Grenze International Journal of Engineering and Technology, Jan Issue

Precision Agriculture with Crop Yield Prediction

K.Srikanksha Manaswini[1], A.Jhansi Swetha[2], K.Jahnavi[3] and G.Kalyani[*4]
[1-4]Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India
Email: {manaswinikopparthy, alapatijhansiswetha2002, jahnavikadavakollu, kalyanichandrak } @gmail.com

***Abstract*—Farming is the foundation of any country's monetary design and has a basic impact on the Indian economy. Farmers can utilize innovation to help them in figuring out what to cultivate and how to cultivate. It helps farmers to choose the correct crop and increase productivity. Farmers can make better marketing decisions if they have an accurate and high. confidence forecast of their annual production, allowing them to sell their products at the best possible price. Hence in this paper, we are building a model for the prediction of crop yield using machine learning algorithms. Crop yield prediction depends on different factors like genotype, natural factors, etc., due to which it is a challenging task. Our application deals with regression analysis to forecast the value of the dependent variable. The dataset we used consists of four sets of data like yield performance, management data, weather data, and soil data. We have used four types of regressors namely Multilinear Regression, Logistic Regression, Support Vector Regressor, and Artificial Neural Network Regressor to predict the yield. We have considered the Mean Absolute Error as an evaluation metric for our project and compared all the models with respect to it. The best result is obtained by the ANN Regressor algorithm.**

***Index Terms*— Agriculture, Crop Yield Prediction, Machine learning, Regressors, Mean Absolute Error.**

I. INTRODUCTION

Agricultural improvement is essential to offer needs for a developing non-agricultural labor force, raw substances for commercial production, and tax revenue to help the rest of the economy's improvement, and provide an increasing marketplace for domestic manufacturing. Over the past agriculture, one of the most concerns for farmers has been how to rise crop productivity. What are the most effective methods for increasing crop yield per acre? What factors have the greatest impact on crop yield?

Machine learning allows computers to learn without having to be explicitly programmed. Machine learning and deep learning techniques can be used to implement various applications such as Crop Recommendation [1], Fertilizer Recommendation [2], Soil fertility prediction [3] and Disease detection [4],[5]. Crop yield refers to the number of grains produced by a specific land piece. It is measured in kilograms per hectare or bushels per acre. The typical harvest yield per section of land is utilized to evaluate a rancher's rural result on a specific field during a given period.

Crop yield prediction depends on multiple features like natural factors, crop genotype, management practices, and their interactions. So machine learning provides a platform to implement crop yield prediction by using various algorithms. So based on soil and atmospheric conditions predicting yield from the crop is our goal. The agricultural yield can be predicted using our application. The crop yield will be predicted using a machine learning algorithm based on the input data provided. To empower the application to foresee, it is to be trained in

Grenze ID: 01.GIJET.9.1.96

light of the past information about the harvest yield. Subsequent to training the application forecasts can be made.

The main objectives of our work include studying the different parameters which affect the yield of the crop, extracting the important features, developing a model for yield prediction, evaluating the model, and comparing the model with other models. Yield prediction is very useful for farmers who need to make quick decisions. Farmers can use a precise harvest creation forecast model to assist them with sorting out what to cultivate and accordingly when to cultivate.

II. Review of Literature

In [6], Convolutional neural networks and recurrent neural networks are used in this study to develop a deep learning model for forecasting yields based on environmental data and management strategies. This study looked at four different categories of data: management, weather, yield performance, and soil. During each of the three validation years, the CNN RNN model significantly outperforms the RF, DFNN, and LASSO models in every metric.

In [7] Crop genotype, yield performance, and environmental data were used in the training. They used the difference between their two trained deep neural networks—one for yield and another for check yield—to forecast the yield difference. This model greatly outperforms other common methods such as SNN(Shallow neural networks), Lasso, and regression tree, according to their computational results The results showed that environmental factors had a bigger influence on crop production than genotype.

In [8] In this study, they have evaluated the density of corn kernels in a picture of corn ears and used that estimate to predict the frequency of kernels. For feature extraction, they used the VGG-16 Convolutional Neural Network, which integrates feature maps from different network scales to make it resistant to changes in image scale. In order to generate pseudo-density maps and train the model, they also used a semi-supervised technique.

In [9] The aim of this work is to forecast the average crop production of two distinct crops, namely corn and soybean using a series of satellite photos taken before to harvest. They introduced YieldNet, a new convolutional neural network framework that uses a novel deep learning architecture to predict corn and soybean yields. According to numerical results, the suggested technique correctly forecasts corn and soybean production from one to four months earlier than harvests.

In [10] To identify and analyze the characteristics and algorithms used in the crop yield prediction study, they conducted a Systematic Literature Review (SLR). Temperature, rainfall, and soil composition are the most frequently used features while Artificial Neural Networks is the most commonly used method in these models, as per the analysis.

III. Proposed Architecture and Methodology

A. Proposed Architecture.

Fig. 1 shows the design of the proposed work. The crop dataset which is taken consists of yield performance, management data, weather data, and soil data. In the preprocessing stage, feature selection techniques are applied. After preprocessing step, the data is partitioned into training data(80%) and testing data (20%). Regression models such as Multilinear Regression, Logistic Regression, Support Vector Regressor, and Artificial Neural Network Regressor are given training data. The testing data is used to evaluate the model. Metrics like as MSE, MAE, RMSE, Variance Score, and Accuracy are calculated to assess the model performance. The model can be applied to real-world data if the error is minimum.

B. Proposed Methodology of the Work

Data Preprocessing

Data preprocessing is used to improve the model's performance. It's a crucial step in building effective Machine Learning models. In the preprocessing stage, various feature selection techniques are applied. In the model, multilinear regression and logistic regression we have used a feature selection method in which the features with constant values are removed. Out of a total of 394 independent features, 13 features are of constant values. So these features have been removed.

In the Support Vector Regressor, we have used the principal component analysis(PCA) feature selection method. Principal Component Analysis (PCA) is an unsupervised learning method for reducing the number of features in machine learning.

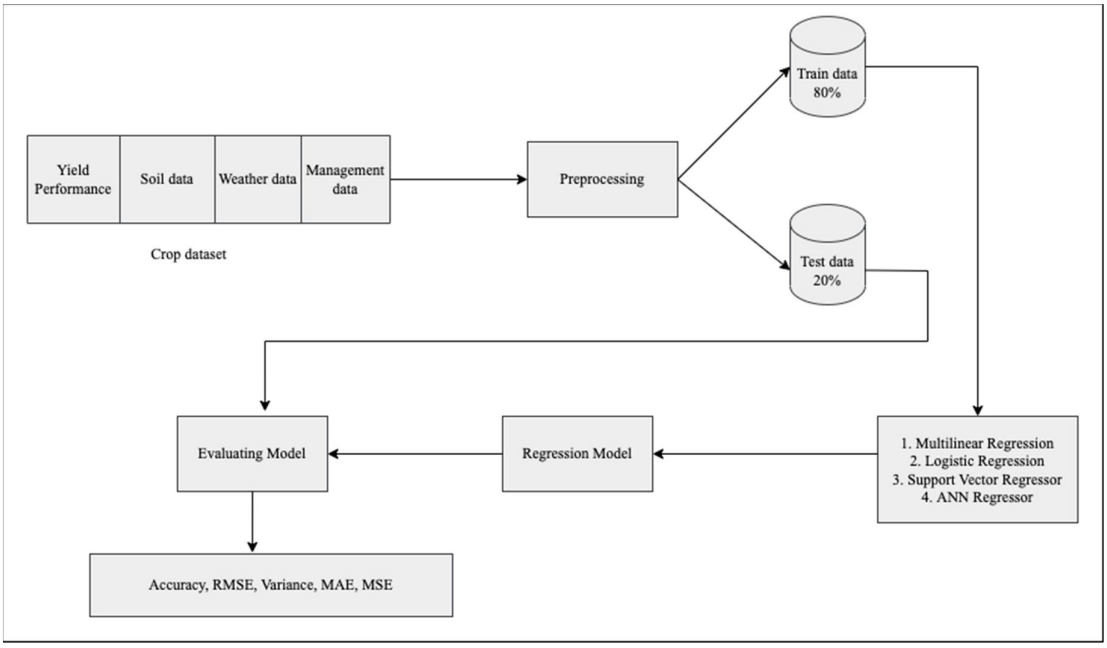

Figure 1. Architecture of the Regressor Model

C. Algorithms

The following are the algorithms used to build models for the crop yield prediction

Algorithm: Multilinear Regression

Input:

- Data set, D, consisting of training tuples along with the class label;
- Feature list;
- Feature_selection method;

Output: A multilinear regression model

Method:

1. Extracting dependent and independent Variables
2. For independent variables apply the Feature_selection method(D, feature_list) to find the best features
3. We will fit our regression model to the training set
4. Test the performance of the model using the test data
5. Calculate variance, Mse to measure the model performance
6. $explainedVariance(y, \hat{y}) = 1 - \frac{Vary - \hat{y}}{Vary}$
7. $MSE = \frac{1}{n} * \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$
8. If explainedVariance is 1 then it is a perfect prediction

Algorithm: Logistic Regression

Input:

- Data set, D, consisting of training tuples along with the class label;
- Feature list;
- Feature_selection method;

Output: A logistic regression model

Method:

1. Extracting dependent and independent Variables
2. For independent variables apply the Feature_selection method(D, feature_list) to find the best features
3. We will fit our regression model to the training set
4. Test the performance of the model using the test data
5. Confusion matrix is built to check the accuracy

6. $acc = \frac{1}{|Te|} * \sum_{x \in Te} I[\hat{c}(x) - c(x)]$

Algorithm: Support Vector Regressor
Input:

- Data set, D, consisting of training tuples along with the class label;
- Feature list;
- Feature_selection method;

Output: A support vector regressor model
Method:
1. Read the Dataset(D)
2. For attributes apply the Feature_selection method(D, feature_list) to find the best features
3. Apply Feature Scaling on Dataset (D) which helps to normalize the data within a particular range.
4. We will fit the SVM regressor to the training set
5. Test the performance of the model using the test data
6. Calculate variance, Mse to measure the model performance
7. $explainedVariance(y, \hat{y}) = 1 - \frac{Vary - \hat{y}}{Vary}$
8. $MSE = \frac{1}{n} * \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$
9. If explainedVariance is 1 then it is perfect prediction

Algorithm: ANN Regressor
Input:

- Data set, D, consisting of training tuples along with the class label;
- Feature list

Output: ANN Regressor model
Method:
1. Input the dataset to the model
2. Split the Dataset(D) into features and target
3. Scale the dataset using z-score normalization
4. Train the Neural Network model with three layers namely Input Layer, Hidden Layers, Output Layer
5. for i:=1 to epochs
5.1 Calculate the net input to be forwarded to the next unit
5.2 Apply an activation function to it
5.3 Calculate the output of that next unit
5.4 Update the weights by back-propagating the error
6. Loss functions are applied
7. Output the predicted test data using the trained model

D. Loss Function

The loss functions we used in the regressors are -

1) Mean Squared Error (MSE)

$$MSE = \frac{1}{n} * \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \quad (1)$$

MSE is a frequently used model evaluation metric in regression models. It is the mean of the squared difference of predicted and actual observations. It just cares about the average amount of error, regardless of its direction.

2) Mean Absolute Error (MAE)

$$MAE = \sum_{i=0}^{n} \frac{1}{n} * |y_i - \hat{y}_i| \quad (2)$$

The MAE is the average of the absolute deviations of predictions and actual observations. This, like MSE, measures the magnitude of errors without considering the direction of the errors.

IV. Discussion on Experimental Investigations

E. Dataset

The dataset which we considered has four sets: yield in bu/ac, management data, weather, and soil data. We would estimate the yield of soyabean crop based on the three sets of data. We have taken a dataset from a paper [1]. The soybean dataset has data for 9 states (25,345 samples) with 395 columns and 25346 rows.
The four sets of data contain the following features shown in Table I**.**

TABLE I. TYPES OF DATA USED FOR YIELD PREDICTION

Management data	Week by week combined percent of planted fields in each state, starting from April of every year
Weather data	Weather variables, namely rainfall, maximum temperature, minimum temperature, and vapor pressure
Soil data	Soil variables like clay percentage, available water content, pH of the soil, organic matter percentage, sand percentage, etc.
yield	Yield data contains observed average yield for soybean

E. Discussion on Results

Google Colab is the environment in which the model was built and validated. It's a free notebook environment that allows you to connect to Google Drive. Many machine learning libraries are pre-installed in Colab, and we can use the import keyword followed by the library name to bring them into our notebook. The libraries used in this project are Pandas, Numpy, sklearn, Matplotlib, Keras, Seaborn, etc.
In the model, multilinear regression we have used a feature selection method in which the features with constant values are removed. Out of a total of 394 independent features, 13 features are of constant values. So these features have been removed. Then the data with the remaining features are divided into training and testing data. Now the model is constructed and trained with the training data. The validation is done with the testing data and obtained a MAE of 3.903 and a MSE of 27.06.
In the model, Logistic regression, firstly we have converted the target variable into categorical values and then divided the dataset into training and testing data. Then the model is constructed and trained using the training data and validation is done with testing data and obtained an accuracy of 0.92. In terms of errors, the mean absolute error obtained is 6.388 and the mean squared error is 0.07.
In the Support Vector Regressor, we have used PCA feature selection method and then the process is done and obtained a MAE of 3.90 and MSE of 20.33.
In the Artificial neural networks regressor model after splitting the data into train and test data, we have taken one input layer, one output layer, and five hidden layers. We have run the algorithm with two types of metrics and obtained a 2.833 MAE with 250 epochs and a 16.55 MSE with 250 epochs. We have taken the default learning rate as 0.1 and implemented the process. The tabular representation of the MAE, MSE for different algorithms are in Table II.

TABLE II. MAE AND MSE FOR THE ALGORITHMS USED

Algorithm Name	Mean Absolute Error	Mean Squared Error
Multilinear Regression	3.903	27.06
Logistic Regression	6.388	0.07
Support Vector Regressor	3.290	20.33
ANN Regressor	2.833	16.55

The training and testing error for the ANN Regressor model is shown in Fig. 2. The error graph in Fig. 2 is with 250 epochs. Fig. 2(a) shows graphical representation of mean absolute error with respect to epochs. At epoch-1 the validation error is 6.304 and the training error is 18.3101 and at epoch-250 the validation error is nearly 2.833 and the training loss is 0.935. Fig. 2(b) shows graphical representation of mean squared error with respect to epochs. At epoch-1 the validation error is 69.4614 and the training error is 539.8401 and at epoch-250 the validation error is nearly 1.2692 and the training loss is 17.3807. The error decreases as the number of epochs increases.

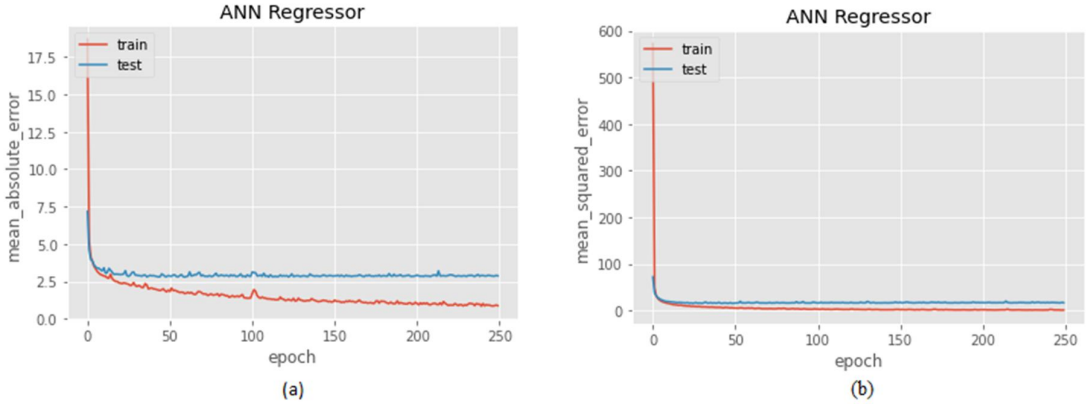

Figure 2. MAE and MSE with respect to epochs in ANN Regressor

The ROC curve from the logistic regression model is depicted in Fig. 3. ROC curves (Receiver Operating Characteristic) are useful tools for assessing and fine-tuning classification models. The test is more accurate when the curve in the plot is closer to the top left corner.

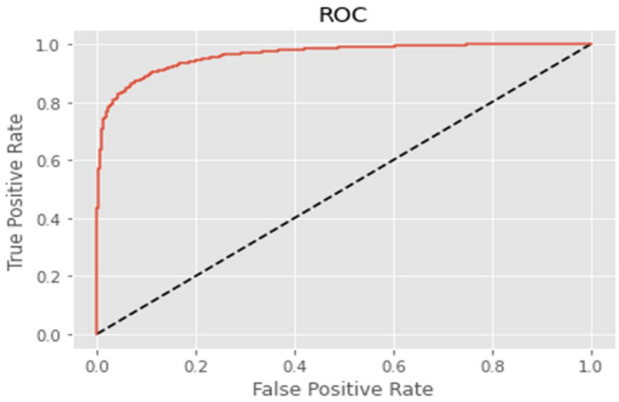

Figure 3. ROC curve for Logistic Regression

F. Comparison of Results

We have considered the MAE and MSE as evaluation metrics for our project and compared all the models with respect to them. In the four regressors, the best result was obtained by the ANN Regressor algorithm with an MAE of 2.833. Fig. 4 shows Graphical Representation of the MAE and MSE of various Regressor Algorithms.

V. CONCLUSIONS

Crop yield prediction is done using machine learning techniques based on weather, soil, and management data. Machine learning is a decision-making tool that helps in predicting crop yields, as well as for deciding which crops to cultivate and when to cultivate them. Farmers will benefit from the implementation of this project since it will help them reduce losses and increase crop yields, allowing them to expand their agricultural resources. This will not only help farmers in deciding which crop to cultivate in the next season, but it will also aid in the

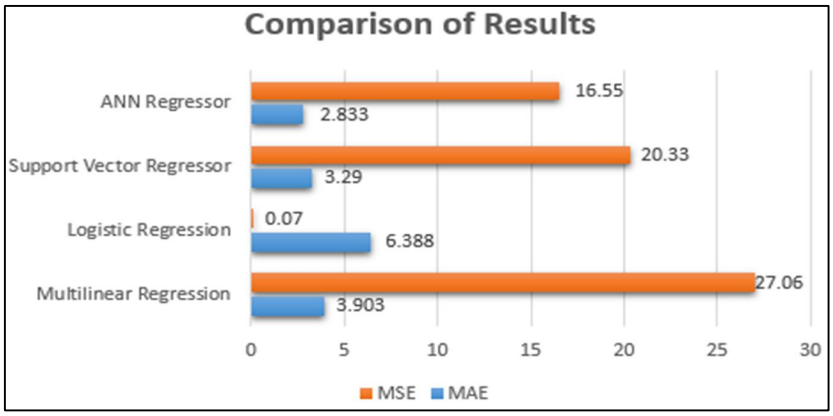

Figure 4. Comparison of Results

technological and agricultural divide. Presently the model is constructed for a particular crop named soybean. We used four algorithms of regression in machine learning over the dataset which we have taken. The best features are extracted by removing all the constant features in the dataset and also used pca then the model is constructed. The model was evaluated using evaluation metrics like MSE, MAE. The best obtained results are for the ANN regressor. This can be further extended and used for different types of crops. Also, we have taken certain values as a dataset to predict the crop yield, these can be changed to images and further study can be done over it.

REFERENCES

[1] Yadav, Jagdeep & Chopra, Shalu & M, Vijayalakshmi. (2021). SOIL ANALYSIS AND CROP FERTILITY PREDICTION USING MACHINE LEARNING. International Research Journal of Computer Science. 8. 32-40. 10.26562/irjcs.2021.v0802.001.

[2] 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.

[3] T G, Keerthan Kumar. (2019). Random Forest Algorithm for Soil Fertility Prediction and Grading Using Machine Learning. 9. 1301-1304. 10.35940/ijitee.L3609.119119.

[4] Janakiramaiah, B. et al. 'Intelligent System for Leaf Disease Detection Using Capsule Networks for Horticulture'. 1 Jan. 2021 : 6697 – 6713.

[5] Subramanian, Malliga, L.V., Narasimha Prasad, B., Janakiramaiah, A., Mohan Babu and VE, Sathishkumar, Hyperparameter Optimization for Transfer Learning of VGG16 for Disease Identification in Corn Leaves Using Bayesian Optimization, Big Data, volume 10-3, 215-229, 2022, doi =10.1089/big.2021.0218.

[6] Khaki, Saeed, Lizhi Wang, and Sotirios V. Archontoulis. "A CNN-RNN framework for crop yield prediction." Frontiers in Plant Science 10 (2020): 1750.

[7] Khaki, Saeed, and Lizhi Wang. "Crop yield prediction using deep neural networks." *Frontiers in plant science* 10 (2019): 621.

[8] Saeed Khaki, Hieu Pham, Ye Han, Andy Kuhl, Wade Kent, Lizhi Wang, "DeepCorn: A semi-supervised deep learning method for high-throughput image-based corn kernel counting and yield estimation", Knowledge-Based Systems, Volume 218, 2021, 106874, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2021.106874.

[9] Khaki, S., Pham, H. & Wang, L. "Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning". *Sci Rep* **11,** 11132 (2021). https://doi.org/10.1038/s41598-021-89779-z

[10] Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal, " Crop yield prediction using machine learning: A systematic literature review",Computers and Electronics in Agriculture,Volume 177, 2020, 105709,ISSN0168-1699,https://doi.org/10.1016/j.compag.2020.105709.