Crop Yield Prediction through different Machine Learning Algorithm

*

1st Tanmay Jain
Computer Science
Bennett University
Greater Noida, India
tanmayjain84@gmail.com

2nd Vasanthagokul S
Computer Science and Engineering
Sri Ramakrishna Engineering College
Coimbatore, India
vasanth.1801247@srec.ac.in

3rd Shaik Sazid
Computer Science)
SRKIT Engineering College
Vijayawada, India
sazid.sunny@gmail.com

4th Arsh Srivastava
Mechanical Engineering
NIT Silchar
Silchar, India
arshsrvp2006@gmail.com

5th Anjali Yadav
Computer Science And Engineering
Bennett University
Greater Noida, India
e19soe807@bennett.edu.in

***Abstract*—Precise yield estimation is crucial in agriculture. Remote Sensing (RS) frameworks are extra commonly applied in constructing choice help devices for modern cultivating frameworks to enhance yield even as diminishing working prices and natural effect. The number of crops inside the area is one of the principle additives for determining crop yield. This counting task is to be done carried out using a human rather than a computer and is hence time-consuming In this paper, we propose an green approach that uses computer vision and precisely count the vegetation in a digital photograph, in addition to we've also devised algorithms which could correctly determine the yield on the idea of given elements along with soil precipitation, Area, humidity Index and greater such elements which performs a crucial position in figuring out the yield. We have used various algorithms such as random forest regression, Support Vector Machine, CNN and Deep Neural network in this paper and worked on the above problem.**

***Index Terms*—Yield Prediction; object counting; computer vision; deep learning; convolution neural network; deep neural networks; regression; SVM;**

I. Introduction

Farming is a major source of income for many people in developing countries. In addition, agricultural growth has been more rapid than growth in the non-agricultural sectors in recent years in many countries. The two types of index products are parametric and sample-based. Examples of parametric indices in insurance include weather (with triggers based on variables such as rainfall, temperature, humidity, wind speed, etc.), flooding (water levels and durations triggers), wind speed (velocity and duration triggers) and seismic activity (Richter scale triggers).Test-based records incorporate territory based yield protection and test-based domesticated animals file protection. Zone yield protection is basically a put choice on the normal yield for a creation in a locale/region. On the off chance that the zone is adequately huge, region yield protection isn't powerless to moral danger issues, since the activities of an individual rancher will have no observable effect on the zone normal yield. Region yield protection likewise has moderately low exchange costs since there is no compelling reason to set up and check explicit ranch yields for each safeguarded unit nor is there any need to direct on-ranch misfortune alteration.

Bennett University

II. Data-set Used

Three types of datasets are used in this paper two of which are in the csv format and one is the image data.

A. Data-set Description and Source

First the data-set is taken from the website of Government of India which consist of following fields area, production, state and season. This data consists of data of Andhra Pradesh state for 10 years. Second data-set was taken from GitHub[1] which consisted of more number of features such as Humidity, Precipitation, NDVI, Pressure, Temperature,Wind Speed etc. We have two years of Wheat data. These data are relocated to specific latitudes and longitudes.

Columns 1-5 in the file provide information on location and time. After columns 5 are raw features, like NDVI or wind speed. Day in Season is a calculated feature defining how many days since the start date of the season have occurred. The yield is the label, the value that should be predicted. The County Name, State, and Date are excluded from training as this will result in over-fitting.. Third data-set consisted of the images of paddy plants taken from the UAV(Drone) which consisted of ten images. [3]

III. Research Methodology

As this is a regression problem, basically we have applied three different algorithms namely Support Vector Regression

III-B, Random Forest III-A, Deep Neural Network III-C. In the third data-set which consists of image data we have used Convolution Neural Network III-D

Data Preprocessing is done on the CSV data.The Categorical data is encoded through one-hot encoding. Then the data is normalized through Min-Max scalar and the missing values in the particular column are replaced with the mean of that column. Python pandas library is used for pre-processing. At last the data is splitted into train(0.7) and test(0.3) before applying the different models.

Data preprocessing is done on the image dataset. Since there was no available dataset to download, we had to create our own dataet by hand labelling number of crops per image. We made 10 such images for use which were all different from each other. We had created a csv file with the image ID and its respecive number of crops in the corresponding column. We normalised the varying image sizes into a 400x400 pixel size so that feeding the images into the model is easy. We split the images into 8 images for training and 2 images for testing. [2]

A. Random Forest Regression

The random forest model is a sort of added substance model that settles on forecasts by joining choices from a succession of base models.

$$g(x) = f0(x) + f1(x) + f2(x) + ... \quad (1)$$

where the last model g is the sum of basic base models fi. Here, each base classifier is a decision tree. This expansive procedure of utilizing different models to acquire better prescient execution is called model ensembling. In arbitrary backwoods, all the base models are built autonomously utilizing an alternate sub-test of the information.

We have imported RandonForestRegressor from sklearn library and computed the given scores for comparison. [3]

B. Support Vector Regression

A support vector machine builds hyperplanes in a high or limitless dimensional space, which can be utilized for arrangement, relapse, or different assignments like exceptions discovery.

The Support Vector Regression (SVR) utilizes indistinguishable standards from the SVM for order, with just a couple of minor contrasts. On account of relapse, an edge of resistance (epsilon) is set in estimate to the SVM which would have just mentioned from the issue. However, other than this reality, there is likewise a progressively entangled explanation, the calculation is increasingly confused along these lines to be taken in thought. Notwithstanding, the fundamental thought is consistently the equivalent: to limit mistake, individualizing the hyperplane which augments the edge, remembering that piece of the blunder is endured. We have imported Support Vector Regressor from sklearn library and computed the given scores for comparison. [3]

C. Deep Neural Network

A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. We had imported a sequential neural network from keras for designing our own model. We had used multiple layers of dense fully connected layers so that accuracy is high. We have trained the 2 different neural networks on the 2 datasets respectively for 50 epochs.

For the first Neural Network for the first dataset, it consisted of 1 dense input layer with relu activation function, 3 hidden layers with relu activation function and the output layer was a dense layer with one output node with linear activation function. The network used for the second dataset had a total of 5 dense hidden layers in place of the 3 used earlier. [5]

D. Convolution Neural Network

A convolution neural network (CNN) is a form of artificial neural network that is specifically designed to process pixel data for image recognition and processing. In image processing, the Convolution Neural Network is Robust. And CNN is also Artificial Intelligence, which uses Deep Learning to perform both generative and descriptive tasks. [4]

In this analysis we had to build a convolution neural network which regresses and gives a linear output that we can use as the problem that we are solving is not a classification problem but rather a regression model. [2]b1

In out CNN-Regression model, we have one input layer which accepts an image of 400x400 size, 1 convolution 2d layers with a 3x3 filter and a max-pooling 2d layer with stride 2x2, 2 convolution 2d layers with a 3x3 filter and a max-pooling 2d layer with stride 2x2, 2 convolution 2d layers with a 3x3 filter and a max-pooling 2d layer with stride 2x2, a flatten layer and 3 fully connected dense layers and finally an output layer which has 1 output with linear activation function [3]

E. Performance Evaluation

[3] The performance was analysed on the basis of various factors such as:

- Mean Absolute Error:Absolute Error is the amount of error in your measurements. It is the difference between the measured value and "true" value. The Goal is to minimize the MSE.
- Mean Squared Error: mean squared error (MSE) measures the average of the squares of the errors that is, the average of the squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The Goal is to minimize the MSE.
- R2-Score:It is (total variance explained by model) / total variance. So if it is 100, the two variables are perfectly correlated, i.e., with no variance at all. A low value would show a low level of correlation, meaning a regression model that is not valid, but not in all cases. Lower the R2 score the more accurate the model will be, But it is not accurate in all cases.

Mean Absolute Error is taken as the base for comparing the different models.

F. Results

All the parameters given in III-E are calculated for the comparison between different models. All the results given below are calculated for the second data-set In case of Random Forest regression the results are as follows:

- Mean Squared Error: 64.986
- Mean Absolute Error: 6.160
- R2 Score: -0.355

In case of Support Vector regression the results are as follows:

- Mean Squared Error: 76.452
- Mean Absolute Error: 6.869
- R2 Score: -0.912

In case of Deep Neural Network the results are as follows:

- Mean Absolute Error: 2.162
- loss: 11.148
- Validation Absolute Error: 2.049
- Validation loss: 10.961

In case of the image data-set CNN is applied the results are:

- Mean Absolute Error: 9.974
- loss: 145.040
- Validation Absolute Error: 13.771
- Validation loss: 221.351

All the given results can be visualized in section III-H

G. Conclusion

From the results given in section III-F we can conclude that Deep Neural Networks is preferably the best model for predicting the crop yield as it has the lowest Mean Absolute error. The same can be visualized from the graphs given in section III-H

In case of counting the crops from the images the accuracy is not very good due to the unavailability of sufficient data-set. as we had only 10 images but in feature the accuracy can be improved and the model can be made more accurate. The graph relating the MAE and epochs can be seen in section III-H

H. Figures and Tables

a) Neural Network: We can see from the given figures as the epochs in the case of Neural Networks increases the error decreases and at the end we are achieving the model with least MAE and max Accuracy.

"Fig. 3",. Represent the deep neural network

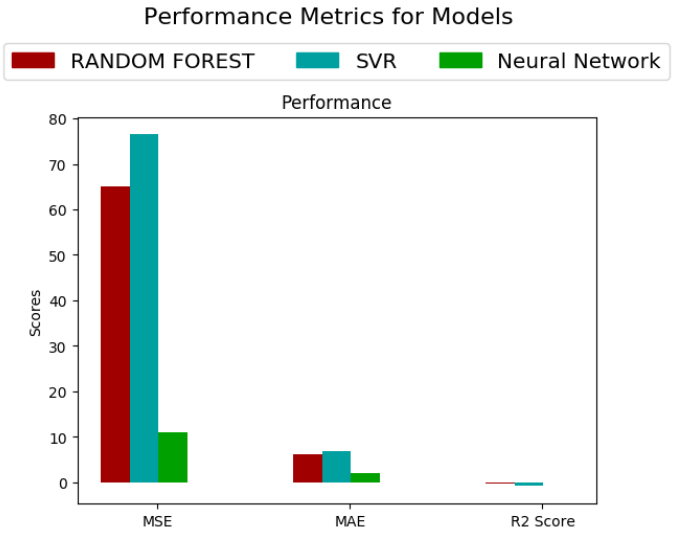

Fig. 1. MAE vs Epoch : CNN Regression

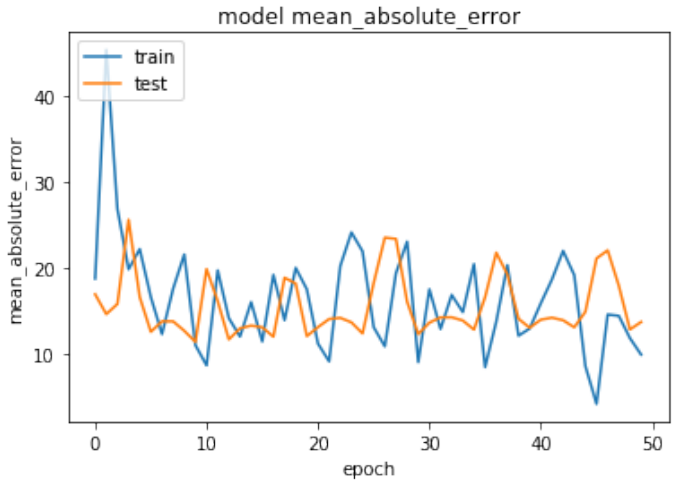

Fig. 2. MAE vs Epoch : CNN Regression

REFERENCES

[1] Yuan, Jun Ni, Bingbing Kassim, Ashraf. (2014). Half-CNN: A General Framework for Whole-Image Regression.

[2] Keras, Regression, and CNNs, Adrian Rosebrock, January 28 2019, unpublished

[3] Choudhury, A. Jones, J.. (2014). Crop yield prediction using time series models. Journal of Economics and Economic Education Research. 15. 53-68.

[4] Alaslani, Maram Elrefaei, Lamiaa. (2019). Transfer Learning with Convolutional Neural Networks for IRIS Recognition. International Journal of Artificial Intelligence Applications. 10. 49-66. 10.5121/ijaia.2019.10505.

[5] Training Recurrent Neural Networks, Ilya Sutskever, PhD Thesis, 2012.

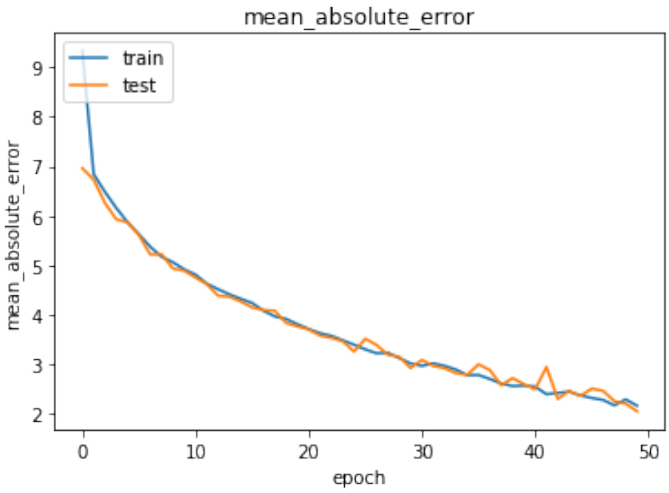

Fig. 3. MAE vs Epoch : RNN