[1]A. Zakiuddin Ahmed,
[2]Dr. T. Abdul Razak

Soil Classification using the Stacking Ensemble Learning Technique for Crop Agronomy

Abstract: - The main objective of the research work is to implement stacking ensemble learning techniques for classification of the soil types of a given region to determine the most appropriate crop to cultivate using proper irrigation systems and suitable fertilizers. Soil is a major factor in crop agronomy, and India has a various types of soil including red, black, sandy, alluvial, forest and mountain soil. The agricultural yield mainly relies on the type of soil, season (Kharif, Rabi and Zaid), irrigation method (sprinkle, surface, drip) and appropriate fertilizers. The proposed approach is being used to classify and analyze the soil of a particular region with the intent to enhance the yield of agriculture. It also helps agronomists in forecasting which crop might be preferable to cultivate and also suggesting the suitable fertilizers and irrigation systems (avoid wastage of water) to be adopted. In this research paper, different types of soil are classified (regarding cultivation) through our proposed Stacking Ensemble Learning (classification technique) by using artificial intelligence and machine learning techniques. The resulting decision tree serve as valuable tool for farmers and agricultural practitioners to understand the optimize crop selection based on prevalent soil conditions. The proposed method uses three base classifiers (KNN, Random forest and XGBoost) and a meta_learner (AdaBoost) to create an Ensemble model. Compared to existing works (SVM, KNN, Decision tree and Bayesian Model algorithms), the soil classification result using our proposed stacking ensemble learning approach to decision tree is more accurate.

Keywords: Ensemble Learning, stacking, soil dataset, machine learning, decision tree.

I. Introduction:

Agriculture is the essential entity for our country’s economy because large sections of the population are living in rural areas, and they are completely dependent on agriculture for their livelihood. Machine Learning techniques are being used in analyzing a large dataset; establish classification models, to discover precious patterns. Machine learning is built on the techniques of statistics and computer science that enables computers in constructing models, from sample data, to computerize the processes of making decisions based on the given soil dataset. Machine Learning learns the models in existing data, and then uses a model that recognizes a pattern in new data and makes the predictions. Data preprocessing techniques are applied to datasets to prepare the data for the classification process, which has to be carried out efficiently.

Decision Tree is the influential tools used for the classification of large datasets. It belongs to the supervised classification family which facilitates to interpret the any given dataset easily. The primary objective to build a training model with the aid of decision tree that will be used to predict the classes accurately. In decision models every branch is associated with an attributes and every leaf node corresponds to the class label. The results of existing methods (SVM, KNN Decision Tree, and Bayesian model algorithms) and the current work is compared to show the ensemble learning approach to decision tree algorithm generates more accurate results than the existing works discussed in this research paper.

It is possible to classify the soil from the perspective of material and resource. Soil testing can be performed in the commercial laboratories which propose more type of tests, targeting a group of minerals and compounds present in soil. The benefit of working with a soil test laboratory is familiar with the local soil attributes. This is helpful in the expert’s abilities to assess which tests are most likely to offer valuable insight about the soil.

II. Literature Survey:

D. Mieye et.al [1], suggested the agriculturalists to cultivate the most productive crops based on the types of soil and he prevalent climatic condition (cultivation seasons) of that region.

[1] Research Scholar, Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Affiliated to Bharathidasan University, Tiruchirappalli.
[2]Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Affiliated to Bharathidasan University, Tiruchirappalli

Alobaidiet.al [2], studied some of the appropriate ensemble learning models and also learned that a comprehensive analysis of eight ensemble models is performed using the five machine learning algorithms and three ensemble approaches. It has been observed that the result improved the prediction.

Ramesh Babu Palepu et al., [3], investigated agricultural soil for resolution-making in numerous issues related to the agricultural domain. Numerous data mining tools and techniques were assumed for analyzing soil type and its feature, which will be more tolerable to cultivate the suitable crop. When compared to established statistical analysis, a larger dataset gives better results, which may also enhances the validation of appropriate pattern.

P.Surya et al [4], tested with various regression techniques with the help of collected agriculture soil datasets and devised an algorithm for the effective prediction of agricultural yield for various crops in Tamil Nadu and particularly in and around Erode district, based on soil type classification.

Ashwini Rao et al.,[10] used many machine learning models to examine how the soil of a particular geographic area has been categorized using classification algorithms like support vector machine (SVM) and K-Nearest Neighbor(KNN)and also derived an algorithm to recommend the most suitable crops to be cultivated.

III. Proposed System:

Ensemble Learning is a machine learning models that merges numerous base models to produce a single ultimate predictive model. Ensemble learning approach could be used in many classification problems that could produce better results compared to any other method. Ensemble Learning selects a collection (Ensemble) of hypothesis and unites their predictions. Ensemble learning approach generates a group of base-learners which when combined; multiple learning models (classifiers) are firmly constructed for the classification of data items for the given dataset. In the area of machine learning and statistics, the stacking ensemble learning technique attempts to produce improved outcome of the predictive model, by improving their efficiency and accuracy in the results. The Ensemble Learning technique merges multiple models and produces better predictive power. For example, a group of people is likely to build better decisions compared to individuals when group members are from different domains.

Weighted average ensemble method

For this research paper soil dataset named “***soilmain.csv***” which consists of **attributes** (nutrient OC, K, Mn, Cu, Fe,P, pH, Zn, EC and physical properties – texture, color, humidity etc) and a total number of ***5200 instances*** are being used as input to both existing and proposed algorithms. Initially, 80% of training samples (***4160 instances)*** are being trained to build a classifier model, and 20% of the test data (1040 of unknown class label) are being employed in order to forecast the type of soil. Table 1 given below describes the nutrient attributes of the soil data, but soil has physical properties like texture, color, structure etc., also considered while classifying the soil types.

Table-1: Nutrient attribute descriptions of Soil dataset

Attributes	Description
OC	Organic Carbon, %
K	Potassium , ppm(parts per million)
Mn	Manganese
Cu	Copper
Fe	Iron ppm
P	Phosphorous
pH	pH value of the soil
Zn	Zinc, ppm
EC	Electrical conductivity, deciSiemens per metre (dS/m)

Mathematical models for defining Stacking ensemble learning meta- classifier/algorithm for suitable crop agronomy:

Decision tree classifier is defined as C_{x_i,y_j} where $i = 0,1,\ldots,5$ and $j = 1,2,3,4$ and its ensemle prediction is

$$\sum_{n}^{1} \square\, C_{x_i,y_j} = \{x_i \quad i = 0,1,\ldots,5 \;\; y_j \quad j = 1,2,3,4\} \tag{1}$$

The weight D_{x_i,y_j} is classiied along with the samples are defined as

$$D_{x_i,y_j} = \frac{\sum_{\mathrm{n}}^{1} \square C_{x_i,y_j}}{n} exp\, exp\left(tx_i\, , ty_j\right) \tag{2}$$

where n is the numer of iterations and t be the training data set. The roundoff error rate of classfier r_e is calculated

$$r_e = \frac{1}{2} log\left(\frac{1-e_t}{e_t}\right) \tag{3}$$

Here e_t referred as error in training data set. Now apply the AdaBoost algorithm (meta-classifier) based on weights and re-assigned error (minimized error).
Meta-classifier (AdaBoost ensemble) algorithm:

Input:

Soil Data Set: soilmain.csv
Features of Soil Data Set $SD \ = \{x_i, y_j\}$
$SD_x \ = \{x_i\}_{i=0}^{5}$
$SD_y \ = \{y_i\}_{j=1}^{4}$

x_1 = Class-0	y_1 = Precision
x_2 = Class-1	y_2 = Reccall
x_3 = Class-2	y_3 = F1-score
x_4 = Class-3	y_4 = Support
x_5 = Class-4	
x_6 = Class-5	

Procedure:

1: Initialize the weight based on the number of iterations.
2: Build the training data set based on e_t.
3 : Calculate the weight of the classifier using (2)
4 : Find the round off error according to (1) by using (3).
5 : Initialize the predictor variables
The roundoff error is further approximated as

$$r_{e_a} = \frac{1}{n}\left[\sum_{1}^{\mathrm{t}} \mathrm{e}\, C_{x_i,y_j} - D_{x_i,y_j}\right]$$

The proposed stacking ensemble learning model:

Our proposed algorithm uses 3 base-classifiers (KNN, Random Forest and XGBoost) and a ***meta_classifier*** *(AdaBoost classifier/Algorithm)* discussed above. This algorithm is implemented using machine learning tool Python.

Pseudo code Algorithm for our proposed Stacking ensemble learning technique is given below:

Step 1: Load the Soil dataset “Soilmain.csv” and carry out the necessary preprocessing.
Step 2 : Divide 80% of provided soil datasets into the training set and 20% to the test set

x-train, x-test, y-train, y-test = split_train_test (x, y, train-size =80%, test-size=20%)
Standardize the features after splitting the given dataset.

Step 3: Define the base_classifers (KNN, Random forest, XGBoost) and initialize to store the predicated values of base_classifiers.

knn_classifier1=KNeighborsClassifier (n_neighbors =5)
rf_classifier2 =RandomForestClassifier (n_estimate=100,random_state=42)
xgb_classifier3=XGBClassifier (n_estimate=100,random_state=42)

step 4 : For each base classifier, adhere to the following
(i) Train the base_ classifiers using the training sets.

(ii) Make use of test data to make predictions.
(iii) Store the results of predictions in the base_prediction list.

Step 5: Construct the stacking ensemble model as meta_learner (AdaBoost algorithm) using the above *base_classifiers*; and also initialize list to store the predicted values as follows:

base_classifiers={('knn',knn_classifier1),('rf', rf_classifier2), ('xgb', xgb_classifier3)
stacking_classifier (AdaBoost)= estimater= base_classifiers, find_eatimator=AdaBoostClassifier)

Step 6: Train the meta_ classifier (AdaBoost Algorithm) using the base_prediction as input data and actual labels are from the test dataset. And Fit the stacking ensemble learning model
Step 7: Make the predictions using the trained meta_learner on test data. The meta-models prediction is described by $P_{meta}(X) = f(P_1(X), P_2(X), ...P_k(X))$ where f is the meta model's function.
Step 8 : Make use the metrics precision, recall, F-measure and accuracy to evaluate the performance of the current (proposed) model.
Step 9 : Evaluate the appropriate crops for each type of soil to predicted; and guide to make use of appropriate fertilizers and suitable irrigation system for the identified crops.

The individual models that are modeled or embedded with additional pertinent parameters or models are taken into consideration by the ensemble set. Once the embedded model is identified, then its relative weighted average parameter or model has constructed to identify the optimized parameter and this leads to predict or classify the existing models or parameters from the training data set. This classification task requires machine learning algorithm, involves the accuracy and identify the mislead data. Based on the confusion matrix depicted below in Table-2, the accuracy is arrived.

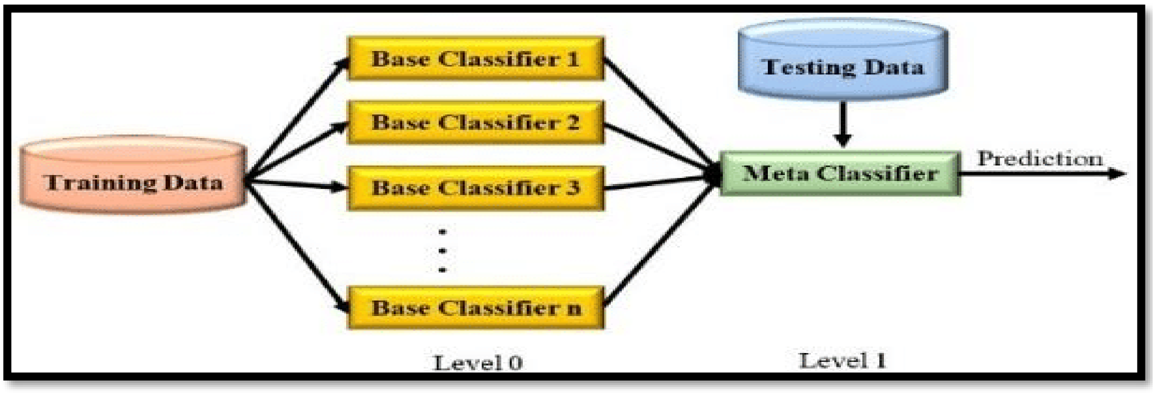
Table-2 : Model of Confusion Matrix

Prediction	**True Class Actual Positive (1)**	**Actual Negative (0)**
Predicted Positive (1)	True Positive (TP)	False Positive (FP)
Predicted Negative (0)	False Negative (FN)	True Negative (TN)

Here ,
Accuracy $= \frac{TP+TN}{TP+TN+FN+FP}$ (Likewise we can calculate other parameter also)

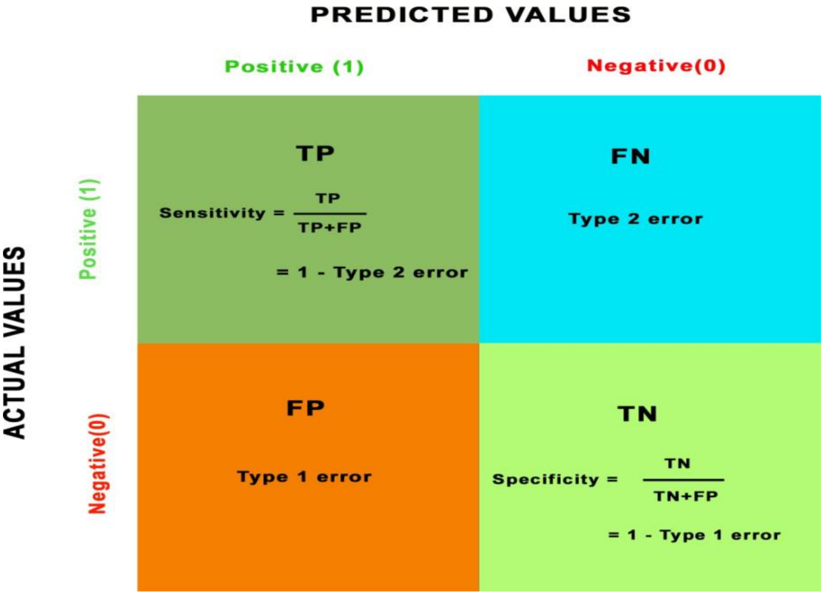
The general architecture of stacking ensemble learning technique with our current work of soil type classification is shown in the below figure.

Figure-1 Architecture of Stacking Ensemble Learning Technique



In the above architectural diagram there are two levels, in the level 0 base-classifiers used in our work are KNN, Random Forest and XGB algorithms . And in Level 1 , AdaBoot algorithm is used as meta-classifier.

Table-3: Generalized structures of computing actual values and predicted values



		PREDICTED VALUES	
		Positive (1)	Negative(0)
ACTUAL VALUES	Positive (1)	TP Sensitivity = $\frac{TP}{TP+FP}$ = 1 - Type 2 error	FN Type 2 error
	Negative(0)	FP Type 1 error	TN Specificity = $\frac{TN}{TN+FP}$ = 1 - Type 1 error

IV. Result and Discussion:

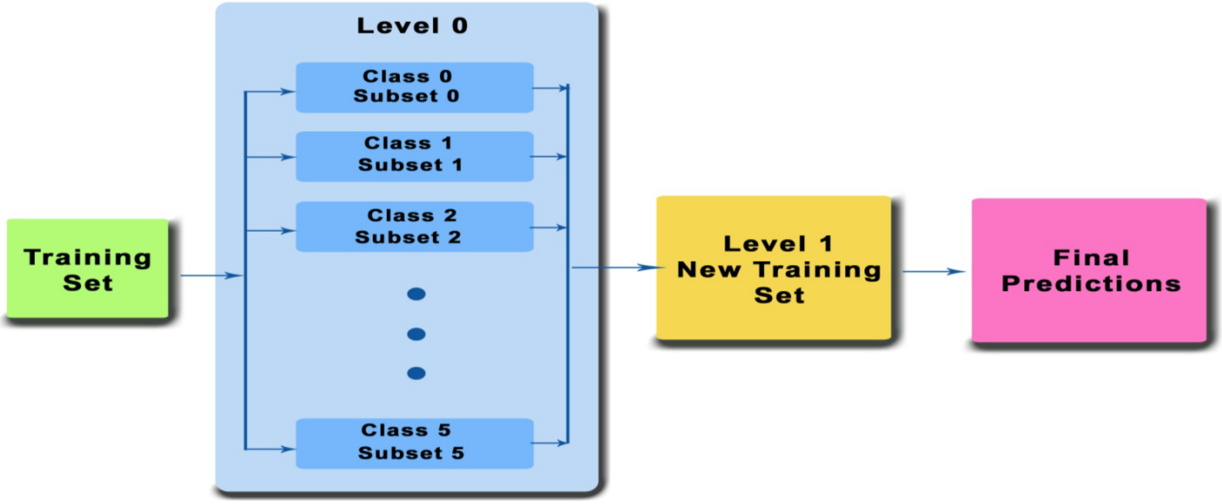
The ensemble was structured with an AdaBoost meta-learner orchestrating the fusion of base-classifiers used in this work. The synergy among these diverse algorithms aimed to enhance the overall prediction performance of the soil classification model and the results are shown in the below table (Table -4). The resulting model exhibited improved generalization and robustness, showcasing the efficacy of ensemble learning in the complex domain of soil classification.

Table- 4: The Outcome of the proposed method

	Precision	Recall	F1-score	Support
Class-0	0.94	0.87	0.90	468
Class-1	0	0	0	0
Class-2	0	0	0	0
Class-3	0.86	0.85	0.87	416
Class-4	0	0	0	0
Class -5	0.71	0.98	0.68	156
micro-average	0.86	0.87	0.88	1040
macro-average	0.39	0.45	0.42	1040
weighted-average	0.87	0.89	0.88	1040
Overall Accuracy is	**0.887935**			

The four metrics (Accuracy, Precision, Recall and F1-score) used in the proposed algorithm; and they are associated with classification of soil type of a particular geographical area. Finally, these four metrics outline the base to make the final evaluation. The Below Figure 2 represents the soil classification structure for our class o to class 5.

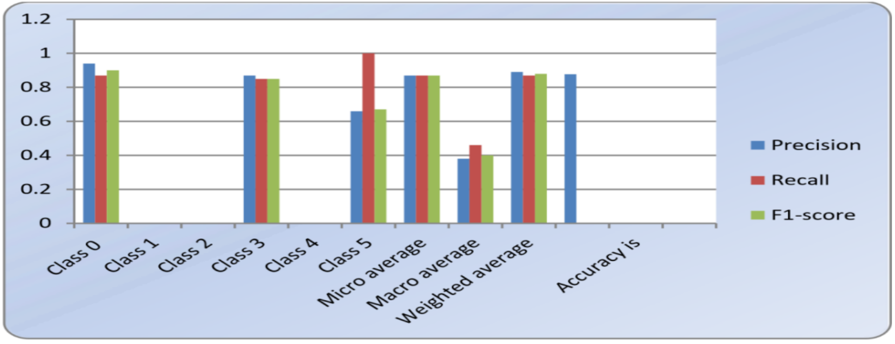
Figure-2: Soil classification structure for class 0 to class 5



The corresponding confusion matrix is defined as[70 2 4 30], sensitivity = 0.9722, specificity = 0.9375, and the overall accuracy of class-0, class-3 and class-5 is 0.89.

The results of the current (proposed) algorithm are show in the above Table-4, and it is visually shown in Figure- 3 using the data visualization tool, a bar chart. Out of 6 Soil types, our proposed algorithm founds 3 soil types namely, Class-0 (Red soil), Class-3 (Literate soil) and Class-5(Alluvial soil). The proposed work (algorithm) predicts the 89% accuracy, it's actually a better result compared to the previous works (algorithms). The above Table-4 also shows the values of micro average, macro average and weighted average.

Figure-3 Graphical representation of proposed Method



Based on the soil type classified and prevailing climatic condition, the farmers are suggested to cultivate one or two suitable crops from the list of crops. The Table-5 given below shows the suitable crops can be cultivated against each soil types predicted by the proposed model. Figures 4 and 5 define the soil classification against true predicted classes and false predicted classes.

Figure- 4 : Soil Classification – True Predicted Classes

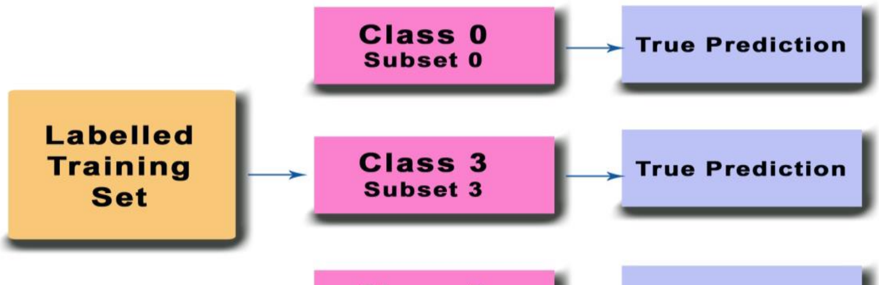
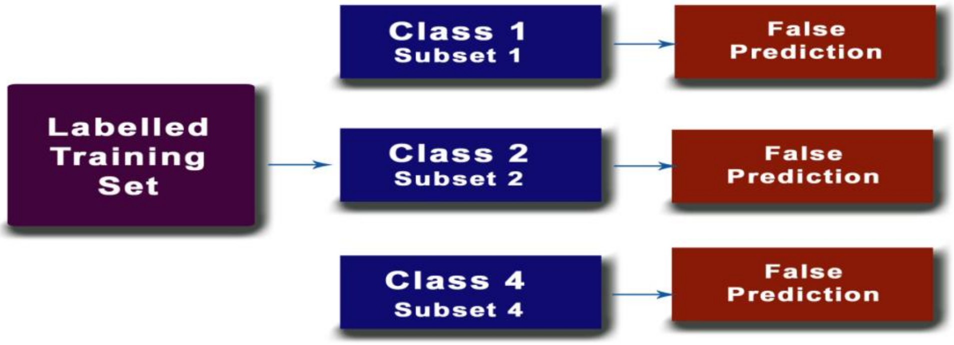


Figure- 5 : Soil Classification – False Predicted Classes



Once the Soil type has been classified using our proposed method, then we can predict the appropriate crop to cultivate based on the climatic condition (season such as Kharif, Rabi and Zaid) using a simple algorithm which is shown in below table (Table- 5).

Table-5: Recommended crops for each the type of soil

Type of soil	Season	Recommended crops to cultivate
Class-0 (Red Soil)	Kharif	Rice, maize, cotton, Black gram, Groundnut
	Rabi	Wheat, bajra, rice, oil seeds
	Zaid	watermelon, musk melon, cucumber, tomato
Class-3 (Literate Soil)	Kharif	tea, coffee, cotton, rice , pulses.
	Rabi	Wheat, coffee, pepper
	Zaid	Coconut, pulses, vegetables.
Class -5 (Alluvial Soil)	Kharif	Rice, maize, Sorghum, Groundnut, sunflower, soybean
	Rabi	Wheat, Barley, mustard, chickpeas, oil seeds
	Zaid	Cucumber, Bitter ground, Ridge ground, Bottle gourd, Okra

The agronomists have to be educated in order to apply suitable irrigation systems (drip, sprinkler, furrow, mulching etc) and fertilizers (Nitrogen, potassium NPK , organic manure etc) to the above mentioned crops to yield the more production in accordance with the Soil type and Season (Kharif, Rabi and Zaid). The following table (Table-6) shows the suitable fertilizers and irrigation system recommended for each type of crops predicted.

Table-6 : Suitable fertilizers and Irrigation System

Crop	Suitable Fertilizers	Irrigation system
Rice	Nitrogen, Phosphorus, Potassium (NPK) fertilizers	Flood irrigation, Drip irrigation (if possible)
Wheat	Nitrogen, Phosphorus, Potassium (NPK) fertilizers	Drip / sprinkler irrigation
Millets	Organic manure, Nitrogen-rich fertilizers	Rainfed, drip irrigation
Cotton	Phosphatic fertilizers, organic manure	Drip / Furrow irrigation
Vegetables	Balanced NPK fertilizers, Organic amendments	Drip , sprinkler irrigation
Groundnut	Nitrogen-rich fertilizers, organic matters	Drip, Sprinkler irrigation
Coconut	Organic manures, Balanced NPK	Drip/Basin irrigation ,
Soybean	Nitrogen-rich fertilizers, Phosphatic fertilizers	Drip irrigation, Furrow irrigation

Tea / Coffee	Organic manures and Nitrogen-rich fertilizers	Drip irrigation , Sprinkler irrigation

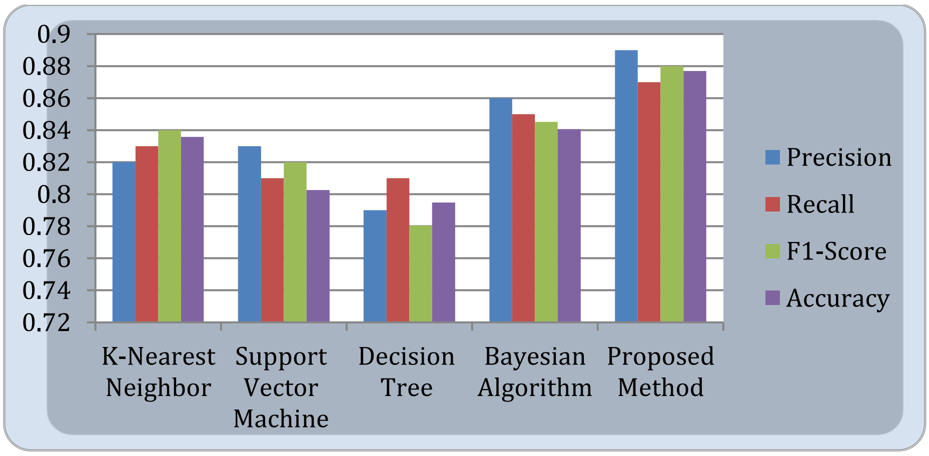
Comparison between existing and proposed models:

The simplest way we can comprehend the effectiveness of model we propose is to compare its result with the results of our previous algorithms used in our research paper. The final results are furnished in below table (Table-7). The table's data are also shown graphically using Bar chart in the figure below (figure-5), help us to easily understand the performance of existing algorithms with the proposed algorithm as shown in the table as well as in the figure.

Table-7: Comparison between existing and proposed models

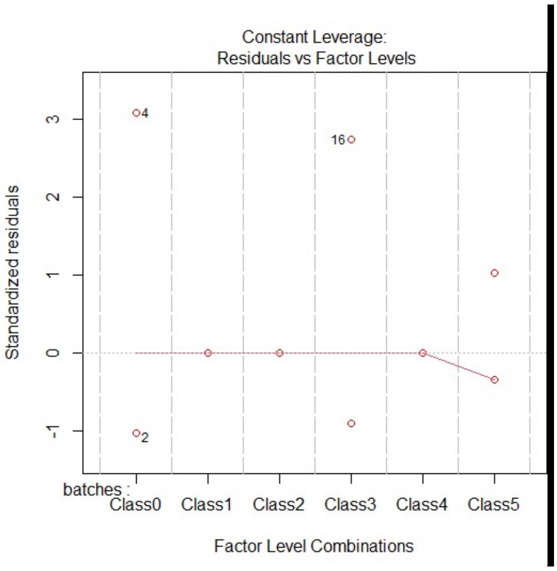
Algorithms	Precision	Recall	F1-Score	Accuracy
KNN	0.84	0.82	0.83	0.83569
SVM	0.83	0.81	0.82	0.80254
Decision Tree	0.79	0.81	0.7805	0.79477
Bayesian Models	0.86	0.85	0.84521	0.840501
Proposed Method	**0.89**	**0.87**	**0.88**	**0.87695**

Figure-6 : Comparisons of existing and proposed algorithms



Figures 7 to 10 explains the concept of residual Vs factor values, residual Vs fitted values, residual Vs quartiles and residual Vs factor combinations.

Figure-7: Residual Vs Factor Values   **Figure-8: Residual Vs Fitted Values**



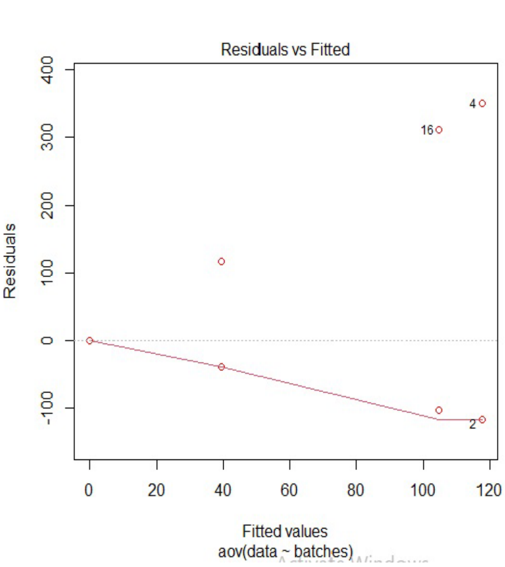
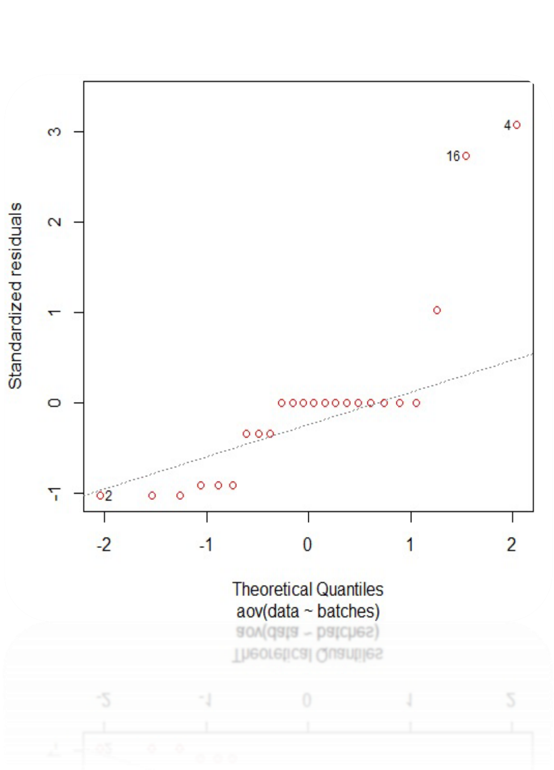
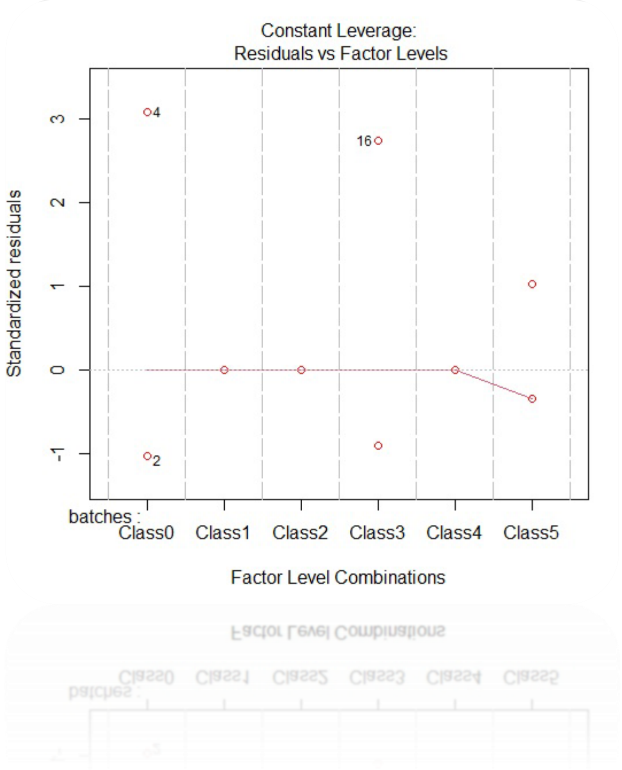


Figure -9: Residual Vs Quartiles **Figure-10: Residuals Vs Factor Combinations**





V. Conclusion:

The application of ensemble learning techniques with decision tree in agricultural field is an innovative research idea. Our proposed stacking ensemble learning approach has classified the various types of soil for given dataset of a specific area, and facilitating agronomists in identifying the right crops to grow across different climatic conditions (Cultivation seasons) . In this research paper, the utilization of stacking ensemble learning for constructing decision trees has proven to be a robust methodology for accurate soil classification. By extending the scope to include tailored recommendation for crop cultivation, fertilizers and irrigation system, this study provides a comprehensive guide for farmers to optimize their agricultural practices. Then the result of our current work is compared with the previous (existing) works and both results given in the table format, as well as e depicted using the visualizing tool bar-chart for the better understanding of the classification of soil type in the result and discussion section of this paper. Finally we can understand that our current work predicts accurate result. The results predicted using the stacking ensemble learning can be improved using the machine learning algorithm called multilayer neural network algorithm.

References:

1. D. Mienye and Y.Sun, "A Survey of Ensemble Learning : Concepts, Algorithms, Applications, and Prospects," in *IEEE Access*, vol. 10, pp. 99129-99149, 2022, doi: 10.1109/ACCESS.2022.3207287.
2. Alobaidi M, H. Meguid, M.A,Chebana, “Ensemble learning for classification of Soil Liquefactions”, CSCS Annual Conference, June 12-15, 2019
3. Ramesh BabuPalepu and Rajesh Reddy Muley, “ Analysis of Agricultural Soil by using Data Mining Technique” International Journal of Engineering Science and Computing, Volume:7, Issue. 177, October 2017, ISSN:2321-3361
4. P.Surya and Dr.L.Laurence Aroquiaraj, "Crop yields prediction in Agriculture using Data Mining Predictive Analytic" International Journal of Research and Analytical Review (IJRAR), vol.5, Issue 4, pp. 783-787, ISSN 2348- 1269.
5. Vrushal Milan Dolas, Uday Joshi, "A Novel Approach for Classification of Soil and Crop Prediction", International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.7, Issue 3,pp.20-24, March 2018, ISSN 2320-0887.
6. Rushika Gadge, Pooja more, Juilee Kulkarni, Sachee Nene, Priya , “Prediction of Crop yields using Machine Learning Techniques”, International Journal Research Journal of Engineering and Technology (IRJET), volume: 05, issue: 02, pp. 2237-2239, Feb 2018, e-ISSN: 2895-0056.6.
7. A.Mythili and N.Saranya, “Classification of Soil Type and Crop Suggestion using Machine Learning Technique”, International Journal of Engineering Research & Technology (IJERT),Vol. 9 Issue 02, pp. 671-673 February-2020, ISSN: 2278-0181.
8. Chandan , Ritula Thakur, ”Recent Trends of Machine Learning in Soil Classification: A Review”**,** International Journal of Computational Engineering Research (IJCER), vol: 08, Issue. 9, pp. 25-31, September – 2018, ISSN (e): 2250 – 3005.
9. D. Ashok Kumar and N. Kannathasan,” A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, pp. 422-426, May 2011, ISSN (Online): 1694-0814.
10. Ashwini Rao, Janhavi , Abhishek Gowda N.S, and Mrs.Rafeqa Beham,”Machine Learning Models in Soil Classification and crop detection”, International Journal of Scientific Research & Development, Volume:4, Issue: 1,pp.792-794, April 2016, ISSN (online): 2321-0613.
11. Yin, J., Medellín-Azuara, J., Escriva-Bou, A., and Liu, Z. (2021). “Bayesian machine learning ensemble approach to quantify model uncertainty in predicting groundwater storage change”, *Sci. Total Environ.* 769, 144715. doi:10.1016/j.scitotenv.2020.144715.
12. M. Ragab, A. M. Abdel Aal, A. O. Jifri, and N. F. Omran, “Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques,” *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 6241676, 2021.
13. S. B. Keser and S. Aghalarova, “Hela: a novel hybrid ensemble learning algorithm for predicting academic performance of students,” *Education and Information Technologies*, pp. 1–32, 2021.
14. S. Zian, S. A. Kareem, and K. D. Varathan, “An empirical evaluation of stacked ensembles with different meta-learners in imbalanced Classification,” *IEEE* Acce ss, vol. 9, 2021.

.