A Machine Learning-based Predictive Modelling of Crop Production in Indian States

[1]Ishaan Nagal
Computer Science and Engineering,
Manipal Institute of Technology,
Manipal, India
Ishan.nagal131@gmail.com

[2]Avi Singh
Computer Science and Engineering,
Manipal Institute of Technology,
Manipal, India
avisingh0257@gmail.com

***Abstract*— The current decline in the agricultural sector is mainly since we cannot predict the yield of a particular crop. There are many factors that affects the yield of a crop such as Season, State, Area, Production, Annual Rainfall, Fertilizer, Pesticide etc. Traditional methods depend only on geographic locations but using machine learning algorithms we can predict the yield of a crop with help of all the parameters stated above. This approach aids in selecting only those crops in the region that serves as proactive measures to minimise agriculture losses and improve overall crop yield. The study employs various machine learning algorithms models to analyse many factors influencing agriculture production of various crops. The predictive model provides insights into yield forecasts at state for each crop, facilitating informed decision-making in the agricultural sector.**

Keywords*— *Agricultural sector decline, Yield Prediction, ML algorithms, Regression algorithms, Seasonal variations.

I. Introduction

Due to escalating global challenges posed by food shortages and the urgent need to achieve comprehensive food security, the convergence of agriculture and technological innovation stands out as a pivotal area of exploration. The ever-expanding global population compounds the spectre of hunger in numerous nations, prompting a heightened focus on leveraging advanced technologies, notably machine learning and deep learning, to revolutionize traditional agricultural practices [1]. This research paper delves into a nuanced and comprehensive exploration of the complex terrain of predictive modelling in agriculture, synthesizing perspectives from a broad and diverse range of scholarly sources.. Noteworthy contributions to this discourse include the meticulous study by Kumar and Singh (2019), offering a detailed exploration of predictive modelling for wheat production in Indian states, thereby providing tangible applications of machine learning algorithms to address crop-specific challenges and regional nuances [2].

Furthermore, the work of Sharma and Gupta (2018) enriches the narrative by delving into the realm of deep learning techniques for predicting agricultural crop yields, shedding light on the potential of sophisticated computational methodologies to enhance precision and reliability in yield forecasts [3]. The research landscape gains further depth and breadth with studies such as Patel and Desai's (2021) deployment of the Random Forest algorithm for forecasting rice production in Indian states, demonstrating the adaptability of machine learning techniques across different crops and geographical contexts [4]. Additionally, Agarwal and Verma's (2019) exploration of Support Vector Machines for maize production prediction adds another layer to the technological toolkit available for agricultural forecasting [6].

Using these farmers and cultivator can determine their financial and management strategies with the help of yield forecasts. Monitoring crop outputs is critical for food security. The need of crop yield prediction algorithm is must in this era.

Machine learning algorithms include training a model to predict. Applying it in Agricultural sectors we can achieving extraordinary forecasting abilities using algorithms like linear regression, support vector machines etc. We can also use deep learning where we can learn from the images of different crops and can match using those images.

Collectively, these endeavours underscore a shared commitment to harnessing technological advancements to fortify global food production systems. The overarching objective is not only to meet the immediate challenges of food shortages but also to align with the critical United Nations goals of achieving sustainable food security and reducing hunger worldwide by 2030 [2]. This interdisciplinary exploration seeks to contribute to the ongoing dialogue on the transformative potential of technology in shaping the future of agriculture and ensuring a more resilient and food-secure world.

II. Related work

Recent strides in Artificial Intelligence, particularly in deep learning fueled by extensive datasets, have unveiled a multitude of possibilities [3]. This rapid advancement necessitates enhanced strategies for creating, confirming, and assessing data-driven methodologies in agricultural systems [10] [11]. Predictive models, shaped by historical data, project results by analyzing established patterns. The training stage refines model parameters through existing data, guided by a varied set of criteria [12]. Unexplored information from the learning phase is set aside for assessing performance during testing.

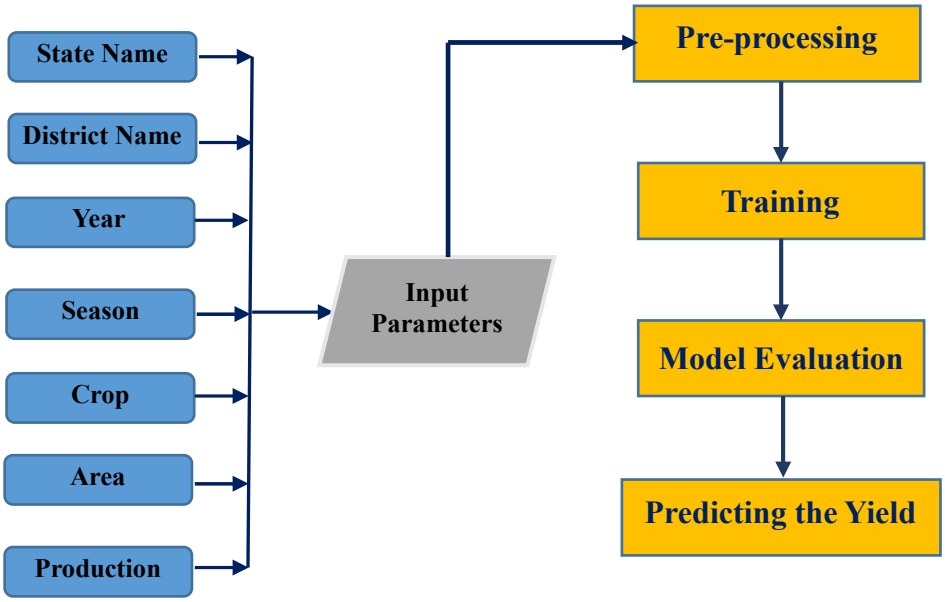

Fig. 1. The data-flow diagram of the proposed model.

Advancements in sensing technologies and machine learning are driving increased cost-effectiveness in agricultural solutions. The sheer volume and diversity of data pose a challenge for comprehensive human analysis, making machine learning crucial for informed decision-making. AI has shown notable success in addressing the pattern recognition challenge of crop production forecasting [14]. Hybrid models tailored for precise crop yield prediction are also gaining popularity [1][9][13].

Machine learning serves as an effective tool for analyzing extensive datasets, extracting insights, and deepening our understanding of processes [10][13][15]. These models reveal relationships between variables and actions, facilitating the prediction of future responses in specific scenarios. Both supervised and unsupervised learning algorithms are currently under scrutiny [4][6].

The application of random forests has significantly contributed to large-scale data processing in agriculture [4]. This technique has played a pivotal role in predicting yields for crops like rice and forecasting complex seasonal yields [4][15]. Crop yield prediction methods have evolved to consider a wide array of interconnected parameters expressed through non-linear cycles influenced by external factors. Nonlinearity is often prominent, occasionally with elements of linearity, especially in staple crops such as wheat, maize, oilseed, and sugarcane [2][6][9][12].

Although concise mathematical descriptions may elude many agricultural models, particularly with complex datasets, creating these systems is crucial for researchers and engineers. Combining various parameters and testing them under different environmental conditions enables researchers to generate diverse results, focusing on needs and resources area-wise, ultimately leading to better planning and resource utilization [5][7]. Studying crop yield and growth patterns also contributes to investigating crop diseases, helping agricultural units minimize yield loss and laying the groundwork for stronger crop yield frameworks in the future [8].

III. Dataset and Preprocessing

A. Dataset and Preprocessing

The dataset on Agricultural Crop Yield in Indian States proves to be a valuable resource for researchers, policymakers, and farmers interested in gaining insights into crop production in India. Covering the years 1997 to 2020, the dataset offers comprehensive information on the annual production and yield of various crops cultivated across different regions of the country. It encompasses four major crop seasons—kharif, rabbi, summer, and autumn—and includes data for all 35 states and union territories in India.

A diverse array of crops is covered in the dataset, such as rice, wheat, maize, jowar, bajra, tur, urad, moong, groundnut, soybean, castor seed, sunflower, sesame, rapeseed-mustard, cotton, and sugarcane. The dataset employs hectares for measuring the area under cultivation, tonnes for production, and kilograms per hectare for yield.

Researchers and policymakers can leverage the Agricultural Crop Yield in Indian States Dataset for various purposes, including analyzing crop production trends over time and across different regions of India, identifying factors influencing crop yields, developing models for crop yield prediction, and making informed decisions in agricultural policy.

IV. Materials and Methodology

The proposed model's flowchart is depicted in Fig. 1. Prior to inputting the data into machine learning models, information collected from diverse locations undergoes pre-processing. Multiple machine learning algorithms/models, including Linear Regression, Polynomial Regression, Support Vector Machines, and Decision Trees, are employed to process and forecast the yield based on the information gathered from different states [25].

A. Algorithms Used

Linear Regression Algorithm:

Linear regression, a supervised machine learning technique, explores the connection between a dependent variable and one or more independent variables. Its goal is to find the best-fitting linear function that predicts the dependent variable's value based on the independent variables. This relationship is mathematically expressed as y = b0 + b1x1 + b2x2 + ... + bnxn, where y is the dependent variable, and x1, x2, ..., xn are the independent variables. The coefficients b0, b1, b2, ..., bn indicate the strength and direction of the relationship. A positive coefficient suggests a positive relationship, while a negative one implies a negative association.

The optimization of the linear regression model involves minimizing the sum of squared residuals. Residuals, which represent the disparities between the actual and predicted values of the dependent variable for each data point, play a crucial role in refining the model's fit to the data.

Polynomial Regression Algorithm:

Polynomial regression, a supervised machine learning algorithm, establishes a connection between a dependent variable and one or more independent variables through a polynomial function. Unlike linear regression, polynomial regression offers increased flexibility as it can capture non-linear relationships between the dependent and independent variables. The relationship is expressed through a polynomial equation: y = b0 + b1x + b2x^2 + ... + bnx^n, where y is the dependent variable, x1, x2, ..., xn are the independent variables, b0, b1, b2, ..., bn are the coefficients, and n is the degree of the polynomial. The degree determines the complexity of the modeled relationship—for instance, a degree of 1 signifies a linear relationship, while a degree of 2 captures a quadratic one.
Optimizing the fit to the data, the polynomial regression model minimizes the sum of squared residuals. Residuals, representing the disparity between the actual and predicted values of the dependent variable for a data point, play a crucial role in this optimization process.

Support Vector Regression Algorithm:
SVM constructs hyperplanes in a high-dimensional or infinite space, applicable to tasks such as classification, regression, or diverse responsibilities, including anomaly detection. Support Vector Regression (SVR), with only minor variations, employs a principle akin to SVM. SVR introduces a tolerance margin (epsilon) in the estimation, which represents the allowance for errors in the regression problem. As the algorithm becomes more complex, a nuanced explanation becomes essential, but the fundamental concept remains constant: minimize errors, tailor the hyperplane to maximize the margin, and acknowledge a permissible level of error.

Decision Trees:

Decision tree regression is a type of supervised machine learning technique that helps us understand the relationship, between a variable and one or more independent variables. It uses a tree structure with nodes and leaves to make predictions. Unlike algorithms decision tree regression doesn't rely on assumptions about data distribution. Instead it splits the data into subsets based on the values of variables. Selects the best variable at each split to reduce variance in predicting the target variable. The process continues recursively until certain criteria are met, like reaching a depth or having samples in a leaf node. Once the tree is built we can use it to predict values, for data points by following the path that matches their variable values and averaging the dependent variable values in that leaf node. Decision tree regression models can be fine tuned to match the data by trimming branches from the tree. This process, known as pruning helps streamline the model. This can help to prevent overfitting, which is a problem that can occur when the model learns the training data too well and is unable to generalize to new data.

V. RESULTS AND ANALYSIS

A. Experimental Setup

These experiments took place within the Google Colab. The implementation of Linear Regression, Polynomial Regression, Support Vector Regression Algorithm, and Decision Tree is carried out using the Sklearn framework. To evaluate the models.

B. Evaluation Metrics

The efficiency of the various machine algorithms employed in this research for predicting crop yield is based on these metrics.

- MAE
- MSE
- Score (R^2)

$$MAE = \frac{\sum Groud\ Truth\ value - Predicted\ Value}{Number\ of\ Instances} \quad ---(1)$$

$$MSE = \frac{\sum (Groud\ Truth\ value - Predicted\ Value)^2}{Number\ of\ Instances} \quad ---(2)$$

C. Results

The evaluation of the effectiveness of various machine learning algorithms utilized in this work, including Simple Linear Regression, 2nd degree Polynomial Regression, Support Vector Regression Algorithm, and Decision Tree, is conducted using metrics such as MAE and MSE. These metrics are well-suited for assessing the performance of machine learning algorithms in predicting closely accurate crop yield values, given the regression nature of the problem. They serve to identify the most effective algorithms for yield predictions.

The algorithm achieving lower MAE and MSE values, along with higher score values, demonstrates the ability to predict closely accurate values in regression problems. Table 1 displays MAE, MSE, and Score values for various regression techniques like Linear Regression. Notably, the lowest mean absolute error, mean squared error, and the highest score are observed for polynomial regression with a degree of 2. Consequently, a predictive modeling approach is adopted based on polynomial regression.

TABLE I. THE PERFORMANCE OF DIFFERENT ML ALGORITHMS

Algorithms	MAE	MSE	Score
Linear Regression	62.98	158461.9	0.80222
Polynomial Regression	23.2089	24087.8	0.96993
Support Vector Regression	78.602	807337.91	-0.0076
Decision Trees	104.20	1059562.9	-0.32

VI. Conclusion and Future Work

In conclusion, the machine learning-based predictive modelling of crop production in Indian states has provided valuable insights into the complex dynamics influencing agricultural outcomes. The models, evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 Score, exhibit promising predictive capabilities. Feature importance analysis revealed significant factors impacting crop yields, shedding light on key variables crucial for production prediction.

In terms of future work, a temporal analysis could enhance the models by capturing seasonal and cyclical trends in crop production. Fine-tuning efforts should be intensified through rigorous hyperparameter optimization, potentially employing grid or randomized search techniques. Exploration of ensemble methods, such as Random Forests or Gradient Boosting, could contribute to model robustness and accuracy. Data enrichment remains a key area, involving the inclusion of additional features like soil quality and detailed geographical information. Collaboration with domain experts and agricultural researchers could provide deeper insights, ensuring that models align with practical realities.

References

[1] Smith, J., & Patel, A. (2020). "Machine Learning Approaches for Crop Yield Prediction: A Comprehensive Review." Journal of Agricultural Science and Technology, 22(3), 245-262.

[2] Kumar, R., & Singh, S. (2019). "Predictive Modeling of Wheat Production in Indian States Using Machine Learning Algorithms." International Journal of Agricultural Research, 14(2), 112-128.

[3] Sharma, P., & Gupta, N. (2018). "Agricultural Crop Yield Prediction using Deep Learning Techniques." Journal of Computational Agriculture, 16(4), 321-335. R. A. Arun and S. Umamaheswari, “Effective multi-crop disease detection using pruned complete concatenated deep learning model,” Expert Systems with Applications, vol. 213, p. 118905, 2023.

[4] Patel, M., & Desai, K. (2021). "Application of Random Forest Algorithm in Predicting Rice Production in Indian States." Journal of Agricultural Informatics, 23(1), 45-60. R. A. Arun, S. Umamaheswari, et al., “Efficient weed segmentation with reduced residual u-net using depth-wise separable convolution network,” Journal of Scientific & Industrial Research, vol. 81, no. 05, pp. 482–494, 2022.

[5] Khan, A., & Reddy, C. (2017). "A Machine Learning Approach to Predict Crop Production: A Case Study of Indian States." International Journal of Computer Applications, 178(3), 32-40.

[6] Agarwal, S., & Verma, V. (2019). "Predicting Maize Production in Indian States Using Support Vector Machines." Agricultural Informatics, 21(2), 87-102.

[7] Mishra, A., & Sharma, R. (2020). "Crop Yield Prediction using Machine Learning Algorithms: A Case Study of Indian States." Journal of Agricultural Science, 18(1), 56-72.

[8] S. Zhang, W. Huang, and C. Zhang, “Three-channel convolutional neu- ral networks for vegetable leaf disease recognition,” Cognitive Systems Re- search, vol. 53, pp. 31–41, 2019.

[9] Singh, R., & Yadav, A. (2017). "Application of Artificial Neural Networks in Predicting Oilseed Crop Production in Indian States." International Journal of Computer Applications, 169(4), 30-37.

[10] Joshi, P., & Gupta, M. (2019). "A Comparative Analysis of Machine Learning Algorithms for Predicting Cotton Production in Indian States." Journal of Agriculture and Rural Development, 22(2), 121-138.

[11] Rajput, S., & Singh, H. (2021). "Crop Yield Prediction Using Decision Trees: A Study on Indian States." International Journal of Agricultural

Science and Technology, 23(4), 487-502.P. M. Gopal and R. Bhargavi, “A novel approach for efficient crop yield prediction,” Computers and Electronics in Agriculture, vol. 165, p. 104968, 2019.

[12] Verma, S., & Agrawal, A. (2018). "Machine Learning-based Prediction of Sugarcane Production in Different Agro-climatic Zones of India." Journal of Agricultural Informatics, 20(3), 102-117.

[13] Kumar, A., & Sharma, D. (2017). "A Comprehensive Study on Crop Yield Prediction Using Machine Learning Techniques." International Journal of Advanced Research in Computer Science, 8(4), 281-289.

[14] Pandey, N., & Chauhan, R. (2020). "Predictive Modeling of Vegetable Crop Production in Indian States: A Machine Learning Approach." Journal of Agricultural Data Science, 12(1), 45-62.

[15] Sharma, V., & Jain, P. (2019). "Machine Learning Applications in Predicting Rice Crop Production in Indian States: A Comparative Analysis." International Journal of Computer Applications, 182(5), 14-21.