International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

CROP PRODUCTION PREDICTION

Prof. Pallavi Patil, Ms. Ayushi Patil, Ms. Gayatri Khandalkar, Ms. Tejal Patil, Ms. Sneha Jadhav

Department Of Computer Engineering, Genba Sopanrao College Of Engineering, Balewadi, Pune

---

Abstract - Over 50% of India's population relies on agriculture for their survival, making it the foundation of the Indian economy. Weather, climate, and other environmental changes have evolved into a serious threat to the continued health of agriculture. The decision-support tool for crop production prediction (CPP), which includes supporting decisions about which crops to produce and what to do during the growing season of the crops, is provided by machine learning (ML), which plays a vital part in this process. The current study examines a systematic review that extracts and synthesises the CPP features, and it also explores a number of approaches that have been created to examine crop production prediction using artificial intelligence techniques.

Reduction in relative error and lower crop yield prediction accuracy are the Neural Network's main drawbacks. Similar to this, supervised learning algorithms failed to recognise the nonlinear relationship between input and output variables, which presented a challenge during the selection, grading, or sorting of fruits. To establish an accurate and effective model for crop classification, including crop yield estimation based on weather, crop disease, classification of crops based on the growing phase, etc., numerous investigations were advised.Prior to harvest, crop output predictions would aid farmers and policymakers in deciding on the best course of action for marketing and storage.

Key Words: (Agriculture, Area, Rainfall, Temperature ,Crop Production, Pesticides, Machine Learing)

1.INTRODUCTION:

Since it is essential to both human and animal survival in India, agriculture is the foundation of the country's economy. The need for agricultural products will skyrocket as the world population, which was estimated at 1.8 billion in 2009, is expected to reach 4.9 billion by 2030. The need for agricultural products will increase as the world's population grows, necessitating effective farmland development and an increase in crop output. Higher agricultural crop output is the primary goal of crop yield estimation, and this goal is pursued by utilising numerous tested models. Due to its effectiveness in a number of fields, including forecasting and fault detection, machine learning is being employed all over the world.

1.1.OVERVIEW:

ML offers a number of efficient techniques that are used to identify the relationship between input and result in yield and crop prediction. For example, smart irrigation systems, crop disease prediction, crop selection, weather forecasting, and determining the minimal support price are all examples of machine techniques used in agriculture. This study examines the advantages and disadvantages of the various ML-based agriculture techniques.

1.2.PROBLEM STATEMENT:

India's response to climate change, Over the past 20 years, factors like soil composition, pesticide use, and others have had a significant negative impact on the performance of the majority of agricultural crops. estimating the crop based on productivity, temperature, pesticide use, and rainfall. The project will assist farmers in choosing which crop to plant for the best yield.

2.BODY OF PAPER:

2.1.DATASET:

Source: Kaggle .

2.2.DATASET DETAILS:

INFORMATION:

Understanding global agricultural output is essential for tackling issues with food security and minimising the effects of climate change in light of the ongoing growth of the human population.

Predicting crop yield is a significant agricultural issue. For making judgements regarding agricultural risk management and generating forecasts for the future, it is vital to understand that agricultural productivity is primarily influenced by weather conditions (rain, temperature, etc.), pesticides, and reliable information about past crop yield.

There are 28242 rows and 8 characteristics in this collection. The data set used is shown in general in fig. 3.1, which is provided below. Figure (a) displays the names of each attribute, the number of entries, and the attribute's data type. Target property in this case is hg/ha_yield.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28242 entries, 0 to 28241
Data columns (total 8 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Unnamed: 0                      28242 non-null  int64
 1   Area                            28242 non-null  object
 2   Item                            28242 non-null  object
 3   Year                            28242 non-null  int64
 4   hg/ha_yield                     28242 non-null  int64
 5   average_rain_fall_mm_per_year   28242 non-null  float64
 6   pesticides_tonnes               28242 non-null  float64
 7   avg_temp                        28242 non-null  float64
dtypes: float64(3), int64(3), object(2)
memory usage: 1.7+ MB
```

Figure:(a)

2.3 .SYSTEM ARCHITECTURE:

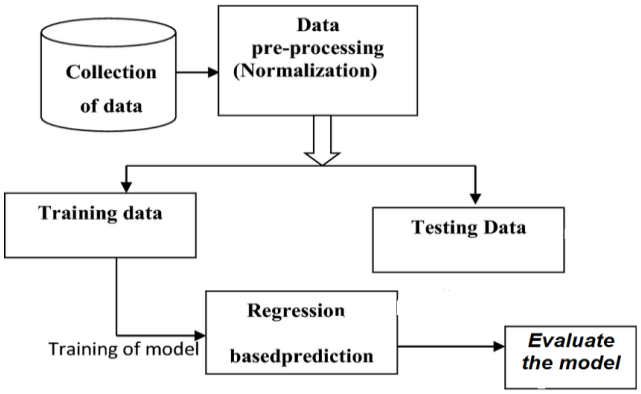


Fig:(B) System Architecture.

2.4.DATA COLLECTION:

The systematic process of obtaining observations or measurements is known as data collection. Although methods and objectives may vary throughout fields, the general data collection procedure is basically the same. Consider the following before you start gathering data:

- The purpose of the study;
- The kinds of data you'll gather;
- The techniques and protocols you'll employ to gather, store, and process the data.

The researchers must specify the data sources, data types, and methodologies used during data gathering. We'll quickly discover that are a variety of data collection techniques.

2.5.DATA PREPROCESSING:

Any project involving machine learning must start with data. Data collection, selection, preparation, and transformation are all part of the difficult second stage of a project's implementation.

Data for this research was obtained via Kaggle. Temperature, rainfall, area, pesticides, and yield (production) all have independent data sets, and these data sets are combined to support the problem statement. The kind of data used determines the kind of prediction that must be made. Our data is identified as such here.

There are various datatypes in the used data. Figure .(a) shows that several properties have data types of object. The attribute has categorical values, therefore Since we cannot build the model on this dataset in this circumstance, we execute a single hot encoding to fix the issue. Data can be converted using one hot encoding as a means of getting a better prediction and preparing the data for an algorithm. By using one-hot, we create a new categorical attribute for each category value and give it a binary value of 1 or 0. A binary vector is used to represent each integer value. The index is denoted by a 1 and all values are zero. As a result, no more objects exist.
property in order to begin building the model. After one hot encoding, there are 116 total attributes and 28241 rows.

Once the one-hot encoding is complete, the missing values in this data set are checked to ensure that there are none. Here, we use standardscaler to do standardisation. When your data has a Gaussian distribution, scaling of features is done via standardisation. It works best for: Normalisation can put all the data on a same scale, regardless of whether they are categorical, numerical, textual, or time series data. As a result, it is frequently utilised for scoring while developing or updating a predictive model.

2.6.DATA VISUALIZATION:

When presented in pictorial form, a lot of information is simpler to comprehend and process. The process of converting enormous data sets into a statistical and graphical representation is known as data visualisation. Making data more understandable and accessible is a crucial objective of data science and knowledge discovery methodologies. Due to its capability to show results at the end of the machine learning process, data visualisation has gained popularity recently. However, it is also increasingly being utilised as a tool for exploratory data analysis prior to applying machine learning models.

It offers a clearer grasp of how the attributes are related to one another.

For example, we have an attribute named area that contains the countries; this attribute can be represented in the form of a pie chart, making it simpler to understand which countries are present in our data set and in what proportion the entries are present. In our project, we have used a variety of visualisation tools to show various instances of our data.

Here, we utilise a pie chart for the area, a box plot for the average rainfall, and a histogram for the production of hg/ha_yield. Heatmap is a data visualization technique.

2.7.DATA SPLITTING:

When data is divided into two or more subgroups, this is known as data splitting. A two-part split often involves testing or evaluating the data in one part and training the model in the other. Data splitting is a crucial component in data science, especially when building models from data. This method makes it possible to create data models and the procedures that employ them, such machine learning.

In a straightforward two-part data split, the training data set is utilised to train and create models. Training sets are frequently used to estimate various parameters or to evaluate the performances of various models. After the training is complete, the testing data set is used. To ensure that the final model operates properly, the training and test sets of data are compared. Data is frequently divided into three or more sets when it comes to machine learning. The dev set, which is one more set with three sets, is used to alter the settings for the learning process. The size of the initial data pool or the number of predictors in a predictive model may have an impact on how the data should be divided; there is no defined rule or metric for this. The data for our project is divided 70:30, or 70% training data and 30% testing data. Data sets should be split up such that there is a large amount of training data available. For instance, the ratio of training to testing data may be 80/20 or 70/30. The data determines the precise ratio. Data splitting is frequently used in machine learning to prevent overfitting. In that situation, a machine learning model fits its training data too well and is unable to consistently fit new data. Usually, the initial data in a machine learning model is divided into three or four sets. The training set, the development set, and the testing set are the three sets that are frequently used:

The portion of data used to train the model is known as the training set. In order to improve any of its parameters, the model must observe and learn from the training set. The data that was tested in the final model and contrasted with the earlier data sets is known as the "testing set." The testing set serves as an assessment of the chosen algorithm and mode.

2.8. EVALUATING THE MODEL:

Model evaluation involves measuring an ML model's performance to determine which one performs the best and best matches the given challenge. Model evaluation is essential to ensuring that a model operates properly and optimally when used in production.

The model is considered accurate if accuracy is high. If the accuracy is low, it indicates that the model's predictive ability is weak or that it is not the ideal model to apply to the issue. Thus, ML experts are aware of the requirement to enhance the model to raise its accuracy.

In this project we have used evaluation n methods like r2 score, RMSE (Root Mean Squared Error) and MAE(Mean Absolute Error).Here, r2 score represents the accuracy which varies between 0-1, more the r2 score closer to 1 better the accuracy. Whereas, RMSE and MAE represents the error given by the model.

2.9.ALGORITHM IMPLEMENTED:

2.9.1.REGRESSION:

Regression analysis is a group of statistical procedures used in statistical modelling to determine the relationships between a dependent variable (often referred to as the "outcome" or "response" variable, or a "label" in machine learning jargon), and one or more independent variables (often referred to as "predictors," "covariates," "explanatory variables," or "features").

In some circumstances, regression analysis can be used to infer causal links between the independent and dependent variables. Regression analysis is commonly used for prediction and forecasting, where its use substantially overlaps with the subject of machine learning. Regressions by themselves, it should be noted, only illuminate connections between a dependent variable and a group of independent variables in a given dataset.

2.9.2.DECISION TREE:

By segmenting the feature variables into small zones, each of which will have one prediction, the Decision Tree Regressor attempts to forecast a continuous target variable. One continuous variable will be used first, followed by two continuous variables.Regression tasks are performed using the decision tree regressor, a machine learning method. Each internal node of the tree represents a test on a feature, each branch indicates the test's result, and each leaf node represents a prediction as the algorithm builds a model in the shape of a decision tree.

2.9.3.SUPPORT VECTOR REGRESSOR:

A machine learning approach called support vector regression (SVR) is utilised for regression tasks. The approach, which is based on the idea of support vector machines (SVMs), looks for a function that matches the data as closely as possible while simultaneously containing mistakes within a predetermined range. Instead of classifying points, SVR predicts a continuous output value for a new input based on the input's position in relation to the hyperplane to modify the SVM for regression. The margin between the hyperplane and the support vectors is used as a tuning parameter to regulate the trade-off between fitting the data and generalisation. The predicted value is the distance between the input and the hyperplane.

2.9.4.RANDOM FOREST REGRESSOR:

The machine learning algorithm known as the random forest regressor is utilised for regression tasks. It is an ensemble method that creates a single model from several decision tree regressors.A number of input features can be used with the random forest regressor, which can handle both linear and nonlinear data.
Given that each tree is trained on a randomly chosen portion of the data, the random forest regressor can be successful in minimising the effects of outliers in the data**.** The most pertinent features for predicting the target variable can be found using the random forest regressor's measure of feature relevance.

2.9.5 ADABOOST REGRESSOR:

A machine learning algorithm called the Adaboost regressor is utilised for regression tasks. The Adaboost ensemble approach, on which the algorithm is based, combines a number of weak learners into a strong learner .Adaboost regressor works by iteratively training a series of regression models on the training data, with each model aiming to fix the mistakes produced by the one before it. Higher weights are given to examples that are challenging to predict when the algorithm distributes weights to each training instance based on their misclassification rate.

The Adaboost regressor can be used with a wide range of base estimators and can handle both linear and nonlinear data.By giving examples that are challenging to forecast a larger weight, the Adaboost regressor can be useful in minimising the effects of outliers in the data.Since each succeeding model concentrates on fixing the faults of the preceding model rather than training on the complete dataset, the Adaboost regressor can be computationally effective.

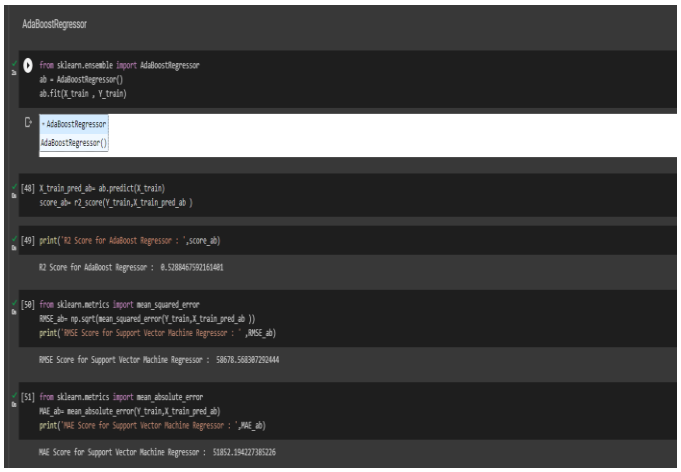
2.9.6 MULTI-LAYER PERCEPTRON REGRESSOR:

A common artificial neural network (ANN) used in machine learning for regression tasks is the MLP (Multilayer Perceptron) Regressor. It has several layers of nodes, each of which functions as a computational unit that processes input from the layer before and computes an output based on its internal parameters and activation function**.** During training, the MLP Regressor iteratively modifies the nodes' weights and biases to reduce the discrepancy between the anticipated output and the actual output. Backpropagation, a method for accomplishing this, entails calculating the gradient of the loss function with respect to the weights and biases and updating them as necessary.

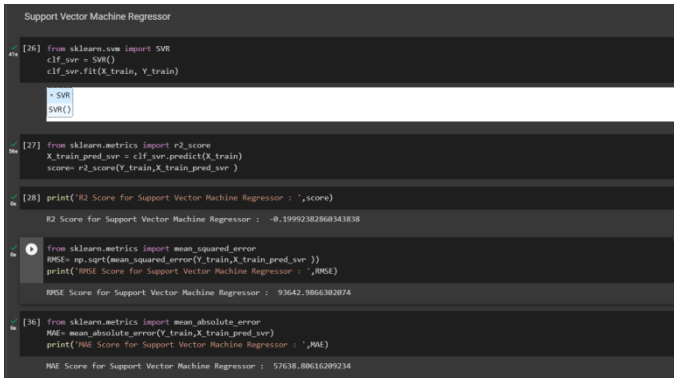
3.RESULTS:

Algorithm used	R2 Score	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
Adaboost regressor	0.528	51852.1942	58678.5683
Support vector regressor (SVR)	-0.199	57638.8061	93642.9866
Random forest regressor	0.819	23736.2256	36345.5453
Decision tree regressor	0.934	14261.7472	21813.0737
MLP (Multilayer Perceptron)	0.337	46590.4725	69558.9380

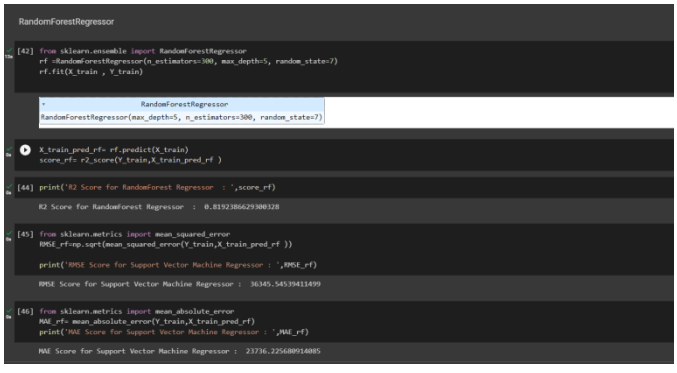
1.Adaboost regressor:



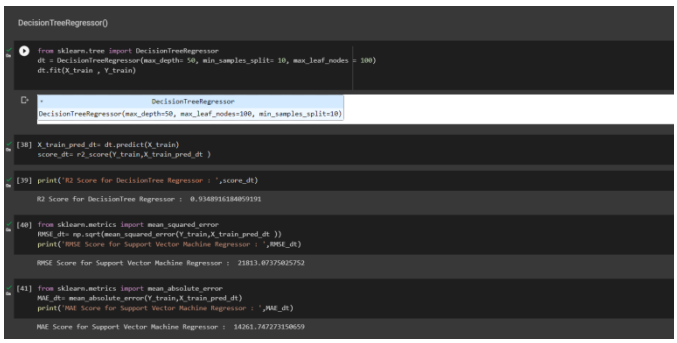
2. SVR :



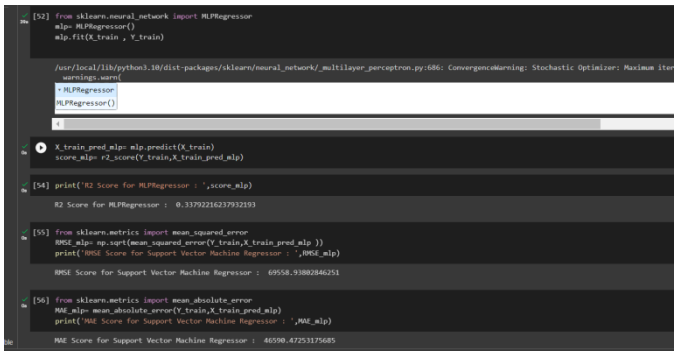
3. Random Forest Regressor:



4. Decision Tree Regressor:



5.MLP:



3. CONCLUSIONS:

In conclusion, employing machine learning (ML) techniques to estimate crop yield has proven to be a potential way to raise agricultural productivity. Large volumes of data may be analysed using ML models, which can then provide precise forecasts about crop output while accounting for diverse elements including soil, climate, and pesticides.

With varied degrees of effectiveness, a number of ML methods, including regression, decision trees, neural networks, and support vector machines, have been used to predict crop yield. The particular crop, location, and data resources all influence the ML model selection as well as the data input properties**.** Overall, by enhancing crop management and lowering crop losses, the application of ML in agricultural production prediction has the potential to revolutionise agriculture. To assure their usefulness and viability, more study is required to enhance ML models and incorporate them into real-world agricultural systems.

4.REFERENCES:

- S. D. Kumar, S. Esakkirajan, S. Bama, and B. Keerthiveena, "A microcontroller-based machine vision approach for tomato grading and sorting using SVM classifier," Microprocessors and Microsystems, vol. 76, pp.103090, 2020.
- P. Tiwari, and P. Shukla, "Crop yield prediction by modified convolutional neural network and geographical indexes," International Journal of Computer Sciences and Engineering, vol. 6, no. 8, pp. 503-513, 2018.
- P. Sivanandhini, and J. Prakash, "Crop Yield Prediction Analysis using Feed Forward and Recurrent Neural Network," International Journal of Innovative Science and Research Technology, vol. 5, no. 5, pp. 1092-1096, 2020.
- N. Nandhini, and J. G. Shankar, "Prediction of crop growth using machine learning based on seed," Ictact journal on soft computing, vol. 11, no. 01, 2020.
- International Journal of Computer Sciences and Engineering, vol. 6, no. 8, pp. 503-513, 2018. P. Tiwari and P. Shukla, "Crop yield prediction by modified convolutional neural network and geographical indexes."
- J. Prakash and P. Sivanandhini, "Crop Yield Prediction Analysis Using Feed Forward and Recurrent Neural Network," International Journal of Innovative Science and Research Technology, vol. 5, no. 5, 2020, pp. 1092–1096.
- "Prediction of crop growth using machine learning based on seed," Ictact journal on soft computing, vol. 11, no. 01, 2020; N. Nandhini and J. G. Shankar .
- Priya and U. Muthaiah Balamurugan, M. employing a machine learning algorithm to predict agricultural yield. Engineering Science Research Technology International Journal .