JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 07, ISSUE 17, 2020

CLASSIFICATION OF SOIL PH CLASSES BASED ON SOIL FACTORS USING MACHINE LEARNING TECHNIQUES

Ramji, M[1] and Rajarathinam, A[2]

[1,2]Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, India

7jipgm@gmail.com[1] and rajarathinam@msuniv.ac.in [2]

https://orcid.org/0000-0002-4491-3093 and https://orcid.org/0000-0002-3245-3181

Abstract

Soil classification is a method for organising and classifying data concerning soil. In order to define lands in a straightforward, consistent, and understandable manner, which is crucial for plantation and agricultural decision-making this category of soil was created. The current approach to determining soil type is time-consuming and primarily dependent on agricultural specialists. The use of machine learning is anticipated to improve soil classification and recommend the required factors. With a total accuracy of 82.19% for Nitrogen (N), it has trouble correctly predicting the High category. With an overall accuracy of 69.53%, and the efficacy of the categorization model to predict pH levels varies among different soil factors.

Keywords: LDA, CART, RF, KNN, SVM, Soil parameter, Classification and Accuracy.

1. Introduction

Soil is crucial for the development of agriculture. The health of the soil and the crops it supports are supported by a healthy soil profile that allows air and water to circulate easily into and through it. Soil holds onto water for agricultural growth, which also facilitates movement for equipment and animals. The majority of the components required for plant growth are ingested by roots from the soil. The soil type can be determined using a variety of techniques, including conventional techniques, knowledge, and technology. Predicting soil behaviour is made easier by understanding soil classification. It is possible to predict how effectively a soil will grow crops by looking at its behaviour. Currently, soil-related categorization algorithms can be designed using machine learning (ML) and deep learning (DL) models, which have just become available. For predicting soil moisture, soil nutrient content, and soil types. A group of five classifiers, including random forest (RF), support vector machine (SVM), K-nearest neighbour (KNN), linear discriminant analysis (LDA), and classification and regression

tree (CART), were used to categorize based on soil macronutrient levels and soil fertility indices. The class label was then evaluated on a scale of high, low, and medium based on their numerical value (Kayad et al., (2021), Zeraatpisheh et al., (2022), and Mehrjardi et al., (2022)). The Machine learning tests was evaluated by R programming tool (Ver. 4.0.5).

The goal of this work is to create and assess a classification model for predicting the quantity of pH in soil samples across three different classes (high, medium, and low) based on the different soil factors.

2. Review of related works

Several researchers have recently used machine learning techniques in the soil domain. The following is a summary of the application of several ML approaches during the last few years in the field of crop recommendation prediction from soil analysis.

Raza, (2008) studied the situation of Karnataka's soil health. The various soil types were analysed and categorised using various machine learning approaches. Using the tree-based models Decision Tree (C5.0) and Random Forest (RF), this research study for the classification of soil types was carried out SVM (Support Vector Machines) and XGBOOST (eXtreme Gradient Boosting). While execution times for various models varied and Random Forest had the most efficient computation time with approximately comparable accuracy, accuracy and Kappa values suggested XGBOOST performed the best.

Coopersmith et al., (2014) used information from NEXRAD regarding the precipitation over time in their analysis. K-Nearest Neighbors (KNN), Decision Trees (DTs), and enhanced perceptrons were employed as classification algorithms. Since the prediction was utilized to determine whether or not the soil would be ready, the issue was essentially binary in nature. With an accuracy of 93%, (KNN) outperformed the other two algorithms, and was followed by the enhanced perceptron algorithm.

According to Sirsat et al., (2018) for numerous important soil nutrients, including organic carbon (OC), phosphorus pentoxide (P2O5), iron (Fe), manganese (Mn), and zinc (Zn), pedotransfer functions that automatically predict the village-wise soil fertility indices have been developed using a wide range of regression techniques. The extremely random regression trees (extraTrees) produced the best results. Using machine learning approaches

like Support Vector Regression (SVR), Ensembled Regression (ER), and Neural Network (NN), the soil attributes were predicted. The outcomes demonstrated that Ensembled Regression (ER) performed better than SVM and NN (Singhatiya & Ghosh, (2018)).

Pant et al., (2020) applied several machine learning techniques, including Support Vector Machine, Logistic Regression, Linear Discriminant Analysis (LDA), and K-NN, used soil physical and chemical characteristics and macronutrients to predict the soil quality categorization and soil fertility. The data set was labelled using the unsupervised K-Means approach. SVM was discovered to have a good accuracy rating of 96.62%.

A machine learning algorithm has been suggested by Malik et al., (2021) to estimate fertility and agricultural output. The author's data for three crops—potato, chilli pepper, and tomato—were the basis for the study. Crop yields were calculated using the Decision Trees classifier, the Naive Bayes algorithm, and the K-Nearest Neighbour method.

The new ensemble regression crop prediction model beat a number of supervised machine learning and sophisticated ensemble learning approaches, according to Iniyan et al., (2022). The sophisticated ensemble regression crop prediction model outperformed numerous supervised machine learning and sophisticated ensemble learning approaches in terms of anticipated yield.

Comparing model-averaging methods for forecasting the spatial distribution of soil attributes was done by Kaya et al., (2022). Their findings demonstrated that the best methods for predicting soil qualities were ANNs and random forest-based classifiers. However, it was discovered that the applicability of learning-based models varied in different use situations, and the study did not provide an inference of soil attributes for the best preservation of soil quality.

3. Materials and Methods

3.1. Material

The dataset considered for usage in the given proposed work is a crop recommendation dataset primarily comprising of soil properties, along with the N, P, K, pH, Temperature, Humidity and Rainfall. An open source dataset is obtained from the Kaggle (https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset). Totally 2200 samples were used based on the recommended twenty two different crops.

3.2. Methods

A supervised machine learning approach called classification employs the algorithm to classify new observations after it has learned from the data input that was provided to it. Leonel, (2019). With the use of a collection of example data that have been classified, this method determines what data should be recognized. The construction of a classifier involves two stages. The training set must choose which parameter to concentrate on and how to merge the many types of data into a single form of data during the training phase. When it comes to testing, the set will be evaluated by applying it to test data with a defined aim and contrasting it with chosen data. For more information, the testing set will generate a result that indicates how long it takes to interpret each piece of data with precision and determine whether it has a high level of accuracy or not. In figure 1, described proposed structure of the work.

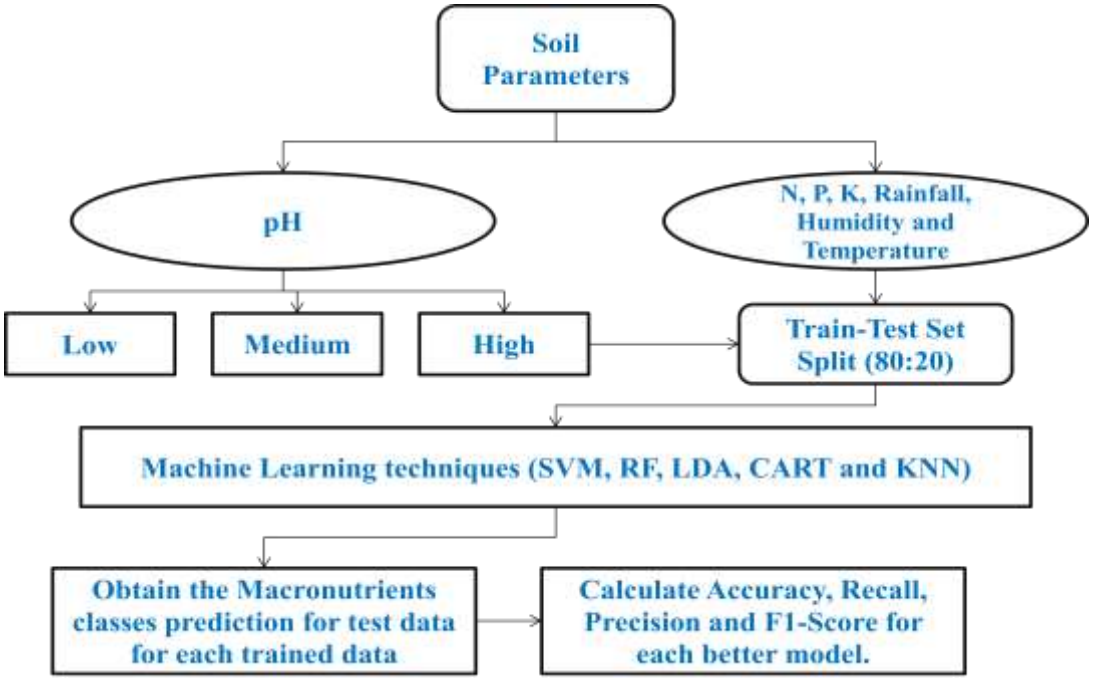

Figure 1: Proposed structure

3.3. Linear discriminant analysis

The goal of LDA is to classify cases into three or more categories using continuous or dummy categorical variables as predictors. The term DA (Fisher, (1936)) refers to numerous types of analyses. DA is the most popular statistical technique to classify individuals or observations into no overlapping groups, based on scores derived from a suitable “statistical

decision function" constructed from one or more continuous predictor variables. While investigating the differences between the groups or categories, the necessary step is to identify the attributes with most contributions to maximum reparability between known groups or categories in order to classify a given observation in to one of the groups. For that purpose, DA successively identifies the linear combination of attributes known as canonical discriminant functions (equations) that contribute maximally to group separation. Predictive DA addresses the question of how to assign new cases to groups.

DA involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is:

$$D = a + v_1 X_1 + v_2 X_2 + v_3 X_3 + ... + v_i X_p$$

Where D is the discriminant function, v_i is the discriminant coefficient or weight for variable, X_i is the independent score for the i^{th} variable, *a* represent the constant and p is the number of predictor variables.

3.4. Random Forest

A random forest is just a collection of trees used for prediction. For each tree in the model, random vectors are created and then sampled. When a lot of trees are created and the most well-liked class is chosen, the result is a random forest. By creating numerous decision trees, random forest aids in the minimization of variance error. The Random Forests classifier is, by definition, a mixture of many classifiers that are organized in a tree structure. It can also be expressed as follows:

$$(h(x, \theta_k),\ k = 1, ...)$$

Herein θ_k i.e Random vectors which are independent and identically distributed. The classes' input is X, and the main notion is that each tree casts a vote to determine which classes are most popular by incorporating the x. (Breiman, (2001)).

3.5. Support vector machine

The (SVM) splits the classes by introducing a hyperplane between them in its binary form, acting as a classifier. By increasing the margin distance between the points and the hyperplane, the classes are split. The support vector closest point is the hyperplane. One

versus one and one versus all concepts are frequently used in multi-class problems. For each class pair contained in the data, a single classifier is generated using the One versus One approach. The One versus All strategy, on the other hand, builds a variety of classifiers that function to separate a specific class from all other classes. Every time a new observation is taken into account, the classifier with the best decision-making function is selected. A greater number of hyper planes are built using multiclass (SVM) Sun et al., (2018).

3.6. K Nearest Neighbors

One of the most fundamental yet crucial classification methods in machine learning is K-Nearest Neighbors. It falls under the category of supervised learning and is widely used in intrusion detection, data mining, and pattern recognition. Since it is non-parametric and makes no underlying assumptions about the distribution of the data (unlike other algorithms like GMM, which assume a Gaussian distribution of the input data), it is extensively applicable in real-world applications. We are provided some prior information (also known as training data), which organises coordinates according to an attribute. Euclidean distance, which can be determined using the following equation, is used by default in KNN approaches Short and Fukunaga (1981).

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

where, p and q are subjects to be compared with n characteristics.

3.7. Classification and Regression tree

The threshold value of an attribute is used to divide the nodes in the decision tree into sub nodes. The Gini Index criterion is used by the CART algorithm to find the sub nodes with the best homogeneity. The training set is the root node, which is divided into two by taking the best attribute and threshold value into account. Additionally, the subsets are divided according to the same rationale. This continues until the tree has either produced all of its potential leaves or found its last pure sub-set. Tree pruning is another name for this.

The formula of the Gini Index is as follows:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

where, 'p_i' is the probability of an object being classified to a particular class (Daniya et al., (2020)).

4. Results and discussion

The tests compare the outputs of the recommended datasets for soil nutrients and factors of crops using LDA, CART, KNN, SVM and RF. Accuracy and kappa are two performance indicators used to assess the study's findings. This section discusses the results of the experiment's five algorithms, which were classified using overall measures and two performance measurements.

Table 1: pH category information

Soil Parameters	Category	Ratings	%	N
Potential of Hydrogen	Low	Below 6	26.05	573
	Medium	6 to 7	52.7	1159
	High	above 07	21.3	468

Table 1 shows the Nutrient rating, percentage of available category with percentage (Krishnaveni, et al., (2014)) of potential of Hydrogen for plants growth. If pH levels are low then the soil fertility level is low. In this dataset, middle classes found a higher percentage of pH (52.7%) than the other two groups. High category percentage of pH was found to be 21.3% respectively. It is advised to add more fertiliser to the soil to increase its fertility due to its pH content.

4.1. Soil factor functions

- ✓ **Potential of Hydrogen (pH)** is the solubility, mobility, and bioavailability of trace elements is governed by soil pH, which determines how they are transported by plants.
- ✓ **Nitrogen (N)** is a very important and needed for plant growth. It is found in healthy soils, and give plants the energy to grow, and produce fruit or vegetables. Nitrogen is actually considered the most important component for supporting plant growth.
- ✓ **Phosphorus (P)** is one of the major plant nutrients in the soil. It is a constituent of plant cells, essential for cell division and development of the growing tip of the plant.
- ✓ **Potassium (K)** helps plants make strong stems and keep growing fast. It's also used to help fight disease.

- ✓ **Soil temperature** directly affects plant growth. Most soil organisms function best at an optimum soil temperature. Soil temperature impacts the rate of nitrification. It also influences soil moisture content, aeration and availability of plant nutrients.
- ✓ Soil moisture or **Humidity** plays an important role in agricultural monitoring, drought and flood forecasting, forest fire prediction, water supply management, and other natural resource activities.
- ✓ **Rainfall** is a major component of the water cycle and is responsible for depositing most of the fresh water on the Earth. It provides water for hydroelectric power plants, crop irrigation, and suitable conditions for many types of ecosystems.

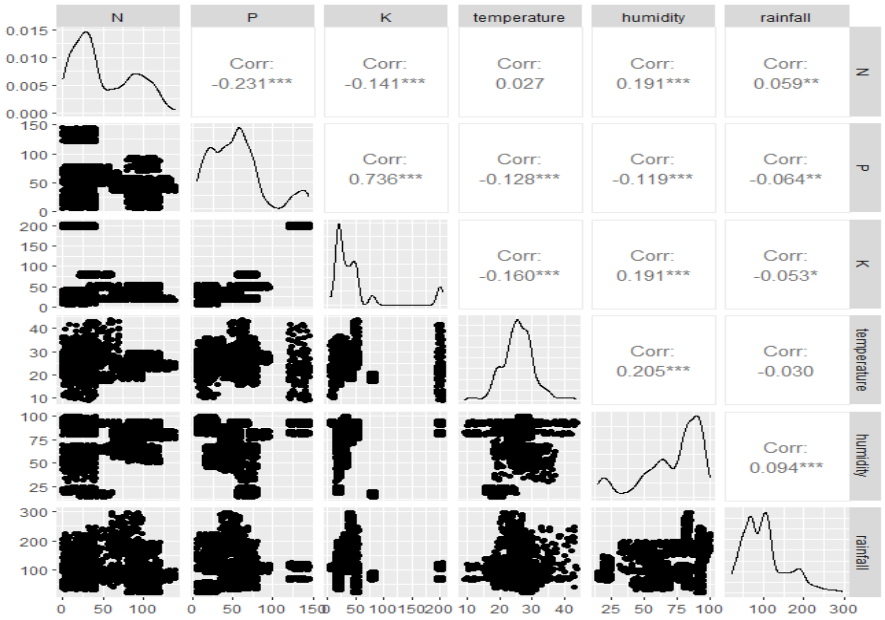

Figure 2: Correlation between different soil factors

In the figure 2, each factors distribution is displayed diagonally. The bivariate scatter plots with a fitted line are shown on the bottom of the diagonal, and the correlation coefficient and significance level of the independent variables are shown as stars on the top. Each degree of significance has a corresponding symbol: symbols ("*", "*", "**", "***") are equivalent to the p-values (0, 0.001, 0.01, 0.05). K and P are highly positively correlated (0.736) compared to the other factors. N and P are highly negatively correlated (-0.231).

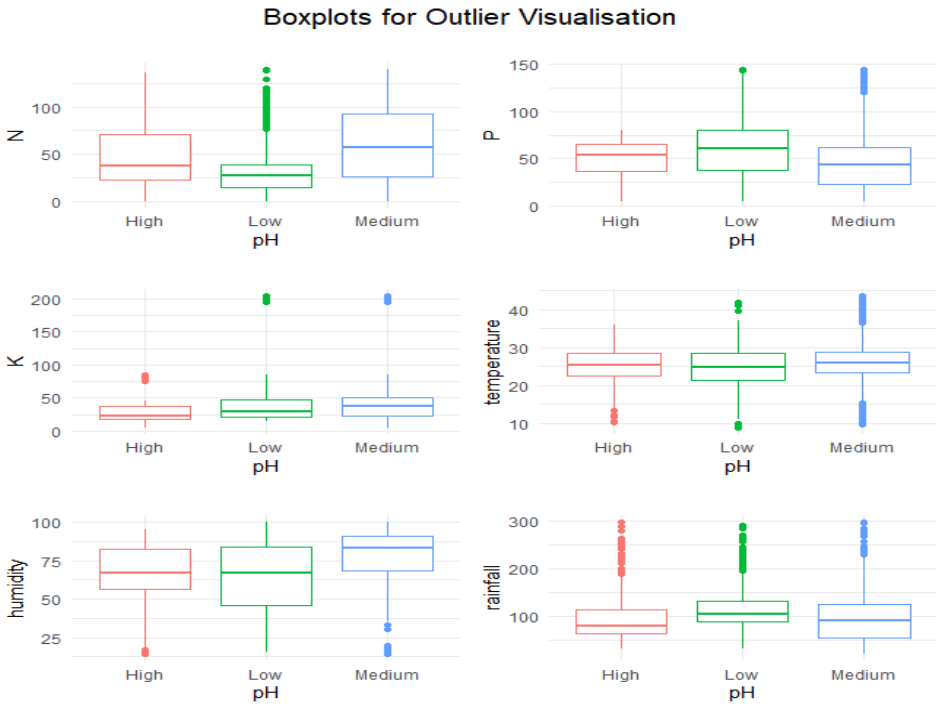

Figure 3: Outlier identification

The box plot (figure 3), where the box is constructed with the median value of the data set, graphically depicts the existence of outliers. Outliers are indicated by the upper or lower values of this box plot, with the centre line of the box representing the data median. The three varied colours stand for the three categories of the pH levels (Low, Medium and High).

Table 2: Accuracy and Kappa values for different models

Model	LDA	CART	KNN	SVM	RF
Accuracy	0.57	0.60	0.61	**0.64**	0.60
Kappa	0.20	0.26	0.34	**0.38**	0.34

The table 2 lists the success rates of various models for predicting pH levels in a certain situation. Accuracy and Kappa coefficient are the two evaluation measures that are employed, and it is clear which model will provide the best fit. According to Table 2 results for five different classification models, the Support Vector Machine classifier has a high accuracy compared to other models in pH. Figure 4 high (0.64 and 0.38) was demonstrated by accuracy and the kappa values of SVM. Thus, the SVM model can choose which extra categorization to apply to the soil factors.

ISSN- 2394-5125 VOL 07, ISSUE 17, 2020

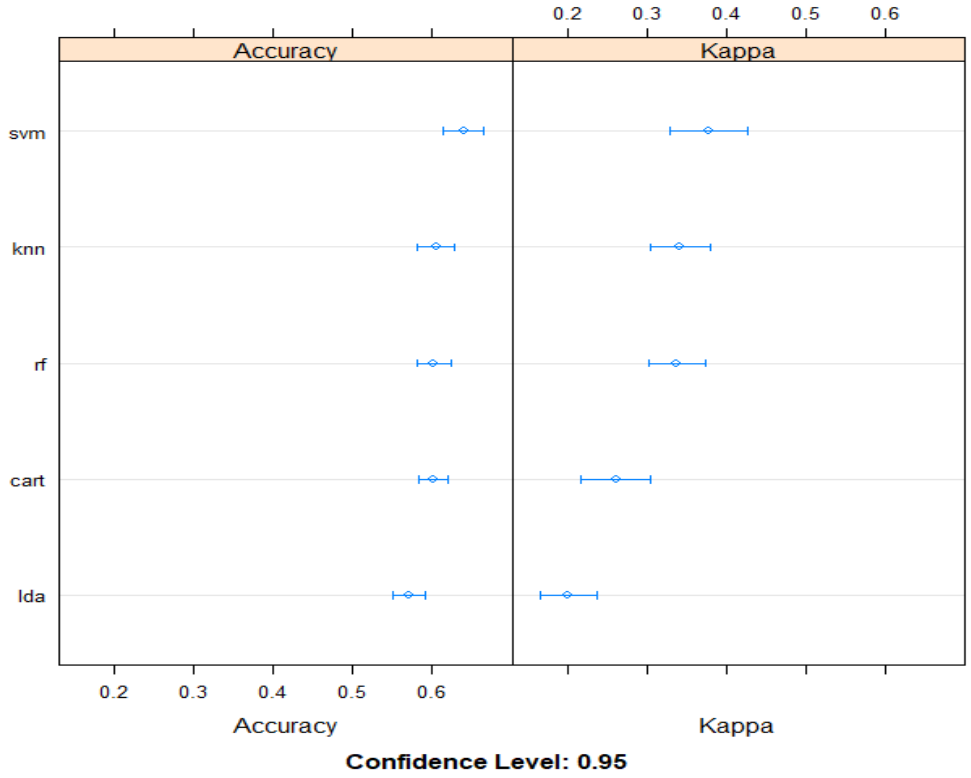

Figure 4: Accuracy and Kappa values

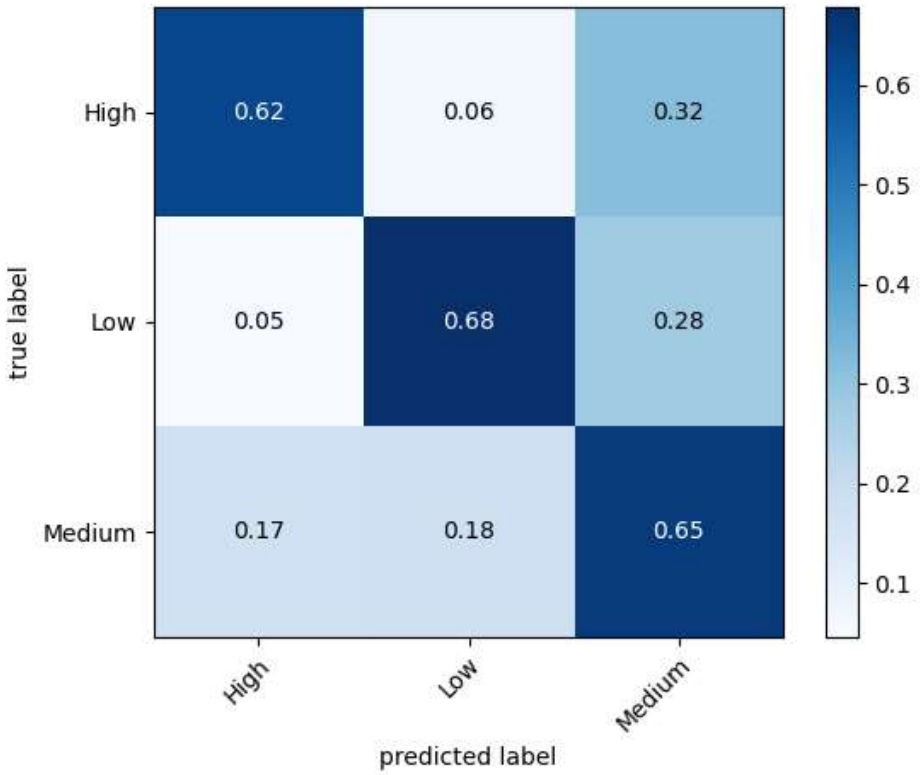

Figure 5: Confusion matrix

The confusion matrix plot for the pH classes is shown in Figure 5; where the column represents the target class and the row represent the projected class (Output Class). The diagonal cells relate to appropriately categorised observations. The off-diagonal cells are associated with observations that were misclassified. In each cell, the number of observations and the proportion of all observations are displayed. In the case of pH, 62% of samples are correctly categorised as high, whereas 32% and 6% are incorrectly categorised in medium and low categories. 68% and 65% of samples were properly categorised as having low and medium pH categories.

Table 3: Overall measurements

Macronutrients	Class	N	Accuracy	Precision	Recall	F1 score	Overall accuracy
pH	High	93	**82.19%**	0.62	0.42	0.50	69.53%
	Low	114	81.05%	0.68	0.52	0.59	
	Medium	231	66.89%	0.65	0.81	0.72	

In order to predict the soil factors level pH in three separate classes: high, medium, and low the performance metrics of a classification model are presented in the table 3. Along with the overall accuracy, the measurements include Accuracy, Precision, Recall, and F1 score. It is important that the model illustrates differing levels of effectiveness for various nutrients and classes. With a total accuracy of 82.19% for Nitrogen (N), it has trouble correctly predicting the High category. With an overall accuracy of 69.53%, the model's advantages and disadvantages in predicting nutrient levels are highlighted by these findings, emphasizing the necessity of additional improvement and optimization for precise nutrient assessment.

5. Conclusion

This study has shown that is possible to predict the right pH category levels for a specific soil using amount of soil information. Based on the nutritional levels, the pH was divided into three separate classes. Typically, pH was taken into consideration. The five machine learning techniques have been studied in detail includes SVM, CART, RF, LDA, and KNN. The generated models SVM model can choose which extra categorization to apply through pH. With a total accuracy of 82.19% for Nitrogen (N), it has trouble correctly predicting the High category. With an overall accuracy of 69.53%, and the efficacy of the categorization model to predict pH levels varies among soil factors. These

findings highlight the necessity of individualized enhancements to the model's predictive skills, notably in the precise classification of high and low nutrient levels, to increase its efficiency in nutritional assessment and related applications.

Reference

[1]. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[2]. Coopersmith, E. J., Minsker, B. S., Wenzel, C. E., & Gilmore, B. J. (2014). Machine learning assessments of soil drying for agricultural planning. *Computers and Electronics in Agriculture*, *104*, 93–104. https://doi.org/10.1016/j.compag.2014.04.004

[3]. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[4]. Iniyan, S., & Jebakumar, R. (2022). Mutual information feature selection (MIFS) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression (MSER). *Wireless Personal Communications*, *126*(3), 1935-1964.

[5]. Kaya, F., Keshavarzi, A., Francaviglia, R., Kaplan, G., Başayiğit, L., & Dedeoğlu, M. (2022). Assessing machine learning-based prediction under different agricultural practices for digital mapping of soil organic carbon and available phosphorus. *Agriculture*, *12*(7), 1062.

[6]. Kayad, A., Sozzi, M., Gatto, S., Whelan, B., Sartori, L., & Marinello, F. (2021). Ten years of corn yield dynamics at field scale under digital agriculture solutions: A case study from North Italy. *Computers and Electronics in Agriculture*, *185*(106126), 106126. https://doi.org/10.1016/j.compag.2021.106126

[7]. Krishnaveni, M., Kalimuthu, R., Ponraj, K., Magesh, P., Lavanya, K., & Shyni, G. (2014). Nutrient analysis of soil samples collected from Yercaud road, Salem district, Tamil Nadu, India. *International Journal of Pharmaceutical Sciences Review and Research*, *26*(1), 216–217.

[8]. Leonel, J. (2019, October 9). Classification methods in machine learning. Medium. https://medium.com/@jorgesleonel/classification-methods-in-machine-learning-58ce63173db8

[9]. Malik, P., Sengupta, S., & Jadon, J. S. (2021, January). Comparative analysis of soil properties to predict fertility and crop yield using machine learning algorithms. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 1004-1007). IEEE.

[10]. Pant, H., Lohani, M. C., Bhatt, A., Pant, J., & Joshi, A. (2020). Soil Quality Analysis and Fertility Assessment to Improve the Prediction Accuracy using Machine Learning Approach. *International Journal of Advanced Science and Technology*, *29*(3), 10032–10043.

[11]. Raza Ansari, S. (2018). *Application of Machine Learning Techniques for Soil Type Classification of Karanataka (Doctoral dissertation).*

[12]. Short, R., & Fukunaga, K. (1981). The optimal distance measure for nearest neighbor classification. *IEEE Transactions on Information Theory*, *27*(5), 622–627. https://doi.org/10.1109/tit.1981.1056403

[13]. Singhatiya, S., & Ghosh, D. S. (2018). Performance Evaluation of Artificial Intelligence on Soil Property Detection. *Smart Moves Journal Ijoscience*, *4*(10), 5. https://doi.org/10.24113/ijoscience.v4i10.166

[14]. Sirsat, M. S., Cernadas, E., Fernández-Delgado, M., & Barro, S. (2018b). Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods. *Computers and Electronics in Agriculture*, *154*(September 2017), 120–133. https://doi.org/10.1016/j.compag.2018.08.003

[15]. Sun, Y., Feng, X., & Yang, L. (2018). Predicting tunnel squeezing using multiclass support vector machines. *Advances in Civil Engineering*, *2018*, 1–12. https://doi.org/10.1155/2018/4543984

[16]. Taghizadeh-Mehrjardi, R., Khademi, H., Khayamim, F., Zeraatpisheh, M., Heung, B., & Scholten, T. (2022). A comparison of model averaging techniques to predict the spatial distribution of soil properties. *Remote Sensing*, *14*(3), 472. https://doi.org/10.3390/rs14030472

[17]. Zeraatpisheh, M., Garosi, Y., Reza Owliaie, H., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., & Xu, M. (2022). Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. *Catena*, *208*(105723), 105723. https://doi.org/10.1016/j.catena.2021.105723

[18]. Daniya, T., Geetha, M., & Kumar, K. S. (2020). Classification and regression trees with gini index. *Advances in Mathematics: Scientific Journal*, *9*(10), 8237-8247.