COMPUTATIONAL PREDICTION MODEL FOR PEPPER YIELD PREDICTION USING SUPPORT VECTOR REGRESSION

Akhil Wilson, Hemalatha N* and Raji Sukumar
akhilwilson999@gmail.com, , rajivinod.a@gmail.com, hemalatha@staloysius.ac.in*

Abstract: The yield prediction is the one of the challenging problem in agriculture. Here in this research work we have predicted the yield of Pepper in the state of Kerala, India. With the help of Machine Learning and by considering the soil properties, micro climatic condition and area of the Pepper we have predicted the yield. Here we have used Linear Regression and Support Vector Regression algorithms in order to predict the pepper yield. Experimental results gave best accuracy of 97.685% for Support Vector Regression.

Keywords: Agriculture, Pre-processing, Analysis, Regression, prediction

I Introduction

The information that crops offer is turned into profitable decisions only when efficiently managed. Current advances in data management are making Smart Farming grow exponentially as data have become the key element in modern agriculture to help producers with critical decision-making. Valuable advantages appear with objective information acquired through sensors with the aim of maximizing productivity and sustainability. This kind of data-based managed farms rely on data that can increase efficiency by avoiding the misuse of resources and the pollution of the environment.

Data-driven agriculture, with the help of robotic solutions incorporating artificial intelligent techniques, sets the grounds for the sustainable agriculture of the future. This paper reviews the current status of advanced farm management systems by revisiting each crucial step and studied various soil, climatic properties which affects the yield of pepper. This model helps the farmers to take better decisions and it reduces the differences between the actual and expected yield. We cannot predict the future yield accurately but using this machine learning techniques we can reduce the differences between the expected yield and the actual yield.

2 Methodology

In this scenario machine learning techniques were used for the prediction.

2.1 Dataset

The data set is collected from the Department of Soil Survey and Soil Conservation Government of Kerala, Planning and Economic Affairs Department Government of Kerala, Kerala agriculture university, Meteorological department Government of India. It is collected by Focus Group. The data studies the Area and the Yield in all the block panchayats in the state of Kerala, India. The whole data is about the pepper.

2.2 Features Used

The dataset has features such as 'District, Blocks, Soil Types, Organic Carbon, Phosphorous, Potassium, Manganese, Boron, Copper, Iron, Sulphur, Zinc, Soil PH, Temperature, Humidity, Precipitation, Crop, Area, Yield. The Data in xlsx format. The dataset has the observations from all 14 districts of Kerala.

2.3 Work Flow

Here for this research work the data is collected from the various government resources of Kerala. The feature involves micro climatic conditions, Soil properties and area of the paddy. Then the data cleaning process which involve detection of missing value, Skewness, outliers and their treatments. Using normalization techniques, the data transformations applied to the dataset. By using the principal component analysis, data reduction performed on the dataset. The data is splits into training and testing. Then after the successive splitting of the data the different regression algorithms such as Linear Regression, and Support Vector Regression algorithms applied in order to predict the pepper yield. Based on the accuracies obtained from different models, the best algorithm is selected for the model deployment. The Accuracies obtained from the different model is compared in order to find the best algorithm for model deployment. For this entire research work the python environment is used. Entire workflow is depicted in Figure 1.

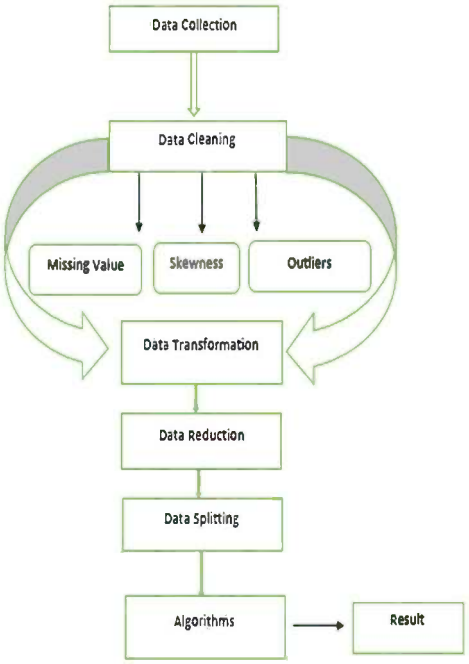

Fig 1: Architecture Diagram

2.4 Pre-processing

2.4.1 Checking Missing Value

The missing values in the data affects the accuracy of the model. The algorithms will not support if data has the missing values in it's the first step of pre-processing is checking null values in the dataset. If there exist, the missing data we need to treat them. There are different deletion and imputation techniques available in order to treat the missing values in the dataset. Here in this pepper-dataset we have checked the missing value using isnull() command in python.

2.4.2 Checking Skewness

Skewness is the degree of asymmetry observed in a probability distribution. The skewness does not detect the outliers but it can give the direction to the outlier in the dataset. Using the skew () command in python we can easily detect the skewness exist in the data. The distplots of the different features in the dataset will give a visualization of distributions in the dataset. Using different transformations, we can treat the skewness in the data. Here we have used the box-cox transformation to treat the skewness.

2.4.3 Data Transformation

Here in this particular research work we have used normalization techniques for data transformations. In this technique the minimum value is converted into zero and the maximum value is converted into one. All other values in the dataset is converted into a decimal between zero and one. The label encoding also performed in the dataset. Here some of the features such as State, District, Blocks, Soil types are in non-numerical form. We need to convert these features into numerical form in order to apply the machine learning algorithms. The LabelEncoder() in python converts the categorical value to the numerical form based on alphabetical order of the text.

2.4.4 Data Reduction

Since the dataset contains many features we can perform dimensionality reduction techniques in this particular dataset. For Dimensionality reduction we have used the Principal Component Analysis(PCA) which transform the data from the high dimension to the lower dimension. This process will not lose much information in the data.it holds the most of the information in the data even after the dimensionality reduction using PCA.

2.5 Data Splitting

This is the step which includes training and testing of input data. The data which is loaded is divided into two sets, such as training and test data, with a division ratio of 80% or 20%, such as 0.8 or 0.2. In the learning phase, a classifier is used to form the available input data. In this step, create the classifier's support data and preconceptions to approximate and classify the function. During the test phase, the data is tested. We can say that the final data is formed during pre-processing and is processed by the machine learning module. The split was made for training and testing of the system. So the process is considered as essential for any supervised machine learning or data science application. Here the train test splitting is taking place randomly.

2.6 Algorithms

2.6.1 Linear Regression

In this linear regression the dataset contains the dependent and independent variables. Here we have a set of input values we can call it as x and from the input values the output values (y) is predicted. In the linear regression we can say that both the input values(x) and output values(y) are numeric values. The equation for the simple linear regression is

$$y = ax + b \qquad \text{eq(1)}$$

y-dependent variable

a-slope

x- independent variable

b-y intercept

The scatter plot helps to determine the relationship between variable. Checking this relationship between the variable is important in the linear model. If there is no increasing or decreasing trends in the scatter plot, then we can say that the fitting of this regression model will not give a useful model.

2.6.2 Support Vector Regression(SVR)

Support vector regression is a part of the Support Vector Machine (SVM). SVM can be applied to both the classification and regression problems. When the SVM is applied to the regression problem then it is called the Support Vector Regression. In SVM, the algorithm finds the hyper plane in an N dimensional space, where N is nothing but the number of dependent variables. Here the concept of marginal plane and hyper plane exists. The marginal lines are chosen so that they cover all the data or allow some violation. The marginal lines which cover all the data is called hard margin and the marginal line which allows for some violations are called soft margin. We can say that the support vector regression is sensitive to the outliers because the margin includes the outliers. If we consider the soft margin in SVR then it will be similar to the linear regression. Another advantage of the support vector regression is it has the ability to incorporate non linearity. So we can say that this algorithm is one of the powerful regression algorithm.

3.Results and Discussions

3.1 Pre-processing results

```
1 data1.isnull().sum()
```

```
District                    0
Bloacks                     0
Soil Types                  0
Organic Carbon(%)           0
Phosphorous(Kg/Ha)          0
Pottassium(Kg/Ha)           0
Manganese(ppm)              0
Boron(ppm)                  0
Copper(ppm)                 0
Iron(ppm)                   0
Sulphur(ppm)                0
Zinc(ppm)                   0
Soil PH                     0
Temparature(deg.celsius)    0
Humidity                    0
Precipitation(in)           0
Crop                        0
Area(Ha)                    0
Yield(Tones)                0
```

Fig 2: Checking Missing values

It is found that there are no missing values present in the data and therefore no need to treat the missing data here (Figure 2). Similarly the skewness in the data is checked using skew() function in python and the result is given in Figure 3.

```
1 data_num.skew()
```

```
Organic Carbon(%)           1.831031
Phosphorous(Kg/Ha)         12.295663
Pottassium(Kg/Ha)           0.752533
Manganese(ppm)              2.416460
Boron(ppm)                  4.406709
Copper(ppm)                 2.755706
Iron(ppm)                   2.131298
Sulphur(ppm)                4.066535
Zinc(ppm)                   2.417999
Soil PH                    12.328762
Temparature(deg.celsius)   -0.562687
Humidity                   -1.148633
Precipitation(in)          -1.225586
Area(Ha)                    3.629138
Yield(Tones)                3.665552
```

Fig 3: Checking The skewness in data

It shows that there exists skewness in the data. So we used box-cox transformation to treat this skewness. The resultant skewness after the transformation is given in Figure 4.

```
Organic Carbon(%)           0.092932
Phosphorous(Kg/Ha)         -0.396601
Pottassium(Kg/Ha)          -0.013161
Manganese(ppm)              0.010392
Boron(ppm)                  0.124288
Copper(ppm)                 0.000970
Iron(ppm)                   0.034329
Sulphur(ppm)                0.001081
Zinc(ppm)                   0.002723
Soil PH                    -0.098701
Temparature(deg.celsius)   -0.021672
Humidity                   -0.003643
Precipitation(in)          -0.342416
Area(Ha)                   -0.049400
Yield(Tones)               -0.022793
```

Fig 4: After box-cox transformation

The transformation sucessfully coverts the data.We have also plotted the boxplots to detect the outliers (Figure 5).

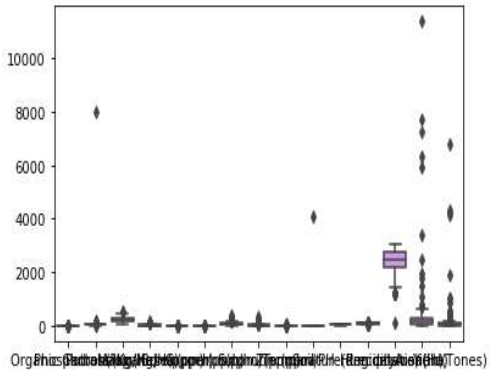

Fig 5: Boxplot to detect outliers

It is found that there exist outliers in the data. There is lesser number of outliers present in the data so we have used winsorizing method here in order to treat the outliers in the dataset. The resultant boxplot after the treatment is shown in Figure 6.

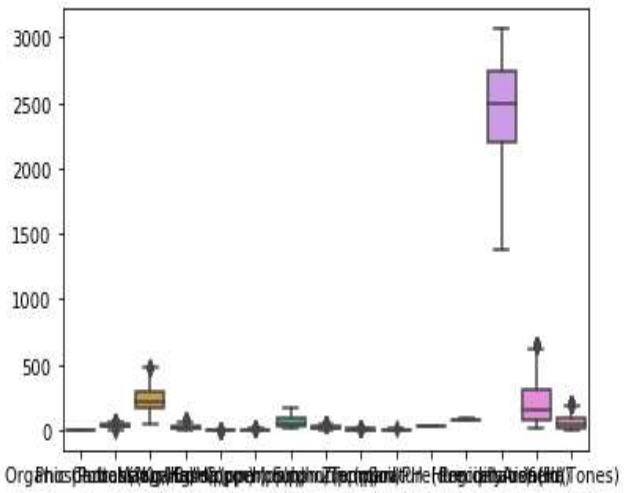

Fig 6: After Winsorizing treatment

Data is normalized using the normalization techniques and the result is given in Figure 7.

Organic Carbon(%)	Phosphorous(Kg/Ha)	Pottassium(Kg/Ha)	Manganese(ppm)	Boron(ppm)	Copper(ppm)	Iron(ppm)	Sulphur(ppm)	Zinc(ppm)	Soil PH	Ten
0.003066	0.200961	0.670190	0.045730	0.000236	0.001061	0.033468	0.003891	0.003402	0.007647	
0.000339	0.006061	0.054538	0.004451	0.000355	0.000505	0.004023	0.014048	0.000883	0.001786	
0.000766	0.002111	0.021881	0.004204	0.001068	0.000212	0.004606	0.012193	0.000593	0.001694	
0.000742	0.002044	0.021193	0.004071	0.001034	0.000205	0.004461	0.011810	0.000575	0.001640	
0.000100	0.000732	0.017767	0.001857	0.000285	0.000205	0.005491	0.008408	0.000203	0.827748	

Fig 7: Normalization

The data reduction is performed by using the principal component analysis and the result is given in Figure 8.

	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5
0	0.899864	0.365897	0.495852	0.152094	-0.093563
1	0.222823	-0.128194	-0.039755	-0.036860	0.024319
2	-0.023183	-0.060702	-0.073519	-0.030083	-0.004752
3	0.101289	-0.108687	-0.071432	-0.041470	0.031676
4	0.170790	0.402146	-0.423405	0.664804	0.133976
5	-0.098704	-0.020138	-0.078558	-0.028034	-0.014666
6	-0.050645	-0.030523	0.004637	-0.002792	0.033456
7	0.018428	-0.061904	0.005171	-0.010725	0.062198
8	-0.014643	-0.047182	0.005156	-0.006839	0.045473

Fig 8: Data reduction using PCA

3.2 *Visualization results*

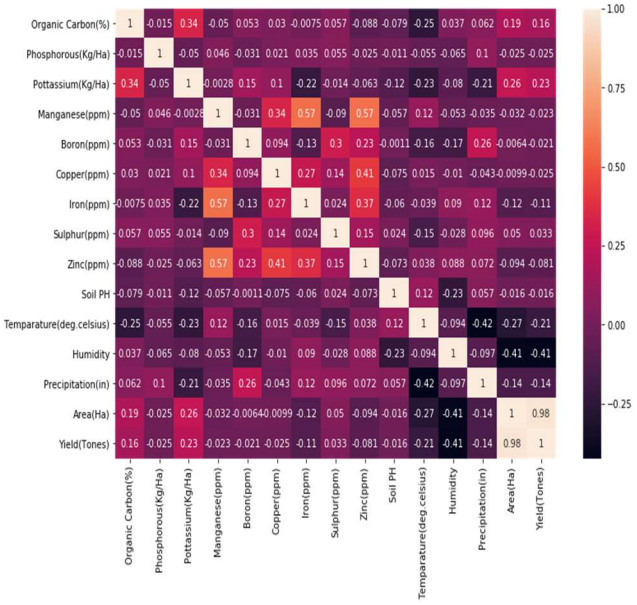

Fig 9: Heat map

The heat map indicates the correlation between the variables (Figure 9). Some of the variables are positively correlated and some them are negatively correlated. And the area has a high correlation with the yield.

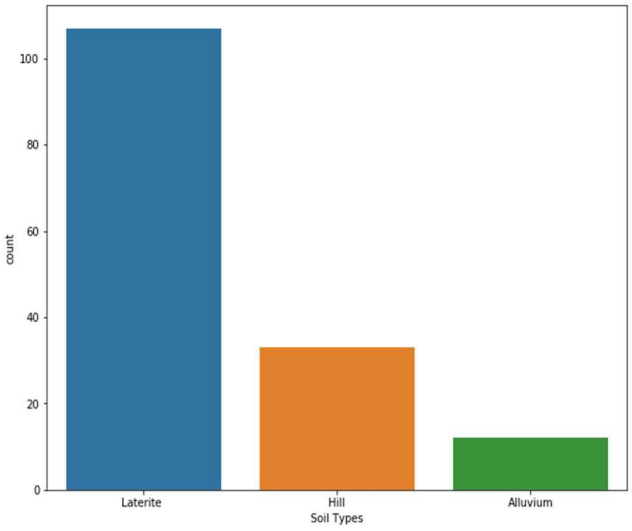

Fig 10: Bar chart of soil types

Figure 10 shows that the Laterite soil is present in most of the places in Kerala and most of the crop can be cultivated in this type of soil. The contents in this soil supports all the crops not only pepper but also other crops in Kerala.

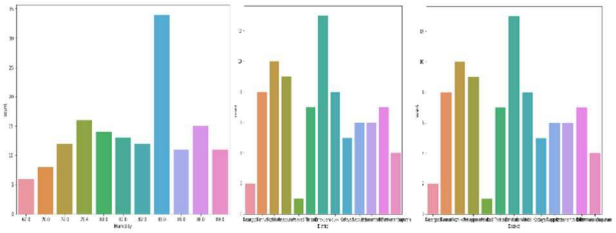

Fig 11: Plot of Humidity, precipitation, Temperature

From Figure 11 we have 83.0 as the average Humidity which occurs in Kerala when comparing with the Humidity's in all the districts in Kerala. 2500(mm) is the most occurring precipitation in entire Kerala. It is also found that 26.2-degree Celsius temperature is the most occurring temperature in Kerala which is suitable for pepper.

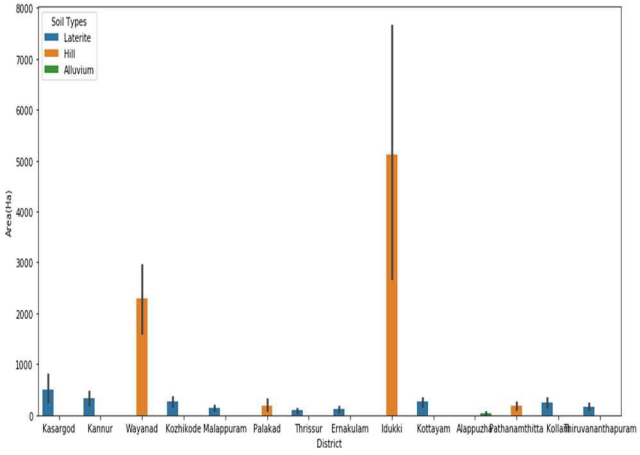

Fig 12: Bar chart showing district vs area for pepper

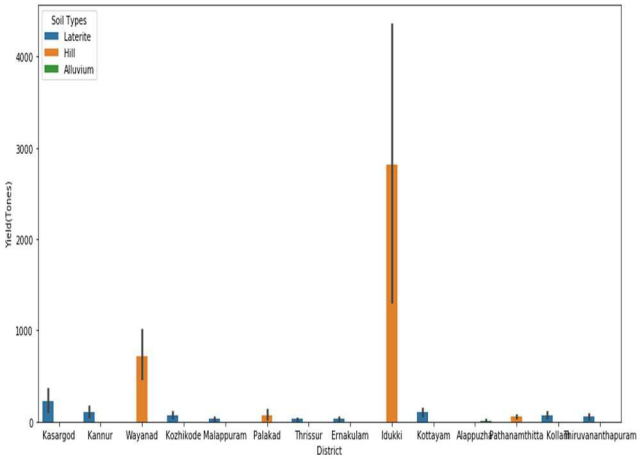

Fig 13: Bar chart- district vs yield for pepper

From Figure 12, it is clear that in Idukki and Wayanad has more lands used for pepper crop. In Idukki more than 5000 hector is used for pepper and in Wayanad more than 2000 hector used for pepper crop. Alappuzha give very less yield for pepper. From the Figure 13, we can observe that pepper is giving more yield with Hill soil. The alluvium soil is not that much suitable for the pepper.

Table 1: District vs Area of pepper

District	Area(Ha)
Kasargod	2991.9
Kannur	3643.8
Kozhikode	3245.96
Wayanad	9143.99
Malappuram	2106.19
Palakad	2543.38
Thrissur	1517.12
Ernakulam	1674.87
Idukki	40966.63
Kottayam	2852.74
Alappuzha	564.14
Pathanamthitta	1481.51
Kollam	2693.57
Thiruvanathapuram	1823.43

Table 2: District vs Yield of pepper

District	Yield(Tonnes)
Kasargod	1383.189
Kannur	1164.296
Kozhikode	893.623
Wayanad	2881.936
Malappuram	440.475
Palakad	842.041
Thrissur	442.089
Ernakulam	409.308
Idukki	22532.291
Kottayam	1160.716
Alappuzha	110.67
Pathanamthitta	467.483
Kollam	809.416
Thiruvanathapuram	619.655

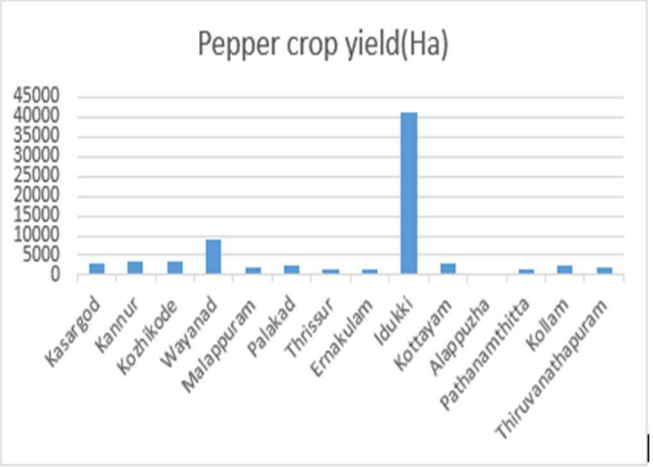

Fig 14: Bar chart for district Vs Yield

From Table 1 and 2, it is found that in the district Idukki the production of pepper is high.When analysing the trends it is found that the iduki has cultivated the pepper in a larger area and a high quantity of the yield obtained from this district (Table 2).The climatic conditions and the soil properties in this district is suitable for the pepper.So this was the trend in area and yield for pepper in the year 2018-2019 in Kerala.A total of 77249.23 hector pepper was cultivated in 152 block panchayats in Kerala and obtained a total of 34157.188 tonnes of pepper was obtained in the area in the year 2018-2019 (Figure 14). Based on this past data and present climatic, soil properties we can predict the yield of the pepper using machine learning model.

3.3Algorithm results

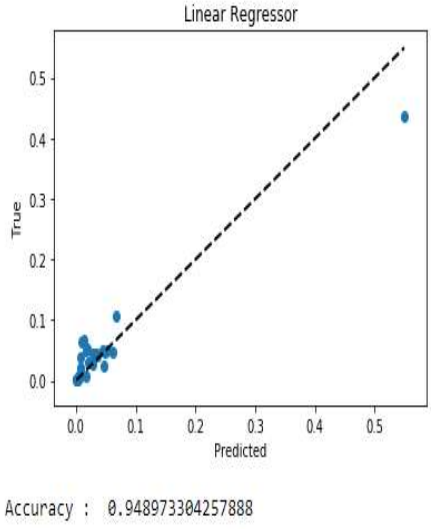

Fig 15: Accuracy plot of Linear Regression

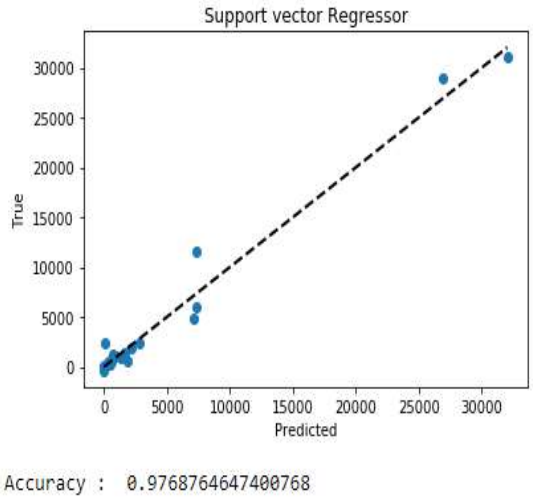

Fig 16: Accuracy plot of SVR

The linear regression is working with 94.89% accuracy and the support vector regression working with 97.68% accuracy (Figure 15 and 16). Therefore, we can say that here the support vector regression is the best algorithm in order to predict the yield of pepper.

4.Conclusion

Based on the past and present micro-climatic and soil properties available, we have tried machine learning techniques to predict the future pepper yield. To predict the yield of pepper, machine learning algorithms were tried out in this research work. It was found that Support Vector Regression gave the best accuracy of 97.68%. Hence, this algorithm can be used for model deployment.

5. Acknowledgement

The authors extend their appreciation to the Deputyship of RESEARCH AND INNOVATION wing of TECHTERN Pvt. Ltd. through SMART-AGRO research and for providing all support for this research work with the project number TTRD-DS-05-2021. This is also an extension of a post-doctoral research program under Kannur University.

References

[1] Hegde, Niranjan G., et al. "Survey paper on agriculture yield prediction tool using machine learning." *Int. J* 5 (2017): 36-39.

[2] Priya, P., U. Muthaiah, and M. Balamurugan. "Predicting yield of the crop using machine learning algorithm." *International Journal of Engineering Sciences & Research*

Technology 7.1 (2018): 1-7.

[3] Hassan, Fazal Mahmud, et al. *Agricultural yield and profit prediction using data analysis techniques*. Diss. 2018.

[4] Josephine, B. Manjula, et al. "Crop Yield Prediction Using Machine Learning." *International Journal Of Scientific & Technology Research* 9.02 (2020).

[5] Kumar, Arun, Naveen Kumar, and Vishal Vats. "Efficient crop yield prediction using machine learning algorithms." *International Research Journal of Engineering and Technology* 5.06 (2018): 3151-3159.

[6] Bondre, Devdatta A., and Santosh Mahagaonkar. "Prediction of crop yield and fertilizer recommendation using machine learning algorithms." *International Journal of Engineering Applied Sciences and Technology* 4.5 (2019): 371-376.