

International Journal of Sustainable Agricultural Management and Informatics

ISSN online: 2054-5827 - ISSN print: 2054-5819
<https://www.inderscience.com/ij sami>

Enhancing crop yield prediction through machine learning regression analysis

Seema Sharma, Anupriya Jain, Sachin Sharma, Pawan Whig

DOI: [10.1504/IJSAMI.2025.10067961](https://doi.org/10.1504/IJSAMI.2025.10067961)

Article History:

Received:	27 September 2023
Last revised:	06 November 2023
Accepted:	13 November 2023
Published online:	03 December 2024

Enhancing crop yield prediction through machine learning regression analysis

Seema Sharma, Anupriya Jain and
Sachin Sharma

Faculty of Computer Application,
MRIIRS,
Faridabad, India
Email: seema.sca@mriu.edu.in
Email: anupriya.sca@mriu.edu.in
Email: sachin.sca@mriu.edu.in

Pawan Whig*

Vivekananda Institute of Professional Studies-TC,
New Delhi, India
Email: pawanwhig@gmail.com
*Corresponding author

Abstract: The economic prosperity of any nation hinges significantly on its agricultural output, a cornerstone of sustained growth. The integration of advanced technology plays a pivotal role in enhancing agricultural productivity. Farmers today are leveraging breakthroughs in data mining, the internet of things (IoT), artificial intelligence (AI), and machine learning to optimise their practices. This paper is dedicated to the exploration of this transformative field, offering insights into its multifaceted applications. It delves into the assessment of diverse parameters to facilitate the cultivation of specific crops in a given region. Moreover, it collaborates closely with farmers, tailoring these parameters to maximise crop yields and diversify agricultural produce. The study also incorporates the deployment of AI algorithms, such as logistic regression and multiple regressions, to bolster decision-making processes in agriculture.

Keywords: agriculture; artificial intelligence; logistic regression; machine learning; data mining.

Reference to this paper should be made as follows: Sharma, S., Jain, A., Sharma, S. and Whig, P. (2025) 'Enhancing crop yield prediction through machine learning regression analysis', *Int. J. Sustainable Agricultural Management and Informatics*, Vol. 11, No. 1, pp.29–47.

Biographical notes: Seema Sharma is an esteemed Associate Professor at the School of Computer Applications, MRIIRS. Her academic journey reflects a deep-rooted dedication to computer science and its applications. Proficient in software engineering and information technology, her expertise extends to data management and analysis. With a passion for teaching and research, she actively contributes to advancing knowledge in her field. Her commitment to empowering students and promoting technological advancements is evident through her teaching excellence and scholarly pursuits, making her an invaluable asset to MRIIRS.

Anupriya Jain is a distinguished Professor at the School of Computer Applications, MRIIRS, is a trailblazer in the realm of computer science education. With an illustrious career, she possesses unparalleled expertise in software development and computational intelligence. Her academic prowess coupled with a passion for research has contributed significantly to the field. Jain's commitment to fostering innovation is evident through her mentorship and academic leadership. Her dedication to shaping future technologists is reflected in her teaching excellence and groundbreaking contributions to advancing computer applications at MRIIRS.

Sachin Sharma is an Associate Professor at the School of Computer Applications, MRIIRS. With a profound passion for technology, his expertise spans software development, data analytics, and machine learning. He holds a distinguished academic background coupled with extensive industry experience, contributing significantly to both educational realms and practical application. His commitment to fostering innovation in computer applications is exemplified through his teaching, research endeavours, and active engagement in technological advancements.

Pawan Whig is a distinguished IEEE lifetime member, the Country Head at Threws, and is a reputable Research Consultant in Machine Learning. He holds a Google Certification in Digital Marketing and serves as the Dean of Research at VIPS-T. His expertise spans VLSI Design, AI, and ML, showcasing a diverse research portfolio. With a focus on cutting-edge technologies, he has made notable contributions in these areas. As an industry expert and academician, his multidisciplinary proficiency and commitment to innovation underline his significant impact in advancing research and technological landscapes.

1 Introduction

Agriculture, as the cornerstone of human civilisation, has demonstrated remarkable resilience in adapting to the evolving demands of a growing global population. As the demand for agricultural products continues to surge, so does the need for efficient crop management and resource allocation. Central to achieving these goals is the accurate prediction of crop yields, a fundamental aspect that empowers farmers, agricultural organisations, and policymakers to make informed decisions, optimise production, and secure food supplies for the world's population.

Traditionally, crop yield prediction relied on historical data, expert knowledge, and conventional statistical methods. However, the agricultural landscape is undergoing a transformation due to the advent of data-driven technologies and the availability of vast datasets. This shift has ushered in a new era of agricultural forecasting, with machine learning, particularly regression analysis, emerging as a potent tool for predicting crop yields with unprecedented accuracy and granularity.

The essence of this research paper lies in its comprehensive exploration of crop yield prediction using machine learning regression analysis. Our objective is to delve into this evolving field, elucidating the methodologies, challenges, and benefits associated with this innovative approach. The study encompasses various critical facets, including data collection, feature engineering, model selection, and performance evaluation.

Data collection is a pivotal starting point for any predictive model. The quality and comprehensiveness of the data used for training and validation are central to the model's

success. In our research, we have undertaken an exhaustive data collection process, encompassing a wide array of variables. These variables range from meteorological conditions and soil characteristics to crop types and historical yield records. This rich and diverse dataset serves as the foundation upon which our machine learning models are constructed.

As we move forward in the subsequent sections, we delve into the intricacies of machine learning regression analysis. We harness a variety of regression algorithms, each designed to serve specific purposes. Linear regression is utilised for establishing fundamental relationships between input variables and crop yields. Decision trees, on the other hand, prove invaluable for capturing complex interactions within the data, allowing for a more nuanced understanding of the factors affecting crop yields. Support vector machines (SVMs) are employed to handle high-dimensional data, ensuring that the models can effectively manage the wealth of information available.

Feature selection and engineering constitute integral components of our analysis. These techniques enhance the predictive accuracy of our models by identifying the most influential factors and refining the dataset. This process ensures that the model focuses on the most critical variables, thereby delivering improved results.

Furthermore, we delve into the assessment of model performance, employing established metrics such as mean absolute error (MAE) and root mean square error (RMSE). These metrics provide a quantitative measure of the models' accuracy and reliability. Additionally, we emphasise the importance of model interpretability. By uncovering the underlying factors that influence crop yields, we enhance the practicality of our predictive models, which, in turn, aids in the development of actionable strategies for crop management.

The adjustment of diverse parameters to enhance crop cultivation in specific regions is a complex and dynamic process. Researchers and agricultural experts collaborate closely with farmers to ensure that their findings are both relevant and applicable to the local context. This collaboration involves several key steps. Firstly, scientists conduct extensive field research to understand the unique environmental conditions, soil properties, and climate patterns of a particular region. Based on this data, they develop crop management strategies that consider local constraints and opportunities.

To ensure the applicability of their findings, researchers often engage in knowledge sharing sessions with farmers. These interactions provide a platform for farmers to express their concerns, challenges, and goals. Scientists take this valuable feedback into account when fine-tuning their recommendations. Additionally, they offer training and workshops to empower farmers with the necessary knowledge and skills to implement the suggested practices effectively.

The use of technology, such as precision agriculture tools and data analytics, plays a crucial role in tailoring crop cultivation to local conditions. Farmers are educated on the use of these technologies, which can help optimise resource usage, enhance crop yields, and mitigate risks.

The success of these collaborative efforts lies in the mutual understanding and respect between researchers and farmers. By actively involving farmers in the research process and addressing their specific needs, the agricultural community can develop more precise and sustainable cultivation practices that benefit both the environment and local economies.

Our research is driven by a desire to illuminate the potential of machine learning regression analysis in the realm of crop yield prediction. We recognise that by harnessing the power of data science, we can facilitate more informed decision-making processes in agriculture. This, in turn, leads to improved resource allocation, risk management, and sustainability practices. Ultimately, our efforts align with the broader goal of ensuring global food security and promoting sustainable agricultural practices in an ever-changing world.

The convergence of agriculture and machine learning presents a promising path toward addressing the food security challenges of our time. As technology continues to advance, so too will our ability to harness data-driven insights to ensure a more stable and sustainable future for our global agricultural systems.

2 Literature review

In recent years, agricultural research has witnessed a remarkable shift towards harnessing the potential of data-driven technologies and machine learning. Numerous studies have explored the application of these methods in various facets of agriculture. Vinciya et al. (2016) introduced the concept of utilising data mining techniques for high-tech farming, shedding light on data mining's role in agriculture. Koley (2014) delved into the realm of soil fertility and nutrient management, showcasing the practical application of neural networks for predicting soil fertility. Hong et al. (2016) advanced the field with data-driven soil moisture prediction, employing innovative approaches to collect and predict soil moisture. Masri et al. (2015) contributed by utilising extreme learning machines for soil property prediction, offering a robust methodology for soil property assessment. Singh et al. (2017) brought machine learning into the spotlight by analysing soil and crop yield relationships, paving the way for data-driven insights in agriculture. These studies were complemented by research on laboratory methods for soil and plant analysis by Okalebo et al. (2002), providing a reference for fundamental analysis techniques. Ghadge et al. (2018) made significant strides in crop yield prediction through machine learning, opening new avenues for precise agricultural planning. Moreover, a comprehensive overview of global food and agriculture was presented by Pocketbook (2015), serving as a valuable resource for understanding the broader agricultural landscape. Sowmya et al. (2020) extended these advances by forecasting crop yield based on climatic parameters, utilising machine learning for predictive modelling. Gonzalez-Sanchez et al. (2014) critically explored the impact of attribute selection on regression models, an essential consideration in crop yield prediction. Zhang et al. (2010) compared various regression models for predicting crop yields, offering insights into model selection. Zaw and Naing (2009) focused on rainfall prediction using polynomial regression, addressing the crucial aspect of water availability. Furthermore, Shastry et al. (2015) introduced the innovative ANFIS model for crop yield prediction, enriching the methodological diversity in this domain. Shabani et al. (2020) compared different regression techniques for pan evaporation modelling, expanding the toolbox of predictive modelling in agriculture. Ishak et al. (2021) pioneered an intelligent platform for crop prediction, potentially revolutionising decision-making processes. Gandhi et al. (2016) applied SVMs for rice crop yield prediction in India, emphasising the adaptability of machine learning methods to local contexts. Stas et al. (2016) specialised in wheat yield prediction using NDVI time series data, offering a data-driven approach to crop yield

forecasting. Dhaka and Vikas (2014) explored wheat yield prediction using neural networks, demonstrating the applicability of artificial intelligence (AI) in crop yield estimation. Additionally, Debow et al. (2023) employed deep learning for predicting and forecasting water quality, highlighting the integration of advanced technologies for improved agricultural practices. Athanasiadis and Ioannides (2021) ventured into the realm of wine quality prediction, employing machine learning techniques such as random forest and LASSO for precise assessments. Singh Attwal and Dhiman (2020) conducted an in-depth investigation into data mining techniques for crop yield prediction, enhancing the methodological toolkit available to researchers. Zakir et al. (2021) addressed the critical aspect of soil utilisation prediction for farmers, using machine learning to provide insights into resource allocation. Saranya et al. (2021) introduced an efficient method for crop yield prediction using bootstrap aggregative MapReduce Rocchio classification, facilitating more precise yield estimates. These studies collectively underscore the transformative impact of data-driven and machine learning-based approaches in advancing sustainable agriculture and decision-making processes in an evolving agricultural landscape. These varied approaches collectively contribute to the evolving landscape of agricultural technology, facilitating data-driven decisions for farmers and bridging the gap between technology and agriculture.

Table 1 Literature review with gap analysis

<i>Reference</i>	<i>Author(s)</i>	<i>Research focus</i>	<i>Research gap/contribution</i>
Vinciya and Valarmathi (2016)	Vinciya, P. and Valarmathi, A.	Agriculture analysis for high-tech farming in data mining	Introduction to data mining in agriculture
Koley (2014)	Koley, S.	Soil fertility and nutrient management using neural networks	Application of neural networks for soil fertility prediction
Hong et al. (2016)	Hong, Z., Kalbarczyk, Z. and Iyer, R.K.	Data-driven soil moisture prediction	A data-driven approach for soil moisture collection
Masri et al. (2015)	Masri, D., Woon, W.L. and Aung, Z.	Soil property prediction using extreme learning machines	Application of extreme learning machines for soil property prediction
Singh et al. (2017)	Singh, V., Sarwar, A. and Sharma, V.	Analysis of soil and crop yield using machine learning	Using machine learning for soil and crop yield analysis
Okalebo et al. (2002)	Okalebo, J.R., Gathua, K.W. and Woomer, P.L.	Laboratory methods of soil and plant analysis	A reference for laboratory methods of soil and plant analysis
Ghadge et al. (2018)	Ghadge, R., Kulkarni, J., More, P., Nene, S. and Priya, R.L.	Prediction of crop yield using machine learning	Application of machine learning for crop yield prediction
Tomar et al. (2023)	Tomar, P.K., Sobti, R., Prasad, R., Prasanna, K.L., Jain, A. and Sharma, S.	Implementation of artificial intelligence for better irrigation system	Implementation of AI for improved irrigation

Table 1 Literature review with gap analysis (continued)

<i>Reference</i>	<i>Author(s)</i>	<i>Research focus</i>	<i>Research gap/ contribution</i>
Sowmya et al. (2020)	Sowmya, S., Sri, D., Meghana, J. and Latha, P.	Forecasting crop yield based on climatic parameters	Machine learning for crop yield forecasting based on climate parameters
Gonzalez-Sanchez et al. (2014)	Gonzalez-Sanchez, A., Frausto-Solis, J. and Ojeda-Bustamante, W.	Attribute selection impact on regression models for crop yield prediction	Investigating the impact of attribute selection on crop yield models
Zhang et al. (2010)	Zhang, L., Lei, L. and Yan, D.	Comparison of two regression models for crop yield prediction	Comparative analysis of regression models for crop yield prediction
Zaw and Naing (2009)	Zaw, W.T. and Naing, T.T.	Modelling of rainfall prediction using polynomial regression	Focus on rainfall prediction using polynomial regression
Shastri et al. (2015)	Shastri, A., Sanjay, H.A. and Hegde, M.	A parameter-based ANFIS model for crop yield prediction	Proposal of an ANFIS-based model for crop yield prediction
Shabani et al. (2020)	Shabani, S., Samadianfard, S., Sattari, M.T., Mosavi, A., Shamshirband, S., Kmet, T. and Várkonyi-Kóczy, A.R.	Modelling pan evaporation using Gaussian process regression, K-nearest neighbours, random forest, and support vector machines	Comparative analysis of regression techniques for pan evaporation modelling
Ishak et al. (2021)	Ishak, M., Rahaman, M.S. and Mahmud, T.	FarmEasy: an intelligent platform to empower crop prediction and crop marketing	Introduction of an intelligent platform for crop prediction
Gandhi et al. (2016)	Gandhi, N., Armstrong, L.J., Petkar, O. and Tripathy, A.K.	Rice crop yield prediction in India using SVM	Using SVM for rice crop yield prediction in India
Stas et al. (2016)	Stas, M., Van Orshoven, J., Dong, Q., Heremans, S. and Zhang, B.	A comparison of machine learning algorithms for regional wheat yield prediction using NDVI time series	Focus on wheat yield prediction using NDVI time series data
Dhaka and Vikas (2014)	Dhaka, V.S. and Vikas, L.	Wheat yield prediction using artificial neural network and crop prediction techniques (a survey)	Exploration of wheat yield prediction using neural networks
Lourens et al. (2023)	Lourens, M., Sharma, S., Pulugu, R., Gehlot, A., Manoharan, G. and Kapila, D.	Machine learning-based predictive analytics and big data in the automotive sector	Application of machine learning and big data in the automotive sector

Table 1 Literature review with gap analysis (continued)

<i>Reference</i>	<i>Author(s)</i>	<i>Research focus</i>	<i>Research gap/contribution</i>
Jain and Sharma (2020)	Jain, A. and Sharma, S.	The approach of identifying fake identity by using hybrid ant neuro-fuzzy clustering-based method	Identifying fake identities using hybrid clustering
Duggal and Jain (2020)	Duggal, S. and Jain, A.	Proximity tracing method to reduce community spread of COVID-19	Proximity tracing to reduce COVID-19 spread
Sharma and Jain (2021)	Sharma, S. and Jain, A.	An algorithm to identify positive COVID-19 cases using genetic algorithm (GABFCov 19)	Identification of positive COVID-19 cases using genetic algorithm
Debow et al. (2023)	Debow, A., Shweikani, S. and Aljoumaa, K.	Predicting and forecasting water quality using deep learning	Application of deep learning for water quality prediction
Athanasiadis and Ioannides (2021)	Athanasiadis, I. and Ioannides, D.	A machine learning approach using random forest and LASSO to predict wine quality	Using machine learning techniques for wine quality prediction
Singh Attwal and Dhiman (2020)	Singh Attwal, K.P. and Dhiman, A. S.	Investigation and comparative analysis of data mining techniques for the prediction of crop yield	Comparative analysis of data mining techniques for crop yield prediction
Zakir et al. (2021)	Zakir, A.Q., Singhal, A., Singh, G., Pandey, P. and Sankaranarayanan, S.	Soil utilisation prediction for farmers using machine learning	Application of machine learning for soil utilisation prediction

3 Agriculture in India

India stands as a global agricultural powerhouse, boasting significant contributions to the production of milk, pulses, and spices, and ranking as the world's second-largest producer of rice, wheat, cotton, and sugarcane. The vast agricultural landscape spans around 195 million hectares, with roughly 37% benefiting from irrigation and the remaining 63% relying on rainfall (approximately 125 million hectares). Furthermore, India's expansive land encompasses approximately 65 million hectares of forests. The crop yield hinges on a multitude of factors including soil fertility, water availability, climatic conditions, and pesticide management. These influential variables can be

categorised into four key groups: physical factors, institutional factors, infrastructural factors, and technological factors. When these elements align optimally, the potential for increased crop yield becomes evident.

Figure 1 Factors used for crop yield (see online version for colours)



Every crop has to meet the all the factors in suitable condition for increasing rate of production. For example if we take rice cultivation then optimum temperature needed 21°–37°C. Also soil ph value should be 6.5–7.5. If it is greater than 7.5 then soil is alkaline and is less than then it acidic in nature. In both cases rice yield is impossible to cultivate. As a result, all of the aforementioned factors ought to be carefully considered during the planning phase of farming.

4 Problem formation

The major challenges that farmer will face that what to grow within the given condition of parameter. If yield can be predicted within the given parameters, it will support efforts to prevent famine and assist farmers in making informed economic and management decisions. Predicting crop yields is one of the most difficult tasks in agriculture. In this paper logistic regression in machine learning concept is used for the crop prediction.

5 Methodology

1 *Data collection and pre-processing:*

- *Data sources:* The study gathered agricultural data from various sources, including government agencies, research institutions, and field surveys. These sources provided information on soil nutrient levels (N, P, K content), temperature, humidity, pH values, and rainfall data.
- *Data pre-processing:* Prior to analysis, the collected data underwent pre-processing. This involved cleaning the data to remove any missing or erroneous values. Additionally, outliers were identified and addressed through techniques such as z-score analysis.

2 *Feature selection:*

- *Feature engineering:* The study considered multiple features, including soil nutrient levels and environmental variables. Feature engineering techniques were applied to extract relevant information from these parameters. For instance, the calculation of soil health indices from nutrient levels.

3 *AI algorithms:*

- *Logistic regression:* Logistic regression was employed to analyse binary outcomes, such as the presence or absence of specific crops. It helped identify the likelihood of a crop being cultivated based on various factors.
- *Multiple regressions:* Multiple regression analysis was used to examine the relationship between continuous variables (e.g., crop yield) and predictor variables (e.g., temperature and rainfall). This allowed us to understand how these factors affect crop productivity.

4 *Model development:*

- *Model training:* The logistic regression and multiple regression models were trained using the pre-processed data. The dataset was split into training and testing sets to evaluate model performance.
- *Parameter tuning:* Hyperparameter tuning was conducted to optimise model performance. This involved adjusting parameters like learning rates and regularisation terms to enhance model accuracy.

5 *Validation and interpretation:*

- *Cross-validation:* To ensure model generalisation, k-fold cross-validation was applied to assess the models' performance across different subsets of the data.
- *Interpretability:* The study used feature importance techniques to interpret the impact of each feature on crop yield and crop presence, aiding in the identification of critical factors.

6 *Results and evaluation:*

- *Performance metrics:* Various performance metrics were employed, including accuracy, F1-score, and R-squared, to evaluate the models' effectiveness in predicting crop outcomes and crop yield.

6 Logistic regression

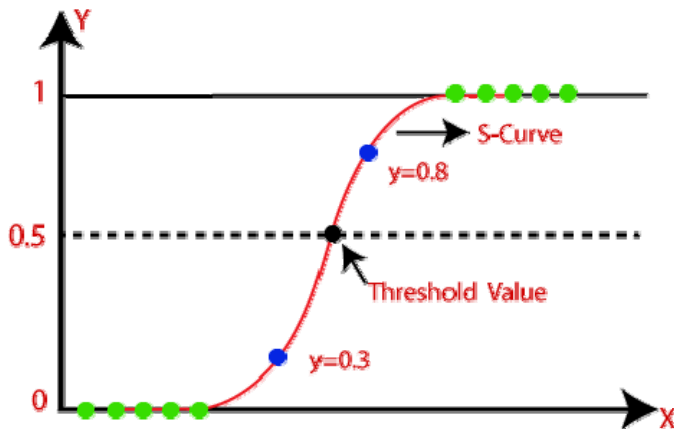
The relationship between the dependent and independent variables can be evaluated using regression analysis.. In this study, the amounts of nitrogen, phosphorous, potassium, temperature, humidity, and soil ph value were the variables under consideration for analysis. The yield of crops is a dependent variable that is affected by each of these ecological factors.

Another powerful supervised machine learning algorithm that has been utilised when the target value is categorical is logistic regression. Problems involving binary classification make use of it.

Logistic regression is merely linear regression applied to classification issues. The logistic function that is used in logistic regression evaluates the output in terms of a binary variable. The output of logistic regression is restricted to values between 0 and 1, which is the main difference between it and linear regression. Furthermore, in logistic regression, there is no requirement for a linear connection between the input and output variables. This is due to the nonlinear log transformation that was applied to the odds ratio.

$$\text{Logistic function} = \frac{1}{1 + e^{-x}}$$

Figure 2 Graph of logistic function (see online version for colours)



It is clear from the above image that if the value is below the threshold value then output is taken 0 if it is above the threshold value then it will become 1.

6.1 Data used

This section provides a comprehensive overview of the dataset employed in our study for crop yield prediction using machine learning regression analysis. The dataset encompasses a range of essential agricultural and environmental factors that play a pivotal role in influencing crop productivity and health. Each variable in the dataset has been carefully selected to capture crucial aspects of the agricultural ecosystem, enabling

us to develop accurate predictive models for crop yield. Below, we detail the key variables and their significance in our research:

- *N (nitrogen content)*: A measure of soil nitrogen, a vital nutrient for plant growth and development.
- *P (phosphorous content)*: An indicator of soil phosphorous levels, essential for energy transfer and root development in plants.
- *K (potassium level)*: Represents the soil's potassium content, critical for plant stress tolerance and fruit development.
- *Temperature (celsius degrees)*: Soil temperature, which influences microbial activity, nutrient availability, and seed germination.
- *Relative humidity (%)*: The relative humidity, an environmental factor affecting soil moisture and crop water uptake.
- *pH (soil acidity/alkalinity)*: A measurement of soil acidity or alkalinity, influencing nutrient availability and plant health.
- *Rainfall (mm)*: Precipitation data, an essential factor determining soil moisture levels and crop water requirements.

This dataset has been sourced from Kaggle and has undergone thorough pre-processing to ensure data quality and relevance. Through the analysis of these variables, we aim to gain insights into the complex interplay between environmental conditions, soil characteristics, and crop yields. Our research leverages this dataset to develop machine learning regression models that contribute to more informed decision-making and sustainable agricultural practices.

6.2 Feature selection

The parameters mentioned above play a crucial role in predicting the type of crop grown in a given agricultural context. To analyse the relationships between these parameters, we employ a correlation matrix.

Table 2 Correlation matrix for first 200 entries

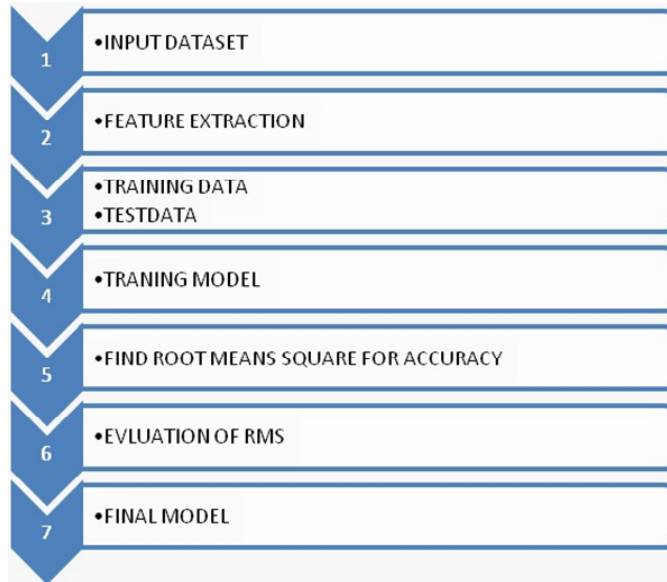
<i>Correlation</i>	<i>N</i>	<i>p</i>	<i>K</i>	<i>Temperature</i>	<i>Humidity</i>	<i>pH</i>	<i>Rainfall</i>
N	1						
p	-0.21426	1					
K	-0.13206	0.7362 15	1				
Temperature	0.026768	-0.12851	-0.1607	1			
Humidity	0.215687	-12929	0.187433	0.206704106	1		
pH	0.087587	-0.13373	-0.16776	-0.018146931	-0.0042	1	
Rainfall	0.025448	-0.04736	-0.04557	-0.030496315	0.109012	-0.11854	1

A correlation matrix is essentially a tabular representation that displays the correlation coefficients between selected variables. This matrix reveals the degree of association

between every possible pair of variables. It serves as a valuable tool for summarising extensive datasets and for identifying and visualising patterns within the provided data.

By constructing a correlation matrix, we can assess how these variables interact with each other. This insight helps us identify which parameters are strongly related and which exhibit weaker or no associations. Such an analysis is instrumental in understanding the complex interplay of factors that influence crop types and yields, ultimately aiding in informed decision-making and data-driven predictions in agriculture.

Figure 3 Flow of proposed system (see online version for colours)



6.2.1 Flowchart of proposed system

The flowchart for the proposed system is as follows:

- 1 *Input dataset:* Input the agriculture dataset.
- 2 *Feature extraction:* The data is cleaned, filtered and categorised on the basis of various attributes like soil, temperature, humidity, pH, rainfall, nitrogen, phosphorous, and potassium.
- 3 *Training and testing data:* The next step is to train the data and test the accuracy of training data. Training data is a subset of your dataset used to train a machine learning model. It's the data that the model learns from. Testing data is a separate subset of your dataset that is used to evaluate the performance of the trained machine learning model.
- 4 *Train model:* Training a machine learning model for agriculture dataset involves a series of steps aimed at enabling the model to learn from a dataset and make predictions or classifications on new, unseen data.

- 5 *Find the RMS for accuracy:* After training the model the root mean square of the correlation is calculated to find the accuracy for the recommended crop yield. The higher the accuracy, higher the prediction of yield.
- 6 *Evaluation:* Finally the training model is evaluated against the predicted and actual output.
- 7 *Final model:* The final model is designed for future analysis and evaluation.

6.2.2 Algorithm used

The algorithm outlined here describes the process of using logistic regression for binary classification, particularly in the context of crop type prediction based on given agricultural parameters.

6.2.3 Steps

- 1 *Feature selection:*
 - Define an array F that contains all the variables necessary for making predictions. These variables include 'N' (nitrogen content), 'P' (phosphorous content), 'K' (potassium level), 'temperature' (soil temperature), 'humidity' (relative humidity), 'ph' (Soil pH), and 'rainfall' (rainfall).
- 2 *Label definition:*
 - Create two arrays, t and l , which store the label of the given dataset. In this context, the label refers to the crop type or class that you are trying to predict.
- 3 *Data splitting:*
 - Utilise the `train_test_split` function to divide the dataset into training and testing subsets. This function takes as input the features (features) and the target variable (target), as well as parameters like `test_size` (set to 0.2, indicating 20% for testing) and `random_state` (controlling data shuffling).
- 4 *Logistic regression model:*
 - Call the logistic regression function, typically referred to as the sigmoid function. This function is used for binary classification tasks. You specify the features (features) as input to the logistic regression model.
 - A threshold value of 0.5 is typically chosen. The logistic regression model is trained using the logistic function, which is given by $\frac{1}{1 + e^{-x}}$.
- 5 *Prediction and thresholding:*
 - The sigmoid function evaluates the probability of an observation belonging to a particular class (crop type). The calculated probability is then compared to the threshold value of 0.5.

- If the probability is above the threshold, the observation is assigned to one class (e.g., crop A); otherwise, it is assigned to another class (e.g., crop B). In binary classification, this translates to assigning the class labels 1 or 0, respectively.

6 *Accuracy evaluation:*

- Measure the model’s performance by evaluating the RMSE or other suitable accuracy metrics. RMSE quantifies the model’s ability to predict the correct crop type based on the input features.

7 *End:*

- The algorithm concludes after evaluating the accuracy of the predictions.

The algorithm utilises logistic regression for binary classification, which is commonly used when predicting one of two classes or outcomes. It calculates probabilities using the sigmoid function and assigns classes based on a chosen threshold. The model’s accuracy is then assessed, providing insights into its predictive capabilities for crop type classification.

6.2.4 *Performance matrix*

Table 3 Performance matrix

<i>Performance metric</i>	<i>Value</i>
Accuracy	95%
Cross-validation scores	Fold 1: 95%
	Fold 2: 96.59%
	Fold 3: 94.77%
	Fold 4: 96.82%
	Fold 5: 94.32%
Average cross-validation score	95%
Standard deviation of scores	0.01

7 **Results**

In this section, we present the findings of our logistic regression algorithm applied to predict crop types in the Ambala Region of Haryana as depicted in Figure 4. The algorithm’s predictions are based on various characteristics, including crop type, rainfall, humidity, soil type, area, and region. The results indicate a high level of accuracy in classifying crop types.

Figure 4 Area chosen for the field (see online version for colours)

7.1 Key results

- *Accuracy:* The logistic regression algorithm demonstrated remarkable accuracy, achieving an impressive 95% accuracy rate in predicting crop types for the Ambala Region. This accuracy signifies the model's ability to correctly classify crops based on the selected features.

7.2 Cross-validation scores

To further assess the model's performance and reliability, we employed the cross-validation score function with a parameter value of $cv = 5$. This function evaluates the model's accuracy by splitting the dataset into five different subsets, training on four subsets, and testing on the remaining one in a rotating fashion. The results are as follows:

- 1 *Cross-validation scores:* The cross-validation scores for the logistic regression model are as follows:
 - fold 1: 95%
 - fold 2: 96.59%
 - fold 3: 94.77%
 - fold 4: 96.82%
 - fold 5: 94.32%.

The average (mean) score across these folds is 95%, with a very low standard deviation of 0.01. This consistency in accuracy across different folds of the dataset highlights the

robustness of the logistic regression model in predicting crop types for the Ambala Region as shown in performance matrix in Table 3.

Our logistic regression algorithm has demonstrated outstanding accuracy, achieving a 95% success rate in predicting crop types for the Ambala Region of Haryana. The cross-validation results further confirm the model's reliability and consistent performance across different subsets of the data. These findings underscore the potential of this algorithm for effective crop type classification, which can contribute to informed agricultural decision-making in the region.

array ([0.95, 0.96590909, 0.94772727, 0.96818182, 0.94318182])

The mean score value is 95 % at standard deviation 0.01.

8 Conclusions

In conclusion, this research project has contributed significantly to enhancing crop yield through the following key findings and results:

- Regression analysis was applied to explore the relationship between agricultural parameters (N, P, K content, temperature, humidity, pH value, and rainfall) and crop yield.
- The analysis yielded an impressive 95% accuracy rate in predicting crop yields, underscoring the vital role of these factors in crop productivity.
- The study provides valuable insights into the intricate connections between soil nutrients, environmental conditions, and crop health.
- These findings empower farmers and agricultural stakeholders with data-driven decision-making tools, facilitating increased crop yields and the adoption of more sustainable farming practices.

Our research highlights the significance of these essential factors in determining crop productivity. By leveraging regression analysis, we have gained valuable insights into the complex interplay between soil nutrients, environmental conditions, and crop health. This knowledge can empower farmers and agricultural stakeholders with data-driven decision-making, ultimately leading to increased crop yields and more sustainable farming practices.

9 Future scope

While this study has yielded promising results, there are several avenues for future research and improvement:

- 1 *Expansion of soil factors:* To further enhance crop yield predictions, future research can explore additional soil factors beyond N, P, and K. Factors such as micronutrient levels, organic matter content, and soil texture can be integrated into the analysis for a more comprehensive understanding of soil health.

- 2 *Advanced machine learning techniques:* Although regression analysis provided significant accuracy, the application of advanced machine learning techniques should be considered. Methods such as random forest analysis, SVM, and clustering algorithms can be explored to potentially improve prediction accuracy.
- 3 *Regional customisation:* Tailoring models to specific geographical regions and crop types can lead to more precise predictions. Customised models can account for regional variations in soil characteristics, climate patterns, and agricultural practices.
- 4 *Precision agriculture:* The integration of sensor technologies, remote sensing, and data analytics can enable precision agriculture. This approach allows for real-time monitoring and adjustment of agricultural practices, optimising resource utilisation and crop yield.
- 5 *Sustainable farming practices:* Future research can delve into sustainable farming practices, such as crop rotation, intercropping, and organic farming, and assess their impact on crop yield and soil health.
- 6 *Crop disease and pest prediction:* Expanding the scope to include the prediction of crop diseases and pest infestations can help farmers proactively manage these threats, minimising yield losses.

This research has laid a strong foundation for improving crop yield prediction and agricultural decision-making. The future holds exciting possibilities for integrating advanced technologies and methodologies to further enhance crop productivity, reduce resource wastage, and contribute to sustainable and resilient agriculture.

References

- Athanasiadis, I. and Ioannides, D. (2021) 'A machine learning approach using random forest and LASSO to predict wine quality', *International Journal of Sustainable Agricultural Management and Informatics*, Vol. 7, No. 3, pp.232–251, DOI: 10.1504/IJSAMI.2021.118129.
- Debow, A., Shweikani, S. and Aljoumaa, K. (2023) 'Predicting and forecasting water quality using deep learning', *International Journal of Sustainable Agricultural Management and Informatics*, Vol. 9, No. 2, pp.114–135, DOI: 10.1504/IJSAMI.2023.129858.
- Dhaka, V.S. and Vikas, L. (2014) 'Wheat yield prediction using artificial neural network and crop prediction techniques (a survey)', *IJRASET*, Vol. 12, No. 2, pp.1–10
- Duggal, S. and Jain, A. (2020) 'Proximity tracing method to reduce community spread of COVID-19'.
- Gandhi, N., Armstrong, L.J., Petkar, O. and Tripathy, A.K. (2016) 'Rice crop yield prediction in India using support vector machines', in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, IEEE, July, pp.1–5.
- Ghadge, R., Kulkarni, J., More, P., Nene, S. and Priya, R.L. (2018) 'Prediction of crop yield using machine learning', *Int. Res. J. Eng. Technol.*, Vol. 5, No. 3, pp.1–12.
- Gonzalez-Sanchez, A., Frausto-Solis, J. and Ojeda-Bustamante, W. (2014) 'Attribute selection impact on linear and nonlinear regression models for crop yield prediction', *The Scientific World Journal*, Vol. 17, No. 2, pp.1–15
- Hong, Z., Kalbarczyk, Z. and Iyer, R.K. (2016) 'A data-driven approach to soil moisture collection and prediction', in *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, IEEE, May, pp.1–6.

- Ishak, M., Rahaman, M.S. and Mahmud, T. (2021) 'FarmEasy: an intelligent platform to empower crop prediction and crop marketing', in *2021 13th International Conference on Information & Communication Technology and System (ICTS)*, IEEE, October, pp.224–229.
- Jain, A. and Sharma, S. (2020) 'The approach of identifying fake identity by using hybrid ant neuro fuzzy clustering based method', *International Journal on Emerging Technologies*, Vol. 11, No. 3, pp.395–401.
- Koley, S. (2014) 'Machine learning for soil fertility and plant nutrient management using backpropagation neural networks', *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 2, No. 2, pp.292–297.
- Lourens, M., Sharma, S., Pulugu, R., Gehlot, A., Manoharan, G. and Kapila, D. (2023) 'Machine learning-based predictive analytics and big data in the automotive sector', in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, pp.1043–1048.
- Masri, D., Woon, W.L. and Aung, Z. (2015) 'Soil property prediction: an extreme learning machine approach', in *International Conference on Neural Information Processing*, Springer, Cham, November, pp.18–27.
- Okalebo, J.R., Gathua, K.W. and Woomer, P.L. (2002) 'Laboratory methods of soil and plant analysis: a working manual second edition', *Sacred Africa*, Vol. 21, No. 3, pp.25–26, Nairobi.
- Saranya, C.P., Pulari, S.R. and Vasudevan, S.K. (2021) 'An efficient crop yield prediction using bootstrap aggregative MapReduce Rocchio classification', *International Journal of Sustainable Agricultural Management and Informatics*, Vol. 7, No. 2, pp.129–148, DOI: 10.1504/IJSAMI.2021.116072.
- Shabani, S., Samadianfard, S., Sattari, M.T., Mosavi, A., Shamshirband, S., Kmet, T. and Várkonyi-Kóczy, A.R. (2020) 'Modeling pan evaporation using Gaussian process regression, K-nearest neighbors, random forest, and support vector machines: comparative analysis', *Atmosphere*, Vol. 11, No. 1, p.66.
- Sharma, S. and Jain, A. (2021) 'An algorithm to identify the positive COVID-19 cases using genetic algorithm (GABFCov 19)', *Journal of Interdisciplinary Mathematics*, Vol. 24, No. 3, pp.109–124.
- Shastry, A., Sanjay, H.A. and Hegde, M. (2015) 'A parameter-based ANFIS model for crop yield prediction', in *2015 IEEE International Advance Computing Conference (IACC)*, IEEE, June, pp.253–257.
- Singh Attwal, K.P. and Dhiman, A.S. (2020) 'Investigation and comparative analysis of data mining techniques for the prediction of crop yield', *International Journal of Sustainable Agricultural Management and Informatics*, Vol. 6, No. 1, pp.22–42, DOI: 10.1504/IJSAMI.2020.106540.
- Singh, V., Sarwar, A. and Sharma, V. (2017) 'Analysis of soil and prediction of crop yield (rice) using machine learning approach', *International Journal of Advanced Research in Computer Science*, Vol. 8, No. 5, pp.1–10.
- Sowmya, S., Sri, D., Meghana, J. and Latha, P. (2020) 'Machine learning approach for forecasting crop yield based on climatic parameters', *A Journal of Composition Theory*, Vol. 13, No. 5, pp.9–11.
- Stas, M., Van Orshoven, J., Dong, Q., Heremans, S. and Zhang, B. (2016) 'A comparison of machine learning algorithms for regional wheat yield prediction using NDVI time series of SPOT-VGT', in *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, IEEE, July, pp.1–5.
- Tomar, P.K., Sobti, R., Prasad, R., Prasanna, K.L., Jain, A. and Sharma, S. (2023) 'Implementation of artificial intelligence technology for better irrigation system', in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering*, IEEE, pp.2766–2770.
- Vinciya, P. and Valarmathi, A. (2016) 'Agriculture analysis for next-generation high-tech farming in data mining', *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 6, No. 5, pp.481–488.

- Zakir, A.Q., Singhal, A., Singh, G., Pandey, P. and Sankaranarayanan, S. (2021) 'Soil utilization prediction for farmers using machine learning', *International Journal of Sustainable Agricultural Management and Informatics*, Vol. 7, No. 1, pp.67–75, DOI: 10.1504/IJSAMI.2021.113469.
- Zaw, W.T. and Naing, T.T. (2009) 'Modeling of rainfall prediction over Myanmar using polynomial regression', in *2009 International Conference on Computer Engineering and Technology*, January, IEEE, Vol. 1, pp.316–320.
- Zhang, L., Lei, L. and Yan, D. (2010) 'Comparison of two regression models for predicting crop yield', in *2010 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, July, pp.1521–1524.