Tuijin Jishu/Journal of Propulsion Technology
ISSN:1001-4055
Vol. 44 No.3 (2023)

A Study on Prediction for Crop Area, Production, and Yield Analys is Using Machine Learning Approaches

J. Karthikeyan[1], Dr. A. Murugan[2]

[1]*Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India*

[2]*Assistant Professor, Department of Computer Science, Periyar Arts College, Cuddalore, (Deputed from Annamalai University, Annamalainagar) Tamil Nadu, India*

Abstract

Analyzing crop area, production, and yield typically involves data collection from farmers, the application of remote sensing technologies like satellite imagery, and on-site surveys. This gathered data is then used to assess crop performance, identify trends, and offer recommendations to improve agricultural practices.Data mining finds extensive application across diverse sectors such as business, healthcare, finance, marketing, and scientific research, enabling the extraction of knowledge and insights from data that may not be readily discernible using conventional approaches.This paper considers Crops Production between 2006 to 2011. The machine learning approaches which is used to analysis and predict the dataset usinglinear regression, multilayer perceptron, SMOreg, random forest, random tree, and REP tree**.** Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

Keywords: *Machine learning, crop area, production, yield analysis, decision tree, correlation coefficient, and test statistics.*

1. Introduction and Literature Review

Machine learning and data mining in the analysis of crop area, production, and yield, agricultural stakeholders can bolster their decision-making, streamline resource utilization, boost efficiency, and promote sustainable farming practices. These technologies foster a data-centric approach to agriculture, ultimately resulting in more effective crop management and heightened food security.

The present research aims to illuminate the role of machine learning in agriculture by conducting an extensive review of recent scholarly literature. This review is based on keyword combinations such as "machine learning" alongside "crop management," "water management," "soil management," and "livestock management," following PRISMA guidelines. Only journal papers published between 2018 and 2020 were considered. The findings suggest that this topic spans various disciplines, promoting international convergence research. Notably, crop management takes center stage, with a plethora of machine learning algorithms utilized, with Artificial Neural Networks standing out for their efficiency. Maize and wheat, along with cattle and sheep, were the most investigated crops and animals, respectively. Additionally, a variety of sensors, including those on satellites and unmanned ground and aerial vehicles, have been employed to gather reliable input data for data analysis. It is anticipated that this study will serve as a valuable guide for all stakeholders, raising awareness of the potential benefits of employing machine learning in agriculture and fostering more systematic research in this domain [1].

Soil classification is based on test reports, focusing on features related to boron (B), organic carbon (OC), potassium (K), phosphorus (P), available boron (B), and soil pH. Cross-validation is conducted in ten steps, with 10% of data used for validation in each step. The fast-learning classification method, extreme learning method (ELM), is employed to identify soil micronutrients. Various activation functions are used to optimize the methodology. The analysis results in the classification of nutrients and the proposal of optimal soil conditions for different regions. The study reveals that soils in Tamil Nadu have normal electrical conductivity, a red color, are rich in potassium (35% of samples), nitrogen (80% of samples), and sulfur (75% of samples), while they exhibit sufficiency or deficiency in magnesium, boron, zinc, and copper [2].

A method is proposed for classifying soil based on macro and micro nutrients, predicting the crop type suitable for the soil. Multiple machine learning algorithms, including K-Nearest Neighbour (k-NN), Bagged tree, Support vector machine (SVM), and logistic regression, are employed [3].

The primary objective of this work is to explore soil supplements using data mining classification techniques. A substantial dataset of soil nutrient status is collected from the Department of Agriculture, Cooperation and Farmers Welfare, covering various states. The paper focuses on specific districts in Tamil Nadu, analyzing soil nutrients, including Nitrogen, Phosphorus, Potassium, Calcium, Magnesium, Sulfur, Iron, Zinc, and more, using Naïve Bayes, Decision Tree, and a hybrid approach of Naïve Bayes and Decision Tree. Classification algorithms are compared based on accuracy and execution time [4].

Author introduces an approach that employs various machine learning techniques to predict crop yields based on soil analysis. Data for crop yield prediction is obtained from different regions of Jammu District, focusing on macro-nutrients and micro-nutrients. Several classifier algorithms are utilized in a classification framework [5].

Data mining is a valuable tool for extracting hidden insights from large pre-existing databases. In this paper, weather data is used to determine whether conditions are conducive to playing golf. Several classification algorithms are applied, with the Random Tree algorithm outperforming others with an accuracy of 85.714% [6].

Machine learning techniques are applied to predict Mustard crop yield from soil analysis. The study employs five supervised machine learning techniques: K-Nearest Neighbor (KNN), Naïve Bayes, Multinomial Logistic Regression, Artificial Neural Network (ANN), and Random Forest, evaluating their performance based on accuracy, recall, precision, specificity, and f-score [7].

The project presents a machine learning-based solution for analyzing soil properties and grading soil based on nutrient content. It uses regression algorithms for predicting soil rankings and classification algorithms for crop recommendations. Random Forest achieves the highest accuracy score [8].

The study explores the effects of soil particle size on element estimation using hyperspectral imaging, developing models to predict soil element content. Results indicate varying precision in predicting different elements, with some elements predicted more accurately in sieved soils. Grinding the soil samples did not significantly affect the precision of prediction [9].

The paper proposes an efficient assessment of groundwater levels, rainfall, population, food grains, and enterprises data using stochastic modeling and data mining. Novel data assimilation analysis is introduced to predict groundwater levels effectively [10] and [11].

The research involves chronic disease data analysis and presents the performance of five different decision tree algorithms. The M5P decision tree approach stands out as the best algorithm for building the model compared to other decision tree approaches [12].

2. Backgrounds and Methodologies

A data mining decision tree is a widely used machine learning technique for classification and regression tasks. It visually depicts a sequence of decisions and their possible outcomes in a tree-like structure. Each internal node represents a decision based on a specific feature, and each branch corresponds to the potential result of that

decision. The tree's leaf nodes represent the final decision or the predicted outcome.The "CART" (Classification and Regression Trees) algorithm is the most used algorithm for building decision trees [13].

2.1 Linear Regression

Linear regression is a statistical technique employed to comprehend and forecast the connection between two variables by discovering the optimal straight line that most effectively aligns with the data points. It aids in ascertaining how alterations in one variable correspond to changes in another, proving valuable for predictions and trend recognition.

The core idea of linear regression is to find the best-fitting straight line (also called the "regression line") through a scatterplot of data points. This line represents a linear equation of the form:

$$y = m_x + b \quad \ldots (1)$$

Where:

- y is the dependent variable (the one you want to predict or explain).
- x is the independent variable (the one you're using to make predictions or explanations).
- m is the slope of the line, representing how much
- y changes for a unit change in x.

b is the y-intercept, indicating the value of y when x is 0.

2.2 Multilayer Perception

A Multilayer Perceptron (MLP) is an artificial neural network consisting of multiple layers of interconnected nodes or neurons. It's a fundamental architecture in deep learning and is used for various tasks, including classification, regression, and more complex tasks like image recognition and natural language processing. The architecture of an MLP typically includes three types of layers:

i. **Input Layer:** This layer consists of neurons receiving input data. Each neuron corresponds to a feature in the input data, and the values of these neurons pass through the network.
ii. **Hidden Layers:** These layers come after the input layer and precede the output layer. They are called "hidden" because their activations are not directly observed in the final output.
iii. **Output Layer:** This layer produces the network's final output. The number of neurons in the output layer depends on the problem type.

2.3 SMO

SMO stands for "Sequential Minimal Optimization," an algorithm used for training support vector machines (SVMs), machine learning models commonly used for classification and regression tasks. The SMO algorithm is particularly well-suited for solving the quadratic programming optimization problem that arises during the training of SVMs.

Step 1. **Initialization:** Start with all the data points as potential support vectors and initialize the weights and bias of the SVM.
Step 2. **Selection of Two Lagrange Multipliers:** In each iteration, the SMO algorithm selects two Lagrange multipliers (associated with the support vectors) to optimize.
Step 3. **Optimize the Pair of Lagrange Multipliers:** Fix all the Lagrange multipliers except the selected two, and then optimize the pair chosen to satisfy certain constraints while maximizing a specific objective function.
Step 4. **Update the Model:** After optimizing the selected pair of Lagrange multipliers, update the SVM model's weights and bias based on the new values of the Lagrange multipliers.
Step 5. **Convergence Checking:** Check for convergence criteria to determine whether the algorithm should terminate.
Step 6. **Repeat:** If convergence hasn't been reached, repeat steps 2 to 5 until it is.

2.4Random Forest

Random Forest is a popular machine learning ensemble method for classification and regression tasks. It is an extension of decision trees and is known for its high accuracy, robustness, and ability to handle complex datasets. Random Forest is widely used in various domains, including data science, machine learning, and pattern recognition.The main idea behind Random Forest is to create an ensemble (a collection) of decision trees and combine their predictions to make more accurate and stable predictions. The following steps describe what Random Forest works like.

- ❖ Bootstrap Aggregating (Bagging)
- ❖ Decision Tree Construction
- ❖ Voting for Classification, Averaging for Regression

The key advantages of Random Forest are:

- ❖ Reduced overfitting
- ❖ Robustness
- ❖ Feature Importance

Steps involved in Random Forest

Random Forest is an ensemble learning method combining multiple decision trees to make more accurate and robust predictions for classification and regression tasks. The steps involved in building a Random Forest are as follows:

Step 1. Data Bootstrapping
Step 2. Random Feature Subset Selection
Step 3. Decision Tree Construction
Step 4. Ensemble of Decision Trees
Step 5. Out-of-Bag (OOB) Evaluation
Step 6. Hyperparameter Tuning (optional)

2.5Random Tree

In machine learning, a Random Tree is a specific type of decision tree variant that introduces randomness during construction. Random Trees are similar to traditional decision trees but differ in how they select the splitting features and thresholds at each node.The primary goal of introducing randomness is to create a more diverse set of decision trees, which can help reduce overfitting and improve the model's generalization performance. Random Trees are commonly used as building blocks in ensemble methods like Random Forests.The critical characteristics of Random Trees are as follows:

- ❖ Random Feature Subset
- ❖ Random Threshold Selection
- ❖ No Pruning
- ❖ Ensemble Methods

Steps involved in Random Tree

Step 1. Data Bootstrapping:
Step 2. Random Subset Selection for Features:
Step 3. Decision Tree Construction:
Step 4. Voting (Classification) or Averaging (Regression):

2.6REP Tree

REP (Repeated Incremental Pruning to Produce Error Reduction) Tree is a machine learning algorithm for classification and regression tasks. A decision tree-based algorithm constructs a decision tree using a

combination of incremental pruning and error-reduction techniques.The key steps involved in building a REP Tree are as follows:

- ❖ Recursive Binary Splitting
- ❖ Pruning
- ❖ Repeated Pruning and Error Reduction

Steps involved in REP Tree

REP Tree (Repeated Incremental Pruning to Produce an Error Reduction Tree) is a machine learning algorithm for classification and regression tasks. It is an extension of decision trees that incorporates pruning to reduce overfitting and improve the model's generalization performance. Below are the steps involved in building a REP Tree.

Step 1. Recursive Binary Splitting
Step 2. Pruning
Step 3. Repeated Pruning and Error Reduction
Step 4. Model Evaluation

2.7 Accuracy Metrics

The predictive model's error rate can be evaluated by applying several accuracy metrics in machine learning and statistics. The basic concept of accuracy evaluation in regression analysis is comparing the original target with the predicted one and using metrics like R-squared, MAE, MSE, and RMSE to explain the errors and predictive ability of the model [14]. The R-squared, MSE, MAE, and RMSE are metrics used to evaluate the prediction error rates and model performance in analysis and predictions [15] and [16].

R-squared (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The values from 0 to 1 are interpreted as percentages. The higher the value is, the better the model is.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad \dots (2)$$

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad \dots (3)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad \dots (4)$$

Relative Absolute Error (RAE) is a metric used in statistics and data analysis to measure the accuracy of a forecasting or predictive model's predictions. It is particularly useful when dealing with numerical data, such as in regression analysis or time series forecasting.

$$RAE = \frac{\sum|y_i - \hat{y}_i|}{\sum|y_i - \bar{y}|} \quad \dots (5)$$

Root Relative Squared Error (RRSE) is another metric used in statistics and data analysis to evaluate the accuracy of predictive models, especially in the context of regression analysis or time series forecasting.

$$RRSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}} \quad \dots (6)$$

Equation 3 to 7 are used to find the model accuracy, which is used to find the model performance and error. Where Y_i represents the individual observed (actual) values, $\hat{Y}_i$ represents the corresponding individual

predicted values, $\bar{Y}$ represents the mean (average) of the observed values and Σ represents the summation symbol, indicating that you should sum the absolute differences for all data points.

3. Numerical Illustrations

The corresponding dataset was collected from the open souse Kaggle data repository. The crop area, production, yield analysis datasetincludes5 parameters whichhavedifferent categories of data likecrop, year, area, production, yield[17]. Adetailed description of the parameters is mentioned in the following Table 1.

Table 1. Crop area, production, yield analysis sampledataset

Crop	Year	Area	Production	Yield
Total Foodgrains	2006 - 07	128.5	158.8	123.6
Rice	2006 - 07	168.5	200.8	119.2
Wheat	2006 - 07	115	131.6	114.4
Jowar	2006 - 07	120.7	124.3	103
Bajra	2006 - 07	94.5	136.4	144.3
Maize	2006 - 07	156.1	198.8	127.3
Ragi	2006 - 07	60.2	56.1	93.2
Small millets	2006 - 07	50.9	53.9	106
Barley	2006 - 07	72.8	88.1	121

Table 2: Machine Learning Models with Correlation coefficient

ML Approaches	Area	Production	Yield
Linear Regression	0.9344	0.9975	0.9978
Multilayer Perceptron	0.9383	0.9913	0.9885
SMOreg	0.9584	0.9976	0.9981
Random Forest	0.9481	0.9837	0.9926
Random Tree	0.9405	0.9841	0.9928
REP Tree	0.9301	0.9808	0.9928

Table 3: Machine Learning Models with Mean Absolute Error

ML Approaches	Area	Production	Yield
Linear Regression	10.1034	7.1632	5.6931
Multilayer Perceptron	10.6984	16.0502	15.0350
SMOreg	7.9784	6.4856	5.6287
Random Forest	8.7458	20.8051	11.7638
Random Tree	9.2702	21.3886	12.2439

REP Tree	10.0911	23.1812	11.9434

Table 4: Machine Learning Models with Root Mean Squared Error

ML Approaches	Area	Production	Yield
Linear Regression	16.0588	13.9538	9.8709
Multilayer Perceptron	16.0387	26.2495	23.109
SMOreg	12.9531	14.2499	9.0844
Random Forest	14.274	36.5733	19.2086
Random Tree	15.3106	34.8845	17.7054
REP Tree	16.5736	38.2554	17.6972

Table 5: Machine Learning Models with Relative Absolute Error (%)

ML Approaches	Area	Production	Yield
Linear Regression	30.8990	8.9193	11.0527
Multilayer Perceptron	32.7186	19.9848	29.1891
SMOreg	24.4002	8.0755	10.9277
Random Forest	26.7471	25.9053	22.8385
Random Tree	28.3509	26.6318	23.7705
REP Tree	30.8613	28.8639	23.1870

Table 6: Machine Learning Models with Root Relative Squared Error (%)

ML Approaches	Area	Production	Yield
Linear Regression	35.7620	7.0743	6.6422
Multilayer Perceptron	35.7173	13.3079	15.5501
SMOreg	28.8459	7.2244	6.1129
Random Forest	31.7875	18.5418	12.9255
Random Tree	34.0959	17.6856	11.9140
REP Tree	36.9086	19.3946	11.9085

Table7: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Area	Production	Yield
Linear Regression	0.3800	0.0700	0.0500
Multilayer Perceptron	3.1500	3.0100	2.9900

SMOreg	0.6600	0.7400	0.8700
Random Forest	0.2500	0.0500	0.0500
Random Tree	0.1000	0.1000	0.1000
REP Tree	0.2200	0.2200	0.2200

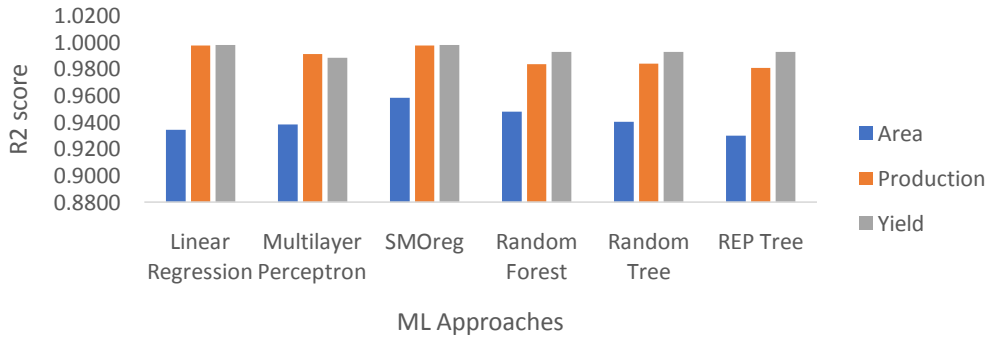

Fig. 1. R2 Score for Machine Learning Approaches

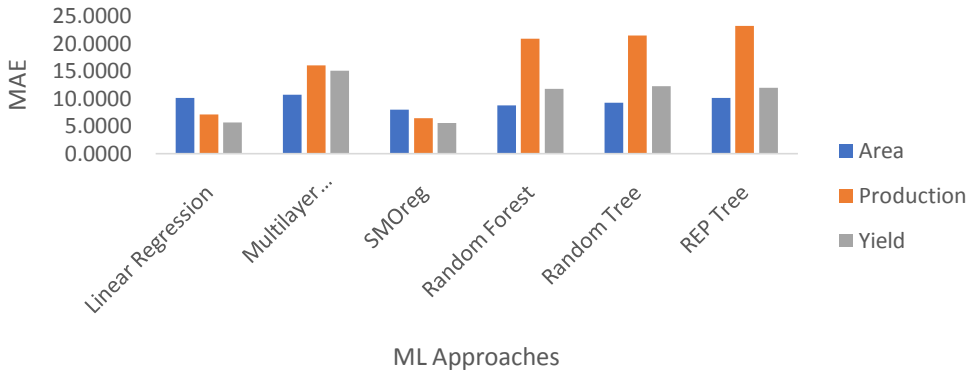

Fig. 2.Machine Learning Models with MAE

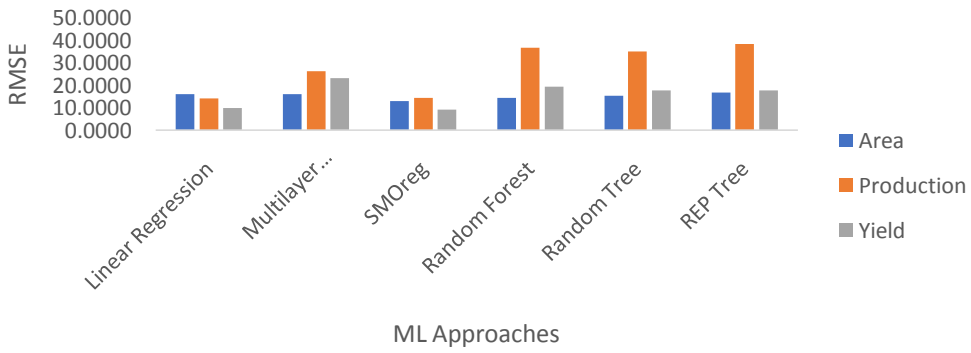

Fig. 3. Machine Learning Models with RMSE

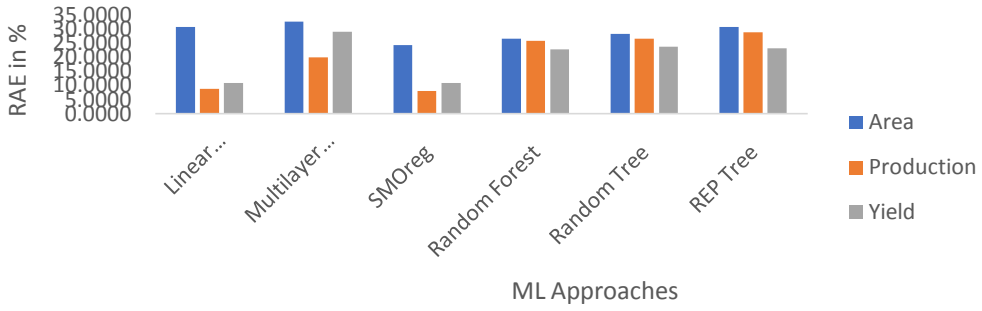

Fig.4. Machine Learning Models with RAE (%)

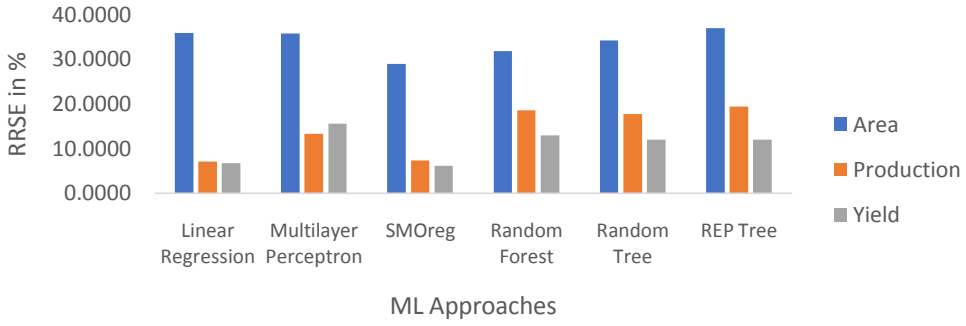

Fig. 5. Machine Learning Models with RRSE (%)

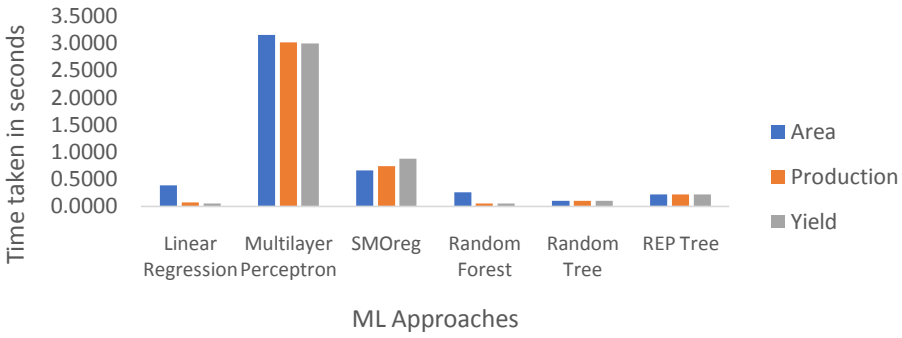

Fig. 6. Machine Learning Models and its Time Taken to Build the Model (Seconds)

4. Results and Discussion

Table 1 provides an explanation of five key parameters, encompassing various data categories, such as crop, year, area, production, yield, and diverse agricultural product information. Utilizing this dataset, we applied six distinct machine learning approaches, namely linear regression, multilayer perceptron, SMOreg, random forest, random tree, and REP tree, to unveil hidden patterns and determine the most influential parameter for future predictions. Comprehensive results and numerical representations are presented across Table 1 to Table 7 and Figure 1 to Figure 6.

These findings are derived from Equation 2, Table 2, and Figure 1, which are employed to compute the R2 score and correlation coefficients across the five parameters. The numerical evidence suggests substantial variations between these parameters. Notably, when using area, production, and yield, all six machine learning approaches exhibit a robust positive correlation, approaching 0.9.

We gauge model errors by employing the Mean Absolute Error (MAE) as defined in Equation 3, incorporating six different machine learning algorithms. Across the board, all seven ML approaches yield the best error performance, with an average MAE of approximately 5.6. These outcomes are illustrated in Table 3 and Figure 2.

The Root Mean Square Error (RMSE) is utilized to quantify the disparities between predicted and actual values, in line with Equation 4. In this context, each machine learning approach exhibits superior error performance, with an average RMSE of around 9. Corresponding numerical representations can be found in Table 4 and Figure 3.

Furthermore, the Relative Absolute Error (RAE), determined using Equation 5, is employed to measure accuracy by comparing predicted and actual values in percentage terms. This study involves six ML classification algorithms, with linear regression and REP Tree displaying the highest error rates. The remaining four ML approaches showcase exceptional performance and minimal error, with analogous results in RRSE. Corresponding numerical illustrations are available in Table 5 to 6 and Figure 4 to 5.

In the realm of time efficiency, a critical aspect of machine learning, Table 7 and Figure 6 reveal noteworthy insights. Multi-layer perceptrons are the most time-consuming, whereas Random Tree, REP Tree, and Random Forests require the least time to build models. Linear regression and SMOreg also stands out as a minimal time-consuming approach for model development. These observations are mirrored in the visual representations.

5. Conclusion and Future Research

In conclusion, it is imperative to address the limitations of our model. These constraints encompass considerations related to micro and macro nutrient data, such as area, production, and yield, as well as model-specific factors that may contribute to potential underperformance. Additionally, computational constraints that may have influenced model development should be acknowledged. For future research, we propose several avenues for improvement. This includes exploring additional data sources to enrich the dataset, investigating more effective algorithms or hyperparameters, and fine-tuning the model to enhance its overall performance. This research offers valuable insights to the Department of Agriculture and related stakeholders, aiding them in developing the agriculture sector.

6. Reference

[1] Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D. and Bochtis, D., 2021. Machine learning in agriculture: A comprehensive updated review. Sensors, 21(11), p.3758.

[2] BlesslinSheeba, T., Anand, L.D., Manohar, G., Selvan, S., Wilfred, C.B., Muthukumar, K., Padmavathy, S., Ramesh Kumar, P. and Asfaw, B.T., 2022. Machine Learning Algorithm for Soil Analysis and Classification of Micronutrients in IoT-Enabled Automated Farms. Journal of Nanomaterials, 2022.

[3] Saranya, N. and Mythili, A., 2020. Classification of soil and crop suggestion using machine learning techniques. Int J Eng Res Technol, 9(02), pp.671-673.

[4] Manjula, E. and Djodiltachoumy, S., 2017. Data mining technique to analyze soil nutrients based on hybrid classification. International Journal of Advanced Research in Computer Science, 8(8).

[5] Singh, V., Sarwar, A. and Sharma, V., 2017. Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach. International Journal of Advanced Research in Computer Science, 8(5).

[6] Rajesh, P. and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using the Weka tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243.

[7] Pandith, V., Kour, H., Singh, S., Manhas, J. and Sharma, V., 2020. Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. Journal of Scientific Research, 64(2), pp.394-398.

[8] Keerthan Kumar, T.G., Shubha, C. and Sushma, S.A., 2019. Random forest algorithm for soil fertility prediction and grading using machine learning. Int J InnovTechnolExplorEng, 9(1), pp.1301-1304.

[9] Malmir, M., Tahmasbian, I., Xu, Z., Farrar, M.B. and Bai, S.H., 2019. Prediction of soil macro-and micro-elements in sieved and ground air-dried soils using laboratory-based hyperspectral imaging technique. Geoderma, 340, pp.70-80.

[10] Rajesh, P., Karthikeyan, M. and Arulpavai, R., 2019, December. Data mining approaches to predict the factors that affect the groundwater level using a stochastic model. In AIP Conference Proceedings (Vol. 2177, No. 1). AIP Publishing.

[11] Rajesh, P. and Karthikeyan, M., 2019. Data mining approaches to predict the factors that affect agriculture growth using stochastic models. International Journal of Computer Sciences and Engineering, 7(4), pp.18-23.

[12] Rajesh, P., Karthikeyan, M., Santhosh Kumar, B. and Mohamed Parvees, M.Y., 2019. Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience, 16(4), pp.1472-1477.

[13] Kohavi, R., &Sahami, M. (1996). Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286).

[14] Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/

[15] S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, “Root mean square error (RMSE): A comprehensive review,” International Journal of Applied Mathematics and Statistics, vol. 59, no. 1, pp. 42–49, 2019.

[16] Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566

[17] https://www.kaggle.com/code/jocelyndumlao/crop-yield-variation-across-states/input?select=datafile+%282%29.csv