Analysis and Prediction for Crop Yield Variations across States in India 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
Email: thalapathik80@gmail.com

[2]Assistant Professor, Department of Computer Science, Periyar Arts College, Cuddalore, (Deputed from Annamalai University, Annamalainagar) Tamil Nadu, India
Email: drmuruganapcs@gmail.com

Abstract:

The fluctuations in crop yields among Indian states stem from a intricate interaction between natural and human-made elements. Grasping these factors is of paramount importance for policymakers and farmers as they endeavor to make informed choices and tackle the issues associated with attaining food security and ensuring agricultural sustainability in the nation.Data mining finds widespread application in domains like business, science, finance, and marketing for the purpose of revealing concealed patterns and insights that guide decision-making and enhance diverse operational aspects. The typical process comprises stages such as data collection, data preprocessing, data transformation, model construction, and evaluation.This paper considers crop yield variations across states in India-related dataset like crop, state, cost of cultivation labour (hectare) a2+fl, cost of cultivation (hectare) c2, cost of production (quintal) c2, yield (quintal/ hectare). The machine learning approaches which is used to analysis and predict the dataset usingLinear Regression, Multilayer Perceptron, SMOreg, M5P, 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 yield variations across states in India, decision tree, correlation coefficient, and test statistics.

DOINUMBER:10.48047/NQ.2022.20.19.NQ99447 NEUROQUANTOLOGY2022;20(19):4850-4859

1. INTRODUCTION AND LITERATURE REVIEW

Exploring crop yield fluctuations among Indian states through the application of data mining and machine learning methods has the potential to yield valuable insights and facilitate the enhancement of agricultural practices and policies.Machine learning approaches encompass methods and strategies within the realm of artificial intelligence for constructing algorithms and models that empower computers to glean knowledge and make predictions or decisions from data.Data mining approaches encompass methods and techniques employed to delve into extensive datasets, reveal patterns, and extract valuable insights and knowledge from the data.

The potential of handheld LIBS for determining the mass fractions of major

nutrients (Ca, K, Mg, N, P) and trace nutrients (Mn, Fe) in soil. It also assessed other soil parameters like humus content, soil pH, and plant-available P. Various multivariate regression methods (PLSR, Lasso, and GPR) were used for calibration and prediction, with the best results achieved for Ca, K, Mg, and Fe. Lower concentrations of Mn and the limited data for N and P influenced the results. Soil parameters not directly tied to a single element, such as pH, were also predicted, with Lasso and GPR outperforming PLSR [1].

A system that incorporates agricultural data to suggest suitable crops using a voting-based ensemble classifier algorithm. The system aims to improve crop yield prediction, encourage crop rotation, and assist with farmer-friendly fertilization decisions. The system achieved an accuracy of approximately 92% [2].

Significant models for most targeted nutrients (S, P, B) can be produced with R-squares ranging from 40% to 85%. Nutrients such as Mn, Zn, Al, B, and Na were identified as critical for predicting crop yield based on a comparison with the OFRA field trial database. Improving prediction accuracy in Africa involves collecting more training samples, harmonizing measurement methods, and using more detailed covariates specific to the region [3].

Author explain aimed to establish a connection between element content in soils, plant leaves, and changes in selected chlorophyll a fluorescence parameters to detect plant stress early. Machine-learning methods, including principal component analysis, hierarchical k-means, and self-organizing maps, were used to analyze data. The results indicated patterns related to nutrient deficiency and chlorophyll fluorescence parameters. This approach can be informative and cost-effective for detecting plant stress [4].

A system that uses IoT devices to collect information on soil nutrient levels, temperature, season, soil type, fertilizer usage, and water pH. Data is analyzed using principal component analysis (PCA) and machine learning algorithms (LR, DT, RF) to forecast crop yield and recommend suitable fertilizers. The system aims to enhance crop production and support smart farming [5].

Data mining techniques were applied to a weather dataset, which was used to predict whether weather conditions were conducive to playing golf. Seven classification algorithms were used, with the Random Tree algorithm achieving the highest accuracy at 85.714% [6].

The paper outlines a smart farming system that utilizes data mining techniques, satellite information, soil testing reports, and clustering algorithms to make informed decisions based on weather changes, crop growth stages, water usage, and fertilizer recommendations. The system aims to increase agricultural productivity by managing farm operations effectively [7].

An intelligent soil nutrient and pH classification system that uses weighted voting ensemble deep learning. The system employs deep learning models (GRU, DBN, BiLSTM) and a weighted voting ensemble to classify soil nutrient and pH levels. Hyperparameter optimization is performed using the MRFO algorithm. The system demonstrated improved performance in soil nutrient and pH classification [8].

The article discusses the challenge and opportunities of Big Data in agriculture. It emphasizes the importance of big data characteristics, such as volume, velocity, variety, and veracity. Agricultural economists are highlighted as uniquely positioned to contribute to the research and outreach agenda on Big Data, addressing policy, farm management, supply chain, consumer demand, and sustainability issues [9].

Author suggest uses stochastic modeling and data mining approaches to assess groundwater levels, rainfall, population, food grains, and enterprises data. Data assimilation analysis is proposed to predict groundwater levels effectively. The study emphasizes the efficiency and accuracy of the approach [10][11].

The paper presents an analysis of chronic disease data using five classification algorithms. The M5P decision tree approach is identified as the best algorithm for building the model, surpassing 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.

2.4 M5P

M5P is a machine learning algorithm used for regression tasks. It is an extension of the decision tree-based model called M5, which Ross Quinlan developed. The M5 algorithm combines decision trees and linear regression to create more accurate and flexible regression models.M5P, specifically, stands for M5 Prime. It enhances the original M5 algorithm to improve its predictive performance. M5P uses a tree-based model to divide the data into subsets based on feature values recursively and then fits linear regression models to each of these subsets. The result is a piecewise linear regression model, where different linear regressions are used for other regions of the input feature space.

Steps involved in the M5P

Step 1. Building the initial decision tree (M5 model): Recursive Binary Splitting and Pruning (optional)
Step 2. Linear Regression Model: Leaf Regression Models and Model Parameters
Step 3. Piecewise Linear Regression: Piecewise Prediction
Step 4. Model Evaluation:Training and Testing.

2.5 Random 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)

2.6 Random 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):

2.7 REP 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

2.8 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 yield variations across states in Indiadataset include 6 parameters whichhavedifferent categories of data like crop, state, cost of cultivation labour (hectare) a2+fl,cost of cultivation (hectare) c2,cost of production (quintal) c2, yield (quintal/ hectare)[17]. Adetailed description of the parameters is mentioned in the following Table 1.

Table 1. Crop yield variations across states in Indiasampledataset

Crop	State	Cost of Cultivation Labour	Cost of Cultivation (Hectare) C2	Cost of Production (Quintal) C2	Yield (Quintal / Hectare)

		(Hectare) A2+FL			
GROUNDNUT	Tamil Nadu	22507.86	30393.66	2358	11.98
GROUNDNUT	Gujarat	22951.28	30114.45	1918.92	13.45
GROUNDNUT	Maharashtra	26078.66	32683.46	3207.35	9.33
MAIZE	Bihar	13513.92	19857.7	404.43	42.95
MAIZE	Karnataka	13792.85	20671.54	581.69	31.1
MAIZE	Rajasthan	14421.46	19810.29	658.77	23.56
MAIZE	Uttar Pradesh	15635.43	21045.11	1387.36	13.7

Table 2: Machine Learning Models with Correlation coefficient

ML Approaches	Correlation Coefficient
Linear Regression	0.9366
Multilayer Perceptron	0.9857
SMOreg	0.9301
M5P	0.9497
RandomForest	0.9716
RandomTree	0.9412
REPTree	0.9336

Table 3: Machine Learning Models with Mean Absolute Error and Root Mean Squared Error

ML Approaches	MAE	RMSE
Linear Regression	59.8411	91.5539
Multilayer Perceptron	17.5909	40.9509
SMOreg	40.5769	92.1798
M5P	32.5812	78.6659
RandomForest	22.1715	58.5198
RandomTree	29.9158	82.6235
REPTree	29.3952	88.8744

Table 4: Machine Learning Models with Relative Absolute Error (%) and Root Relative Squared Error (%)

ML Approaches	RAE (%)	RRSE (%)
Linear Regression	41.5729	37.0332
Multilayer Perceptron	12.2208	16.5645
SMOreg	28.1897	28.1897
M5P	22.6349	31.8201
RandomForest	15.4030	23.6711
RandomTree	20.7832	33.4209
REPTree	20.4215	35.9494

Table5: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Timetaken (seconds)

Linear Regression	0.1500
Multilayer Perceptron	0.3300
SMOreg	0.0700
M5P	0.0700
Random Forest	0.0800
Random Tree	0.0100
REP Tree	0.0100

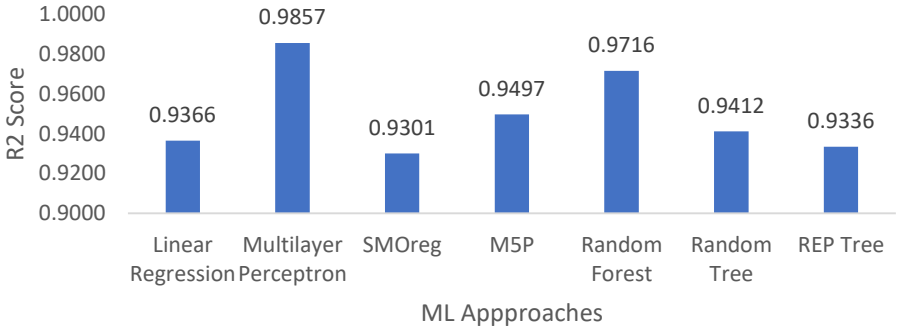

Fig. 1. R2 Score for Machine Learning Approaches

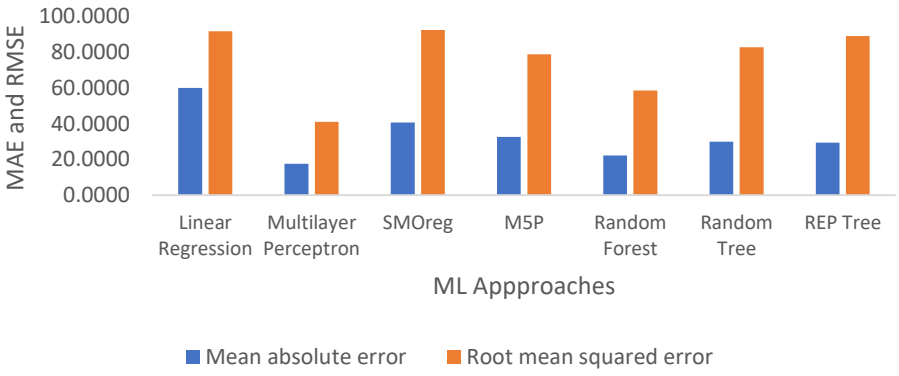

Fig. 2.Machine Learning Models with MAE and RMSE

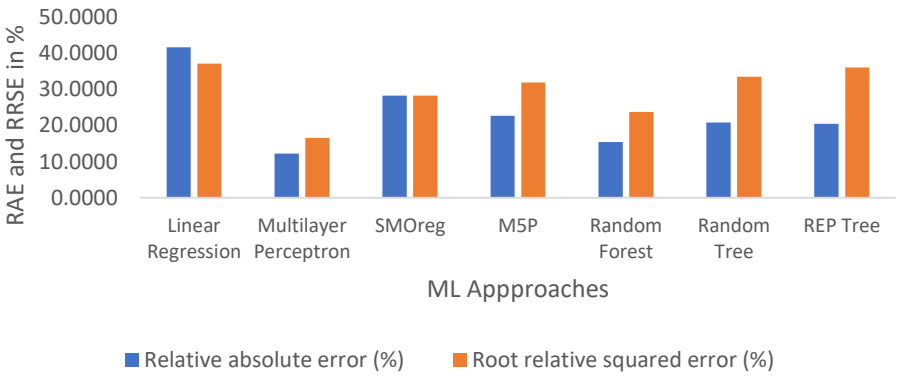

Fig. 3. Machine Learning Models with RAE (%) and RRSE (%)

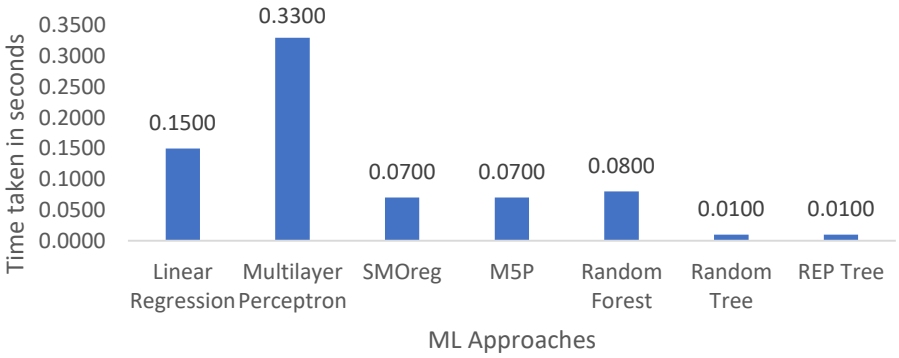

Fig. 4. Machine Learning Models and its Time Taken to Build the Model (Seconds)

4. RESULTS AND DISCUSSION

In Table 1, we elucidate six parameters encompassing various data categories, including crops, states, cultivation labor cost (per hectare) denoted as a2+fl, cultivation cost (per hectare) labeled as c2, production cost (per quintal) denoted as c2, yield (quintals per hectare), model price, and diverse agricultural product information. Based on the dataset, it becomes apparent that we employed six additional machine learning approaches, namely linear regression, multilayer perceptron, SMOreg, M5P, random forest, random tree, and REP tree, to unveil hidden patterns and determine which parameter exerts the most influence on future predictions. The results and numerical representations are presented across Table 1 to Table 5 and Figure 1 to Figure 4.

These findings are rooted in Equation 2, Table 2, and Figure 1, which facilitate the computation of the R2 score and correlation coefficient by comparing these six parameters. The numerical data indicates substantial variations among the parameters. In this context, the utilization of seven distinct machine learning approaches yields a strong positive correlation, approaching 0.9, particularly when examining yield (quintals per hectare).

We gauge model errors using Mean Absolute Error (MAE) as depicted in Equation 3. This analysis encompasses six machine-learning algorithms. Remarkably, all seven ML approaches exhibit superior error performance, with an average error value of approximately 17. The Root Mean Square Error (RMSE), which measures the disparity between predicted and actual values, is calculated using Equation 4. Once again, all ML approaches consistently demonstrate commendable error performance, averaging around 40. You can find detailed numerical representations in Table 3 and Figure 2.

Relative Absolute Error (RAE) quantifies accuracy through Equation 5, offering a percentage-based comparison between predicted and actual values. This study employs seven ML classification algorithms. Notably, linear regression stands out with the highest error rate in solving this problem, while the remaining six ML approaches exhibit superior performance with minimal error. This pattern is echoed in the Relative Root Square Error (RRSE), as demonstrated in the numerical representations within Table 4 and Figure 3.

Furthermore, time efficiency plays a pivotal role in machine learning approaches. As indicated in Table 5 and Figure 4, the multilayer perceptron emerges as the most time-consuming solution for addressing this challenge. In contrast, M5P, Random Tree, REP Tree, and Random Forests require the least time to build their respective models. Linear regression and the SMOreg approach also prove efficient in constructing models with minimal time requirements. These observations are consistent with the visual representations provided.

5. CONCLUSION AND FUTURE RESEARCH

In conclusion, it is imperative to address the limitations of our model, encompassing factors such as crop selection,

state-specific considerations, cultivation labor cost (per hectare) denoted as a2+fl, cultivation cost (per hectare) labeled as c2, production cost (per quintal) denoted as c2, and yield (quintals per hectare). Moreover, we must acknowledge data-related intricacies, particularly regarding yield (quintals per hectare), as well as model-specific variables that may contribute to potential underperformance. Additionally, computational constraints that might have influenced the model's development should be considered.Looking ahead, we propose several potential enhancements and avenues for future research. These include the exploration of additional data sources, an in-depth investigation of more effective algorithms and hyperparameters, and fine-tuning the model to enhance its overall performance. This research offers valuable insights for the Department of Agriculture and related stakeholders seeking to optimize the agricultural sector.

6. REFERENCE

[1]. Erler, A., Riebe, D., Beitz, T., Löhmannsröben, H.G. and Gebbers, R., 2020. Soil nutrient detection for precision agriculture using handheld laser-induced breakdown spectroscopy (LIBS) and multivariate regression methods (PLSR, Lasso and GPR). Sensors, 20(2), p.418.

[2]. Archana, K. and Saranya, K.G., 2020. Crop Yield Prediction, Forecasting and Fertilizer Recommendation using Voting Based Ensemble Classifier. SSRG Int. J. Comput. Sci. Eng, 7, pp.1-4.

[3]. Hengl, T., Leenaars, J.G., Shepherd, K.D., Walsh, M.G., Heuvelink, G., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E. and Wheeler, I., 2017. Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. Nutrient Cycling in Agroecosystems, 109(1), pp.77-102.

[4]. Kalaji, H.M., Bąba, W., Gediga, K., Goltsev, V., Samborska, I.A., Cetner, M.D., Dimitrova, S., Piszcz, U., Bielecki, K., Karmowska, K. and Dankov, K., 2018. Chlorophyll fluorescence as a tool for nutrient status identification in rapeseed plants. Photosynthesis Research, 136(3), pp.329-343.

[5]. Najeeb Ahmed, G. and Kamalakkannan, S., 2022. Developing an IoT-Based Data Analytics System for Predicting Soil Nutrient Degradation Level. In Expert Clouds and Applications (pp. 125-137). Springer, Singapore.

[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]. Chandak, P.P. and Agrawal, A.J., 2017. Smart farming system using data mining. International Journal of Applied Engineering Research, 12(11), pp.2788-2791.

[8]. Escorcia-Gutierrez, J., Gamarra, M., Soto-Diaz, R., Pérez, M., Madera, N. and Mansour, R.F., 2022. Intelligent agricultural modelling of soil nutrients and ph classification using ensemble deep learning techniques. Agriculture, 12(7), p.977.

[9]. Coble, K.H., Mishra, A.K., Ferrell, S. and Griffin, T., 2018. Big data in agriculture: A challenge for the future. Applied Economic Perspectives and Policy, 40(1), pp.79-96.

[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+%281%29.csv