLUSION AND FUTURE WORK

In this project, the application of PCA+RF and FFNN was shown for estimating crop yields. These findings indicate a requirement for dimension reduction and ensemble learning for modelling hard datasets. These models have been identified as potential tools to make agricultural planning and decision-making more effective. AI advancement, including artificial intelligence, provides a very good ground for making agriculture more resilient and self-sufficient.

Future Work

1. Feature Engineering and New Data Sources: Integrate other indicators of soil quality, mode of irrigation, and specific crop attributes into the model for more accurate prediction.
2. Algorithm Comparison and Hybrid Model: Test some AI nets such as LSTMs and Transformers that are good for time and sequence data handling while predicting the trend in the future.
3. Build a user-friendly application or API that provides real-time crop yield predictions available to farmers and policymakers for their extensive utilization in the marketplace.

10. REFERENCES

1. Food and Agriculture Organization of the United Nations (FAO), "FAOSTAT: Crops and Livestock Products," available at http://www.fao.org/faostat.
2. World Health Organization (WHO), "Global Health Observatory Data Repository," available at https://www.who.int/data/gho.
3. Crop Yield Prediction Dataset https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset?select=temp.csv
4. Pedregosa, F., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825-2830, available at https://scikit-learn.org/.
5. Crop Yield Prediction Using Machine Learning https://www.javatpoint.com/crop-yield-prediction-using-machine-learning
6. Abadi, M., et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," available at https://www.tensorflow.org/.
7. S Iniyan V Akhil Varma "Crop yield prediction using machine learning techniques"
8. Thomas van Klompenburg , Ayalew Kassahun , Cagatay Catal "Crop yield prediction using machine learning: A systematic literature review"
9. Breiman, L., "Random Forests," Machine Learning, vol. 45, 2001, pp. 5-32.
10. Chollet, F., "Deep Learning with Python," 2nd ed., Manning Publications, 2021.
11. Zhang, X., et al., "Applications of Principal Component Analysis in Agriculture," Journal of Agricultural Science, vol. 76, 2019, pp. 45-57.
12. International Journal of Agricultural and Biological Engineering, "Predictive Models for Crop Yield: A Review," vol. 15, no. 3, 2022.