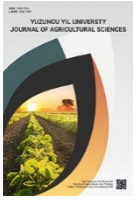
Yuzuncu Yil University
Journal of Agricultural Sciences
(Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi)
https://dergipark.org.tr/en/pub/yyutbd
ISSN: 1308-7576
e-ISSN: 1308-7584



Research Article

Yield Prediction and Recommendation of Crops in India's Northeastern Region Using Machine Learning Regression Models

Nisha SHARMA[1], Mala DUTTA*[2]

[1,2]Department of Computer Technology, Assam down town University, Guwahati- 781026, Assam, India

[1]https://orcid.org/0000-0002-4315-8225, [2]https://orcid.org/0000-0001-9560-0751

*Corresponding author e-mail: maladuttasid@gmail.com

Article Info

Received: 03.07.2023
Accepted: 02.10.2023
Online published: 15.12.2023
DOI: 10.29133/yyutbd.1321518

Keywords

Crop yield prediction,
Ensemble Learning,
Machine Learning,
Recommender system,
Regression

Abstract: Agriculture has a big impact on society because it is essential for a large percentage of our food. The issue of hunger is getting worse because of a growing population in many nations, resulting in food shortages or insufficiencies. To meet the world's food needs, it is ever more crucial to provide crop protection, conduct detailed land surveys, and predict crop yields. To calculate the estimated number of crops that are produced in a year, this research focuses on the use of machine learning techniques to predict crop yield and recommend crops with the highest yield and profitability in the Northeast region of India. The crop market's fluctuations in prices may be controlled with the aid of this information. To estimate agricultural crop yields, this study accurately evaluates a range of machine learning regression models, such as Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost (eXtreme Gradient Boosting), and AdaBoost. With a 0.98 R^2 score for the XGBoost and 0.96 for the Random Forest, they performed better than the other models. By evaluating crop yields along with their corresponding market prices and costs, we have determined their profitability. As a result, we have provided recommendations for the top five profitable crops in India's Northeastern region.

To Cite: Sharma, N, Dutta, M, 2023. Yield Prediction and Recommendation of Crops in India's Northeastern Region Using Machine Learning Regression Models. *Yuzuncu Yil University Journal of Agricultural Sciences*, 33(4): 700-708.
DOI: https://doi.org/10.29133/yyutbd.1321518

1. Introduction

In the agricultural industry, crop production prediction is a crucial issue. Every farmer strives to understand crop output and whether it lives up to their expectations, which involves assessing the farmer's prior experience with the crop to anticipate the yield. To manage agricultural risk, accurate crop history data is essential. In order to effectively plan and make decisions regarding resources, which are critical to agriculture's role in supplying the world's food needs, crop yield predictions must be correct. In this context, machine learning becomes a useful technique to improve agricultural yield prediction models.

Machine learning algorithms can produce precise projections of crop yields for certain regions or farms by recognizing complex patterns and relationships within this information. The results of applying machine learning algorithms to a data collection of sugarcane crop information from Karnataka, India, are demonstrated by (Renuka and Sujata, 2019). Regression approaches are a useful technique to produce a forecast for the area. The regression method may be utilized for crop forecasts for the area with satisfactory results, it is a useful tool for yield prediction. The R^2 statistics result is considered as good for crop production prediction (Shastry et al., 2017). Using a variety of machine learning techniques, a method for predicting the classification of production that depends on macro- and micronutrients is presented (Singh et al., 2017). The algorithm’s precision was improved with the use of stacking regression (Potnuru et al., 2020). After analyzing the soil dataset category is predicted (Paul et al., 2015). It is determined that the crop yield is a classification rule from the expected soil category. For predicting crop yield, naive Bayes and k-Nearest Neighbour algorithms are employed. The indicated future study involves developing effective models utilizing other classification methods like support vector machines.

To improve classification performance, diversity measurements depending on the correlation between errors are employed to compute the classifiers' correlation approach (Yiang, 2011). Using Random Forest, (Everingham et al., 2016) proposed a method for forecasting sugarcane production. The parameters used in this work include the biomass index, climate data (such as rainfall, radiation levels, and temperature), and production data from the two years prior. In a study by (Deepa et al., 2023), the idea is to provide farmers with internet access through a smartphone application that predicts yield. In the GPS, the user enters the location and soil type. Algorithms can anticipate agricultural yields for crops chosen by the user as well as select the list of crops that will provide the greatest profits. Crop productivity estimates are made using a variety of different machine learning methods. A 95% accuracy percentage was achieved by random forest model among them. Two distinct strategies were put out (Ung and Mittrapiyanuruk, 2018) for separating the information about the plot's characteristics from the category of sugarcane production at the plot level. The strategies are based on using ensemble models Gradient Boost and Random Forest. The yield prediction model’s accuracy was improved by modifying the bias, weight, and optimizer in a multilayer perceptron neural network. The suggested model predicts crop yield using an ANN with a three-layer neural network (Kale and Patil, 2019). The requirements and strategies for developing a precision agricultural software model are discussed in the paper (Babu, 2013). It thoroughly examines precision farming's fundamentals.

In (Kumar et al., 2015), the elements that influence crop choices, such as rate of production, price in the market, and governmental policies, are analyzed. This study suggests a method that fixes the selection issue of crops and raises the crop's net yield rate. It proposes that a season's worth of crops be chosen while taking the weather, crop type, soil type, and water density into account. The research (Savla et al., 2015) compares the classification algorithm’s yield prediction capabilities in precision agriculture. These algorithms are used to predict the production of a soybean crop using data that has been gathered over several years. In this study, various models were employed for the yield prediction techniques. Here, the Bagging technique of ensemble learning is considered the best algorithm for yield prediction. A machine learning (ML) model is designed to forecast agricultural output. The data was gathered and educated using supervised machine learning with six different regression models to estimate crop yields. Random Forest Regressor which is an ensemble model fared better than the other models, with an MAE of 468.16 and a Cross-Validation score of 0.6087 (Panigrahi, 2023). Machine learning techniques are used to generate recommendations for crops based on geological and climatic characteristics. The dataset for the five different crops, including rice, ragi, gram, potato, and onion, has been considered when designing the recommendation crop system (Garanayak et al., 2021).

We have proposed a recommendation system that will predict how much crop will be gathered from a given agricultural area and recommend the crops having the highest crop yield value and profitability. This approach makes predictions using a variety of data sources, including area, production, previous yield records, and other relevant elements. This study uses regression models to accurately predict agricultural production in the future for the Northeast region. Regression techniques such as XGBoost, Gradient Boost, Random Forest, AdaBoost, Decision Tree, and Linear Regression have been used for predicting crop yield**.**

2. Material and Methods

2.1. Framework for our proposed model

The proposed methodology for our proposed recommendation model has been illustrated below in Figure 1**.**

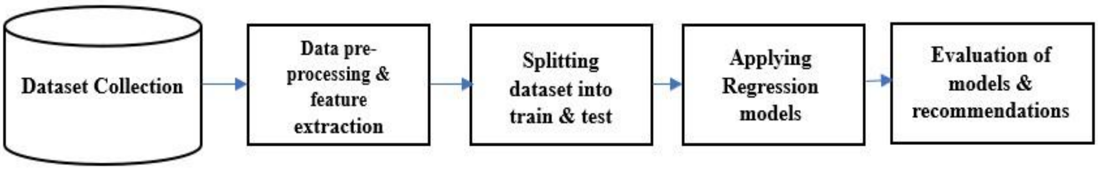

Dataset Collection
Data pre-processing & feature extraction
Splitting dataset into train & test
Applying Regression models
Evaluation of models & recommendations

Figure 1. Framework for methodology.

2.1.1. Data collection

The dataset utilized in this research originates from (Kumar, 2018), and has been specially selected to address the unique agricultural context of the Northeastern area of India. A thorough study of crop-related trends and dynamics in this area is made possible by the dataset's wide temporal range, which extends from 1997 to 2015. This dataset is a useful tool for understanding the dynamics of the region's agriculture, improving predictions of crop yields, and making suggestions for the development of a sustainable agricultural sector. Parameters in the dataset are as follows: State Name- this field indicates the name of the Indian state from where the Northeastern region agricultural data was gathered, District Name- this field indicates the district level data of each state, Crop Year- this field indicates the year in which the data was recorded, Season- this field indicates the season for understanding the pattern of crops in different seasons, Crop- this field indicates the crop name cultivated in the respective years, Area- The area parameter indicates the amount of land (in hectares (ha)) that is being used to grow a specific crop in a given year and district. This metric is important for yield calculations as it helps in understanding the scope of agricultural activity, and Production- this metric is used to describe the overall agricultural output (measured in metric tons (MT)) for the selected crop, district, and year. It is an essential variable for crop yield analysis because it is directly related to the agricultural productivity of the area.

2.1.2. Data pre-processing

In this stage, several measures were undertaken to ensure that the data was correctly prepared for analysis. This included dealing with data ranges, missing numbers, and identifying important features. The model was implemented using high-level programming language Python which is known for its readability and simplicity. We have used the specific version 3.8.5 of Python 3.x series which includes various useful libraries such as Scikit-Learn, NumPy, and Pandas. The tool that we have used to run the Python program is Jupyter Notebook. It is an interactive open-source web tool that enables us to create and share documents with real-time code, equations, visuals, and text. Jupyter Notebook is a flexible tool that integrates interactive scripting, documentation, and visualization into a single environment. It is frequently used by data scientists, researchers, educators, and experts in domains where data analysis and interactive computing are crucial for tasks including data exploration, machine learning, scientific research, and collaborative work. The Pandas library's isnull() method was used to address missing values in the dataset. This approach identified any null values that were present in the dataset. Once identified, the same library's fillna() function was used to replace the missing data. The fillna() function was used to replace missing values in columns containing numerical characteristics with the mean values of the associated columns.

2.1.3. Feature Selection using correlation matrix

The correlation matrix employing Pearson correlation is used here for feature selection. A statistical tool called Pearson correlation is used to express the linear relationship between two continuous variables. A correlation matrix (Figure 2) was created to examine the relationships between the parameters in the provided dataset. The correlation matrix demonstrates favorable relationships between yield and production as well as production and area. These findings suggest that greater agricultural areas and higher levels of output are often associated with higher yields. Understanding these interactions can help influence decision-making processes in agricultural practices linked to optimizing yield, production techniques, and land utilization. Therefore, the crop yield which we have calculated using the area and production that were provided in the dataset is considered an important parameter for our prediction analysis. After the process of feature selection, with a division ratio of 75% or 25%, the loaded data is divided into two sets of training and testing.

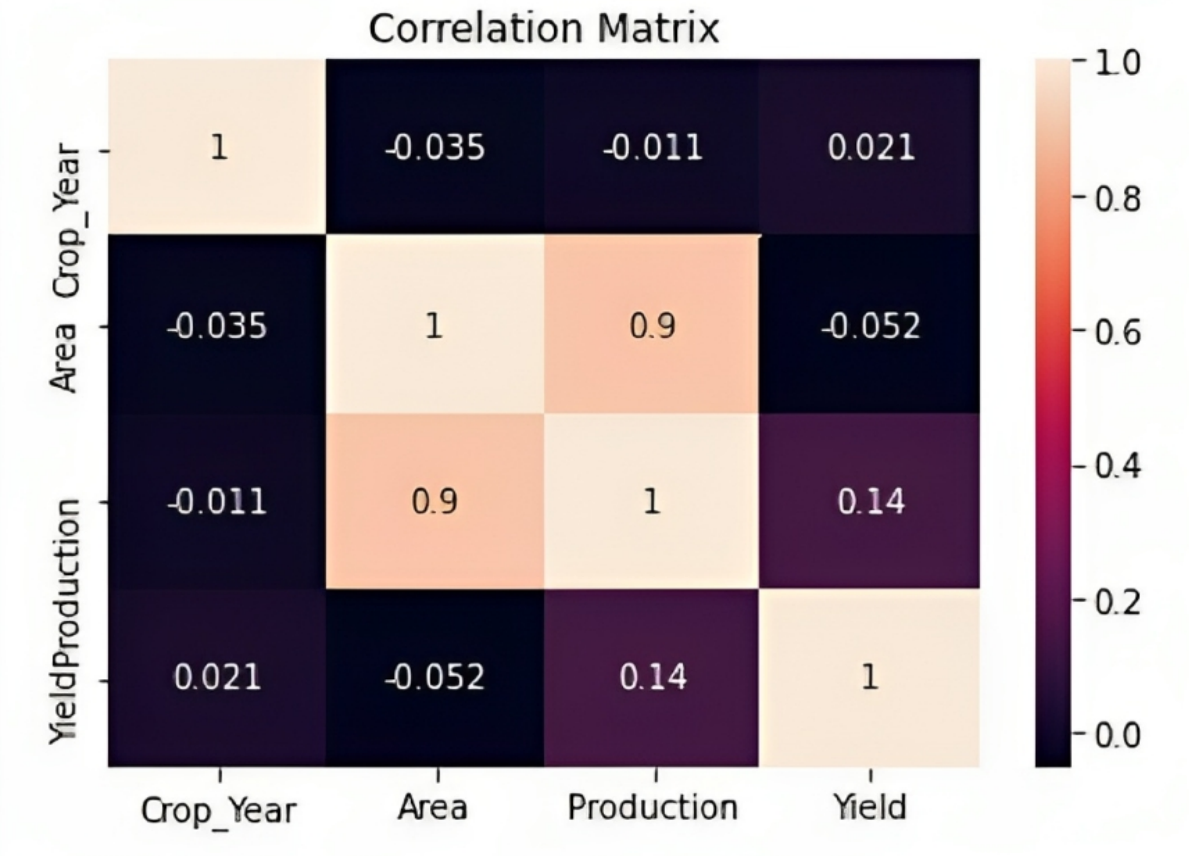


Figure 2. Correlation matrix.

2.1.4. Machine learning models used

For the purpose of predicting crop yield, we have used regression models and ensemble models. A regression model is used to examine the relationship between a target or dependent variable and one or more independent variables or characteristics. A regression model's main objective is to understand and predict the values of the dependent variable based on the values of the independent variables. And, a machine learning technique known as an ensemble model combines the predictions of various base models (commonly referred to as "weak learners") to produce a more reliable and precise prediction. The concept underlying ensemble learning is that by combining the forecasts of various models, the resulting ensemble model can frequently outperform any single base model. The regression models- Decision Tree, Linear Regression, and ensemble models- Random Forest, GradientBoost, XGBoost (eXtreme Gradient Boosting), and AdaBoost are used in this research. Brief descriptions of all the models used in our proposed approach are given below:

Linear Regression: Linear regression is used to find the relation between a dependent variable and one or more independent variables.

Random forest regressor: It is an ensemble model with a collection of decision trees. Each tree is built using a randomly selected subset of the data and features.

Gradient Boost: An ensemble model, Gradient Boosting incrementally assembles a group of ineffective prediction models, typically decision trees, to produce an effective predictive model. It operates by fitting new models iteratively to the residual errors of the previous models while attempting to reduce the mistakes caused by the prior models.

XGBoost: Extreme gradient boosting (XGBoost) Regressor is an advanced version of gradient boosting that builds extremely precise regression models. It includes various improvements to enhance efficiency and handle complex datasets.

Decision Tree: It is a popular supervised machine-learning technique used which can be used in regression and classification problems. Each internal node represents a feature, each branch represents a decision or rule based on that attribute, and results are represented by the leaf nodes.

AdaBoost: An ensemble model Adaptive Boosting or AdaBoost operates by initially assigning equal weights to each training instance's weights and then using the data to train a weak model. The weighted errors of the preceding models are the focus of the next models.

2.1.5. Evaluation of model

Several evaluation metrics, such as R^2 Score, Root Mean Squared Error (RMSE), cross-validation (CV), and Mean Absolute Error (MAE), are used to evaluate the model's performance.

Mean Absolute Error: It is calculated by adding up the absolute differences between each observation's actual and estimated values and dividing by the total number of observations.

$$\textbf{Formula for MAE,} \quad \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'| \tag{1}$$

Here (1),
n is the number of observations
y_i represents the observation's actual values
y_i' represents the estimated or predicted values

RMSE: It calculates the average squared difference between the estimated and actual values. It gives a measurement of the overall prediction error of an estimator.

$$\textbf{Formula for RMSE,} \quad \sqrt{\sum_{i=1}^{n}\frac{(y_i' - y_i)^2}{n}} \tag{2}$$

Here (2),
n is the number of observations
y_i represents the observation's actual values
y_i' represents the estimated or predicted values

R^2 Score: The coefficient of determination, also referred to as R^2 score, is a statistical metric used to evaluate the quality of fit of a regression model.

$$\textbf{Formula for } \mathbf{R^2}, \quad 1\text{-}\frac{SSR}{SST} \tag{3}$$

Here (3),
SSR represents sum squared regression which is the square of the residuals.
SST is the total sum of squares which is the sum of the data's distance from the mean squared.

Cross_validation score: Here, the available dataset is split up into subsets or folds and the score is determined by averaging the results from each iteration.

3. Results and Discussion

We have proposed a recommendation model that can predict agricultural yields for several crops in the Northeast region of India and recommend crops having the highest crop yields and profitability. The data was utilized to train six different regression models, including Linear Regression, Decision Tree Regression, GradientBoost Regression, Random Forest Regression, AdaBoost Regression, and XGBoost Regression, in order to provide accurate predictions of crop yields. R^2 Score, Root Mean Squared Error (RMSE), cross-validation (CV), and Mean Absolute Error (MAE) for the models are presented in Table 1.

Table 1. Accuracy of the regression models in terms of MAE, RMSE, R^2 Score, and CV Score

Models	RMSE	MAE	R^2 Score	CV Score
Random Forest	2475	607	0.96	0.73
AdaBoost	5422	2421	0.85	0.53
Decision Tree	3128	683	0.95	0.66
XGBoost	2090	652	0.97	0.75
GradientBoost	3426	1018	0.94	0.72
Linear Regression	5039	2380	0.87	0.65

Figure 3 shows the MAE and RMSE for trained models. The RMSE value, shown in blue on the graph, is lowest for XGBoost Regression, followed by Random Forest and Decision Tree. The orange graph displays the MAE value for Random Forest Regression, which is the lowest, followed by XGBoost Regression.

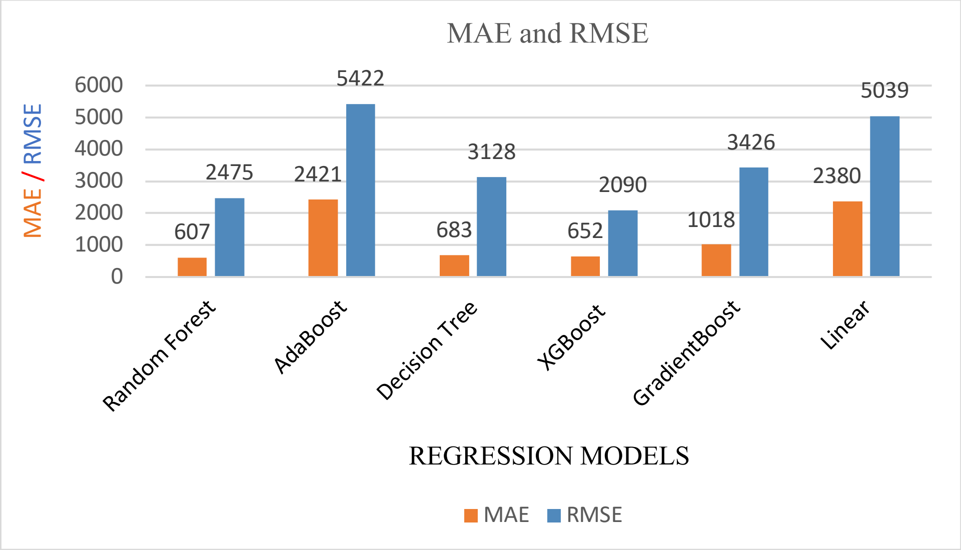


Figure 3. Graphical representation of MAE and RMSE.

The R^2 and Cross-Validation scores for the trained Regression models are graphically depicted in Figure 4. The optimum R^2 score is 1. The orange graph displays the R^2 score, and the blue graph displays the cross-validation score. XGBoost has the highest cross-validation score, followed by Random Forest, Gradient Boost, and Decision Tree. The XGBoost has the highest R^2 Score followed by Random Forest and Decision Tree.

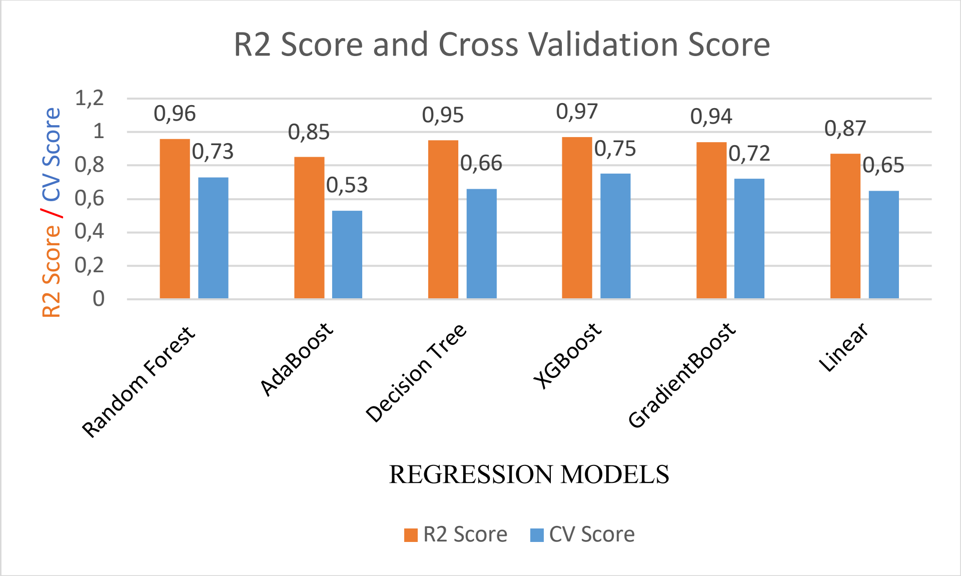


Figure 4. Graphical representation of R2 Score and CV Score.

Table 2. Accuracy of the XGBoost before and after parameter optimization

Models	**Before parameter optimization**	**After parameter optimization**
RMSE	2090	1972
MAE	652	548
R2 Score	0.97	0.98
CV Score	0.75	0.77

We further analyze and apply hyperparameter optimization to the XGBoost model as its accuracy was higher than any other model's. The accuracy was greatly enhanced once the model's parameters were optimized as shown in Table 2. In Table 2. We have represented the evaluation result for XGBoost before and after applying hyperparameter tuning for parameter optimization.

The model is further used to determine the crop having a high yield, the top five crops having high yield are presented in Table 3.

Table 3. The top 5 crops have high yields in terms of metric tons per hectare

Crop_Name	**Highest Average Yield (MT/ha)**
Sugarcane	24.374
Cabbage	11.833
Banana	10.626
Tapioca	10.472
Ginger	9.0000

Then we tried to find out the most profitable crop by considering the cost and market price of the crops having higher yield value (Table 3). The data for the market price of crops are collected from the website commodityonline.com and average cost price of the crops Ginger, Tapioca, Cabbage, Banana, and sugarcane are respectively collected from the following sources (Asia Farming, 2023**)**, (Times of India, 2019), (Agri farming, 2023), (Patowary et al., 2022) and (Government of India, 2023). Also, a complete link of all the sources is given in the reference section. Rank wise recommendation of crops based on their profitability is shown in Figure 5.

The findings (Table 2) show that while sugarcane has the highest yield, it also has the lowest profitability (Figure 5). This finding draws attention to an important aspect of agricultural decision-making. It might not be the best course of action to focus solely on yield data when choosing crops. Instead, a thorough analysis that considers both yield and profitability becomes necessary. Farmers and decision-makers can choose which crops to grow more intelligently by taking into consideration both criteria, ensuring not just good yields but also long-term financial gains.

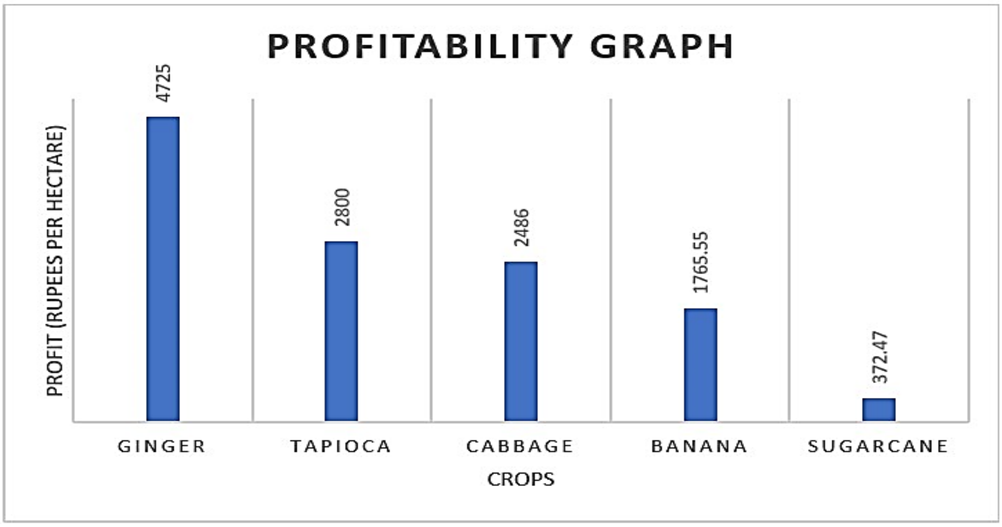


Figure 5. Crops recommendations rank wise.

4. Conclusion

The purpose of this research study is to create a recommendation model using machine learning techniques that can predict crop yield and recommend crops with the highest average yield and profitability. The dataset pertaining to the different crops grown in the Northeast region is considered by the recommendation system. The dataset for these crops is first preprocessed, and multiple regression models such as Decision Tree Regression, Linear Regression, and ensemble models, including XGBoost, Gradient Boost, Random Forest, and AdaBoost, are then employed to predict the yield and its accuracy. XGBoost outperforming the other models is then further enhanced through hyperparameter optimization to enhance the accuracy of our model. Ginger, Tapiaco, cabbage, banana, and sugarcane are the crops recommended rank wise for their maximum crop output and profitability. The highest R^2 Score that could be obtained using the models provided above is 98%. An advanced ensemble technique can be investigated in the future to increase the accuracy of yield prediction and crop recommendation.

References

Agri Farming. (2023). Cabbage Cultivation: Income, Cost, Profit - A Project Report. Agrifarming. in. Retrieved September, 13,2023, from https://www.agrifarming.in/cabbage-cultivation-income-cost-profit-project-report.

Asia Farming. (2023). Ginger Farming Business Plan: A Comprehensive Guide for Successful, Profitable Cultivation and Harvesting. Asia Farming. Retrieved September, 11, 2023. https://www.asiafarming.com/ginger-farming-business-plan-a-comprehensive-guide-for-successful-profitable-cultivation-and-harvesting.

Babu, S. (2013). *A Software model for precision agriculture and marginal farmers* Paper presented at the IEEE Global Humanitarian Technology of Conference: South Asia satellite (GHTC-SAS), Trivandrum, India. http://dx.doi.org/10.1109/GHTC-SAS.2013.6629944.

Deepa M., Sowmiya, V., Tamizhan, E., Venkat V.M.P., & Ranjani, S. (2023). Crop recommender system Based on Machine Learning. *International Journal for Innovative Research in a multidisciplinary field* https://doi.org/10.2015/IJIRMF/202303020)

Everingham, Y., Sexton, J., Skocaj, D., & Bamber, G. I. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, *36*(2), 27-35. http://dx.doi.org/10.1007/s13593-016-0364-z.

Garanayak, M., Sahu, G., Mohanty, S. N., & Jagadev, A. K. (2021). Agricultural recommendation system for crops using different machine learning regression methods. *International Journal of Agricultural and Environmental Information Systems (IJAEIS), 12*(1), 1-20. http://doi.org/10.4018/IJAEIS.20210101.oa1.

Government of India. (2023). Press Information Bureau, Government of India. https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1935899#:~:text=291.975%2Fqtl%20for%20sugarcane%20in,157%2Fqtl.

Kale, S. S., & Patil, P.S. (2019). *A Machine learning approach to predict crop yield and success rate* paper presented at IEEE Pune Section International Conference (PuneCon), Pune, India, 2019, 1-5. https://doi.org/10.1109/PuneCon46936.2019.9105741.

Kumar, P. (2018). India Crop Production - State wise. https://data.world/thatzprem/agriculture-india. Retrieved March, 04, 2023.

Kumar, R., Singh, M.P, Kumar, P., & Singh J.P. (2015). *Crop selection method to maximize crop yield rate using machine learning techniques* Paper presented at International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM). https://doi.org/10.1109/ICSTM.2015.7225403.

Panigrahi, B., Kathala, K.C.R., & Sujatha, M. (2023). *Machine Learning based Comparative Approach to Predict the Crop Yield using Supervised Learning with Regression Models* paper Presented at International Conference on Machine Learning and Data Engineering. https://doi.org/10.1016/j.procs.2023.01.241.

Patowary, M., Kumar, S., & Singh, V. (2022). A Study on Marketing aspects of Banana in Goalpara District of Assam. *IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), 15*(5), 01-08. https://doi.org/10.9790/2380-1505010108.

Paul, M., Vishwakarma, S.K., & Verma, A. (2015). *Analysis of Soil behavior and Prediction of Crop Yield using Data Mining Approach* Paper presented at International Conference of Computational Intelligence and Communication Networks. https://doi.org/10.1109/CICN.2015.156.

Potnuru, N. S., Pinapa V. S., Bollu, A.L., & Jabber, B. (2020). *Crop Yield Prediction based on Indian Agriculture using Machine Learning* Paper presented at 2020 International Conference for Emerging Technology (INCET). Belgaum, India. 1-4. https://doi.org/10.1109/INCET49848.2020.9154036

Renuka, & Terdal, S. (2019). Evaluation of Machine learning algorithms for Crop Yield Prediction. *International journal of engineering and advanced Technology*. pp 4082-4086 8(6). http://www.doi.org/10.35940/ijeat.F8640.088619.

Savla, A., Dhawan, P., Bhadada H., Israni, N., Mandholia, A., & Bhardwaj, S. (2015). *Survey of Classification algorithms for formulating yield prediction accuracy in precision agriculture* Paper presented at Innovations in Information, Embedded, and Communication Systems (ICIIECS). Coimbatore, India. 1-7. https://doi.org/10.1109/ICIIECS.2015.7193120.

Shastry, A., Sanjay, H. A., & Bhanusree, E. (2017). Prediction of crop yield using regression techniques. *International Journal of Soft Computing*, *12*(2), 96-102. DOI: 10.36478/ijscomp.2017.96.102

Singh, V., Sarwar, A., & Sharma, V. (2017). Analysis of soil and prediction of crop yield (Rice) using machine learning approach. *International Journal of Advanced Research in Computer Science*, *8*(5), 1254-1259.

Times of India. (2019). *Erode tapioca farmers reap profit after price shoots up*. Times of India. Retrieved September, 11, 2023. https://timesofindia.indiatimes.com/city/salem/erode-tapioca-farmers-reap-profit-after-price-shoots-up/articleshow/69934516.cms.

Ung, P. C., & Mittrapiyanuruk, P. (2018). *Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques* Paper presented at 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), Nakhonpothom, Thailand. 1-6. https://doi.org/10.1109/JCSSE.2018.8457391.

Yang, L. (2011). Classifiers selection for ensemble learning based on accuracy and diversity. *Procedia Engineering*, *15*, 4266-4270. https://doi.org/10.1016/j.proeng.2011.08.800.