Ahliya Journal of Business Technology and MEAN Economies **Vol. 01 Issue 01 (2024)**
[Published by Palestine Ahliya University] **ISSN: 3007-9691**

Predicting Crop Yield Productivity Using Machine Learning Algorithms: A Comparison of Linear and Non-linear Approaches

Murad Zeer[1], Mutaz Rasmi Abu Sara[1], Jawad H Alkhateeb[2] and Mohammad F. J. Klaib[3]

[1] Faculty of Engineering and Information Technology, Palestine Ahliya University (Palestine)
✉ muradzeer@paluniv.edu.ps
✉ moutaz.a@paluniv.edu.ps

[2] College of Computer Engineering and Science, Prince Mohammad Bin Fahd University, Khobar 31952 (Saudi Arabia)
✉ jalkhateeb@pmu.edu.sa

[3] Intelligent Systems Engineering, Middle East University (Jordan)
✉ mklaib@meu.edu.jo

Received:25/05/2024 Accepted:17/06/2024 Published:15/07/2024

Abstract:

Predicting crop yield productivity is crucial for farmers and the agricultural sector to gain insights into crop productivity and returns. With advancements in technology and artificial intelligence, predicting crop yield using machine learning algorithms has become an important innovation. This study aimed to predict crop yield productivity using various machine learning algorithms and techniques. The dataset was sourced from Kaggle, preprocessed, and analyzed using linear algorithms such as Linear Regression, LASSO, and Ridge, as well as non-linear algorithms including SVR, KNN Regressor, and Polynomial Regression. Mean Squared Error (MSE) was computed to evaluate algorithms performance. Comparing the efficacy of linear versus non-linear algorithms on the dataset revealed that non-linear algorithms outperformed linear ones, indicating that the dataset's non-linear nature. Therefore, non-linear machine learning algorithms like SVR, KNN Regressor, and Polynomial Regression were recommended for better accuracy. Among these, KNN Regressor performed the best with an MSE of 0.00025, followed by SVR and Polynomial Regression with MSE values of 0.00142 and 0.0024, respectively.

Keywords: *Crop Yield Productivity; Machine Learning; Regression; SVR; Regression KNN; LASSO; Ridge; Polynomial Regression.*

Ahliya Journal of Business Technology and MEAN Economies Vol. 01 Issue 01 (2024)
[Published by Palestine Ahliya University] ISSN: 3007-9691

التنبؤ بإنتاجية المحاصيل باستخدام خوارزميات التعلم الآلي: مقارنة بين النهج الخطي وغير الخطي

مراد الزير[1]، معتز أبو سارة[2]، جواد الخطيب[3] محمد كليب[3]

[1] كلية الهندسة وتكنولوجيا المعلومات، جامعة فلسطين الأهلية (فلسطين)

muradzeer@paluniv.edu.ps

moutaz.a@paluniv.edu.ps

[2] كلية هندسة وعلوم الحاسوب، جامعة الأمير محمد بن فهد، الخبر 31952 (المملكة العربية السعودية)

jalkhateeb@pmu.edu.sa

[3] هندسة النظم الذكية، جامعة الشرق الأوسط (الأردن)

mklaib@meu.edu.jo

تاريخ الاستلام:2024/05/25 تاريخ القبول:2024/06/17 تاريخ النشر:2024/07/15

ملخص:

يُعدّ التنبؤ بإنتاجية المحاصيل أمرًا بالغ الأهمية بالنسبة للمزارعين والقطاع الزراعي للحصول على نظرة ثاقبة لإنتاجية المحاصيل وعوائدها. مع التقدّم في التكنولوجيا والذكاء الاصطناعي، أصبح التنبؤ بإنتاجية المحاصيل باستخدام خوارزميات التعلم الآلي ابتكارًا مهمًا. تهدف هذه الدراسة إلى التنبؤ بإنتاجية المحاصيل باستخدام خوارزميات وتقنيات التعلم الآلي المختلفة. تم الحصول على مجموعة البيانات من Kaggle، وتمت معالجتها مسبقًا وتحليلها باستخدام خوارزميات خطية، مثل: Linear Regression وLASSO وRidge، بالإضافة إلى خوارزميات غير خطية، مثل: SVR وKNN Regressor وPolynomial Regression. تم حساب متوسط الخطأ التربيعي (MSE) لتقييم أداء الخوارزميات. كشفت مقارنة فاعلية الخوارزميات الخطية مقابل الخوارزميات غير الخطية في مجموعة البيانات أن الخوارزميات غير الخطية تفوقت على الخوارزميات الخطية، مما يشير إلى الطبيعة غير الخطية لمجموعة البيانات. ولذلك، يُوصى باستخدام خوارزميات التعلم الآلي غير الخطية مثل SVR وKNN Regressor وPolynomial Regression للحصول على دقة أفضل. من بين هذه، كان أداء KNN Regressor هو الأفضل مع MSE بقيمة 0.00025، يليه SVR والانحدار متعدد الحدود بقيم MSE تبلغ 0.00142 و0.0024، على التوالي.

الكلمات المفتاحية: *إنتاجية المحاصيل؛ التعلم الآلي؛ الانحدار؛ SVR؛ الانحدار KNN؛ LASSO؛ Ridge؛ الانحدار متعدد الحدود.*

1. Introduction:

Predicting crop yield productivity is crucial both from a research and practical standpoints, involving a complex set of steps aimed at estimating returns and intensifying efforts for sustainable use of natural resources (Van Klompenburg et al., 2020; Elavarasan & Vincent, 2020; Paudel et al., 2021). Financially and quantitatively estimating crop productivity is essential for determining appropriate strategic plans for export policies and imports in the agricultural sector, aiming to increase agricultural income. Given its significance, technological advancements, particularly artificial intelligence, have been leveraged, with machine learning algorithms playing a prominent role (Rashid et al., 2021).

Traditional regression poses challenges in handling such complexities, but with the advent of machine learning, these challenges have been addressed, minimizing their negative impact (Hu et al., 2023). Machine learning contributes by facilitating predictive operations through establishing relationships between input variables and outputs, employing methods to learn prediction strategies based on inputs (Shahhosseini et al., 2021). Predictive techniques using machine learning help form a clear picture of crop productivity and suggest avenues for improvement to enhance farmers' profitability (Kumar et al., 2020). Despite the availability of various machine learning algorithms, determining their superiority requires handling specific datasets to achieve optimal accuracy (Iniyan et al., 2023). To ensure accurate prediction, high-quality inputs related to factors such as soil, weather, and others are essential in determining yield (Kheir et al., 2024), highlighting the preference for machine learning due to its capability to handle complex relationships between crop features and yield (Morales & Villalobos, 2023).

Previous studies have employed machine learning algorithms for predicting crop yield productivity, as evidenced by studies such as Abbas et al. (2020), Jhajharia et al. (2023), Hu et al. (2023), Shafi et al. (2023), and Manjunath & Palayyan (2023). Hence, this study aims to predict crop yield productivity using six machine learning algorithms, three linear (Linear Regression, LASSO, Ridge) and three non-linear (SVR, Regression KNN, Polynomial Regression).

2. Previous Studies:

Several previous studies have focused on predicting crop yield productivity using machine learning algorithms, each employing different methodologies and algorithms.

Abbas et al. (2020) studied potato tuber yield prediction using linear regression, KNN, SVR, and Elastic Net (EN). The study concluded that SVR outperformed the other algorithms in predicting potato tuber yield.

Jhajharia et al. (2023) applied various algorithms including Lasso regression, SVM, Gradient Boosting, LSTM, and Random Forest for predicting random forest crop yield. The study found that Random Forest performed the best, followed by SVM.

Hu et al. (2023) compared a Bayesian ensemble model (BM) with several machine learning algorithms such as SVM, Random Forests, and others for yield prediction. The Bayesian ensemble model was superior to the other algorithms tested in the study.

Shafi et al. (2023) evaluated three machine learning algorithms: Extreme Gradient Boosting (XGB), Random Forest, and LASSO. The study demonstrated that LASSO achieved the best performance compared to the other algorithms evaluated.

Manjunath & Palayyan (2023) investigated five machine learning algorithms: Random Forest, Linear Regression, SVM, Decision Tree, and XGBoost for crop yield prediction. Their study concluded that a hybrid model combining Decision Tree (DT), Random Forest (RF), and XGBoost achieved the highest accuracy among the models tested.

These studies collectively highlight the diverse applications of machine learning algorithms in predicting crop yield productivity, showcasing the variability in algorithm performance based on the specific crop and dataset characteristics.

3. Methodology:

3.1 Data:

The study utilized a dataset sourced from Kaggle (https://www.kaggle.com), comprising 2200 rows distributed across 22 different agricultural crops. The dataset consists of 9 features, including 8 inputs and one output variable, labeled as "Yilde". The sample characteristics are detailed in Table 1:

Table 1: Dataset Description

	Description
Source	Kaggle
Shape	(2200 rows × 9 Columns)
Input	Nitrogen, Phosphorus, Potassium, Temperature, Humidity, pH_Value, Rainfall, Crop
Output	Yield
Category	Crop = (22)

3.2 Data Preprocessing:

3.2.1 Missing Value Detection:

Missing values in datasets are known to adversely impact model quality and statistical outcomes. Detecting and handling missing values are crucial for decision-making regarding their treatment. In this study, missing values (NaN) were checked for in each cell of the dataset, and it was found that there were no missing values presented.

3.2.2 Outlier Detection:

Outliers are data points that significantly deviate from the expected behavior of the system (Siddiqi et al., 2023). They are identified using various methods, including Boxplot analysis. Outliers were detected in the features "Phosphorus", "Potassium", "Temperature", "Humidity", "ph_Value", and "Rainfall". These outliers are illustrated in Figure 1.

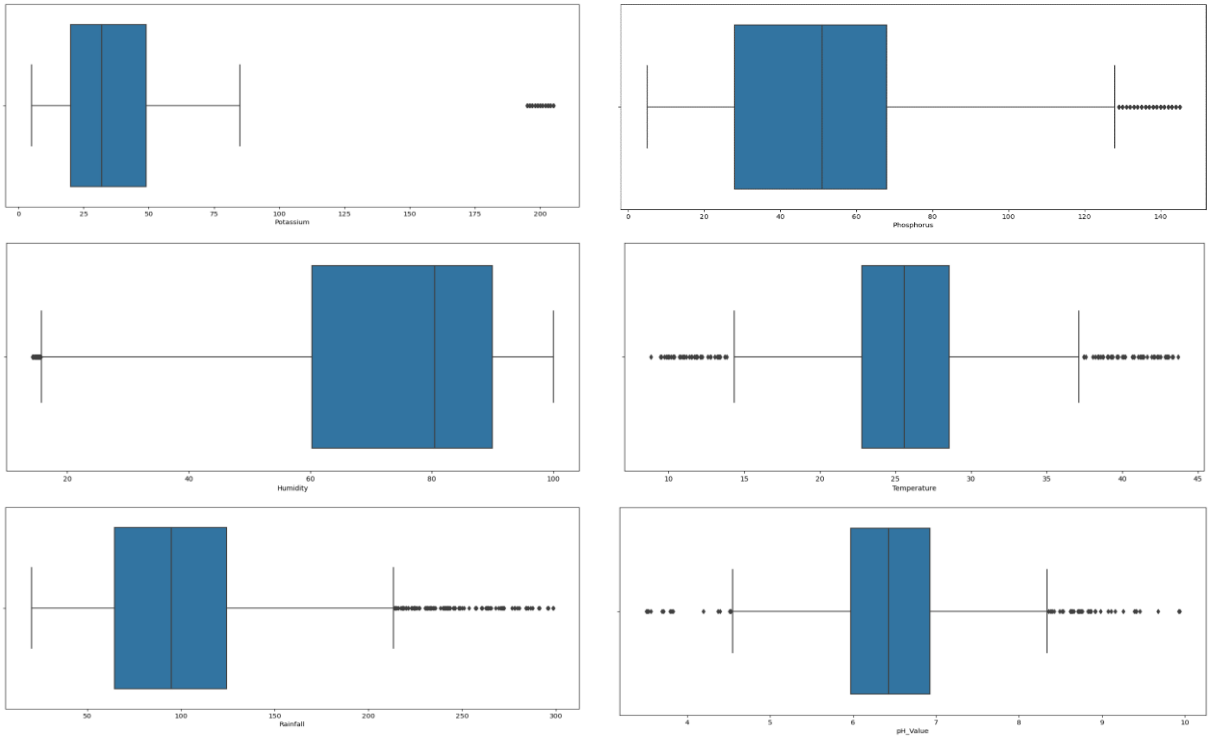

Figure (1): Outliers in Dataset Features

The outliers in the dataset features were processed by transforming the skewness of the data to achieve a symmetric distribution. Logarithmic transformation was applied to columns "Phosphorus" and "Potassium" to reduce skewness and enhance distribution consistency. Additionally, square root transformation was utilized for the "Humidity" column, which exhibited leftward skewness. Outliers in the "Yield" were handled using Interquartile Range (IQR) method, where values exceeding Q3 were compensated with Q3 and values below Q1 were compensated with Q1. This approach ensures that all dataset values are preserved across all 22 agricultural crops, maintaining data integrity for consistent prediction processes.

3.3.3 Scaling:

The "Yield" values were scaled using MinMaxScaler to transform them into the range of [0-1]. This scaling method helps prevent large values from dominating over smaller ones during statistical modeling of the data, thereby improving performance. The scaling transformation is performed according to the following equation:

$$Scaled\ (X_{Yield}) = \frac{x - \min\ (x)}{\max(x) - \min\ (x)} \quad (1)$$

After applying this equation, Yield is transformed to be within the range of 0 to 1, where the value 0 represents the minimum for Yield and the value 1 represents the maximum.

3.3.4 Principal Component Analysis (PCA):

PCA is used to reduce dimensions by projecting each data point onto a small number of principal components that capture the most effective variance (Dash et al., 2023). This results in lower-dimensional data with maximum possible variance retained. In this study, dimensionality reduction was achieved using two principal components, ordered according to their contribution in explaining the data variance.

3.3.5 One-Hot-Encoding:

The categorical feature "Crop", which represents 22 agricultural crops without a specific order, was encoded into numerical form using One-Hot-Encoding. This transformation handled "Crop" as a categorical variable by encoding it into a matrix format.

3.3 Train-Test Split:

In machine learning models, datasets are typically split into training and testing sets. It's common practice to allocate (80%) of the data for training and (20%) for testing (Joseph & Vakayil, 2022). In this study, the data was divided into training and testing sets with an 80% training and 20% testing ratio.

3.4 Algorithms:

In this study, six algorithms were utilized, categorized into two sections: linear algorithms, which include Linear Regression, LASSO, and Ridge; and non-linear algorithms, comprising SVR, KNN Regressor, and Polynomial Regression.

3.4.1 Linear Regression:

Linear regression is a mathematical approach for examining the relationship between the studied variables and understanding the expected effects between independent and dependent variables (Maulud & Abdulazeez, 2020).

3.4.2 Ridge Regression:

Ridge regression is an extension of linear regression, using the same prediction process as ordinary least squares. Ridge tests parameters (W) to predict training data while fitting an additional constraint where these parameters are very small, meaning values of W are close to zero (Müller & Guido, 2016).

3.4.3 LASSO Regression:

LASSO (Least Absolute Shrinkage and Selection Operator) is an alternative to Ridge regression, constraining parameters to be close to zero but in a slightly different manner. This algorithm zeros out some coefficients, effectively disregarding certain features entirely to facilitate model interpretation (Müller & Guido, 2016).

3.4.4 Support Vector Regression (SVR):

SVR operates similarly to Support Vector Machine (SVM) in aiming to minimize errors and efficiently predicting time series data (Bathla, 2020).

3.4.5 KNN Regressor:

Regression KNN is commonly used for prediction due to its simplicity, calculating the average output values of the nearest K neighbors (Öngelen & İnkaya, 2023).

3.4.6 Polynomial Regression:

Polynomial regression is utilized to adjust linear regression to handle nonlinear relationships between variables. It incorporates multiple degrees of variables to better express nonlinear relationships within data.

4. Results and Discussion:

Six machine learning algorithms were employed to predict the crop yield in this study. Mean Squared Error (MSE) was calculated to determine the superiority of the suitable algorithm for predicting crop yield, according to the following equation:

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \quad (2)$$

Where y_i represents the actual values of crop yields, $\hat{y}_i$ denotes the predicted values from the model, and n is the number of points in the testing set. After calculating MSE for each algorithm, the algorithm with the lowest value is considered superior, indicating that the model predicts more accurately and closely resembles the actual values.

R^2 (R-squared) was computed to evaluate the model's performance during training and testing. It expresses the proportion of the variance in the dependent variable (Yield) that is explained by the model, and is defined by the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{2}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{2}(y_i - \bar{y}_i)^2} \quad (3)$$

Where y_i represents the actual values of Yield, $\hat{y}_i$ denotes the predicted values from the model, $\bar{y}_i$ is the average of the actual Yield values, n is the number of data points in the testing dataset, and R2 (R-squared) ranges between 0 and 1. A value closer to 1 indicates that the model explains the variance in the actual data more accurately.

Table (2): MSE results for the algorithms used in the study

	(MSE) for Training Set	(MSE) for Test Set	Training Set Score (R^2)	Test Set Score (R^2)
Linear regression	0.0236	0.0206	0.468	0.4737
Ridge (Alpha: 1000)	0.024	0.0213	0.46	0.457
Lasso (Alpha: 0.1)	0.023	0.0209	0.466	0.465
SVR	0.0013	0.00142	0.97	0.96
KNN Regressor (k=85)	0.0004	0.00025	0.99	0.99
Polynomial Regression (degree=2)	0.00266	0.0024	0.94	0.938

In this study, six different algorithms were used to predict crop yield, and their performance was evaluated using two main metrics: Mean Squared Error (MSE) and R-squared (R^2). The goal of this evaluation was to understand each algorithm's ability to predict yield accurately and interpret the data. Starting with linear algorithms such as Linear Regression, and other regression techniques like Ridge and Lasso, these algorithms showed poor performance. They achieved MSE values ranging between 0.02 and 0.03 for both training and test sets. The R-squared values were approximately 0.46-0.47, indicating a low ability to predict yield and explain variance in the data.

On the other hand, non-linear algorithms such as SVR, KNN Regressor, and Polynomial Regression demonstrated excellent performance. Comparatively, SVR and KNN Regression achieved very low MSE values (around 0.001-0.0004) and high R-squared values exceeding 0.95 for both training and test sets. This suggests their high capability to adapt to the data and provide accurate predictions of crop yield. The figure 2 illustrates the MSE results for the algorithms used in the study.

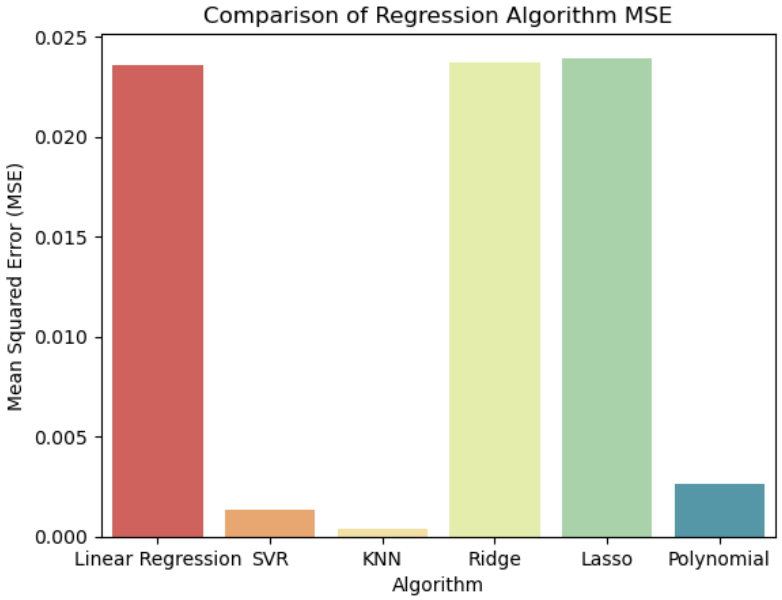

Figure (2): Comparison of Algorithms Used in the Study

The results illustrate varied performance among the six algorithms used to predict agricultural crop yield, necessitating a deep understanding of the different methods that handle data and provide predictions.

4.1 Linear Algorithms and Underfitting:

The linear algorithms used in the study, such as Linear Regression, Ridge, and Lasso, are characterized by their simplicity in interpreting relationships between variables. While they provide acceptable results, they may suffer from the phenomenon of underfitting at times. This means they may not capture the true complexities in the data, especially if the relationship between variables is non-linear or more complex than can be represented linearly. For instance, if there are non-linear interactions between environmental variables such as temperature, humidity, and fertilizers, linear algorithms may not be sufficient to provide accurate predictions, figure 3 shows the MSE results for the used algorithms.

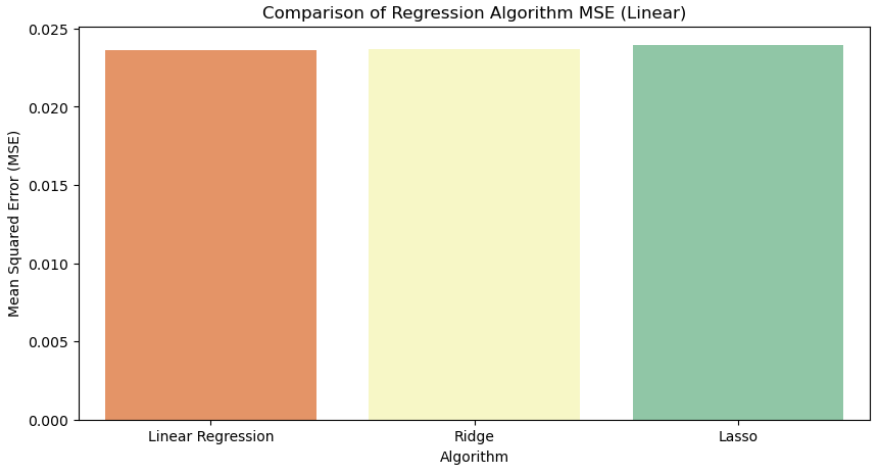

Figure (3): Comparison of linear Algorithms

4.2 Non-linear Algorithms:

SVR, KNN Regression, and Polynomial Regression: These non-linear algorithms outperformed linear algorithms in experimental results. Non-linear algorithms are typically more capable of handling complexities and non-linear relationships in data. For example, SVR uses mathematical models that allow flexible adaptation to data in various ways based on statistical support, enabling high accuracy in predictions. On the other hand, KNN Regression relies on a simple idea of averaging values from nearest neighbors, effectively managing complexities and non-linear patterns.

Polynomial Regression, another technique, offers greater flexibility in representing non-linear complexities due to its ability to model high-order interactions between variables. By incorporating higher-order terms, Polynomial Regression can provide accurate models that surpass linear models in predicting data.

From figure 4, it becomes evident that KNN Regressor achieved the best results, followed by SVR and then Polynomial Regression, with R^2 values for training sets being 0.99, 0.97, and 0.94 respectively.

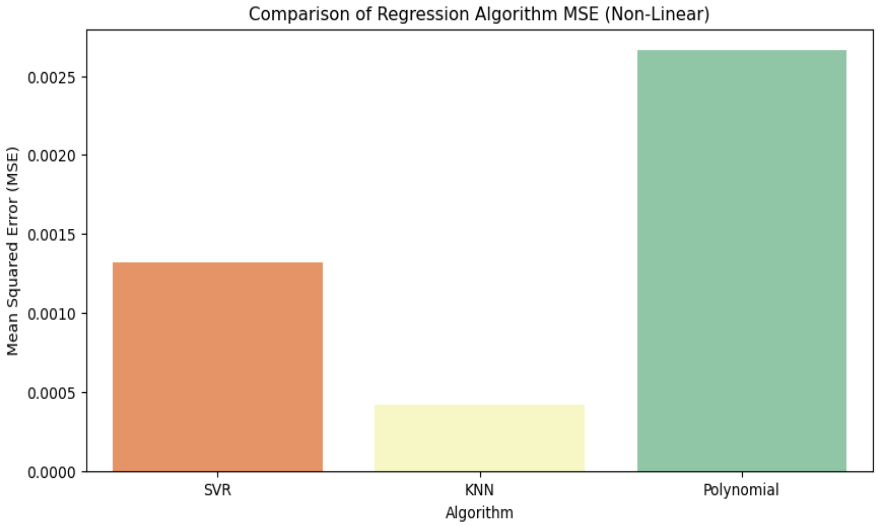

Figure (4): Comparison of MSE for Non-linear Algorithms

Non-linear algorithms demonstrated superiority over linear algorithms in this study due to their capability to handle non-linear complexities in agricultural data. This highlights the importance of selecting an appropriate model that fits the nature and complexities of the data to achieve optimal predictive performance.

5. Conclusion:

The results obtained from using six different algorithms to predict agricultural crop productivity showed that non-linear algorithms such as SVR, Regression KNN, and Polynomial Regression outperformed linear algorithms like Linear Regression, Lasso, and Ridge in delivering better performance. This superiority stems from their ability to handle complexities and non-linear relationships between variables, thereby achieving higher accuracy in crop productivity predictions. Regression KNN emerged with the best results, followed by SVR and then Polynomial Regression, with R^2 values for training sets being 0.99, 0.97, and 0.94 respectively.

6. Recommendations:

The study recommends conducting future studies on the same dataset using advanced artificial intelligence algorithms. Researchers should carefully select algorithms that are suitable for the dataset used. The study identified underfitting in the use of linear algorithms, which does not imply preprocessing issues but rather their limited ability to handle the dataset's non-linearity. This underscores the appropriateness of non-linear algorithms for handling such data.

References:

Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, *10*(7), 1046.

Bathla, G. (2020, November). Stock Price prediction using LSTM and SVR. In *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 211-214). IEEE.

Elavarasan, D., & Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE access*, *8*, 86886-86901.

Dash, C. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, *6*, 100164.

Hu, T., Zhang, X., Bohrer, G., Liu, Y., Zhou, Y., Martin, J., ... & Zhao, K. (2023). Crop yield prediction via explainable AI and interpretable machine learning: Dangers of black box models for evaluating climate change impacts on crop yield. *Agricultural and Forest Meteorology*, *336*, 109458.

Iniyan, S., Varma, V. A., & Naidu, C. T. (2023). Crop yield prediction using machine learning techniques. *Advances in Engineering Software*, *175*, 103326.

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *218*, 406-417.

Joseph, V. R., & Vakayil, A. (2022). SPlit: An optimal method for data splitting. *Technometrics*, *64*(2), 166-176.

Kheir, A., Nangia, V., Elnashar, A., Devakota, M., Omar, M., Feike, T., & Govind, A. (2024). Developing automated machine learning approach for fast and robust crop yield prediction using a fusion of remote sensing, soil, and weather dataset. *Environmental Research Communications*.

Kumar, Y. J. N., Spandana, V., Vaishnavi, V. S., Neha, K., & Devi, V. G. R. R. (2020, June). Supervised machine learning approach for crop yield prediction in agriculture sector. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)* (pp. 736-741). IEEE.

Manjunath, M. C., & Palayyan, B. P. (2023). An Efficient Crop Yield Prediction Framework Using Hybrid Machine Learning Model. *Revue d'Intelligence Artificielle*, *37*(4).

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140-147.

Morales, A., & Villalobos, F. J. (2023). Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, *14*, 1128388.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.".

Öngelen, G., & İnkaya, T. (2023). LOF weighted KNN regression ensemble and its application to a die manufacturing company. *Sādhanā*, *48*(4), 246.

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, *187*, 103016.

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE access*, *9*, 63406-63439.

Shafi, U., Mumtaz, R., Anwar, Z., Ajmal, M. M., Khan, M. A., Mahmood, Z., ... & Jhanzab, H. M. (2023). Tackling food insecurity using remote sensing and machine learning based crop yield prediction. *IEEE Access*.

Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific reports*, *11*(1), 1606.

Siddiqi, S., Qureshi, F., Lindstaedt, S., & Kern, R. (2023). Detecting Outliers in Non-IID Data: A Systematic Literature Review. *IEEE Access*. 70333-70352.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709.