Computational Model for Pepper Yield Prediction Using Support Vector Regression

Akhil Wilson[1], RajiSukumar[2], Hemalatha N[3]

Lulu Financial Group, India [1], Techtern Pvt Ltd, Kannur[2], St Aloysius College of Management and Information Technology, India[3]

Corresponding author: Hemalatha N, Email: hemalatha@staloysius.ac.in

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 percent for Support Vector Regression.

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

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 779–788. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-73

1. Introduction

The information that crops offer is turned into profitable decisions only when efficiently managed [1]. 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 [2-5]. 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 technique 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 and Yield. The Data in xlsx format. The dataset has the observations from all 14 districts of Kerala.

2.3 Workflow

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 involves detection of missing values, Skewness, outliers and their treatments. Using normalization techniques, the data transformations is performed on the dataset. Further, the data is split 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, python environment is used. Entire workflow is depicted in Fig. 1.

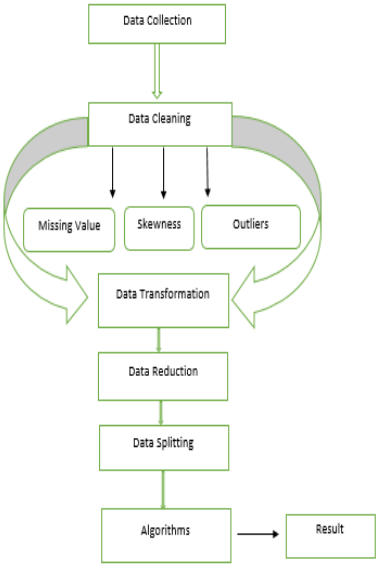

Fig 1. Architectural diagram.

2.4 Data Pre-processing

(i) 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. The first step of pre-processing is checking null values in the dataset. If missing data exists, then 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.

(ii) 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 existing 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.

(iii) 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 is 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.

(iv) 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 transforms the data from 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.

(v) 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 train 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.

3. Algorithms

3.1 Linear Regression: In this algorithm the dataset contains the dependent and independent variables. Here we have a set of input values which can be called 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. Equation for the simple linear regression is

$$y=ax+b \text{ `} \quad (1)$$

where, y-dependent variable, a-slope, x- independent variable and 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.

3.2 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.

4. Results and Discussion

In the present work, data pre-processing was carried out with respect to the data., results of which are discussed in the sub-sections. Visualization results are depicted in another sub-section and finally algorithmic results in last section.

4.1 Pre-processing Results

Data was tested for missing values, skewness, outliers and dimension reduction.

```
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 was found that there are no missing values present in the data and therefore no need to treat the missing data (Fig. 2). Similarly the skewness in the data was checked using skew() function in python and the result given in Fig. 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 skewness in data

Fig. 3 shows that there exists skewness in the data. So here we have used box-cox transformation to treat this skewness. The resultant skewness after the transformation is given in Fig. 4. The transformation sucessfully converts the data.

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

We have also plotted the boxplots to detect the outliers (Fig. 5). It was found that there existed outliers in the data. Outliers present in the data was less in number so Winsorizing method was used to treat the outliers in the dataset. The resultant boxplot after the treatment is shown in Fig. 6.

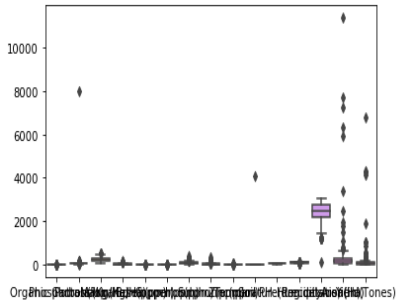

Fig 5. Boxplot to detect outliers

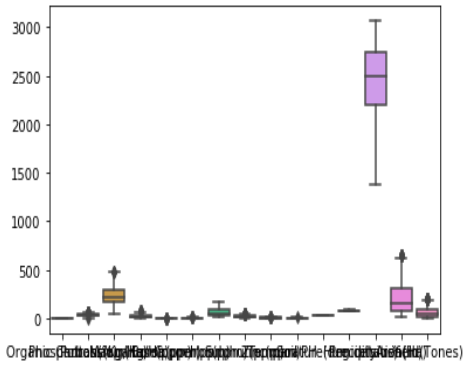

Fig 6. After Winsorizing treatment

Data is normalized using the normalization techniques and the result is given in Table 1.

Table 1. Table showing Normalization.

Organic carbon	Phosphorus	Pottassium	Manganese	Boron	Copper	Iron	Sulphur	Zinc	Soil PH
0.00306	0.200961	0.67019	0.045	0.0002	0.0010	0.03	0.003	0.003	0.007
0.00033	0.006061	0.0545	0.00445	0.0003	0.0003	0.004	0.014	0	0.001
0.00076	0.002111	0.021881	0.004204	0.0010	0.0002	0.004	0.012	0.0005	0.001
0.00074	0.002044	0.02119	0.0040	0.0010	0.0002	0.004	0.011	0.0005	0.0016
0.00010	0.000732	0.01776	0.001857	0.0002	0.0002	0.0041	0.008	0.0002	0.827

The data reduction is performed by using the principal component analysis and the result is given in Fig. 7.

Table 2. Data reduction using PCA.

	Principal component 1	Principal component 2	Principal component 3	Principal component 4	Principal component 5
0	0.8999	0.3658	0.4958	0.1520	-0.0935
1	0.2228	-0.1282	-0.039	-0.03687	0.02431
2	-0.0232	-0.0607	-0.0735	-0.0300	-0.0047
3	0.1013	-0.1086	-0.0715	-0.0414	0.0316
4	0.1707	0.4021	-0.4234	0.6648	0.1339
5	-0.0987	-0.0201	-0.0785	-0.0280	-0.0146
6	-0.0506	-0.0305	0.0046	-0.0027	0.0334
7	0.0185	-0.0619	0.0052	-0.0107	0.0621
8	-0.0146	-0.0471	0.0052	-0.0068	0.0454

5. Visualization Results

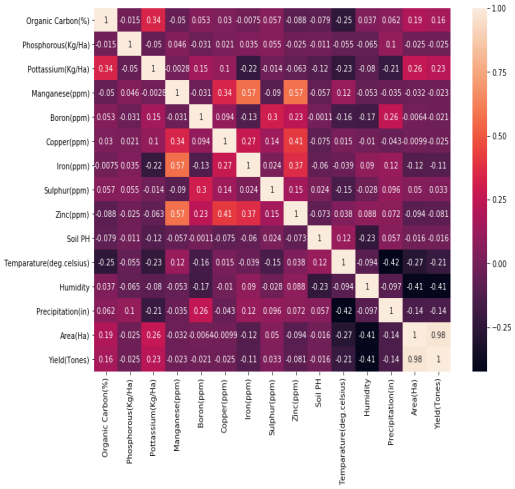

Fig 8. Heat Map

The heat map indicates the correlation between the variables (Fig. 8). Some of the variables are positively correlated and some are negatively correlated. And the area has a high correlation with the yield.

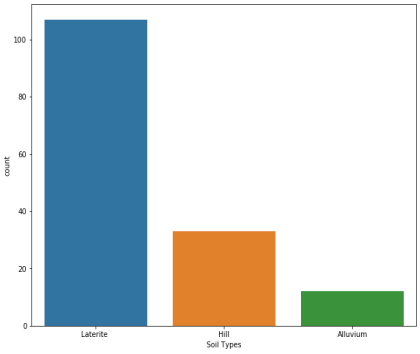

Fig 9. Bar chart of soil types

Fig 9 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.

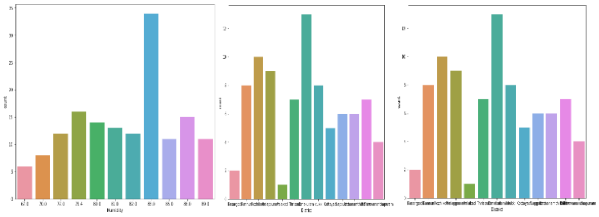

Fig 10. Plot of Humidity, precipitation, Temperature

From Fig. 10 we have 83.0 as the average Humidity which occurs in Kerala when compared with the Humidity's in all the other districts in Kerala. 2500(mm) is the most occurring precipitation and 26.2-degree Celsius temperature is the most occurring temperature in Kerala which is suitable for pepper.

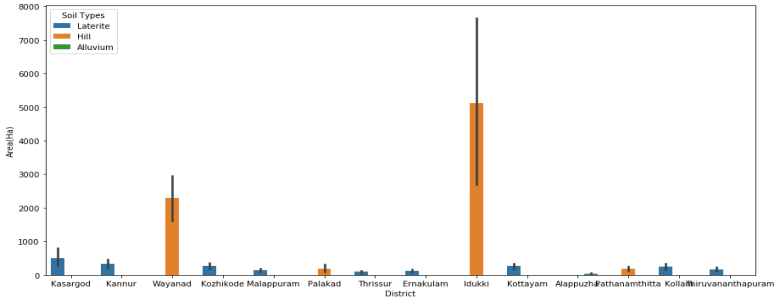

Fig 11. Bar chart showing district vs area for pepper

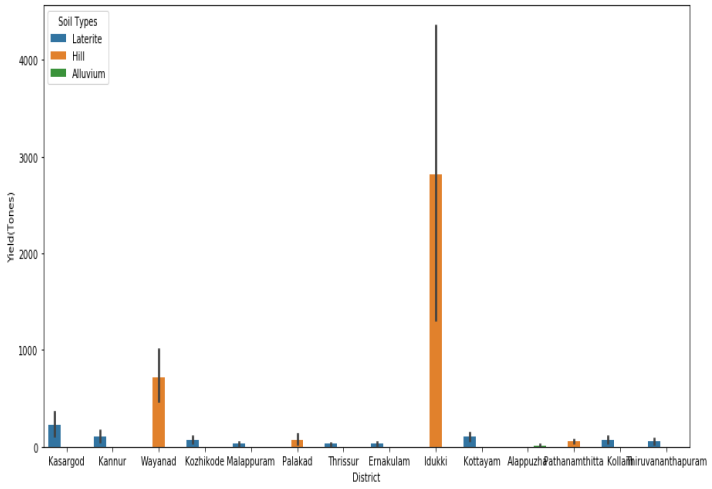

Fig 12. Bar chart showing district vs yield for pepper

From Fig. 11, it is clear that Idukki and Wayanadhas 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 Fig. 12, we can observe that pepper is giving more yield with Hill soil. Alluvium soil is not suitable at all for production of pepper.

Table 1. District vs Area & Yield of pepper

District	Area (Ha)	Yield (Tons)
Kasargod	2991.9	1383.2
Kannur	3643.8	1164.3
Kozhikode	3245.9	893.7
Wayanad	9143.9	2881.9
Malappuram	2106.1	440.5
Palakkad	2543.4	842.04
Thrissur	1517.1	442.1
Ernakulam	1674.8	409.3

Idukki	40966.6	22532.3
Kottayam	2852.7	1160.7
Alappuzha	564.1	110.7
Pathanamthitta	1481.5	467.5
Kollam	2693.5	809.4
Thiruvananthapuram	1823.4	619.7

From Table 3, it was found that in the district Idukki the production of pepper is high.When analysing the trends it was found that the Idukki has cultivated the pepper in a larger area and a high quantity of the yield was obtained from this district (Table 3).The climatic conditions and the soil properties in this district are suitable for the pepper. From the Table we have, a total of 77249.23 hector pepper cultivated in 152 block panchayats in Kerala and a total of 34157.188 tons of pepper obtained from the cultivated area. Based on the past data and present climatic, soil properties we can predict the yield of pepper using machine learning regression model.

6. Algorithm Results

In the present work, machine learning regression model *viz.*,Linear regression and Support Vector regression were used and were trained using the data. Trained model was later validated with test data. Linear regression model has an accuracy of 94.89% and support vector regression working with 97.68% accuracy (Fig. 13 and 14). Therefore, we can conclude that Support Vector regression is the best algorithm in order to predict the yield of pepper.

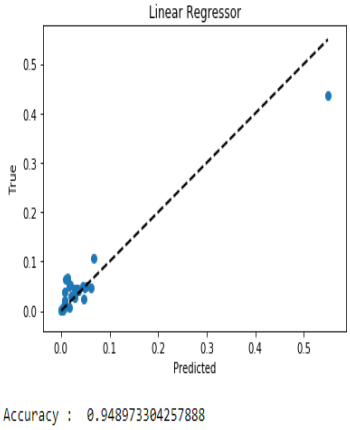

Fig 13. Accuracy plot of Linear Regression

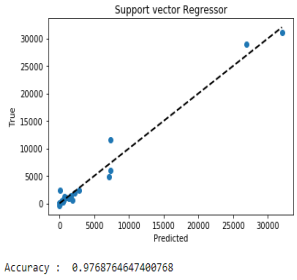

Fig 14. Accuracy plot of Support Vector Regression

7. Conclusions

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 regression 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. In our future work, would like to further enhance the model with more data and also develop a web server for general use.

8. 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., (2017). Survey paper on agriculture yield prediction tool using machine learning. International Journal of Advance Research in Computer Science and Management Studies 5:36-39.

[2] Priya, P., Muthaiah,U., and Balamurugan, M. (2018). Predicting yield of the crop using machine learning algorithm. International Journal of Engineering Sciences & Research Technology 7(1):1-7.

[3] Josephine, B. Manjula, et al. (2020). Crop Yield Prediction Using Machine Learning. International Journal of Scientific & Technology Research 9(2):2102-2106.

[4] Kumar, Arun, Naveen Kumar, and Vishal Vats. (2018). Efficient crop yield prediction using machine learning algorithms. International Research Journal of Engineering and Technology 5(6):3151-3159.

[5] Bondre, Devdatta A., and SantoshMahagaonkar. (2019). Prediction of crop yield and fertilizer recommendation using machine learning algorithms. International Journal of Engineering Applied Sciences and Technology 4(5): 371-376.