Machine Learning Based Oilseed Crop Yield Forecasting with Recommendation System Using Organic Manures in Tamil Nadu

Mithra C.
Research Scholar, Dept. of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Annamalai Nagar, Tamil Nadu, India

A. Suhasini
Professor, Dept. of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Annamalai Nagar, Tamil Nadu, India

Abstract

Agriculture is critical to India's economy. The most serious threat to food stability is population growth. Population growth increases demand, forcing farmers to produce more to keep up. Crop yield prediction technology can help ranchers increase productivity and efficiency. For the cultivation of oilseed crop yield, proper manure rates are required. When nutrients are scarce or over-fertilized, yields suffer and the environmental burden increases. To address these concerns, our proposed work employs machine learning techniques to predict the yield of oilseed crops grown with organic manure, as well as the amount and type of agricultural manure to be used for a specific crop in various districts of Tamil Nadu. The training set includes actual yield data from 1961 to 2007, while the validation set includes data from 2008 to 2019. The results of the proposed algorithm are compared to those of other machine learning algorithms, namely bagging, random forest, linear regression, and naive bayes which have accuracy rates of 98.5%, 96.5%, 94.5%, and 92.5%, respectively. According to the study, bagging (Bootstrap Aggregation) outperforms other algorithms for crop yield prediction, whereas boosting algorithms outperform other algorithms for recommendation systems to determine which crop to plant, which type of organic manure to use, and how much manure to use in a specific area and time.

Keywords: *Crop Yield Forecasting, Organic manures, Dose of manures, Data mining, Machine learning, Recommendation system.*

1. INTRODUCTION

Agriculture is widely regarded as the primary source of employment for the majority of India's population, with roughly 70% of the country's population living in rural areas and relying on agriculture for a living. Furthermore, when compared to other occupations, the agricultural sector contributes nearly 50% of India's GDP [22]. Agriculture and related industries account for the majority of the local economy in Tamil Nadu. Despite the fact that 93% of farmers are small and marginal, more than two-thirds of rural households in the state rely heavily on agriculture for a living. The world's largest producer of oilseeds is India. Various oilseeds are grown, accounting for approximately 12% of the country's total cropped area. Groundnut accounts for 28% of total oil seed production in Tamil Nadu and 29% of total oil seed production in the country. Due to rising demand for oilseeds and supply stagnation, an approximate fertilizer for the crop is suggested to increase yield. Fertilizers have greatly increased agricultural productivity, assisting India in transitioning from a food-scarce to a food-sufficient region. It is widely acknowledged that increased

reliance on agricultural chemicals such as inorganic fertilizers has had negative environmental consequences as well as decreased soil fertility. The use of chemical fertilizers alone increased crop yield in the initial years but adversely affect sustainability [14,15]. Organic manures such as farm yard manure, vermicompost poultry manure, pressmud, sheep manure, and crop residues are thought to be a nutrient storehouse for plant growth[12]. Organic manures improve soil, physical, chemical, and biological properties, which have a direct impact on moisture retention, nutrient conservation, and other soil properties that improve fertility, productivity, and water-holding capacity [16]. It takes a long time to use experimental crops in the field to predict yields, making it difficult for ranchers to choose the best crop at any given time. To address these issues, traditional (experimental) agricultural yield methods are being phased out in favour of computerised yield prediction methods [11]. Data mining is the process of sorting through large datasets to identify patterns and relationships that can aid in data analysis and resolution. Data mining techniques are important in the production of oilseed crops because they provide access to large amounts of data while also providing ideas for forecasting future trends [13].

Furthermore, farmers can use data analysis to help them with real-time crop health monitoring, predictive analytics for future yields, and resource management decisions based on proven trends. Machine learning is a sophisticated predictive analytics tool that has been widely used to create decision support systems in a variety of fields, including finance, marketing, and, most recently, agriculture [21]. Machine learning in agriculture is promising because it allows farmers, policymakers, and other agricultural stakeholders to make more informed decisions. Machine learning applications in agriculture will increase the efficiency with which resources are used for cultivation, harvesting, and livestock production [4,18]. Predicting crop yield is one of agriculture's most difficult tasks [19,20]. It is essential in global, regional, and field decisions. Decisions about soil, weather, the environment, and crop fields. Crop yield is predicted using soil, meteorological, environmental, and crop parameters. As a result, a decision support system based on Graphical User Interface (GUI) is developed to assist farmers in determining the type and quantity of manures to use for a specific crop at a specific time in the near future [14]. Model evaluation is accomplished by analyzing the performance of a machine learning model, as well as its strengths and weaknesses, using various evaluation metrics. Precision, recall, F-score, and specificity are some of the metrics used to assess model performance.

In this paper, we look at four machine-learning models that show how future crop yield can be predicted using attributes such as temperature, humidity, soil type, and area, among others, to improve crop yield prediction accuracy. The boosting algorithm helps farmers determine the type and amount of organic manure to use for a specific crop. It is represented using a graphical user interface to allow ranchers to identify oilseed crop yield information. The current paper discusses data collection, data preprocessing, and feature selection before comparing them to four machine learning algorithms: bagging, random forest, linear regression, and nave bayes to determine which algorithm is best suited for crop yield prediction using organic manure. Many trials were conducted for each of the four distinct algorithms to determine whether or not the accuracy rates had changed.

2. Related Work

[20] tested four machine learning algorithms to predict crop yield for potatoes, sunflower, spring barley, and soft wheat: ridge regression, K-NN, support vector regression, and Gradient-Boosted decision trees. These methods belong to different classes of algorithms based on how they learn the relationships between features and labels. Furthermore, the few feature selection algorithms used for machine learning models to predict as accurately as possible are random forest, recursive feature elimination with LASSO, and mutual information. [10] conducted a meta-analysis in China to assess the impact of manure application on crop yield and soil attributes. Manure application increased yield by 7.6% when compared to inorganic fertilisers, and productivity increased with longer-term manure application, increasing by 27.7% when the applications lasted more than 10 years. As a result, an Ordinary Least Squares (OLS) regression analysis was performed to estimate the interaction between soil parameter changes and yield in soils with added manure. [7] used the proposed machine learning algorithm, boosted regression tree, to compare manure and synthetic fertiliser, accounting for 39% of manure, 21% of synthetic fertiliser, and 40% of soil properties providing variation in relative yield. These findings suggest that manure application is a viable strategy for regulating crop yields due to the improvement in soil fertility.

[5] proposed and compared a deep-learning-based RNN-LSTM model to other models, including ANN, RF, and multi-variate Linear regression, and demonstrated that the RNN-LSTM model outperforms other models for crop yield prediction. [28] proposed that four types of machine learning methods successfully predicted biomass bio-oil yield using biomass composition and pyrolysis conditions. In terms of prediction, the random forest model outperformed the SVM, DT, and MLR. According to the findings, optimal parameters selected using a genetic algorithm-based approach have a significant impact on bio-oil yield.

[27] compared a multi-layer perceptron based on deep learning to other machine learning algorithms such as random forest, decision tree, K-nearest neighbour, Ordinary Least Squares, and Support vector regression. In addition, hyperparameter optimization was performed to improve yield estimation. The accuracy of yield estimation provided by DLMMLP is satisfactory. [25] examined the performance of three machine learning algorithms: linear regression, decision tree, and random forest, and discovered that the random forest model produces significant R2 and MSE values. As a result, the random forest algorithm predicts crop yields well. With 33 experimental data, [8] suggested developing a traditional ANN model as well as a novel least squares Support vector Machine (LS-SVM) model. The results show that the LS-SVM model outperforms the traditional ANN model in terms of predicting performance and robustness for the modelling study of the cattle manure pyrolysis process and other similar processes. SVR, RF, Extreme learning machine, ANN, and DNN were the five regression models used in [9]. When compared to other algorithms, the DNN and RF models produced promising results.

[17] proposed ridge and lasso regression, CART, KNN, SVM, XGB, and RF as fine-tuned machine learning models. R2, RMSE, and MAE metrics were used to compare the algorithm statistically. Using a grid search approach, the best-performing model (RF) was fine-tuned once more for the bias-variance

trade-off. RF performed the best in terms of goodness-of-fit. The RF method was then used to pick out the key variables and interactions. [24] proposed an agribot - an intelligent interactive interface that uses data mining and machine learning techniques to help farmers decide which crop to grow in a given year. The NLP technique is used to implement it. The system is designed in such a way that farmer input queries about the agricultural context can be received in audio format, making farmer interaction more user-friendly. KNN, DT, and RF were the three machine learning techniques used. When compared to other machine learning algorithms, Random Forest outperforms them. [3] proposed crop yield prediction in relation to rainfall. The proposed method of evaluation outperformed other existing methods because it evaluates all regression techniques, including linear regression, polynomial regression, support vector regression, decision tree regression, and Extreme Gradient Boosting regression, for two crops of four individual states. The MSE technique is used to validate the model's performance. [26] proposed a yield prediction model for rice and wheat crops that combined an ecological distance algorithm with crop yield predictors. When the proposed model was compared to the existing algorithm, the intelligent model based on EDA produced better prediction accuracy. [6] proposed a model for forecasting wheat crop yield based on data mining classification algorithms and stepwise linear regression. WEKA and SPSS tools were used in this model prediction. Weather and crop data were used as factors. The study found that MLP and additive regression produced better results than other algorithms. [1] proposed four machine learning algorithms, LR, EN, KNN, and SVR, which were used to predict potato tuber yield from soil and crop properties proximal sensing data. The SVR models outperformed all other models.

[23] demonstrated that applying manure can be an effective way of restoring microbial biomass loss caused by intensive NPK application. Variations in response, however, are determined by specific manure types, application rates, local climate, and inherent soil properties. The results of the RF models revealed that the most important factors controlling microbial biomass response to manure application were likely manure type, application rate, and soil initial properties. [2,18] found that combining poultry manure with tillage increased grain output by 39.5% when compared to tillage alone. Manure-Zero tillage methods increased grain yields by 15% when compared to manure-mechanized tillage methods. As a result, the organic manure application outperforms other algorithms on the field.

3. Methodology

Figure 3 depicts the overall architecture of the proposed model, which employs four machine learning algorithms: bootstrap aggregation, RF, LR, NB, and DT. It was also compared to other algorithms to see which one performed the best. Pycharm Community Edition 2022.2.3.64 was used for the study. The Bootstrap Aggregation algorithm was used to create a recommendation system for end users using oilseed yield data obtained from the Official Government Website, which included soil, meteorological, yield, and organic manure data, among other things. The algorithm divides yield into two broad categories: high yield and low yield. The final decision is made after developing a model based on anticipated targets. Crop yield forecasting enables more precise production planning and decision making. A recommendation system (GUI) is also included

in the proposed model to assist farmers in determining the best manure ratio and crop type for a given season. They provide a very useful framework for outlining options and investigating the potential consequences of those options.

3.1 Oilseed crop

India is the fourth-largest producer of oilseeds in the world. It accounts for 20.8% of total global cultivable land and 10% of total global production. Groundnut, safflower, sunflower, sesame, mustard, and castor are among the oilseeds grown in the country. Rainfed farming by small farmers accounts for nearly 72% of the oilseed area, resulting in low productivity. However, by introducing cutting-edge production technologies, a breakthrough in oilseed production was achieved. As a result, production of oilseeds increased from 108.3 lakh tonnes in 1985-86 to 365.65 tonnes in 2020-21. Over the last five years, India's oilseed production has increased. In 2020-2021, the country's output was 365.65 lakh tonnes, a 10% increase over the previous year.

3.1.1 Importance of organic manure

Organic manure will inevitably be used to meet crop nutrient needs in the future because it not only increases yield but also preserves the soil's physical, chemical, and biological qualities. Organic sources that can be incorporated into the soil are becoming increasingly scarce. Organic manure gradually mineralizes and releases vital minerals that have been locked up, improving soil fertility while also increasing crop output and quality [16].

3.1.2 Advantages of organic manures

Organic manure contains all of the nutrients that plants require, but only in small amounts. It improves the structure and texture of the soil as well as its ability to retain water. When compared to mineral fertilizers, it is less expensive. Furthermore, it helps to preserve oil by increasing fertility and productivity.

3.2 Agricultural Dataset

• The dataset for this study came from the following sources:

• The Department of Meteorological Centre India provides access to weather datasets such as temperature, humidity, and rainfall.

• The types and quantities of organic manure have been obtained for Agricultural University Departments.

• The weather atlas website is used to collect environmental parameters such as sunshine.

• Datasets for various oilseed yields are gathered from ICRISTAT, the Tamil Nadu Government Website (www.data.govt), and the University Department of Agriculture.

•Key environmental factors considered in this study include soil temperature, pH, rainfall, humidity, and the minimum and maximum temperatures of a specific location and area. Some agronomic parameters are also included, such as textures (red loamy, clay loam, deep red loam, and so on) and seasons. Furthermore, crop yield prediction takes into account manure types such as farmyard manure, poultry manure, sheep manure, vermicompost, neem seed cake, and so on, as well as quantity and NPK soil nutrient content.

This study considered the following crops:

- Castor
- Coconut
- Rapeseed
- Groundnut
- Safflower

- Other oilseed crops

3.2.2 Dataset Description

As input, data gathered from various sources is fed into the model. For the above oilseed crops, a set of data is initially collected in all districts of Tamil Nadu, including parameters such as state name, district name, area, productivity, organic manure, and so on. The information in this.csv file was gathered between 1961 and 2019. The final dataset contains 1012 records and 25 attributes.

3.3 Preprocessing

Preprocessing is required before applying any machine learning technique to a dataset. Raw data is frequently gathered from various sources. The raw data contains information that is incomplete, inconsistent, or out of date. As a result, before processing, redundant data must be filtered. The data series provided contains a large number of 'NA' values, which can be filtered in Python by replacing missing values with an average value. Outliers are removed using a robust scalar technique. The data is then transformed to make it more accessible. To ensure that all values fall within a study range, the final dataset is normalised. Equation 1 depicts the simplest normalisation (constant factor normalisation) formulae. This method is used to normalise data to a factor ranging from 0 to 1. Figure 1 represents the box plot representation of outlier elimination.

$$X^{|} = X/K \quad \text{---------1}$$

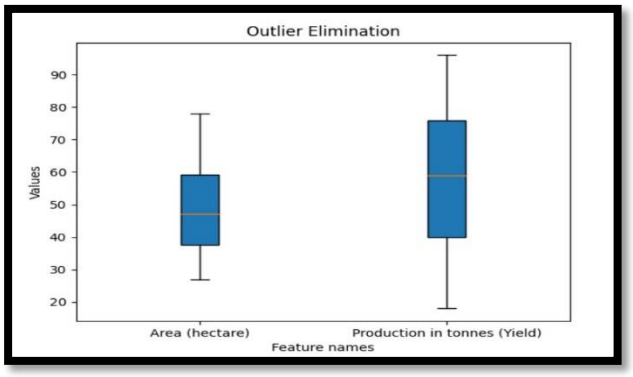
Where,

X denotes the raw value

X| denotes the normalized value

K is a numeric value

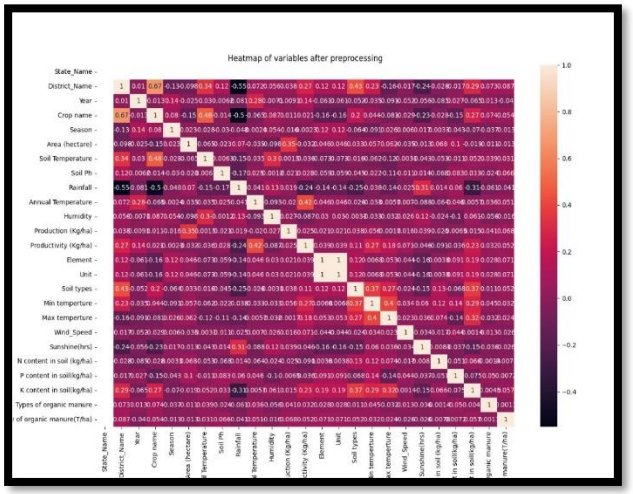
Fig 1: Box plot representation for outlier elimination



3.4 Data Analysis

After the raw data has been preprocessed, it must be ensured through inspection, cleansing, transformation, and designing processes to provide useful information and conclusions, as well as enable decision-making to move forward with the proper understanding of the dataset. When outliers are discovered in the data, a box plot graph must be created for easy understanding. Figure 2 shows the variable heatmap after preprocessing.

Fig 2: Heatmap of variables after preprocessing



3.5 Dimensionality reduction

High-level factors influencing prediction accuracy must be carefully selected in order to make accurate predictions. Many feature selection techniques are used, including LDA, PCA, and factor analysis. Factor analysis is the best choice for this study because it allows you to transform and compress the dataset while keeping only the most important features. This dataset contains a total of 25 features. The PCA technique was used to select 20 critical features in this case. To select 17 critical feature subsets, the LDA technique was used. The 13 critical feature subsets were chosen using the factor analysis technique. Among all of these dimensionality reduction techniques, factor analysis feature subsets provide the most accurate results. The optimal feature subset was determined by feeding these feature subsets into the bagging method.

3.6 Training and testing model

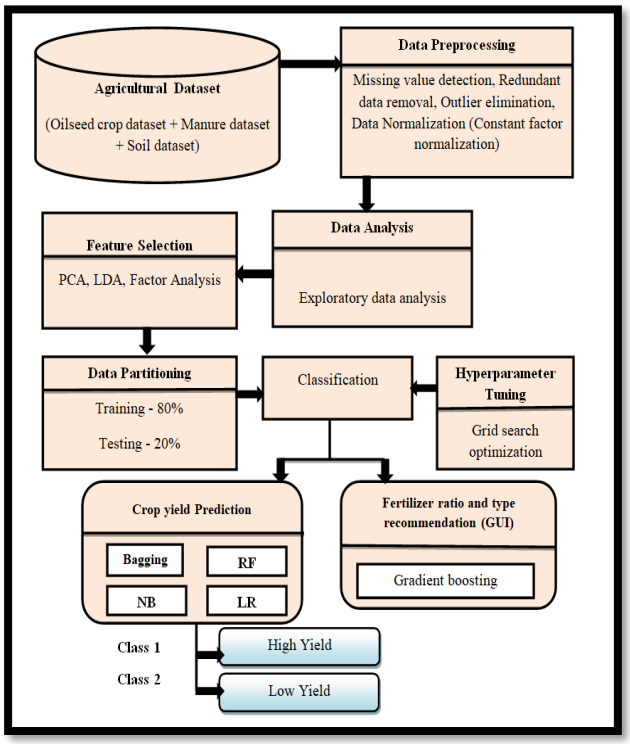
The dataset can be divided into training and testing sets during the preprocessing phase. We divided the dataset into 80% and 20% training and testing groups, respectively. This stage of model development is crucial. The training dataset is used to build the model, and the testing dataset is used to validate it. As an outcome, we used the training dataset to fit the model. As a result, we fit the model with training data and test its accuracy with testing data.3.7 Prediction algorithm After the data has been separated, the model is created and trained. Creating a machine learning model to understand the pattern necessitates the use of a machine learning algorithm and training data. We use a number of machine learning algorithms in this case, all of which are well-known supervised learning algorithms with clear and concise representations.

Comparison of accuracy of the proposed model with existing ones

Table 1. Accuracy of proposed models

Models	Accuracy
Bagging	**98.5**
RF	96.5
LR	94.5
NB	92.5

Fig 3: Architecture diagram for oilseed crop yield prediction and manures recommendation system

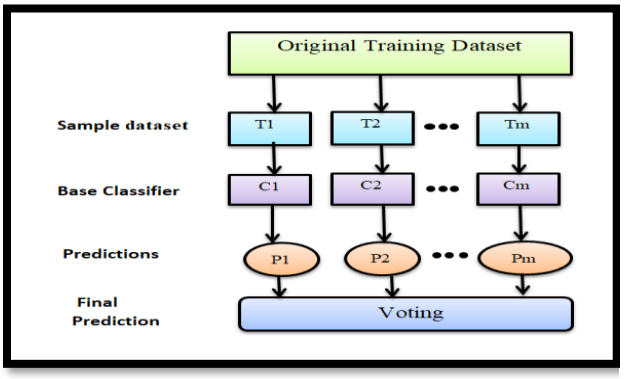


- Classification model for oilseed crop yield prediction

There are 1012 records in the oilseed crop yield dataset for six different crops. After preprocessing, the oilseed crop yield prediction decreases by 979 records. The training set is made up of 779 records, while the testing test is made up of the remaining 200 records. We created a machine learning model to forecast

oilseed crop yield. All of the proposed algorithms, including bagging, linear regression, and naive bayes classifiers, are compared. The bagging algorithm outperforms the others in terms of forecasting oilseed crop yield. Pycharm is a model training platform that uses machine learning algorithms.

Fig 4: Process flow of bagging algorithm



Bagging is a type of ensemble machine learning method that combines the results of multiple learners to improve performance. These algorithms operate by dividing the training set and running it through various machine learning models, then combining their predictions when they return to generate an overall prediction for each instance in the original data using the majority voting technique. Figure 4 depicts the bagging algorithm's process flow.

Algorithm:

The steps involved in developing a crop yield classification model.

Input: An experimental dataset containing weather, crop, soil, and manure information

Output: Crop Yield Forecasting using the experimental dataset

Method:

Step 1: Collection of data and feature analysis

a) Gather, organise, and format the data:

Using the model requires more than just raw data. To achieve the desired results, data must be collected, stored, and organised.

b) Examine and select features:

Following preprocessing, the data is evaluated to produce useful information and conclusions in order to proceed with proper knowledge of all variables. Following dimensionality reduction, the factor analysis method is used to select essential feature sets. The selected features are then processed using machine learning techniques..

Step 2: Divide the data into two groups

The training set has the most data and will be used to train the vast majority of the samples that will result in the yield. The training set includes nearly 80% of the samples collected. The final piece of data is used by the testing set to determine how well the system works.

Step 3: Training sets for classification

The complexity of the problem will determine the model system, and the structure must be chosen accordingly. The construction, design, and structure of the training set can be changed during training.

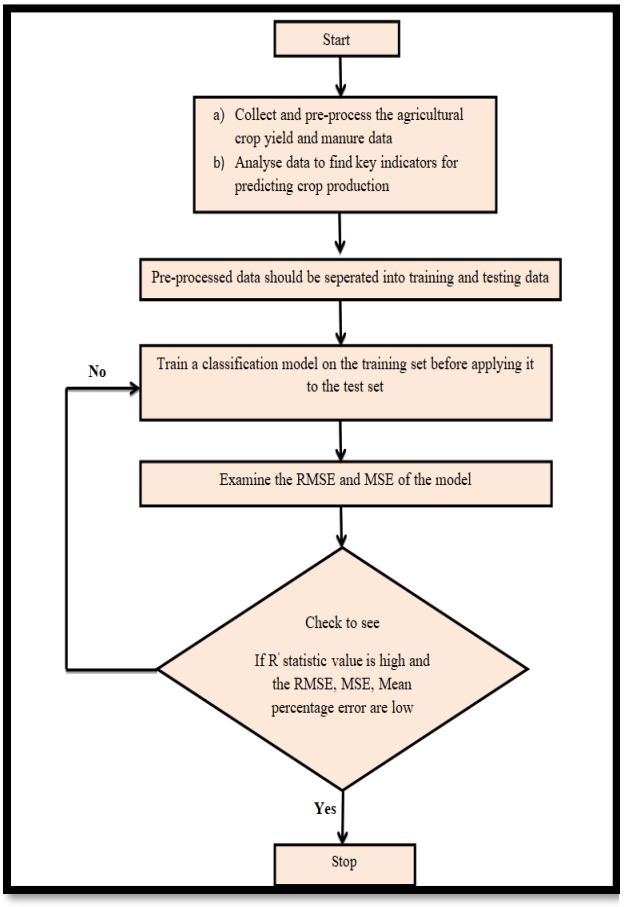
Step 4: Determine the RMSE, R2 statistic and MSE values for each model

Calculate the MSE and RMSE values by repeating the trained classification model on the test set. Compare the results with different classification models. The model with different classification models. The best crop yield prediction model has the lowest MSE and RMSE values, as well as the highest R2 statistic value. The flow chart for the classification technique used to forecast crop yield is shown in Figure 5.

Step 5: Predict Yield

When new input is provided, the trained model is used to predict the output. The trained model was saved as a file in order to be estimated with new data. Before being tested on the testing dataset, these models were properly trained on the training dataset. This prediction model uses machine learning techniques to learn properties from training data in order to make accurate predictions.

Fig 5: Flow diagram for classification methodology for predicting oilseed crop yield



3.8 Prediction results

Figure 6 depicts a comparison of actual and predicted values for all crops in Tamil Nadu.

Fig.6 Comparison of actual value versus predicted value for all crops in Tamil Nadu

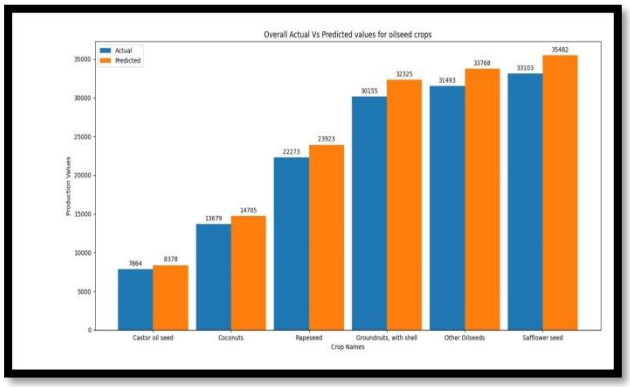


Table 2. Absolute error calculation for all crop yield prediction

Crop name	Actual value (ha)	Predicted value (ha)	Absolute error (ha)	Absolute error (kg/ha)
Castor	7864	8378	514	0.514
Coconut	13679	14705	1026	1.026
Rapeseed	22273	23923	1650	1.65
Groundnut	30155	32325	2170	2.17
Other oilseeds	31493	33768	2275	2.27
Safflower	33103	35482	2379	2.37

3.9 Error calculation for various classification algorithms

The following table shows the mean square error and root mean squared error formulae,

Table 3. Formulae for error calculation

MSE	RMSE
$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}}$
i- variable i	i- variable i
n - number of data points	N -number of non-missing data pointsY_i
Y_i-observed values	Y_i - actual observations time series
$\hat{Y}_i$- predicted values	$\hat{Y}_i$ - estimated time series

The mean square error for all machine learning algorithms is depicted in figure 7 below,

Fig.7. MSE of proposed models

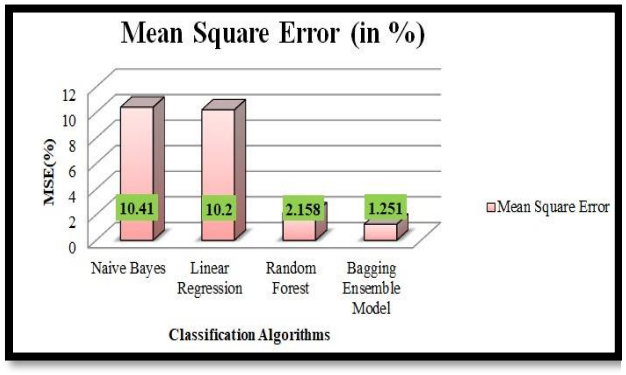
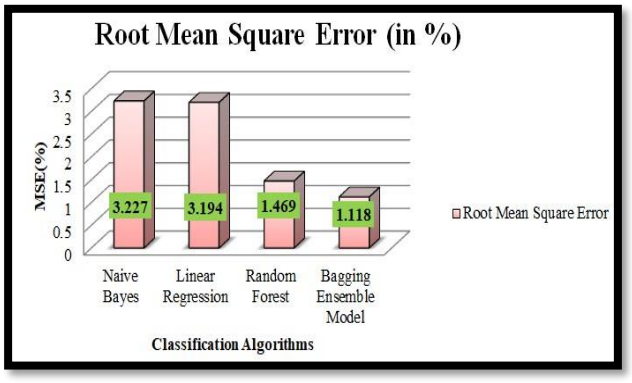


Figure 8 depicts the root mean square error for all the machine learning algorithms

Fig.8. RMSE of proposed models



3.9.1 Comparison with different models

We obtained a 98.5% accuracy rate, indicating that this model predicts yield more accurately. In terms of accuracy, the bagging technique

(Bootstrap Aggregation) outperformed other models. This is due to model and structure changes made during training. Table 1 compares the accuracy of various proposed algorithms, while Figure 9 depicts a graphical comparison of machine learning model accuracy.

4. Evaluation Metrics

Table 4: Formulae for evaluation metrics

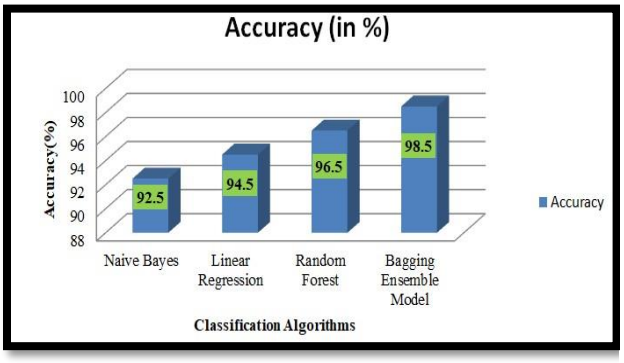
Accuracy	Recall	Precision	F - Measure	Specificity
TP + TN/TP + TN+FP+FN	TP/TP+FN	TP/TP + FP	2 * Prec * Recall/ Prec +Recall	TN/TN + FP

Where, TP represents True positive, TN represents True Negative, FP represents False Positive, and FN represents False Negatives. There are numerous methods for measuring performance. Accuracy, precision, recall, and F-measure are some of the most popular metrics.

4.1 Accuracy

The accuracy of the classifier is simply how often it predicts correctly. It is calculated by dividing the total number of predictions by the number of correct predictions. The accuracy of all machine learning algorithms is compared in Figure 9.

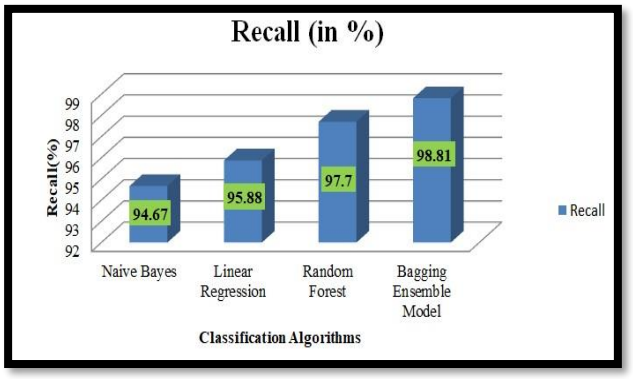
Fig.9. Comparison of accuracy for all proposed models



4.2 Recall

The recall is calculated as the ratio of correct detections to total positive samples. Figure 10 compares the recall values of all machine learning.

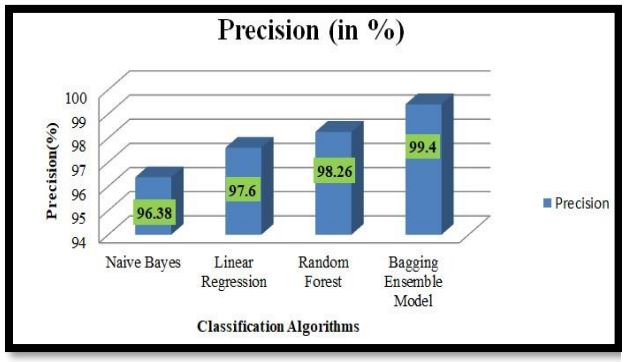
Fig.10 Comparison of recall for all proposed models



4.3 Precision

Precision is defined for a given label as the ratio of true positives to predicted positives. The precision values of all machine learning algorithms are compared in Figure 11.

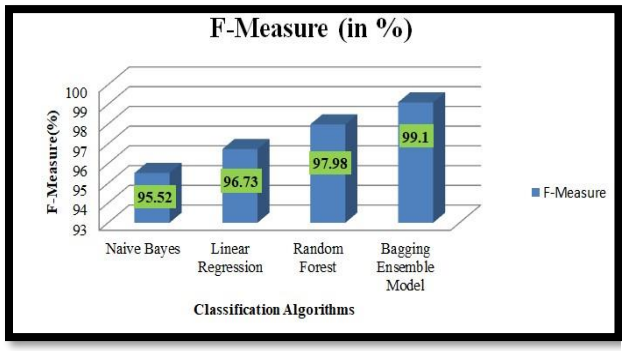
Fig.11. Comparison of precision for all proposed models



4.4 F-measure

The F-measure is the harmonic mean of precision and recall. Figure 12 compares the F-measure values of all machine learning algorithms

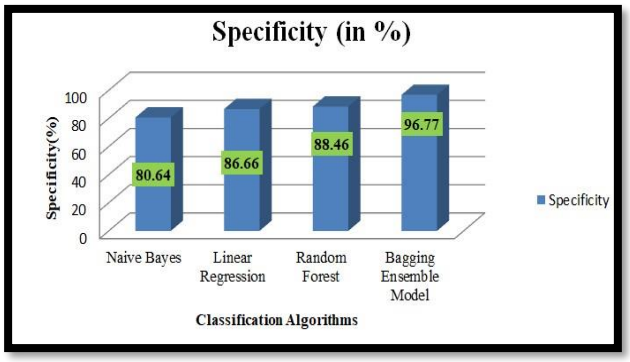
Fig.12. Comparison of F-measure for all proposed models



4.5 Specificity

Specificity is defined as the ratio of true negatives to the total number of true negatives and false positives. Figure 13 compares the specificity values of all machine learning algorithms

Fig.13 Comparison of specificity for all proposed models



4.6 Execution time for all machine learning algorithms

Figure 14 compares the training execution time of the proposed algorithms,

Fig.14. Comparison of training execution time for all proposed models

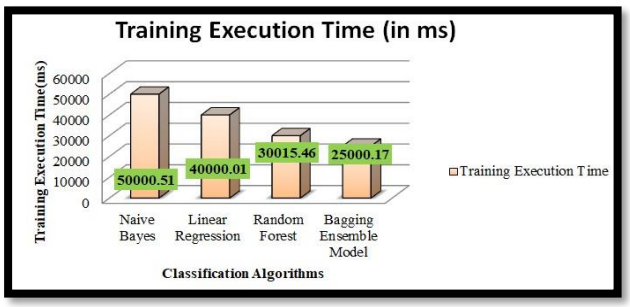
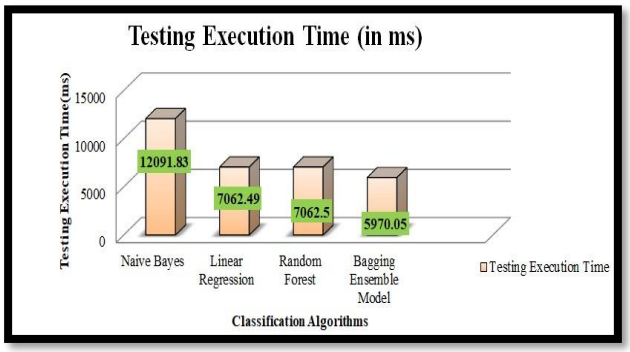


Fig.15. Comparison of testing execution time for proposed models



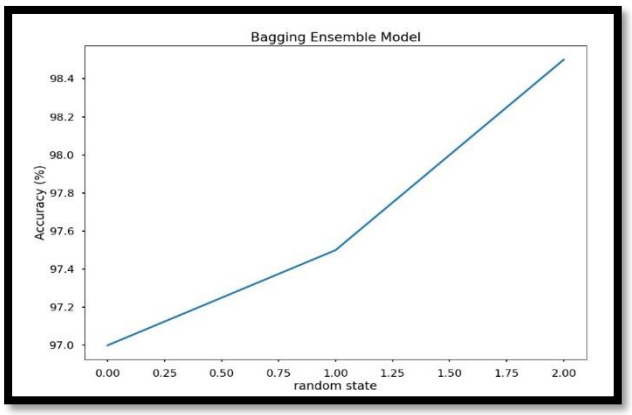
5. Results and Discussion

5.1 Overall observations of the proposed algorithms

5.1.1 Observations on Bagging:

The parameters chosen for the study are estimator, n estimators, max samples, max features, bootstrap, bootstrap features, oob score, warm start, n jobs, random state, verbose, and base estimator range. During the research, the following observation was made. The accuracy improves as the random state parameter value is increased. Many trails were constructed. For the parameter "random state," with an assumed value range of 0 to 2, the top three trials were considered. Iterating through different random state range values improves model performance from 97% to 98.5%. The accuracy begins to deteriorate below the value of 1 for the random state; again, if the random state is changed, the values seen will vary. Estimator, n estimators, max samples, max features, bootstrap, bootstrap features, oob score, warm start, n jobs, random state, verbose, and base estimator range are the parameters chosen for the study. The following observation was made during the research. As the random state parameter value is increased, the accuracy improves. Many trails have been built. The top three trials were considered for the parameter "random state," with an assumed value range of 0 to 2. Model performance improves from 97% to 98.5% when iterating through different random state range values. The random state's accuracy begins to deteriorate below the value of 1; again, if the random state is changed, the values seen will vary.

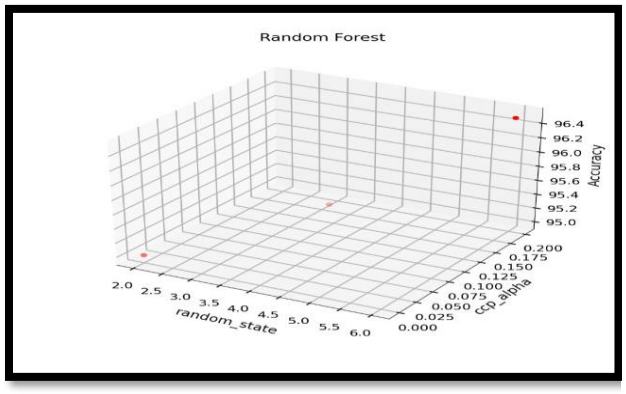
Fig 16. Observations on the parameter for Bagging Ensemble Model



5.1.2 Observations on Random Forest:

The study's parameters include n estimators, criterion, max depth, min samples split, min samples leaf, min weight fraction leaf, max features, max-leaf nodes, min impurity decrease, bootstrap, oob score, n jobs, random state, verbose, warm start, class weight, ccp alpha, and max samples using the sklearn library. During the research, the following observation was made. Accuracy tends to increase as the value for random state decreases and the number of estimators increases. The resulting accuracies were 95%, 95.5%, and 96.5% for the parameters "random state" and "ccp alpha" with assumed values of 2,4,6 and 0.0, 0.1, 0.2, respectively. Figure 17 depicts the observations on random forest algorithm parameters.

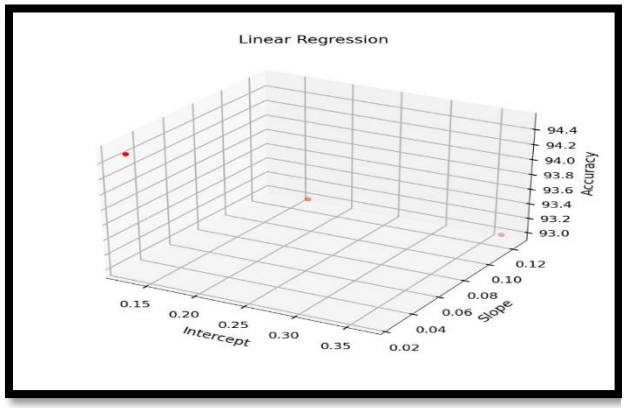
Fig.17 Observations on the parameter for Random Forest



5.1.3 Observations on Linear regression:

An independent variable, two dependent variables (x1 and x2), the slope (m), and the intercept are among the study's parameters. As the slope (m) and intercept values decrease, so does the accuracy. Three trials were performed with varying intercept values of 0.36, 0.21, and 0.12 and slope(m) values of 0.12, 0.08, and 0.02 decreased, the accuracy increased while the other parameters remained constant, and the resulting accuracies were 93%, 93.5%, and 94.5%, respectively. Figure 18 depicts the observations on linear regression algorithm parameters.

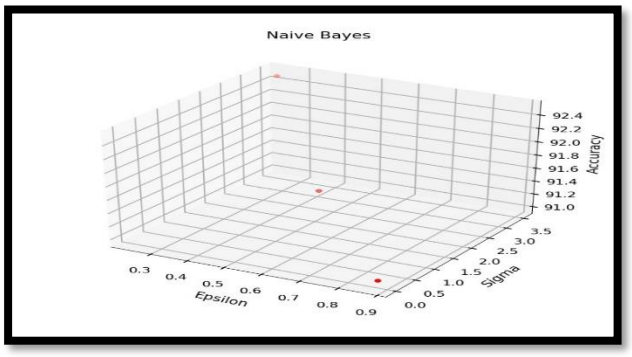
Fig.18 Observations on the parameter for Linear Regression



5.1.4 Observations on Naive Bayes:

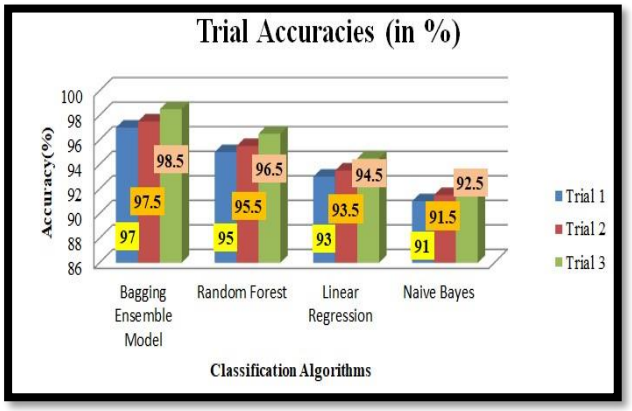
Among the parameters taken into account for the study are nb alpha, priors, smoothing, epsilon, sigma, and theta. Many trials were carried out. The top three trials were chosen. If the values of the parameter "sigma" increase with varying values of 0, 1.7, 3.5, and similarly for the parameter "epsilon," if the values decrease with varying values of 0.87,0.54,0.23, while remaining constant for the other parameters. The resulting accuracies of the three trials were 91%, 91.5%, and 92.5%. Figure 19 depicts the parameter observations for the naive bayes algorithm.

Fig.19 Observations on the parameter for Naive Bayes



The figure 20 depicts the observations of trial accuracies for proposed models,

Fig.20 Trail accuracies for proposed models

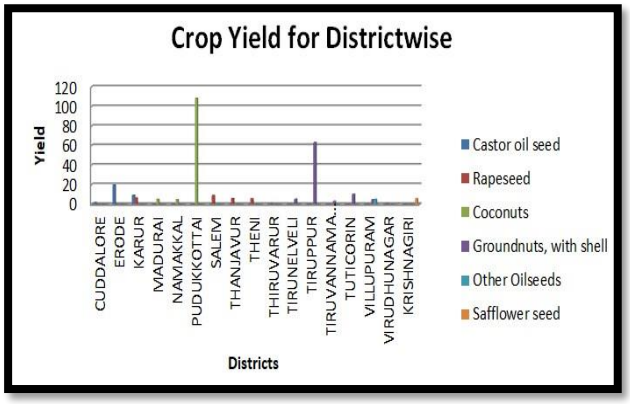


5.1.6 Discussion based on District wise crop yield

The goal of this paper is to comprehend the analysis of location-specific oilseed crop yields, which will be handled by a machine learning algorithm. For this study, a dataset in.csv format was considered. In this scenario, the training test consumes 80% of the data while the validation set consumes 20%. The model's accuracy was determined after successful training and testing, indicating how well the model performed in forecasting the yield. Figure 22 depicts a graphical user interface for predicting crop yield in the future. Figure 21 depicts a summary of all oil seed crop production districts in Tamil Nadu. According to the statistics collected between 1961 and 2019,

- Erode has a higher proportion of castor oilseed production
- Salem has a higher ratio of rapeseed production
- Pudukottai has a higher proportion of coconut proportion
- Tirupur has a greater proportion of groundnut oilseed production
- Villupuram has a larger proportion of other oilseed production
- Krishnagiri has a larger proportion of safflower production

Fig.21 District-wise crop yield statistics



5.1.7 Recommendation system for manure and oilseed crop yield

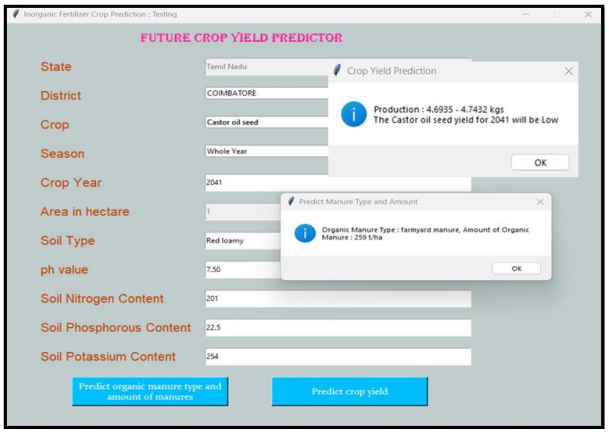
5.1.8 GUI Creation

The study helps farmers decide which crop to grow in a specific area at a specific time, as well as whether or not it will be profitable during forecasts. It also indicates low or high yield with ranges to help ranchers or end users make successful selections while saving time and accuracy. Users can share the district name, state name, crop, season, and crop year to predict area, soil type, pH value, and soil nitrogen, phosphorous, and potassium content. After entering these attribute values, the user can use the "Predict crop yield range" button to predict the future yield of a specific crop with a high or low yield rate. This suggestion system's "Predict organic manure type and amount of manure" button also assists users in forecasting separate quantities of NPK to be taken for a specific crop as well as the manure type to be used for a specific region. The result is calculated by taking into account a range of values based on the average of all prediction errors for each crop. Based on the prediction error, the formula presented below is used to calculate the yield range for each crop.

Predicted value ± Predicted error	------2

In figure 16, the crop production result range for the Coimbatore district's castor oilseed crop is calculated for the entire year based on the average of prediction errors of all castor crops in a specific location, and the low and high rate is estimated by taking the mean of each crop based on the records in the dataset. This system also helps with manure type and quantity recommendations for oilseed crops in a specific area at a specific time. The crop is considered low yielding if the prediction value is less than the mean score, and high yielding if it is greater than the mean.

Fig.22. Recommendation systems for crop yield and predictor



The visualization of GUI data can also be accomplished by plotting yield variables with varying parameters. Data visualisation, such as graphs or figures, helps with idea capture and comprehension. Figure 22 depicts the primary goal of GUI data visualisation. The prediction module facilitates the discovery of patterns, correlations, and outliers in large datasets. The graph above depicts the district's relationship with yield.

6. Conclusions

The researchers looked at crop yield forecasting algorithms that took temperature, season, and location into account. Rainfall, temperature, and other variables such as season, location, and organic manure data can be used to forecast yield in a specific district. When all factors are considered, the bootstrap aggregation technique outperforms all others. The number of parameters in the dataset is increased to improve accuracy. When compared to other prediction algorithms such as random forest, linear regression, and naive bayes, bagging is found to be the best. Because the database contains a much larger number of variables, the predictions are more accurate. This work will assist farmers in reducing risk and maximising crop yields in order to improve their agricultural resources.

We forecasted future crop yields using soil test results and organic manure dosage in this study. We have also developed a recommendation system for farmers to determine the best crop to cultivate in the coming season, as well as recommendations for ranchers on manure type and quantity. This will not only assist farmers in determining the best crop to cultivate for the upcoming season, but it will also aid in bridging technological and agricultural divides. Our work is limited in that yield is only implemented in 30 districts of Tamil Nadu and not in other states. Our project's future work aims to include regional languages such as Tamil, Telegu, Hindu, Kannada, Malayalam, and others in the graphical user interface that benefits farmers across the country. Furthermore, Natural Language Processing (NLP) can be used to collect farmer queries via voice mode and provide the desired result to ranchers via a graphical user interface (GUI). This voice mode query system makes it simple

for uneducated farmers to access the recommendation system.

Table 5. Abbreviations

S. No	Name	Abbreviation
1	NPK	Nitrogen Phosphorus Potassium
2	RMSE	Root Mean Squared Error
3	MSE	Mean Squared Error
4	RF	Random Forest
5	LR	Linear Regression
6	SVM	Support Vector Machine
7	KNN	K- Nearest Neighbors
8	MAE	Mean Absolute Error
9	API	Application Programming Interface
10	LASSO	Least Absolute Shrinkage and Selection Operator
11	GBRT	Gradient Boosted Regression Trees
12	ANN	Artificial Neural Network
13	GUI	Graphical User Interface
14	ICRISTAT	International Crops Research Institute for the Semi-Arid Tropics
15	LDA	Linear Discriminant Analysis
16	ML	Machine Learning
17	DT	Decision Tree

References

[1] Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy, 10(7). https://doi.org/10.3390/AGRONOMY10071046

[2] Agbede, T. M., & Ojeniyi, S. O. (2009). Tillage and poultry manure effects on soil fertility and sorghum yield in southwestern Nigeria. Soil and Tillage Research, 104(1), 74–81. https://doi.org/10.1016/j.still.2008.12.014

[3] Antony, B. (2021). Prediction of the production of crops with respect to rainfall. Environmental Research, 202(June), 111624. https://doi.org/10.1016/j.envres.2021.111624

[4] Aworka, R., Cedric, L. S., Adoni, W. Y. H., Zoueu, J. T., Mutombo, F. K., Kimpolo, C. L. M., Nahhal, T., & Krichen, M. (2022). Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries. Smart Agricultural Technology, 2(March), 100048. https://doi.org/10.1016/j.atech.2022.100048

[5] Bali, N., & Singla, A. (2021). Deep Learning Based Wheat Crop Yield Prediction Model in Punjab Region of North India. Applied Artificial Intelligence, 35(15), 1304–1328. https://doi.org/10.1080/08839514.2021.1976091

[6] Bhojani, S. H., & Bhatt, N. (2020). Wheat crop yield prediction using new activation functions in neural network. Neural Computing and Applications, 32(17), 13941–13951.

https://doi.org/10.1007/s00521-020-04797-8

[7] Cai, A., Xu, M., Wang, B., Zhang, W., Liang, G., Hou, E., & Luo, Y. (2019). Manure acts as a better fertilizer for increasing crop yields than synthetic fertilizer does by improving soil fertility. Soil and Tillage Research, 189(February 2018), 168–175. https://doi.org/10.1016/j.still.2018.12.022

[8] Cao, H., Xin, Y., & Yuan, Q. (2016). Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach. Bioresource Technology, 202, 158–164. https://doi.org/10.1016/j.biortech.2015.12.024

[9] Chergui, N. (2022). Durum wheat yield forecasting using machine learning. Artificial Intelligence in Agriculture, 6, 156–166. https://doi.org/10.1016/j.aiia.2022.09.003

[10] Du, Y., Cui, B., zhang, Q., Wang, Z., Sun, J., & Niu, W. (2020). Effects of manure fertilizer on crop yield and soil properties in China: A meta-analysis. Catena, 193(April). https://doi.org/10.1016/j.catena.2020.104617

[11] Feng, P., Wang, B., Liu, D. L., Waters, C., Xiao, D., Shi, L., & Yu, Q. (2020). Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agricultural and Forest Meteorology, 285–286(January), 107922. https://doi.org/10.1016/j.agrformet.2020.107922

[12] Guo, L., Wu, G., Li, Y., Li, C., Liu, W., Meng, J., Liu, H., Yu, X., & Jiang, G. (2016). Effects of cattle manure compost combined with chemical fertilizer on topsoil organic matter, bulk density and earthworm activity in a wheat-maize rotation system in Eastern China. Soil and Tillage Research, 156, 140–147. https://doi.org/10.1016/j.still.2015.10.010

[13] Hammer, R. G., Sentelhas, P. C., & Mariano, J. C. Q. (2020). Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. Sugar Tech, 22(2), 216–225. https://doi.org/10.1007/s12355-019-00776-z

[14] Haq, Z. U., Ullah, H., Khan, M. N. A., Raza Naqvi, S., Ahad, A., & Amin, N. A. S. (2022). Comparative study of machine learning methods integrated with genetic algorithm and particle swarm optimization for bio-char yield prediction. Bioresource Technology, 363(August), 128008. https://doi.org/10.1016/j.biortech.2022.128008

[15] Jiang, W., Xing, Y., Wang, X., Liu, X., & Cui, Z. (2020). Developing a sustainable management strategy for quantitative estimation of optimum nitrogen fertilizer recommendation rates for maize in Northeast China. Sustainability (Switzerland), 12(7), 1–11. https://doi.org/10.3390/su12072607

[16] Luo, G., Li, L., Friman, V. P., Guo, J., Guo, S., Shen, Q., & Ling, N. (2018). Organic amendments increase crop yields by improving microbe-mediated soil functioning of agroecosystems: A meta-analysis. Soil Biology and Biochemistry, 124(May), 105–115. https://doi.org/10.1016/j.soilbio.2018.06.002

[17] Nayak, H. S., Silva, J. V., Parihar, C. M., Krupnik, T. J., Sena, D. R., Kakraliya, S. K., Jat, H. S., Sidhu, H. S., Sharma, P. C., Jat, M. L., & Sapkota, T. B. (2022).

Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in Northwest India. Field Crops Research, 287(July). https://doi.org/10.1016/j.fcr.2022.108640

[18] Obsie, E. Y., Qu, H., & Drummond, F. (2020). Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. Computers and Electronics in Agriculture, 178(September), 105778. https://doi.org/10.1016/j.compag.2020.105778

[19] Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. Agricultural Systems, 187(June 2020), 103016. https://doi.org/10.1016/j.agsy.2020.103016

[20] Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., & Athanasiadis, I. N. (2022). Machine learning for regional crop yield forecasting in Europe. Field Crops Research, 276(November 2021), 108377. https://doi.org/10.1016/j.fcr.2021.108377

[21] Prasad, N. R., Patel, N. R., & Danodia, A. (2021). Crop yield prediction in cotton for regional level using random forest approach. Spatial Information Research, 29(2), 195–206. https://doi.org/10.1007/s41324-020-00346-6

[22] Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction. IEEE Access, 9, 63406–63439. https://doi.org/10.1109/ACCESS.2021.3075159

[23] Ren, F., Sun, N., Xu, M., Zhang, X., Wu, L., & Xu, M. (2019). Changes in soil microbial biomass with manure application in cropping systems: A meta-analysis. Soil and Tillage Research, 194(January), 104291. https://doi.org/10.1016/j.still.2019.06.008

[24] Sawant, D., Jaiswal, A., Singh, J., & Shah, P. (2019). AgriBot - An intelligent interactive interface to assist farmers in agricultural activities. 2019 IEEE Bombay Section Signature Conference, IBSSC 2019, 2019Januar, 3–8. https://doi.org/10.1109/IBSSC47189.2019.8973066

[25] Singh Boori, M., Choudhary, K., Paringer, R., & Kupriyanov, A. (2022). Machine learning for yield prediction in Fergana valley, Central Asia. Journal of the Saudi Society of Agricultural Sciences, xxxx. https://doi.org/10.1016/j.jssas.2022.07.006

[26] Tian, L., Wang, C., Li, H., & Sun, H. (2020). Yield prediction model of rice and wheat crops based on ecological distance algorithm. Environmental Technology and Innovation, 20, 101132. https://doi.org/10.1016/j.eti.2020.101132

[27]Tripathi, A., Tiwari, R. K., & Tiwari, S. P. (2022). A deep learning multi-layer perceptron and remote sensing approach for soil health based crop yield estimation. International Journal of Applied Earth Observation and Geoinformation, 113(July), 102959. https://doi.org/10.1016/j.jag.2022.102959

[28] Ullah, Z., khan, M., Raza Naqvi, S., Farooq, W., Yang, H., Wang, S., & Vo, D.

V. N. (2021). A comparative study of machine learning methods for bio-oil yield prediction – A genetic algorithm-based features selection. Bioresource Technology, 335(May), 125292. https://doi.org/10.1016/j.biortech.2021.125292