Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Outlier Detection And Denoising By The Measures Of Dispersion With Naïve Bayes To Predict Soil Fertility

Raynukaazhakarsamy[1] , Dr. J. G. R. Sathiaseelan[2]

[1]Research Scholar, Department of Computer Science, Bishop Heber College (Autonomous), Trichy – 620 017, Affiliated to Bharathidasan University, Tiruchirappalli – 620 024, Tamil Nadu, India.

[2.] Associate Professor & Head, Department of Computer Science, Bishop Heber College (Autonomous), Trichy – 620 017, Affiliated to Bharathidasan University, Trichy – 620 024, Tamil Nadu, India.

Abstract:
Farming is one of the essential segments of Indian Economy and it funds about 17% of Gross Domestic Product (GDB). Growing crops more than millennia without thinking often about renewing has prompted consumption and depletion of soil supplements bringing about their low usefulness. To work on production, compound composts are added. Adequate measure of manures should be added for the improvement of good yield and simultaneously the normal nature of soil stays perfect. The soil should be identified as fertile or not fertile to carry out the process of manuring. Methods/Statistical analysis: It is essential to remove redundant data, replace missing values and outliers from the soil dataset. Removing duplicates and replacing missing values are performed using Discretized Naïve Bayes (DNBayes) and Outliers are detected and replaced using Discretized Naïve Bayes with Quartiles (DNBQ). Findings: The performance of the proposed system is analyzed in terms of different types of errors like KS, MAE, RMSE, RAE and RRSE along with TPR, FPR, Precision, Recall, F1-Score, ROC and classification accuracy. Prediction of soil fertility based on DNBayes and DNBQ achieved 85% and 87% of classification accuracy. Novelty/Applications: The proposed approaches will benefit soil scientists in decision making to help farmers in predicting fertile soil for sugarcane cultivation.

Keywords: Discretized Naïve Bayes; Denoising; Outlier Detection; Inter Quartile Range; Soil Fertility

1. Introduction

Soils are the most important resources[1,2] in agriculture and their protection, maintenance, and improvement is being critical now days due to the usage of chemical in agricultural land. Three main nutrients nitrogen (N), phosphorus (P) and potassium (K) are necessary for the plants to grow. Soil analysis much needed to determine the amount of nutrients, composition and other particulars. Tests are generally carried out for to measure fertility and indicate deficiencies that are taken for the sake of reparation. The soil analysis laboratories are available with the relevant technical literature on miscellaneous aspects of soil assessment, included test methodologies and making fertilizer recommendations.

It assists farmers in determining the use of fertilizers and Farm manure must be applied in a variety of stages the crop's growing cycle. The current procedure for examining the nutrient of a soil is a troublesome undertaking. One needs to give the soil sample to the Soil Survey Department to investigate the fertility of the soil. This strategy for breaking down soil supplement level is truly tedious and subsequently a large portion of the ranchers can't get a greatest yield during collecting, prompting ghastly circumstances. This is a significant issue. Recent research has been directed to gauge and guide soil supplement substance in huge regions utilizing hyper-ghastly methods, nonetheless, it is difficult to acquire exact assessments.

In this research work section 2 discusses about the related works, section 3 talks about the machine learning algorithms, section 4 details the outlier detection, section 5 frames the proposed work, section 6 elaborates the experimental results with its graphical representation and section 7 concludes with future scope.

2. Related Work

The impairment or weakening of the material due to response with the atmosphere is called corrosion[3]. The eroding of soil can source natural tragedies such as landslides. Soil corrosivity[4] is a severe challenge in industrial and infrastructure events. Soil data analysis[5] is done using various algorithms and predictions method. They have presented a comparative study on Naïve Bayes and J48 (C4.5). Graphic methods of histogram and box plot[6] are combined to discover the outliers from the soil dataset of phosphorus (P).

Attribute selection[7] is done with the help of Fast Correlation Based Filter method. A scheme is designed that utilizes a set of data pre-processing events consists of choosing proper attributes and discretizing functions[8].

Farming data[9] have been used to find optimum elements that increase the crop production using the data mining technique like Partition Around Medoids (PAM), Clustering Large Applications (CLARA) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is modified and used to aggregate data by district wise with similar temperature, rain and soil type.

An automated system is developed to classify soils based on fertilization. Once the fecundity class labels have been obtained with automated system, they conducted comparative research on various classification[10] technologies using WEKA. This research has implemented a healthy practical application of the method of linear regression by forecasting an obscure property of the soil test. The findings of this study substantially lower price of these tests, which will save Indian people a lot of effort and time.

Missing values and outliers are regularly faced during the records series segment of observational or investigational research performed in all arenas of soil dataset[11]. Missing values can stand up from records loss in addition to dropouts and nonresponses of the look at partakers. The presence of missing values leads to a littler test estimate than aiming and in the long run negotiations the unwavering quality of the study outcomes. It also can create unfair consequences whilst inferences about a populace are drawn primarily based totally on the sort of trial, undermining the trustworthiness of the data.

A pre-application pace to the machine learning[12] is effective for reducing dimensionality, eliminating non-significant data thereby increasing accuracy. As a part of the pre-treatment procedure, missing records are both unnoticed in desire of simplicity or substituted with substituted values envisioned with a numerical method[13]. In general, missing value study takes into account efficiency, the management of missing data and the consequent analytical complexity, and the bias among missing and observed values. Another difficultly is outlier analysis that represents the values that lying outside the distribution pattern of soil dataset. An outlier may have diverse sources such as; goofs in measuring, off-base decimal focuses, blunder in copying, and accidental estimation of a part of different population.

3. Machine Learning Algorithms

3.1 Naïve Bayes

Naive Bayes classifiers are scalable and act as a conditional probability model, dataset instance is signified by a vector $Y = \{ y_1, y_2 \ldots y_m \}$ representing m features or independent variables. It assigns each probability from eq. (1). Each possible outcome C_k is classes.

$$p(C_k \, | y_1 . y_2, \ldots . y_m) \tag{1}$$

When number of features are larger or the values accepted for features are many, the basic model is infeasible. Using the algorithm, the basic model is reformulated as,

$$p(C_k \mid y) = \frac{p(C_k)p(y|C_k)}{p(y)} \tag{2}$$

$p(C_k \mid y)$ - Posterior probability
$p(C_k)$ - Prior probability of class
$p(y)$ - Prior probability of predictor

3.2 K Nearest Neighbor Algorithm

The K Nearest Neighbor algorithm[14] is a kind of supervised learning method that is utilized for classification and regression. This method may also be considered to fill in mislaid values and resample datasets. K-Nearest Neighbor deliberates data points to estimate the class or continuous value for a different datapoint. We use whole training occurrences to forecast outcome for hidden data, instead of taking weights from preparation data to expect output (as in model-based algorithms). The learning procedure is suspended until a estimation is mandatory on the new element, and the model is not learned using training data beforehand. The mapping function in KNN does not have a pre-set form.

Algorithm:

Step 1: Choose the Kth neighbor's number.
Step 2: Decide the Euclidean space among K neighbours using the eq. (3).
Step 3: Consider the K next neighbours based on the Euclidean space obtained.
Step 4: Sum the quantity of data points in every category among these k neighbours.
Step 5: Allocate the different data points to the group with the highest amount of neighbours.
Step 6: The model is complete.

$$\text{Euclidean distance } = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (3)$$

3.3 Support Vector Machine (SVM)

SVM is a supervised machine learning technique[15] that can be utilized to solve classification and regression problems. The goal of SVM is to discover a hyperplane with the greatest margin in an N-dimensional space that differentiates data points in separate two classes to perform classification. Hinge loss is a loss function that helps in margin maximization. The cost function is given by the eq. (4)

$$c(x, y, f(x)) = \begin{cases} 0 \\ 1 - y * f(x) \end{cases} \qquad (4)$$

If the projected and actual values have the same sign, the cost is zero otherwise, the loss value is calculated. A regularization factor is also added to the cost function. The regularization parameter's objective is to strike a compromise among margin maximization and loss. The cost functions look like this after totaling the regularization option as shown in eq. (5).

$$\min_w \lambda \|w\|^2 + \sum_{i=1}^{m}(1 - y_i)\langle x_i, w\rangle \qquad (5)$$

3.4 Proportional K-Interval Discretization (PKID)

PKID helps to normalize the amount and dimension of discretized gaps to the quantity of training existences, hence discovers an appropriate trade-off between the bias and variance of the probability evaluation for NB classifiers.

While implementing PKID, it is important to follow rules given below:

– Discretization is restricted to identified values of a numeric features. Unknown values can be simply ignored.

– For certain features, dissimilar training occurrences may hold same values. It is important to continuously keep the same values in a single interval. Thus, preferably every interval should contain precisely √N instances, the exact dimension of each interval may differ.

– Set N training examples with recognized values of a numeric attribute, grasp [√N] as the ordinary dimension of the discretized pause. Smaller size of intervals is not permitted. Only larger sizes are allowed because of the occurrence of same values or to accommodate the last interval when its size is among [√ N] and [√ N] × 2.

3.5 Inter Quartile Range (IQR)

IQR is a extensively recognized technique to discover outliers in data. Using this IQR, the complete dataset is divided into four equivalent sectors, or quartiles.

The IQR formula helps in calculating the alterations among the third and first quartile. The IQR formula trials the variability, based on separating an ordered set of data into quartiles. Quartiles are three values or cuts that split each respective part as the first, second, and third quartiles, represented by IQ1, IQ2, and IQ3.

IQ1 - cut in the first half of the rank-ordered data set
IQ2 - median value of the set
IQ3 - cut in the second half of the rank-ordered data set.

IQR Formula

Interquartile range = Upper Quartile – Lower Quartile
$IQ_2=IQ_3–IQ_1$
Here,
IQR = Interquartile range (IQR = IQ_2)
IQ_1 = (1/4)[(n + 1)]th term)
IQ_3 = (3/4)[(n + 1)]th term)
n = number of data points

The uses of IQR are given below:

- ✓ Unlike range, IQR states where the majority of data lies and is thus desired over range.
- ✓ IQR can be helpful to discover outliers in a data set.
- ✓ Gives the principal tendency of the data.

4. Outlier Detection

The following steps help us to find the
IQR:

- It is important to arrange the data points in rising order.
- IQ_2 denotes the median of the data. When the amount of data points is odd, the central term is (n+1)/2 and whereas the amount of data points is even, the median is the mean of the two middle points.
- IQ_1 indicates the median of the data points to the left of the median found in step 2.
- IQ_3 signifies the median of the data points to the right of the median found in step 2.
- IQR = $IQ_2=IQ_3–IQ_1$
-

Steps involved in Outlier Detection using IQR

1. IQ_1 - Lower Quartile (25% Quartile)
2. IQ_3 - Upper Quartile (75% Quartile)

3. IQR = $IQ_3 - IQ_1$ (Inter Quartile Range)
4. Outlier1 = IQ_3 + OF * IQR (Maximum Threshold Value)
   if x > outlier1 then,
   x = outlier1
5. Outlier2 = IQ_3 – OF * IQR (Minimum Threshold Value)
   if x < outlier2 then,
   x = outlier2
6. Median = Max & Min default value

Note:

1. IQR - Inter Quartile Range
2. OF - Outlier Factor = 3.0 (fixed)

5. Proposed Methodology

The proposed system includes Discretized Naïve Bayes (DNBayes) and Discretized Naïve Bayes with Quartiles (DNBQ). They are detailed in this section.

5.1 Discretized Naïve Bayes (DNBayes)

The proposed DNBayes (Figure 1) is a combination of Proportional k-Interval Discretization (PKID) and Naïve Bayes (NB). It initially considers 128 samples of input elements. First step is to remove the redundant data. If the any redundant sample occurs it is removed otherwise all the samples could be taken for further process. Here the second step indicates the replacement of missing values. Mean value would be considered for numerical attributes whereas Mode values are considered for nominal attributes. The third step is to apply PKID filter to discretize the dataset into equal number of bins. The final step is to apply the concept of Naïve Bayes. In case of NB, pH attribute is utilized for condition checking.

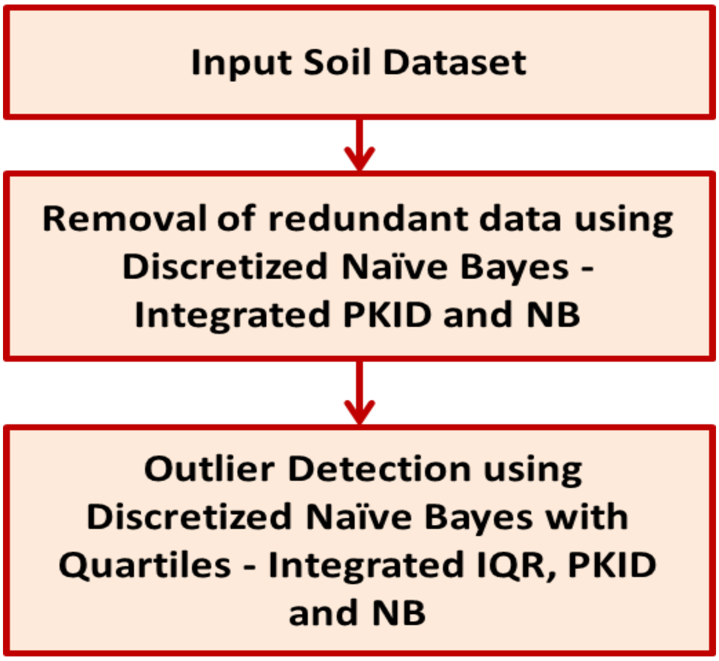
Input Soil Dataset
Removal of redundant data using Discretized Naïve Bayes - Integrated PKID and NB
Outlier Detection using Discretized Naïve Bayes with Quartiles - Integrated IQR, PKID and NB

Figure 1: Proposed Framework

5.2 Discretized Naïve Bayes with Quartiles (DNBQ)

The concept of DNBQ (Figure 1) is an integration of Proportional k-Interval Discretization (PKID), Naïve Bayes (NB) and Inter Quartile Range (IQR).

Step 1: The input samples are attained after removing redundant values if occur from the whole dataset.
Step 2: Replacement of missing values is carried out by taking mean for numeric attributes and mode for nominal attributes.
Step 3: IQR (Inter Quartile Range) is implemented on the soil dataset statistically using Excel to identify and replace outliers then imported in WEKA for further process
Step 4: The concept of PKID filter is implemented to divide the given dataset into equal number of bins.
Step 5: The Naïve Bayes considers pH attribute that is for condition checking process.

6. Results and Discussion

Real-Time dataset of soil samples collected from Rajshree Sugars and Chemicals Ltd., Theni is taken for this research work. Sugarcane soil dataset is divided into training and testing samples with the ratio of 60% and 40% respectively. Test dataset with the size of 128 samples is utilized for this research work and experimental results are produced. Following 10 attributes namely, pH, EC, OC, N, P, K, S, Fe, Mn and Zn are considered. Table 1 shows the Statistical report for soil attributes. The experimental results of both proposed works DNBayes and DNBQ are given below

Table 1: Error Metrics for DNBayes

Algorithms & Error Metrics	NB	KNN	SVM	DNBayes
KS	0.0692	0.2479	0.0899	0.6978
MAE	0.3907	0.3632	0.3835	0.2017
RMSE	0.6211	0.5959	0.6192	0.3246
RAE	0.8048	0.7481	0.7898	0.4112
RRSE	1.2606	1.2095	1.2569	0.6553

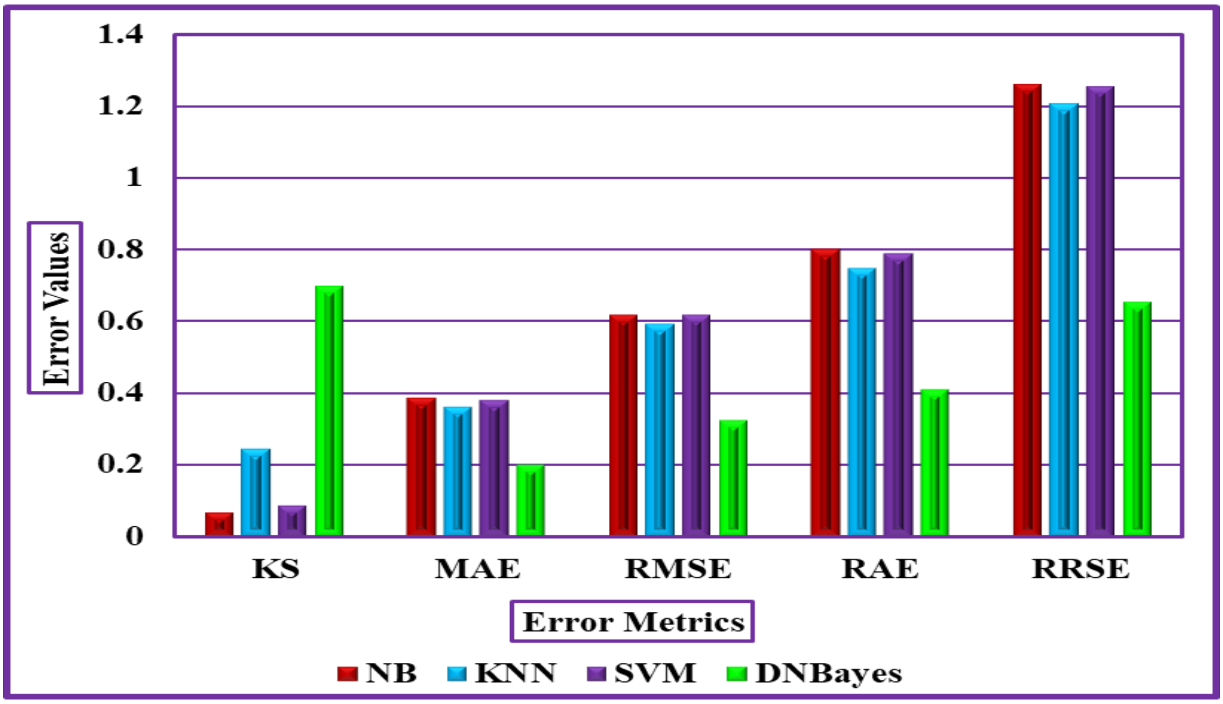

Figure 2: Error Metrics for DNBayes

Table 2: Evaluation Metrics for DNBayes

Algorithms & Evaluation Metrics	NB	KNN	SVM	DN Bayes
TPR	0.609	0.639	0.617	0.852
FPR	0.549	0.394	0.538	0.152
Precision	0.684	0.636	0.701	0.852
Recall	0.609	0.639	0.617	0.852
F1-Score	0.494	0.637	0.508	0.852
ROC	0.748	0.607	0.539	0.928

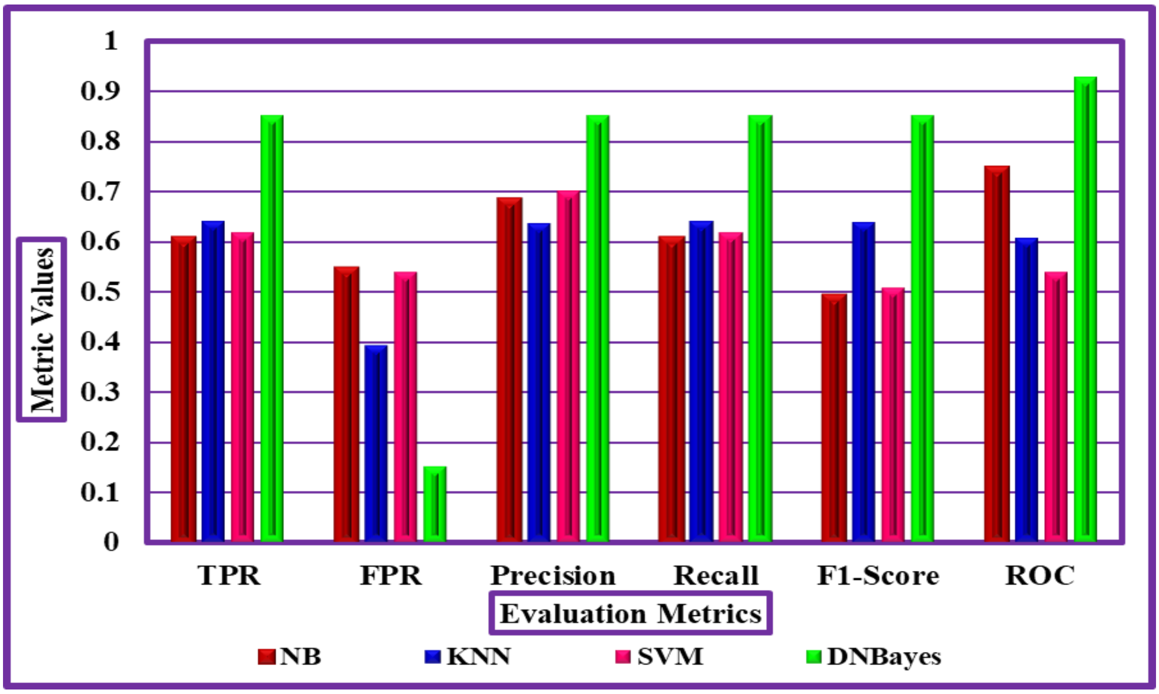

Figure 3: Evaluation Metrics for DNBayes

Table 3: Classification Accuracy for DNBayes

Metrics & Algorithms	No. of Samples		Classification Accuracy (%)
	Correctly Classified	Incorrectly Classified	
NB	78	50	61
KNN	82	46	64
SVM	79	49	62
DNBayes	109	19	85

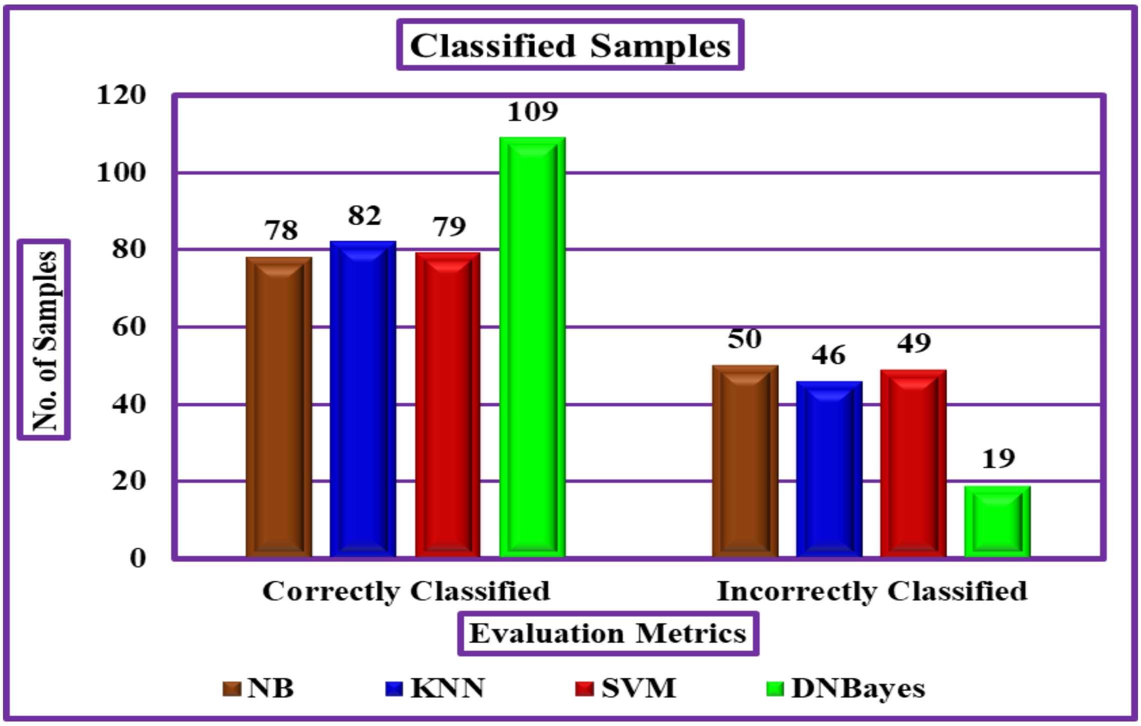

Figure 4: Classified Samples for DNBayes

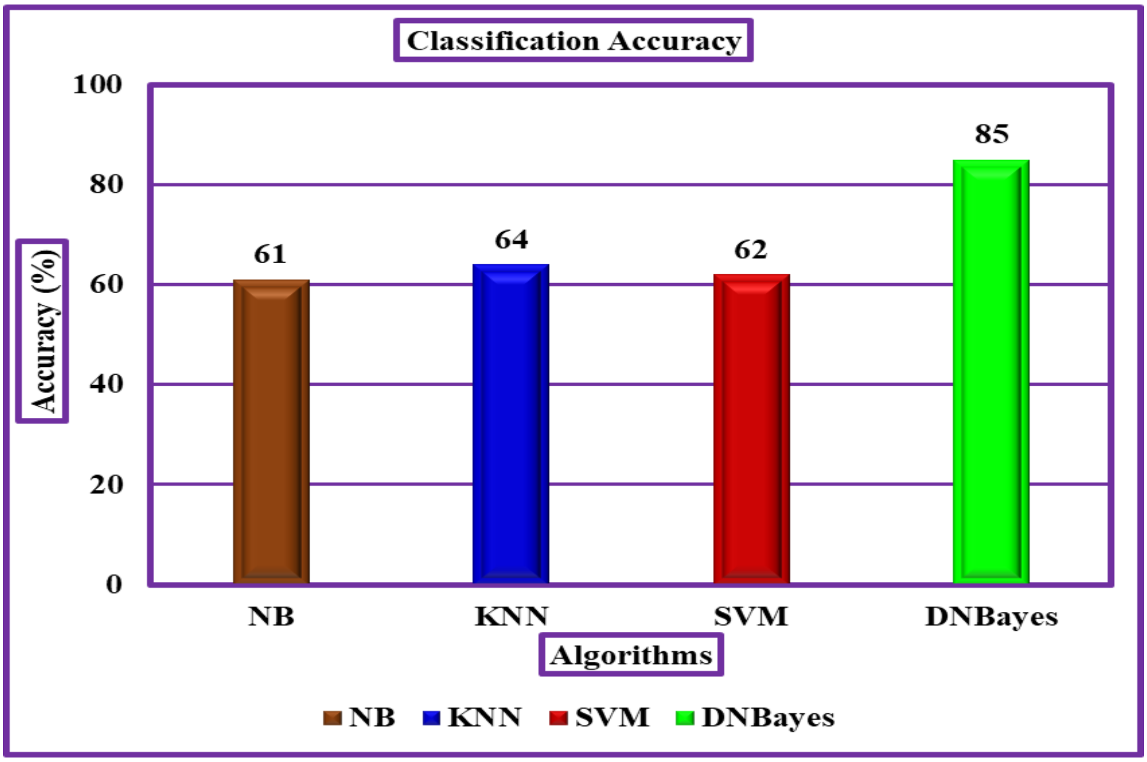

Figure 5: Classification Accuracy for DNBayes

Experimental results of Kappa Statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and evaluation metrics like True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1-Score, Receiver Operator Characteristic (ROC), number of samples classified and classification accuracy based on DNBayes are shown in (Table 1 - 3) also its graphical representation is shown in (Figure 2 – 5). The results reveal that DNBayes outperforms other existing algorithms NB, KNN and SVM by producing 85% of classification accuracy in a real-time dataset with 128 samples.

Table 4: Comparison of Soil Attributes for Outlier Detection

Metrics & Soil Attributes	Standard Deviation		Mean Values	
	DNBayes	DNBQ	DNBayes	DNBQ
pH	1.5	0.499	7.9	7.753
EC	1.4	0.126	0.46	0.24
OC	32.3	0.307	6.1	0.6
N	27.6	15.295	89.2	86.366
P	90.5	12.173	33.6	22.084
K	170.1	151.427	222.8	215.305
S	16.3	15.028	19.6	19.04
Fe	18.7	6.296	13.9	8.84
Mn	4.01	1.947	5.2	4.038
Zn	1.8	0.885	1.7	1.415

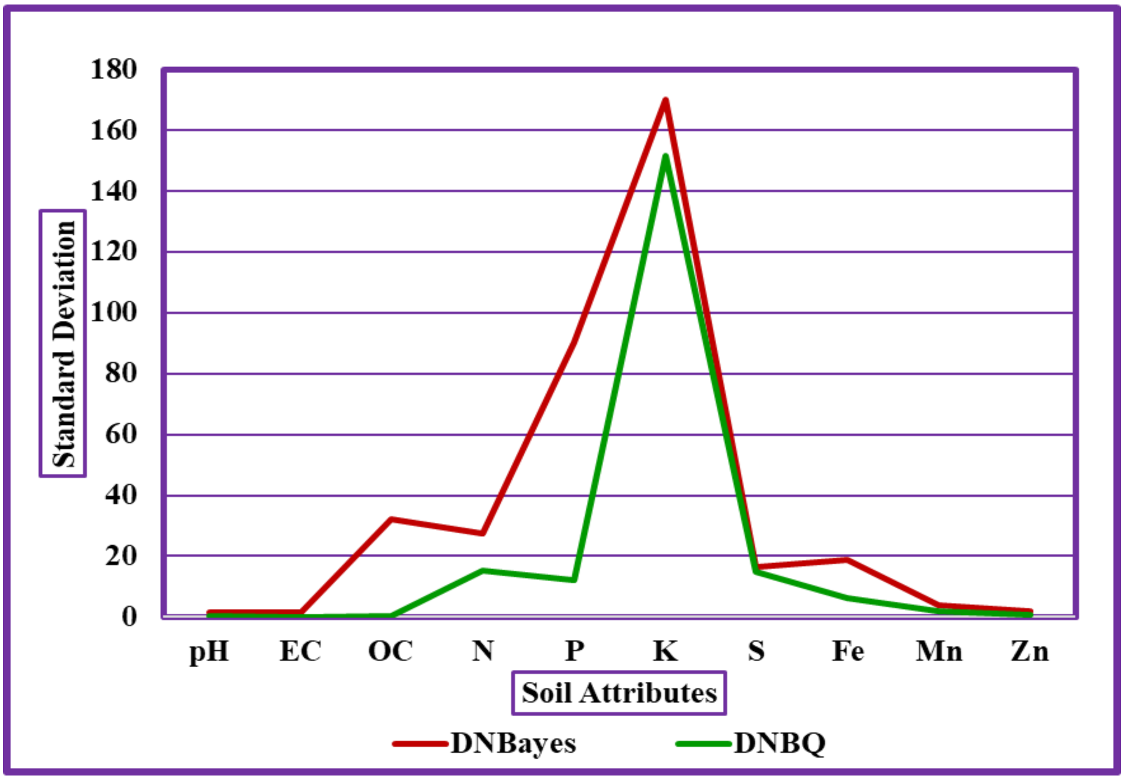

Figure 6: Standard Deviation of Soil Attributes

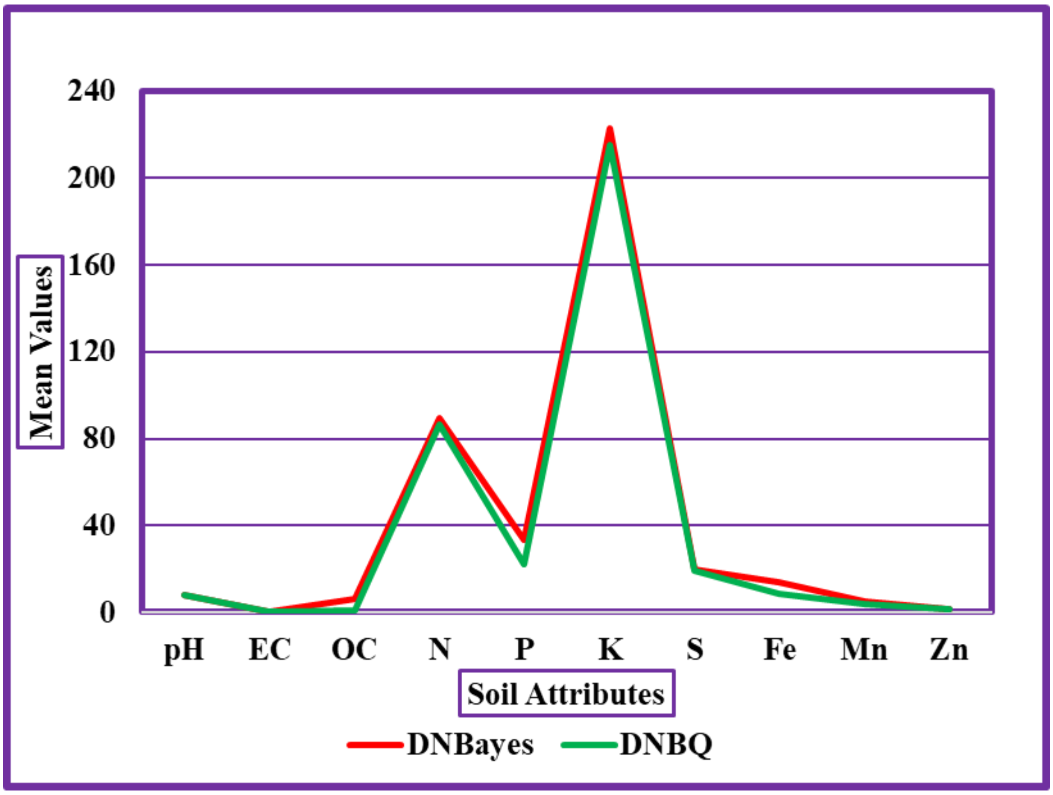

Figure 7: Mean Values of Soil Attributes

Table 4 shows the experimental results about the measures of dispersion for soil attributes after performing outlier detection and replacement using DNBQ. This is proved by comparing the results of standard deviation and mean values of soil attributes with outliers by DNBayes and without outliers by DNBQ with its graphical representations shown in Figure 6 & 7. The output reveals that standard deviation and mean values are reduced in DNBQ which significantly reduces the misclassification error rate and simultaneously improves the success rate by increasing the classification accuracy that are shown in the following results.

Table 5: Error Metrics for DNBQ

Algorithms & Error Metrics	NB	DNBayes	DNBQ
KS	0.0692	0.6978	0.7343
MAE	0.3907	0.2017	0.1924

RMSE	**0.6211**	**0.3246**	**0.3202**
RAE	**0.8048**	**0.4112**	**0.3962**
RRSE	**1.2606**	**0.6553**	**0.6499**

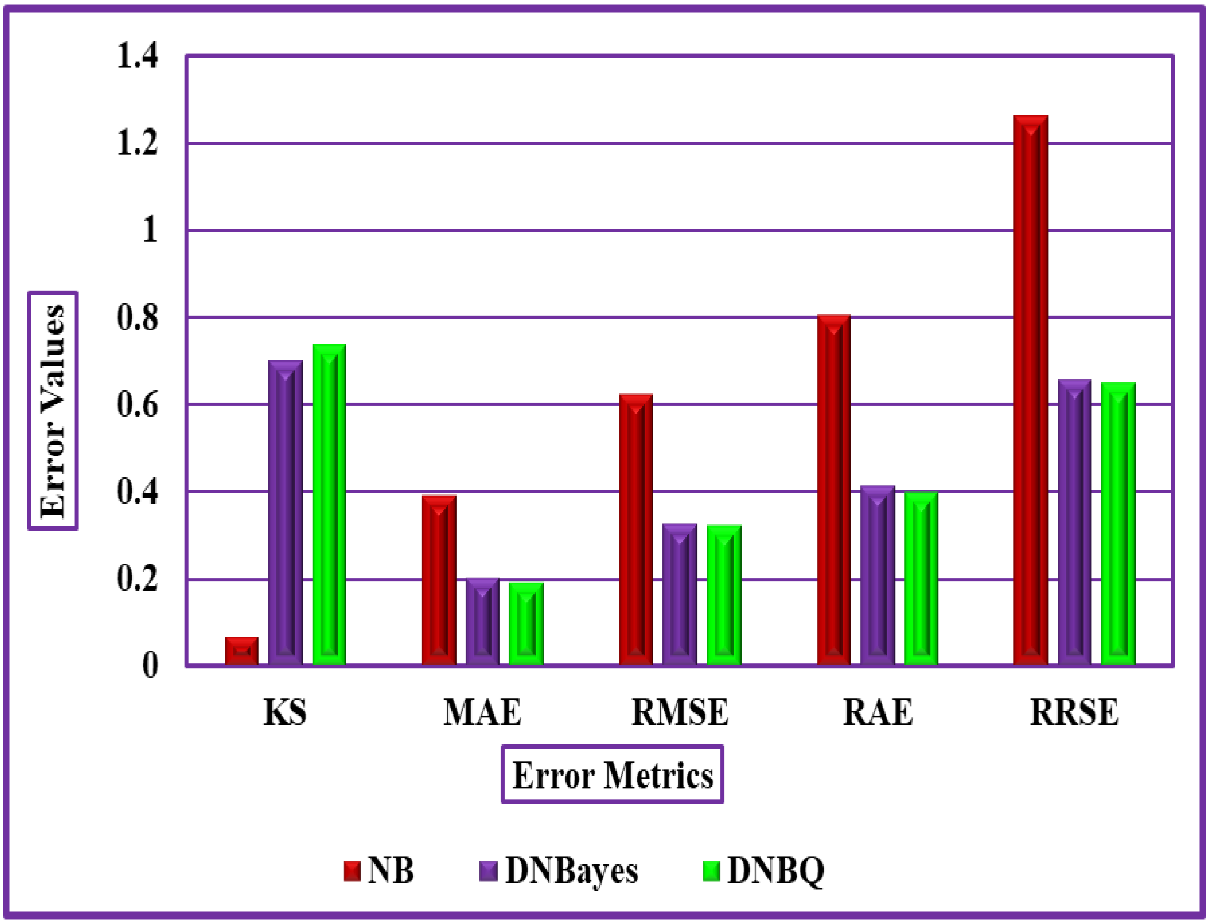

Figure 8: Error Metrics for DNBQ

Table 6: Evaluation Metrics for DNBQ

Algorithms & Evaluation Metrics	**NB**	**DNBayes**	**DNBQ**
TPR	**0.609**	**0.852**	**0.872**
FPR	**0.549**	**0.152**	**0.144**

Precision	**0.684**	**0.852**	**0.872**
Recall	**0.609**	**0.852**	**0.872**
F1-Score	**0.494**	**0.852**	**0.872**
ROC	**0.748**	**0.928**	**0.946**

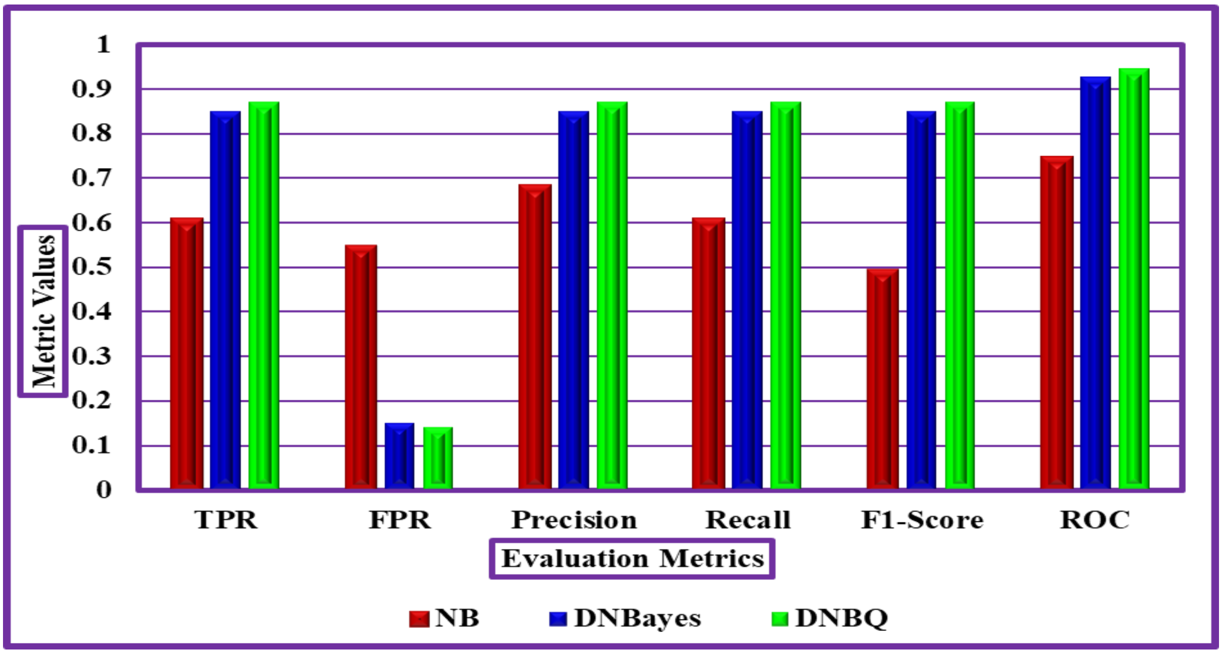

Figure 9: Evaluation Metrics for DNBQ

Table 7: Classified Samples and Classification Accuracy for DNBQ

Metrics & Algorithms	**No. of Samples**		**Classification Accuracy (%)**
	Correctly Classified	**Incorrectly Classified**	
NB	**78**	**50**	**61**
DNBayes	**109**	**19**	**85**
DNBQ	**111**	**17**	**87**

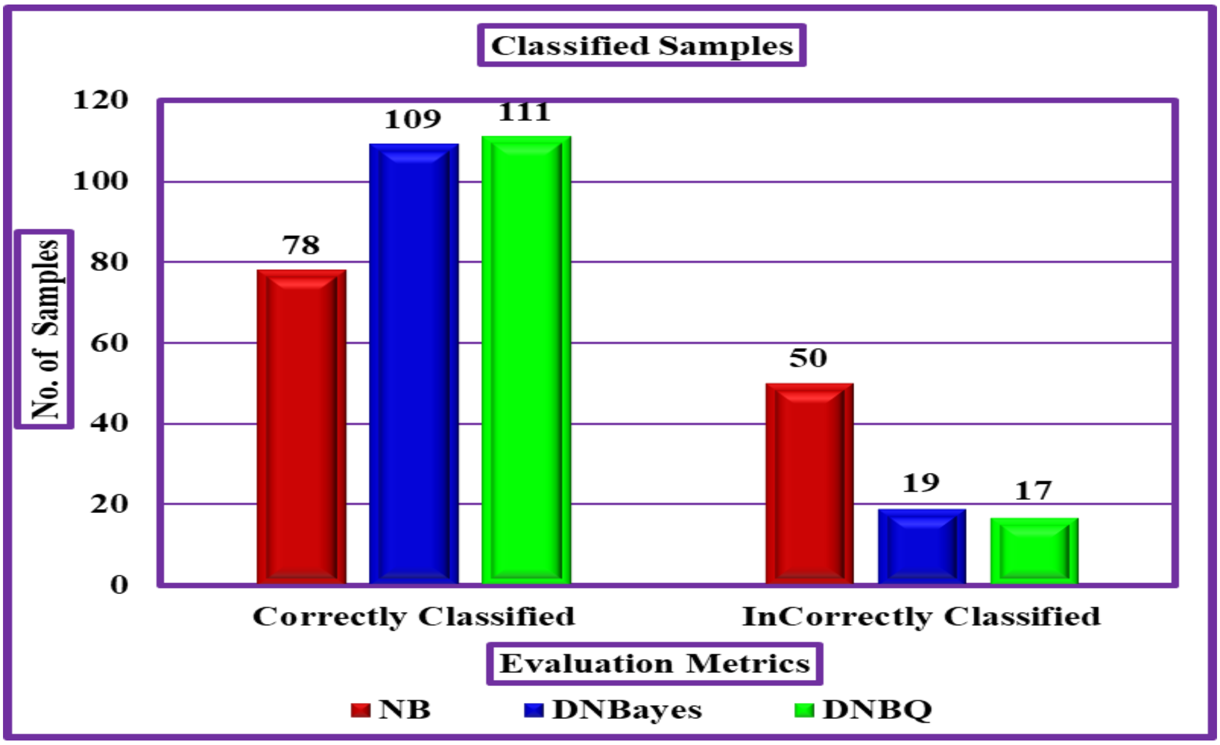

Figure 10: Number of Classified Samples for DNBQ

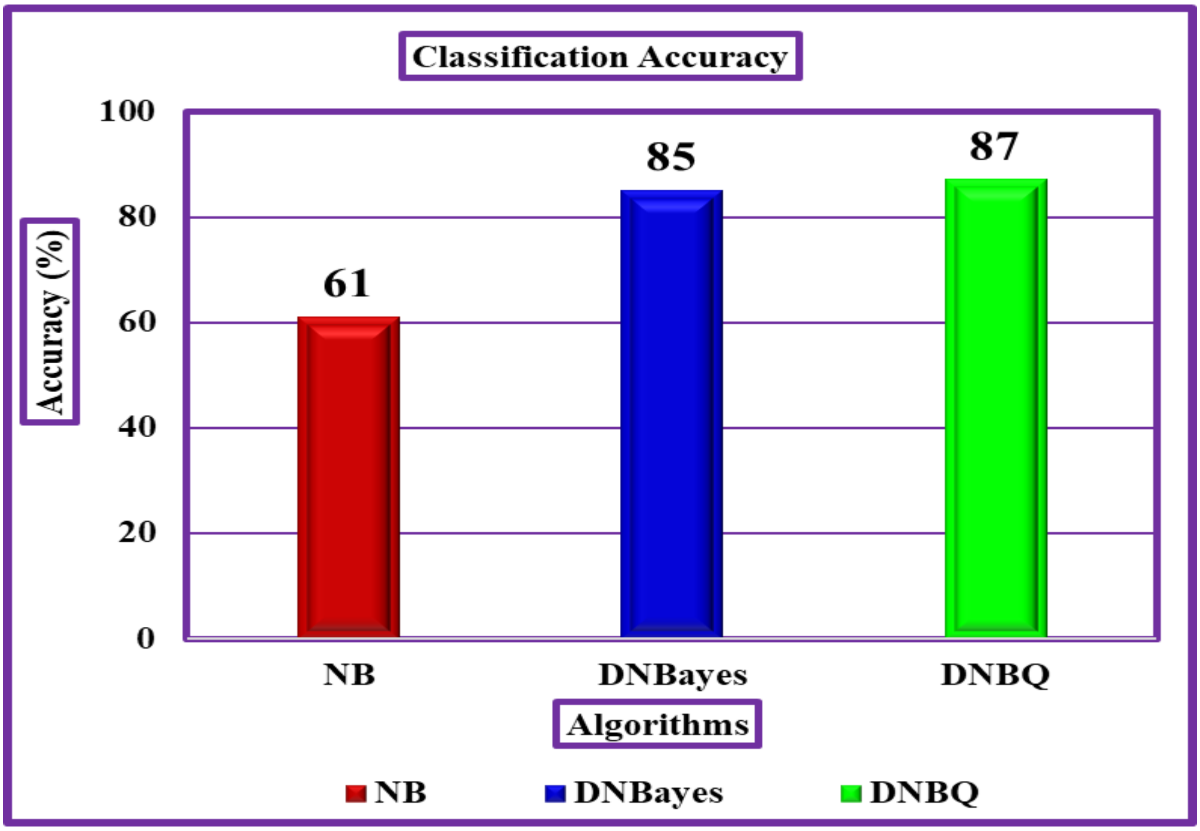

Figure 11: Classification Accuracy for DNBQ

Experimental results of Kappa Statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and evaluation metrics like True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1-Score, Receiver Operator Characteristic (ROC), classified samples and classification accuracy based on DNBQ are shown in (Table 5 - 7) also its graphical representation is shown in (Figure 8 - 11). The results reveal that DNBQ outperforms DNBayes by producing 87% of classification accuracy in a real-time dataset with 128 samples without outliers.

7. Conclusion and Future scope

Several methods are adopted for denoising the soil dataset to predict and classify as fertile and non-fertile soil. The proposed methodologies DNBayes and DNBQ provide the promising results as 85% and 87% respectively in terms of classification accuracy whereas NB, KNN and SVM give 61%, 64% and 62% of classification accuracy. Experimental results obtained for real-time soil dataset of 128 testing samples and have been proved that DNBQ outperforms DNBayes with respect to outlier detection and replacing the samples using various evaluation metrics. This supports soil scientists for decision making to help farmers in sugarcane cultivation. In future this work can be carried over with feature engineering for extracting relevant attributes to increase classification accuracy.

References

1. Dong, X., Tian, J., Zhang, R. H., He, D. X., & Chen, Q. M. (2017). Study on the Relationship between Soil Emissivity Spectra and Content of Soil Elements. Guang pu xue yu Guang pu fen xi= Guang pu, 37(2), 557-565.
2. Foth, H. D., & Ellis, B. G. (2018). Soil fertility. CRC Press.
3. Duraipaandiyaan, A. P., Sathyamoorthy, S., Vishnuvardhan, M., & Devi, S. S. (2021, May). An IoT Based System for Monitoring the Environment. In Journal of Physics: Conference Series (Vol. 1916, No. 1, p. 012162). IOP Publishing.
4. Ratih, I. D., Retnaningsih, S. M., & Dewi, V. M. (2021, April). Classification of soil quality using K-Nearest Neighbors methods. In IOP Conference Series: Earth and Environmental Science (Vol. 739, No. 1, p. 012011). IOP Publishing.
5. Baskar, S. S., Arockiam, L., & Charles, S. (2013). Applying data mining techniques on soil fertility prediction. International Journal of Computer Applications Technology and Research, 2(6), 660-662.
6. Fu, W., Zhao, K., Zhang, C., Wu, J., & Tunney, H. (2016). Outlier identification of soil phosphorus and its implication for spatial structure modeling. Precision Agriculture, 17(2), 121-135.
7. Gholap, J., Ingole, A., Gohil, J., Gargade, S., & Attar, V. (2012). Soil data analysis using classification techniques and soil attribute prediction. arXiv preprint arXiv:1206.1557.
8. Deshmukh, D. H., Ghorpade, T., & Padiya, P. (2015, January). Improving classification using pre processing and machine learning algorithms on NSL-KDD dataset. In 2015 International Conference on Communication, Information & Computing Technology (ICCICT) (pp. 1-6). IEEE.
9. Majumdar, J., Naraseeyappa, S., & Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: application of big data. Journal of Big data, 4(1), 1-15.

10. Devi, S. S., Gowtham, V., & Sharon, B. J. (2021, May). Daily stress classification using functional near infrared spectroscopy. In Journal of Physics: Conference Series (Vol. 1916, No. 1, p. 012161). IOP Publishing.
11. Barman, U., & Choudhury, R. D. (2020). Soil texture classification using multi class support vector machine. Information processing in agriculture, 7(2), 318-332.
12. Devi, S. S., Karthika, G. H., & Deepika, M. (2021, May). Machine Learning based Classification for Heart Disease Identification. In Journal of Physics: Conference Series (Vol. 1916, No. 1, p. 012174). IOP Publishing.
13. Taher, K. I., Abdulazeez, A. M., & Zebari, D. A. (2021). Data Mining Classification Algorithms for Analyzing Soil Data. Asian Journal of Research in Computer Science, 17-28.
14. Khan, M., Ding, Q., & Perrizo, W. (2002, May). k-nearest neighbor classification on spatial data streams using P-trees. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 517-528). Springer, Berlin, Heidelberg.
15. Liu, Y., Wang, H., Zhang, H., & Liber, K. (2016). A comprehensive support vector machine-based classification model for soil quality assessment. Soil and Tillage Research, 155, 19-26