A Comparative Study Of Different Algorithms Used To Predict The Crop, Its Yield And Price

Pratiksha Pawar
Pimpri Chinchwad College of Engineering, Pune
pratiksha1132@gmail.com

Vishwajeet Shinde
Pimpri Chinchwad College of Engineering, Pune
vishwajeetshinde2001@gmail.com

Aniket Raut
Pimpri Chinchwad College of Engineering, Pune
aniketraut.rr@gmail.com

Saloni Suke
Pimpri Chinchwad College of Engineering, Pune
salonisuke7@gmail.com

Under Guidance of- Prof. Sagar Salunke
Department of Computer Engineering
PCET's
Pimpri Chinchwad College of Engineering, Pune

Abstract:

India is considered an agricultural land and many people have agriculture as their occupation. So India is in dire need of having Crop and its yield as well as its price prediction. Based on soil pH, Rainfall, humidity, temperature, and various factors we can predict the result. Our system will suggest the best suitable crop for a particular land-based on content and weather parameters. So there are various algorithms and techniques that can be taken into consideration like Decision tree Regressor, Random Forest, Particle Swarm Optimization (PSO)-Back Propagation (BP) Neural Network Model, K- Nearest Neighbor (KNN). After comparing the algorithms our aim is to find the best suitable algorithms for prediction which will lead us to find a proper crop according to a given set of factors.

Keywords:

Machine Learning, Decision tree Regressor, Random Forest, Particle Swarm optimization, Back propagation, PSO, BP, Neural Network, Crop Recommendation, Crop Price Prediction, Yield Prediction, KNN.

Introduction:

Agriculture is the backbone of India. As we know, food stands first in the basic need of survival, and the agriculture sector needs to be given the highest preference in development. In recent times, it has become inevitable to use technology to create awareness about cultivation. The seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. The Indian agriculture sector accounts for 18 percent of Indian agriculture gross domestic product (GDP) and provides employment to 50% of the country's workforce.

Machine learning (ML) approaches are used in many fields, ranging from supermarkets to evaluate the behaviour of customers to the prediction of customer's phone use. Machine learning has also been used in agriculture for several years. Crop yield prediction is one of the challenging problems in precision agriculture, and many models have been proposed and validated so far. This problem requires the use of several datasets since crop yield depends on many different factors such as climate, weather, soil, use of fertilizer, and seed variety.

Crop Prediction is a need in a country like India where the farmers consider the traditional methods while choosing the crop. In India many of the farmers don't even know about the soil

factors like pH level of soil, phosphorus and nitrogen level of soil, fertility .And with that atmospheric factors like temperature, humidity, climate, rainfall etc. They just go with their previous one, instead of it we will suggest they test their soil and then go with the crop which is suitable for it. And by considering these soil and atmospheric factors, here we built one machine learning model which will recommend crops, predict crop yield and after that predict price of yield. We have involved four major algorithms to get the system with less error and with more accuracy. We are going to use the below algorithms:

1) Decision Tree Regressor

The Decision Tree is a predictive model that works by testing conditions at each tree level, and moving down the tree where different decisions are identified. The situation depends on the application and the decision-making outcome. Decision tree algorithms include the C4.5, CART, and ID3.

2) Random Forest

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. It can be used for both classification and regression tasks.

3) PSO-BP neural network

PSO-BP neural network is a model made up of a combination of BP neural network and PSO algorithm. The PSO algorithm helps in optimization and BP neural networks in prediction. The weights of the BP neural network are mapped to particles of the PSO algorithm. The particles which have the best weights and threshold are assigned to the BP neural network for further processing.

4) K- Nearest Neighbour (KNN)

The KNN is a nonparametric, supervised learning methodology which uses training sets to classify data points into specific categories. Information is collected from all educational cases, as well as the correlations in basic classifications, based on a new case. The training dataset is checked for the greatest number of previous cases (neighbour) k, and new instances (x) estimated by summing up the output attributes for the k cases.

Literature Survey:

The proposed system [1] recommends the best suitable crop for particular land by considering parameters such as annual rainfall, temperature, humidity and soil pH. Among these parameters annual rainfall is predicted by the system itself by using previous year data with the SVM algorithm and other parameters have to be entered by the user. In the output section the system displays a suitable crop, required seeds/acre, market price and approximate yield of the recommended crop and also the system takes NPK values in the input section to display the required NPK for the recommended crop.

The proposed system [2] recommended the suitable yield for a particular crop by contemplating factors like water, UV, fertilizer by using ANN along with K-fold validation. Major challenge in agriculture is to increase the production in the farm and deliver it to the end customers with best possible price and good quality. [3]The paper says, vast research has been done and several attempts are made for application of Machine learning in agricultural fields. Proposed model i.e. Decision tree got more efficiency in finding better yield for crops.

In this paper crop price forecasting service was made for business purposes .The algorithms used in the following paper are autoregressive integrated moving average

(ARIMA), the partial least square (PLS), artificial neural network (ANN) .Past prices are considered as a training dataset. Further, a new addition was made to PLS which was combined with response surface methodology (RSM) to investigate non-linear relationships between past prices. Error in percentage form is calculated for each algorithm which is further used for comparison purposes [4].

Back propagation neural network is one of the most widely used neural network. Further to increase the accuracy Particle swarm optimization algorithm is used for optimization purposes. To calculate accuracy, the sum of absolute difference (SAD) and mean square error (MSE) are used. Comparison between traditional BP neural network and PSO-BP neural network as per the factors [5].

The paper [6] discusses creating a recommendation system using the concept of precision agriculture. Data mining is a very important part for providing the insights about organic and non-organic factors. The algorithms described in this paper are Random tree, CHAID, K-Nearest Neighbour and Naive Bayes. Various important soil factors are taken into consideration like pH, colour, texture etc. and crops taken are banana, cotton, sorghum etc. from Madurai district.

Various standard machine learning algorithms are used to predict or recommend crops. Among the selected algorithms, the KNN algorithm gives 92% of accuracy [7].

K-Nearest Neighbour [8] can be used for both classification and regression. K-Nearest Neighbours is a non-complex algorithm which stores all the available cases and classifies new cases based on some similarity measure.

Algorithmic Study:

This section describes the algorithm based working of the model for all the different techniques mentioned below:

3.1 Decision Tree Regressor using feature selection

3.1.1 Algorithm for Decision Tree Regressor:

1) Start.
2) Download the dataset from the Kaggle for crop prediction in CSV format.
2) Read the crop prediction data from the .csv file using pandas.read_csv in data.
3) The data now contains key values which are the temperature, humidity, pH, rainfall, and label.
4) Changing the categorical data to values like rice=0, wheat= 1, etc.
5) Dividing the data for training and testing purposes in which 80% is assigned for training purposes and 20% for testing.
6) Training data to be fed to the Decision Tree Regressor.
7) Performing different optimization techniques on the Decision Tree Regressor.
8) Checking the accuracy of the model on the basis of the test data and calculating the loss function.
9) Stop.

3.1.2 Features:

1. Less Efforts

Compared to other algorithms, decision trees require less effort for data preparation during pre-processing.

2. No need for normalization and scaling

A decision tree does not require normalization of data and scaling of data as well.

3. No effect of missing data

Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.

4. Intuitive Model

A Decision tree model is very intuitive and easy to explain to technical teams as stakeholders.

3.1.3 Challenges

1. Splitting of data

A decision tree will be unable to make accurate predictions if the data that are provided to split out a prediction are unrelated to the data that it has been trained on.

2. Over fitting

Decision Trees are prone to over fitting. That's why they are rarely used and instead other tree-based models are preferred like Random Forest and XGBoost .

3.2 Random Forest

3.2.1 Algorithm for Random Forest

1) Start
2) Read the data from CSV file using pandas in data
3) Data contains the key values like state name, district name, crop year, season, crop, Area, production.
4) Pre-processing on the state name, season, and the crop is needed to be done which is converting the categorical data into numerical data.
5) Dividing the data into two parts training and testing in such a way that 90% of the data is taken by training data and the rest 10% is taken for testing data.
6) Feeding the training data to the input layer of the random forest.
7) Calculating the loss function using mean squared error based on the output of the random forest output layer.
8) The testing of the neural network is to be done after it is trained on the training data set.
9) Determining the accuracy of the model.
10) Storing the neural model in h5 format for further use.
11) Stop

3.2.2 Features

1. Less computationally expensive

Random forest is less computationally expensive compared to other algorithms.

2. No need for monotonic transformation and scaling

The random forest doesn't need to do any monotonic transformation and scaling of data as well.

3. Good accuracy and easy to use

The random forest means decrease impurity and mean decrease accuracy for feature selection.

4. Clustering and locating outliers

Random forest handles outliers by essentially binning them for balancing error in class population unbalanced data sets

3.2.3 Challenges

1. Random forest requires much time for training as it combines a lot of decision trees to determine the class.
2. The Range of prediction a Random forest can make is bound by the highest and lowest labels in the training data.

3.3 PSO-BP Neural network

3.3.1 Algorithm for PSO-BP neural network

1) Read data from datasets containing data of gasoline price, glyphosate price etc.
2) The input nodes are initiated with prices of green pepper, compound fertilizer etc. and output is green pepper price.
3) Other parameters like learning rate, training number etc. are set.
4) The PSO process begins and the finest weights and threshold obtained is then delivered to BP neural network for training.
5) The BP neural network is trained using around 85% of the whole data and the rest 15% is used for testing purposes.

6) After the training of data the model is then tested.
7) To measure the accuracy of the model sum of absolute difference (SAD) and mean square error (MSE) are calculated.

3.3.2 Features:

1. Accuracy
The PSO-BP neural network model gives high accuracy as compared to traditional BP-neural networks.
2. High problem Solving ability
The model can solve difficult irregular mapping problems.
3. Wide range.
The model can be applied to solve a variety of large or small range problems.

3.3.3 Challenges:

1. PSO-BP neural network is complex and not easy to use.
2. The model takes a lot of time for training as the PSO algorithm and BP neural network is used.

3.4 K Nearest Neighbours (KNN)

3.4.1 Algorithm for KNN

1) The input dataset is a comma separated values file containing the soil dataset, which has to be subjected to pre-processing.

2) Input dataset is subject to various pre-processing techniques such as filling of missing values, encoding of categorical data and scaling of values in the appropriate range.

3) The acquired input is stored in a variable. The variable reaches out to the crop dataset.

4) The input and the data samples are mapped and the input locates itself into a crop class by using the KNN algorithm of classification in machine learning.

5) The data sample includes the N, P, K and pH for a wide range of crops and fruits grown on the fertile soil of India.

6) As a result of mapping, the dataset returns the suitable crop for the particular soil parameter.

7) The obtained result is displayed to the user through an interface.

3.4.2 Features:

1. Lazy Learning
KNN algorithm is a lazy learning algorithm that takes zero time to learn because it only stores data from the training part.
2. Dimensionality of Input
From feature selection reduces the dimensionality of the input feature space.

3.4.3 Challenges:

1. Computationally expensive
KNN requires high memory need to store all of the training data so it can be computationally expensive.
2. In case of large data the KNN prediction stage might get slow.

Datasets:

1. As Kaggle is widely used for crop datasets, the first dataset we are going to take for Crop Prediction is taken from the Kaggle by the name of Crop Recommendation Dataset. It contains the data regarding Nitrogen (ratio of Nitrogen content in soil), Phosphorus (ratio of Phosphorus content in soil), Potassium (ratio of Potassium content in soil), Temperature (in degree Celsius), Humidity (relative humidity in %), pH (pH value of

the soil), rainfall (in mm), label (which type of crop is grown).

	A	B	C	D	E	F	G	H
1	N	P	K	temperatu	humidity	ph	rainfall	label
2	90	42	43	20.87974	82.00274	6.502985	202.9355	rice
3	85	58	41	21.77046	80.31964	7.038096	226.6555	rice
4	60	55	44	23.00446	82.32076	7.840207	263.9642	rice
5	74	35	40	26.4911	80.15836	6.980401	242.864	rice
6	78	42	42	20.13017	81.60487	7.628473	262.7173	rice
7	69	37	42	23.05805	83.37012	7.073454	251.055	rice
8	69	55	38	22.70884	82.63941	5.700806	271.3249	rice
9	94	53	40	20.27774	82.89409	5.718627	241.9742	rice
10	89	54	38	24.51588	83.53522	6.685346	230.4462	rice
11	68	58	38	23.22397	83.03323	6.336254	221.2092	rice

Fig 1. Dataset for crop prediction analysis

2. Dataset is the most crucial aspect that makes algorithm training possible. The better the collection of the dataset the better the accuracy will be. The Second dataset is taken from Kaggle by the name Agricultural Production in India. The yield prediction module dataset contains the following columns: State, District, Crop, Season, Area and Production as these are some major factors that crops depend on.

State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
Andaman and	NICOBARS	2000	Kharif	Arecanut	1254	2000
Andaman and	NICOBARS	2000	Kharif	Other Khar	2	1
Andaman and	NICOBARS	2000	Kharif	Rice	102	321
Andaman and	NICOBARS	2000	Whole Yea	Banana	176	641
Andaman and	NICOBARS	2000	Whole Yea	Cashewnu	720	165
Andaman and	NICOBARS	2000	Whole Yea	Coconut	18168	65100000
Andaman and	NICOBARS	2000	Whole Yea	Dry ginger	36	100
Andaman and	NICOBARS	2000	Whole Yea	Sugarcane	1	2
Andaman and	NICOBARS	2000	Whole Yea	Sweet pot	5	15
Andaman and	NICOBARS	2000	Whole Yea	Tapioca	40	169

Fig 2. Dataset for yield prediction

Proposed solution:

System uses machine learning to make predictions of Crop, Crop Price and Crop yield prediction. It uses historical data and information to gain experiences and generate a trained model by training it with the data. Proposed system then makes the output prediction. Accuracy of the classifier will be better if the data collection is better.

Data Collection:

1) The atmospheric humidity, temperature, soil moisture, soil pH, area, sunlight and P ,N content of soil are sent to the database for the prediction of Crop. Fig. 1 indicates a dataset for Crop Prediction which we are going to use for prediction.
2) To receive a good harvest, the above conditions should be satisfied. It is needed to have a certain temperature, humidity, soil pH, and soil moisture for a plant to be grown healthy.
3) Yield prediction is based on various factors like state, soil quality, area, season, crop production. Fig. 2 is a dataset for Yield Prediction which consists of these mentioned factors.
4) Compound fertilizer, glyphosate and No.93 gasoline price and green pepper price prediction is considered as data for PSO-BP neural network.

Data Processing Using Machine Learning :
(Dealing with missing values, data cleaning, and train/test split)

1) Crop Prediction : Decision Tree Algorithm, KNN algorithm

After getting the data appropriate for the provided algorithm we can start working on the algorithms for predicting the crop suitable according to conditions.

2) Price Prediction: PSO-BP Neural network Algorithm

The Algorithm is discussed in the algorithmic study [3.3] section for starting the process flow for prediction .Further the model is trained considering 85% data and tested using 15 % of data.The accuracy of model is significantly increased since optimization algorithm i.e PSO is used against the traditional Back propagation neural network.

3) Yield Prediction: Random Forest Algorithm

After taking value as an input from data collection to the algorithm, it returns the corresponding crop yield prediction.

Conclusion and future scope:

Our paper facilitates a comparative and well-formulated study of different machine learning algorithms to efficiently predict some agricultural predictions. For predicting crops on the basis of humidity, pH, rainfall we have studied Decision Tree and another algorithm for prediction is KNN which has majorly Soil related factors.

For yield prediction on the basis of environmental factors and area, Random forest gives the highest prediction with more accuracy. Also for crop price prediction PSO-BP neural network is the most accurate result as it has optimization by PSO and prediction of BP neural network.

Reference:

[1] Nischitha K , Dhanush Vishwakarma , Mahendra N , Ashwini , Manjuraju M.R, International Journal of Engineering Research & Technology (IJERT), Vol. 9 Issue 08, August-2021

[2] F. F. Haque, A. Abdelgawad, V. P. Yanambaka and K. Yelamarthi, "Crop Yield Prediction Using Deep Neural Network," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp

[3] Design And Implementation Of Crop Yield Prediction Model In Agriculture- International Journal Of Scientific & Technology Research Volume 8, Issue 01, January 2020.

[4]. Yung-Hsing Peng, Chin-Shun Hsu, and Po-Chuang Huang “Developing Crop Price Forecasting Service Using Open Data from Taiwan Markets” ,IEEE ,Nov. 20-22, 2015

[5]. YE Lu, LI Yuping, LIANG Weihong, SONG Qidao, LIU Yanqun, QIN Xiaoli(2015)Vegetable Price Prediction Based on PSO-BP Neural Network,8th International Conference on Intelligent Computation Technology and Automation.

[6] S.Pudumalar*, E.Ramanujam*, R.Harine Rajashreeń, C.Kavyań, T.Kiruthikań, J.Nishań, “Crop Recommendation System for Precision Agriculture” ,2016 IEEE Eighth International Conference on Advanced Computing (ICoAC)

[7]Shilpa Mangesh Pande1 , Dr. Prem Kumar Ramesh2 , Anmol3 , B.R Aishwarya4 , Karuna Rohilla5 , Kumar Shaurya, “Crop Recommender System Using Machine Learning Approach” , Proceedings of the Fifth International Conference on Computing Methodologies and Communication (IEEE 2021).

[8] Shravani , Uday Kiran , Yashaswini J , and Priyanka , “Soil Classification And Crop Suggestion Using Machine Learning”, International Research Journal of Engineering and Technology (IRJET). June 2020

[9] S. Veenadhari ,Dr. Bharat Misra , Dr. CD Singh, “Machine learning approach for forecasting crop yield based on climatic parameters”, International Conference on Computer Communication and Informatics, Jan , 2019

[10] Yanghui Kang , Mutlu Ozdogan , Xiaojin zhu , Zhiwei Ye , Christopher Hain and Martha Anderson, Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest,19 May 2020

[11] Archana Gupta , Dharmil Nagda , Pratiksha Nikhare ,Atharva Sandbhor, “Smart crop Prediction using IOT and Machine Learning”,International Research Journal of Engineering and Technology (IRJET), 2020

[12] Dr.A.K.Mariappan, Ms C. Madhumitha, Ms P. Nishitha, Ms S. Nivedhitha , “Crop Recommendation System through Soil Analysis Using Classification in Machine Learning” , International Journal of Advanced Science and Technology , Jun 2020

[13] A. Nigam, S. Garg, A. Agrawal and P. Agrawal, "Crop Yield Prediction Using Machine Learning Algorithms," 2019 Fifth International Conference on Image Information Processing (ICIIP), 2019.

[14] Manoj G S1, Prajwal G S2, Ashoka UR3, Prashant Krishna4, Anitha P, “Prediction and Analysis of Crop Yield using Machine Learning Techniques” , International Journal of Engineering Research & Technology (IJERT)-2020.