Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 2, pp. 213–223
DOI : 10.47974/JSMS-1248

Enhancing agriculture productivity with machine learning in crop yield optimization

Dinesh Bhoyar *
Department of Electronics & Telecommunication
Yashwantrao Chavan College of Engineering
Nagpur
Maharashtra
India

Sachin Jain †
Department of Computer Science
Oklahoma State University
Stillwater
United States

Jagdish D. Kene §
Department of Electronics and Communication Engineering
Shri Ramdeobaba College of Engineering and Management
Nagpur
Maharashtra
India

Y. A. Suryawanshi ‡
Department of Electrical Engineering
Yeshwantrao Chavan College of Engineering
Nagpur
Maharashtra
India

* *E-mail:* **dinesh.bhoyar23@gmail.com (Corresponding Author)**
† *E-mail:* **sachin.jain@okstate.edu**
§ *E-mail:* **kenejd@rknec.edu**
‡ *E-mail:* **yogesh_surya08@rediffmail.com**

A. R. Bhagat Patil @
Department of Computer Technology
Yeshwantrao Chavan College of Engineering
Nagpur
Maharashtra
India

Shivkumar J. Karale #
Department of Computer Technology
Yeshwantrao Chavan College of Engineering
Nagpur
Maharashtra
India

Abstract

The majority of the GDP of an agriculture-based economy comes from farming. This initiative was inspired by the rising suicide rates among farmers, which may be related to poor crop yields. The field of agriculture is now seriously threatened by changes in the climate and other environmental factors. For this problem to be solved effectively and practically, machine learning is a crucial strategy. Estimating agricultural output based on historical information such as Ph, humidity temperature, rainfall, N, P, K. We used Machine Learning method to achieve this. We constructed and compared a number of various machine learning algorithms and ultimately settled on the Random Forest Algorithm, which provided an accuracy of 97.87%.

Subject Classification: *68M11.*

Keywords: *Crop prediction, Maximum yield, Machine learning, Agriculture, Random forest, Logistic regression, Decision tree classifier, GUI.*

1. Introduction

The main emphasis was given to the agriculture field considering the importance of production in India, although it is very important in every corner of the world, In terms of farm output, India is rated second globally [1]. About 41.49% of all workers were employed in the agricultural and related industries in 2020, which contributed 18.4% of the GDP [2]. The financial contribution of the agriculture sector to India's GDP is steadily shrinking.

@ *E-mail:* **`arbhagatpatil@gmail.com`**
# *E-mail:* **`sjkarale@gmail.com`**

A total of 10,000,881 people of maharashtra dependent on farming sector consisting of farmers and agricultural labourers. The farmers committed suicide during the year 2021, around to 6.6% of total suicides victims in the country [3].

Using technology to raise awareness about farming has been a need in recent years. Food insecurity is brought on by seasonal climate changes that are harmful to basic resources including soil, water, and air. In order to solve this issue, we suggest a system that would allow farmers to choose crops based on economic and environmental considerations in order to maximize their output, which will subsequently assist them fulfill the nation's growing demand for food supplies. The suggested system makes use of machine learning to make informed estimations. To ensure that farmers receive the most possible yield from their crops, the system will offer crop recommendation and crop selection depending on weather conditions that are optimal for the crop. The system analyses variables like rainfall, temperature, N, P, K, and humidity to predict suitable crops.

Predicting the right crop is a significant agricultural issue. Artificial intelligence's branch of machine learning describes a machine's capacity to imitate intelligent human behavior.

2. Literature Survey

In [4], the authors proposed a mobile application system for farmers in Maharashtra and Karnataka, using GPS to identify their location and inputting their area and soil type. Machine Learning algorithms like SVM, ANN, Random Forest (RF), were used for forecasting crop yield, with RF having the highest accuracy of 95%. The system also suggests the best fertilizer timing for optimal yield.

The system proposed by Kamath et al. [6] uses Random Forest, to predict crop yield based on historical data, incorporating soil agronomic parameters and seasonal variations. Train and test data is in the ratio 80% and 20%, respectively. The system is accessible via a website with four functional modules: Crop, Soil, Weather, and Predict. The Predict module allows users to input district region, crop type, soil type, and cultivating area to generate estimated crop yield.

This framework put forward by M.S. Minu et al. [7] focuses on climate estimating, crop yield expectation and harvest cost determination. Grouping strategies like RF,SVM, Logistic Reasoning(LR) and ANN are utilized. The information of anticipated yield is available for the cultivators through a web application. This guides them to choose the yield they might want to plant for the expected year. The web application gives a discussion to the ranchers to stock the merchandise without mediators thereby empowering them to acquire the most cost for their items.

In their study, the authors of. [8] proposed a system that utilizes calculations such as Multiple Linear Regression to identify the pattern among the data. Once identified, it is processed according to the given conditions to suggest the best feasible crops based on the prevailing environmental conditions. This system only requires the user's location and recommends various beneficial crops, providing farmers with a decision on which crop to grow.

The objective of [9] was to develop a model for crop yields, which can help farmers make informed decisions about which crops to sow based on the climate and soil nutrients. They found that (RF) algorithm achieved the highest accuracy of 99.32% when using both Entropy and Gini Criterion. In contrast, the (KNN) algorithm had the lowest accuracy of 97.04%, and Decision Tree Classifier had accuracy levels between KNN and RF. The study recommends further data collection from fields to improve the classification of crop types using other algorithms such as (ANN) or (CNN).

3. Methodology

The proposed system uses machine learning to anticipate the crop, and Python programming language is used. To gather experience and create a trained model by using the data, machine learning leverages past

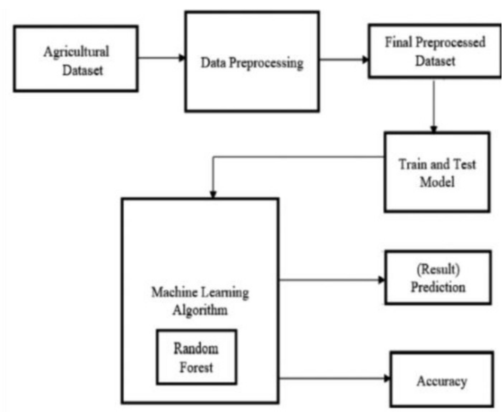

Figure 1

Block diagram of model

data and knowledge. The output of this model is then predicted. The accuracy of the model will increase with the quality of the dataset collected. Regression and classification techniques used in machine learning have been found to outperform a number of statistical models. The implementation of proposed system shown in Figure 1. The various steps involved are explained below.

1. Temperature 2. Rainfall 3. Humidity 4. Ph 5. Nitrogen 6. Phosphorous 7. Potassium levels in the soil.

Geographical elements including soil chemistry, rainfall, terrain, soil type, temperature, and others have a huge impact on how much food can be produced. These elements significantly increase crop output.

Steps involved in the implementation:

1. Data collection: Our datasets were obtained from the following sources: The rainfall, temperature, and humidity data was collected, where value of N,P, K, and pH was referenced from [18].This data was found in [15] with the data credits given to the Indian government. These variables' data were supplied as input. Figure 2. shows a sample of input dataset

N	P	K	temperature	humidity	ph	rainfall	label
90	42	43	20.87974371	82.00274423	6.502985292	203.9832	rice
85	58	41	22.00094	80.31964408	7.038096361	226.6555374	rice
60	55	44	23.00445915	82.3207629	7.840207144	263.9642476	rice
74	35	40	26.49109635	80.15836264	6.980400905	242.8640342	rice
78	42	42	20.13017482	81.60487287	7.628472891	262.7173405	rice
69	37	42	23.05804872	83.37011772	7.073453503	251.0549998	rice
69	55	38	22.70883798	82.63941394	5.70080568	271.3248604	rice
94	53	40	20.27774362	82.89408619	5.718627178	241.9741949	rice
89	54	38	24.51588066	83.5352163	6.685346424	230.4462359	rice

Figure 2

Snapshot of dataset

2. Preprocessing: Processing raw data to make it suitable for machine learning is the goal of preprocessing. The method consists of data formatting, cleansing, and sampling.
3. Data: Finding out how each parameter affects the target parameter can also be done with the use of pre-processing shown in Figure 3.

```
▾ Data Cleaning

[12] df.P = df.P.fillna(value = 0)
     df.K = df.K.fillna(value = 0)

[13] df.temperature = df.temperature.fillna(value = df.temperature.mean())
     df.humidity = df.humidity.fillna(value = df.humidity.mean())
     df.ph = df.ph.fillna(value = df.ph.mean())
     df.rainfall = df.rainfall.fillna(value = df.rainfall.mean())

[14] df.label = df.label.fillna(value = df.label.mode()[0])

df.isna().sum()

N              0
P              0
K              0
temperature    0
humidity       0
ph             0
rainfall       0
label          0
dtype: int64
```

Figure 3

Data Cleaning

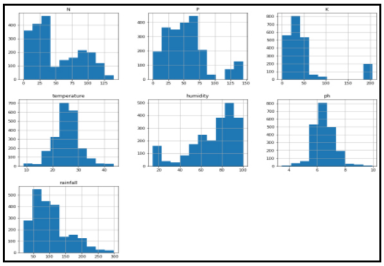

Figure 4

Histogram of each feature with a default bin:10

A new feature called water usage was used which was calculated based on the amount of rainfall as low, medium, and high. The threshold values are 150,250 where values less than 150mm of rainfall is considered as low water usage and above 250 mm as high-water usage. The following is shown in Figure 4,5.

	N	P	K	temperature	humidity	ph	rainfall	label	Water Usage
0	90	42.0	43.0	20.879744	82.002744	6.502985	203.983200	rice	2
1	85	58.0	41.0	22.000940	80.319644	7.038096	226.655537	rice	2
2	60	55.0	44.0	23.004459	82.320763	7.840207	263.964248	rice	0
3	74	35.0	40.0	26.491096	80.158363	6.980401	242.864034	rice	2
4	78	42.0	42.0	20.130175	81.604873	7.628473	262.717340	rice	0

Figure 5

water usage in dataset

4. Splitting a dataset: A dataset that will be used for machine learning should be divided into training and testing categories. The method consists of data formatting, cleansing, and sampling.
5. Model training: After separating the gathered data into train and test sets using preprocessing, we moved on to model training. The algorithm was fed with training data during this process.

4. Comparison of algorithms

Table 1 shows a comparison of the three models. It can be inferred that the advantages of random forest over the rest is high.

5. Testing and model assessment: This stage aims to create the most straightforward model capable of accurately and quickly formulating a target value. Model tuning is one way a data scientist can accomplish this. For an algorithm to work at its best, model parameters must be optimized. It is done to make sure that the model is not over fitting or creating too high of a variance 6. Predict yield: The outcome on fresh input is predicted using the trained model. Here, we've stored the learned model in a file so it can forecast fresh input.

In this approach, the trained Machine Learning model created in Jupiter is used and the model is tested using a dataset. This prediction model includes the random forest method, which develops attributes from training data and uses that data to produce predictions. The datasets are scrutinized to guarantee correctness. Our crop prediction and optimal yield system provides improved production planning and decision-making.

The Graphical User Interface for the proposed model using pickle and flask web frame. Html, CSS was used to display and create styles as shown in Figure 6.

Table 1

Comparison of various models of Machine Learning

Algo	**Description**	**Benefits**
RANDOM FOREST	It is fabricating a multitude of decision tress at training time.	Over-fitting is less of an issue. Pruning is not needed.
K-Nearest Neighbour (KNN)	Non-parametric path for regression and classification. The input consists of, k closest training models in the feature space.	Tests production output of different crops in previous years.
Artificial Neural Networks (ANN)	Non-parametric path for regression and classification. The input years. consisting of, k closest training models in the feature space.	It is an intelligent system that can deceive complex data.
DECISION TREE	All possible outcomes of a decision are illustrated by a branching method by a graph. Accredits any possible outcomes, so that decision can be automated.	To classify an obscure sample, its characteristics values are tested against the decision tree.

Steps:

1. Take 5 inputs from the user: N, P,K, temperature, humidity, Ph, rainfall. The expected (valid) range of these input values, along with their measurement units is mentioned on the webpage, to make sure there is no miscommunication with the user.
2. These values are given as input to the machine learning implementation in the backend
3. The algorithm returns the most suitable crop to the end user on the server.

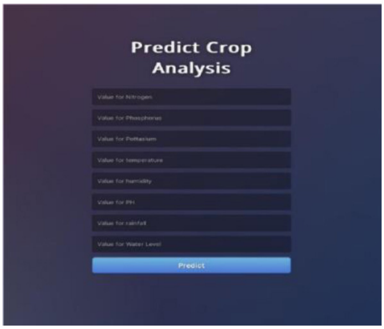

Figure 6

GUI of crop prediction system

5. Experimental result and analysis

Accuracy: For accuracy measurement the total number of forecasts divided by the number of right predictions, accuracy is determined as shown in Table 2.

Table 2

Accuracy of proposed models

Model	Accuracy
Logistic Regression	93.63
Decision tree	95.45
Random Forest	97.87

It should be noted that Random forest gains the maximum accuracy and outperforms all other models whereas RF algorithm attained high accuracy by taking into account several algorithms when forecasting the

Fig. 7a

Confusion matrix of logistic regression

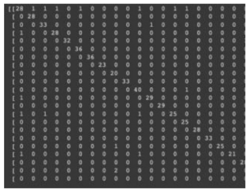

Fig. 7b

Confusion matrix of decision tree classifier

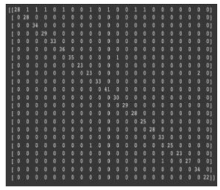

Fig. 7c

Confusion matrix of random forest algorithm

yield. This is due to the fact that it builds the decision tree for each set of training data, merge the various trees into a single one, and estimates the yield by taking into account the average value of the tree.

A confusion matrix is used for the performance evaluation of a classification model by comparing predicted and actual values. It consists of four quadrants: True Positive, False Positive, False Negative, and True Negative (TN).The matrix can be used to calculate various evaluation metrics. We infer that the matrix of RF algorithm has least number of false values demonstrating a high decisive model. Below graphics shows in Figure 7confusion matrices of all the three models

A classification report as shown in Figure 8 is a type of evaluation report generated after training an ML algorithm for a classification task. It provides a summary of the performance of the algorithm on the test set, including the varied evaluation metrics, and support for each class label. The report also includes the overall model accuracy and the macro and micro averages of the metrics. It is a useful tool for performance comparison of varied algorithms and selecting the best one for a given task. Graphic 14 shows the classification report of logistic regression.

2. Future Scope

More features and precise datasets improve the performance of the system. Future scope of this proposed model will concentrate on periodically updating the datasets to provide reliable forecasts, and the procedures can be automated. Along with predicting the best crop, an approximate yield of the crop can be predicted based on the area available. This will supplement the farmers' knowledge of the variation of production of each crop and thus make a well- informed decision.

	precision	recall	f1-score	support
apple	0.97	0.78	0.86	36
banana	0.97	1.00	0.98	28
blackgram	0.93	0.74	0.82	34
chickpea	0.97	0.97	0.97	29
coconut	1.00	0.94	0.97	33
coffee	0.97	0.97	0.97	36
cotton	0.90	0.97	0.93	36
grapes	1.00	1.00	1.00	23
jute	0.79	0.88	0.83	25
kidneybeans	0.97	1.00	0.99	33
lentil	0.90	0.93	0.92	41
maize	0.96	0.87	0.91	30
mango	1.00	0.97	0.98	29
mothbeans	0.76	1.00	0.86	28
mungbean	0.96	1.00	0.98	25
muskmelon	1.00	0.96	0.98	28
orange	0.97	1.00	0.99	33
papaya	1.00	0.96	0.98	26
pigeonpeas	1.00	0.96	0.98	23
pomegranate	0.90	0.96	0.93	28
rice	0.91	0.91	0.91	34
watermelon	0.96	1.00	0.98	22
accuracy			0.94	660
macro avg	0.94	0.94	0.94	660
weighted avg	0.94	0.94	0.94	660

Figure 8

Classification report of logistic regression

An application can be developed that enables the user to log in and get the predicted output. This will enable the user to keep a record of the predictions. The application can be trained to provide statistical information regarding each set of feature values.

References

[1] "Contribution of Agriculture Sector towards GDP", 3 August (2021). [En línea]. Available:https://www.pib.gov.in/PressReleasePage.aspx?PRID=1741942.

[2] "Employment in agriculture (% of total employment) (modeled ILO estimate) – India," [En línea]. Available: https://data.worldbank.org/indicator/SL.AGR.EMPL.ZS ?locations=IN.

[3] "FAO in India," [En línea]. Available: https://www.fao.org/india/fao-in-india/india-at-a-glance/en/.

[4] P. S. Anwesha Borthakur, "Agricultural Research in India: An Exploratory Study," *International Journal of Social Science & Interdisciplinary Research* (2012).

[5] S. Raja, B. Sawicka, Z. Stamenkovic y G. Mariammal, "Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers," IEEEE Access, vol. 10, pp. 23625-23641, January (2022).

[6] A. Jain, "Analysis of Growth and Instability in Area, Production, Yield and Price of Rice in India," Social Change and Development, vol. 15, no 2, pp. 46-66 (2018).

[7] U,. M. P.Priya, "Predicting yield of the crop using machine learning algorithm," (2018).

[8] P. P., S. S., S., S. S. Pallavi Kamath, "Crop yield forecasting using data mining," KeAi Chinese Roots Global Impact, (2021).

[9] G. M. B., K. M., S. M., S. S. K. S. Saranya C P, "A SURVEY ON CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHMS," IJRAR, vol. 7, no 1, pp. 494-498, March (2020).

[10] A. K. ,. C. C. Thomas van Klompenburga, "Crop yield prediction using machine learning: A systematic literature review," *ScienceDirect Computers and Electronics in Agriculture*, vol. 177, (2020).

[11] K. Priyadharsini, J. R. Dinesh Kumar, N. Udaya S Susmaa Rao and S. Yogarajalakshmi, "AI- ML Based Approach in Plough to Enhance the Productivity," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, pp. 1237-1243 (2021), doi: 10.1109/ICICV50876.2021.9388634.

[12] A. Ingle, "Crop Recommendation Dataset," [Online]. Available: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset?select=Crop_recommendation.csv.