A Comparative Machine Learning Approach for Predicting Agricultural Productivity under Climate Variability

Priti S. Dighe[1], Santosh K. Khule[2]

[1]Assistant Professor, Department of Computer Science, Padmashri Vikhe Patil College of Arts, Science and Commerce, Pravaranagar, Email ID: priti.bawake@gmail.com
[2]Assistant Professor, Department of Computer Science, Padmashri Vikhe Patil College of Arts, Science and Commerce, Pravaranagar, Email ID: santoshkhule11@gmail.com

Abstract
Agriculture is highly dependent on climatic and environmental conditions, making agricultural productivity vulnerable to changes in temperature, rainfall, humidity, and other weather-related factors. Increasing climate variability creates difficulties for farmers and agricultural planners in estimating future crop productivity and managing resources efficiently. Conventional statistical approaches may have limitations when the relationship between climate variables and crop productivity is nonlinear and involves interactions among several factors. Machine learning provides an alternative approach for learning complex relationships from historical agricultural and climatic data.
This research proposes a comparative machine learning framework for predicting agricultural productivity under climate variability using Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Random Forest Regression (RFR), and Extreme Gradient Boosting (XGBoost) Regression. Climatic variables such as temperature, rainfall, humidity, and solar-related indicators, together with agricultural variables such as crop type, cultivated area, and historical yield, can be used as explanatory variables. Agricultural productivity is considered the target variable. The proposed framework consists of data collection, preprocessing, exploratory analysis, feature selection, model development, prediction, and performance evaluation. The models are evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). The study is designed to determine which regression approach provides the most reliable prediction while also examining the relationship between climate variability and agricultural productivity. The proposed framework can support agricultural planning, climate adaptation, resource allocation, and data-driven decision-making.

Keywords: Climate Change, Climate Variability, Agricultural Productivity, Machine Learning, Multiple Linear Regression, Decision Tree, Random Forest, XGBoost, Crop Yield Prediction.

Introduction

Agriculture plays an important role in food security, rural development, employment, and economic growth. Crop productivity, however, depends on several interacting factors, including temperature, precipitation, soil conditions, irrigation, crop characteristics, and agricultural management practices. Among these factors, climate conditions are particularly important because variations in temperature and rainfall can directly influence crop growth and final yield.
Climate change has increased the importance of understanding the relationship between climate conditions and agricultural production. Increasing temperatures, altered rainfall patterns, droughts, floods, and other climate extremes can affect crop development and productivity. The magnitude of these effects varies according to crop, region, soil, management practices, and the timing and severity of climate events.
Traditional approaches to agricultural prediction often rely on statistical models. Multiple Linear Regression, for example, can estimate the relationship between a dependent variable and several independent variables. However, agricultural systems frequently exhibit nonlinear relationships. A small increase in temperature may have little effect within an optimal temperature range but may substantially reduce productivity once a crop exceeds its heat tolerance. Machine learning techniques can capture such nonlinear patterns. This research therefore proposes a comparative framework involving Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression. The objective is to compare their predictive capabilities using the same agricultural and climatic dataset.

2. Problem Statement

Agricultural productivity is affected by changing climatic conditions, making accurate productivity prediction increasingly challenging. Temperature fluctuations, irregular rainfall, humidity variations, and other environmental changes can influence crop growth and yield.
A simple statistical model may not adequately capture nonlinear relationships between climate variables and agricultural productivity. Consequently, there is a need for a machine-learning-based approach that can analyze historical agricultural and climate data and provide reliable productivity predictions.
The main problem addressed by this research is: How effectively can machine learning regression algorithms predict agricultural productivity using climate and agricultural variables, and which of the selected algorithms provides the best predictive performance?

3. Research Objectives
To develop and evaluate a machine learning framework for predicting agricultural productivity under climate variability.
To collect historical agricultural and climate data; identify important climate and agricultural variables; preprocess and integrate the datasets; develop MLR, DTR, RFR, and XGBoost models; compare their predictive performance; identify influential variables; and determine the most suitable algorithm for the prediction task.

4. Research Questions
RQ1: Can climate and agricultural variables be used to predict agricultural productivity using machine learning?
RQ2: How does Multiple Linear Regression perform compared with tree-based regression methods?
RQ3: Does Random Forest Regression provide better predictions than Decision Tree Regression?
RQ4: Can XGBoost improve prediction performance compared with the other selected algorithms?
RQ5: Which climate variables have the greatest influence on agricultural productivity?
5. Research Hypotheses
H0: There is no significant difference in the predictive performance of Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression for agricultural productivity prediction.
H1: There is a significant difference in the predictive performance of at least one of the four machine learning regression algorithms.
H2: Climate variables such as temperature and rainfall significantly contribute to the prediction of agricultural productivity.

6. Literature Review
Machine learning has become an important approach for agricultural prediction because agricultural datasets often contain numerous environmental, climatic, and production-related variables. Crop-yield prediction studies commonly use weather, soil, vegetation, and management-related features.
Recent reviews report frequent use of Linear Regression, Random Forest, Gradient Boosting, Decision Tree, and other machine learning approaches for crop-yield prediction. Evaluation commonly uses RMSE, MAE, and R^2.
Climate-impact studies have found that temperature changes can negatively affect yields of several major crops, while the magnitude and direction of effects depend on crop and location. These findings support combining climate variables with agricultural variables in predictive models.
Random Forest and XGBoost are particularly suitable for this study because ensemble tree methods can model nonlinear relationships and interactions among predictors. Nevertheless, the best algorithm should be established empirically rather than assumed in advance.

7. Proposed Research Methodology
The proposed methodology consists of data collection, data integration, preprocessing, exploratory data analysis, feature selection, train-test splitting, model development, evaluation, and comparative analysis.
Data Collection → Climate and Agricultural Data → Data Integration → Data Preprocessing → Exploratory Data Analysis → Feature Selection → Train-Test Split → MLR / DTR / RFR / XGBoost → Model Evaluation → Model Comparison → Best Prediction Model.

8. Dataset
A suitable implementation can combine agricultural production or yield information with climate observations. FAOSTAT is a suitable public source for agricultural production, harvested area, and yield information. Climate observations may be obtained from a reliable meteorological or climate-data source.
Suggested input variables include Year, Crop Type, Temperature, Rainfall, Humidity, Solar Radiation, Soil Moisture, Cultivated Area, Previous Yield, and Irrigation information where available.
The target variable is Agricultural Productivity or Crop Yield, for example measured in tonnes per hectare. The exact variables should depend on the final dataset.

9. Data Preprocessing
Raw agricultural and climate data may contain missing values, inconsistent units, duplicate records, outliers, and categorical variables. Preprocessing is therefore required before model training.
Missing numerical values may be handled using appropriate imputation or interpolation. Duplicate records should be removed. Categorical variables such as crop type or region can be encoded numerically. Extreme observations should be investigated to distinguish genuine climate events from data errors.
Feature scaling can be applied where required. Tree-based models generally do not require scaling, whereas scaling can be incorporated for models that are sensitive to feature magnitude.

10. Exploratory Data Analysis
Exploratory Data Analysis is used to understand the dataset before modeling. Recommended analyses include productivity distributions, temperature-versus-productivity plots, rainfall-versus-productivity plots, humidity-versus-productivity plots, correlation analysis, year-wise productivity trends, and climate-variable trends.
Correlation analysis can provide an initial indication of relationships between individual variables and productivity.

However, correlation alone does not establish causation and may not capture nonlinear relationships.

11. Feature Selection

Feature selection is used to identify variables that contribute meaningfully to productivity prediction. Possible methods include correlation analysis, Random Forest feature importance, XGBoost feature importance, Recursive Feature Elimination, and domain knowledge.

Potential features include temperature, rainfall, humidity, soil moisture, cultivated area, previous yield, and crop type. The final feature set should be determined from the actual dataset.

12. Machine Learning Algorithms

12.1 Multiple Linear Regression

Multiple Linear Regression is used as the baseline model. Its general form is $Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta nXn + \varepsilon$, where Y is agricultural productivity, Xi are predictor variables, βi are coefficients, β0 is the intercept, and ε is the error term. MLR provides an interpretable reference for comparing nonlinear models.

12.2 Decision Tree Regression

Decision Tree Regression recursively divides observations according to predictor values and predicts the target from the resulting regions. It can capture nonlinear relationships and interactions. Important hyperparameters include max_depth, min_samples_split, and min_samples_leaf.

12.3 Random Forest Regression

Random Forest Regression combines predictions from multiple decision trees trained using different samples and feature subsets. It can model nonlinear relationships, is relatively robust, and can provide feature-importance estimates. Important hyperparameters include n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features.

12.4 XGBoost Regression

XGBoost, or Extreme Gradient Boosting, builds decision trees sequentially so that later trees improve errors made by earlier trees. It is suitable for structured agricultural data with nonlinear relationships and interactions. Important hyperparameters include n_estimators, learning_rate, max_depth, subsample, colsample_bytree, reg_alpha, and reg_lambda.

13. Model Training

The dataset can be divided into training and testing portions. A common starting point is 80% training and 20% testing. For a time-dependent agricultural prediction problem, however, a chronological split is preferable because it more realistically represents learning from historical observations and predicting later observations.

Where sufficient data are available, k-fold cross-validation can be used on the training data for model and hyperparameter selection. The test set should remain untouched until final evaluation.

14. Performance Evaluation

Mean Absolute Error (MAE) = $(1/n)\ \Sigma\ |yi - \hat{y}i|$. MAE represents the average absolute prediction error; lower values indicate better performance.

Mean Squared Error (MSE) = $(1/n)\ \Sigma\ (yi - \hat{y}i)^2$. MSE gives greater weight to large errors; lower values indicate better performance.

Root Mean Squared Error (RMSE) = $\sqrt{[(1/n)\ \Sigma\ (yi - \hat{y}i)^2]}$. RMSE is expressed in the same unit as the target variable; lower values indicate better prediction.

$R^2 = 1 - [\Sigma(yi - \hat{y}i)^2 / \Sigma(yi - \bar{y})^2]$. R^2 measures the proportion of target variation explained by the model; higher values indicate better performance.

15. Proposed Experimental Comparison

The four algorithms should be evaluated using the same data split and evaluation metrics. The final table should contain the actual MAE, MSE, RMSE, and R^2 values obtained from the experiments.

Model-selection rule: the preferred model should generally have low MAE, MSE, and RMSE and high R^2. The final winner should not be declared until the actual dataset has been processed and the models have been evaluated.

16. Results and Discussion

Multiple Linear Regression provides an interpretable baseline and can indicate whether the productivity relationship is approximately linear. Its performance may be limited when climate-productivity relationships are nonlinear.

Decision Tree Regression can represent nonlinear relationships and variable interactions, but an unconstrained tree may overfit. Random Forest reduces some of this instability by aggregating many trees and can also provide feature-importance information.

XGBoost builds trees sequentially to improve previous errors and can model complex nonlinear relationships. It may perform strongly on structured agricultural datasets, but its performance depends on data quality and hyperparameter tuning.

The actual numerical results, feature rankings, and statistical comparisons must be calculated from the selected dataset. No accuracy or error values should be invented.

17. Comparative Analysis
MLR: simple and interpretable, but assumes a linear relationship. DTR: captures nonlinear relationships, but can overfit. RFR: robust ensemble approach and useful for feature importance, but less interpretable than MLR. XGBoost: powerful nonlinear ensemble method, but requires careful tuning.
The comparison will determine whether greater model complexity actually improves predictive accuracy for the selected agricultural dataset.

18. Feature Importance Analysis
Feature importance should be calculated after model training. Random Forest can provide impurity-based importance, while XGBoost can provide model-based importance measures. For more robust interpretation, SHAP values can also be considered.
A final feature-importance table should contain the actual ranking and importance values generated from the selected dataset. The ranking should not be predetermined.

19. Advantages of the Proposed System
The framework combines agricultural and climatic information; compares conventional and ensemble machine learning approaches; models nonlinear relationships; provides quantitative productivity predictions; identifies influential variables; supports agricultural planning; and provides an objective basis for comparing algorithms.

20. Limitations
The quality of predictions depends on the availability and reliability of historical data. A model trained for one geographic region may not generalize to another. Important variables such as fertilizer application, irrigation, crop varieties, planting dates, and pest attacks may be unavailable. Extreme climate events may be poorly represented in historical data. Ensemble models are also less straightforward to interpret than linear regression.
Machine learning predictions should not automatically be interpreted as causal evidence. Establishing causal climate impacts requires appropriate research designs and controls.

21. Future Scope
Future research can investigate Artificial Neural Networks, LSTM, CNN, and transformer-based models. Satellite-derived vegetation indicators such as NDVI, EVI, and LAI can be incorporated. Explainable AI methods such as SHAP can be used to explain individual predictions. Future climate scenarios can be incorporated to estimate productivity under different warming and precipitation conditions. Regional models and real-time IoT-based prediction systems are additional possibilities.

22. Conclusion

This research proposes a comparative machine learning framework for predicting agricultural productivity under climate variability. The framework combines historical agricultural information with climate-related variables and evaluates Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression.
MLR provides an interpretable baseline, while DTR captures nonlinear relationships. RFR improves robustness by combining multiple decision trees, and XGBoost uses gradient boosting to progressively reduce prediction errors. The proposed methodology enables objective comparison using MAE, MSE, RMSE, and R^2.
The framework can contribute to data-driven agricultural planning by estimating productivity under varying climatic conditions. Future work can strengthen the approach through remote sensing, soil and irrigation data, explainable AI, and future climate scenarios.

References

1. Hu, T., Zhang, X., Khanal, S., & Wilson, R. (2024). Climate change impacts on crop yields: A review of empirical findings, statistical crop models, and machine learning methods. Environmental Modelling & Software, 179, 106119.
2. Crop yield prediction using machine learning: An extensive and systematic literature review. Smart Agricultural Technology, 10, 100718, 2025.
3. Jabed, M. A., & Murad, M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. Heliyon, 10(24), e40836.
4. Machine learning approaches for crop yield prediction: A systematic literature review. Remote Sensing Applications: Society and Environment, 43, 102151, 2026.
5. Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., & Srinivasan, K. (2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. Computers and Electronics in Agriculture, 155, 257–282.
6. Food and Agriculture Organization of the United Nations. FAOSTAT: Food and Agriculture Data.