Eco. Env. & Cons. 29 (April Suppl. Issue) : 2023; pp. (S358-S372)
Copyright@ EM International
ISSN 0971–765X

DOI No.: http://doi.org/10.53550/EEC.2023.v29i02s.059

Fertilizer type and quantity recommendation to increase oilseed crops yield prediction with inorganic fertilizers using machine learning algorithms

Mithra C[1] and A. Suhasini[2]

Department of Computer Science and Engineering, Faculty of Engineering and Technology (FEAT), Annamalai University, Annamalai Nagar, Tamil Nadu, India

(Received 16 October, 2022; Accepted 15 December, 2022)

ABSTRACT

Agriculture contributes significantly to India's economy. The most serious threat to food security is population growth. Population growth increases demand, forcing farmers to produce more to increase supply. Crop yield prediction technology can help farmers to increase their output. Optimal fertilizer dose are required for boosting oilseed crop yield cultivation. However, when nutrients are scarce or over-fertilization occurs, yields are considerably lowered and the environmental burden is increased. To address these issues, our proposed work employs machine learning techniques in the prediction of crop yield using inorganic fertilizer as well as the amount and type of agricultural fertilizer to be used for a specific crop in various districts of Tamil Nadu. Actual yield data from 1961 to 2007 is used as a training set, and data from 2008 to 2019 is used as a validation set. The results of the proposed algorithm are compared with those of the other machine learning algorithms namely random forest, linear regression, support vector machine, and naive bayes with an accuracy rate of 94%, 91.33%, 88.4% and 75.56% respectively are observed. According to the study, random forest results outperform other algorithms for crop yield prediction, and the decision tree algorithm works better for recommendation systems. The research also helps farmers by providing a recommendation system for determining which crop to plant and which type of inorganic fertilizer and how much quantity of fertilizer to use in a specific area and time. The proposed study also seeks to examine different observations for each method by changing parameters to see if the varying parameter influences the accuracy rate or not.

Key words : *Crop yield prediction, NPK fertilizer, Dose of fertilizer, Data mining, Machine learning, Recommendation system*

Introduction

In India, the main source of employment is in the agricultural industry and its allied businesses. Most of the people in rural areas still dependprimarily on agriculture as their source of survival and 82% of such farmers are marginal and small-scale (Kanuru *et al*., 2021). On the other hand, farmers are uninformed of the importance of cultivating crops in the proper season and location. In this case, increasing crop quality and yield is a major challenge and is possible by determining crop adaptability and yield by utilising a variety of production influencing factors (Kamath *et al*., 2021). Data mining is a method of gathering previously unknown anticipated information from large databases. Mining data aids in the analysis of future patterns and characteristics, allowing businesses to make better decisions. Data analy-

([1]Research Scholar, [2]Professor)

sis is the data to generate relevant insights and conclusions (Kamath *et al*., 2021). Agriculture has long been seen as a natural fit for big data (Hampannavar *et al*., 2018). Machine learning is important because it has a decision-making system for crop yield forecasting, which incorporates assisting with decisions on which crops to grow and what practices to use during the crop growth period (Kamath *et al*., 2021; Meng *et al*., 2021). One of the primary benefits of machine learning techniques is their ability to solve huge non-linear problems autonomously using datasets from multiple sources. Crop yield forecasting is one of agriculture's most difficult tasks. It is crucial in decision-making at global, regional, and field levels. Crop yield is predicted using soil, meteorological, environmental, and crop parameters (Gopal and Bhargavi, 2019). This type of prediction assists ranchers in making the correct decision at a right time. Therefore the maximization of the yield aids in the growth of the economy also. Fertilizers play a critical role in crop yield and have the potential to increase crop yields, and benefits farmer income and food protection (Sarkar *et al*., 2018). Fertilizer usage increases crop yields, but their overuse hardens the soil, diminishes fertility, strengthens insecticides, pollutes air and water, and emits greenhouse gases, posing health and environmental risks (Meza-Palacios *et al*., 2020). Therefore, soil nutrients namely nitrogen (N), phosphorus (P) and potassium (K) are highly necessary for plant growth (Kumari and Saritha, 2017). As a result, a recommendation system is created to assist farmers in determining the type and ratio of fertilizer to use for a specific crop at a specific time (Hampannavar *et al*., 2018). High levels of noise, inaccurate data, outliers, biases, and missing data can significantly lower a model's predictive ability. Hence, machine learning based models might estimate crop yield based on farm parameters, social factors, and meteorological inputs. Precision, recall, RMSE and MSE are the evaluation metrics used to evaluate the prediction of oilseed crop yield.

In this study, we examine four machine learning models that demonstrate how future crop yield can be predicted using attributes such as humidity, temperature, soil type, area, and so on to improve crop yield prediction accuracy. The present paper mentions data collection, pre-processing, and feature selection and compares it to four machine learning algorithms namely RF, LR, SVM, and NB to determine which algorithm is best suited for crop yield prediction using inorganic fertilizer. The work also recommends the type and quantity of inorganic fertilizer for a specific season of oilseed crops. In addition, multiple attempts were performed for each of the four different algorithms to determine whether there is any change in the accuracy rates or not.

Literature Survey

Bondre and Mahagaonkar, (2019) proposed a machine learning algorithm namely SVM and RF to predict crop yield and recommend fertilizer for agricultural land with the help of yield data, location and fertilizer data. Naresh *et al.*, (2020) focused mainly on crop yield prediction using naive bayes algorithm and KNN for fertilizer recommendation based on features like temperature, rainfall, soil features, etc recommend fertilizer ratio. Divya *et al*., (2019) proposed crop suitability and fertilizers recommendation using a data mining algorithm. They used NPK contents for fertilizer suggestions for wheat crops and used an ontology-based recommendation system. This work helps in recommending suitable fertilizers and price predictions for each inorganic fertilizer. For evaluation, the precision metric is used.

Bharath *et al*., (2022) analyzed crop yield prediction and fertilizer usage based on using the KNN algorithm. The attributes like soil nutrients (NPK) and climatic variables like temperature, and rainfall were selected for the prediction of yield of crops and in addition, they also predicted the revenue expenses for the selected crops. Manjunath *et al*., (2020) designed a system in the form of android- based application that helps third-party users to predict the yield of crops based on weather and soil data. Multiple Linear Regression is the most suitable technique for the prediction of crop yield.

Kuturu *et al*., (2020) suggested a machine learning model named RF helps in predicting crop yield based on the attributes like soil type, temperature, humidity, water level, spacing depth, soil pH, season, and fertilizer. The performance of the system was evaluated by a few metrics namely MSE, MAE, R squared, RMSE, and accuracy to predict the yield of the crop. Mounika *et al*., (2021) investigated various related attributes like percentage of nutrients in the soil, climatic variables from API, type of soil to predict the crop yield and to provide proper recommendations to the farmers. Patil, (2020) worked mainly to increase the revenue of the ranchers as well as to raise the yield of the crop to grow for a

particular period and to provide fertilizer ratio and type forecasting information. To accomplish this, an efficient machine learning algorithm namely Random forest and back propagation algorithms are used for crop yield and fertilizer prediction.

Jahan and Shahariar, (2020) proposed a model named decision tree to forecast the fertilizer recommendation of maize. In this work, they have taken the image dataset for processing and classified the data into four categories. Finally, 93% accuracy is achieved in the decision tree algorithm to predict fertilizer prediction. Qin *et al*., (2018) proposed a machine learning model for predicting corn's economic optimal nitrogen rate. The 4 ML models were taken into consideration namely Linear Regression, LASSO regression, and Gradient boost regression tree (GBRT). Among all those algorithms, ridge regression outperformed other models. The model performances were evaluated by MAE and R^2 models. Archana and Saranya, (2020) suggested data mining techniques for crop yield prediction and fertilizer recommendation. KNN algorithm is proposed to forecast the crop yield to make farmers decide efficiently. Abhang *et al*., (2018) used soil analysis for crop fertility prediction. The proposed system uses a classification algorithm to predict suitable crops based on the pH of the soil and other climate variables. Pande *et al*., (2021) explained that among several Machine Learning algorithms, SVM and ANN were used to forecast crop yield. The algorithms namely KNN, Multivariate Linear Regression (MLR), RF, and ANN are employed. With 95% accuracy, the RF model provided the best results. The algorithm also makes a recommendation on when to apply fertilizers to increase the yield of the crops.

R and John Aravindhar, (2021)estimated the amount of fertilizers needed for banana is predicted using the regression method and three soil nutrients for crop growth namely N, P, K. The amount of NPK that soil naturally contains varies from place to place. Chauhan and Chaudhary, (2021) developed a machine learning-based recommendation model which recommends the best crop to produce and the right amount of fertilizer to grow. SVM algorithm outperformed when compared to other machine learning algorithms such as RF, and KNN for the crop and fertilizer recommendation. K and K.G, 2020; Ali, (2021) proposed an ML model for crop yield prediction and fertilizer recommendation systems. In this, many machine learning algorithms were performed namely SVM, KNN, RF, and voting-based ensemble classifiers were used. Among them, voting based ensemble classifier achieved superior results. Coulibali *et al.*, (2020) proposed site-specific machine learning predictive fertilization modes for potato crops in Eastern Canada. The model was conducted from 1979 to 2017. The model compared predictions from the hierarchical Mitscherlich model, KNN, RF, Neural network, and Gaussian process. The most potential algorithm to support choices that reduce financial or agronomic risks stands out as Gaussian processes.

Methodology

Figure 1 depicts the overall architecture of the proposed model, which uses two ML algorithms namely Random forest and Decision tree. In addition, it was also compared with three Ml algorithms namely LR, NB, and SVM. In this work, Pycharm Community Edition 2022.2.3.64 was used to conduct the research. The random forest algorithm is applied to oilseed yield data from the Official Government Website, which includes soil data, meteorological data, yield data, and inorganic fertilizer data, etc. In this, the RF algorithm was used to predict the crop yield and a decision tree was used to build a recommendation system for the end users. A random forest algorithm is a supervised machine learning algorithm for classification and regression problems. We know that a forest is made up of many trees, and the more trees there are, the more robust the forest is. The algorithm divides the yield attribute into two broad categories namely high and low. The final decision is made after generating a model based on the anticipated targets. Crop yield prediction enables more accurate production planning and decision-making. The suggested model also incorporates a recommendation system (GUI) that uses a decision tree technique to assist farmers in determining the recommended fertilizer ratio and crop type for a given season. The main reason to choose a decision tree is that it assists us in deciding between several options. They provide a highly effective structure for laying out options and investigating the potential outcomes of those options.

Oilseed Crops

Oilseed crops are recognized as those whose oil is the most valuable component of the seed, being utilized for both edible and industrial purposes. There is also considerable vegetable oil produced as a

byproduct of extraction for other components as is the case with corn oil. Oil serves primarily as a source of energy and carbon precursors in germinating seeds. Synthesis of storage lipids occurs in the seed and, thus, oil composition is genetically determined by the embryo, and the relative weight of the embryo to endosperm and seed coat determines oil content. It is generally accepted that there is a negative relationship between protein and oil content. Oil and protein constituents are synthesized at different rates and times during oilseed development. Variation in nitrogen fertility during seed development and maturation affects the synthesis of fatty acids and, therefore, their final proportions in the oils of mature seeds. Since not only oil composition but oil content as well is affected by nitrogen availability in soil, this can affect oil utilization and the value of specific oilseed crops. While nitrogen-limiting situations generally reduce total oilseed production and, hence, oil yield per acre, there are few instances where the crop quality is reduced by inadequate nitrogen availability (Kumari and Saritha, 2017)

Importance of chemical fertilizers

Chemical fertilizers have been widely used to achieve maximum productivity in conventional agricultural systems. The continuous and excessive utilization of chemical fertilizers plays a major role, directly and/or indirectly. The rising global population and land resource limitations were the main reasons for the use of pesticides and chemical fertilizers to maximize crop productivity. This intensive utilization is reflected directly and/or indirectly in the ecosystem. Nitrogen is often the most limiting factor in crop production. Hence, the application of fertilizer nitrogen results in higher biomass yields and protein yields. In oil seed crops, protein levels are increased upon nitrogen fertilization. Phosphorus is an important primary nutrient and enhances root growth thereby facilitating the absorption of water and nutrients from deeper layers. Phosphorus stimulates not only root growth but also hastens the maturity of oilseed crops. The P requirement of oilseeds and pulses is relatively high as it plays an important role in plant metabolism. Potassium increases yields and improves the quality of agricultural produce. Potassium also enhances the ability of plants to resist diseases, insect attacks, cold and drought stresses, and other adverse conditions (Subramaniam *et al.*, 2014).

Agricultural Dataset

The sources for the datasets utilized in this research are mentioned below,

- The Department of Meteorological Centre India provides access to weather datasets which includes temperature, humidity, rainfall, etc.
- Fertilizer types and quantity information has been obtained from Agricultural University Departments.
- Climatic factors such as sunshine are gathered from the weather atlas portal.
- Various oilseed yield datasets are gathered from ICRISTAT, the Tamil Nadu Government Website (www.data.govt), and the University Department of Agriculture (Ali, 2021; Chitdeshwari *et al.*, 2017; Subramaniam *et al.*, 2014)

This study takes some important climatic variables, such as soil temperature, pH, rainfall, humidity, and the minimum and maximum temperatures of a specific location and area. Some soil parameters such as textures (red loamy, clay loam, deep red loam, etc.) as well as different seasons are included. In addition, fertilizer (NPK) soil nutrient content data, quantity, and types are also taken into consideration for the prediction of crop yield.

The following oilseed crops were considered for this study,

- Castor
- Coconut
- Rapeseed
- Groundnut
- Safflower
- Other oilseed crops

Dataset Description

The data collected from various sources are provided as input to the model. For the above oilseed crops in all districts of Tamil Nadu, a set of data is initially collected that includes parameters such as state name, district name, humidity, productivity, inorganic fertilizer type, and so on. This .csv dataset was compiled between 1961 and 2019. The final dataset contains 1012 records with 28 attributes.

Preprocessing

Preprocessing is required before applying any machine-learning technique to a dataset. Data collected from various sources is common in raw form. The raw data contains information that is missing, incon-

sistent, or outdated. As a result, before processing, redundant data must be filtered. The data series provided contains a huge 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 facilitate data access. To verify that all values fall within a study range, the final dataset is normalized. Equation 1 depicts the normalization technique formulae (Z-score). Data is normalized to a factor ranging from 0 to 1 after the z-score technique is applied.

$$x' = (x - u) / sd \quad .. (1)$$

where,
x is raw value,
x' is the normalized value
u is the mean of the values
sd is the standard deviation of the values

Data analysis

After preprocessing the raw data, the data must be assured by the process of inspection, cleansing, transformation, and designing to produce meaningful information and conclusions and support decision-making to carry forward with a proper grasp of the dataset.

Dimensionality reduction

To make reliable predictions, high-level factors that influence prediction accuracy must be carefully chosen. There are many feature selection techniques accessible, but Linear Discriminant analysis (LDA) is the one that works best for this research because it helps to transform and compress the dataset only with essential features. This dataset has a total of 28 features. The 21 critical feature subsets were chosen using the LDA technique. The optimal feature subset was chosen by feeding these feature subsets into the random forest method. The criteria chosen included humidity, rainfall, location, production, and others. When those attributes were incorporated into statistical models and machine learning algorithms, the model's classification accuracy got enhanced.

Training and testing model

The dataset can be segregated into training and testing sets during the preprocessing stage. We partitioned the dataset into 80% for training and 20% for testing. This is a crucial step in the model's development. The model is trained using the training dataset, and the model is validated using the testing dataset. As a result, we fit the model with the training dataset. As a result, we fit the model with training data and test its accuracy with testing data.

Prediction algorithm

After the data has been segmented, the model is generated and trained. The action of training a machine learning model necessitates the use of a machine learning algorithm and training data to comprehend the pattern. In this case, we employ a variety of machine learning algorithms that are well-known supervised learning algorithms with a clear and concise representation.

Comparison of accuracy of the proposed model with existing ones

Table 1. Accuracy of proposed models

Models	Accuracy
RF	96.38
DT	96.38
LR	91.30
SVM	88.40
NB	82.69

Classification model for oilseed yield prediction

The oilseeds crop yield dataset includes 1012 records for 6 crops. After preprocessing, the oilseed crop yield prediction tends to reduce by 976 records. After that, the training set contains 781 records, while the testing set consists of the remaining 231 records. The machine learning model we created to predict the yield of oilseed crops. All of the proposed algorithms, including RF, LR, SVM, and NB classifiers are compared. Among these models, the RF algorithm has been found accurate to forecast oilseed crop yield. Pycharm is a platform used for developing a trained model with machine learning algorithms.

A random forest algorithm is a supervised machine learning algorithm that is widely used in classification and regression problems. We know that a forest is made up of many trees, and the more trees there are, the more robust the forest is. Similarly, the more trees in a random forest algorithm, the greater its accuracy and problem-solving capability. RF is a classifier that uses the average of several decision trees on different subsets of a given dataset to improve its predictive accuracy. It is built on the idea of ensemble learning, which is the process of com-

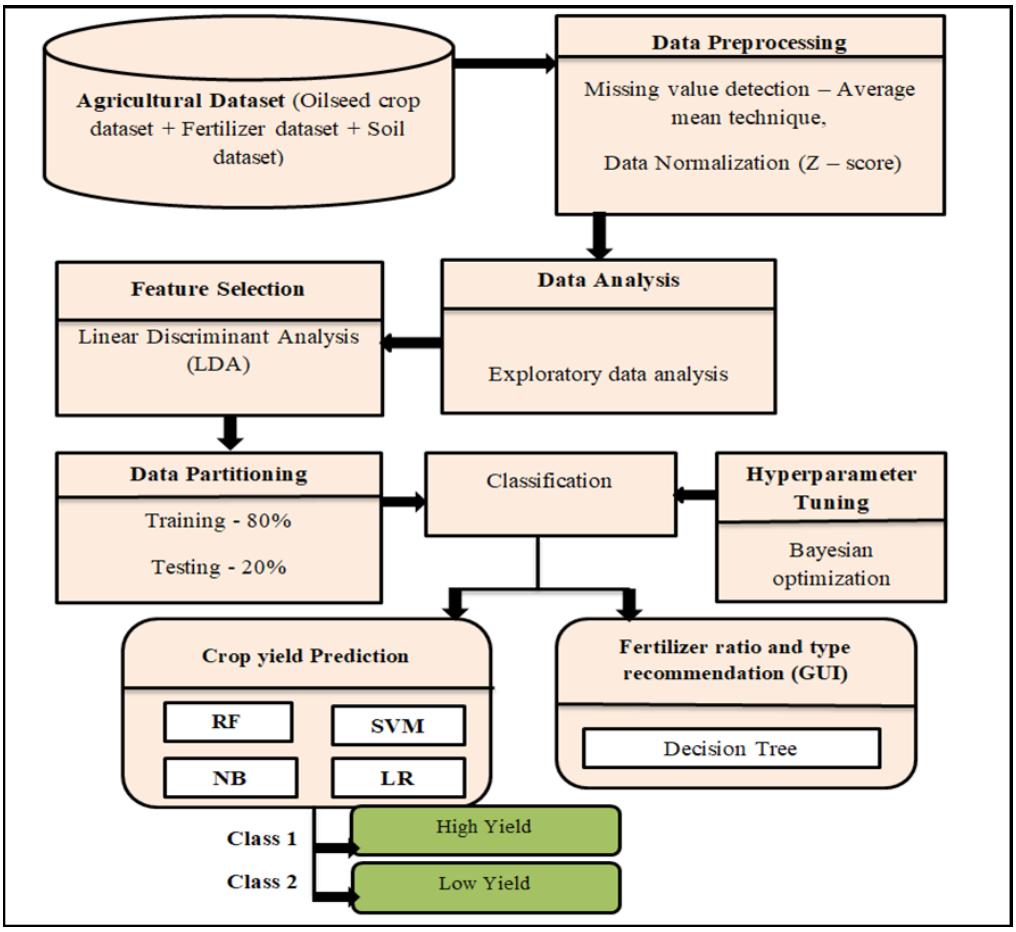

Fig. 1. Architecture diagram for theoilseed crop yield prediction and fertilizer recommendation

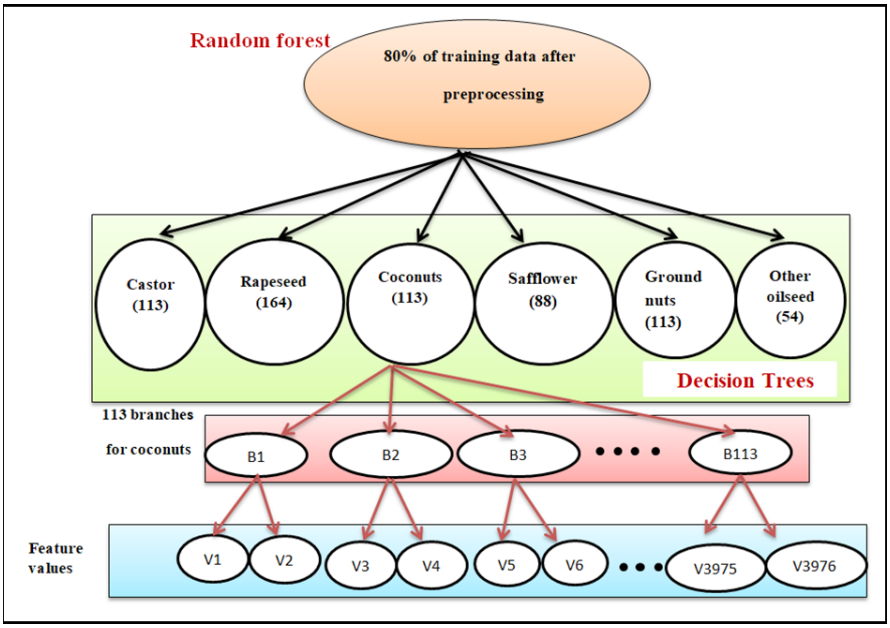

Fig. 2. Process flow of Random Forest algorithm

bining multiple classifiers to solve a complex problem and improve the performance of the model. Figure 2 explains the process flow of the random forest algorithm,

Algorithm: The steps involved in creating a classification model for the crop yield dataset

Input: An experimental dataset of weather, crop, soil, and fertilizer data (NPK)

Output: Crop yield prediction for the experimental dataset

Method:

Step 1: Data collection and feature analysis

a) Gather, arrange and format the data

The model generally requiresraw data for processing. It is necessary to gather the data, store it when needed, and arrange it such that the desired outcomes are attained.

b) Analyze and choose features

After preprocessing, the data is evaluated to produce useful information and conclusions to progress with the proper expertise of all the variables. Once the dimensionality reduction is completed, essential feature sets are chosen by employing the LDA method. ML techniques are then employed to process the chosen features.

Step 2: Separate the data into two groups

The training set will contain the most information and will be used to train the majority of the examples that will result in the yield. Approximately

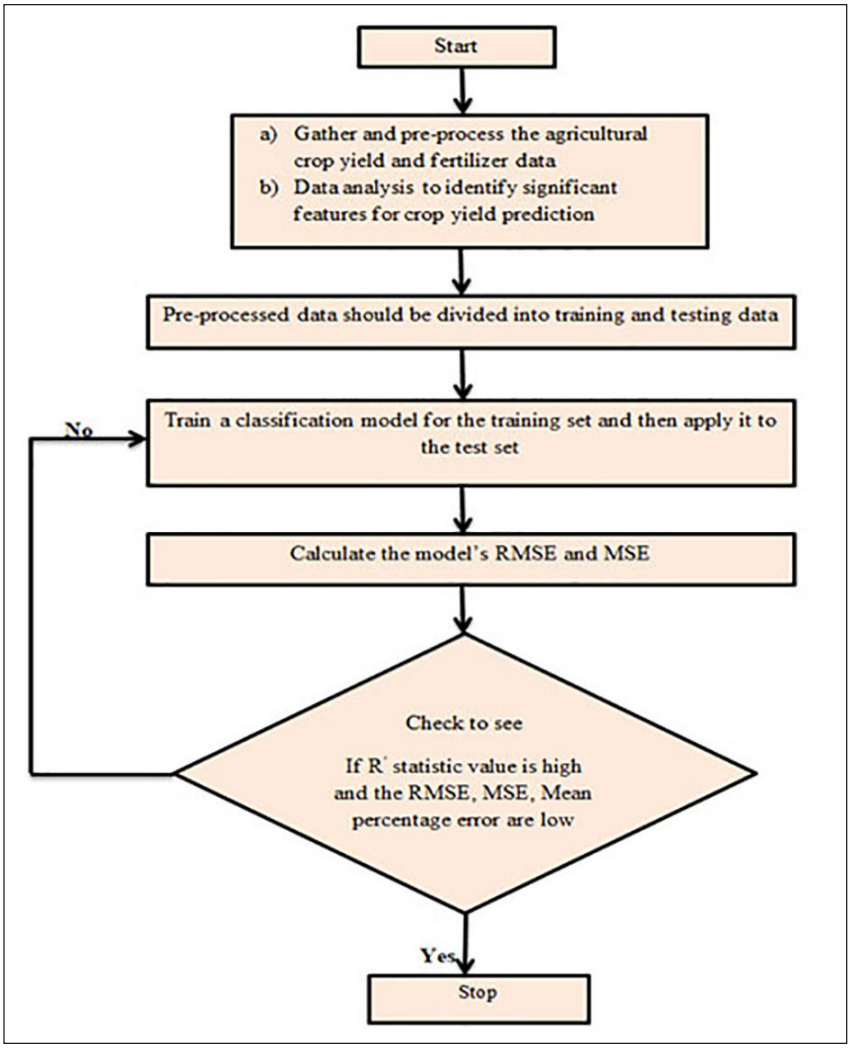

Fig. 3. Flow diagram of classification methodology for the oilseed crop yield prediction

80% of the collected samples are used in the training set. The testing set makes use of the final piece of data to evaluate how effectively the system functions.

Step 3: Classification of trained sets

The complexity of the problem will determine the model system, and the structure must be chosen accordingly. During training, it is possible to adjust the construction, modeling, and structure during training set.

Step 4. Calculate each model's RMSE, R^2 statistic, and MSE values

Run the trained classification model on the test set again and compute the MSE and RMSE values. Compare the results with different classification models. The best crop yield prediction model has the lowest MSE and RMSE values as well as the highest R^2 statistic value. The flow chart for the classification technique used to forecast crop yield is shown in Figure 3.

Predict Yield

When new input is provided, the trained model is used to predict the output. The trained model was saved as a file so that it could be estimated using new input. These models were trained properly on the training dataset and tested on the testing dataset. This prediction model employs machine learning that learns the properties from training data to make accurate predictions.

Prediction results

Figure 4 below explains the comparison of actual value and predicted value for all crops in Tamil Nadu.

Error calculation for various classification algorithms

The formulae for mean square error and root mean squared error is displayed below,

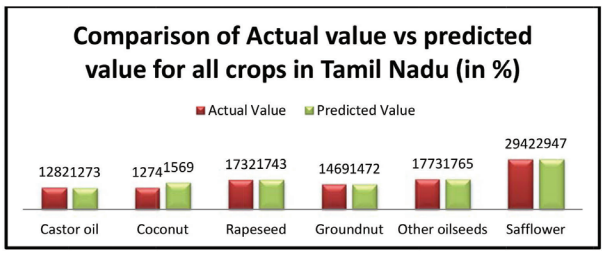

Fig. 4. Comparison of actual value versus predicted value for all crops in Tamil Nadu

Table 2. Absolute error calculation for all crop yield prediction

Crop name	Actual value	Predicted value	Absolute error (in %)
Castor	1282	1273	0.09
Coconut	1274	1569	2.95
Rapeseed	1732	1743	0.11
Groundnut	1469	1472	0.03
Other oilseeds	1773	1765	0.08
Safflower	2942	2947	0.05

Figure 5 below represents the mean square error for all machine learning algorithms,

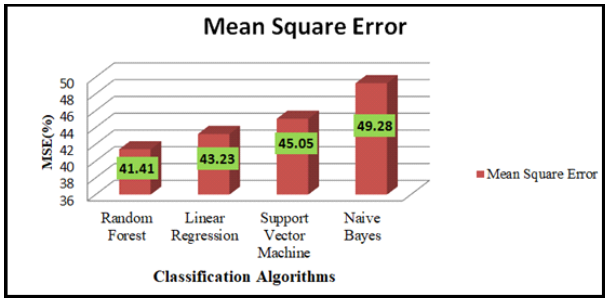

Fig. 5. MSE of proposed models

Figure 6 given below represents the root mean square error for all machine learning algorithms,

Comparison with different models

We got an accuracy of 94%, indicating that this

Table 3. Formulae for error calculation

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

Table 4. Formulae for evaluation metrics

Accuracy	Recall	Precision	Specificity
$\frac{TP + TN}{TP + TN + FP + FN}$	$\frac{TP}{TP + FN}$	$\frac{TP}{TP + FP}$	$\frac{TN}{TN + FP}$

TP- True Positive, TN - True Negative, FP - False Positive, FN- False Negative

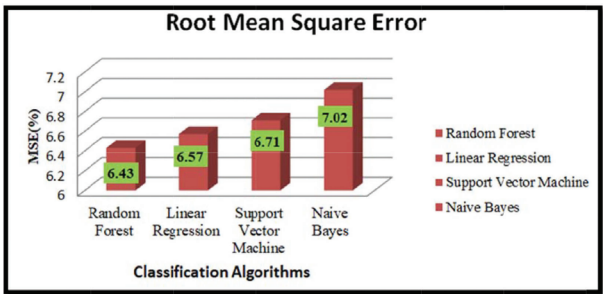

Fig. 6. RMSE of proposed models

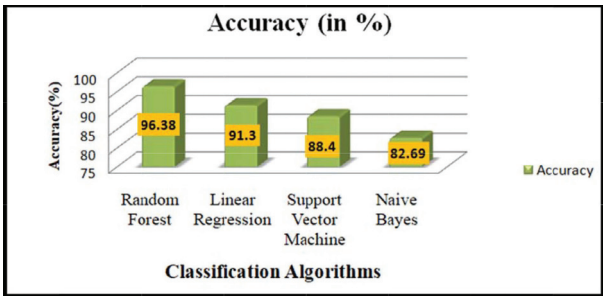

Fig.7. Comparison of accuracy for all proposed models

model is better at predicting yield. In terms of accuracy, the random forest algorithm outperformed other models. This is due to model and structural changes made during training. Table 1 compares the accuracy of various algorithms, and figure 7 depicts a graphical comparison of machine learning model accuracy.

Evaluation Metrics

There are numerous ways to evaluate performance. Some of the most popular metrics are accuracy, precision, recall, and the confusion matrix. The confusion matrix is frequently used to describe the performance of a classification model on a set of test data for which the true values are known. It is calculated for all four machine learning models namely RF, LR, SVM, and NB, and it is noted that RF works better when compared to other machine learning algorithms.

Accuracy

Accuracy is simply how frequently the classifier predicts correctly. It is defined as the number of correct predictions divided by the total number of predictions. The accuracy of all machine learning algorithmsis compared in Figure 7.

Recall

The recall is a ratio of the number of correct detections to the total number of positive samples. The recall values of all machine learning algorithms are compared in Figure 8.

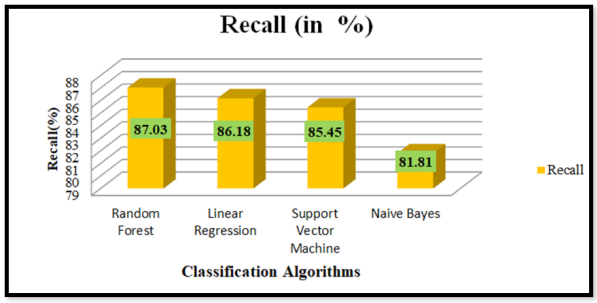

Fig. 8. Comparison of recall for proposed models

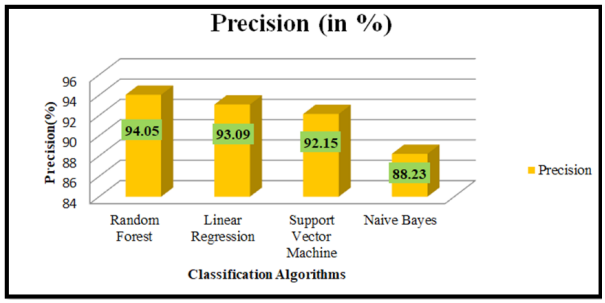

Fig. 9. Comparison of recall for proposed models

Precision

Precision is defined as the ratio of true positives to predicted for a given label. The precision values of all machine learning algorithms are compared in Figure 9.

F - measure

The harmonic mean of precision and recall is called f-measure. F-measure values of all machine learning algorithms are compared in figure 10.

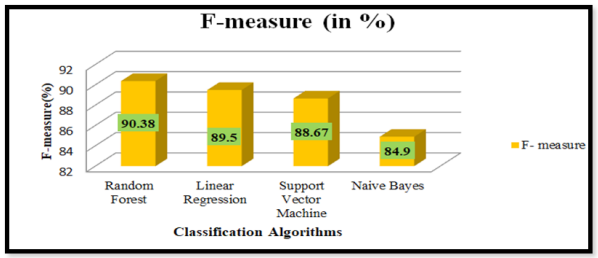

Fig. 10. Comparison of recall for proposed models

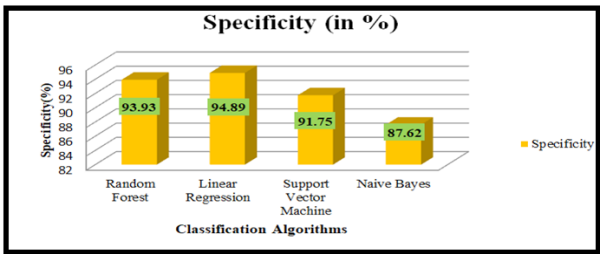

Fig. 11. Comparison of specificity for proposed models

Specificity

The ratio of true negatives to the total number of true negatives and false positives is known as specificity. The specificity values of all machine learning algorithms are compared in Fig. 11.

Execution time for all machine learning algorithms

The training execution time of the proposed algorithms is compared in Fig. 12.

The comparison of the testing execution time of the proposed algorithm is shown in Figure 13.

Results and Discussion

Overall observations of proposed algorithms

Three trials namely trial 1, trial 2, and trial 3 were conducted for each algorithm namely RF, LR, SVM, and NB.

Observations for RF

The various parameters considered for the study

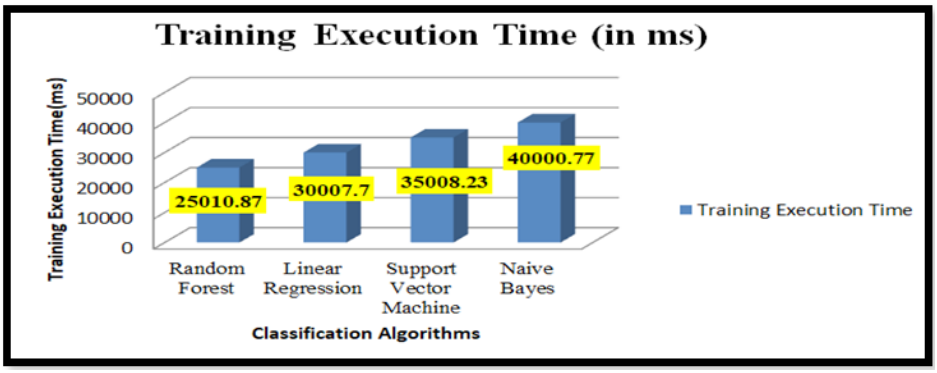

Fig. 12. Comparison of training execution time for proposed models

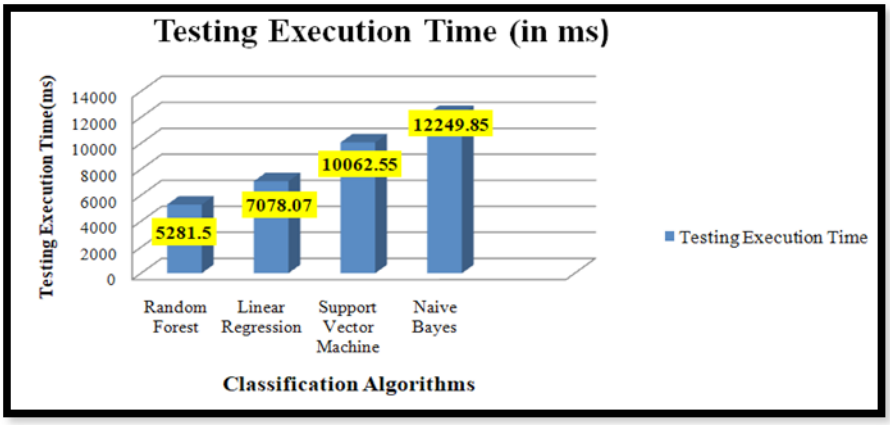

Fig. 13. Comparison of testing execution time for proposed models

arerf_no_branch, estimators, min and max sample, max features, min, and max leaf length, max_depth, random state,and forest size. During the study, the following observation was made. When the value for random state increases, accuracy tends to decrease. Three trials were done for the parameter "random_state" with an assumed value of 2,1,0.4 and their resulting accuracies were 71.87%, 90.78%, and 97.84% respectively.

Observations for LR

The various parameters considered for the study include independent var(y) and two dependent variables (x_1 and x_2) and as the slope (m) increases, the accuracy rate gets increases. Similarly, as the slope(m) value decreases, the accuracy gets decreased. Three trials were done with varying values of 0.12,0.7,0.9 for slope (m) while retaining the same values for other parameters and their resulting accuracies were 79.42%, 91.35%, and 93.47% respectively.

Observations for SVM

The various parameters considered for the study include kernel, regularization (c), and gamma variable. During the study, the following observations were made. The parameter 'kernel' has 3 types namely linear, polynomial and radial basis functions. The parameter with kernel type 'RBF' proved to be better than the other kernel types. Considering the regularization(c) term, the smaller value of 'c' creates a smaller margin hyperplane and a larger value of 'c' creates a larger-margin hyperplane. Subsequently, a lower value of gamma will loosely fit the training dataset, whereas a higher value of gamma will exactly fit the training dataset. Three trials were made with varying values of 'c' (regularization parameter) are assumed to be 2.5,2,1 and the values of gamma are 0.2,0.35,0.1 and the resulting accuracies were found to be 66.32%, 80.62%, 90.84% respectively.

Observations for NB

The various parameters considered for the study include alpha, priors, smoothing, epsilon, sigma, and theta. During the study, the following observation was made. As the value of alpha increases, the accuracy rate gets increases. Three trials were made for 'α' with varying values of 0.75, 0.80, 0.90, and accuracy rates of 56.62%, 75.32%, and 87.15 respectively. From the observation, it is concluded that, the greater the 'α' the greater the accuracy. Figure 14 depicts the trial accuracies ofthe proposed algorithms.

Discussion based on District wise crop yield

The goal of this paper is to comprehend the location-specific oilseed crop yield analysis, which will then be handled by a machine learning algorithm. For this study, a dataset in .csv format was considered. In this scenario, training uses 80% of the data and 20% of the data for validation. The model's accuracy was determined after successful training and testing indicating how well the model performed in forecasting yield. Figure 16 depicts a graphical user interface for predicting crop yield in the future. Figure 15 depicts a summary of all oilseed crop production districts in Tamil Nadu.

According to the statistics collected between 1961 and 2019,

- Erode has relatively more castor oilseed production

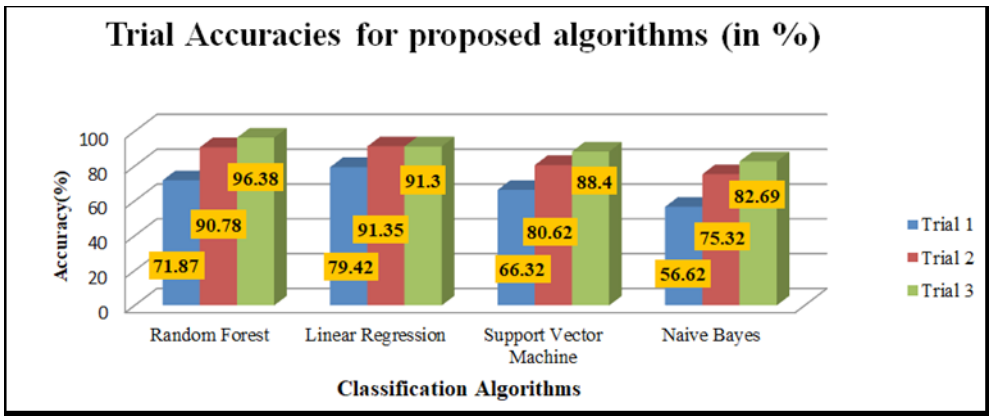

Fig. 14. Trail accuracies for proposed models

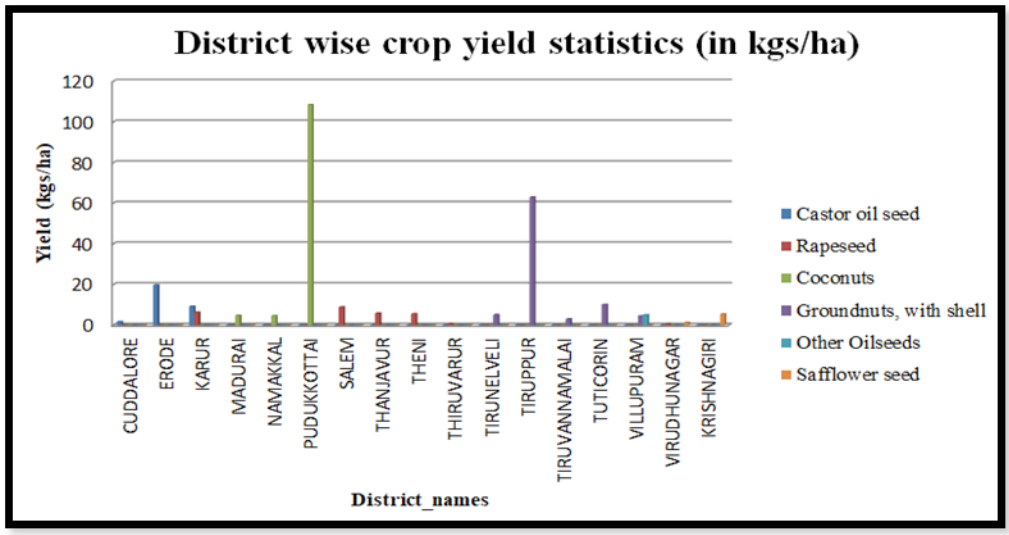

Fig. 15. District-wise crop yield statistics

- Salem has relatively more rapeseed production
- Pudukottai has relatively more coconut production
- Tirupur has relatively high groundnut oilseed production
- Villupuram has relatively high other oilseed production
- Krishnagiri has relatively more safflower production

Recommendation system for fertilizer and crop yield

GUI Creation

The study aids farmers in choosing which crop to grow in a specific area at a particular time and also provides information indicating whether it will be profitable or not during forecasts. Furthermore, it indicates low or high yield with ranges to help ranchers or end-users make successful selections,

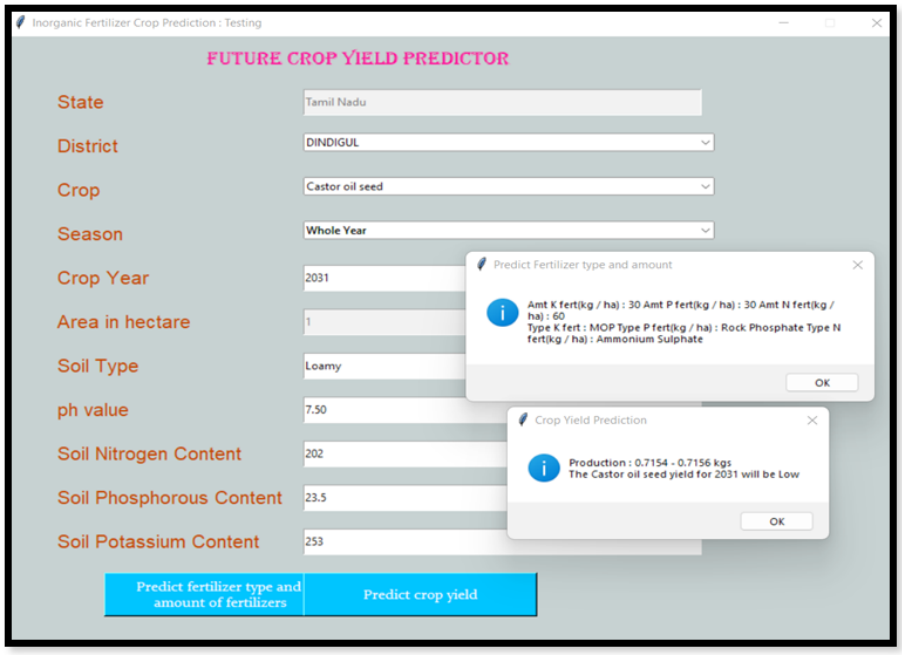

Fig. 16. Recommendation systems for crop yield and fertilizer predictor

allowing them to save time and accuracy. Users can share the district name, state name, state name, crop, season, and crop year to predict area, soil type, pH value, and soil content of Nitrogen, Phosphorous, and potassium by using the prediction module. Following the entry of these attribute values, the user can click the 'predict crop yield' button to estimate the yield of a particular crop in the future along with the yield rate classification as high or low. In addition, this recommendation system also helps the users with the "Predict fertilizer type and amount of fertilizers" button to forecast separate quantity intake (NPK) to be taken for a particular crop along with the fertilizer type to be used for a particular region. The result is obtained by taking into account the range of values based on the average of all prediction errors for each crop. The formula presented below is used to calculate the yield range based on the prediction error of each crop.

Predicted value ± Predicted error .. (2)

In Figure 16, for the Dindigul district, castor oilseed crop for the entire year, the crop production result range is calculated based on the average of prediction errors of all castor crops in a particular location, and then the low and high rate is estimated by taking the mean of each crop based on the records in the dataset. This recommendation system also helps in fertilizer type and quantity suggestions for oilseed crops in a particular area in a particular period. If the prediction value is less than the mean score, it is considered a low-yield crop and if the prediction value is greater than the mean score, it is considered a high-yield crop.

GUI data visualization is also accomplished by plotting yield variables with different parameters. Converting data into visual contexts such as graphs or figures helps humans capture and comprehend ideas. Figure 16 depicts the primary goal of GUI data visualization. The prediction module simplifies the identification of patterns, correlations, and outliers in large datasets. The graph above depicts the district's relationship with the yield.

Conclusion

This study looked at machine learning algorithms for crop yield forecasting that used temperature, season, and location as inputs. Rainfall, temperature, and other variables such as season, location, and fertilizer data can be used to forecast yield in a particular district. When all factors are taken into account, the random forest classification technique tends to be the best classifier among all other classifiers. Using the dataset with more parameters improves accuracy. RF is found to be the best prediction algorithm when compared to other prediction algorithms such as LR, SVM, and NB. Our database contains a much larger number of variables, resulting in more accurate predictions. The emergence of this work will assist farmers in reducing risk and maximizing crop yields to improve their agricultural resources.

In this paper, we predicted future crop yields using soil test results and inorganic fertilizer dosage. We have also developed a recommendation system for farmers to determine the best crop to cultivate in the coming season, as well as fertilizer type and quantity recommendations for ranchers. This will not only assist farmers in determining the best crop to cultivate in the coming season, but it will also aid in bridging technological and agricultural divides. The limitation of our work is that yield is only implemented for 30 districts in Tamil Nadu and not for other states. Our project's future work aims to include regional languages in the graphical user interface such as Tamil, Telegu, Hindi, Kannada, Malayalam, and others which benefit farmers across the country.

Abbreviations

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

Conflicts of Interest

The manuscript has not already been published or submitted to another journal for publication consideration. This paper is not known to involve any conflicts of interest.

References

Abhang, K., Chaughule, S., Chavan, P. and Ganjave, S. 2018. Soil Analysis and Crop Fertility Prediction. *International Research Journal of Engineering and Technology*. 05(03): 3106–3108. https://www.chem.purdue.edu/gchelp/howtosolveit/

Ali, S.M. 2021. Machine Learning based Crop Recommendation System for Local Farmers of Pakistan. *Revista Gestão Inovação e Tecnologias.* 11(4): 5735–5746. https://doi.org/10.47059/revistageintec.v11i4.2613

Archana, K. and Saranya, K.G. 2020. Crop yield prediction, forecasting and fertilizer recommendation using Data mining algorithm. *International Journal of Computer Science Engineering*.

Bharath, S.M., Manoj, S., Adhappa, P., Patagar, P.L. and Bhaskar, R. 2022. Crop Yield Prediction with Efficient Use of Fertilizers. *Lecture Notes in Electrical Engineering*. 783(July):937–943. https://doi.org/10.1007/978-981-16-3690-5_87

Bondre, D. A. and Mahagaonkar, S. 2019. Prediction of Crop Yield and Fertilizer Recommendation Using Machine Learning Algorithms. *International Journal of Engineering Applied Sciences and Technology.* 04(05): 371–376. https://doi.org/10.33564/ijeast.2019.v04i05.055

Chauhan, G. and Chaudhary, A. 2021. Crop Recommendation System using Machine Learning Algorithms. *Proceedings of the 2021 10th International Conference on System Modeling and Advancement in Research Trends, SMART 2021*, 3307: 109–112. https://doi.org/10.1109/SMART52563.2021.9676210

Chitdeshwari, T., Santhi, R., Radhika, K., Sivagnanam, S., Hemalatha, S., Dey, P. and Subba Rao, A. 2017. GPS and GIS BASED Soil Fertility Mapping for Cuddalore District of Tamil Nadu. *Madras Agricultural Journal.* 104(7–9): 251. https://doi.org/10.29321/maj.2017.000054

Coulibali, Z., Cambouris, A. N. and Parent, S. É. 2020. Site-specific machine learning predictive fertilization models for potato crops in Eastern Canada. *PLoS One* (Vol. 15, Issue 8 July). https://doi.org/10.1371/journal.pone.0230888

Gosai, Dhruvi, Raval, Chintal, Nayak, Rikin, Jayswal, Hardik, Patel, Axat, 2021. *Crops yield prediction and efficient use of fertilizers using machine learning*. (2021). 8(2): 1539–1545.

Divya, K.V., Jatti, A., Joshi, P.R. and Krishna, S.D. 2019. Progress in Advanced Computing and Intelligent Engineering. In:*Progress in Advanced Computing and Intelligent Engineering*. (Vol. 714). Springer Singapore. https://doi.org/10.1007/978-981-13-0224-4

Gopal, P. S. M. and Bhargavi, R. 2019. A novel approach for e ffi cient crop yield prediction. *Computers and Electronics in Agriculture.* 165(July): 104968. https://doi.org/10.1016/j.compag.2019.104968

Hampannavar, K., Bhajantri, V. and Totad, S. G. 2018. Prediction of Crop Fertilizer Consumption. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, 1–5. https://doi.org/10.1109/ICCUBEA.2018.8697827

Jahan, N. and Shahariar, R. 2020. Predicting fertilizer treatment of maize using decision tree algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*. 20(3): 1427–1434. https://doi.org/10.11591/ijeecs.v20.i3.pp1427-1434

K, A. and K.G, D.S. 2020. Crop Yield Prediction, Forecasting and Fertilizer Recommendation using Voting Based Ensemble Classifier. *International Journal of Computer Science and Engineering.* 7(5): 1–4. https://doi.org/10.14445/23488387/ijcse-v7i5p101

Kamath, P., Patil, P., Sushma, S.S. and S. S. 2021. Crop yield forecasting using data mining. *Global Transitions Proceedings.* 2(2): 402–407. https://doi.org/10.1016/j.gltp.2021.08.008

Kanuru, L., Tyagi, A. K., Aswathy, S., Fernandez, T. F., Sreenath, N. and Mishra, S. 2021. Prediction of Pesticides and Fertilizers using Machine Learning and Internet of Things. *2021 International Conference on Computer Communication and Informatics, ICCCI 2021*, 1–6. https://doi.org/10.1109/ICCCI50826.2021.9402536

Katuru, K.H., Kishan, S.R. and Dasari, S.B. 2020. Predicting Crop yield and Effective use of Fertilizers using Machine Learning Techniques. *International Journal of Innovative Technology and Exploring Engineering*. 9(7): 1288–1292. https://doi.org/10.35940/ijitee.g5911.059720

Kumari, M.S. and Saritha, J. D. 2017. Effect of phosphorus fertilizers on oil seed crops. *Agriculture Update.* 12(Special-3): 749–754. https://doi.org/10.15740/has/au/12.techsear(3)2017/749-754

Manjunath, M., Venkatesha, G. and Dinesh, S. (n.d.). *Agricultural Crop Yield Prediction and Efficient Use of Fertilizer Using Machine Learning*. 5(1): 1–13.

Meng, L., Liu, H., Ustin, S. L. and Zhang, X. 2021. Predicting maize yield at the plot scale of different fertilizer systems by multi-source data and machine learning methods. *Remote Sensing.* 13(18). https://doi.org/10.3390/rs13183760

Meza-Palacios, R., Aguilar-Lasserre, A. A., Morales-Mendoza, L. F., Rico-Contreras, J. O., Sánchez-

Medel, L. H. and Fernández-Lambert, G. 2020. Decision support system for NPK fertilization: a solution method for minimizing the impact on human health, climate change, ecosystem quality and resources. *Journal of Environmental Science and Health - Part A Toxic/Hazardous Substances and Environmental Engineering*. 55(11): 1267–1282. https://doi.org/10.1080/10934529.2020.1787012

Naresh, V., Vatsala, B.R. and Raj, C.V. 2020. Crop Yield Prediction and Fertilizer Recommendation. *International Journal for Research in Engineering Application & Management*. 10(6): 135–138. https://doi.org/10.35291/2454-9150.2020.0452

Pande, S. M., Ramesh, P. K., Anmol, A., Aishwarya, B. R., Rohilla, K. and Shaurya, K. 2021. Crop Recommender System Using Machine Learning Approach. *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, *Iccmc*. 1066–1071. https://doi.org/10.1109/ICCMC51019.2021.9418351

Patil, L. 2020. Crop Yield Prediction on the Basis of Soil Composition using Machine Learning Algorithms. XVI(201): 201–205.

Qin, Z., Myers, D.B., Ransom, C.J., Kitchen, N.R., Liang, S.Z., Camberato, J.J., Carter, P.R., Ferguson, R. ., Fernandez, F.G., Franzen, D.W., Laboski, C.A.M., Malone, B.D., Nafziger, E.D., Sawyer, J.E. and Shanahan, J. F. 2018. Application of machine learning methodologies for predicting corn economic optimal nitrogen rate. *Agronomy Journal*. 110(6): 2596–2607. https://doi.org/10.2134/agronj2018.03.0222

R, J.S. and John Aravindhar, D. 2021. Fertilizer Estimation using Deep Learning Approach. *Nveo - Natural Volatiles & Essential Oils Journal | NVEO*. 8(4): 5745–5752. https://www.nveo.org/index.php/journal/article/view/1237

Sarkar, Uditendu, Banerjee, Gouravmoy and Ghosh, Indrajit, 2018. A Machine Learning Based Fertilizer Recommendation System for Paddy and Wheat in West Bengal.Springer Professional. *Wasser Und Abfall*. 20(5): 63. https://doi.org/10.1007/s35152-018-0064-x

Subramaniam, S. M., Santhi, R. and Swaminathan, H. 2014. An Appraisal of Available Nutrients Status and Soil Fertility Mapping for Salem District of Tamil Nadu An Appraisal of Available Nutrients Status and Soil Fertility Mapping for Salem District of Tamil Nadu. February 2017.