Type and quantity of organic manures recommendation and yield prediction of oilseed crops using machine learning algorithms

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

[1]Research Scholar, Dept. of Computer Science and Engineering, Faculty of Engineering And Technology, Annamalai University, Annamalai Nagar, Tamil Nadu, India, mithrac.official@gmail.com*

[2]Professor, Dept. of Computer Science and Engineering, Faculty of Engineering And Technology, Annamalai University, Annamalai Nagar, Tamil Nadu, India, suha_babu@yahoo.com

1 Corresponding author*

Abstract

Agriculture is essential to the Indian economy. Population growth faces the most serious threat to food stability. Population growth raises demand, facing farmers to produce more to keep up with the demand. Crop yield prediction technology can assist ranchers in improving efficiency and productivity. Correct manure rates are required for the cultivation of oilseed crop yield. When nutrients are scarce or over-fertilized, yields are significantly reduced and the environmental burden is enhanced. To resolve these concerns, our proposed work employs machine learning techniques in the prediction of the yield of oilseed crops using organic manure as well as the amount and type of agricultural manure to be used for a specific crop in different districts of Tamil Nadu. The training set consists of actual yield data from 1961 to 2007 and the validation set consists of data from 2008 to 2019. The proposed algorithm's results are compared to those of other machine learning algorithms namely bagging, random forest, linear regression and naive bayes with 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, while the boosting algorithms perform better for recommendation systems for determining 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 Prediction, Organic manures, Dose of manures, Data mining, Machine learning, Recommendation system

1.Introduction

Agriculture is considered to be the main source of employment for the majority of India's population, with approximately 70% of the country's population living in rural areas and relying on agriculture for a living. Furthermore, the agricultural sector contributes nearly

50% of India's GDP when compared to other occupations [22]. In Tamil Nadu, agriculture and its affiliated industries constitute the majority of the local economy. With 93% of farmers being small and marginal but more than two-thirds of rural households in the state still rely heavily on agriculture for their livelihood. India is the world's largest producer of oilseeds. Various oilseeds are grown, accounting for approximately 12% of the total cropped area in the country. Among all oilseed production in Tamil Nadu, Groundnut occupies 28% of the total area and 29% of the total production of oilseeds in the country. The approximate fertilizer for the crop is suggested to boost the yield due to the rising demand for oilseeds and the stagnation in their supply. Agricultural productivity has benefited greatly from fertilizers, which have helped India transition from a food-scarce to a food–sufficient region. It is widely obvious that greater reliance on agricultural chemicals such as inorganic fertilizers has resulted in negative environmental repercussions as well as reduced 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 considered to be a storehouse of various nutrients necessary for plant growth[12]. Organic manures not only increase yield but also 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]. Using experimental crops in the field to anticipate yields takes a lot of time, which makes it difficult for ranchers to choose the best crop at any given time. To overcome these issues, traditional (experimental) agricultural yield methods are being replaced by computerized yield prediction methods for better prediction [11]. Data mining is the process of sorting through large datasets to identify patterns and relationships that can aid in the resolution through data analysis. Data mining techniques are important in the production of oilseed crops because they allow access to large amounts of data as well as give ideas to forecast future trends[13].

In addition, the use of data analysis can assist farmers in real–time crop health monitoring, predictive analytics for future yields and resource management decisions based on proven trends. Machine learning is an advanced predictive analytics tool that has been widely used to create a decision support system in many fields such as finance, marketing and most recently agriculture [21]. The use of machine learning in agriculture is promising because it assists farmers, policymakers and other agricultural stakeholders in making informed decisions. Machine learning applications in agriculture will improve the efficient use of resources for cultivation and harvesting as well as livestock production [4,18]. Predicting crop yield is one of agriculture's most challenging undertakings [19,20]. It is critical in global, regional and field decision-making. Soil, meteorological, environmental and crop field decision-making. Soil, meteorological, environmental and crop parameters are used to forecast crop yield. As a result, a decision support system based on Graphical User Interface (GUI) is generated to help farmers decide the type and quantity of manures to use for a specific crop at a specific time shortly[14]. Model evaluation is achieved by analyzing a machine learning model's performance, as well as its strengths and weaknesses by using various evaluation metrics. Some of the metrics used to evaluate model performance are precision, recall, F-score and specificity.

In this paper, we consider four machine-learning models that exhibit how future crop yield can be predicted using attributes like temperature, humidity, soil type, area and so on to improve crop yield prediction accuracy. The boosting algorithm assists farmers in determining the type and quantity of organic manure to use for a specific crop. It is represented using GUI to allow ranchers to identify oilseed crop yield information. The current paper discusses data collection, preprocessing of data and feature selection before comparing them to four machine learning algorithms namely bagging, random forest, linear regression and naïve bayes to determine which algorithm is the best suited for crop yield prediction using organic manure. For each of the four distinct algorithms, many trials were carried out to assess if the accuracy rates had changed or not.

2. Related Work

[20] evaluated four machine learning algorithms namely ridge regression, K-NN, support vector regression and Gradient-Boosted decision trees to predict crop yield for potatoes, sunflower, spring barley and soft wheat. Based on how they learn the relationships between features and labels, these methods represent different classes of algorithms. In addition, random forest, recursive feature elimination with LASSO and mutual information are the few feature selection algorithms used for the machine learning models to predict as accurately as possible.[10] performed a meta-analysis to evaluate the impact of manure application on crop yield and the related soil attributes in china. In comparison to inorganic fertilizers, manure application considerably enhanced the yield by 7.6% and productivity rose with longer-term manure application, rising by 27.7% when the applications lasted more than 10 years. Hence, an Ordinary Least Squares (OLS) regression analysis was examined to estimate the interaction between changes in soil parameters and yield in soils with added manure. [7] have compared manure with synthetic fertilizer by the proposed machine learning algorithm namely boosted regression tree which accounts for 39% of manure, 21% of synthetic fertilizer and 40% of soil properties providing variation in relative yield. Due to the improvement in soil fertility, these findings suggest that manure application is a viable strategy for regulating crop yields.

[5] proposed and compared a deep-learningbased RNN-LSTM model with other models namely ANN, RFand multi-variate Linear regression and proved that RNN- LSTM model outperforms well when compared to other models for the prediction of crop yield. [28] proposed that using biomass composition and pyrolysis conditions, four types of machine learning methods successfully predicted biomass bio-oil yield. The random forest model performed better in terms of prediction than the SVM, DT and MLR. The analysis revealed that optimal parameters chosen using a genetic algorithm-based approach have a significant impact on bio-oil yield.

[27] compared deep learningbased multi-layer perceptron with other machine learning algorithms namely random forest, decision tree, K-nearest neighbor, Ordinary Least Squares, and Support vector regression. In addition, hyperparameter optimization was done for the betterment of yield estimation. DLMMLP provides satisfactory yield estimation

accuracy. [25] compared the performance among three machine learning algorithms namely linear regression, decision tree,and random forest, and found random forest model produces significant R^2 and MSE values. Hence, the random forest algorithm performs well in crop yield prediction. [8] recommended developing a traditional ANN model as well as a novel least squares Support vector Machine (LS-SVM) model with 33 experimental data. The findings demonstrated that the intelligent modeling approach named LS-SVM model outshines the traditional ANN model in terms of predicting performance and robustness for the modeling study of the cattle manure pyrolysis process and other similar processes. [9] used five regression models namely SVR, RF, Extreme learning machine, ANN, and DNN. DNN and RF models showed encouraging results when compared to other algorithms.

[17] suggested a set of fine-tuned machine learning models namely ridge and lasso regression, CART, KNN, SVM, XGB, and RF. The algorithm was statistically compared using R^2, RMSE, and MAE metrics. The best-performing model (RF) was fine-tuned once more for the bias-variance trade-off using a grid search approach. In terms of goodness-of-fit, RF performed the best. The RF method was then used to select important variables and interactions. [24] proposed an agribot – an intelligent interactive interface to assist farmers using data mining and machine learning techniques to decide which crop to choose for a particular year. It is implemented using the NLP technique. The system is designed in a manner that facilitates receiving responses to farmer input queries about the agricultural context in audio format, making farmer interaction more user-friendly. The three machine learning techniques were used namely KNN, DT, and RF. Random Forest provides better accuracy when compared with other machine learning algorithms. [3] suggested the prediction of the yield of crops concerning rainfall. The proposed method of evaluation was better than other existing methods of evaluation as it evaluates all the regression techniques namely linear regression, polynomial regression, Support vector regression, Decision tree repression, and Extreme Gradient Boosting regression for two crops of 4 individual states. The performance of the model is validated utilizingthe MSE technique. [26] proposed a model which combined an ecological distance algorithm along with crop yield predictors to construct a yield prediction model for rice and wheat crops. The proposed model was compared with the existing algorithm and the intelligent model using EDA produces higher prediction accuracy. [6] proposed the model of predicting the wheat crop yield using data mining classification algorithms and stepwise linear regression. In this model prediction, WEKA and SPSS tools were used. The factors used were weather and crop data. The study shows that the results of MLP and additive regression were better when compared to other algorithms. [1] proposed four machine learning algorithms namely LR, EN, KNN and SVR were used to predict potato tuber yield from proximal sensing data of soil and crop properties. All other models were outperformed by the SVR models.

[23] showed that manure application can be an effective way to restore microbial biomass loss caused by intensive NPK application. However, variations in response are determined by specific manure types, application rates as well as local climate and inherent soil properties. The results of RF models revealed that manure type, application rate, and soil initial properties were likely the most important factors controlling microbial biomass response to manure application. [2,18] suggested the application of poultry manure along

with tillage enhanced grain output by 39.5% when compared to tillage alone. When compared to manure-mechanized tillage methods, the manure-Zero tillage methods enhanced grain yields by 15%. Hence, the organic manure application gives better on-field performance when compared to other algorithms.

3. Methodology

The overall architecture of the proposed model is depicted in figure 3, which employs four machine learning algorithms namely bootstrap aggregation, RF, LR, NB, and DT. Furthermore, it was compared with each other to determine which algorithm performs best. Pycharm Community Edition 2022.2.3.64 was used to conduct the research. The Bootstrap Aggregation algorithm is applied to oilseed yield data obtained from the Official Government Website, which includes soil, meteorological, yield, and organic manure data, etc., and boosting algorithm was used to create a recommendation system for end users. The algorithm categorizes yield into two broad categories namely high and low yield. After developing a model based on anticipated targets, the final decision is made. Crop yield prediction allows for more precise production planning and decision-making. The proposed model also includes a recommendation system (GUI) that employs a boosting technique to help farmers determine the best manure ratio and crop type for a given season. They offer a very useful framework for outlining options and exploring the possible repercussions of those options.

3.1 Oilseed crop

India is the world's fourth-largest producer of oilseeds. It accounts for 20.8% of the total global cultivation area and 10% of global production. Oilseeds grown in the country includes groundnut, safflower, sunflower, sesame, mustard, and castor. Almost 72% of the oilseed area is restricted to rainfed farming by small farmers, resulting in low productivity. However, a breakthrough in oilseed production was achieved by introducing cutting-edge production technologies. As a result, oilseed production increased from 108.3 lakh tonnes in 1985-86 to 365.65 tonnes in 2020-21. Oilseed production in India got increased over the last 5 years. The country's production in 2020-2021 was 365.65 lakh tonnes, a 10% increase over the previous year.

3.1.1 Importance of organic manure

In the future, sustainable agriculture will inevitably use organic manure to meet crop nutrient needs because they not only increase yield but also preserve the soil's physical, chemical, and biological qualities. Organic sources that can be incorporated into the soil are getting harder to find. Organic manure slowly mineralizes and releases vital minerals that have been locked up, helping to improve soil fertility while also boosting crop output and quality [16].

3.1.2 Advantages of organic manures

Organic manure contains all of the nutrients that plants require but in limited quantities. It improves the soil's structure, and texture and can retain water. It is cost-effective when compared to mineral fertilizers. Furthermore, it also aids in the preservation of oil by increasing fertility and productivity.

3.2 Agricultural Dataset

- The dataset used in this study was obtained from the following sources:
- The Department of Meteorological Centre India allows access to weather datasets that include temperature, humidity, rainfall, and so on.
- Organic manure types and quantity details have been obtained for Agricultural University Departments.
- Environmental parameters like sunshine are collected from the weather atlas webpage.
- Various oilseed yield datasets are gathered from ICRISTAT, the Tamil Nadu Government Website (www.data.govt) as well as from the University Department of Agriculture.

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

This study considered the following crops:

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

3.2.2 Dataset Description

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

3.3 Preprocessing

Before applying any machine learning technique to a dataset, preprocessing is required. Data gathered from various sources are frequently in raw form. The raw data contains incomplete, inconsistent, or out-of-date information. As a result, redundant data must be filtered before processing. The provided data series contains a large number of 'NA' values that can be filtered in python by replacing missing values with an average value. A robust scalar technique is used to remove outliers. The data is then transformed to make it easier to access. The final dataset is normalized to ensure that all values fall within a study range. The simplest normalization (Constant factor normalization) formulae are depicted in equation 1. This technique is used to normalize 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

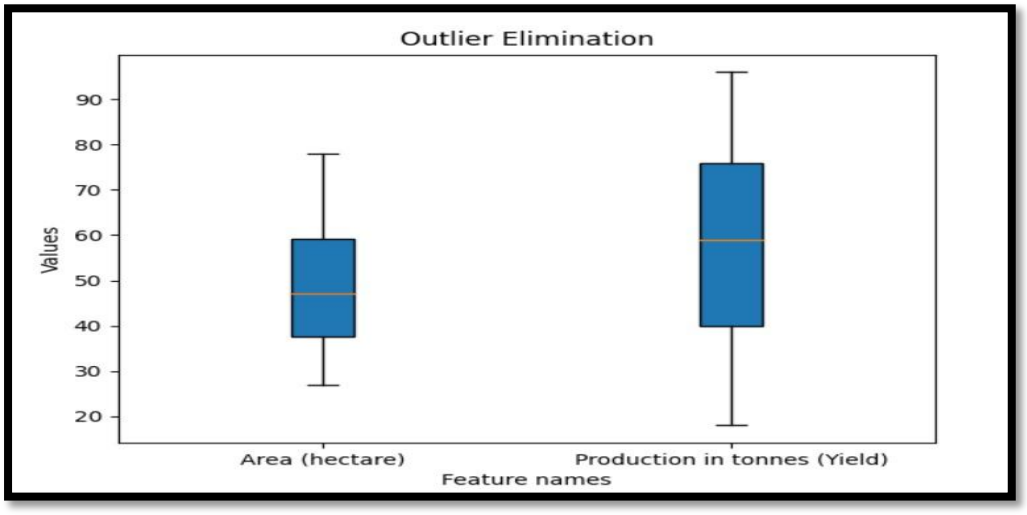

Fig 1: Box plot representation for outlier elimination

3.4 Data Analysis

After the raw data has been preprocessed, it must be ensured through the processes of inspection, cleansing, transformation, and designing 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 found in the data, a box plot graph must be generated for easy comprehension. Figure 2 depicts the heatmap of variables after preprocessing.

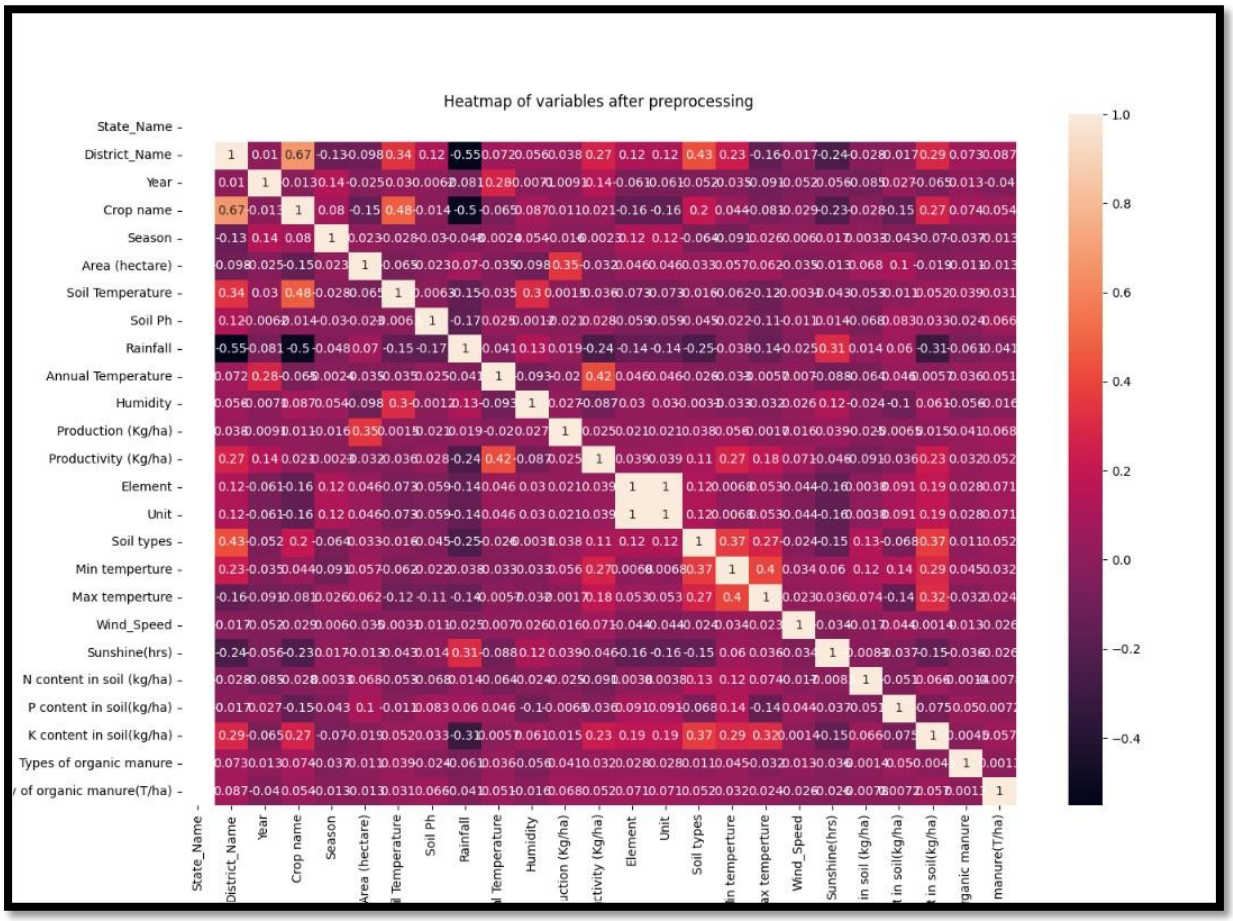

Fig 2: Heatmap of variables after preprocessing

3.5 Dimensionality reduction

To make accurate predictions, high-level factors influencing prediction accuracy must be carefully selected. Many feature selection techniques such as LDA, PCA, and factor analysis are chosen. Factor analysis is the best option for this research because it helps to transform and compress the dataset only with the most important features. This dataset contains 25 features in total. In this, the PCA technique was used to select 20 critical features. The LDA technique was used to select 17 critical feature subsets. The factor analysis technique was used to select the 13 critical feature subsets. Among all these dimensionality reduction techniques, the factor analysis feature subsets provide significant accuracy. By feeding these feature subsets into the bagging method, the optimal feature subset was determined. Humidity, rainfall, location, production, and other factors were considered. When those characteristics were incorporated into statistical models and machine learning algorithms, the classification accuracy of the model got improved.

3.6 Training and testing model

During the preprocessing phase, the dataset can be split into training and testing sets. We divided the dataset into training and testing groups of 80% and 20% respectively. This phase of the model's development is critical. The training dataset is used to create the model, while the testing dataset is used to verify the model. We used the training dataset to fit the model as an outcome. As a result, we use training data to fit the model and testing data to evaluate its accuracy.

3.7 Prediction algorithm

The model is created and trained after the data has been segregated. To understand the pattern, creating a machine learning model requires the usage of a machine learning algorithm and training data. In this case, we use several machine learning algorithms, 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

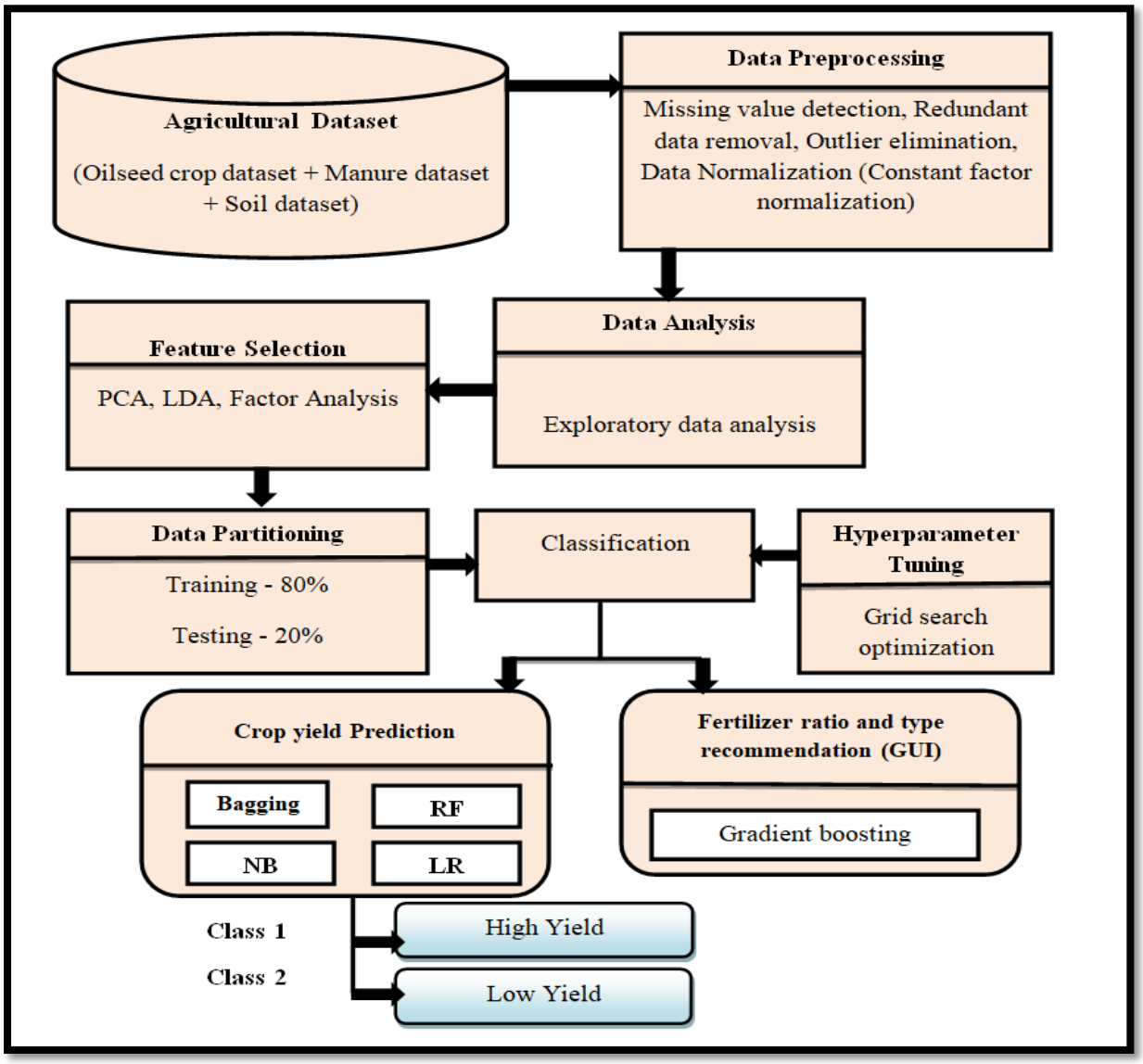

Fig 3: Architecture diagram for oilseed crop yield prediction and manures recommendation system

Classification model for oilseed crop yield prediction

The oilseed crop yield dataset contains 1012 records for 6 different crops. The oilseed crop yield prediction tends to decrease by 979records after preprocessing. The training set then contains 779records, while the testing test contains the remaining 200records. We developed a machine learning model to predict the yield of oilseed crops. All of the proposed algorithms are compared including bagging, linear regression, and naive bayes classifiers. Among all these models, the bagging algorithm is effective at forecasting oilseed crop yield. Pycharm is a platform for training a model using machine learning algorithms.

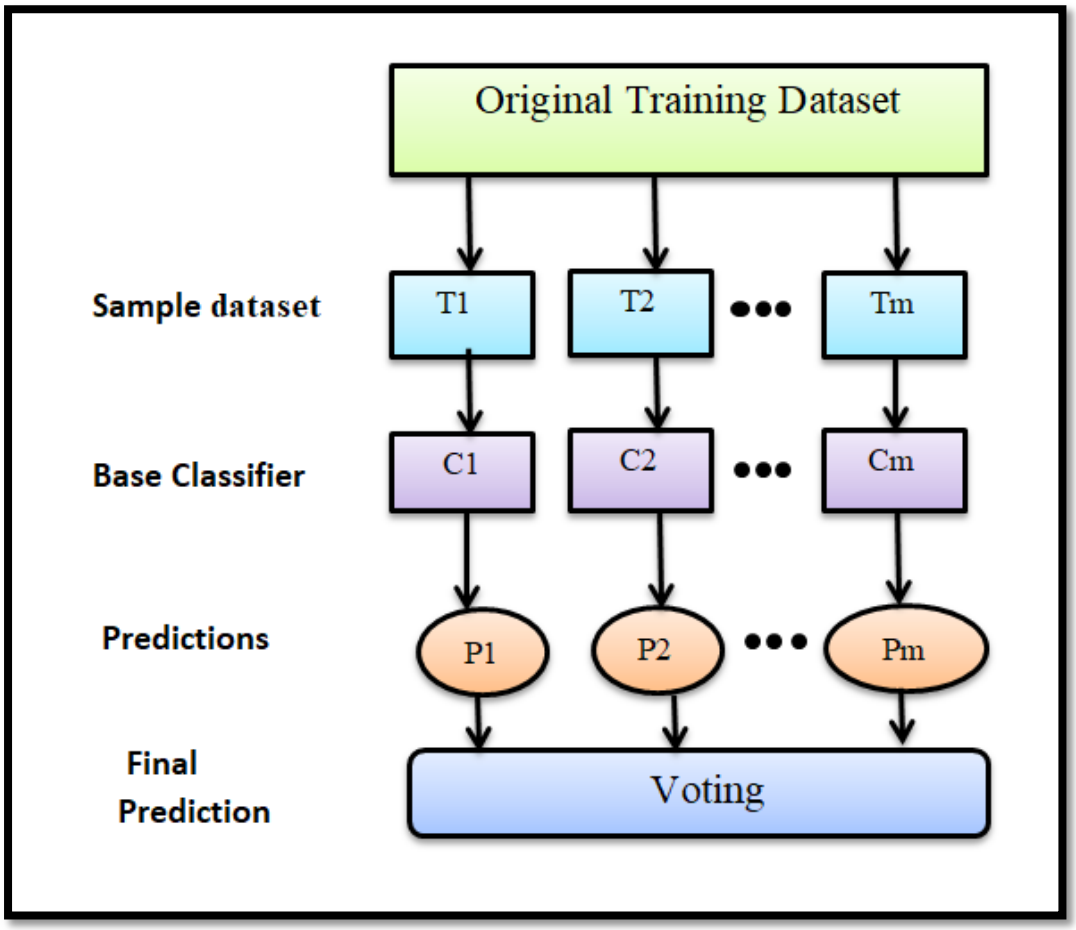

Fig 4: Process flow of bagging algorithm

A sort of ensemble machine learning method called bagging combines the results from numerous learners to enhance performance. These algorithms work by subdividing 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 by majority voting technique. Figure 4 explains the process flow of the bagging algorithm.

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) Collect, organize and format the data:
Using the model necessitates more than just raw data. It is necessary to collect data, store it when needed and organize it to achieve the desired results.
b) Examine and select features:

After preprocessing, the data is evaluated to produce useful information and conclusions 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 contains the most information and will be used to train the vast majority of the samples that will result in the yield. Nearly 80% of the samples collected are used in the training set. The testing set makes use of the final piece of data to determine how well the system works.

Step 3:Training sets for classification

The model system will be determined by the problem's complexity and the structure must be chosen accordingly. During training, it is possible to change the construction, design, and structure of the training set.

Step 4: Determine the RMSE, R^2 statistic and MSE values for each model

Repeat the trained classification model on the test set and calculate the MSE and RMSE values. Compare the outcomes with various classification models. The model with the various classification models. The model with the best crop yield prediction has the lowest MSE and RMSE values as well as the highest R^2 statistic value. Figure 5 depicts the flow chart for the classification technique used to forecast the crop yield.

Step 5: Predict Yield

The trained model is used to predict the output when new input is provided. The trained model was saved as a file so that it could be estimated with new data. These models were properly trained on the training dataset before being tested on the testing dataset. To make accurate predictions, this prediction model employs machine learning techniques that learn properties from training data.

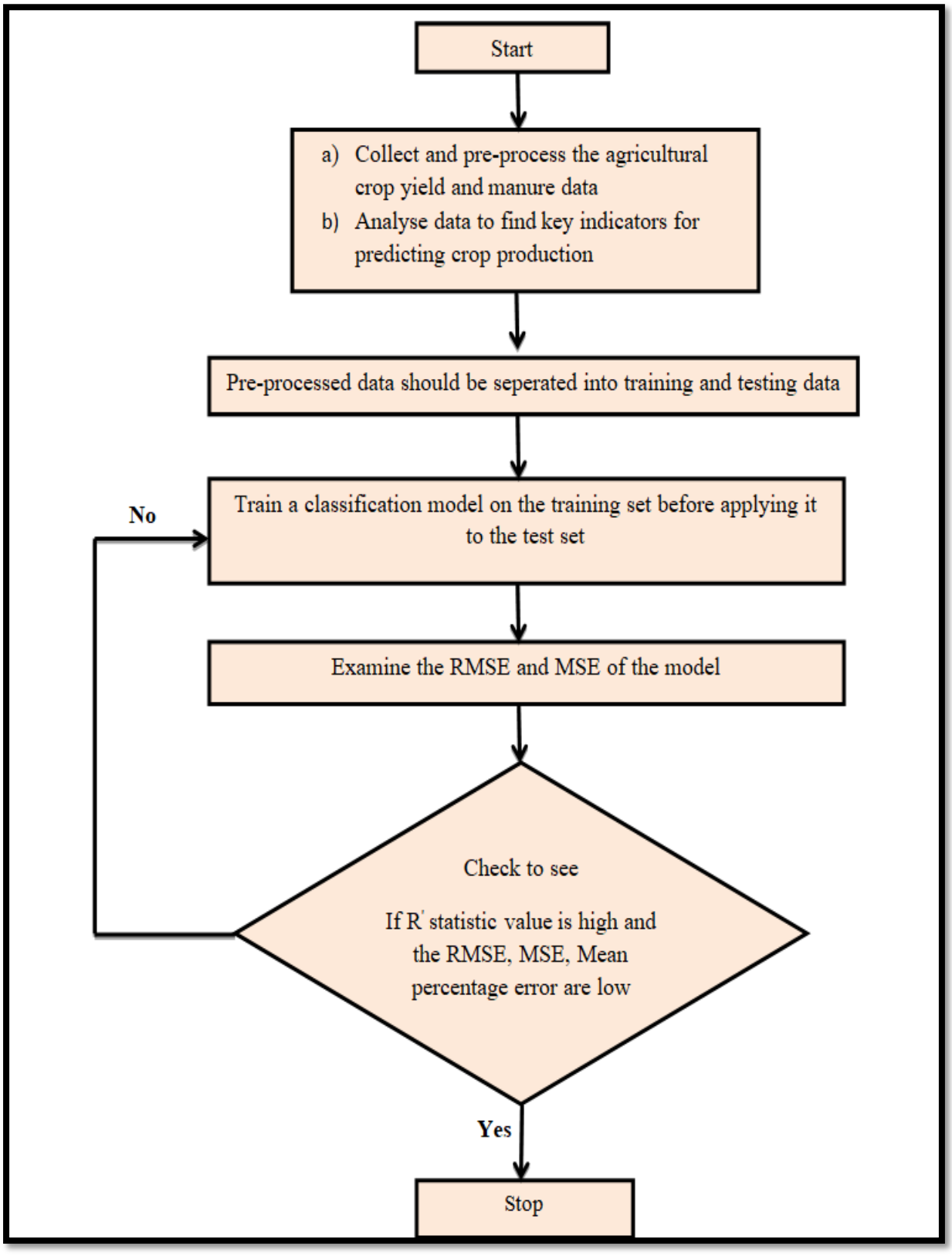

Fig 5: Flow diagram for classification methodology for predicting oilseed crop yield

3.8 Prediction results

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

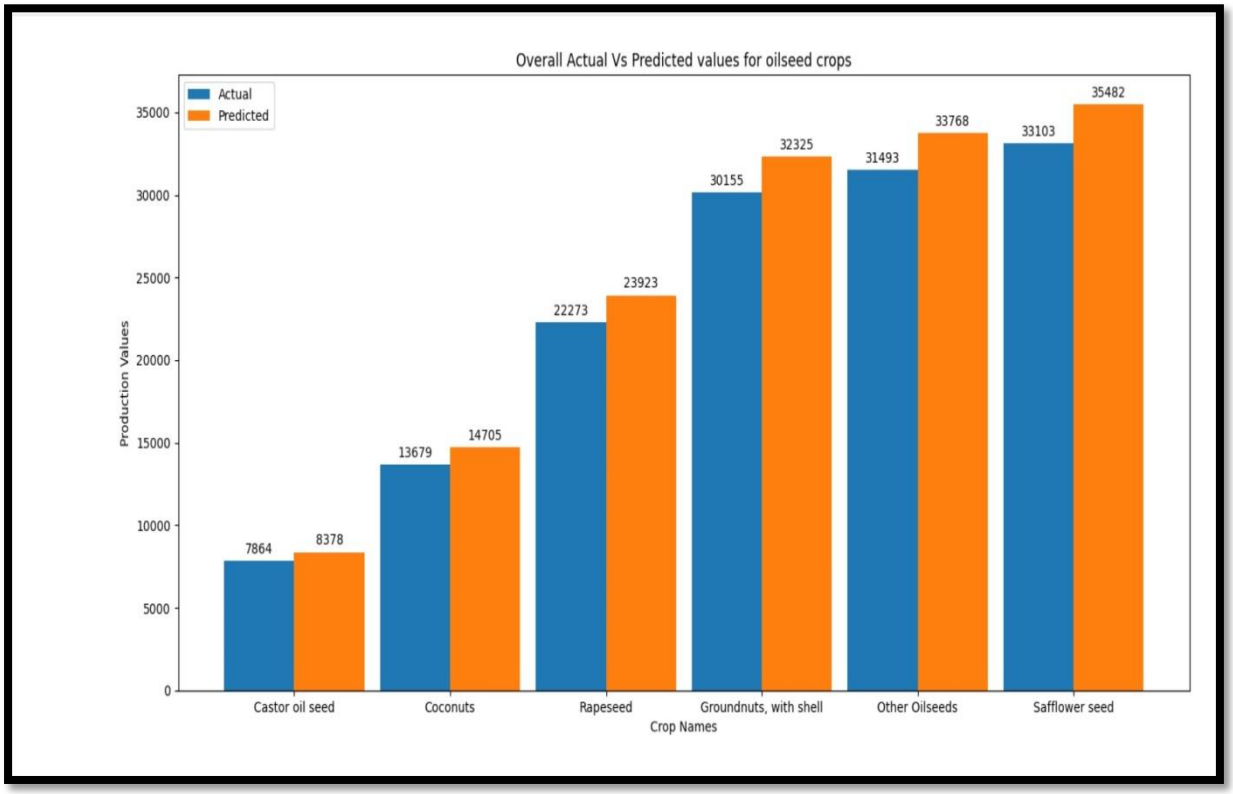

Fig.6 Comparison of actual value versus predicted value for all crops in Tamil Nadu

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

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

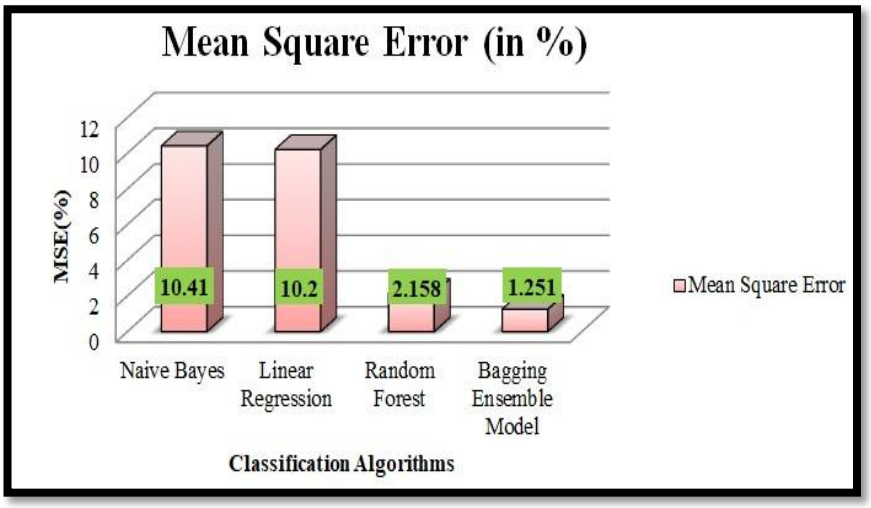

Fig.7. MSE of proposed models

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

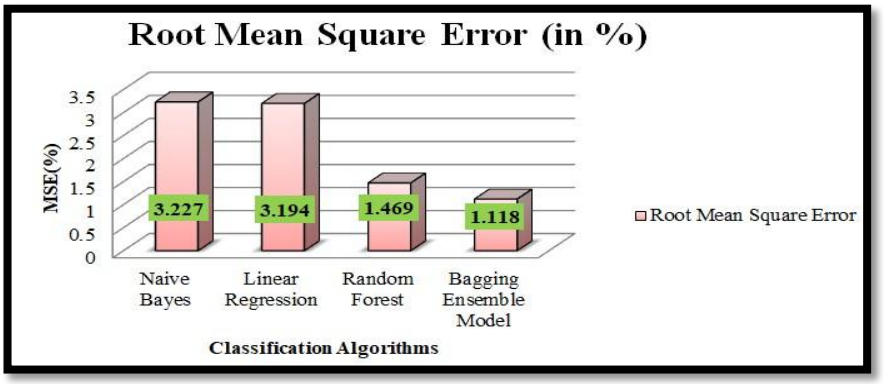

Fig.8. RMSE of proposed models

3.9.1 Comparison with different models

We obtained a **98.5%** accuracy rate, indicating that this model is more accurate at predicting yield. The bagging technique (Bootstrap Aggregation) outperformed other models in terms of accuracy. This is due to changes made to the model and structure during training. Figure 9 depicts a graphical comparison of machine learning model accuracy while Table 1 compares the accuracy of various proposed algorithms.

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 classifier's accuracy is simply how often it predicts correctly. It is calculated by dividing the number of correct predictions by the total number of predictions. Figure 9 compares the accuracy of all machine learning algorithms**.**

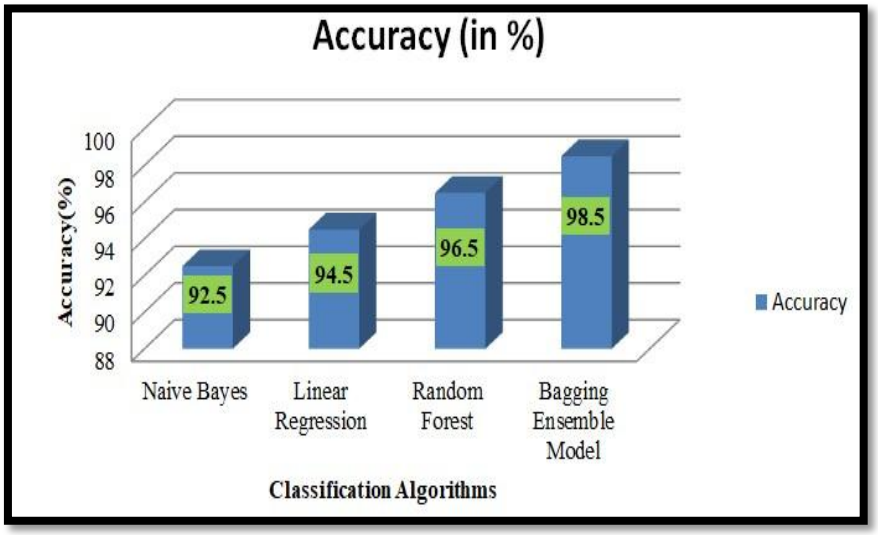

Fig.9. Comparison of accuracy for all proposed models

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.

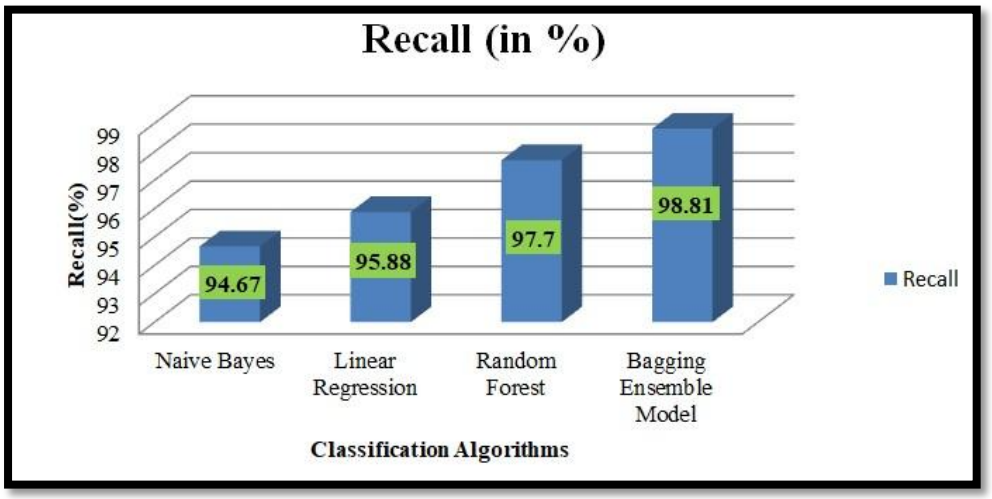

Fig.10 Comparison of recall for all proposed models

4.2 Precision

For a given label, precision is defined as the ratio of true positives to predicted positives. Figure 11 compares the precision values of all machine learning algorithms.

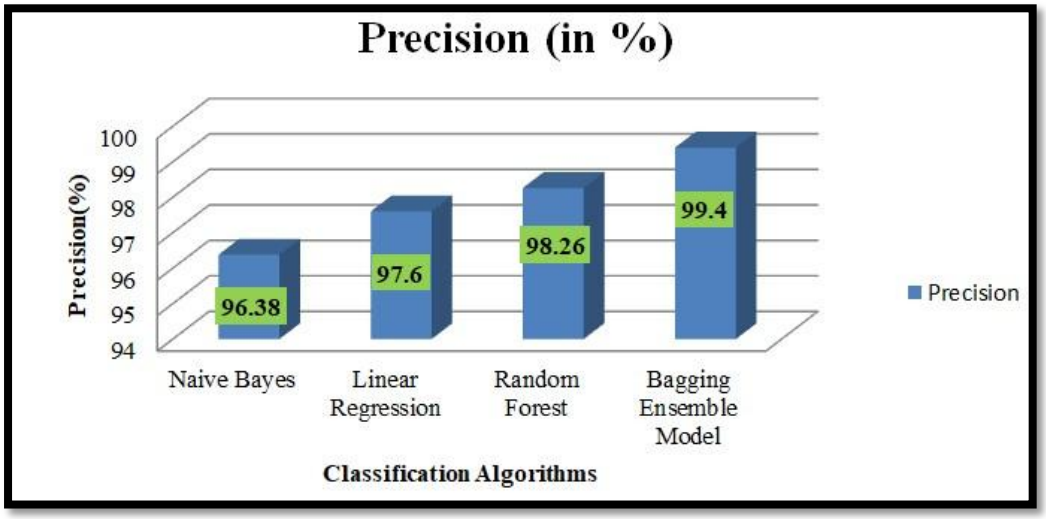

Fig.11. Comparison of precision for all proposed models

4.3 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

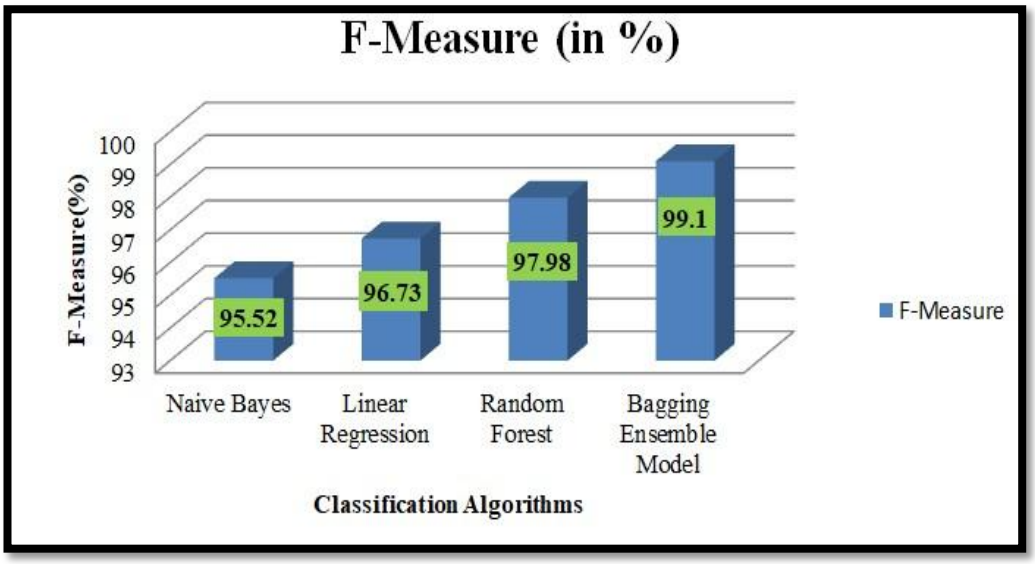

Fig.12. Comparison of F-measure for all proposed models

4.4 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

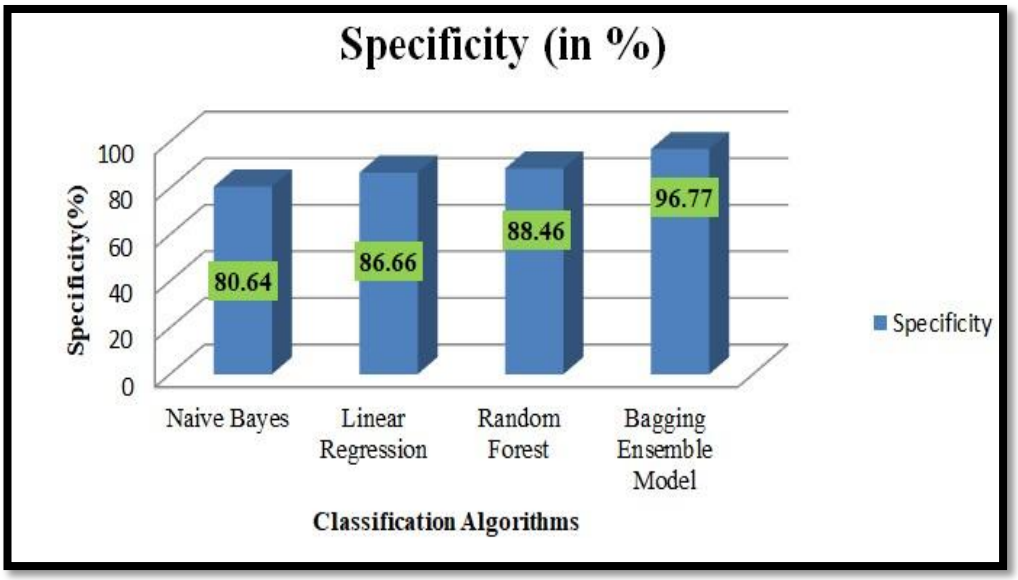

Fig.13 Comparison of specificity for all proposed models

4.5 Execution time for all machine learning algorithms

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

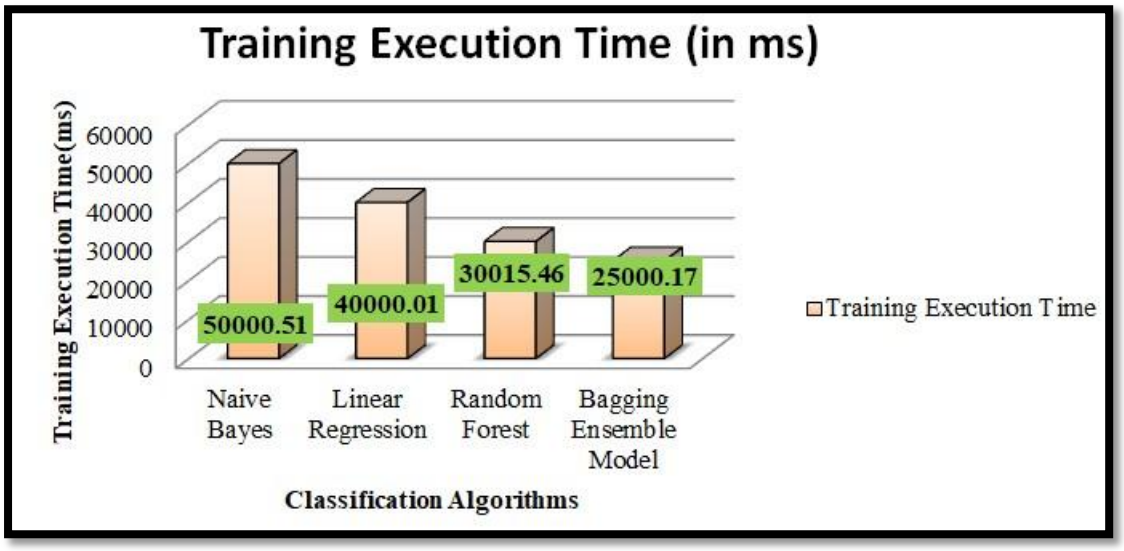

Fig.14. Comparison of training execution time for all proposed models

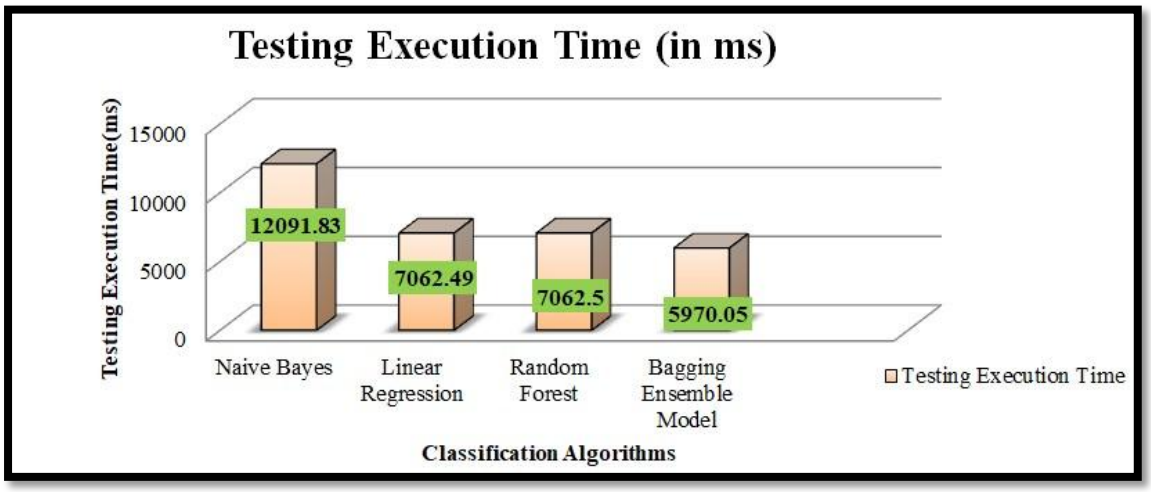

Fig.15. Comparison of testing execution time for proposed models

5. Results and Discussion

5.1 Overall observations of the proposed algorithms

5.1.1 Observations on Bagging:

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 study. When the value for the random state parameter is increased, the accuracy increases. Many trails were built. The top three trials were considered for the parameter "random state" with an assumed value range of 0 to 2. We can see an increase in model performance from 97% to 98.5% by iterating through different random state range values. Below the value of 1 for the random state, the accuracy begins to deteriorate; again, if you change the random state, the valuesseen will vary. In this case, we see an improvement in accuracy when the parameter values get increased.Figure 16 depicts the observations of the bagging ensemble model.

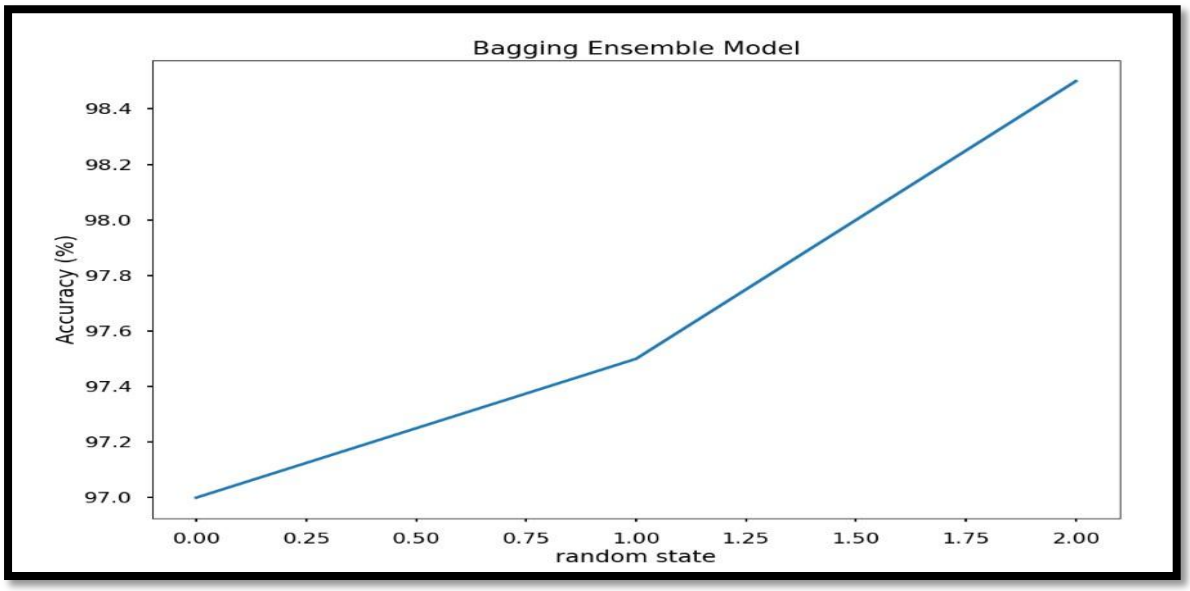

Fig 16. Observations on the parameter for Bagging Ensemble Model

5.1.2 Observations on Random Forest:

The various parameters considered for the study are 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, max samples using sklearn library. The following observation was made during the study. When the value for random state decreases and the number of estimators increases, accuracy tends to rise. Three trials were conducted for the parameters "random state" and "ccp_alpha" with assumed values of 2,4,6 and 0.0, 0.1, 0.2 respectively, and the resulting accuracies were 95%, 95.5%, and 96.5%.Figure 17 depicts the observations on parameters of random forest algorithm.

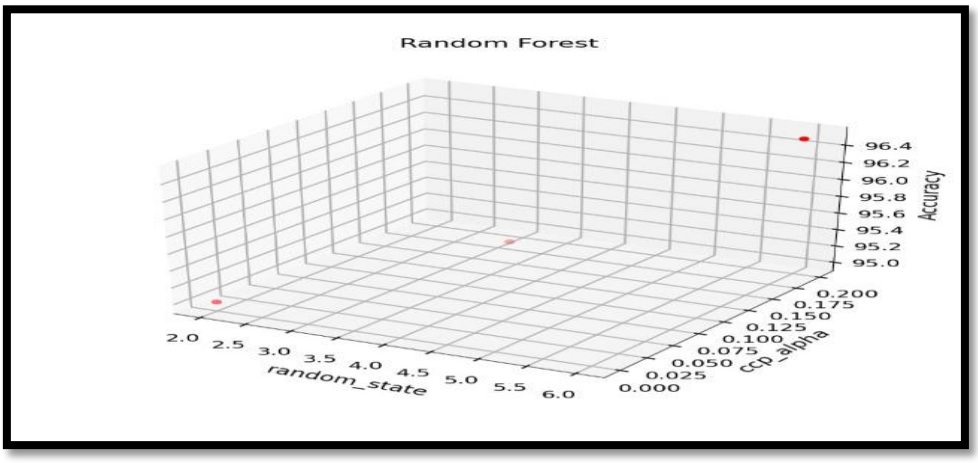

Fig.17 Observations on the parameter for Random Forest

5.1.3 Observations on Linear regression:

The study's parameters include an independent variable, two dependent variables (x1 and x2), the slope (m), and the intercept. The accuracy increases as the slope (m) value and

intercept value get decrease. Three trials were performed with varying values of 0.36, 0.21, 0.12 for the parameter intercept and varying values of the slope(m) values 0.12, 0.08, 0.02 decreased, the accuracy increased while keeping the other parameters constant, and their resulting accuracies were 93%, 93.5%, and 94.5%, respectively.Figure 18 depicts the observations on parameters of linear regression algorithm.

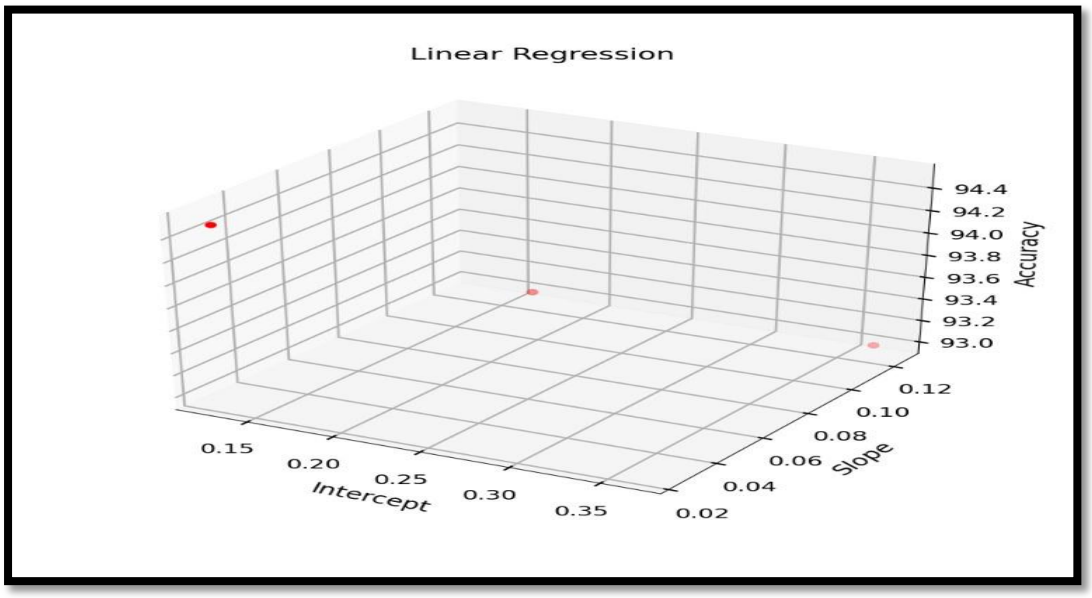

Fig.18 Observations on the parameter for Linear Regression

5.1.4 Observations on Naive Bayes:

The various parameters considered for the study include nb_alpha, priors, smoothing, epsilon, sigma, and theta. Many trials were performed. Out of which top three trials were taken. 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 get decreased with varying values of 0.87,0.54,0.23 while retaining the same values for other parameters. Then their resulting accuracies were 91%, 91.5%, and 92.5%of the three trials are shown. Figure 19 depicts the observations on the parameter for naive bayes algorithm.

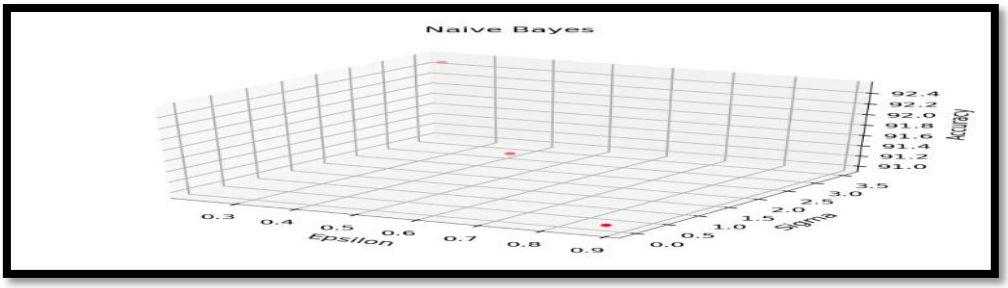

Fig.19 Observations on the parameter for Naive Bayes

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

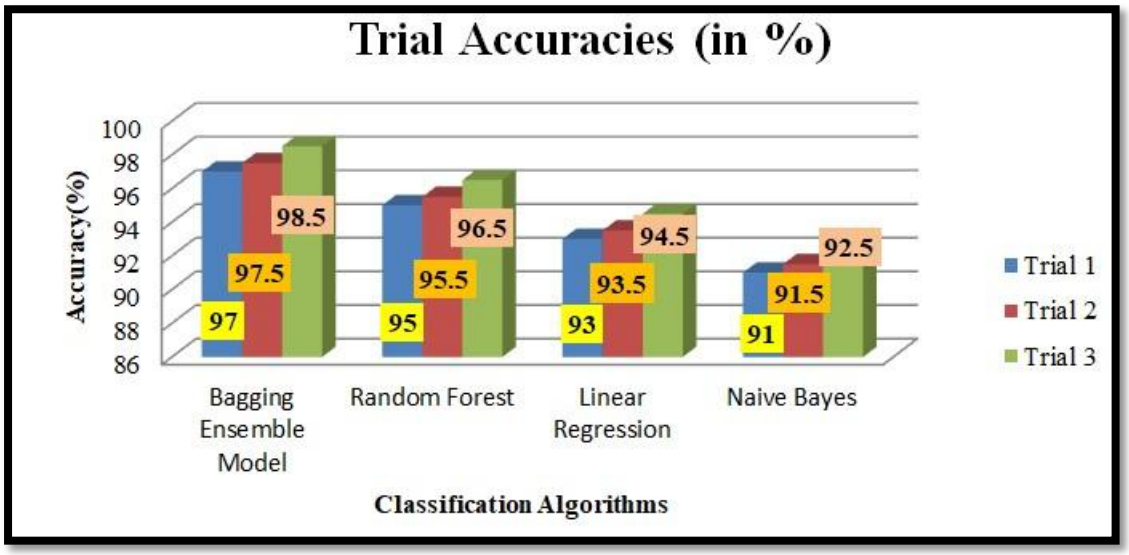

Fig.20 Trail accuracies for proposed models

5.1.6 Discussion based on District wise crop yield

The purpose of this paper is to understand location-specific oilseed crop yield analysis, which will be handled by a machine learning algorithm. A dataset in .csv format was considered for this study. In this scenario, the training test uses 80% of the data, and the validation set uses 20% of the data. After successful training and testing, the model's accuracy was determined indicating how well the model performed in forecasting the yield. Figure 22 depicts a graphical user interface for predicting future crop yield. Figure 21 depicts a summary of all Tamil Nadu oilseed crop production districts.

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

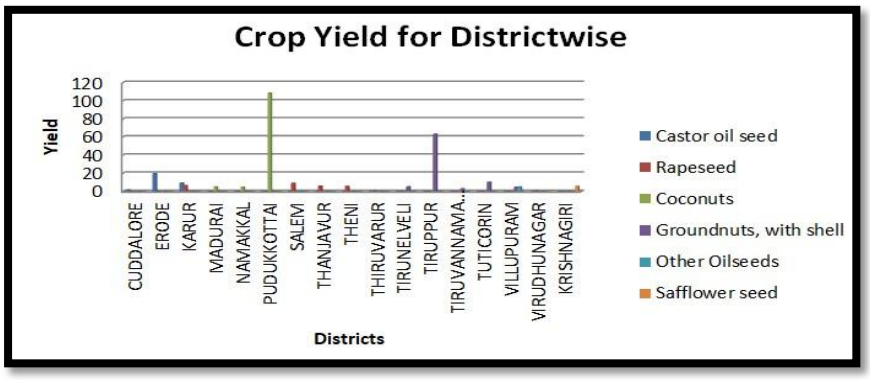

Fig.21 District-wise crop yield statistics

5.1.7 Recommendation system for manure and oilseed crop yield

5.1.8GUI Creation

The study assists farmers in deciding which crop to grow in a specific area at a specific time, as well as providing information on whether it will be profitable or not during forecasts. Furthermore, it indicates low or high yield with ranges to assist ranchers or end users in making successful selections, saving time and accuracy. The prediction module allows users to share the district name, state name, crop, season, and crop year to predict area, soil type, pH value, and soil content of nitrogen, phosphorous, and potassium. After entering these attribute values, the user may use the "Predict crop yield range" button to determine the yield of a specific crop in the future with a high or low yield rate. The "Predict organic manure type and amount of manure" button on this suggestion system also assists users in forecasting separate quantity intake of NPK to be taken for a specific crop as well as the manure type to be used for a specific region. The outcome is obtained by considering 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 for each crop based on the prediction error.

Predicted value ± Predicted error	----------(2)

In figure 16, the crop production result range for the Coimbatore district's castor oilseed crop for the entire year is calculated based on the average of prediction errors of all castor crops in a specific 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 aids in manure type and quantity recommendations for oilseed crops in a specific area at a specific time. If the prediction value is less than the mean score, the crop is considered as low yielding and if it is greater than the mean, the crop is considered as high yielding.

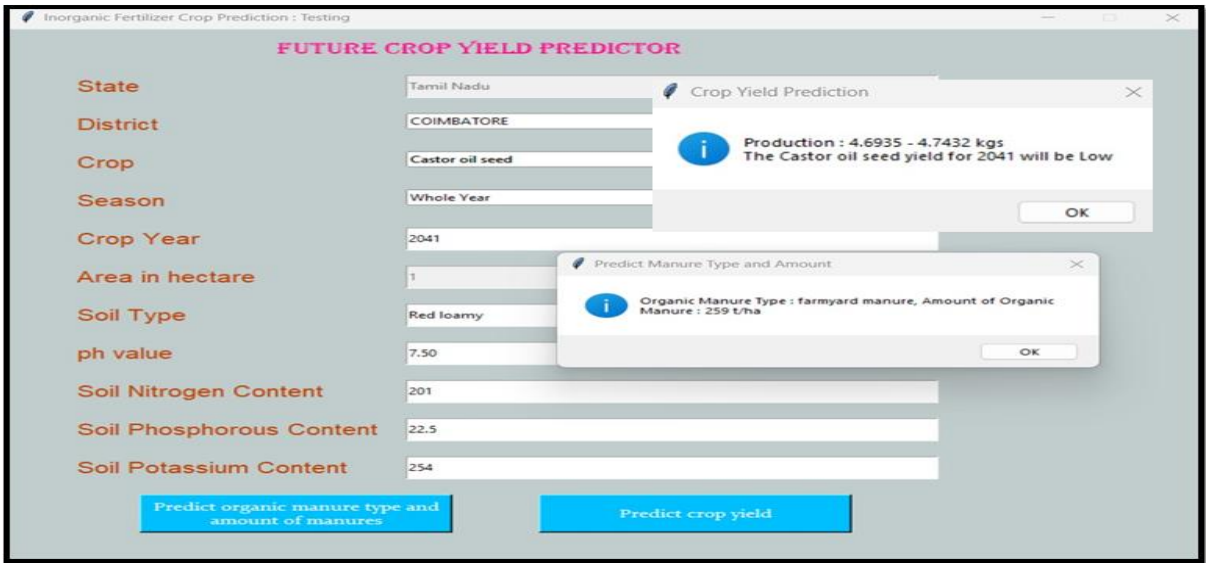

Fig.22. Recommendation systems for crop yield and predictor

GUI data visualization can also be achieved by plotting yield variables with varying parameters. Data visualization such as graphs or figures aids in the capture and comprehension of ideas. The primary goal of GUI data visualization is depicted in figure 22. The prediction module makes it easier to find patterns, correlations, and outliers in large datasets. The graph above depicts the relationship between the district and yield.

6. Conclusions

This study looked at crop yield forecasting algorithms that used temperature, season, and location as inputs. Forecasting yield in a specific district can be done using rainfall, temperature, and other variables such as season, location and organic manure data. When all factors are considered, the bootstrap aggregation technique is the best classifier among all others. Increasing the number of parameters in the dataset improves accuracy. Bagging is found to be the best prediction algorithm when compared to other prediction algorithms such as random forest, linear regression, and naive bayes. The database contains a much larger number of variables, resulting in more accurate predictions. The creation of this work will aid farmers in reducing risk and maximizing crop yields to improve their agricultural resources.

In this work, we used soil test results and organic manure dosage to forecast future crop yields. We have also developed a recommendation system for farmers to determine the best crop to cultivate in the coming season as well as manure type and quantity recommendations for ranchers. This will not only help farmers determine the best crop to cultivate in the coming season, but it will also help bridge technological and agricultural divides. The limitation of our work is that yield is only implemented in 30 districts of Tamil Nadu but not in other states. The future work of our project aims to include regional languages such as Tamil, Telegu, Hindu, Kannada, Malayalam, and others in the graphical user interface which helps the farmers benefit across the country. In addition, Natural Language Processing (NLP) can be utilized to acquire farmer queries via voice mode and provide the desired result to ranchers via GUI. This voice mode query system enables uneducated farmers to easily 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