Crop Prediction Using ML

Himanshi Goel[1], Aishwaryendra Narayan[2], Aman Singh[3], Anusthan Pandey[4], Shashwat Singh[5]
[1,2,3,4,5]*Department of Information Technology, Raj Kumar Goel Institute of Technology, Ghaziabad U.P.*

***Abstract*-The significance of agriculture in India's economy is widely acknowledged. This research paper focuses on predicting crop yields across the country, covering a wide range of crops. What distinguishes this study is its unique capability to forecast agricultural production for any chosen year, using easily understandable factors such as state, district, season, and area. To accomplish this, the article employs various regression techniques, including the notable Kernel Ridge, Lasso, and ENet algorithms. These advanced statistical methods serve as the foundation of the paper's prediction methodology, facilitating accurate estimations of crop output.**

Keywords: Crop yield prediction, Lasso, Kernel Ridge, ENet, Stacked Regression, Machine Learning (ML).

I. INTRODUCTION

In order to enhance comprehension and facilitate the study of the diverse range of crops cultivated in India, they are frequently classified into distinct categories or orders. This classification system aids researchers, policymakers, and farmers in conducting more efficient analyses and examinations of crops. By grouping crops into orders, it becomes simpler to identify both the similarities and differences among them, comprehend their specific cultivation needs, and formulate suitable strategies for their growth and maintenance. The categorization of crops into orders is typically based on shared characteristics, such as botanical attributes, growth patterns, duration of cultivation, climatic preferences, nutrient requirements, or economic importance [1].

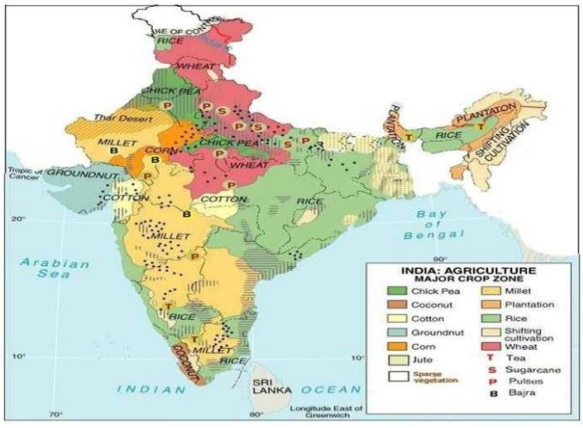

Fig.1.Famous Categories of crops over states in India(based on Season)

The dataset utilized in this study comprises more than 2.5 million data points, encompassing significant variables such as State, District, Crop, Season, Year, Area, and Product. Figure 1 visually depicts the geographical regions and ecosystems of India, along with the typical crop sequencing observed throughout the year. To enhance the accuracy of yield estimations and minimize errors, advanced regression techniques, including Lasso, ENet, Kernel Ridge, and model combining (mounding), were employed. These sophisticated approaches resulted in improved projections and refinement of the analytical process. The essay is structured into four main components: Literature Review, Methodology, Conclusion, and Future Work. The Literature Review section provides a comprehensive background for the study by examining previous research and relevant intellectual contributions pertaining to the subject matter.

II. BACKGROUND STUDY

The CRY algorithm, introduced by AnantharaM.G. et al. in February 2013, is a predictive model utilized for agricultural datasets. It employs beehive clustering techniques to forecast crop yields, with a specific focus on rice, sugarcane, and paddy yields in India. The researchers investigated parameters such as crop type, soil type, soil pH value, moisture, and crop sensitivity to develop their predictions [2].

In their research conducted in August 2019, Chawla, I. et al. combined fuzzy logic with statistical time series models to predict crop output. They considered variables such as temperature and rainfall, which are crucial factors influencing crop growth. The outcomes of their prediction were categorized into "good yield" and "very good yield," indicating satisfactory or exceptionally high crop output based on their forecast. The incorporation of fuzzy logic allowed the researchers to account for the uncertainty and imprecision inherent in agricultural systems [3].

Chaudhari, A. N. et al., in their study conducted in August 2018, aimed to enhance crop yield prediction accuracy by integrating three algorithms: clustering k-means, Apriori, and Bayes. By considering factors like cultivation area, rainfall patterns, and soil type, their system was designed to determine the most suitable crops for cultivation in specific areas [4].

In December 2017, Gandge, Y. conducted a study focused on colorful crops and applied various machine learning methods to the prediction process. However, further information about the specific techniques or outcomes of this study is not provided.

In a study conducted by Petkar, O. in July 2016, different techniques were explored to determine the most suitable methods for various crops. The research aimed to identify which approaches would yield the best results for each specific crop [5].

In a study conducted by Armstrong, L. J. and colleagues in July 2016, Artificial Neural Networks (ANNs) were employed to forecast rice yield in specific regions of Maharashtra, India. The study focused on climate variables such as temperature, rainfall, and reference crop evapotranspiration within a certain range. Historical data from the Indian Government's database spanning 1998 to 2002 was used for the analysis [6].

Petkar, O. introduced a decision support system in July 2016, which served as an interface for input provision and error handling in the context of rice crop yield prediction. This system allowed users to interact with the model by providing input parameters and addressing potential errors or uncertainties in the predictions [7].

In December 2018, Chakrabarty, A. et al. conducted a study on crop prediction in Bangladesh, specifically focusing on the cultivation of rice (three different types), jute, wheat, and potatoes [9].

Mariappan, A.K. and colleagues conducted research on rice crop statistics in Tamil Nadu, India. The study explored various fundamental aspects of climate that influence rice cultivation [10].

Kalbande, D.R. and colleagues employed Support Vector Regression (SVR), Multi Polynomial Regression, and Random Forest Regression in 2018 to forecast sludge yield. These regression models were evaluated using parameters such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) values [11].

III. METHODOLOGY

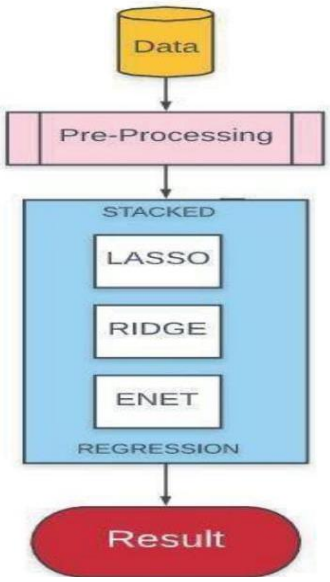

Fig.2 Process chart of the research project

A. Pre-processing:

The provided dataset for sludge yield contains several 'NA' values, which typically indicate missing or null values. In Python, a pre-processing step was applied to handle these missing values. This process involves filtering the dataset to appropriately handle the 'NA' values. Techniques such as removing rows or columns with 'NA' values or imputing them with estimated values based on other data points can be employed.

B. Stacked Regression:

The described process refers to a form of ensemble learning known as stacking with averaging. It involves combining multiple models and incorporating a meta-model to enhance overall predictive performance.

Step 1: The training dataset is divided into two sets - the "train" set and the "holdout" set. Step 2: The base models (e.g., Support Vector Regression, Multi Polynomial Regression, Random Forest Regression) are trained using the training set. Step 3: The trained base models are tested using the holdout set. Step 4: The predictions obtained from the base models on the holdout set are used as inputs for the meta-model. The meta-model, represented as the advanced-position learner (e.g., Lasso Regressor), is trained using these holdout set predictions as features and the actual target variable values.

The training data is divided into five folds or subsets, either randomly or using a specific sampling strategy. In

each iteration, each base model is trained using four out of the five folds, while the remaining fold serves as the holdout or test set. The base models make predictions on the holdout fold, and these predictions are considered as the output of the base models for that particular fold. The predictions from all the base models on the holdout folds (across iterations) are combined to form meta-features. These meta-features, along with the original features, are used as inputs for the meta-model. The meta-model is trained on the complete dataset, incorporating both the original features and the meta-features generated from the base models. The meta-model learns to make predictions based on these combined features and the actual target variable.

Figure 2 represents the meta-model (in this case, the Lasso Regressor), which operates on the combined features. The operation of stacked regression is illustrated in Figure 3.

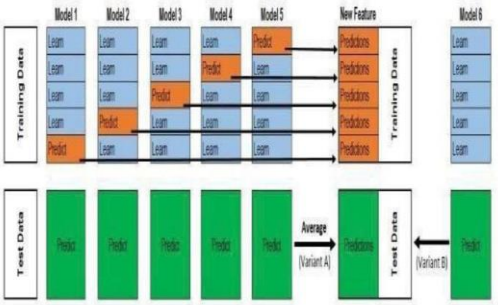

Fig. 3. Stacked Regression

C. Output:

In this design, the root mean square error (RMSE) is utilized as the performance metric. RMSE measures the differences between predicted values and actual values in a regression problem. When the models were applied individually, the ENet model exhibited an error of approximately 4, the Lasso model had an error of around 2, and the Kernel Ridge model had an error of about 1. However, after combining or stacking the models (referred to as "mounding" in this context), the overall error reduced to below 1.

To obtain the predictions depicted in Figure 4, the stacked regression approach was employed, incorporating the outputs of the individual models. The exact process and details of how the predictions were obtained are not provided, but it can be inferred that the stacked model's combined predictions contributed to the results presented in Figure 4.

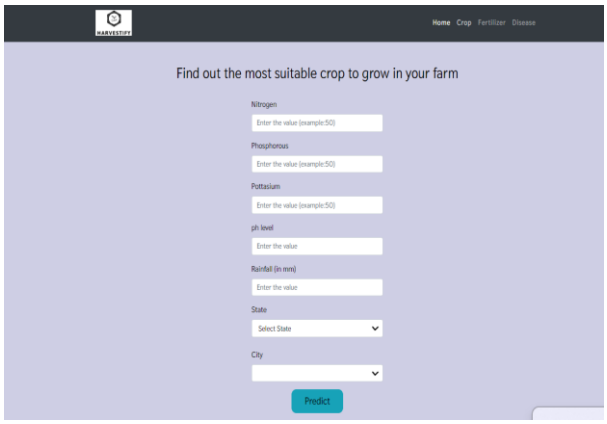

Fig. 4. Interface of Web APP

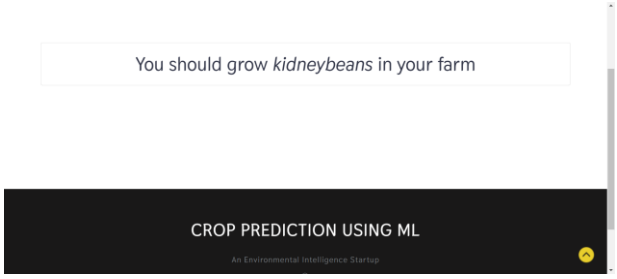

Fig. 5. Crop Predicted by System (Output)

IV. CONCLUSION AND FUTURE WORK

The integration of stacked regression in our model has resulted in significant improvements compared to using individual models separately. By inputting variables such as State, City, pH level, Rainfall, Nitrogen, Phosphorus, and Potassium, our model can predict the most suitable crops based on these inputs. Currently, the model is accessible as an online application, allowing users to input the relevant parameters and receive crop recommendations. However, as part of our future work, we plan to develop a dedicated mobile application that can be easily accessed and utilized by farmers. This will provide a more user-friendly and convenient platform for farmers to benefit from the crop recommendation system. Additionally, we aim to enhance the accessibility and usability of the system by translating the entire application into regional languages. This will enable farmers from diverse linguistic backgrounds to effectively utilize the application without facing language barriers.

REFERENCE

[1] “data.gov.in.”[Online].Available:https://data.gov.in/
[2] Ananthara, M. G., Arunkumar, T., & Hemavathy, R. (2013, February). CRY—an improved crop yield

prediction model using bee hiveclustering approach for agricultural data sets. In 2013International Conference on Pattern Recognition, Informatics and Mobile Engineering (pp. 473-478). IEEE.
[3] Bang, S., Bishnoi, R., Chauhan, A. S., Dixit, A. K., & Chawla, I. (2019, August). Fuzzy Logic based Crop Yield Prediction using Temperature and Rainfall parameters predicted through ARMA, SARIMA, and ARMAX models. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1- 6). IEEE.
[4] Bhosale, S. V., Thombare, R. A., Dhemey, P. G., & Chaudhari, A. N. (2018, August). Crop Yield Prediction Using Data Analytics and Hybrid Approach. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.
[5] Gandge, Y. (2017, December). A study on various data mining techniques for crop yield prediction. In 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT) (pp. 420-423). IEEE.
[6] Gandhi, N., Petkar, O., & Armstrong, L. J. (2016, July). Rice crop yield prediction using artificial neural networks. In 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 105-110). IEEE.
[7] Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016, July). Rice crop yield prediction in India using support vector machines. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-5). IEEE.
[8] Gandhi, N., Armstrong, L. J., & Petkar, O. (2016, July). Proposed decision support system (DSS) for Indian rice crop yield prediction. In 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 13-18). IEEE.
[9] Islam, T., Chitty, T. A., & Chakrabarty, A. (2018, December). A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh. In 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 1-6). IEEE.
[10] Mariappan, A. K., & Das, J. A. B. (2017, April). A paradigm for rice yield prediction in Tamilnadu. In 2017 IEEE Technological Innovationsin ICT for Agriculture and Rural Development (TIAR) (pp. 18-21). IEEE.
[11] Shah, A., Dubey, A., Hemnani, V., Gala, D., & Kalbande, D. R. (2018). Smart Farming System: Crop Yield Prediction Using Regression Techniques. In Proceedings of International Conference on Wireless Communication (pp. 49-56). Springer, Singapore.
[12] https://github.com/Gladiator07