Crop Recomendation System using Support Vector Machine considering indian dataset

Tapas Kumar Mishra*, Sambit Kumar Mishra, Kanaparthi Jeevan Sai, Shreyas Peddi, Lakshmi Pranav Kakumanu, Nukala Snehith Reddy and Manideep Surusomayajula

SRM University-AP, Amaravati, Andhrapradesh, India.
*kmtapas@gmail.com

Abstract. Since a long years agriculture is considered as a major profession for livelihoods of the Indian. Still, agriculture is not profitable as many farmers take the worse step as they can not survive from the burden of loans. So, one such place where there is yet large scope to develop is agriculture. In comparison to other countries, India has the highest production rate in agriculture. However, still, most agricultural fields are underdeveloped due to the lack of deployment of ecosystem control technologies. Agriculture when combined with technology can bring the finest results. Crop yield is a highly complex trait determined by multiple factors such as rainfall, temperature, fertilizers, pesticides, ph level, environment, and their interactions. Predicting the crop selection/yield in advance of its harvest would help the policymakers and farmers for taking appropriate measures for farming, marketing and storage. Thus, in this paper we propose a crop selection using machine learning technique as support vector machine (SVM) and polynomial regression. This model will help the farmers to know the yield of their crop before cultivating onto the agricultural field and thus help them to make the appropriate decisions. It attempts to solve the issue by building a prototype of an interactive prediction system. Accurate yield prediction is required to be done after understanding the functional relationship between yield and these parameters because along with all advances in the machines and technologies used in farming, useful and accurate information about different matters also plays a significant role in it. In this paper, we have simulated SVM and polynomial regression technique to predict which crop can yield better profit. Both of the models are simulated comprehensively on Indian dataset and an analytical report has been presented.

Keywords: Crop Selection · Machine Learning · Indian Agriculture · Prediction .

1 Introduction

Agriculture is the backbone of indian economy. In India, most of the crops depend on the weather conditions. For example, rice cultivation mainly depends on the rainfall. Now-a-days all the seasonal moments are not the same as the

previous. We could not even predict whether there would be any floods or any water scarcity in the future. So in order to maximize the crop production prediction of various aspects of crop are required based on the weather conditions in the locality. Yield prediction is an important agricultural problem. Usually, the farmers used to predict their yield from the previous year's yield. As it is discussed earlier, we could not predict the yield based on last year's outcome due to many factors like crop stress, soil impurity, floods, pesticides, pests and diseases. Here we are going to use some existing mathematical models. As farmers are growing the hybrid products that the soil generally is not supportable but they are using pesticides and growing those. So the quality of the soil decreases. And hence we could not predict for those crops. Due to these abundant inventions people are concentrated on cultivating hybrid crops where they lead to an unhealthy life. Now-a-days, modern people can take the help of technology in various dimensions to grow crops. Thus, this paper aims at predicting the suitable crop type which can be selected by the farmers to grow that is the need of the today's generation. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. Data mining is also useful for predicting crop yield production [5, 2].

1.1 Road Map

In this paper we have focused to predict the annual yielding of a particular crop so that the crop would be producing maximum crop, can be selected by the farmer for farming. The crop yielding is to be predicted by using various machine learning algorithms such as: KNN, SVM, linear regression, polynomial regression. Further, we will make a quantitative analysis of the accuracy. To predict the yielding three parameters will be used, i.e.rainfall at different geographical locations, crop type, season taken from the dataset given by Indian Council of Agricultural Research (ICAR) [1].

Table 1. dataset model

CropType	MinRainfallRequired	Season	Yield
X			Y
—	—	—	—

2 State of the art

This problem is identified before a couple of years. After that many attempts has been made by the researchers throughout the globe. However, there exist many limitations among the farmers and the technological support by the regions and states. Some of the directions for this issue are described here as follows.

	CROP	SEASON	MeanRainfallRequired	MeanYield
0	Rice	Kharif	901.830769	2204.5000
1	Rice	Rabi	157.059172	3211.1875
2	Rice	Total	146.790532	2301.1250
3	Wheat	Rabi	97.884539	2962.1250
4	Jowar	Kharif	821.280128	1055.1875
5	Jowar	Rabi	140.953092	751.3125
6	Jowar	Total	70.170475	878.4375
7	Bajra	Kharif	831.371474	1079.1875

Fig. 1. Sample Dataset

The authors in [6] mainly focuses on predicting the yield of the crop based on the existing data by using the Random Forest algorithm. The data represents the scenario of Tamilnadu. Random forest algorithm is used for accurate crop yield prediction. This random forest algorithm is used for maintaining high accuracy and precision. They have used adaptive boosting(meta-algorithm) for increasing the efficiency of the model. They have suggested that implementation of decision trees is quite easier than that of random forest. But when considering the accuracy, random forest is the best choice and K-means algorithm is used to forecast the pollution factor in the atmosphere. The data is taken from [www.data.gov.in] and data about climate change was gathered from [www.imd.gov.in]. Random forest classifier is used for both the regression analysis and classification analysis. So this algorithm is used for both classification and regression.

In [8], the authors have predicted crop yield using decision tree classifier. They used rainfall, perception, production, temperature data to construct a random forest which is a collection of decision trees using $\frac{2}{3}$ rd of the data and they tested using $\frac{1}{3}$ rd of the data. Usually, decision tree classifiers uses greedy approach, where an attribute chooses at first step can't be used anymore which can give better classification if used in later steps. Also it overfits the training data which can give poor results for unseen data to overcome which they have combined results from different models to get a better result.

In [7], they have taken a dataset containing soil type, soil Ph, Humidity, Temperature, Rainfall, Wind, Production, Cost of Production and annual yield of that region for past 10-12 years and a decision tree classifier model has been implemented on the data for crop yield and K-Nearest neighbours has been applied for prediction of rainfall with 76.8% accuracy for crop yield prediction and 89.4% accuracy for rainfall prediction.

In [4], the authors used a deep neural network model to predict four crop yields namely: Aus-rice, Aman rice, Boro rice, Jute, Wheat and Potato using rainfall data, land types, chemical fertilizers, soil information. The DNN model is compared with RF, SVM and LR. DNN outperforms than other models with highest accuracy rate of 98% (Aus rice), 95% (Aman rice), 96% (Boro rice), 97% (Potato), 96% (Wheat) and 94% (Jute).

The authors in [9] investigated the crop suggestion model based on soil classification using machine learning techniques. The study proposed an SVM based model to suggest crops which are specific to soil conditions. The proposed SVM model outperforms KNN and bagged trees with 95% of accuracy.

In [10], the authors investigated the rice yield prediction performance of KNN, decision tree(DT) and Naive Based(NB) using 11 parameters of micronutrients and macronutrients. The prediction accuracy for Naive Based is 98%, DT is 94% and KNN is 97% is achieved. The study concluded that NB achieved better prediction rate and was suitable for rice yield prediction using soil parameters.

In [3] proposed a crop recommender model for farmers using machine learning models. The prediction model is prepared using ANN and the model performance is compared against DT, KNN, RF. ANN achieved a highest of 91% than other models. The crop suitability is predicted using rainfall, soil type, soil conditions, temperature and geographical location.

3 Existing techniques

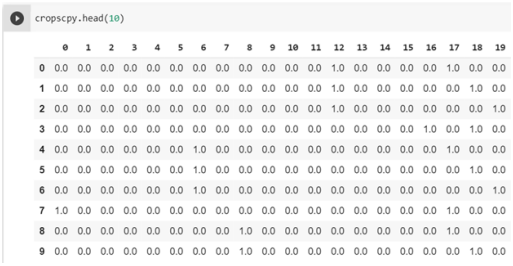
3.1 K-NearestNeighbour

KNN stands for K-Nearest Neighbour. In this supervised learning algorithm which is used for classification, we classify based on how it's neighbours are classified. It stores all it's previous cases and classifies new ones based on how similar it is to the previous cases. Here, K signifies the amount of neighbours we take for comparing the distance. We find the distance between new and previous cases based on the equation 1. The distance D using minkowski equation is:

$$D(X_1, X_2) = \left[\sum |X_1 - X_2|^{1/p}\right]^p \tag{1}$$

Here, value of p is taken as 2 for minkowski distance. Further, X_1 represents coordinates of the neighbour nodes and X_2 represents coordinates of the new node. By calculating all the k-nearest point distances we can get a decision based on those coordinates. Consider the fig 1 above where you can see three X variable columns and one Y variable column. In those X columns we have two categorical variables and one numerical variable and Y is also a numeric variable. So before applying formula we have to convert the categorical variables into numerical before training and it is easy to display the mathematical model and to train the model too.

The categorical rows are transformed into the numerical rows by applying one hot encoding which is shown in the Fig. 2. Further,the crops column is

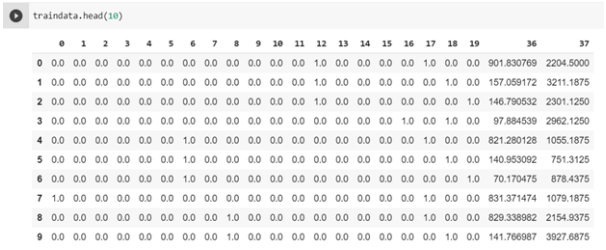


```
cropscpy.head(10)
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
4	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
7	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Fig. 2. Dataset after applying the one hot encoding

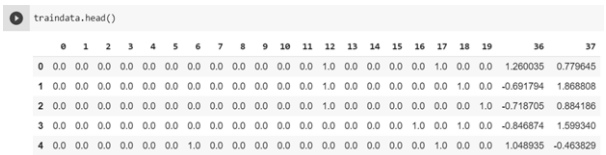
represented by columns 0-16 and the season column is represented by columns 17-19 while applying the KNN. In addition to this, column 36 is MeanRainFall-Required which is appended after pre processing the dataset and column 37 is YieldOfTheCrop which is to be predicted.



```
traindata.head(10)
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	36	37
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	901.830769	2204.5000
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	157.059172	3211.1875
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	146.790532	2301.1250
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	97.884539	2962.1250
4	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	821.280128	1055.1875
5	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	140.953092	751.3125
6	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	70.170475	878.4375
7	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	831.371474	1079.1875
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	829.338982	2154.9375
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	141.766987	3927.6875

Fig. 3. Dataset used for training the model



```
traindata.head()
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	36	37
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.260035	0.779645
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	-0.691794	1.868808
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	-0.718705	0.884186
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	-0.846874	1.599340
4	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.048935	-0.463829

Fig. 4. Dataset after applying Standard Scalar

3.2 REGRESSION

Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x). Regression analysis is majorly used for prediction

purposes as it provides predicted entities as a function of the dependent entities. The main goal of regression is the construction of an efficient model to predict the dependent attributes from a bunch of attribute variables. The regression model is to build a mathematical equation that defines y as a function of the x variables. Next, this equation can be used to predict the outcome (y) based on new values of the predictor variables (x). Data generated as output in this regression was used in model verification and analysis.

Some types of regression models used in this prediction are:

Linear Regression Linear regression algorithm shows a linear relationship between a dependent (Y) and one or more independent (X) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. The linear regression model provides a straight line representing the relationship between the variables. Consider the Fig. 5 presented below:

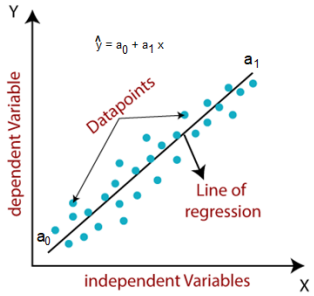


Fig. 5. Representing Linear Regression

Where, Y= dependent variable (Target Variable), X= independent variable (predictor Variable) a_0= intercept of the line (Gives an additional degree of freedom) a_1 = Linear regression coefficient (scale factor to each input value). ϵ = random error The values for x and y variables are training datasets for Linear Regression model representation. Consider the Fig 1 above where you can see three X variable columns and one Y variable column. In those X columns we have two categorical variables and one numerical variable and Y is also a numeric variable. So before applying formulae we have to convert the categorical variables into numerical before training and it is easy to display the mathematical model and to train the model too. The categorical rows become into these numerical rows and instead of Crops column we can use 0-16 columns and instead of season we can use 17-19 columns while applying our model.

Polynomial Regression A regression of y on x may be a polynomial regression of y on x if the ability of independent variable power is greater than one. Such an equation is known as the polynomial regression equation.

$$Y = a + b * X^2 \tag{2}$$

In this polynomial regression technique, the best fit line is a curve line but not a straight line. That curved line fits into the data points.

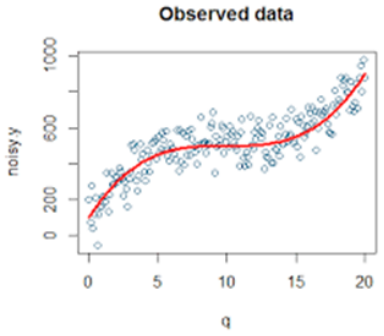


Fig. 6. Representing Polynomial Regression

Support Vector Machine It is a supervised learning algorithm that we can use for both classification and regression tasks. It can be used in our project in such a way that it can classify which soil and atmospheric conditions are suited for growing a crop and provide maximum yield to the farmers. In general, we will have many independent features(in our case features include rainfall in that area, crop and season) to determine the dependent feature and in turn as there are many features the dimension of the dataset will be more so as complexity. And classify the dataset based on linear or non linearly separable hyperplane. Based on it we can classify the crop in a reasonable way. Hyperplane is the decision boundary that classifies data points into respective categories either in one dimensional or multidimensional. In two dimension its a line and in three dimension its a plane and more than that its a hyperplane.

4 Simulation and Analysis

We have simulated the models using suport vector machine and polynomial regression extenssively on the given dataset [1]. The simulation setup and models on this above dataset is as follows.

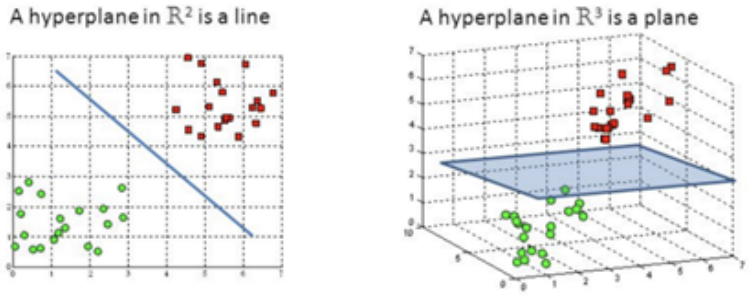


Fig. 7. Representing Support Vector Machine

4.1 Simulation

Simulation using Support Vector Machine) In this algorithm, we have created a support vector machine model with parameters $c = 1$, $\epsilon = 0.1$, $kernel = rbf$, where c is used for optimising the hyperplance and the value of c is inversely propertional to the margin of the hyperplance, ϵ is the error rate, $kernel$ is used for non-linear separation of the data, in our model we have used radial basis function (rbf). We converted the gathered data into integers using one hot encoding and train the model with input parameters (X) such as crop type, season and rainfall and give us output parameter (Y) crop yield. In support vector machine when the model is made we make planes and each input is called support vector points and we calculate the distance between the point and the plane using the formula $-\gamma|(|x_1 - x_2|)|^2$ where $\|x_1 - x_2\|$is the euclidean distance and x_1 is a point on the model and x_2 is an input point, γ is a hyper parameter used for curvature of the kernel. Using the distance we calculate the similarity of x_1 and x_2 to get the resultant value.

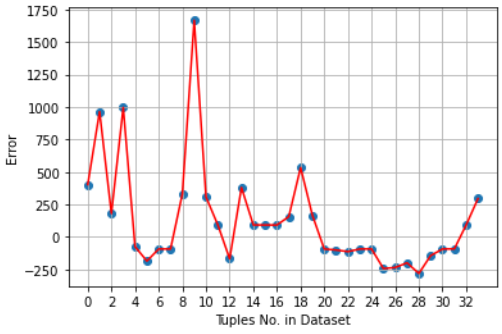


Fig. 8. Error Plot for SVM Regression Model

Simulation of Polynomial Regression Generally we know that polynomial regression is the relationship between the independent variable x and the dependent variable y is modelled as n^{th} degree polynomial. For using this regression in this paper we have changed parameters to numbers and we have created a dataset related to this project. Now for calculating value of crop we have taken two things those are type of crop and season and convert it into one-hot encoding format and calculating the distance using the equation 3 of polynomial regression.

$$Y = b_0 + b_1 * x_1 + b_2 * x_2^2 + b_3 * x_3^3 + \ldots + b_n * x_n^n + \epsilon \tag{3}$$

wehre, Y is targeted output and b_0, b_1, ..., b_n are regression coefficients and x is input variable and

e is the residual error. In our data set we have three features they are: 1) Season, 2) Crop, 3) Mean rainfall required, 4) Meain yield. So, according to the features we can form equation is:

$$Y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \epsilon \tag{4}$$

where, x_1 = CROP, x_2= SEASON, x_3= MeanRainfallRequired, Y = MeanYield. Here we are taking degree of the equation as one only because if we take degree above than one it will be overfitted as in our dataset less number of duplicate occurances of records are there.

Using one-hot encoding we have encoded $CROP$ and $SEASON$ the crop became column-17 and season became three colums on total they are encoded into 20 columns and 21st column is minimum rainfall required. Thus, $b_0 - b_{16}$ Belongs to Crop, $b_{17} - b_{19}$ Belongs to Season and output of the equation is $PredictedYield$.

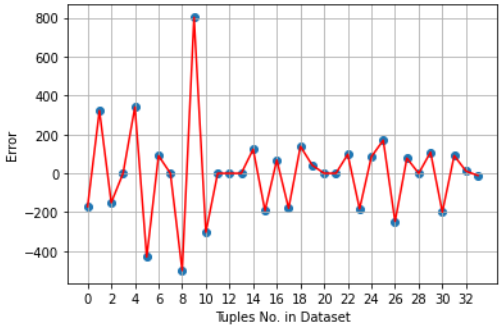


Fig. 9. Error Plot for Polynomial Regression Model

4.2 Analysis of Models

Here we have trained our dataset with Support Vector Machine and Polynomial Regression. It is observed that polynomial regression shows better accuracy when it is simulated with our dataset. further, it is observed that when dataset can not be regressed linearly kernels like RBF kernel, Gaussian Kernel, etc. are used to separate them. So whenever the dataset is non linear and when we need to separate it by a kernel SVM will be good. Furthermore, whenever the relationship between X and Y is linear or non linear and if the dataset is large, polynomial regression may result better accuracy.

Table 2. Accuracy Analysis

Model	Accuracy
SupportVectorMachine	78.92762726425498
PolynomialRegressor(Degree 1)	93.92827101446136

5 Conclusion

In this paper, we have shown that we can predict the yield of the crop based on the crop-type, season, rainfall approximately. If we increase the parameters in the dataset, the accuracy may increase. We have used SVM model and polynomial regression for predicting the crop yield. This would help the farmers to increase their decision making capability which will increase their profit. Using these ideas, the agricultural sector can be transformed to a profitable sector which can gain interest among the new generation. Further, it is observed that prediction accuracy of the different models may vary by different datasets. Thus, in future a hybrid model with more attributes may be designed for robust prediction.

References

1. of Agricultural Research (ICAR), I.C.: Field Crop varieties released (Central Release).
urlhttps://data.gov.in//catalog//field-crop-varieties-released-central-release (2014), [Online; accessed 21-February-2021]
2. Chlingaryan, A., Sukkarieh, S., Whelan, B.: Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and electronics in agriculture **151**, 61–69 (2018)
3. Doshi, Z., Nadkarni, S., Agrawal, R., Shah, N.: Agroconsultant: Intelligent crop recommendation system using machine learning algorithms. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). pp. 1–6. IEEE (2018)

4. Islam, T., Chisty, T.A., Chakrabarty, A.: A deep neural network approach for crop selection and yield prediction in bangladesh. In: 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). pp. 1–6. IEEE (2018)
5. van Klompenburg, T., Kassahun, A., Catal, C.: Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture **177**, 105709 (2020)
6. Kumar, R., Singh, M., Kumar, P., Singh, J.: Crop selection method to maximize crop yield rate using machine learning technique. In: 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM). pp. 138–145. IEEE (2015)
7. Patil, A., Kokate, S., Patil, P., Panpatil, V., Sapkal, R.: Crop prediction using machine learning algorithms. International Journal of Advancements in Engineering & Technology **1**(1), 1–8 (2020)
8. Priya, P., Muthaiah, U., Balamurugan, M.: Predicting yield of the crop using machine learning algorithm. International Journal of Engineering Sciences & Research Technology **7**(1), 1–7 (2018)
9. Rahman, S.A.Z., Mitra, K.C., Islam, S.M.: Soil classification using machine learning methods and crop suggestion based on soil series. In: 2018 21st International Conference of Computer and Information Technology (ICCIT). pp. 1–4. IEEE (2018)
10. Singh, V., Sarwar, A., Sharma, V.: Analysis of soil and prediction of crop yield (rice) using machine learning approach. International Journal of Advanced Research in Computer Science **8**(5) (2017)