RESEARCH ARTICLE

Adoption of Machine Learning Methods for Crop Yield Prediction-based Smart Agriculture and Sustainable Growth of Crop Yield Production – Case Study in Jordan

Muneer Nusir[1], Mohammad Alshirah[2], Sahar Al Mashaqbeh[3], Rayeh Alghsoon[4]

1 Prince Sattam Bin Abdulaziz University
2 Al al-Bayt University
3 Hashemite University
4 Al-Ahliyya Amman University

Funding: This project was funded by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University award number 2023/01/26504

Potential competing interests: No potential competing interests to declare.

Abstract

Crop yield prediction is significant for global food security and economic systems. Numerous algorithms for machine learning have been utilized to support crop yield prediction due to the increasing complexity of factors influencing plant growth. Machine learning (ML) models are quite tedious because the models of ML for agriculture-based are complex. This study combines several models to build a sturdy and accurate model. Linear regression predicts a measurable response using various predictors and assumes a linear relation between the response variable and predictors. This research study explores the adoption of machine learning methods for crop yield prediction and their potential to support sustainable growth of crop yields. The dataset was collected from two main sources: i) the Department of Statistics Jordan and ii) the climate change knowledge portal, which is used to train the proposed model; and the availability of large datasets has cleared the path for the application of ML techniques in crop yield prediction. Nine ML regression analysis algorithms were tested to predict the crop yield; more than one algorithm gave very good results in prediction. XGBoost, multiple linear regression, Random forest, and Lasso regression give low mean squared errors of 0.092, 0.024, 0.023, and 0.023. Crop prediction may be remarkably useful from ML algorithms, but there are many challenges. One of these challenges is the quality of the data and the data volume, where machine learning algorithms need large data. Further, because of the intricacy of agriculture systems, developing ML models can be challenging. In this research study, the strengths of optimization and machine learning are integrated to build a new predictive model for crop yield prediction. The developed integrated model in this study contributes to increasing the efficiency of crop production, and reducing prices when food shortages are found. In addition, the proposed model supports the crop prediction process, where crop prediction has a vital role in agricultural planning and procedures for making decisions. ML algorithms are an essential instrument for decision assistance for crop prediction, either in supporting decisions on the suitable to grow. The algorithm's performance may be improved by applying more innovative techniques. The developed model helps policymakers on precise forecasts, to make suitable evaluations of imports and exports to strengthen food security nationwide.

Muneer Nusir[1,*], **Mohammad Alshirah**[2], **Sahar Al Mashaqbeh**[3], and **Rayeh Alghsoon**[4]

[1]*Assistant Professor, Prince Sattam Bin Abdulaziz University, Alkharj, Riyadh, Saudi Arabia*

[2]*Associate Professor, Al al-Bayt University, Mafraq, Jordan*

[3]*Assistant Professor, the Hashemite University, Zarqa, Jordan*

[4]*Part-time Lecturer, Amman Al-Ahliyya University, Amman, Jordan*

[*]Corresponding author: Muneer Nusir (e-mail: moneer.techno@gmail.com)

Keywords: Crop Yield Prediction, Lasso regression, Machine Learning, Random forest, Regression, Smart Agriculture, and XGBoost.

I. Introduction

The escalating challenge in the global agricultural field is to strive to meet the increasing food, feed, fiber, and bioenergy requirements. It faces climate change averseness and declines in natural resources [1][2]. Agriculture must therefore find innovative solutions to increase productivity while minimizing its impacts on the environment and ensuring food security. Adopting ML methods for crop yield prediction is a pivotal issue as the crop yield to which humanity is growing is increasingly prone to threats from various sources [3]. The gap between the availability and desire for food is significant and increasing with time [4].

By 2050, food security is predicted to become one of the greatest global challenges; this coincides with the world population reaching 9.7 billion, equivalent to 20.6% of the current population [5][6]. It implies that up to 30%–40% of the food produced goes bad in developing countries before finding consumers' hands [7].

The rapid growth in population and the effects of climate change generating severe weather conditions have increased pressure on global agriculture to produce more [2]. There is a growing need for creative, sustainable methods that can improve yield while simultaneously ensuring food quality and safety. The nature of the ML algorithms used in this paper is believed to be a particularly interesting mechanism. It provides a more transparent and differentiated solution than those developed by traditional econometric models, decision trees, and even previous works related to the issue [8][9].

In this context, yield forecasting is a crucial duty for agriculture. The crop yield prediction strategies and processes are nonlinear and change with time [10]. An accurate prediction supports creating strategic policies before reaching a crisis level in the agricultural economy caused by a food shortage. Improving crop yield prediction can increase agricultural output and lead to economic consequences. Traditionally, experts in farm management and agricultural economics relied heavily on farming-associative economic factors or historical yield data[11][12].

These data on their own are insufficient since they do not precisely capture how economic models, which are based on

agronomic principles, make yield predictions. Researchers in the area of forecasting agricultural yields have tried to resolve this lack of precise knowledge about the link between economic theory and field-level conditions by incorporating both field features' details (i.e. year, average yield, planted area, harvested area, production, and crop type) and against the prediction of crop production's attributes (i.e., temperature mean, temperature maximum, temperature minimum, and precipitation). However, variability between fields causes much of the difficulty in making accurate predictions. Recent developments in ML techniques and remote sensing data offer great potential to increase yield predictions by exploiting spatial field-level variability [13][14].

Over the last ten years, ML approaches have been conducted on a variety of agricultural systems to deliver more accurate solutions. This is mainly because ML techniques can handle very complicated nonlinear agricultural issues [15][16]. Instead of assuming the functional form, probability distribution, or smoothness of the data model, as classic statistical approaches do, ML makes no such assumptions [8][17].

According to [8], ML approaches can ascertain the correlation between independent and dependent variables through data analysis. ML approaches rely on structures that are non-parametric and semi-parametric, and their validity is determined by precision in predictions. If the class and other properties fulfill certain probability distributions, a non-parametric technique does not need any previous presumptions on those distributions' shape [9]. A proposed model is developed to efficiently predict the harvest production surpassing by conserving the valid data distribution with a precision of 93.7% [10] using a deep learning algorithm.

In this research, several regression models for predicting the yield of crops like wheat, cotton, and lentils are applied depending on soil weather and crop parameters. ML methods are conducted to train the models. To predict crop yield, linear regression models, lasso regression, random forest regression, XGboost, SVM regression, decision tree analysis, ridge regression, Elastic Net regression and polynomial regression are utilized in this research study. The main aim of the present study is conducted and achieved through the following objectives:

- To predict the crop yield, help the future conditions of the production level of the yield, and then help the farmers to avoid loss.
- To determine if the integrated model can achieve better crop prediction and which model provides more accurate prediction.
- To evaluate several ML algorithms and identify which technique is the most precise prediction of point-scale yield.

Overall, the structure of the proposed research offers an extensive exploration of the topic, and the literature review elaborates on the review of the literature to identify research gaps. Research methodology is achieved through regression models based on ML methods, followed by a case study that provides practical applications and an in-depth discussion of the findings, leading to a well-supported conclusion.

II. Literature Review and Related Work

A. Effectiveness of Factors Influencing Crop Yield

Understanding the relation between soil properties and characteristics for agriculture processes and crop yield production is crucial. Crop yield is influenced by a variety of weather and soil parameters, making it a complex phenomenon. In addition to uncontrollable factors such as climate and soil conditions, many controllable factors affect crop yield, including farming practices, types and amounts of fertilizers used, and irrigation frequency. Given this complexity, it is crucial to measure the influence of these different factors on crop yield. The following section reviews research conducted in this area.

The impact of machine learning on various industries and the utilized methodologies in different research studies have been presented [18]. The study by Majumder *et al*., (2020) investigated how different land use practices affect surface temperature variations and their negative impacts on rice and wheat crop yields. Conducted across three distinct climatic regions in Punjab, the research classified satellite data into four primary land use and land cover (LULC) categories: water, vegetation, built-up areas, and plain soil. The findings revealed that areas transitioning from agriculture, plain soil, and forests to urban development experienced a rise in temperatures. Additionally, the Normalized Difference Vegetation Index (NDVI) was found to have a positive correlation with rice and wheat yields but a notably negative correlation with Land Surface Temperature (LST).

The impact of various climatic factors on wheat yield in Northwest India was reviewed by[19]. The study analyzed key elements such as daily temperature, precipitation, variability of groundwater, and index of evapotranspiration. It was found that a rise in the number of days with temperatures surpassing 35°C during the wheat maturation phase resulted in lower yields. Moreover, the depletion of groundwater and surface water for irrigation, caused by insufficient rainfall during the wheat-growing season (November–March), worsened the situation. As a result, the combination of high temperatures, severe water scarcity, and reduced irrigation significantly contributed to the decline in wheat yields. Additionally, the study discussed and analyzed several classification algorithms. The results demonstrated a high classification accuracy, with the Bayes Net algorithm achieving 99.59% and the Naïve Bayes Classifier and Tree algorithms reaching 99.46% accuracy[18].

B. Machine Learning Techniques for Predicting Crop Yield

Crop yield production poses a significant challenge in precision agriculture, primarily due to the complex interplay of factors such as weather, soil conditions, climate, and fertilizer usage [20]. Accurate and timely crop yield forecasting before harvest is crucial, yet it remains a difficult task for researchers because of the wide variety of influencing factors. However, recent advancements in machine learning have demonstrated considerable promise in overcoming this challenge. By analyzing various features, machine learning tools can identify patterns and correlations within datasets, leading to improved yield predictions. For these models to be effective, they must be trained on comprehensive datasets collected from relevant sources [21].

ML algorithms are generally divided into two primary categories: supervised and semi-supervised learning. Predicting

crop yield is a crucial component of the decision-making process, as it aids farmers in planning and making informed decisions for the future [4].

Accurate crop yield prediction can assist farmers in determining what to grow and the optimal timing for planting[10]. By identifying factors and areas that could lead to adverse growing conditions, yield predictions help reduce losses. Additionally, these predictions can be used to assess and optimize growing conditions, potentially enhancing crop growth [22]. The following section reviews a range of machine-learning techniques employed for predicting crop yield across different crop varieties. Various approaches and tools have been employed to explore the relationships between crop yield and soil properties [23]. Multiple regression models and correlations are commonly utilized to achieve these goals [24].

ML serves as a vital decision-support tool for crop yield prediction. This research involves performing and training various algorithms, including linear regression, LASSO regression, random forest regression, decision tree regression, polynomial regression, ridge regression, Elastic Net regression, and Extreme Gradient Boosting (XGBoost). An explanation of these algorithms is provided in the following section.

1. Linear regression is a statistical technique used to model the relationship between a numerical outcome and one or more explanatory variables. When multiple inputs are involved, an iterative method can be used to optimize the coefficient values by minimizing the model's error on the training data. This optimization process, known as Gradient Descent, starts with random coefficients and adjusts them iteratively to reduce the error. The process continues until the sum of squared errors is minimized or no further improvement can be achieved [25].
2. Decision tree regression is a technique that utilizes a flowchart-like tree structure to represent decisions and their potential outcomes, including input costs and benefits. It falls under supervised learning algorithms and can handle both continuous and categorical output variables. When employed to predict continuous-valued outputs, it is known as a regression tree. This method is particularly effective in scenarios where the relationship between variables is non-linear. However, a notable drawback of decision tree regression is the risk of overfitting [26].
3. Random Forest is a supervised algorithm applicable for both classification and regression tasks. It works by combining multiple decision trees through a technique called Bootstrap Aggregation, or bagging. Each decision tree is trained on a different sample of the data, which helps reduce the overall variance compared to individual trees. The final prediction is based on the collective output of all the trees, with the result in regression being obtained by averaging the outputs of each tree [27].
4. XGBoost is a form of ensemble learning that systematically combines the predictive strengths of multiple models to create a single, more accurate model. In XGBoost, the base learners might individually perform poorly on certain predictions, but when combined, their errors tend to offset each other, allowing the more accurate predictions to prevail. This cumulative effect of combining models leads to improved final predictions [28].
5. Polynomial regression, introduced by Drucker *et al*., (1997), extends traditional regression analysis to handle non-linear relationships between a dependent variable and one or more independent variables. While simple linear regression is suitable for linear relationships, polynomial regression allows for more complex, non-linear relationships by fitting a polynomial equation to the data. This approach can be used in both simple and multiple regression scenarios to

capture the intricacies of variable interactions that linear models might miss.

6. Ridge regression is a technique used to tackle multicollinearity issues in data. While Ordinary Least Squares (OLS) estimates are unbiased, they can have large variances, resulting in observed values that may deviate significantly from the true values. Ridge regression addresses this problem by reducing standard errors and stabilizing the estimates. However, unlike some other methods, ridge regression does not set any coefficients to zero, so it cannot simplify the model or make it more parsimonious [29].
7. Support Vector Machine (SVM) regression aims to find a hyperplane in an N-dimensional space (where N is the number of features) that best separates data points into distinct classes. Although multiple hyperplanes could separate the classes, the goal is to identify the one that maximizes the margin. By maximizing this margin, the model improves its ability to classify future data points with greater accuracy and confidence [30].
8. Lasso regression, introduced by Tibshirani, (1996), addresses the limitation of ridge regression by performing variable selection. While ridge regression reduces standard errors without setting any coefficients to zero, Lasso regression shrinks less important regression coefficients to exactly zero. This results in a simplified and more parsimonious model. For improving the accuracy of crop production forecasts, Lasso can be particularly effective as it selects the most important predictor variables, reduces the number of predictors in the model, and enhances the efficiency of identifying key predictors.
9. Elastic Net Regression, introduced by Zou and Hastie, (2005), was developed to overcome the limitations of both ridge and LASSO regression. While LASSO is effective at eliminating variables by setting some coefficients to zero, and ridge regression is better suited for handling highly correlated variables, neither method is ideal when dealing with a large number of variables with unknown degrees of correlation. Elastic Net addresses this by combining the penalties of both LASSO and ridge regressions, making it more suitable for estimating functions in situations where variable correlation is complex or unknown.

A yield forecast model for rice crops was created using four distinct techniques. The models were built using 15 years of meteorological and crop yield data and validated with a three-year dataset. Their performance was assessed using various metrics. The experimental analysis revealed that the Artificial Neural Networks (ANN) model outperformed the others, demonstrating its high effectiveness and suitability for predicting rice yield in the Udham Singh Nagar (USN) district of Uttarakhand.

A proposed work utilizes a recurrent neural network deep learning algorithm to forecast crop yield, effectively predicting yields with very high accuracy while preserving the original data distribution [10]. Additionally, a ML model for predicting farm production was developed by [31]. This model involved training six different supervised regression techniques. Among these, Random Forest Regression showed the best performance, achieving a Mean Absolute Error (MAE) of 468.16 and a Cross-Validation Score of 0.6087, surpassing the other models.

ML Classification Models were used by Anakha *et al*., (2021) to forecast crop output based on various factors. The study employed several classifiers. The authors emphasize that precision farming should prioritize quality over environmental factors. The research covered 14 districts in Kerala, detailing various aspects. Users can register using a mobile app,

input their location and region, and access the system. Among the models, Logistic Regression achieved an accuracy of 87.8%, Naïve Bayes reached 91.58%, and the Random Forest Classifier, utilizing Bagging, had the highest accuracy at 92.81%. The Random Forest model's accuracy on test data was 91.34%. The research aims to enhance the efficiency and effectiveness of crop cultivation.

Wheat crop yield estimation was performed using a Support Vector Regression (SVR) model[32]. The study tested several models, including nine base learner models and two ensemble models. SVR exhibited the highest learning efficiency among the nine models tested. Despite the higher cost of ensemble models, they did not significantly enhance accuracy. Additionally, increasing the amount of training data improved performance across all models. In another study, the effectiveness of machine learning techniques for predicting corn yield in Iowa State was examined using remote sensing data. The results indicated that machine learning methods are effective for yield estimation, with Deep Learning (DL) delivering particularly stable and favorable outcomes [33].

In this paper, several regression models are performed and trained. The proposed research methodology to develop the model is explained in the following section.

III. Research Methodology

Due to the complexity of machine learning models for the agricultural-based, creating the machine learning models is quite tedious. The modeling problem can be reflected as either a classification or a regression-based problem[34]. This study uses the agriculture dataset from the Department of Statistics in Jordan and weather data to train the proposed model [35].

This dataset from the Department of Statistics includes 5 crop details from 1999 to 2022. There are a total of six columns; year, average yield, harvested area, planted area, crop type and production. Production is the target variable; the prediction of crop production or yield depends on the remaining 5 attributes in the dataset. Data from the second source contains temperature mean (T-mean, ∘C), maximum temperature T-max in ∘C, minimum temperature T-min in ∘C, and precipitation (Prcp, mm) of Jordan from 1999 to 2022. The description of the dataset is given in Table 1. The dataset is preprocessed before exploratory data analysis. Planted area and harvested area are the number of acres each year from 1999 till 2019, average yield represents yield or production from harvested, and production is presented in metric tons. Temperature is in centigrade and precipitation is in millimeters. Exploratory data analysis (EDA) of the dataset is discussed in detail in the result analysis section.

Machine learning models can be leveraged in agriculture for early detection of crop disease identification, crop yield prediction, weather forecasting, crop price prediction, and species identification [36]. Results from machine learning models are vital information for farmers in informed decision-making at each step in agriculture. The general steps followed by farmers are as follows; the first step is about the selection of crops. The second step is regarding the preparation of the land, the third step is seed sowing and the fourth Step is about irrigation & fertilizing. After that the crop maintenance step is started then the harvesting step begins and finally the post-harvesting activities start [37]. The

executed methodology for this research is described in Figure 1. Step 1 is data collection from multiple sources, step 2 is data preprocessing where data is cleaned by removing null values and outliers, step 3 is exploratory data analysis, in step 4 the data is split into test data and train data, step 5 is training the machine learning model of training dataset and step 6 is result analysis after testing the model in the test dataset.

The following steps explain the process of collecting data, processing, exploring, splitting data, and training data.

1) Data Collection

First-step data is collected from the website of the Department of Statistics Jordan and the climate change knowledge portal. Data is downloaded in comma-separated value (CSV) format. The Department of Statistics provides data and reports on the socioeconomic aspects of Jordan, including the environment, agriculture, and much more.

Table 1. Dataset Description

	Column name	Datatype	Values
1	Year	Date	Year
2	Planted area	Floating point	Dunum or Acre
3	Harvested area	Floating point	Dunum or Acre
4	Average yield	Floating point	Metric Tons per Dunum
5	Production	Floating point	Metric Tons
6	Crop	String	Wheat, barley, lentil, chickpea, vetch
7	Temperature Minimum	Floating point	Minimum temperature of the year
8	Temperature Maximum	Floating point	Maximum temperature of the year
9	Temperature Mean	Floating point	The mean temperature of the year
10	Precipitation	Floating point	Millimeter (mm)

2) Data Preprocessing

Preprocessing of data after collection. Python programming is used in this study and Jupyter Notebook is the development environment. The files including EDA and machine learning models are uploaded to GitHub for public access [38]. During data preprocessing the dataset is loaded as a pandas data frame. The first step is to check for any null values and zero variance columns. The crop column is encoded from string to binary for EDA, one hot encoding technique is applied to convert this column which results in the addition of new columns separate for each crop; crop Wheat, crop_Barley, crop_Chickpea, crop_Lentil and crop Vetch. With the addition of these columns, the total number of columns changed from 10 to 13. Encoding allows the machine learning algorithm to perform better [39]. The planted area and harvested area columns represent acres of land where a crop is planted and harvested, and the production column presents metrics tons of crop yield, these values are large floating point numbers. Large numbers affect the machine learning model performance, thus these columns are transformed with smart transform and log transform. Smart transform is applied on independent variables, and log transformation on dependent variable ‘Production’. After

preprocessing the data next step is to conduct exploratory data analysis (EDA).

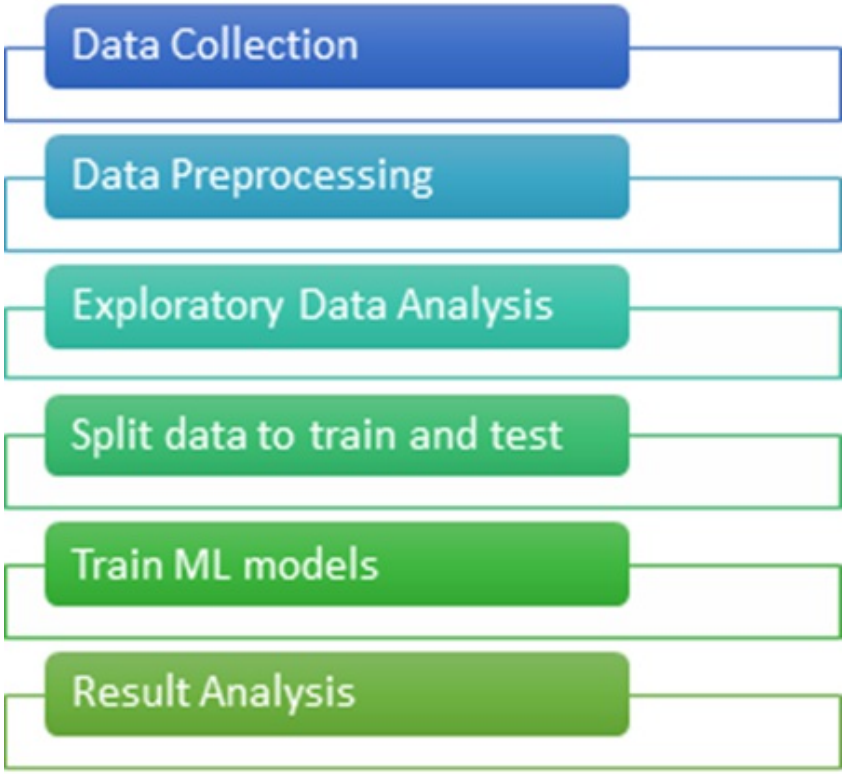

Figure 1. Proposed Research Methodology for Crop Yield Prediction

3) Exploratory Data Analysis (EDA)

EDA presents the basic information, data interpretability, and visualization of the data to make it clear and simple to understand [39]. Production and average yield of each crop by year (2000-2020) can be visualized with data of each crop in Figure 2.

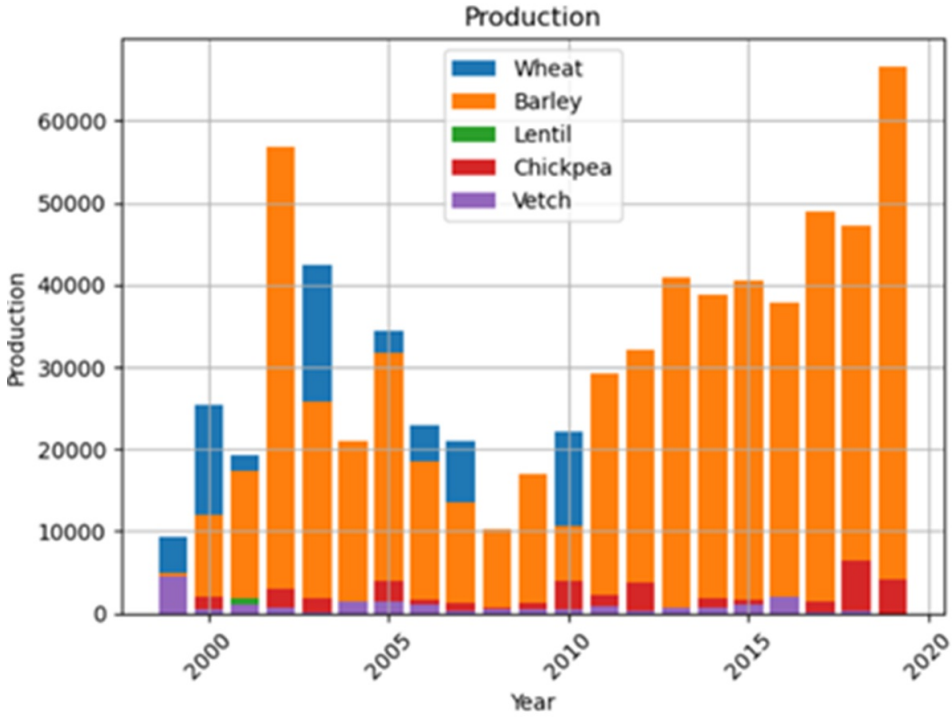

Figure 2. Yearly Production of Crops

The dataset here is of 13 columns among which the 'Production' column is the target column which is being predicted or measured using the machine learning models, and the remaining 12 columns are independent columns which are used in predicting and measuring the target column. All of the columns contribute to predicting the production by calculating the correlation between the independent and target variables. To find the correlation, Pearson's Correlation Coefficient method is applied and the results are shown in Table 2. Pearson correlation coefficient measures the strength of the linear relationship between 2 quantitative variables. If the value of this coefficient is close to -1 or 1, this indicates a strong linear relation between these two variables. If the value is close to 0 indicates poor correlation. Equation 1 represents Pearson's correlation coefficient. Harvested area and Planted area columns contribute the most in predicting the production.

$$r = (n * \Sigma xy - \Sigma x * \Sigma y) / sqrt((n * \Sigma x^2 - (\Sigma x)^2) * (n * \Sigma y^2 - (\Sigma y)^2 \qquad (1)$$

Table 2. Correlation coefficients

Year	0.162050
Planted Area	0.732012
Harvested Area	0.935576
Temp_mean	0.050504
Temp_min	0.093477
Temp_max	0.007114
Precipitation	0.169525
Average Yield	-0.117689
crop_Barley	0.634870
crop_Chickpea	-0.284775
crop_Lentil	-0.338342
crop_Vetch	-0.325990
crop_Wheat	0.314238
Production	1.000000

4) Split data to train and test

After preprocessing the data and before building the ML model, the data is split into training and testing sets to measure the model's performance [40]. Machine learning algorithms are used to train the models. The split dataset is categorized into four groups: X-test, X-train, Y-test, and Y_train. X_train and y_train are used to train the model. X_test is the input to the model which predicts and generates the results as y_pred. Predicted output is compared with y_test to evaluate the model performance [41]. To predict production, multiple machine learning algorithms are applied in this research, including linear multiple regression, ridge regression, lasso regression, polynomial regression, elastic net regression, random forest regression, SVM regression, decision tree regression, and XGBoost.

5) Training machine learning models

Multiple linear regression (MLR) is a powerful algorithm for understanding the relation between variables. It establishes a linear relation between dependent and independent variables. When the linear regression model is trained it can be utilized to predict production with new data points. Outliers in the data will skew the results significantly. MLR models the relationship between the multiple independent variables and the dependent variable. Equation 2 represents multiple linear regression, where $\hat{y}$ (y-hat) represents the predicted value of the dependent variable based on independent variables, β_0 (beta-naught) is the y-intercept, which represents the predicted value of dependent variable (y) when all the independent variables are zero, β_1 (beta-one) to β_n (beta-n) are the regression coefficients for each independent variable (x_1 to x_n) and ε (epsilon) represents the error term.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \qquad (2)$$

Lasso regression extends the principles of linear regression by integrating a regularization term to deal with problems such as overfitting and variable selection. The Lasso regression formula is represented in equation 3 [42].

minimize: $1/2 * \Sigma(y_i - \beta_0 - \Sigma\beta_j * x_{ij})^2 + \lambda * \Sigma|\beta_j|$ (3)

Polynomial regression is a statistical technique that models the relation between dependent and independent variables using a polynomial function of degree n, e is the error term as shown in equation 4. Polynomial regression, unlike linear regression, can capture complex and non-linear patterns, rather than assuming a linear relationship.

$y = b0 + b1*x + b2*x^2 + b3*x^3 + ... + bn*x^n + e$ (4)

Ridge Regression is a statistical technique utilized to analyze multiple regression data that is affected by multicollinearity. The objective is to improve the accuracy of prediction and stability of the model by combining a penalty term into the ordinary least squares (OLS) estimation. The formula for the ridge regression is equation 5.

$\beta = (X^T X + \lambda I)^{-1} X^T y$ (5)

Elastic Net regression is a regularization method that incorporates the advantages of Ridge and Lasso regression. The objective is to overcome the limits of each unique method while capitalizing on their advantages. As shown in equation 6 λ1 and λ2 are the regularization parameters.

$minimize: 1/2 * \Sigma(y_i - \beta_0 - \Sigma\beta_j * x_{ij})^2 + \lambda1 * \Sigma|\beta_j| + \lambda2/2 * \Sigma\beta_j^2$ (6)

Support Vector Machines (SVM) are typically utilized for classification, however, in this case, they are being applied to regression problems. The aim is to identify a hyperplane that optimally accommodates the dataset while accounting for a specific margin of error. The initial component in the objective function shown in equation 7 represents the complexity of the model, whereas the next component quantifies the degree of error. The regularization parameter C determines the balance between accurately fitting the data and the model complexity.

$minimize: 1/2 ||w||^2 + C * \Sigma(max(0, |y_i - w^T x_i - b| - epsilon)$ (7)

Decision tree regression is a non-parametric technique that builds a model like a tree to represent decisions and their potential outcomes, which includes chance events, costs of resources, and utility. The algorithm initiates by identifying the optimal feature to divide the data. The data is separated into subsets according to the split. This procedure is iteratively performed for each subset, generating new nodes and branches until the end condition is satisfied. Although decision tree regression does not have a specific formula, its algorithmic methodology makes it a robust and easily understandable solution for a wide range of regression situations.

Random forest is one of the powerful supervised machine learning algorithms (classification and regression tasks are performed). Random forest builds multiple decision trees at the training time and generates outputs of the classification and regression for the individual trees. Greater tree density in a forest leads to more accurate predictions [21][42]. Random Forests is a non-parametric advanced classification and regression tree (CART) analysis method. CART includes multiple decision trees; it combines the predictions from different decision trees in the forest as shown in equation 8 [33].

$Minimize: (1/2N) \Sigma(yi - \hat{y}i)^2 + \lambda \Sigma|\beta j|$ (8)

XGBoost, short for eXtreme Gradient Boosting, is an ensemble learning technique using gradient boosting. The model is constructed by iteratively including decision trees, each tree attempting to enhance the predictions made by the preceding ones. Equation 9 shows the fundamental concept of XGBoost. $\hat{y}(x)$ is the predicted value for a new data point (x), $f_0(x)$ is an initial prediction, $\Sigma f_t(x)$ is the sum of predictions from all the T decision trees in the ensemble. $f_t(x_t)$ represents the prediction made by the decision tree and T is the total number of trees in the XGBoost model [43].

$$\hat{y}(x) = \Sigma f_t(x) = f_0(x) + \Sigma f_t(x_t) \text{ where } t = 1 \text{ to } T \qquad (9)$$

IV. Results and Discussion

This study emphasizes on security of the food supply chain in Jordan, particularly examining the trends and factors affecting crop yields. By employing exploratory data analysis (EDA), the research aims to uncover patterns and relationships among various agricultural variables without relying on pre-established hypotheses.

Key aspects of the study include:

- Planted and Harvested Areas: The study compares the planted and harvested areas of different crops, with a specific focus on wheat and barley. These crops show significant differences in their planted and harvested areas compared to other crops.
- Trends Over Time: The research highlights a concerning trend of decreasing planted areas since 1999. This decline prompts a need for increased plantation efforts and a deeper investigation into the underlying causes.
- Annual Production: The study measures the annual production of various crops in metric tons per year, identifying barley and wheat as the most highly produced crops in Jordan.

The findings aim to inform agricultural policies and practices, help to define the challenges faced by the agriculture sector, and improve crop yields in the future.

Exploratory data analysis (EDA) reveals patterns and relationships among variables without the need for prior hypotheses. This study identifies various relationships among data variables. Figure 3 illustrates the planted and harvested areas of crops, highlighting significant differences between wheat and barley compared to other crops. Notably, the planted area has been decreasing since 1999, underscoring the need to increase plantation efforts and investigate the causes behind this decline. The annual production of different crops is measured in metric tons per year, with barley and wheat being the most highly produced crops in Jordan.

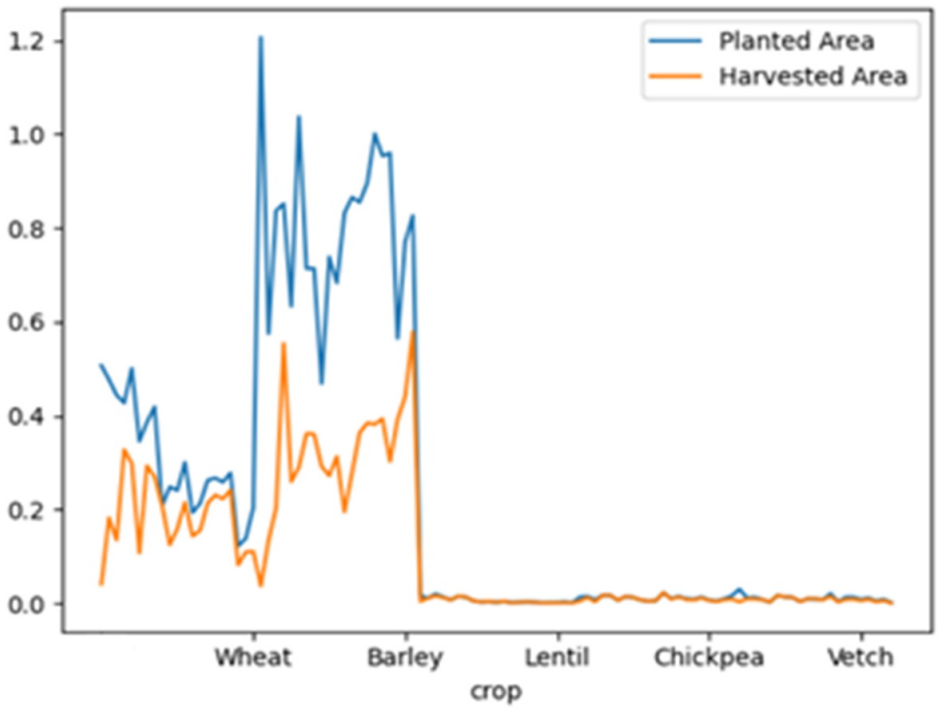

Figure 3. Comparison of planted area and harvested area

Various error metrics were applied, including (MSE), (MAE), R-squared, (RMSE) and (MAPE) to assess the performance of the ML model. The results of these error metrics for all models are tabulated in Table 3. Multiple linear regression models were employed to predict production based on the linear relation between the target variable and independent variables. The model attained an R-squared value of 0.994 and MSE of 0.024.

The results indicate that the model captures a significant portion (99.4%) of the variance in the dependent variable. Lasso regression, a type of linear regression, incorporates both feature selection and regularization techniques into the model training process. The cost function is modified by adding an L1 penalty term, which penalizes the total absolute values of the model coefficients. The Mean Squared Error (MSE) of the Lasso regression model is 0.023, and the R-squared value is 0.995. This high R-squared value signifies that the model's predictions on the test data exhibit substantial accuracy. This implies that the current model setup accurately represents the relationship between features and production. Prediction errors over the test years using several algorithms are analyzed and the comparison for each algorithm is tabulated in Table 3.

Table 3. Result of multiple models

Method	MSE	MAE	MAPE	R-Squared	RMSE
Multiple linear regression	0.024	0.106	1.27%	0.994	0.156
Lasso regression	0.023	0.104	1.23%	0.995	0.153
Polynomial Regression	0.238	0.087	3.34%	0.981	0.296
Ridge Regression	0.035	0.126	1.45%	0.992	0.187
ElasticNet Regression	1.308	0.995	13.73%	0.725	1.143
SVM Regression	0.271	0.368	4.66%	0.943	0.521
Decision Tree Regression	0.243	0.368	5.19%	0.948	0.493
Random forest	0.023	0.108	1.57%	0.993	0.153
XGBoost	0.092	0.236	3.25%	0.980	0.303

The Random forest revealed significantly higher performance compared to the linear regression model. With an Out-of-Bag (OOB) score of 0.733, it shows a high level of generalization ability, indicating effective performance on unseen data. The mean squared error (MSE) of 0.023 is significantly lower than that of the decision tree regression model, reflecting more precise and accurate predictions on average. Additionally, the R-squared value of 0.993 proposes that the model explains a significant portion of the variability in the target variable.

These findings highlight the importance of choosing an appropriate model for the data. While linear regression is often used as an initial approach, it may not be suitable for capturing complex relationships. In contrast, random forest is more robust against non-linear relationships and noisy data, leading to a significantly better fit in this context.

The XGBoost method achieved an R-squared value of 0.980, indicating that it explains 98% of the variance in the target variable (production) based on the independent features used. An R-squared value of 0.980 is considered excellent, suggesting that the model captures a significant portion of the factors influencing production. Additionally, the (MSE) for the XGBoost model is 0.092.

Lasso regression and random forest regression exhibit exceptionally low (MSE) and high R-squared values. These results suggest that multiple models can effectively perform on the given dataset. Figure 4 presents a comparison of the (MSE) across all models, while Figure 5 illustrates the R-squared scores.

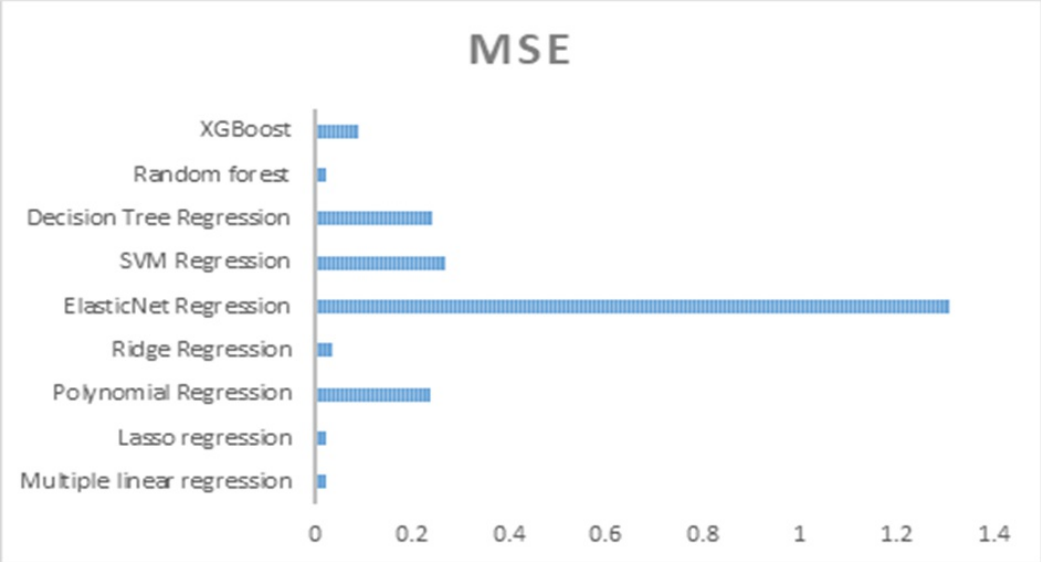

Figure 4. Mean Squared Error for the applied Algorithms

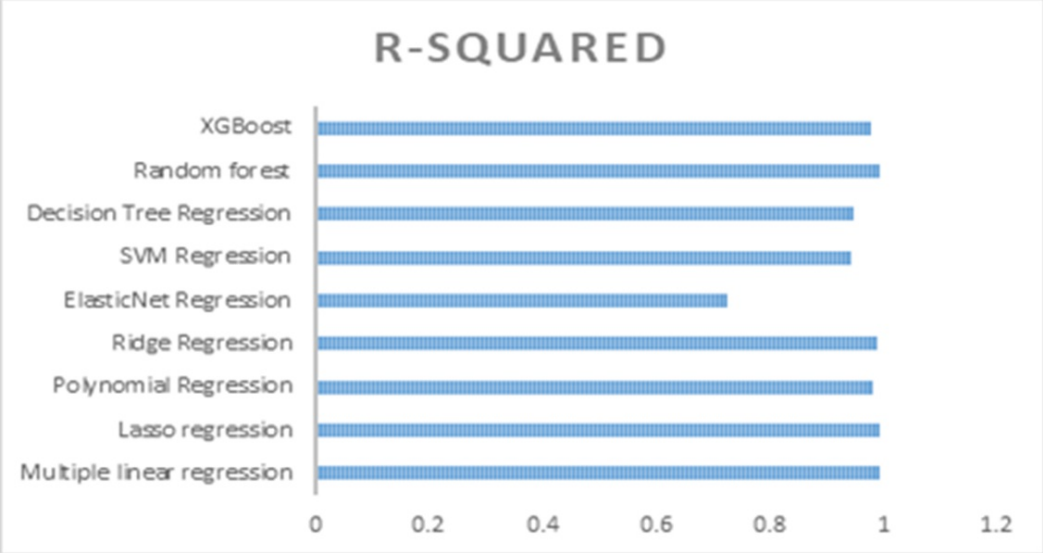

Figure 5. R-squared Error for the applied Algorithms

Future scenario predictions are tested using dependent variable data generated through the moving averages method, utilizing the average of the past five years. The dataset includes data up to 2019. The columns for planted area, harvested area, and average yield data from 2020 to 2026 are generated using moving averages, with temperature data also produced by the same method. The regression analysis model predicts future production based on this input data generated through the moving averages method (Basso and Liu, 2019).

Crop production from 2020 to 2026 has been predicted, with the results presented in Table 4. And Figure 6. Crop production forecasts are primarily based on in-season variables, statistical regressions between historical yields and field surveys. While field surveys remain the predominant method for forecasting crop yields in most countries, statistical regression using historical data is also significant in predicting crop yields (Chen and Guestrin, 2016).

Table 4. Prediction of crop production with a linear regression model

Year	Wheat	Barley	Lentil	Chickpea	Vetch
2020	175652.7	868176.6	22444.83	10285.62	7979.839
2021	184896.8	926955.3	13924.49	14804.36	8070.362
2022	200690.4	1247908	5489.831	14031.92	9654.13
2023	183149.2	1267007	2593.026	15373.58	5493.662
2024	166359.4	1299932	1873.166	11293.65	7236.073
2025	130390.9	1131512	18935.23	8763.661	6627.644
2026	80990.01	815371.2	31518.21	6792.26	4089.778

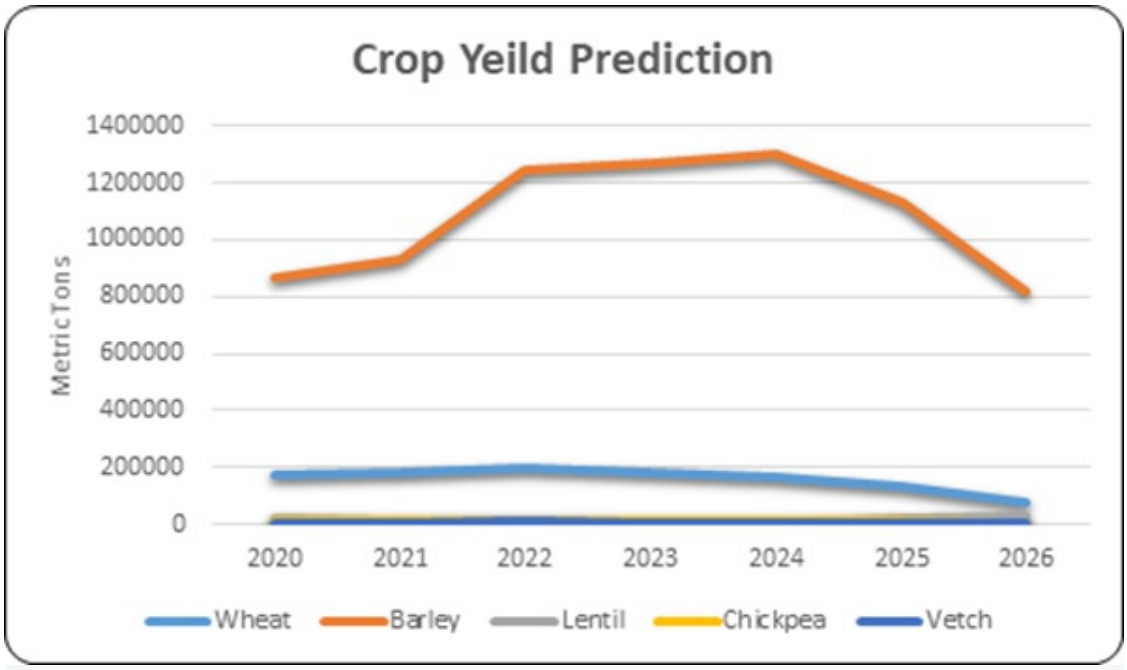

Figure 6. Prediction of crop yield (2020-2026)

V. Conclusion and Future Works

Predicting crop yields is essential for global food security and the economy. Given the increasing complexity of factors influencing plant growth, various machine-learning algorithms have been employed to support crop yield prediction. However, the complexity of these models makes their development quite challenging, and creating machine learning models for agricultural purposes is a tedious task due to the complexity involved.

This study integrates multiple models to create a robust and precise prediction system. Linear regression, which predicts a measurable response using various predictors and assumes a linear relationship between the response variable and predictors, is one of the methods used. Additionally, the adoption of machine learning methods for crop yield prediction and their potential to support sustainable growth is an exciting area of research. These findings could provide valuable insights for enhancing agricultural productivity.

The research gathered extensive datasets from the Department of Statistics Jordan and the Climate Change Knowledge Portal for training the proposed model. The utilization of large datasets has significantly enhanced the application of machine-learning techniques in predicting crop yields. Nine machine learning regression algorithms were thoroughly tested, and the results were quite promising. Notably, XGBoost, multiple linear regression, random forest, and Lasso regression exhibited impressive performance, achieving low mean squared errors of 0.092, 0.024, 0.023, and 0.023, respectively.

Future studies should highlight the potential of machine learning to revolutionize crop yield prediction and support sustainable agricultural practices. The future research trends in crop yield prediction using machine learning are quite promising and diverse. A further study could assess the long-term effects of integrating IoT and remote sensing, combining Internet of Things (IoT) devices, and remote sensing technologies with machine learning models to gather real-time data on soil conditions, weather, and crop health. This integration can enhance the accuracy and timeliness of yield predictions. Furthermore, it is crucial to conduct research aimed at determining climate change adaptation by developing models that predict crop yields under various climate change scenarios. This involves integrating climate models and long-term weather forecasts to assess the impact of changing climatic conditions on crop production. Additionally, there is a necessity to concentrate on leveraging big data analytics to manage the substantial amount of data generated from diverse sources, including satellite imagery, weather stations, and agricultural sensors.

This research has raised several important questions that demand further investigation. Here are the key areas of focus: First, it is crucial to explore advanced techniques such as deep learning and hybrid models. Models for deep learning, in particular recurrent neural networks (RNNs) and convolutional neural networks (CNNs), are increasingly essential for capturing complex patterns in large datasets. Second, we must emphatically utilize machine learning to drive precision agriculture practices, optimizing resource use (e.g., water, fertilizers) and minimizing environmental impact while maximizing crop yields. Third, it is imperative to incorporate genomic data to comprehensively understand the genetic factors influencing crop yields, thereby enabling the development of crop varieties that are more resilient to environmental stresses. Finally, there is an urgent need to focus on making ML models more interpretable and transparent to yield valuable insights for helping farmers and agricultural stakeholders understand the decision-making process of the models and trust their predictions.

About the Authors

Muneer Nusir, I was awarded my PhD degree from Brunel University-London (England) in July 2015, and my research interest is in Digital Service Co-design and Human-centered design. I have been an Assistant professor (Feb 2016-present) at the Information Systems Department of Prince Sattam Bin Abdulaziz University, Kingdom of Saudi Arabia. I attended a number of training courses and member of the IEEE-branch of Alkharj. I have been publishing several research Papers (International Conferences & refereed Journals). My research interest falls within Digital Service Co-design, Human-centered Design, Business Intelligence, Health Informatics, Smart city and business process variability

management. I have teaching, practical and research experience. My experience in teaching incorporates different universities in Jordan, the UK and Saudi Arabia.

Mohammad Alshirah Ph.D., is an Associate Professor at the Information Systems Department at Al al-Bayt University. He received his BSc. in Software Engineering from the Hashemite University, Jordan in 2007 and received his MSc in Computer Science from Al al-Bayt University, Jordan in 2010. He obtained his PhD in Software Engineering at the Department of Computer Science of the University of Leicester in the United Kingdom in 2016. He has a variety of academic and professional qualifications and experience. His research interests include, but are not limited to, Machine Learning, Software Usability and User Experience.

Sahar Al Mashaqbeh received a B.Sc. degree in Mechatronics engineering and an M.Sc. degree in Industrial Engineering from the Hashemite University in 2006 and 2011, respectively, and another M.Sc. in quality improvement engineering and Ph.D. degrees in industrial engineering from the University of Bradford, U.K., in 2019 and 2020, respectively. She currently serves as an assistant professor at the Hashemite University, Jordan. Furthermore, since November 2021, she has been working as an Honorary Visiting Research Fellow with the University of Bradford.

Rayeh Alghsoon Holds a Bachelor's degree (2012) and a Master's degree (2018) in Computer Engineering from the University of Jordan. A part-time lecturer at the University of Jordan (2012-2015) and Amman Al-Ahliyya University (2021-2024), a research assistant at Jordan University of Science and Technology (2018-2019), and currently serves as the Assistant Executive Secretary of the Association of Agricultural Research Institutions in the Near East and North Africa (AARINENA).

Statements and Declarations

This project was funded by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University award number 2023/01/26504

Other References

- A. Majumder, P. K. Kingra, R. Setia, S. P. Singh, and B. Pateriya, “Influence of land use/land cover changes on surface temperature and its effect on crop yield in different agro-climatic regions of Indian Punjab,” Geocarto Int., vol. 35, no. 6, pp. 663–686, 2020, doi: 10.1080/10106049.2018.1520927.
- H. Drucker, C. J. C. Surges, L. Kaufman, A. Smola, and V. Vapnik, “Support vector regression machines,” Adv. Neural Inf. Process. Syst., no. May 2018, pp. 155–161, 1997.
- R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.
- H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

- V. Anakha, A. S. Jinsu Mani, R. Mathew, and V. Williams, “A Prediction of Crop Yield using Machine Learning Algorithm,” Int. J. Eng. Res. Technol., vol. 9, no. 13, pp. 1072–1077, 2021, doi: 10.1109/ICOEI51242.2021.9452742.
- dosweb, “The Crop Production in Jordan,” 2023. https://dosweb.dos.gov.jo/agriculture/crops-statistics/

References

1. ^*E. J. Rifna, M. Dwivedi, D. Seth, R. C. Pradhan, P. K. Sarangi, and B. K. Tiwari, “Transforming the potential of renewable food waste biomass towards food security and supply sustainability,” Sustain. Chem. Pharm., vol. 38, no. 101515, 2024.*
2. a, b *V. Kumar et al., “Emerging challenges for the agro-industrial food waste utilization: A review on food waste biorefinery,” Bioresour. Technol., vol. 362, no. 127790, 2022.*
3. ^*D. Paudel et al., “Machine learning for large-scale crop yield forecasting,” Agric. Syst., vol. 187, no. October 2020, p. 103016, 2021, doi: 10.1016/j.agsy.2020.103016.*
4. a, b *M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour, and E. H. M. Aggoune, “Internet-of-Things (IoT)-Based Smart Agriculture: Toward Making the Fields Talk,” IEEE Access, vol. 7, pp. 129551–129583, 2019.*
5. ^*A. Ghosh, A. Kumar, and G. Biswas, “Exponential population growth and global food security: challenges and alternatives,” Bioremediation Emerg. Contam. from Soils, pp. 1–20, 2024.*
6. ^*T. Qureshi, M. Saeed, K. Ahsan, A. A. Malik, E. S. Muhammad, and N. Touheed, “Smart Agriculture for Sustainable Food Security Using Internet of Things (IoT),” Wirel. Commun. Mob. Comput., vol. 2022, 2022.*
7. ^*Biz4Intellia, “An end-to-end IoT Business Solution,” 2023. https://www.biz4intellia.com/*
8. a, b, c *G. S. Mittal and J. Zhang, “Prediction of freezing time for food products using a neural network,” Food Res. Int., vol. 33, no. 7, pp. 557–562, 2000, doi: 10.1016/S0963-9969(00)00091-0.*
9. a, b *E. J. Gonzalez-Sanchez, O. Veroz-Gonzalez, G. L. Blanco-Roldan, F. Marquez-Garcia, and R. Carbonell-Bojollo, “A renewed view of conservation agriculture and its evolution over the last decade in Spain,” Soil Tillage Res., vol. 146, no. PB, pp. 204–212, 2015, doi: 10.1016/j.still.2014.10.016.*
10. a, b, c, d *D. Elavarasan and P. M. Durairaj Vincent, “Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications,” IEEE Access, vol. 8, pp. 86886–86901, 2020, doi: 10.1109/ACCESS.2020.2992480.*
11. ^*J. P. Bharadiya, N. T. Tzenios, and M. Reddy, “Predicting Crop Yield Using Deep Learning and Remote Sensing,” J. Eng. Res. Reports, vol. 24, no. 12, pp. 29–44, 2023, doi: 10.9734/jerr/2023/v24i12858.*
12. ^*J. Moersdorf, M. Rivers, D. Denkenberger, L. Breuer, and F. U. Jehn, “The Fragile State of Industrial Agriculture: Estimating Crop Yield Reductions in a Global Catastrophic Infrastructure Loss Scenario,” Glob. Challenges, vol. 8, no. 1, pp. 1–17, 2024, doi: 10.1002/gch2.202300206.*
13. ^*P. Muruganantham, S. Wibowo, S. Grandhi, N. H. Samrat, and N. Islam, “A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing,” Remote Sens., vol. 14, no. 9, 2022, doi: 10.3390/rs14091990.*
14. ^*A. Sharifi, “Yield prediction with machine learning algorithms and satellite images,” J. Sci. Food Agric., vol. 101, no. 3,*

pp. 891–896, 2021, doi: 10.1002/jsfa.10696.

15. ^*N. Tantalaki, S. Souravlas, and M. Roumeliotis, "Data-Driven Decision Making in Precision Agriculture: The Rise of Big Data in Agricultural Systems," J. Agric. Food Inf., vol. 20, no. 4, pp. 344–380, 2019, doi: 10.1080/10496505.2019.1638264.*
16. ^*X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," Comput. Electron. Agric., vol. 121, pp. 57–65, 2016, doi: 10.1016/j.compag.2015.11.018.*
17. ^*J. Khazaei, G. R. Chegini, and M. Bakhshiani, "A novel alternative method for modeling the effects of air temperature and slice thickness on quality and drying kinetics of tomato slices: Superposition technique," Dry. Technol., vol. 26, no. 6, pp. 759–775, 2008, doi: 10.1080/07373930802046427.*
18. a, b*E. Elbasi et al., "Crop Prediction Model Using Machine Learning Algorithms," Appl. Sci., vol. 13, no. 16, 2023, doi: 10.3390/app13169288.*
19. ^*S. Mukherjee, A. Mishra, and K. E. Trenberth, "Climate Change and Drought: a Perspective on Drought Indices," Curr. Clim. Chang. Reports, vol. 4, no. 2, pp. 145–163, 2018, doi: 10.1007/s40641-018-0098-x.*
20. ^*X. Xu et al., "Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China," Ecol. Indic., vol. 101, no. July 2018, pp. 943–953, 2019, doi: 10.1016/j.ecolind.2019.01.059.*
21. a, b*T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," Comput. Electron. Agric., vol. 177, no. January, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.*
22. ^*H. Burdett and C. Wellen, "Statistical and machine learning methods for crop yield prediction in the context of precision agriculture," Precis. Agric., vol. 23, no. 5, pp. 1553–1574, 2022, doi: 10.1007/s11119-022-09897-0.*
23. ^*Y. Miao, D. J. Mulla, and P. C. Robert, "Identifying important factors influencing corn yield and grain quality variability using artificial neural networks," Precis. Agric., vol. 7, no. 2, pp. 117–135, 2006, doi: 10.1007/s11119-006-9004-y.*
24. ^*D. G. B. Alexandra N. Kravchenko, "Correlation of Corn and Soybean Grain Yield with Topography and Soil Properties," Agron. J., vol. 92, no. 1, 2000, doi: https://doi.org/10.2134/agronj2000.92175x.*
25. ^*A. Morales and F. J. Villalobos, "Using machine learning for crop yield prediction in the past or the future," Front. Plant Sci., vol. 14, no. March, pp. 1–13, 2023, doi: 10.3389/fpls.2023.1128388.*
26. ^*P. Sharma, P. Dadheech, N. Aneja, and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning," IEEE Access, vol. 11, no. September, pp. 111255–111264, 2023, doi: 10.1109/ACCESS.2023.3321861.*
27. ^*L. BREIMAN, "Random Forests," Mach. Learn., vol. 45, pp. 5–32, 2001, doi: 10.1007/978-3-030-62008-0_35.*
28. ^*Y. Li et al., "A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering," Int. J. Appl. Earth Obs. Geoinf., vol. 118, no. January, p. 103269, 2023, doi: 10.1016/j.jag.2023.103269.*
29. ^*P. P. Jorvekar, S. K. Wagh, and J. R. Prasad, "Predictive modeling of crop yields: a comparative analysis of regression techniques for agricultural yield prediction," Agric. Eng. Int. CIGR J., vol. 26, no. 2, pp. 125–140, 2024.*
30. ^*S. Suthaharan, Machine Learning Models and Algorithms for Big Data Classification. Hamburg: Springer, 2016.*
31. ^*B. Panigrahi, K. C. R. Kathala, and M. Sujatha, "A Machine Learning-Based Comparative Approach to Predict the*

Crop Yield Using Supervised Learning with Regression Models," Procedia Comput. Sci., vol. 218, no. 2022, pp. 2684–2693, 2023, doi: 10.1016/j.procs.2023.01.241.

32. ^*E. Kamir, F. Waldner, and Z. Hochman, "Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods," ISPRS J. Photogramm. Remote Sens., vol. 160, no. November 2019, pp. 124–135, 2020, doi: 10.1016/j.isprsjprs.2019.11.008.*
33. a, b*J. H. Jeong et al., "Random forests for global and regional crop yield predictions," PLoS One, vol. 11, no. 6, pp. 1–15, 2016, doi: 10.1371/journal.pone.0156571.*
34. ^*D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: A survey," Comput. Electron. Agric., vol. 155, no. October, pp. 257–282, 2018, doi: 10.1016/j.compag.2018.10.024.*
35. ^*Climate Change Knowledge Portal, "Jordan's climate," 2023. https://climateknowledgeportal.worldbank.org/country/jordan/climate-data-historical*
36. ^*K. M. F. Elsayed, T. Ismail, and N. S. Ouf, "A Review on the Relevant Applications of Machine Learning in Agriculture," Ijireeice, vol. 6, no. 8, pp. 1–17, 2018, doi: 10.17148/ijireeice.2018.681.*
37. ^*V. Meshram, K. Patil, V. Meshram, D. Hanchate, and S. D. Ramkteke, "Machine learning in agriculture domain: A state-of-art survey," Artif. Intell. Life Sci., vol. 1, no. August, p. 100010, 2021, doi: 10.1016/j.ailsci.2021.100010.*
38. ^*M. Nusir, "Research Data /Agriculture-Jordan," 2024. https://drive.google.com/file/d/1lguMLagDm3eQ0QNQyAQUnEe5oUm-WZFV/view*
39. a, b*F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," Comput. Stat., vol. 37, no. 5, pp. 2671–2692, 2022, doi: 10.1007/s00180-022-01207-6.*
40. ^*J. W. Tukey, Exploratory Data Analysis. 1977. doi: 10.1007/978-3-642-41714-6_52111.*
41. ^*https://scikit-learn.org/, "train_test_split," 2018. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html*
42. a, b*T. W. Crowther et al., "Mapping tree density at a global scale," Nature, vol. 525, no. 7568, pp. 201–205, 2015, doi: 10.1038/nature14967.*
43. ^*T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second. Springer, 2009.*