This is a previous version of the article published in Marine Georesources & Geotechnology. 2023. https://doi.org/10.1080/1064119X.2023.2251025

GRADIENT BOOSTING TREES WITH BAYESIAN OPTIMIZATION TO PREDICT ACTIVITY FROM OTHER GEOTECHNICAL PARAMETERS

Esteban Díaz[1], Giovanni Spagnoli [2*]

[1] Departamento de Ingeniería Civil. Escuela Politécnica Superior, Universidad de Alicante, P.O. Box 99, E-03080 Alicante, Spain, esteban.diaz@ua.es

[2] DMT GmbH & Co. KG, Am TÜV 1, 45307 Essen, Germany, giovanni.spagnoli@dmt-group.com ORCID: 0000-0002-1866-4345

* Corresponding author.

Abstract

Clay swell potential can be classified based on the value of activity and it is defined as the ratio of plasticity index to clay content as a percentage. This paper outlines the investigation into how activity correlates with other key properties of clayey soils. Specifically, four approaches were evaluated for predicting activity using: a) liquid limit (LL), specific surface area (SSA), cation exchange capacity (CEC) and clay content; b) LL, SSA and CEC; c) LL; and d) SSA and CEC. For this purpose, a database of 104 samples was collected from which 35 machine learning algorithms were trained. Gradient Boosting Trees showed the highest prediction accuracy in the four approaches and, to improve its prediction performance, a Bayesian Optimization was applied to tune theirs hyperparameters, resulting in the final models. The performance of the developed models was evaluated, showing prominent results with exceptionally good metrics, except in the approach from SSA and CEC where the trained algorithm was not capable of predicting activity with confidence (R^2=0.46). This algorithm can predict activity using only LL with high accuracy (R^2=0.94), and when combined with SSA and CEC, the precision is further enhanced (R^2=0.96). Finally, a variable importance analysis was performed, indicating LL is the variable with the greatest influence in predicting activity.

Keywords: machine learning; liquid limit; specific surface area; cation exchange capacity; clay content; activity.

1. Introduction

Key properties of clays such as Atterberg limits, Specific Surface Area (SSA) and Cation Exchange Capacity (CEC), are important in geotechnical engineering and in particular for characterizing expansive soils. Expansive soils are a very problematic matter in several fields of civil engineering, posing significant amount of damage (Jones Jr and Holtz 1973). Expansive soils are those in which the variation of water content results in a large volume change (Kar 2021). The construction on these soils often generates serious issues, especially when lightweight structures are built. There are several approaches to estimate the expansive potential of a soil. Expansive soils can be identified by the Atterberg limits, clay content, or a combination of both. Skempton (1953) proposed the concept of activity, i.e. the ratio of plasticity index (PI) to clay fraction content, which can be utilized as an index property to establish the swelling potential of expansive soil. Peck et al. (1974) correlated PI with the expansion potential. Zapata et al. (2006) suggested that considering the % passing 75μm improved the prediction. Different authors developed other types of tests to identify expansive soils (e.g. Lambe 1960; Yao et al. 2004). SSA (Chittoori and Puppala 2011) and CEC (Mitchell and Soga 2005; Nelson et al. 2015) are also used to indirectly identify expansive soils. According to Low (1987), the surface's level of hydration has a significant impact on how clays behave. This makes the SSA of a clayey soil extremely significant, and it has been proposed that SSA can be used to forecast the engineering behaviour of fine-grained soils (e.g. Warkentin 1972). The principal clay minerals, such as montmorillonite, illite, and kaolinite, have a significant impact on the SSA of soils, which reflects the consistency traits as well as the clay concentration (Spagnoli and Shimobe 2019). Muhunthan (1991) attempted a rheological correlation between SSA and LL. CEC and the soil's swelling characteristics are strongly connected. With an increase in CEC, there is also an increase in soil swelling (Christidis 1998). The higher the presence of smectitic clay minerals (i.e. expansive clays), the more the clay swells. According to Al-Rawas (1999), the cations are what regulate how expansive soils are. Therefore, both LL and CEC are variables that regulate the soils' propensity to swell (Spagnoli and Shimobe 2019). Chittori and Puppala (2011) suggested that

since SSA of smectitic soils is higher than those of kaolitic soils, SSA can be used to predict the expansion potential, as the water holding capacity is higher.

Although activity is not a complex or difficult parameter to obtain, it is a relevant property of clayey soils correlated with swelling potential. Several authors have proposed a correlation between soil activity and other geotechnical parameters (e.g. Polidori 2009; Spagnoli and Shimobe 2019). However, in general, the proposed geotechnical correlations consider only two parameters at the same time (input and output). Machine learning (ML) techniques have progressively emerged as an alternative approach to address numerous geotechnical challenges (e.g. Díaz et al. 2018; Díaz et al. 2021; Díaz and Tomás 2021; Phoon and Zhang 2023; Salvatore et al. 2022; Wang et al. 2020; Zhang et al. 2022a; Zhang et al. 2023; Zhang et al. 2022b). Motivated by these advantages, this paper employs machine learning techniques to explore potential correlations between soil activity and other properties of clayey soils (LL, SSA, CEC and clay content). To achieve this, the study focused on four approaches based on various properties associated with activity, comparing their outcomes to evaluate the performance of each. The objective is to understand clearly which properties can be utilized to determine activity with an appropriate degree of accuracy. To this end, four prediction models to forecast activity were developed using a dataset of experimental results. With this dataset, a comparative study of various ML algorithms was carried out. The algorithms that performed best underwent Bayesian optimization to determine the appropriate model hyperparameters. The results of the final tuned models in predicting activity are then evaluated and discussed. Finally, an importance analysis of the variables is conducted to identify the most important parameters in the prediction of activity.

2. Database

104 data points regarding clay content <2μm, SSA (m^2/g), CEC (meq/100g), and LL obtained by means of the Casagrande cup were acquired from 12 different publications and a single datapoint belonging to the authors. The activity value was obtained from the data. Pure clays and natural clays were selected, in order to have a relatively high heterogeneity in the data. Specifically data from (Arifin 2008), Cerato (2001), Cerato and Lutenegger (2002), Cerato and Lutenegger (2004),

Cerato and Lutenegger (2005), Marcial (2013), Mishra et al. (2012), Santhoshkumar et al. (2016), Schwing et al. (2013), Sivapullaiah et al. (2008), Spagnoli et al. (2013) and Zhang et al. (2003) were used.

2.1. Statistical description

Some descriptive statistical information is provided in Table 1. Skewness and kurtosis are two measures of the shape of a distribution in statistics. Skewness measures the degree of asymmetry in a distribution, while kurtosis measures the degree of peakedness or flatness. Skewness is defined as the third standardized moment of a distribution and can be positive (skewed right), negative (skewed left), or zero (symmetric). Positive skewness indicates that the tail of the distribution is longer on the right-hand side, while negative skewness signifies that the tail is longer on the left-hand side. Kurtosis is defined as the fourth standardized moment of a distribution and can be high (leptokurtic) indicating that the distribution has a sharp peak and fat tails, or low (platykurtic) kurtosis signifying a flatter distribution with fewer outliers (Wackerly et al. 2014).

Figure 1 presents the box plots for the data. A box plot is a visual representation of the distribution of a dataset that shows the median, quartiles, and outliers. It consists of a rectangular box, which extends from the first quartile (Q1) to the third quartile (Q3), with a vertical line inside that represents the median. The distance between the upper and lower edges of the box, known as the interquartile range (IQR), contains the middle 50% of the data. The whiskers extend from the box to the minimum and maximum values within a certain range of the dataset. The range is typically set at 1.5 times the IQR (Gelman and Hill 2006). Figure 1 shows several outliers, which are data points outside of this range and are plotted individually as dots outside the whiskers. Figure 2 shows selected histograms of the data. The goodness of fit for the probability distributions is conducted using the Anderson-Darling (AD) test. The AD is a statistical test based on the idea of comparing the cumulative distribution function (CDF) of the sample data to the CDF of the theoretical distribution being tested (Hollander, et al., 2015). Notably, with the exception of clay

content, none of the data adheres to the normal distribution. AD gives more weight to the tails than does the Kolmogorov-Smirnov (KS) test (Stephens 1974).

Statistics	Clay content (%) <2µm	SSA (m^2/g)	CEC (meq/100g)	LL (%)	Activity
Count	104	104	104	104	104
Mean	56.039	285.459	37.834	181.361	2.189
Median	57	225.5	28.25	74	0.99099
Mode	36.2	15	2	42	0.44
Minimum	10.2	11	0.8	24	0.2
Maximum	100	800	120.9	678	8.0132
Skewness	-0.18	0.49	0.81	1.14	1.15
Kurtosis	-0.61	-1.12	-0.135	-0.132	-0.053
Standard Deviation	20.23	235.25	31.07	180.77	2.24

Table 1. Descriptive statistics of the data analyzed.

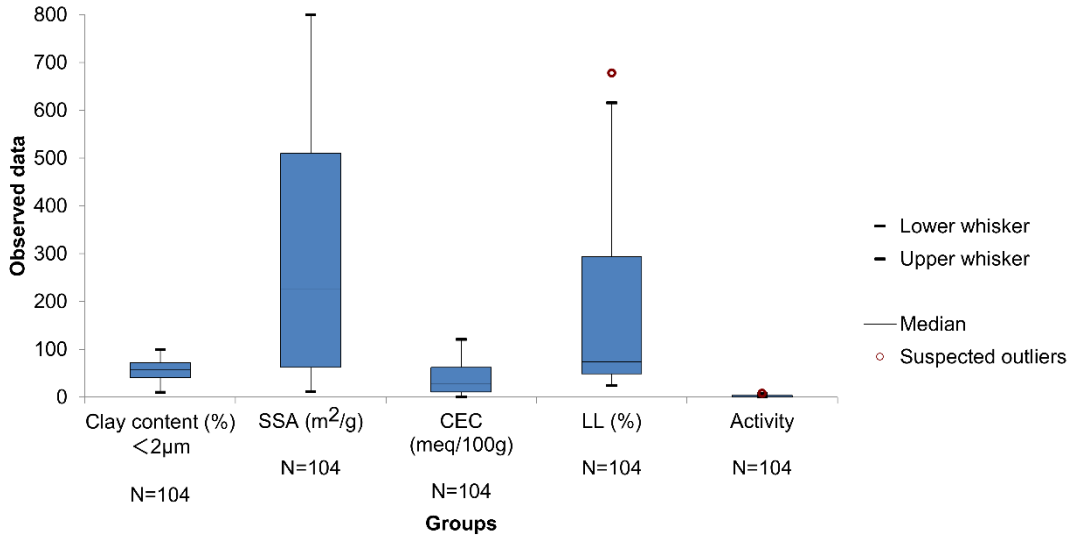


Figure 1. Box plots of the data analyzed.

For the data presented in Table 1, the clay content values follow a 2-Param Weibull distribution, while the remaining parameters follow a 3-Param Weibull distributions (see histograms in Figure 2). Additionally, a normality test has been performed for all data sets. A normality test is a statistical test used to determine whether a given set of data comes from a normally distributed population. Normality tests are used to check the assumption of normality, which is often made in statistical inference procedures such as hypothesis testing and confidence interval estimation

(Devore et al. 2013). The normality test used was the Shapiro-Wilk test, which is based on the deviation between the observed data and the expected normal distribution.

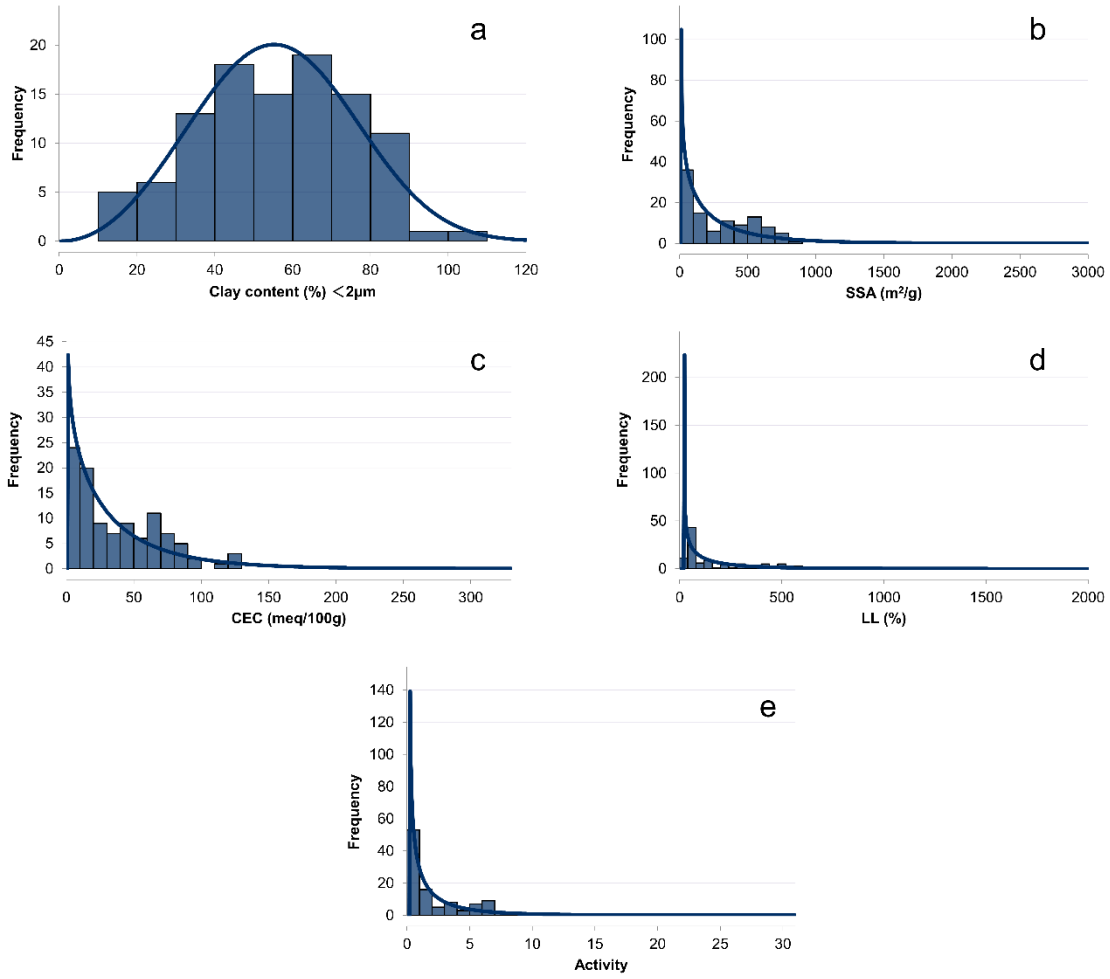
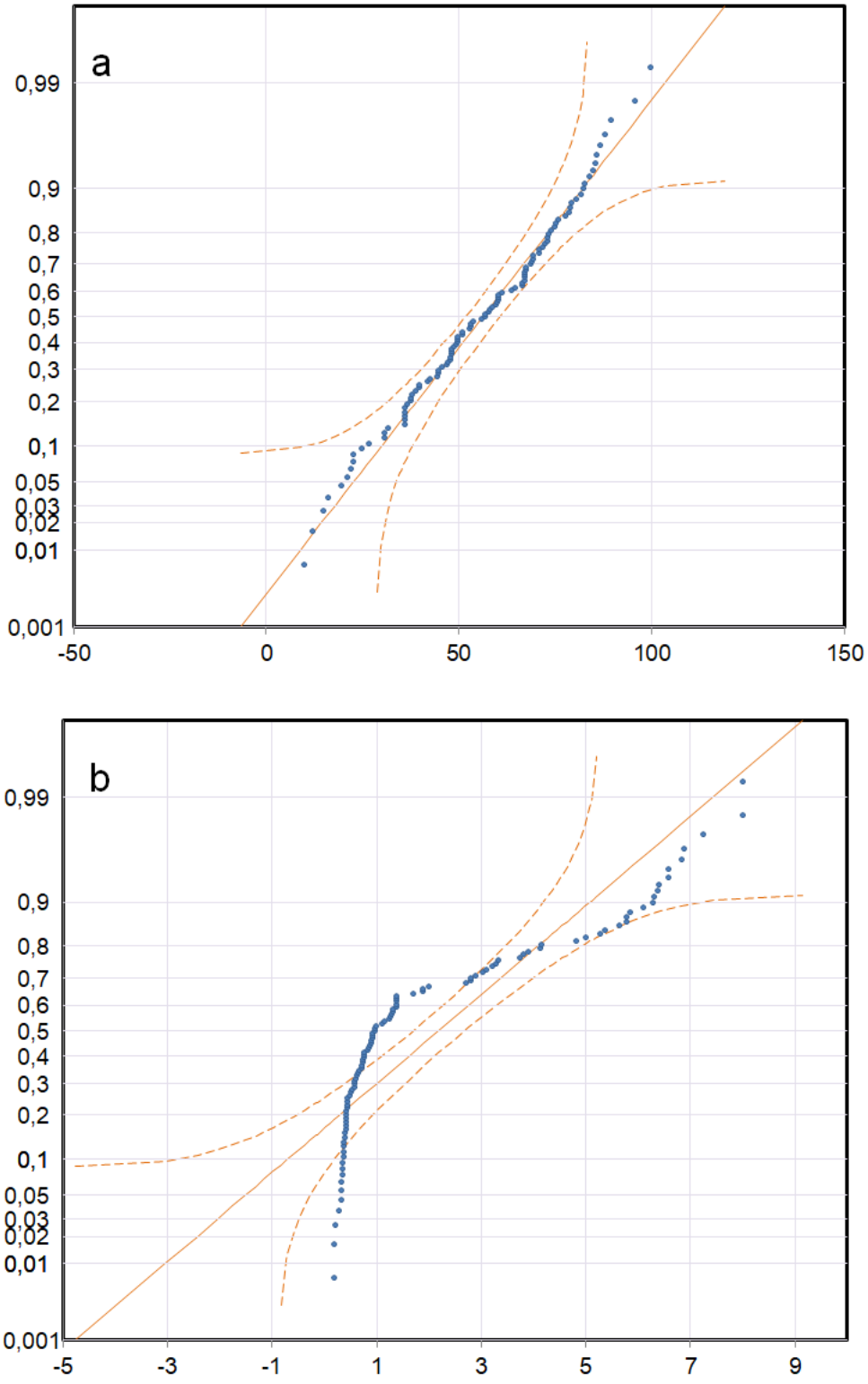


Figure 2. Histograms for the data from Table 1 for a) clay content, b) SSA, c) CEC, d) LL, and e) activity

Figure 3 shows the probability plots for clay content and activity, as an example. A probability plot is a graphical technique used in statistics to assess whether a set of data follows a particular distribution. It compares the ordered values of a dataset to the expected values of a theoretical distribution. If the data follows the distribution, the plot should show a roughly straight line. If the data does not follow the distribution, the plot will show deviations from the straight line (Moore and McCabe 1989). In a probability plot, the x-axis represents the expected values of the

theoretical distribution being tested, while the y-axis represents the ordered values of the dataset being tested. The points on the plot are plotted based on their rank order, such that the smallest observation is plotted at the far left and the largest observation is plotted at the far right. The closer the points are to the straight line, the better the fit of the data to the theoretical normal distribution being tested. The probability plots for CEC, SSA and LL are similar to Figure 3b. It is possible to observe that while the data follow a straight line for clay content (Figure 3a), this is not the case for activity (Figure 3b).



145

146 *Figure 3. Probability plots for a) clay content and b) activity. To note the difference considering*

147 *their normality path.*

148

3. Methodology.

With the 104 samples described in the previous section, the four variables previously presented were used to predict activity of clayey soils. Specifically, four approaches have been studied taking into account different variables (Table 2) and these predict activity from: 1) LL, clay content, SSA and CEC, 2) LL, SSA and CEC, 3) LL, and 4) SSA and CEC.

Approach 1	Approach 2	Approach 3	Approach 4
LL	LL	LL	SSA
Clay content	SSA	-	CEC
SSA	CEC	-	-
CEC	-	-	-

Table 2. Variables considered in each of the approaches considered in the prediction of activity.

3.2. Machine learning methods

3.2.1. Gradient boosting trees

Gradient boosting trees are a powerful family of machine learning algorithms for performing gradient descent on decision trees using the boosting ensemble learning method. The main idea behind them is to combine iteratively several simple models (i.e. weak learners) to obtain a model with enhanced prediction accuracy (i.e. strong learner). Boosting algorithms were initially proposed for classification tasks (Freund 1995; Freund and Schapire 1996; Schapire 1990). Friedman (2001) expanded the boosting to regression tasks by creating the gradient boosting machines method (GBM). The boosting method adjusts the weights of the training sample according to the last iteration and assigns more weight to observations that are difficult to predict and less weight to those that have already been well managed. It can be understood as a numerical optimization algorithm aiming to find an additive model that maximally reduces the loss function. The GBM algorithm builds successive decision trees to fit one training example at a time (the tree that best reduces the loss function). As it fits each new sample, it updates its knowledge of which features are important for the prediction of future samples. It starts with an initial estimation for model parameters and iteratively enhances these estimations until a required level of accuracy has been achieved or some other stopping criterion has been satisfied. Specifically, in regression problems, the algorithm begins by initializing the model with a first prediction, which is a decision

tree that maximally minimizes the loss function (mean squared error in regression), then at each stage a new decision tree is fitted to the existing residual and added to the prior model to update the previously obtained residuals. The algorithm keeps on iterating until the prefixed maximum number of iterations is reached. This process is called stage wise, meaning that at each new stage, the decision trees included into the model at previous steps remain unchanged. With the process of fitting decision trees to the residuals, the algorithm is enhanced in the zones where it does not perform well. Four hyperparameters mainly govern behaviour of the GBM: 1) the learning rate, 2) the number of boosting stages to perform, 3) the number of features to consider when looking for the best split, and 4) the maximum depth of the individual regression estimators.

3.2.2. Bayesian hyperparameter optimization

The performance of the machine learning algorithms is strongly determined by the model parameters (i.e. hyperparameters) which need to be set before training. In this study, the optimal hyperparameters were established applying Bayesian optimization (Shahriari et al. 2015; Snoek et al. 2012) which is a general technique for function optimization. Bayesian optimization builds a probability model based on the previous evaluation results of the target for finding the value that minimizes the objective function. Bayesian hyperparameter optimization was performed using BayesSearchCV from the Scikit-optimize package on Python (Head et al. 2020). This method uses stepwise Bayesian Optimization to discover the most promising hyperparameters in the problem-space. This optimization method was chosen for hyperparameter tuning due to its efficiency (Eriksson et al. 2019) and because it has been proven to be superior to other optimization algorithms on many optimization benchmark functions (Jones 2001). Finally, it must be remarked that this technique has been extensively used for machine learning hyperparameter tuning in geotechnics (e.g. Li et al. 2022; Zhang et al. 2021).

3.2.3. Verification and evaluation of the machine learning models

In this paper, a training set was employed to choose and build the predictive models and a test set was used to examine the trained models in each of the four approaches analysed. The coefficient

of determination (R^2), the mean square error (MSE) and the mean absolute error (MAE), were applied with the aim of evaluating the reliability of the algorithms considered and to interpret the correspondence between predictions and observed values. The definition of R^2, MSE and MAE is expressed by (Equations 1 to 3):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

where y is the measured value, $\hat{y}$ is the model predicted value, $\bar{y}$ is the average of the measured values, and n is the number of samples in the training or testing sets.

On the other hand, k-fold cross-validation (Stone 1974), was used in the Bayesian optimization. In the k-fold cross-validation approach, the dataset is randomly shuffled and then divided into k folds. k-1 folds are used to train the model and the remaining fold (the test set) is employed for the evaluation. The process is repeated k times, and performance of the model is evaluated by the mean prediction error of k sub-datasets. In this study, the optimization was done with a 5-fold cross-validation.

3.2.4. Data preparation. Feature rescaling

To improve the performance of the machine learning algorithms, the dataset was pre-processed. In particular, the input parameters of the dataset were standardized using the min-max scaling, which involves rescaling the range of features to scale the range in [0, 1]. Using this data rescaling method, the impact of parameters with different scales on the algorithm performance can be minimised.

3.2.5. Machine learning algorithm selection

35 ML algorithms were built and trained in each of the four approaches considered, using the Scikit-learn package (Pedregosa et al. 2011) which is the most useful and robust library for machine learning in Python. The initial selection included: SVR, Random Forest Regressor, Extra Trees Regressor, AdaBoost Regressor, NuSVR, Gradient Boosting Regressor, K-Neighbors Regressor, Histogram-based Gradient Boosting Regressor, Bagging Regressor, MLP Regressor, Huber Regressor, Linear SVR, Ridge CV, Bayesian Ridge, Ridge, Linear Regression, Transformed Target Regressor, Lasso CV, Elastic Net CV, Lasso Lars CV, Lasso Lars IC, Lars CV, Lars, SGD Regressor, RANSAC Regressor, Elastic Net, Lasso, Orthogonal Matching Pursuit CV, Passive Aggressive Regressor, Gaussian Process Regressor, Orthogonal Matching Pursuit, Decision Tree Regressor, Dummy Regressor, Lasso Lars and Kernel Ridge.

3.3. Model conception

To choose the optimal model to predict activity in each of the four approaches considered, the next phases were followed:

1. Building a database, collecting data from different research papers. In this phase, 104 samples of clayey soils containing values of LL, clay content, SSA, CEC and activity were gathered.
2. Rescaling of the input variables.
3. Application of 35 Machine Learning Algorithms: Using the chosen inputs for each of the considered approaches.
4. Identification of the best models, considering R^2 as the main statistical performance indicator.
5. Optimizing the best models for each of the four approaches using Bayesian optimization.
6. Assessing the predictive capability of the four selected models considering the test set.
7. Performing a feature importance analysis of the selected models to find the inputs with a higher influence on the predictions.

4. Application and results

4.1. Predictive comparisons among different algorithms

In this section, from the 35 machine learning algorithms considered, a selection process is carried out, choosing R^2 as the reference metric. Table 3 displays the results of this analysis, but only highlights the top ten results for each of the considered approaches. From the analysis of the comparative study, it can be deduced that for the four approaches, the algorithm with the best performance is the Gradient Boosting Regressor Trees (GBRT).

Ranking	Approach 1		Approach 2		Approach 3		Approach 4	
	Algorithm	**R^2**	**Algorithm**	**R^2**	**Algorithm**	**R^2**	**Algorithm**	**R^2**
1	Gradient Boosting Regressor	0.98	Gradient Boosting Regressor	0.96	Gradient Boosting Regressor	0.95	Gradient Boosting Regressor	0.45
2	Extra Tree Regressor	0.96	Extra Trees Regressor	0.94	Bagging Regressor	0.94	Passive Aggressive Regressor	0.44
3	Gaussian Process Regressor	0.95	Decision Tree Regressor	0.93	AdaBoost Regressor	0.94	Extra Trees Regressor	0.43
4	Random Forest Regressor	0.93	Random Forest Regressor	0.93	SVR	0.94	MLP Regressor	0.42
5	Linear SVR	0.93	RANSAC Regressor	0.93	Random Forest Regressor	0.94	Linear SVR	0.42
6	Huber Regressor	0.93	Linear SVR	0.93	Decision Tree Regressor	0.93	AdaBoost Regressor	0.41
7	Ridge	0.93	Huber Regressor	0.93	XGB Regressor	0.93	Poisson Regressor	0.40
8	Passive Aggressive Regressor	0.92	Bagging Regressor	0.93	Nu SVR	0.93	Gamma Regressor	0.40
9	Elastic Net CV	0.92	AdaBoost Regressor	0.92	Linear SVR	0.93	SGD Regressor	0.36
10	RANSAC Regressor	0.92	Lars	0.92	RANSAC Regressor	0.93	Random Forest Regressor	0.36

Table 3. Ranking of the first ten algorithms with the best coefficient of determination according to the analysis of the 35 initially selected.

Since the GBRT selected models are decision tree-based algorithms and these are fairly insensitive to the scale of the features, the developed models were compared with and without

rescaling. The results without scaling of features suggested that it had virtually no impact on the GBRT results.

4.2. Bayesian optimization

With the four GBRT models previously selected, a Bayesian optimization process was performed to find the hyperparameters that give the models better predictive accuracy. In this Bayesian searching process, the R^2 was chosen as the reference metric. The cross-validation folds were created by a stratified fold with a splitting number of 5 and an iteration number of 70. With this process, the hyperparameters were established using cross-validation on the training set and the predictions were performed on the test set. The values of the hyperparameters obtained in this optimization process for each of the approaches taken into consideration are shown in Table 4.

Hyperparameter	**Approach 1**	**Approach 2**	**Approach 3**	**Approach 4**
Learning rate	0.07427	0.01259	0.00194	0.00118
Maximum depth	2	4	2	1
Number of estimators	1013	894	1881	1830
Number of features	None	None	None	2

Table 4. Values of the hyperparameters in the final GBRT models obtained by Bayesian optimization for each of the approaches considered.

4.3. Performance analysis

Following the fine-tuning of the hyperparameters, the selected models with the best performance in each one of the approaches considered were analysed with the test dataset. In Table 5 is included a summary of the performance metrics obtained in the training and test sets of the definitive tuned models.

Approach	**R^2**		**MAE**		**MSE**	
	Training	**Test**	Training	**Test**	Training	**Test**
1	0.999	**0.984**	0.02	**0.19**	0.01	**0.08**
2	0.999	**0.964**	0.02	**0.28**	0.01	**0.16**
3	0.934	**0.944**	0.40	**0.42**	0.34	**0.26**
4	0.667	**0.459**	1.01	**1.26**	1.71	**2.45**

290 *Table 5. Summary of accuracy parameters in the train and test sets of the models for each*
291 *approach.*

292

293 Figure 4 examines the relationships between the models' predictions and the actual activity with
294 a linear regression analysis. For a perfect fit, all data should fall along a 1:1 line, as the model
295 outputs would be equal to the measured values. An excellent fit was obtained for the first two
296 approaches with values of R^2>0.96. A slightly lower level of accuracy is obtained in the approach
297 3 although it still gets good metrics (R^2>0.94). Approach 4 does not offer good results and it can
298 be concluded that there is no reliable way to predict activity from CEC and SSA because less than
299 half of the variance in the outcome variable is explained by the model (R^2=0.46).
300 On the other hand, from the analysis of the MAE values in the test set, the proposed GBRT
301 deviates, on average, from the predictions by ±0.19, ±0.28, ±0.42 and ±1.26 respectively for each
302 of the respective approaches. It should be noted when analyzing these values that approximately
303 40% of the test set has activity values exceeding 2.5, reaching up to values of nearly 7 (Figure 4).
304 The values of the first three approaches show accurate models, but in the fourth approach the
305 value is too high, showing a notable dispersion in the prediction of activity. Approach 1 yielded
306 the best predictive performance and includes as predictor variables clay content (directly related
307 to activity) and LL (associated with activity through PI). Nevertheless, Approach 2, which no
308 longer includes clay content, and Approach 3 (using only LL) also presented outstanding
309 performance metrics. This is an important observation, as this algorithm enables the prediction of
310 soil activity solely based on LL without significant errors, or even minimizing these errors if SSA
311 and CEC are additionally available alongside LL, rendering further soil properties unnecessary.

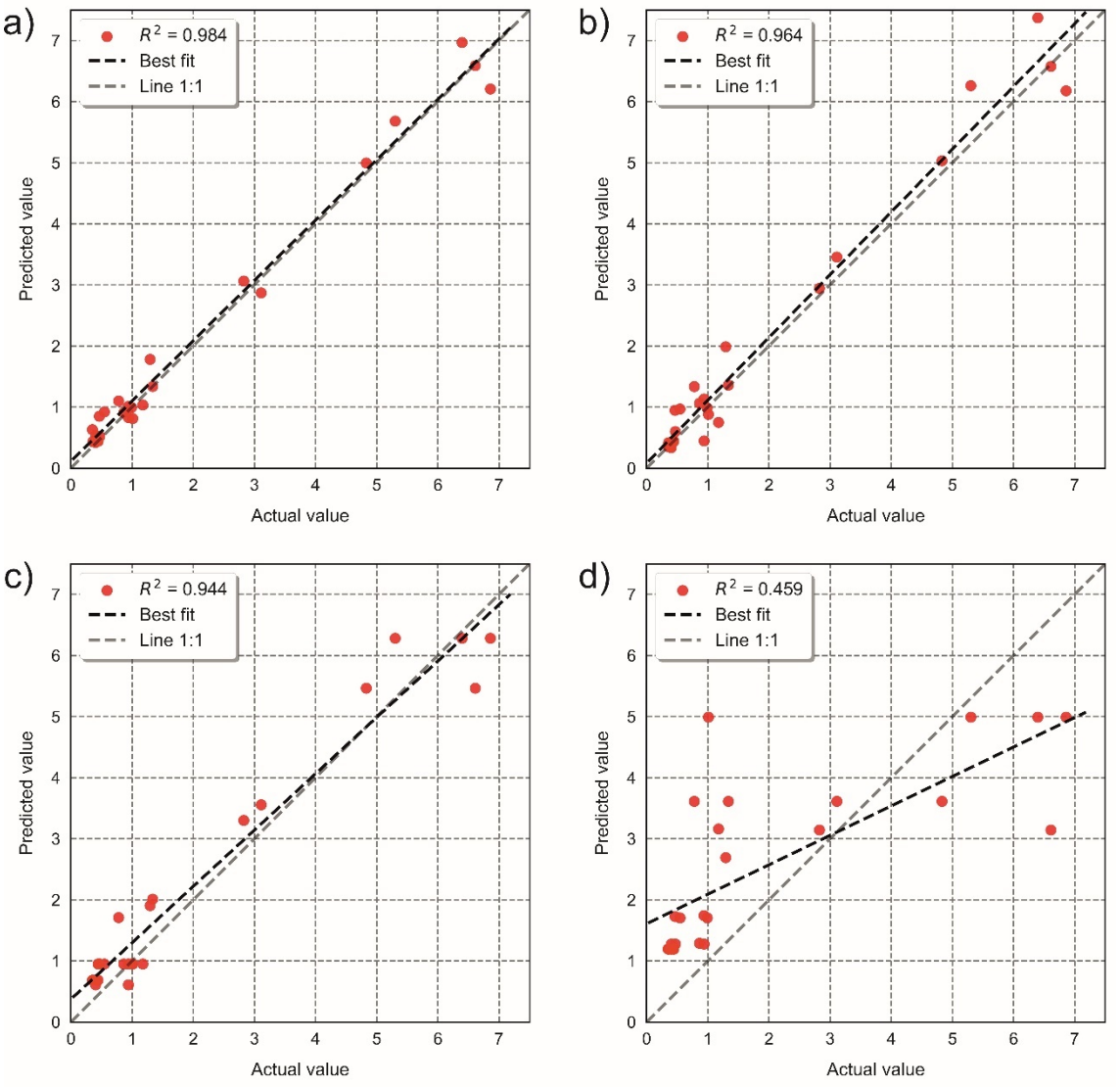


Figure 4. Performance of the selected models for the training and test datasets in each of the approaches considered, a) activity from LL, clay content, SSA and CEC, b) activity from LL, SSA and CEC, c) activity from LL, and d) activity from SSA and CEC.

Figure 5 presents the residuals of the GBRT model predictions for each of the four approaches, which are calculated as the difference between predicted and observed values. Additionally, the figure showcases the quantile-quantile (Q-Q) plots for these residuals. A Q-Q plot contrasts the quantiles of a given dataset with the quantiles of a theoretical probability distribution, in this case, the Gaussian distribution.

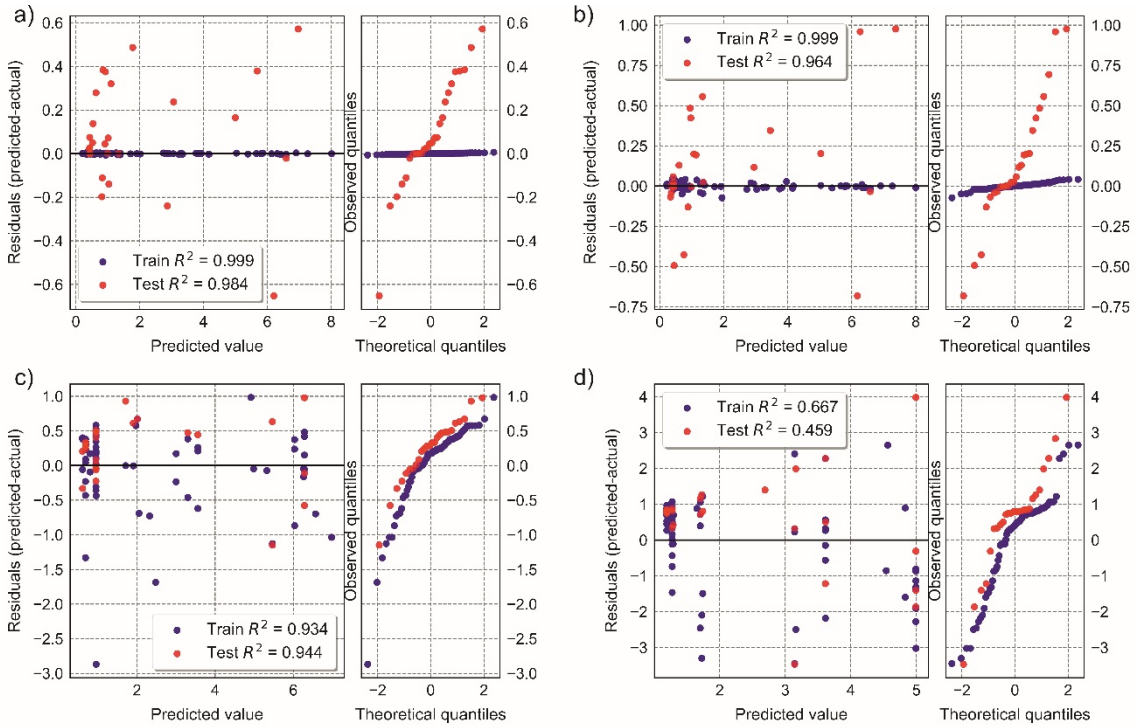


Figure 5. Residuals and Q-Q plots of ach of the approaches considered, a) activity from LL, clay content, SSA and CEC, b) activity from LL, SSA and CEC, c) activity from LL, and d) activity from SSA and CEC.

From examining the Q-Q plots, it can be inferred that in the test dataset, the residuals follow a distribution close to Gaussian although with a slight tendency towards positive residuals in the first two approaches. In approaches 3 and 4, the residuals move away from the Gaussian distribution, showing a light negative skewed distribution. These trends to Gaussian distributions are desirable since the algorithms, on average, predict the values with low error, and there are no extreme deviations in the predictions.

The evolution of the predicted activity in the test dataset for the four approaches considered is shown in Figure 6, compared with the measured values. Approaches 1, 2, and 3 accurately capture the evolution of actual values. In these 3 approaches, there are no areas where the algorithm tends to have less reliable predictions, both in the maximum and in the minimum values of activity, the difference between the predicted and the actual values is relatively small. Instead, in approach 4,

the predictive power is less, and the algorithm seems not to capture well the prediction of the high values and it can fail the prediction to identify large values of activity.

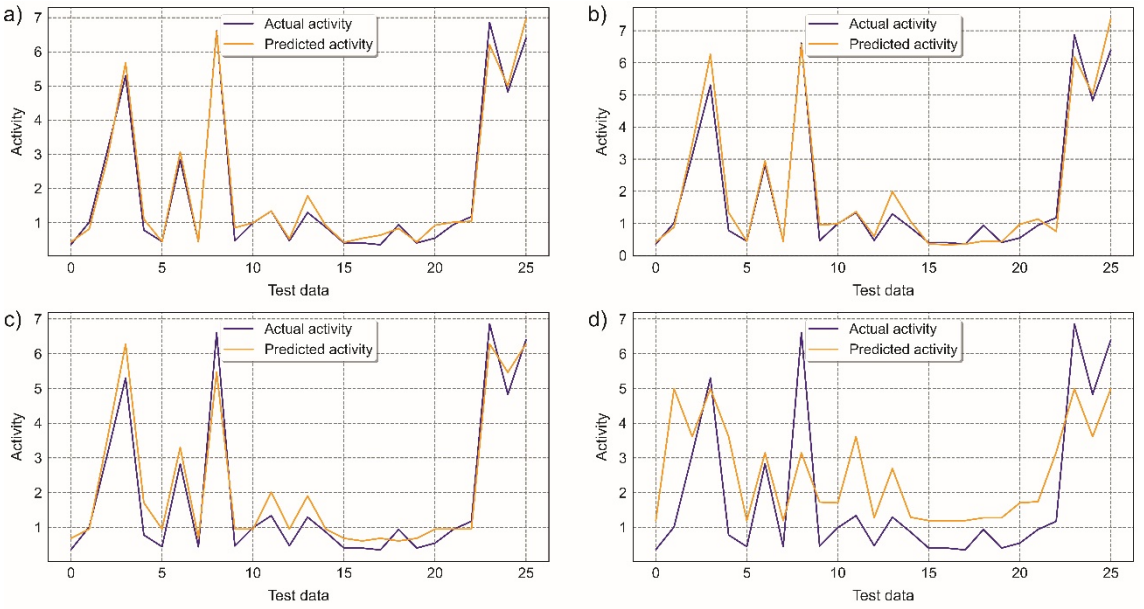


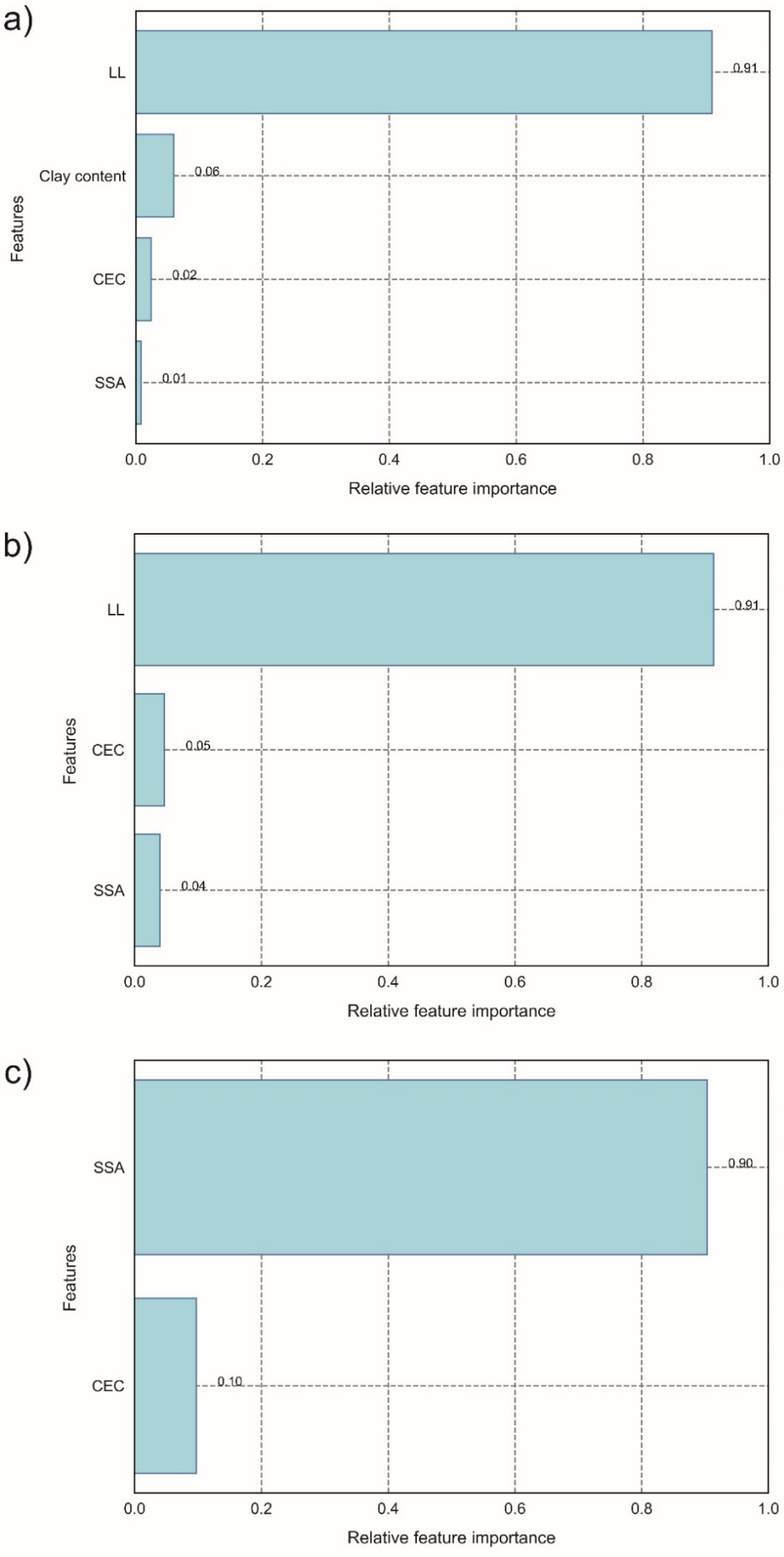
Figure 6. Activity prediction results for the test dataset in each of the approaches considered, a) activity from LL, clay content, SSA and CEC, b) activity from LL, SSA and CEC, c) activity from LL, and d) activity from SSA and CEC.

4.4. Variable importance in the prediction of activity

The feature importance is an important tool for the model interpretability, providing an evaluation of the predictive capacity of the input variables which can help to know the contributions of these to the output of the model. The trained GBRT models can automatically calculate feature importance, which can be obtained through the Gini Importance or mean decrease in impurity (Breiman et al. 1984). This method determines each feature importance as the sum over the number of splits in all the trees that include a specific feature, proportionally to the number of samples it splits.

Figure 7 shows the acquired importance order of the input variables in the approaches considered, taking into consideration that the approach 3 only considers one variable, so this analysis is not necessary. A higher value compared to another means greater importance of a feature for making

a prediction. In the developed GBRT models, LL is clearly the most important feature variable in the approaches 1 and 2, and the rest of the variables have a less impact on the prediction of activity. This is an interesting observation, as in approach 1 our model gives greater importance in the prediction to LL than to clay content, which is a variable that by definition is proportional to the value of activity. However, LL is also related to plasticity index and therefore to activity, and this strong relationship has already been shown in numerous works (e.g. Spagnoli and Shimobe 2019). In the approach 4, SSA has a significant impact in the prediction of activity, while the contribution of CEC is small.


a)
LL
Clay content
CEC
SSA
Features
0.91
0.06
0.02
0.01
0.0
0.2
0.4
0.6
0.8
1.0
Relative feature importance
b)
LL
CEC
SSA
Features
0.91
0.05
0.04
0.0
0.2
0.4
0.6
0.8
1.0
Relative feature importance
c)
SSA
CEC
Features
0.90
0.10
0.0
0.2
0.4
0.6
0.8
1.0
Relative feature importance

371

Figure 7. Importance ranking of the input parameters in the activity prediction for each of the approaches considered, a) activity from LL, clay content, SSA and CEC, b) activity from LL, SSA and CEC and c) activity from SSA and CEC.

5. Conclusions.

In geotechnical engineering, dealing with expansive soils is crucial due to their challenging swelling and shrinking behaviors. This study introduces a novel approach, utilizing GBRT to predict activity in clayey soils, a property related to swelling potential. Such predictive capabilities have direct applications in pre-anticipating soil behavior, leading to safer and more efficient construction and infrastructure planning. To achieve this aim four approaches to predict activity were considered using different input variables: 1) LL, clay content, SSA and CEC, 2) LL, SSA and CEC, 3) LL, and 4) SSA and CEC. This research paper has offered several key insights and contributions:

1. GBRT outperformed the 35 algorithms that were initially selected and evaluated. These algorithms were subsequently tuned using a Bayesian optimization process, obtaining the definitive algorithms for the four approaches considered.
2. In the first three approaches, models yield high prediction accuracy. Approach 1, with its inclusion of clay content (linked to activity) and LL (related via PI), had the highest predictive accuracy ($R^2 = 0.98$).
3. The notable performance of Approaches 2 and 3 suggests that soil activity can be reliably predicted using only LL ($R^2 = 0.94$), with even greater precision when combined with SSA and CEC ($R^2 = 0.96$).
4. The fourth approach, i.e. predicting activity from SSA and CEC, did not show satisfactory results in any of the models analysed, obtaining an R^2 value of 0.46, being able to conclude that there is no reliable way to predict activity from SSA and CEC.
5. The conducted feature importance analysis indicated that LL is the most influential variable in predicting activity for Approaches 1 and 2, with other variables having a lesser impact. In Approach 4, SSA is the primary contributor to the prediction of activity.

Finally, the potential for scalability and adaptability of the proposed algorithms might be subject to further detailed investigation. By incorporating a larger dataset, the proposed algorithms could be refined, potentially extending the relevance of the current work. This could also create the way for accounting for the effects of additional variables, marking a direction for future research.

Availability statement

The data supporting the research results can be accessed at https://huggingface.co/datasets/EstebanDC/activity-clays or can be obtained by contacting the corresponding author.

References

Al-Rawas, A.A. 1999. The factors controlling the expansive nature of the soils and rocks of northern Oman. Engineering Geology, **53**, 327-350.

Arifin, Y.F. 2008. *Thermo-hydro-mechanical behavior of compacted bentonite-sand mixtures: an experimental study*. https://nbn-resolving.org/urn:nbn:de:gbv:wim2-20081006-14288. doi: https://doi.org/10.25643/bauhaus-universitaet.852.

Breiman, L.I., Friedman, J.H., Olshen, R.A. & Stone, C.J. 1984. Classification and regression trees (cart). Encyclopedia of Ecology, **40**, 582-588.

Cerato, A. 2001. Influence of specific surface area on geotechnical characteristics of fine-grained soils. Unpublished MSc Thesis, Department of Civil and Environmental Engineering, University of Massachusetts.

Cerato, A.B. & Lutenegger, A.J. 2002. Determination of surface area of fine-grained soils by the ethylene glycol monoethyl ether (EGME) method. Geotechnical testing journal, **25**, 315-321.

Cerato, A.B. & Lutenegger, A.J. 2004. Determining intrinsic compressibility of fine-grained soils. Journal of Geotechnical and Geoenvironmental Engineering, **130**, 872-877.

Cerato, A.B. & Lutenegger, A.J. 2005. Activity, relative activity and specific surface area of fine-grained soils. *Proceedings of the International Conference on Soil Mechanics and Geotechnical Engineering*. AA Balkema Publishers, 325.

Chittoori, B. & Puppala, A.J. 2011. Quantitative estimation of clay mineralogy in fine-grained soils. Journal of Geotechnical and Geoenvironmental Engineering, **137**, 997-1008.

Christidis, G.E. 1998. Physical and chemical properties of some bentonite deposits of Kimolos Island, Greece. Applied Clay Science, **13**, 79-98.

Devore, J.L., Farnum, N.R. & Doi, J.A. 2013. Applied statistics for engineers and scientists. Cengage Learning.

Díaz, E., Brotons, V. & Tomás, R. 2018. Use of artificial neural networks to predict 3-D elastic settlement of foundations on soils with inclined bedrock. Soils and Foundations, **58**, 1414-1422, doi: https://doi.org/10.1016/j.sandf.2018.08.001.

Díaz, E., Pastor, J., Rabat, Á. & Tomás, R. 2021. Machine learning techniques for relating liquid limit obtained by Casagrande cup and fall cone test in low-medium plasticity fine grained soils. Engineering Geology, **294**, 106381.

Díaz, E. & Tomás, R. 2021. Upgrading the prediction of jet grouting column diameter using deep learning with an emphasis on high energies. Acta Geotechnica, **16**, 1627-1633, doi: https://doi.org/10.1007/s11440-020-01091-8.

Eriksson, D., Pearce, M., Gardner, J., Turner, R.D. & Poloczek, M. 2019. Scalable global optimization via local bayesian optimization. Advances in neural information processing systems, **32**.

Freund, Y. 1995. Boosting a weak learning algorithm by majority. Information and computation, **121**, 256-285.

Freund, Y. & Schapire, R.E. 1996. Experiments with a new boosting algorithm. *icml*. Citeseer, 148-156.

Friedman, J.H.J.A.o.s. 2001. Greedy function approximation: a gradient boosting machine. 1189-1232.

Gelman, A. & Hill, J. 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge university press.

Head, T., Kumar, M., Nahrstaedt, H., Louppe, G. & Shcherbatyi, I. 2020. scikit-optimize/scikit-optimize: v0. 8.1. Zenodo.

Jones, D.R. 2001. A taxonomy of global optimization methods based on response surfaces. Journal of global optimization, **21**, 345-383.

Jones Jr, D.E. & Holtz, W.G. 1973. Expansive soils-the hidden disaster. Civil engineering, **43**.

Kar, K.K. 2021. Handbook of Fly Ash. Butterworth-Heinemann.

Lambe, T.W. 1960. The Character and Identification of Expansive Soils: A Report Completed for the Technical Studies Program of the Federal Housing Administration. Federal Housing Administration.
Li, D., Liu, Z., Xiao, P., Zhou, J. & Jahed Armaghani, D. 2022. Intelligent rockburst prediction model with sample category balance using feedforward neural network and Bayesian optimization. Underground Space, **7**, 833-846, doi: https://doi.org/10.1016/j.undsp.2021.12.009.
Low, P.F. 1987. Structural component of the swelling pressure of clays. Langmuir, **3**, 18-25.
Marcial, D. 2013. Measuring water retention properties of a series of bentonite clays a wide range of suctions. Advances in unsaturated soils. Taylor Francis Group, London, 135-140.
Mishra, A.K., Ohtsubo, M., Li, L.Y. & Higashi, T. 2012. Influence of various factors on the difference in the liquid limit values determined by Casagrande's and fall cone method. Environmental Earth Sciences, **65**, 21-27.
Mitchell, J.K. & Soga, K. 2005. Fundamentals of soil behavior. John Wiley & Sons New York.
Moore, D.S. & McCabe, G.P. 1989. Introduction to the Practice of Statistics. WH Freeman/Times Books/Henry Holt & Co.
Muhunthan, B. 1991. Liquid limit and surface area of clays. Geotechnique, **41**, 135-138.
Nelson, J.D., Chao, K.C., Overton, D.D. & Nelson, E.J. 2015. Foundation engineering for expansive soils. John Wiley & Sons.
Peck, R.B., Hanson, W.E. & Thornburn, T.H. 1974. Foundation engineering. John Wiley & Sons.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, **12**, 2825-2830.
Phoon, K.-K. & Zhang, W. 2023. Future of machine learning in geotechnics. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, **17**, 7-22.
Polidori, E. 2009. Reappraisal of the activity of clays. Activity chart. Soils and Foundations, **49**, 431-441.
Salvatore, E., Modoni, G., Spagnoli, G., Arciero, M., Mascolo, M.C. & Ochmański, M. 2022. Conditioning clayey soils with a dispersant agent for Deep Soil Mixing application: laboratory experiments and artificial neural network interpretation. Acta Geotechnica, **17**, 5073-5087.
Santhoshkumar, T., Abraham, B., Sridharan, A. & Jose, B. 2016. Role of bentonite in improving the efficiency of cement grouting in coarse sand. Geotechnical Engineering Journal of the SEAGS & AGSSEA, **47**, 1-8.
Schapire, R.E. 1990. The strength of weak learnability. Machine learning, **5**, 197-227.
Schwing, M., Chen, Z., Scheuermann, A. & Wagner, N. 2013. Dielectric properties of a clay soil determined in the frequency range from 1 MHz to 40 GHz. *Proc 10th Int. Conf. Electromagn. Wave Interact. Water and Moist Substances*, 242-250.
Shahriari, B., Swersky, K., Wang, Z., Adams, R.P. & De Freitas, N. 2015. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, **104**, 148-175.
Sivapullaiah, P., Prasad, B.G. & Allam, M. 2008. Methylene blue surface area method to correlate with specific soil properties. Geotechnical testing journal, **31**, 503-512.
Skempton, A. 1953. The colloidal activity of clays. Selected papers on soil mechanics, **1**, 57-61.
Snoek, J., Larochelle, H. & Adams, R.P. 2012. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, **25**.
Spagnoli, G., Feinendegen, M. & Rubinos, D. 2013. Modification of clay adhesion to improve tunnelling excavation. Proceedings of the Institution of Civil Engineers-Ground Improvement, **166**, 21-31.
Spagnoli, G. & Shimobe, S. 2019. A statistical reappraisal of the relationship between liquid limit and specific surface area, cation exchange capacity and activity of clays. Journal of Rock Mechanics and Geotechnical Engineering, **11**, 874-881.
Stephens, M.A. 1974. EDF statistics for goodness of fit and some comparisons. Journal of the American statistical Association, **69**, 730-737.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological), **36**, 111-133.
Wackerly, D., Mendenhall, W. & Scheaffer, R.L. 2014. Mathematical statistics with applications. Cengage Learning.
Wang, L., Wu, C., Tang, L., Zhang, W., Lacasse, S., Liu, H. & Gao, L. 2020. Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method. Acta Geotechnica, **15**, 3135-3150.
Warkentin, B. 1972. Use of the liquid limit in characterizing clay soils. Canadian Journal of Soil Science, **52**, 457-464.
Yao, H.-l., Yang, Y. & Cheng, P. 2004. Standard moisture absorption water content of soil and its testing standard. ROCK AND SOIL MECHANICS-WUHAN-, **25**, 856-859.
Zapata, C., Houston, S., Houston, W. & Dye, H. 2006. Expansion index and its relationship with other index properties. *Unsaturated Soils 2006*, 2133-2137.
Zhang, W., Gu, X., Tang, L., Yin, Y., Liu, D. & Zhang, Y. 2022a. Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. Gondwana Research, **109**, 1-17.
Zhang, W., He, Y., Wang, L., Liu, S. & Meng, X. 2023. Landslide Susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing. Geological Journal.
Zhang, W., Wu, C., Zhong, H., Li, Y. & Wang, L. 2021. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geoscience Frontiers, **12**, 469-477, doi: https://doi.org/10.1016/j.gsf.2020.03.007.
Zhang, W., Zhang, R., Wu, C., Goh, A.T. & Wang, L. 2022b. Assessment of basal heave stability for braced excavations in anisotropic clay using extreme gradient boosting and random forest regression. Underground Space, **7**, 233-241.
Zhang, Y., Qu, Y., Liu, G. & Wu, S. 2003. Engineering geological properties of Miocene hard clays along the middle line of the North–South Diversion Water Project in China. Bulletin of Engineering Geology and the Environment, **62**, 213-219.