Crop Yield Prediction Using Machine Learning

B. SHIVANI[1], A. MANISHA[2], G. EESHWAR[3], K. BANU TEJA PHANINDRA SINGH[4], B. SNEHITHA[5]

[1, 2, 3, 4, 5] *Dept. of Information Technology, TKR College of Engineering and Technology, Hyderabad, Telangana*

Abstract*— Agriculture is the field that plays an important role in improving our country's economy. Selecting every crop is very important in agriculture planning. Crop yield prediction is an important Agricultural problem. Every farmer always tries to know how much yield will be produced and whether it meets their expectations. Predicting the crop yield well ahead of its harvest would help the farmers for taking appropriate measures. Several agricultural yield prediction and modeling systems have been created in the past, with varying degrees of effectiveness. Previously, yield prediction was calculated by examining farmers' experience with a specific crop. In the agricultural area, climate and other environmental changes have become a big challenge. Machine learning algorithms' predictions will assist farmers in deciding which crop to plant in order to achieve maximum production.

I. INTRODUCTION

India is a country where agriculture and agriculture-related industries are the dominant sources of living for the people. Agriculture is a large part of the country's economy. Aside from that, India suffers from natural calamities such as floods and droughts, which result in crop losses. While harvesting crops, one's strategy should be spot on, taking into account elements such as season, soil moisture, and weather conditions, as well as when to harvest the crop to receive the best yield.

In recent years, using technology to improve cultivation awareness has become unavoidable. Seasonal climatic conditions are also shifting, causing harm to critical assets like land, water, and air, leading to food insecurity. In one scenario, agricultural yields are always falling short of demand, enabling the development of a smart system to address the issue of declining crop productivity.

Agricultural yields are continually falling short of demand, and there is a need for a smart system that can solve the problem of decreasing crop yield. Steel plows, seed drills, barrows, hoes, and other upgraded implements have only recently begun to be adopted by Indian farmers to a limited extent. Traditional farming methods are to blame for the country's low agricultural productivity.

II. LITERATURE SURVEY

This segment discusses how many researchers have worked on various Machine Learning algorithms for Crop yield predictions. *Ms. Shreya V. Bhosale, Ms. Ruchita A. Thombare, Mr. Prasanna G. Dhemey, Ms. Anagha N. Chaudhari Proposed* "Crop Yield Prediction Using Data Analytics and Hybrid Approach" in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) on 1 August and published by Institute of Electrical and Electronics Engineers (IEEE)**.** They used three algorithms namely clustering k-means, Apriori, and Naïve Bayes algorithm, then they hybridized the algorithms for better efficiency of yield prediction and they consider parameters like Land Area, Rainfall, Soil type, Season, Crop name, Production, and Year, District.

All of the findings from the three algorithms are combined in the final Accumulation model and provided to the GUI model. Based on the results of the Apriori and Nave Bayes algorithms, the Graphical User Interface then offers crop names for greater crop yield and estimates crop yield in quintals. The graphical results of K-means and Nave Bayes for crop analysis in defined rainfall are represented in the final depiction.

Aditya Shastry, Sanjay H A, and Madhura Hedge proposed " A Parameter based ANFIS Model for crop

yield prediction" in 2015 IEEE International Advance Computing Conference (IACC) and published by the Institute of Electrical and Electronics Engineers (IEEE).They used Fuzzy logic, Adaptive Neuro-Fuzzy Inference System (ANFIS), and Multiple Linear Regression and input parameters like biomass, extractable soil water (ESW), radiation, and rain. These models are used for wheat yield prediction only. The ANFIS model is trained and validated for the test set using a training dataset. The method is repeated until the RMSE value drops. The RMSE value for wheat crop prediction using Fuzzy Logic is 6.4251, the RMSE value for Multiple Linear Regression is 9.252, and the RMSE value for ANFIs is 3.3282. Based on RMSE values, the results of the three prediction models, Fuzzy logic, ANFIS, and Multiple Linear Regression, are compared. The ANFIS model proved more accurate than Fuzzy logic and Multiple Linear Regression in predicting wheat yield

S.Veenadhari, Dr.Bharat Misra, and Dr.CD Singh proposed "Machine learning approach for forecasting crop yield based on climatic parameters" in International Conference on Computer Communication and Informatics (ICCCI -2014), Jan 2014 and published by the Institute of Electrical and Electronics Engineers (IEEE)**.** They used the C4.5 algorithm to find the most influencing climatic parameters on the crop yields of selected crops in selected districts of Madhya Pradesh. The selected crops in this study are Soybean, Maize, Paddy, and Wheat. Climatic parameters such as rainfall, maximum temperature, minimum temperature, Potential Evapotranspiration, Cloud cover, and Wet day frequency were also collected from various sources.

Ratchaphum Jaikla, Sansanee Aueph anwiriyakul, and Attachai Jintrawet "Rice Yield Prediction using a Support Vector Regression method" in 2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology and published by Institute of Electrical and Electronics Engineers (IEEE). They found that SVR's performance in predicting rice grain weight is comparable to DSSAT4 based on the experimental data. They also identified that the MAPE of grain weight computed from the model and a human expert is 2.94 percent, whereas the MAPE of DSSAT4 and a human expert is 2.91 percent, indicating that, while the MAPE of this model is higher than DSSAT4, the error from this model is still within acceptable limits.

III. PROPOSED SYSTEM

Step 1: Define the Problem Statement's goal.
We must first determine what needs to be projected. In this paper, the goal is to forecast agricultural yields.

Step 2: Collecting data
Once we know what kinds of data we need, we need to know how to get it. Data can be gathered manually or by web scraping. We require a dataset containing information on crop types, seasons, area, state, and production for our paper.

Step 3: Data Preparation
We nearly never have the correct format for the data we collect. Missing values, redundant variables, duplicate values, and other irregularities will be found in the data collection. Such irregularities must be removed because they can lead to incorrect computations and predictions. As a result, at this point, we check the data set for any irregularities and correct them immediately.

Step 4: Exploratory Data Analysis
Exploratory Data Analysis, or EDA, is the machine learning brainstorming step. Understanding the patterns and trends in the data is the goal of data exploration. All of the useful insights are drawn at this point, and the relationships between the variables are recognized.

For example, while estimating yield, we know that if the crop is cultivated in the right spot at the right time, we may expect a high yield. At this point, such interconnections must be recognized and mapped.

Step 5: Building a Machine Learning Model
The Machine Learning Model is built using all of the insights and patterns discovered during Data Exploration. The data set is always separated into two parts, training data and testing data, at this stage. The model will be built and analyzed using the training

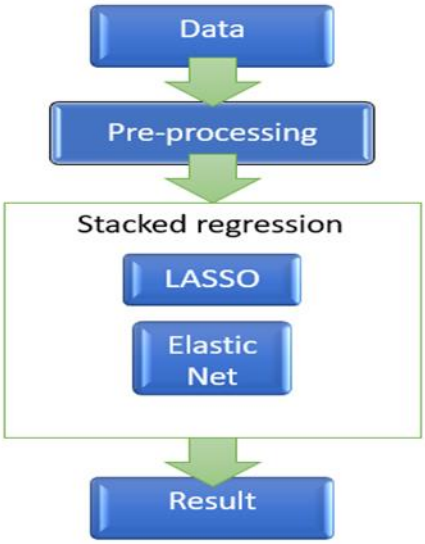
data. The model's logic is based on the Machine Learning Algorithm that is currently in use.

The sort of problem we're trying to solve, the data set, and the problem's complexity all play a role in selecting the proper algorithm.

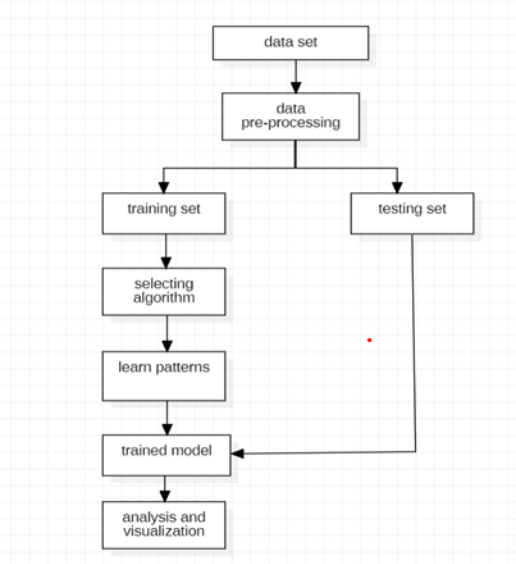
Step 6: Model Evaluation & Optimization
It is finally time to put the model to the test once it has been built using the training data set. The testing data set is used to determine the model's efficiency and accuracy in predicting the outcome. Any further model improvements can be applied once the accuracy has been calculated. To increase the model's performance, techniques such as parameter adjustment and cross-validation can be applied.

Step 7: Predictions
The model is then used to make predictions after it has been validated and modified. Crop yield production based on user inputs is the ultimate output.

IV. SYSTEM DESIGN



- Data Flow Diagram:



Data preprocessing:
The transformations made to our data prior to feeding it to the algorithm are referred to as pre-processing. Data preprocessing is a method for transforming raw data into a clean data set. anytime data is acquired from various sources, it is obtained in raw format, which makes analysis difficult.
Steps in data preprocessing:

Step-1: removing null values
One of the most crucial steps is to remove null values from the dataset. These null values have a negative impact on the performance and accuracy of any machine learning method. As a result, it is essential to eliminate null values from the dataset before running any machine learning algorithm on it.

To eliminate null values from the dataset, we will utilize the (pandas)Python library.

Step-2: handling character data
We typically deal with datasets containing numerous labels in one or more columns in machine learning. Words or numbers can be used as labels.

Label encoding is the process of turning labels into a numeric form that is machine-readable. Machine learning algorithms can then determine how those labels should be used in a more efficient manner. In supervised learning, it is a crucial pre-processing step for the structured dataset.

A Dataset is selected which is suitable for our requirements. We need to remove duplicate records from the dataset by using python libraries like the panda, and NumPy. If there are any null values in the dataset then we need to replace null values with the mean, mode, or median of a particular column or with Zero or One. To handle character data we need to use methods like label encoding. Data scaling must be used for ranging the values in the dataset.

The detailed procedure is as follows

- The data set is divided into two different parts. One is for training(train dataset), and another is for testing(test dataset).
- We need to train the models(LASSO, Elastic Net) with the first part, i.e. train the dataset. Then we need to test the models with the second part, i.e. test dataset.
- the results of the testing are the predicted values. These predicted values are then provided as inputs for the meta-model(LASSO regressor), for better predictions. This process is known as stacking.

The predicted values of each model are compared with the original values to know the efficiency of the model.

V. IMPLEMENTATION

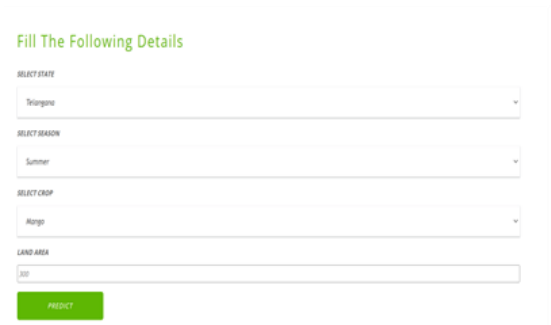
Sample code for splitting the dataset into train and test dataset:

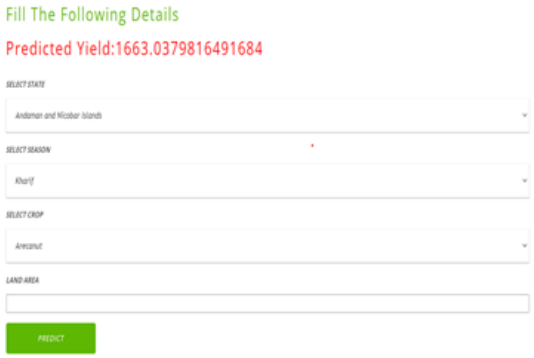
```
x = data1.drop("Production",axis=1)
y=data1['Production'].values
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.33, random_state=42)
print("x_train :",x_train.shape)
print("x_test :",x_test.shape)
print("y_train :",y_train.shape)
print("y_test :",y_test.shape)
```

Testing is the process of examining a system or its component(s) with the goal of determining if it meets the requirements. Simply, testing is the process of running a system to find any gaps, faults, or missing requirements that do not match the actual needs.

Test Cases

S. NO	TEST NAME	INPUTS	EXP O/P	ACTUAL OUTPUT	RES
1	Collect and load the dataset	Load the crop yield dataset collected from Kaggle	Load crop yield dataset	Successfully crop yield dataset is loaded	pass
2	Feature selection	Crop yield dataset	Only the important features are selected	Successfully important features are selected	pass
3	Splitting dataset into training and testing set	Dataset with selected features	spitted to training and testing set	successfully spitted to training dataset and testing dataset	pass
4	Training model with machine learning algorithms	Training dataset	Trained model	Successfully trained machine learning model	pass
5	Testing ML model with testing dataset	Testing dataset	testing dataset validated	successfully validated testing data set	pass
6	Print error rate of the algorithms used	Testing results	RMSE values of the algorithms used are displayed	Successfully RMSE values of the algorithms used are displayed	pass





VI. CONCLUSION

This approach is aimed to fix the rising rate of farmer suicides while also supporting them in becoming more financially secure. The goal of this study is to use machine learning techniques to estimate crop yield. The accuracy is calculated using a variety of machine learning approaches. Appropriate datasets were gathered, evaluated, and trained using machine learning technology.

The r2 score of LASSO is 0.005263685700963694 and the r2 score of Elastic Net is 0.005240518964385488. After using the stacking regressor the r2 score is 0.005298486910471301. Farmers would benefit from accurate predictions of various crop yields throughout different districts. This helps to boost the Indian economy by increasing crop output rates

REFERENCES

[1] Aditya Shastry, Sanjay H A, and Madhura Hedge (2015) “A Parameter based ANFIS Model for crop yield prediction”

[2] M. Gunasundari Ananthara, Dr. T. Arun Kumar, Ms. R. Hemavathy (2013) “CRY – An improved Crop Yield Prediction model using Bee Hive Clustering Approach for Agricultural data sets”

[3] A. Majid Awan and Mohd. Noor Md. Sap (2006) “An Intelligent System Based on Kernel Methods for Crop Yield Prediction”

[4] Ratchaphum Jaikla, Sansanee Auephanwiriyakul, and Attachai Jintrawet (2008) “Rice Yield Prediction using a Support Vector Regression method”

[5] Ms. Shreya V. Bhosale, Ms. Ruchita A. Thombare, Mr. Prasanna G. Dhemey, Ms. Anagha N. Chaudhari (2018) “Crop Yield Prediction Using Data Analytics and Hybrid Approach”

[6] S.Veenadhari, Dr.Bharat Misra, Dr.CD Singh (2014) “Machine learning approach for forecasting crop yield based on climatic parameters”