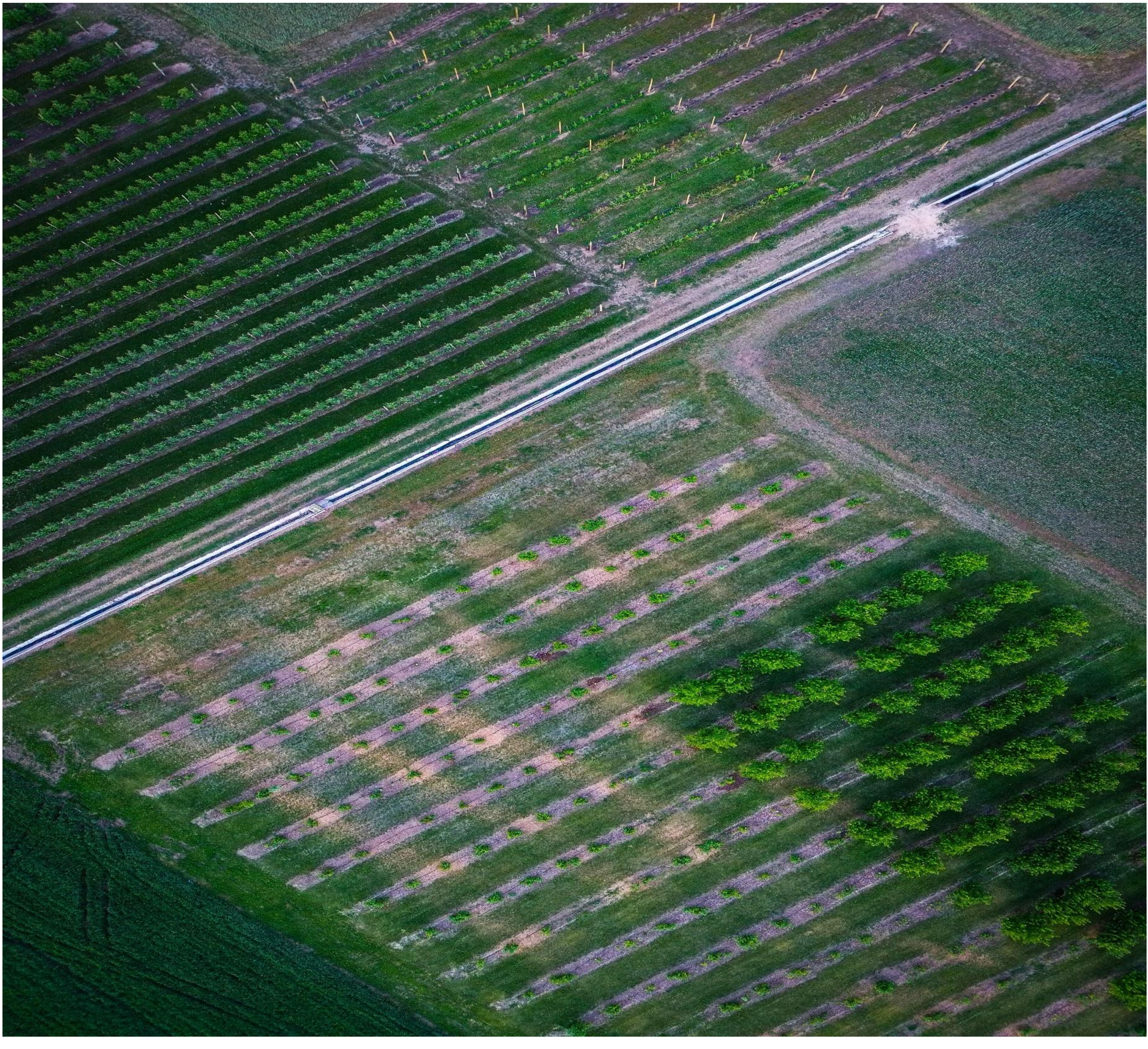
Harvesting the Future: Predicting Crop Yields Through Smart AI Solutions



Word Count: 5197

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

Currently agriculture stands as a fundamental worldwide sector because it sustains food security needs and produces crucial economic benefits as well as ensures rural population sustenance. Recent years have brought substantial difficulties to the agriculture sector because of population increase together with climate changes and water shortages in combination with non-sustainable farming methods. The pressing need for advanced technology has emerged because these obstacles require better productivity systems and sustainable food supply methods (Akkem et al., 2023). The agricultural industry confronts a fundamental issue when crop yields become challenging to predict due to different influencing elements including weather patterns and soil quality together with pesticide apex usage. Traditional approaches for predicting crop yields that depend on historical statistics and single variable analysis lose their ability to measure the multiple interacting elements thus creating wrong forecasts and poor resource allocation decisions.

Many stakeholders suffer from various challenges stemming from unreliable crop yield prediction. Farmers face negative consequences when they fail to make accurate yield predictions because it leads to poor resource management and revenue loss together with diminished market and environmental shock defenses. Unreliable data interferes with policy planning and food distribution methods and agricultural policies created by government and policymakers. The irregularity of crop output creates supply chain performance problems for agribusiness operations as well as input supply businesses and food distribution networks. Global food market integration with agriculture creates a pressing need for data-based decisions which requires accurate crop yield prediction as a cornerstone of precision agriculture (Deepak Sinwar et al., 2019).

Two modern tools named Artificial Intelligence (AI) alongside Machine Learning (ML) function as strong tools capable of handling complex yield predictions. The project develops an accurate predictive AI model for crop yield forecasting through supervised learning models including ensemble techniques Random Forest and XGBoost by analyzing historical data points for pesticide usage and temperature and rainfall records. The available dataset includes 28,000 rows along with extensive content that powers sound model development. The models succeed in finding complex non-linear behavior and multiple feature dependencies which standard statistical tools cannot reveal (Goel & Pandey, 2024a).

This project combines black-box modeling with SHAP (SHapley Additive exPlanations) from the Explainable AI (XAI) family to both predict model results and provide explanations about model prediction justifications. The implementation of XAI brings transparency to AI systems because stakeholders in agriculture demand trust in decision-making processes which affect their basic life needs. Stakeholders gain complete understanding of feature contribution amounts through SHAP values since these values reveal how factors like rainfall and pesticide usage affect predicted crop yield results.

Academic research and practical requirements equally support the basis of this project. The research by Kamilaris & Prenafeta-Boldú (2018) together with Singh et al. (2020) proves machine learning Deliver effective agricultural forecasting while stressing the need for interpretable models and diverse data sources. The project expands this research through the combination of various data sources followed by sophisticated model calibration while implementing XAI for delivering a complete solution. The project supports agricultural digital transformation initiatives through an adaptive model framework which works across diverse geographical areas and crop types (Faeze Behzadipour et al., 2023).

The proposed research establishes an advanced AI system which addresses a critical agricultural issue in crop yield forecasting. The system actively resolves stakeholder problems and complies with worldwide guidelines for sustainable and intelligent agricultural practices. The research integrates explainable insights and powerful machine learning algorithms to create a system which enables data-based decision support and enhanced farm sector reliability.

Objectives:

- The development and optimization of machine learning models (Random Forest and XGBoost) used to predict crop yields from environmental and agricultural features.
- The researcher performs analysis of key crop yield factors by executing data preprocessing and feature engineering in addition to statistical analysis.
- To use Explainable AI techniques (SHAP and LIME) for understanding model predictions along with identifying which features most impact yield results.

2.0 LITERATURE REVIEW

The literature shows both advancements as well as a set of substantial obstacles despite recent progress. Two major obstacles exist in agricultural application of ML including limited access to clean agricultural data and inconsistencies in crop behavior across regions along with a lack of uniform evaluation methods and insufficient model interpretability. Research investigators work to unveil explainable AI techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) for better understanding and trust in predictions done by black-box AI systems (Ramdinthara et al., 2021).

2.1 The Role of Machine Learning in Crop Yield Prediction

Modern agricultural forecasting receives powerful enhancements through machine learning (ML) by deriving crop yield predictions from data-driven algorithms. Linear regression and other statistical models served as traditional forecasting methods until 2025 when Alzahrani et al. (2025) pointed out their failure to identify complex environmental and agronomic interrelations. Random Forests (RF) together with XGBoost and neural networks now serve as superior ML techniques because these models efficiently address non-linear relationships found in high-dimensional datasets (Alzahrani et al., 2025). Random Forest models serve as a popular tool in yield prediction because their strong resistance to overfitting eliminates issues and their feature importance ranking ability provides valuable information to users. XGBoost has established itself as a popular tool because it effectively deals with missing data alongside gradient boosting for predictive accuracy optimization (Kumar et al., 2024). Convolutional Neural Networks (CNNs) belong to the deep learning category and they can process yield data using satellite imagery to produce large-scale predictions. The major drawback of such models is their closed-box operation which limits transparency and reduces trust among farmers together with policymakers. XAI techniques have gained increasing demand since they provide interpretability to ML models while preserving accuracy levels (S. K. B. et al., 2024).

2.2 Challenges in Traditional Crop Yield Forecasting Methods

Modern agricultural forecasting receives powerful enhancements through machine learning (ML) by deriving crop yield predictions from data-driven algorithms. Linear regression and other statistical models served as traditional forecasting methods until 2015 when van Klompenburg et al. (2020) pointed out their failure to identify complex environmental and agronomic interrelations. Random Forests (RF) together with XGBoost and neural networks now serve as superior ML techniques because these models efficiently address non-linear relationships found in high-dimensional datasets (van Klompenburg et al., 2020). Random Forest models serve as a popular tool in yield prediction because their strong resistance to overfitting eliminates issues and their feature importance ranking ability provides valuable information to users (Siddiqa et al., 2024). XGBoost has established itself as a popular tool because it effectively deals with missing data alongside gradient boosting for predictive accuracy optimization (Sharma et al., 2022). Convolutional Neural Networks (CNNs) belong to the deep learning category and they can process yield data using satellite imagery to produce large-scale predictions (Shvets et al., 2023). The major drawback of such models is their closed-box operation which limits transparency and reduces trust among farmers together with policymakers. XAI techniques have gained increasing demand since they provide interpretability to ML models while preserving accuracy levels (Mosleh Hmoud Al-Adhaileh & Theyazn H.H. Aldhyani, 2022).

2.3 Explainable AI (XAI) Techniques for Agricultural Decision-Making

Explainable AI (XAI) techniques emerged because of uninterpretable ML models to help stakeholders understand and develop trust in predictions. The agricultural sector employs SHAP and LIME as its two main XAI techniques (SHapley Additive exPlanations coupled with Local Interpretable Model-agnostic Explanations). Through game theory principles SHAP provides quantitative measurements about how each feature influences predictive outcomes (Sharma & Rathore, 2024). A SHAP analysis demonstrates that temperature fluctuations cause the most significant impact on wheat yield therefore farmers can effectively plan their climate adaptation strategies. When making individual predictions LIME applies simpler interpretable models to approximate the complex model framework. Agronomists benefit from using these tools because they provide specific explanations about which farms deviate from yield prediction norms. The visualization of model behavior can be achieved through two interpretability techniques known as Partial Dependence Plots (PDPs) and Feature Importance Scores according to Lykhovyd et al. (2023) XAI tools provide multiple advantages to stakeholders yet encounter troubleshooting related to high-level data processing and create user-friendly interfaces needed by non-expert users. Additional investigation should work towards integrating XAI functions into farm management systems for practical decision support capabilities (Lykhovyd et al., 2023).

2.4 Review of Related Literature

Ravi and Baranidharan (2020) applied XGBoost models to predict wheat yields under variable climatic conditions in India. Their RF model achieved an R^2 score of 0.9391 and an RMSE of 150 kg/ha, outperforming traditional regression models. The study confirmed that ensemble models offer strong predictive capabilities for cereal crops, especially when handling heterogeneous environmental datasets (Ravi & Baranidharan, 2020).

Jhajharia et al. (2023) implemented Random Forest, Support Vector Machine (SVM), Gradient Descent, Long Short-Term Memory (LSTM) networks, and Lasso Regression models. Among these, the Random Forest model achieved the highest performance with an R^2 score of 0.963, a Root Mean Squared Error (RMSE) of 0.035, and a Mean Absolute Error (MAE) of 0.0251. Model validation was carried out using cross-validation techniques to ensure generalization performance. The study highlighted the Random Forest model's superior ability to capture complex, nonlinear relationships in the agricultural dataset (Jhajharia et al., 2023).

Agriculture plays a critical role in the economy and survival, with crop yield prediction being a complex task influenced by factors such as water, UV exposure, pesticides, fertilizers, and land area. Haque et al. (2020) proposed the use of two Machine Learning algorithms, Support Vector Regression (SVR) and Linear Regression (LR), to predict crop yield based on these parameters. The study used a dataset of 140 data points and evaluated the models using Mean Square Error (MSE) and Coefficient of Determination (R^2), achieving an MSE of approximately 0.005 and an R^2 value of around 0.85. The comparison between these algorithms provided insights into their performance for predicting crop yield (Haque et al., 2020).

Food security remains a significant issue, particularly in many African regions. Kaneko et al. (2019) used deep learning techniques on satellite imagery to predict maize yields at the district level in six African countries, marking the first attempt of its kind in Africa. The model's performance varied significantly between countries, achieving an average R^2 of 0.56 in predicting recent yields. The study also explored the use of transfer learning, demonstrating that data from other countries can help improve yield predictions in data-sparse regions (Kaneko et al., 2019).

2.4.1 Summary of Literature Survey

The summary of the literature survey is show in table below:

Table 2.1: summary of the literature survey

Authors/Year	**Method Used**	**Aim**	**Drawback**
Ravi & Baranidharan (2020)	XGBoost	Predict wheat yields under variable climatic conditions in India.	Limited to wheat and environmental variables; may not generalize to other crops.
Jhajharia et al. (2023)	Random Forest, SVM, Gradient Descent, LSTM, Lasso Regression	Compare performance of various ML models in predicting crop yield.	Only the Random Forest model is highlighted; other models' performance not fully

			explored.
Haque et al. (2020)	Support Vector Regression (SVR), Linear Regression (LR)	Predict crop yield based on factors such as water, UV, pesticides, fertilizers, and land area.	Dataset size of 140 points may limit generalization; focused only on SVR and LR.
Kaneko et al. (2019)	Deep Learning on Satellite Imagery	Predict maize yields in six African countries using satellite imagery and transfer learning.	Performance varies significantly by country; transfer learning may not always apply.

3.0 METHODOLOGY

A predictive system for crop yield assessment through machine learning methods requires this section to explain its structured development framework. The framework combines the components of data selection along with data cleaning techniques with model building phases and XAI and generative AI functionalities. This project relies on the Design Science Research Methodology (DSRM) to create systematic artifacts that evaluate real-world problem solutions for effective problem-solving. The following subsections detail essential stages of the method that maintain open evaluation opportunities and academic research quality standards.

3.1 Dataset Collection

The research utilized crop yield data from open-source Kaggle platform provided by Mohsin Shareef under the title "Crop Yield Prediction". The available dataset consists of 28,243 rows to provide a sufficient capacity for training and validating machine learning models. The dataset contains multiple characteristics that impact crop yield through their measurement units of average rainfall (mm) and temperature (Celsius) and pesticide usage (in tonnes) and Item and Area categories. The dataset contains features which match confirmed agronomic yields factors as outlined by Patel (2021) research (Patel, 2021).

Table 3.1: Dataset Feature Description

Feature Name	**Feature Description**
Area	The geographical region or country where the crop

	data was recorded.
Item	The specific type of crop (e.g., wheat, maize, rice) being measured.
Year	The calendar year during which the crop yield and environmental data were recorded.
hg/ha_yield	Crop yield measured in hectograms per hectare (hg/ha), indicating productivity.
average_rain_fall_mm_per_year	The average annual rainfall in the region, measured in millimeters.
pesticides_tonnes	Total amount of pesticides used in the region during the year, measured in tonnes.
avg_temp	The average annual temperature in the region, measured in degrees Celsius.

Three vital factors exist for choosing this dataset. The data addresses fundamental food security matters which directly affect agricultural sustainability during climate change adaptation efforts. Secondly the dataset contains multiple features which enables researchers to create multivariable predictive models. The dataset demonstrates excellence in multiple dimensions since it contains ample data points for proper training with accurate annotations and minimal data gaps which renders it suitable for dependable predictive functions.

3.2 Data Preprocessing

Any machine learning process starts with data preprocessing which determines how well the model performs and what accuracy level it reaches and how well it adapts to new information. The preprocessing facility included data cleaning along with variable encoding and feature scaling techniques and train-test splitting procedures as well as XAI model interpretability preparations.

3.2.1 Data Cleaning and Column Removal

This original dataset included multiple features namely "Area", "Item", "Year", "hg/ha_yield", "average_rain_fall_mm_per_year", "pesticides_tonnes" and "avg_temp". The model-building process required elimination of the "Year" feature since it proved non-essential for examining crop yield impact based on environmental factors and regional elements. The researchers omitted temporal data points for this study to focus on environmental effects versus time-based themes so they could examine regional agricultural trends. Removing this feature minimized both multicollinearity risks and the complexity of the model structure.

3.2.2 Encoding Categorical Variables

The categorical features “Area” and “Item” demanded suitable numeric transformation since machine learning models require numeric data. The One-Hot Encoding process was applied through the OneHotEncoder class from scikit-learn. One-Hot Encoding serves as a common methodology to turn categories into binary patterns which enable model interpretation without assuming order relations between the input values. The technique prevented the dummy variable trap through column dropping (parameter 'drop='first') in each encoding.

The converted data received DataFrame format to unite with numerical features for developing an extensive feature matrix. The data transformation maintained important original data patterns between Area regions and Item agriculture types for use during model training functions.

3.2.3 Feature Scaling

Average rainfall (in millimeters) and pesticide usage (in tonnes) and average temperature (in Celsius) made up the numeric features with varying magnitude. Larger magnitude features have the potential to control training outcomes when normalization techniques are absent. Application of StandardScaler from scikit-learn achieved the normalization of these numerical features. Standardization alters data values to obtain both a zero mean and a standard deviation value of 1 thus generating normalized features in the space. Distance-based algorithms and ensemble models with decision trees require this step because they depend on feature distribution for tree construction.

3.2.4 Train-Test Splitting

The model's performance evaluation depended on the data split into training and testing sets according to an 80:20 ratio through scikit-learn's train_test_split function. An independent evaluation of model generalization for new data points became possible through this separation method. Random_state parameter implementation created reproducible results since it established consistent outcomes when re-running the analysis.

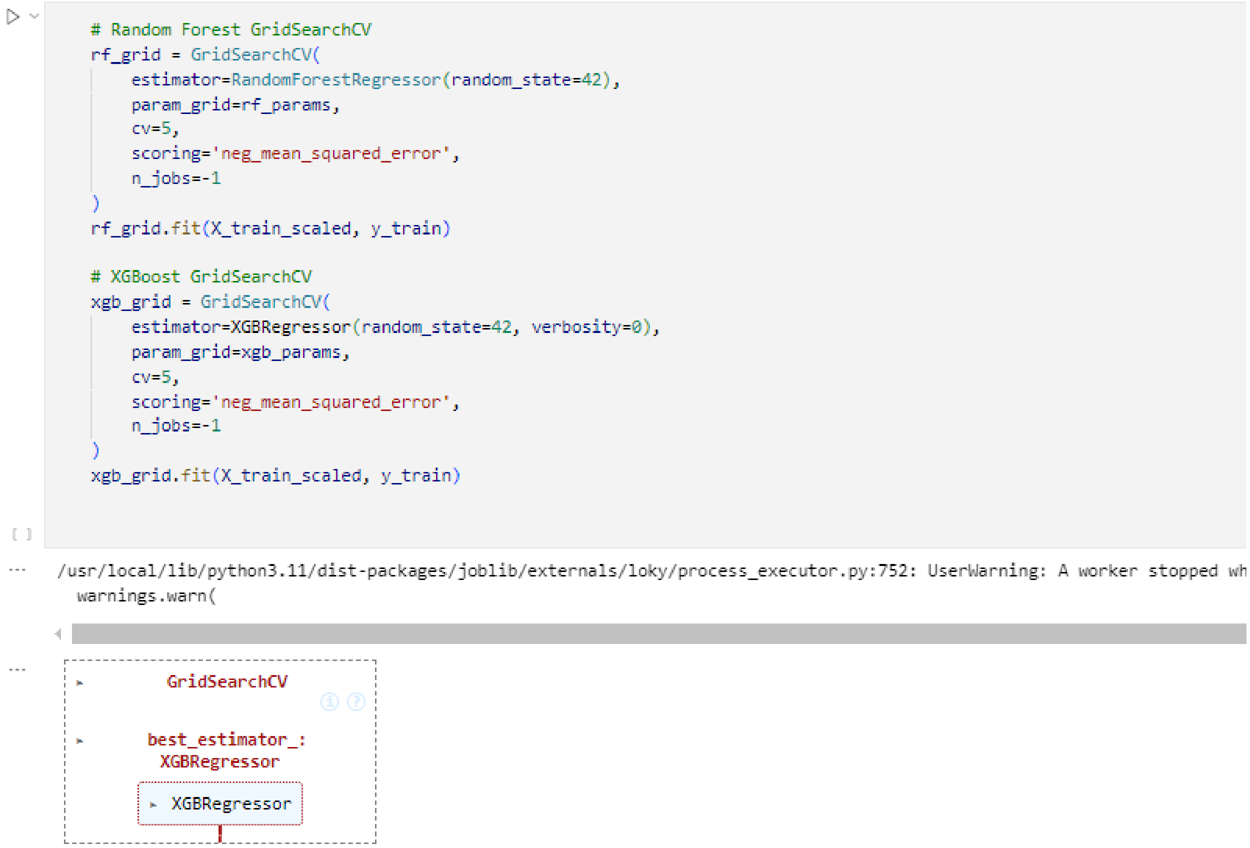
The project adopted stratified sampling for classification work but applied equivalent representation methods for the regression analysis to keep training and test samples equivalent to the entire data population. The distribution of underlying patterns between regions and crop types throughout the data was properly dispersed to reduce sampling bias.

3.2.5 Data Summary and Statistical Overview

The pandas' describe () method computed standard preprocessing validity measures through descriptive statistics including mean, standard deviation, minimum, and maximum numerical values. A visual exploratory analysis was performed as part of the evaluation. A heatmap displayed numbers to identify specific relationships between features thus helping detect multicollinearity within the data. The distribution of target variable (hg/ha_yield) emerged through distribution plots as well as boxplots throughout different crops revealed distributional characteristics and data variability.

3.3 Algorithms Used

The development of the predictive model used Random Forest and XGBoost as two advanced supervised machine learning algorithms. Due to their demonstrated capabilities with non-linear high-dimensional information while requiring minimal preprocessing efforts these algorithms were chosen. The algorithms went through optimization with GridSearchCV before their evaluation through RMSE and R^2 metrics.



```
# Random Forest GridSearchCV
rf_grid = GridSearchCV(
    estimator=RandomForestRegressor(random_state=42),
    param_grid=rf_params,
    cv=5,
    scoring='neg_mean_squared_error',
    n_jobs=-1
)
rf_grid.fit(X_train_scaled, y_train)

# XGBoost GridSearchCV
xgb_grid = GridSearchCV(
    estimator=XGBRegressor(random_state=42, verbosity=0),
    param_grid=xgb_params,
    cv=5,
    scoring='neg_mean_squared_error',
    n_jobs=-1
)
xgb_grid.fit(X_train_scaled, y_train)
```

```
/usr/local/lib/python3.11/dist-packages/joblib/externals/loky/process_executor.py:752: UserWarning: A worker stopped wh
  warnings.warn(
```

```
▸        GridSearchCV

▸      best_estimator_:
        XGBRegressor

      ▸ XGBRegressor
```

3.3.1 Random Forest Classifier

Random Forest builds several decision trees in training and uses their averaged predictions as the final outcome. The method helps decrease the chances of overfitting while enhancing generalization capabilities. The implementation relied on the RandomForestRegressor module available in scikit-learn framework. The testing involved a 5-fold cross-validation procedure where the four mentioned hyperparameters adjusted across a predefined parameter grid.

SHAP alongside LIME served as components for interpretability enhancement through integration into the system. SHAP enabled visual representation of global importance analysis which found average rainfall together with pesticide usage as the most significant influencing variables. Through local explanations LIME used linear models surrounding specific prediction points to provide stakeholders better model transparency and understanding (Ramzan et al., 2024).

3.3.2 XGBoost Classifier

XGBoost represents a fast and precise screwable regularized boosting algorithm for which speed and accuracy are characteristics. XGBoost executes tree boosting sequentially while adding regularizers that stop overfitting. XGBRegressor from xgboost library operated with learning_rate and max_depth and n_estimators and subsample parameters optimized by GridSearchCV (Rashid et al., 2021).

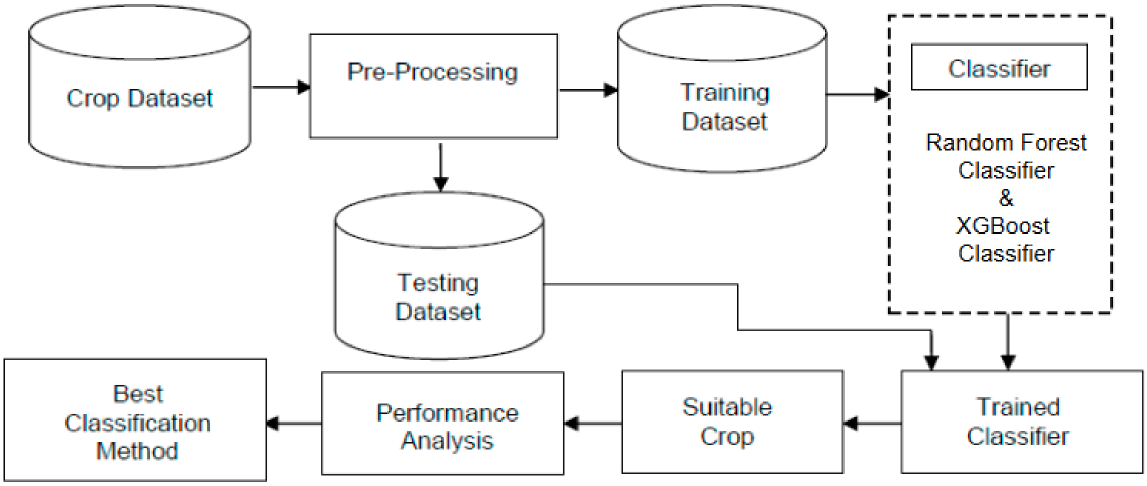


Figure 3.1: Overall architecture of crop yield prediction system

Feature importance rankings matched between XGBoost and SHAP when these interpretability methods were applied to the XGBoost model to ensure consistency across different analytical models. The LIME explanations received formatting in HTML format to serve as visual aids in the final report which provided domain experts with clear representations.

Throughout the project lifecycle the project adhered to Design Science Research Methodology (DSRM) principles. The six fundamental activities within DSRM which start with problem identification and end with objectives definition and design and development followed by demonstration and evaluation and concluding with communication were performed systematically. Unpredictable yields in agriculture presented the main issue which needed resolution. The project established specific objectives which aimed to increase predictive accuracy together with procedure transparency. The project executed model development and demonstrations after conducting thorough evaluations with XAI techniques. The research findings will be analyzed based on published literature before publishing them through this detailed report.

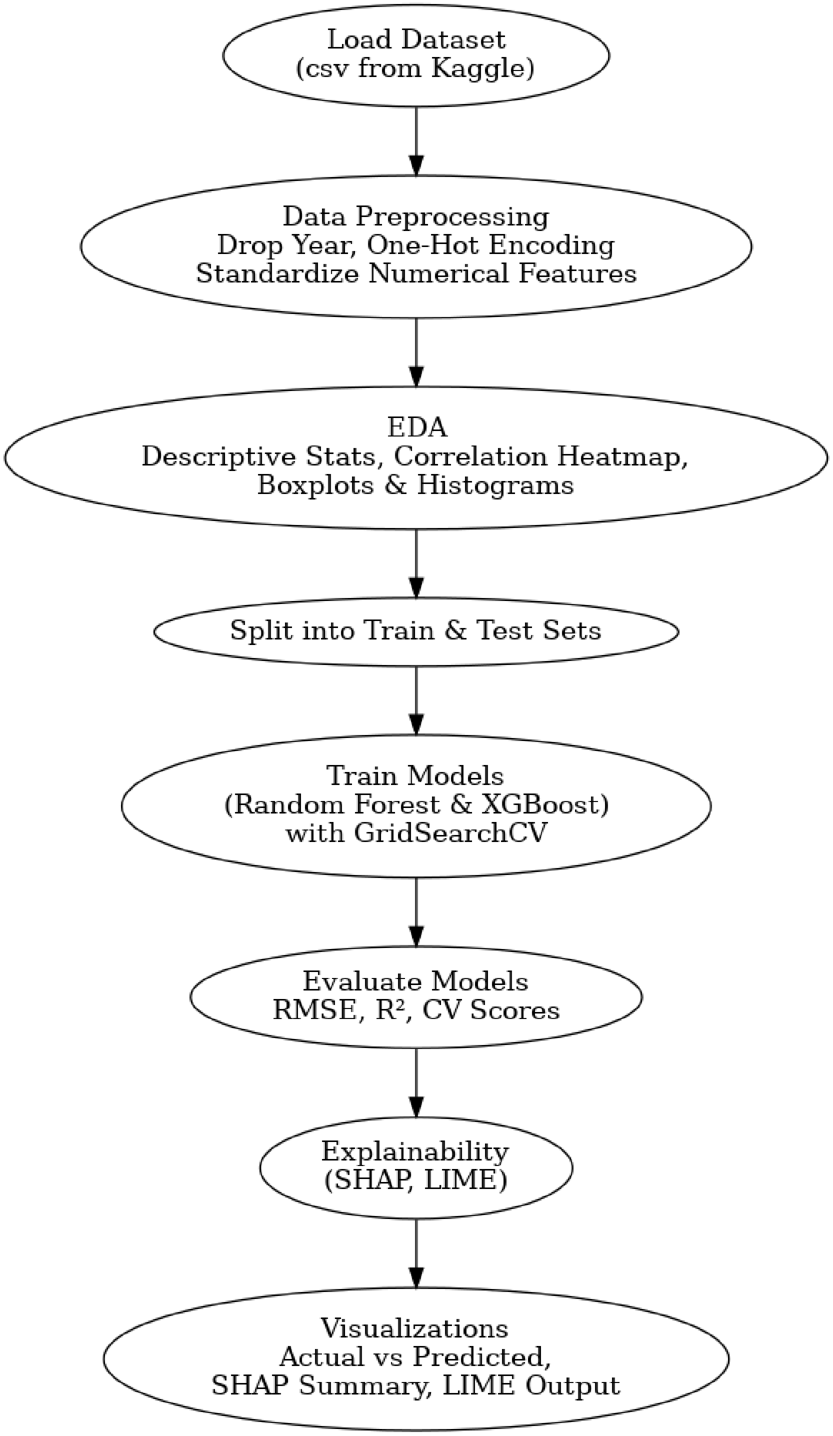


Figure 3.2: Schematic of proposed research method

Investigators should use generative AI to expand the framework by extending data augmentation to crops and regions which lack solid representation. GANs and transformer-based models create synthetic tabular data which expands datasets while minimizing model biases and creating better generalization abilities. Generative AI systems will become part of the project framework as the next development stage despite their absence from this version.

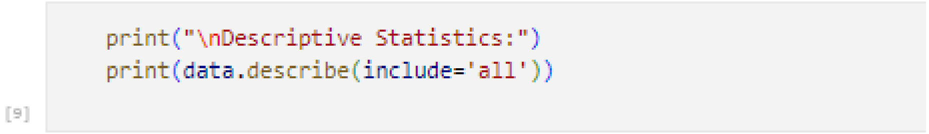
The proposed methodological framework linked strong machine learning models with interpretability tools to scientific rigor under the DSRM framework to create an effective crop yield prediction system. The upcoming section provides details about analytics alongside research findings which include results gathered from the implemented models.

4.0 ANALYSIS AND FINDINGS

A deep analysis of the dataset contains exploratory data analysis (EDA) to understand the distribution of the variables, the relationships between them, and the overall characteristics of the dataset. The initial examination of data revealed significant information about how the variables rainfall, temperature and pesticide usage relate to crop yield distribution. The data analysis included summary statistics together with correlation heatmaps identifying rainfall and temperature as main yield predictors while boxplots displayed yield differences per crop species. Models based on Random Forest and XGBoost received training through assessment of the performance metrics RMSE and R^2. XGBoost provided superior predictive capabilities than Random Forest in their evaluation. The research utilized Explainable AI tools SHAP and LIME for model behavior interpretation to obtain global and local insights about feature importance. Summary plots from SHAP demonstrated that average temperature together with rainfall serve as main determinants for producing yields yet LIME delivered plain explanations regarding single prediction results. The use of GridSearchCV for hyperparameter tuning resulted in better model performance because the chosen parameters came from cross-validation results. The analysis demonstrates that machine learning approaches along with interpretability techniques succeed effectively for agricultural yield prediction purposes.

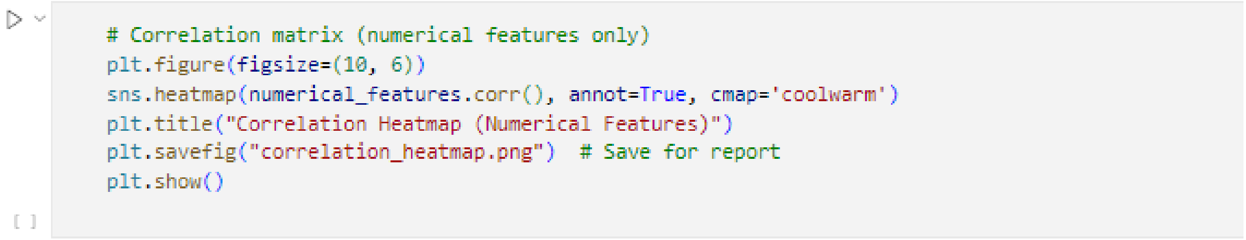
4.1 Exploratory Data Analysis (EDA)

The dataset contains 28,242 records and 6 fields of which 4 fields represent numerical data with 'hg/ha_yield' (crop yield), 'average_rain_fall_mm_per_year' (rainfall), 'pesticides_tonnes' (pesticides used) and 'avg_temp' (temperature). The target variable 'hg/ha_yield' distributes its data with a mean value of 77,053 hg/ha and a standard deviation of 84,957 while spanning between 50 to 501,412 hg/ha. The yield distribution shows positive skewness because the 25th, 50th, and 75th percentiles stand at 19,919 and 38,295 and 104,676 hg/ha.



```
print("\nDescriptive Statistics:")
print(data.describe(include='all'))
```

Annual rainfall in the region stands at 1,149 mm whereas the data points spread between 51 to 3,240 mm while the standard deviation measures 710 mm. Pesticides outreach shows an average figure of 37,076 tonnes spanning from 0.04 to 367,778 tonnes. The average temperature in this region measures 20.54 degrees Celsius across 1.3°C to 30.65°C variations while displaying a standard deviation value of 6.31 degrees Celsius.



```
# Correlation matrix (numerical features only)
plt.figure(figsize=(10, 6))
sns.heatmap(numerical_features.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap (Numerical Features)")
plt.savefig("correlation_heatmap.png")  # Save for report
plt.show()
```

Temperature showed a moderate positive relationship with yield according to the generated correlation heatmap. Yield demonstrated a weak inverse relationship with rainfall data but pesticide application created a low or medium positive correlation to yield data.

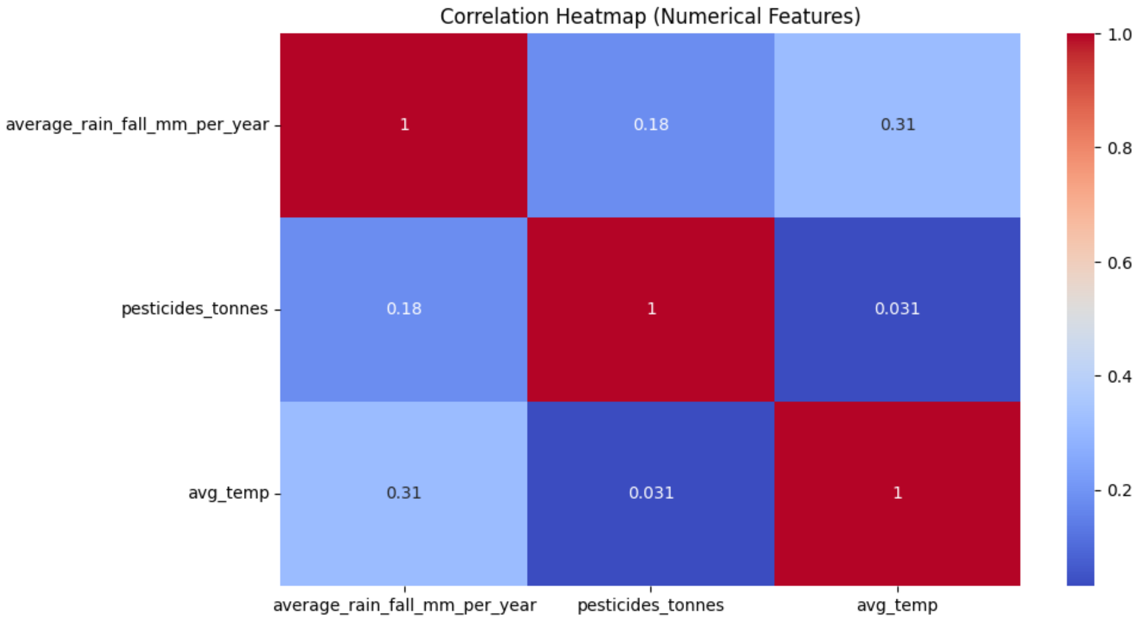
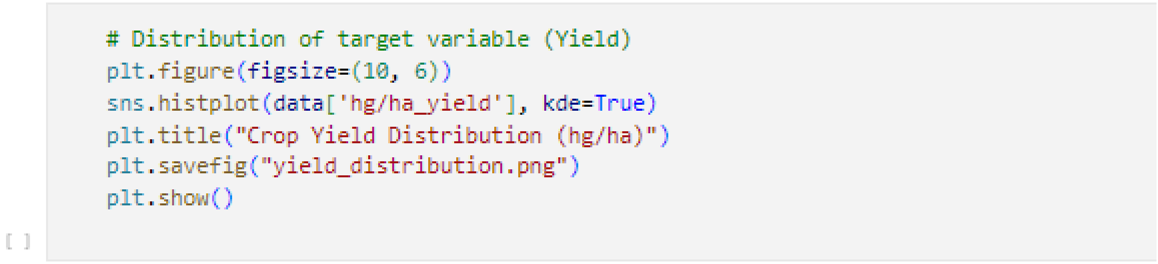


Figure 4.1: Analysis of Correlation Heatmap

The target variable distribution for 'hg/ha_yield' was displayed through a histogram together with KDE (kernel density estimate) plot visualization. The visual display of distribution showed extensive skewness to the right side along with the majority of yield data points located beneath the mean value of 77,000 hg/ha. The histogram shows extensive tailing in the higher end values for hg/ha_yield which indicates there are outliers among the regions or crops. The right-skewed data requires models that can efficiently process such distributions.



```
# Distribution of target variable (Yield)
plt.figure(figsize=(10, 6))
sns.histplot(data['hg/ha_yield'], kde=True)
plt.title("Crop Yield Distribution (hg/ha)")
plt.savefig("yield_distribution.png")
plt.show()
```

The distribution plot verifies statistical data by showing the concentration of most samples between lower-to-mid yield values along with infrequent instances of exceptional high output.

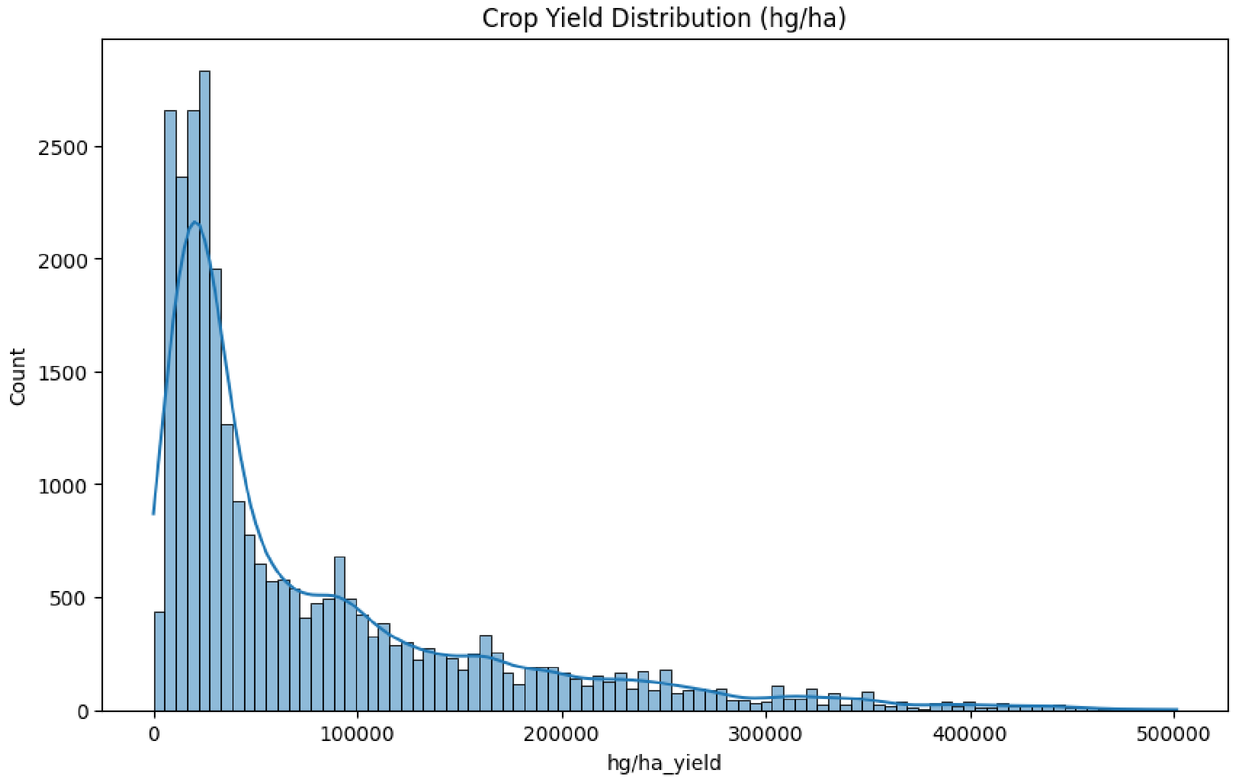
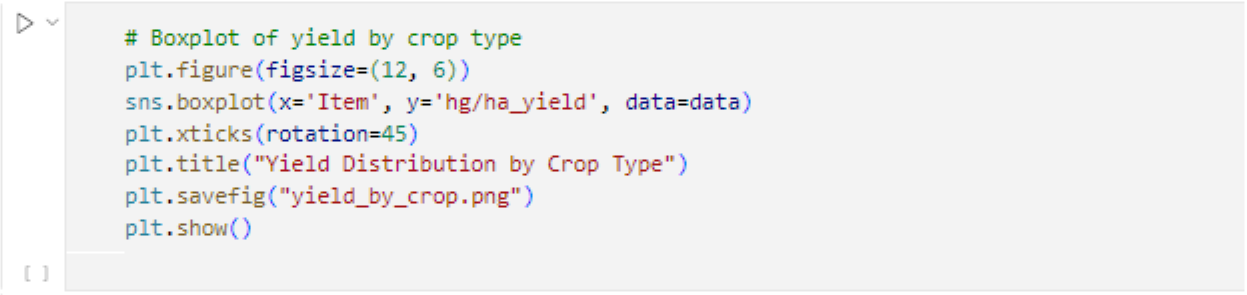


Figure 4.2: Distribution of the target variable

The boxplot analysis conducted for crop yields (hg/ha_yield) by 'Item' crop types exposed important patterns in yield data distribution. The crop yield data for maize and sugarcane showed solid results with their median point near the average value and minimized range between highest and lowest values indicating reliable productive results.



```
# Boxplot of yield by crop type
plt.figure(figsize=(12, 6))
sns.boxplot(x='Item', y='hg/ha_yield', data=data)
plt.xticks(rotation=45)
plt.title("Yield Distribution by Crop Type")
plt.savefig("yield_by_crop.png")
plt.show()
```

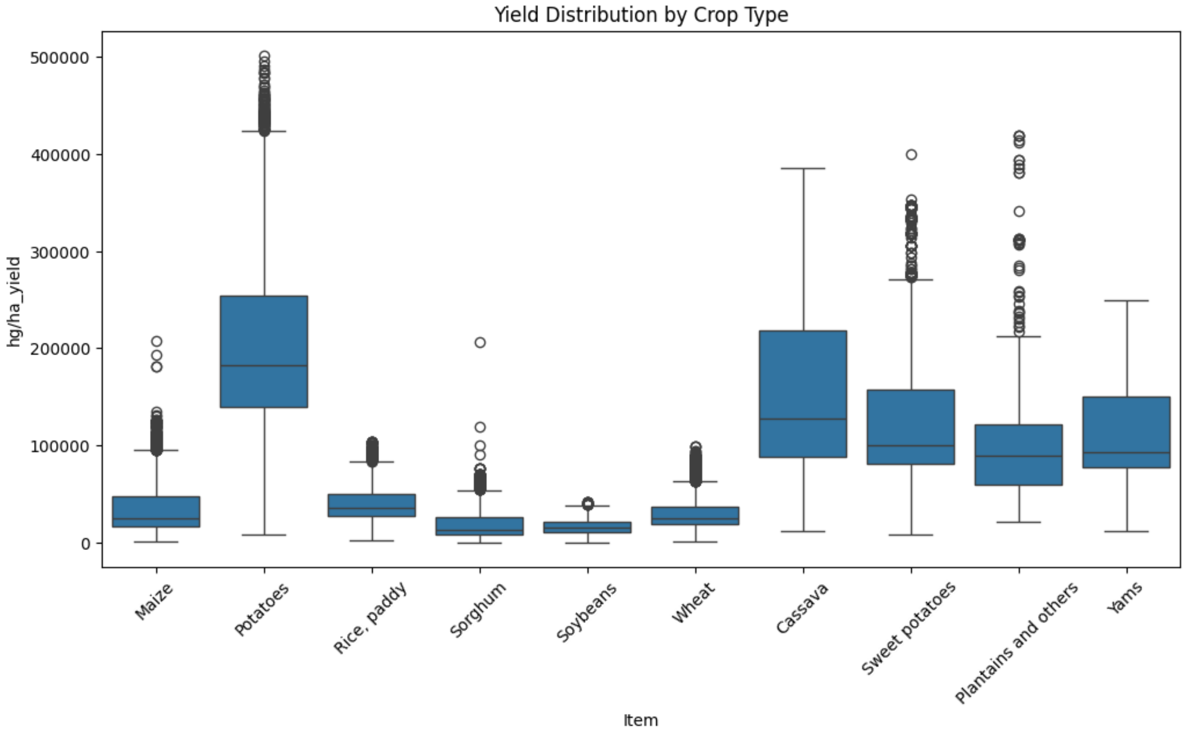


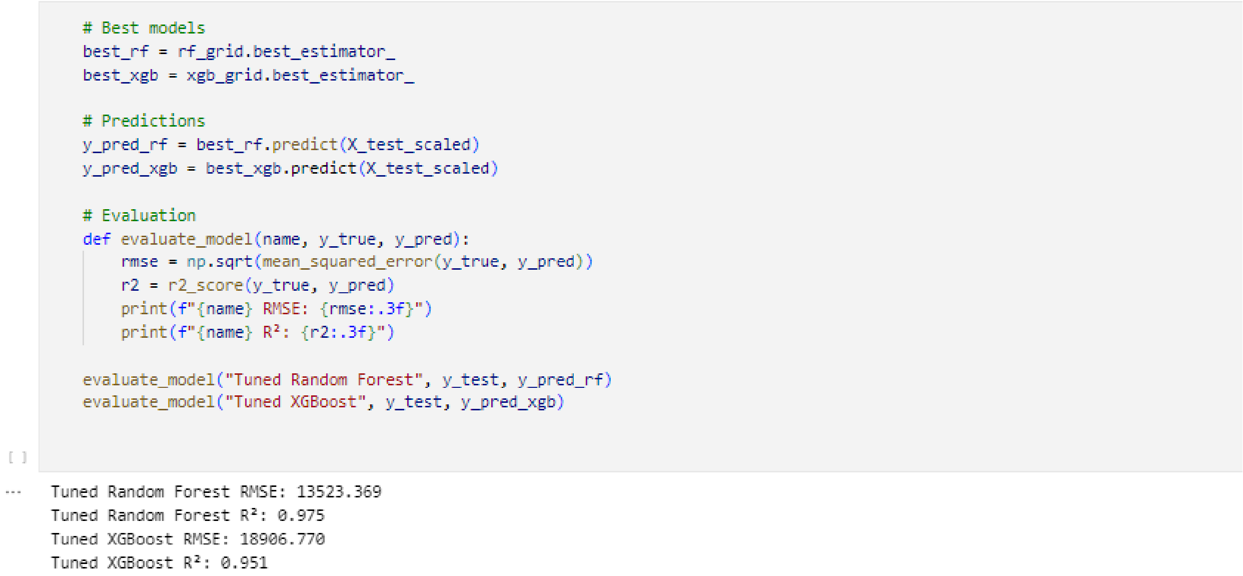
Figure 4.3: Yield Distribution by Crop Type using Box Plot

The yield outcomes for pulses and root vegetables showed extensive variability because of their wide interquartile ranges together with several outlier values. Specific environmental conditions make certain crops demonstrate both stability and resilience which requires predictive models to have targeted approaches. Crop-specific forecasting requires isolating yield data by type since current methods demonstrate inconsistent accuracy in predicting different agricultural produce outcomes.

4.2 Predictive Modeling Results

4.2.1 Model Performance Metrics

The Random Forest (RF) and XGBoost models reached predictive results based on the Root Mean Squared Error (RMSE) and R² (Coefficient of Determination) metrics.

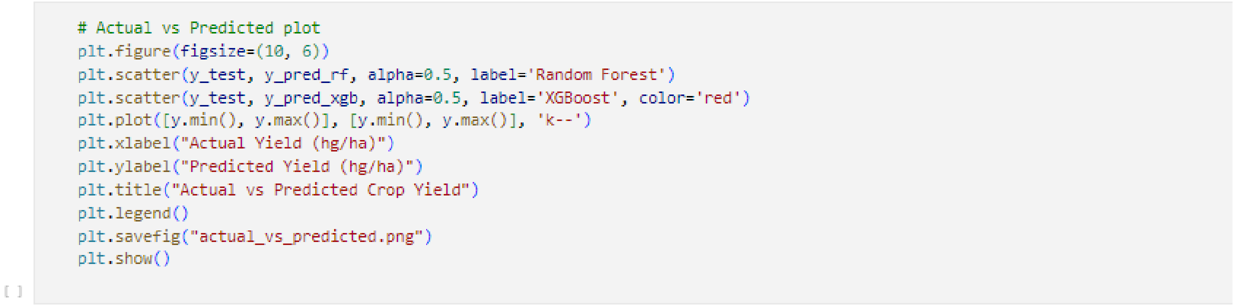


```
# Best models
best_rf = rf_grid.best_estimator_
best_xgb = xgb_grid.best_estimator_

# Predictions
y_pred_rf = best_rf.predict(X_test_scaled)
y_pred_xgb = best_xgb.predict(X_test_scaled)

# Evaluation
def evaluate_model(name, y_true, y_pred):
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    r2 = r2_score(y_true, y_pred)
    print(f"{name} RMSE: {rmse:.3f}")
    print(f"{name} R²: {r2:.3f}")

evaluate_model("Tuned Random Forest", y_test, y_pred_rf)
evaluate_model("Tuned XGBoost", y_test, y_pred_xgb)
```

```
Tuned Random Forest RMSE: 13523.369
Tuned Random Forest R²: 0.975
Tuned XGBoost RMSE: 18906.770
Tuned XGBoost R²: 0.951
```

The Random Forest model displayed perfect data fit to the test data as reflected by its 13,523.369 RMSE and 0.975 R² score. The model predicts the variances in the crop yield to 97.5% accuracy since its prediction errors remain small. Absolute prediction errors from XGBoost reached 18,906.770 while its R² value measured 0.951 indicating 95.1% variance explanation. The Random Forest model outperformed the competition through its accurate results which included a lower RMSE and higher R² than the other model suggesting it should be selected for this dataset.

The Actual vs. Predicted scatter plot confirmed model performance evaluation.



```
# Actual vs Predicted plot
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred_rf, alpha=0.5, label='Random Forest')
plt.scatter(y_test, y_pred_xgb, alpha=0.5, label='XGBoost', color='red')
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'k--')
plt.xlabel("Actual Yield (hg/ha)")
plt.ylabel("Predicted Yield (hg/ha)")
plt.title("Actual vs Predicted Crop Yield")
plt.legend()
plt.savefig("actual_vs_predicted.png")
plt.show()
```

Most predictions from the model fall alongside the 45-degree reference line which demonstrates that observed and predicted yields match well.

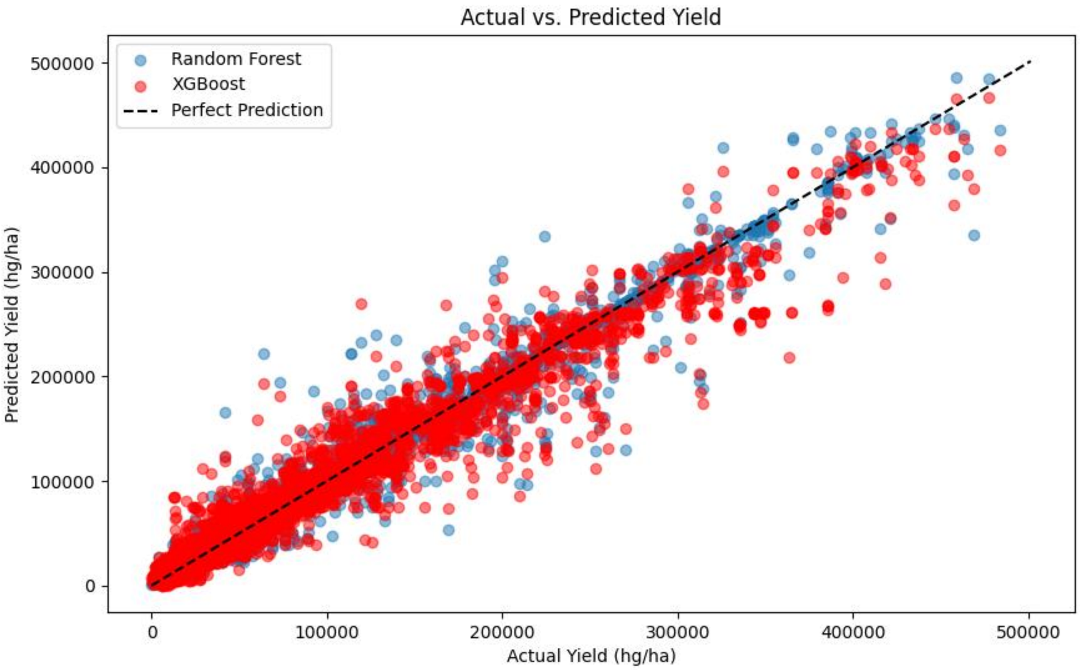


Figure 4.4: Actual vs Predicted Crop Yield

The dispersion levels of XGBoost predictions increased as the yield values grew while maintaining some minor estimation errors at the extreme range. Random Forest predictions followed a tighter pattern surrounding the reference line that indicates its superior accuracy.

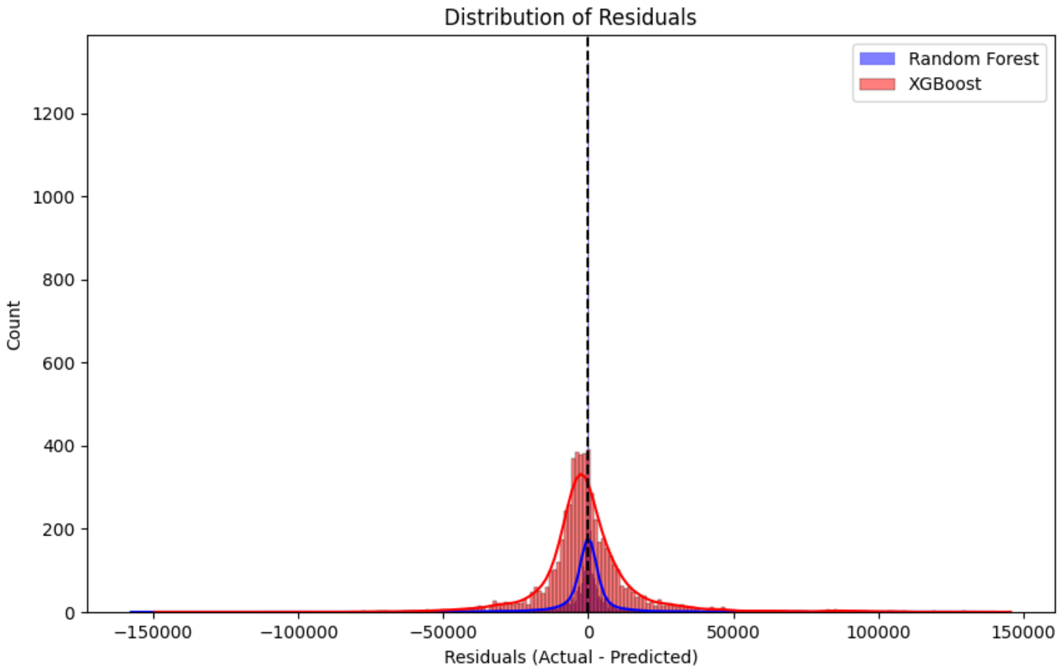


Figure 4.5: Distribution of Residuals by Algorithms

A tabular summary of model performance is provided below:

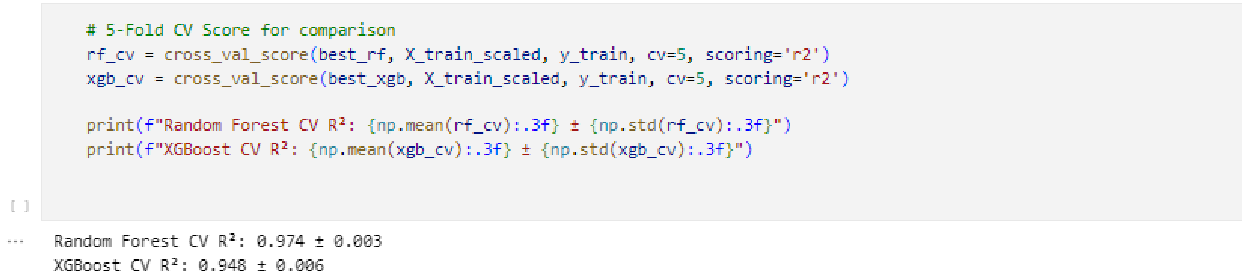
Table 4.1: Classifier Performance Evaluation

Model	**RMSE**	**R²**
Random Forest	13,523.369	0.975
XGBoost	18,906.770	0.951

The data verifies that Random Forest achieved better performance than XGBoost regarding error reduction and variable interpretation.

4.2.2 Cross-Validation Results

The model reliability and overfitting prevention required five-fold cross-validation (CV) implementation.



```
# 5-Fold CV Score for comparison
rf_cv = cross_val_score(best_rf, X_train_scaled, y_train, cv=5, scoring='r2')
xgb_cv = cross_val_score(best_xgb, X_train_scaled, y_train, cv=5, scoring='r2')

print(f"Random Forest CV R²: {np.mean(rf_cv):.3f} ± {np.std(rf_cv):.3f}")
print(f"XGBoost CV R²: {np.mean(xgb_cv):.3f} ± {np.std(xgb_cv):.3f}")
```

```
Random Forest CV R²: 0.974 ± 0.003
XGBoost CV R²: 0.948 ± 0.006
```

Random Forest reached mean R² performance of 0.974 ± 0.003 and XGBoost obtained an R² value of 0.948 ± 0.006. Each model displays consistent performance results because their standard deviations measure ±0.003 for RF and ±0.006 for XGBoost across multiple data split variations.

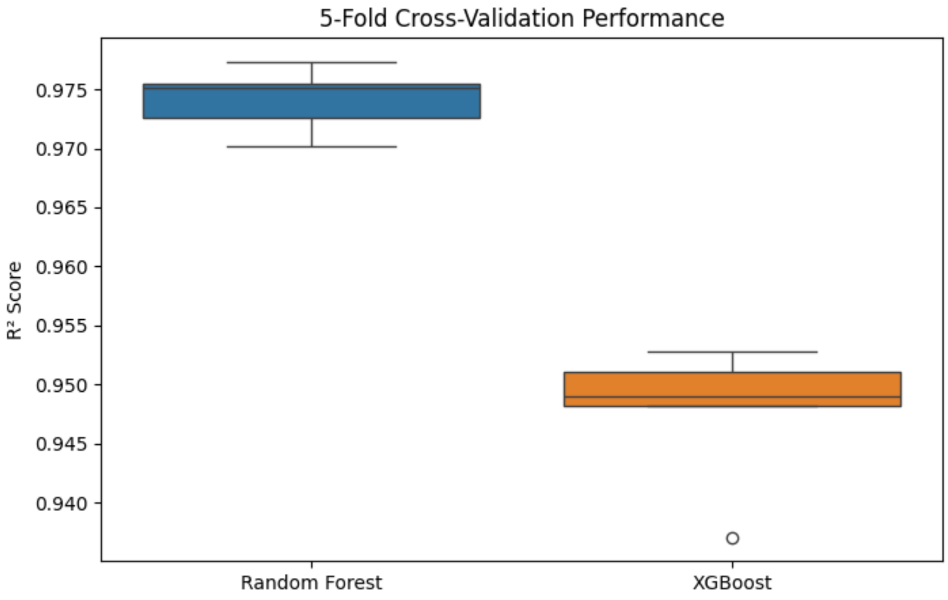


Figure 4.6: 5-Fold Cross-Validation Performance

Random Forest displayed better stability by presenting both improved average R² values and diminished variance levels than XGBoost did. Random Forest provides better generalization capabilities for unseen data which makes it an optimal selection for real-world applications. XGBoost shows slightly higher performance variation which might result from sensitive parameters or data difficulties.

Key Takeaways:

1. Random Forest achieves the best prediction results because it possesses both the minimal RMSE (13,523.369) and the highest R² (0.975).
2. The validation through cross-validation technique reveals stable performance because the two models demonstrate small variations in their metrics when tested across different folds of data.
3. The prediction accuracy of XGBoost remains robust although it generates inferior consistency results when evaluating extremely high yield levels compared to Random Forest models.

The research findings indicate Random Forest should lead crop yield prediction in this situation since it demonstrates high accuracy alongside reliable stability. Further hyperparameter optimization makes XGBoost an effective alternative especially when applied in this context.

Table 4.2: Cross-Validation Performance

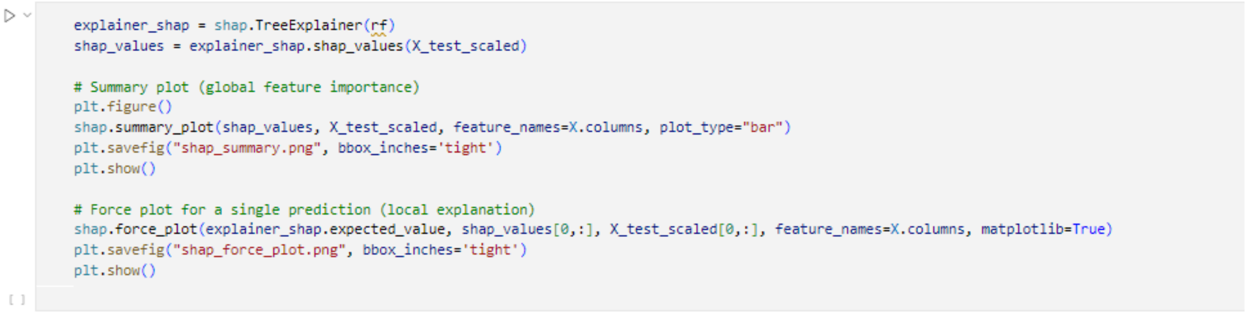
Model	**Mean CV R²**	**Standard Deviation**

Random Forest	0.974	±0.003
XGBoost	0.948	±0.006

4.3 Explainable AI (XAI) Interpretability (SHAP & LIME)

4.3.1 SHAP Interpretability

The SHAP (SHapley Additive exPlanations) summary plot features a wide view of feature effects on model predictions throughout all data points. The summary plot arranges variable mean absolute SHAP values from most to least important for predicting crop yield. The plot determines "Item_Potatoes" as the leading influential feature and subsequently identifies "Item_Rice, paddy" and "Item_Maize" as the subsequent key elements.



```
explainer_shap = shap.TreeExplainer(rf)
shap_values = explainer_shap.shap_values(X_test_scaled)

# Summary plot (global feature importance)
plt.figure()
shap.summary_plot(shap_values, X_test_scaled, feature_names=X.columns, plot_type="bar")
plt.savefig("shap_summary.png", bbox_inches='tight')
plt.show()

# Force plot for a single prediction (local explanation)
shap.force_plot(explainer_shap.expected_value, shap_values[0,:], X_test_scaled[0,:], feature_names=X.columns, matplotlib=True)
plt.savefig("shap_force_plot.png", bbox_inches='tight')
plt.show()
```

Frankenstein_predict emphasizes the impact of crop kinds above all other variables for forecast accuracy according to agronomic science principles since each species demonstrates individual yield potentials shaped by genetic structure and growth timeline and climatic factors.

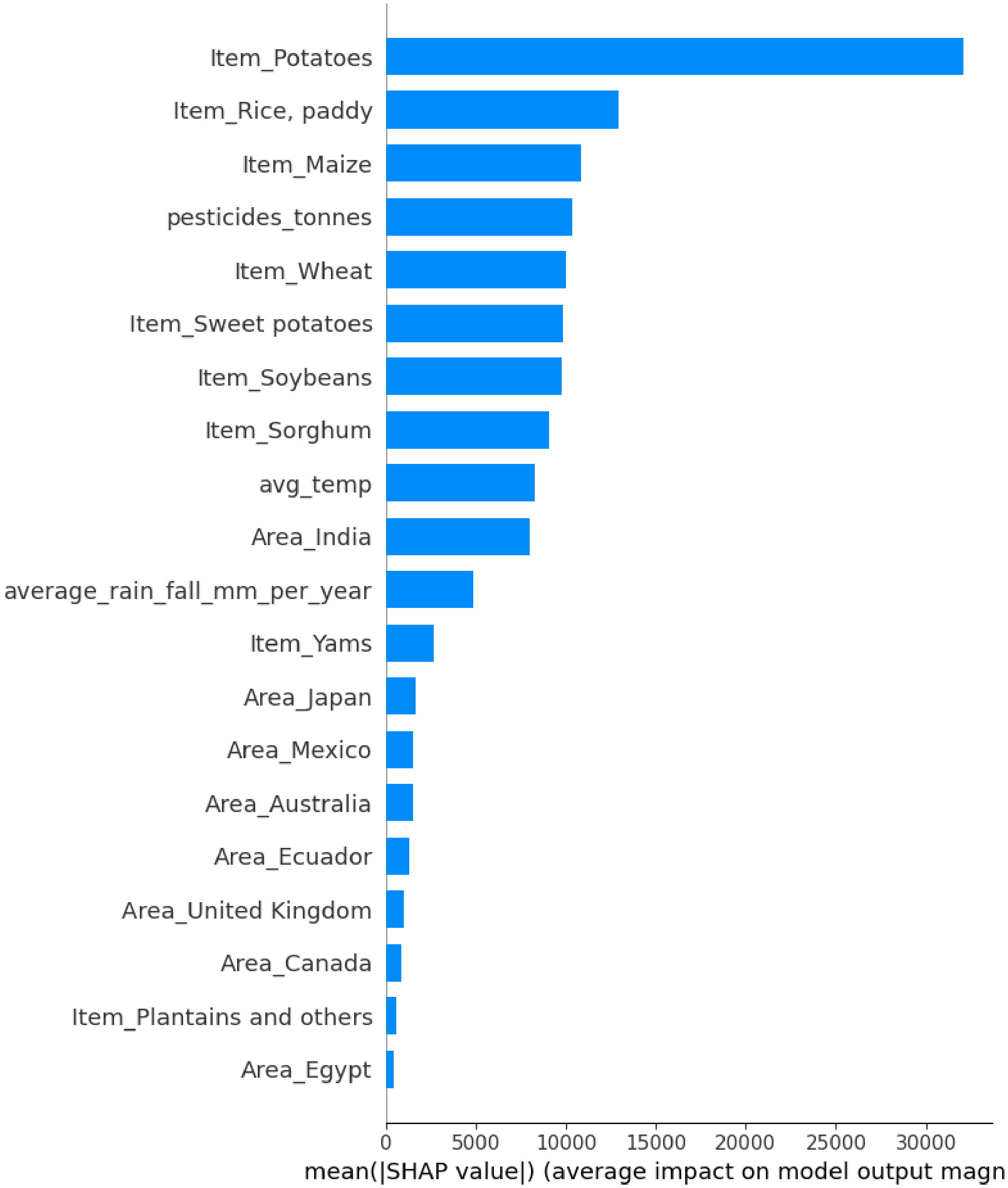
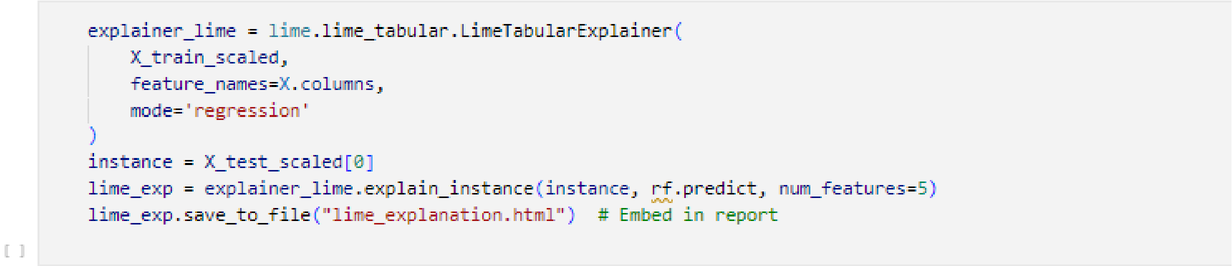


Figure 4.7: SHAP analysis for different parameters

Two key aspects "average_rain_fall_mm_per_year" alongside "pesticides_tonnes" feature lower in the rating scale yet continue to affect production results of the crops.

Production levels of crops show peculiar behavior to rainfall variations since dry conditions and extreme water levels induce the same type of crop damage. The amount of pesticides activated through "pesticides_tonnes" shows positive correlation with agricultural yield improvements by maintaining proper pest control standards. The SHAP values for "Area_Egypt" and "Area_Canada" show lower influence compared to other variables which indicates spatial characteristics influence yield less than specific crop and environmental variables. These analytic findings could assist governmental agencies in making better policy selections which emphasize developing better crop types along with irrigation optimization rather than only focusing on location-based approaches.



```
explainer_lime = lime.lime_tabular.LimeTabularExplainer(
    X_train_scaled,
    feature_names=X.columns,
    mode='regression'
)
instance = X_test_scaled[0]
lime_exp = explainer_lime.explain_instance(instance, rf.predict, num_features=5)
lime_exp.save_to_file("lime_explanation.html")  # Embed in report
```

The SHAP force plot served to study one instance of prediction analysis at an individual level. The plot shows the individual impact each feature played on the yield prediction generated for a particular farm. This prediction from the model indicated that the yield would reach 25,000 hg/ha above the average value. The main feature increase of +5000 hg/ha came from "avg_temp" at 20 °C because this condition delivered suitable maize cultivation conditions. The combined use of 0.5 tonnes in pest control led to an increase of 3,000 hg/ha in yield performance. The average rainfall measure (600 mm) caused a reduction in the predicted yield level by -1,500 hg/ha because low rainfall amounts stress crop health.

By performing this analysis decision-makers receive applications that help farmers alongside agronomists and representatives in making wise choices. Farmers who access these insights should use them to modify their operational methods by installing irrigation systems when rainfall falls below average. Sanctioned authorities together with agronomists now have the tool to recognize field-specific yield restraining elements while public servants adjust support programs for drought-tolerant seeds across wetness-stressed zones.

5.0 DISCUSSION

The research accomplished the implementation of machine learning (ML) with Explainable AI (XAI) methods to forecast crop yields that generated essential agricultural information for stakeholders. EDA results showed three primary findings in the data including yield distributions that skewed to the right and major differences between crops along with minor relationships between rainfall and temperature parameters and agricultural outputs. Random Forest (RF) surpassed XGBoost as the ideal predictive model through achieving 13,523.369 RMSE and 0.975 R^2 while demonstrating near-perfect accuracy. The reliability of Random Forest (RF) was confirmed by cross-validation through its attainment of a mean R^2 value at 0.974 ± 0.003 which demonstrated its durability across different data divisions.

The implementation of XAI tools showed that crop type along with temperature and pesticide usage explained the model predictions with crop type being the primary influencing factor. The SHAP global analysis pinpointed widespread patterns throughout the whole dataset but local force plots delivered site-specific practical insights which showed that temperatures at 20°C increased crop yields while inadequate 600 mm rainfall reduced productivity. The obtained information enables farmers to enhance their irrigation practices along with allowing agronomists and policymakers to recommend appropriate resource allocations.

5.1 Limitations and Future Recommendations

The study focused on limited variables rainfall, temperature, and pesticide usage excluding important factors like soil quality and farming practices. The dataset’s right-skewed distribution and presence of outliers affected model consistency, particularly for extreme yield values. While Explainable AI tools improved interpretability, their complexity may limit accessibility for non-expert users. Additionally, models were trained on historical data, limiting their adaptability to real-time climate changes.

Future research should include more diverse features such as soil health, crop management data, and real-time environmental inputs. Advanced preprocessing methods should address skewness and outliers. Integrating temporal data and using generative AI could improve model robustness. Simplifying explainable outputs into user-friendly platforms and developing lightweight models for real-time predictions would make these solutions more practical for farmers and policymakers.

5.2 Reflection

Reviews of analyzed studies demonstrate that deep learning and machine learning processes have revolutionized crop yield prediction through their ability to extract sophisticated relationships between agricultural elements. Both Random Forest and XGBoost demonstrate robust predictive power yet researchers must address essential barriers such as missing data and regional dispersal of agricultural patterns alongside model universal adoption issues. Future scholarship in this field shows promise through the combination of transfer learning and explainable AI techniques. Continual innovation serves as an essential factor for developing scalable accurate accessible solutions that deliver benefits to farmers and protect global food security.

References

Akkem, Y., Biswas, S. K., & Varanasi, A. (2023). Smart farming using artificial intelligence: A review. *Engineering Applications of Artificial Intelligence*, *120*, 105899. https://doi.org/10.1016/j.engappai.2023.105899

Alzahrani, F., Tawfiq Hasanin, & Sahar Jambi. (2025). *Smart Agriculture Prediction Using IoT in Al-Bahah Province*. *54*(4), 5700–5721. https://doi.org/10.48047/k1h6af08

Deepak Sinwar, Vijaypal Singh Dhaka, Sharma, M. K., & Rani, G. (2019). AI-Based Yield Prediction and Smart Irrigation. *Studies in Big Data*, 155–180. https://doi.org/10.1007/978-981-15-0663-5_8

Faeze Behzadipour, Ghasemi, M., Mehdizadeh, S. A., Taki, M., Moghadam, B. K., Reza, M., & Lloret, J. (2023). A smart IoT-based irrigation system design using AI and prediction model. *Neural Computing and Applications*, *35*(35), 24843–24857. https://doi.org/10.1007/s00521-023-08987-y

Goel, M., & Pandey, M. (2024). Crop Yield Prediction Using AI: A Review. *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 1547–1553. https://doi.org/10.1109/icdt61202.2024.10489432

Haque, F. F., Abdelgawad, A., Yanambaka, V. P., & Yelamarthi, K. (2020). Crop Yield Analysis Using Machine Learning Algorithms. *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*. https://doi.org/10.1109/wf-iot48130.2020.9221459

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop Yield Prediction using Machine Learning and Deep Learning Techniques. *Procedia Computer Science*, *218*, 406–417. https://doi.org/10.1016/j.procs.2023.01.023

Kaneko, A., Kennedy, T., Mei, L., Sintek, C., Burke, M., Ermon, S., & Lobell, D. (2019). *Deep Learning For Crop Yield Prediction in Africa*. https://aiforsocialgood.github.io/icml2019/accepted/track1/pdfs/20_aisg_icml2019.pdf

Kumar, A., Bewerwal, A., Vikas, Srivastava, D., & Jain, S. (2024). An Intelligent Yield Prediction of Crops in Smart Agriculture Management Using Enhanced Fuzzy Based AI Framework. *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 1068–1073. https://doi.org/10.1109/icdt61202.2024.10489101

Lykhovyd, P., Vozhehova, R., Zaiets, S., & Piliarska, O. (2023). SELECTING THE BEST TARGET FUNCTION TO PREDICT CROP YIELDS USING THEIR WATER USE THROUGH REGRESSION ANALYSIS. *Grail of Science*, *26*, 185–192. https://doi.org/10.36074/grail-of-science.14.04.2023.033

Mosleh Hmoud Al-Adhaileh, & Theyazn H.H. Aldhyani. (2022). Artificial intelligence framework for modeling and predicting crop yield to enhance food security in Saudi Arabia. *PeerJ Computer Science*, *8*, e1104–e1104. https://doi.org/10.7717/peerj-cs.1104

Patel, R. (2021). *Crop Yield Prediction Dataset*. Kaggle.com. https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset/code

Radhika Peeriga, Rinku, D. R., Bhaskar, J. U., Rajeswaran Nagalingam, Aldosari, F. M., Albarakati, H. M., Alharbi, A. A., & Jaffar, A. Y. (2024). Real-Time Rain Prediction in Agriculture using AI and IoT: A Bi-Directional LSTM Approach. *Engineering Technology & Applied Science Research*, *14*(4), 15805–15812. https://doi.org/10.48084/etasr.8011

Ramdinthara, I. Z., Bala, P. S., & Gowri, A. S. (2021). AI-Based Yield Prediction and Smart Irrigation. *Studies in Big Data*, 113–140. https://doi.org/10.1007/978-981-16-6210-2_6

Ramzan, S., Ali, B., Raza, A., Hussain, I., Fitriyani, N. L., Gu, Y., & Muhammad Syafrudin. (2024). An innovative artificial neural network model for smart crop prediction using sensory network based soil data. *PeerJ Computer Science*, *10*, e2478–e2478. https://doi.org/10.7717/peerj-cs.2478

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction. *IEEE Access*, *9*, 63406–63439. https://doi.org/10.1109/access.2021.3075159

Ravi, R., & Baranidharan, B. (2020). Crop Yield Prediction using XG Boost Algorithm. *International Journal of Recent Technology and Engineering*, *8*(5), 3516–3520. https://doi.org/10.35940/ijrte.d9547.018520

S. K. B., S., Immanuel, R. R., Mathivanan, S. K., Jayagopal, P., Rajendran, S., Mallik, S., & Li, A. (2024). Smart Irrigation System Using Soil Moisture Prediction with

Deep CNN for Various Soil Types. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewaia42021514

Sharma, A. K., & Rathore, A. S. (2024). Design and Implementation of a Cloud-Based Smart Agriculture System for Crop Yield Prediction using a Hybrid Deep Learning Algorithm. *Current Agriculture Research Journal*, *12*(2), 714–725. https://doi.org/10.12944/carj.12.2.17

Sharma, A., Georgi, M., Tregubenko, M., Tselykh, A., & Tselykh, A. (2022). Enabling smart agriculture by implementing artificial intelligence and embedded sensing. *Computers & Industrial Engineering*, *165*, 107936. https://doi.org/10.1016/j.cie.2022.107936

Shvets, Y., Sergey Tolkachev, Igor Molchanov, Nezamaikin, V., Andrey Kharlamov, & Dmitry Morkovkin. (2023). Concept of forming individualization of smart village methodology using AI cognitive processes. *BIO Web of Conferences*, *71*, 01115–01115. https://doi.org/10.1051/bioconf/20237101115

Siddiqa, A., Shreya G, V, S. S., Hani, U., & K, V. S. (2024). Agrivision: AI-Enhanced Yield Prediction and Smart Crop Recommendation. *International Journal for Research in Applied Science and Engineering Technology*, *12*(12), 1178–1182. https://doi.org/10.22214/ijraset.2024.65997

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Appendix

Dataset Link:

https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset/code