e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-5 (October 2021)
https://doi.org/10.31033/ijemr.11.5.22

A Comparative Study of Different Algorithms used to Predict the Crop, its Yield and Price

Pratiksha Pawar[1], Vishwajeet Shinde[2], Aniket Raut[3], Saloni Suke[4] and Prof. Sagar Salunke[5]
[1]Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, INDIA
[2]Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, INDIA
[3]Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, INDIA
[4]Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, INDIA
[5]Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, INDIA

[1]Corresponding Author: pratiksha1132@gmail.com

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 try to predict and recommend crop by considering some soil and atmospheric parameters. So there arevarious 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

I. INTRODUCTION

Agriculture is the backbone of India. As we know, food is a basic necessity for survival, thus agricultural development must be given the highest priority. 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 Provides employment about 50% in our India and is responsible for 18 percent of Indian agriculture gross domestic product (GDP.

Machine learning (ML) overtures are usedin many fields. Machine learning has also been used in agriculture for several years. Crop yield prediction has become one of the important necessities in today's agricultural world for better production. Thus, it turns into challenging problems in meticulous agriculture. As a consequence many models have been proposed and validated so far. This problem is necessary to employ several datasets since crop yield depends on many different factors such as climate, use of fertilizer, weather, soil, area, season 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 toget 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 model which gives predictions based on testing conditions of trees ateach and at every level different decisions are identified. It has various sub trees so the problem is divided and solved by using algorithms.

2) Random Forest

Random forest is an elastic, easy to handle machine learning algorithm that produces, even without hyper-parameter tuning, a splendid result most of the time. Random forest frames multiple decision trees and merges them together to come by adapted 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. In this model optimization is done by PSO algorithm and prediction is done by BP neural network. The weights considered for BP Neural Network are plotted as particles of PSO algorithms. 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, and new instances estimated by summing up the output attributes for the k cases.

II. 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. The 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 result of the given paper we get a suitable crop, market price, and the system takes NPK values as an input to display required NPK for recommended crops.

The proposed system [2] recommended the suitable yield for a particular crop by contemplating factors like water, UV, fertilizer byusing 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 beendone and several attempts are made for application of Machine learning in agricultural fields. Proposed model i.e. Decision tree got moreefficiency 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 optimizationalgorithm 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 thefactors [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 KNNalgorithm gives 92% of accuracy [7].

K-Nearest Neighbour [8] is the best algorithm 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.

III. 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.

Download the dataset from the Kaggle for cropprediction 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 trainingpurposes and 20% for testing.

6) Training data is forwarded to the Decision Tree Regressor.

7) Performing different optimization techniques on the Decision Tree Regressor.

8) Checking the accuracy of the model on thebasis of the test data and calculating the loss function.

9) Stop.

3.1.2 Features

1) Less Efforts

In contrast 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 ofdata and scaling of data as well.

3) No effect of missing data

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

4) Intuitive Model

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

3.1.3 Challenges

1) Splitting of Data

If the data provided to split out the prediction are not related to trained data then a decision tree willmake correct predictions.

2) Over fitting

Decision Trees are inclined to over fitting. That'swhy 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 indata

3) Data contains the key values like state name, district name, crop year, season, crop, Area, production.

4) For converting the categorical data into numerical data, pre-processing on the state name, season, and the crop is needed to be done.

5) Dividing the data into two parts training and testing in such a way that 90% of the data is takenby training data and the rest 10% is taken for testing data.

6) Feed 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 randomforest output layer.

8) The testing of the neural network is to be doneafter it is trained on the training data set.

9) Determining the accuracy of the model.

10) Storing the neural model in h5 format forfurther use.

11) Stop

3.2.2 Features

1) Less computationally expensive

Random forest is less computationally expensive as compared to other algorithms.

2) No need for monotonic transformation andscaling

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

3) Good accuracy and easy to use

The random forest concentrates on improving purity of nodes and mean diminishing accuracyfor feature selection.

4) Clustering and locating outliers

Random forest essentially categorizing outsiders for balancing error in class population unbalanceddata sets.

3.2.3 Challenges

1) Random forest requires much time for training as it combines a lot of decision trees to determinethe class.

2) The Range of prediction a Random forest can make is bound by the highest and lowest labels inthe 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 outputis green pepper price.

3) Other parameters like learning rate, trainingnumber 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 thentested.

7) To measure the accuracy of the model sum of absolute difference (SAD) and mean square error (MSE) are calculated.

3.3.3 Features

1) Accuracy

The PSO-BP neural network model gives highaccuracy as compared to traditional BP-neuralnetworks.

2) High problem Solving ability

The model can solve difficult irregular mappingproblems.

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 noteasy 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) Start

2) The input dataset is a comma separated values file containing the soil dataset.

3) Dataset is taken from kaggle for crop recommendation. Also we will change categoricaldata into numerical data.

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

5) Samples of input data are mapped and the input finds itself into a crop class by using the KNN algorithm of classification in machine learning.

6) The data sample contains some of the soil and atmospheric parameters like N, P, K and pH for a wide range of crops and fruits grown on the fertilesoil of India.

7) After mapping, result of the algorithm shows that the dataset returns the suitable crop for the particular soil parameter.

8) Then obtained result is displayed to the userthrough an interface.

9) Stop.

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

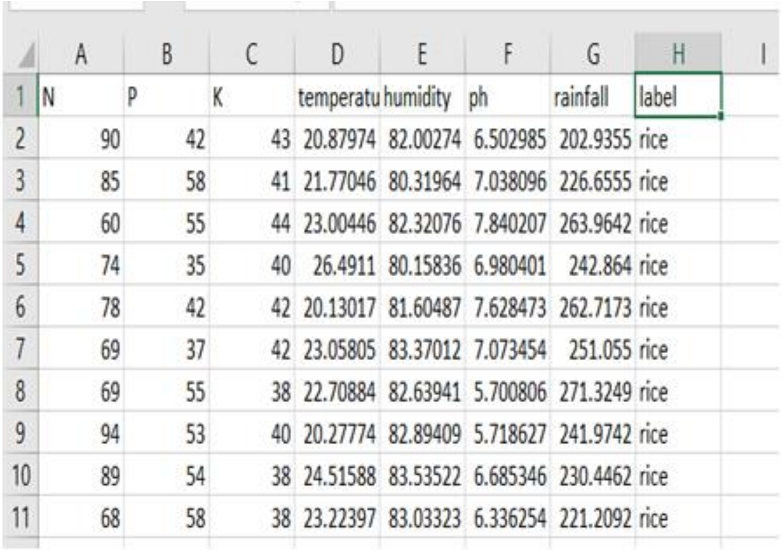
1) Computationally expensive

KNN requires high memory need to store all ofthe training data so it can be computationally expensive.

In case of large data the KNN prediction stage might get slow.

IV. DATASETS

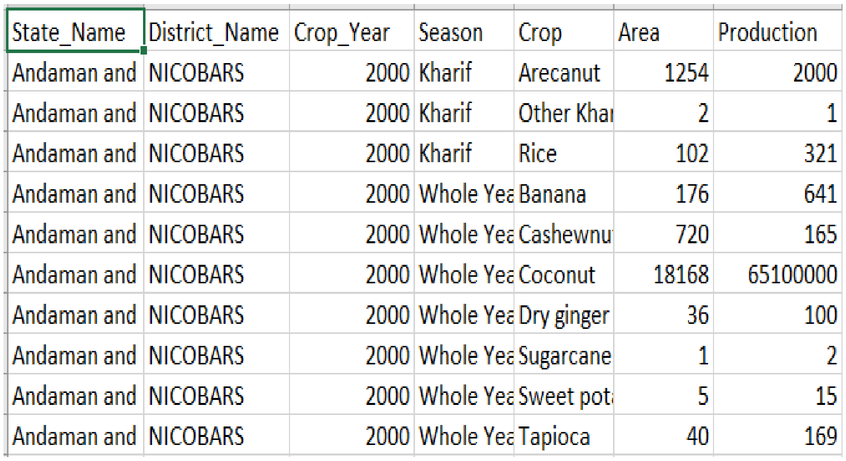
1) As Kaggle is widely used for crop datasets, the first dataset we are going to take for Crop Prediction is taken from the Kaggle by the name of Crop Recommendation Dataset. It contains the data regarding Nitrogen (ratio of Nitrogen content in soil), Phosphorus (ratio of Phosphorus content in soil), 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).This dataset is used in Decision Tree Regressor and KNN.



	A	B	C	D	E	F	G	H	I
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	

Figure 1: Dataset for crop prediction analysis

Dataset is the most important factor 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. This dataset is used in Random Forest algorithm.



State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
Andaman and	NICOBARS	2000	Kharif	Arecanut	1254	2000
Andaman and	NICOBARS	2000	Kharif	Other Khai	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

Figure 2: Dataset for yield prediction

V. 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 suitableaccording to conditions.

2) Price Prediction: PSO-BP Neural networkAlgorithm

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 isused 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.

VI. 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.

REFERENCES

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

[2] F. F. Haque, A. Abdelgawad, V. P. Yanambaka, & K. Yelamarthi. (2020). Crop yield prediction using deep neural network. In: *IEEE 6th World Forum on Internet of Things (WF-IoT).*

[3] Sangeeta & Shruthi G. (2020). Design and Implementation of crop yield prediction model in agriculture. *International Journal of Scientific & Technology Research 8*(01), 544-549.

[4] Yung-Hsing Peng, Chin-Shun Hsu, & Po- Chuang Huang. (2015). Developing crop price forecasting service using open data from taiwan markets. *IEEE.*

[5] YE Lu, LI Yuping, LIANG Weihong, SONG Qidao, LIU Yanqun, & QIN Xiaoli. (2015). Vegetable price prediction based on PSO-BP neural network. In: *8th International Conference on Intelligent Computation Technology and Automation.*

[6] S.Pudumalar, E.Ramanujam, R.Harine Rajashreeń, C.Kavyań, T.Kiruthikań, & J.Nishań. (2016). crop recommendation system for precision agriculture. *IEEE Eighth International Conference on Advanced Computing (ICoAC).*

[7] Shilpa Mangesh Pande, Dr. Prem Kumar Ramesh, Anmol, B.R Aishwarya, Karuna Rohilla, & Kumar Shaurya. (2021). 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, & Priyanka. Soil classification and crop suggestion using machine learning. *International Research Journal of Engineering and Technology (IRJET).*

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

[10] Yanghui Kang, Mutlu Ozdogan, Xiaojin zhu Zhiwei Ye, Christopher Hain, & Martha Anderson. (2020). *Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in*

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-5 (October 2021)
https://doi.org/10.31033/ijemr.11.5.22

the USMidwest.

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

[12] Dr.A.K.Mariappan, Ms C. Madhumitha, Ms P. Nishitha, & Ms S. Nivedhitha. (2020). Crop recommendation system through soil analysis using classification in machine learning. *International Journal of Advanced Science and Technology*.

[13] A. Nigam, S. Garg, A. Agrawal, & P. Agrawal. (2019). Crop yield prediction using machine learning algorithms. *2019 Fifth International Conference on Image Information Processing (ICIIP)*.

[14] Manoj G S, Prajwal G S, Ashoka UR, Prashant Krishna, & Anitha P. (2020). Prediction and analysis of crop yield using machine learningtechniques. *International Journal of Engineering Research & Technology (IJERT)*.