Bangladesh J. Environ. Sci., Vol. 46, 134-139, 2024 @ BAED ISSN 1561-9206

A COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR CROP YIELD PREDICTION IN BANGLADESH

[1]Jannatunnaher, [1]M. Sultana and [2]I. A. Badhan
[1]Department of Computer Science and Engineering, Begum Rokeya University, Rangpur
[2]Department of Electrical and Electronic Engineering, Begum Rokeya University, Rangpur

ABSTRACT

Agriculture is often regarded as the principal means of ensuring food provision worldwide. Besides being the supplier of food, this sector has contributions to supply raw materials in different industries in our country. In light of declining crop production and food shortages around the world, one of the most important criteria in agriculture today is predicting future yields and selecting the right crop for the right land at the right time. Therefore, in this study we have proposed a method that will help us by predicting crop production based on previous year data analysis on some predictive parameters using machine learning. Although several studies have been conducted in this area, most of them are not in the context of Bangladesh. In this study, we applied Support Vector Machine (SVM), Random Forest (RF) and Lasso algorithms for crop yield prediction using our collected dataset of major rice crops (Aush, Aman, and Boro). The dataset contains the weather factors (temperature, rainfall, humidity) data from 2015 to 2022 as predictors. A comparative analysis of machine learning algorithms is performed based on evaluation metrics (MAE, MSE, RMSE) values. RF shows the best crop yield prediction results with least error values for two types of crops: Aman (MAE = 0.02511, MSE = 0.00138, RMSE = 0.0371) and Boro (MAE = 0.02576, MSE = 0.00112, RMSE = 0.03352). However, Lasso notes the least error values for Aus which account for MAE of 0.030260, MSE of .00124 and RMSE of 0.03524. From our experiments, we can estimate that RF demonstrates the optimum result while taking into account several predicting properties and performance metrics.

Key words: Rice crop, machine Learning, SVM, RF, Lasso

Introduction

Agriculture is considered as a crucial industry in Bangladesh concerning GDP. During the fiscal year (FY) 2021-22, the agriculture industry saw a growth rate of 3.05%, which increased to 3.37% in the subsequent financial year 2022-23. Based on preliminary calculations, the agriculture industry is projected to see a growth rate of 3.21% in FY 2023–24 (Bangladesh Economic Review, 2024). In FY 2022-23, in all four sub-sectors of agriculture, the growth rate of crops, horticulture and fisheries, animal husbandry, forestry and allied services, increased compared to the previous fiscal year, but it is estimated that this year every one of the sub-sectors will decrease. So, the importance is being given to increasing the production of alternatives to the imported crops and increasing the quantity of export-oriented crops will be given more importance according to the previous plan (NAEP, 1996). Improving crop yield prediction is crucial for enhancing agricultural efficiency. Farmers and stakeholders may anticipate educated decisions about resource allocation, planting strategies, and harvest planning, which can eventually result in increased yields and reduced resource wastage (Elbasi *et al*., 2023). Advancements in machine learning and data analytics have created new prospects for enhancing agricultural output estimates. Since the last few years, researchers have been focusing heavily about the utilization of machine learning, image processing and information-focused technologies in various sectors of agriculture such as automatic prediction of plant diseases, correct crop identification for a field, forecasting of crop yields to stabilize food supply and so on (Liakos *et al*., 2018). A significant amount of study was conducted to identify appropriate crops for a particular crop field considering the current climate to increase the yield rate (Yesugade *et al*., 2019; Tom, 2020). Researchers have also emphasized on forecasting crop yields to stabilize the amount of food supply for a country (Tamasiga *et al*., 2023). Andrew Crane-Droesch combined two algorithms, such as Back

Propagation Network (BPN) along with Kohonen Self-Organizing Map (Kohenon's SOM) for predicting produce from agriculture depending on soil and fertilizer parameters, which can assist the farmers (Crane-Droesch, 2018). Our research leverages innovative technologies to aid for enhancement of our agronomy. By comparing the results of multiple algorithms, we can discover the most fruitful schemes and potentially come upon novel ways to address challenges to predict yield of quality and quantity of produce. The availability of agricultural data, including historical records, weather data, and sensor data, makes this data science approaches increasingly relevant (Liang and Shah, 2023). Different varieties of Aus, Aman, and Boro have been identified by researchers as capable of being cultivated year-round to address the food scarcity issue of Bangladesh (Rahman *et al*., 2020). Overall, the motivation of our research lies in the proficient to revolutionize agriculture by harnessing the power of algorithms and data analysis of major crops like rice in Bangladesh to drive better outcomes for farmers, consumers, and the environment. It addresses critical challenges in agriculture, making it more sustainable, efficient, and resilient amid global food security apprehensions.

Materials and Methods

This study's methodology consists of two main tasks: gathering pertinent data and applying diverse machine learning algorithms to forecast the cultivation. According to the fundamental rule of machine learning, the data is split in an 80:20 ratio for training and testing purposes. The results of the machine learning algorithms, like Support Vector Machine (SVM), Random Forest (RF) and Lasso, are analyzed using assessment criteria like mean squared error (MSE), mean absolute error (MAE) and root mean square error (RMSE). The manifest overview of the crop yield forecasting system is shown in Fig. 1.

Dataset Collection: The collected dataset contains the history of major rice crops (Aus, Aman and Boro) production for different districts of Bangladesh from the time period 2015 to 2022. The data has been collected from the National Aeronautics and Space Administration (NASA), the yearbook of the Bangladesh Agricultural Development Corporation (BADC), the Ministry of Agriculture (Bangladesh) and the Bangladesh Bureau of Statistics (BBS). In the dataset, there are a total of eight attributes, such as year, district, precipitation, highest temperature, lowest temperature, humidity, area and production.

Machine Learning Techniques

Support Vector Machine (SVM): SVM is a mathematical constructor, an algorithm designed to optimize a certain mathematical function when applied to a designated dataset. SVM-based classification may be articulated using four fundamental concepts: separating hyperplane, maximum-margin hyperplane, soft margin, and kernel function (Noble, 2006). The basic goal of the SVM method is to identify the optimal hyperplane in an N-dimensional space that intercepts the data points into distinct attribute classes (Meyer and Wien, 2012). The hyperplane seeks to preserve the maximum buffer between the closest points of different classes. The overall number of distinct characteristics dictates the dimension of the hyperplane; for example, if there are two features, the hyperplane becomes a line and it transforms into a 2D plane for three input features. Support Vector Machines (SVMs) have been successfully employed to address real data processing challenges involving large-scale datasets (Moguerza and Muñoz, 2006).

The equation for the linear hyper plane can be written as:

$$w^T(x) + b = 0 \quad \text{..... (1)}$$

Where, w=normal vector of the hyperplane

b = intercept and bias term of the hyperplane equation

Lasso Algorithm: Lasso (Least Absolute Shrinkage and Selection Operator) is considered as an assessment of regression technique that simultaneously conducts parameter selection and regularizes in order to boost the forecasting precision and interpretability of the resultant statistical model (Emmert-Streib

and Dehmer, 2019). Lasso regression effectively manages outlier-affected datasets, produces very sparse solutions, and addresses large-scale problems due to its probabilistic framework. We offer an effective method with assured global convergence for iteratively reweighted least-squares (IRLS) regression procedures. This establishes a new framework for sparse regression models within the extensive category of IRLS models, encompassing robust regression and logistic regression. Performance assessments on several standard benchmark datasets demonstrate this model's advantages over similar methods (Roth, 2004). Although Lasso is a prominent technique for sparse, high-dimensional regression problems, determining the lasso solution remains a computationally challenging endeavor when the quantity of variables significantly exceeds the overall amount of data points. (Genovese *et al*., 2012).

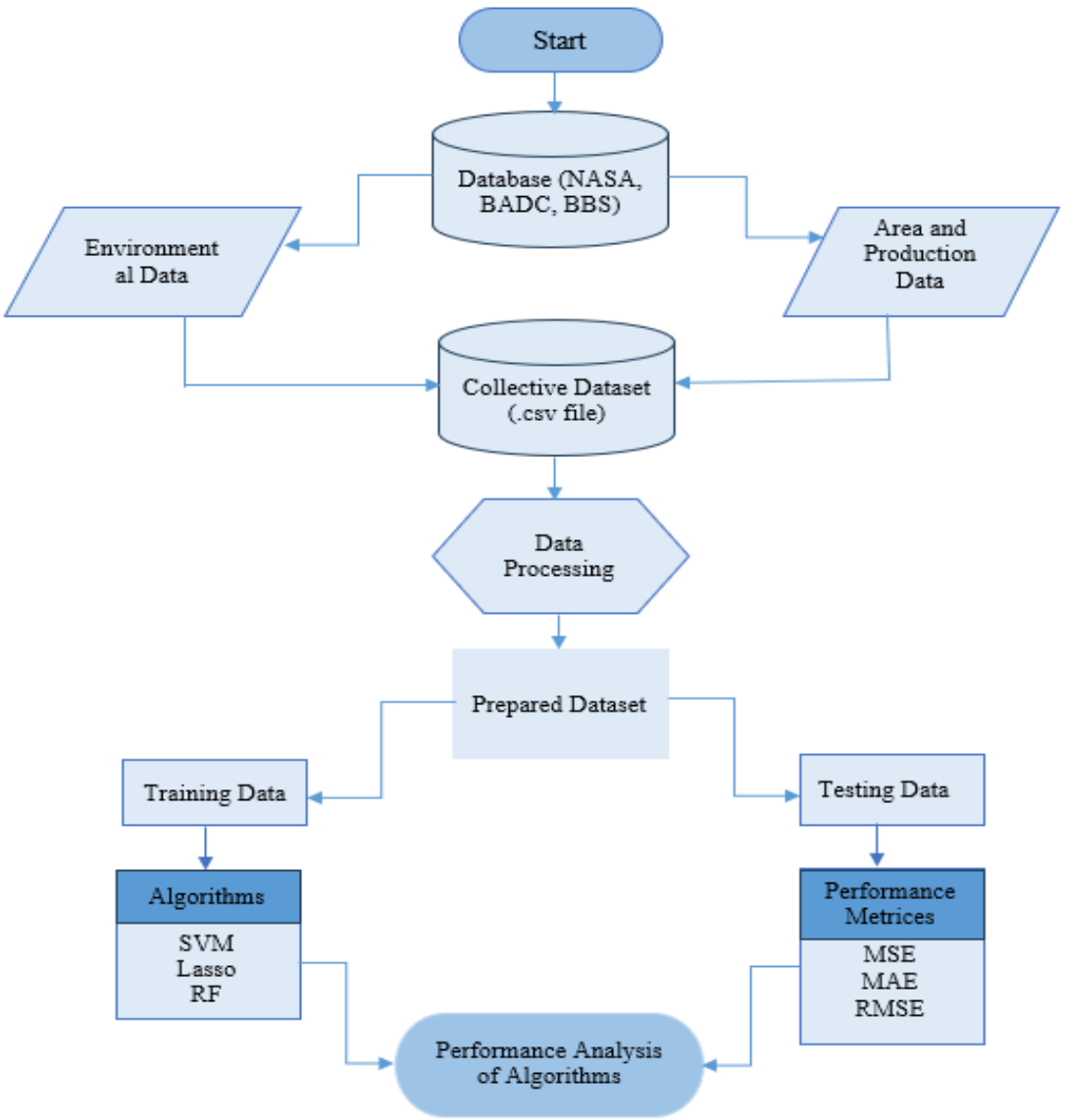
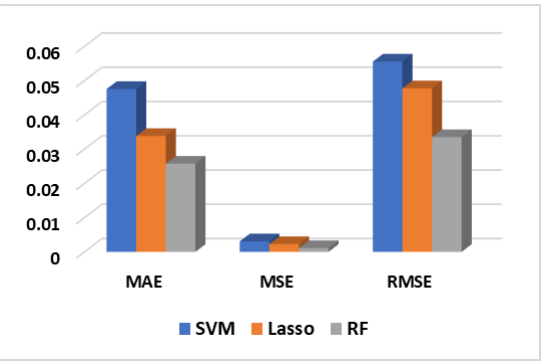


Fig. 1. Overview of the crop yield prediction using machine learning techniques

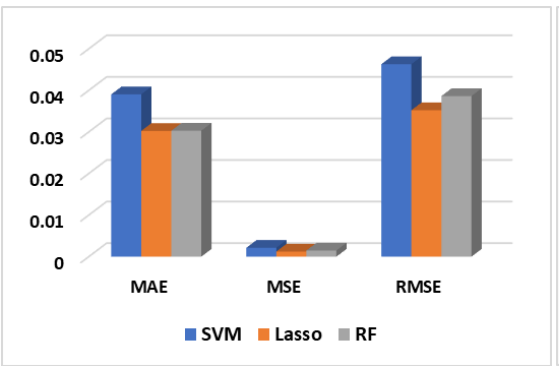
Random Forest: The random forest (RF) approach blends several randomized decision trees and averages their predictions in order to ensure superior performance and has proved incredibly effective as a both versatile classification as well as regression method (Biau and Scornet, 2016). This technique constructs numerous decision trees by introducing randomization in two key ways: firstly, it employs bootstrap sampling and secondly, randomization occurs at decision nodes. This process is iterated multiple times, typically 100 to 1000, resulting in the creation of a random forest (Rigatti, 2017). An increased quantity of trees in a random forest method can lead to higher accuracy and eliminate overfitting problems (Schonlau and Zou, 2020).

Results and Discussion

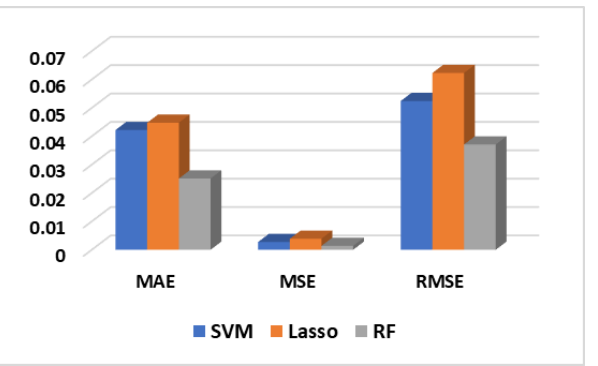
The production rate of Aus, Aman and Boro is not always constant. It changes based on various criteria such as land area, environmental data variations, etc. The study represented the prediction of production of three variations of rice crops utilizing three distinct ensemble-based machine learning algorithms. Table 1 displays the error-based evaluation metric values for three of our investigated machine learning algorithms in estimating the harvest. The comparative analysis of the algorithms' effectiveness is assessed depending on the error values measured during the crop yield prediction. Fig. 2 represents error values (evaluation metrics) for the machine learning methods. Here Blue represents SVM, Apricot-Orange represents Lasso and Grey represents RF algorithm. The gray bar in the figure indicates the least amount of error (MAE, MSE, RMSE) values obtained by applying the RF algorithm. Among all three types of rice crop yield prediction, the superiority of RF is shown compared to the other investigated algorithms in two types of crops (Aman and Boro). In the case of Aus, the Lasso algorithm showed a better result than the other two. The overall second-lowest evaluation error values are shown by the Lasso algorithm.



Boro prediction



Aus prediction



Aman prediction

Fig. 3. Comparative analysis of error measures for machine learning algorithms

Table 1. Assessment values for our investigated algorithms in yield forecasting of three primary rice varieties

Crops	Algorithm	MAE	MSE	RMSE
Aus	SVM	0.03905	0.00214	0.046308
	Lasso	0.03026	0.00124	0.03524
	RF	0.0303	0.00149	0.038636
Aman	SVM	0.04221	0.00275	0.05242
	Lasso	0.04478	0.00388	0.0623
	RF	0.02511	0.00138	0.0371
Boro	SVM	0.04755	0.00309	0.05555
	Lasso	0.03385	0.00228	0.04775
	RF	0.02576	0.00112	0.03352

Discussion

In this research, we investigated a predictive system using SVM, RF and Lasso algorithms for forecasting of crop yield employing our collected dataset. This entails evaluating the degree of correspondence between anticipated values and actual observed crop yields, as well as comparing the efficacy of each algorithm with suitable assessment criteria, which include MAE, MSE and RMSE, to figure out the optimal model for forecasting crop yields. The researchers also performed several agricultural yield prediction tasks employing various machine learning and deep learning techniques. The authors investigated and concluded that random forest is superior to polynomial regression and decision tree to forecast crop production of a given area (Gowda and Reddy, 2020). In this (Singha and Swain, 2021) paper, production of rice and potato crops were speculated using SVM, ANN and deep neural network (DNN) algorithms based on indicators like climate, soil and agricultural production from 2017 to 2018. It was found that DNN showed the best crop yield prediction results with 98% accuracy for both rice (RMSE = 0.20 tons/ha and $R^2 = 0.98$) and potato (RMSE = 0.95 tons/ha and $R^2 = 0.97$) crops. Data mining technique, linear regression (LR) showed the highest predictive result compared to k-nearest neighbor (KNN) for annual yield of three crops: cotton, sugarcane, and turmeric (Devika and Ananthi, 2008). An experiment was conducted to project the yields of Irish potatoes and maize utilizing acquired data predictors, such as temperature and precipitation and resulted in the RF model performing the best, with RMSE values for maize and potatoes of 129.9 and 510.8, respectively, and R^2 values of 0.817 and 0.875 for the same crops than the other two, polynomial regression and SVM (Kuradusenge *et al*., 2023). The final results of a separate investigation indicated that the decision tree classifier and random forest regression methods exhibited superior accuracy for crop recommendation and yield prediction, respectively, among other classifications and regression algorithms (Sundari *et al*., 2022). Similarly, we employed the SVM, Lasso and RF algorithms using collected data from different organizations in Bangladesh on three types of rice and it resulted in the RF algorithm with the best results of RMSE of 0.03352 in the case of boro rice, which in fact is comparable to the existing literature. Therefore, this research provides a substantial contribution to the development of predictive systems that are beneficial for farmers and our national government planning.

Conclusion and Future Work

This research provides an opportunity to explore and understand the interactions between different algorithms and agricultural data, contributing to the broader scientific knowledge in the field of precision agriculture. Accurate crop yield predictions can serve as valuable decision support tools for farmers, agricultural consultants, and policymakers, helping them make more informed choices in managing agricultural activities. In the future, the research can be extended to encompass multiple agricultural seasons to capture the variability in crop yield predictions across different years and weather conditions. This can help to identify long-term trends and improve the robustness of predictions. It also extends research to predict not only crop yield but also crop quality parameters such as nutritional value, taste and shelf life, which are too incredibly important to both producers and consumers.

References

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197-227.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *13*(11), 114003.

Devika, B., & Ananthi, B. (2008). Analysis of crop yield prediction using data mining technique to predict annual yield of major crops. *International Research Journal of Engineering and Technology*, 1460. https://www.irjet.net/archives/V5/i12/IRJET-V5I12275.pdf.

Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., Shdefat, A., & Saker, L. (2023). Crop Prediction Model Using Machine Learning Algorithms. *Applied Sciences, 13(16),* 9288. https://doi.org/10.3390/app13169288.

Emmert-Streib, F., & Dehmer, M. (2019). High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction, 1(1)*, 359-383.

Genovese, C. R., Jin, J., Wasserman, L., & Yao, Z. (2012). A comparison of the lasso and marginal regression. *The Journal of Machine Learning Research*, *13*(1), 2107-2143.

Gowda, S., & Reddy, S. (2020). Design and implementation of crop Yield Prediction model in agriculture. *International Journal of Scientific & Technology Research 8(01)*, 544.

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A. & Uwamahoro, A. (2023). Crop yield prediction using machine learning models: Case of Irish potato and maize. *Agriculture*, *13*(1), 225.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors (Basel, Switzerland)*, 18(8), 2674.

Liang, C., & Shah, T. (2023). IoT in agriculture: The future of precision monitoring and data-driven farming. *Eigenpub Review of Science and Technology*, 7(1), 85-104.

Meyer, D., & Wien, F. T. (2012). Support vector machines. *The Interface to libsvm in package e1071. e1071 Vignette*.

Moguerza, J. M., & Muñoz, A. (2006). Support vector machines with applications.

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565-1567.

Rahman, M. M., Islam, M. J., & Anwar, M. P. (2020). Evaluating the potentials of boro and aus rice varieties as a substitute to the short duration rice varieties in aman season. *Fundamental and Applied Agriculture, 5(1)*, 108-115.

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31-39.

Roth, V. (2004). The generalized LASSO. *IEEE transactions on neural networks*, *15*(1), 16-28.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3-29.

Singha, C., & Swain, K. C. (2021). Rice and potato yield prediction using artificial intelligence techniques. *Studies in big data*, 185–199. https://doi.org/10.1007/978-981-16-6210-2_9.

Sundari, V., Anusree, M., & Swetha, U. (2022). Crop recommendation and yield prediction using machine learning algorithms. *World Journal of Advanced Research and Reviews*, *14*(3), 452-459.

Tamasiga, P., Onyeaka, H., Bakwena, M., Happonen, A., & Molala, M. (2023). Forecasting disruptions in global food value chains to tackle food insecurity: The role of AI and big data analytics–A bibliometric and scientometric analysis. *Journal of Agriculture and Food Research*, 14, 100819.

Tom, K. (2020). Crop Prediction Using Machine Learning. *International Journal of Future Generation Communication and Networking, 13(3)*, 1896–1901.

Yesugade, K. D., Chudasama, H., Kharde, A., Mirashi, K., & Muley, K. (2019). Crop Suggesting System Using Unsupervised Machine Learning Algorithm. *International Journal of Computer Sciences and Engineering, 7(3)*, 322–325. https://doi.org/10.26438/ijcse/v7i3.322325.