Using Random Forest Algorithm In Agricultural Production Forecasting

Author Details:
[(1)]**Trinh Van Chung –** IT Faculty - Nguyen Trai University, Hanoi
[(2)]**Nguyen Viet Quang -** Grade 12 Maths - Tran Phu High School for the Gifted - Hai Phong

Abstract
Forecasting agricultural output plays an important role in government policies. In recent years, Vietnam has always had an oversupply crisis. Accurate forecasting of agricultural output helps state agencies be proactive in finding output for products and measures to reserve when supply exceeds demand. In case demand exceeds supply, it will push households to increase productivity or add more crops as well as increase acreage to be able to meet the needs of the agricultural market.
There are many methods to forecast a continuous quantity such as linear regression, co-integrated moving average (ARIMA) autoregressive models, and artificial neural networks... This paper uses the random forest regression tool to predict and compare with the multiple linear regression method. Using India's agricultural data we found that the Random Forest method gave much better results. We will use the Random Forest tool to forecast Vietnam's agricultural output when sufficient data is collected.
Keywords: *machine learning, Random Forest algorithm, agricultural production, forecasting*

1. INTRODUCTION

In developing countries, many people consider agriculture the main source of income. In recent years, agricultural growth has been driven by environmental and technical innovations. In addition, the use of information technology can increase the accuracy of decision-making, and thus farmers can profit in the best way. Therefore, the process and techniques of data mining related to agriculture need to be paid more attention to and used to be able to find knowledge from that data to make optimal decisions in agricultural production and consumption.

Vietnam is a fairly populous country with a population of more than 100 million people and is the second largest rice exporter in the world. Agriculture is an important sector affecting the economy of Vietnam, it contributes up to 13.5% of Vietnam's gross domestic product (GDP) and provides jobs for about 50% of the Vietnamese population. However, the production of crops is affected by different factors, such as soil type, rainfall, seed quality, etc. Therefore, a prediction system is needed to minimize loss and maximize crop yield and profit for farmers.

This project's scope is to investigate a crop profile dataset for the agricultural industry using machine learning techniques. We focus on implementing crop yield prediction systems using machine learning techniques by performing analysis on agricultural datasets.

2. METHODOLOGY

In agriculture, Machine Learning is considered a new field, as much work has been done with the help of machine learning in the field of agriculture. There are different philosophies put forward and evaluated by researchers worldwide in the field of agriculture and related sciences.

CH. Vishnu Vardhan Chowdary, and Dr. K. Venkataramana [1], developed the ID3 algorithm to improve the crop yield quality of tomatoes, and it was implemented on the PHP platform, and the dataset was used in CSV format. Temperature, area, humidity, and tomato yield were the various parameters used in this study. Sujatha and P. Isakki [2] used data mining techniques for prediction: this model works based on different parameters like crop name, soil area, soil type, pH value, grain type, and water and also predicts crop outbreak and disease. In this way, one has the right to choose crops depending on climate data and required parameters. N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy [3], proposed SVM to predict the crop yield of rice.

2.1. Decision Tree algorithm

2.1.1. Entropy

Given a discrete variable x taking the values with probabilities $p_1, p_2, ..., p_n$ (respectively) $0 \le p_i \le 1$; $\sum_{i=1}^{n} p_i = 1$.

Denote $p = (p_1, p_2, ..., p_n)$. The entropy of the p-distribution is

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i)$$

The entropy function has its minimum value if there is a value of $p_i = 1$, and its maximum value if all p_i are equal.

2.1.2. ID3 Algorithm

The sum of the weights of the entropy at the leaf nodes after building the decision tree is considered the loss function of that decision tree. The weights here are proportional to the number of data points assigned to each node. The job of the algorithm is to find reasonable ways of division (reasonable order of attribute selection) so that the final loss function has as small a value as possible. This is achieved by selecting an attribute such that if it is used for division, the entropy at each step is reduced by a maximum.

Consider a problem with different C classes, working with a non-leaf node with data points belonging to a set S, |S| = N. Among the N points, there are N_c points of class c. The probability that a data point falls into class c is approximately equal to $\frac{N_c}{N}$. The **entropy** at this node is calculated as:

$$H(S) = -\sum_{c=1}^{C} \frac{N_c}{N} \log\left(\frac{N_c}{N}\right)$$

Assuming the selected attribute is x, the points in S are divided into child nodes S_1, S_2, K, S_k.

Define $H(x,S) = \sum_{k=1}^{K} \frac{m_k}{N} H(S_k)$ as the entropy weighted sum of each child node, information gain $G(x,S) = H(S) - H(x,S)$

The objective is to choose x such that $G(x,S)$ max, or $H(x,S)$ min

2.1.3. CART algorithm

This algorithm is quite similar to ID3 but uses the Gini index instead of information gain.

Gini at each node: $Gini = 1 - \sum_{i=1}^{C} (p_i)^2$, where C is the number of classes to be classified, ni is the number of elements of the i[th] class, N is the total number of elements in that node, $p_i = \frac{n_i}{N}, \sum_{i=1}^{C} p_i = 1$.

The Gini function has its minimum value if there is a value of $p_i = 1$, and its maximum value if all p_i are equal.

$Gini\_index = Gini(p) - \sum_{k=1}^{K} \frac{m_k}{M} Gini(c_k)$ is the Gini index, similar to information gain.

To avoid the case of overfitting (the model is very accurate on the training set but inaccurate on the test set), there are several methods: using a stop condition or pruning methods.

2.2. Random Forest algorithm

The Random Forest algorithm is done by selecting a set with n data and k random attributes from the dataset and using the Decision Tree algorithm with the selected dataset.

The Random Forest algorithm will include many decision trees, each tree is built using the Decision Tree algorithm on different data sets and using different attribute sets. Then the prediction results of the Random Forest algorithm will be aggregated from the decision trees.

2.3. Multiple regression model

Assuming the population to be studied has N elements, a randomly selected n-element part of the population is called a sample. The results that are inferred from the observed sample are called estimators. The regression function and the regression model with the parameters to be estimated are called the sample regression function and the sample regression model.

Sample regression model:

$$\hat{Y} = \hat{b}_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + ... + \hat{b}_{k+1} X_{k+1} + e, e: u$$

Sample regression function:

$$\hat{Y} = \hat{b}_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + ... + \hat{b}_{k+1} X_{k+1}$$

where $\hat{b}_1, \hat{b}_2, ..., \hat{b}_{k+1}$ are the estimates of the corresponding coefficients in the population regression model.

The regression function is the estimate of the population regression function, and the value of the sample regression function is the estimate for the expectation of the dependent variable. To estimate close to the observed value of the dependent variable, we use the method of ordinary least square (OLS-ordinary least square). That is, $\hat{b}_1, \hat{b}_2, ..., \hat{b}_{k+1}$ are estimated that the value of the sum of squares of the differences between the observed values is minimal.

Error function $S = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{b}_1 - \hat{b}_2 X_{i,2} - \hat{b}_3 X_{i,3} - ... - \hat{b}_{k+1} X_{i,k+1}\right)^2$

We need to find the set of parameters $\hat{b}_1, \hat{b}_2, ..., \hat{b}_{k+1}$ so that the error function *S* reaches the minimum value.

The parameter set $\hat{b}_1, \hat{b}_2, ..., \hat{b}_{k+1}$ is the solution of the system of equations:

$$S = \sum_{i=1}^{n} \left(Y_i - \hat{b}_1 - \hat{b}_2 X_{i,2} - \hat{b}_3 X_{i,3} - ... - \hat{b}_{k+1} X_{i,k+1}\right)^2$$

$$\begin{cases} \frac{\partial S}{\partial \hat{b}_1} = 0 \\ \frac{\partial S}{\partial \hat{b}_2} = 0 \\ \vdots \\ \frac{\partial S}{\partial \hat{b}_{k+1}} = 0 \end{cases} \Rightarrow \begin{cases} 2\sum_{i=1}^{n}\left(Y_i - \hat{b}_1 - \hat{b}_2 X_{i,2} - \hat{b}_3 X_{i,3} - \ldots - \hat{b}_{k+1} X_{i,k+1}\right)(-1) = 0 \\ 2\sum_{i=1}^{n}\left(Y_i - \hat{b}_1 - \hat{b}_2 X_{i,2} - \hat{b}_3 X_{i,3} - \ldots - \hat{b}_{k+1} X_{i,k+1}\right)(-X_{i,2}) = 0 \\ \vdots \\ 2\sum_{i=1}^{n}\left(Y_i - \hat{b}_1 - \hat{b}_2 X_{i,2} - \hat{b}_3 X_{i,3} - \ldots - \hat{b}_{k+1} X_{i,k+1}\right)(-X_{i,k+1}) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n}(\hat{b}_1 + \hat{b}_2 X_{i,2} + \hat{b}_3 X_{i,3} + \ldots + \hat{b}_{k+1} X_{i,k+1}) = \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n}(\hat{b}_1 + \hat{b}_2 X_{i,2} + \hat{b}_3 X_{i,3} + \ldots + \hat{b}_{k+1} X_{i,k+1}) X_{i,2} = \sum_{i=1}^{n} Y_i X_{i,2} \\ \vdots \\ \sum_{i=1}^{n}(\hat{b}_1 + \hat{b}_2 X_{i,2} + \hat{b}_3 X_{i,3} + \ldots + \hat{b}_{k+1} X_{i,k+1}) X_{i,k+1} = \sum_{i=1}^{n} Y_i X_{i,k+1} \end{cases}$$

$$\Rightarrow \begin{cases} \hat{b}_1 \sum_{i=1}^{n} 1 + \hat{b}_2 \sum_{i=1}^{n} X_{i,2} + \ldots + \hat{b}_{k+1} \sum_{i=1}^{n} X_{i,k+1} = \sum_{i=1}^{n} Y_i \\ \hat{b}_1 \sum_{i=1}^{n} X_{i,2} + \hat{b}_2 \sum_{i=1}^{n} X_{i,2} X_{i,2} + \ldots + \hat{b}_{k+1} \sum_{i=1}^{n} X_{i,k+1} X_{i,2} = \sum_{i=1}^{n} Y_i X_{i,2} \\ \vdots \\ \hat{b}_1 \sum_{i=1}^{n} X_{i,k+1} + \hat{b}_2 \sum_{i=1}^{n} X_{i,2} X_{i,k+1} + \ldots + \hat{b}_{k+1} \sum_{i=1}^{n} X_{i,k+1} X_{i,k+1} = \sum_{i=1}^{n} Y_i X_{i,k+1} \end{cases} \quad (*)$$

$$Let\ X = \begin{bmatrix} 1 & X_{12} & \cdots & X_{1,k+1} \\ 1 & X_{22} & \cdots & X_{2,k+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,2} & \cdots & X_{n,k+1} \end{bmatrix} \Rightarrow X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1,k+1} & X_{2,k+1} & \cdots & X_{n,k+1} \end{bmatrix}$$

$$\Rightarrow X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1,k+1} & X_{2,k+1} & \cdots & X_{n,k+1} \end{bmatrix} \begin{bmatrix} 1 & X_{12} & \cdots & X_{1,k+1} \\ 1 & X_{22} & \cdots & X_{2,k+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,2} & \cdots & X_{n,k+1} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} X_{i,2} & \cdots & \sum_{i=1}^{n} X_{i,k+1} \\ \sum_{i=1}^{n} X_{i,2} & \sum_{i=1}^{n} X_{i,2} X_{i,2} & \cdots & \sum_{i=1}^{n} X_{i,2} X_{i,k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{i,k+1} & \sum_{i=1}^{n} X_{i,2} X_{i,k+1} & \cdots & \sum_{i=1}^{n} X_{i,k+1} X_{i,k+1} \end{bmatrix}$$

and $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ M \\ Y_n \end{bmatrix}$ $b = \begin{bmatrix} b_1 \\ b_2 \\ M \\ b_{k+1} \end{bmatrix}$ $\hat{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ M \\ \hat{b}_{k+1} \end{bmatrix}$

It follows that $X'Y = \begin{bmatrix} 1 & 1 & L & 1 \\ X_{12} & X_{22} & L & X_{n,2} \\ M & M & K & M \\ X_{1,k+1} & X_{2,k+1} & K & X_{n,k+1} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ M \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i X_{i,2} \\ M \\ \sum_{i=1}^{n} Y_i X_{i,k+1} \end{bmatrix}$

And then (*) becomes: $X'X\hat{b} = X'Y$

If $\det(X'X) \neq 0$ we have the solution $\hat{b} = (X'X)^{-1}X'Y$ is the estimate for the coefficient of the multiple regression model.

3. DATA

We have a database of more than 240 thousand records of agricultural production in India including data on the province, district, year, crop, area, and production. The productivity of each different land is different, the different seasons are different. The data, therefore, provide important information for forecasting output. However, the data here lacks rainfall, wind speed, and fertilizer amount, so the forecast results are somewhat inaccurate.

	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0
...	...	...	...	...	...	...	...
246086	West Bengal	PURULIA	2014	Summer	Rice	306.0	801.0
246087	West Bengal	PURULIA	2014	Summer	Sesamum	627.0	463.0
246088	West Bengal	PURULIA	2014	Whole Year	Sugarcane	324.0	16250.0
246089	West Bengal	PURULIA	2014	Winter	Rice	279151.0	597899.0
246090	West Bengal	PURULIA	2014	Winter	Sesamum	175.0	88.0

246091 rows × 7 columns

Figure 1. Data used for the study

Categorical variables have multiple values for each attribute. We encode the data into independent variables and take the output as the dependent variable. Take 80% data for training and 20% data for testing.

	0	1	2	3	4	5	6	7	8	9	...	673	674	675	676	677	678	679	680	681	682
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	2000.0	1.0	2.0	1254.0
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	0.0	0.0	0.0	0.0	2000.0	1.0	74.0	2.0
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	0.0	0.0	0.0	0.0	2000.0	1.0	95.0	102.0
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	2000.0	4.0	7.0	176.0
4	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	2000.0	4.0	22.0	720.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
242356	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	2014.0	3.0	95.0	306.0
242357	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	2014.0	3.0	102.0	627.0
242358	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	2014.0	4.0	106.0	324.0
242359	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	2014.0	5.0	95.0	279151.0
242360	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	2014.0	5.0	102.0	175.0

242361 rows × 683 columns

Figure 2. Data has been coded for forecasting

	0	1	2	3	4	5	6	7	8	9	...	673	674	675	676	677	678	679	680	681	682
50925	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	2007.0	2.0	119.0	5467.0
54763	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	2012.0	2.0	92.0	830.0
101919	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	2012.0	4.0	87.0	9.0
158521	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	2010.0	2.0	63.0	13.0
225597	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	2002.0	2.0	59.0	117.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
178440	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	2005.0	1.0	25.0	619.0
137884	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	2013.0	4.0	87.0	30.0
111609	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	2008.0	1.0	25.0	134.0
145776	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	2005.0	0.0	43.0	17956.0
220185	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	2009.0	3.0	116.0	1.0

48473 rows × 683 columns

Figure 3. Data used for testing

4. RESULTS

We use the linear regression model and Random Forest model to predict agricultural production based on geographical location, crop season, and cultivated area. The results by the linear regression model give much more inaccurate results even with many negative predictors (while the output is non-negative).

For the random forest model, the accuracy is much higher. Here are the first 20 values of the forecast results using different models:

	y_Randomforest	y_DecisionTree	y_LinearRegression	y_test
0	4466.9700	4355.00	-3.789347e+05	4529.0
1	458.7627	308.00	-1.236674e+05	425.0
2	120.9400	95.00	4.430467e+05	254.0
3	4.1060	5.00	7.654590e+05	7.0
4	144.8500	134.00	1.873790e+05	130.0
5	24.9600	20.00	1.963607e+04	21.0
6	132066.5200	132208.00	5.580524e+05	98392.0
7	203.5200	152.00	-1.042517e+06	182.0
8	536.4831	680.00	1.082600e+05	835.0
9	327.1000	283.00	-9.710170e+05	190.0
10	13975.8825	7476.00	3.912524e+05	13796.0
11	1905.5770	1471.00	-4.498811e+05	1802.0
12	151.6000	133.00	-6.514147e+05	132.0
13	14.4800	0.00	2.100696e+05	0.0
14	223.8700	271.00	9.817232e+04	284.0
15	405.9800	197.00	-6.063158e+05	224.0
16	21.1800	6.00	1.198891e+06	6.0
17	913.4663	988.51	3.512192e+07	873.0
18	14.7300	15.00	-4.025861e+05	15.0
19	23.3090	25.00	-1.033151e+06	25.0

Figure 3. Forecast results by Randomforest method, Decision Tree, Linear Regression, and actual observed value - y_test.

According to the results of Table 2, we see that the Random Forest method gives much more accurate results than other methods.

	y_Randomforest	y_DecisionTree	y_LinearRegression	y_test
0	4466.9700	4355.0	-3.789347e+05	4529.0
1	458.7627	308.0	-1.236674e+05	425.0
2	120.9400	95.0	4.430467e+05	254.0
3	4.1060	5.0	7.654590e+05	7.0
4	144.8500	134.0	1.873790e+05	130.0
...	...	...	...	...
48468	227.4300	200.0	5.510765e+05	212.0
48469	183.0730	175.0	7.630898e+05	182.0
48470	367.9580	54.0	-1.965410e+05	42.0
48471	16385.2600	14023.1	-1.349594e+05	16070.6
48472	1.0000	1.0	-1.159327e+06	1.0

48473 rows × 4 columns

Figure 4. Forecast results for the test set consisting of 48473 observations

	MAE_Randomforest	MAE_DecisionTree	MAE_LinearRegression
0	175488.726964	194612.12545	1.611241e+06

Figure 5. The average absolute error of 3 methods

Figure 5 gives the average absolute error (MAE) of the 3 methods compared to the test data set. This result shows that the Random Forest method has the smallest MAE.

5. CONCLUSION

This study tested three methods to forecast crop yields in different regions and seasons. The methods shown have different accuracy, in which the popular and traditional method, the linear regression method, is not suitable for forecasting in this data set. The decision tree method and the Random Forest method give more accurate results. The Random Forest method gives results that the absolute error is 10 times smaller than that of the linear regression method.

In the next study, we will collect data from Vietnam to make forecasts and possibly policy implications from the forecasts for Vietnamese agriculture.

6. REFERENCES

i. *CH. Vishnu Vardhanchowdary, Dr.K.Venkataramana, Tomato Crop Yield Prediction using ID3, March 2018, IJIRT Volume 4 Issue 10 pp,663-62.*

ii. *R. Sujatha and P. Isakki, A study on crop yield forecasting using classification techniques 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, 2016, pp. 1-4.*

iii. *N. Gandhi, L. J. Armstrong, O. Petkar and A. K. Tripathy, Rice crop yield prediction in India using support vector machines 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), KhonKaen, 2016, pp. 1-5.*

iv. *N. P. Sastra and D. M. Wiharta, ―Environmental monitoring as an IoT application in building smart campus of UniversitasUdayana, in Proc. Int. Conf. Smart Green Technol. Elect. Inf. Syst. (ICSGTEIS), Oct. 2016, pp. 85–88.*

v. *M. Suganya., Dayana R and Revathi. R, Crop Yield Prediction Using Supervised Learning Techniques, International Journal of Computer Engineering and Technology, 11(2), 2020, pp. 9-20.*