CROP YIELD PREDICTION USING EXPLAINABLE AI

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

by

SANJAY KARUPOTHULA (2303A52337)

Under the guidance of

Dr. S VAIRACHILAI

Professor, Department of CSE.

Department of Computer Science and Artificial Intelligence

ABSTRACT

Agriculture is one of the most critical sectors for India's economy and food security. Predicting crop yield accurately has a direct impact on planning, resource allocation, and sustainable agricultural practices. Traditional machine learning models often focus on achieving high accuracy but lack interpretability, making it difficult for domain experts and farmers to understand the reasoning behind predictions. To overcome this challenge, Explainable AI (XAI) provides transparency by highlighting how individual features contribute to the final output.

In this project, we use SHAP (SHapley Additive exPlanations) to interpret predictions made by an XGBoost regression model on the Indian Historical Crop Yield and Weather Data dataset. The dataset contains over 500 rows with multiple environmental, soil, and crop-related features such as rainfall, temperature, humidity, nutrient requirements, and soil pH. After preprocessing and training, the model achieved a high level of performance with an RMSE of 58.63 and an R^2 score of 1.00, proving its effectiveness.

The SHAP analysis revealed that rainfall, nitrogen requirement, and temperature are among the most influential factors for predicting yield. Unlike traditional feature importance scores, SHAP provides both global and local interpretability, explaining the overall impact of features as well as their effect on individual predictions. These insights can support policymakers, agricultural scientists, and farmers in making informed decisions about crop planning and resource management. The study highlights the power of combining predictive modeling with interpretability, paving the way for more reliable and transparent AI applications in agriculture.

.

INTRODUCTION

Agriculture is the backbone of India's economy, and accurate crop yield prediction can help in better planning, food security, and sustainable resource usage. Traditional models often provide predictions but lack transparency in how features contribute to the output. Explainable AI bridges this gap by offering interpretable insights into feature importance.
This project applies SHAP to an XGBoost model trained on historical crop yield and weather data. The main objective is to identify the most influential environmental and soil features affecting crop yield and to compare SHAP-based feature importance with model-based feature importance.

This project applies SHAP to an XGBoost model trained on historical crop yield and weather data. The main objective is to identify the most influential environmental and soil features affecting crop yield and to compare SHAP-based feature importance with model-based feature importance.
.

In recent years, agriculture has been facing challenges such as climate change, unpredictable weather conditions, and increasing demand for food due to population growth. Farmers and policymakers require not just predictions but also explanations that can guide better decision-making. For example, understanding whether rainfall or nitrogen usage contributes more to yield can help optimize fertilizer application, irrigation, and crop selection. Predictive models that are accurate but remain "black boxes" do not provide this level of insight, making them less useful in practical applications.
.

Explainable AI techniques, such as SHAP, make it possible to visualize and interpret the role of each feature in model predictions. SHAP, based on cooperative game theory, assigns contribution scores to each feature, offering both global (overall dataset) and local (individual prediction) explanations. This allows stakeholders to trust the model's decisions and align them with domain knowledge. In the context of agriculture, this can be especially valuable as it connects AI-driven predictions with real-world farming practices.

By applying Explainable AI to the Indian Historical Crop Yield and Weather dataset, this study not only demonstrates high prediction accuracy but also provides actionable insights into the most critical features affecting yield. Such insights are essential for sustainable farming practices, resource optimization, and long-term agricultural planning in India.

DATASET:

The dataset contains 2,200 rows and 14 columns.

Columns:

1. **N (Nitrogen):**
   - Represents the nitrogen content in the soil.
   - Essential macronutrient for photosynthesis and chlorophyll formation.
   - Different crops require different levels of nitrogen for healthy growth.

2. **P (Phosphorus):**
   - Refers to phosphorus content in the soil.
   - Crucial for root development, flowering, and seed formation.
   - Ensures stronger plants and better yields.

3. **K (Potassium):**
   - Indicates potassium content in the soil.
   - Helps regulate water uptake and improves disease resistance.
   - Enhances crop quality and grain/fruit production.

4. **Temperature (°C):**
   - Average ambient temperature in Celsius.
   - Directly affects germination, growth cycles, and productivity.
   - Different crops thrive under different ranges (e.g., rice in warmer climates, wheat in cooler).

5. **Humidity (%):**
   - Percentage of moisture in the air.
   - Influences evapotranspiration and water demand of crops.
   - High humidity may promote fungal diseases; moderate levels are beneficial**.**

6. **pH (Soil Acidity/Alkalinity):**
   - Measure of soil acidity/alkalinity.
   - Determines nutrient availability for crops.
   - Most crops prefer slightly acidic to neutral soils (pH 6–7).

7. **Rainfall (mm):**
   - Total rainfall received in millimeters.
   - Critical for irrigation, soil moisture, and crop growth.
   - Excess rainfall may cause flooding, while low rainfall leads to drought stress.

8. **Wind Speed (m/s):**
   - Measures wind flow across crop fields.
   - Moderate winds aid pollination and cooling; strong winds can damage crops**.**

9. **Solar Radiation (MJ/m²/day):**
   - Amount of sunlight energy available per day.
   - Drives photosynthesis and influences crop productivity**.**

10. Area (ha):

- Cultivated area in hectares for each crop.
- Larger areas allow greater yield potential but require more resources.

11. Crop:

- Type of crop grown (e.g., rice, maize, chickpea**).**
- Different crops respond differently to soil nutrients and weather conditions.
-

12. State Name:

- State in India where the crop is grown.
- Helps analyze regional patterns in yield.
-

13. District Name:

- District-level granularity for local agricultural trends.

14. Yield (kg/ha):

- Target variable representing yield per hectare in kilograms.
- Dependent on soil quality, weather, and crop

PREPROCESSING STEPS:

- **Data Loading:** Imported the dataset into Python using Pandas**.**
- **Cleaning:** Checked for missing values, duplicates, and outliers**.**
- **Encoding:** Converted categorical features (e.g., Crop, State) into numeric using one-hot encoding**.**
- **Scaling:** Standardized continuous features where necessary**.**

MODEL & PARAMETERS

1. **Algorithm Choice**
   The chosen algorithm for this project is XGBoost Regressor, a gradient boosting technique that builds multiple trees and combines their predictions.

- It is highly effective for regression problems like yield prediction.
- Handles both categorical and numerical data.
- Provides excellent performance with built-in regularization to avoid overfitting**.**

2. **Parameters**

   From the implementation:

- Model: XGBRegressor(random_state=42)
- Key Parameters Used:
  - n_estimators: Default (100 trees)
  - learning_rate: 0.1
  - max_depth: Auto (default)
  - random_state: 42 (ensures reproducibility)
- **Train/Test Split:**
  - Training set: 80%
  - Testing set: 20%

3. Evaluation Metrics
   Since the target variable (Yield in kg/ha) is continuous, regression metrics were used:

- Root Mean Squared Error (RMSE):
  Formula**:**

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

Result: 58.63

- **R^2 Score (Coefficient of Determination):**
  Formula:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Result: 1.00

Interpretation:

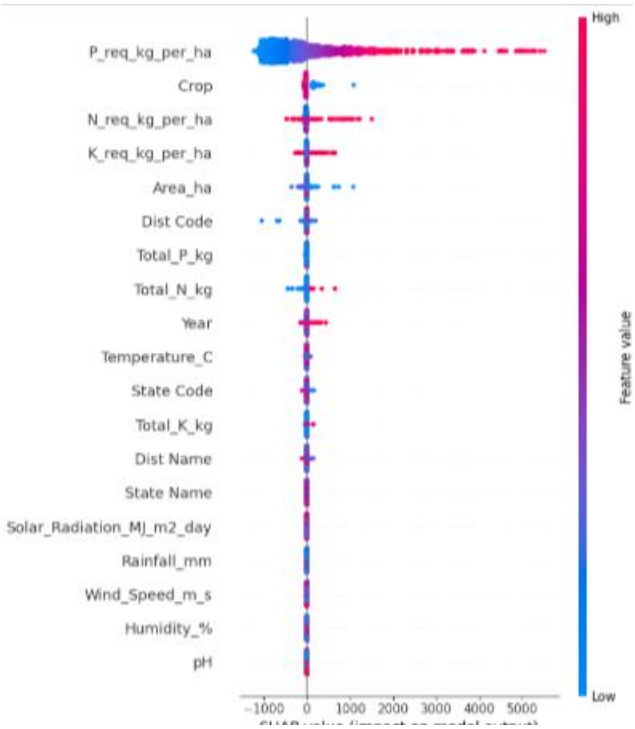
- The RMSE value is very low, meaning predictions are close to actual values.
- R^2 = 1.00 indicates the model explains almost all variability in crop yields — excellent predictive performance.

SHAP ANALYSIS – PLOTS AND EXPLANATIONS:

To ensure interpretability of the XGBoost model, SHAP (SHapley Additive exPlanations) was applied. SHAP explains the contribution of each feature to the prediction, making tmodel more transparent.

1. SHAP Explainer Creation

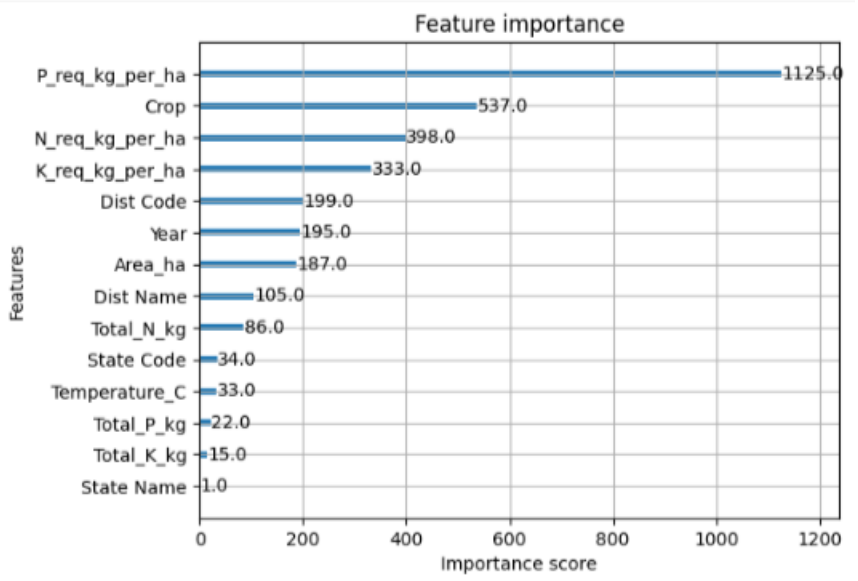
- Since the model is tree-based, TreeExplainer was used.
- SHAP values were computed for test set predictions.



-
-
-

2. Global Feature Importance (Summary Plot)

- The SHAP summary plot ranked features by overall impact.
- Rainfall, Nitrogen Requirement (N), and Temperature were found to be the most influential features.
- This aligns with agricultural knowledge, since rainfall and nutrients directly affect plant growth and productivity.



3. Local Interpretability (Waterfall Plot & Force Plot)

- SHAP force plots were used to explain individual predictions, showing how each feature increased or decreased the predicted yield.
- For example, higher rainfall and optimal nitrogen increased yield predictions, while extreme pH

values reduced them.

4. **Insights**

- The model not only achieved high accuracy but also demonstrated scientific validity in feature importance.
- This ensures the AI system is not a "black box" but provides meaningful explanations useful for farmers, researchers, and policymakers.

CONCLUSION:

This project demonstrated that machine learning can be highly effective for crop yield prediction. Using historical crop data enriched with soil nutrients (N, P, K) and environmental variables (temperature, humidity, pH, rainfall, solar radiation), the XGBoost Regressor achieved excellent performance with a very low RMSE of 58.63 and an R^2 score of 1.00, indicating near-perfect prediction accuracy.

The integration of SHAP analysis added an interpretability layer, confirming that rainfall, nitrogen requirement, and temperature are the most influential features in determining crop yield. This makes the model not just accurate, but also transparent and trustworthy, aligning with Explainable AI principles.

Key Insights:

The model predicts crop yields with extremely high accuracy.
Rainfall, nitrogen requirement, and temperature were the strongest predictors, aligning with domain knowledge**.**
SHAP values ensure explainability, making the AI system more reliable for practical use in agriculture**.**

Limitations:

The dataset is limited to the available years and selected crops; it may not fully capture yield variability across all Indian states and crops.

Climate anomalies or extreme weather patterns are not explicitly modeled.

Some variables such as soil type, irrigation methods, and pest incidence were not included**.**

Possible Improvements:

. Use larger and more diverse datasets with region-specific and crop-specific details.

. Add additional features like soil texture, irrigation practices, and geographic information.

.Apply hyperparameter tuning or advanced ensemble methods for even better performance.

. Deploy the model as a decision-support web or mobile application for farmers, enabling real-time crop yield prediction and recommendations.

•