Indian J. Applied & Pure Bio. Vol. 38(1), 276-292 (2023). **ISSN: 0970-2091**
A web of Science Journal

The oilseed crop yield prediction using Machine learning Algorithms

[*1]C. Mithra, [*2]A. Suhasini and [*3]S. Jothilakshmi

[1&2]Department of Computer Science and Engineering, Annamalai University,
Annamalai Nagar-608001 (India)
[*1]mithrac.official@gmail.com
[*2]suha_babu@yahoo.com
[3]Department of Information Technology, Annamalai University,
Annamalai Nagar-608001 (India)
[*3]jothi.sekar@gmail.com

Abstract

Agriculture provides significant economic support in India. Population growth is the most serious threat to food security. Population growth raises demand, requiring farmers to produce more to increase supply. Crop yield prediction technology can assist farmers in producing more yields. The primary goal of this study is to forecast oilseed crops yield for various districts of Tamil Nadu state. Machine learning classification algorithms are used to forecast oilseed crops yield. The actual yield data from 1961 to 2007 years are used as a training set and from 2008 to 2019 as a validation set. The results of the proposed algorithm are compared with those of existing algorithms namely linear regression, Support Vector Machine, Naive Bayes, and K-Nearest Neighbour, and an accuracy of 86.3%, 83.33%, 80.3%, and 77.56% respectively is observed. The actual and predicted oilseed crop yields from 1961 to 2019 are analysed for the yield forecast model, and the error percentages with underestimated and overestimated predictions for the various districts of Tamil Nadu are calculated. According to the study, the results of the linear regression are found to be superior to those of other algorithms. The study also assists farmers by providing a recommendation system to decide which crop to plant in a specific area and time.

Key words : Oilseed crop, yield prediction. Machine learning algorithm, Support vector Machine Naive Bayes, MLP.

[1]Research Scholar, [2]Professor, [3]Associate Professor

Agriculture is India's predominant career, and the country's economy is entirely depending on it and more than 70% of its rural households nevertheless rely mostly on agriculture. As a result, many farmers have started to include new technologies and strategies into their farming operations to improve their yields. Farmers, on the other hand, are unaware of the significance of cultivating crops at the right time and area. In this situation, figuring out crop adaptability and yield using multiple factors that influence production can improve crop quality and yield, resulting in better financial growth and profitability[13]. Data mining techniques are required for accomplishing realistic and powerful solutions to this problem. Agriculture has long been a natural target for big data. Environmental conditions, soil variability, input stages, rainfall, *etc*. have made it even more important and seek assistance when making crucial farming decisions[22]. Machine learning (ML) plays an important role as it has a decision support system for crop yield prediction, which includes assisting with decisions on what crops to grow and what practices to follow during the crop's growing season[16]. This paper investigates various machine learning techniques used in crop yield estimation and provides an in-depth evaluation of their accuracy. The proposed algorithm for oilseed crop yield prediction attempts to combine the strengths of the aforementioned approaches while avoiding their limitations. Methods are used to convert raw data from customers into useful information. Most farmers anticipate higher yield in the ensuing season based on their long-term field experience with specific crops. In addition, they don't get a fair rate for their crops. This usually happens due to obsolete methods of irrigation or poor crop selection, but it can also happen when crop yields are less than expected. Most of the time farmers do not gain the anticipated crop yield due to a variety of factors[13]. The crop yield dataset is made up of many additives. The most desirable crop yield may be estimated by analysing the soil and atmosphere for a specific region to increase crop production. Data with a high level of noise, erroneous data, outliers, biases, and incomplete datasets can notably reduce the predictive power of models. The RMSE and MSE are used for data validation[1], which indicates that an interactive linear regression model could predict crop yield based on-farm conditions, social factors, and climatic inputs. In this research work, we analyse four machine learning models which demonstrate, how the future crop yield can be predicted based on attributes like humidity, temperature, soil type, area etc. to improve crop yield prediction accuracy. The current article discusses the details about data collection, pre-processing, feature selection and compares it with four machine learning algorithms namely Linear Regression, K-Nearest Neighbour, Support Vector Machine and Naive Bayes to know which algorithm effectively works for crop yield prediction.

Related works :

Prediction model was primarily based on a random forest algorithm to analyse the future agricultural crop yield. The factors mainly used in this research area are based on soil type, pH value, fertilizer, temperature, precipitation, and humidity. The evaluation of the model has been done by comparing the proposed algorithm with the XG Boost classifier, KNN, and logistic regression and found that the proposed algorithm gives better accuracy[1]. Hybrid deep learning[8] based crop yield prediction system for paddy crops using a deep belief network and fuzzy neural network system. This system mainly focuses on processes like foreseeing crop yield, plant diseases examination, crop enhancement, *etc.*

Hybrid MLR-ANN model[10] based on machine learning and AI to predict appropriate paddy crop yield. Planting area, irrigation area, fertilizer usage, and irrigation specifications were all considered. Rainfall, maximum temperature, and minimum temperature are indications of climatic variables. In their research, 5 features from the crop dataset were deemed vital for prediction out of the 15 features in the crop dataset. Using evaluation metrics, the hybrid model's prediction accuracy is compared to ANN, MLR, SVR, KNN, and random forest models. Linear algorithms (LDA and LR)[6] work better when compared to non-linear algorithms (NB, SVM, KNN) for the prediction of maize yield in Eastern and Southern Africa.

A model that offers crop suggestions namely paddy, wheat, and maize to farmers in choosing the right soil to enhance crop yield. The results obtained from the models are compared with the performance metric R square score. The decision tree regression is found to be best for crop yield prediction[15].

A model for predicting wheat crop yield based on data mining classification algorithms and stepwise linear regression[4]. The study found that MLP and additive regression performed better than other algorithms. Combining crop modeling and machine learning improves crop yield predictions in the corn belt of the United States. It has also been established that soil and water variables could improve machine learning yield prediction in the corn belt of the central United States[21]. Determined relevant attributes and pre-processed the data using big data techniques and machine learning algorithms to predict which algorithm suits best for crop yield prediction[18]. To assist farmers in determining which crop to grow in a particular season and forecasting whether or not that would be profitable[21]. RNN[3] model based on deep learning to predict wheat crop yield. They found that Multivariate Linear Regression (MLR) is less effective when compared to the proposed algorithms. The most needed features for accurate crop yield predictions[19]. According to the findings, FFS features combined with random forest can be used to accurately predict crop yield.[11] determined advanced ensemble regression to predict the crop yield prediction based on the phenotype factors which include precipitation, solar radiation, maximum and minimum temperature, *etc.* As a result, the proposed model outperformed several supervised machine learning and advanced learning ensemble algorithms. To estimate autumn crop yields,[7] compared the performance of SVR, RF, and DNN models based on R squared, RMSE, and MAE. It was determined that DNN performed the best, followed by SVR and RF, in predicting autumn crop yield with the highest accuracy.

A method was proposed for evaluation that is better than other existing methods of evaluation[2]. It evaluated all the regression techniques namely linear regression, polynomial regression, support vector regression, decision tree regression, random forest, and XGB Regression for the two crops of four individual states. A DNN approach[14] that provides superior prediction accuracy with RMSE being 12% of the average yield. A method to increase the yield of crops. The K-means[5] clustering algorithm allowed for a faster search in less time. The Apriori algorithm assisted in adding up the specified location, and the Naive Bayes algorithm was used to implement a system that predicted the crop name and accurately estimate yield in a specific farm.

A model named ANN[12] algorithm to predict the crop yield for Maharashtra. They concluded that Linear Regression algorithms can be replaced with ANN methods to give accuracy for crop prediction. The regression algorithms[17] viz., kernel ridge, lasso, and ENet algorithms and stacking regression to predict the yield. According to the results, stacked regression is better than other models. [9](*Crop Yield and Rainfall Prediction in Tumakuru District Using Machine Learning*, 2018) carried out a study using various machine learning algorithms like linear regressions, SVM, KNN, and decision trees. Out of which, SVM is found to have the highest efficiency when compared to other algorithms for future crop yield prediction.

Proposed methodology :

The proposed model's overall architecture, which employs a linear regression algorithm, is described in fig. 2.

The research in this paper was carried out with the aid of PyCharm Community Edition 2019.1.3 64. The LNR vital classification algorithm is applied to data collected from the official government website, which contains oilseed yield as well as soil, temperature, and humidity. Linear regression is a supervised learning-based machine learning algorithm. It performs regression tasks for predicting datasets based on a given independent variable. This algorithm classifies the yield attribute into high and low. After the generation of a linear regression algorithm, based on the expected targets, the final decision is made. Crop yield forecasting allows for better production plans and decision-making. The proposed system includes a prediction module based on a data mining algorithm namely linear regression, which is used to predict yields of major oilseed crops in Tamil Nadu based on historical data. The GUI has also been created to help the end-users to analyze the future yield prediction.

Agricultural dataset :

The following are the sources for the datasets used in this study,

- The weather dataset is obtained from the Department of Meteorological Centre India
- Different Oil seed yield dataset was collected from ICRISAT, Tamil Nadu government website (www.data.govt) as well as from the University Department of Agriculture, Tamil Nadu.
- Environmental variables like sunshine were collected from the weather atlas.

This research considers some essential climatic variables namely soil temperature, soil pH, rainfall, humidity, sunshine, and the minimum and maximum temperatures of a particular place and area. Some agronomic parameters of the soil such as texture viz. red loamy, red sandy calcareous, saline coast, deep red loam, alluvium, non-calcareous brown, deep red soil, calcareous black, black, red sandy soil, non-calcareous red, lateritic, coastal alluvium, clay loam, saline coastal alluvium, etc., as well as different seasons are included.

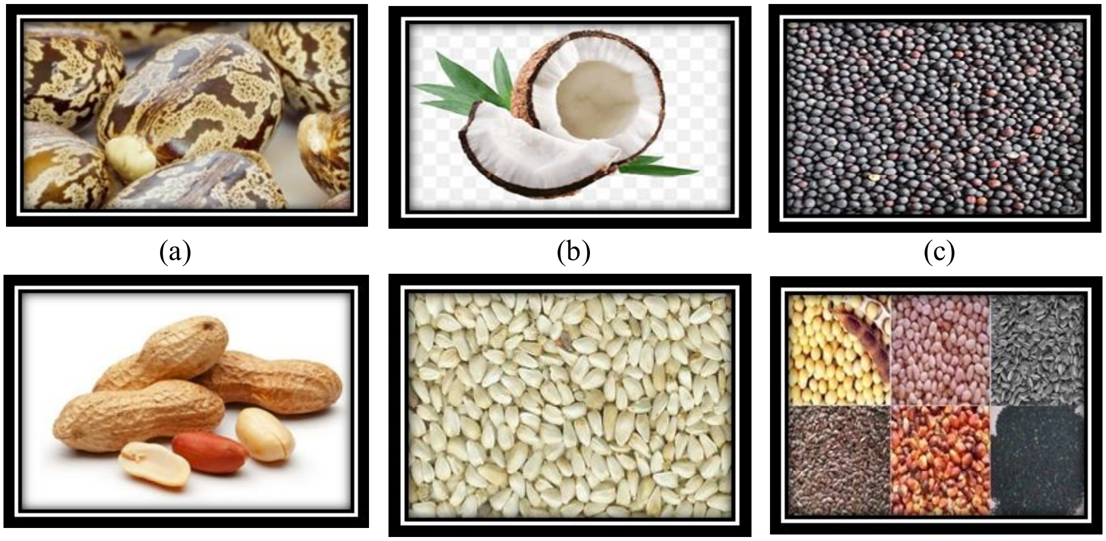
The following crops have been considered for this research:

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

Crop details :

The Directorate of Oilseeds Development (Ministry of Agriculture and Farmers Welfare, Government of India (https://oilseeds.dac.gov.in/)[23] has reported the oilseed production in Tamil Nadu for the year 2019-2020 and 2020-2021. The details of selected crops included in this study are presented below.

Oilseed production increased from 361.009 lakh tonnes in 2020-21 as a result of the TMOP's concerted efforts (4th Advest). This is due to an increase in area as well as an increase in productivity from 1224 kg/ha to 1254 kg/ha during 2019-20 and 2020-21, respectively. Due to the improving environmental conditions and support from the government for oilseed production/developmental programs and policies, India produced 361.009 lakh tonnes of oilseeds during 2020-2021, followed by 332.192 lakh tonnes during 2019-2020, with a record productivity level of 1284 kg/ha during 2017-2018 and 1254 kg/ha during 2020-2021. In the last two years, 2019-2020 as well as 2020-2021, India had the highest growth rates[23]



(a) (b) (c)

(d) (e) (f)

Fig. 1: oilseed Crops (a) Castor (b) Coconut (c) Rapeseed (d) Groundnut (e) Safflower (f) Other oilseeds

a) Castor crop growth rate in Tamil Nadu:

Castor is a significant non-edible oilseed crop in India, with enormous industrial and commercial values. With an area of 13,900 hectares, Tamil Nadu is an important castor growing state in India. Salem, Namakkal, Erode, Dharmapuri, and Perambalur are the major castor-producing districts. It is worth noting that, when compared to other crops, castor has experienced the highest growth rate in terms of production and productivity over the last two decades. The development of hybrid castor has resulted in spectacular progress in terms of area, production, and productivity in castor. During the Kharif season in 2019-2020, Castor produced 0.016 lakh tonnes on an area of 0.053 lakh hectares with a yield rate of 312 kg/ha in Tamil Nadu.

b) Coconut crop growth in Tamil Nadu:

Secondary oilseeds sources such as coconut are established by combining an edible group of plantation crops. This perennial crop is grown on 4.40 lakh hectares in Tamil Nadu, with a yearly production was estimated 52,140 lakh nuts and a productivity of 11,560 nuts for each hectare in 2019-2020. Coimbatore, Tiruppur, Thanjavur, and Dindigul are the major coconut cultivating districts.

c) Rapeseed crop growth in Tamil Nadu:

Rapeseed oilseed crop is one of the major primary sources of oilseed edible group with an area and yield of 0.001 lakh hectares and 233 kg/ha respectively during 2019-2020.

d) Groundnut with shells growth in Tamil Nadu :

Groundnut oilseeds crop is one of the major primary sources of oilseed that has been established by combining edible group during 2019-2020, with an area, production, and yield of 2.092 lakh hectares, 5.185 lakh tonnes, and 2479 kg/ha in Kharif season, respectively. In the Rabi season, the area, production, and yield were 2.092 lakh hectares, 5.185 lakh tonnes, and 2479 kg/ha, respectively.

e) Safflower crop growth in Tamil Nadu :

Safflower oilseed crop is now a minor commercial crop, with approximately 800,000 tonnes produced each year. Safflower oilseeds crop is one of the major primary sources of oilseed edible group having no significant yield during 2019-2020.

Dataset Description :

Data for these crops and variables are provided as input. Initially, a set of data is collected that encompasses parameters such as state name, district name, humidity, temperature, productivity, production, and so on, for the above oilseed crops in all districts of Tamil Nadu. This CSV dataset was compiled between 1961 and 2019. The final dataset comprises of 1012 records with 19 attributes.

Pre-processing :

Before the application of any machine learning technique on a dataset, some pre-processing needs to be done. The data collected from various sources are often in raw form. The raw data contains information that is incomplete, inconsistent, or obsolete. As a result, such redundant data must be filtered prior to processing. The supplied information series contains a large number of ‘NA’ values, which can be filtered in Python by replacing missing values with an average value. A robust scalar is being used to remove outliers. Then, the data transformation is done to make the data access easy. The final dataset is normalized to ensure that all values fall within a consistent range. The formulae for the normalization technique are shown in Equation 1. (min-max scalar). After applying the min-max scalar, data gets normalized to a factor from 0 to 1.

$$\boxed{X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)}} \quad (1)$$

Data Analysis :

Following raw data pre-processing, the data must be ensured through the processes of inspecting, cleansing, transforming, and modeling to arrive at useful information, and conclusions, and assist decision-making to proceed further with a clear understanding of the dataset.

Dimensionality Reduction :

The choosing of high-level features that contribute to prediction accuracy is critical in achieving accurate prediction. There are various feature selection algorithms available but, the best utilized for this research is principal component analysis (PCA) as it helps to transform the dataset into a compressed form. There are a total of 19 features in this dataset. The PCA algorithm was used to select 13 essential feature subsets. These feature subsets were fed into the linear regression algorithm to determine the best feature subset. Humidity, rainfall, area, production, and other factors were chosen. Once those characteristics were applied to machine learning algorithms and statistical models, they improve the classification accuracy of the model.

Training and Test model :

During the pre-processing stage, the dataset can be divided into training and testing sets. We divided the dataset into parts, with 80% for training and 20% for testing. This is a critical step in creating the model. The training dataset is used to train the model, and the testing dataset is used to validate the model. As an outcome, we fit the model with training data and evaluate it with testing data to determine the model’s accuracy.

Prediction algorithm :

The next step is to generate and train the model after the data has been split. To comprehend the pattern, the action of training a machine learning model necessitates the use of a machine learning algorithm and training data. In this case, we use many machine learning algorithms which are well known supervised learning algorithms with a simple representation.

Comparison of accuracy of proposed model with existing ones :

Table-1. Accuracy of proposed model

Models	Accuracy(in %)
LR	***86.27***
SVM	83.33
Naive Bayes	80.39
KNN	77.56

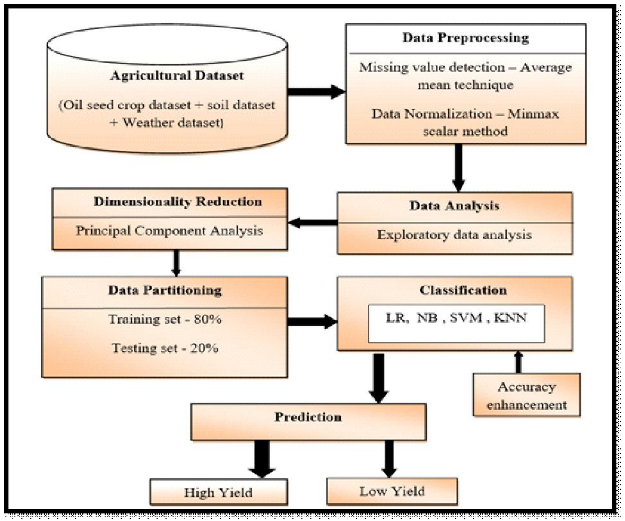


Fig. 2. Methodology adopted to predict oilseed crops yield

Regression model for oilseeds yield prediction :

The oilseeds crop yield dataset consists of a total of 6 crops with 1012 records. Out of which 810 records are considered as training set and the remaining 202 records are considered for the testing set. The machine learning model is developed for oilseed crop yield prediction. All the proposed algorithms namely LR, KNN, NB, and SVM classifiers are compared with one another. Among all these models, linear regression is found to predict the oilseed crop yield accurately. In LR, the prediction of the dependent attributes is done using a single independent attribute and not collectively.

Linear regression is a statistical tool for predicting the exact numerical value that a variable may assume to estimate the future crop yield. PyCharm is a platform for developing a trained model with machine learning algorithms. The regression analysis predicts an entity as a function of independent entities. It also determines the relationship between independent and dependent variables[16].

Algorithm: The steps involved in creating a regression model for a crop yield forecast.

Input: An experimental dataset of weather, crop, and soil data

Output: Crop yield prediction for the experimental dataset.

Method:

Step 1:*a) Collect, format, and organize the data:*Working with the model requires more than just raw data. The information must be gathered, stored according to the necessity, and organized in such a way that appropriate results are obtained.

b) Analyse and select features: After preprocessing, the data gets analyzed into beneficial information and conclusions to proceed further with a clear understanding of all variables. After that, dimensionality reduction is carried out. By applying the PCA algorithm, essential feature sets were selected. The selected features are further used in machine learning algorithms for processing.

Step 2: *The data must be divided into two groups:*The training set will contain the most information and will be used to train the majority of the examples to produce the yield. Approximately, 80% of the samples collected are used as part of the training set. The testing set employs the remaining measure of information to test how well the system performs.

Step 3: *Regression on trained sets:* The model system is dependent on how complex the problem is, and the structure must be chosen accordingly. The construction, modeling and structure can be adjusted during training.

Step 4: *Determine the RMSE, R^2Statistic, and MSE values for each model* Run the trained regression model on the test set and calculate the MSE and RMSE values again. Compare the values with various regression models. The best model for crop yield prediction is the one with the lowest MSE and RMSE values and the highest R2statistic value. The Figure 3 depicts the flow chart for the regression methodology used to predict crop yield.

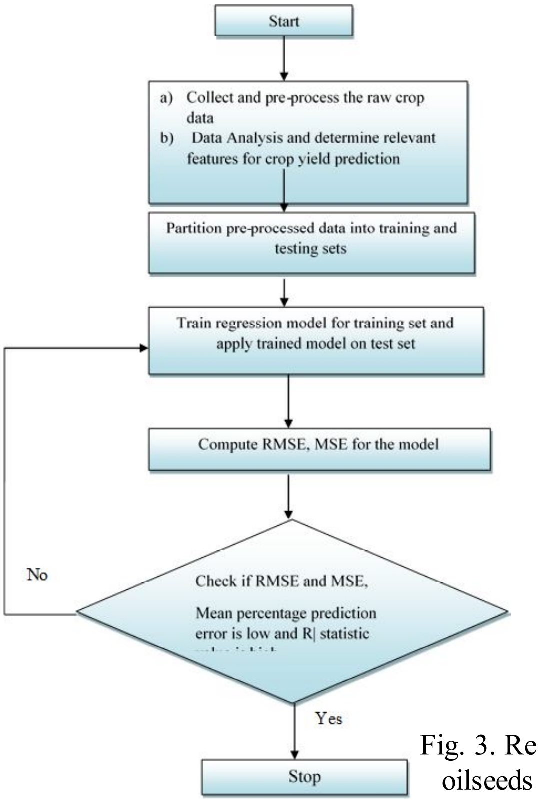

Start
a) Collect and pre-process the raw crop data
b) Data Analysis and determine relevant features for crop yield prediction
Partition pre-processed data into training and testing sets
Train regression model for training set and apply trained model on test set
Compute RMSE, MSE for the model
No
Check if RMSE and MSE,
Mean percentage prediction error is low and R| statistic
Yes
Stop

Fig. 3. Regression methodology for oilseeds crop yield prediction

Predict Yield :

The trained model is used to predict the output when provided with new input. We saved the trained model as a file so that it could be estimated using new input. These models have been properly trained using the training dataset and evaluated using the testing dataset. To make accurate predictions, this prediction model involves an ML algorithm that learns properties from training data.

Prediction results :

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

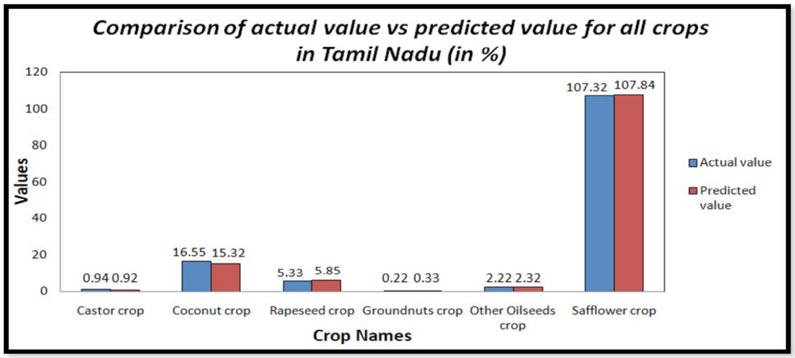


Fig. 4. Graphical representation of a comparison of actual and predicted values for all crop yields

Table-2: Absolute Error Calculation for all crop yield prediction

Crop Name	Actual Value(in %)	Predicted Value (in %)	Absolute Error (in %)
Castor	0.94	0.92	0.02
Coconut	16.55	15.32	1.23
Rapeseed	5.33	5.85	0.52
Groundnut	0.22	0.33	0.11
Other Oilseeds	2.22	2.32	0.10
Safflower	107.32	107.84	0.52

Error calculation for various classification algorithms :

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

Table-3. Formulae for error calculation

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

The Fig. 5 below represents the mean square error for all machine learning algorithms,

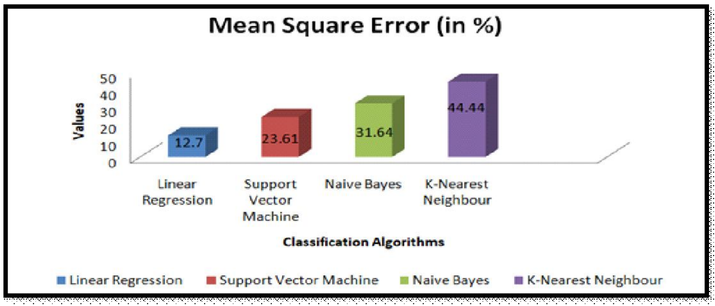


Fig 5. Mean Square Error of proposed models

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

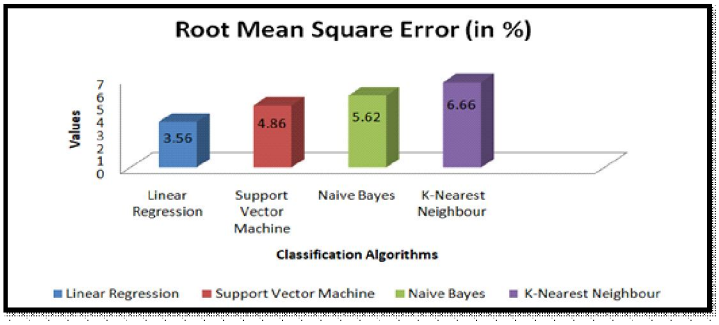


Fig. 6. Root Mean Squared Error of proposed models

Comparison with different models :

We achieved an accuracy of 86.27 percent, indicating that this model is more effective at predicting yield. The linear regression algorithm surpassed other models in terms of accuracy. This is due to adjustments made in the model and structure during training. Table 1 compares the accuracy of different algorithms, and figure 7 depicts a graphical comparison of the accuracy of machine learning models.

Evaluation Metrics :

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			

There are many ways for measuring performance. Accuracy and confusion matrix are some of the most popular metrics. The Confusion matrix is often used to describe the performance of a classification model on a set of test data for which the true values are known and is calculated for all the four machine learning models namely LR, SVM, NB, KNN and it is noted that linear regression works better (true positive value of 89, true negative value of 81, false positive value of 12 and false negative value of 15) when compared to other machine learning algorithms.

Accuracy :

Accuracy simply measures how often the classifier predicts correctly. It is defined by the ratio of the number of correct predictions to the total number of predictions. The Fig 7 below compares the accuracy of all machine learning algorithms.

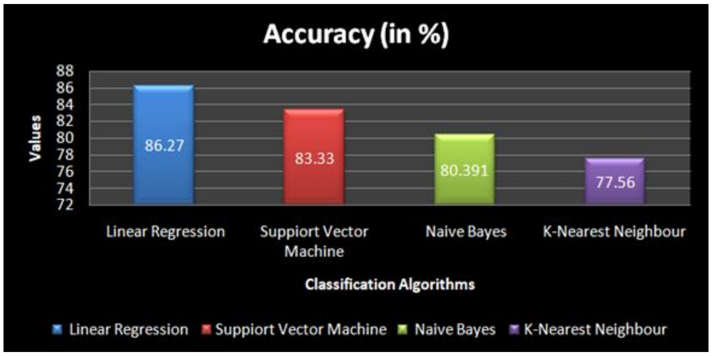


Fig. 7. Comparison of accuracy with proposed ML models

Recall :

The recall is a ratio of the correct detection over the total number of actual positive samples. The Fig 8 below compares the recall values of all machine learning algorithms.

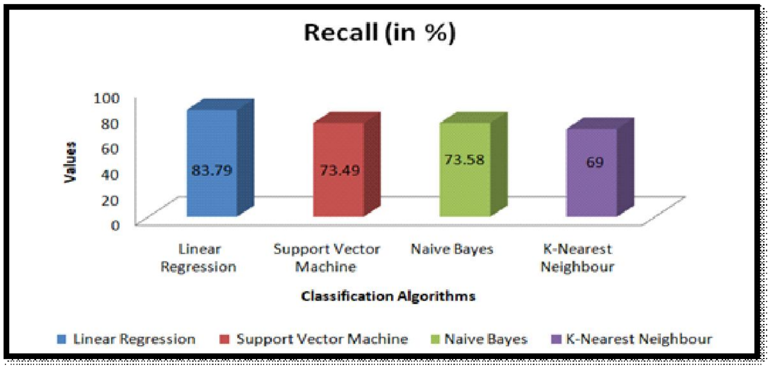


Fig. 8. Comparison of recall for proposed models

Precision :

Precision for a label is defined as the ratio of the number of true positives to the number of predicted positives. The Fig 9 presented below compares the precision values of all machine learning algorithms.

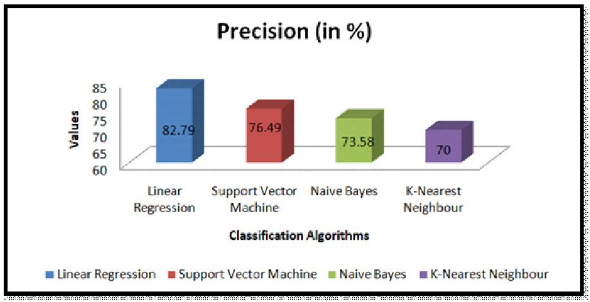


Fig. 9. Comparison of precision for proposed models

F- Measure :

F- Measure is the harmonic mean of precision and recall. The Fig 10 given below compares the F-measure values of all machine learning algorithms.

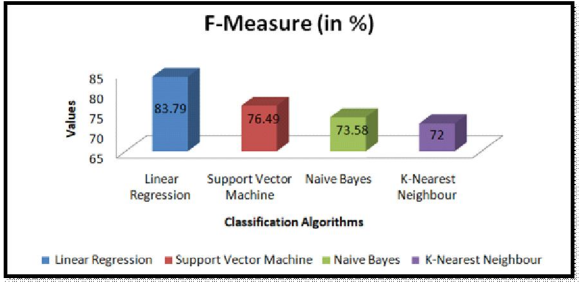


Fig. 10. Comparison of F-measure for proposed models

Specificity :

Specificity is the ratio of the true negatives to the total number of true negatives and false positives. The Fig 11 below compares the specificity values of all machine learning algorithms.

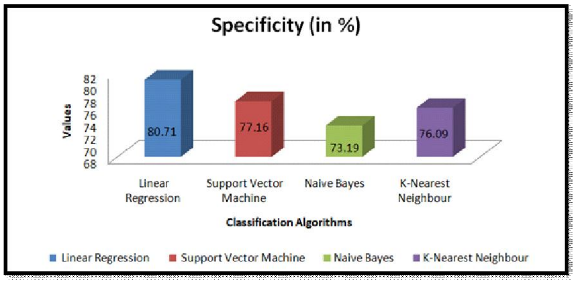


Fig. 11. Comparison of specificity for proposed models

Execution time for all machine learning algorithms :

The Fig. 12 below provides a comparison of the training execution time of the proposed algorithms.

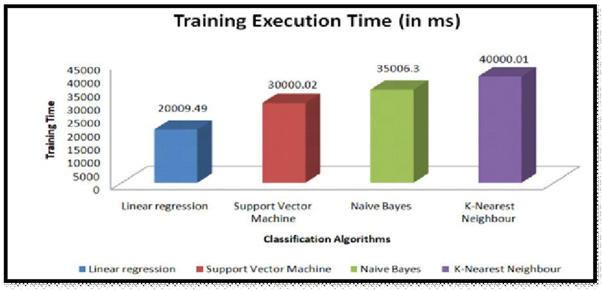


Fig. 12. Comparison of training execution time for proposed models

The Fig 13 below provides a comparison of the testing execution time of the proposed algorithms,

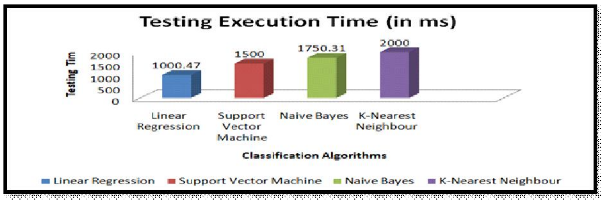


Fig. 13. Comparison of testing execution time for proposed models

The purpose of this paper is to understand the location-specific oilseed crop yield analysis, which is then handled using a machine learning algorithm. A dataset in .csv format was considered for this study. In this case, 80% of the data is used for training, while the remaining 20% is used for validation. Following successful training and testing, the model's accuracy was determined, indicating the model's performance in predicting yield. We have designed a graphical user interface that predicts the future yield of crops as shown in fig. 15. The summary of all oilseed crop production districts in Tamil Nadu is shown below in fig. 14.

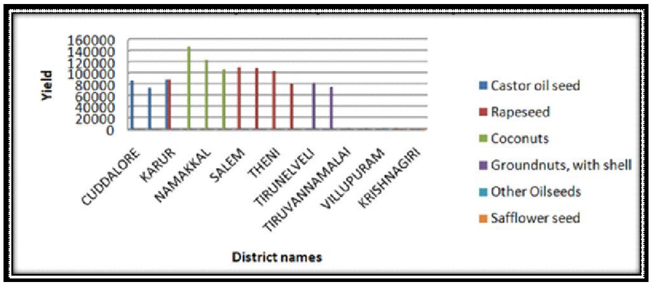


Fig. 14. Oilseed crop yield (in kgs/ha) for all districts in Tamil Nadu

According to the statistics observed from 1961 to 2019,

- Erode has relatively more castor oilseed production
- Salem has relatively more rapeseed production
- Pudukkottai has relatively more coconut production
- Tirupur has relatively high groundnut oilseed production
- Villupuram has relatively high other oilseed production
- Krishnagiri has relatively more Safflower oilseed production

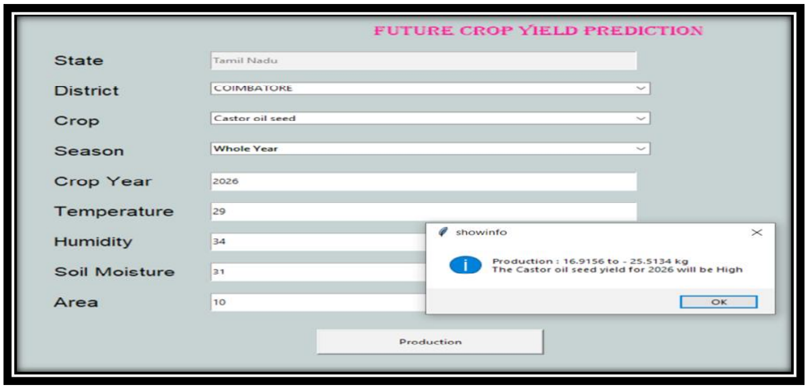
Prediction module :

GUI Creation: The study helps farmers decide which crop to cultivate in a particular region at a specific time and whether it will be lucrative or not during forecasting. It also provides specific details by stating whether the crop is profitable or not. It can also indicate high or low yields with ranges to guide farmers or end-users in making profitable decisions, allowing them to save time and money.

This prediction module enables users to share the district name, crop name, season, area, temperature, crop year, humidity, and soil moisture. After entering these attribute values, the user can click the production button to see the estimated yield ranges in kilograms, as well as the yield rate classification as high or low in the future. The result is obtained by considering the range of values based on the average of all prediction errors for each particular crop. The formula presented below is used for the calculation of the range of yield which is based on the prediction error of each crop,

$$\boxed{\text{Predicted range} = \text{Predicted value} \pm \text{Predicted error}} \quad (3)$$

In fig. 15 below, for the Coimbatore district, castor crop for the whole year, the 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 yield rate is estimated by taking the mean of each crop based on the records in the dataset. If the prediction value is below the mean score, then it is considered to be a low yield crop and if the prediction value is above the mean score, then it is considered to be a high yield crop.



FUTURE CROP YIELD PREDICTION

State	Tamil Nadu
District	COIMBATORE
Crop	Castor oil seed
Season	Whole Year
Crop Year	2026
Temperature	29
Humidity	34
Soil Moisture	31
Area	10

Production : 16.9156 to - 25.5134 kg
The Castor oil seed yield for 2026 will be High

Fig. 15. Recommendation systems for crop yield predictor

Data Visualization of GUI is also accomplished by plotting yield variables with various parameters. The practice of converting data into visual contexts, such as graphs or figures, aids humans in capturing and comprehending ideas. The primary goal of data visualization is depicted in Fig. 15. The prediction module makes it easier to identify patterns, correlations, and outliers in huge datasets. The graph above depicts the relationship between the district and the yield.

This paper addressed machine learning algorithms for crop yield forecasting using temperature, season, and location as inputs. Rainfall, temperature, and other factors such as season and location can also be used to forecast yield in a specific district. When all factors are considered, linear regression is indeed the best classifier. The use of a dataset with more parameters enhances the accuracy. Linear regression is found to be the best prediction algorithm when compared to other algorithms such as KNN, NB, and so on. Our database includes a significantly greater number of variables, resulting in more accurate predictions. This work's emergence will aid farmers in reducing risk and maximizing crop yields to improve their agricultural resources.

This will help farmers not only to determine the correct crop to cultivate in the coming season but will also help bridge the 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. The future work of our project aims to include regional languages such as Tamil, Telugu, Hindi, Kannada, Malayalam, and others in the graphical user interface, which will benefit farmers across the country.

Abbreviations :

Table-5. Abbreviations

S. No	Name	Abbreviation
1	KNN	K -Nearest Neighbour
2	NB	Naive Bayes
3	MLR	Multiple Linear Regression
4	ANN	Artificial Neural Network
5	SVR	Support Vector Regression
6	RF	Random Forest
7	MSE	Mean Squared Error
8	RMSE	Root Mean Squared Error
9	DNN	Deep Neural Network
10	MAE	Mean Absolute Error
11	XGB	Extreme Gradient Booster
12	PCA	Principal Component Analysis
13	LR / LNR	Linear Regression
14	TMOP	Technology Mission on Oilseeds and Pulses
15	ML	Machine learning
16	MLP	Multi Layer Perceptron
17	GUI	Graphical User Interface
18	FFS	Forward Feature Selection

References :

1. Ansarifar J., L. Wang and S.V. Archontoulis (2021) *Scientific Reports*, *0123456789*, pp.1–14.
2. Antony B. (2021). *Environmental Research*, *202* (June), 111624.
3. Bali N. and A. Singla (2021) Dee Learning Based Wheat Crop Yield Prediction Model in Punjab Region of North India Deep Learning Based Wheat Crop Yield Prediction Model in Punjab Region of

North India, *Appl. Artif. Intell.*, pp.1–25.

4. Bhojani S. H. and N. Bhatt (2020) *Neural Computing and Applications., 32:* 13941–13951.
5. Bhosale M. S. V. (2018) Crop Yield Prediction Using Data Analytics and Hybrid Approach. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–5.
6. Chipindu W. M. L, and I. N. S. Mkuhlani (2020) *SN Applied Sciences*, *2*(5): 1–14.
7. Dang C., Y. Liu, H. Yue, J. Qian and R . Zhu (2020) *Can. J. Remote Sens.*, pp. 1–20.
8. Elavarasan D. and P. M. D. Raj (2021) *Neural Computing and Applications*, *33:* 13205–13224.
9. Girish L. (2018) Crop Yield and Rainfall Prediction in Tumakuru District using Machine Learning, Nation Conference on Technology on rural development, pp. 61–65.
10. Gopal P.S.M. and R. Bhargavi (2019) *Appl. Artif. Intell., 33*(7): 621–642.
11. Jebakumar S. I. R. (2022) *Wireless Personal Communications*, *0123456789. 126:* 1935–1964.
12. Kale S. S. (2019) *A Machine Learning Approach to Predict Crop Yield and Success Rate*. pp. 1–5.
13. Kamath P., P. Patil, S. Shrilatha and S. Sowmya (2021) *Global Transitions Proceedings*, *2*(2): 402–407.
14. Khaki S. and L. Wang (2019) *Crop Yield Prediction Using Deep Neural Networks 10*(May), pp. 1–10.
15. Mahajan J. (2021) *International Journal of Information Technology. 13:* pp. 1441–1448 .
16. Majumdar, J., S. Naraseeyappa, and S. Ankalaki, (2017). *Journal of Big Data. 20:* 1-15.
17. Nishant P. S., P. S. Venkat, B. L. Avinash and B. Jabber (2020) *Crop Yield Prediction based on Indian Agriculture using Machine Learning*, pp. 5–8.
18. Palanivel K. (2019) *International Journal of Computer Engineering and Technology, 10*(03): 110–118.
19. S, M.G.P. and R. Bhargavi (2019) *Applied Artificial Intelligence*, *33*(7): 621–642.
20. Sana A. K. K, A, B. A. Bhat, S. Kumar, and N. Bhat (2021) An Efficient Algorithm for Predicting Crop using Historical Data and Pattern Matching Technique, *Glob. Transitions Proc.*, pp. 0–12.
21. Shahhosseini M., G. Hu, I. Huber, and S.V. Archontoulis (2021) *Sci. Rep.*, no. 0123456789, pp. 1–15.
22. Shastry A., H.A. Sanjay and E. Bhanusree (2017) *Prediction of Crop Yield Using Regression Techniques*, *12*(2): 96–102.
23. Oilseeds: Directorate of Oilseeds Development (Ministry of Agriculture and Farmers Welfare, Government of India (https://oilseeds.dac.gov.in/)