Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

Research Article

An Experimental Analysis of Traditional Machine Learning Algorithms for Maize Yield Prediction

Souand P.G. Tahi[1*], Castro G. Hounmenou[1], Vinasetan Ratheil Houndji[1,2], Romain Glèlè Kakaï[1]

[1] Laboratory of Biomathematics and Forest Estimations, Faculty of Agronomic Sciences, University of Abomey-Calavi, 04 BP 1525 Cotonou, Benin
[2] Institute of Training and Research in Computer Science, University of Abomey-Calavi, 01 BP 526 Cotonou, Benin
E-mail: souandtahi@gmail.com

Received: 19 February 2024; **Revised:** 29 March 2024; **Accepted:** 16 April 2024

Abstract: Maize plays a significant role in the African diet and is one of the main staple foods in many parts of the continent. Accurate yield estimations ensure an adequate food supply, contributing to food security and reducing the risk of food shortages. They also enable market planning and price setting. Machine learning is well known as one of the most advanced statistical methods for predicting crop yields. This paper provides extensive experiment results of machine-learning models on maize production. Thirteen basic supervised learning algorithms classified into classic and ensemble learning are compared using three datasets of different sizes and from various sources (Kaggle, Zenodo). These datasets are from three main origins: experimentation, specifically covering crop data with 240 observations; predictions on crop yield from the FAO (Food and Agriculture Organization) and World Data Bank with 4,121 observations; and historical data from China with 975 observations. The metrics used to evaluate the models are the coefficient of determination, the mean absolute error, the root mean square error, and the explained variance score. Moreover, permutation importance is used on the best models to identify the most relevant predictors for the models according to the data. The results show that extremely randomized trees (ERT) and extreme gradient boosting (XGBoost) are more suitable for predicting maize yield with a coefficient of determination between 0.75 and 0.96 and 0.73 and 0.96, respectively. With the other metrics, the ERT model shows a low performance. Its training time varies between 2,547 and 7,814 seconds as obtained from a computer with characteristics of HP core i5, CPU @ 1.00 GHz, 1.9 GHz, and 8 GB RAM under 134 Windows 10. ERT and XGBoost are best suited to these databases of varying dimensions, making them perfect for predicting maize yield and streamlining decision-making processes.

***Keywords*:** classic machine learning, ensemble learning, maize, yield prediction, secondary data

MSC: 62P12, 68T05

1. Introduction

Maize is a vital crop in various regions (Africa, Latin America, and some Asian countries) [1]. It is a source of phytochemicals and is the world's third most cultivated and consumed cereal after wheat and rice [2]. Its annual production is estimated at 1,162.35 million tons, achieved through average productivity of 5.75 tons per hectare [3]. It has a lot of

DOI: https://doi.org/10.37256/cm.5420244481

calories and is a rich source of fiber, protein, and carbs [4]. This propriety makes it a special player in the fight against food insecurity. According to Tech et al. [5], 9.7 billion people will inhabit the planet by 2050. This strong growth leads to a high request for higher agricultural production. To meet this demand, productivity and production methods must be improved. According to Singh [6], traditional agriculture confronts several difficulties, including floods, droughts, crop illnesses, storage, etc. These challenges need to be addressed to ensure the sustainability of agriculture. Precision agriculture can manage the challenges to reduce environmental impact and maximize yields. Machine learning (ML) is a promising strategy that enables machines to learn and develop without being explicitly programmed; it is essential to precision farming [7]. Rao et al. [8] claim that ML opens up new possibilities for big data research in many agricultural sectors. It might be advantageous for managing crops, caring for animals, maintaining the health of the soil, managing water, etc. It is a method that promises to guarantee sustainable agriculture and food security in an expanding global community. Nowadays, accurately predicting crop yields is one of the key components of sustainable agriculture.

Various machine-learning models have been used to predict crop yields. However, the accuracy of these models depends on multiple factors such as input variables, the number of observations, and hyperparameters. In a recent study, Ruan et al. [9] used proximal sensing and meteorological data to create an in-season wheat yield forecast model at the field scale. They used two feature selection techniques and eleven statistical and machine learning (ML) regression algorithms and concluded that RF and XGBoost had the best overall performance ($R^2 = 0.74 \sim 0.78$). To estimate maize yield, Sarijaloo et al. [10] employed a variety of models, including neural networks, gradient boosting machines, random forests, adaptive boosting, XGBoost, and decision trees. It was discovered that the XGBoost model could accurately predict maize yield. Ahmad et al. [11] established a model for predicting maize yields that considered temporal variance in maize yields. He named the Normalized Difference Vegetation Index (NDVI) and Land Surface Temperature (L.S.T) variables as essential remote sensing-derived inputs. A study by Dhaliwal et al. [12] compared different models such as Partial least squares regression (PLS), Multivariate adaptive regression splines (MARS), multiple linear regression (LR), Regularized regression, and Random forest (RF) to understand the temporal and spatial heterogeneities in sweet maize yield. The input variables considered included time components, spatial components, genetic factors, crop management practices, and weather and soil parameters. Random Forest (RF) provided the best predictions with a lower root mean square error (RMSE = 3.29 Mt/ha). To assess the efficacy of various input variables in predicting yield, Meng et al. [13] forecast maize yield at the plot level from 1994 to 2007 using various data sources, including monthly climate, satellite data, and soil data. The results show that using all its data sets with random forests (RF) and AB (adaptive boosting) can achieve better yield prediction performance ($R^2 = 0.85 \sim 0.98$). Reddy [14], Kalimuthu et al. [15], and Abbas [16] have all made accurate crop yield predictions using machine learning techniques.

This paper presents an extensive experiment of machine learning models for maize production. We evaluate the adaptability of ML techniques for yield prediction and determine the most convincing technique for maize yield prediction.

The study compares the performance of various machine learning techniques to demonstrate their potential for maize yield prediction. The rest of the paper is organized as follows. Section 2 presents the methodology adopted for data collection, pre-processing, analysis, and model evaluation criteria. The results are presented and discussed in Section 3, and Section 4 concludes.

2. Methodology

This session describes the methodology used to evaluate machine learning models for maize production. It covers the dataset used, the pre-processing steps, the analysis techniques, the metrics used for model evaluation, and predictor importance analysis.

2.1 *Data collection*

For data collection, keywords such as “maize yield prediction”, “corn yield prediction”, “corn dataset”, and “maize datasets” were used to search for datasets from Kaggle and Zenodo repositories. Only datasets containing maize yield variables or maize productivity were considered. Three different agricultural datasets containing information about maize

yield were collected. The collected data included experimental data from Mexico, historical data from 91 countries worldwide from 1990 to 2013 obtained from sources like the Food and Agriculture Organization (FAO) and the World Bank, and historical data concerning environmental parameters, pollution, and maize yield in China. The datasets are summarized in Table 1.

Table 1. Datasets summary

Dataset	Variables	Description of variables used	Data size	Provenance
Crop yield prediction	Areas, year, pesticide, temperature, rainfall	Year, pesticide, temperature, rainfall	4121	Kaggle https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset
Cover crop and irrigation impacts on weeds and maize yield	Year, site, plot, irrigations, total amount water, treatment, cover crop biomass, weed biomass, number of corn cobs, weight of corn cobs weed biomass at critical period, weed count before corn harvest	Year, site, plot, irrigations, total amount of water, treatment, cover crop biomass harvested, weed biomass produced during cover crop growing period, number of corn cobs per two 1 meter rows in the weedy subplot, weight of corn cobs, weed biomass at critical period, weed count before harvest.	240	Zenodo https://zenodo.org/record/5905378
Marked impacts of pollution mitigation on crop yields in China	Year, province, temperature 2m at plant, temperature 2m growing season, temperature 2m harvest season, precipitation plant season, precipitation growing season, precipitation harvest season, aerosol optical depth for plant season, aerosol optical depth for growing season,aerosol optical depth for harvest, surface ozone for plant season, surface ozone for growing season, surface ozone for harvest, maize yield.	Year, province, temperature 2m at plant, temperature 2m growing season, temperature 2m harvest season, precipitation plant season, precipitation growing season, precipitation harvest s eason, aerosol optical depth for plant season, aerosol optical depth for growing season, aerosol optical depth for harvest, surface ozone for plant season, surface ozone for growing season, surface ozone for harvest, maize yield.	975	Zenodo https://zenodo.org/record/7232790

The database input variables are grouped into climatic, edaphic, water stress, irrigation, pesticide factors, treatment, and collection year, with data sizes of 240, 975 and 4,121 entries. The "Marked Impacts of Pollution Mitigation on Crop Yields in China" database contains historical data from various Chinese provinces from 1980 to 2018. It includes climatic parameters and pollution factors collected in different provinces at different stages of maize production (planting, growing, and harvest). The "Cover crop and irrigation impacts on weeds and maize yield" data comes from an experiment that evaluated the water requirements to produce a Winter Cover (barley, Austrian winter pea, and mustard) with sufficient biomass for weed suppression during maize growth at two New Mexico sites. Predictor categories included cover crop type, irrigation, weed quantity, and water stress characteristics. The "Crop yield prediction" data is derived from World Bank and FAO websites in several countries between 1990 and 2016. It includes pesticide, climate, and yield variables. The yield and pesticide variables exhibit very little dispersion around the mean (Figure 1a). Similar observations were noted for the variables weed count before corn and weed biomass for the growing season" in the dataset "impact of cover crops and irrigation on weeds and maize yield" (Figure 1b). Regarding the dataset "Market impact of pollution mitigation on crop yield in China", most variables showed minimal dispersion around the mean (Figure 1c). For a comprehensive analysis of the various variables, descriptive statistics of the data were presented in Table 4 in the supplementary file.

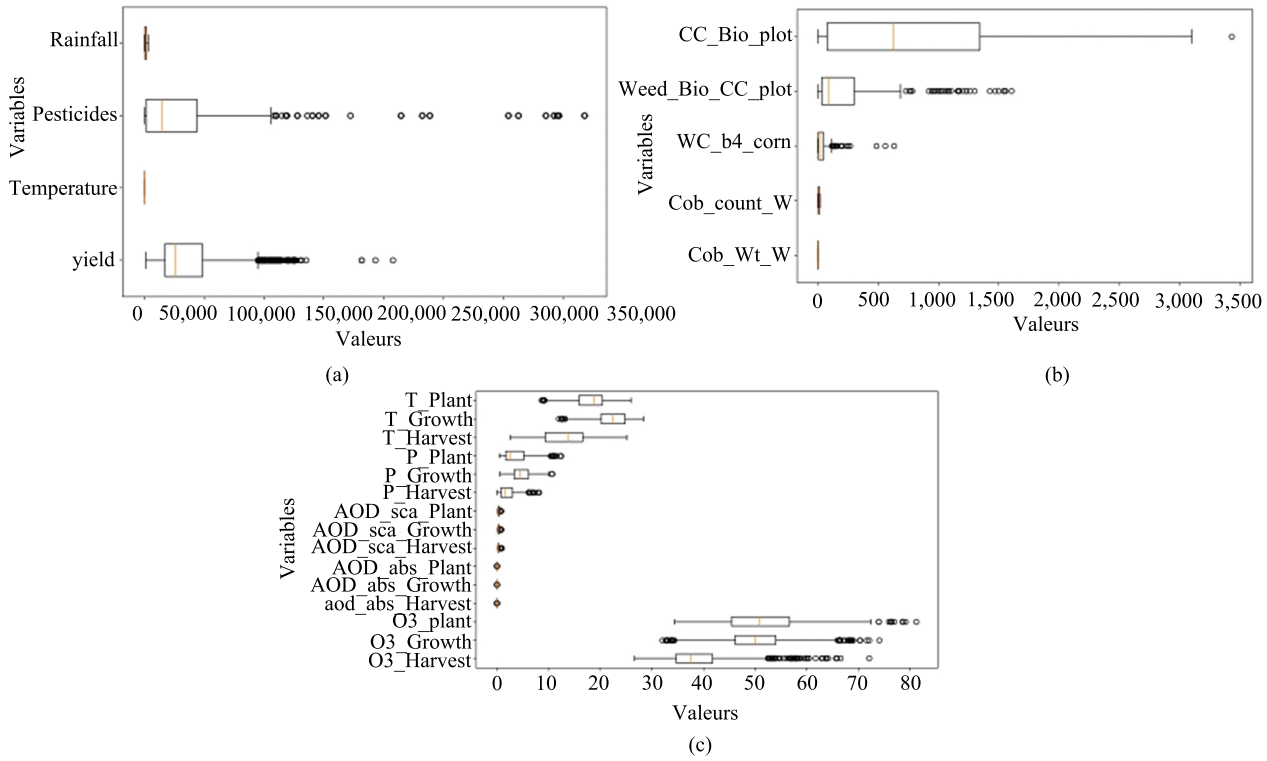


Figure 1. Variable distribution. (a): Crop yield prediction importance, (b): Cover crop and irrigation impacts on weeds and maize yield; (c): Marked impacts of pollution mitigation on crop yields in China

2.2 *Data pre-processing*

The data collected were subjected to a complete pre-processing activity before being used for modeling. Firstly, missing data was removed from the various datasets. Based on the correlation matrices (Tables 1-3 in supplementary data) the input variables were selected using a selection threshold of 0.8 in absolute value. A correlation of 80% or more suggests that the retained variable already captures a significant portion of the information in the removed variable. This approach reduced data redundancy and allowed feature selection for modeling. Afterward, the outliers were eliminated with interquartile range (IQR) after identifying them with the boxplot, encoded categorical variables into binary variables, normalized input variables with the *MinimaxScaler* technique on the *scikit-learn* library, and partitioned data into training and test sets. Seventy percent 70% of the data within each dataset was used to train models, and the remaining 30% to test their performance. This pre-processing ensures data quality, reliability, and compatibility, preparing data for accurate and robust maize yield prediction modeling. However, before data pre-processing, the yield variables of the various data are adjusted to the kilogram to facilitate the interpretation of the results.

2.3 *Models used*

Crop yield prediction involves the use of supervised learning techniques. In this work, thirteen (13) basic supervised machine-learning models categorized into two groups were applied to the three datasets to forecast maize production (Table 2). These models included classical learning "support vector machines (SVM), K-nearest neighbors (KNN), multiple linear regressions (LR), ridge regressions (RR), least absolute shrinkage and selection operations (LASSO), decision trees (DT)" and ensemble learning "adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), gradient boosting regressions (GBR), light gradient-boosting machines (light GBM), extremely randomized trees (ERT), random forest (RF), and Bagging regression (BR)". These models are implemented using the scikit-learn package with

Python 3.9.13. The description of the models used is summarized in Table 2. The best models were obtained after hyperparameter optimization, which involved using grid search to optimize the hyperparameters defined in the models. The selected hyperparameters for each model are also presented in the Table 2. The scatter plots were performed under the Matplotlib library to appreciate the relationship between actual and predicted variables. Then, the best four models derived from the datasets were represented. The best model obtained for each dataset is used to identify the top 10 essential variables contributing to maze yield prediction. To achieve this, we employ the variable importance permutation technique. This method assesses an error's impact by permuting a given feature's values. If the permutation of values results in a significant change in error, it signifies the importance of that feature for the model. Moreover, the accumulated local effects of "alibi" packages also evaluated the predictor impact.

2.4 *Model evaluation and analysis of predictor importance*

Various metrics were used to assess the models' performance and identify the best one for maize prediction. These metrics examine the models' final performance by contrasting expected and actual results. They include the coefficient of determination (R-square), mean absolute error (MAE), root mean square error (RMSE), and explained variance score (EVS). Models were also evaluated using the execution time function (train time) and predicted time (computing time)in Python. Analyses were performed on a computer with characteristics of HP core i5, CPU @ 1.00 GHz, 1.9 GHz, and 8 GB RAM under Windows 10. The coefficient of adjustment is the proportion of the dependent variable the model explains, and the mean absolute error (MAE) is the mean distance between model predictions and actual values. The RMSE calculates the standard deviation of the residuals. EVS is the variation of the model's output for which the predictors can account. We presented below the mathematical expressions for these measures in which n is the number of observations, y_i is the actual maize yield, and $\hat{y}_i$ is the predicted maize yield. $Var(y)$ is the variance of prediction errors of actual maize yield values. The EVS and Rsquare metrics are best when the score is close to 1, while the other metrics are best when their score tends toward 0. The best model identified from each dataset was used to determine the important variables through the permutation technique. The method consists of measuring the effect of permuting the values of a variable on the model's performance. It quantifies the contribution of each variable to model prediction and selects the most informative variables to improve performance. The graphical representation of the top 10 most important variables was presented, highlighting their significance for the models.

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

$$MAE = \frac{\sum(|y_i - \hat{y}_i|)}{n}, \tag{2}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3}$$

$$EVS = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{Var(y)}. \tag{4}$$

Table 2. Model descriptions

Categories	Algorithms	Description	Hyperparameters	hyperparameters values tested
Classical learning	Support vector regression	SVM looks for an optimal hyperplane in a multidimensional space that can efficiently separate data points of different classes. The hyperplane is chosen to maximize the margin between the nearest points of each class, thus enabling better differentiation between types in the feature space. The size of the hyperplane depends on the number of features used for classification.	C: Regularization Parameter; Gamma: gamma parameter; Kernel: kernel functions	C': 0.933, 0.0001, 0.01, 0.1, 0.5, 1, 10, 100, 1000; gamma: 1, 0.1, 0.01, 0.001, 0.0001, 0.5: kernel: linear', sigmoid.
	KNN	The K-nearest neighbors (KNN) algorithm determines the probability of the test data belonging to specific classes by considering the 'K' nearest training data points. The class with the highest probability is chosen. For regression problems, the predicted value is the mean of the 'K' selected training points.	K: number of neighbors; Weights: weight parameters	n_neighbors: 20 to 30, weights: uniform, distance
	Linear regression	A statistical technique is used to model the relationship between a dependent variable and one or more independent variables using a linear equation. It estimates the value of the dependent variable as a function of the independent variables, assuming a linear relationship between them.		
	Ridge Regression	RR is a model-fitting method used to analyze all data with multicollinearity, performing L2 regularization. In multicollinearity, least squares are unbiased, and variances are oversized, resulting in a significant difference between predicted and actual values.	Alpha: Constant that multiplies the L2 term	alpha: 0.01,0.02,0.03,0.04,0.0001, 0.0003, 0.001, 0.05, 0.09, 0.1, 0.3, 1, 2, 3, 4, 5
	LASSO	Lasso is a popular type of regularized linear regression that includes an L1 penalty. It reduces the coefficients of input variables that contribute almost nothing to prediction. This penalty allows certain coefficient values to reach zero, effectively eliminating input variables from the model and achieving a kind of automatic feature selection.	Alpha: Constant that multiplies the L1 term	alpha: 0.01,0.02,0.03,0.04,0.0001, 0.0003, 0.001, 0.05, 0.09, 0.1, 0.3, 1, 2, 3, 4, 5
Ensemble learning	Decision tree	DT is a technique that uses a tree structure to make decisions based on data features. It recursively partitions the feature space into smaller subsets using criteria such as the Gini index or entropy to build a sequence of hierarchical decision rules for predicting the class of a new. example.	max_depth: maximum depth of the tree; min_sample_leaf: minimum number of samples required to be at a leaf node; max_feature: number of features to consider when looking for the best split	max_depth: min_sample_leaf: max_feature:
	Gradient Boosting Regression	GBR adjusts the weights of the training examples iteratively, focusing on the residuals of previous predictions. At each stage, a new model is added to predict the remaining residuals and the previous models' predictions are combined to form the final prediction. This gradually improves prediction accuracy by reducing residual errors and capturing complex relationships between input and target variables.	learning_rate: Weight applied to each classifier at each boosting iteration; Subsample: fraction of samples to be used for fitting the individual base learners; n_estimators: maximum number of estimators; max_depth: maximum depth of the tree	learning_rate: 0.01,0.02,0.03,0.04, 0.0001, 0.0003, 0.001, 0.05, 0.09, 0.1, 0.3, 1; subsample: 0.5, 0.7, 1.0, 0.9, 0.5, 0.2, 0.1; n_estimators: 1500, 1, 5, 10, 100, 500, 1000; max_depth: 3, 5, 7, 8, 19, 20,4,6,8,10

Table 2. (cont.)

Categories	Algorithms	Description	Hyperparameters	hyperparameters values tested
Ensemble learning	AdaBoost	AdaBoost combines several classifiers to increase accuracy. The AdaBoost classifier builds a strong classifier by combining several poorly performing classifiers to obtain a strong, high-precision classifier. It defines the weights of the classifiers and trains the data sample at each iteration to guarantee accurate predictions of unusual observations. Any machine learning algorithm can be used as a base classifier if it accepts weights on the training set.	n_estimators: maximum number of estimators; learning_rate: Weight applied to each classifier at each boosting iteration	n_estimators:1, 5, 10, 100, 1000, 200, 500, learning_rate: 0.01,0.02,0.03,0.04, 0.0001, 0.0003, 0.001, 0.05, 0.09, 0.1, 0.3, 1
	XGBoost	XGBoost (eXtreme Gradient Boosting) is a machine learning algorithm that combines decision trees sequentially to create a powerful predictive model. It uses gradient boosting with specific optimization methods to improve model accuracy and efficiently handle large datasets	min_child_weigth: minimum weighted fraction of the sum total of weights; n_estimators: maximum number of estimators; gamma: regulation parameter which controls the minimum reduction in loss	min_child_weight 1, 5, 8,10, learning_rate:0.01,0.02,0.03,0.04, 0.0001, 0.0003, 0.001, 0.05, 0.09, 0.1, 0.3, 1, n_estimators":1, 5, 10, 100, 1000, 200, 500, gamma: 0.5, 1, 1.5, 2, 5
	Light GBM	Gradient-boosting framework based on decision trees to increase model efficiency and reduce memory usage. Light GBM (Light Gradient Boosting Machine) is a popular open-source framework for gradient boosting. It is designed to handle large-scale data sets faster than other popular gradient-boosting frameworks, such as XGBoost and CatBoost.	learning_rate: Weight applied to each classifier at each boosting iteration; n_estimators: maximum number of estimators; n_leave: number of leaves	min_child_weight: 1, 5, 8,10; learning_rate:0.01,0.02,0.03,0.04, 0.0001, 0.0003, 0.001, 0.05, 0.09, 0.1, 0.3, 1 n_estimators:1, 5, 10, 100, 1000, 200, 500, n_leaves:1, 3, 4, 5, 31
	Random Forest	Combination of predictive trees. Each tree is built by selecting sets of random variables and data samples. Robust against noise. RF efficiently handles high-dimensional data sets, and avoids overfitting.	n_estimators: maximum number of estimators, 'max_depth': Maximum number of levels in tree, 'max_features': Number of features to consider at every split	n_estimators: 1, 5, 10, 100, 1000, 200, 500, max_features: 1 to 10, max_depth: 1 to10
	Extremely randomized tree	ERT adds more trees to dataset subsamples and uses majority voting. This strategy reduces variance. The technique determines the optimum threshold as a splitting rule by applying random thresholds for each feature in the subsamples.	n_estimators: maximum number of estimators, 'max_depth': Maximum number of levels in tree	n_estimators: 1500, 1, 5, 10, 100, 500, 1000, max_depth: 3, 5, 7, 8, 19, 20,4,6,8,10
	Bagging regression	Bagging combines the results of many learners to improve performance. It splits the training set into subsets and subjects them to machine-learning models. It combines their predictions when they generate a global forecast for each instance of the original data.	n_estimators: maximum number of estimators,	n_estimators: 1, 5, 10, 100, 1000, 200, 500,

3. Results and discussion

Three datasets with various input variables were subjected to applying a total of thirteen machine learning techniques. The metrics derived from the models built are listed in Table 3 while optimum hyperparameters of models for each dataset are presented in Table 4. The coefficient of determination exceeded 90% for "Cover crop and irrigation impacts" data for models LM, ERT, XGBoost, RR, and LASSO. In addition, the ERT model recorded low values for the MAE and RMSE metrics. It better described the variance score of maize yield than the other models. However, its training time was significant. Furthermore, in the context of this dataset, the XGBoost and RR models demonstrated stronger predictive performance, yielding R-squared coefficients of 0.925 and 0.920, respectively. The observed accuracy of the RR model using 'cover crop and irrigation impacts' data attests to its robustness as a regularized linear regression model, thanks to the incorporation of regularization that penalizes the coefficients of independent variables. This precise and reliable methodology provides an explicit approach to preventing overfitting and maintaining model stability. The findings align with previous research conducted by Qin et al. [17], which explored LR, RR, LASSO, and GBR techniques for predicting the economically optimal nitrogen rate in maize production. In addition, a study by Sun et al. [18] predicts end-of-season tuber yield and tuber set in potatoes using in-season UAV-based hyperspectral imagery and six machine learning models. The authors obtained optimal results with the ridge regression model, which yielded an R-squared value of 0.65.

The model's ERT, XGBoost, BR, GBR, and Light GBM demonstrated strong performance when applied to the 'Crop yield prediction' dataset. However, ERT, GBR, and XGBoost exhibited the highest correlation coefficients, each achieving R-squared values of 0.966. The prediction errors gave the lowest values for the ERT model (RMSE = 4,273.668 kg/ha, and MAE = 2,292.56 kg/ha). Furthermore, it recorded a considerable predictive variance (EVS = 0.967) and outperformed the XGBoost and GBR models in execution time. The findings are corroborated by Cao et al. [19], who compared LM, XGBoost, RF, and SVM algorithms to increase winter wheat yield in northern China using satellite, climate, and S2S atmospheric prediction data. XGBoost demonstrated the highest skill level when utilizing S2S predictions as inputs, achieving an R-squared value of 0.85. In a similar, Mariadass et al. [20] proposed using XGBoost for annual crop yield predictions, leveraging a dataset with temperature, rainfall, and pesticide variables. Their experiments yielded an R-squared value of 98%. It was shown that XGBoost is a helpful model for maize prediction [21]. Using the "Marked impacts of pollution mitigation on crop yields in China" dataset, the coefficient of determination exceeded 70% for the ERT, XGBoost, light GBM, and GBR models. However, ERT and XGBoost observed the highest coefficient of determination and explained variance scores. Furthermore, the ERT model achieved the lowest error regarding RMSE and MAE, while XGBoost observed a longer execution time. The findings reported by Li et al. [22] regarding soybean yield prediction using the XGBoost model align with an R-squared value of 0.85. In their study, the authors used a combination of spectral, meteorological, and soil data to establish a framework for county-level soybean yield prediction.

In addition, classical models exhibit high performance compared to ensemble methods, although ERT and XGBoost remain superior. This could be attributed to the scale of the data or the challenge of capturing non-linear relationships between variables in less complex datasets. Comparing XGBoost and ERT models to the ensemble models used, the ensemble methods generally demonstrate acceptable performance across all three datasets. However, ERT and XGBoost models notably outperform them. The training time of these ensemble models is typically shorter than that of XGBoost and ERT models, except for the GBR model, which exhibits longer training time. The performance of the ensemble models can be attributed to their capability to leverage multiple decision trees to minimize errors and optimize yield and their adaptability to complex variable relationships. However, the AdaBoost model underperforms compared to most ensemble methods for maize prediction. The effectiveness of this model is more closely dependent on the input variables used. Models utilizing climatic parameters have shown the best performance, followed by those relying on experimental data. Conversely, combining climatic and pollution parameters at different stages of maize development yielded lower performance.

A key finding that emerges from the study is that the ERT consistently outperformed all other models across the examined datasets. This model has proven a robust and valuable asset for maize yield prediction. It is a strong and helpful model for maize yield production. The findings from Tyler et al.'s [23] research align with the results. Indeed, the authors investigated historical patterns in sorghum productivity throughout the United States using a substantial

dataset of Sorghum bicolor yield and environmental variables. Through the application of machine learning approaches, including multiple linear regression (MLR), random forest (RF), and extremely randomized trees (ERT), they concluded that ERT exhibited the highest performance in terms of predictive accuracy. Gao et al. [24] developed predictive models for optimizing fertilization decisions in maize, rice, and soybean cultivation. They employed a range of machine learning algorithms, including Random Forest (RF), XGBoost, Support Vector Regression (SVR), Multiple Linear Regression (MLR), Artificial Neural Networks (ANN), and Extra Trees Regressor (ERT). These models were constructed using historical data encompassing crop performance, soil nutrient levels, and fertilization characteristics. ERT model demonstrated good accuracy in simulating crop yields, achieving an R-squared value of 0.749. In addition, Zhang et al. [25] described using machine learning to predict winter wheat leaf water content using multi-temporal crop canopy models and vegetation indices. The authors used hyperspectral images to estimate crop yields with reasonable accuracy. The results showed that ERT can effectively predict the water content of wheat leaves with an R-square equal to 0.88.

The effectiveness of the ERT model in predicting maize yields for the three data sets can be explained by features intrinsic to the algorithm. Indeed, ERT randomly constructs several decision trees. This randomization in the construction of the trees helps to reduce the potential bias and improve the generalisability of the model. As a result, it is less likely to overfit the training data, making it a reliable choice for predicting maize yields. Also, the ERT model can easily handle complex and heterogeneous datasets, which is standard in agriculture, where data can come from various sources. Its ability to generate a decision tree and combine their predictions reinforces the model's stability. The random variations introduced into each tree help reduce the model's overall variance, which is particularly beneficial in regression tasks where it is essential to minimize prediction errors. The model can capture non-linear relationships between the input and target variables, which can be crucial for modeling complex phenomena such as crop growth. These features make it a valuable tool for agricultural researchers. Nevertheless, this work has certain limitations that should be taken into consideration.

The study focused mainly on evaluating the performance of classical and ensemble learning methods. A comparison with deep learning methods was not undertaken, which could be an exciting avenue for future research. In addition, the performance of the models was evaluated on three specific datasets. More is needed to generalize the results better and strengthen the robustness of the conclusions. We recommend extending this analysis to various maize datasets from multiple sources. Moreover, the "crop yield prediction" data primarily aggregates and provides limited country-specific information. It is crucial to have region-specific data to enhance the accuracy and relevance of results. This approach allows for the consideration of local nuances, which can have a significant impact on agricultural yields.

Table 3. Models performance based on *Rsquare*: Coefficient of determination, RMSE: Root mean square error, MA: mean absolute error, EVS: explained variance score, Time = temps for models training

Models	Datasets	LR	Lasso	SVR	KNN	RR	DT	RF	GBR	GBM	XGB	ADB	BR	ERT
MAE	Cover crop and irrigation impact	0.230	0.243	0.265	0.934	0.230	0.361	0.262	0.250	0.300	0.230	0.307	0.319	0.202
EVS		0.923	0.912	0.894	0.095	0.923	0.814	0.902	0.903	0.877	0.927	0.872	0.863	0.939
RMSE		0.307	0.324	0.364	1.103	0.307	0.469	0.342	0.344	0.382	0.297	0.391	0.407	0.272
R square		0.920	0.911	0.888	-0.030	0.920	0.814	0.901	0.900	0.877	0.925	0.871	0.860	0.937
Training time (s)		0.005	0.541	4.296	0.557	0.564	38.123	879.671	8920.403	100.562	2547.399	329.119	0.913	324.491
computing time (s)		0	0.0001	0.01	0.06	0	0	0.002	0.09	0	0.001	0.2	0.01	0.04
MAE	Crop yield prediction	13700.183	13700.183	11613.641	6892.495	13703.806	3789.766	4022.946	2465.445	2747.827	2469.361	10786.845	2808.548	2275.697
EVS		0.417	0.417	0.495	0.773	0.417	0.893	0.923	0.963	0.956	0.963	0.642	0.948	0.967
RMSE		17865.154	17865.154	17067.161	11142.520	17865.749	7661.481	6481.684	4484.380	4925.549	4484.799	14078.062	5364.749	4273.668
R square		0.417	0.417	0.468	0.773	0.417	0.893	0.923	0.963	0.956	0.963	0.638	0.947	0.967
Training time (s)		0.453	0.606	339.443	2.655	1.517	64.383	890.527	98896.890	202.468	7614.951	419.510	10.917	322.376
computing time (s)		0.0009	0	0.189	0.02	0	0	0.05	0.008	0.05	0.02	0.09	0.37	0.09
MAE	Marked impacts of pollution mitigation on crop yields in China	690.809	685.004	758.697	861.927	689.751	798.136	660.979	573.878	596.896	582.676	806.962	625.996	554.778
EVS		0.631	0.634	0.526	0.439	0.632	0.479	0.644	0.711	0.703	0.734	0.520	0.687	0.752
RMSE		916.928	912.402	1038.683	1129.951	915.836	1089.003	899.712	812.591	822.836	778.410	1047.211	845.024	752.348
R square		0.630	0.634	0.525	0.438	0.631	0.478	0.644	0.710	0.702	0.733	0.518	0.686	0.751
Training time (s)		0.687	6.428	27.877	0.735	0.418	82.875	1762.111	85416.091	177.103	3782.927	181.277	49.636	231.354
computing time (s)		0	0	0.01	0	0	0	0.015	0.016	0	0	0.02	0.06	0.094

Table 4. Optimum models hyperparameters

Models	Crop yield prediction data	Cover crop data	Marked impacts of pollution mitigation data
LR	-	-	-
Lasso	alpha:0.001	alpha:0.0001	alpha:0.05
SVR	C: 1000; gamma: 1; kernel: linear	C: 1000; gamma: 1; kernel: linear	C: 1000; gamma: 1; kernel: linear
KNN	n neighbors:20 ; weight: Distance	n neighbors:20 ; weight: Distance	n neighbors: 26 ; weight: Distance
RR	alpha:0.03	alpha:0.03	alpha:0.01
DT	max depth: 20; max features: 3; min sample leaf:2	max_depth: 12; max features: 4; min sample leaf:5	max_depth: 18; max features:12; min sample leaf:20
RF	max depth: 9; max features: 3; min sample leaf:10	max_depth: 5; max features: 3; min sample leaf:10	max_depth: 9; max features:9; min sample leaf:31
GBR	learning rate:1, max_depth:19, n estimators: 100, subsample: 0.5	learning rate:0.02, max_depth:4, n estimators: 500, subsample: 0.1	learning rate:0.03, max_depth:4, n estimators: 1500, subsample: 0.5
Light GBM	Learning rate:0.09, n estimators: 100, number of leaves:31	Learning rate:0.09, n estimators: 200, number of leaves:5	Learning rate:0.03, n estimators: 1000, number of leaves:31
XGBoost	gamma: 5, learning rate: 0.04, min child weight: 1, n estimator: 1000	gamma: 0.5, learning rate: 0.3, min child weight: 8, n estimator: 100	gamma: 0.5, learning rate: 0.1, min child weight: 10, n estimators: 500
BR	n estimators: 200	n estimators: 100	n estimators: 500
adaboost	learning rate: 0.04, number of estimators: 100	learning rate: 0.03, number of estimators: 500	learning rate: 1, number of estimators: 200
ERT	max depht: 19, n estimators: 500	max depht: 5, n estimators: 500	max depht: 20, n estimators: 1000

The predicted and actual values of the four best models derived from the different datasets were represented in the following graphs. The ERT and XGBoost models obtained from the "cover crop dataset" (Figure 2) and the "crop yield prediction dataset" (Figure 3) showed a positive link between predicted and actual variables. Both models estimated the actual maize yield values with high accuracy. Only the predicted values of the ERT model were more correlated with actual values in the pollution dataset. However, compared with the XGBoost, LightGBM, and GBR models, it can estimate maize yield more accurately (Figure 4). These results confirmed that ERT was a good model for maize yield prediction.

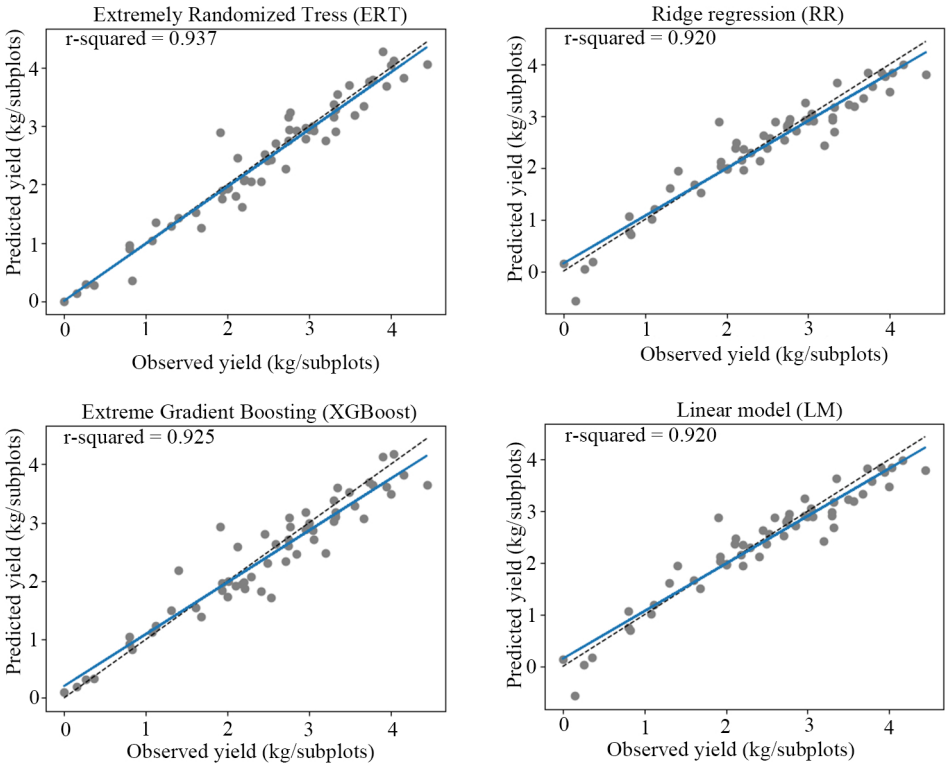


Figure 2. Scatter plots between the actual and predicted variables of the four best methods of the cover crop data

3.1 *Important variables*

The results for the ten most important variables contributing to yield prediction for the three datasets are shown in Figure 5. The variables "precipitation at growth season" and "aerosol optical depth for growth" were identified as the main contributors to yield prediction. These variables strongly influenced the model with the dataset "Marked impacts of pollution mitigation on crop yields in China". For the "Cover crop and irrigation impacts on weeds and maize yield", the variables "number of corn cobs per two 1 meter rows" and "year" had a significant effect on maize yield prediction. In contrast, the variables "temperatures" and "humidity" had a more significant impact on the model of the "crop yield prediction" dataset. Their impact on maize yield prediction was particularly significant. This suggests that these variables played a major role in the model's performance, specifically for these datasets. This information is valuable for understanding the key factors affecting maize yield in the context of this study.

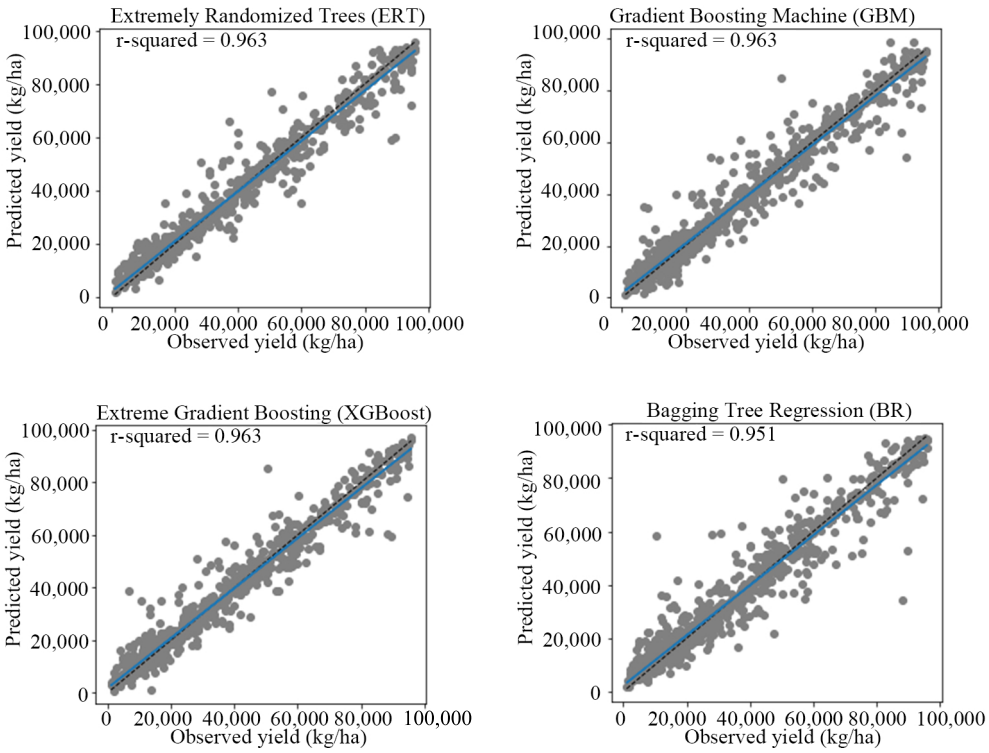
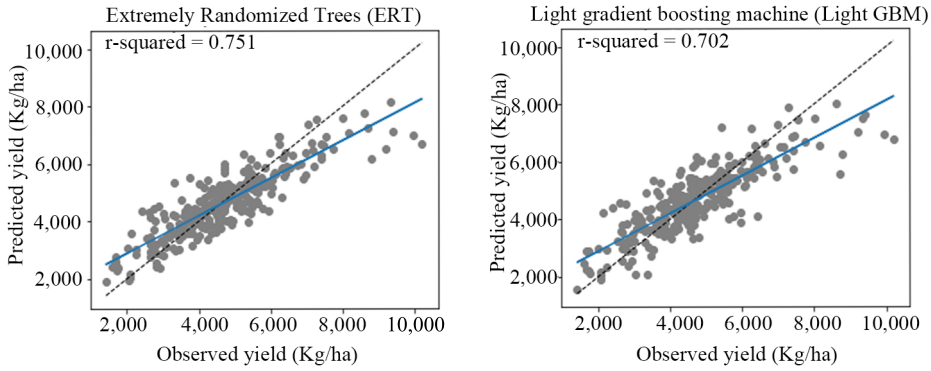


Figure 3. Scatter plots between the actual and predicted variable of the four best crop yield prediction datasets



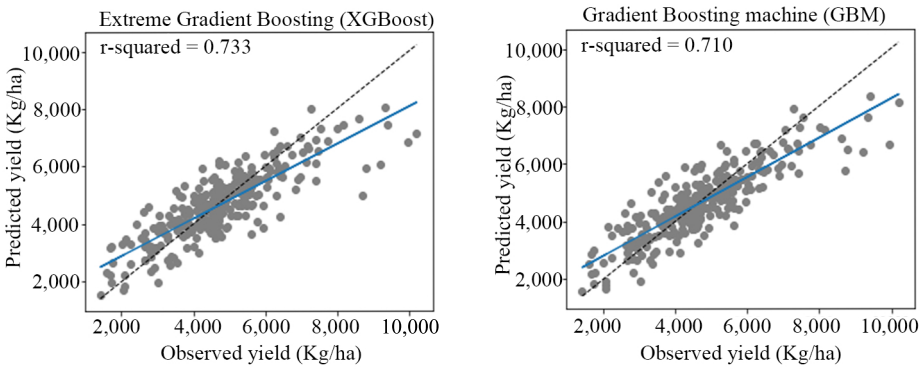


Figure 4. Scatter plots between the actual and predicted variables of the four best methods of Marked impacts of pollution mitigation on crop yields in China dataset

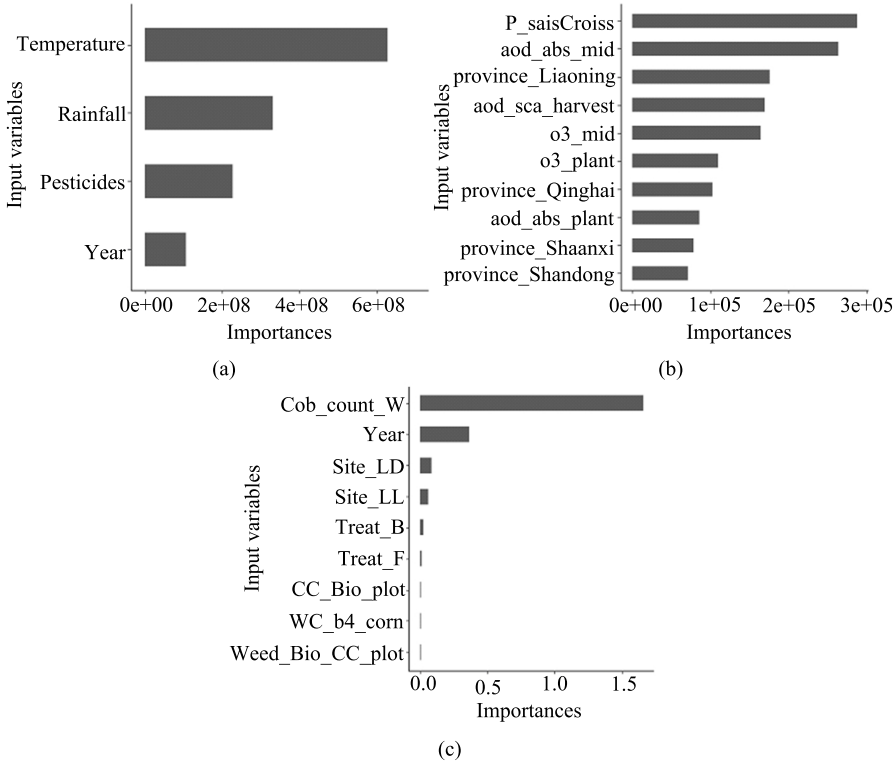


Figure 5. Importance variables of the models. (a): Crop yield prediction importance, (b): Marked impacts of pollution mitigation on crop yields in China, (c): Cover crop and irrigation impacts on weeds and maize yield. P_saisc_roiss: precipitation at growth season; $aod_sca_harvest$: Aerosol optical depth for harvest season using scale, $O3_mid$: Ozone in growth season, $O3_plant$: ozone in plant season, $Province_Qinghai$: Qinghai province, aod_abs_plan: Aerosol optical depth for plant season without scale, $Province_Shaanxi$: Shaanxi province, $Province_Shandong$: Shandong province, cob_count_w: number of corn cob, $Site_LD$:leyendecher site, $site_LL$: Los luno site, $Treat_B$: Treatment with barley, $treat_F$: treatment with fallow, cc_bio_plot: cover crop biomass harvested, wc_b4c_plot: weed biomass for growing season

Using accumulated local effects on each dataset and keeping other variables constant, it is observed that an increase in year or pesticide usage resulted in higher maize yield, whereas an increase in temperature reduced crop yield (Figure 6a). The impact of precipitation varies continuously. With cover crop yield data, an increase in the number of corn cobs leads to an augmentation in maize yield (Figure 6b) while year decreases yield. The yield on the Leyendecker site was

low compared to the Los Lunas site, which shows a high yield. Concerning pollution impact data, it is observed that an increase in precipitation during the growth phase, as well as AOD (aerosol optical depth) during this period, increase yield (Figure 6c). The province of Liaoning had a low yield, whereas the ozone increase during growth made a constant yield.

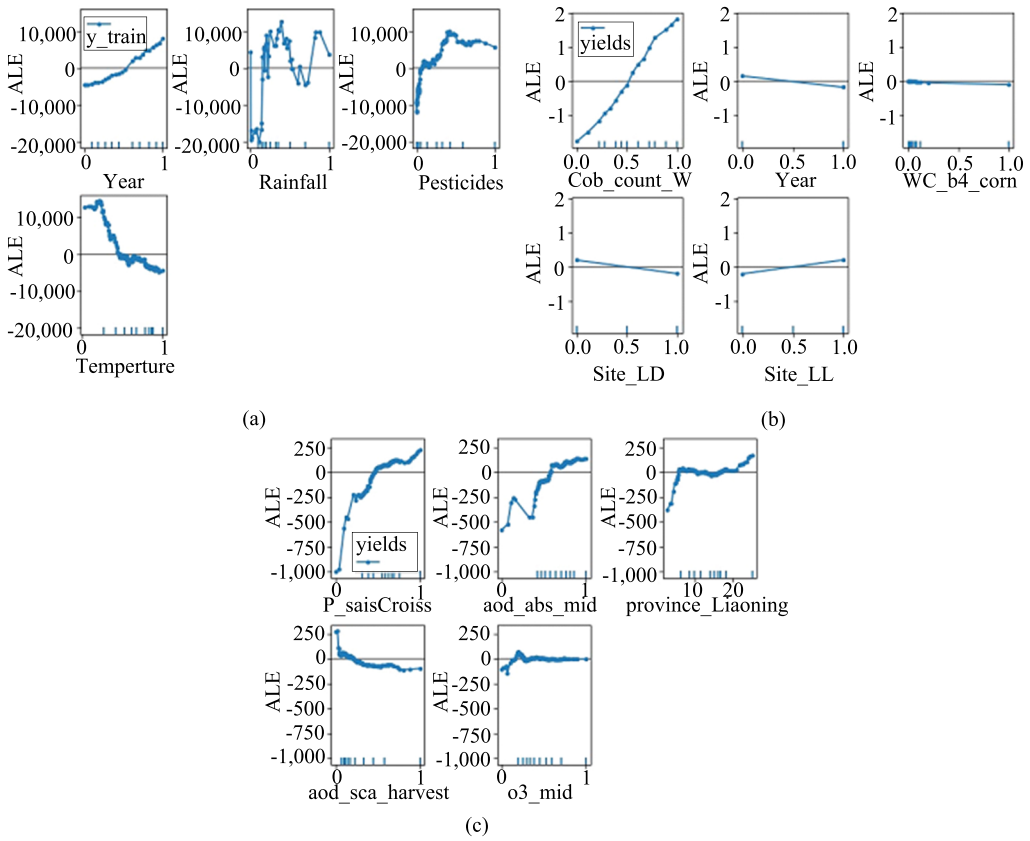


Figure 6. Accumulated local effect on model predictors. (a): Crop yield prediction importance, (b): Cover crop and irrigation impacts on weeds and maize yield; (c): Marked impacts of pollution mitigation on crop yields in China

4. Conclusion

Machine learning algorithms are becoming increasingly popular for estimating crop production. This study provided extensive experiment results of machine-learning models on maize production using three different data sets. The results, on the datasets used, show that ERT and XGBoost consistently outperformed competing models regarding coefficient of determination and explained variance. ERT shows superior performance with the error evaluation metrics, characterized by lower errors. While most ensemble methods share similarities, ERT stands out through their wholly random and independent tree construction and reduced sensitivity to noisy data. This feature helps mitigate overall overfitting. Although it is generally considered a faster ensemble model than others, it is crucial to consider processing time. Based on the variable permutation analysis, it is evident that “precipitation during the growth season”, “the number of corn cobs per two 1-meter rows”, and “humidity” are the primary variables that play a significant role in yield prediction across various datasets. However, The results cannot be generalized as they require further validation using additional datasets.

Acknowledgment

This work was supported by:

- German Academic Exchange Service (DAAD),

- Scholarship Program in Artificial Intelligence for Development (AI4D) Africa, funded by the International Development Research Centre (IDRC) and the Swedish International Cooperation Agency (SIDA), and managed by the African Centre for Technology Studies and Development (ACTS).

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Shiferaw B, Prasanna BM, Hellin J, Bänziger M. Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security*. 2011; 3: 307-327. Available from: https://doi.org/10.1007/S12571-011-0140-5/TABLES/3.

[2] Shah TR, Prasad K, Kumar P. Maize-A potential source of human nutrition and health: A review. *Cogent Food and Agriculture*. 2016; 2: 1166995. Available from: https://doi.org/10.1080/23311932.2016.1166995.

[3] FAO. *Statistics, food and agriculture organization of the united nations*. Rome, Italy: FAO; 2020.

[4] Shiferaw B, Prasanna BM, Hellin J, Bänziger M. Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security*. 2011; 3: 307-327. Available from: https://doi.org/10.1007/s12571-011-0140-5.

[5] Teh D, Rana T. The use of internet of things, big data analytics and artificial intelligence for attaining UN's SDGs. *Handbook of big data and analytics in accounting and auditing*. Singapore: Springer Nature Singapore; 2023. p.235-253. Available from: https://doi.org/10.1007/978-981-19-4460-4_11.

[6] Singh S, Jain P. Applications of artificial intelligence for the development of sustainable agriculture. *Agro-biodiversity and agri-ecosystem management*. Singapore: Springer Nature Singapore; 2022. p.303-322. Available from: https://doi.org/10.1007/978-981-19-0928-3_16.

[7] Vaishali P, Haneet K, Surjeet S, Jatinder M, Vinod S. Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. *Journal of Scientific Research*. 2020; 64: 394-398. Available from: https://doi.org/10.37398/jsr.2020.640254.

[8] Rao EP, Rakesh V, Ramesh KV. Big data analytics and artificial intelligence methods for decision making in agriculture. *Indian Journal of Agronomy*. 2021; 66: 279-287.

[9] Guojie R, Xinyu L, Fei Y, Davide C, Ata-UI-Karim T, Xiaojun L, et al. Improving wheat yield prediction integrating proximal sensing and weather data with machine learning. *Computers and Electronics in Agriculture*. 2022; 195: 106852. Available from: https://doi.org/10.1016/j.compag.2022.106852.

[10] Sarijaloo FB, Porta M, Taslimi B, Pardalos PM. Yield performance estimation of corn hybrids using machine learning algorithms. *Artificial Intelligence in Agriculture*. 2021; 5: 82-89. Available from: https://doi.org/10.1016/j.aiia.2021.05.001.

[11] Ahmad I, Singh A, Fahad M, Waqas MM. Remote sensing-based framework to predict and assess the interannual variability of maize yields in Pakistan using landsat imagery. *Computers and Electronics in Agriculture*. 2020; 178: 105732. Available from: https://doi.org/10.1016/j.compag.2020.105732.

[12] Dhaliwal DS, Williams MM. Sweet corn yield prediction using machine learning models and field-level data. *Precision Agriculture*. 2024, 25(1): 51-64. Available from: https://doi.org/10.1007/s1119-023-10057-1.

[13] Meng L, Liu H, Ustin SL, Zhang X. Predicting maize yield at the plot scale of different fertilizer systems by multi-source data and machine learning methods. *Remote Sensing*. 2021; 13: 3760. Available from: https://doi.org/10.3390/rs13183760.

[14] Reddy D, 2021 MK. Crop yield prediction using machine learning algorithm. In *5th International Conference on and Undefined 2021*. Japan; 2021. Available from: https://doi.org/10.1109/ICICCS51141.2021.9432236.

[15] Kalimuthu M, Vaishnavi P, Kishore M. Crop prediction using machine learning. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. 2020. Available from: https://doi.org/10.1109/ICSSIT48917.2020.9214190.

[16] Abbas F, Afzaal H, Farooque AA, Tang S. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*. 2020; 10: 1046. Available from: https://doi.org/10.3390/AGRONOMY10071046.

[17] Qin Z, Myers DB, Ransom CJ, Kitchen NR, Liang SZ, Camberato JJ, et al. Application of machine learning methodologies for predicting corn economic optimal nitrogen rate. *Agronomy Journal*. 2018; 110: 2596-2607. Available from: https://doi.org/10.2134/AGRONJ2018.03.0222.

[18] Sun C, Feng L, Zhang Z, Ma Y, Crosby T, Naber M, et al. Prediction of end-of-season tuber yield and tuber set in potatoes using in-season UAV-based hyperspectral imagery and machine learning. *Sensors*. 2020; 20: 5293. Available from: https://doi.org/10.3390/S20185293.

[19] Cao J, Wang H, Li J, Tian Q, Niyogi D. Improving the forecasting of winter wheat yields in northern China with machine learning-dynamical hybrid subseasonal-to-seasonal ensemble prediction. *Remote Sensing*. 2022; 14: 1707. Available from: https://doi.org/10.3390/RS14071707.

[20] Mariadass DAL, Moung EG, Sufian MM, Farzamnia A. Extreme gradient boosting (XGBoost) regressor and shapley additive explanation for crop yield prediction in agriculture. *IEEE*. 2022; 2022: 219-224. Available from: https://doi.org/10.1109/ICCKE57176.2022.9960069.

[21] Rao M, Dangeti S, Amiripalli SS. An efficient modeling based on XGBoost and SVM algorithms to predict crop yield. *Advances in Data Science and Management*. 2022; 86: 565-574. Available from: https://doi.org/10.1007/978-981-16-5685-9_55.

[22] Li Y, Zeng H, Zhang M, Wu B, Zhao Y, Yao X. A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. *International Journal of Applied Earth Observation and Geoinformation*. 2023; 118: 103269. Available from: https://doi.org/10.1016/j.jag.2023.103269.

[23] Huntington T, Cui X, Mishra U, Scown CD. Machine learning to predict biomass sorghum yields under future climate scenarios. *Biofuels, Bioproducts and Biorefining*. 2020; 14(3): 566-577. Available from: https://doi.org/10.1002/bbb.2087.

[24] Gao J, Zeng W, Ren Z, Ao C, Lei G, Gaiser T, et al. A fertilization decision model for maize, rice, and soybean based on machine learning and swarm intelligent search algorithms. *Agronomy*. 2023; 13(5): 1400. Available from: https://doi.org/10.3390/agronomy13051400.

[25] Zhang J, Zhang W, Xiong S, Song Z, Tian W, Shi L, et al. Comparison of new hyperspectral index and machine learning models for prediction of winter wheat leaf water content. *Plant Methods*. 2021; 17: 1-14. Available from: https://doi.org/10.1186/s13007-021-00737-2.