Predictive modeling of crop yields: a comparative analysis of regression techniques for agricultural yield prediction

Priti Prakash Jorvekar[1*], Sharmila Kishor Wagh[2], Jayashree Rajesh Prasad[3]

(1. Research scholar, Smt. Kashibai Navale College of Engineering, SPPU university, Pune, India;
2. Professor, MES College of Engineering, Pune, SPPU University, Pune, India;
3. Professor, School of Engineering, MIT Art, Design and Technology University, Pune, India)

Abstract: Crop yield prediction plays a key role in modern agriculture, it enables farmers to make decisions about resource distribution, crop production management, and marketing business strategies. Regression models are extensively used for crop yield prediction. The performance of different regression techniques may vary depending on various factors such as the dataset, features, and modeling assumptions. In this paper, a comparative study was conducted to evaluate and compare the performance of different regression models for agriculture crop yield prediction. Collected a comprehensive dataset encompassing historical crop yield data, weather parameters and pesticides data features from various agricultural regions, then applied and compared various regression models, including Linear Regression (LR), K Nearest neighbor Regression (KNR), Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Gradient Boosting Regression (GBR),Linear Model Lasso Regression, Elasticnet Regression, Ridge Regression to predict crop yields for various crops. This study involved evaluating the performance of these regression models based on several performance metrics, including R ^{2} score, Root Mean Squared Error(RMSE), Mean Squared Error(MSE), Mean Absolute Error(MAE), Median Absolute Error(Median AE), Explain variance score and computing time. The results of our study provide insights into the comparative performance of different regression models for crop yield prediction in agriculture. Determined that the performance of the regression models varies crop type, area, and dataset used. Overall, the random forest regression model demonstrated the best performance in terms of R^2, followed by K Nearest neighbor with hyper parameter tuning and decision tree regression. However, the choice of the most suitable regression model may also depend on other factors such as the interpretability and computational efficiency requirements of the application. Our research findings contribute to the existing literature on crop yield prediction in agriculture and afford treasured information for farmers, policymakers, and researchers to make conversant conclusions about the selection of appropriate regression models for crop yield prediction in their specific contexts. Further research could explore the combination of different regression models or the integration of other ML techniques to better the $R^2$and robustness of crop yield prediction models in agriculture.

Keywords: K Nearest neighbor Regression, SVR, Linear Regression Decision Tree Regression, Random Forest Regression, Elasticnet, Ridge Regression

Citation: P. P. Jorvekar, S. K. Wagh, and J. R. Prasad. 2024. Predictive modeling of crop yields: a comparative analysis of regression techniques for agricultural yield prediction. Agricultural Engineering International: CIGR Journal, 26 (2):125-140.

1 Introduction

Crop yield prediction is a fundamental aspect of modern precision farming as it provides valuable information for farmers to make decisions about crop management practices, resource allocation, distribution and market business strategies. Accurate crop yield prediction can help optimize agricultural production, mitigate risks, and ensure sustainable farming practices.

Traditionally, crop yield prediction has been based

Received date: 2023-11-23**Accepted date:**2024-01-14

[*]**Corresponding author: Priti Prakash Jorvekar,** Research scholar, Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, India , 411041. Email: pritiprmjorvekar@gmail.com.

on empirical knowledge, experience, and historical data. However, with advancements in technology and the availability of large datasets, machine learning techniques, specifically regression models, have extended popularity for crop yield prediction. Regression models allow for the identification of patterns and relationships between various factors such as weather parameters, pesticides crop data, and historical yield data to make predictions about future crop yields.

There are diverse regression models that can be used for crop yield prediction, each with its own potencies and limitations. LR, SVR, DTR and RFR are among the commonly used regression techniques in agricultural applications. But the performance of these models may differ subject to dataset, features, and modeling assumptions.

Given diversity of regression models and their potential implications for agricultural decision-making, it is significant to do comparative study to assess and compare the performance of different regression models for crop yield prediction in agriculture. Such a study can provide insights into the relative strong point and limitations of different models, identify the most suitable model for a particular context, and contribute to the development of effective crop yield prediction methodologies.

In this research, a comparative study of regression models for yield prediction in agriculture domain was presented. Collected a comprehensive dataset comprising historical crop yield data, weather parameters and other relevant features from multiple agricultural regions. Applied and evaluated various regression models, including LR, SVR, DTR, RFR to predict crop yields for multiple crops and compared the performance of these models based on several performance metrics and conducted sensitivity analysis to assess their robustness and reliability under different scenarios.

The findings of this study have implications for farmers, policymakers, and researchers involved in agricultural decision-making. By comparing the performance of different regression models, the main aim to provide insights into the suitability and effectiveness of these models for crop yield prediction in agriculture. This research contributes to the existing literature on crop yield prediction and provides valuable information for improving agricultural decision-making and optimizing crop production practices.

2 Related work

Crop yield prediction is a decisive aspect of modern precision agriculture that has significant implications for farmers, policymakers, and other stakeholders. Accurate crop yield prediction can help enhance resource distribution, improve crop management routines, mitigate risks, and enhance market strategies. It can enable farmers to make decisions about planting schedules, irrigation, fertilization and harvesting, resulting in higher yields, reduced costs, and increased profitability.

Traditional methods of crop yield prediction, based on empirical knowledge and historical data, may not always provide accurate and reliable results due to the complex interactions among various factors affecting crop growth and yield. However, with advancements in technology and the availability of large datasets, machine learning techniques, particularly regression models, have emerged as promising tools for crop yield prediction.

The motivation of this research paper is to conduct a comparative analysis and study of regression models for crop yield prediction in agriculture to fill the existing research gap and provide evidence-based insights for farmers, policymakers, and researchers. By evaluating and comparing the performance of different regression models, the aim to contribute to the advancement of crop yield prediction methodologies and provide practical guidance for selecting the most suitable model for a particular agricultural context. Our research findings can help improve agricultural decision-making, optimize crop production practices, and ultimately contribute to sustainable and profitable agriculture.

3 Literature survey

Table1 shows that dataset, important features, algorithm used, result evaluation parameters of various studies in crop yield prediction using regression.

Table1 Literature survey

No.	Algorithm used	Dataset	Parameters	Evaluation Parameters	Source of the reference
1	LR	FAO	Year, annual rainfall, food price index, production, yield, area under irrigation	R^2＝0.7	Sellam and Poxovammal, 2016
2	MLR,ANN,SVR,RFR,KNR, hybrid MLR-ANN	Thirty years of paddy crop Department of Economics and Statistics, Government of Tamil Nadu	Temperature (minimum, maximum and average), canal length, tube wells, production, rain fall, solar radiation, nitrogen, phosphorus and potassium	RMSE=0.051, MAE=0.041, R^2=0.99	Maya and Bhargavi,2019
3	Quadratic, interactions, linearpolynomial	wheat, maize and cotton yields dataset	Biomass, diywater content, grain protein, grain size,ESW,soil water	R^2=0.6, RMSE=0.3	Aditya et al., 2017
4	KNR	Rice wheat dataset	Crop, district, season, yield	R^2=0.80, RMSE=0.5	Ramesh et al., 2019
5	Ridge, lasso, ER	Indian Government Repository	Season, production, area, year, state, district	RMSE=4	Potnuruet al., 2020
6	KNR,DNN,ANN	FAO	Soil and metrological parameters	R^2=8, RMSE=5.1	D,Jayanarayana and Kumar, 2021
7	LR	Sanliurfa, Turkey	Year, satellite data	R^2=0.87	Yunus and Polat,2023
8	LR,EN,KNR,SVR	Prince Edward Island and three fields of New Brunswick USA	Soil and crop properties	RMSE=5.97	Farhat et al., 2020
9	RFR,SVR	Missouri, USA	RGB imaginary after growth ofplant,temperature,precipitation	R^2=0.720, RMSE=15.9	Maimaitijiang et al., 2020
10	SVR,RFR,MPR	USDA for the state of Iowa	Yield,temperature, precipitation data	RMSE=5.48, MAE=3.57, MAE=1.58, R^2=0.968	Ayush et al., 2018
11	RFR,PR,SVR	Potatoes and Maize dataset for Musanze, Rwanda	crop production and metrological parameters	RMSE=510.8, MAE=129.9, R^2=0.875	Martin Kuradusenge, et al., 2023
12	RFR	Study area southeastern part of Germany	Satellite-based crop biomass, solar radiation,and temperature	RMSE=8, MAE=1.6, R^2=14.3	Maninder et al., 2023
13	LR	ICRISA rice, wheat, and pearl millet	Climate and crop production data	RMSE=0.2	Balsher et al., 2023
14	MLR,DTR,GBR,EN, lasso	district-level data in Maharashtra	crop type, season, area of the field, Temperature, Rain fall, humidity, soil type.	RMSE=5.2, MAE=0.35, R^2=8.95	Iniyan et al., 2023
15	SKN	Field in Bedfordshire, UK	soil data, satellite imagery crop growth data	RMSE=0.026	Pantazi, 2016
16	SVR	Field in Australia	Climatic data, Satellite image data	R^2=0.7	Elisa et al., 2020
17	SVR, RFR, DTR	paddy crop, 5 years data, Tamilnadu, India	Yield data, Climatic data	RMSE=0.41, MAE=0.58, R^2=0.38	Dhivya et al., 2018
18	NB, MLP, SARIMA	cocoa crop in southwest Nigeria	Yield data, Climatic data	RMSE=143.13, MAE=47.66, MAPE=15.51, MSE=20.487	Sunday et al., 2023
19	RFR, XGBoost,KNR,LR	southern district of India	Area, Temperature, rainfall, season	MAE=0.78	Aruvanshet al.,2019
20	GBR,RFR, DTR	Study area of India	Yield data, Climatic data	MAE=0.57	Kasi and Kumar,2023
21	SVR, KNN, RFR, lasso	South China, 30 years data	Productiondata, yield data, N,P,K	RMSE=0.051, MAE=0.041, R^2=0.99	Mamunuret al.,2021
22	RFR, DTR, polynomial regression	Field in India	Rainfall,temperature, humidity, soil pH, soil type	RMSE=0.56	Shruthi and Sangeeta,2020

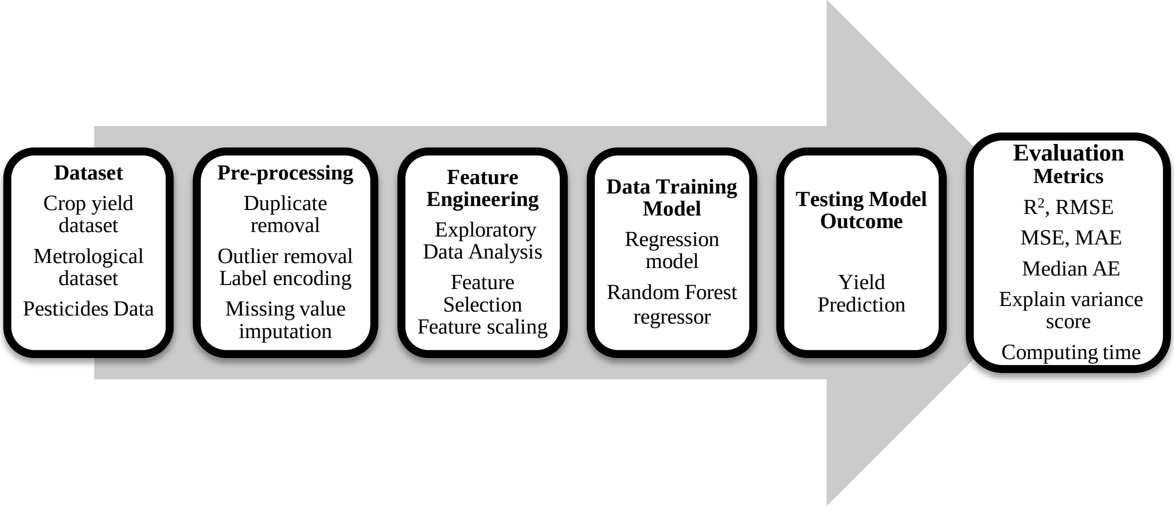


Figure 1 Architectural diagram

4 Implementation

4.1 Data collection

Data collection involves collecting relevant data for crop yield prediction. Crop yield prediction dataset was taken from various sources for 23 years with 101 different countries. Details of dataset parameters and source from where it has taken is as follows:

4.1.1 Crop yield dataset

This dataset downloaded from Food and Agriculture Organization (FAO) website, It includes item, domain code, year code, domain, area code, year, element code, element, item code, unit, value(hgha^{-1}yield) from year 1961 to 2016.

4.1.2 Metrological dataset

Crop yield is majorly depending on metrological data (climatical data) that is Rainfall and temperature data. Rain fall data collected from world data bank which includes area, year, average rain fall per year, average temperature.

4.1.3 Pesticides dataset

Pesticides dataset downloaded from FAO website which includes domain, area, element, item, year, value pesticides tones.

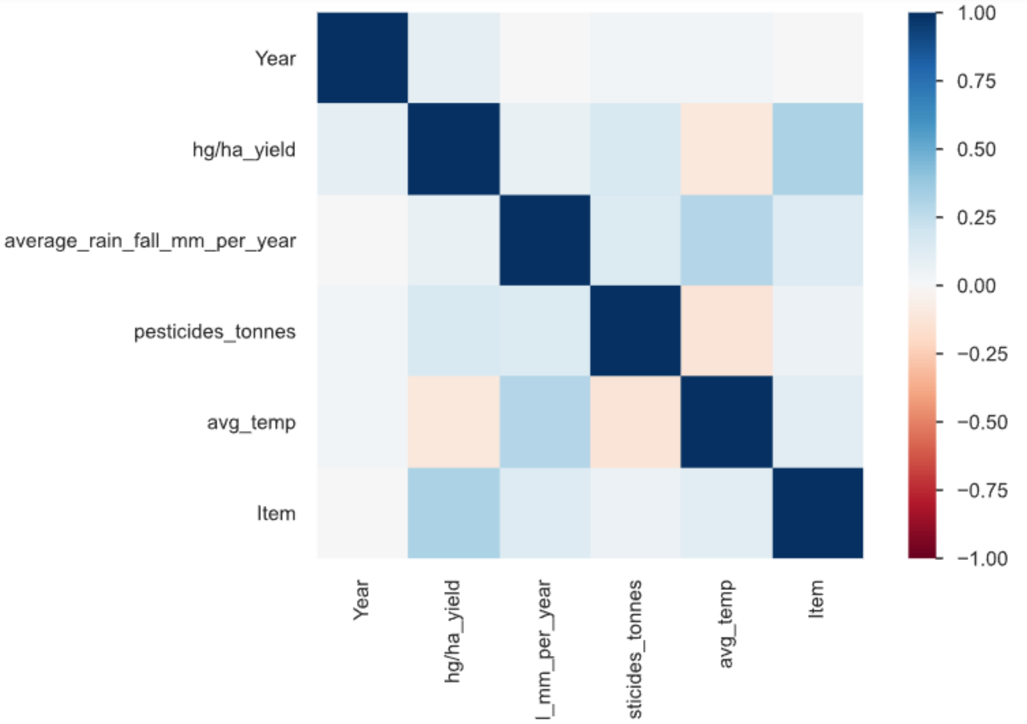


Figure 2 Heatmap

4.2 Data preprocessing

Data pre-processing is the techniques to convert dataset to clean and process dataset , for this study gathered data from various sources so it is collected in raw format , it needs some preprocessing which involves: data merging based on primary key area , missing data imputation, encoding for categorical features as regression models cannot operate on label data directly, outlier detection and removal , scale down the features to one certain range do that can remove the unwanted noise from the dataset before sending it to training the regression model.

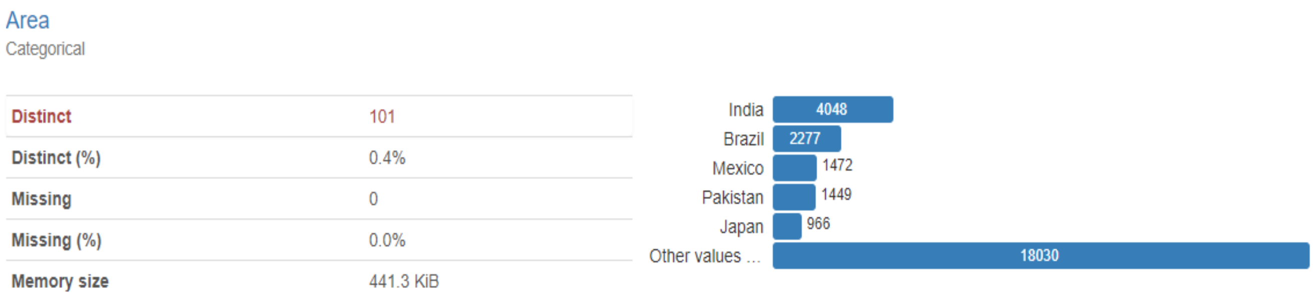
4.3 Feature selection and exploratory data analysis

Based on feature importance's feature selection techniques taken most relevant features for this study which includes year, area, hgha^{-1}yield, Average rain fall per year, pesticides tones, average temperature for 23 years data for 101 different countries.

Dataset includes 7 features, Detail explanation of features used in study as follows:

4.3.1 Area

Area is categorical feature for country names where 101 distinct counties names are present in dataset.



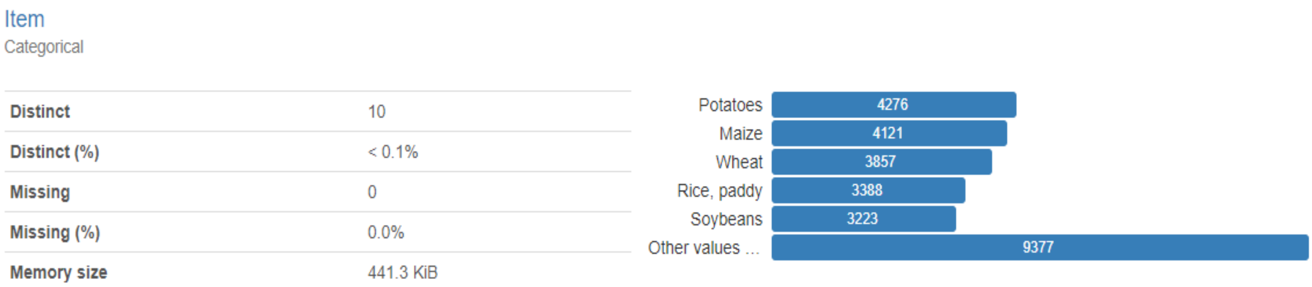
Area
Categorical

Distinct	101
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	441.3 KiB

Figure 3 EDA for area feature

4.3.2 Item

Item is categorical feature where 10 different crop names are present such as maize, potatoes, rice, paddy, sorghum, soybeans, wheat, cassava, sweet potatoes, yams.



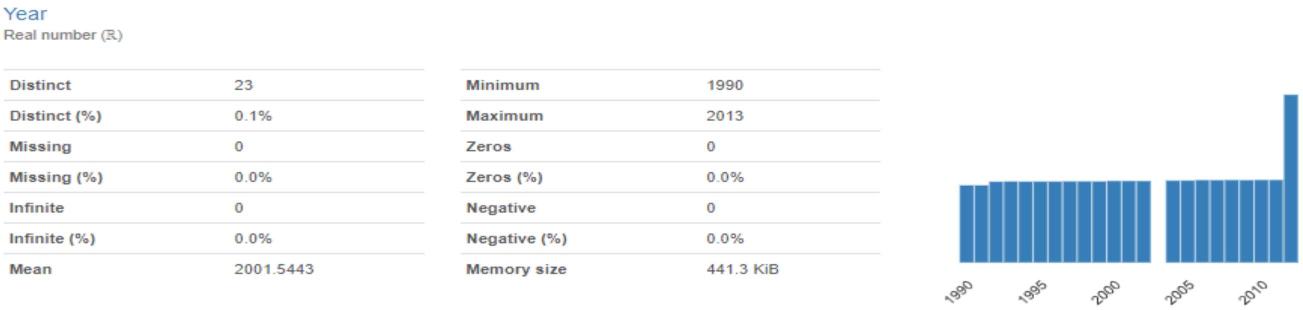
Item
Categorical

Distinct	10
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	441.3 KiB

Figure 4 EDA for item feature

4.3.3 Year

This study uses 23 years of data starting from 1990 to 2023.



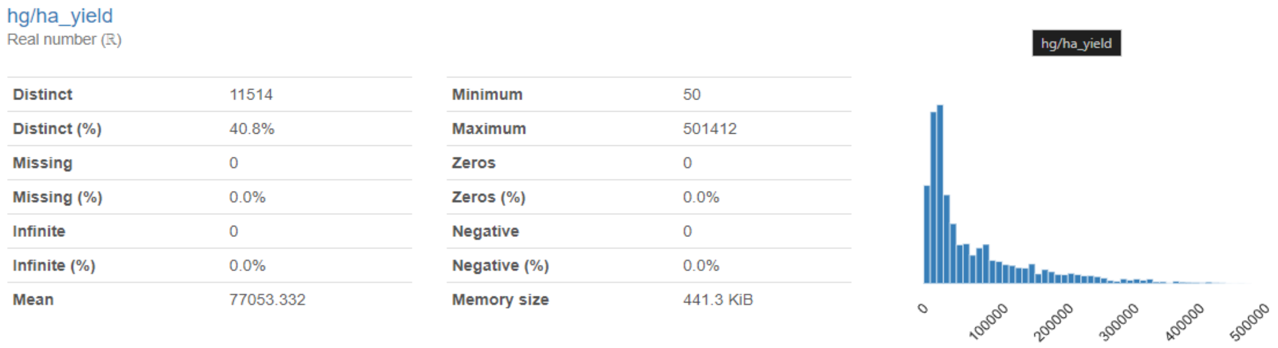
Year
Real number (ℝ)

Distinct	23	Minimum	1990
Distinct (%)	0.1%	Maximum	2013
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2001.5443	Memory size	441.3 KiB

Figure5EDA for year feature

4.3.4 Yield

Yield is the dependent or the target feature of the study, unit used to measure the yield is hg ha^{-1} yield.



hg/ha_yield
Real number (ℝ)

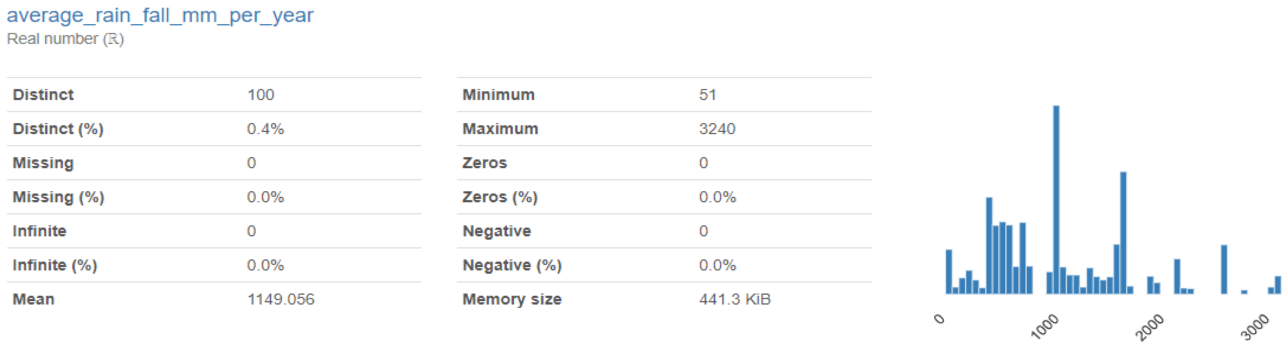
Distinct	11514
Distinct (%)	40.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	77053.332

Minimum	50
Maximum	501412
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	441.3 KiB

Figure 6 EDA for yield feature

4.3.5 Average rain fall per year

Average rain fall is the important feature of metrology, and it is measured in mm per year. For the current study minimum average rain fall is 51 mm per year and maximum rainfall is 3240 mm per year.



average_rain_fall_mm_per_year
Real number (ℝ)

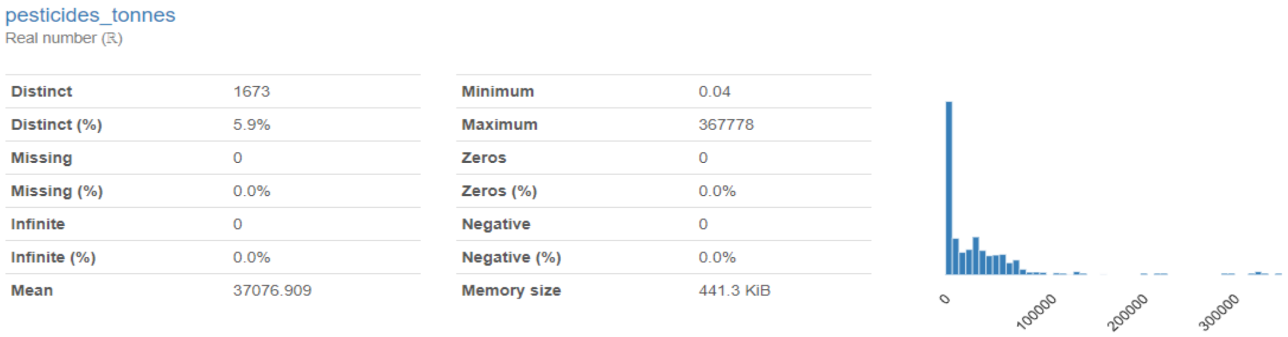
Distinct	100
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1149.056

Minimum	51
Maximum	3240
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	441.3 KiB

Figure 7 EDA for average rain fall per year feature

4.3.6 Pesticides

In this study pesticides information used, unit to measure the pesticides per year is tones per year.

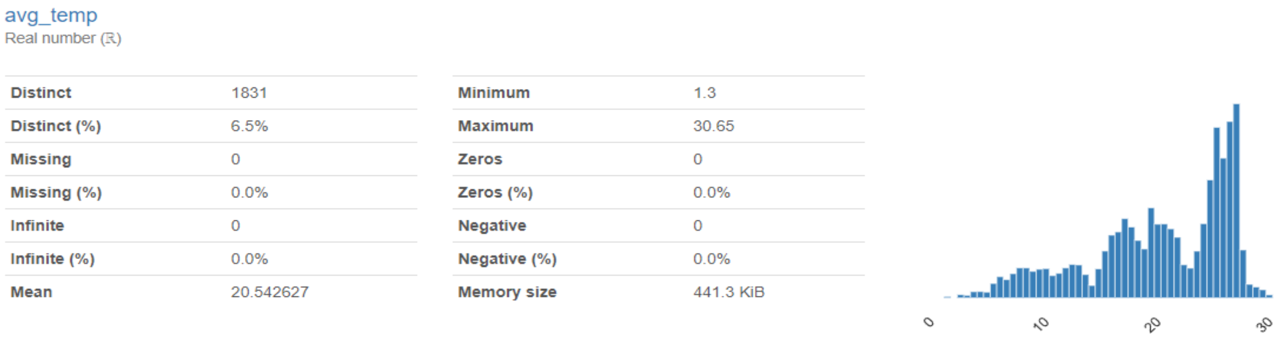


pesticides_tonnes
Real number (ℝ)

Distinct	1673
Distinct (%)	5.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	37076.909

Minimum	0.04
Maximum	367778
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	441.3 KiB

Figure 8 EDA for pesticides feature



avg_temp
Real number (ℝ)

Distinct	1831
Distinct (%)	6.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	20.542627

Minimum	1.3
Maximum	30.65
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	441.3 KiB

Figure 9 EDA for average temperature feature

4.3.7Average temperature

Average temperature is the important feature of metrology, and it is measured in Celsius. For the current study minimum temperature is 1.3 degree Celsius per year and maximum temperature is30.65 degree Celsius per year.

4.4Feature scaling

In EDA author observed in the dataset, it contains highly instable magnitudes of data, its ranges and units, The features with extreme range will weight a lot additional in the distance calculations than features of low magnitude, to defeat this need to take all the features to similar level of magnitude So for that in this study used Standard Scaler feature scaling techniques to scale down the values in specific range. It is a popular preprocessing technique used in ML to scale the features of a dataset. It transforms the features of a dataset to have zero mean and unit variance, which can be useful for certain, algorithms that are sensitive to the scale of the features.

4.5Algorithm used

Crop yield prediction has a target feature yield in hector which is continues in nature so here need to apply various regression type algorithms. This section involves training and evaluating a regression model using pre-processed and engineered data. Commonly used regression techniques for crop yield prediction may include LR, SVR, DTR, RFR among others. In this study, trained the model based on 13 different regression model which includes some with default parameters and some with hyper parameter tuning. Details explanation of each model as follows:

4.5.1 Linear regression

Linear regression (Sellam and Poxovammal, 2016) is used statistical technique for modeling the association between a dependent variable crop yield and 7 different independent variables year, area, average rain fall per year, pesticides tones, average temperature in a linear manner.

Considering a regression problem, the target variable is represented by *Y* and the predictor variables (features) are represented by X_1, X_2, ..., X_n.

The linear regression model is the relationship between the independent variables and the target variable is linear and can be represented by a linear grouping of the predictor variables, weighted by regression coefficients:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n \times X_n + \varepsilon \qquad (1)$$

where:

Y is the variable you want to predict.

X_1, X_2, ..., X_n are the denote the variables used to make predictions.

β_0, β_1, β_2, ..., β_n are the regression coefficients that determine the contribution of each predictor variable to the predicted value of Y.

ε is the represents the error term, which is the variability in Y that is not explained by the predictor variables.

4.5.2 K-neighbors regression

K-neighbors regression (KNR) (Maya and Bhargavi, 2019; Ramesh et al.,2019;D.Jayanarayanaand Kumar, 2021) is a supervised ML algorithm used for regression tasks. It is a type of KNN algorithm. KNR is used to predict continuous target variables based on the values of their KNN in the training dataset. In KNR, the algorithm works by finding the KNN of a given input data point in the training dataset built on a distance metric, such as Euclidean distance or Manhattan distance. The predicted value for the target variable of the input data point is then calculated as the average (or weighted average) of the target variable values of its *k*-nearest neighbors. The value of *k*, known as the "*k*" hyperparameter, determines the number of neighbors used for prediction and is typically set by the user.

4.5.3 K-neighbors regression with hyper parameter tuning

The choice of k, the distance metric, and other hyper parameters in KNR can significantly impact its performance. Therefore, hyper parameter tuning is an important step in optimizing the performance of the model. This study uses gridsearchCV hyper parameter tuning to get optimized values for following mentioned parameters:

K- neighbors regressor (n_neighbors=9, *p*=1, weights='distance')

Where, p parameter determines the order of the Minkowski distance. When p = 1, it's the first-order (Manhattan) distance, it indicates the Manhattan distance (also known as L1 norm), which is the sum of the absolute differences of the coordinates.

4.5.4 SVR

Support vector regression, which is a ML algorithm used for regression tasks. SVR is a type of SVM that is used to predict continuous numeric values. SVR is particularly useful for solving problems where there are complex relationships between input features and target outputs, and when the data may not be linearly separable.

In SVR, the goal is to find a function that best fits the data while also minimizing the prediction errors. The algorithm finds a hyper plane that has the largest margin from the data points, while also allowing for a certain tolerance for errors. This margin is determined by a parameter called the epsilon (ε) which is specified by the user. The points that fall within the margin or violate the margin (i.e., the support vectors) are used to construct the regression function.

Given a training dataset of n samples, each represented by an input vector x_i of size m and a corresponding target value y_i, SVR seeks to find a function *f(x)* that approximates the mapping from input space to output space. The SVR model can be defined as:

$$f(x)= w^T x + b \tag{2}$$

where:

w is a weight vector orthogonal to the hyper plane;

b is the bias term.

4.5.5 SVR with hyper parameter tuning

Hyper parameter tuning is an important step in optimizing the performance of the model here in study selected values of C, gamma and kernel manually to improve the results. Hyper parameters used to improve the performance of SVR are as follows: C=100, gamma=1, kernel='linear'.

4.5.6 Decision tree regression with hyper parameter tuning

Decision tree regression in scikit-learn provides several hyper parameters that can be tuned to optimize its performance. This study uses GridSearchCV hyper parameter tuning to get optimized values for following mentioned parameters.

Decisiontreeregressor(criterion='mse', max_depth=7, max_features='auto', min_samples_leaf=7, min_samples_split=0.1)

4.5.7 Decision tree regression

The DTR is a ML algorithm used for regression jobs, it involves predicting a continuous target variable based on independent features. A DTR operates by recursively separating the data into splits based on the values of input independent variables, and then predicting average target variable values in each subset. The goal is to create a tree-like structure where the data is split into homogeneous subsets with respect to the target variable.

$$f(x) = \Sigma\, w_i \times I(x \in R_i) \tag{3}$$

where:

f(x) is Target value for the input vector x.

R_i represents a region in the input space defined by a specific path through the decision tree.

w_i is the predicted value associated with region R_i.

I is a function which returns 1 if the state inside the parentheses is true otherwise it has to return 0.

The decision tree model splits the input space into regions based on binary decisions made at each internal node of the tree. These decisions are based on threshold values for specific features. As we traverse down the tree, each internal node compares the value of a particular feature in the input vector x to a threshold value and determines the appropriate branch to follow based on the result of the comparison. Eventually, to reach a leaf node that corresponds to a specific region in the input space, and the predicted value associated with that leaf node is used as the final prediction for the input vector x.

4.5.8 Random forest regression

RFR is ML algorithm that is under the category of ensemble methods, specifically the RF algorithm, used for regression tasks. It is a popular algorithm for solving regression problems in data science and machine learning due to its ability to handle complex data and mitigate over fitting.

The main idea behind Random Forest Regression is to combine the estimates of multiple decision tree models to obtain a more accurate and robust prediction. During prediction, the Random Forest Regression algorithm aggregates the predictions of all the individual trees in the forest to obtain the final prediction. This aggregation is done by taking the mean (for regression tasks) of the individual tree predictions.

$$f(x) = \Sigma f_k(x) \quad (4)$$

where:

f(x) is Target value for the input vector *x*.

$f_k(x)$ is Target value from the k-th DT in the RF.

The prediction from each decision tree, $f_k(x)$, is obtained using the decision tree Equation3 described earlier. The decision tree splits the input space into areas based on the values of different features, and each region corresponds to a leaf node in the tree. The predicted value associated with a leaf node is used as the prediction for the input vector x. In the random forest, the predictions from all the decision trees are averaged to obtain the final prediction.

4.5.9 Random forest regression with hyper parameter tuning

This study uses Grid Search CV hyper parameter tuning to get optimized values for following mentioned parameters.

Random Forest Regressor (max_depth=8, min_samples_leaf=4, min_samples_split=0.01, n_estimators=1000)

4.5.10 Gradient boosting regression

GBR works by merging multiple weak learners, typically decision trees, into a strong learner. The algorithm builds an initial model and then iteratively adds subsequent models to correct the errors of the previous models. The errors are minimized by adjusting the target values of the training examples based on their residuals, which are the differences between the predicted values and the actual values. The gradient descent optimization technique is used to find the optimal values for the model parameters during the training process.

$$f(x) = \Sigma \gamma_k \times h_k(x) \quad (5)$$

where:

f(x) is target value for the input vector x.

γ_k is rate or step size associated with the k-th weak learner.

$h_k(x)$ is the predicted value from the k-th weak learner.

In this, the algorithm builds an ensemble of weak learners sequentially. Each weak learner is trained to correct the errors made by the previous weak learners. At each iteration, the algorithm fits a weak learner, such as a DT, to the negative gradient of the loss function with respect to the current prediction. The prediction from each weak learner, $h_k(x)$, is combined with a weight γ_k, which determines the contribution of that weak learner to the final prediction.

4.5.11 Linear model lasso regression

LR (Potnuru et al., 2020)is a variation of LR that introduces a penalty term called L1 regularization or Lasso regularization. It adds a penalty term to the linear regression objective function, which is the absolute value of the coefficients multiplied by a hyper parameter (alpha). Lasso Regression can shrink the less important features to exactly zero, which makes it useful for feature selection and can help to mitigate over fitting.

The Lasso Regression, which is a linear model with L1 regularization, is used for feature selection and regularization in linear regression tasks. It adds a penalty term to the ordinary least squares (OLS) cost function, encouraging the model to select a sparse set of features by promoting the coefficients of irrelevant features to be zero.

The equation for the Lasso Regression model can be expressed as follows:

$$y = w_0 + w_1x_1 + w_2x_2 + \ldots + w_px_p \quad (6)$$

where:

y represents the target variable;

$w_0, w_1, w_2, \ldots, w_p$ are the coefficients (weights) associated with each feature;

$x_1, x_2, \ldots, x_p$ represent the feature values.

The lasso regression model aims to minimize the following cost function:

$$cost = (1/2n)\ \Sigma(y_i - (w_0 + w_1x_{1i} + w_2x_{2i} + ... + w_px_{pi}))^2 + \alpha\ \Sigma|w| \quad (7)$$

where:

n is the number of samples in the dataset.

α is the regularization parameter that controls the strength of the regularization. It determines the trade-off between the fit to the training data and the magnitude of the coefficients. A higher *α* leads to more regularization and encourages more coefficients to become zero.

The L1 regularization term, *Σ*|*w*|, encourages sparsity by promoting feature selection and shrinking the coefficients of irrelevant features towards zero.

4.5.12 Elasticnet regression

ElasticNet (Potnuru et al., 2020)Regression is a machine learning algorithm that combines features of both Lasso Regression and Ridge Regression. It is a linear regression algorithm with a penalty term that is a linear combination of L1 (Lasso) and L2 (Ridge) regularization. ElasticNet is useful for regression tasks when there are multiple features with potentially high correlation, and it aims to mitigate the limitations of both Lasso and Ridge regressions. It is used for feature selection and regularization in linear regression tasks, offering a balance between the Lasso and Ridge regression methods.

$$y = w_0 + w_1x_1 + w_2x_2 + ... + w_px_p \quad (8)$$

where:

y represents the target variable.

w_0, w_1, w_2, ..., w_p are the coefficients weights associated with each feature.

x_1, x_2, ..., x_p represent the feature values.

The ElasticNet Regression model aims to minimize the following cost function:

$$cost = (1/2n)\ \Sigma(y_i - (w_0 + w_1x_{1i} + w_2x_{2i} + ... + w_px_{pi}))^2 + \alpha_1\ \Sigma|w| + (1/2)\alpha_2\ \Sigma w^2 \quad (9)$$

where:

n is the number of samples in the dataset.

α_1 and α_2 are the regularization parameters that control the strength of the L1 and L2 regularization terms, respectively. They determine the trade-off between the fit to the training data, the sparsity of the model, and the magnitude of the coefficients.

The L2 regularization term, Σw^2, encourages small but non-zero coefficients, improving the stability and robustness of the model.

4.5.13 Ridge regression

Ridge *R* (Potnuru et al., 2020) is another variation of LR that introduces a penalty term called L2 regularization or Ridge regularization. It adds a penalty term to the LR objective function, which is the squared sum of the coefficients multiplied by a hyper parameter. RR shrink the coefficients towards zero, but it does not force them to exactly zero, which makes it useful for reducing multi co linearity in the data.

$$y = w_0 + w_1x_1 + w_2x_2 + ... + w_px_p \quad (10)$$

where:

y is the target variable;

w_0, w_1, w_2, ..., w_p are the coefficients weights associated with each feature.

x_1, x_2, ..., x_p are feature values.

The Ridge R model aims to minimize the following cost function:

$$cost = (1/2n)\ \Sigma(y_i - (w_0 + w_1x_{1i} + w_2x_{2i} + ... + w_px_{pi}))^2 + \alpha\ \Sigma w^2 \quad (11)$$

where:

n is number of samples in the dataset.

α is the regularization parameter that controls the strength of the regularization. It determines the trade-off between the fit to the training data and the magnitude of the coefficients. A higher α leads to more regularization, shrinking the coefficients towards zero.

The L2 regularization term, Σw^2, is the sum of the squared values of the coefficients. It penalizes large coefficients and encourages the model to have smaller and more balanced coefficients across all features.

5 Evaluation metrics

Model Evaluation: This component involves evaluating the performance of the trained regression model using evaluation metrics, such as R^2score, root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE), median absolute error (MedianAE), explain variance score (VS) and computing time. Model evaluation helps assess the accuracy and reliability of the model’s predictions and may involve fine-tuning or optimizing the model parameters for better performance.

Table 2Regression evaluation metrics

Sr.no	Model	R2	RMSE	MSE	MAE	MedianAE	VS	Computing Time
0	LR	0.754	42709.75	1824122475	29683.317	20143.567	0.754	0.12
1	KNR	0.956	18115.16	328158823	8298.9	2861.2	0.956	5.0
2	KNRT	0.961	16923.08	286390452.9	7480.933	2296.571	0.961	89.10
3	SVR	-0.203	94445.57	8919965471	57572.467	26159.981	0.002	117.71
4	SVRT	0.691	47874.36	2291954712	26347.516	10458.397	0.7	90.11
5	DTRT	0.699	47256.14	2233143114	29267.976	14374.413	0.699	0.18
6	DTR	0.96	17286.36	298818201.7	6123.364	335	0.96	0.63
7	RFR	0.973	14031.78	196890834.1	5705.45	1447.72	0.973	28.76
8	RFRT	0.881	29688.92	881431780.4	18814.116	11588.915	0.881	126.33
9	GBR	0.866	31491.85	991736434.5	20296.47	11977.044	0.866	7.33
10	LASSO	0.754	42713.92	1824478960	29660.381	20077.233	0.754	11.44
11	ER	0.251	74516.14	5552654783	54543.087	43842.369	0.251	11.58
12	RIDGE	0.754	42725.76	1825490592	29633.088	20020.949	0.754	11.67

6 Results and discussions

This research, ML regression algorithms played an important role in achieving this experimentation, this experiment is done on personal DELL laptop, which has a configuration of Intel(R) Core (TM) i5-8265U CPU @ 1.60GHz1.80 GHz, and 16 GB RAM. Importance of the project results are examined in this section. In this project, author has chosen 3 files of dataset which are in .csv format of data have considered for the experiment. Did the experimentation on 13 different regression algorithms. Figure 10 shows, different evaluation parameters used for result analysis.

Figure 10a shows computing time comparison plot, LR takes a shorter time than DTR comes second than DTRT comes first than other regression algorithms.

Figure 10b shows R^2 score comparison plot, Random Forest Regression has a highest R^2 score 0.973, K Neighbors Regression with hyper parameter tuning comes second with R^2 score 0.961 and Decision Tree Regression comes third withR^2score 0.96, K Neighbors Regression comes fourth withR^2 score 0.956.

Figure 10c shows, RMSE comparison plot, the RFR model achieved a lower RMSE compared to the Regression model. This indicates that the RFR model has better predictive accuracy.

Figure 10d shows, MSE comparison plot, the RFR, KNRT, DTR Regression model achieved a lower MSE compared to the other Regression model. This indicates that the RFR, KNRT, DTR model have better predictive accuracy.

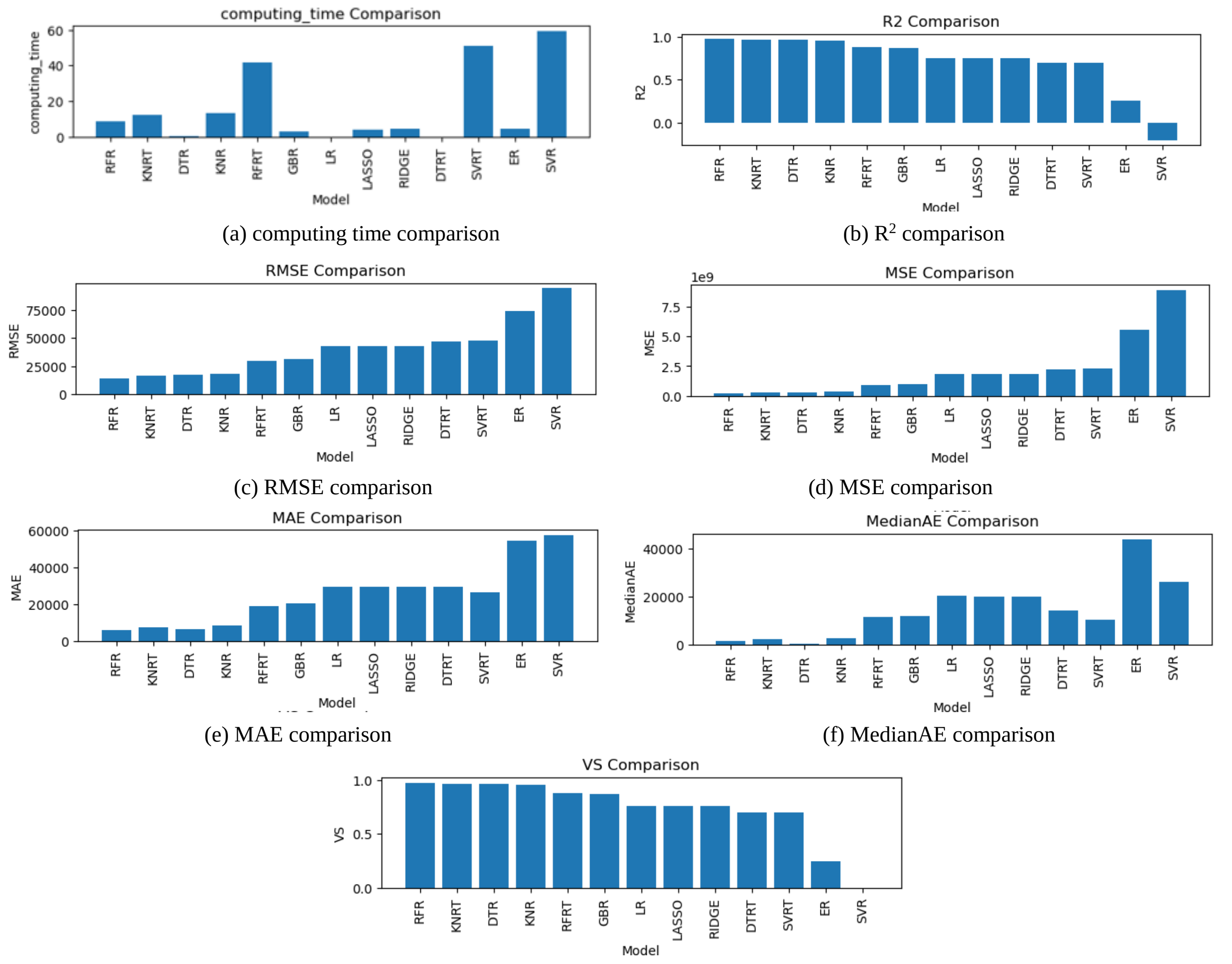
Figure 10e shows MAE comparison plot, the RFR, DTR, KNRT, KNR Regression model achieved a lower MAE score compared to the other Regression model. This indicates that the RFR, DTR, KNRT, KNR model have better predictive accuracy.

Figure 10f shows Median AE comparison plot, the RFR, DTR, KNRT, KNR Regression model achieved a lower Median AE score compared to the other Regression model. This indicates that the RFR, DTR, KNRT, KNR model have better predictive accuracy.

Figure 10g shows, VS comparison plot, comparing the variance scores of different regression models helps determine which model provides a better fit to the data and explains a larger proportion of the variance in the target variable. A higher variance score indicates a better-performing model RFR, KNRT, KNR have higher variance score as compared with other regression models.

Overall, these comparison plots provide insights into the performance of different regression models in terms of computational time, R^2 score, RMSE, MSE, MAE, MedianAE, and variance score. They can help researchers and practitioners select the most suitable regression model for their specific prediction tasks.

Graphical comparison analysis for 13 different regression model for 8 different regressions metric as follows:



(a) computing time comparison

(b) R^2 comparison

(c) RMSE comparison

(d) MSE comparison

(e) MAE comparison

(f) MedianAE comparison

(g) VS comparison

Figure 10 Computing time comparison, R^2 comparison, RMSE Comparison, MSE comparison, MAE comparison, Median comparison, VS comparison

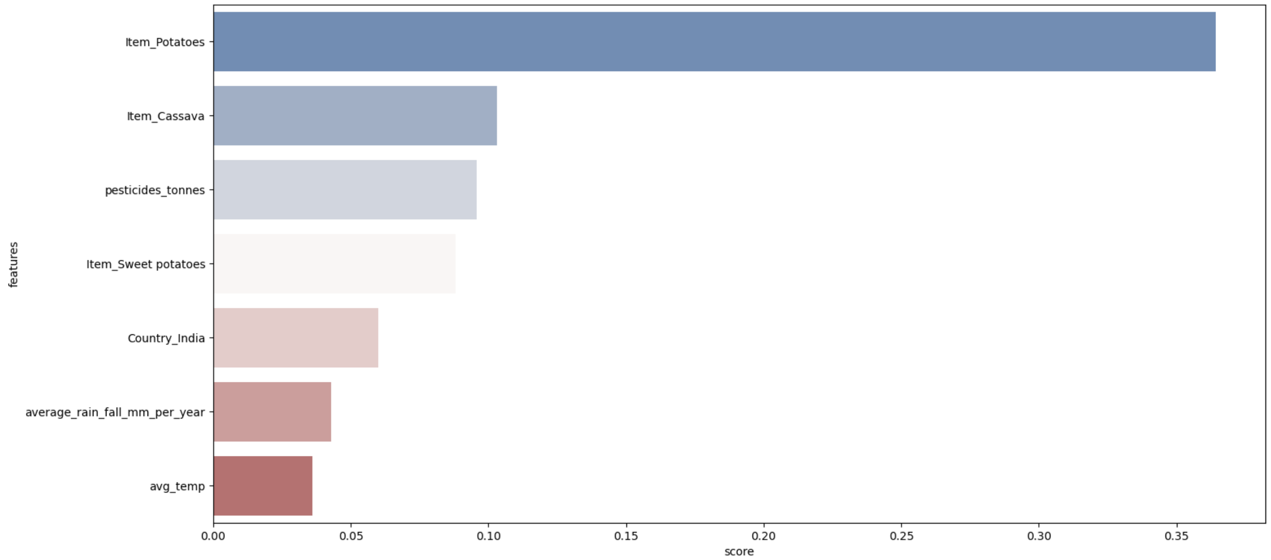


Figure 11 Important factors that affects crops

Figure 11 illustrates the key factors influencing crop outcomes. Among these factors, potatoes exhibit the highest importance in the decision-making process of the model, as they represent the most abundant crop in the dataset. Following potatoes, cassava demonstrates a significant impact on crop yield, ranking as the second most influential feature, primarily attributed to the effects of pesticides. Additionally, sweet potatoes exhibit notable importance in the dataset, particularly in terms of crop yield. The production location of the crop also holds implications, with India boasting the largest overall crop yield in the dataset. Furthermore, average rainfall and average air temperature have substantial effects on crop yield, aligning with the initial assumption. These features significantly influence the expected crop yield within the model. It is worth noting that the dataset encompasses 101 countries, with India leading in terms of highest crop yield production, followed by Brazil, Mexico, and others in descending order.

Table 3Highest yield production countries

Item	
Cassava	0.924311
Maize	0.889169
Plantains and others	0.803042
Potatoes	0.909158
Rice, paddy	0.896501
Sorghum	0.807015
Soybeans	0.836912
Sweet potatoes	0.848899
Wheat	0.924299
Yams	0.927155

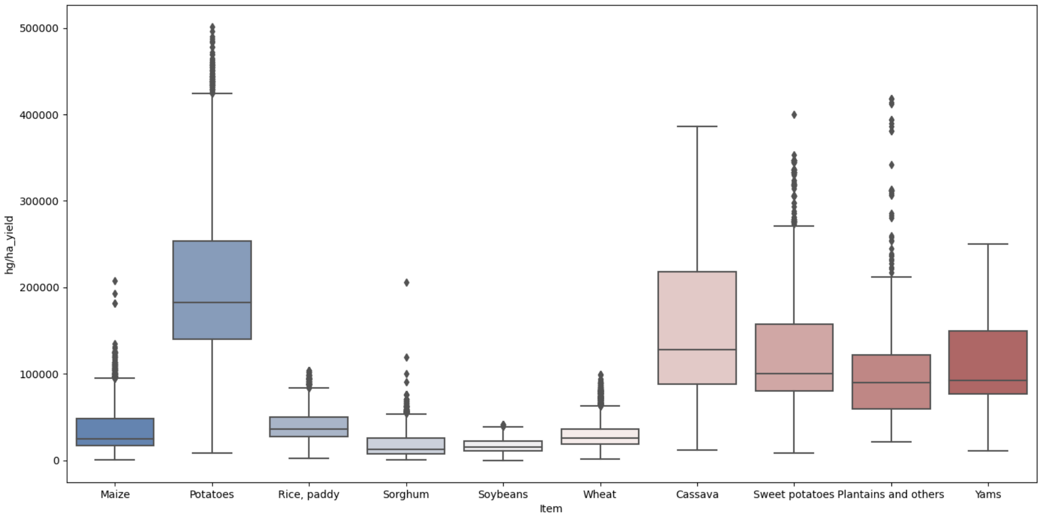


Figure 12 Boxplot for yield for each crop

Figure 12 shows the yield for each item that is crop like maize, potatoes, rice, paddy, soybeans, wheat, cassava, sweet potatoes, others and yams. Out of all these items, Potatoes has a highest yield then cassava, sweet potatoes and so on. Figure 13 shows the relationship between the actual and predicted yield values for the crop. Each data point represents an instance and the closer the points align to a diagonal line, the better the model's predictions match the actual values.

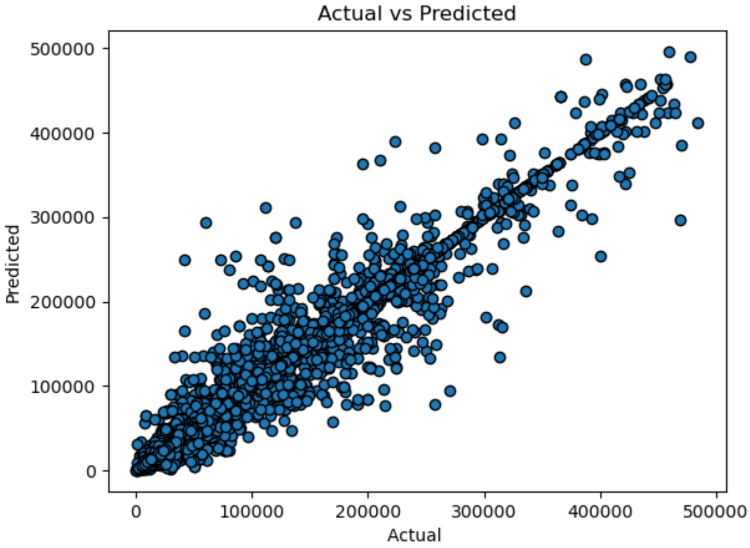


Figure 13 Actual vs. predicted results for crop yield prediction

7 Conclusion

In conclusion, a performance-based comparison of regression models for crop yield prediction in agriculture is a valuable research area that can contribute to the advancement of agricultural decision-making and resource allocation. Through rigorous evaluation and comparison of different regression models, researchers can gain insights into their strengths, weaknesses, and applicability in various agricultural contexts. However, this research area also faces challenges related to data quality, model complexity, generalizability, and model evaluation.

After doing the performance-based comparison on 13 different regression model such as LR, KNR, SVR, DTR, RFR, GBR, Linear Model Lasso Regression, Elasticnet Regression, Ridge Regression also used GridSerach CV hyper parameter tuning technique on following regression models to improve the model performance KNR, DTR, RFR, SVR To evaluate performance of regression model used R ^{2}score, RMSE, MSE, MAE, Median SE, Explain variance score and computing time. Experimental results shows Random Forest Regression has a highest R^2 score 0.973, K Neighbors Regression with hyper parameter tuning comes second with R^2 score 0.961 and Decision Tree Regression comes third withR^2 score 0.96, K Neighbors Regression comes fourth withR^2 score 0.956.

A performance-based comparison of regression models for crop yield prediction has the potential to greatly benefit the agriculture industry. It empowers farmers and policymakers to make informed decisions regarding crop management practices, resource allocation, and risk mitigation. By conducting further research in this field, we can advance the development of more accurate and reliable crop yield prediction models. Ultimately, this will contribute to improved crop productivity, sustainability, and enhanced food security for the future.

8 Challenges and future scope

8.1 Challenges

Data quality and availability: One of the challenges in conducting a performance-based comparison of regression models for crop yield prediction is the quality and availability of data. Agricultural data, such as weather parameters, soil characteristics, and crop management practices, may be sparse, noisy, or inconsistent, which can affect the accuracy and reliability of the prediction models.

Model complexity and interpretability: Different regression models may have varying levels of complexity, which can impact their interpretability and practical applicability in real-world agricultural settings. Some complex models, such as deep learning techniques, may achieve high predictive accuracy but can be challenging to interpret and explain to stakeholders, such as farmers or policymakers.

Generalizability and scalability: Crop yield prediction models need to be able to generalize and scale across different regions, crops, and growing seasons to be practical and useful for farmers and policymakers. However, achieving high generalization and scalability can be challenging due to the inherent variability in agricultural systems, including differences in soil types, weather patterns, and crop management practices.

Model selection and evaluation: Choosing the most appropriate regression model for a specific agricultural context can be challenging due to the vast number of available models with varying assumptions, algorithms, and parameters. Additionally, evaluating the performance of different models requires careful consideration of appropriate evaluation metrics, cross-validation techniques, and statistical significance tests.

8.2 Future scope

Improved data collection and integration: Future research can focus on improving the quality and availability of agricultural data, including the integration of diverse data sources such as remote sensing, drones, and IoT devices. This can enhance the accuracy and reliability of crop yield prediction models.

Advanced model development: Research can explore the development of advanced regression models, such as ensemble methods, Bayesian approaches, and hybrid models, to further improve the accuracy and interpretability of crop yield prediction models. Additionally, incorporating domain-specific knowledge, such as crop physiology and phenology, can enhance the predictive capability of the models.

Model interpretability and explain ability: Future research can focus on developing techniques to improve the interpretability of complex regression models, such as deep learning techniques, to gain trust and acceptance among stakeholders. This can involve techniques such as model visualization.

Decision support systems: Crop yield prediction models can be integrated into decision support systems that provide actionable insights and recommendations to farmers and policymakers. Future research can focus on developing user-friendly decision support systems that are tailored to the needs of different agricultural stakeholders, considering factors such as local context, user preferences, and usability.

Real-time and dynamic prediction: Research can explore the development of real-time and dynamic crop yield prediction models that can adapt to changing weather conditions, crop growth stages, and management practices. This can enable farmers to make timely and informed decisions for optimizing crop yield based on current conditions, resulting in improved yield prediction accuracy and resource allocation.

In conclusion, a performance-based comparison of regression models for crop yield prediction in agriculture faces challenges related to data quality, model complexity, generalizability, and model evaluation. However, there are opportunities for future research to address these challenges and further enhance the accuracy, interpretability, and applicability of crop yield prediction models, leading to improved decision-making in agriculture.

References

Aditya, S.,H.A.Sanjay, andE. Bhanusree. 2017.Prediction of crop yield using regression techniques. *International Journal of Soft Computing*, 12(2):96-102.

Aruvansh,N., S.Garg, A. Agrawal, and P.Agrawal. 2019. Crop yield prediction using machine learning algorithms.In *2019 Fifth International Conference on Image Information Processing (ICIIP), IEEE*-130, Shimla, India, 15-17 November 2019

Ayush,S., A. Dubey,V. Hemnani, D. Galaand, and D. R.Kalbande.2018. Smart farming system: crop yield prediction using regression techniques. In Book:*Proceedings of International Conference on Wireless Communication*,(1)*:*49-56.

Balsher,S.S., Z.Mehrabi, and N.Ramankutty, and M. Kandlikar. 2023. How can machine learning help in understanding the impact ofclimate change on crop yields.*Environment Research Letters*, 18(2023): 024008.

D.Jayanarayana, R., andR. Kumar. 2021.Crop yield prediction using machine learningalgorithm. In *Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS),1466-1470.*Madurai, India, May 6-8th, 2021

Dhivya,E., D.Vincent,V.Sharma,A.Zomaya,and K.Srinivasan.2019. Forecasting yield by integrating agrarian factor sand machine learning models: Asurvey.*Computers and Electronics in Agriculture*,155:257-282.

Elisa,K.,F.Waldner, and Z.Hochman. 2020. Estimating wheat yields in Australia using climate records,satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*,160(2):124-135.

Farhat,A.,H.Afzaal,A. Farooque,andS.Tang. 2020. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, 10(7):1046.

Iniyan,S.,A. V.Verma,and C.T.Naidu. 2013. Crop yield prediction using machine learning techniques.*Advance in Engineering Software*,175: 103326.

Kasi,L. R.,and A. P. S.Kumar.2023. Machine learning techniques for weather based crop yield prediction. In*Third International Conference on Artificial Intelligence and Smar tEnergy(ICAIS), IEEE*,1-6. Coimbatore, India, Feb2-4th, 2023.

Maimaitijiang, M.,V.Sagan,P.Sidike,S.Hartling,F.Esposito, andF.B.Fritschi.2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment*, 237:111599.

Mamunur, R.,B. S. Bari,Y. Yusup, M. A. Kamaruddin, and N. Khan. 2021. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction.*IEEE*,4: 1-37

Maninder,S.D., T.Dahms,C. Kuebert-Flock, T.Rummler, J.Arnault,I. Stefan-Dewenter,and T. Ullmann. 2023. Integrating random forest and crop modeling improves the cropyield prediction of winter wheatand oil seed rape.*Frontiers in Remote Sensing,* 3: 1-19

Martin Kuradusenge, Eric Hitimana, Damien Hanyurwimfura. 2023. Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *MDPI, Agriculture*, 13(225):1-19

Maya, G. P.S., andR.Bhargavi.2019. A novel approach for efficient crop yield prediction. *Computers and Electronics in Agriculture,* 165: 104968.

Pantazi, X. E. D.Moshoua, T.Alexandridis, R.L.Whettonc,and A.M. Mouazen.2016. Wheat yield prediction using machine learning and advanced sensing techniques.*Computers and Electronics in Agriculture,* 121:57-65.

Potnuru,S. N., P.S.Venkat, B.L. Avinash, and B.Jabber.2020.Crop yield prediction based on Indian agriculture using machine learning. *International Conference for Emerging Technology(INCET),* 15. Belgaum, India, 05-07 June 2020.

Ramesh,A. M., V. Rajpurohit, and S.Shweta. 2019. Crop yield prediction using machine learning techniques. In*2019 5th International Conference for Convergence in Technology (I2CT*),1-4. Bombay, India, Mar 29-31th, 2019.

Sellam, V., andE. Poxovammal. 2016. Prediction of crop yield using regression analysis. *Indian Journal of Science and Technology,* 9(38): 1-5.

Shruthi, G.,andR.Sangeeta.2020. Design and implementation of crop yield prediction model in agriculture.*International Journal of Scientific andTechnology Research,*8(1): 544.

Sunday,S. O., E.A.Olajubu, and D. Olanikel. 2023.An ensemble deep learning approach for predicting cocoa yield, Heliyon.*Cell Press,*9(4):e15245.

Yunus,K., and N.Polat. 2023. A linear approach for wheat yield prediction by using different spectral vegetation indices.*International Journal of Engineering and Geosciences,*8(1): 52-62.