ANALYSIS AND PREDICTION FOR AGRICULTURE DATASET WITH WEATHER CONDITIONS AND SOIL NUTRIENTS LEVEL USING MACHINE LEARNING CLASSIFICATION APPROACHES

S. Dhanavel[1*], A. Murugan[2]

Article History: Received: 16.01.2023 **Revised:** 01.03.2023 **Accepted:** 14.04.2023

Abstract

Machine learning classification constitutes a subset of both artificial intelligence and data science. It involves training models to categorize or classify data points into predefined classes based on their distinct attributes or features. The core objective of classification is to empower the model to discern patterns and correlations within the data, enabling it to correctly assign new and unseen data points to their appropriate classes. This research, taking into consideration agriculture dataset with soil nutrient-related parameters like nitrogen (N), phosphorus (P), and potassium (K) in soil, and weather-related pieces of information like temperature, humidity, pH, rainfall, and their class label agriculture products. In this paper, we utilize the machine learning approaches to find the future prediction with accuracy parameters using logistic regression, multilayer perceptron, simple logistic, SMO, decision stump, hoeffding tree, J48, LMT, random forest, random tree, and REPtree. Numerical illustrations are also provided to prove the results and discussion.

Keywords: Data Mining, Machine Learning, Decision Tree, Classifications, Micro and Macro Nurtrients.

[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
Email: [1*]dhanavel2008@gmail.com, [2]drmuruganapcs@gmail.com

[*]Corresponding Authors: S. Dhanavel
[1*]Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India
Email: [1*]dhanavel2008@gmail.com

DOI: 10.31838/ecb/2023.12.1.571

1. Introduction

In classification scenarios, the input data encompasses a collection of features representing individual data points' quantifiable traits or characteristics. These features serve as the basis for the model to generate predictions about the category to which each data point belongs. The classes, in turn, signify the discrete categories or labels that the model strives to allocate to the data points. Constructing a classification model encompasses several pivotal stages: data collection and preparation, feature extraction and selection, model selection, model training, model evaluation, hyperparameter tuning, and model deployment. Classification finds extensive utility in diverse domains, encompassing image recognition, natural language processing, fraud detection, medical diagnosis, sentiment analysis, and more. Its effectiveness hinges on factors such as the caliber and extent of the training data, algorithmic selection, and precise parameter tuning to attain accurate and dependable predictions. The term "Correctly Classified Instances" represents a concept used in evaluating machine learning models to assess their performance. Calculating Correctly Classified Instances is part of the overall model evaluation process. Incorrectly Classified Instances refer to the instances or data points in a machine learning model's evaluation or testing dataset that the model classifies incorrectly. In simpler terms, these are instances where the model's predictions do not align with the actual target or ground truth values.

Literature Review

The researchers explain various concepts related to the micro and macro nutrients and the authors provides review of various data mining techniques used on agriculture soil dataset for fertilizer recommendation. Mainly I focused on various soil parameters like Fe, S, Zn, Cu, N and Ph value etc. In this survey, we also describe some Agriculture problems that can be solved by using data mining techniques [1]. The research presents a brief analysis of crop yield prediction using Multiple Linear Regression (MLR) technique and Density based clustering technique for the selected region i.e. East Godavari district of Andhra Pradesh in India [2]. Research aimed to assess these new data mining techniques and apply them to the various variables consisting in the database to establish if meaningful relationships can be found [3]. The role of data mining in perspective of soil analysis in the field of agriculture and also confers about several data mining techniques and their related work by several authors in context to soil analysis domain. The data mining techniques are of very up-to-the-minute in the area of soil analysis [4]. The potential of handheld LIBS for the determination of the total mass fractions of the major nutrients Ca, K, Mg, N, P and the trace nutrients Mn, Fe was evaluated. Additionally, other soil parameters, such as humus content, soil pH value and plant available P content, were determined. Since the quantification of nutrients by LIBS depends strongly on the soil matrix, various multivariate regression methods were used for calibration and prediction. These include partial least squares regression (PLSR), least absolute shrinkage and selection operator regression (Lasso), and Gaussian process regression (GPR). The best prediction results were obtained for Ca, K, Mg and Fe. The coefficients of determination obtained for other nutrients were smaller. This is due to much lower concentrations in the case of Mn, while the low number of lines and very weak intensities are the reason for the deviation of N and P. Soil parameters that are not directly related to one element, such as pH, could also be predicted. Lasso and GPR yielded slightly better results than PLSR. Additionally, several methods of data pretreatment were investigated [5]. Crop yield prediction and forecasting will increase the agricultural production. Periodical crop rotation will improve the soil fertility. This system supports farmer friendly fertilization decision making. The accuracy of this system was around 92% [6]. Predictions were produced for 15 target nutrients: organic carbon (C) and total (organic) nitrogen (N), total phosphorus (P), and extractable—phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg), sulfur (S), sodium (Na), iron (Fe), manganese (Mn), zinc (Zn), copper (Cu), aluminum (Al) and boron (B). Model training was performed using soil samples. An ensemble model was then created for each nutrient from two machine learning algorithms—random forest and gradient boosting, as implemented in R packages ranger and xgboost—and then used to generate predictions in a fully-optimized computing system. Cross-validation revealed that apart from S, P and B, significant models can be produced for most targeted nutrients (R-square between 40–85%). Further comparison with OFRA field trial database shows that soil nutrients are indeed critical for agricultural development, with Mn, Zn, Al, B and Na, appearing as the most important nutrients for predicting crop yield [7].

The machine learning systems uses IoT devices to gather information such as soil nutrient level, temperature of atmosphere, season of the atmosphere, soil type, fertilizer used and water pH level periodically. Further, the data gathered from the sensor will be passed to a principal component analysis (PCA), which are used to reduce features in order to obtain a better prediction level. Also, ML algorithms such as linear regression (LR), decision trees (DT) and random forest (RF) are implemented to forecast and classify the crop yield from the previous data based on soil nutrient degradation level and recommend suitable fertilizer for every particular crop [8]. Soil properties (including physical, chemical, and biological properties) and the characteristics of the spatial soil data are first introduced. Spatial clustering techniques are then summarized in five different categories. Soil data analysis using spatial clustering is reviewed in four categories of agricultural applications: agricultural production management zoning, comprehensive assessment of soil and land, soil and land classification, and correlation study for agro-ecosystem. The traditional clustering algorithms generally work well, and prototype-based clustering methods are more preferred in practice. Some machine learning models can be further introduced into the spatial clustering algorithms for better accommodation to various characteristics of soil dataset [9]. RF is a familiar machine learning decision tree algorithm that belongs to supervised learning methods. In these approaches working principles based on classification and regression. RF is generally called ensemble learning, which is used to combine different classifiers to solve various problems with enhanced performance of the model. The Random Forests classifier compared to others is the best classifier for capable of precisely classifying the huge amount of data. RF decision tree approaches mainly focused learning procedure for classification and regression methods, it will be creating many decision trees and level of the tree at training time for outputs the class with classes output from single trees [10]. The destinations have been assented of solidness in paddy advancement and to expand the development of creation in a maintainable way to meet the nourishment prerequisite for the developing populace. In any farming fields, it for the most part, happens that at whatever point the choices in regards to different methodologies of arranging is viewed as, for example, season-wise rainfall, region, production and yield rate of principal crops, and so forth. In this paper, it is proposed to discover the forecast level of concentration in paddy improvement for different years of time series data utilizing stochastic model approach. Numerical examinations are outlined to help the proposed work [11] In the recent times, there has been an increasing demand for efficient strategies in the field of data assimilation about groundwater. Data mining process is a discovery of hiding information that utilizes the prediction efficiently by stochastic sensing concept. This paper proposes an efficient assessment of groundwater level, rainfall, population, food grains and enterprises dataset by adopting stochastic modeling and data mining approaches. Firstly, the novel data assimilation analysis is proposed to predict the groundwater level effectively. Experimental results are done and the various expected ground water level estimations indicate the sternness of the approach [12].

2. Methods and Background

Kappa statistic: The Kappa statistic, also called Cohen's Kappa or simply Kappa, is a statistical metric utilized to assess the level of agreement between two or more raters or classifiers when assigning categorical ratings or labels to items. It goes beyond considering agreement by chance alone. The Kappa statistic is represented on a scale from -1 to 1. A Kappa value of -1 signifies perfect disagreement between the raters or classifiers. A Kappa value of 0 indicates agreement that is no better than chance. A Kappa value of 1 implies perfect agreement between the raters or classifiers. The calculation of Kappa employs the formula:

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e}$$

$$P_o = \frac{\text{Number of items with agreement}}{\text{Total number of items}}$$

$$P_e = \sum \frac{\text{Total count in row } \times \text{Total count in column}}{\text{Total number of items}}$$

Where, P_o denotes the observed agreement, i.e., the proportion of items on which raters or classifiers agree. P_e represents the expected agreement, i.e., the agreement expected by chance.

Logistic Regression

Logistic Regression is a statistical method used for binary classification, which means it's used to predict the probability of an observation belonging to one of two classes (usually labeled as 0 and 1). It's a type of regression analysis that's particularly suited for categorical outcome variables. The formula for logistic regression involves the logistic function (also known as the sigmoid function) to transform the linear combination of input features into a value between 0 and 1, representing the predicted probability of the positive class. The formula is as follows:

$$P\left(Y = \frac{1}{X}\right) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

P(Y=1/X) is the probability that the dependent variable Y is the binary outcome equal to 1 given the input features $X_1 + X_2 + \cdots + X_n$. e is the base of the natural logarithm. $\beta_0 + \beta_1 + \cdots + \beta_n$ are the coefficients that need to be estimated from the training data. $X_1 + X_2 + \cdots + X_n$. are the input features. Logistic regression is often implemented using optimization algorithms to find the best-fitting coefficients that minimize the prediction error.

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:

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.

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.

Output Layer: This layer produces the network's final output. The number of neurons in the output layer depends on the problem type.

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.

Decision Stump

A Decision Stump is a simple machine learning model that serves as a weak learner, often used in ensemble learning methods like boosting. It's a basic model that makes decisions based on a single feature (input) and a threshold value. Despite its simplicity, when combined with other decision stumps or more complex models, decision stumps can contribute to building stronger predictive models. Here's how a Decision Stump works:

Step 1. Input Feature: A Decision Stump focuses on a single feature from the input data.

Step 2. Threshold: The model selects a threshold value for the chosen feature.

Step 3. Prediction: For each data point, the Decision Stump compares the value of the chosen feature with the threshold.

Step 4. Decision Rule: The decision rule of a Decision Stump can be expressed as follows:

a. If feature value < threshold, predict one class.

b. If feature value >= threshold, predict the other class.

Hoeffding Tree

A Hoeffding Tree, also known as VFDT (Very Fast Decision Tree) or Incremental Decision Tree, is a machine learning algorithm designed for online, incremental learning on streaming data. It's beneficial when you have large volumes of data that are continuously arriving and you want to update your model in real-time without retraining the entire dataset. Here's a simplified overview of how the Hoeffding Tree algorithm works:

Step 1. **Initialization:** Start with an empty tree.

Step 2. **Data Arrival:** As new data instances arrive, they update the decision tree. Instead of storing and processing all the data, only a small random sample is used for making decisions.

Step 3. **Splitting Nodes:** When a node reaches a threshold in terms of the number of instances it has seen, it evaluates the quality of its current split.

Step 4. **Leaf Node Prediction:** If a split is unnecessary or a decision is made to stop growing the tree, the node becomes a leaf node and predicts the majority class of instances seen so far.

Step 5. **Adaptation:** As more data arrives, the tree may adapt by adjusting its structure based on the new information.

J48

J48, also known as C4.5, is a popular decision tree algorithm used for classification tasks in machine learning and data mining. It was developed by Ross Quinlan and is an extension of the earlier ID3 (Iterative Dichotomiser 3) algorithm. J48 is widely used due to its effectiveness, ease of use, and ability to handle both categorical and numerical attributes. Here are the key features and steps of the J48 algorithm:

Step 1. Attribute Selection: J48 uses a top-down, recursive approach to build the decision tree.

Step 2. Splitting Nodes: Once an attribute is selected, the dataset is split into subsets based on the attribute's values.

Step 3. Recursion: The algorithm then recursively applies the same process to each subset.

Step 4. Pruning: After the tree is fully grown, J48 performs pruning to remove branches that do not contribute significantly to the predictive accuracy.

Step 5. Handling Missing Values: J48 can handle missing attribute values by distributing instances with missing values to all branches proportionally based on the distribution in the training data.

Step 6. Post-Pruning: J48 also supports post-pruning, which involves removing branches from the tree after it has been fully constructed.

Step 7. Leaf Node Prediction: Each leaf node of the tree corresponds to a class label.

LMT

LMT (Logistic Model Trees) is a machine learning algorithm that combines decision trees with logistic regression to create a hybrid model for classification tasks. It aims to harness the strengths of both decision trees and logistic regression, mitigating their individual weaknesses. LMT was introduced as an alternative to traditional decision trees and has shown promise in improving predictive performance and interpretability. Here's how the LMT algorithm works:

Step 1. Decision Tree Generation: LMT starts by constructing a decision tree using a top-down, recursive approach like traditional decision trees.

Step 2. Leaf Node Transformation: Unlike regular decision trees, LMT does not assign class labels directly to the leaf nodes.

Step 3. Predictions: When a new instance is presented to the LMT model, it traverses the decision tree to determine the appropriate leaf node.

Random Forest

Random Forest is a powerful ensemble learning algorithm used for both classification and regression tasks. It's based on the concept of bagging (Bootstrap Aggregating) and utilizes multiple decision trees to create a robust and accurate predictive model. Here's how the Random Forest algorithm works:

Step 1. Bootstrapped Sampling: The algorithm starts by creating multiple subsets of the training data through random sampling with replacement.

Step 2. Random Feature Selection: When building each decision tree, Random Forest further introduces randomness by considering only a subset of the available features at each split.

Step 3. Decision Tree Construction: For each bootstrapped dataset, a decision tree is constructed.

Step 4. Voting or Averaging: For classification tasks, the predictions from each decision tree are combined through majority voting; for regression tasks, they are averaged.

Random Tree

A "Random Tree" could refer to different things depending on the context. It might refer to a decision tree that has been built using some form of randomness, or it could be a term used in a specific domain or framework. Without more context, it's challenging to provide a precise answer. However, I can offer a couple of interpretations that might be relevant:

Step 1. Randomized Decision Tree: A Random Tree might be referring to a decision tree constructed using randomness, similar to how Random Forest uses random sampling of data and features.

Step 2. Specific Framework: Depending on your machine learning or data analysis framework, "Random Tree" could be a specific term or concept introduced within that framework.

REPTree

REPTree, short for "Reduced Error Pruning Tree," is a decision tree algorithm primarily used for classification tasks in machine learning. It is designed to create decision trees while incorporating a reduced-error pruning technique to avoid overfitting. The algorithm was introduced as a part of the WEKA machine learning software. Here's how the REPTree algorithm works:

Step 1. Tree Construction: REPTree follows a recursive approach to build a decision tree. It starts by selecting the best attribute to split the data based on metrics like information gain or gain ratio.

Step 2. Recursive Splitting: The algorithm examines potential attribute splits at each node and chooses the one that maximizes the selected splitting criterion.

Step 3. Reduced Error Pruning: After the tree is fully grown, REPTree performs reduced-error pruning to eliminate branches that do not contribute significantly to the tree's accuracy.

Step 4. Prediction: Once the tree is pruned, it can be used for making predictions.

Numerical Illustrations

Dataset: The crop recommendation system contains various nutrient-related information on the levels of nitrogen (N), phosphorus (P), and potassium (K) in soil, as well as weather-related pieces of information like temperature, humidity, pH, rainfall, and their class variable namely agriculture products mentioned the impact on the growth of 22 crops. The dataset [13] can be used to make data-driven recommendations for achieving optimal nutrient and environmental conditions to improve crop yield. The corresponding dataset indicates 2200 instances and eight parameters. Table 1 shows only some significant examples like rice, maize, black gram, pomegranate, banana, coconut, and cotton.

Table 1: Dataset for crop recommendation with weather and nutrients

N (kg/ha)	P (kg/ha)	K (kg/ha)	Temperature (Celsius)	Humidity (%)	pH (Value)	Rainfall (mm)	Agriculture Products
90	42	43	20.8797	82.0027	6.5030	202.9355	rice
85	58	41	21.7705	80.3196	7.0381	226.6555	rice
60	55	44	23.0045	82.3208	7.8402	263.9642	rice
74	35	40	26.4911	80.1584	6.9804	242.8640	rice
78	42	42	20.1302	81.6049	7.6285	262.7173	rice
71	54	16	22.6136	63.6907	5.7499	87.7595	maize
61	44	17	26.1002	71.5748	6.9318	102.2662	maize
80	43	16	23.5588	71.5935	6.6580	66.7200	maize
73	58	21	19.9722	57.6827	6.5961	60.6517	maize
61	38	20	18.4789	62.6950	5.9705	65.4384	maize
56	79	15	29.4844	63.1992	7.4545	71.8909	blackgram
25	62	21	26.7343	68.1400	7.0401	67.1510	blackgram
42	61	22	26.2727	62.2881	7.4187	70.2321	blackgram
42	73	25	34.0368	67.2111	6.5019	73.2357	blackgram
44	58	18	28.0364	65.0660	6.8144	72.4951	blackgram
2	24	38	24.5598	91.6354	5.9229	111.9685	pomegranate
6	18	37	19.6569	89.9370	5.9376	108.0459	pomegranate
8	26	36	18.7836	87.4025	6.8048	102.5185	pomegranate
37	18	39	24.1470	94.5111	6.4247	110.2317	pomegranate
20	27	41	20.5134	92.5168	5.7001	110.5764	pomegranate
91	94	46	29.3679	76.2490	6.1499	92.8284	banana
105	95	50	27.3337	83.6768	5.8491	101.0495	banana
108	92	53	27.4005	82.9622	6.2768	104.9378	banana
86	76	54	29.3159	80.1159	5.9268	90.1098	banana
80	77	49	26.0543	79.3965	5.5191	113.2297	banana
2	40	27	29.7377	47.5489	5.9546	90.0959	mango
39	24	31	33.5570	53.7298	4.7571	98.6753	mango
21	26	27	27.0032	47.6753	5.6996	95.8512	mango
25	22	25	33.5615	45.5356	5.9774	95.7053	mango
20	19	35	34.1772	50.6216	6.1139	98.0069	mango
18	30	29	26.7627	92.8606	6.4200	224.5904	coconut
37	23	28	25.6129	94.3139	5.7401	224.3207	coconut
13	28	33	28.1301	95.6481	5.6870	151.0762	coconut
2	21	35	25.0289	91.5372	6.2937	179.8249	coconut
10	18	35	27.7980	99.6457	6.3820	181.6942	coconut
133	47	24	24.4023	79.1973	7.2313	90.8022	cotton
136	36	20	23.0960	84.8628	6.9254	71.2958	cotton
104	47	18	23.9656	76.9770	7.6334	90.7562	cotton
133	47	23	24.8874	75.6214	6.8274	89.7605	cotton
126	38	23	25.3624	83.6328	6.1767	88.4362	cotton

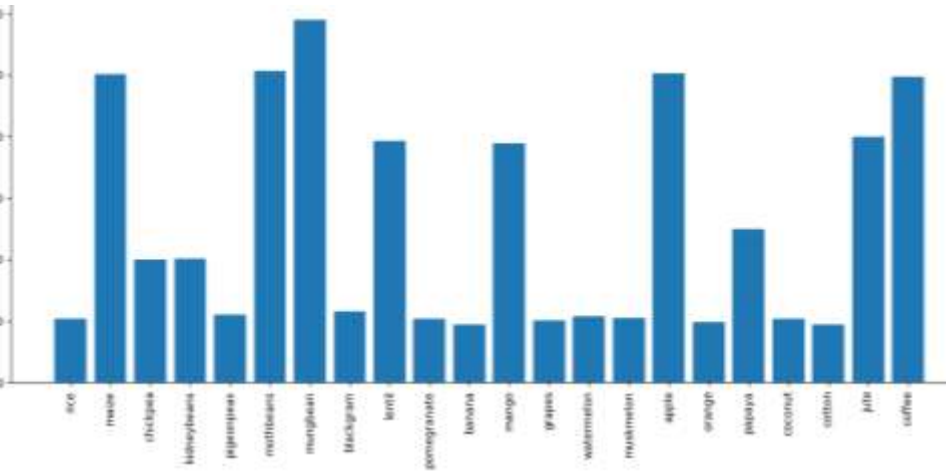

Fig. 1. Data analysis for agriculture products with nitrogen (N)

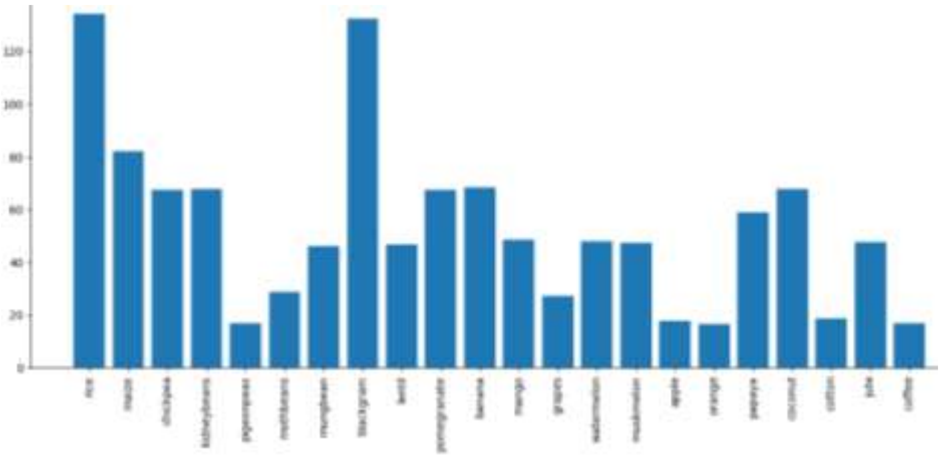

Fig. 2. Data analysis for agriculture products with phosphorus (P)

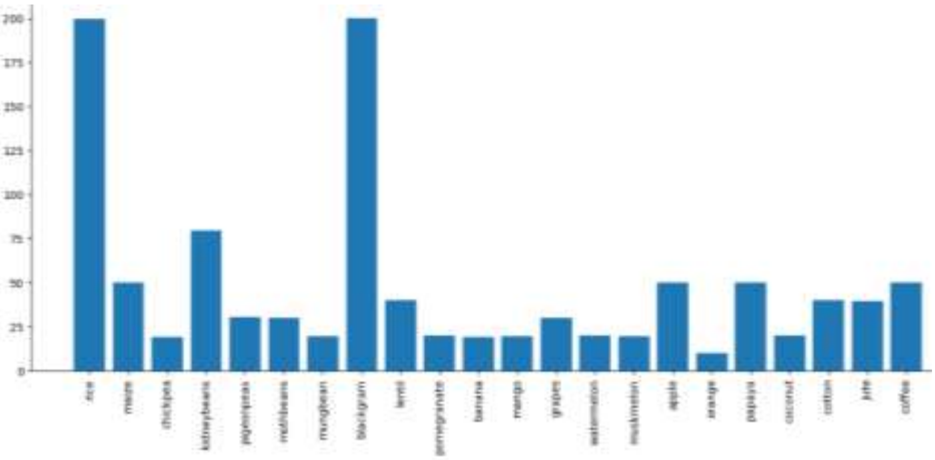

Fig. 3. Data analysis for agriculture products with potassium (P)

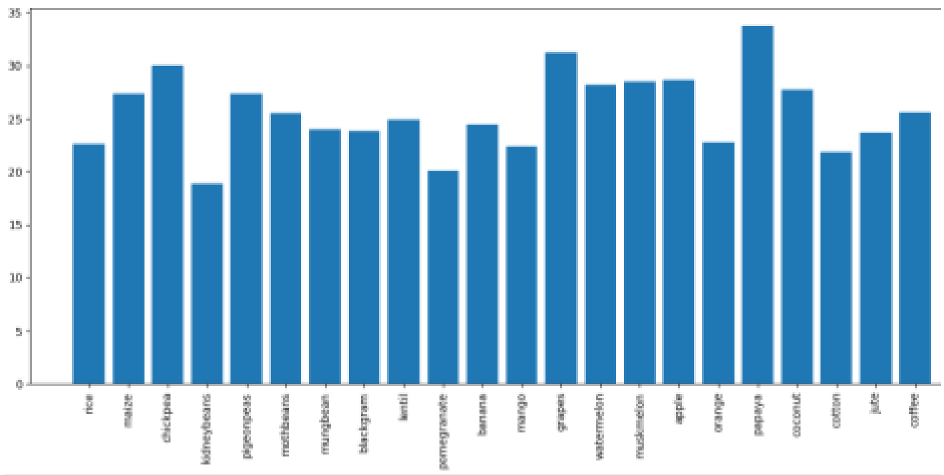

Fig. 4. Data analysis for agriculture products with temperature

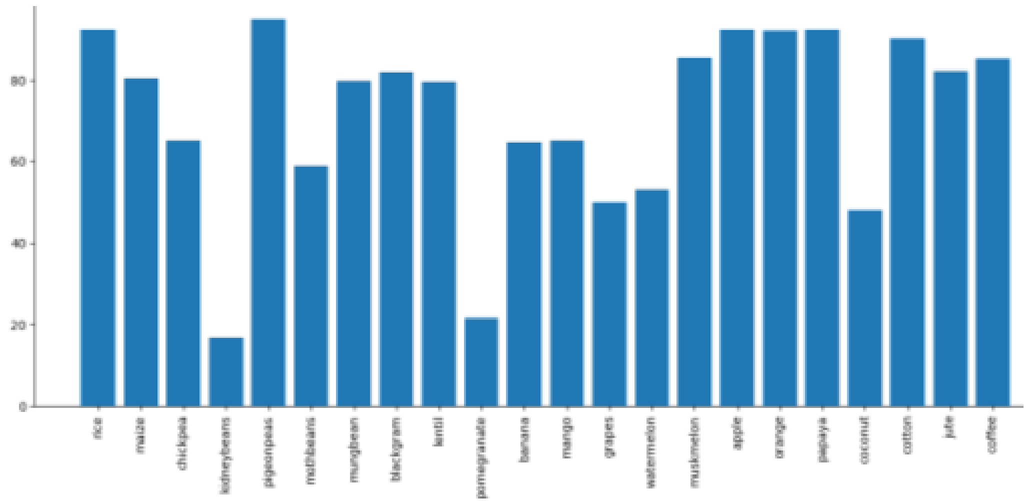

Fig. 5. Data analysis for agriculture products with humidity

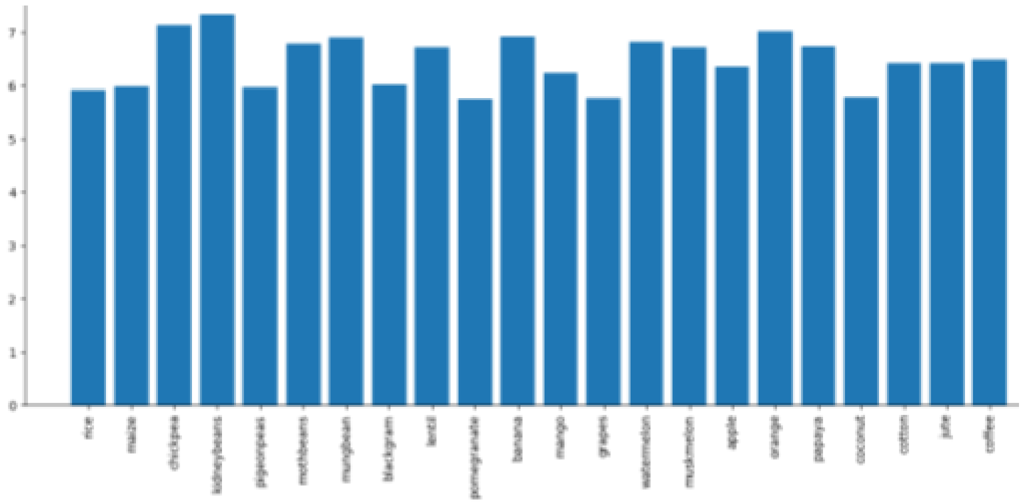

Fig. 6. Data analysis for agriculture products with pH value

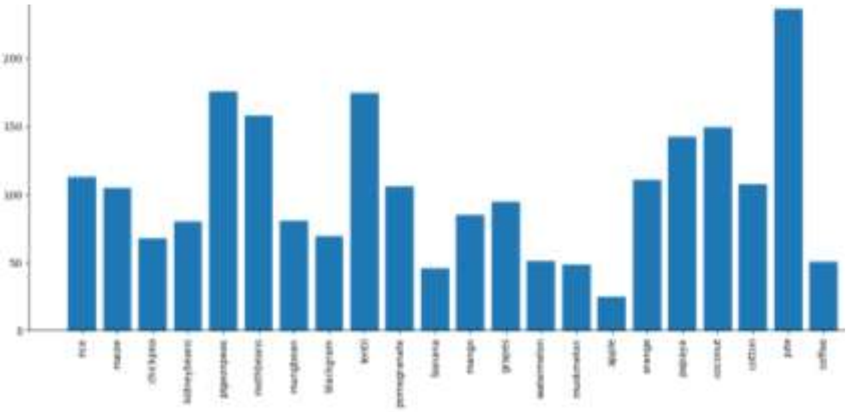

Fig. 7. Data analysis for agriculture products with rainfall

Table 2: Number of correctly classified and incorrectly classified instances and their percentages

Machine Learning Approaches	Correctly Classified Instances	Incorrectly Classified Instances	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)
Logistic Regression	2154	46	97.9091	2.0909
Multilayer Perceptron	2161	39	98.2273	1.7727
Simple Logistic	2159	41	98.1364	1.8636
SMO	2141	59	97.3182	2.6818
Decision Stump	199	2001	9.0455	90.9545
Hoeffding Tree	2189	11	99.5000	0.5000
J48	2181	19	99.1364	0.8636
LMT	2156	44	98.0000	2.0000
Random Forest	2186	14	99.3636	0.6364
Random Tree	2165	35	98.4091	1.5909
REPTree	2143	57	97.4091	2.5909

Table 3: Machine learning approaches with kappa statistic, accuracy performance, and time taken to build the models

Machine Learning Approaches	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)	Time taken (seconds)
Logistic Regression	0.9781	0.0021	0.0401	2.4683	19.2554	43.1900
Multilayer Perceptron	0.9814	0.0040	0.0350	4.5682	16.7902	36.3100
Simple Logistic	0.9805	0.0025	0.0348	2.9376	16.6870	4.6700
SMO	0.9719	0.0827	0.2000	95.2515	96.0338	1.1200
Decision Stump	0.0471	0.0828	0.2035	95.4321	97.6998	0.0400
Hoeffding Tree	0.9948	0.0009	0.0198	1.0304	9.5254	0.4600
J48	0.9910	0.0009	0.0274	1.0702	13.1765	0.2000
LMT	0.9790	0.0025	0.0368	2.8747	17.6742	18.3900
Random Forest	0.9933	0.0026	0.0220	2.9395	10.5624	0.9900
Random Tree	0.9833	0.0014	0.0380	1.6667	18.2574	0.0200
REPTree	0.9729	0.0030	0.0454	3.4587	21.7884	0.0700

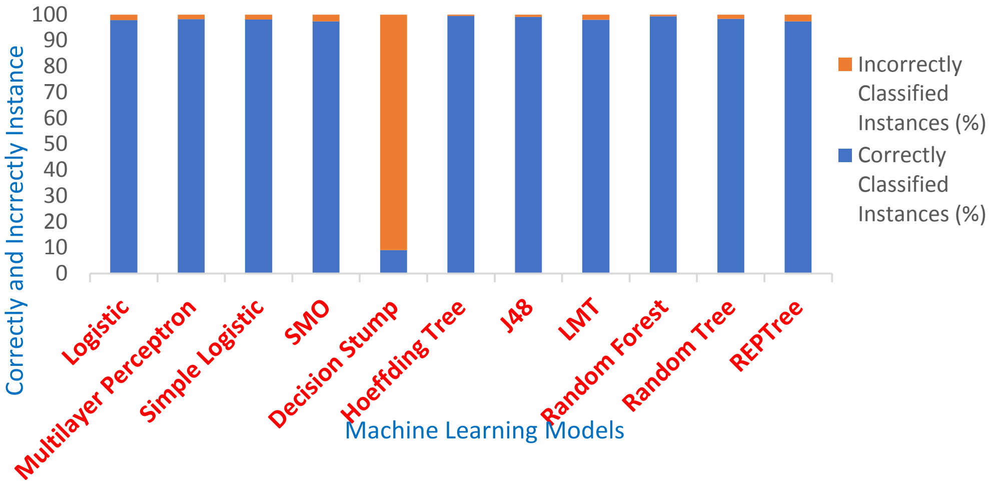

Fig. 8. Correctly Classified and Incorrectly Classified Instances (%)

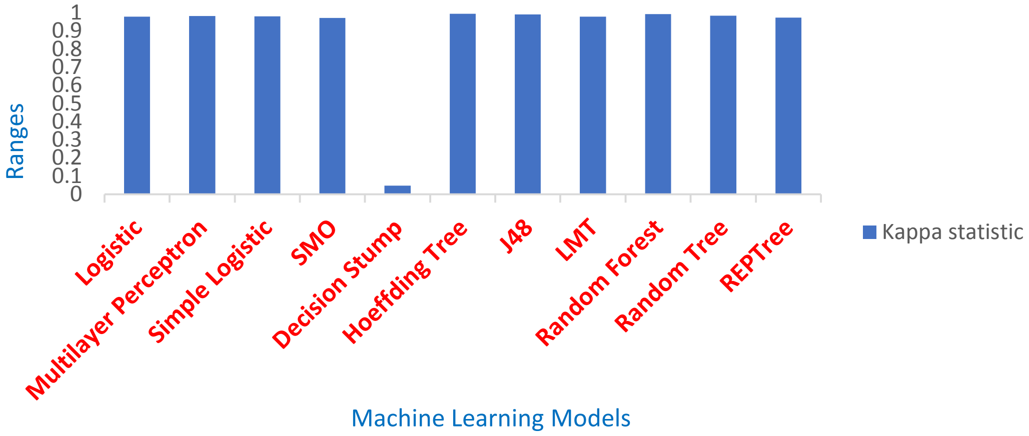

Fig. 9. Machine Learning Approaches with Kappa Statistic

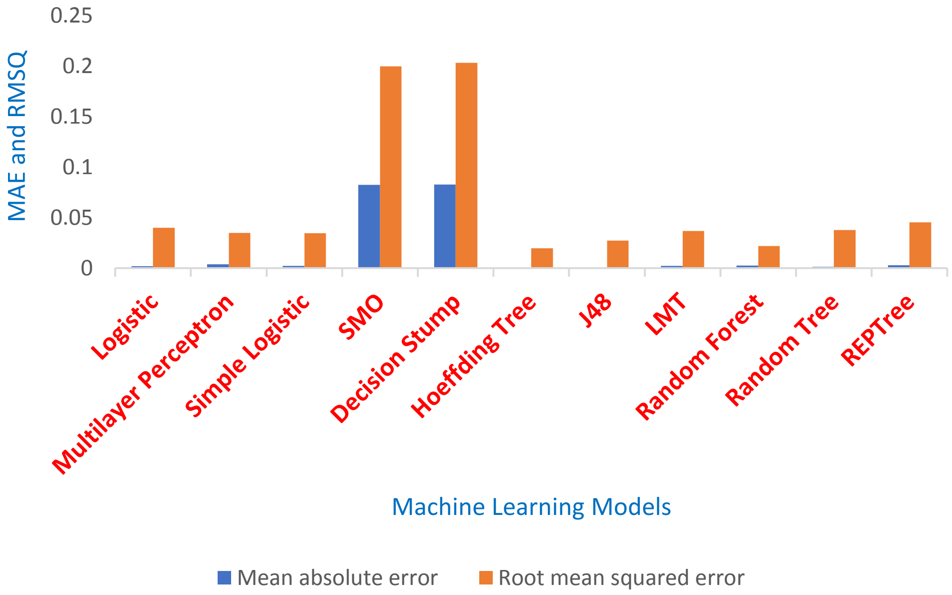

Fig. 10. Mean Absolute Error and Root Mean Squared Error

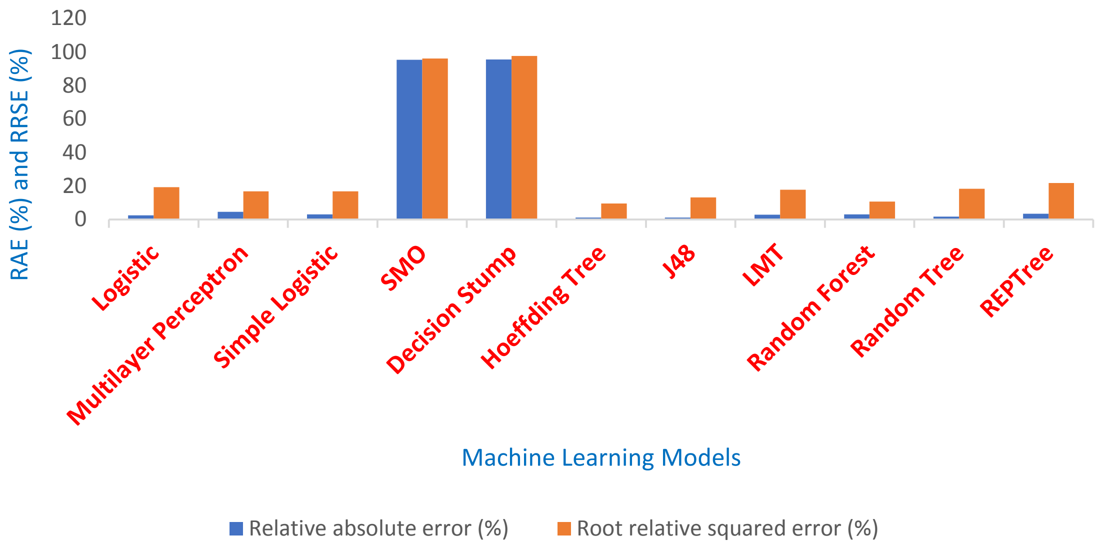

Fig. 11. Relative Absolute Error and Root Relative Squared Error (%)

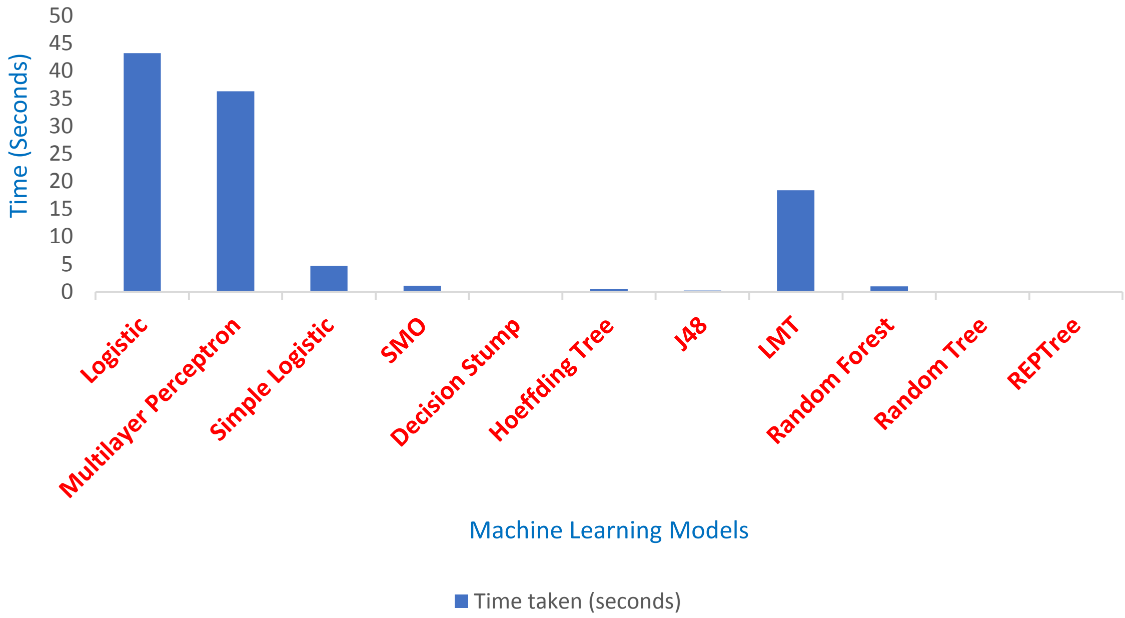

Fig. 12. Time taken to Build the Machine Learning Models (in seconds)

3. Result and Discussion

This research focused on crop recommendation systems, including nutrient levels and weather conditions. levels of nitrogen (N), phosphorus (P), and potassium (K) in soil, as well as weather-related pieces of information like temperature, humidity, pH, rainfall, and their class variable, namely agriculture products, mentioned the impact on the growth of 22 crops. The related sample dataset is indicated in Table 1. Data analysis is used to visualize various nutrient levels with agriculture products, namely agriculture products with nitrogen (N), phosphorus (P), and potassium (P). The related results and discussions are shown in Fig. 1, Fig. 2, and Fig. 3. Similarly, agriculture products with temperature, humidity, pH values, and rainfall are shown in Fig. 4, Fig. 5, Fig. 6, and Fig. 7. Based on numerical illustrations, Table 2 indicates the number of correctly and incorrectly classified instances and their percentages. In this research, 11 machine learning algorithms, namely decision stump return, correctly classified instances at 9% and poorly classified cases at 90%. The remaining 10 ML classification algorithms return the correctly classified instance range between 97% to 99.5%.

The related results and discussions are shown in Table 2 and Fig. 8.

The Kappa statistic is represented on a scale from -1 to 1. Kappa value of -1 signifies perfect disagreement between the raters or classifiers. In this case, except for the decision stump, the remaining ML algorithms return the Kappa statistics values of nearly 1, which means they accept the arguments. The related results and discussions are shown in Table 3 (Column 2) and Fig. 9. Mean Absolute Error (MAE) is a metric commonly used to measure the accuracy of a predictive model, particularly in the context of regression tasks. In this case, all the weather and nutrient parameters have nearly 0 error rates for using MAE test statistics. Similarly, the Root Mean Squared Error (RMSE) is another standard metric used to evaluate the performance of predictive models, particularly in regression tasks. Like Mean Absolute Error (MAE), RMSE measures the accuracy of predictions. In these cases, the error rate is also nearly 0. Both the MAE and RMSE also returned almost 0. Table 3 (Columns 3 and 4) and Fig. 10 show the related results and discussions. Relative Absolute Error (RAE) and Root Relative Absolute Error (RRAE) are variations of the absolute error metrics (MAE and RMSE) that consider the scale of the actual values in the evaluation of a predictive model. In these cases, SMO and Decision stump return the error rate is high, and the remaining ML classification approaches return low error statistics. Table 3 (Columns 5 and 6) and Fig. 11 show the corresponding results and discussion. Every research considers time as a primary factor. In this case, the time taken to build the machine-learning approaches returns acceptable time for processing the outcomes. Logistic regression and multilayer perception take maximum time compared to the other ML classification approaches. Table 3 (Column 7) and Fig. 12 show the diagrammatical representation.

4. Conclusion and Further Studies

In summary, our research endeavors aimed to investigate the impact of this research, which is beneficial for the former and the Department of Agriculture and for awareness of the weather conditions and nutrient levels implications for agriculture development. In this research taking consideration in to 8 parameters and 11 machine learning classification approaches. Based on results and discussion, most ML approaches return better results with test statistics. However, it's essential to acknowledge the limitations of our study. Our analysis was constrained by the available dataset's granularity, which occasionally hindered a more nuanced exploration of certain factors. Furthermore, the study primarily focused on specific predictions for all the parameters. urban context may not be fully generalizable to diverse geographical and cultural settings. In the future, consider other machine learning approaches with test statistics to improve the accuracy and reduce the time complexity.

5. References

6. Jethva, J.M., Gondaliya, N. and Shah, V., 2018. A review on data mining techniques for fertilizer recommendation. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, IJSRCSEIT, 3(1).
7. Ramesh, D. and Vardhan, B.V., 2015. Analysis of crop yield prediction using data mining techniques. International Journal of research in engineering and technology, 4(1), pp.47-473.
8. Raorane, A.A. and Kulkarni, R.V., 2012. Data Mining: An effective tool for yield estimation in the agricultural sector. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 1(2), pp.1-4.
9. Palepu, R.B. and Muley, R.R., 2017. An analysis of agricultural soils by using data mining techniques. Int. J. Eng. Sci. Comput, 7(10).
10. 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.
11. 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.
12. 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.
13. 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.
14. Gao, H., 2021, January. Agricultural Soil Data Analysis Using Spatial Clustering Data Mining Techniques. In 2021 IEEE 13th International Conference on Computer

Research and Development (ICCRD) (pp. 83-90). IEEE.

15. Rajesh, P. and Karthikeyan, M. 2017. A comparative study of data mining algorithms for decision tree approaches using WEKA tool. Advances in Natural and Applied Sciences, 11(9), pp. 230-243.
16. Rajesh, P. and M. Karthikeyan, 2019. Prediction of Agriculture Growth and Level of Concentration in Paddy - A Stochastic Data Mining Approach. Advances in Intelligent Systems and Computing, Springer, Vol. 750, pp.127-139.
17. Rajesh, P., Karthikeyan, M. and Arulpavai, R., 2019, December. Data mining approaches to predict the factors that affect the groundwater level using stochastic model. In AIP Conference Proceedings, (Vol. 2177, No. 1). AIP Publishing.
18. https://www.kaggle.com/code/ysthehurricane/crop-recommendation-system-using-lightgbm/input