Journal of Geoscience and Environment Protection, 2026, 14(5), 123-136
https://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Detection of Soil Chemical Profiles Using Supervised Learning: A Comparison of Ensemble Algorithms

Ahmed Babacar Sarr[1,2*], Mapathé Ndiaye[1,2], Sabou Sarr[1,2], Abdoulaye Cisse[1,2], Ndiouga Camara[1,2]

[1]Engineering Science Department, University Iba Der THIAM, Thiès, Sénégal
[2]Laboratoire de Mécanique et Modélisation, Thiès, Sénégal
Email: *ababacar.sarr@univ-thies.sn

How to cite this paper: Sarr, A. B., Ndiaye, M., Sarr, S., Cisse, A., & Camara, N. (2026). Detection of Soil Chemical Profiles Using Supervised Learning: A Comparison of Ensemble Algorithms. *Journal of Geoscience and Environment Protection, 14,* 123-136.
https://doi.org/10.4236/gep.2026.145009

Received: April 6, 2026
Accepted: May 25, 2026
Published: May 28, 2026

Abstract

Geophysical prospecting comprises a set of methods used to measure variations in a physical field or in the Earth's chemical potential. It plays a crucial role in subsurface exploration to characterize these heterogeneities. This enables the location of buried structures, lithologies or geological features to be determined. The aim of this study is to evaluate the effectiveness of machine learning algorithms in soil classification based on chemical properties. A series of samples was collected from various regions of Senegal and examined in the laboratory to determine their chemical compositions. Data pre-processing, ranging from cleaning missing values and handling inconsistencies to normalization and checking statistical consistency, was carried out on the data prior to the application of machine learning algorithms to generate classification reports. The results obtained following data processing show that ensemble methods, particularly random forests, are effective classifiers, with the X-gradient classifier and the bagging classifier achieving the highest classification accuracy of over 98%. These results demonstrate the relevance of using machine learning algorithms as tools for soil classification, complementing geotechnical studies. However, it should be noted that this study was conducted on samples whose chemical composition is known, and which were collected in an environment conducive to producing a specific chemical content depending on the soil type and region.

Keywords

Soil Classification, Geochemistry, Cluster, Geochemistry Factors, Supervised Learning Techniques

1. Introduction

There is growing interest in the research and development of machine learning and data mining techniques designed to facilitate geophysical studies and decision-making in geoscience, such as the classification of soils based on their geophysical parameters (Lim et al., 2020). Generally speaking, statistical learning methods are applied to field data in order to identify broad trends at both large-scale and local levels, thereby assessing the behavior of specific subsets. These results can be used in future scenarios to establish a classification system or to strengthen the methodological approach. These trained models can be used to assist engineers in guiding their decision-making (Reich et al., 1996; Kumar et al., 2025). Furthermore, these models can uncover previously unknown correlations between variables and the output, thereby improving knowledge and understanding of the soil. These correlations may lead to better interpretations or strategies for the utilization of the subsoil. Given that predictive models calculate forecasts based on information relating to a soil sample, but whose constituent minerals are common to that soil, they constitute promising tools for the purpose of soil classification. The potential of predictive models lies in their ability to generalize from training data. Despite their weakness in applying the same rules, they can process larger and more complex data whilst picking up on subtleties. This is because, in certain situations, they can prove counterintuitive (Naser, 2026). Predictive models rely on training data and depend on the quantity and quality of that data (Johnson & Dasu, 2003). A model extracts the existing signal from the data whilst ignoring noise. The measured data contain certain imperfections, some of which are related to the hardware and others to the system. Pre-processing these measurements involves cleaning and transforming the raw data into a coherent and usable signal. This improves the ability of predictive models to extract the actual signal from the measurements. Various methods exist depending on the specific requirements for improving the data (Varga & Kurko, 2010), such as imputation, outlier handling or scaling to prevent class dominance. As predictive models are based on digital data, choices must be made regarding encoding methods, dimensionality, or feature engineering (Zhao et al., 2018). These techniques aim to improve the performance of predictive models.

2. Material

2.1. Data

The data comes mainly from Senegal, The Gambia and Mali. These neighboring countries share a common geological heritage. This closely related geological heritage is linked to the geological continuity of a single basin. **Table 1** shows the distribution of acquisitions by country. These acquisitions are not made uniformly, but rather at sites carefully selected for their mineralogical potential.

The data collected comes from different countries but pertains to a single continuous geological zone. This zone consists of clay, sand, and granite, as shown in **Table 2**, which presents the quantities collected by soil type.

Table 1. Sample sizes by country.

	Sample	Pourcentage	%
Sénégal	406	87.12446352	%
Mali	8	1.716738197	%
Gambie	52	11.15879828	%
Total	466	100	%

Table 2. Sample sizes by soil type.

	Sample	Pourcentage	%
Argile Noire	130	27.89699571	%
Granite	123	26.39484979	%
Argile Blanche	72	15.45064378	%
Argile Rouge	68	14.59227468	%
Sable	73	15.66523605	%
Total	466	100	%

The samples collected reflect the diversity of soils, colors and, above all, the types of materials found in these different regions. This variation is explained by the fact that a large proportion of the data collected, particularly in the south, comes from lacustrine environments. These environments are characterized by a meandering river network, but also by significant plant decomposition of living organisms, making these soils very rich in organic matter. These data allow us to observe the ranges of variation in composition according to soil type and their mineral content. The diversity of the soils highlights their mineralogical variability depending on the conditions of formation. The data include, amongst other things: the content of silica (SiO_2), alumina (Al_2O_3), titanium monoxide (TiO_2), quicklime (CaO), sodium oxide (Na_2O), hematite (Fe_2O_3), magnesia (MgO) and potassium oxide (K_2O). Using these data, we calculated: the silica saturation index (IndSatSilice), soil alkalinity (indAlcalin), the proportion of iron relative to other oxides (IndFer), the basic potential (SomOxyBasik) and the total composition (Compotot).

In our study, the response variables correspond to the five soil types defined based on their mineralogical characteristics. The indices used provide additional information that facilitates a better understanding of the soils and their behavior.

This classification is based on soil type characteristics, without taking the region of origin into account. The Precambrian basement is an outcrop of hard rock located in eastern Senegal. This sedimentary basin provides a wealth of information on the soil science of Senegal. It stretches from Mauritania to Guinea-Bissau.

2.2. Data Collection and Analysis

As part of this study, data were collected from several targeted regions in Senegal,

The Gambia and Mali. Sampling sites were selected along riverbanks, in areas particularly representative of alluvial deposits and local hydrogeological dynamics. The collection operations were carried out using a pick-up truck. The tools used included a GPS, a spade, a pickaxe, an auger and bags for packaging the collected materials. Sampling depths varied depending on the material but generally ranged from 30 cm to 1.5 m. Sampling was carried out according to spatial variability and soil friability. As part of this study, scalar robust normalization was used to control outliers, standardize the scales of the variables and improve the convergence of the algorithms. Exploratory statistical analyses enable the data to be visualized in order to understand the influences and dominance of the variables. These collected samples are predominantly rich in clay, granite and sand. These are the major soil types present in the area. Table 3 shows the overall statistics for the data collected from the sites. It is on the basis of this data that the scaling will be applied.

Table 3. Mean and variance of the measured chemical parameters.

Statistique	SiO_2	Al_2O_3	TiO_2	CaO	MgO	Fe_2O_3	K_2O	Na_2O
mean	64.74	18.86	1.05	0.55	0.43	2.24	0.84	0.86
std	12.13	6.91	0.62	2.45	1.92	3.26	1.24	0.74

The compressive strength of samples taken from different soils ranges from very dense/rigid soil to very soft rock, with typical compressive strength values ranging from 0.95 to 5 MPa. Soil analysis plays a crucial role in nutrient management. Samples taken are subjected to laboratory analysis to determine their chemical composition. Chemical soil analysis serves to quantify the nutrients that should be available to plants and for human needs. Soil analysis involving the characterization of its constituents and the composition of its inorganic phases is referred to as chemical analysis. A chemical analysis involves the identification of several parameters such as: pH, redox potential, organic matter content, total nitrogen, calcium carbonate content, available phosphorus, boron, metals (Cu, Zn, Mn and Fe), exchangeable potassium, calcium, magnesium, sodium, and anion content (NO_3^-, SO_4^{2-}, PO_4^{3-}, Cl^-) (Tripathi et al., 2018). The above parameters are generally measured by soil analysis laboratories. They determine the soil category based on its mineralogical composition. The content of chemical elements is determined by inductively coupled plasma atomic emission spectroscopy (ICP-AES), with soil samples first being treated with acids. Table 1 shows the overall statistics for the data collected from the sites. It is on the basis of this data that the scaling will be applied.

3. Methodology

Classification can be defined as a mapping of features to an object. A classification

function $y = f(X)$ is responsible for classifying data points into hyperplanes. The inputs are represented by a vector X and the output by a class label y. Using instances of X and y, supervised learning attempts to train a classification model. The aim of the function is to map features to outputs. Due to the size of the raw data and the requirements of the study, it is split into batches for training, evaluation and validation.

The supervised classification task using machine learning is divided into four general stages:

- ✓ Pre-processing, which involves exploring the data,
- ✓ Training, which familiarizes the model with the classification features,
- ✓ Evaluation, which measures the model's performance,
- ✓ Validation, which assesses the model's ability to generalize.

For this study, the data were first subjected to various preprocessing steps and then divided as follows: 70% for the training set, 20% for the test set, and 10% for the validation set.

The strength of machine learning models lies in these hyperparameters. An optimal search helps to identify a more effective model. An unbiased evaluation quantifies the model's ability to classify samples not used during training. In other words, it assesses how well the model generalizes. Performance metrics for classifiers or regression models are used to assess predictive ability.

3.1. Random Forest Classifier

Random forests are based on the principle known as the wisdom of crowds. They combine multiple decision trees built on samples obtained via bootstrapping and subsets of features (Bose & Bose, 2025). The idea is to produce robust predictions by aggregating the results of the subsets known as trees. For classification, this involves a majority vote, and for regression, an average. This approach is robust to outliers in various fields such as geophysics, geology and geochemistry. Visual geological mapping is slow and biased; random forests enable the production of more reliable maps useful for targeting areas of interest (Darijani et al., 2022). Choosing the number of trees to ensure a balance between interpretability and algorithmic complexity remains one of the major challenges. Splitting the data into trees aims to reduce node impurity by using measures such as the Gini index or entropy. The optimal split is the one that maximizes the information gain.

3.2. Logistic Regression

Logistic regression is a primarily binary classification algorithm. It transforms a linear combination of variables into a probability using the sigmoid function. This probability is used to determine whether an example belongs to a given class. The characteristic parameters of a given sample are used to predict categorical outcomes (Giasson et al., 2006). For multi-class problems, it is extended using the one-versus-all strategy. This strategy generates a model for each class using the softmax function (Bewick et al., 2005). These extensions allow a probability vec-

tor, which sums to 1, to be assigned to several classes simultaneously whilst maintaining the model's interpretability. However, performance depends on the choice of strategy and the quality of the data, particularly for imbalanced classes.

3.3. Gaussian Naïve Bayes

The Naive Bayes algorithm is a simple and effective classifier, well-suited to high-dimensional data such as that encountered in soil classification, particularly for food crops (Kom & Kom, 2024). Its Gaussian version (GNB) assumes that the features are continuous and follow a normal distribution. This simplifies calculations but limits its applicability to real-world data, which do not adhere to this assumption of a normal distribution. To overcome this limitation, a variant known as Stable Naive Bayes (SNB) replaces the Gaussian distribution with stable distributions, capable of modelling asymmetry and outliers in real-world data. These stable distributions generalize the normal distribution whilst retaining its property of stability under addition. This makes them suitable for real-world or non-symmetric data. This version is defined by four main parameters: α, which controls the tails or outliers; β, the asymmetry of the data; γ, their scale; and δ, the location. This offers increased flexibility for feature analysis (Zeng & Pinsky, 2025). This extension of Naïve Bayes makes it more robust when dealing with complex distributions. Thus, SNB retains Bayesian simplicity whilst improving accuracy on real-world data.

3.4. Adaboost Classifier

This is an ensemble algorithm that combines weak classifiers to produce a more powerful predictive model. It operates through successive iterations, with each model being trained on weighted data. It demonstrates greater accuracy in lithology prediction (Sun et al., 2024). The subsequent iteration focuses on the misclassified samples from the previous one. The key hyperparameters of this estimator include the choice of base classifier, the number of iterations, and the learning rate, which controls the influence of each model. It is primarily used for binary or multi-class classification tasks. Its strength lies in its simplicity and efficiency, although it is sensitive to outliers in the measured data.

3.5. Gradient Boosting Classifier

Gradient Boosting is an ensemble method based on the principle of sequentially building weak models, known as bootstrapping. This process aims to correct the prediction errors of previous training runs. These models have improved the ability to predict the risk of landslides (Yasin et al., 2025). There is a version for regression, but in the case of classification, the algorithm used is the Gradient Boosting Classifier. Thus, compared to other algorithms, Gradient Boosting has demonstrated good performance based on wave velocities, density and natural gamma to target mineral deposits (Atita et al., 2022). Where each new decision tree is trained to improve the separation between classes by following the gradient of the loss

function. Both models apply the same logic, but regression tasks aim to progressively reduce the residual prediction error. The final model, in both classification and regression cases, is a weighted combination of several weak trees. This allows complex and non-linear relationships in real-world data to be captured. The key parameters are the loss function, the number of trees and the learning rate. These parameters must be carefully tuned to avoid overfitting. GBC and GBR offer great flexibility and remarkable performance.

3.6. X Gradient Boosting Classifier

XGB models are optimized variants of Gradient Boosting, designed to be faster and more efficient in terms of computational cost and algorithmic complexity on large datasets. As with the ensemble methods mentioned, they rely on the sequential construction of weak decision trees. Each new tree corrects the errors of the previous ones by following the gradient of the loss function. XGBoost introduces regularization mechanisms (L1 and L2) to reduce the risk of overfitting whilst improving the model's generalization. In classification, it assigns probabilities to classes. Based on class probabilities, it makes a final decision, whereas for regression, the decision is based on an optimization of numerical predictions. For estimating above-ground biomass, XGBR has demonstrated its potential in terms of accuracy compared to other methods (Pham et al., 2020). Key parameters include those for gradient descent as well as metaheuristic optimization. This study demonstrates the ability of XGBC to predict improvements in academic achievement through time management (Li, 2024). These results demonstrate that ensemble models are renowned for their performance, flexibility and ability to handle complex and large-scale data. In summary, XGBC and XGBR offer a combination of power, speed and robustness, which explains their widespread adoption in research, particularly in geoscience.

3.7. Voting and Bagging Classifier

The Voting Classifier is a technique based on the principle of the wisdom of crowds as applied to supervised models. It combines algorithms to produce predictions that are more robust than those obtained by a single model trained on the entire dataset. The decision is then made by majority vote (hard voting) or by combining probabilities (soft voting). Key parameters include the type of estimators and the weights assigned to them. Weighted voting is based on differential evolution, which improves classification accuracy and offers strong generalization ability as well as broad applicability (Zhang et al., 2014). Bagging is based on the principle of stabilizing models by combining multiple predictors trained on different batches of samples drawn from the same dataset. Each subsample is obtained through bootstrap sampling, which introduces diversity and robustness into the constructed models. Bagging is a technique designed simply to reduce the prediction error of machine learning algorithms by reducing the variance of unstable prediction methods (Liu et al., 2020).

4. Result and Discussion

In geochemistry, chemical elements act as indicators that provide clues about the physical, chemical or biological properties that led to the formation of the rock. The ratios between certain elements can indicate their relative abundance within a rock. Table 4 shows the contributions of the scaled data for these markers in the soil.

Table 4. Mean and variability of components by soil type.

Soil_Type	SiO_2		AL_2O_3		TiO_2		CaO		MgO		Fe_2O_3		K_2O		Na_2O	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Argile Blanche	73.70	15.44	13.49	9.41	1.37	0.58	0.27	0.17	0.08	0.07	2.36	5.01	0.12	0.05	0.47	0.12
Argile Noire	57.75	6.55	23.25	3.79	1.29	0.31	0.29	0.09	0.26	0.16	1.82	0.60	0.32	0.08	0.59	0.09
Argile Rouge	68.32	16.78	11.87	6.13	1.44	1.27	1.53	4.98	0.30	0.25	11.03	8.47	0.24	0.09	0.73	0.20
Granite	72.33	9.06	14.61	3.08	0.34	0.36	1.02	4.33	0.97	3.68	1.70	1.75	2.43	1.55	1.65	1.09
Sable	73.32	9.17	13.81	8.37	0.64	0.16	1.78	0.85	0.45	0.23	2.42	0.55	0.33	0.07	0.60	0.20

From this data, we will focus on three reports:

1) Silica saturation index = $SiO_2/(SiO_2 + Al_2O_3 + Fe_2O_3)$,

2) Alkalinity index = $Na_2O + K_2O/CaO + MgO$,

3) Iron index = $Fe_2O_3/(SiO_2 + Al_2O_3 + Fe_2O_3)$,

4) Basic oxide index = $CaO + MgO + Na_2O + K_2O$.

These indices characterize the cation saturation state of soil colloids. They have a significant influence on variations in the number of adsorbed cations. This makes it possible to compare the basic saturation index and the cation saturation index across a wide range of values. These indices enable the chemical state of a soil to be characterized, which determines the soil's saturation. The indices provided indicate a methodological approach to quantifying soil saturation. They serve as indicators of soil moisture tension depending on their concentration. The multivariate datasets contain numerous variables that characterize the richness of the geochemical data. This richness enables us to explain complex soil behaviors that cannot be understood from a single observation. Multivariate methods allow us to simultaneously study variations in composition observed across multiple sites. It is highly likely that the observed variations are linked to a smaller number of underlying causes. By exploiting the correlations present in the data, we have been able to identify more parsimonious underlying models. The histograms on the diagonal show the distribution of each oxide according to soil type. Alumina (Al_2O_3) is concentrated in clays, reflecting the presence of aluminosilicates.

Iron (Fe_2O_3) and titanium (TiO_2) oxides are more pronounced in red clays, as shown in Figure 1. From the scatter plots, we can identify the major oxides, which are effective markers for soil differentiation. Silica distinguishes sands from clays. Meanwhile, the content of alumina and/or iron oxide indicates variations in the

color of the clay or its formation conditions. Alumina, titanium oxide and iron oxide are strongly correlated. Meanwhile, sodium oxide and potassium oxide are also strongly correlated.

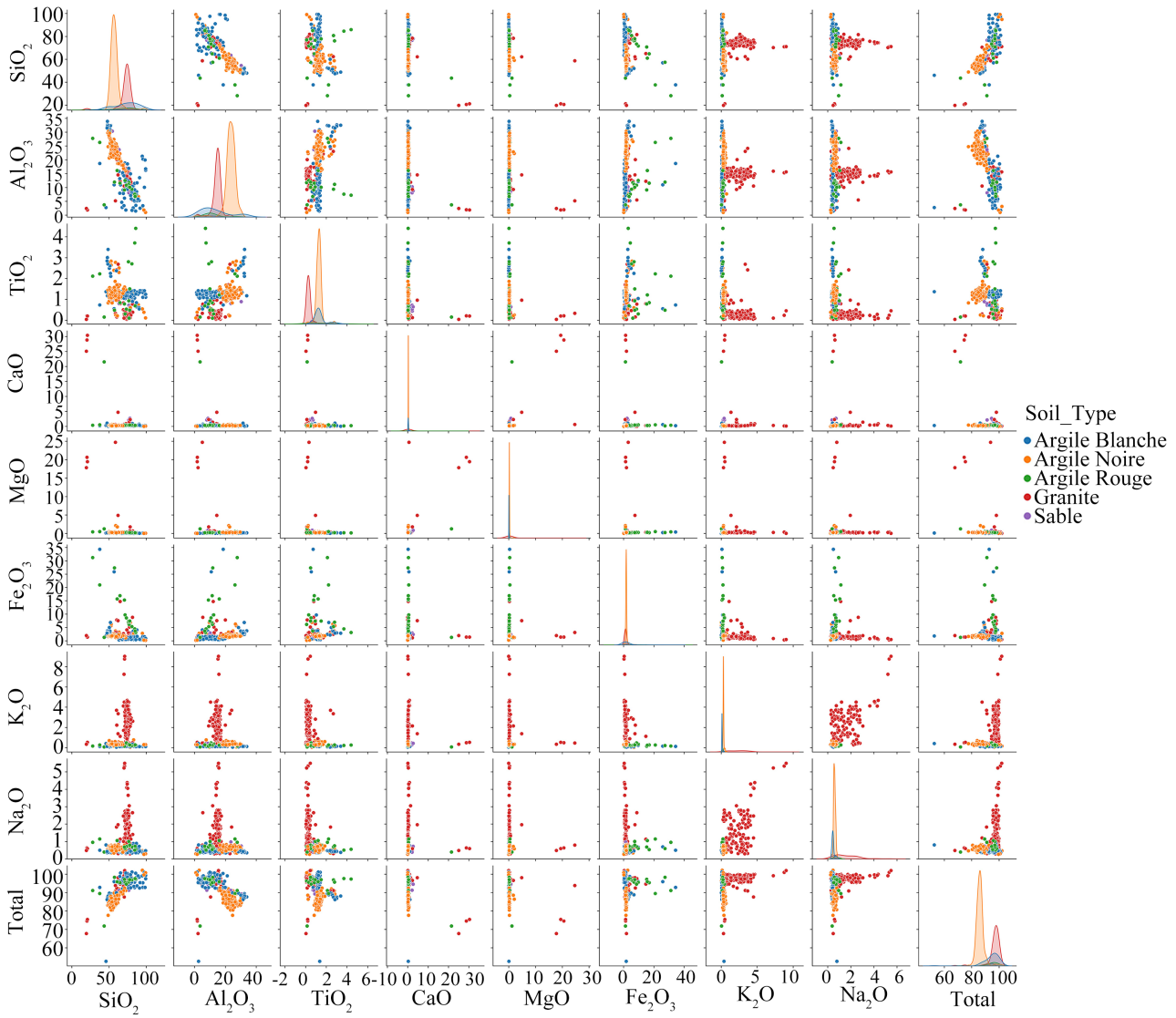

Figure 1. Multivariate analysis of geochemical signals.

4.1. Factor Analysis

Before applying the machine learning model, the features must be rescaled through normalization. This involves rescaling the features in the dataset so that they are centered around the median using interquartile rescaling. Robust scaling algorithms enable features to be scaled in a way that is resilient to outliers in the measured data. This method is very similar to the MinMax scaling method, but its distinctive feature is that it uses interquartile ranges rather than the minimum and maximum values used in MinMax scaling. This is a scaling algorithm that reduces the influence of outliers by using the median and the extremes of the data

based on the quantile range. For this scaling, each value of x is calculated relative to the first quartile Q_1 and then normalized relative to the interquartile range, as shown in the following equation:

$$x = \frac{x_i - Q_1}{Q_1 - Q_3}$$

An examination of the factors to interpret the underlying properties they represent shows, as illustrated in Figure 2, the following contributions by groups of variables:

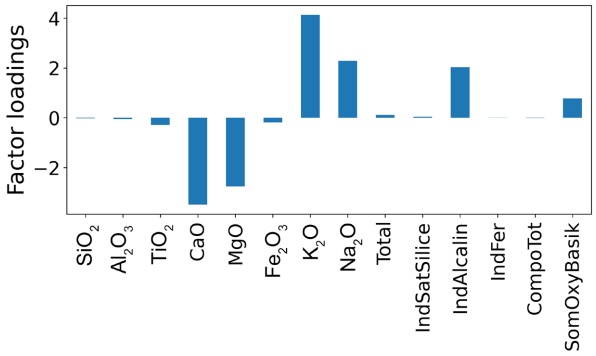

Figure 2. Factors interpretation.

- The variable components Fe_2O_3, MgO, CaO and IndFer reflect a dimension linked to coloring oxides and ferric indices. Soils with high levels of these components are often associated with the clays typical of central Senegal.
- The Al_2O_3, K_2O and Na_2O components indicate basic soils. These soils are the opposite of those dominated by compounds with a high iron index, reflecting a chemical contrast between ferric environments and the basic environments of the south.

These factors indicate ferric, basic, alkaline or siliceous soil types, distinguishing between soils rich in iron oxides and those that are more basic, magnesium-rich or rich in sodium and potassium oxides. In practice, this reveals an important geochemical dimension for soil classification. In this study, exploratory factor analysis is used to reduce the number of sample characteristics and identify the underlying factors.

4.2. Soil Clustering

Factor analysis of the data reduced the initial set of nine chemical characteristics to a set of five. These characteristics, grouped into adjacent sub-groups, enable four new characteristics to be derived for core samples based on their shared geochemical traits. These groups represent similar geochemical facies. Cluster analysis is a suitable approach for assigning a common facies label to similar samples. This clustering makes it easy to distinguish between samples. This batch analysis is a form of classification that falls broadly under the category of unsupervised machine learning. It is an approach used to infer a structure from the data meas-

ured within the dataset itself. Figure 3 illustrates the extent to which the silica and alumina content in soils in Senegal significantly influences their characterization. The data are reduced to principal components using the K-means approach, as shown in Figure 4. Each sample is thus assigned to a cluster representing geochemical facies. Examination of the signatures reveals characteristics specific to each cluster; for example, the iron index in the case of red clay soils. All these various transformations were applied to the raw data before it was divided into training, test, and validation sets.

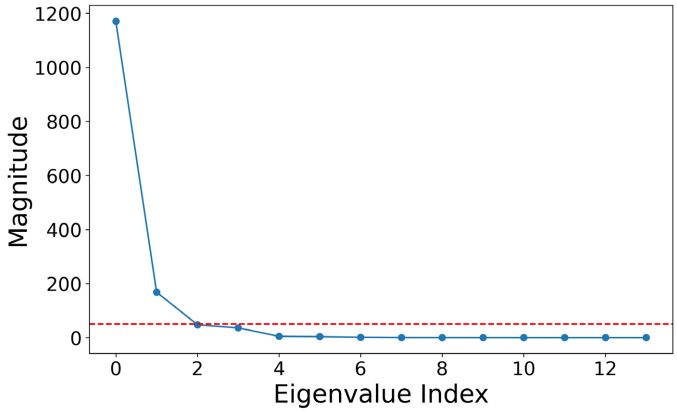

Figure 3. Retention threshold of the eigenvalue distribution.

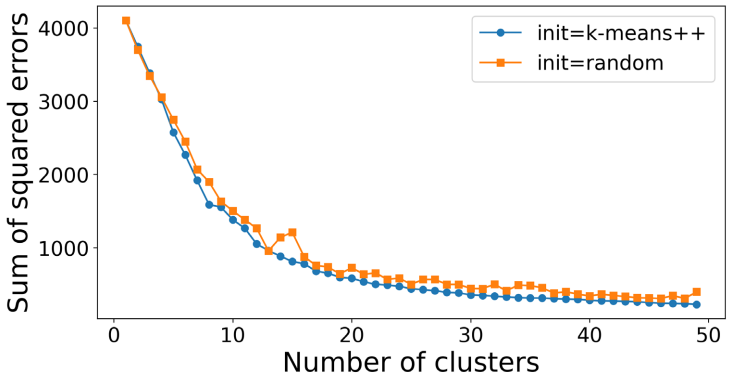

Figure 4. Comparison of k-means++ and random initialization.

4.3. Comparison of Classification Model Performance

In the context of the experiment, classification accuracy depends on the estimators' hyperparameters and their tuning. Table 5 compares the classification results, summarizing the performance of our models in terms of precision, recall, F1-score and accuracy. We observe that the random forest achieves the best overall classification performance with a precision of 99% and an accuracy of the same value. It is followed by X gradient boosting and the bagging classifier, which are very close behind (98% overall precision each). Logistic regression and the gradient boosting classifier follow closely in terms of overall precision. The Adaboost estimator provides the lowest overall accuracy in our classification.

Table 5. Comparative analysis of supervised methods.

	Setting	P	R	F1	Acc
Regression Logistique (rl)	C = 0.9	97	98	98	97
Random Forest (rf)	n_estimators = 200	99	99	99	99
Bagging Classifier (bgc)	n_estimators = 200	97	98	98	98
Adaboost Classifer (abc)	Algorithme = SAMME	64	76	78	76
Gradient Boosting Classifier (gbc)	Log_loss	85	92	94	97
X gradient Boosting Classifier (xgbc)	StritifiedKfold	91	97	94	98
Voting Classifier (vc)	rl, rf, gnb,	96	95	95	95

k = kernel; Acc = Accuracy; P = Precision; R = Recall; F1 = F1-score.

5. Conclusion

Over the past decade, geoscientific modelling approaches have attracted increasing interest, gradually shifting from statistical models and "box" models to Earth system models. However, statistical models often rely on a limited number of environmental factors to describe physical or chemical processes. Due to dynamic processes linked to climate and human activities, these parameters are non-linear as they vary considerably from one spatio-temporal domain to another. In this study, we introduced an ensemble algorithm framework to address the classification problem; this helps to facilitate soil recognition criteria by individual models and enhances the ability of specific environmental variables used in these models to better characterize the processes. The results lead to the following conclusions:

1) Ensemble learning methods have clearly demonstrated their potential for soil classification. Compared with traditional ensemble approaches, random forests achieved higher values for overall accuracy, precision, recall and F1-score in the estimation of soil classification parameters.

2) Multimodal datasets contain a high degree of similarity, which leads the model to make classification errors. Reducing this similarity through an in-depth exploration of soil properties would improve classification accuracy.

3) Soil is defined by these mechanical, physical or chemical properties. They help to improve the models, but rely on the effectiveness of machine learning classifiers. These classifiers require significant human intervention, such as the tuning of hyperparameters.

4) Models based purely on deterministic statistical data may appear to be incompatible with the known physical laws governing the Earth. This leads to the conclusion that the performance of such models does not accurately reflect the rheological reality of soil properties. The determination of physical stress by parameter would be an excellent tool for validating estimators in geoscientific classification.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Atita, O., Durrheim, R., & Saffou, E. (2022). Evaluation of Machine Learning Algorithms for the Classification of Lithology Using Geophysical Logs. In *NSG2022 4th Conference on Geophysics for Mineral Exploration and Mining* (pp. 1-5). European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.202220121

Bewick, V., Cheek, L., & Ball, J. (2005). *Statistics Review 14: Logistic Regression. Critical Care, 9,* Article Number 112. https://link.springer.com/article/10.1186/cc3045

Bose, S., & Bose, S. (2025). Random Forests: The Wisdom of Crowds in Action. *Journal of Emerging Trends in Computer Science and Applications, 1,* 67-91. https://doi.org/10.65525/jetcsa.v1i1.5

Darijani, M., Farquharson, C. G., & Perrouty, S. (2022). A Random Forest Approach to Predict Geology from Geophysics in the Pontiac Subprovince, Canada. *Canadian Journal of Earth Sciences, 59,* 489-503. https://doi.org/10.1139/cjes-2021-0089

Giasson, E., Clarke, R. T., Inda Junior, A. V., Merten, G. H., & Tornquist, C. G. (2006). Digital Soil Mapping Using Multiple Logistic Regression on Terrain Parameters in Southern Brazil. *Scientia Agricola, 63,* 262-268. https://doi.org/10.1590/s0103-90162006000300008

Johnson, T., & Dasu, T. (2003). *Data Quality and Data Cleaning: An Overview.* https://doi.org/10.1145/872874.872875

Kom, N. S., & Kom, M. (2024). *Classification System for Soil Types Suitable for Food Crops using Naïve Bayes Method. Sistemasi: Jurnal Sistem Informasi, 13,* 1102-1113. https://doi.org/10.32520/stmsi.v13i3.3956 **https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/24**

Kumar, S., Machireddy, J. S., Sankaran, T., & Sholapurapu, P. K. (2025). Integration of Machine Learning and Data Science for Optimized Decision-Making in Computer Applications and Engineering. *Journal of Information Systems Engineering and Management, 10,* 748-759. https://jisem-journal.com/index.php/journal/article/view/8990

Li, S. (2024). Exploring the Impact of Time Management Skills on Academic Achievement with an XGBC Model and Metaheuristic Algorithm. *International Journal of Advanced Computer Science and Applications, 15,* 107-119. https://doi.org/10.14569/ijacsa.2024.0150710

Lim, C. S., Mohamad, E. T., Motahari, M. R., Armaghani, D. J., & Saad, R. (2020). Machine Learning Classifiers for Modeling Soil Characteristics by Geophysics Investigations: A Comparative Study. *Applied Sciences, 10,* Article 5734. https://doi.org/10.3390/app10175734

Liu, H., Jia, J., & Gong, N. Z. (2020). On the Intrinsic Differential Privacy of Bagging. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Main Track*, Yokohama, 7-15 January 2021, 2730-2736. https://doi.org/10.24963/ijcai.2021/376

Naser, M. Z. (2026). When to Use Machine Learning? And Which Problems Stand to Benefit? *Urban Lifeline, 4,* Article No. 3. https://doi.org/10.1007/s44285-025-00054-3

Pham, T. D., Le, N. N., Ha, N. T., Nguyen, L. V., Xia, J., Yokoya, N. et al. (2020). Estimating Mangrove Above-Ground Biomass Using Extreme Gradient Boosting Decision Trees Algorithm with Fused Sentinel-2 and ALOS-2 PALSAR-2 Data in Can Gio Biosphere Reserve, Vietnam. *Remote Sensing, 12,* Article 777. https://doi.org/10.3390/rs12050777

Reich, Y., Medina, M. A., Shieh, T., & Jacobs, T. L. (1996). Modeling and Debugging Engineering Decision Procedures with Machine Learning. *Journal of Computing in Civil Engineering, 10,* 157-166. https://doi.org/10.1061/(asce)0887-3801(1996)10:2(157)

Sun, Y., Pang, S., & Zhang, Y. (2024). Application of Adaboost-Transformer Algorithm for Lithology Identification Based on Well Logging Data. *IEEE Geoscience and Remote Sensing Letters, 21,* 1-5. https://doi.org/10.1109/lgrs.2024.3372513

Tripathi, D., Alonso-Pérez, M. O., & Tiwari, D. K. (2018). Rapid Diagnosis of Soil Nutrients Using Microscopic Techniques. *Microscopy and Microanalysis, 24,* 680-681. https://academic.oup.com/mam/article/24/S1/680/6945654

Varga, M., & Kurko, K. (2010). Some Methods of Data Improvement in EIS. In *2010 32nd International Conference on Information Technology Interfaces (ITI)* .

Yasin, K. C., Niasari, S. W., Nukman, M., Sudarmaji,, Irnaka, T. M., Hartantyo, E. et al. (2025). Integration of Geoscientific Data Mining and GBM Hyperparameter Optimization to Improve the Accuracy of Landslide Hazard Prediction. In *2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)* (pp. 367-370). IEEE. https://doi.org/10.1109/caibda65784.2025.11183569

Zeng, X., & Pinsky, E. (2025). Stable Distribution Naive Bayes Achieves Higher Accuracy than Traditional Naive Bayes Classification. *International Journal on Cybernetics & Informatics, 14,* 71-86. https://doi.org/10.5121/ijci.2025.140205

Zhang, Y., Zhang, H., Cai, J., & Yang, B. (2014). A Weighted Voting Classifier Based on Differential Evolution. *Abstract and Applied Analysis, 2014,* 376950. https://onlinelibrary.wiley.com/doi/10.1155/2014/376950

Zhao, Z., Shi, D., Huo, H., & Fang, T. (2018). Feature Encoding Methods Evaluation Based on Multiple Kernel Learning. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing (pp. 209-213).* ACM. https://doi.org/10.1145/3195106.3195152