EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

An Efficient Crop Yield Prediction System Using Machine Learning

Debabrata Swain[1, *], Sachin Lakum [1], Samrat Patel[1], Pramoda Patro [2] and Jatin[1]

[1]Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India
[2]Department of Mathematics, K L University, Hyderabad, Telangana, India

Abstract

Farming is considered the biggest factor in strengthening the economy of any country. It also has significant effects on GDP growth. However, due to a lack of information and consultation, farmers suffer from significant crop losses every year. Typically, farmers consult agricultural officers for detecting crop diseases. However, the accuracy of predictions made by agricultural officers based on their experience is not always reliable. If the exact issues are not identified at right time then it results in a heavy crop loss. To address this issue, Computational Intelligence, also known as Machine Learning, can be applied based on historical data. In this study, an intelligent crop yield prediction algorithm is developed using various types of regression-based algorithms. The Crop Yield Prediction Dataset from the Kaggle repository is used for model training and evaluation. Among all different regression methods Random Forest has shown the better performance in terms of R2 score and other errors.

Keywords: Farming, Regression, Crop Prediction, Mean Absolute Error (MAE), Root mean square Error (RMSE), R2 Score

Received on 01 December 2024, accepted on 01 March 2024, published on 07 March 2024

doi: 10.4108/eetiot.5333

*Corresponding author. Email: debabrata.swain7@yahoo.com

1. Introduction

Farming is always considered as the strongest pillar to support the economy of any country. According to the report of World Bank, Agriculture contributes 4% growth of global GDP and 25% for the developing countries GDP. According to World Economic Forum, in India around 58% of population are dependent on farming as their primary source of income. In farming/agriculture, farmers mostly face a lot of challenge due to various external factors such as scarcity in environmental conditions, improper domain knowledge, soil quality, etc. Because of all these adversarial conditions, every year farmer suffers with the problem of every crop loss. In India, from 2017 to 2019 around 8.5% of crop loss happened due to different external factors [2]. Recently the entire world has experienced the great loss in economy of Sri Lanka due to improper practices adoption in farming. This kind of economic breakdown mostly results in the suicide of the farmers. Due to heavy loss in farming, every year around 4.3% farmers lost their lives [1].

This kind of human disaster can be predicted if the amount of crop yield prediction can be done in advance. This prediction can help a farmer to adopt any precautionary measure to save the crop loss to a larger extent. Nowadays Machine learning helps in various domains of the society to do future prediction using a set of collected present data. Machine Learning is a branch of Artificial Intelligence that enables a computer to build an analytical model based on historical data. Regression is a machine learning based method that calculates a continuous uncertain outcome using present data. In this work a noble crop yield production has been done based on different agricultural features/data. Different regression-based models such as lasso regression, ridge, random forest and KNN are applied on the agricultural dataset available in Kaggle repository.

2. Literature Review

Nigam et al. [3] have developed a crop yield production system using Machine Learning. The dataset used by them is restricted to the southern parts of India. The features considered temperature, rainfall, etc. The different models used are Random Forest (RF), K-Nearest Classifier, XGBoost Classifier, etc. Out of all the models The Random Forest Classifier has shown better performance.

Champaneri et al. [4] have used Supervised Learning algorithms to predict crop farming yield. The data used by them only focuses on the districts of Maharashtra which was collected by the Indian government.

By using Random Forest Classifier, they developed a prediction system which is hosted on the Internet. Abbas et al. [5] have made use of proximal sensing techniques which helps to identify different types of crops produced. This paper has shown that their area of interest in predicting potato yield through four Machine Learning algorithms. The data they used was collected from around 40 locations investigating every field four times during the growing season. As mentioned before the four algorithms used are Linear Regression, Elastic Net, Support Vector Regressor (SVR) and k-Nearest Neighbour (KNN). The highest mean accuracy obtained by them was through using Support Vector Regressor (SVR).

Sellam V. and Poovammal E. [6] have used the data collected by the Ministry of Agriculture to perform the analysis. The data is spread over the span of over 10 years from 2000 all the way to 2011. They have made use of Linear Regression and applied it three times for finding a relation between the yield and other parameters for rice crop. They made use of MATLAB environment to implement regression which gave them an R2 value of 70%.

Nishant et al. [7] have used Regression machine learning techniques on the data which they collected from the Indian Government Repository. Three regression models used and achieved a RMSE of less than 1%. The final model was then deployed as a web application.

Ghosh et al.'s 2023 study focuses on [8] "Water Quality Assessment Through Predictive Machine Learning", highlighting the use of machine learning for analyzing and predicting water quality parameters. In "Unraveling the Heterogeneity of Lower-Grade Gliomas," Rahat, Ghosh,[9] and colleagues (2023) delve into deep learning-assisted segmentation and genomic analysis of brain MR images, offering new insights into this medical condition. Potato Leaf Disease Recognition and Prediction using Convolutional Neural Networks," by Ghosh, Rahat [10] and team (2023), showcases the application of convolutional neural networks in accurately identifying diseases in potato leaves. Mandava, Vinta, Ghosh, and Rahat's [11] research presents "An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular Disease in Bangladeshi Population", integrating advanced AI techniques for health predictions. The 2023 study by Mandava et al., titled "Identification and Categorization of Yellow Rust Infection [12] in Wheat through Deep Learning Techniques", applies deep learning methods to detect and categorize wheat infections effectively. Hasim, Rahat, Ghosh,[13] and colleagues' 2023 article, "Using Deep Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh", explores AI-based solutions for diagnosing rice-leaf diseases. Deciphering Microorganisms through Intelligent Image Recognition", authored by Khasim, Ghosh, Rahat,[14] and others in 2023, discusses the use of machine learning and deep learning in identifying microorganisms through advanced image recognition techniques. The 2023 study by Mohanty, Ghosh [15] Rahat, and Reddy, "Advanced Deep Learning Models for Corn Leaf Disease Classification", focuses on the application of deep learning in classifying diseases in corn leaves based on a field study. Alenezi[16] and team's 2021 research, "Block-Greedy and CNN Based Underwater Image Dehazing for Novel Depth Estimation and Optimal Ambient Light", investigates novel CNN-based methods for enhancing underwater image clarity and depth estimation.

3. Proposed Work

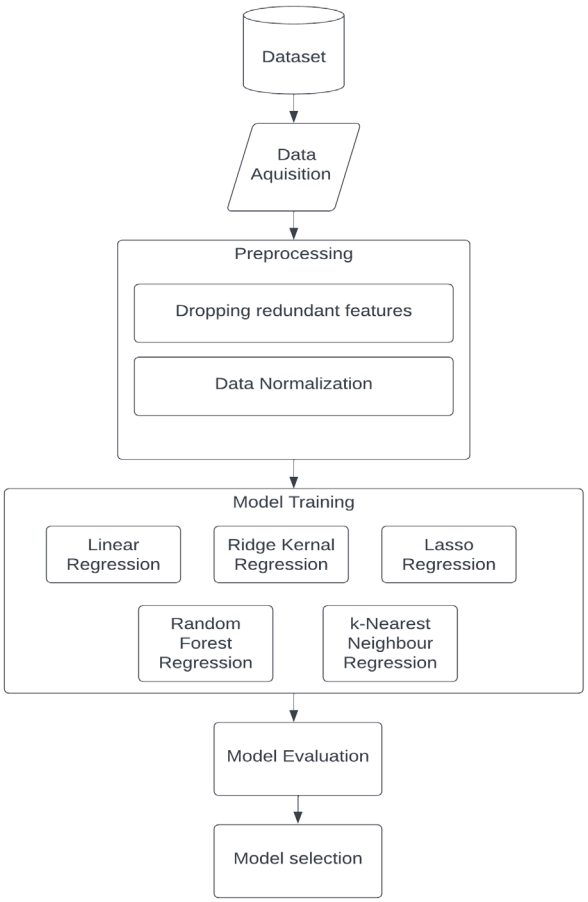

Figure 1. Proposed System

3.1 Data Collection

The dataset is containing agricultural data about various countries around the world like India, Mexico, Brazil, United Kingdom, etc available in the Kaggle repository [8]. The various features present in the dataset are such as crop type, average temperature, average rainfall, amount of pesticides used in tonnes and a target feature which is the amount of yield produced.

3.2 Data Preprocessing

3.2.1 Chi-squared test

Chi-squared test is performed to determine if there is sufficient associativity between two variables [9]. It is often used in categorical data to determine how much the dependent variable is actually dependent on the given feature as shown in equation 1. Thus, chi-Square test was performed which concluded in the feature 'Year' being redundant and was thus dropped.

$$\chi 2 = \frac{\sum(O_i - E_i)^2}{E_i} \quad (1)$$

In equation 1, O_i is the Observed value and E_i is the Expected value.

3.2.2 Normalization

Afterwards, due to the data having a widespread in values, data normalization was performed [10]. For this, Min-Max Scaler was deployed. As a result, the data scaled in the range of 0 and 1. The scaled values calculated using equation 2 improve the performance of algorithms.

$$x_{scaled} = \frac{x - x_{min}}{x_{\max} - x_{min}} \quad (2)$$

3.2.3 Train-Test Split

Finally, the data was divided into two parts: training and testing. The training part consisted of 70% of the total data and testing was made up of the rest. The split was done using the same random state which ensures reproducibility leading to consistent results. The split was also done with shuffle enabled which ensures that the permutation for training and testing sample selection will be random which prevents any bias which could occur due to the order of the samples.

3.3 Machine Learning Algorithms

3.3.1 Linear Regression:

Linear Regression is a regression method which predicts the value of dependent variable based on one or more independent variable [11]. As seen in equation 3 it follows a linear relationship between dependent variable and independent variable. It predicts a continuous value. The formula for linear regression is as:

$$Y_i = \beta_o + \beta_1 . x_i \quad (3)$$

, where β0 is the intercept and β1 is the regression coefficient. This model's predicted values can be evaluated

through certain metrics such as MAE, MSE, RMSE and R-squared.

3.3.2 Kernel Ridge Regression (KRR):

It is a non-parametric method of regression which obtains the relationship between the expected values and the outcomes of the predictor variables [12]. It is used for both non-linear and linear relationships. Equation 4 shows the formula for coefficient vector associated with the predictors.

$$\beta = (X^T X + \lambda I)(X^T y) \quad (4)$$

It is also estimated based on RMSE, MSE, MAE and Goodness Fit.

3.3.3 Lasso Regression:

Lasso regression is a regularization method of regression model to avoid overfitting and predict accurate results [13]. It uses a shrinkage function to shrink the data values towards a central point. A cost function and loss function to predict future values. The formula for lasso cost function is given below in equation 5.

$$Cost = (Y - X\beta)^T (Y - X\beta) \quad \textbf{(5)}$$

3.3.4 Random Forest Regression:

Random Forest Regressor is an Ensemble learning model which draws random decision trees and averages the multiple outputs/predictions to obtain an accurate result [14]. Decision trees with some bootstrapping creates the random forest regressor to predict the values. It uses several parameters to predict the expected value like no. of decision trees, maximum depth of the decision trees, no. of iterations, etc.

3.3.5 K-Nearest Neighbor Regression:

K-nearest Neighbors regression (KNN regressor) is a regression model which groups same or nearby values and predicts a value [15]. KNN regressor is a non-parametric algorithm which does not support underlying data. KNN regressor aims to create new groups of similar features and predict the output values [16]. It calculates the distance between two data points and measure the distance to group them [17]. It tries to minimize the group values and predict required results. The formula for calculating Euclidean distance is shown in equation 6.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

4. Result and Discussion

The features represented in our dataset predict the value of the crop yield produced. Amongst the above-mentioned algorithms, Random Forest Regressor has shown the better result in terms of R^2 score, RMSE and MAE. The model has shown R^2 score value as 0.98, RMSE as 0.45, and MAE value as 0.1088. The formula for R-squared and Mean Absolute Error are given in equation 7 and 8 respectively.

$$R^2 = 1 - \frac{Sum\ Squared\ Regression\ (SSR)}{Total\ sum\ of\ squares\ (SST)} \quad \textbf{(7)}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | x_i - x | \quad (8)$$

5. Conclusion

There is a noticeable impact on crop yield production (measured in tonnes) due to factors such as pesticide usage, average rainfall (in millimetres), average temperature, and the type of crop grown under specific conditions. Our analysis reveals that Potato (Solanum tuberosum) is the predominant crop cultivated across vast areas over the years.

By conducting a rigorous analysis, we obtained an R-squared value of 0.98, indicating a strong relationship between the aforementioned factors and crop yield. This high R-squared value suggests that approximately 98% of the variability in crop yield can be explained by rainfall, temperature, and pesticide use combined.

References

[1] https://www.downtoearth.org.in/news/climate-change/india-lost-crops-on-18-million-hectares-to-extreme-floods-from-2017-2019-govt-75506

[2] https://ncrb.gov.in/sites/default/files/chapter-2A%20farmer%20suicides.pdf

[3] Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019). Crop Yield Prediction Using Machine Learning Algorithms. 2019 Fifth International Conference on Image Information Processing (ICIIP).

[4] Champaneri, M., Chachpara, D., Chandvidkar, C., & Rathod, M. (2016). Crop yield prediction using machine learning. Technology, 9, 38.

[5] Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. Agronomy, 10(7), 1046.

[6] Sellam, V., & Poovammal, E. (2016). Prediction of crop yield using regression analysis. Indian Journal of Science and Technology, 9(38), 1-5.

[7] P. S. Nishant, P. Sai Venkat, B. L. Avinash and B. Jabber, " Crop Yield Prediction based on Indian Agriculture using Machine Learning," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, (pp. 1-4).

[8] Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S., Mohanty, S.N. (2023). Water Quality Assessment Through Predictive Machine Learning. In: Intelligent Computing and Networking. IC-ICN 2023. Lecture Notes in Networks and Systems, vol 699. Springer, Singapore. https://doi.org/10.1007/978-981-99-3177-4_6

[9] Rahat IS, Ghosh H, Shaik K, Khasim S, Rajaram G. Unraveling the Heterogeneity of Lower-Grade Gliomas: Deep Learning-Assisted Flair Segmentation and Genomic

Analysis of Brain MR Images. EAI Endorsed Trans Perv Health Tech [Internet]. 2023 Sep. 29 [cited 2023 Oct. 2];9.https://doi.org/10.4108/eetpht.9.4016

[10] Ghosh H, Rahat IS, Shaik K, Khasim S, Yesubabu M. Potato Leaf Disease Recognition and Prediction using Convolutional Neural Networks. EAI Endorsed Scal Inf Syst [Internet]. 2023 Sep. 21https://doi.org/10.4108/eetsis.3937

[11] Mandava, S. R. Vinta, H. Ghosh, and I. S. Rahat, "An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular Disease in Bangladeshi Population", EAI Endorsed Trans Perv Health Tech, vol. 9, Oct. 2023.https://doi.org/10.4108/eetpht.9.4052

[12] Mandava, M.; Vinta, S. R.; Ghosh, H.; Rahat, I. S. Identification and Categorization of Yellow Rust Infection in Wheat through Deep Learning Techniques. EAI Endorsed Trans IoT 2023, 10. https://doi.org/10.4108/eetiot.4603

[13] Khasim, I. S. Rahat, H. Ghosh, K. Shaik, and S. K. Panda, "Using Deep Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh", EAI Endorsed Trans IoT, vol. 10, Dec. 2023 https://doi.org/10.4108/eetiot.4579

[14] Khasim, H. Ghosh, I. S. Rahat, K. Shaik, and M. Yesubabu, "Deciphering Microorganisms through Intelligent Image Recognition: Machine Learning and Deep Learning Approaches, Challenges, and Advancements", EAI Endorsed Trans IoT, vol. 10, Nov. 2023. https://doi.org/10.4108/eetiot.4484

[15] Mohanty, S.N.; Ghosh, H.; Rahat, I.S.; Reddy, C.V.R. Advanced Deep Learning Models for Corn Leaf Disease Classification: A Field Study in Bangladesh. Eng. Proc. 2023, 59, 69.https://doi.org/10.3390/engproc2023059069

[16] Alenezi, F.; Armghan, A.; Mohanty, S.N.; Jhaveri, R.H.; Tiwari, P. Block-Greedy and CNN Based Underwater Image Dehazing for Novel Depth Estimation and Optimal Ambient Light. Water 2021, 13, 3470. https://doi.org/10.3390/w13233470

[17] https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset

[18] Swain D, Mehta U, Bhatt A, et al. A robust chronic kidney disease classifier using machine learning. Electronics 2023; 12(1): 212. doi: 10.3390/electronics12010212

[19] H. W. Herwanto, A. N. Handayani, A. P. Wibawa, K. L. Chandrika and K. Arai, "Comparison of Min-Max, Z-Score and Decimal Scaling Normalization for Zoning Feature Extraction on Javanese Character Recognition," 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Malang, Indonesia, 2021, pp. 1-3, doi: 10.1109/ICEEIE52663.2021.9616665.

[20] A. Lakshmanarao, M. N. Kumar, K. S. V. Ratnakar and Y. Satwika, "Crop Yield Prediction using Regression Models in Machine Learning," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 423-426, doi: 10.1109/ICAAIC56838.2023. 10141462.

[21] J. He, L. Ding, L. Jiang and L. Ma, "Kernel ridge regression classification," 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 2014, pp. 2263-2267, doi: 10.1109/IJCNN.2014.6889396.

[22] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 2016, pp. 18-20, doi: 10.1109/ICACA.2016.7887916.

[23] P. Dong, H. Peng, X. Cheng, Y. Xing, X. Zhou and D. Huang, "A Random Forest Regression Model for Predicting Residual Stresses and Cutting Forces Introduced by Turning IN718 Alloy," 2019 IEEE International Conference on Computation, Communication and Engineering (ICCCE), Fujian, China, 2019, pp. 5-8, doi: 10.1109/ICCCE48422.2019.9010767.

[24] M. Atanasovski, M. Kostov, B. Arapinoski and M. Spirovski, "K-Nearest Neighbor Regression for Forecasting Electricity Demand," 2020 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), Niš, Serbia, 2020, pp. 110-113, doi: 10.1109/ICEST49890.2020.9232768.

[25] Kumar, S., Neware, N., Jain, A., Swain, D., Singh, P. (2020). Automatic Helmet Detection in Real-Time and Surveillance Video., Machine Learning and Information Processing. Advances in Intelligent Systems and Computing, vol 1101. Springer, Singapore. https://doi.org/10.1007/978-981-15-1884-3_5

[26] Swain, Drdebabrata & Satapathy, Santosh & Acharya, Biswaranjan & Shukla, Madhu & Gerogiannis, Vassilis & Kanavos, Andreas & Giakovis, Dimitris. (2022). Deep Learning Models for Yoga Pose Monitoring. Algorithms. 15. 403. 10.3390/a15110403.

[27] E. Brilliandy, H. Lucky, A. Hartanto, D. Suhartono and M. Nurzaki, "Using Regression to Predict Number of Tourism in Indonesia based of Global COVID-19 Cases," 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS), IPOH, Malaysia, 2022, pp. 310-315, doi: 10.1109/AiDAS56890.2022.9918731.

[28] S. A. Septianingrum, M. Alfian Dzikri, M. A. Soeleman, P. Pujiono and M. Muslih, "Performance Analysis of Multiple Linear Regression and Random Forest for an Estimate of the Price of a House," 2022 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2022, pp. 415-418, doi: 10.1109/iSemantic55962.2022.9920454.

[29] J. Qi, J. Du, S. M. Siniscalchi, X. Ma and C. -H. Lee, "Analyzing Upper Bounds on Mean Absolute Errors for Deep Neural Network-Based Vector-to-Vector Regression," in IEEE Transactions on Signal Processing, vol. 68, pp. 3411-3422, 2020, doi: 10.1109/TSP.2020.2993164.